UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS
REPORT 107

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK
BERICHT 107

ANALYSIS AND EVALUATION OF CELL IMPUTATION

ISMO HORPPU

JYVÄSKYLÄ
2008

Editor: Pekka Koskela
Department of Mathematics and Statistics
P.O. Box 35 (MaD)
FI-40014 University of Jyväskylä
Finland

ISBN 978-951-39-2859-9
ISSN 1457-8905

University Printing House
Jyväskylä 2008

Abstract

The overall objective of this thesis is to analyse and evaluate predictions of missing data values using cell imputation, in which one divides data into disjoint partitions (cells) and fills in the missing data values by local estimates. In this thesis cells are formed with the help of clustering algorithms. Imputations using classical methods such as linear regression and nonparametric regression are included for comparison.

The analysis and evaluation of imputation is done in a multiobjective way, where distribution level and unit level are considered. In the literature people typically considered either distribution level or unit level, which is not sufficient in order to distinguish between the performance of various imputation methods. In addition, computational complexities of various methods are numerically analysed here. Computational complexity is important when methods are used with huge data sets, such as censuses.

Theoretical analysis consists of deriving results for distribution level and unit level error quantities. At distribution level the first two moments are mainly considered. The bias and the variance of the first moment estimator and the bias of the second moment are derived. An analysis on the behaviour of Kolmogorov-Smirnov statistic is also developed. At unit level, mean squared error quantities are computed. Both finite sample and asymptotic (limiting) results are derived.

Simulation studies and empirical studies with real-world data sets show that the proposed cell methods perform well. Two of the proposed methods are able to preserve multimodal conditional distribution, whereas the competing methods fail in this. The best competing methods are the 1-nearest neighbour method and the linear regression methods. However, the computational complexity of the nonparametric regression methods is inferior compared to the cell method for large sample sizes. Further, the linear regression method should be used only if the linearity assumption holds well enough.

A link between the theory and numerical studies is included. An example given here shows that an approximation of variance for the mean estimator of imputations done using the cell method works in practice.

Author

Ismo Horppu
Department of Mathematics and Statistics
University of Jyväskylä
Finland

Supervisor

Pasi Koikkalainen, Dr.Tech.
Department of Mathematics and Statistics
University of Jyväskylä
Finland

Reviewers

Professor Lasse Holmström
Department of Mathematical Sciences
University of Oulu
Finland

Professor Ray Chambers
Centre for Statistical and Survey Methodology
School of Mathematics and Applied Statistics
University of Wollongong
Australia

Opponent

Nicholas T. Longford, Ph.D.
SNTL, Reading, England,
and Departament d'Economia i Empresa
Universitat Pompeu Fabra
Barcelona, Spain

Acknowledgements

I wish to express my sincere gratitude to my supervisor, Pasi Koikkalainen, Dr.Tech., who has given me more insight and expanded my knowledge on both statistics and computer science. I have greatly benefited from his expertise and constant support during this work.

I am thankful to the reviewers, Professor Lasse Holmström from the University of Oulu and Professor Ray Chambers from the University of Wollongong for their comments and criticism. I also wish to mention Steve Legrand, Ph.D., who was responsible for the proofreading.

I also want to express my thanks to all my co-workers in the Laboratory of Data Analysis and in the Department of Mathematics and Statistics of the University of Jyväskylä. Especially I like to mention my colleagues Anssi Lensu, Ph.D., Jouni Raitamäki, Ph.D., Markku Mielityinen, Ph.Lic., and Michael Haranen, M.Sc. I have learned a lot about statistics from Professors Antti Penttinen, Hannu Oja, and Stefan Geiss. In addition, the seminars which were organized during my studies have been useful. I wish also to thank Janne Kujala, Ph.D., from the Department of Mathematical Information Technology of the University of Jyväskylä, Professor Seppo Laaksonen from the University of Helsinki/Statistics Finland, and Pasi Piela, Ph.Lic., from Statistics Finland.

My special thanks to the following people: Y. L. Lai for copying and sending me Chapter 4 of his thesis [59], Theo Gasser and Sinikka Kohler for publication [28], Robert Santos for publication [90], and Klaus Hess for report [37].

This research took place at the Department of Mathematical Information Technology, Laboratory of Data Analysis, University of Jyväskylä, and at the Department of Mathematics and Statistics, Laboratory of Data Analysis, University of Jyväskylä. Financial support was provided by EUROSTAT (during the EurEdit project), Computational Methods of Information Technology (ComMIT) Graduate School, and Laboratory of Data Analysis, University of Jyväskylä. I appreciate all of these sources for financing this work.

I wish to express my deepest gratitude to Heini for having been so patient and supportive. All my friends deserve thanks for their encouragement and support. My brother, Timo, and my parents, Sirpa and Hannu, from whom I have received constant support, deserve special acknowledgement. Also, my brother has deepened my insight on music and life in various ways.

Jyväskylä, January 2008

Ismo Horppu

Contents

Chapter 1

Introduction

This chapter is organized as follows. We start with a discussion about the background of the research, which leads to the description of the research problem. A motivation for our objectives is given next, including some related observations made during the research. Then the main results of the study are summarized. Finally, the structure of the thesis and the contributions of the author are described.

It should be noted that the author and the supporting research team come from the area of computer science, rather than that of statistical sciences. Therefore we do not have any orthodox views about any best practices in statistical analyses. Instead, this work should be though as a data engineering approach.

1.1 Background

Many real-world data sets are incomplete. There are various reasons for this state of affairs, known as missingness in statistics. Some of the reasons are: nonresponse in surveys and censuses, malfunctioning measurement devices, and errors in data transfers. A typical consequence of missing data is that analyses that are more complicated than those complete data sets are needed for incomplete data sets. This means that more time is consumed in the analysis, and experienced and educated people are required for it. This easily leads to higher costs. But in practice only limited resources are available to handle large and complex data sets. Therefore one must have efficient methods to achieve results under practical constraints.

The background of the current work is based on the development of new methods for incomplete data. This was started in the EurEdit[1] project, which began in 2000 and was finished in 2003. The participants[2] were national statistical agencies, universities, and enterprises. The project was funded by the Statistical Office of the European Communities (Eurostat). The aim was, whenever possible, to detect errors in data, correct errors using editing rules, and replace missing data values by predictions using imputation methodology. The work was done with data sets

[1] See website www.cs.york.ac.uk/euredit/ for details

[2] see the appendix for this chapter

that were collected by national statistical agencies. Such data sets include surveys and censuses. The research objectives were to develop new methods for editing and imputation, and to compare new and classical methods.

In EurEdit we were the only partner that was using clustering methods for imputation. Our methods were based on the tree-structured self-organizing map (TS-SOM) [55, 56], which is better known as a neural network algorithm. Our approach with TS-SOM was to divide data into clusters and perform imputation more or less separately in each of the data subsets. This is essentially the idea that was proposed by Santos [90], who refers to these methods as cell imputation methods.

During the EurEdit project we developed several cell imputation methods, where the missing values inside a cell were selected either from

a) cluster mean,

b) cluster mean with added noise,

c) donor from cluster, or

d) nearest neighbour within a cluster.

Due to limited resources we were not able to analyse the developed methods, nor were we able to conduct extensive experiments to find the best practices for the methods. Thus, we should consider the EurEdit results as preliminary studies of new ideas.

From the results of the EurEdit project it was noticed that the TS-SOM methodology was relatively good with some data sets, such as the Danish Labour Force Survey (DLFS), while the methodology was not particularly good for data sets such as UK Annual Business Inquiry (UK ABI/survey) and UK Sample of Anonymized Records (UK SARS/census). Therefore a re-evaluation is necessary to find out why the methodology might not work on a particular case.

In this thesis we shall examine the imputation methods under more complicated missing-data mechanisms and the analysis is focused on the properties of the methods. In addition to experiments we want to have an analytical insight. So far, there has not been any deeper analytical study of the developed new methods. Thus there is gap in theory, which needs to be studied. With a theory we expect to understand the features, including any pitfalls, of the methods that are important for real-world applications.

1.2 Lessons from EurEdit

The idea of EurEdit was to evaluate methods using real-world data sets. Several data sets were employed with many incomplete variables. To evaluate the performance of different methods, many criteria were used, including measures for distributional accuracy such as Kolmogorov-Smirnov statistic (KS), and unit level errors such as squared unit level error (DL2), and absolute unit level error (DL1). See reference [10] for details on the statistics.

Two kinds of data sets were used in the project: development data sets and evaluation data sets. In the development phase incomplete data sets and corresponding true data sets were available. However, in the evaluation phase the partners had to complete incomplete evaluation data sets without knowing corresponding true data sets. The completed data sets were sent to Office for National Statistics (ONS) for evaluation.

In EurEdit each partner conducted a couple of imputation experiments with their methods, which then were evaluated by ONS. While this is a rather objective way of doing evaluations, it is also problematic in many ways. It was not possible to eliminate the role of good or bad luck from the results. Especially, the variation in the imputation results was unknown. Furthermore, each partner had to decide by themselves how to optimize their methods. While some looked for minimal errors in distribution, others were optimizing unit level imputations. As a consequence it is difficult to say which of the methods is better than another. This is discussed in the evaluation report [116].

The lessons from EurEdit, on which the current thesis tries to improve, can be summarized as follows:

- Computed evaluation statistics were based on a single imputation without repetitions and the results included no information about statistical confidence. Because the reliability of the evaluation results was not assessed in the EurEdit, we cannot be fully certain about whether the results of a method in a certain setting were good or bad just by an accident. To be of practical usability it is important to know how a method performs on average, not just on a single experiment.
- There was only a limited possibility to vary imputation model parameters. This leads to a conclusion that some methods were artificially good in some evaluations and artificially bad in others.
- The missing-data mechanism used was mostly MCAR(missing complete at random, see section 2.1 for details), which is unrealistic. As a consequence many real differences and thus, practical potentials, of various methods were not clearly visible in the results of the project.
- There was no theoretical insight into the methods.

Now one may wonder whether some of the data sets used in the EurEdit project have been used in this thesis also. The answer is no, for two reasons: 1) The EurEdit project has ended, and the data sets used in it are no longer available, and 2) it would be unfair to finetune just the TS-SOM methodology for the data sets used in the project.

Overall, we expect that we shall obtain an understanding of how the methods perform in different situations. This then would allow us to apply the proposed methods in a most appropriate way for real-world applications.

1.3 Research problem

The primary objective of this thesis is to analyse cell imputation methods. This is done both analytically and empirically, and the focus is to find out what the differences between the methods are. The advantages and disadvantages of the methods are studied from a practical viewpoint in order to find the conditions under which one method is better than another. In this context the first question is:

> **Why should we analyse imputation methods?**

To be more specific, we hope that the analysis of imputation methodology reveals us how the missingness and imputation methodology affects our statistical estimates, predictions and conclusions. We hope to obtain an understanding about the types of errors which are made when using different ways to handle an incomplete set of observations.

It should be noted that there is no single criteria for judging when one imputation method is better than another. It depends on one's viewpoint. For example, one may try either to preserve the unit level properties or distributional properties of a data set. On a unit level one tries to, for instance, minimize the (squared) prediction error between single observations, while on a distributional level one is interested about distributional measures such as the bias of the first and the second moment. This discussion leads to the following question:

> **What are the causes of imputation errors, and how these are seen on a unit level and on a distribution level?**

It is well-known that if one manages to minimize the squared prediction error at the unit level then the variance estimator becomes biased downwards. In this thesis we go further. The degree of bias is related to the expectation of conditional variance of the target and the proportion of missing data values. The minimization of unit level errors decreases the variance of mean estimators, and thus increases their precision. On the other hand, if distribution level errors are minimized, then the squared prediction error increases. This increase is related to the expectation of the variance of the target and the proportion of missing data values. These are just the first glimpses on the results that are given in this thesis.

Usually there is a big gap between theoretical results and practice. Often this is due to simplifications that are required for the theoretical study. To decrease the gap we shall ask:

> **What are the causes of imputation variances for a finite random sample on n observations?**

This is a difficult question. In theory we can simplify things a lot when n goes to infinity. We may also study things when we are given a fixed data of n observations. But as we shall see, both of these simplifications lead to unrealistic conclusions when compared to errors in the real world. Thus the main challenge of the current thesis is to analyse imputation when assuming only that n observations are sampled from a superpopulation. We then compare the imputation results with superpopulation parameters.

1.4 Observations

During research we had many unsuccessful attempts to meet our objectives. This experience can be discussed in terms of observations which we think are of general interest. The observations are about

- analytical computations for incomplete random sample of n observations,
- conditionalized viewpoints to analytical computations,
- the fairness of the comparison of the methods, and
- computational complexity of the methods.

These topics are discussed with more detail in the following.

1.4.1 Exact analysis and approximations

In our first attempt, we tried to obtain the properties of the methods without using approximations. However, we observed that the exact analytical computations for incomplete samples are time demanding, complicated, and sometimes very difficult or even impossible to do. Therefore we decided to use approximations, which are mainly based on the Taylor series expansions and the laws of large numbers. The benefits of approximate results are that they are faster to derive and often easier to interpret, because approximations are shorter than exact derivations. However, approximate results hold well only for very large sample sizes, and it easily happens that the magnitude of an approximation error cannot be solved. And worse, there are problems in the interpretation of results as well. Although the approximative equation of the studied property of a method is shorter than the exact one, it often happens that the causes for the behaviour of the method are hidden. We cannot always say, for example, what increases the bias of an imputation, despite the fact that we have an approximative equation for it.

1.4.2 Conditional properties

The reason for the difficulty of interpretation was that we tried to analyse the whole imputation procedure under a single approximate equation. In other words we made analytical formulas over a joint distribution of all possible random factors, such as the number of observations, uncertainty of model parameters etc. With this approach different random terms got mixed in the way that it was impossible to say how the user of the method could control the method, and it was virtually impossible to compare advantages with disadvantages between any two methods.

The solution for this was based on an observation that it is more clever to compute formulas under three conditionalization levels. It means that we study the role of imputation under a given model and a given data set, uncertainty of imputations with given model and a given training data set, and properties of random sample from the superpopulation.

1.4.3 When is it fair to draw conclusions

The third observation is that one should be very careful in order to obtain a fair comparison between the methods. In other words, one can compare the imputation performance of two methods only under the same assumptions. This is a major challenge for an overall evaluation, which includes both the estimation of the model and its use in the imputation task. While the overall performance can be analysed for standard methods such as linear regression, this is not easy for nonparametric methods such as neural networks. A major problem is the loss of identifiability of new methods which prevents us from defining exactly what kind of variability is caused by the estimation phase of the model. Only empirical estimates can be given. Therefore some comparisons are fair only when the analysis is conditionalized for a given (fixed) model, and the role of the model is given via external (empirical) evaluation.

1.4.4 Practical issues with computationality

The last observation is about the computational complexity of imputation. There are methods with desirable statistical properties that cannot be used in practice because of computational problems. Methods whose requirements for computer memory or time to complete the task grow clearly more than linearly with respect to the number of observations are often good for academic exercises only. With real-world data sets, such as census data, we need good computational performance. Therefore we include computational time requirement as an evaluation criteria in this work too.

1.5 Main results and claims

There are several types of results. One is the overall study of cell imputation, which is supported by both theoretical and empirical investigations. Another is a long list of observations about situations where the behaviour of cell imputation can be explained in detail. To summarize the results we note especially the following:

i) An equation and an algorithm to compute analytical error estimates of cell imputation. For example, for a given clustering, we may estimate the variance of the completed mean of data for a given sample size n. This is given in Chapter 6 (Algorithm 6.4.2). However, due to time limitations variance estimate is not used in empirical studies in Chapters 7, 8 and 9. The reason for this is that the final details of the error estimates were completed after the empirical studies.

ii) The relative roles of the causes of imputation errors are demonstrated in various examples. Especially interesting is the difference in errors caused by the fixed sample and random sample of n observations. See, for example, Sections 4.2.2 and 6.4.3.

iii) Sometimes a noisy model (overtrained model) is good in imputation. There are several theoretical results and empirical examples demonstrating that nonparametric methods can achieve the same performance in terms of marginal moments using either stiff models with added noise or flexible models without noise.

iv) Multimodal missingness is not a problem for nonparametric methods. Both nearest neighbour and cell imputation can impute incomplete multimodal data in terms of marginals, but cell imputation is better when data is conditionalized as seen in Section 7.2.

v) The advantages of cell imputation are most visible when missingness mechanisms are complex. In our examples we see MAR and NMAR type cases, where cell imputation outperforms other methods.

vi) In Chapter 9 new results about the role of model parametrizations are given. For example, by changing the amount of added noise in imputations, the performance can be shifted from unit level to distribution level.

vii) In the empirical part (Chapters 8 and 9) repeated sampling is used to obtain reliable estimates of real-world imputation performances. The methods are ranked according to their Pareto optimality.

viii) Computational differences of the methods are studied in detail.

The results can be concluded by a claim

Cell imputation is a good and practical way to do imputation.

Some of the advantages of the proposed methods are that they are nonparametric and contain a tunable bias vs variance tradeoff. Using simple methods inside the cells, we can show good imputation performance with respect to our evaluation criteria under both the MCAR and MAR type of missingness. Most notably the proposed methodology provides good enough performance simultaneously, on both distributional and on unit level measures, while standard methods tend to do well only on one type of criteria. From the practical viewpoint we can show that cell imputation is easily implemented using Tree-Structured Self-Organizing Maps [55, 56] which is a computationally efficient version of Kohonen's SOM algorithm [54].

The nonparametric nature of the proposed methodology implies flexible modelling. Therefore the proposed method is capable of modelling data sets where the parametric form is not known beforehand. As a consequence of flexibility we need to solve a so-called bias-variance dilemma which is a tradeoff between squared prediction bias and prediction variance. A very flexible model has a low prediction bias and high variance of prediction, whereas a stiff model has a high prediction bias and low prediction variance. In our case the bias and variance tradeoff can be controlled quite easily by the user.

The disadvantage of the proposed new methodology is that the parameters of the model used in the methods are not identifiable. Non-identifiability of parameters is a problem, because derivation of statistical properties, such as the variance of parameter estimates, is difficult or even impossible. The non-identifiability problem is typically encountered in mixture modelling and with other flexible modelling methods such as neural networks.

There exists evidence for practical usability of the new methods. The experiments done in the Euredit project give some indication that the methods may be usable in practice. Secondly, the new empirical studies conducted for this thesis indicate the same. This indicates that the proposed methodology is practically safe and easy to use for complex data sets whose properties are not fully understood.

1.6 Structure of the thesis

This thesis has ten chapters as shown in Figure 1.1. In addition to Introduction and Conclusion there are three methodological chapters (4, 5, and 6) and three empirical chapters (7, 8, and 9). The review of methodology is given in Chapters 2 and the setup of the problem is described in Chapter 3.

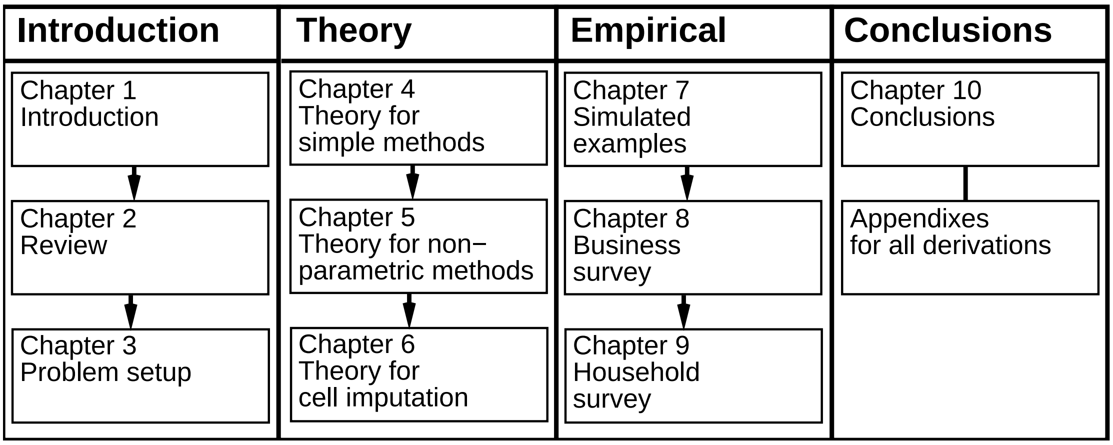

Introduction
Theory
Empirical
Conclusions
Chapter 1 Introduction
Chapter 2 Review
Chapter 3 Problem setup
Chapter 4 Theory for simple methods
Chapter 5 Theory for non–parametric methods
Chapter 6 Theory for cell imputation
Chapter 7 Simulated examples
Chapter 8 Business survey
Chapter 9 Household survey
Chapter 10 Conclusions
Appendixes for all derivations

Figure 1.1: Organization of Chapters.

The main results of this thesis are given in Chapters 4, 6, 7, 8, and 9. These describe the setup of the problem, the main theoretical work about cell imputation, and the actual evaluation. Chapters 4 and 5 complete the theoretical work by giving formulas for methods that are used in methodological comparisons, but there is no clear interpretation of the results.

There are many approximations in this thesis. To improve the clarity of presentation, the derivations of the approximations (justifications) are moved to appendixes. Only some illustrative examples are left in the main text. This has an additional benefit since it gives more room for derivations, can then be given very explicitely, using small and easily verifiable steps. As a consequence of this the appendix forms a set of background data, containing about 100 pages of formulas for anyone who likes to check them.

Below, a more detailed list of chapter contents:

Chapter 2 begins with a description of missing-data mechanisms. Then common estimation methods for complete and incomplete data are described. Imputation and its properties and drawbacks are summarized. The difference between single and multiple imputation is briefly discussed about. Finally, imputation model classes and related work done by other people are summarized.

Chapter 3 defines the framework under which the research problem is studied. The notation which is used in this thesis is described in the beginning. This is followed by an introduction to a simple practical scenario. Then a theoretical setting is concretized by giving details of the scenario. The measures, on which the actual evaluation statistics is based in this thesis, are defined. Conditionalizations, approximations, and decompositions, which are used to ease the interpretation of results, are introduced.

Chapter 4 contains the results for simple methods and linear regression. Two numerical studies are included. In the first study we investigate preservation of first moment. The second study is about analysis of variance sources such as imputation noise vs. sampling errors.

Chapter 5 contains an analysis of nonparametric regression. Nearest neighbour imputation and kernel regression are studied. The chapter contains a study about mean squared errors and flexible models.

Chapter 6 is about cell imputation. K-Means, Self-Organizing Map, and TS-SOM clustering algorithms are briefly introduced, likewise their application in imputations is described. In addition to analytical results two numerical studies are included. In the second study the impact of a clustering algorithm on imputations is investigated.

Chapter 7 evaluates imputation methods using carefully designed experiments, in which we focus on some specific phenomena. The chapter contains four studies: a study about the role of a missing-data mechanism, another one about imputation of multimodal distribution, a study about a classification task with multivariate covariate, and a study about computational properties.

Chapter 8 is a step towards a more practical situation (as compared to Chapter 7): a simplified case study is presented. We have formed a finite population by cleaning a real-world sample from missing-data values and special values. Imputation of continuous turnover of enterprises is evaluated. The study utilizes a small and medium-sized enterprises data set from the UK.

Chapter 9 contains experiments under the most practical situation (missingness in covariates and presence of special values). The data set used in the experiments is Quarterly Labour Force Survey. Imputations of almost continuous variable AGE and categorical variable SEX (of person) are evaluated.

Chapter 10 contains the conclusions of the thesis. In addition, some ideas for future research, which may be useful for other researchers, are summarized.

1.7 Contributions of the author

This work is a result of co-operation between the author and the supervisor. The problem and the imputation methodology as well as many of the given examples were proposed by the supervisor. The role of the author was to do the theoretical work and conduct the empirical experiments. Thus

- All the derivations (proofs) were done by the author.
- All the experiments that are presented in this thesis were conducted by the author.

The theoretical setup of computed properties and examples was done jointly. Some general ideas came from the supervisor, but the actual details were developed by the author.

The empirical evaluations were done almost solely using the NDA (Neural Data Analysis) software, which has been developed in our laboratory (see Häkkinen [42] for more details). The original versions of the TS-SOM methods for imputation were developed by the supervisor, but these were partially rewritten by the author. In addition the author has made all the simulation macros used in this thesis.

All the conclusions that are presented here were drawn up jointly by the author and the supervisor. The supervisor has been quite helpful in the preparation of the final text.

Chapter 2

Review of estimation from incomplete data and imputation of missing values

This chapter presents a short introduction to the general problem of handling missing data. The actual problems that are studied in this thesis are introduced in the next chapter.

The chapter begins with a specification of missing data mechanisms. Then basic estimation methodology is reviewed for both complete and incomplete data. This is followed by a short historical review about imputation. Finally an overview of the imputation models and strategies used in this thesis is given.

2.1 Missing-data mechanisms

This section is about missing-data mechanisms. Rubin and Little [62] have defined three classes of them: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). The mechanisms define the dependency between missingness indicators and data.

To formalize the idea let $\mathbf{d}_{\mathsf{n}\times p}$ be a partially unknown true data matrix which has p variables and n observations. The true data $\mathbf{d}$ is assumed to be a random iid sample from the joint distribution of variables. Missing data can be specified via a realization of missingness indicators:

$$M_{j,i} = \begin{cases} 1, & \text{if } d_{ji} \text{ is missing,} \\ 0, & \text{otherwise.} \end{cases}$$

In the MCAR mechanism $\mathbf{M}$ is independent of both the observed and the unobserved part of sample, thus

$$\Pr(\mathbf{M}|\mathbf{d}) = \Pr(\mathbf{M}|\mathbf{d}^{obs}, \mathbf{d}^{mis}) = \Pr(\mathbf{M}). \tag{2.1}$$

where $\mathbf{d}^{obs}$ means observed values and $\mathbf{d}^{mis}$ missing values.

When the MCAR mechanism applies standard methodology based on iid random samples can be used to build an estimator for missing data. This makes analytical studies easiest to do. Unfortunately the MCAR mechanism typically is considered to be unrealistic for real-world data sets.

In the MAR mechanism the probability of missingness depends on the observed part of the sample, but not on the unobserved part. Formally we may write it as

$$\Pr(\mathbf{M}|\mathbf{d}) = \Pr(\mathbf{M}|\mathbf{d}^{obs}, \mathbf{d}^{mis}) = \Pr(\mathbf{M}|\mathbf{d}^{obs}). \tag{2.2}$$

This implies that we should use the observed data to predict missing values. It is also very likely that the distribution of missing data is different from that of observed data.

One should note that in the past a different name for missing-data mechanism was used. For example, Santos writes [90]

> ”Data are said to be missing at random (MAR) for a specified variate Y if the respondent stratum is a simple random sample of size R from the total population.”

However, in the current literature this clearly corresponds to the MCAR mechanism as defined by Little and Rubin. The point of the remark is to emphasize that one should not confuse missing-data mechanisms in older and newer publications.

The NMAR mechanism is potentially the most complicated missing-data mechanism as the probability of $\mathbf{M}$ may depend on both the observed and the unobserved part of a sample. In some cases, there is an exact logical rule that can be used to find missing values. If such a rule cannot be found, and the missingness seems totally arbitrary, it might be impossible to impute the missing values. In the case of statistical modelling, when external knowledge is available about the dependencies between the missing values as well as about the dependencies between missing and observed data, one might use the Markov or Gibbs processes for imputation. This is, however, outside of the scope of the current thesis.

In this thesis we do not need to identify missing-data mechanisms from data, because we define them in the setup of our experiments. For the interest of the reader, we briefly summarize the possibilities for missingness identification next. It is generally impossible to detect an exact missing-data mechanism for an incomplete data set, but there exist some identification methods, mainly tests, that are able to distinguish between MCAR and non-MCAR mechanisms.

Likelihood based tests have been proposed by Fuchs (1982)[27] for contigency tables, and by Little (1988)[61] for multivariate normal data. A nonparametric test has been proposed by Diggle (1989)[19] for preliminary screening. Rideout and Diggle (1991)[83] have proposed a parametric test which requires the modelling of the missing-data mechanism. Chen and Little (1999)[13] have generalized Little's (1988)[61] basic idea of constructing test statistics. They avoid distributional assumptions, whereas Little (1988)[61] assumed normal data. Some specific tests for linear regression models have been proposed too. Simon and Simonoff (1986)[95] have written an article in which they describe tools for MAR diagnostic and for other purposes. They make no assumptions about the nature of the missing value process. Simonoff has introduced (1998)[96] a test to detect non-MCAR mechanisms. His diagnostics are based on standard outlier and leverage-point regression diagnostics. Recently Toutenburg and Fieger (2001)[105] introduced methods to analyse and detect non-MCAR processes for missing covariates. They use an outlier detection to identify non-MCAR cases.

2.1.1 Ignorability and other assumptions

Assuming that distribution of $\mathbf{D}$ can be parametrized with $\boldsymbol{\theta}^*$, distribution of response indicators $\mathbf{M}$ can be written in the form of $f(\mathbf{M}|\mathbf{d}, \boldsymbol{\psi}^*, \boldsymbol{\theta}^*)$, where $\boldsymbol{\psi}^*$ is some set of parameters for the missingness. Without further assumptions the missing data problem is not solvable, at least when considering maximum likelihood or Bayesian inferences (see Sections 2.3.2 and 2.3.3). One should note that one can always do imputations, but the quality of imputations is dependent on the type of missingness. Analysis of impact of imputation is likely to be difficult in a general missingness case.

The most usual simplification is to assume that the missingness mechanism is ignorable, i.e.:

a) missing-data mechanism is missing at random (MAR) and

b) parameters $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$ are distinct such that the joint parameter space $\Omega_{\boldsymbol{\theta}^*,\boldsymbol{\psi}^*}$ equals to $\Omega_{\boldsymbol{\theta}^*} \times \Omega_{\boldsymbol{\psi}^*}$.

However, as stated by Little and Rubin [62]

> ”MAR is typically regarded as the more important condition here, in the sense that if the data are MAR but distinctness does not hold, inference based on the ignorable likelihood is still valid from the frequency perspective, but not fully efficient.”

Further, there exists some other factorizations of likelihood for some models and incomplete data patterns which result in simpler suboptimization tasks [62].

We do not assume ignorability throughout this thesis. Instead we shall assume that the fully observed part of data $\mathbf{D}^{train}$ is iid drawn from an observed distribution $f_{Y^{obs},\boldsymbol{X}^{obs}}$ and the incomplete part of data $\mathbf{D}^{test}$ is iid drawn from an incomplete data distribution $f_{Y^{mis},\boldsymbol{X}^{mis}}$. These notations are explained in Chapter 3, and the required assumptions are described within the appropriate context.

2.1.2 Missingness and statistical analyses

In statistical analysis there are basically three possible ways to handle missing information:

i) omit (discard) the incomplete part of data

ii) replace the missing values with the imputed ones

iii) use a specific estimation methodology for the incomplete data

In the first alternative all incomplete cases (data records) are discarded. One version of this is called listwise deletion (LD). This is viable for simple problems, where missigness does not dominate the outcome due to the number amount of missing values or/and due to the differences between the observed and missing populations. Especially in multivariate cases, where any number of variables may hold missing values, it easily happens that the number of complete records may not be enough for a reliable estimation. King et al. estimate that the LD approach is used in approximately 94% of publications in political science field when any one variable remains missing after filling in guesses for some [51]. King et al. consider publications recent to year 2001. There are issues with LD. If the missing-data mechanism is not MCAR then LD is likely to yield biased estimators, possibly highly so. Myrtveit et al. write [69]

> ”Specifically, we need to understand the limitations of the seemingly innocent and widely used, and abused, LD.”

In the quote ”we” refers to researchers in empirical software engineering.

Sometimes a slight improvement to LD is available, allowing a more efficient use of information, for instance, if one estimates pairwise statistics, such as covariances. In pairwise deletion (PD) one uses all the available pairs of observations. The good news is that pairwise deletion does not introduce bias in MCAR cases, if the statistics is originally unbiased. The bad news is that one has to be careful when using pairwise computed statistics with univariate statistics. For example, in computation of correlation one needs covariance and variance. A problem may arise if the sample sizes used for covariance and variance are different. In such case the estimated correlation coefficients can be out of the valid range. Further, one may encounter problems even if one uses the same sample sample size for both covariance and variance. Namely, the computed correlation matrix may not be positive definite. See [62] for details on these issues.

The second alternative, imputation, which also is the main topic of this thesis, is most applicable in cases where a complete data set is required for unspecified future analyses. That situation is typical, e.g., in statistical offices which offer data services for a large community of users and disciplines. As it is impossible to find optimal incomplete data estimators for all purposes using limited resources, the only sensible solution is to complete the missing values by reasonably good predictions. An introduction to imputation methodology follows immediately after the reviews of estimation from complete data set and the third alternative, which follow next.

Theoretically the best alternative for statistical analyses from incomplete data is to develop specialized estimators for the purpose. The drawback is the complexity of the task. Although several methods exist, there are many common problems for which there are no ready-to-use tools available. Then one must either develop one, or return to alternatives i) or ii) as introduced above.

2.2 Estimation from complete data set

In estimation one uses an estimator function to obtain values for model parameters from observed data. Thus the estimator is a function of random sample data, and the output of an estimator is an estimate for a parameter. For concreteness, we describe the least squares (LS) estimation, the maximum likelihood (ML) estimation, and the Bayesian approach.

Least squares estimation

In the least squares (LS) estimation one minimizes the sum-of-squared errors. As an example we define LS estimation for $Y = g^*(\mathbf{x}) + \epsilon$, where ϵ is zero mean noise. The model is a general regression function $g(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a set of estimates of parameters.

Let our data be a set of observed pairs $\{y_j, \mathbf{x}_j\}_{j=1}^{\mathsf{n}}$ of univariate response Y with covariates $X_1, \ldots, X_{p-1}$. In LS-regression one tries to find the functional relationship between Y and $\boldsymbol{X}$ using minimization of sum of square error between observed y_j and modelled response $g(\mathbf{x}_j|\boldsymbol{\theta})$.

$$\mathrm{SSE}(\boldsymbol{\theta}) = \sum_{j=1}^{\mathsf{n}} \Big(y_j - g\big(\mathbf{x}_j|\boldsymbol{\theta}\big)\Big)^2,$$

Thus the estimate is $\boldsymbol{\theta}^{LS} = \mathrm{argmin}_{\boldsymbol{\theta}}\, \mathrm{SSE}(\boldsymbol{\theta})$. The regression function g needs to be defined in order to derive an optimum solution. For many flexible models a closed form solution does not exist, and iterative optimization methods need to be applied.

As a historical insight we note that perhaps the earliest form of the least squares estimation was published by Legendre in 1805. This was followed by a publication by Gauss in 1809. They both considered least squares estimation with a linear regression model. However, according to Gauss he had been using LS estimation before Legendre. More details can be found from [8], for example.

Maximum likelihood estimation

The concept of maximum likelihood (ML) was introduced by Fisher between 1912 and 1922 (for details see [1]). In maximum likelihood estimation one assumes a parametrized distribution for data $\mathbf{D}$. Then, one defines a likelihood function $L(\boldsymbol{\theta}|\mathbf{d})$ for parameters with a given data set. The idea is to maximize likelihood function $L(\boldsymbol{\theta}|\mathbf{d})$ such that the estimate $\boldsymbol{\theta}$ for parameters $\boldsymbol{\theta}^*$ is

$$\boldsymbol{\theta}^{ML} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\, L(\boldsymbol{\theta}|\mathbf{d}).$$

In practice it often turns out to be more easier to maximize log-likelihood, especially when deriving analytical results and in case of exponential distributions.

As an example, ML estimation of regression function parameters is considered next. Let data $\mathbf{d}$ contain n observations of response Y with a given covariate $\boldsymbol{X}$, and assume that the observations of response are independently sampled with means $g^*\big(\mathbf{x}_j|\boldsymbol{\theta}^*\big), j = 1, \dots, \mathsf{n}$. Due to the factorization of independent observations

$$L(\boldsymbol{\theta}|\mathbf{d}) \propto \prod_{j=1}^{\mathsf{n}} f\Big(y_j|\mathbf{x}_j, \boldsymbol{\theta}\Big).$$

The corresponding log-likelihood function is then

$$\log L(\boldsymbol{\theta}|\mathbf{d}) = \log \prod_{j=1}^{\mathsf{n}} f\Big(y_j|\mathbf{x}_j, \boldsymbol{\theta}\Big) = \sum_{j=1}^{\mathsf{n}} \log f\Big(y_j|\mathbf{x}_j, \boldsymbol{\theta}\Big).$$

Note that this leads to LS estimation when

$$f(y_j|\mathbf{x}_j, \boldsymbol{\theta}) \propto \exp\Bigg(- \Big(y_j - g(\mathbf{x}_j|\boldsymbol{\theta})\Big)^2 \Bigg).$$

Bayesian approach

The philosophical roots of Bayesian inference go back to Thomas Bayes, whose essay was published in 1764 by Price. At that time his work was mainly ignored by most, but the main idea (the Bayes's theorem) was rediscovered by Laplace (see [99] for details). In the Bayesian approach a data set is considered as fixed and an unknown parameter $\boldsymbol{\theta}$ is considered as a realization from an unobserved distribution. This is quite different from classical statistics in which a data set is considered to be random over repetitions of samplings and a parameter is considered as fixed. Given a data set one may write the distribution of the parameter as

$$f(\boldsymbol{\theta}|\mathbf{d}) = \frac{f(\mathbf{d}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int f(\mathbf{d}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

where $f(\mathbf{d}|\boldsymbol{\theta})$ is the likelihood and $f(\boldsymbol{\theta})$ is the prior distribution of the parameter. The distribution $f(\boldsymbol{\theta}|\mathbf{d})$ is called posterior distribution. Prior information about the parameter $\boldsymbol{\theta}$ is incorporated to Bayesian analysis through prior distribution, which is set before the analysis is carried out. Prior distribution needs not to be proper, thus it need not integrate to value 1. However, inferences about the parameter are done from the posterior distribution which must be proper. Posterior corresponds to prior distribution updated by the data set. As an example, one may compute a maximum a posteriori (MAP) estimate which is defined as

$$\boldsymbol{\theta}^{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, f(\boldsymbol{\theta}|\mathbf{d}).$$

One can find more details on Bayesian inference in [30].

The Bayesian approach and the maximum likelihood method both utilize likelihood. However, in Bayesian inference prior knowledge is also incorporated in the analysis. These methods lead to the same solution in two situations. First is when the prior has infinite support and the number of observations approaches infinity. Secondly, maximum a posteriori estimate equals the maximum likelihood estimate always when prior distribution is uniform with infinite support.

2.3 Estimation from incomplete data

In the following we have a short review of estimation from incomplete data. This includes basic ideas for

a) weighted estimators

b) missing data likelihood estimation, and

c) Bayesian approach and simulation.

The first idea, weighted estimators, is to adjust complete observations with weights to compensate for the missing ones. The second alternative is based on the Maximum likelihood concept that can be generalized for incomplete data. Likelihood may be decomposed into parts corresponding to observed data and the conditional distribution of the missing data given the observed data and the parameters. Then the likelihood is maximized with a suitable method. The third approach is based on Bayesian inference, where estimates are computed from the posterior distribution of the parameters of observed data and missing-data indicators. The Bayesian approach may also be used for imputation. The idea is to draw from the predictive posterior distribution of missing data given observed data. However, in practice draws from the predictive distribution may be difficult to do. There exist some simulation methods to ease in this task.

2.3.1 Weighting methods

In weighting methods the estimators are adjusted for nonequal sampling. Perhaps the best known weighting method is the so-called Horvitz-Thompson (H-T) estimator [39], where each observation j is weighted by the inverse of its sampling probability π_j. Thus observations Y_j are replaced with weighted observations Y_j, w_j, where $w_j = \pi_j^{-1}$ and the corresponding sample size is $n_w = \sum_{j=1}^{\mathsf{n}} w_j$. This idea applies naturally to missing data as well. We only need to consider missingness as weighted sampling, where sampling probability π_j is replaced by the sampling probability multiplied by the conditional probability for response given unit was sampled. Typical examples include:

The Horvitz-Thompson estimator [39] of a finite population total is defined as

$$\hat{T}^{H-T} = \sum_{j=1}^{\mathsf{n}} Y_j \pi_j^{-1} = \sum_{j=1}^{\mathsf{n}} Y_j w_j.$$

For population mean the H-T estimator takes the form

$$\hat{\mu}^{H-T} = \frac{1}{n_w} \sum_{j=1}^{\mathsf{n}} Y_j w_j,$$

In practice conditional probabilities for the response given unit was sampled are not known. There are multiple approaches to estimate them such as weighting class estimators, propensity weighting, and weighted generalized estimation equations. For simplicity, the details are omitted here. The descriptions of the methods are given in [62].

One should note that in the sampling theory, in which weighting estimators are common, a capital letter followed by an index typically denotes (non-random) population value and a small letter followed by an index denotes random observation. In the current thesis we do not use that notation.

2.3.2 Likelihood based methods

Likelihood estimation can be extended for incomplete data also. Direct methods are based on incomplete data likelihoods, while another popular approach is based on iterative refinement toward expected complete likelihood. Both approaches can be seen as a generalization of the maximum likelihood principle.

Application of maximum likelihood (ML) to incomplete data is theoretically straightforward. However, in practice one may encounter some analytical problems as some equations which are needed in estimation may be difficult to derive. This difficulty arises typically when the missing-data mechanism is complicated. The construction of Expectation Maximization is often considered easier than that of the direct missing likelihood model. Therefore it is possible that with the EM method more complicated missing-data problems may be solved in practice. Usually both the direct ML and EM assume that data missingness is ignorable, as described in Section 2.1.1. We begin by introducing maximum likelihood, after which EM is described.

Missing data likelihood and its maximization

Maximum likelihood was applied analytically to estimation from incomplete data as early as 1932 by Wilks [111]. A lot of work has been done with maximum likelihood (ML) estimators of the first two moments of a population from an incomplete data. Wilks derived the ML estimators of the parameters (means, standard deviations, and correlation) of bivariate normal population from incomplete sample. Interest remained high and more research work followed by such authors as Lord (1951), Hartley (1958), Edgett (1956), Anderson (1957), Nicholson (1957), Hocking and Smith (1968), and Morrison (1971) [64, 34, 21, 3, 75, 38, 74] among others. Most of the research work was restricted on bivariate or trivariate gaussian distribution and specific missing-data patterns. The idea of Hartley in [34] was to simplify and unify maximum likelihood computations and estimates from incomplete data.

The basic idea, as described by Schafer [91], is that for any incomplete data $\mathbf{d}^{inc}$ the parametrized complete data density can be written as

$$f_{\boldsymbol{Y}}(\mathbf{y}|\boldsymbol{\theta}) = f_{\boldsymbol{Y}}(\mathbf{y}^{obs}, \mathbf{y}^{mis}|\boldsymbol{\theta}) = f_{\boldsymbol{Y}^{obs}}(\mathbf{y}^{obs}|\boldsymbol{\theta}) f_{\boldsymbol{Y}^{mis}|\boldsymbol{Y}^{obs}}(\mathbf{y}^{mis}|\mathbf{y}^{obs}, \boldsymbol{\theta}).$$

The log likelihood can then be written as

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{y}^{obs}) + \log f_{\boldsymbol{Y}^{mis}|\boldsymbol{Y}^{obs}}(\mathbf{y}^{mis}|\mathbf{y}^{obs}, \boldsymbol{\theta}),$$

where $l(\boldsymbol{\theta}|\mathbf{y}^{obs})$ is the observed data likelihood and $\log f_{\mathbf{Y}^{mis}|\mathbf{Y}^{obs}}(\mathbf{y}^{mis}|\mathbf{y}^{obs},\boldsymbol{\theta})$ is the likelihood for conditional distribution of the missing data given observed data and parameters.

An advantage of the maximum likelihood approach is that it is sometimes possible to calculate explicit estimates under an assumption that mathematical computations related to the problem are not too difficult. Typically one defines the model for $l(\boldsymbol{\theta}|\mathbf{y}^{obs})$ and $f_{\mathbf{Y}^{mis}|\mathbf{Y}^{obs}}(\mathbf{y}^{mis}|\mathbf{y}^{obs},\boldsymbol{\theta})$ and solves the maximum likelihood problem using a suitable method. The development of information technology and the advantages in stochastic optimization techniques has made this approach feasible for complex likelihood functions as well.

Expectation Maximization

As stated earlier maximum likelihood inference for complicated missing-data mechanisms can be quite difficult to apply. Expectation-Maximization (EM, introduced by Dempster, Laird, and Rubin [18]) algorithm provides a way to simplify the problem.

Assuming ignorable (MAR) missing-data mechanism, for simplicity, likelihood can be written in the form of

$$L(\boldsymbol{\theta}|\mathbf{y}) = \int f_{\mathbf{Y}^{obs},\mathbf{Y}^{mis}}(\mathbf{y}^{obs},\mathbf{y}^{mis}|\boldsymbol{\theta})d\mathbf{y}^{mis}.$$

Instead of maximizing the full likelihood directly the idea of the EM algorithm is to maximize the expected log likelihood, which is also known as the Q-function

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \int \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) f_{\mathbf{Y}^{mis}|\mathbf{Y}^{obs}}(\mathbf{y}^{mis}|\mathbf{y}^{obs},\boldsymbol{\theta}^t)d\mathbf{y}^{mis},$$

where $\boldsymbol{\theta}^t$ is the previous estimate of parameters $\boldsymbol{\theta}^*$. This leads to iteration, which in well-defined problems is shown to converge to a stationary global maximum of full likelihood (see [112] for more details).

But, not all problems are well defined. There can be multiple modes, saddle points, likelihood ridges, and boundary issues [91]. On the other hand, other estimation methods must also face the issues.

The EM procedure is described in Algorithm 2.1. At first, an initial parameter estimate is selected and a convergence criterion is set. In the E step the conditional expectation of ”missing data” given the observed data and current estimated parameters is computed. The quotation marks are used because EM does not necessarily impute missing values themselves. Little and Rubin have stated [62] that

> “The key idea of EM, which delineates it from the *ad hoc* idea of filling in missing values and iterating, is that ”missing data” are not Y_{mis} but the functions of Y_{mis} appearing in the complete-data loglikelihood $l(\boldsymbol{\theta}|Y)$.”

However, there are many ad hoc iterative imputation-estimation methods which in fact are EM algorithms for models where the complete data loglikelihood $l(\boldsymbol{\theta}|\mathbf{y}^{obs},\mathbf{y}^{mis})$

is linear in $\mathbf{y}^{mis}$ [62]. Note that the EM algorithm does not provide a variance estimate. However, Meng and Rubin have introduced a supplemental EM (SEM) algorithm which does this [68].

Algorithm 2.1 The EM algorithm

1. Set initial parameter estimate $\boldsymbol{\theta}^{(1)}$ and convergence criterion δ to some small positive value close to zero.

2. **E-step:** compute the expected complete-data loglikelihood if $\boldsymbol{\theta}$ were $\boldsymbol{\theta}^{(t)}$:
$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int l(\boldsymbol{\theta}|\mathbf{y}^{obs}, \mathbf{y}^{mis}) f(\mathbf{y}^{mis}|\mathbf{y}^{obs}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}) d\mathbf{y}^{mis}.$$

3. **M-step:** maximize the expected complete-data loglikelihood to solve $\boldsymbol{\theta}^{(t+1)}$:
$$\mathcal{Q}(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \text{ forall } \boldsymbol{\theta}.$$

4. If $||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}|| > \delta$ then repeat from Phase 2.

2.3.3 Bayesian approach

Bayesian inference under general missing-data mechanism is obtained from the distribution $f(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}, \mathbf{M})$, where $\mathbf{M}$ is the response indicator matrix. The inference is related to maximum likelihood inference from full likelihood L_{full}. However, prior information is incorporated in the analysis formally as

$$f(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}^{obs}, \mathbf{M}) \propto f(\boldsymbol{\theta}, \boldsymbol{\psi}) L_{full}(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}^{obs}, \mathbf{M}),$$

where L_{full} is any function proportional to $f(\mathbf{y}^{obs}, \mathbf{M}|\boldsymbol{\theta}, \boldsymbol{\psi})$ which is defined as

$$f(\mathbf{y}^{obs}, \mathbf{M}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\mathbf{y}^{obs}, \mathbf{y}^{mis}|\boldsymbol{\theta}) f(\mathbf{M}|\mathbf{y}^{obs}, \mathbf{y}^{mis}, \boldsymbol{\psi}) d\mathbf{y}^{mis}.$$

As before, inference becomes simpler if the missing-data mechanism is ignorable. Under that assumption

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}^{obs}, \mathbf{M}) &\propto f(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{y}^{obs}) f(\boldsymbol{\psi}) L(\boldsymbol{\psi}|\mathbf{y}^{obs}, \mathbf{M}) \\ &\propto f(\boldsymbol{\theta}|\mathbf{y}^{obs}) f(\boldsymbol{\psi}|\mathbf{y}^{obs}, \mathbf{M}). \end{aligned}$$

Thus, inferences of $\boldsymbol{\theta}$ can be based on the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y}^{obs})$. It should be noted that the definition of ignorable mechanism for Bayesian inference is stronger than for maximum likelihood inference. Namely, a priori independence of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ requires distinctness of parameter spaces $\Omega_{\boldsymbol{\theta}}$ and $\Omega_{\boldsymbol{\psi}}$. More details on Bayesian inference for incomplete data set can found in [62] and [30].

Bayesian methodology may also be used for imputation. In Bayesian imputation the idea is to draw from the posterior distribution $f(\mathbf{y}^{mis}|\mathbf{y}^{obs})$. However, in practice draws from the previous distribution are difficult to do. There exist some methods which may be utilized for imputation. Little and Rubin list data augmentation and the Gibbs sampler, which both are iterative methods and suitable for ignorable missing-data mechanism. For simplicity we describe data augmentation only. Details on Gibbs sampling for imputation can be found in [62].

Data augmentation

Data augmentation was invented by Tanner and Wong [100]. It is an iterative simulation method. The original algorithm consists of a multiple imputation step at each iteration. However, Little and Rubin consider a definition of data augmentation that is slightly different from the one of Tanner and Wong. The algorithm steps considered by Little and Rubin are the following:

I step: draw $\mathbf{y}^{mis,(t+1)}$ from $f(\mathbf{y}^{mis}|\mathbf{y}^{obs}, \boldsymbol{\theta}^{(t)})$, where t denotes iteration number

P step: draw $\boldsymbol{\theta}^{(t+1)}$ from $f(\boldsymbol{\theta}|\mathbf{y}^{obs}, \mathbf{y}^{mis,(t+1)})$.

The first step corresponds to an imputation step, whereas the second one is a posterior step. These two steps are typically simpler than drawing from the posterior distribution of $\mathbf{y}^{mis}$ given $\mathbf{y}^{obs}$. The two steps are repeated sufficiently long. This iterative process can be shown to yield, in the limit, a draw from the joint posterior distribution of $\mathbf{y}^{mis}, \boldsymbol{\theta}$ given $\mathbf{y}^{obs}$. In some sense this method is similar to the EM-algorithm. As an exception, in the first step the draws are done from a conditional distribution, whereas in the EM-algorithm conditional means are used. As a consequence, covariance-matrix estimations require no magnitude correction factors as they do in EM. As a pitfall, estimation efficiency of the mean is lost. However, this can be compensated by averaging the results of multiple repetitions of the whole algorithm, but with a computational cost.

2.4 Imputation

This thesis is about imputation, the idea of which is to replace missing data values by their predictions. It should be noted that replacement of missing data values may be considered as a data reconstruction problem, which is explained in detail, for example, in reference [71]. Before describing imputation in more detail we briefly summarize its history. A more complete overview of the subject can be obtained by reading the references [46, 62, 115, 89, 2].

The earliest ideas on imputation are probably those by Allan and Wishart in 1930. Rubin [86] writes that

> “The formulae giving the least squares estimates for one missing cell in randomized block and latin square designs were first given by Allan and Wishart (1930).”

The Rubin's paper [86] consists of a non-iterative algorithm for filling in missing data values in the analysis of any variance design. Other early ideas are refereed in the introduction part of Cheng's paper [14]. Cheng writes

> "Yates (1933), Bartlett (1937), and Healy and Westmacott (1956) introduced the ideas of filling in the least squares estimates of all missing values in the analyses of variance and covariance."

Hartley's 1956 paper contains a unique formula for filling in missing values in the analysis of variance for any design [33]. However, in case of multiple missing values the formula has to be used iteratively. In 1960 Buck defined a regression imputation method for the estimation of the covariance-variance matrix of any population from incomplete data [7]. Imputation was used in applications in 1950s and 1960s for example by Jaszi, Phillips, and Wharton [43, 77, 110]. After the 1960s other researchers, including Rockwell and Fellegi and Holt, became involved with imputation [84, 23]. According to Zhao [115] the major reason for the use of imputation was to achieve complete data. Rubin introduced multiple imputation in 1987 [87]. Next we will summarize the desirable properties of imputation, after which the best known imputation procedures are introduced.

2.4.1 Desirable properties and drawbacks of imputation

We refer to a good review paper of imputation written by Kalton and Kasprzyk [46] and to a paper written by Jinn [44].

The desirable properties of the imputation are [46]

- "First... it aims to reduce biases in survey estimates from missing data...",
- "Second, by assigning values at the micro level and thus allowing analyses to be conducted as if the data set were complete, imputation makes analyses easier to conduct and results easier to present. Complex algorithms to estimate population parameters in the presence of missing data (e.g., the Expectation-Maximization algorithm of Dempster, Laird and Rubin, 1977) are not required.", and
- "Third, the results obtained from different analyses are bound to be consistent, a feature which need not apply with an incomplete data set.".

These viewpoints are especially important in statistical offices. Non-response in surveys and censuses may cause considerable bias to analyses if it is not taken into account properly. Therefore the first property is very practical. Secondly, maximum likelihood and expectation maximization approaches require more trained people, who may not be able to apply them due to mathematical difficulties. One of the advantages of the third property is that a covariance matrix from an imputed data set is likely to be proper, which is not necessary true for covariance matrix estimated directly from incomplete data set.

Naturally there are some drawbacks as well.

- Imputation “does not necessarily lead to estimates that are less biased than those obtained from the incomplete data set” [46],
- ”there is also a risk that analysts may treat the completed data set as if all the data were actual responses, thereby overstating the precision of the survey estimates.” [44], and
- “Even if the biases of univariate statistics are reduced, the relationship between variables may be distorted.” [44].

In the first drawback estimates refer to survey estimates such as mean and total. A solution for the second drawback is to take imputation uncertainty into account. This may be accomplished by multiple imputation.

2.4.2 Single imputation

The idea of single imputation is to replace each missing data value by a single prediction.

The main advantages of single imputation are that it is simple, yields a complete data set, and requires no additional storage space for the data set. Therefore standard complete-data methods of analysis can be applied after single imputation. Rubin mentions also that a major advantage of single imputation is that data collector’s knowledge may be incorporated in imputations [87].

A data collector has benefits over a typical user of a data set. The collector may have better information about and understanding of the process that creates the missing data. This is an especially important factor in cases where the collector (for example the Census Bureau) has more information available for imputation than would be available for public-use on the resultant data bases. Secondly, the collector is likely to have greater resources for analysis than a typical user of a data set.

A disadvantage of single imputation is that application of complete-data methods to imputed data sets treats missing values as if they were known. Treating missing values as if they were known is problematic. Inferences based on the imputed data set will be too sharp, because additional variability due to unknown values is not being taken into account. Quantities such as correlations that depend on variabilities can be badly biased. Rubin describes another problematic issue of single imputation [87]

> “when non-response is not really understood, no account is being taken of the uncertainty arising from not knowing which nonresponse models for imputation are appropriate”

2.4.3 Multiple imputation

In this thesis we do not consider multiple imputation in analytical or empirical studies. However, because it is widely referred to in publications and used in applications, we consider it important to make the reader aware of it by offering a brief description here.

The first seeds for multiple imputation were sown by Rubin in the late 1970s [88], and a lot of development ensued. The idea in multiple imputation is to first produce several, say k, versions of the completed data sets with an imputation model. Secondly, normal complete-data set analyses are done for each of the k data sets. Finally, the results are combined using Rubin's rules.

The advantages of multiple imputation include the two advantages of single imputation: it allows the use of complete-data methods of analysis and the incorporation of the data collector's knowledge. The second advantage is actually enhanced. Data collectors can use their knowledge for reflecting uncertainty about which values to impute. However, there are also three important advantages compared to single imputation.

First of all, multiple imputation increases the efficiency of estimation. Efficiency here refers to the reduction of imputation related variance. Secondly, under certain assumptions one can use a combination of standard complete data methods to do inferences on imputed data. For example, it is rather straightforward to obtain estimates of imputation variances. The third advantage is that one can study the sensitivity of inferences to different models for non-response. The reader is suggested to read the book by Rubin for more details [87].

The disadvantages compared to single imputation are that more work is needed to produce the completed data sets, and more space is required to store data. However, these disadvantages are not severe when the value of k is reasonable (i.e., not too large). The value of k typically has to be increased as a function of fractions of the missing information, otherwise the multiple imputation is not fully satisfactory [87].

2.5 Model classes and imputation strategies

The emphasis of this thesis is in a scenario where incomplete data is replaced with a single imputed data set. Thus we are not directly involved with multiple imputation. In addition, we are interested in cell imputation, where imputation is done in subsets or clusters of data. Roughly put, the idea is to divide data into clusters (cells) that are made of similar observations and apply imputation in each of these clusters more or less independently. Optimally data variation inside the cells would correspond to unexplained noise, while variation between the cells could be explained by observed covariates.

The above viewpoint leads to several open questions, including

- How is imputation done inside the cells?
- How is the cell partitioning (clustering) obtained?
- How can we evaluate the results?

To answer these, several imputation models are used together with a couple of different imputation strategies. More formally, we consider model assisted imputation in the form of

$$Y^{imp} = g(\mathbf{x}^{obs}|\boldsymbol{\theta}) + \hat{\epsilon}(\mathbf{x}^{obs}),$$

where $g(\mathbf{x}^{obs}|\boldsymbol{\theta})$ is the approximative model for our data and $\hat{\epsilon}(\mathbf{x}^{obs})$ represents our estimate of the unexplained noise. For models $g(\mathbf{x}^{obs}|\boldsymbol{\theta})$ we use either

i) data mean $g(\mathbf{x}^{obs}|\boldsymbol{\theta}) = \mu^{obs}$, where covariates are ignored, and which serves as a baseline for comparative evaluations

ii) regression methods, where we use both parametric and nonparametric approaches, namely linear regression, nearest neighbour regression and kernel regression.

iii) clustering methods of type $g(\mathbf{x}^{obs}|\boldsymbol{\theta}) = \mu_{\mathrm{b}(\mathbf{x}^{obs})}$, where $\mu_1, \mu_2, \ldots, \mu_{\mathsf{n}_c}$ are cluster (cell) centroids and $\mathrm{b}(\mathbf{x}^{obs})$ is the selector of the "best" centroid for observation $\mathbf{x}^{obs}$.

For noise $\hat{\epsilon}(\mathbf{x})$ we use three imputation strategies

M (mean): where noise is omitted, and which we call (model) mean imputation

R (random): where noise is simulated, implying that $\hat{\epsilon}(\mathbf{x}^{obs})$ is taken as an iid sample from a noise model $f_{\hat{\epsilon}(\mathbf{x}^{obs})}(e|\mathbf{x}^{obs})$

D (donor): random donor, where $\hat{\epsilon}(\mathbf{x}^{obs})$ is picked randomly with replacements from a set of centered observed values, $\{y_k - g(\mathbf{x}^{obs}|\boldsymbol{\theta})\}, y_k \in \mathbf{d}^{train}$

Since imputation strategies M, R and D can be combined with any of the models in categories i), ii) and iii), we shall organize the current study according to model types. When there is no danger of misunderstandings we refer to these types as mean imputation, regression imputation, and cell imputation.

2.5.1 Mean and random donor imputation

In mean imputation the mean of observed data is substituted as a replacement for missing data values. In random donor imputation the missing data values are replaced by drawing observed data values. Drawings may be done with or without replacements. In this thesis we consider samples with replacement. These baseline methods do not utilize any covariate information thus they are expected to perform poorly in regression tasks.

The results for mean and random donor imputation methods are well known in the literature (for example see [62] for details). However, there exist some results for random donor imputation which are not available in basic literature. Namely, Lai's thesis of 1998 [59] consists of derivation of some analytical properties of mean estimator based on multivariate random hot deck under MCAR and special MAR response mechanisms. Some of the Lai's formulas are also available in a paper written by Shen and Lai in 2001 [94], which concerns the quality of life data.

2.5.2 Regression imputation

Regression methods are an obvious choice for any modelwise evaluation of a methodology. Assuming that response Y can be explained by fully observed covariates $X_1, \ldots, X_{p-1}$, some regression model could be an optimal way to solve imputation problems. There is also a lot of easily obtainable knowledge about regression methods. Especially the theory of linear regression is well established (see for example [66]). Therefore linear regression is included in our study as well. For wider perspective we consider also some nonlinear and nonparametric alternatives such as nearest neighbour regression [72] and kernel regression [70]. In general all regression methods are likely to perform better than simple mean and random donor imputation methods under non-MCAR missing-data mechanisms.

Basic results of linear imputation can be found from standard references such as [62]. A recent example of analyses can be found from the PhD thesis of Zhao [115]. He assumes a MAR missing-data mechanism and a fixed design, i.e., the observations of explanatory variables are fixed. The biases of imputation estimators for the first two moments are derived, as is the variance of the estimator for the first moment. Some interesting publications related to imputation and estimation of the linear regression model have been authored, e.g., by Toutenburg, Skrivastava, Shalabh, and Jinn [102, 103, 104, 107, 44].

We are not aware of any thorough analyses of k-nearest neighbour methods (k-NN) in the context of imputation. Chen and Shao (1997) have introduced results for 1-nearest neighbour imputation in a survey framework [12]. They give an approximation for bias and variance of the imputed mean of a population. An exact bias is also derived for some specific distributions, and it is shown that the empirical distribution function of the imputed data converges asymptotically to the true distribution function. Cheng (1994) has derived the asymptotic distribution for the mean of the imputed data for Nadaraya-Watson (NW) kernel regression [14]. Cheng remarks that the asymptotic distribution is also the same for k-nearest neighbour regression. However, covariate is assumed to be univariate. In another work Hruschka et al [40] have used K-Means clustering to optimize computational complexity of k-NN regression imputation. Their idea is similar to binning used to optimize speed in kernel regression methods (see for example [41] for details).

One problem with nearest neighbour methods is their computational complexity which limits the applicability of the method to small data sets. One possible solution is proposed by Hruschka et al. [40], where data is first compressed with K-means clustering. Then the nearest neighbour imputation canditates are searched from the smaller set of the cluster centroids instead of the original observations. Empirical studies show that the method is suitable, i.e., its results are comparable with imputations provided by k-NN in a number of regression and classification tasks.

Some of our methods are based on smoothing. In regression smoothing a weighted mean of response values over a local neighborhood is used as a prediction. The idea is similar to the borrow-strength[†] technique, which is used in small area estimation (SAE) [82]. The idea is that the neighborhood of a small area is used to increase the number of observations. As a consequence estimation variance is typically decreased, but squared bias is increased. In the imputation context this technique has not been studied much. Titterington and Mill (1983) introduced a kernel-based method for density estimation from incomplete data [101]. That paper contains consistency results for a density estimate, and the method has been used for imputation as an application. Cheng (1994) deals with the estimation of expectation of a variable from incomplete data using a nonparametric (kernel) method [14].

2.5.3 Cell imputation

Cell imputation requires three phases. First, cells are constructed. This may be done using clustering methods, such as K-Means [32], SOM [54], or mixture modelling [24], but we may also use external knowledge like gender or age categories for the task. Second, all observations, including incomplete ones, are associated to cells using observed data. This may require the use of a probabilistic classifier. Finally, the missing data values are imputed using information that is assigned to the corresponding cell. This can be done using previously introduced methods, for example mean or donor imputation.

Some analytical research have been done on cell imputation. A report written by Santos (1981) [90] consists of analytical derivations for cell mean and cell random hot deck methods. His study also includes a baseline and two linear regression methods. He considers a general mechanism, a MCAR mechanism, and a MCAR within clusters missing-data mechanism. The cells are assumed to be fixed, i.e., based on some deterministic division of the cell covariate. Santos assumes a finite population sampling framework and presents large sample biases of covariance and variance estimators based on imputation. Kalton and Kish (1981) [47] used clustering in order to reduce variance, due to the hot deck procedure, to neglible level. They stratified the respondents by their target values into equal-sized stratas. Kim and Fuller [50] study analytical properties of the mean estimator based on the fractional hot deck imputation within cells. There are also some empirical studies where K-Means or SOM clustering has been used in the imputation method.

Self-organizing maps have been used for prediction of missing data values in some applications. However, all the studies that we know of are empirical. Fessant et al. compare three different cell imputation methods based on SOM [25]. Their application consists of imputation of missing values in a French personal transport data set. Rallo et al. apply SOM to impute missing data values in an application in the chemical engineering field [80]. Their imputations are based on a SOM prototype imputation, thus an incomplete data set is assigned to a winning node in which its missing data values are replaced by the corresponding values of a prototype vector.

[†] We thank Ray Chambers for this observation.

Cottrell and Letrémy give ideas about how to apply Kohonen's algorithm for an incomplete data set [15]. They also discuss imputation of missing data in three applications. SOM based multiple imputations are empirically tested by Rallo et al. [81]. Their application consists of two industrial processes.

Finally we want to note that there are some imputation methods based on classification and regression tree models [6] which are functionally equivalent to cell imputation. Imputation methods utilising classification and regression trees have been used in some applications. Regression tree based imputation methods were used in the EurEdit project [11]. In addition, Creeli and Krotki have compared some imputation methods which are based on tree methods [17].

2.6 Summary

Missing-data mechanisms were defined in the beginning of the chapter. Statistical properties, such as bias of estimators computed from the imputed data may be crucially affected by the missing-data mechanism. The concept of ignorability of missing-data mechanism was introduced. Ignorability typically makes maximum likelihood and Bayesian analyses from incomplete data easier. Methods for identification of missing-data mechanism were briefly described.

The basic estimation methodology was reviewed for both complete and incomplete data. The least squares, maximum likelihood, and Bayes approaches were considered for complete data. Estimation alternatives for incomplete data were described. The methods included weighting, maximum likelihood, Bayes approach and simulation. In this thesis we utilize neither the Bayes nor the simulation approach.

Imputation, the main subject of this thesis, was described. A brief review of the history of imputation was given. Some desirable properties and drawbacks of imputation were listed. Single and multiple imputation approaches were also described. In this thesis we are not directly involved with multiple imputation.

Finally, model classes and imputation strategies were introduced. Model classes included mean imputation (and random donor), regression imputation, and cell imputation. The main focus in this thesis is on analyses of cell imputation. However, the first two types mentioned serve as comparative methods. Previous research work related to model types was described.

Chapter 3

The setup of the problem and basic decompositions

This chapter defines the research problem and motivates the reader about possible practical uses of the study. Our description begins with a real-world scenario, where an incomplete data is replaced with an imputed one. A key note is that unless one knows exactly what the future use of the imputed data set is, it is impossible to say what the best imputation method for the task is. Therefore we are interested in methods that perform reasonably well according to many different viewpoints. Our hypothesis is that cell imputation is one of such methods.

To quantify our hypothesis, several measures of imputation performance are introduced. We consider how missingness and imputation methodology affect statistical estimates, predictions and conclusions. More specifically we are interested in the trade-off between the preservation of the properties of data on the unit level and those on the distribution level. The assumptions for and the conditions of the measures are described in this chapter. An insight about the practical relevance of our limitations is also explained.

Recall, from Section 2.5, our taxonomy of different imputation models and strategies. Our aim here is to describe how we are to compare those models against each other. This leads to a combination of theoretical and simulation studies in Chapter 7.

Finally, we describe the theoretical framework for the thesis and its relation to simulation studies and real-world data.

Before describing the real-world scenario, the notation which is used in this thesis, is defined.

3.1 Notation

We need to distinguish between parameters, random variables, estimators, and fixed values. Table 3.1 illustrates the notation used for the following quantities: expectation and variance of target Y, expectation and variance of covariate $\boldsymbol{X}$, expectation of Y given $X = \mathbf{x}$ (conditional mean), and expectation of variance of Y given $X = \mathbf{x}$ (over the distribution of covariate $\boldsymbol{X}$).

Parameter	R.v./distribution	Estimator	Fixed value
μ^*	$Y \sim f_Y$	$\hat{\mu}$	μ
τ^*	$Y \sim f_Y$	$\hat{\tau}$	τ
$\overline{\boldsymbol{X}}^*$	$\boldsymbol{X} \sim f_{\boldsymbol{X}}$	$\hat{\overline{\boldsymbol{X}}}$	$\overline{\boldsymbol{X}}$
$\boldsymbol{\Sigma}^*_{\boldsymbol{X}}$	$\boldsymbol{X} \sim f_{\boldsymbol{X}}$	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$	$\boldsymbol{\Sigma}_{\boldsymbol{X}}$
$g^*(\mathbf{x})$	$Y\|\boldsymbol{X} \sim f_{Y\|\boldsymbol{X}}$	$\hat{g}(\mathbf{x})$	$g(\mathbf{x})$
v^*	$\epsilon_{\|\mathbf{x}} = Y_{\|\mathbf{x}} - g^*(\mathbf{x}) \sim f_{\epsilon\|\mathbf{x}}(e)$	$\hat{v}$	v

Table 3.1: Illustration of notation. R.v. is the abbreviation for random variable.

The notations for data set, variable, and observations are depicted in Table 3.2. Note that the fixed value of the realization of j:th noise term is denoted by e_j, whereas random quantity is denoted by ϵ_j.

Random quantity	Fixed value/realization
$\mathbf{D}$	$\mathbf{d}$
Y	y
Y_j	y_j
ϵ_j	e_j

Table 3.2: Notation for data set, variable, and observations.

For some quantities it is also necessary to distinguish the data set to which the quantity is associated. This is denoted by a superscript as illustrated for the mean estimators in Table 3.3. Further, in many results of this thesis it is also necessary to mark the imputation method and strategy. This is denoted by superscript ^{method,strategy}. As an example $\hat{\mu}^{comp,B,M}$ is the mean estimator which is computed using baseline imputation strategy.

Symbol	Description
$\hat{\mu}$	Estimator is associated to true data set.
$\hat{\mu}^{obs}$	Estimator is associated to observed data set (completely observed observations).
$\hat{\mu}^{mis}$	Estimator is associated to missing data set.
$\hat{\mu}^{imp}$	Estimator is associated to imputed data set.
$\hat{\mu}^{comp}$	Estimator is associated to completed data set.

Table 3.3: Example of notation which is used to depict to which data set the quantity is associated.

Subscripts are used in indexing (of observations, cells, or covariate dimension) and for giving details. Examples of the use of subscripts include:

- Indexing

  i) dimension indexing X_u,
  ii) observation index x_j (paired with the dimension index: $x_{j,u}$), and
  iii) cell index $\hat{\mu}_i^{obs}$ or cell index pair $h_{i,l}$, where $h_{i,l}$ is a quantity between cells i and l.

- Details about matrix size etc.

  i) dimensions of matrix or number of observations on which quantity is based:

     a) $\mathbf{A}_{\mathsf{n}\times p}$ matrix with n rows and p columns,
     b) $\mathbf{d}_{\mathsf{n}}$ data set with n observations,

  ii) referred variable: for example $\mathbf{\Sigma}^*_{\mathbf{X}}$, and
  iii) number of cells: n_c.

3.2 A real-world scenario

This thesis is not from the viewpoint of one method in one specific setting. Instead we are seeking to obtain an understanding on a more complex problem, i.e.:

> **How good is the imputation with respect to unknown future uses of data?**

To concretise the idea we have made a simple practical scenario that applies to some working processes in statistical offices. To put it shortly, we consider a situation where an incomplete data set is imputed and is then sent to an analyst. This is depicted in Figure 3.1. The imputer receives an incomplete data and returns a completed one. This data may then be used for several types of application dependent analyses. We do not except that the analyst is an expert of statistics with incomplete data, and therefore he/she must have a complete data set that is good enough for practical situations. The analysis of a completed sample is supposed to be done independently of imputations. As is the case in the real-world, the analyst may not know that the sample has been imputed.

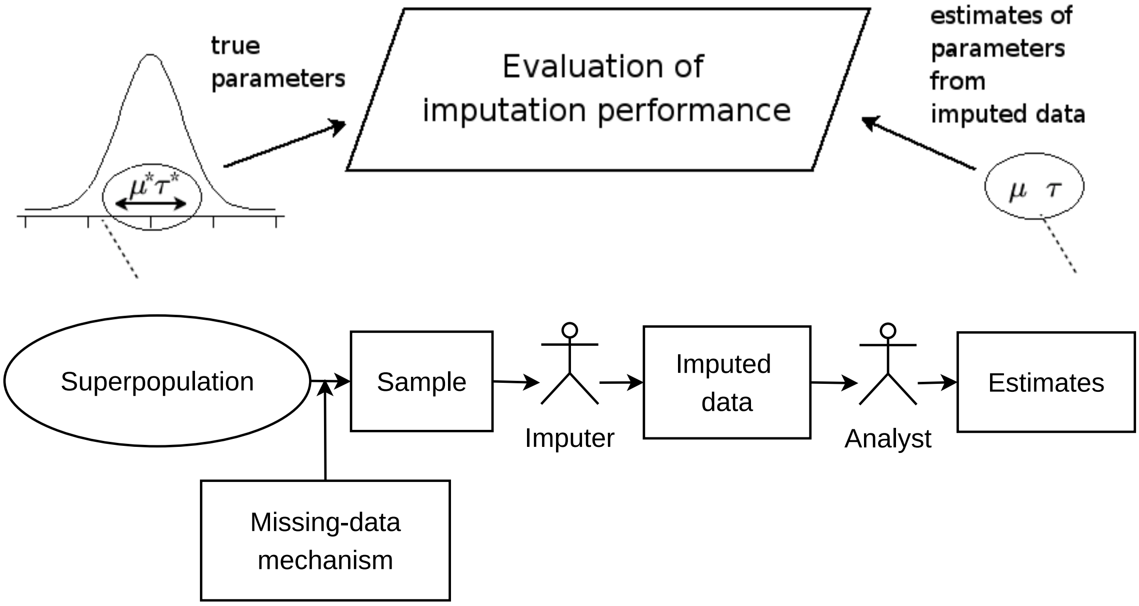


Figure 3.1: Data flow chain.

For simplicity we assume throughout this thesis that missingness is limited to one variable only, which is denoted by Y, while fully observed variables are denoted by covariates $X_1, X_2, \ldots, X_{p-1}$. This limitation simplifies theoretical analyses, and allows us to write missingness via a univariable random response variable R, whose realisation is the indicator

$$r_j = \begin{cases} 0, & \text{when element } d_{j1} \in \mathbf{d} \text{ is missing, and} \\ 1, & \text{otherwise.} \end{cases}$$

where $\mathbf{d}$ is the realization of a true data set.

Let the probability for missingness be p^*, formally $\Pr(R_j = 0) = p^*$. One should not confuse p^* and p which have different meanings. Now distributions of Y and $\boldsymbol{X}$ for random true data $\mathbf{D}$ can be written as a two component mixture of the observed part and the missing part. Formally, the joint distribution can be decomposed as

$$f_{Y,\boldsymbol{X}}(y, \mathbf{x}) \quad = \quad (1-p^*) f_{Y^{obs},\boldsymbol{X}^{obs}}(y, \mathbf{x}) + p^* f_{Y^{mis},\boldsymbol{X}^{mis}}(y, \mathbf{x}). \qquad (3.1)$$

The reader should note that here we have neither restricted the framework to be the selection model

$$f(R, Y, \boldsymbol{X}) = f(R|Y, \boldsymbol{X}) f(Y|\boldsymbol{X}) f(\boldsymbol{X})$$

nor the (pattern)-mixture model

$$f(R, Y, \boldsymbol{X}) = f(R|\boldsymbol{X}) f(Y|R, \boldsymbol{X}) f(\boldsymbol{X}),$$

as described in references [63, 62, 87]. See Appendix A3.4.1 for details. Instead, both models are possible, when $f(Y, \boldsymbol{X})$ is written in terms of observed and missing distributions. It should be noted that the global missingness p^* does not apply for measures that are conditionalized with $\mathbf{x}$. For those cases we use notation $p^*_{\mathbf{x}}$ instead. We shall use Equation (3.1) to decompose our evaluation statistics as explained later. The random data matrix $\mathbf{D}$ of n observations can be defined as

$$\mathbf{D} = \{Y_j, \boldsymbol{X}_j\}_{j=1}^{\mathsf{n}}, \text{ where } Y_j, \boldsymbol{X}_j \overset{iid}{\sim} f_{Y,\boldsymbol{X}}(y, \mathbf{x}).$$

Given realized data matrix $\mathbf{d}$, we denote the observed vector of responses as $\mathbf{r} = [r_1, r_2, \ldots, r_{\mathsf{n}}]^T$. The application of $\mathbf{r}$ to realized true data $\mathbf{d}$ gives us a realized incomplete data set $\mathbf{d}^{inc}$, where some values of d^{inc}_{j1} are missing. Once we have the realizations of $\mathbf{D}$ and $\boldsymbol{R}$, the incomplete data matrix $\mathbf{d}^{inc}$ can be given in a reorganized form

$$\mathbf{d}^{inc} = [d^{inc}_{ji}]_{\mathsf{n}\times p} \left[\begin{array}{c|ccc} y_1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{\mathsf{n}^{obs}} & x_{\mathsf{n}^{obs},1} & \vdots & x_{\mathsf{n}^{obs},p-1} \\ \hline ? & x_{\mathsf{n}^{obs}+1,1} & \cdots & x_{\mathsf{n}^{obs}+1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ ? & x_{\mathsf{n},1} & \cdots & x_{\mathsf{n},p-1} \end{array}\right],$$

where ? denotes a missing value and n^{obs} is the number of observed values. Thus first n^{obs} cases are complete and the rest $\mathsf{n}^{mis} = \mathsf{n} - \mathsf{n}^{obs}$ are incomplete. The reorganization is applied also to the true data matrix $\mathbf{d}$. To prevent confusion with $\mathbf{d}$ the reorganized true data matrix is denoted as $\mathbf{d}^{true}$. Further, to ease indexing of response indicators for completed data set $\mathbf{D}^{comp}$, which is defined in Section 3.3, we redefine r_j to be

$$r_j = \begin{cases} 0, & \text{when element } y_j \text{ is missing in } \mathbf{d}^{inc}\text{, and} \\ 1, & \text{otherwise.} \end{cases}$$

From now on r_j refers to the response indicator for incomplete data set $\mathbf{d}^{inc}$ unless otherwise stated. Thus there is no risk of confusing it with indicators for original data set $\mathbf{d}$.

In the block form we use the following notations for partial matrices

$$\begin{aligned} \mathbf{d}^{inc} &= \left[\begin{array}{c|c} \mathbf{d}^{obs}_Y & \mathbf{d}^{obs}_{\boldsymbol{X}} \\ \hline ? & \mathbf{d}^{mis}_{\boldsymbol{X}} \end{array}\right] = \left[\begin{array}{c} \mathbf{d}^{train} \\ \hline \mathbf{d}^{mis} \end{array}\right] \\ \mathbf{d}^{true} &= \left[\begin{array}{c|c} \mathbf{d}^{obs}_Y & \mathbf{d}^{obs}_{\boldsymbol{X}} \\ \hline \mathbf{d}^{mis}_Y & \mathbf{d}^{mis}_{\boldsymbol{X}} \end{array}\right] = \left[\begin{array}{c} \mathbf{d}^{train} \\ \hline \mathbf{d}^{test} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{d}_Y & \mathbf{d}_{\boldsymbol{X}} \end{array}\right], \end{aligned}$$

where the names of matrices are obvious except $\mathbf{d}^{train}$ and $\mathbf{d}^{test}$, which are borrowed from the community of statistical pattern recognition. Usually $\mathbf{d}^{train}$ refers to training data, which is used in model building, while $\mathbf{d}^{test}$ is used for testing the model performance. In our case $\mathbf{d}^{test}$ is the incomplete part of true observations, and it can be used to "test" the performance of imputation methodology. Such a partitioning is useful in the analysis of imputations, since it allows the decomposition into observed and unobserved parts as $\mathbf{d}^{obs} = \mathbf{d}_Y^{obs} \cup \mathbf{d}_{\boldsymbol{X}}^{obs} \cup \mathbf{d}_{\boldsymbol{X}}^{mis}$. One should note that random train and test matrices and stochastics of their elements are defined as

$$\begin{aligned}
\mathbf{D}^{train} &= \{Y_j, \boldsymbol{X}_j\}_{j=1}^{\mathsf{N}^{obs}} && \text{where } Y_j, \boldsymbol{X}_j \overset{iid}{\sim} f_{Y^{obs}, \boldsymbol{X}^{obs}}, \text{ and} \\
\mathbf{D}^{test} &= \{Y_j, \boldsymbol{X}_j\}_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} && \text{where } Y_j, \boldsymbol{X}_j \overset{iid}{\sim} f_{Y^{mis}, \boldsymbol{X}^{mis}}.
\end{aligned}$$

We can concretise the theoretical setting as follows

i) True data $\mathbf{d}$ is an iid sample from an unknown distribution. The observed incomplete data $\mathbf{d}^{inc}$ is a result of a random missingness pattern that follows some distribution $f_R(r|\mathbf{I})$, where $\mathbf{I}$ may depend on data.

ii) Training data $\mathbf{d}^{train}$ is an iid sample from $f_{Y,\boldsymbol{X}|R=1}(y, \mathbf{x}) = f_{Y^{obs}, \boldsymbol{X}^{obs}}(y, \mathbf{x})$ and test data $\mathbf{d}^{test}$ is an iid sample from $f_{Y,\boldsymbol{X}|R=0}(y, \mathbf{x}) = f_{Y^{mis}, \boldsymbol{X}^{mis}}(y, \mathbf{x})$. One should note that this assumption implies that the missing-data mechanism can belong to the NMAR class.

iii) Data consists of a fixed number of n observations of variable Y and covariates $X_1, \ldots, X_{p-1}$, where all X_i are fully observed and some of the values of Y are missing. The number of missing values N^{mis} is assumed to be a random variable with mean $\mathbb{E}[\mathsf{N}^{mis}] = \mathsf{n}p^*$ where $p^* \in (0, 1)$, and given data $\mathbf{d}^{inc}$, we have realization $\mathsf{N}^{mis} = \mathsf{n}^{mis}$.

iv) Distributions of Y and X_i are unknown, but we know that the first two moments of Y do exist

$$\begin{aligned}
\mathbb{E}[Y] &= \mu^* \\
\mathbb{V}\text{ar}[Y] &= \tau^*.
\end{aligned}$$

The above assumptions also imply a requirement for the existence of conditional moments:

$$\begin{aligned}
\mathbb{E}[Y|\boldsymbol{X} = \mathbf{x}] &= g^*(\mathbf{x}) \text{ and} \\
\mathbb{V}\text{ar}[Y|\boldsymbol{X} = \mathbf{x}] &= v^*(\mathbf{x}),
\end{aligned}$$

where $v^*(\mathbf{x})$ is usually assumed to be constant $v^*(\mathbf{x}) = v^*$.

Furthermore, theoretical study requires several assumptions, the most common of which are described below. In the analysis, the existence of corresponding moments for the covariate $\boldsymbol{X}$ is required. The moments are denoted as

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}] &= \overline{\boldsymbol{X}}^*, \text{ and} \\
\mathbb{V}\text{ar}[\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X} - \overline{\boldsymbol{X}}^*)(\boldsymbol{X} - \overline{\boldsymbol{X}}^*)^T] = \boldsymbol{\Sigma}_{\boldsymbol{X}}^*,
\end{aligned}$$

where we assume that the variances $\mathbb{V}\text{ar}[X_i],\ i = 1, \ldots, p-1$ are non-zero. Existence of the fourth central moment of Y is also required, to ensure that the second moment of the variance estimator, corresponding to the observed part of sample, exists. We do not derive the variances of variance estimators in this thesis, which is explained later. However, the assumption is still important. Estimation of variance makes no sense if the variance of an underlying estimator does not exist, because the distribution of the estimator does not degenerate as the sample size goes to infinity.

The assumptions on the existence of central moments exclude some superpopulations. As a consequence some marginal distributions of Y and $\boldsymbol{X}$ are excluded from the analysis. For example, existence of the first and second moments, which are required to make the use of the first sample moment sensible, excludes the Cauchy distribution family. Neither of the two moments exist for Cauchy distributions. Secondly, existence of the fourth moment excludes also some distributions. As an example, there are parameter combinations for Pareto distributions which result in the inexistence of the fourth moment. In practice these moment assumptions are required in many other statistical analysis too. Therefore, we do not consider these to be too restrictive for practicality.

Finally, recall decomposition of $f(Y, \boldsymbol{X})$ in Equation (3.1). It implies, see Appendix A3.4.2 for details, that a conditional mean can be written as

$$g^*(\mathbf{x}) = \mathbb{E}[Y|\boldsymbol{X} = \mathbf{x}] \quad = \quad (1 - p^*_{\mathbf{x}})g^{*obs}(\mathbf{x}) + p^*_{\mathbf{x}}g^{*mis}(\mathbf{x}), \tag{3.2}$$

where $p^*_{\mathbf{x}} = \Pr(R = 0|\mathbf{x})$ and $g^{*obs}(\mathbf{x})$ and $g^{*mis}(\mathbf{x})$ are the conditional means for observed and missing Y values at $\mathbf{x}$. In the MAR or MCAR case $g^{*obs}(\mathbf{x}) = g^{*mis}(\mathbf{x}) = g^*(\mathbf{x})$.

3.3 About evaluation

Recall our definitions for (reordered) true and incomplete realized data sets $\mathbf{d}^{true}$ and $\mathbf{d}^{inc}$. In imputation we replace $\mathbf{d}^{inc}$ with the (randomly) completed data set $\mathbf{D}^{comp}$. In the completed data set the missing values of $Y_j^{comp} \in \mathbf{D}^{comp}$ are given with a model $g(\mathbf{x}|\boldsymbol{\theta})$ as follows

$$Y_j^{comp} = \begin{cases} Y_j^{obs}, & \text{if } r_j = 1, \quad (Y_j \text{ is observed}) \\ Y_j^{imp} = g\big(\boldsymbol{X}_j^{mis}|\boldsymbol{\theta}\big) + \hat{\epsilon}\big(\boldsymbol{X}_j^{mis}\big), & \text{when } r_j = 0 \quad (Y_j \text{ is missing}). \end{cases}$$

The evaluation of imputation performance is now a study of distributional and pointwise and unit level properties of an imputation method.

All our evaluations are conditionalized. By this we mean that there is some set of background assumptions, denoted as $\mathcal{Q}$, under which the analysis is done. Most commonly we assume that the number of observations is fixed to n, while the number of missing values N^{mis} is random. The value of N^{mis} is ensured, via technical assumptions, to be such that all the computed estimates do exist. Details of these conditionalizations are given in Section 3.5.

The aim of the current evaluation is to answer questions about imputation performance. We like to know

i) if an imputation method is good enough in general,

ii) under what conditions it is good, and

iii) how different methods compare against each other?

None of the questions can be answered directly, but we can find partial explanations by developing measures of imputation performance. These measures are related to true data, $\mathbf{D}^{true}$, and the underlying superpopulation of Y as follows

1) Distributional (aggregate level) measures are related to first and second moments of the distribution of Y, namely

   - expectation $\mu^* = \mathbb{E}[Y]$, and
   - variance $\tau^* = \mathbb{V}\mathrm{ar}[Y]$.

2) Prediction measures are related to $Y|\mathbf{x}$. On the level of sample $\mathbf{D}^{true}$ this is related to the unit level predictions of Y_j given $\mathbf{x}_j$.

In other words, we use both distributional and unit level information in our evaluations.

In relation to actual imputation, the above measures are computed from the completed data set $\mathbf{d}^{comp}$, which contains n observations of the variable Y^{comp}. This implies that our evaluation must be done by studying the sample estimates of $y_j^{comp} \in \mathbf{d}^{comp}$:

$$\mu = \frac{1}{\mathsf{n}} \sum_{j=1}^{\mathsf{n}} y_j^{comp}$$

$$\tau = \frac{1}{\mathsf{n}-1} \sum_{j=1}^{\mathsf{n}} \left(y_j^{comp} - \mu\right)^2, \text{ and}$$

as well as the direct values of y_j^{comp} on the unit level measures. With respect to these measures, the aim of the imputation is to get estimates μ, τ, and y_j^{comp} as close to μ^*, τ^*, and y_j as possible. But closeness itself does not tell us what the imputation performance is, because the differences between μ and μ^* are affected by several reasons

a) estimation errors when computing μ from finite sample

b) imputation errors caused by the model and strategy

c) conditions on data and evaluation measures.

Since we are mainly interested in case b), an extra caution must be taken when analyzing the results of our studies. Namely, because a) and c) can dominate the results, which diminishes the role of imputation methodology.

3.4 Evaluation statistics

The actual evaluation statistics that is used in this thesis is based on following measures

1) Biases and variances of the first moment estimator of Y

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}] &= \mathbb{E}[\hat{\mu}^{comp} - \mu^*|\mathcal{Q}]\\ \mathbb{V}\text{ar}[\hat{\mu}^{comp}|\mathcal{Q}] &= \mathbb{E}[(\hat{\mu}^{comp} - \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}])^2|\mathcal{Q}]\end{aligned}$$

2) Bias of the second moment estimator of Y

$$\mathbb{B}\text{ias}[\hat{\tau}^{comp}|\mathcal{Q}] = \mathbb{E}[\hat{\tau}^{comp} - \tau^*|\mathcal{Q}]$$

3) Expected mean squared error of predictions for Y_j

$$\mathbb{E}[\hat{mse}(Y^{comp})|\mathcal{Q}] = \mathbb{E}\left[\frac{1}{\mathsf{N}^{mis}}\sum_{j=1}^{\mathsf{n}}\left(Y_j^{comp} - Y_j\right)^2|\mathcal{Q}\right],$$

One should note that the divider is random quantity N^{mis} and not fixed n. This makes sense as we want to measure the average error per prediction (observed Y values are not predicted). However, we may have problems with interpretations.

In addition, we sometimes study other properties as well, for example the Kolmogorov-Smirnov distance

4) $\mathbb{E}[\hat{K}_\infty|\mathcal{Q}] = \mathbb{E}\left[\sup_y |F_{Y^{mis}}(y) - \hat{F}_{Y^{imp}}(y)|\Big|\mathcal{Q}\right],$

where $F_{Y^{mis}}$ is the cumulative distribution function of Y^{mis}, and $\hat{F}_{Y^{imp}}$ is an estimator for the empirical cumulative distribution function of Y^{imp} which is based on data $\mathbf{D}_Y^{imp}$.

Due to mathematical difficulty, we omit analytical results that concern $\mathbb{V}\text{ar}[\hat{\tau}^{comp}|\mathcal{Q}]$. Other researchers, such as Zhao[115] and Santos[90], have also done so. To explain why it is feasible to do that, Santos lists three reasons do [90]:

- "variances and mean squared errors of covariance estimator (and for complex statistics in general) are complicated and tedious to derive",
- "results are complex and difficult to interpret", and
- "for large samples variances of most statistics are small".

The above evaluation measures 1), 2) and 3) can often be rewritten in another form, depending on conditionalizations $\mathcal{Q}$. In our formulas population mean μ^* is usually not dependent on $\mathcal{Q}$, and thus

$$\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}] = \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}] - \mu^*.$$

Another thing to note is that the mean squared prediction errors are sometimes at some given point $\boldsymbol{X}^{mis} = \mathbf{x}_0$ for a missing Y value, instead of data set $\mathbf{d}$. This quantity is defined as

$$5)\ \ \mathrm{mse}(Y^{imp}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}) = \mathbb{E}\Big[(Y^{imp}_{|\mathbf{x}_0} - Y^{mis}_{|\mathbf{x}_0})^2 | \boldsymbol{X}^{mis} = \mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}\Big],$$

where prediction $Y^{imp}_{|\mathbf{x}_0} = \hat{g}(\mathbf{x}_0|\hat{\boldsymbol{\theta}}) + \hat{\epsilon}(\mathbf{x}_0)$. Integration is done over the joint distribution of target and prediction given point $\mathbf{x}_0$, the number of missing data values n^{mis}, and sample size n. The reason for conditionalizing by the number of the missing data values is to fix the training sample size and thus make the results from the literature easy to apply. Error measure 5) is related to error measure 3), which becomes apparent in Section 3.8.3. However, it is easier to compute and interpret than $\mathbb{E}[\hat{mse}(Y^{comp})|\mathcal{Q}]$ as there is less randomness.

3.4.1 About interpretations

When doing interpretations about evaluation statistics, one should keep in mind that none of the measures alone can tell us what the performance of a particular imputation method is. Some of the reasons for this are:

- Mean imputation, which reduces $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}]$, and therefore improves μ estimate, easily increases the bias of $\hat{\tau}^{comp}$.
- Some sources of variation in $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}]$ are due to a small sample size rather than the imputation method. It might be a good idea to compare the results to the natural estimation variance $\mathbb{V}\mathrm{ar}[\hat{\mu}]$. Thus instead of population mean μ^*, we should compare the results against estimate $\hat{\mu} = \frac{1}{\mathsf{n}}\sum_{j=1}^{\mathsf{n}} Y_j$.
- Minimization of prediction errors like $\hat{mse}(Y^{comp})$ tends to increase errors in distributional measures, and vice versa.

There are also different levels of interpretations. Most notably we include the results for finite versions of conditionalized estimators with respect to the number of observations n. Asymptotic versions are included when mathematically feasible. Usually asymptotic versions are simpler to read and interpret, but they lack information of the imputation behaviour with a small sample size. In addition, asymptotic results are often approximated using the Taylor series, which may lead to problems in interpretations. A further aid for interpretations is based on conditionalizations which are described next.

3.5 Conditionalizations: $\mathcal{Q}_1$, $\mathcal{Q}_2$, $\mathcal{Q}_3$ and $\mathcal{Q}_P$

There are two reasons for conditionalizations: interpretation and technical. Technical reasons are due to complications in mathematical analyses. In most occasions these represent purely pathological cases of no real interest for the reader. In conditionalization we focus our interest on more common situations. A typical example is the variance of an estimator in a cell that has only one observation. Surely this is not interesting, and we limit our analyses for cells that have always a reasonable number of observations for the current study.

Conditionalizations that aim for better interpretations are more interesting. To summarise, we have three main levels of these:

1) $\mathcal{Q}_1 = \{\mathsf{n}\}$, where only the number of observations is fixed. Thus data $\mathbf{D}$, with n observations, is sampled from the superpopulation, and number of missing data values N^{mis} is sampled from its distribution. The imputation model is then estimated from the data, and imputation is done using the newly obtained estimators for Y^{imp}. Sampling N^{mis} is connected to drawing the response pattern $\boldsymbol{R}$ from its distribution.

2) $\mathcal{Q}_2 = \{\mathsf{n}, \mathbf{d}^{train}_{\mathsf{n}^{obs}}, g(\mathbf{x}|\boldsymbol{\theta})\}$, where the sample size, training data, and imputation model remain the same but the test data still varies. One should observe that conditionalising by sample size and training data fixes also the number of missing data values.

3) $\mathcal{Q}_3 = \{\mathbf{d}^{train}_{\mathsf{n}^{obs}}, \mathbf{d}^{test}_{\mathsf{n}^{mis}}, g(\mathbf{x}|\boldsymbol{\theta})\}$, where everything is fixed except the response indicators and randomness of imputation, which is sampled from $f_{\hat{\epsilon}(\mathbf{x})}$. One should note that conditionalization by train and test data sets fixes the number of observations n and the number of missing data values $\mathsf{N}^{mis} = \mathsf{n}^{mis}$. Further, the sum of response indicators is constant but the indicators are random. However, the randomness of the indicators has no effect on the statistics which we compute.

The role of the above conditionalizations is to limit the sources of variation in such a way that on the last level, $\mathcal{Q}_3$, we know that the variation must be due to unexplained randomness $\hat{\epsilon}(\mathbf{x})$. At the second level $\mathcal{Q}_2$ the variation is due to $\hat{\epsilon}(\mathbf{x})$ and partially to the observed part of data. Finally, at the first level randomness is due to data, to the number of missing-data values (missing-data pattern), to the model, and to unexplained randomness. One should note that at levels $\mathcal{Q}_2$ and $\mathcal{Q}_3$ quantities $\hat{g}(\mathbf{x})$ and $\hat{\epsilon}(\mathbf{x})$ are independent, which allows us to decompose some of the computed properties in a useful manner.

As an example, the connections for the conditionalized expectations of the first moment are as follows. The expectation of $\hat{\mu}^{comp}$ at the third level can be written as

$$\begin{aligned}\mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_3] &= \mathbb{E}[\hat{\mu}^{comp}|\mathbf{d}^{train}, \mathbf{d}^{test}, g(\mathbf{x}|\boldsymbol{\theta})] \\ &= \int \mu f_{\hat{\mu}^{comp}|\mathcal{Q}_3}\Big(\mu|\mathbf{d}^{train}, \mathbf{d}^{test}, g(\mathbf{x}|\boldsymbol{\theta})\Big) d\mu.\end{aligned}$$

One should notice that at the third level the randomness of $\hat{\mu}^{comp}$ comes from the noise terms which are described via noise distribution $f_{\hat{\epsilon}(\mathbf{x})}(e|\mathbf{x})$. The connection between the second and the third level can be derived as

$$\begin{aligned}\mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_2] &= \mathbb{E}[\hat{\mu}^{comp}|\mathsf{n},\mathbf{d}^{train},g(\mathbf{x}|\boldsymbol{\theta})]\\ &= \int\left[\int \mu f_{\hat{\mu}^{comp}|\mathcal{Q}_3}\Big(\mu|\mathbf{d}^{train},\mathbf{d}^{test},g(\mathbf{x}|\boldsymbol{\theta})\Big)d\mu\right]f(\mathbf{d}^{test})d\mathbf{d}^{test}\\ &= \int \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_3]f(\mathbf{d}^{test})d\mathbf{d}^{test}.\end{aligned}$$

Hence the link between the third and the second level is simple. One just needs to integrate the third level result with respect to the distribution of the partially observed part of data. Similarly one can derive the connection between the first and the second level

$$\begin{aligned}\mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_1] &= \mathbb{E}[\hat{\mu}^{comp}|\mathsf{n}]\\ &= \int \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_2]f\Big(\mathsf{n}^{mis},\mathbf{d}^{train},g(\mathbf{x}|\boldsymbol{\theta})\Big)d\Big(\mathsf{n}^{mis},\mathbf{d}^{train},g(\mathbf{x}|\boldsymbol{\theta})\Big).\end{aligned}$$

The benefit of these derivations becomes apparent when the first level equations become difficult to interpret. Such situations are typically seen with the evaluation of cell imputation (Chapter 6). Via conditionalizations we get at least some understanding about the role and magnitude of noise terms, training data, etc.

3.5.1 Technical conditionalization $\mathcal{Q}_P$

As noted earlier, sometimes our evaluation is conditionalized at point $\boldsymbol{X}^{mis} = \mathbf{x}_0$. Similarly, some imputation models like those that are based on kernel and nearest neighbour are easiest to study at a given point $\mathbf{x}_0$. In the case of imputation we may select $\mathbf{x}_0$ to be $\mathbf{x}^{mis} \in \mathbf{d}_{\boldsymbol{X}}^{mis}$, which leads us to conditionalization $\mathcal{Q}_P = \{\mathsf{n}, \mathbf{d}_{\boldsymbol{X}}^{mis}\}$. Remark that conditionalising by $\mathbf{d}_{\boldsymbol{X}}^{mis}$ fixes the number of missing data values $\mathsf{N}^{mis} = \mathsf{n}^{mis}$.

In other words, training data $\mathbf{D}^{train}$, and thus the imputation model, and $\mathbf{D}_Y^{mis}$ are allowed to be random, but the statistics is computed for a given point $\mathbf{x}^{mis}$ and for a fixed sized training data. Although this leads to some problems in interpretations, it has the benefit that many results from the literature are directly applicable, especially concerning nonparametric methods [72]. Remark that the prediction target is $Y^{mis} \in \mathbf{D}_Y^{mis}$.

Although conditionalizations $\mathcal{Q}_P$ are not comparable with $\mathcal{Q}_3$ and $\mathcal{Q}_2$, we can compute $\mathcal{Q}_1$ by integrations like

$$\mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_1] = \int \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_P]f(\mathsf{n}^{mis},\mathbf{d}_{\boldsymbol{X}}^{mis})d(\mathsf{n}^{mis},\mathbf{d}_{\boldsymbol{X}}^{mis}).$$

3.6 Presentation of results via theorems, approximations, proofs, justifications, and simulations and experiments

In this thesis we want results which have practical interpretation. This means that the results for a finite sample size n have to be derived. In the literature asymptotic results ($\mathsf{n} \to \infty$) are often given. However, this is not realistic in practice because the sample size n is finite or often quite small. Derivation of results for a finite n leads to problems because of mathematical difficulty which is partly due to random number of observations and technical conditionalisations. We have dealt with this issue by decompositions (see Section 3.7 for example) and approximations. Rigorous treatment of random number of observations is quite complicated even for some specific estimation problem with large n (see for example [97]). Decompositions allow easier interpretation, whereas approximations make the derivation of results more feasible. Empirical evaluation is included to verify that our approach works.

Next we introduce the terminology used for approximations and describe how it corresponds to the terminology for exact results. The terminology for approximations is used to distinguish approximative results, and their derivations, from exact results.

3.6.1 Theorems and proofs vs. approximations and justifications

The correspondence between the terminology used in exact results and approximative results is shown in Table 3.4 where approximation is similar to theorem, consequence to corollary, intermediate step to lemma, and justification to proof. The difference between exact and approximative results is that derivation of approximative results is mathematically less exact. However, some approximative results (at least the asymptotic ones) might be turned into exact results, provided a careful analysis of the approximation error terms were done.

Exact results	Approximative results
Theorem	Approximation
Corollary	Consequence
Lemma	Intermediate step
Proof	Justication

Table 3.4: Corresponce between the terminology for exact results and that of approximative results.

3.6.2 About some error terms

In some of the given approximative results there are higher order terms which are inversely proportional to the sample size squared. One may wonder whether these terms are important. There are situations in which such terms are significant in practice. For example the higher order term may have considerably higher constant (bias) than the lower order term which is inversely proportional to the sample size (recall that order notation supresses constants). As a result the higher order term may be important, at least for small sample sizes. An example of such behavior is given in Appendix A4.4.

3.6.3 About simulations and experiments

Another viewpoint to methodological evaluation is obtained by simulations and experiments. When easily done, we shall compare experiments with theory. Otherwise, simulations and experiments allow us to see how methods behave in practice. When presenting results, especially in Chapters 7, 8 and 9, the best results for a given statistic are underlined, and some that are close to it are written in boldface. There was also an attempt to see if the best results are significantly different than others, and it seems that this is the case, probably because of the large number of repetitions in our simulations.

3.7 Basic decompositions for moments

Our methodology is strongly based on decompositions, where superpopulation is divided between observed and missing information. The lemma below gives the decompositions of the first two moments of a superpopulation for Y. The lemma follows directly from the previous assumptions ii) and iii).

Lemma 3.1 *Decomposition of superpopulation moments.*
For $Y_j \in \mathbf{D}$,

a) the first moment of Y_j can be decomposed as

$$\mu^* = (1-p^*)\mu^{*obs} + p^*\mu^{*mis}, \tag{3.3}$$

where p^* is the probability of non-response, μ^{*obs} is the expectation of observed $Y_j^{obs} \in \mathbf{D}_Y^{obs}$ and μ^{*mis} is the expectation of $Y_j^{mis} \in \mathbf{D}_Y^{mis}$.

b) the variance of $Y_j \in \mathbf{D}$ may be decomposed as

$$\begin{aligned} \tau^* &= \underbrace{(1-p^*)\tau^{*obs} + p^*\tau^{*mis}}_{\text{within variance}} \\ &\quad + \underbrace{p^*(1-p^*)(\mu^{*mis} - \mu^{*obs})^2}_{\text{between variance}}, \end{aligned} \tag{3.4}$$

where τ^{*obs} is the variance of the observed Y_j^{obs} values and τ^{*mis} is the variance of the unobserved Y_j^{mis} values.

There are two clusters in the previous decompositions: one corresponding to response and other to non-response. Therefore the decomposition of variance (Equation 3.4) consists of the variance within clusters (the first two terms) and the variance between clusters (the last term). In case of two clusters the variance between clusters can be written as one term as shown previously. A partial proof of Lemma 3.1 is given next.

Proof 3.2 *Partial proof of Lemma 3.1.*

a)

$$\begin{aligned}
\mu^* &= \mathbb{E}[Y_j] = \mathbb{E}\big[\mathbb{E}[Y_j|R_j]\big] \\
&= \Pr\big(R_j = 1\big)\mathbb{E}[Y_j|R_j = 1] + \Pr\big(R_j = 0\big)\mathbb{E}[Y_j|R_j = 0] \\
&= \Pr(R_j = 1)\mu^{*obs} + \Pr\big(R_j = 0\big)\mu^{*mis} \\
&= (1 - p^*)\mu^{*obs} + p^*\mu^{*mis},
\end{aligned}$$

where p^* is the probability for non-response, μ^{*obs} is the expectation of the observed $Y_j^{obs} \in \mathbf{D}_Y^{obs}$, R_j is the response indicator for the random true data set $\mathbf{D}$, and μ^{*mis} is the expectation of $Y_j^{mis} \in \mathbf{D}_Y^{mis}$. This decomposition is based on the fact that $f_Y(y) = (1 - p^*)f_{Y^{obs}}(y) + p^* f_{Y^{mis}}(y)$.

b) using similar technique as in i) (see proof in Appendix A3.2.2 for details), we can decompose the variance of $Y_j \in \mathbf{D}$ as

$$\begin{aligned}
\tau^* &= \mathbb{V}\mathrm{ar}[Y_j] \\
&= \mathbb{E}\big[\mathbb{V}\mathrm{ar}[Y_j|R_j]\big] + \mathbb{V}\mathrm{ar}\big[\mathbb{E}[Y_j|R_j]\big] \\
&= \underbrace{(1 - p^*)\tau^{*obs} + p^*\tau^{*mis}}_{\text{within variance}} + \underbrace{p^*(1 - p^*)(\mu^{*mis} - \mu^{*obs})^2}_{\text{between variance}},
\end{aligned}$$

where R_j is the response indicator for random true data set $\mathbf{D}$, τ^{*obs} is the variance of the observed Y_j^{obs} values and τ^{*mis} is the variance of the unobserved Y_j^{mis} values.

□

Decompositions similar to the ones above can be written for imputed data set $\mathbf{D}^{comp}$ also. Thus trivially

Lemma 3.3 *Decomposition of the estimators for the first two moments of Y^{comp}.*

a)

$$\hat{\mu}^{comp} = \frac{1}{n}\Big(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathsf{N}^{mis}\hat{\mu}^{imp}\Big), \text{ and} \tag{3.5}$$

$$\hat{\tau}^{comp} = \underbrace{\frac{\mathsf{N}^{obs}-1}{n-1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{n-1}\hat{\tau}^{imp}}_{\text{within variance}} + \underbrace{\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{n(n-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2}_{\text{between variance}}, \tag{3.6}$$

where

$$\hat{\mu}^{obs} = \frac{1}{\mathsf{N}^{obs}} \sum_{Y^{obs}\in\mathbf{D}_Y^{obs}} Y^{obs},$$

$$\hat{\mu}^{imp} = \frac{1}{\mathsf{N}^{mis}} \sum_{Y^{imp}\in\mathbf{D}_Y^{imp}} Y^{imp},$$

$$\hat{\tau}^{obs} = \frac{1}{\mathsf{N}^{obs}-1} \sum_{Y^{obs}\in\mathbf{D}_Y^{obs}} (Y^{obs}-\hat{\mu}^{obs})^2, \text{ and}$$

$$\hat{\tau}^{imp} = \frac{1}{\mathsf{N}^{mis}-1} \sum_{Y^{imp}\in\mathbf{D}_Y^{imp}} (Y^{imp}-\hat{\mu}^{imp})^2.$$

Proof: *Trivial for $\hat{\mu}^{comp}$. A bit more complicated for $\hat{\tau}^{comp}$ (see Appendix A3.2.1 for details).* □

As $\mathsf{n} \to \infty$ we get the following approximations

Approximation 3.4 *Approximations for asymptotics of $\hat{\mu}^{comp}$ and $\hat{\tau}^{comp}$. Limits for expectations of $\hat{\mu}^{comp}$ and $\hat{\tau}^{comp}$ can be approximated as*

a)

$$\lim_{n\to\infty} \mathbb{E}[\hat{\mu}^{comp}|n] \approx (1-p^*)\mu^{*obs} + p^*\mu^{*imp},$$

where p^* is the probability of missingness and μ^{*imp} is the limiting mean of imputed values $Y_j^{imp} \in \mathbf{D}_Y^{imp}$.

b)

$$\lim_{n\to\infty} \mathbb{E}[\hat{\tau}^{comp}|n] \approx \underbrace{(1-p^*)\tau^{*obs} + p^*\tau^{*imp}}_{\text{within variance}} + \underbrace{p^*(1-p^*)(\mu^{*imp}-\mu^{*obs})^2}_{\text{between variance}},$$

where p^*, τ^{*obs}, and μ^{*obs} are as before and τ^{*imp} is the limiting variance of the imputed values.

The above approximation is derived by i) applying the first order Taylor approximation to both quantities and ii) taking the limit and assuming that the approximation error (corresponding to the Taylor remainder term) goes to zero as $\mathsf{n} \to \infty$. The details are given in Appendix A3.3.

Although everything seems quite trivial now, the reader should note that the solution requires us to

i) state what μ^{*imp} and τ^{*imp} are. This may be difficult if the imputation mechanism is complicated, and

ii) compute the solutions with random number of N^{mis} missing values.

3.8 Basic decompositions for evaluation statistics

In order to say how a certain imputation method performs and why, we shall decompose the final evaluation statistics so that the causes of behaviour are easier to read. To do this we apply our previouly introduced ideas. For simplicity, in the current context, we omit the most rigorous notations of conditions that are required to ensure that the above estimates exist. When required, such conditions are discussed in the following chapters and in the appendixes.

3.8.1 Bias of 1^{st} moment

Equations (3.3) and (3.5) allow us to decompose the bias of the first moment $\hat{\mu}^{comp}$ in the following form:

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}] &= \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}] - \mu^* \qquad (3.7)\\
&= \underbrace{\underbrace{\mathbb{E}\left[\Big(1 - \frac{\mathsf{N}^{mis}}{\mathsf{n}}\Big)\hat{\mu}^{obs}|\mathcal{Q}\right]}_{\text{Sample estimate}} - \underbrace{(1-p^*)\mu^{*obs}}_{\text{True of observed}}}_{A}\\
&\quad \underbrace{+\underbrace{\mathbb{E}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\hat{\mu}^{imp}|\mathcal{Q}\right]}_{\text{Impute estimate}} - \underbrace{p^*\mu^{*mis}}_{\text{True of missing}}}_{B} \quad .
\end{aligned}$$

Term A in the decomposition is not really about imputation methodology since it is due to sampled estimates of the observed part of the data. For this reason we shall refer to it as sample bias. The second term B is more interesting as it measures the role of the imputed data with respect to true data. Thus it tells us about the imputation methodology. Further simplification is possible when the conditionalizations are specified.

Decomposition is simplified and most usable when $\mathcal{Q} = \mathcal{Q}_1$:

$$\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}_1] = \mathbb{E}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\hat{\mu}^{imp}|\mathcal{Q}_1\right] - p^*\mu^{*mis} + O(\mathsf{n}^{-1}),$$

where term $O(\mathsf{n}^{-1})$ is due to technical reasons as explained in Appendix A3.3. However, one should note that expectation may be difficult to compute or the result may be difficult to interpret.

By applying the first order Taylor approximation and taking limit $\mathsf{n} \to \infty$ (assuming the approximation error goes to zero) we get

$$\begin{aligned}
\lim_{\mathsf{n}\to\infty} \mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}_1] &\approx p^*\Big(\lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\mu}^{imp}|\mathcal{Q}_1] - \mu^{*mis}\Big)\\
&= p^*(\mu^{*imp} - \mu^{*mis}).
\end{aligned}$$

3.8.2 Bias of the 2^{nd} moment

Using similar techniques as before, the bias of the second moment $\hat{\tau}^{comp}$ may be decomposed using (3.4) and (3.6) as:

$$\begin{aligned}
& \mathbb{B}\text{ias}[\hat{\tau}^{comp}|\mathcal{Q}] = \mathbb{E}[\hat{\tau}^{comp}|\mathcal{Q}] - \tau^* = A + B + C \qquad (3.8) \\
= \quad & \underbrace{\mathbb{E}\Big[(1 - \frac{\mathsf{N}^{mis}}{\mathsf{n}-1})\hat{\tau}^{obs}|\mathcal{Q}\Big] - (1-p^*)\tau^{*obs}}_{A} + \underbrace{\mathbb{E}\Big[\frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}|\mathcal{Q}\Big] - p^*\tau^{*mis}}_{B} \\
+ \quad & \underbrace{\mathbb{E}\Big[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs} - \hat{\mu}^{imp})^2|\mathcal{Q}\Big] - p^*(1-p^*)(\mu^{*obs} - \mu^{*mis})^2}_{C}.
\end{aligned}$$

As before term A is due to the sampling estimates of τ^{*obs} from the observed data. The role of the imputation methodology is again most directly seen in term B, where $\hat{\tau}^{imp}$ is compared with true τ^{*mis} of the imputed part of the data. The additional term C is due to cross terms that become important when missingness is not completely random or if the estimated imputation model does not fit to the data well.

Again the above decomposition is simplified when $\mathcal{Q} = \mathcal{Q}_1$:

$$\mathbb{B}\text{ias}[\hat{\tau}^{comp}|\mathcal{Q}_1] = B + C + O(\mathsf{n}^{-1}).$$

Yet the result is still somewhat difficult to interpret with finite n, because the estimation variances are mixed into the terms. A more interpretable decomposition based on the first order Taylor approximation, when mathematically feasible (the required limits may be difficult to compute), may be written at the first level as

$$\begin{aligned}
& \lim_{\mathsf{n}\to\infty} \mathbb{B}\text{ias}[\hat{\tau}^{comp}|\mathcal{Q}_1] \approx p^*(\lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\tau}^{imp}|\mathcal{Q}_1] - \tau^{*mis}) \\
+ \quad & p^*(1-p^*)\Big(\lim_{\mathsf{n}\to\infty} \mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp})^2|\mathcal{Q}_1] - (\mu^{*obs} - \mu^{*mis})^2\Big) \\
= \quad & \underbrace{p^*(\tau^{*imp} - \tau^{*mis})}_{A} + \underbrace{p^*(1-p^*)\Big((\mu^{*obs} - \mu^{*imp})^2 - (\mu^{*obs} - \mu^{*mis})^2\Big)}_{B},
\end{aligned}$$

where term A is the asymptotic difference between the variances of imputed and missing Y values. The term measures how well an imputation method has preserved the second moment of missing Y values. If the imputation method provides (asymptotically) an unbiased second moment estimator then A is zero. Term B measures how well the first moment is preserved by the imputation method. If the imputation method yields (asymptotically) an unbiased first moment estimator then B is zero.

3.8.3 Bias-variance decompositions

In statistical literature it is well known that mean squared error can always be decomposed into (squared) bias, variance and noise components [36]. Now we shall apply this fact in the context of imputation. We begin from a predictive model of missing Y at some point $\boldsymbol{X}^{mis} = \mathbf{x}_0$, where we assume that the target follows model

$$Y_{|\mathbf{x}_0} = g^{*mis}(\mathbf{x}_0) + \epsilon^{mis}_{\mathbf{x}_0}$$

and our predictive model at the same point is

$$\hat{Y}_{|\mathbf{x}_0} = \hat{g}(\mathbf{x}_0) + \hat{\epsilon}_{\mathbf{x}_0},$$

where $\mathbb{E}[\epsilon^{mis}_{\mathbf{x}_0}] = 0$ and $\mathbb{E}[\hat{\epsilon}_{\mathbf{x}_0}] = 0$.

For clarity of formulas we omit the somewhat rigorous notation $Y^{mis}_{|\mathbf{x}_0}$ and $Y^{imp}_{|\mathbf{x}_0}$ here, and thus use $Y_{|\mathbf{x}_0}$ and $\hat{Y}_{|\mathbf{x}_0}$.

The mean squared error at the test point is defined as an expectation

$$\mathrm{mse}(\hat{Y}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}) = \mathbb{E}\Big[(\hat{Y}_{|\mathbf{x}_0} - Y_{|\mathbf{x}_0})^2 | \mathsf{n}^{mis}, \mathsf{n}\Big],$$

where integration is done with respect to $\hat{Y}$ and Y given $\mathbf{x}_0$, number of missing data values n^{mis}, and sample size n. Conditionalising by n^{mis} and n ensures that the size of the training data set, which is used to form the imputation model, is fixed. This allows us to easily apply the results from the literature, especially for nonparametric methods. Mean squared error can be decomposed as follows:

Theorem 3.5 *Decomposition for mean squared error* $\mathrm{mse}(\hat{Y}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n})$.

$$\begin{aligned}\mathrm{mse}(\hat{Y}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}) &= \Big(\underbrace{\mathbb{E}[\hat{g}(\mathbf{x}_0)|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}] - g^{*mis}(\mathbf{x}_0)}_{\text{imputation bias at } \mathrm{x}_0}\Big)^2 + \underbrace{\mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}}]}_{\text{imputation variance at } \mathrm{x}_0} \\ &\quad + \underbrace{\mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}_0}]}_{\mathrm{v}^{*\mathrm{mis}}(\mathbf{x}_0),\ \text{target noise at } \mathrm{x}_0}.\end{aligned}$$

Proof:

$$\begin{aligned}\mathrm{mse}(\hat{Y}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}) &= \mathbb{E}_{\hat{Y},Y|\mathbf{x}_0,\mathsf{n}^{mis},\mathsf{n}}[\hat{Y}^2_{|\mathbf{x}_0} - 2\hat{Y}_{|\mathbf{x}_0}Y_{|\mathbf{x}} + Y^2_{|\mathbf{x}_0}] \qquad (3.9) \\ &= \mathbb{E}[\hat{Y}^2_{|\mathbf{x}_0} + \mathbb{E}[\hat{Y}_{|\mathbf{x}_0}] - \mathbb{E}[\hat{Y}_{|\mathbf{x}_0}] + 2\mathbb{E}[\hat{Y}_{|\mathbf{x}}]\hat{Y}_{|\mathbf{x}_0} - 2\mathbb{E}[\hat{Y}_{|\mathbf{x}_0}]\hat{Y}_{|\mathbf{x}_0}] \\ &\quad +\mathbb{E}[Y^2_{|\mathbf{x}_0} + \mathbb{E}[Y_{|\mathbf{x}_0}] - \mathbb{E}[Y_{|\mathbf{x}_0}] + 2\mathbb{E}[Y_{|\mathbf{x}}]Y_{|\mathbf{x}_0} - 2\mathbb{E}[Y_{|\mathbf{x}_0}]Y_{|\mathbf{x}_0}] \\ &\quad -2\mathbb{E}[\hat{Y}_{|\mathbf{x}_0}Y_{|\mathbf{x}_0}] \\ &= \mathbb{E}[(\mathbb{E}[\hat{Y}_{|\mathbf{x}_0}] - \hat{Y}_{|\mathbf{x}_0})^2] + \mathbb{E}[(\mathbb{E}[Y_{|\mathbf{x}}] - Y_{|\mathbf{x}_0})^2] \\ &\quad +\mathbb{E}[\hat{Y}_{|\mathbf{x}_0}]^2 - 2\mathbb{E}[\hat{Y}_{|\mathbf{x}_0}Y_{|\mathbf{x}_0}] + \mathbb{E}[Y_{|\mathbf{x}_0}]^2 \qquad || \quad * \\ &= \mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0}] + \mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}_0}] + (\mathbb{E}[\hat{Y}_{|\mathbf{x}_0}] - \mathbb{E}[Y_{|\mathbf{x}_0}])^2 - 2\mathbb{C}\mathrm{ov}[\hat{Y}_{|\mathbf{x}_0}, Y_{|\mathbf{x}_0}] \\ &= (\mathbb{E}[\hat{Y}_{|\mathbf{x}_0}] - \mathbb{E}[Y_{|\mathbf{x}_0}])^2 + \mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0}] + \mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}_0}] \\ &= \Big(\underbrace{\mathbb{E}[\hat{g}(\mathbf{x}_0)|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}] - g^{*mis}(\mathbf{x}_0)}_{\text{imputation bias at } \mathrm{x}_0}\Big)^2 \\ &\quad + \underbrace{\mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}}]}_{\text{imputation variance at } \mathrm{x}_0} \\ &\quad + \underbrace{\mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}_0}]}_{\mathrm{v}^{*\mathrm{mis}}(\mathbf{x}_0),\ \text{target noise at } \mathrm{x}_0},\end{aligned}$$

*where at * we have applied the fact that* $\mathbb{E}[\hat{Y}Y] = \mathbb{E}[\hat{Y}]\mathbb{E}[Y] + \mathbb{C}\mathrm{ov}[\hat{Y}, Y]$. *The covariance term* $\mathbb{C}\mathrm{ov}[\hat{Y}_{|\mathbf{x}_0}, Y_{|\mathbf{x}_0}]$ *is zero because* $\hat{Y}$ *and* Y *are conditionally independent given* $\mathbf{x}_0$. □

Further, one should note that imputation variance can be decomposed as

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0,\mathsf{n}^{mis},\mathsf{n}}] &= \mathbb{V}\mathrm{ar}[\hat{g}(\mathbf{x}_0) + \hat{\epsilon}_{\mathbf{x}_0}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}] \qquad (3.10)\\ &= \underbrace{\mathbb{V}\mathrm{ar}[\hat{g}(\mathbf{x}_0)|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}]}_{\text{variance of conditional mean estimate}} + \underbrace{\mathbb{V}\mathrm{ar}[\hat{\epsilon}_{\mathbf{x}_0}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}]}_{\text{imputation noise, } \hat{\mathrm{v}}(\mathbf{x}_0)}\\ &\quad + \underbrace{2\mathbb{C}\mathrm{ov}[\hat{g}(\mathbf{x}_0), \hat{\epsilon}_{\mathbf{x}_0}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}]}_{\text{cross term}}.\end{aligned}$$

One should note that the cross term above is often zero. The reason for this is that the expectation of the estimated noise term is typically zero. Further, the conditional mean estimate and the noise term often are conditionally independent given training data (recall the chain rule of covariance).

As before, we can conditionalize the result for better intepretation. For example, if we want to study the role of simulated noise $\hat{\epsilon}_{|\mathbf{x}}$ we should fix the model $\hat{g}(\mathbf{x})$. Then $\mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0,\mathsf{n}^{mis},\mathsf{n}}] = \mathbb{V}\mathrm{ar}[\hat{\epsilon}_{|\mathbf{x}_0,\mathsf{n}^{mis},\mathsf{n}}]$, thus with given model $g(\mathbf{x})$ we get

$$\begin{aligned}\mathrm{mse}\big(\hat{Y}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}, g(\mathbf{x})\big) &= \big(\underbrace{g(\mathbf{x}_0) - g^{*mis}(\mathbf{x}_0)}_{\text{model bias}}\big)^2 + \underbrace{\mathbb{V}\mathrm{ar}[\hat{\epsilon}_{|\mathbf{x}_0,\mathsf{n}^{mis},\mathsf{n}}]}_{\hat{\mathrm{v}}(\mathbf{x}_0)}\\ &\quad + \underbrace{\mathbb{V}\mathrm{ar}[\epsilon^{mis}_{|\mathbf{x}_0}]}_{\mathrm{v}^{*\mathrm{mis}}(\mathbf{x}_0)}. \qquad (3.11)\end{aligned}$$

Decomposition (3.11) reveals us that prediction error can be minimized by decreasing simulated noise ($\mathbb{V}\mathrm{ar}[\hat{\epsilon}_{|\mathbf{x},\mathsf{n}^{mis},\mathsf{n}}]$). On the other hand, by decreasing variance we are likely to introduce bias to the 2^{nd} moment of the computed data as shown in Section 3.8.2.

Bias-variance decompositions can also be written on a population level as shown in the next theorem.

Theorem 3.6 *Bias-variance decomposition at population level.*

$$
\begin{aligned}
\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] = {} & \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{variability of conditional mean estimate}} \\
+ \quad & \underbrace{(\mu_{\mathsf{n}}^{*imp}-\mu^{*mis})^2}_{\text{global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
+ \quad & \underbrace{2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]-\mu_{\mathsf{n}}^{*imp}\right)\left(\mu_{\mathsf{n}}^{*imp}-g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{cross term}} \\
+ \quad & \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{expected variance of conditional mean estimate}} + \underbrace{v_{\mathsf{n}}^{*imp}}_{\text{expected imputation noise}} \\
+ \quad & \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[2\mathbb{C}\mathrm{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{cross term}} + \underbrace{v^{*mis}}_{\text{expected target noise}}.
\end{aligned}
\tag{3.12}
$$

The proof of the above theorem follows after a discussion below.

From the first row of Equation (3.12) one can notice a link between the previously defined error measures 3) and 5) (see Section 3.4). In the innermost integration the conditionalizer is $\mathcal{Q}_P = \{\mathsf{n}, \mathbf{d}_{\boldsymbol{X}}^{mis}\}$ (remembering that $\mathbf{x}^{mis} \in \mathbf{d}_{\boldsymbol{X}}^{mis}$). Then after integrating first over the marginal density of $\mathbf{D}_{\boldsymbol{X}}^{mis}$ and then over number of missing data values N^{mis} we are at the $\mathcal{Q}_1$ level. The innermost conditionalizer may seem a bit rigorous, however it just fixes the imputation position and the size of training data (and the total sample size). One should note that the cross term (in the second last row) is typically zero, as in Equation (3.10).

Expected squared imputation bias consists of four terms: squared global bias, variability of estimated model, variability of true conditional mean, and a cross term. Global bias measures the difference between the first moments of imputed Y and missing Y values. The role of the cross term can be significant. The variance terms measure the variability of the estimated and true models. One can only affect the variability of the estimated model. A more stiff model has less variability but on the other hand is likely to yield an increase in the squared global bias. The reason for this is that (too) stiff a model is likely to yield biased predictions. Interpretation of the cross term is complicated, and it is generally not neglible. For example, consider unbiased predictions: $\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathsf{n}^{mis}, \mathbf{x}^{mis}, \mathsf{n}] = g^{*mis}(\mathbf{x}^{mis})$. Then the cross term equals to minus two times variability of $g^{*mis}(\mathbf{x}^{mis})$.

Proof 3.7 *for Theorem 3.6.*

$$
\begin{aligned}
& \mathbb{E}[m\hat{s}e(Y^{comp})|\mathsf{n}] \qquad (3.13)\\
= \; & \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathrm{mse}(Y^{imp}|\boldsymbol{X}^{mis},\mathsf{n}^{mis},\mathsf{n})\Bigg]\\
= \; & \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\bigg(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\bigg)^2\Bigg]\\
& +\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] + \mathbb{V}\mathrm{ar}[\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\\
& +2\mathbb{C}\mathrm{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Bigg] + v^{*mis}\\
= \; & \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\bigg(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\bigg)^2\Bigg]}_{\text{expected squared imputation bias (ESIB)}}\\
& + \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Bigg]}_{\text{expected variance of conditional mean estimate}} + \underbrace{v^{*imp}_{\mathsf{n}}}_{\text{expected imputation noise}}\\
& + \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[2\mathbb{C}\mathrm{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Bigg]}_{\text{cross term}} + \underbrace{v^{*mis}}_{\text{expected target noise}}.
\end{aligned}
$$

Further, one should note that expected squared imputation bias (ESIB) can be decomposed as

$$
\begin{aligned}
& ESIB = \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\bigg(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - \mu^{*imp}_{\mathsf{n}} \qquad (3.14)\\
+ \; & \mu^{*imp}_{\mathsf{n}} - \mu^{*mis} + \mu^{*mis} - g^{*mis}(\boldsymbol{X}^{mis})\bigg)^2\Bigg]\\
= \; & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Bigg]\\
+ \; & \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\bigg(\mu^{*imp}_{\mathsf{n}} - \mu^{*mis} + \mu^{*mis} - g^{*mis}(\boldsymbol{X}^{mis})\bigg)^2\Bigg]\\
+ \; & 2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\bigg(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - \mu^{*imp}_{\mathsf{n}}\bigg)\bigg(\mu^{*imp}_{\mathsf{n}} - g^{*mis}(\boldsymbol{X}^{mis})\bigg)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right] + (\mu^{*imp}_{\mathsf{n}} - \mu^{*mis})^2 + \mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})] \\
+ \quad & 2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[(\mu^{*imp}_{\mathsf{n}} - \mu^{*mis})(\mu^{*mis} - g^{*mis}(\boldsymbol{X}^{mis})\right] \\
+ \quad & 2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - \mu^{*imp}_{\mathsf{n}}\right)\left(\mu^{*imp}_{\mathsf{n}} - g^{*mis}(\boldsymbol{X}^{mis})\right)\right] \\
= \quad & \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{variability of conditional mean estimate}} + (\underbrace{\mu^{*imp}_{\mathsf{n}} - \mu^{*mis}}_{\text{global bias}})^2 + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
+ \quad & \underbrace{2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - \mu^{*imp}_{\mathsf{n}}\right)\left(\mu^{*imp}_{\mathsf{n}} - g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{cross term}},
\end{aligned}
$$

where $\mu^{*imp}_{\mathsf{n}} = \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathsf{n}^{mis},\mathbf{x}^{mis},\mathsf{n}]$.

Result (3.12) follows by plugging decomposition (3.14) into (3.13) □

From the decompositions (3.13) and (3.14) we get an insight to the preservation of the second moment. Namely, the variance of imputed Y values for sample size n is

$$\tau^{*imp}_{\mathsf{n}} \quad = \quad \mathbb{V}\mathrm{ar}_{N^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right] + v^{*imp}_{\mathsf{n}},$$

The above decomposition states that equation $\lim_{\mathsf{n}\to\infty}\tau^{*imp}_{n} = \tau^{*mis}$ (unbiased second moment) may be achieved in multiple ways. For example, one may use a stiff conditional mean estimate with more varying noise terms or one may use a flexible model with less varying noise terms to yield the same τ^{*imp}_{n}.

Finally, for an even better interpretation the limit of expectation of $\hat{mse}(Y^{comp})$ can be computed:

Corollary 3.8 *Asymptotics of* $\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}]$.

$$
\begin{aligned}
& \lim_{\mathsf{n}\to\infty}\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] = \underbrace{\mathbb{V}\mathrm{ar}[g^{*imp}(\boldsymbol{X}^{mis})]}_{\text{limit of variance of conditional mean "estimator"}} \\
& + \underbrace{(\mu^{*imp} - \mu^{*mis})^2}_{\text{limiting global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
& + \underbrace{\lim_{\mathsf{n}\to\infty} 2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - \mu^{*imp}_{\mathsf{n}}\right)\left(\mu^{*imp}_{\mathsf{n}} - g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{limit of cross term}} \\
& + \underbrace{v^{*imp}}_{\text{limit of imputation noise}} + \underbrace{v^{*mis}}_{\text{expected target noise}},
\end{aligned}
$$

where $g^{*imp}(\mathbf{x})$ is the limit of conditional mean estimate at point $\mathbf{x}$.

The proof of above corollary is given after the discussion below.

We can conclude that the limit consists of a squared bias term, asymptotic imputation noise, and expected target noise (irreducible term). The squared bias term is zero for any pointwise consistent conditional mean estimate. One should remember that in the NMAR case for a nonparametric method it is likely that $g^{*imp}(\mathbf{x}) = g^{*obs}(\mathbf{x})$. Therefore in NMAR asymptotic squared bias can be considerable. In the MAR and MCAR cases there are typically no such problem because $g^{*}(\mathbf{x}) = g^{*obs}(\mathbf{x}) = g^{*mis}(\mathbf{x})$. Asymptotic imputation noise can be controlled. However, reducing it (too much) is likely to yield an asymptotically biased s econd moment estimator.

In the computation of limit of $\hat{mse}(Y^{comp})$ we have assumed that the cross covariance term (the fourth row in Equation (3.12) goes to zero as the sample size goes to infinity. This can be considered as a sensible assumption, because for many methods this term is zero even for a finite sample size n.

Proof 3.9 *to Corollary 3.8.*

$$
\begin{aligned}
&\lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] \\
&= \lim_{\mathsf{n}\to\infty} \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|n}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\right)^2\right] \\
&+ \lim_{\mathsf{n}\to\infty} \quad \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \quad\right] + \lim_{\mathsf{n}\to\infty} \quad v_{\mathsf{n}}^{*imp} \\
&+ \lim_{\mathsf{n}\to\infty} \mathbb{E}_{M,\boldsymbol{X}^{mis}|\mathsf{n}}\left[2\mathbb{C}\mathrm{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right] + \lim_{\mathsf{n}\to\infty} \quad v^{*mis} \\
\approx\; &\underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}}\left[\left(g^{*imp}\left(\mathbf{x}^{mis}\right) - g^{*mis}\left(\mathbf{x}^{mis}\right)\right)^2\right]}_{\text{asymptotic expected squared bias (AESB)}} + \underbrace{v^{*imp}}_{\text{asymptotic imputation noise}} \\
&+ \underbrace{v^{*mis}}_{\text{expected target noise}}.
\end{aligned}
\tag{3.15}
$$

One should note that asymptotic expected squared bias (AESB) term can be decom-

posed using Equation (3.14) as

$$\begin{aligned}
AESB &= \lim_{n\to\infty} \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|n}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, n^{mis}, n]\right] \\
&+ \lim_{n\to\infty} (\mu_n^{*imp} - \mu^{*mis})^2 + \lim_{n\to\infty} \mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})] \\
&+ \lim_{n\to\infty} 2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|n}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, n^{mis}, n] - \mu_n^{*imp}\right)\left(\mu_n^{*imp} - g^{*mis}(\boldsymbol{X}^{mis})\right)\right] \\
= &\underbrace{\mathbb{V}\mathrm{ar}[g^{*imp}(\boldsymbol{X}^{mis})]}_{\text{variance of asymptotic conditional mean "estimate"}} + \underbrace{(\mu^{*imp} - \mu^{*mis})^2}_{\text{asymptotic global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
&\underbrace{+ \lim_{n\to\infty} 2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|n}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, n^{mis}, n] - \mu_n^{*imp}\right)\left(\mu_n^{*imp} - g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{limit of cross term}},
\end{aligned} \tag{3.16}$$

where μ^{*imp} is the asymptotic expectation of the imputed Y values. All the other terms are quitet straightforward to interpret but the limit of the cross term is quite complicated.

Corollary 3.8 follows by plugging Equation (3.15) into Equation (3.16) □

3.9 Summary

This chapter introduced the background of the thesis. First, a real-world scenario for potential applications was described. It corresponds to typical working processes in statistical offices. Primary assumptions, which are needed to concretise the theoretical setting, were also desribed.

As a summary, we studied estimators and their properties were introduced. This included estimators for sample moments and the mean squared error. To aid interpretations, several types of conditionalizations were used. These limit variation sources and thus ease the analyses.

Finally basic decompositions were given. Decompositions are based on the roles of observed and missing data, as well as on the role of estimators. Decompositions make the interpretation of results easier, which is vital for the comparison of imputation methods.

Chapter 4

Imputation using simple methods and linear regression

In this chapter we introduce some basic methods that are used as a reference when analysing the imputation performance of cell imputation. The presentation contains descriptions of the methods, and an analysis of their imputation performances in terms of evaluation statistics. The analysis follows previously explained guidelines. An incomplete random data $\mathbf{D}^{inc}$ with n observations is imputed such that the missing values of an univariate real valued target Y^{mis} are replaced with imputed ones Y^{imp}. For each method we like to know how the imputation error, which is measured through evaluation statistics, can be decomposed and explained. This leads to a methodological comparison in Chapter 7 and it also explains the performance of the methods with real-world data in Chapter 8.

The methods that are studied in this chapter are denoted with letters B and L, where

B is short for baseline, and

L indicates linear regression.

Each method can be applied to imputation using three (or two) types of imputation strategies which are denoted with letters M, R and D as follows:

M (mean imputation),

R (simulated random imputation), and

D (donor imputation).

Note that donor strategy is omitted for linear regression.

Using the shorthand notation, we shall denote completed variables as $Y^{comp,model,strategy}$, for example $Y^{comp,L,R}$ denotes that Y^{comp} was imputed with linear regression with added simulated noise. Similarly for the evaluation statistics, concerning some statistic $\boldsymbol{\theta}$ we shall use notation $\boldsymbol{\theta}^{model,strategy}$.

4.1 Baseline methods

Perhaps the simplest way to impute missing values is to put them into the mean of observed data. Since this underestimates the second moment, a natural next step is to add randomness to imputation. We call these as baseline methods, where the completed target variable Y^{comp} given data and response indicators (conditionalisation $\mathcal{Q}_3$ and $\mathbf{r}$) is

$$Y_j^{comp,B} = \begin{cases} y_j, & \text{if } r_j = 1, \quad y_j \text{ is observed} \\ \mu^{obs} + \hat{\epsilon}, & \text{if } r_j = 0, \quad y_j \text{ is missing,} \end{cases}$$

where $\mu^{obs} = \frac{1}{\mathsf{n}^{obs}} \sum_{j=1}^{\mathsf{n}^{obs}} y_j, \quad y_j \in \mathbf{d}^{train}$ and $\hat{\epsilon}$ is a random term, which is chosen according to our imputation strategy (see Section 2.5):

i) M(mean), $\hat{\epsilon}^M = 0$, which implies mean imputation

ii) R(random), $\hat{\epsilon}^R \sim \hat{f}_\epsilon(e)$, simulated noise where $\hat{\epsilon}$ is iid sampled from noise distribution $\hat{f}_\epsilon(e)$. For simplicity we assume normal distribution $N(0, v^{obs})$ for $\hat{\epsilon}$ where the variance estimate comes from $\mathbf{d}^{train}$:

$$v^{obs} = \tau^{obs} = \frac{1}{\mathsf{n}^{obs} - 1} \sum_{j=1}^{\mathsf{n}^{obs}} \left(y_j - \mu^{obs} \right)^2, \qquad y_j \in \mathbf{d}^{train}.$$

iii) D(donor), $\hat{\epsilon}^D \sim \{y_k - \mu^{obs}\}$, where residual is taken from the randomly selected observation $y_k \in \mathbf{d}^{train}$ with $k \in \{1, \dots, \mathsf{n}^{obs}\}$. This is a sampling with replacement, because the donor indeces are selected independently from $\mathbf{d}^{train}$. Of course, this is same as $y_j^{comp} = y_k$, with random $y_k \in \mathbf{d}^{train}$.

The reason for naming these as baseline methods, is that they give us a reference point for imputation performance. We except that a more advanced method should do better than these baselines, at least when the problem is more complex than that of simple MCAR.

The evaluation of the imputation performance of baseline methods is quite straightforward. These are, after all, well known methods in the literature (see references [62] and [90] for an example). In the current context we only need to write the results in a form that is compatible with our evaluation statistics.

The main results are summarized in Theorem 4.1 and Approximation 4.3. Since there is no difficulty in interpreting these results we omit some details such as the results with conditionalisations $\mathcal{Q}_2$ and $\mathcal{Q}_3$.

4.1.1 Preservation of moments

When the missingness mechanism is MCAR, baseline methods are expected to perform well considering the preservation of the first moment. However, the second moment estimators differ due to the differences in imputation strategies. This is quite obvious for mean imputation, but also the estimation of the noise model is a potential source of errors.

The weakness of the baseline methods is revealed when Y depends on covariates and the missingness mechanism is more complicated than MCAR. Then all the moments of $\hat{Y}^{comp}$ are likely to be biased against the true moments of Y.

The reliability of the estimated moments depends on the number of observations, the distribution of Y^{obs}, the distribution of N^{mis}, and the imputation strategy. Especially, the variance of the first moment, $\mathbb{V}\text{ar}[\hat{\mu}^{comp,B}]$, depends on sample size n, variance τ^{*obs}, the proportion of missing observations p^*, and the variance of N^{mis}.

Let us denote the first moment of the completed data as $\hat{\mu}^{comp,B}$, where B stands for the baseline method. The bias of $\hat{\mu}^{comp,B}$ with respect to population mean is:

Theorem 4.1 *Bias of the first moment* $\hat{\mu}^{comp,B}$
The bias of $\hat{\mu}^{comp,B}$ *for* n *observations is*

$$\mathbb{B}\text{ias}[\hat{\mu}^{comp,B}|\mathsf{n}] = p^*(\mu^{*obs} - \mu^{*mis}).$$

The proof of this theorem and justifications of other approximations and a consequence about the preservation of moments are given after Consequence 4.4.

The variance of the first moment depends on the imputation strategy, as specified in Approximation 4.2.

Approximation 4.2 *Approximation for the variance of the first moment* $\hat{\mu}^{comp,B}$
The variance $\mathbb{V}\text{ar}[\hat{\mu}^{comp,B}]$ *with* n *observations is approximately*

$$\mathbb{V}\text{ar}[\hat{\mu}^{comp,B}] \approx \tau^{*obs}\left(\underbrace{\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\text{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}}_{\text{due sampling}} + \underbrace{C}_{\text{due imputation strategy}}\right),$$

where term C *depends on imputation strategy* $\hat{\epsilon}^S$ *as follows:*

$$C = \begin{cases} 0 & :S{=}M \quad \text{(for mean imputation strategy),} \\ \frac{p^*}{\mathsf{n}} & :S{=}R \quad \text{(for simulated random imputation), and} \\ \frac{p^*}{\mathsf{n}}\left(1 - \frac{1}{\mathsf{n}(1-p^*)}\right) & :S{=}D \quad \text{(for random donor).} \end{cases}$$

We can summarize Theorem 4.1 and Approximation 4.2 in such a way that error in the first moment depends on missingness probability p^*, the difference between $\mu^{*obs} - \mu^{*mis}$, sample size n, and data variance τ^{*obs}. Of course in MCAR cases $\mu^{*obs} = \mu^{*mis}$, which makes $\hat{\mu}^{comp}$ to be unbiased.

Approximation 4.3 *Approximation for the bias of the* 2^{nd} *moment* $\hat{\tau}^{comp,B}$
The bias of $\hat{\tau}^{comp,B}$ *for* n *observations is approximately*

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\tau}^{comp,B}|\mathsf{n}] \approx\; & p^*(\tau^{*imp} - \tau^{*mis}) - p^*(1-p^*)(\mu^{*mis} - \mu^{*obs})^2 \\ & + C + O(\mathsf{n}^{-1}),\end{aligned}$$

where imputation variance τ^{*imp} and sampling error C depend on imputation strategy $\hat{\epsilon}^S$. Imputation variance τ^{*imp} and term C depend on imputation strategy $\hat{\epsilon}^S$ as follows:

$$\tau^{*imp} = \begin{cases} 0 & \textit{:S=M} \quad \text{(for mean imputation strategy), and} \\ \tau^{*obs} & \textit{:S=R and S=D} \quad \text{(random and donor imputation).} \end{cases}$$

and sample error C is

$$C = \begin{cases} 0 & \textit{:S=M} \quad \text{(for mean imputation strategy),} \\ \frac{1-p^*}{n}\tau^{*obs} & \textit{:S=R} \quad \text{(for simulated random imputation), and} \\ \frac{n(1-p^*)-1}{n^2}\tau^{*obs} & \textit{:S=D} \quad \text{(for random donor).} \end{cases}$$

When $\mathsf{n} \to \infty$ we get the asymptotics as follows:

Consequence 4.4 *(Approximation to) asymptotical moments for baseline methods Asymptotically we have (approximately) the following*

$$\begin{aligned} \lim_{n\to\infty} \mathbb{B}\text{ias}[\hat{\mu}^{comp,B}|n] &= p^*(\mu^{*obs} - \mu^{*mis}) \\ \lim_{n\to\infty} \mathbb{V}\text{ar}[\hat{\mu}^{comp,B}|n] &\approx 0 \\ \lim_{n\to\infty} \mathbb{B}\text{ias}[\hat{\tau}^{comp,B}|n] &\approx p^*\Big[(\tau^{*imp} - \tau^{*obs}) - (1-p^*)(\mu^{*mis} - \mu^{*obs})^2\Big], \end{aligned}$$

where

$$\tau^{*imp} = \begin{cases} 0 & \textit{:S=M} \quad \text{(for mean imputation strategy) and} \\ \tau^{*obs} & \textit{:S=R and S=D} \quad \text{(random and donor imputation).} \end{cases}$$

The proofs and justifications for the above consequence, approximations and theorem are quite straightforward.

Proof 4.5 *for Theorem 4.1 is quite trivially*

$$\begin{aligned} \mathbb{B}\text{ias}[\hat{\mu}^{comp,B}|n] &= \mathbb{E}[\hat{\mu}^{comp,B}|n] - \mu^* \\ &= \mu^{*obs} - (1-p^*)\mu^{*obs} - p^*\mu^{*mis} \\ &= p^*(\mu^{*obs} - \mu^{*mis}) \qquad \square \end{aligned}$$

Recall the conditionalizations levels that are used in this thesis

$$\begin{aligned} \mathcal{Q}_1 &= \{\mathsf{n}\} \\ \mathcal{Q}_2 &= \{\mathsf{n}, \mathbf{d}^{train}, g(\mathbf{x}|\boldsymbol{\theta})\} \\ \mathcal{Q}_3 &= \{\mathbf{d}^{train}, \mathbf{d}^{test}, g(\mathbf{x}|\boldsymbol{\theta})\}. \end{aligned}$$

with the help of these the justification of the second approximation can be derived quite easily. For clarity this is given for a simulated random imputation strategy.

In justification of Approximation 4.2 one needs to compute expectation $\mathbb{E}[\frac{1}{N^{obs}}]$. Note that the ratio in expectation may be written as $\frac{1}{\mathsf{n}}/\frac{N^{obs}}{\mathsf{n}}$ (the ratio between two mean estimators, which of the first estimator is deterministic). Hence, some asymptotic properties might be derived by applying the Slutsky's theorem (see [92] for details). However, small sample properties are here of interest. A closed form solution is not possible as the distribution function of N^{mis} is unknown. Therefore the second order Taylor approximation, ignoring higher order terms, is applied. Intermediate Step 4.6 gives the approximate value of the expectation. This approximation is actually equal to the approach taken by Kempen and Vliet who consider expectation and the variance of the ratio of estimators [49].

Intermediate Step 4.6 *An approximation for* $\mathbb{E}[\frac{1}{\mathsf{N}^{obs}}]$ *can be given as:*

$$\mathbb{E}\Big[\frac{1}{\mathsf{N}^{obs}}\Big] \approx \frac{1}{\mathsf{n}(1-p^*)} + \Big(\mathsf{n}(1-p^*)\Big)^{-3}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}|\mathsf{n}].$$

Justification: *Using Taylor approximation at* $\mathbb{E}[\mathsf{N}^{obs}]$ *and* $\mathsf{N}^{obs} = \mathsf{n} - \mathsf{N}^{mis}$ *we get*

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{\mathsf{N}^{obs}}\Big] \quad &\approx \quad \mathbb{E}\Big[\frac{1}{\mathbb{E}[\mathsf{N}^{obs}]} + \frac{1}{1!}\big(\frac{\partial}{\partial \mathsf{N}^{obs}}\frac{1}{\mathsf{N}^{obs}}\big)_{\mathsf{N}^{obs}=\mathbb{E}[\mathsf{N}^{obs}]}(\mathsf{N}^{obs} - \mathbb{E}[\mathsf{N}^{obs}]) \\
&\qquad +\frac{1}{2!}\big(\frac{\partial^2}{\partial \mathsf{N}^{obs}\partial \mathsf{N}^{obs}}\frac{1}{\mathsf{N}^{obs}}\big)_{\mathsf{N}^{obs}=\mathbb{E}[\mathsf{N}^{obs}]}(\mathsf{N}^{obs} - \mathbb{E}[\mathsf{N}^{obs}])^2\Big] \\
&= \quad \frac{1}{\mathbb{E}[\mathsf{N}^{obs}]} + \frac{1}{2}\big(2(\mathsf{N}^{obs})^{-3}\big)_{\mathsf{N}^{obs}=\mathbb{E}[\mathsf{N}^{obs}]}\mathbb{V}\mathrm{ar}[\mathsf{N}^{obs}|\mathsf{n}] \\
&\overset{\mathsf{N}^{obs}=\mathsf{n}-\mathsf{N}^{mis}}{\approx} \quad \frac{1}{\mathsf{n}(1-p^*)} + \Big(\mathsf{n}(1-p^*)\Big)^{-3}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}|\mathsf{n}],
\end{aligned}$$

where $\frac{\partial}{\partial \mathsf{N}^{obs}}$ *is ordinary derivative (not stochastic).*

Justification 4.7 *partial justification of Approximation 4.2.*
It is easiest to start deriving first conditionalizing level $\mathcal{Q}_3$ *and then proceed towards* $\mathcal{Q}_1$*. Clearly for*

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_3] = \mathbb{V}\mathrm{ar}\left[\hat{\mu}^{comp,B,M} + \frac{1}{\mathsf{n}}\sum_{j=1}^{\mathsf{n}^{mis}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_3\right] = \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\tau^{obs}.$$

Variance in the second conditionalisation level is same as it is in the third level because covariates are not utilized. The result at the first conditionalization level is derived using the chain rule of variance and second order Taylor approximation (Intermediate Step 4.6). The required quantities are $\mathbb{E}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_2]$ *and* $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_2]$*, thus:*

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathsf{n}] \quad &= \quad \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}] + \mathbb{E}\Big[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\hat{\tau}^{obs}|\mathsf{n}\Big] \\
&= \quad \mathbb{E}\Big[\frac{1}{\mathsf{N}^{obs}}\Big]\tau^{*obs} + \mathbb{V}\mathrm{ar}[\mu^{*obs}|\mathsf{n}] + \mathbb{E}\Big[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\hat{\tau}^{obs}|\mathsf{n}\Big] \\
&\overset{Taylor}{\approx} \quad \tau^{*obs}\left(\frac{1}{\mathsf{n}(1-p^*)} + \big(\mathsf{n}(1-p^*)\big)^{-3}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}|\mathsf{n}] + \frac{p^*}{\mathsf{n}}\right),
\end{aligned}$$

where the second term is of the order $O(\mathsf{n}^{-2})$ *if* N^{mis} *is binomially distributed.*

Further details and complete justifications for all the imputation strategies can be found in Appendix A4.

One should notice that the second order Taylor approximation reveals more details of the behaviour of the variance of $\hat{\mu}^{comp}$ than does the first order approximation (the first term plus the third term). Actually, the first order approximation would be good only for a large sample size. However, typically the first order Taylor approximation is used to compute variance because it is mathematically feasible. However, with baseline methods the second order approximation is easy to compute.

For compactness we limit the justification of Approximation 4.3 again to simulated random imputation strategy. Thus we consider situtation $\mathbb{B}\text{ias}[\hat{\tau}^{comp,B,R}|\mathsf{n}]$. A full version of the justification can be found in Appendix A4.

Justification 4.8 *partial justification of Approximation 4.3.*
Quantity $\hat{\tau}^{imp}$ can be written as

$$\hat{\tau}^{imp} = \frac{1}{n^{mis}-1}\sum_{j=1}^{n^{mis}}\left(\hat{\epsilon}_j^{mis} - \bar{\hat{\epsilon}}^{mis}\right)^2,$$

where $\bar{\hat{\epsilon}}^{mis}$ is the mean of imputation noise terms. At conditionalization level three quantities $\hat{\epsilon}_j^{mis}$ are identically and independently distributed with the expectation zero and variance τ^{obs}. Therefore the expectation of $\hat{\tau}^{imp}$ is τ^{obs}. Now with
$\mathcal{Q}_3 = \{\mathbf{d}^{train}, \mathbf{d}^{test}, g(\mathbf{x}|\boldsymbol{\theta})\}$

$$\begin{aligned}\mathbb{E}[\hat{\tau}^{B,R}|\mathcal{Q}_3] &= \mathbb{E}\Big[\frac{n^{obs}-1}{n-1}\hat{\tau}^{obs} + \frac{n^{mis}-1}{n-1}\hat{\tau}^{imp} + \frac{n^{mis}n^{obs}}{n(n-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2|\mathcal{Q}_3\Big]\\ &= \frac{n^{obs}-1}{n-1}\tau^{obs} + \frac{n^{mis}-1}{n-1}\tau^{obs} + \frac{n^{obs}}{n(n-1)}\tau^{obs}.\end{aligned}$$

Expectation and bias are same in the second level as in the third level, because covariate information is not used. The first level expectation $\mathcal{Q}_1 = \{n\}$ is now

$$\begin{aligned}\mathbb{E}[\hat{\tau}^{B,R}|\mathcal{Q}_1] &= \mathbb{E}_{\mathsf{N}^{mis},\mathbf{D}^{train}|n}\Big[\frac{\mathsf{N}^{obs}-1}{n-1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{n-1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{obs}}{n(n-1)}\hat{\tau}^{obs}\Big]\\ &\approx \frac{n-np^*-1}{n-1}\tau^{*obs} + \frac{np^*-1}{n-1}\tau^{*obs} + \frac{n-np^*}{n(n-1)}\tau^{*obs}\\ &= (1-p^*)\tau^{*obs} + p^*\tau^{*obs} + \frac{1-p^*}{n}\tau^{*obs}\\ &\quad +O(n^{-1}).\end{aligned}$$

Justifications for consequences are quite obvious and are omitted in the current context. Further details are given in the appendixes.

4.1.2 An example: preservation of the first moment

Theorem 4.1 and Approximation 4.3 give us an insight about the imputation performance of baseline methods. It seems that when the number of observations is small errors in the moments of completed data Y^{comp} are caused mainly by a model missmatch and estimation variance. Of course the severity of these errors depends on data and the type of missingness.

In the simplest case, we see that imputation strategy has very little effect to the error in the first moment $\hat{\mu}^{comp,B}$. Instead, errors are mainly due to sampling and MAR types of differences between μ^{*mis} and μ^{*obs}. This is easily seen also when using simulations. In the following example, (see Figure 4.1), target Y given covariate x is defined as

$$Y_{|x} = \frac{1}{500}x^3 + \epsilon,$$

where $\epsilon \sim N(0, 0.15)$ is about 28% of the variance of Y. Now let the missingness be of MAR type, where

$$X^{obs} \sim N(-2, 15), \text{ while } X^{mis} \sim N(2, 15),$$

as depicted in Figure 4.1.

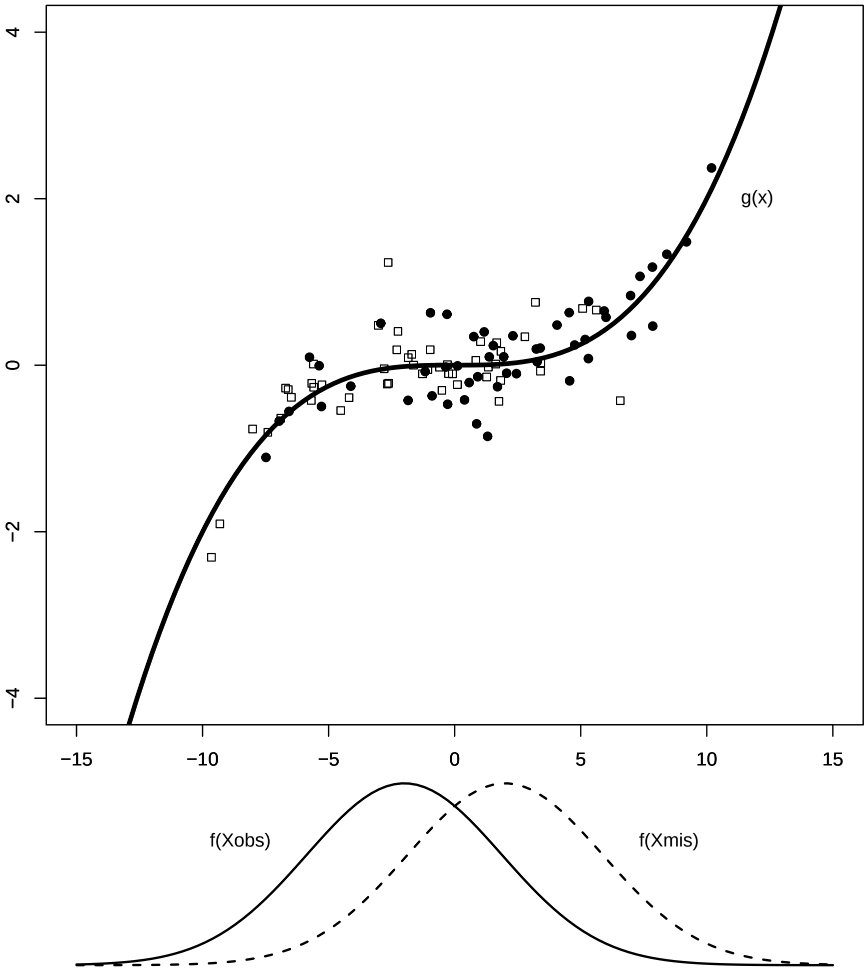


Figure 4.1: True model $g(x)$, marginal distributions of X^{obs} (solid) and X^{mis} (dashed), and a random sample of size 100. Training data is denoted by square plots and draws from the missing population using black dots.

Following our theory we shall take a random sample of n observations, where a random number of N^{mis} observations are missing using approximately $\mathsf{N}^{mis} \sim Bin(\mathsf{n}, \frac{1}{2})$ observations from $Y_{|X^{mis}}$, $X^{mis} \sim f_{X^{mis}}(x)$ and N^{obs} observations from $Y_{|X^{obs}}$, $X^{obs} \sim f_{X^{obs}}(x)$. Now we have everything that is required for the analysis.

Clearly baseline methods are not optimal for this type of data, but we could still get a rather good imputation performance for the first moment. However, the results are quite strongly dependent on the number of observations and the missingness probability (now $p^* = 0.5$).

To compare the simulation results with analytical formulas we need to solve the first two moments of Y^{obs} and Y^{mis}. This requires some analytical computations (see Appendix A4.5 for details). The moments are

$$\begin{aligned}
\mathbb{E}[Y^{obs}] &= -0.196 \\
\mathbb{V}\mathrm{ar}[Y^{obs}] &= 0.49074 \\
\mathbb{E}[Y^{mis}] &= 0.196 \\
\mathbb{V}\mathrm{ar}[Y^{mis}] &= 0.49074, \text{ which leads to} \\
\mathbb{V}\mathrm{ar}[Y] &= 0.49074 + 0.5(1-0.5)(-0.196-0.196)^2 = 0.529156 \approx 0.53
\end{aligned} \tag{4.1}$$

The mean squared error of the first moment can be defined as

$$\mathrm{MSE}[\hat{\mu}^{comp,B}|\mathsf{n}] = \mathbb{B}\mathrm{ias}^2[\hat{\mu}^{comp,B}|\mathsf{n}] + \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B}|\mathsf{n}].$$

Applying Theorem 4.1 with the theoretical moments and the setup of our experiment we see that

$$\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,B}|\mathsf{n}] \approx 0.5(-0.196 - 0.196) = -0.196 \approx -0.2.$$

This bias result was verified by our simulations. However, we omit the result as it is not interesting: roughly a flat line as a function of the sample size.

More interestingly, recall that $\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}|\mathsf{n}] \approx 0.5(1-0.5)\mathsf{n} = 0.25\mathsf{n}$, because N^{mis} is roughly $Bin(\mathsf{n}, 0.5)$ distributed, we get approximations for variances from Approximation 4.2:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathsf{n}] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}] + 0 \approx \frac{0.98}{\mathsf{n}} + \frac{2}{\mathsf{n}^2} \\
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathsf{n}] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}] + \frac{p^*}{\mathsf{n}}\tau^{*obs} \approx \frac{0.98}{\mathsf{n}} + \frac{2}{\mathsf{n}^2} + \frac{0.5}{\mathsf{n}} * 0.49 \\
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}|\mathsf{n}] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}] + \frac{p^*}{\mathsf{n}}\Big(1 - \frac{1}{\mathsf{n}(1-p^*)}\Big)\tau^{*obs} \\
&\approx \frac{0.98}{\mathsf{n}} + \frac{2}{\mathsf{n}^2} + \frac{0.5}{\mathsf{n}}(1 - \frac{1}{\mathsf{n}(1-0.5)}) * 0.49.
\end{aligned}$$

where $\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}] \approx \frac{\tau^{*obs}}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3} \approx \frac{0.49}{0.5\mathsf{n}} + \frac{0.25}{0.125*\mathsf{n}^2} = \frac{0.98}{\mathsf{n}} + \frac{2}{\mathsf{n}^2}$. Further, due to sampling errors the variance for the true data mean estimator $\mathbb{V}\mathrm{ar}[\hat{\mu}^{OPT}|\mathsf{n}]$ is approximately $0.53/\mathsf{n}$. We shall use this as a reference, since variances other than

$\mathbb{V}\mathrm{ar}[\hat{\mu}^{OPT}|\mathsf{n}]$ correspond to too small or large (analytical) confidence intervals for the mean estimator.

Now let $\mathsf{n} = \lfloor 10 * 1.7^k \rfloor$, when $k = 0, \ldots, 9$. Table 4.1 contains the theoretical results which are computed using analytical formulas. The results can be verified by simulations, where data is first sampled from Y for missing and observed parts, and then imputed. To minimize simulation variance this was repeated 8000 times, giving Figures 4.2 and 4.3, where the mean result over all the simulation runs is given.

n	10	17	28	49	83	141	241	410	697	1185
S	k=0					k=5				k=9
R	0.143	0.079	0.046	0.026	0.015	0.009	0.005	0.003	0.002	0.001
D	0.138	0.077	0.046	0.026	0.015	0.009	0.005	0.003	0.002	0.001
M	0.118	0.065	0.038	0.021	0.012	0.007	0.004	0.002	0.001	0.001
OPT	0.053	0.032	0.019	0.011	0.006	0.004	0.002	0.001	0.001	0.000

Table 4.1: Theoretical variances $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B}|\mathsf{n}]$, with 3 digit precision, for imputation strategies, S, and sampling errors from true data $\mathbb{V}\mathrm{ar}[\hat{\mu}^{OPT}|\mathsf{n}]$ (OPT). The variance for the mean of the observed Y values is the same as in mean imputation strategy M.

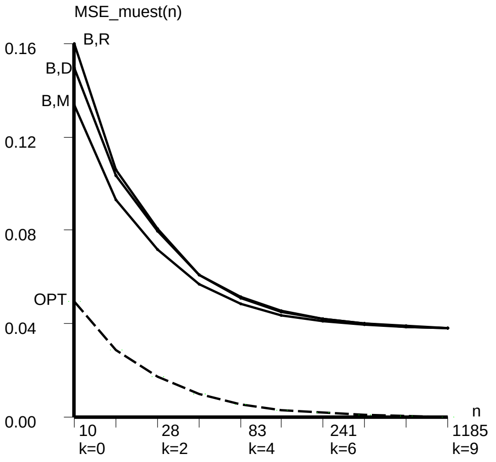


Figure 4.2: Simulated mean squared error of the first moment.

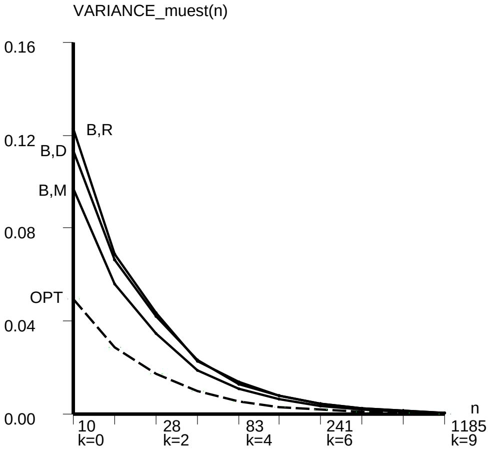


Figure 4.3: Simulated variance of the first moment.

The simulated results (Figure 4.3) correspond to the computed ones (Table 4.1). The theoretical variance results have slightly higher values than the simulated results. One reason for this is that the expectation of N^{mis} is less than $0.5\mathsf{n}$, especially for small sample sizes. This is due to technical conditionalization, as we need to ensure that the estimated quantities exist. Secondly, Taylor approximations which are used to derive the analytical formulas may be inaccurate.

The variance and squared bias ratio is roughly 1:1 for all the three imputation strategies at sample size 28. For a smaller sample size the variance contributes the major part of errors to results. Squared bias dominates the mean squared error for the higher sample size. Therefore the simulated results verify our conclusions that for a small sample size choosing imputation strategy is important when considering the mean squared error of $\hat{\mu}^{comp,B}$. In general, random imputation strategies product a significant increase in variance when compared to the mean imputation strategy. The variance is increased roughly by 20%. The variance for random donor imputation could be reduced by altering the donor sampling strategy.

Variances for all the imputation strategies are larger than for optimal mean $\hat{\mu}^{OPT}$. Thus the corresponding confidence intervals are too large (when compared to the confidence intervals of $\hat{\mu}^{OPT}$). The absolute difference in variances, and thus in confidence intervals, decreases as the sample size grows.

4.1.3 Unit level prediction errors

Baseline methods are maximally stiff, because they ignore the covariates. As a consequence (squared) prediction bias is typically high. Variance of predictions does not depend on statistical properties of observations of covariate. Therefore variance decreases as a function of sample size, and increases as a function of expected proportion of missing data values. These properties are inherited in unit level evaluation statistics. As we can see, baseline methods are most usable at a unit level if the conditional mean of Y given covariate is approximately μ^{*obs}. Approximations 4.9 and 4.10 verify these considerations.

Recall that the conditional mean of Y given $\mathbf{x}$ can be decomposed as $g(\mathbf{x}) = (1-p^*_{\mathbf{x}})g^{*obs}(\mathbf{x}) + p^*_{\mathbf{x}}g^{*mis}(\mathbf{x})$, where $p^*_{\mathbf{x}}$ is the probability for missingness of Y given $\boldsymbol{X} = \mathbf{x}$ and $g^{*obs}(\mathbf{x})$ and $g^{*mis}(\mathbf{x})$ are the conditional means for observed and missing Y values at $\mathbf{x}$ (see Equation 3.2).

Approximation 4.9 *An approximation for the expected mean squared prediction error for n observations*
Expectation of $\hat{mse}(Y^{comp,B})$ with n observations is approximately

$$\begin{aligned}\mathbb{E}[\hat{mse}(Y^{comp,B})|\mathsf{n}] \quad \approx \quad & \underbrace{(\mu^{*obs} - \mu^{*mis})^2}_{\text{global bias}} + \underbrace{\mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\ & \underbrace{+\tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big)}_{\text{expected sampling variance}} \\ & + \underbrace{\quad C\tau^{*obs}\quad}_{\text{expected imputation variance}} + \underbrace{\quad v^{*mis}\quad}_{\text{expected target variance}},\end{aligned}$$

where constant C depends on imputation strategy S:

$$C = \begin{cases} 0 & :S{=}M \quad \text{(for mean imputation strategy),} \\ 1 & :S{=}R \quad \text{(for simulated random imputation), and} \\ 1 - \frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3} & :S{=}D \quad \text{(for random donor).} \end{cases}$$

One should note that the sum of global bias squared and the variability of true model equal to the expected squared bias for baseline methods.

Approximation 4.10 *The mean squared error at a given point $\boldsymbol{X}^{mis} = \mathbf{x}_{mis}$.*
Mean squared error $\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, n^{mis}, n)$ *can be approximated as:*

$$\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, n^{mis}, n) \approx \underbrace{\left(\mu^{*obs} - g^{*mis}(\mathbf{x}^{mis})\right)^2}_{\text{squared bias}} + \underbrace{\frac{1}{n^{obs}}\tau^{*obs}}_{\text{sampling variance}} + \underbrace{C\tau^{*obs}}_{\text{imputation variance}} + \underbrace{v^{*mis}(\mathbf{x}^{mis})}_{\text{target variance}}.$$

where term C depends on imputation strategy S:

$$C = \begin{cases} 0 & :S{=}M \quad \text{(for mean imputation strategy),} \\ 1 & :S{=}R \quad \text{(for simulated random imputation), and} \\ 1 - \frac{1}{n^{obs}} & :S{=}D \quad \text{(for random donor).} \end{cases}$$

Consequence 4.11 *Approximation to asymptotics of* $\mathbb{E}[\hat{mse}(Y^{comp,B})|n]$.
Limit of expectation of $\hat{mse}(Y^{comp,B})$ *is approximately*

$$\lim_{n\to\infty} \mathbb{E}[\hat{mse}(Y^{comp,B})|n] \approx \underbrace{(\mu^{*obs} - \mu^{*mis})^2}_{\text{global bias}} + \underbrace{\mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} + \underbrace{C\tau^{*obs}}_{\text{expected imputation variance}} + \underbrace{v^{*mis}}_{\text{expected target variance}},$$

where constant C depends on imputation strategy S:

$$C = \begin{cases} 0 & :S{=}M \quad \text{(for mean imputation strategy), and} \\ 1 & :S{=}R, S{=}D \quad \text{(for random strategies).} \end{cases}$$

Justification of Consequence 4.11 is straightforward. However, full justifications for Approximations 4.9 and 4.10 are omitted here for compactness. The justifications are given in Appendix A4.3.

4.2 Imputation with linear regression

In this thesis the role of linear regression is similar to that of baseline methods. We use it in comparative evaluations between cell methods and standard algorithms. We hope that in the case of linear data, our advanced methods should not be completely inferior when compared to the optimal linear method. And in the case of complex data, our new methods should do better.

Using the standard linear regression (OLS) [36], our completed target variable Y^{comp} is defined under conditionalizations $\mathcal{Q}_3$ and responses $\mathbf{r}$ by

$$Y_j^{comp,L,S} = \begin{cases} y_j, & \text{if } r_j = 1, \\ \underbrace{\mathbf{x}_j^T \boldsymbol{\beta}^{obs}}_{g_L^{obs}(\mathbf{x}|\boldsymbol{\theta})} + \hat{\epsilon}^S, & \text{when } r_j = 0, \end{cases}$$

where covariate $\mathbf{x}_j = [x_{j,0}, x_{j,1}, \ldots, x_{j,p-1}]^T$, with $x_{j,0} = 1$ and the noise term $\hat{\epsilon}^S$ depends on the imputation strategy

$$\hat{\epsilon}^S \sim \begin{cases} 0, & \text{S=M (mean imputation),} \\ N(0, v^{obs,L,R}) & \text{S=R (simulated random imputation).} \end{cases}$$

Constant $x_{j,0} = 1$ defines the role of intercept β_0^{obs}, while other the parameters $\boldsymbol{\beta}_{-0}^{obs} = [\beta_1^{obs}, \ldots, \beta_{p-1}^{obs}]^T$ define the orientation of the regression surface.

Recall matrix partitioning $\mathbf{d}_{\boldsymbol{X}} = [\mathbf{d}_{\boldsymbol{X}}^{obs}\ \mathbf{d}_{\boldsymbol{X}}^{mis}]$ from Chapter 2. The regression coefficients $\boldsymbol{\beta}^{obs}$ are defined using the standard least squares (OLS) estimation from training data $\mathbf{d}_{\mathsf{n}^{obs}}^{train}$. With a slight change of notation, where

$$\mathbf{d}_{\mathbb{X}} = \mathbf{d}_{\mathbb{X}}^{train} = [\mathbf{1}_{\mathsf{n}^{obs}}\ \mathbf{d}_{\boldsymbol{X},\mathsf{n}^{obs}}^{obs}] = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{\mathsf{n}^{obs},1} & x_{\mathsf{n}^{obs},2} & \ldots & x_{\mathsf{n}^{obs},p-1} \end{bmatrix},$$

and $\mathbf{d}_{\mathbb{Y}} = \mathbf{d}_{\mathbb{Y}}^{train} = \mathbf{d}_{Y,\mathsf{n}^{obs}}^{obs} = [y_1, \ldots, y_{\mathsf{n}^{obs}}]^T$, the regression coefficients are

$$\boldsymbol{\beta}^{obs} = (\mathbf{d}_{\mathbb{X}}^T \mathbf{d}_{\mathbb{X}})^{-1} \mathbf{d}_{\mathbb{X}}^T \mathbf{d}_{\mathbb{Y}}.$$

One should note that above estimate is defined if the matrix $\mathbf{d}_{\mathbb{X}}$ has a full rank, formally $\text{rank}(\mathbf{d}_{\mathbb{X}}) = p$. Technical assumptions ensure this in our analysis.

The random term $\hat{\epsilon}^R$ is drawn iid from Normal distribution $N(0, v^{obs,L,R})$, where the variance is estimated as

$$v^{obs,L,R} = \frac{1}{\mathsf{n}^{obs}} \Big[\mathbf{d}_{\mathbb{Y}} - \mathbf{d}_{\mathbb{X}} \boldsymbol{\beta}^{obs}\Big]^T \Big[\mathbf{d}_{\mathbb{Y}} - \mathbf{d}_{\mathbb{X}} \boldsymbol{\beta}^{obs}\Big].$$

The analysis of the imputation performance of linear regression follows standard references like [62], but is adapted to the current problem setup. For this purpose we define the covariates for the missing part as a matrix

$$\mathbf{d}_{\mathbb{X}}^{mis} = [\mathbf{1}_{\mathsf{n}^{mis}}\ \mathbf{d}_{\boldsymbol{X},\mathsf{n}^{mis}}^{mis}] = \begin{bmatrix} 1 & x_{\mathsf{n}^{obs}+1,1} & x_{\mathsf{n}^{obs}+1,2} & \ldots & x_{\mathsf{n}^{obs}+1,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{\mathsf{n},1} & x_{\mathsf{n},2} & \ldots & x_{\mathsf{n},p-1} \end{bmatrix}.$$

4.2.1 Preservation of moments

The most obvious benefit of linear regression imputation over baseline methods is that the method is expected to perform better under MAR missingness. However, for the preservation of the first two moments of Y, we have less strict conditions, as explained by the following four approximations.

Approximation 4.12 *Approximation for the bias of the first moment $\hat{\mu}^{comp,L}$.*
The approximation for the bias of $\hat{\mu}^{comp,L}$ for n observations is

$$\mathbb{B}\text{ias}[\hat{\mu}^{comp,L}|\mathsf{n}] \approx p^*(\mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathsf{n}]^T \overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}_0^{obs}|\mathsf{n}] - \mu^{*mis})$$

where $_{-0}$ subscript means that all other regression coefficients are picked except intercept term, $\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}]$ is the expected regression coefficients over training data and $\overline{\boldsymbol{X}}^{*mis} = \mathbb{E}[\boldsymbol{X}^{mis}]$ is the expected covariate vector over missing data.

The variance of $\hat{\mu}^{comp,L}$ depends on the sample size, model fit and covariates. Given n, the result is quite complicated. Therefore we shall first see the variance of $\hat{\mu}^{comp,L}$ at conditionalisation level $\mathcal{Q}_2$ where the model, training data and missingness indicators are fixed such that $\mathsf{N}^{mis} = \mathsf{n}^{mis}$.

Approximation 4.13 *Approximation for* $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}, \mathbf{d}^{train}_{\mathsf{n}^{obs}}, \boldsymbol{\beta}^{obs}]$.
The variance of the first moment $\hat{\mu}^{comp,L}$ *given* $\mathcal{Q}_2$ *is approximately*

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathcal{Q}_2] \approx \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Big((\boldsymbol{\beta}^{obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \boldsymbol{\beta}^{obs}_{-0} + C\Big)$$

where

$$C = \begin{cases} 0 & :S{=}M \quad \text{(mean imputation), and} \\ v^{obs,L,R} & :S{=}R \quad \text{(simulated random imputation).} \end{cases}$$

According to Approximation 4.13 there is no variance due to the intercept term at conditionalisation $\mathcal{Q}_2$. The reason for this is that the model is fixed at that level.

The variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}]$ requires rather complex integrations. Therefore we shall give the result in a form of an approximation.

Approximation 4.14 *Approximation for the* $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}]$.
When the variance of $Y^{obs}|\mathbf{x}^{obs}$ *is constant* v^{*obs} *(homoscedastic situation) an approximation for the variance of the first moment* $\hat{\mu}^{comp,L}$ *given* n *observations is*

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}] &= \mathbb{E}_{\mathsf{N}^{mis}, \mathbf{D}^{train}_{\mathsf{N}^{obs}}, \hat{\boldsymbol{\beta}}^{obs}}\Big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathcal{Q}_2]\Big] \\
&\quad + \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}, \mathbf{D}^{train}_{\mathsf{N}^{obs}}, \hat{\boldsymbol{\beta}}^{obs}}\Big[\mathbb{E}[\hat{\mu}^{comp,L}|\mathcal{Q}_2]\Big] \\
&\approx T_1 + T_2 + \underbrace{\qquad C \qquad}_{\text{the variance due to noise modelling}} + \underbrace{O\big(\mathsf{n}^{-1}\big)}_{\text{approximation error}} ,
\end{aligned}$$

where T_1 *is an approximation for* $\mathbb{E}_{\mathsf{N}^{mis}, \mathbf{D}^{train}_{\mathsf{N}^{obs}}, \hat{\boldsymbol{\beta}}^{obs}}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2]]$ *thus*

$$\begin{aligned}
T_1 &= \underbrace{\frac{v^{*obs}}{\mathsf{n}^2}\frac{p^*}{1-p^*}\mathrm{tr}\bigg(\big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\big)^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\bigg)}_{\text{variance due to estimated coefficients}} \\
&\quad + \underbrace{\frac{p^*}{\mathsf{n}}\mathbb{V}\mathrm{ar}\bigg[\mathbb{E}_{\mathbf{D}^{train}_{\mathsf{N}^{obs}}|\mathsf{n}}\Big[\Big((\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}}\Big)^{-1}(\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{Y}\Big]^T\boldsymbol{X}^{mis}\bigg]}_{\text{variability of expected model}}
\end{aligned}$$

and T_2 *is an approximation for* $\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\mathbf{D}^{train}_{\mathsf{N}^{obs}},\hat{\boldsymbol{\beta}}^{obs}}[\mathbb{E}[\hat{\mu}^{comp,L}|\mathcal{Q}_2]]$ *hence*

$$
\begin{aligned}
T_2 = {} & \underbrace{\frac{1}{\mathsf{n}^2}\Big(\mathsf{n}(1-p^*)\tau^{*obs} + (\mu^{*obs})^2\mathbb{V}\mathrm{ar}[\mathsf{N}^{obs}]\Big)}_{\text{sampling variance (due to } \hat{\mu}^{\text{obs}} \text{ and } \mathsf{N}^{\text{obs}})} \\
+ \quad & \underbrace{\frac{1}{\mathsf{n}}v^{*obs}\frac{(p^*)^2}{1-p^*}(\overline{\boldsymbol{X}}^{*mis})^T\Big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\Big)^{-1}\overline{\boldsymbol{X}}^{*mis}}_{\text{imputation variance part 1}} \\
+ \quad & \underbrace{\frac{1}{\mathsf{n}^2}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\left(\mathbb{E}\left[\Big((\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}}\Big)^{-1}(\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{Y}|\mathsf{n}\right]^T\overline{\boldsymbol{X}}^{*mis}\right)^2}_{\text{imputation variance part 2}} \\
+ \quad & \underbrace{\frac{2}{\mathsf{n}^2}\left[\mathsf{n}p^* + \mu^{*obs}\mathbb{E}\left[\Big((\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}}\Big)^{-1}(\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{Y}|\mathsf{n}\right]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\right]}_{\text{cross term (covariance)}},
\end{aligned}
$$

in which

$$
C = \begin{cases}
0 & :S{=}M \ \text{(mean)}, \\
\frac{p^*}{\mathsf{n}}\Bigg(v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\left[\Big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0]\Big)^2\right] \\
\qquad +O\Big(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3}\Big)\Bigg) & :S{=}R \ \text{(random)}
\end{cases}
$$

One should remark that Approximation 4.14 has been derived by approximating the variance of regression coefficients without the intercept term. Underestimation of variance is compensated by the approximation error term. Even approximation for variance computed in Approximation 4.14 is quite complicated. However, one can notice that variance is function of sample size, proportion of missing data values, variance for noise for observed Y values, and on the first two moments of $\boldsymbol{X}^{obs}$ and $\boldsymbol{X}^{mis}$. Provided $\overline{\boldsymbol{X}}^{*mis} = \mathbf{0}$ then a more usable approximation of $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}]$ can be given in the form of the following consequence.

Consequence 4.15 *Simplification of* $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}]$.

Provided $\overline{\boldsymbol{X}}^{*mis} = \mathbf{0}$ *then for* $\mathcal{Q}_1 = \{\mathsf{n}\}$

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}] \;\approx\;& \underbrace{\frac{v^{*obs}}{\mathsf{n}^2}\frac{p^*}{1-p^*}\mathrm{tr}\Big(\big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\big)^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Big)}_{\text{imputation variance part 1 (due to estimated coefficients)}} \\
&\underbrace{+\frac{p^*}{\mathsf{n}}\mathbb{V}\mathrm{ar}\left[\mathbb{E}_{\mathbf{D}^{train}_{\mathsf{N}^{obs}}|\mathsf{n}}\left[\Big((\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}}\Big)^{-1}(\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{Y}\right]^T\boldsymbol{X}^{mis}\right]}_{\text{imputation variance part 2 (variability of expected model)}} \\
&\underbrace{+\frac{1}{\mathsf{n}^2}\Big(\mathsf{n}(1-p^*)\tau^{*obs} + (\mu^{*obs})^2\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\Big)}_{\text{sampling variance (due to } \hat{\mu}^{\mathrm{obs}} \text{ and } \mathsf{N}^{\mathrm{mis}})} + \underbrace{\quad\frac{2}{\mathsf{n}}p^*\quad}_{\text{cross term (covariance)}} \\
&+\underbrace{\qquad C \qquad}_{\text{variance due to noise modelling}} + \underbrace{\quad O\big(\mathsf{n}^{-1}\big)\quad}_{\text{approximation error}},
\end{aligned}$$

where term C *is as earlier.*

Clearly the approximation of variance of $\hat{\mu}^{comp,L}$ goes to zero as sample size goes to infinity.

The bias of the second moment $\hat{\tau}^{comp,L}$ depends on the variance of covariates $\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}$, the model fit and the true simulated noise level, as given in the following approximation.

Approximation 4.16 *Approximation for the bias of* $\hat{\tau}^{comp,L}$.
Bias of $\hat{\tau}^{comp,L}$ *for* n *observations can be approximated with*

$$\begin{aligned}
&\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,L}|\mathsf{n}] \approx \underbrace{p^*\Big(\mathrm{tr}(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) + C - \tau^{*mis}\Big)}_{A} \\
&\underbrace{+\,p^*(1-p^*)\Big[(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 - (\mu^{*obs} - \mu^{*mis})^2\Big]}_{B} \\
&+O(\mathsf{n}^{-1}),
\end{aligned}$$

where term A *is due to difference between the variance of the imputed and missing* Y *values and* B *is due to model missmatch. Term* A *varies for imputation strategies: namely, added imputation variance* C *is*

$$C = \begin{cases}
0 & :S{=}M \;\text{(mean imputation)} \\[2ex]
v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}]\big)^2\Big] & \\
\quad +O\Big(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3}\Big) & :S{=}R \;\text{(random imputation)}
\end{cases}$$

Consequence 4.17 *Approximation for the asymptotic bias of $\hat{\tau}^{comp,L}$.*
Asymptotically we have the following approximation

$$\lim_{\mathsf{n}\to\infty} \mathbb{B}\text{ias}[\hat{\tau}^{comp,L}] \approx \underbrace{p^*\Big((\boldsymbol{\beta}_{-0}^{*obs})^T\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis}\boldsymbol{\beta}_{-0}^{*obs} + Cv^{*obs,L,R} - \tau^{*mis}\Big)}_{A}$$
$$+ \underbrace{p^*(1-p^*)\Bigg\{\Big[\mu^{*obs} - (\boldsymbol{\beta}_{-0}^{*obs})^T\overline{\boldsymbol{X}}^{*mis} - (\beta_0^{*obs})\Big]^2 - (\mu^{*obs} - \mu^{*mis})^2\Bigg\}}_{B}, \tag{4.2}$$

where

$$C = \begin{cases} 0 & :S{=}M \quad \text{(mean imputation), and} \\ 1 & :S{=}R \quad \text{(simulated random imputation).} \end{cases}$$

Term A is bias due to difference between the variance of imputed and missing Y values. The difference between the mean of the imputed and missing Y values is measured by bias term B. Further, $v^{*obs,L,R} = \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{v}^{obs,L,R}]$ is the optimal noise variance parameter over all possible training data.

The justifications for Approximations 4.14-4.16 and Consequence 4.17 are given in Appendix A4.6.

4.2.2 An example: the role of conditionalization levels

As we see from Approximation 4.14, the finite sample error of the first moment of $\hat{\mu}^{comp,L}$ contains many variance components. The actual role of these components is difficult to interpret without conditionalizations $\mathcal{Q}_1$, $\mathcal{Q}_2$, and $\mathcal{Q}_3$. Therefore we shall study the differences between

$$\begin{aligned} Err_1 &= \mathbb{V}\text{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_1], \text{ where } \mathcal{Q}_1 = \{\mathsf{n}\} \\ Err_2 &= \mathbb{E}\Big[\mathbb{V}\text{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2]\Big], \text{ where } \mathcal{Q}_2 = \{\mathsf{n}, \mathbf{d}_{\mathsf{n}^{obs}}^{train}, \boldsymbol{\beta}^{obs}\}, \text{ and} \\ Err_3 &= \mathbb{E}\Big[\mathbb{V}\text{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_3]\Big], \text{ where } \mathcal{Q}_3 = \{\mathbf{d}_{\mathsf{n}^{obs}}^{train}, \mathbf{d}_{\mathsf{n}^{mis}}^{test}, \boldsymbol{\beta}^{obs}\}, \end{aligned}$$

In addition we shall compute the sampling variance

$$\mathbb{V}\text{ar}[\hat{\mu}^{obs}|\mathsf{n}] \approx \tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\text{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big)$$

for comparison purposes.

Note that error component Err_1 contains Err_2 which contains Err_3. Thus we know that

$$\begin{aligned} Err_1 &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathsf{n}] \\ &= Err_2 + \mathbb{V}\mathrm{ar}\Big[\mathbb{E}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2]\Big], \\ Err_2 &= \mathbb{E}\Big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2]\Big] \\ &= \mathbb{E}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\left(\hat{v}^{obs,L,R} + (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \hat{\boldsymbol{\beta}}^{obs}_{-0}\right)\right] \\ &= Err_3 + \mathbb{E}\Big[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \hat{\boldsymbol{\beta}}^{obs}_{-0}\Big]. \end{aligned} \tag{4.3}$$

The numerical values of components Err_1, Err_2, and Err_3 shall reveal us the importance of estimation errors and errors caused by imputation strategy.

Theoretical insight

Computation of Err_1 is done using Approximation 4.14. However, this requires

- superpopulation moments $\mu^{*obs} = \mathbb{E}[Y^{obs}]$ and $\tau^{*obs} = \mathbb{V}\mathrm{ar}[Y^{obs}]$,
- biased slope term $\beta^{*obs,biased}_{-0} = \mathbb{E}\left[\left((\mathbf{D}^{obs}_{\boldsymbol{X}})^T \mathbf{D}^{obs}_{\boldsymbol{X}}\right)^{-1} (\mathbf{D}^{obs}_{\boldsymbol{X}})^T \mathbf{D}^{obs}_{Y} |\mathsf{n}\right]$,
- variance of imputation noise (term C in Approximation 4.14), and
- the data generator from example 4.1.2, page 59 (which is used here too).

Therefore the moments have already been computed (see Equation 4.1 in page 61). The required four quantities are (see Appendix A4.7 for details on derivation)

$$\begin{aligned} \mu^{*obs} &= -0.196 \\ \tau^{*obs} &= 0.49074 \\ \beta^{*obs,biased}_{-0} &= 1051/9500 \\ C &\approx \underbrace{0.15}_{\mathbb{E}_{X^{obs}}[\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs}]]} + \underbrace{69639/475000}_{\text{squared bias}}, \end{aligned}$$

where the squared bias is $\int_{-\infty}^{\infty}(\frac{1}{500}x^3 - \beta^{*obs,biased}_{-0} * x)^2 f_{X^{obs}}(x)dx$.

Given the above quantities error term Err_1 may be computed using Approximation 4.14. However, a bit of work needs to be done in order to compute Err_2 and Err_3. Quantity Err_2 is easiest to compute using the decomposition given in Equation (4.3). By applying the Taylor approximation to the second term in the decomposition of Err_2 and using $\beta^{*obs,biased}_{-0}$ one gets

$$\begin{aligned} \mathbb{E}\Big[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \hat{\boldsymbol{\beta}}^{obs}_{-0}\Big] &\approx \mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}] \\ &\approx \frac{p^*}{\mathsf{n}}\beta^{*obs,biased}_{-0} * 15 * \beta^{*obs,biased}_{-0} \\ &= \frac{0.5}{\mathsf{n}} * \frac{1051}{9500} * 15 * \frac{1051}{9500}. \end{aligned}$$

Error term Err_3 is approximated as

$$\begin{aligned} Err_3 &= \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_i]] = \mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\hat{v}^{obs,L,R}] \\ &\overset{Taylor}{\approx} \mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}]\mathbb{E}[\hat{v}^{obs,L,R}] \approx \frac{p^*}{\mathsf{n}}C = \frac{0.5}{\mathsf{n}} * (0.15 + 69639/475000). \end{aligned}$$

The sampling variance equals approximately to $\frac{0.98}{\mathsf{n}} + \frac{2}{\mathsf{n}^2}$ in this example.

Data generator, the setup of the problem, and imputation model

We shall use the same data generator that was introduced in example 4.1.2 (page 59). However, the sample size range is a bit different, it is specified now as $\mathsf{n} = \lfloor 25 * 1.6^k \rfloor$, where $k = 0, \ldots, 9$. The data cannot be explained by a linear model but unlike with the previous example, it is possible to get a roughly unbiased first moment of Y also with MAR type missingness. The data and an example estimate of the linear model is shown in Figure 4.4, and the scatter between true and imputed values is shown in Figure 4.5.

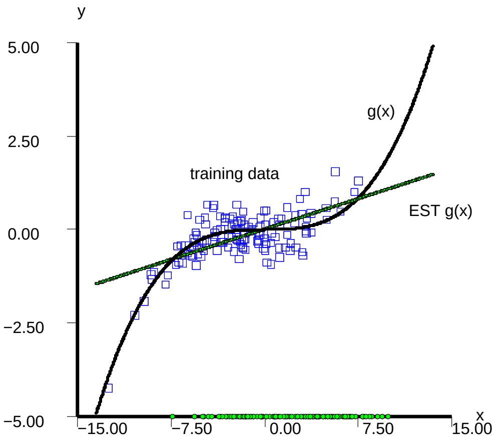


Figure 4.4: True model g(x), estimated linear model EST g(x), training data, and X positions for missing Y values (on X-axis).

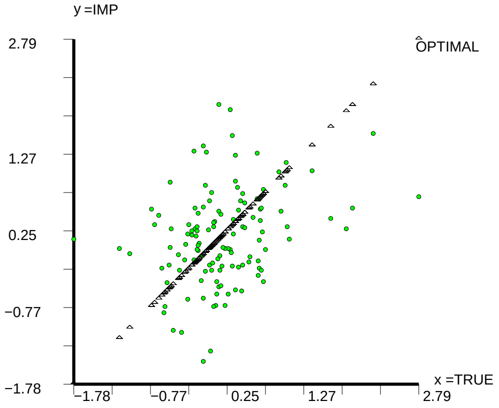


Figure 4.5: Scatter plot of true and imputed Y values. Optimal predictions are shown on diagonal.

The symmetry of scatter in Figure 4.5 indicates unbiasedness, whereas the spread is due to $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathsf{n}]$. Underestimation of missing Y values occur when the value of covariate is approximately between $[-7.5, 0]$, whereas the target values are overestimated when in $[0, 7.5]$. This is seen in Figure 4.4.

Results

The simulated results for conditionalized variances Err_1, Err_2 and Err_3 are summarized in Figure 4.6, and the simulated results for benchmark measure $\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}]$ are given in Figure 4.7. Further, the results are summarized in Table 4.2, that consists of both analytical and simulated results.

The roles of conditionalized variance components $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_1]$, $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2]$ and $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_3]$ are shown in Figure 4.6. Conditionalising by Q_2 drops variance approximately by 75% (when comparing to Q_1). Further, variance is dropped roughly by 88% when conditionalising by $\mathcal{Q}_3$. One can notice that $Err2 \approx 2 * Err3$. Variance increase is highest from the conditionalisation $\mathcal{Q}_2$ to $\mathcal{Q}_1$. This can be expected as many quantities become simultaneously random.

From the simulated results in Table 4.2 and Figures 4.6-4.7 we can see that the sampling variance for $\hat{\mu}^{obs}$ is larger than the variance of $\hat{\mu}^{comp,L,R}$ at $\mathcal{Q}_1$ (Err1) at least for a small sample size. The obvious reasons for this are the use of linear model and the increase of observations in the completed data $\mathbf{d}^{comp}$ against partially observed data $\mathbf{d}^{train}$ that decreases the variance error. This is good news, since also the bias between μ^* and $\hat{\mu}^{comp}$ is decreased. The bias for linear regression with the random strategy is approximately 0.02, and the bias is approximately -0.20 for the baseline mean strategy. Thus, in this example linear imputation clearly decreases imputation variance.

From Table 4.2 one can see that the approximations for $Err2$ and $Err3$ are quite accurate even for small sample size n. Approximation for $Err1$ yields too large a variance (roughly twice as large as it should be). The approximations (Taylor and large sample) which have been used to derive the analytical formula may be inaccurate. Note that absolute values are typically more difficult to get correct than relative values. Therefore we also computed analytically and numerically the relative efficiency

$$\mathrm{Eff}[\hat{\mu}^{comp}|\mathsf{n}] = \frac{\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathsf{n}]}{\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathsf{n}]} = \frac{Err_1}{Err_1 - Err_3},$$

where $\hat{\mu}^{comp,L,M}$ is used as the reference estimator. The efficiency results are given in Table 4.3. Theoretical efficiencies give a slightly optimistic view of the L,R method. Namely, the efficiency there is better (closer to one) than according to simulations.

n	25 k=0	40	64	102	163	262 k=5	419	671	1073	1717 k=9
$Err1$(simulated)	0.0324	0.0198	0.0132	0.0082	0.0053	0.0030	0.0019	0.0012	0.0007	0.0005
$Err1$(theoretical)	0.0602	0.0376	0.0235	0.0147	0.0092	0.0057	0.0036	0.0022	0.0014	0.0009
$Err2$(simul.)	0.0084	0.0053	0.0035	0.0022	0.0014	0.0009	0.0006	0.0003	0.0002	0.0001
$Err2$(theor.)	0.0096	0.0060	0.0038	0.0024	0.0015	0.0009	0.0006	0.0004	0.0002	0.0001
$Err3$(simul.)	0.0042	0.0032	0.0020	0.0013	0.0009	0.0005	0.0003	0.0002	0.0001	0.0001
$Err3$(theor.)	0.0059	0.0037	0.0023	0.0015	0.0009	0.0006	0.0004	0.0002	0.0001	0.0001
$\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}\|\mathsf{n}]$ (theor.)	0.0424	0.0258	0.0158	0.0098	0.0061	0.0038	0.0024	0.0015	0.0009	0.0006
$\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}\|\mathsf{n}]$ (simul.)	0.0359	0.0228	0.0162	0.0094	0.0056	0.0036	0.0021	0.0014	0.0008	0.0005

Table 4.2: Simulated and theoretical conditionalized variances $Err_i = \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_i]], i = 1, 2, 3$ as functions of the sample size $\mathsf{n} = \lfloor 25 * 1.6^k \rfloor$. Variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}]$ (both theoretical and simulated) is included for comparison with Err_1.

n	25 k=0	40	64	102	163	262 k=5	419	671	1073	1717 k=9
$\mathrm{Eff}[\hat{\mu}^{comp}\vert\mathsf{n}]$(simulated)	1.1489	1.1928	1.1786	1.1884	1.2045	1.2000	1.1875	1.2000	1.1667	1.2500
$\mathrm{Eff}[\hat{\mu}^{comp}\vert\mathsf{n}]$(theoretical)	1.1092	1.1094	1.1095	1.1095	1.1010	1.1096	1.1096	1.1096	1.1096	1.1096

Table 4.3: Simulated and theoretical efficiencies as functions of the sample size.

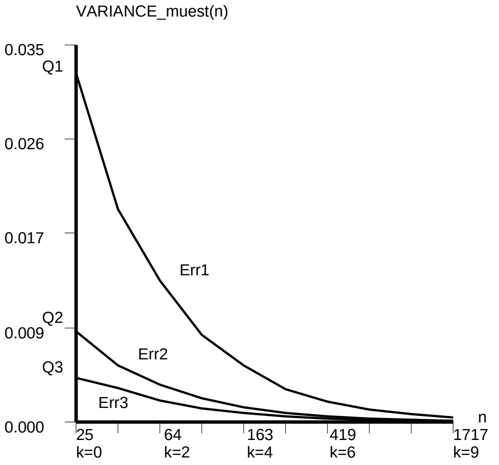


Figure 4.6: Simulated behaviour of decomposed variances $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathsf{n}]$, $\mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2]]$ and $\mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_3]]$ with increasing number of observations.

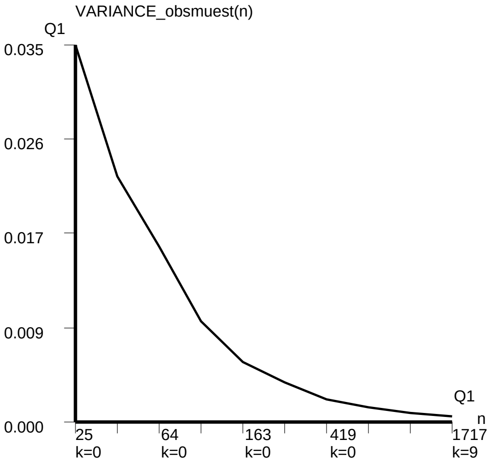


Figure 4.7: Behaviour of sampling variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}]$ with increasing number of observations.

4.2.3 Unit level prediction errors with linear regression

Linear regression methods are more flexible than baseline methods. As a consequence (squared) prediction bias is expected to be lower. There are cases for which the baseline methods perform better. As an example, if a true model is nonlinear and the missing-data mechanism is MAR then linear regression may lead to larger absolute bias of the first moment estimator due to extrapolation errors. Naturally linear regression methods are best at a unit level if $\mathbb{E}[Y^{mis}|\mathbf{x}]$ is close to $\mathbf{x}^T\boldsymbol{\beta}^{*obs}$. These and other considerations are verified by Approximations 4.18 and 4.19.

Approximation 4.18 *Approximation for the expectation of $\widehat{mse}(Y^{comp}|n)$ for linear regression*

Provided variance of $Y^{obs}|\mathbf{x}^{obs}$ is constant v^{*obs} (homoscedastic situation) the expec-

tation of $\hat{mse}(Y^{comp,L})$ with n observations can be approximated as:

$$
\begin{aligned}
\mathbb{E}[\hat{mse}(Y^{comp,L})|\mathsf{n}] \approx\ & \underbrace{\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]}_{\text{variability of approximative model}} \\
+ \quad & \underbrace{\left(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}] - \mu^{*mis}\right)^2}_{\text{global bias}} + \underbrace{\mathbb{V}\text{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
+ \quad & \underbrace{2\left(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}]\right)\left(\mu^{*mis} - g^{*mis}(\overline{\boldsymbol{X}}^{*mis})\right)}_{\text{cross term}} \\
+ \quad & \underbrace{v^{*obs}\left(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\text{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3}\right)\text{tr}(\mathbf{A})}_{\text{expected variance of approximative model predictions}} \\
+ \quad & \underbrace{C}_{\text{expected imputation variance}} + \underbrace{v^{*mis}}_{\text{expected target variance}} + \underbrace{O\left(\mathsf{n}^{-1}\right)}_{\text{approximation error}} ,
\end{aligned}
$$

where $\mathbf{A} = \left(\left(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*mis}(\overline{\boldsymbol{X}}^{*mis})^T\right)\left(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1}\right)$ *and term C depends on imputation strategy S:*

$$
C = \begin{cases}
0 & :S{=}M \text{ mean imputation} \\
v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\left[\left(g^{*obs}\left(\boldsymbol{X}^{obs}\right) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}]\right)^2\right] & \\
\quad +O\left(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\text{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3}\right) & :S{=}R \text{ random imputation}
\end{cases}
$$

In Approximation 4.18 the first three lines of $\mathbb{E}[\hat{mse}(Y^{comp,L})|\mathsf{n}]$ result from the expected squared bias. Further, the expected variance of approximative model predictions (row 4) is formally defined as $\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}[\mathbb{V}\text{ar}[(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T \boldsymbol{X}^{mis} + \hat{\beta}^{obs}_{0}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]]$. Furthermore, one should note the following remarks:

- part 3 of the expected squared bias has been partly approximated using the first order Taylor expansion, and
- expected sampling variance has been approximated using predictions done by the coefficient estimator without the intercept term (the approximation error is used to compensate the lack of the intercept term in the analysis).

Approximation 4.19 *Approximation for* $\text{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n})$.
Provided variance of $Y^{obs}|\mathbf{x}$ is constant v^{*obs} for all $\mathbf{x}$ (homoscedastic situation) the

mean squared error $\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n})$ *can be approximated as:*

$$
\begin{aligned}
\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}) \;\approx\; & \underbrace{\Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis}, \mathsf{n}]^T \mathbf{x}^{mis} + \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}^{mis}, \mathsf{n}] - g^{*mis}(\mathbf{x}^{mis})\Big)^2}_{\text{prediction bias}} \\
& \underbrace{+ (\mathbf{x}^{mis})^T \frac{v^{*obs}}{\mathsf{n}^{obs}} \left(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1} \mathbf{x}^{mis}}_{\text{sampling variance (slopes)}} \\
& + \underbrace{\quad C \quad}_{\text{imputation variance}} + \underbrace{v^{*mis}(\mathbf{x}^{mis})}_{\text{target variance}} + \underbrace{O\big((\mathsf{n}^{obs})^{-1}\big)}_{\text{approximation error}} \quad .
\end{aligned}
$$

where constant C depends on imputation strategy S:

$$
C = \begin{cases}
0 & \textit{:S=M} \text{ (mean imputation),} \\
\underbrace{v^{*obs}}_{\text{expectation of variance of } \mathrm{Y}^{\mathrm{obs}}|\boldsymbol{X}^{\mathrm{obs}}} \\
+ \underbrace{\mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}]\big)^2\Big]}_{\text{expected squared bias}} + \underbrace{O\big((\mathsf{n}^{obs})^{-1}\big)}_{\text{sampling variance}} & \textit{:S=R} \text{ (random imputation).}
\end{cases}
$$

Remark:

- sampling variance has been approximated using the predictions done by the coefficient estimator without the intercept term (approximation error is used to compensate the lack of the intercept term in the analysis).

Lemma 4.20 *Simplification to the trace of the matrix* $\mathbf{A}$ *(expected sampling variance in Approximation 4.18).*
Provided that the second moments and the square of the first moments of $\boldsymbol{X}^{obs}$ *and* $\boldsymbol{X}^{mis}$ *are equal (holds always under MCAR), then*

$$\mathrm{tr}(\mathbf{A}) = p - 1.$$

Recall that the cross term in Approximation 4.18 is close to minus two times the variability of the true model when the MAR missingness and linear model assumption hold. Therefore one can conclude that the expected squared bias (sum of the first three lines of the result) is roughly zero under MAR when the linear model assumption holds, formally

$$\mathbb{E}[Y^{mis}|\mathbf{x}] \approx \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T \mathbf{x} + \mathbb{E}[\hat{\beta}^{obs}_0].$$

The expected variance of approximative model predictions is somewhat complicated to interpret in general case. However, provided that the assumptions of Lemma

4.20 hold its interpretation becomes clear. Then the quantity is $(p-1)v^{*obs}O(\mathsf{n}^{-1})$. Therefore the expected variance is significant if v^{*obs} is large and ratio of dimension of covariate, $p-1$, and sample size n are large. Thus, for a small sample size with many covariates and high residual variance the expected variance is significant. However, asymptotically its impact vanishes. Further, the role of imputation variance (term C) for random strategy is significant if v^{*obs} is large or if the expected squared bias is large. In such case the estimated noise variance is large.

Similar considerations as above hold also for Approximation 4.19. Prediction bias at $\mathbf{x}^{mis}$ is large if conditional mean $g^{*mis}(\mathbf{x}^{mis})$ differs considerably from the expected linear prediction. Further, prediction variance (sampling variance term) is quadratic in position $\mathbf{x}^{mis}$. Secondly, it is controlled by the variance of regression coefficients $\hat{\boldsymbol{\beta}}^{obs}$. The variance of the coefficients is a function of residual variance v^{*obs}, size of training data $\mathbf{D}^{train}$, and the expectation of square of $\boldsymbol{X}^{obs}$. The latter equals to the variance of $\boldsymbol{X}^{obs}$ plus the expectation of $\boldsymbol{X}^{obs}$ squared. Therefore large training data, large variance and expectation of $\boldsymbol{X}^{obs}$ decrease prediction variance. On the other hand, high residual variance v^{*obs} and large norm $||\mathbf{x}^{mis}||$ increase the prediction variance.

Consequence 4.21 *Approximation to asymptotics of* $\mathbb{E}[\hat{mse}(Y^{comp,L})]$.
Limit of expectation of $\hat{mse}(Y^{comp,L})$ *can be approximated as:*

$$\begin{aligned}\lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{mse}(Y^{comp,L})|\mathsf{n}] \approx & \quad \underbrace{(\boldsymbol{\beta}^{*obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \boldsymbol{\beta}^{*obs}_{-0}}_{\text{variability of limit of approximative model}} \\ + & \quad \underbrace{\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} + \beta^{*obs}_0 - \mu^{*mis}\Big)^2}_{\text{asymptotic global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\ + & \quad \underbrace{2\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} + \beta^{*obs}_0\Big)\Big(\mu^{*mis} - g^{*mis}(\overline{\boldsymbol{X}}^{*mis})\Big)}_{\text{cross term}} \\ & + \underbrace{\quad C \quad}_{\text{asymptotic imputation variance}} + \underbrace{\quad v^{*mis} \quad}_{\text{expected target variance}},\end{aligned}$$

where term C depends on the imputation strategy S:

$$C = \begin{cases} 0 & :S{=}M \quad \text{(mean imputation)} \\ & \\ v^{*obs,L,R} & :S{=}R \quad \text{(simulated random imputation)}, \end{cases}$$

in which $v^{*obs,L,R} = \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]$. *The expectation* $\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]$ *is decomposed in Approximation 4.16 (see term C for random strategy).*

For compactness we give next only some ideas about justifications to Approximations 4.18 and 4.19. Full justifications, which are in Appendix A4.8 , are omitted here due to mathematical complexity (technical details) involved. The proof for Lemma 4.20 is given last, and a justification to Consequence 4.21 is omitted as it is straightforward.

The idea behind justifying Approximation 4.18 is to integrate the result of the theorem with respect to the joint distribution of the number of missing data values N^{mis} and $\boldsymbol{X}^{mis}$ given sample size n. In practice, we do this by first integrating over the distribution of $\boldsymbol{X}^{mis}$ and then over N^{mis}.

Approximation needs to be used in order to ease the derivation of Approximation 4.19. In the derivation of squared bias (term C) we assume that $\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis}, \mathsf{n}] \approx \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}] + O\big((\mathsf{n}^{obs})^{-1}\big)$. Further, variance is approximated by computing the variance for predictions done using the coefficients estimate without an intercept term. The lack of the intercept term in the analysis is compensated by the approximation error term. This approximation is quite rough but eased the computation of variance considerably.

Proof 4.22 *Proof to Lemma 4.20 is trivial as:*

$$\begin{aligned}
\mathrm{tr}(\mathbf{A}) &= \mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*mis}(\overline{\boldsymbol{X}}^{*mis})^T\right)\left(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1}\right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{*}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*}(\overline{\boldsymbol{X}}^{*})^T\right)\left(\boldsymbol{\Sigma}^{*}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*}(\overline{\boldsymbol{X}}^{*})^T\right)^{-1}\right) \\
&= \mathrm{tr}\left(\mathbf{I}_{(p-1)\times(p-1)}\right) = p-1 \qquad \square
\end{aligned}$$

4.3 Summary

For baseline methods it was found that reliability of estimated moments depends on the number of observations, the distribution of Y^{obs}, the distribution of the number of missing data values (N^{mis}), and on imputation strategy. Baseline methods are likely to yield high absolute biases of the first two moments under MAR and NMAR missingness. At unit level the baseline methods perform poorly because prediction is too stiff as a function of covariate: it is constant. As a consequence, squared prediction bias will be high. However, prediction variance for mean strategy is low.

A benefit of linear regression methods over baseline methods is that they are expected to perform better under MAR missingness. However, there are cases in which linear regression yields a higher absolute bias than baseline methods in first moment estimator. This is when the linear model assumption does not hold and predictions are badly biased. The bias of the first moment is dependent on how well the linear model assumption holds. Variance of the first moment is a function of sample size, a proportion of missing data values, variance for noise for observed Y values, and of first two moments of $\boldsymbol{X}^{obs}$ and $\boldsymbol{X}^{mis}$. For mean strategy the bias of the second moment depends on how close the variability of the model is to variability $\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]$. Random strategy may be used to reduce bias by modelling the noise terms. At unit level linear regression is likely to perform better than the baseline methods. However, there are exceptions (as in case of the first two moments). Expectation of the mean squared error estimator $\mathbb{E}[\hat{mse}(Y^{comp,L,S})]$

is quite complicated, but under MCAR is slightly simplified. If data generator is approximately linear and the variances of $\boldsymbol{X}^{obs}$ and $Y^{obs}|\boldsymbol{X}^{obs}$ are small or sample size is large then this quantity is low.

To summarize, if the data generator is close to linear then linear regression methods are likely to yield good results. A suitable nonlinear transform of data may render the linear regression methods usable even in nonlinear cases. Finding a transform, if such exists, requires user expertise. Under NMAR missingness, errors due to possible differences between conditional means $\mathbb{E}[Y^{obs}|\mathbf{x}]$ and $\mathbb{E}[Y^{mis}|\mathbf{x}]$ may render the validity of linearity assumption and estimation biases irrelevant.

Chapter 5

Imputation using nonparametric regression

Although the use of nonparametric regression for imputation is not the main topic of this thesis, it is an important family of methods that shares some properties with cell imputation. We shall therefore include nonparametric regression in this study, but we shall not use many pages to explain and interpret the results. As a consequence of this, the approximations of this chapter are more "technically involved" than in other chapters. There are, however, some clarifying examples.

A major problem with our baseline methods and linear regression is the fixed model assumption, which easily leads to a large bias in pointwise predictions. Nonparametric methods may provide an answer to this problem, but there are other benefits as well. From the imputation viewpoint it is especially interesting that we may change the role of variance components, as discussed in Chapter 3, page 51.

In this context we consider two closely related nonparametric regression methods: the K-nearest neighbour regression and kernel regression. In our regression context these can be written in the form of Nadaraya-Watson kernel regression [70, 109] for the observed $\mathbf{d}^{train}_{\mathsf{n}^{obs}\times p}$ sample as

$$g^{obs,K}(\mathbf{x}_0) = \frac{\sum_{j=1}^{\mathsf{n}^{obs}} K\big(\mathbf{x}_0, \mathbf{x}_j\big) y_j}{\sum_{j=1}^{\mathsf{n}^{obs}} K\big(\mathbf{x}_0, \mathbf{x}_j\big)} = \sum_{j=1}^{\mathsf{n}^{obs}} \overline{K}\big(\mathbf{x}_0, \mathbf{x}_j\big) y_j = \sum_{j=1}^{\mathsf{n}^{obs}} w_j y_j, \tag{5.1}$$

where kernel function $K(\mathbf{x}_0, \mathbf{x})$ characterizes the method and $\overline{K}(\cdot,\cdot)$ is normalized kernel.

In kernel regression kernels are selected with fixed width (bandwidth) λ. Later we denote that bandwidth is a function of sample size by $\lambda(\mathsf{n}^{obs})$. Next we assume that the bandwidth parameter is the same for all the components of $\boldsymbol{X}$. This allows us to make a direct link between the kernel and the k-nearest neighbour regressions. Further, this is no restriction in scope of this thesis, because later we analyse univariate X based kernel regression. Typically

$$K(\mathbf{x}_0, \mathbf{x}) = K_\lambda\Big(\mathbf{x}_0 - \mathbf{x}\Big) = \frac{1}{\lambda^{p-1}} K\Big(\frac{\mathbf{x}_0 - \mathbf{x}}{\lambda}\Big), \text{ where} \tag{5.2}$$

where $K(\cdot)$ is the kernel function, $K_\lambda(\cdot) = \frac{1}{\lambda^{p-1}}K(\frac{\cdot}{\lambda})$ is the rescaled kernel with a bandwidth λ. Kernel function K and bandwidth λ define the shape of the kernel. Let $\boldsymbol{\xi} = \mathbf{x}_0 - \mathbf{x}$. Now $\lim_{||\boldsymbol{\xi}||\to\infty} K_\lambda(\boldsymbol{\xi}) = 0$. Next we introduce kernels in the probability density function form. The reason for this is that Rosenblatt's kernel regression results, which are given and applied later, require it [85]. The most typical choices are box kernel

$$K(\boldsymbol{\xi}) = \begin{cases} \frac{1}{v_{p-1}}, & \text{if } ||\boldsymbol{\xi}|| < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{5.3}$$

where v_{p-1} is volume of unit ball in $\mathbb{R}^{p-1}$, and (spherical) Gaussian kernel

$$K(\boldsymbol{\xi}) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^{p-1} \exp(-||\boldsymbol{\xi}||^2/2).$$

In the nearest neighbour case the kernel width λ is defined implicitly by a constant number of k-nearest neighbours. To be more specific the bandwidth is specified by the distance to the k:th nearest neighbour from the regression position $\mathbf{x}_0$. Therefore (generalized) k-nearest neighbour regression formula can be written using the kernel regression formula with the kernel:

$$K(\mathbf{x}_0, \mathbf{x}) = \frac{1}{d_k^{p-1}} K\Big(\frac{\boldsymbol{\xi}}{d_k}\Big), \tag{5.4}$$

where d_k is the distance to the k:th nearest neighbour. With k-nn one has to be more strict about kernel $K(\boldsymbol{\xi})$ than with kernel regression for applying Mack's results [72]. Namely, it has to be a probability density function and its support has to be $(-1, 1)$. The above formulation for a generalized k-nn is adapted from Mack [72].

In this thesis we consider ordinary k-nearest neighbour regression. This is derived by using the box kernel (5.3) in (5.4) that yields the most common form of $k - nn$:

$$g^{obs,N}(\mathbf{x}_0) = \frac{1}{k} \sum_{j \in N_k(\mathbf{x}_0)} y_j,$$

where the set of k-nearest neighbours to point $\mathbf{x}_0$ is defined as

$$N_k(\mathbf{x}_0) = \Big\{ j \Big| \ ||\mathbf{x}_0 - \mathbf{x}_j|| \le ||\mathbf{x}_0 - \mathbf{x}_l|| \Big\}, \ l \notin N_k(\mathbf{x}_0) \ \text{ and } \#(N_k(\mathbf{x}_0)) = k,$$

in which $\#\{A\}$ denotes the cardinality of set A.

In an imputation context kernel (and NN) regression can be applied to estimate the missing values of Y as a function of observed covariates $\mathbf{x}$. Thus

$$Y_j^{comp,K,S} = \begin{cases} y_j, & \text{when } r_j = 1, \\ g^{obs,K}\Big(\mathbf{x}_j^{obs}\Big) + \hat{\epsilon}_j^{K,S}, & \text{when } r_j = 0, \end{cases}$$

where $\hat{\epsilon}^{K,S}$ is an estimate of residual variance and depends on imputation strategies such that

$$\hat{\epsilon}^{K,S} \sim \begin{cases} 0 & \text{:S=M} \quad \text{(mean imputation strategy)}, \\ N(0, v^{obs,K,R}) & \text{:S=R} \quad \text{(simulated random imputation), and} \\ \{y_j - g^{obs,K}(\mathbf{x}_j^{obs})\}_{j=1}^{\mathrm{n}^{obs}} & \text{:S=D} \quad \text{(random donor)}, \end{cases}$$

where $v^{obs,K,R} = \frac{1}{\mathsf{n}^{obs}} \sum_{j=1}^{\mathsf{n}^{obs}} \left(y_j - g^{obs,K}\left(\mathbf{x}_j^{obs}\right)\right)^2$ and in random donor strategy $\hat{\epsilon}$ is randomly drawn from the set $\{y_j - g^{obs,K}(\mathbf{x}_j^{obs})\}_{j=1}^{\mathsf{n}^{obs}}$ with equal probabilities.

A general property that is best seen with kernel (and NN) imputation is that given data $\mathbf{d}$, the variance of imputed values $\mathbb{V}\text{ar}[Y^{imp,K,S}|\mathbf{d}, \mathsf{n}]$ is explained by a mixture of model variance $\mathbb{V}\text{ar}[\hat{g}^{obs,K,S}|\mathbf{d}, \mathsf{n}]$ and residual variance $\mathbb{V}\text{ar}[\hat{\epsilon}^{K,S}|\mathbf{d}, \mathsf{n}]$. Due to the flexibility of nonparametric methods, we may get a good imputation performance, (also in terms of mse), with overestimated (too flexible) models. Thus, overtraining might not be harmful in the imputation context. This will be explained in the forthcoming example.

5.1 Preservation of moments

Due to the nonparametric nature of kernel (and NN) methods it is difficult to derive properties in fine details. However some analyses can be made using the general results of the kernel and nearest neighbour methods (see [72] and [85] for examples).

We can express the bias of the kernel (and NN) methods in the form of the following approximation.

Approximation 5.1 *An approximation for the bias of $\hat{\mu}^{comp,K/N}$*
The bias of the first moment for kernel and k-nn can be approximated with

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] \quad \approx \quad & \underbrace{p^*\Big(\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis}) - g^{*mis}(\boldsymbol{X}^{mis})]\Big)}_{\text{NMAR bias}} \\
+ \quad & \underbrace{\mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}} C\right]}_{\text{estimation bias wrt. } \mathrm{g}^{*\mathrm{obs}}(\mathbf{x}^{\mathrm{mis}})} \quad + \quad \underbrace{\qquad D \qquad}_{\text{bias due to noise estimation}} \\
+ \quad & \underbrace{O(\mathsf{n}^{-1})}_{\text{approximation term}} \quad ,
\end{aligned}$$

where terms $C = \mathbb{E}_{\boldsymbol{X}^{mis}}\left[\mathbb{B}\text{ias}[\hat{g}^{obs,K/N}(\mathbf{x}^{mis})|\mathsf{n}^{mis}, \mathsf{n}]\right]$ *(expected conditional mean estimation bias) and D vary according to kernel/k-nn and imputation strategy as*

$$C = \begin{cases}
\frac{(g^{*obs} f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis}) f''_{X^{obs}}(\overline{X}^{*mis})}{2 f_{X^{obs}}(\overline{X}^{*mis})} \int \xi^2 K(\xi) d\xi \lambda^2(\mathsf{n}^{obs}) & \\
+o\big(\lambda^2(\mathsf{n}^{obs})\big) + O\big((\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs}))^{-1}\big) & (\mathrm{Kernel, p = 2}) \\
& \\
\frac{(g^{*obs} f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis}) f''_{X^{obs}}(\overline{X}^{*mis})}{24 f^3_{X^{obs}}(\overline{X}^{*mis})} \big(k(\mathsf{n}^{obs})/\mathsf{n}^{obs}\big)^2 & \\
+o\big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^2\big) + O\big((k(\mathsf{n}^{obs}))^{-1}\big) & (\mathrm{K-nn, p = 2}) \\
& \\
\frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis}) - g^{*obs}(\overline{\boldsymbol{X}}^{*mis}) Q(f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis})}{2 f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis})(v_{p-1} f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis}))^{2/(p-1)}} (\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)} & \\
+o\Big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)}\Big) + O\big((k(\mathsf{n}^{obs}))^{-1}\big) & (\mathrm{K-nn, p > 2}),
\end{cases}$$

and

$$D = \begin{cases} 0 \\ \qquad\qquad\qquad\qquad :S{=}M/S{=}R \ \ \text{(mean and random strategy)} \\ p^*\mu^{*obs} - \mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\frac{1}{\mathsf{N}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\mathbb{E}_{\mathbf{D}^{train}|\mathsf{n},\mathsf{n}^{mis}}\left[\hat{g}^{obs,K/N}\left(\boldsymbol{X}_j\right)\right]\right] \\ \qquad\qquad\qquad\qquad :S{=}D \ \ \text{(random donor)} \end{cases}$$

where the second derivative of function $h(x)$ is denoted as $h''(x)$, and the product of functions $g(x)$ and $f(x)$ is denoted as $(gf)(x) = g(x)f(x)$,

$$Q(h)(\mathbf{x}) = \sum_{i=1,l=1}^{p-1,p-1}\int_{\mathbb{R}^{p-1}}\xi_i\xi_l\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_l}h(\mathbf{x})I(||\boldsymbol{\xi}||<1)*(1/v_{p-1})d\boldsymbol{\xi}, \tag{5.5}$$

$p-1$ is a dimension of $\boldsymbol{X}$, and v_{p-1} is the volume of the unit ball in $\boldsymbol{X}$ space which is $\mathbb{R}^{p-1}$.

Note that in Approximation 5.1 unit level bias terms require regularity conditions, as described in [72] and [85]. Further, we assume integrability of unit level bias over the distribution of $\boldsymbol{X}^{mis}$. For kernel predictions this means that a trimmed kernel estimate has to be used or more regularity conditions have to be set (see [14] for example).

From Approximation 5.1 one sees that bias consists of NMAR bias, estimation bias, noise term modelling bias, and an approximation term. The bias due to the NMAR mechanism is quite obvious as the true model for Y^{obs} and Y^{mis} may differ. The estimation technique, which is used to construct imputation model $\hat{g}^{obs}(\mathbf{x})$, also contributes to the bias of $\hat{\mu}^{comp}$. Further, random donor strategy for noise terms may also impact the bias.

Even though our approximation is quite rough one sees that the estimation bias increases quadratically as smoothing is increased. This holds for both the kernel and k-nn. One should also note that the effective number of parameters for k-nn is n^{obs}/k, and that the k-nn bias is inversely proportional to it. Namely, when the number of effective parameters grows the model becomes more flexible and thus less biased.

In the non-NMAR case the bias of $\hat{\mu}^{comp}$ is due to unit level prediction biases, which are dependent on the density of $\boldsymbol{X}^{obs}$. From Approximation 5.1 one observes the well-known difference in prediction biases between kernel and k-nn regression. The bias for kernel regression is inversely proportional to the density of $\boldsymbol{X}^{obs}$, and the bias for k-nn is inversely proportional to $f_{\boldsymbol{X}^{obs}}(x)^3$. Thus, the bias of $\hat{\mu}^{comp,K}$ can be considerable if density $f_{\boldsymbol{X}^{obs}}(\overline{X}^{*mis})$ is low.

Consequence 5.2 *Bounds for* $\mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}]$

The following bounds can be derived for the kernel and nearest neighbour methods

$$\begin{aligned} \lim_{\lambda\to\infty}\mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] &= \mathbb{B}\text{ias}[\hat{\mu}^{comp,B}|\mathsf{n}] = p^*(\mu^{*mis}-\mu^{*obs}) \\ \lim_{\lambda\to 0,\mathsf{n}\to\infty}\mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] &= p^*\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})-g^{*mis}(\boldsymbol{X}^{mis})]. \end{aligned}$$

In the first result we do not assume that smoothing is a function of sample size as is assumed in Approximation 5.1. The first result is obvious, since $\lambda \to \infty$ implies that the method will become same as baseline imputation. For k-nn this means that $k \to \mathsf{n}^{obs}$. In the second case imputation follows 1-nearest neighbour, which is undefined for kernel regression unless $\mathsf{n} \to \infty$.

The variance of the first moment depends also on the level of smoothing. Due to mathematical difficulty the result is left in an implicit form. However, it can be subjected to some interpretation as shown after the approximation.

Approximation 5.3 *Approximation for the variance of $\hat{\mu}^{comp,K/N}$*
The variance of the first moment for kernel and k-nn can be approximated with

$$
\begin{aligned}
& \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}] \approx \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[(\frac{\mathsf{N}^{obs}}{\mathsf{n}})^2 \underbrace{\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}^{mis}]}_{\text{sampling variance}} \\
+ \quad & (\frac{\mathsf{N}^{mis}}{\mathsf{n}})^2 \Bigg(\frac{1}{\mathsf{N}^{mis}} \underbrace{\Bigg(\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}} \Bigg[\underbrace{\mathbb{V}\mathrm{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis}, \mathsf{N}^{mis}, \mathsf{n}]}_{A} \Bigg]}_{\text{variance due to conditional mean prediction}} \\
+ \quad & \underbrace{\mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}} \Bigg[g^{*obs}(\boldsymbol{X}^{mis}) + \underbrace{\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis}, \mathsf{n}]}_{B} \Bigg] \Bigg)}_{\text{variance due to conditional mean prediction}} \\
+ \quad & \underbrace{O\Big((\mathsf{N}^{obs})^{-\frac{1}{2}}\Big)}_{\text{due to correlated predictions}} \qquad \Bigg) + \underbrace{2\frac{\mathsf{N}^{obs}\mathsf{N}^{mis}}{\mathsf{n}^2} O\Big((\mathsf{N}^{mis}\mathsf{N}^{obs})^{-\frac{1}{2}}\Big)}_{\text{approximation for cross term (covariance)}} \quad \Bigg] \\
+ \quad & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}} \Bigg[\underbrace{\Big(1 - \frac{\mathsf{N}^{mis}}{\mathsf{n}}\Big)\mu^{*obs} + \frac{\mathsf{N}^{mis}}{\mathsf{n}} \Big(\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})]}_{\mathbb{E}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}^{mis}]} \\
& \underbrace{+\mathbb{E}_{\boldsymbol{X}^{mis}}\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis}, \mathsf{n}] \Big)}_{\mathbb{E}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}^{mis}]} \Bigg] + \underbrace{\qquad C \qquad}_{\text{imputation noise variance}},
\end{aligned}
$$

where terms A, B, and C depend on the estimation method (kernel or k-nn) and on imputation strategy $\hat{\epsilon}^S$ as follows:

$$
A = \begin{cases} \frac{\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs}=X^{mis}]}{f_{X^{obs}}(X^{mis})\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})} \int K^2(\xi)d\xi + o\Big(\frac{1}{\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\Big) & (\mathrm{Kernel, p = 2}), \\ \\ \frac{v_{p-1}\mathbb{V}\mathrm{ar}[Y^{obs}|\boldsymbol{X}^{obs}=\mathbf{X}^{mis}]}{k(\mathsf{n}^{obs})} + o((k(\mathsf{n}^{obs}))^{-1}) & (\mathrm{K-nn, p \geq 2}), \end{cases}
$$

$$B = \begin{cases} \frac{(g^{*obs} f_{X^{obs}})''(X^{mis}) - g^{*obs}(X^{mis}) f''_{X^{obs}}(X^{mis})}{2 f_{X^{obs}}(X^{mis})} \int \xi^2 K(\xi) d\xi \lambda^2(\mathsf{n}^{obs}) & \\ +o(\lambda^2(\mathsf{n}^{obs})) + O((\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs}))^{-1}) & (\mathrm{Kernel, p = 2}), \\ \\ \frac{(g^{*obs} f_{X^{obs}})''(X^{mis}) - g^{*obs}(X^{mis}) f''_{X^{obs}}(X^{mis})}{24 f^3_{X^{obs}}(X^{mis})} (k(\mathsf{n}^{obs})/(\mathsf{n}^{obs}))^2 & \\ +o((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^2) + O((k(\mathsf{n}^{obs}))^{-1}) & (\mathrm{K - nn, p = 2}), \\ \\ \frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\boldsymbol{X}^{mis}) - g^{*obs}(\boldsymbol{X}^{mis}) Q(f_{\boldsymbol{X}^{obs}})(\boldsymbol{X}^{mis})}{2 f_{\boldsymbol{X}^{obs}}(\boldsymbol{X}^{mis})(v_{p-1} f_{\boldsymbol{X}^{obs}}(\boldsymbol{X}^{mis}))^{2/(p-1)}} \left(\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}}\right)^{2/(p-1)} & \\ +o\Big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)}\Big) + O((k(\mathsf{n}^{obs}))^{-1}) & (\mathrm{K - nn, p > 2}), \end{cases}$$

where $Q(h)(x)$ is defined in Equation (5.5) and

$$C = \begin{cases} 0 & \\ & :S{=}M \ \text{(mean imputation)} \\ \\ \frac{p^* v^{*obs}}{\mathsf{n}} + \frac{1}{\mathsf{n}^2} \mathbb{E}_{\mathsf{N}^{mis}} \left[\mathsf{N}^{mis} \mathbb{E}_{\boldsymbol{X}^{obs}} \left[\left(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis}}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs})] \right)^2 \right] \right] & \\ & :S{=}R, S{=}D \ \text{(random and donor strategies)} \end{cases}$$

One can notice from Approximation 5.3 that increasing smoothing will reduce variance due to conditional mean estimate (see term A). This is obvious as the model becomes more stiff. From the theorem one sees also the vulnerability of k-nn regression. Prediction variance for k-nn is inversely proportional to $f_{X^{obs}}(x_0)$ at a point x_0. Therefore if there are regions in the X space in which $f_{X^{obs}}(x)$ is low and $\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs} = x]$ is high then kernel regression may yield higher variance than k-nn (for suitable value of k) due to the conditional mean estimate. The following consequence will give bounds to variance:

Consequence 5.4 *Bounds for* $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}]$

$$\begin{aligned} \lim_{\lambda \to \infty} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B}|\mathsf{n}] \\ \lim_{\mathsf{n} \to \infty} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}] &\approx 0. \end{aligned}$$

Again the first result is obvious (and it requires that smoothing is not a function of sample size), since $\lambda \to \infty$ implies that the method will become the same as baseline imputation. For k-nn this means that $k(\mathsf{n}^{obs}) \to \mathsf{n}^{obs}$. The second case is also assumed to hold exactly: estimator $\hat{\mu}^{comp}$ converges asymptotically to its limit.

The bias of the second moment depends on both the model flexibility (amount of smoothing) and the imputation strategy. Due to technical difficulty the results are given in an implicit form in the following theorem.

Approximation 5.5 *Approximation for bias of $\hat{\tau}^{comp,K/N}$*
Approximate bias can be written as

$$\begin{aligned}
&\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}] \approx p^*(\underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big]}_{\text{variability of expected conditional mean estimate}}\\
+\quad &\underbrace{\qquad C \qquad}_{\text{imputation noise variance}} -\tau^{*mis})\\
+\quad &p^*(1-p^*)\Bigg[(\mu^{*obs}-\mathbb{E}[Y^{imp,K/N}|\mathsf{n}])^2-(\mu^{*obs}-\mu^{*mis})^2\Bigg]\\
+\quad &\underbrace{p^*\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{V}\mathrm{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big]}_{\text{expected sampling variance}}\\
+\quad &\underbrace{\qquad\qquad\qquad O(\mathsf{n}^{-1})\qquad\qquad\qquad}_{\text{sampling variance of }\hat{\mu}^{\mathrm{imp}}\text{ and approximation error (finite sample vs asymptotic)}},
\end{aligned}$$

where term C is

$$C=\begin{cases} 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad :S{=}M \ \text{(mean imputation)} \\ v^{*obs}+\underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{obs}}\Bigg[\Big(g^{*obs}(\boldsymbol{X}^{obs})-\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs})|\boldsymbol{X}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big)^2\Bigg]}_{\text{expected squared bias}} \\ +O(\mathsf{n}^{-1}) \qquad\qquad\qquad\qquad\qquad\qquad :S{=}R,D \ \text{(random strategies)} \end{cases}$$

and $\mathbb{E}[Y^{imp}|\mathsf{n}]=\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\Bigg[\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis},\mathsf{N}^{mis},\mathsf{n}]\Bigg]$ *is*

$$\mathbb{E}[Y^{imp}|\mathsf{n}]\approx\begin{cases} \mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})]+\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\Big[\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big] \\ \qquad\qquad\qquad\qquad\qquad :S{=}M{,}R \ \text{(mean and random)} \\ \\ \mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})]+\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\Big[\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big] \\ +\mu^{*obs}-\mathbb{E}[\frac{1}{\mathsf{N}^{obs}}\sum_{j=1}^{\mathsf{N}^{obs}}\hat{g}^{obs,K/N}\big(\boldsymbol{X}_j\big)] \\ \qquad\qquad\qquad\qquad\qquad :S{=}D \ \text{(random donor)} \end{cases}$$

where $\mathbb{B}\mathrm{ias}[\boldsymbol{X}^{mis}|\mathsf{N}^{mis},\mathsf{n}]$ *is estimation bias with respect to* $g^{*obs}(\boldsymbol{X}^{mis})$*.*

Even though the above approximative bias of $\hat{\tau}^{comp}$ is in an implicit form we can state the following considerations. The bias is affected by sample size, type of conditional mean estimate (kernel or k-nn), and smoothing. Sample size affects the estimation variances and biases. Smoothing has an impact on the variability of the expected conditional mean model, expected sampling variance, and expected prediction $\mathbb{E}[Y^{imp,K/N}|\mathsf{n}]$. If smoothing is increased then the variability of the expected model and sampling variance reduce. However, error in the form of a difference between $\mathbb{E}[Y^{imp,K/N}|\mathsf{n}]$ and μ^{*mis} is likely to grow. For random strategies decreasing smoothing will increase term C, because the expected squared bias grows. This is actually a trade-off between the variability of the conditional mean estimate and the amount of modelled noise terms. The bounds for the bias are given in the following consequence.

Consequence 5.6 *Bounds for* $\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}]$

$$\begin{aligned}
\lim_{\lambda\to\infty} \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}] &= \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,B}|\mathsf{n}] \\
\lim_{\lambda\to 0,\mathsf{n}\to\infty} \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}] &\approx p^*(\mathbb{V}\mathrm{ar}[g^{*obs}(\boldsymbol{X}^{mis})] + C - \tau^{*mis}) \\
&\quad +p^*(1-p^*)[(\mu^{*obs} - \mathbb{E}[g^{*obs}(\boldsymbol{X}^{mis})] - D)^2 \\
&\quad -(\mu^{*obs} - \mu^{*mis})^2],
\end{aligned}$$

where terms C and D depend on imputation strategy $\hat{\epsilon}^S$ as follows:

$$C = \begin{cases} 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad :S{=}M \ \text{(mean)}, \\ v^{*obs} + \lim \mathbb{E}_{\mathsf{N}^{mis}}\mathbb{E}_{\boldsymbol{X}^{obs}}\left[\Big(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis}}[\hat{g}^{obs,K/N}\big(\boldsymbol{X}^{obs}\big)]\Big)^2\right] \\ \qquad\qquad\qquad\qquad\qquad\qquad :S{=}R,D \ \text{(random)}, \end{cases}$$

and

$$D = \begin{cases} 0 \qquad\qquad :S{=}M,R \ \text{(for mean and random imputation)}, \\ \mu^{*obs} - \lim \mathbb{E}_{\mathsf{N}^{mis},\mathbf{D}^{train}}\left[\frac{1}{\mathsf{N}^{obs}}\sum_{j=1}^{\mathsf{N}^{obs}} \hat{g}^{obs,K/N}\big(\mathbf{x}_j\big)\right] \\ \qquad\qquad :S{=}D \ \text{(for random donor)}. \end{cases}$$

In the bounds above we have assumed that random donor strategy behaves asymptotically as simulated random imputation (term C).

Justifications for all the above approximations are given in Appendix A5.1.

5.2 Unit level prediction performance of nonparametric methods

The prediction performance of kernel (and NN) regression has been studied, e.g., by Rosenblatt [85], Mack [72], Gasser and Müller [28, 29]. From the viewpoint of the current thesis these results can be summarized as follows:

Bias and variance of nearest neighbour regression

Assuming n iid observations for training data, the following pointwise bias and variance for k-nn are based on Mack's theorems 1 and 2 [72]. Provided $k = o(\mathsf{n})$,

$\log \mathsf{n} = o(k)$, and suitable regularity conditions hold [72], then bias and variance for (ordinary) k-nearest neighbour can be written as

Theorem 5.7 *Application of Mack's results for* $\mathbb{B}\mathrm{ias}[\hat{g}^N(\mathbf{x}_0)]$ *and* $\mathbb{V}\mathrm{ar}[\hat{g}^N(\mathbf{x}_0)]$.

$$
\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{g}^N(\mathbf{x}_0)] &= \underbrace{g^{*obs}(\mathbf{x}_0) - g^{*mis}(\mathbf{x}_0)}_{\text{NMAR bias}} \\
&\quad + \underbrace{\frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0) - g^{*obs}(\mathbf{x}_0) Q(f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0)}{2 f_{\boldsymbol{X}^{obs}}(\mathbf{x}_0)(v_{p-1} f(\mathbf{x}_0))^{2/(p-1)}} \left(\frac{k}{\mathsf{n}}\right)^{2/(p-1)}}_{\text{estimation bias wrt. } \mathrm{g}^{*\mathrm{obs}}(\mathbf{x}_0)} \\
&\quad \underbrace{+ o\left(\left(\frac{k}{\mathsf{n}}\right)^{2/(p-1)}\right) + O(k^{-1})}_{\text{estimation bias wrt. } \mathrm{g}^{*\mathrm{obs}}(\mathbf{x}_0)} \\
\mathbb{V}\mathrm{ar}[\hat{g}^N(\mathbf{x}_0)] &= \frac{v_{p-1} \mathbb{V}\mathrm{ar}[Y^{obs} | \boldsymbol{X}^{obs} = \mathbf{x}_0]}{k} + o(k^{-1}),
\end{aligned}
$$

where $Q(h)(\mathbf{x}) = \sum_{i=1,l=1}^{p-1,p-1} \int_{\mathbb{R}^{p-1}} \xi_i \xi_l \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_l} h(\mathbf{x}) I(||\boldsymbol{\xi}|| < 1) \frac{1}{v_{p-1}} d\boldsymbol{\xi}$*, and volume of unit ball in* $\mathbb{R}^{p-1}$ *is* $v_{p-1} = \pi^{(p-1)/2} / \Gamma((p-1+2)/2)$.

Theorem 5.7 follows straightforwardly from Mack's theorems 1 and 2 [72]: only the NMAR bias term has been added. One should note that Mack derives the results for generalized k-nn regression (see Equation 5 in [72]). Some work is required to get the above results (see Appendix A5.4 for details). One must also recall that k-nn prediction is likely to be asymptotically biased in the NMAR case. The reason for this is that $g^{*mis}(\mathbf{x}_0)$ and $g^{*obs}(\mathbf{x}_0)$ are likely to be different, and the k-nn prediction converges to the latter one (provided smoothing is properly decreased as assumed by Mack).

Bias and variance of kernel regression

The following pointwise bias and variance results are based on Rosenblatt's theorem 2 [85], and are derived by Mack [72]. When $\lambda(\mathsf{n}) \to 0$ and $\mathsf{n}^{-1}\lambda(\mathsf{n})^{-1} = o(1)$ as $\mathsf{n} \to \infty$ and suitable regularity conditions hold (see [85] for details), then the asymptotic bias and variance for kernel regression with univariate X are as follows:

Theorem 5.8 *Application of Rosenblatt's and Mack's results for* $\mathbb{B}\mathrm{ias}[\hat{g}^K(\mathbf{x})]$ *and* $\mathbb{V}\mathrm{ar}[\hat{g}^K(\mathbf{x})]$.

$$\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{g}^K(x)] &= \underbrace{g^{*obs}(x_0) - g^{*mis}(x_0)}_{\text{NMAR bias}} \\
&+ \underbrace{\frac{(g^{*obs} f_{X^{obs}})''(x_0) - g^{*obs}(x_0) f''_{X^{obs}}(x_0)}{2 f_{X^{obs}}(x_0)} \int \xi^2 K(\xi) d\xi \lambda^2(n)}_{\text{estimation bias wrt. g}^{*\text{obs}}(\text{x}_0)} \\
&+ \underbrace{o\big(\lambda^2(n)\big) + O\big(n\lambda(n)\big)^{-1}\big)}_{\text{estimation bias wrt. g}^{*\text{obs}}(\text{x}_0)} \\
\mathbb{V}\mathrm{ar}[\hat{g}^K(x)] &= \frac{\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs} = x_0]}{f_{X^{obs}}(x_0) n\lambda(n)} \int K^2(\xi) d\xi + o(\frac{1}{n\lambda(n)}),
\end{aligned}$$

where $\lambda(n)$ denotes kernel bandwidth under n observations.

Again only NMAR bias has been added to Rosenblatt's and Mack's original results. For simplicity the study of kernel regression for multivariate X is omitted in the current thesis. For comparison between kernel and k-nn the following corollary gives univariate bias and variance for the nearest neighbour.

Corollary 5.9 *Bias and variance for (ordinary) k-nearest neighbour regression estimate for univariate X.*

$$\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{g}^N(x_0)] &= \underbrace{g^{*obs}(x_0) - g^{*mis}(x_0)}_{\text{NMAR bias}} \\
&+ \underbrace{\frac{(g^{*obs} f_{X^{obs}})''(x_0) - g^{*obs}(x_0) f''_{X^{obs}}(x_0)}{24 f^3_{X^{obs}}(x_0)} (k/n)^2 + o\Big(\big(\frac{k}{n}\big)^2\Big) + O(k^{-1})}_{\text{estimation bias wrt. g}^{*\text{obs}}(\text{x}_0)} \\
\mathbb{V}\mathrm{ar}[\hat{g}^N(x_0)] &= \frac{2\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs} = x_0]}{k} + o(k^{-1}).
\end{aligned}$$

Corollary 5.9 follows from Theorem 5.7. On the other hand, the bias and variance results for generalized k-nn regression for univariate X are given in Table 1 of Mack's paper [72]. The weight function for Mack's formula is $w(v) = 1$ when $|v| < 1$ and 0 otherwise. The corollary follows by substituting $\int v^2 w(v) dv = \frac{1}{3}$ and $\int w^2(v) dv = \frac{1}{2}$ to Mack's Table 1.

One may wonder how big are "kernel integral quantities" in Theorem 5.8 For illustration purpose we next compute them for gaussian $N(0,1)$ kernel.

Lemma 5.10 *Kernel integrals for gaussian $N(0,1)$ kernel.*

In case of univariate gaussian density function, $N(0,1)$, as kernel K following hold

$$\int \xi^2 K(\xi) d\xi = 1,$$

and

$$\int_{-\infty}^{\infty} K^2(\xi) d\xi = \frac{1}{2\sqrt{\pi}} \approx 0.282$$

Proof 5.11 *for Lemma 5.10. The second moment of kernel is* $\int \xi^2 K(\xi) d\xi = 1$ *because variance of* $K = N(0,1)$ *is one. Expectation of squared kernel, formally* $\int K^2(\xi) d\xi$*, is derived next. By using symbolical integration software (Mathematica) one gets*

$$\int_{-\infty}^{\infty} e^{-\xi^2/2} e^{-\xi^2/2} d\xi = \sqrt{\pi}.$$

As a consequence

$$\begin{aligned} \int_{-\infty}^{\infty} K^2(\xi) d\xi &= \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} \right]^2 d\xi \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\xi^2/2} e^{-\xi^2/2} d\xi = \frac{1}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}}. \end{aligned}$$

Approximations for MSE

We are now ready to summarize the results of unit level measures in terms of mean squared error. In the following we have the results for

i) $\mathrm{mse}^{K/N}(Y^{imp}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n})$ which is mse at a given point $\mathbf{x}_0$,

ii) $\mathbb{E}[\hat{mse}(Y^{comp,K/N})|\mathsf{n}]$ where i) is generalized over $f_{\mathbf{X}^{mis}}(\mathbf{x})$, and

iii) $\lim_{n\to\infty,\lambda\to 0} \mathbb{E}[\hat{mse}(Y^{comp,K/N})|\mathsf{n}]$ which is the simpler asymptotic result of ii).

In the imputation context, we are mainly interested about results ii) and iii). Both are, however, best explained via result i) which is explained next.

The computation of mse is based on previously given bias and variance terms. In the imputation context we need to add the role of the added noise term $\hat{\epsilon}^{K/N,S}$, which depends on our imputation strategy. This allows us to write $mse(\mathbf{x})$ in the form of the following theorem.

Approximation 5.12 *Approximative* $mse^{K/N}(Y^{imp}|\mathbf{x}_0, n^{mis}, n)$
Over distribution of training data set with n^{obs} *observations the mean squared error at point* $\mathbf{x}_0$ *can be approximated as follows:*

$$\begin{aligned} mse^{K/N}(Y^{imp}|\mathbf{x}_0, n^{mis}, n) \approx \Big(&\underbrace{g^{*obs}(\mathbf{x}_0) - g^{*mis}(\mathbf{x}_0)}_{\text{NMAR bias}} + \underbrace{\mathbb{B}\mathrm{ias}[\hat{g}^{obs,K/N}(\mathbf{x}_0)|n^{mis}, n]}_{\text{A: estimation bias wrt. } \mathrm{g}^{*\mathrm{obs}}(\mathbf{x}_0)} \\ + \quad &\underbrace{\qquad C \qquad}_{\text{Bias due to noise modelling}} \Big)^2 + \underbrace{\mathbb{V}\mathrm{ar}[\hat{g}^{obs,K/N}(\mathbf{x}_0)|n^{mis}, n]}_{\text{B: imputation model variance}} \\ + \quad &\underbrace{\qquad D \qquad}_{\text{imputation noise variance}} + \underbrace{v^{*mis}(\mathbf{x}_0)}_{\text{target variance}} . \end{aligned}$$

where terms A-D depend on nonparametric estimate (kernel vs k-nn) and imputation strategy as follows

$$A = \begin{cases} \frac{(g^{*obs} f_{X^{obs}})''(x_0) - g^{*obs}(x_0) f''_{X^{obs}}(x_0)}{2 f_{X^{obs}}(x_0)} \int \xi^2 K(\xi) d\xi \lambda^2(n^{obs}) & (\mathrm{Kernel, p = 2}), \\ +o(\lambda^2(n^{obs})) + O\big((n^{obs}\lambda(n^{obs}))^{-1}\big) & \\ & \\ \frac{(g^{*obs} f_{X^{obs}})''(x_0) - g^{*obs}(x_0) f''_{X^{obs}}(x_0)}{24 f^3_{X^{obs}}(x_0)} \big(k(n^{obs})/n^{obs}\big)^2 & (\mathrm{K - nn, p = 2}), \\ +o((\frac{k(n^{obs})}{n^{obs}})^2) + O\big((k(n^{obs}))^{-1}\big) & \\ & \\ \frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0) - g^{*obs}(\mathbf{x}_0) Q(f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0)}{2 f_{\boldsymbol{X}^{obs}}(\mathbf{x}_0)(v_{p-1} f(\mathbf{x}_0))^{2/(p-1)}} (\frac{k(n^{obs})}{n^{obs}})^{2/(p-1)} & (\mathrm{K - nn, p > 2}), \\ +o\Big((\frac{k(n^{obs})}{n^{obs}})^{2/(p-1)}\Big) + O\Big((k(n^{obs}))^{-1}\Big) & \end{cases}$$

$$B = \begin{cases} \frac{\mathbb{V}\mathrm{ar}[Y^{obs} | X^{obs} = x_0]}{f_{X^{obs}}(x_0) n^{obs} \lambda(n^{obs})} \int K^2(\xi) d\xi + o\big(\frac{1}{n^{obs}\lambda(n^{obs})}\big) & (\mathrm{Kernel, p = 2}), \\ & \\ \frac{v_{p-1} \mathbb{V}\mathrm{ar}[Y^{obs} | \boldsymbol{X}^{obs} = \mathbf{x}_0]}{k(n^{obs})} + o\Big((k(n^{obs}))^{-1}\Big) & (\mathrm{K - nn, p \geq 2}), \end{cases}$$

and

$$C = \begin{cases} 0 \\ \qquad\qquad\qquad\qquad :S{=}M{,}R \;\; \text{(mean and simulated random)}, \\ \\ \mu^{*obs} - \mathbb{E}\Big[\frac{1}{n^{obs}} \sum_{j=1}^{n^{obs}} \hat{g}^{obs,K/N}(X_j) | n^{mis}, n\Big] + O(n^{-1}) \\ \qquad\qquad\qquad\qquad :S{=}D \;\; \text{(random donor)}, \end{cases}$$

and

$$D = \begin{cases} 0 \\ \qquad\qquad\qquad\qquad :S{=}M \;\; \text{(mean)}, \\ \\ v^{*obs} + \mathbb{E}_{X^{obs}}\Big[(g^{*obs}(X^{obs}) - \mathbb{E}_{\mathbf{D}^{train}|n^{mis},n}[\hat{g}^{obs,K/N}(X^{obs}|n, n^{mis}])^2\Big] \\ \qquad\qquad\qquad\qquad :S{=}R \;\; \text{(simulated random)}, \\ \\ v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}_{\mathbf{D}^{train}_{n^{obs}}|n^{mis},n}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs}|n, n^{mis}])^2\Big] \\ +\mathbb{E}_{\mathbf{D}^{train}_{n^{obs}}|n^{mis},n}\Big[\Big(\frac{1}{n^{obs}} \sum_{j=1}^{n^{obs}} \Big(Y_j - \hat{g}^{obs,K/N}\big(\boldsymbol{X}_j\big)\Big)\Big)^2\Big] \\ \qquad\qquad\qquad\qquad :S{=}D \;\; \text{(random donor)}. \end{cases}$$

Justifications for this approximation can be found in Appendix A5.2.

From Approximation 5.12 one can notice that the mean squared error at a given point consists of three bias and variance terms. In case of the NMAR mechanism, the difference of models $g^{*obs}(\mathbf{x})$ and $g^{*mis}(\mathbf{x})$ yields an NMAR bias. Nonparametric methods converge towards $g^{*obs}(\mathbf{x})$ provided smoothing is increased as sample size grows. There is also an estimation bias with respect to $g^{*obs}(\mathbf{x})$. This bias vanishes asymptotically provided smoothing is increased. Vulnerability of k-nn method is shown in the estimation bias. Namely, it is inversely proportional to $f_{X^{obs}}(\mathbf{x}_0)^3$. Therefore, if density is low then the estimation bias can be large. Finally, the strategy used for estimation of noise terms may cause bias, and this is the case with random donor. The variance sources are estimated model, imputation noise, and target variability. Vulnerability of kernel regression is shown in the variability of the estimated model. It is inversely proportional to density $f_{X^{obs}}(\mathbf{x}_0)$.

The properties of the mean squared error are inherited to the expectation of the mean squared error which is given in Approximation 5.13. Therefore the above considerations for Approximation 5.12 apply indirectly. Namely, the expected value is derived by integrating the pointwise mean squared error over the distributions of $\boldsymbol{X}^{mis}$ and the number of missing data values N^{mis}. The expected quantity consists of squared imputation bias, variance of conditional mean estimate, imputation noise, cross term, target noise, and terms for approximation. Squared imputation bias can be further decomposed into: variability of expected conditional mean estimate, squared global bias, variability of true model, and cross term.

Approximation 5.13 *Approximation for* $\mathbb{E}[\hat{mse}(Y^{comp,K/N})|\mathsf{n}]$
Expectation of mean square error can be approximated as

$$
\begin{aligned}
& \mathbb{E}[\hat{mse}(Y^{comp,K/N})|\mathsf{n}] \approx \underbrace{A}_{\text{expected squared imputation bias}} \\
+ \quad & \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{B: expected variance of conditional mean estimate}} \\
+ \quad & \underbrace{v_{\mathsf{n}}^{*obs,K/N}}_{\text{C: expected imputation noise}} \\
+ \quad & \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[2\mathbb{C}\mathrm{ov}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{D: cross term}} + \underbrace{v^{*mis}}_{\text{expected target noise}} \\
+ \quad & \underbrace{O\left(\mathsf{n}^{-1}\right)}_{\text{technical term}} \quad ,
\end{aligned}
$$

where the terms are:

$$
\begin{aligned}
A \; = \; & \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[g^{*obs}(\boldsymbol{X}^{mis}) + \mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{variability of expected conditional mean estimate}} \\
& + \underbrace{(\mathbb{E}_{\boldsymbol{X}^{mis}}\left[g^{*obs}(\boldsymbol{X}^{mis})\right] + E - \mu^{*mis})^2}_{\text{global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
& \underbrace{+\, 2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - \mathbb{E}_{\boldsymbol{X}^{mis}}\left[g^{*obs}(\boldsymbol{X}^{mis})\right] - E\right)\right.}_{\text{cross term}} \\
& \underbrace{\left.\left(\mathbb{E}_{\boldsymbol{X}^{mis}}\left[g^{*obs}(\boldsymbol{X}^{mis})\right] + E - g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{cross term}},
\end{aligned}
$$

in which term E is

$$
E = \begin{cases}
\left[\frac{Q(g^{*obs}f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis}) - g^{*obs}(\overline{\boldsymbol{X}}^{*mis})Q(f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis})}{2f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis})(v_{p-1}f(\overline{\boldsymbol{X}}^{*mis}))^{2/(p-1)}}\left(\frac{\mathbb{E}\left[k(\mathsf{N}^{obs})\right]}{\mathsf{n}(1-p^*)}\right)^{2/(p-1)}\right. & \\
\left. +o\Big(\big(\frac{\mathbb{E}[k(\mathsf{N}^{obs})]}{\mathsf{n}(1-p^*)}\big)^{2/(p-1)}\Big)\right] + O\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) & (\mathrm{k-nn, p > 2}), \\
\\
\left[\frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{24f^3_{X^{obs}}(\overline{X}^{*mis})}\Big(\mathbb{E}[k(\mathsf{N}^{obs})]/\mathsf{n}(1-p^*)\Big)^2\right. & \\
\left. +o\Big(\big(\frac{\mathbb{E}[k(\mathsf{N}^{obs})]}{\mathsf{n}(1-p^*)}\big)^2\Big) + O\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big)\right] & (\mathrm{k-nn, p = 2}), \\
\\
\frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{2f_{X^{obs}}(\overline{X}^{*mis})}\int \xi^2K(\xi)d\xi\lambda^2\big(\mathsf{n}(1-p^*)\big) & \\
+o\Big(\lambda^2\big(\mathsf{n}(1-p^*)\big)\Big) + O\Big(\big(\mathsf{n}(1-p^*)\lambda(\mathsf{n}(1-p^*))\big)^{-1}\Big) & \\
& (\mathrm{kernel, p = 2}).
\end{cases}
$$

Terms B-D are the following:

$$
B = \begin{cases}
\frac{v_{p-1}}{\mathbb{E}[k(\mathsf{N}^{obs})]}\left(v^{*obs}(\overline{\boldsymbol{X}}^{*mis}) + \frac{1}{2}\mathrm{tr}\left(\mathbf{H}_{v^{*obs}}\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\right)\right) + o\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) & \\
& (\mathrm{k-nn, p > 2}), \\
\frac{2}{\mathbb{E}[k(\mathsf{N}^{obs})]}\left(v^{*obs}(\overline{X}^{*mis}) + \frac{1}{2}\left(\frac{\partial^2}{\partial x^{mis}\partial x^{mis}}v^{*obs}(x^{mis})\right)_{x^{mis}=\overline{X}^{*mis}}\mathbb{V}\mathrm{ar}[X^{mis}]\right) & \\
+o\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) & (\mathrm{k-nn, p = 2}) \\
\\
\frac{v^{*obs}(\overline{X}^{*mis})}{f_{X^{obs}}(\overline{X}^{*mis})(\mathsf{n}(1-p^*))\lambda(\mathsf{n}(1-p^*))}\int K^2(\xi)d\xi + o\Big(\frac{1}{(\mathsf{n}(1-p^*))\lambda(\mathsf{n}(1-p^*))}\Big) & \\
& (\mathrm{kernel, p = 2}),
\end{cases}
$$

where $\mathbf{H}_{v^{*obs}}$ *is Hessian of* $\mathbb{V}\mathrm{ar}[Y^{obs}|\boldsymbol{X} = \mathbf{x}]$ *and*

$$C = \begin{cases} 0 & \text{:}S\text{=}M \text{ (mean)}, \\ v^{*obs} + \mathbb{E}_{\mathsf{N}^{mis},X^{obs}}\left[(g^{*obs}(X^{obs}) - \mathbb{E}_{\mathbf{D}^{train}_{\mathsf{n}^{obs}}|\mathsf{n}^{mis},\mathsf{n}}[\hat{g}^{obs,K/N}(X^{obs}|\mathsf{n}, \mathsf{n}^{mis}])^2\right] & \\ & \text{:}S\text{=}R,D \text{ (random)} \end{cases}$$

and

$$D = \begin{cases} 0 & \text{:}S\text{=}M/S\text{=}R \quad \text{(mean and simulated random)}, \\ O(\mathsf{n}^{-1}) & \text{:}S\text{=}D \quad \text{(random donor)}. \end{cases}$$

5.2.1 An example: mse and flexible models

The aim of this example is to demonstrate that imputation performance is sometimes unaffected by model fit. As discussed in Chapter 3 we may achieve the same distribution of $\hat{Y}$ with different models. More concretely, we shall study how the relation of the variance of flexible models is related to the variance of residuals of noisy imputation (simulated randomness strategy). For this purpose we shall use (approximately) unbiased k-nearest neighbour models with smoothing that increases from one nearest neighbour to five ($k \in \{1, 3, 5\}$). The results are measured in terms of mse, which in this context is rewritten in the form (see Appendix A5.3 for details):

$$\begin{aligned} \mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] &= \mathbb{E}_{\boldsymbol{X}^{mis}}\left[\mathrm{mse}(Y^{imp}|\boldsymbol{X}^{mis}, \mathsf{n})\right] \\ &= \underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}}\left[\left(\mathbb{E}[\hat{g}^{obs}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\right)^2\right]}_{\text{A: expected squared imputation bias}} \\ &+ \underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}}\left[\mathbb{V}\mathrm{ar}[\hat{g}^{obs}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}]\right]}_{\text{B: expected variance of model}} + \underbrace{v^{*obs}_{\mathsf{n}}}_{\text{C: variance of imputation residuals}} \\ &+ \underbrace{v^{*mis}}_{\text{D: expected variance of target Y}^{\text{mis}}}. \end{aligned}$$

In our experiments we try to keep term A (bias) close to zero. Then the result depends on terms B (model variance) and C (simulated noise) because the term D does not depend on our model or strategy.

Our data is generated from model

$$Y = g(x) + \epsilon,\ \epsilon \sim N(0, 0.15),$$

where

$$g(x) = \frac{1}{1 + \exp(-x)} - 0.5.$$

Observations are selected randomly under MAR type of missingness such that

$$\begin{aligned} &X^{obs} \sim N(-1, 5), \text{ and} \\ &X^{mis} \sim N(1, 1). \end{aligned}$$

An example of data is visualized in Figure 5.1 where $\mathbb{V}\mathrm{ar}[\epsilon]$ represents approximately 50% of $\mathbb{V}\mathrm{ar}[Y]$.

Due to complexity of theoretical results the problem is studied empirically. The values of the studied quantities A-C reveal us the differences between the three imputations with different level of smoothing.

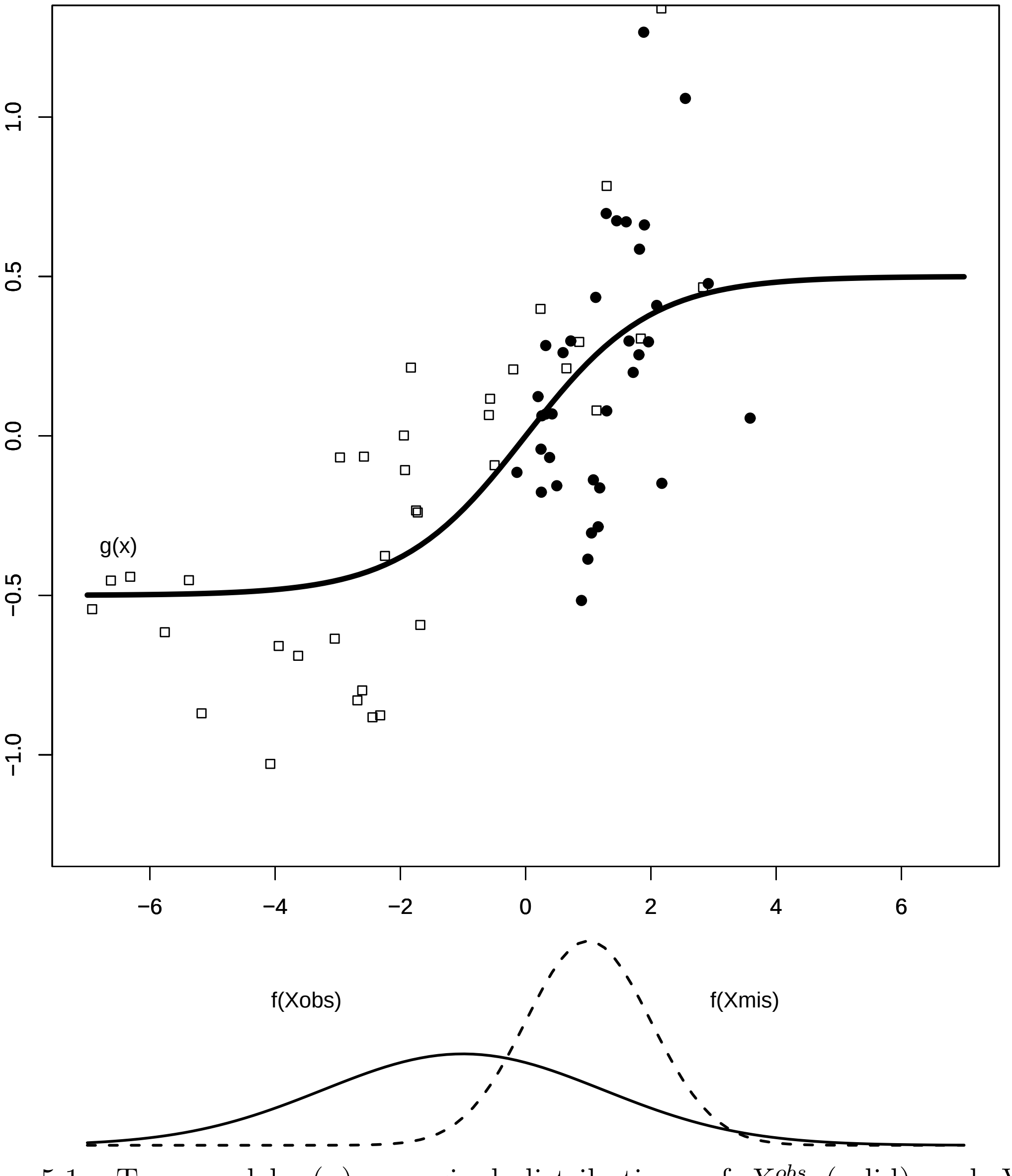


Figure 5.1: True model $g(x)$, marginal distributions of X^{obs} (solid) and X^{mis} (dashed), and random sample of size 70. Training data is denoted by square plots, whereas draws from missing population are denoted by black dots.

Three k-nearest neighbour models with k=1, k=2 and k=5 have been used in imputations, with simulated random strategy such that

$$\hat{Y}^{imp} = \hat{g}^{k-NN}(x^{mis}) + \hat{\epsilon},\ \hat{\epsilon} \sim N(0, \hat{v}^{obs}),$$

where $\hat{v}^{obs}$ was estimated from model residuals

$$\hat{v}^{obs} = \frac{1}{\mathsf{n}^{obs}} \sum_{j=1}^{\mathsf{n}^{obs}} \left(\hat{g}^{k-NN}(x_j^{obs}) - y_j^{obs}\right)^2.$$

The obtained models are visualized in Figure 5.2.

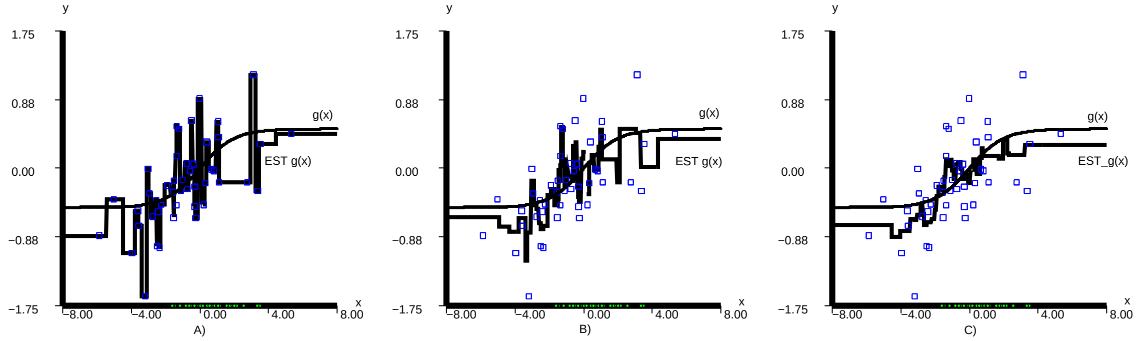


Figure 5.2: Three k-nn models with a) k=1, b) k=2 and c) k=5. Clearly $\hat{v}^{obs} = 0$ for model $k = 1$.

The results of imputation performance are depicted in Table 5.1. All models yield approximately the same MSE and the models are almost unbiased, because term A is close to zero.

The real difference is that model k=1 has zero in term C ($\hat{v}^{obs} = 0$). As the smoothness of the model increases we can see that term B (model variance) decreases and term C (simulated randomness) increases.

Method	**Sample size**	A	B	C	MSE^{total}
k=1	50	0.001	0.153	0.000	0.304
	1033	0.001	0.151	0.000	0.302
k=2	50	0.000	0.074	0.077	0.301
	1033	0.000	0.076	0.075	0.301
k=5	50	0.002	0.033	0.124	0.309
	1033	0.000	0.031	0.120	0.301

Table 5.1: Terms A-C of mean squared error decomposition for three k-nn imputaton methods. A=model bias (squared), B=model variance, C=simulated noise, MSE^{total}=A+B+C+0.15 where the last value is v^{*mis} (irreducible error).

An obvious conclusion is that in terms of $mse(Y^{comp})$ unbiased models with different smoothness can yield the same performance.

5.3 Summary

An advantage of nonparametric kernel methods (and k-NN) is that they are flexible. However, due to the nonparametric nature of kernel methods (and k-NN) it was difficult to derive results in fine detail. Some analyses were possible by applying results from the literature. If missingness is NMAR then a difference between conditional means $g^{*obs}(\mathbf{x})$ and $g^{*mis}(\mathbf{x})$ has an effect on biases (of the first moment estimator or at unit level). Possible bias due to the difference is irreducible even asymptotically, and may dominate results even at small sample sizes. Under MAR or MCAR missingness bias is due to estimation bias. This bias is well known to asymptotically vanish under suitable regularity conditions. However, for small sample sizes this may be an issue. To summarize, if conditional mean is nonlinear then nonparametric methods are especially recommended.

Chapter 6

Cell imputation

This chapter provides the main result of this thesis: a practical way to estimate imputation errors of cell methods. The result is highlighted in a form of an example algorithm in Section 6.4.2 and its usability is tested in an example that follows in Section 6.4.3.

In cell imputation data are divided into subsets (cells), and imputation is done more or less separately for each cell. For successful imputation, cell methods must fullfill two conditions

i) differences between data subsets should reflect predictable differences (by covariates $\boldsymbol{X}$) between the missing values, and

ii) it should be possible to associate an incomplete observation to an approximately correct cell using observed part of data.

Thus the applicability of cell imputation methodology depends on a data and missingness mechanism. Optimally, missing values in the data clusters can be predicted using observed covariates.

There are several ways to build cell models: using side information, categorial covariates or clustering algorithms. In this context our focus is in clustering, but most of our results apply to other types of cell methods as well. These results are written in terms of decompositions, similar to those described in the previous chapters. But a fully detailed analysis of imputation based on clustering algorithms is quite challenging due to the "ad hoc" nature of many algorithms. Therefore some of the results are left in an implicit form.

There are not very many specific publications about cell imputation. One such study was done by Santos [90] in 1981. In addition there are two analytical studies and a number of empirical studies. Santos assumes a finite population sampling framework and presents large sample biases of covariance and variance estimators. He also assumes that the cells are fixed. Kalton and Kish [47], in 1981, used clustering to reduce the variance of donor imputations. Kim and Fuller [50] studied analytical properties of mean estimator based on fractional hot deck imputation within cells in 2004. See [15], [25], [80], and [81] for empirical studies on imputations using self-organizing map and Section 6.2.1 for details of the SOM algorithm.

6.1 An overview of cell imputation in the current thesis

In this thesis three types of clustering approaches are considered for imputation

i) Standard (K-Means type of) clustering, where only completely observed covariates $\boldsymbol{X}^{obs}$ are used to define the clusters. In our applications, two algorithms are used for this, where

   C : is short for ”standard” K-Means clustering [31, 32].

   T : denotes ”standard” TS-SOM algorithms [56], as described in Section 6.2.1

ii) Clustering of joint distribution $f_{Y,\boldsymbol{X}}(y, \mathbf{x})$. There are two ways to do this. Using completely observed part of data or a specific incomplete data training algorithm. We use K-Means for the first option and two variants of the TS-SOM algorithm [57] for the latter option. The options are denoted by

   CJ : K-Means clustering with observed $Y, \boldsymbol{X}$ using completely observed part of data.

   TJ : A new EM-type of incomplete data training algorithm for TS-SOM using all data

   TJ* : TS-SOM clustering similar to K-Means with observed $Y, \boldsymbol{X}$ using complete observed part of data

iii) Smoothed imputation, where cell specific imputation models borrow strength [82] from ”neighboring” clusters. This option applies only to TS-SOM algorithms and is denoted by a small s after the imputation strategy, like

   T,S^s : smoothed TS-SOM imputation with covariate $\boldsymbol{X}$ clusters

   TJ,S^s : smoothed TS-SOM with incomplete training for joint data $(Y, \boldsymbol{X})$,

   where strategy $S \in \{M, R\}$ (smoothed donor imputation is not computed).

6.1.1 Realization of cell imputation

There are many variations of cell imputations. A practical application is done in several steps, which may be implemented in many alternative ways. In this thesis we consider a case where the basic steps are as follows:

1) For a given incomplete data $\mathbf{d}^{inc}$ we use some clustering algorithm to build cells. These cells are parametrized with cell centroids $\mathbf{w}_{\{i\}} = \mathbf{w}_{Y,\{i\}} \cup \mathbf{w}_{\boldsymbol{X},\{i\}}, i = 1, \ldots, \mathsf{n}_c$, where $\mathbf{w}_{Y,\{i\}}$ is the Y part of centroids and $\mathbf{w}_{\boldsymbol{X},\{i\}}$ is the $\boldsymbol{X}$ part.

2) We use some classifier $b(\mathbf{x}'|\mathbf{w}_{\{i\}})$ to associate all observations $\mathbf{x}' \in \mathbf{d}^{inc}$ to cells. Our notation $\mathbf{x}'$ implies that $\mathbf{x}'$ may consist of either covariate $\mathbf{x}$ or both $(y, \mathbf{x})$,

depending on the current situation. Thus for a given data set $\mathbf{d}^{inc}$ we get data divisions such that

$$\Omega_i = \{j|b(\mathbf{x}'_j) = i\}$$

3) Using some imputation strategy, we create a completed data set $\mathbf{d}^{comp}$ such that

$$Y_j^{comp} = \begin{cases} y_j, & \text{if } r_j = 1 \quad (y_j \text{ is observed}), \\ \mu_i^{method} + \hat{\epsilon}_i^{strategy}, & \text{if } r_j = 0 \quad (y_j \text{ is missing}), \end{cases}$$

where μ_i^{method} is the (possible smoothed) mean estimate for Y (cell mean) in the cluster $i = b(\mathbf{x}_j^{obs})$ and randomness follows the imputation strategy as before

i) M (mean), $\hat{\epsilon}_i = 0$

ii) R (random), where $\hat{\epsilon}_i \sim f_\epsilon(e|i)$, typically $f_\epsilon(e|k) = N(0, \tau_i^{method})$

iii) D (donor), where $\hat{\epsilon}_i \sim \{y_{j\in\Omega_i} - \mu_i^{method}\}$,

where τ_i^{method} is the (possible smoothed) variance estimate for Y (cell variance) in the cluster i.

The last step, actual imputation, is basically an application of current baseline methods for subsets of data. In theory we could replace this step with more advanced methods, which would lead to some kind of multilevel approach in imputation.

There are some obvious ways to make variations in Steps 1) and 2). We can have different kinds of clustering methods, as described later. And we can use different types of classifiers to determine how incomplete observations are to be associated to cells. Obviously, the realization of a classifier plays a major role in cell imputation. As a preliminary move towards a deeper study, it is useful to define the following notations.

Let $\mathbf{x}' \in \mathbf{R}^m$ be some realization of random vector $\boldsymbol{X}'$, a soft classifier for class i is a decision function

$$g_i(\mathbf{x}'|\boldsymbol{\theta}) : \mathbf{R}^m \to [0, 1],$$

where $\boldsymbol{\theta}$ is a set of classifier parameters, if any. A typical example of a soft classifier is the Bayes classifier that gives a posterior probability of class i, where

$$g_i(\mathbf{x}') = \frac{f_{\boldsymbol{X}'}(\mathbf{x}'|i)\Pr(i)}{f_{\boldsymbol{X}'}(\mathbf{x}')},$$

where $f_{\boldsymbol{X}'}(\mathbf{x}'|i)$ is the class specific density of $\boldsymbol{X}'$ and $\Pr(i)$ is the prior probability of class i. This also obeys the rather usual condition

$$\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}') = 1, \tag{6.1}$$

since posterior probabilities add to one.

In our implementations we consider two possibilities to convert soft classes into crisp ones. The first is maximum posterior classifier that is defined by

$$b(\mathbf{x}') = \operatorname*{argmax}_l g_l(\mathbf{x}').$$

The second alternative is randomized classifier where crisp class is sampled from categorial distribution

$$b^{\epsilon}(\mathbf{x}') \sim cat\big(g_1(\mathbf{x}'), \ldots, g_{\mathsf{n}_c}(\mathbf{x}')\big),$$

where class $b^{\epsilon} \in \{1, \ldots, \mathsf{n}_c\}$ is selected randomly according to probabilities $g_1(\mathbf{x}'), \ldots, g_{\mathsf{n}_c}(\mathbf{x}')$ under a requirement that condition 6.1 holds.

In both cases the result of classification can be coded in terms of binary indicator vector

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_{\mathsf{n}_c} \end{bmatrix}, \text{ where } c_i = \begin{cases} 1, & \text{if } b(\mathbf{x}') = 1 \\ 0, & \text{otherwise,} \end{cases}$$

and in the case of categorial distribution this is the same as $\mathbf{c} \sim Multin(1; g_1(\mathbf{x}'), \ldots, g_{\mathsf{n}_c}(\mathbf{x}'))$.

One should also note that in many mathematical derivations it is convenient to use soft classifiers $g_i(\mathbf{x}')$ instead of crisp ones. For example, we can define derivates $\frac{\partial g_i(\mathbf{x}')}{\partial \mathbf{x}'}$.

6.2 Brief introduction to K-Means and TS-SOM clustering algorithms

K-Means [31, 32] is perhaps the world's best known clustering algorithm. It is also known as the Lloyd vector quantization method [65]. The basic idea is simple: one tries to minimize distortion measure

$$J = \sum_j \sum_i \left|\left|\mathbf{x}'_j - \mathbf{w}_i\right|\right|_2^2, \; i = \operatorname*{argmin}_l \left|\left|\mathbf{w}_l - \mathbf{x}'_j\right|\right|_2^2$$

by finding the best values for centroids $\mathbf{w}_i, i = 1, \ldots, \mathsf{n}_c$ where n_c is a fixed number of clusters. For distributions the problem can be written as

$$\min_{\mathbf{w}} J' = \sum_l \int_{\mathcal{V}_l} ||\mathbf{x}' - \mathbf{w}_l||_2^2 f_{\boldsymbol{X}'}(\mathbf{x}') d\mathbf{x}',$$

$$\mathcal{V}_l = \left\{\mathbf{x}' \middle| ||\mathbf{x}' - \mathbf{w}_l||_2^2 \leq ||\mathbf{x}' - \mathbf{w}_i||_2^2, \; i \neq l\right\},$$

where $\mathcal{V}_l$ is the Voronoi cell for prototype $\mathbf{w}_l$. The continuous problem is solvable only for simple forms of $f_{\boldsymbol{X}'}(\mathbf{x}')$. With data $\{\mathbf{x}'_j\}_{j=1}^{\mathsf{n}_c}$ one usually uses the following algorithm that can be regarded as a special case of EM-estimation:

1. Initialize randomly
$$\mathbf{w}_i \sim \{\mathbf{x}'_j\}, i = 1, \ldots, \mathsf{n}_c$$

2. Divide data into Voronoi cells
$$\Omega_i = \left\{j \middle| ||\mathbf{x}'_j - \mathbf{w}_i|| \leq ||\mathbf{x}'_j - \mathbf{w}_l||, l \neq i\right\}$$

3. Compute new centroids (cluster means)

$$\mathbf{w}_i^{new} = \frac{1}{n_i} \sum_{j \in \Omega_i} \mathbf{x}'_j, \; n_i = \#\Omega_i$$

4. Check convergence

$$\text{if } ||\mathbf{w}^{new} - \mathbf{w}|| > \delta \text{ then set } \mathbf{w} := \mathbf{w}^{new} \text{ and GOTO 2}$$
$$\text{else stop.}$$

The above algorithm is not guaranteed to converge to global minimum but it usually gives relatively good clusterings. It is, however, recommended that the algorithm is run several times with random initializations to eliminate problems with bad locally optimal solutions. Some statistical properties of the K-Means algorithm may be read from [31, 78, 79].

6.2.1 The self-organizing map (SOM) and its tree-structured variant (TS-SOM)

The TS-SOM [56], which is used in this thesis, belongs to a specific class of self-organizing neural network algorithms. Like the original self-organizing map (SOM) it can be interpreted as an implementation of principal curves and surfaces [35, 60]. The original SOM [54] by Kohonen can be written as a kernel smoothed K-Means algorithm that tries to build a lower-dimensional nonlinear manifold in training data. The smoothing is done along a discretized latent surface that in the SOM terminology defines a neighborhood structure of the centroids. The neighborhood structure of the SOM is depicted in Figure 6.1 for 2-D SOM in 3-D data.

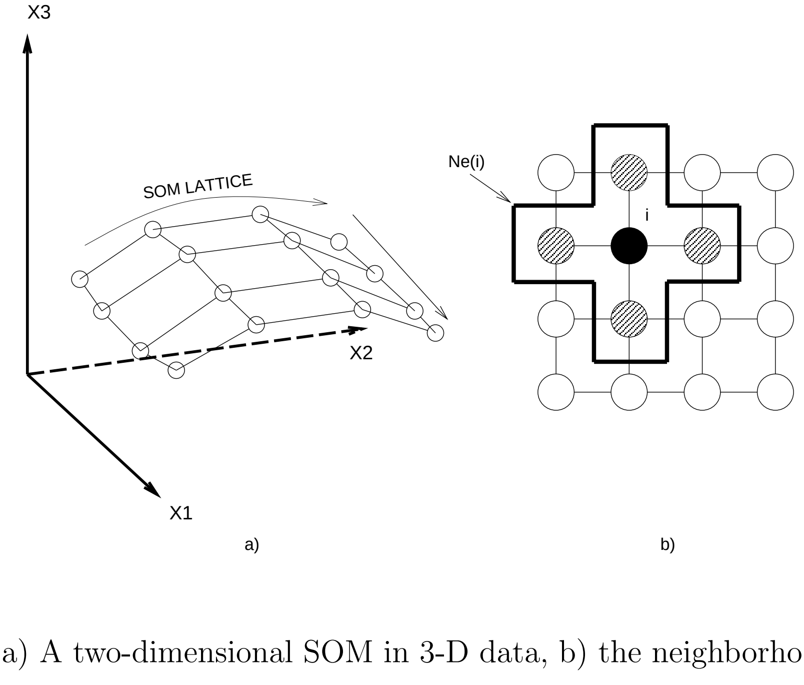


Figure 6.1: a) A two-dimensional SOM in 3-D data, b) the neighborhood of node i.

The SOM lattice which is made of connected data clusters, is a discrete representation of a lower dimensional latent space that is fitted in data. Typically the lattice is defined in a 2D-neighborhood of clusters as shown in Figure 6.1b, which implies that a 2D latent space is fitted in m-dimensional data $\mathbb{x}' \in \mathbb{R}^m$.

A very basic SOM training algorithm can be written as follows

1. Initialize randomly
$$\mathbf{w}_i \sim \{\mathbf{x}'_j\}, i = 1, \ldots, \mathsf{n}_c$$
2. Divide data into Voronoi regions
$$\Omega_i = \left\{ j \middle| ||\mathbf{x}'_j - \mathbf{w}_i|| \leq ||\mathbf{x}'_j - \mathbf{w}_l||, l \neq i \right\}$$
3. Compute centroid mean
$$\overline{\mathbf{x}}'_i = \frac{1}{n_i} \sum_{j \in \Omega_i} \mathbf{x}'_j, \text{ where } n_i = \#\Omega_i$$
4. Do SOM smoothing along neighborhood (latent space)
$$\mathbf{w}_i^{new} = \frac{\sum_l h_{i,l} n_l \overline{\mathbf{x}}'_l}{\sum_l h_{i,l} n_l} \tag{6.2}$$
5. Check convergence

   if $||\mathbf{w}^{new} - \mathbf{w}|| > \delta$ then set $\mathbf{w} := \mathbf{w}^{new}$ and GOTO 2

   else stop.

Step 4 (Equation 6.2) is also known as the Nadaraya-Watson kernel smoother, which in the case of SOM is applied in data clusters over a latent SOM space. The smoothing kernel $h_{i,l}$ defines the latent structure. In the simplest case it is a box kernel in a 2-dimensional neighborhood $Ne(i)$ of node i such that

$$h_{i,l} = \begin{cases} 1, & \text{if } l = i \text{ or } l \in Ne(i), \\ 0, & \text{otherwise.} \end{cases} \tag{6.3}$$

In the original SOM the neighborhood of node i $Ne(i)$ changes over the time of training. Initially it covers more nodes (clusters), which implies stronger smoothing than in the end of training. This idea is somewhat similar to simulated annealing [52, 9], where one first tries to do optimization on a coarse level and later on gets to the finer details. The actual kernel, which is used in this thesis, under the TS-SOM algorithm, is introduced in Equation 6.8 (page 104). The convergence properties and some statistical properties for an SOM algorithm can be found in [114].

The TS-SOM algorithm [56] is a tree-structured variant of SOM that implements decreasing smoothing via a constructive training algorithm, where several SOM networks (layers) are trained with an increasing number of nodes (clusters). When the neighborhood is defined via a constant number of nearest neighbors, the increase of nodes effectively decreases the smoothing. The structure of the TS-SOM is depicted in Figure 6.2. As we can see, there are $2^{(l-1)^D}$ nodes on layer l, where D is the dimension of the SOM lattice.

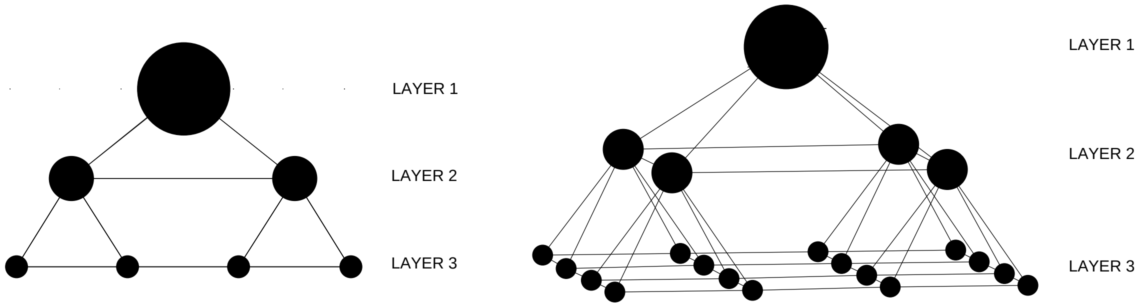


Figure 6.2: The structure of the TS-SOM is made of several SOM layers in a tree structure.

Since TS-SOM is not the topic of the current thesis we shall omit most of the details of the algorithm. All we need to know is that we would normally use TS-SOM like K-Means (or SOM) to define clusters (or smoothed clusters). There is, however, one additional benefit in the current implementation of the TS-SOM. It can be trained with partially observed incomplete data. Thus the Y variate can have a role in the clustering. In Chapter 7 we shall see that this allows one to build imputation models for cases where conditional distribution $f_{Y|\boldsymbol{X}}(y|\mathbf{x})$ is multimodal.

Due to practical reasons we shall limit our analysis to the usage of clustering in imputation. Thus differences between different clustering algorithms are considered to be external information, which could be measured via cluster compactness or quantization rate.

6.2.2 About the use of K-Means and TS-SOM in imputations

As described in Section 6.1 we shall examine three clustering method approaches in cell imputation: a standard method with fully observed covariates $\boldsymbol{X}$, joint clustering using both $\boldsymbol{X}$ and Y, and smoothed clustering using TS-SOM. In all these cases the role of clustering is to define the imputation model by dividing data into subsets Ω_i. The actual imputation is then a union of imputations for subsets of incomplete observations. This can be described in terms of cell means $\{\mu_i\}_{i=1}^{\mathsf{n}_c}$ and cell variances $\{\tau_i\}_{i=1}^{\mathsf{n}_c}$ using the observed part of data.

In standard approach denoted by C (K-Means) and T (TS-SOM) the estimates μ_i and τ_i are computed directly from the observed part of data, e.g.

$$\mu_i \;=\; \mu_i^{obs} = \frac{1}{n_i^{obs}} \sum_{j \in \Omega_i, r_j = 1} y_j^{obs}, \text{ and} \tag{6.4}$$

$$\tau_i \;=\; \tau_i^{obs} = \frac{1}{n_i^{obs} - 1} \sum_{j \in \Omega_i, r_j = 1} \left(y_j^{obs} - \mu_i\right)^2 \tag{6.5}$$

and the cluster is selected to be the nearest one

$$b(\mathbf{x}_j|\mathbf{w}_{\{l\}}) = \underset{l}{\operatorname{argmin}} \, ||\mathbf{x}_j - \mathbf{w}_{\boldsymbol{X},l}||.$$

In "joint"-clustering (denoted by CJ, TJ, and TJ*) clusters are defined for distribution $f_{Y,\boldsymbol{X}}(y, \mathbf{x})$. For K-Means CJ cell i is selected using the whole observation for complete records and the observed covariate part for incomplete records. Crisp classifier is applied as

$$i = b(\mathbf{x}_j'|\mathbf{w}_{\{l\}}) = \begin{cases} \text{argmin}_l\, ||\mathbf{x}_j - \mathbf{w}_{\boldsymbol{X},l}||, & \text{(if } y_j \text{ is missing)}, \\ \text{argmin}_l\, ||(y_j, \mathbf{x}_j)^T - \mathbf{w}_l||, & \text{(if } y_j \text{ is observed)}, \end{cases}$$

However, for TS-SOM a randomized classifier is applied for incomplete records using a sample from the categorial distribution as

$$i = b(\mathbf{x}_j'|\mathbf{w}_{\{l\}}) = \begin{cases} B \sim cat\big(g_i(\mathbf{x}_j|\mathbf{w}_{\boldsymbol{X},1}), \ldots, g_{\mathsf{n}_c}(\mathbf{x}_j|\mathbf{w}_{\boldsymbol{X},\mathsf{n}_c})\big) & \text{(if } y_j \text{ is missing)}, \\ \text{argmin}_l\, ||(y_j, \mathbf{x}_j)^T - \mathbf{w}_l||, & \text{(if } y_j \text{ is observed)}. \end{cases}$$

The smoothing approach applies to TS-SOM only because it requires the concept of neighborhood. The idea is to "borrow strength" from the neighboring cells using kernel smoothing. More formally we replace μ_i and τ_i by

$$\mu_i^s = \frac{\sum_l h_{i,l}^s n_l \mu_i}{\sum_l h_{i,l}^s n_l} \text{ and} \tag{6.6}$$

$$\tau_i^s = \frac{\sum_l h_{i,l}^s n_l \tau_i}{\sum_l h_{i,l}^s n_l}, \tag{6.7}$$

where $h_{i,l}^s$ is the smoothing kernel. We might use the box kernel as in Equation 6.3, but in this thesis we have used

$$h_{i,l} = h_{i,l}^s = \begin{cases} \alpha & \text{if l = i,} \\ \beta & \text{if l} \in \text{Ne(i), l} \neq \text{i} \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

Also in this thesis, the neighborhood $Ne(i)$ is defined to be the nearest neighbors of node i, and kernel weights are kept in constant values $\alpha = 1, \beta = 0.5$.

As a concluding statement it should be noted that sometimes, but not always, cell means μ_i (or μ_i^s) are the same as cell centroids $\mathbf{w}_{Y,i}$. The equality applies only with crisp classifier $b(\mathbf{x}_j'|\mathbf{w}_{\{l\}}^{obs})$ when the joint $(\boldsymbol{X}, Y)$ clustering model is trained using exactly the same data that is used in the computation of cell means μ_i. Yet in practical applications it often is necessary to control the building of clusters by data preprocessing, where original data $\{\mathbf{x}_j, y_j\}_{j=1}^{\mathsf{n}}$ are replaced with preprocessed data $\{\tilde{x}_j, \tilde{y}_j\}_{j=1}^{\mathsf{n}}$. Then classification must be done using preprocessed data $\{\tilde{x}_j, \tilde{y}_j\}_{j=1}^{\mathsf{n}}$, while cell means $\{\mu_i\}_{i=1}^{\mathsf{n}_c}$ and variances $\{\tau_i\}_{i=1}^{\mathsf{n}_c}$ are obtained using "raw data" $\{\mathbf{x}_j, y_j\}_{j=1}^{\mathsf{n}}$, and centroids $\mathbf{w}_{Y,i}$ are not same as cell means μ_i (or μ_i^s).

6.3 Theoretical preliminarities for approximations

The analysis of the performance of cell imputation is more complicated than any of the analyses in the previous chapters. Therefore we must do more simplifications in order to interpret the outcomes of the analyses.

6.3.1 A note about Taylor approximations

As before simplifications are based on Taylor approximations, where some functions $\mathcal{F}(\boldsymbol{Z})$ of multivariate random variables $\boldsymbol{Z}$ are approximated "around" expectations $\mathbb{E}[\boldsymbol{Z}]$. In our case $\mathcal{F}$ is, for example, the first moment of data μ and the random vector $\boldsymbol{Z}$ is the observation vector. For the sake of simplicity, we use mainly "rough" first order linear approximations

$$\mathcal{F}(\boldsymbol{Z}) \approx \mathcal{F}(\mathbb{E}[\boldsymbol{Z}]) + (\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^T \mathcal{F}'(\mathbb{E}[\boldsymbol{Z}]),$$

where $\mathcal{F}'(\boldsymbol{Z})$ denotes vector derivate $\frac{\partial}{\partial \boldsymbol{Z}}\mathcal{F}(\boldsymbol{Z})$.

The problem with using the first order Taylor approximation is that the derived result may be quite inaccurate. As an example, Kempen and Vliet noticed that approximation for variance (of ratio) underestimated true variability [49]. We could apply higher order approximations but they would yield formulas which are complicated and difficult to interpret. An approximate bias of the first moment estimator for unsmoothed methods is likely to be accurate enough. A bias of the second moment estimator, variances of the moment estimator, and the mean squared error at a given point (and other mean squared error results derived from it) are likely to be less accurate. However, the results are readable and quite interpretable, as we wish.

In cell imputation the function $\mathcal{F}(\boldsymbol{Z})$ is typically of a form

$$\mathcal{F}(\boldsymbol{Z}) = \alpha(\boldsymbol{Z}) \sum_{i=1}^{\mathsf{n}_c} \beta_i(\boldsymbol{Z})$$

where quantities $\alpha(\boldsymbol{Z}), \beta_i(\boldsymbol{Z})$ depend on random vector $\boldsymbol{Z}$.

Then, for the sake of example, the computation of imputation statistics like $\mathbb{V}\text{ar}[\mathcal{F}(\boldsymbol{Z})]$ takes the form of

$$\mathcal{F}'(\mathbb{E}[\boldsymbol{Z}])^T \mathbb{V}\text{ar}[\boldsymbol{Z}] \mathcal{F}'(\mathbb{E}[\boldsymbol{Z}]),$$

where $\mathcal{F}'(\mathbb{E}[\boldsymbol{Z}]) = \alpha'(\mathbb{E}[\boldsymbol{Z}]) \sum_{i=1}^{\mathsf{n}_c} \beta_i(\mathbb{E}[\boldsymbol{Z}]) + \alpha(\mathbb{E}[\boldsymbol{Z}]) \sum_{i=1}^{\mathsf{n}_c} \beta_i'(\mathbb{E}[\boldsymbol{Z}])$

It should be noted that for smoothed imputation methods T,M^s/R^s and TJ,M^s/R^s randomness in terms β_i is of ratio form (random quantity divided by random quantity). Typically one handles ratio quantities by computing a second order approximation for expectation and a first order approximation for variance, as was done for example by Kempen and Vliet [49]. However, in order to obtain readable results we apply here the first order Taylor approximation for both quantities.

6.3.2 Priors, posteriors, classifiers, and randomness

The role of classifiers is essential in the analysis of cell imputation. The imputation model is defined as a collection of cells $\{i\} = \{1, \ldots, \mathsf{n}_c\}$, which themselves are defined by classifiers. We may therefore say that the "model" is actually a soft classifier of the type

$$\mathbf{g}(\mathbf{x}'|\mathbf{w}_{\{i\}}) = [g_1(\mathbf{x}'|\mathbf{w}_{\{i\}}), \ldots, g_{\mathsf{n}_c}(\mathbf{x}'|\mathbf{w}_{\{i\}})]^T, \text{ where } \mathbf{w}_{\{i\}} = \mathbf{w}_1, \ldots, \mathbf{w}_{\mathsf{n}_c}.$$

This is also behind "crisp" selection of classes that can be either deterministic

$$b(\mathbf{x}') = \underset{l}{\operatorname{argmax}}\, g_l(\mathbf{x}'|\mathbf{w}_{\{i\}})$$

or randomized

$$b^{\epsilon}(\mathbf{x}') \sim cat\big(g_1(\mathbf{x}'|\mathbf{w}_{\{i\}}), \ldots, g_{\mathsf{n}_c}(\mathbf{x}'|\mathbf{w}_{\{i\}})\big).$$

Because classifiers are trained from data, the classification result is also subject to uncertainty in the estimator of parameters $\hat{\boldsymbol{W}}_{\{i\}}$. This complicates our studies a lot, because everything that is related to cells becomes random. This is illustrated in Figure 6.3. To make our notation clear we need to introduce the following classifiers which are trained from random data

$$\begin{aligned}\hat{b}(\mathbf{x}') &= \underset{l}{\operatorname{argmax}}\, g_l(\mathbf{x}'|\hat{\boldsymbol{W}}_{\{i\}}) \text{ and} \\ \hat{b}^{\epsilon}(\mathbf{x}') &\sim cat\big(g_1(\mathbf{x}'|\hat{\boldsymbol{W}}_{\{i\}}), \ldots, g_{\mathsf{n}_c}(\mathbf{x}'|\hat{\boldsymbol{W}}_{\{i\}})\big).\end{aligned}$$

where $\hat{\boldsymbol{W}}_{\{i\}}$ is the estimator of cell centroids. Thus we have several types of randomness in classification, one caused by our randomized selection b^{ϵ} and one caused by randomness in estimates of parameters $\mathbf{w}_{\{i\}}$. In addition we shall apply classifiers to random data, which can be conditionalized in several ways.

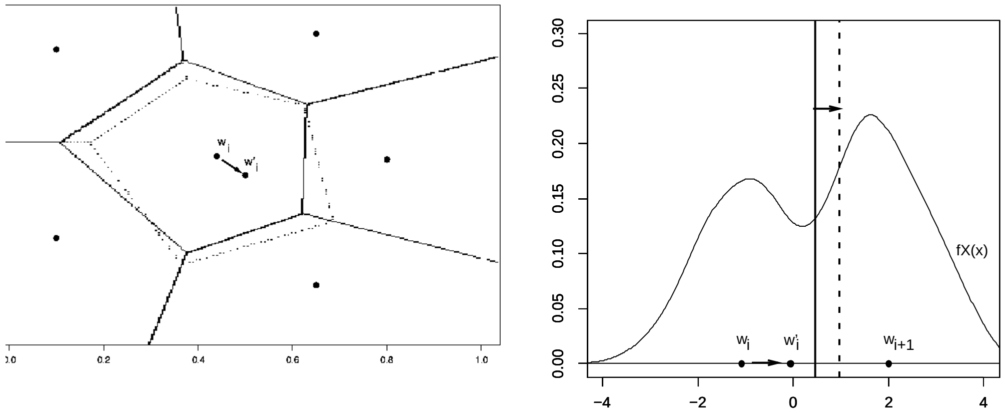


Figure 6.3: Change in estimates of parameters $\mathbf{w}_i$ changes the decision boundaries of Voronoi cells V_i, as well as the cell specific data $f_{\mathbf{X}'}(\mathbf{x}'|i)$ and its realization Ω_i.

Yet another complication is caused by the classification of incomplete data under joint $(Y, \boldsymbol{X})$ clustering. Here we have a possibility of "misclassifications" of training data". In other words we might have a case where the "true" cell of observation pair $(Y^{mis}, \mathbf{X}^{mis})$ should be i but using only covariates $\mathbf{X}^{mis}$ it is classified to some other cell l. In some cases (TS-SOM) this randomness can be written in terms of randomized classifier $\hat{b}^{\epsilon}$, but in the case of joint $(Y, \boldsymbol{X})$ clustering version of K-Means it is an additional source of concern.

In order to make our notation rigid we need to specify how classifiers are applied in various contexts. Assume first that $\mathbf{w}_{\{i\}}$ are fixed and we know the true distribution of data $f_{Y,X}(y,\mathbf{x})$. Then the cell priors are defined by

$$\pi_i = \int_{V_i} f_{Y,X}(y,\mathbf{x}), \text{ where}$$

regions are defined by maximum posterior classifier $b(\mathbf{x}'|\mathbf{w}_{\{i\}})$. Then given $\mathcal{Q}_3 = \{\mathbf{d}^{train}, \mathbf{d}^{test}, \mathbf{w}_{\{i\}}\}$ we define the "correct" classification probability of classifier $b(\mathbf{x})$ using indicator $I_{b(\mathbf{x})=i}$ as

$$\begin{aligned} q_i &= \mathbb{E}[\frac{1}{n_i^{true}} \sum_{j\in\mathbf{d}_i^{true}} I_{b(\mathbf{x}_j^{mis})=i}|\mathcal{Q}_3] \\ &= \frac{1}{n_i^{true}} \sum_{j\in\mathbf{d}_i^{true}} I_{b(\mathbf{x}_j^{mis})=i}, \end{aligned}$$

where $\mathbf{d}_i^{true}$ denotes the part of observations whose true values belong to cell i, and n_i^{true} is the size of $\mathbf{d}_i^{true}$. For randomized classifier $b^\epsilon(\cdot)$ replace $I_{b(\mathbf{x}_j^{mis})=i}$ by $g_i(\mathbf{x}_j^{mis})$. We also define the expected classification probability at conditionalization $\mathcal{Q}_2 = \{\mathsf{n}, \mathbf{d}^{train}, \mathbf{w}_{\{i\}}\}$ as

$$\mathbb{E}[\hat{q}_i|\mathcal{Q}_2] = \mathbb{E}[\frac{1}{N_i^{true}} \sum_{j\in\mathbf{d}_i^{true}} I_{b(\boldsymbol{X}_j^{mis})=i}|\mathcal{Q}_2],$$

Recalling that random observations in true data are iid the above expectation may be computed as

$$\mathbb{E}[\hat{q}_i|\mathcal{Q}_2] = \int_{V_i} \big(b(\mathbf{x}^{mis}) = i\big) f_{\boldsymbol{X}^{mis}|i}(\mathbf{x}^{mis}) d\mathbf{x}^{mis},$$

where $f_{\boldsymbol{X}^{mis}|i}(\mathbf{x}^{mis}|i)$ is the density of observation $\boldsymbol{X}^{mis}$ in cell i for which $\int_{V_i} f_{\boldsymbol{X}^{mis}|i}(\mathbf{x}^{mis}|i) = 1$. For a randomized classifier we need to replace $b(\mathbf{x}^{mis}) = i$ with $g_i(\mathbf{x}^{mis})$.

Probabilities related to π_i and q_i can help us to describe the characteristics of cell imputation. In another level we have actual classification probabilities that depend on our conditionalization level. For a fixed cell structure $\mathbf{w}_{\{i\}}$ that corresponds to our conditionalization level $\mathcal{Q}_3 = \{\mathsf{n}, \mathbf{d}^{train}, \mathbf{d}^{test}, \mathbf{w}_{\{i\}}\}$ we have the posterior probability for single observation $\mathbf{x}_j^{mis}$ and randomized classifier as

$$\Pr(b^\epsilon(\mathbf{x}_j^{mis}) = i|\mathcal{Q}_3) = g_i(\mathbf{x}_j^{mis}|\mathbf{w}_{\{i\}}).$$

For a maximum posterior classifier $b(\cdot)$ the above probability is one if $\mathbf{x}_j^{mis}$ is closest to cell i and zero otherwise.

When these are applied to random variable $\boldsymbol{X}^{mis}$ at $\mathcal{Q}_2$ we have class posteriors over data as

$$\Pr\big(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\big) = \int \Pr(b(\mathbf{x}^{mis}) = i) f_{\boldsymbol{X}^{mis}}(\mathbf{x}^{mis}) d\mathbf{x}^{mis}.$$

For randomized classifier $b^{\epsilon}(\cdot)$ we replace $\Pr(b(\mathbf{x}^{mis}) = i)$ by $g_i(\mathbf{x}^{mis})$. When conditionalization is changed to $\mathcal{Q}_1 = \{\mathsf{n}\}$ the probability is computed over uncertain classifier

$$\Pr(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}) = \int \Pr\big(b(\mathbf{x}^{mis}|\mathbf{w}_{\{i\}}) = i\big) f_{\boldsymbol{X}^{mis}}(\mathbf{x}^{mis}) f_{\hat{\boldsymbol{W}}|\mathsf{n}}(\mathbf{w}) d\mathbf{x}^{mis} d\mathbf{w}.$$

It should be noted that under the assumption of (approximate) MCAR with cells missingness we may approximate

$$\Pr(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2) \approx \frac{\pi_i p_i}{p^*},$$

where p_i is the missingness proportion in cell i and p^* is the probability of missingness. Quantities $p_i, i = 1, \ldots, \mathsf{n}_c$ are defined as

$$p_i = \frac{\int_{V_i} \Pr(R = 0|y, \mathbf{x}) f_{Y,\mathbf{X}}(y, \mathbf{x}) dy d\mathbf{x}}{\int_{V_i} f_{Y,\mathbf{X}}(y, \mathbf{x}) dy d\mathbf{x}},$$

in which $\Pr(R = 0|y, \mathbf{x})$ is the probability of missingness at $y, \mathbf{x}$.

6.4 Preservation of moments

The analysis of cell imputation is quite a challenging problem. Imputation performance depends on the positions of cells in input space and possible uses of smoothers, which add strength to cell estimates. In addition there are two ways to associate (classify) incomplete observations to cells: deterministic and randomized.

To make the results as readable as possible, we start from a fixed model (conditionalization level $\mathcal{Q}_2$). This means that Voronoi regions (cells) are fixed, which eliminates difficult questions about the estimation distribution of Voronoi cells for a given data set with n observations. Later we try to answer these questions by characterizing the properties of cell methods. One should note that the results for smoothed random donor strategy are not included in this study.

The result is given in terms of Approximation 6.1, which is derived from

$$\begin{aligned}\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp}|\mathcal{Q}_2] &= \mathbb{E}\Big[\frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathsf{N}^{mis}\hat{\mu}^{imp}) - \hat{\mu} + \hat{\mu} - \mu^*|\mathcal{Q}_2\Big] \\ &= \mathbb{E}\Big[\frac{1}{\mathsf{n}}(\sum_{i=1}^{\mathsf{n}_c} \mathsf{N}_i^{mis}\hat{\mu}_i^{imp} - \mathsf{N}^{mis}\hat{\mu}^{mis})|\mathcal{Q}_2\Big] + \mathbb{E}[\hat{\mu}|\mathcal{Q}_2] - \mu^*,\end{aligned}$$

in terms of expected values of N_i^{mis} and $\hat{\mu}_i^{imp}$ in cells $i = 1, \ldots, \mathsf{n}_c$.

Approximation 6.1 *Approximation of* $\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp}|\mathcal{Q}_2]$, $\mathcal{Q}_2 = \{n, \mathbf{d}^{train}, \mathbf{w}_{\{i\}}\}$.
The bias of $\hat{\mu}^{comp}$ for n observations, with fixed training data, and fixed imputation

model may be approximated as

$$\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}_2] \approx \underbrace{\frac{1}{\mathsf{n}}(\sum_i \mathbb{E}[\mathsf{N}_i^{mis}|\mathcal{Q}_2]\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] - \mathsf{n}^{mis}\mu^{*mis})}_{\text{bias due to imputation method}}$$
$$+ \underbrace{\frac{1}{\mathsf{n}}(\mathsf{n}^{obs}\mu^{obs} + \mathsf{n}^{mis}\mu^{*mis}) - \mu^*}_{\text{finite sample estimation error}},$$

where $\mathbb{E}[\mathsf{N}_i^{mis}|\mathcal{Q}_2] = \Pr\Big(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big)\mathsf{n}^{mis}$ *and*

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] = \begin{cases} \mu_i^{obs} & : \text{C/CJ/T/TJ(S = M/R/D)} \\ \mathbb{E}[\hat{\mu}_i^{s}|\mathcal{Q}_2] \approx \frac{\sum_l h_{i,l}(\mathsf{n}_l^{obs}+\mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])\mu_l^{obs}}{\sum_l h_{i,l}(\mathsf{n}_l^{obs}+\mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])} & : \text{T/TJ(S} = \text{M}^\text{s}/\text{R}^\text{s}\text{)}, \end{cases}$$

where $\Pr\Big(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big)$ is the probability that random data $\boldsymbol{X}^{mis}$ is associated to the i:th cell. Note that in the above approximation $b(\boldsymbol{X}^{mis})$ is replaced by $b^{\epsilon}(\boldsymbol{X}^{mis})$ for TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods.

The bias of the first moment estimator depends on two quantities: the bias due to imputation method and the finite sample estimation error. The second quantity cannot be affected, whereas the first quantity may be varied by changing the imputation method.

The proof for this and for Approximations 6.2-6.5 and Consequence 6.6 are given in Appendix A6. Approximation 6.1 was derived by applying the first order Taylor approximation. However, it is possible to see that the difference between standard $\boldsymbol{X}$ and joint $(Y, \boldsymbol{X})$ clustering via TS-SOM methods depends on how incomplete observations are classified to cells. To get more insight to quantities $\Pr\Big(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big)$, the distribution of missing data values has to be specified. Such distributions are proposed in Section 6.4.1 together with a more detailed decomposition of probability.

The bias $\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathsf{n}]$ requires quite complex integrations. Therefore we shall summarize the result in a form of Approximation 6.2.

Approximation 6.2 *Approximation of* $\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathsf{n}]$*.*
The bias of first moment $\hat{\mu}^{comp}$ *given* n *observations can be approximated as*

$$\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathsf{n}] \approx p^*(\underbrace{\sum_{i=1}^{\mathsf{n}_c} \Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] - \mu^{*mis}}_{\text{(weighted) difference between mean of imputed and missing Y values}})$$
$$+ \underbrace{O(\mathsf{n}^{-1})}_{\text{approximation error}},$$

where

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] = \begin{cases} \mathbb{E}[\hat{\mu}_i^{obs}|\mathsf{n}] = \mu_i^{*obs} & : \text{C/CJ/T/TJ(S = M/R/D)}, \\ \mathbb{E}[\hat{\mu}_i^{s}|\mathsf{n}] \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}]\mu_l^{*obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}]} & : \text{T/TJ(S} = \text{M}^\text{s}/\text{R}^\text{s}\text{)} \end{cases}$$

where μ_l^{*obs} is the expectation of observed Y values in the l:th cell and in which

$$\begin{aligned}\mathbb{E}\Big[\mathsf{N}_l|\mathsf{n}\Big] &= \mathbb{E}\Big[\mathsf{N}_l^{obs}|\mathsf{n}\Big]+\mathbb{E}\Big[\mathsf{N}_l^{mis}|\mathsf{n}\Big]\\ &\approx \underbrace{\mathsf{n}(1-p^*)\mathrm{Pr}\Big(\hat{b}\big((Y^{obs},\boldsymbol{X}^{obs})^T\big)=l|\mathsf{n}\Big)}_{\text{expected number of complete observations}}\\ &+ \underbrace{\mathsf{n}p^*\mathrm{Pr}\Big(\hat{b}(\boldsymbol{X}^{mis})=l|\mathsf{n}\Big)}_{\text{expected number of incomplete observations}}\quad .\end{aligned}$$

Note that $\hat{b}(\boldsymbol{X}^{mis})$ is replaced by $\hat{b}^{\epsilon}(\boldsymbol{X}^{mis})$ for TS-SOM joint $(Y,\boldsymbol{X})$ clustering methods.

Overall bias is accumulated by the biases within cells, which are weighted by the proportion of imputed data values in the cells. There are two sources for the biases in the cells. First, expectations of observed and missing values of Y in cells may differ under MAR and NMAR mechanisms. Secondly, smoothing may cause bias, even under the MCAR mechanism. However, bias due to these sources may accumulate to neglible value. In our standard cell methods with $\boldsymbol{X}$ clustering, the number of imputed Y values within cells is the same as the correct number of missing Y values belonging to the cells. Therefore there is no "prior bias" for standard clustering, but this is not true for joint cell methods, because the numbers of imputed and missing values may differ considerably. The reason for this is that only covariate information is available for classification. Further, TS-SOM methods use a randomized cell selector $\hat{b}^{\epsilon}(\boldsymbol{X}^{mis})$ which may increase the average number of incorrect classifications.

The variance of $\hat{\mu}^{comp}$ is also affected by the imputation strategy and the random nature of the number of missing data values within the cells. As earlier, we first give the result at conditionalization $\mathcal{Q}_2$, after which an implicit result is given at $\mathcal{Q}_1$. The result is given in the form of Approximation 6.3.

Approximation 6.3 *Approximation of* $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2]$, $\mathcal{Q}_2=\{\mathsf{n},\mathbf{d}^{train},\mathbf{w}_{\{i\}}\}$.

The variance of $\hat{\mu}^{comp}$ for n observations, fixed training data and fixed imputation model can be approximated as

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2] &\approx \underbrace{\frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T\mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]}_{\text{due to randomness of number of missing Y values within cells}}\\ &+ \underbrace{\frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\sum_{i=1}^{\mathsf{n}_c}\mathrm{Pr}\Big(b(\boldsymbol{X}^{mis})=i|\mathcal{Q}_2\Big)\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2]}_{\text{A: due to modelled noise}}\quad ,\end{aligned}$$

where $\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]=(\mathbb{E}[\hat{\mu}_1^{imp}|\mathcal{Q}_2],\ldots,\mathbb{E}[\hat{\mu}_{\mathsf{n}_c}^{imp}|\mathcal{Q}_2])^T$ *in which*

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2]=\begin{cases}\mathbb{E}[\hat{\mu}_i^{obs}|\mathcal{Q}_2]=\mu_i^{obs} & :\mathrm{C/T/CJ/TJ(S=M/R/D)}\quad ,\\ \mathbb{E}[\hat{\mu}_i^{s}|\mathcal{Q}_2]\approx\frac{\sum_l h_{i,l}(\mathsf{n}_l^{obs}+\mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])\mu_l^{obs}}{\sum_l h_{i,l}(\mathsf{n}_l^{obs}+\mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])} & :\mathrm{T/TJ(S=M^s/R^s)}\end{cases}\quad .$$

Term $\boldsymbol{N}^{mis} = (\mathsf{N}_1^{mis}, \ldots, \mathsf{N}_{\mathsf{n}_c}^{mis})^T$ *and term* $\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2]$ *depends on the cell method and on imputation strategy* $\hat{\epsilon}^S$ *as follows:*

$$\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2] = \begin{cases} 0 & : \mathrm{S} = \mathrm{M}/\mathrm{M}^{\mathrm{s}} \quad , \\ \\ \mathbb{E}[\hat{\tau}_i^{obs}|\mathcal{Q}_2] = \tau_i^{obs} & : \mathrm{C/CJ/T/TJ(S = R)} \quad , \\ \tau_i^{obs}(1 - \frac{1}{\mathsf{n}_i^{obs}}) & : \mathrm{C/CJ/T/TJ(S = D)} \quad , \\ \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathcal{Q}_2] = \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l \hat{\tau}_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l}|\mathcal{Q}_2] & : \mathrm{T(S = R^s)} \quad , \\ \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2] \tau_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2]} & \\ \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathcal{Q}_2] = \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l \hat{\tau}_l^{w}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l}|\mathcal{Q}_2] & : \mathrm{TJ(S = R^s)} \quad , \\ \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2] \mathbb{E}[\hat{\tau}_l^{w}|\mathcal{Q}_2]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2]} & \end{cases}$$

in which

$$\begin{aligned} \hat{\tau}_i^{obs} &= \frac{1}{\mathsf{N}_i^{obs}} \sum_{j=1}^{\mathsf{N}_i^{obs}} (Y_{j,i}^{obs} - \hat{\mu}_i^{obs})^2 \\ \mathbb{E}[\hat{\tau}_l^{w}|\mathcal{Q}_2] &= \mathbb{E}[\frac{1}{\mathsf{N}_l^{obs}} \sum_{j=1}^{\mathsf{N}_l^{obs}} (Y_{j,l}^{obs} - \hat{\mu}_l^{s})^2 |\mathcal{Q}_2] \approx \frac{1}{\mathsf{n}_l^{obs}} \sum_{j=1}^{\mathsf{n}_l^{obs}} (y_{j,l}^{obs} - \mathbb{E}[\hat{\mu}_l^{s}|\mathcal{Q}_2])^2, \end{aligned}$$

where $Y_{j,l}^{obs}$ *is the* j*:th random observation of* Y^{obs} *in* l*:th cell and*

$$\mathbb{E}[\hat{\mu}_l^{s}|\mathcal{Q}_2] \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \big(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2]\big) \mu^{obs}}{\sum_l h_{i,l} \big(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2]\big)}.$$

According to Approximation 6.3, the main part of the variance is caused by the first term, which is due to the variance of mean imputations. Its value is controlled by the variance of the number of missing data values within the cells and the weighted sum of smoothed mean estimates. The diagonal and off-diagonal elements of covariance-variance matrix $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$ are of the order $O(\mathsf{n}^{mis})$, because the number of incomplete observations within cells must sum to n^{mis}. As earlier, if one is able to define the distribution of $\boldsymbol{N}^{mis}$, then one could derive the covariance-variance structure of Approximation 6.3, as given in Section 6.4.1. Further, from Approximation 6.3 one can discover the factors causing the variance increase due to modelled noise terms. This quantity is the term A. For random (R) and donor (D) imputation strategies the increase in variance is proportional to ratio $\frac{\mathsf{n}^{mis}}{\mathsf{n}^2}$, classification probabilities and noise variance estimates within cells. Increase of variance due to modelled noise terms can be considerable if $\frac{\mathsf{n}^{mis}}{\mathsf{n}^2}$ or estimated noise terms are large.

Again the derivation of $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_1]$ is rather complex, and therefore we shall give the result in a quite implicit form. The result given in Approximation 6.4 is derived by applying the chain rule of variance as follows

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathsf{n}] = \mathbb{E}\big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2]\big] + \mathbb{V}\mathrm{ar}\big[\mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_2]\big],$$

where Approximation 6.3 is applied to compute $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2]$.

Approximation 6.4 *Approximation of* $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathsf{n}]$.

The variance of the first moment $\hat{\mu}^{comp}$ *given* n *observations can be approximated as*

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathsf{n}] \\
\approx\quad & \underbrace{\mathbb{E}\left[\frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T\mathbb{V}\mathrm{ar}\left[\boldsymbol{N}^{mis}|\mathcal{Q}_2\right]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]\middle|\mathsf{n}\right]}_{\text{due to randomness of test data and classification of incomplete observations}} \\
+\quad & \underbrace{\mathbb{V}\mathrm{ar}\left[\frac{1}{\mathsf{n}}\Big(\mathsf{N}^{obs}\hat{\mu}^{obs}+\mathsf{N}^{mis}\sum_{i=1}^{n_c}\Pr(b(\boldsymbol{X}^{mis})=i|\mathcal{Q}_2)\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2]\Big)\middle|\mathsf{n}\right]}_{\text{due to randomness of training data,imputation model, and number of missing Y values}} \\
+\quad & \underbrace{\frac{p^*}{\mathsf{n}}\sum_{i=1}^{n_c}\Pr(\hat{b}(\boldsymbol{X}^{mis})=i|\mathsf{n})\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]}_{\text{variance due to modelled noise}}\ ,
\end{aligned}$$

where $\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]=(\mathbb{E}[\hat{\mu}_1^{imp}|\mathcal{Q}_2],\ldots,\mathbb{E}[\hat{\mu}_{n_c}^{imp}|\mathcal{Q}_2])^T$ *in which*

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2]=\begin{cases}\mu_i^{*obs} & \\ & :\mathrm{C/CJ/T/TJ(S=M/R/D)}, \\ \mathbb{E}[\hat{\mu}_i^s|\mathcal{Q}_2]\approx\frac{1}{\mathsf{n}}\sum_{i=1}^{n_c}\Pr(b(\boldsymbol{X}^{mis})=i|\mathcal{Q}_2)\frac{\sum_{l=1}^{n_c}h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]\hat{\mu}_l^{obs}}{\sum_{l=1}^{n_c}h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]} & \\ & :\mathrm{T/TJ(S=M^s/R^s)},\end{cases}$$

and terms $\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]$ *depend on cell method and on imputation strategy* $\hat{\epsilon}^S$ *as follows:*

$$\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]=\begin{cases}0 & :\mathrm{S=M/M^s}, \\ \mathbb{E}[\hat{\tau}_i^{obs}|\mathsf{n}]\approx\tau_i^{*obs} & :\mathrm{C/CJ/T/TJ(S=R)}, \\ \approx\tau_i^{*obs}\left(1-\frac{1}{\mathsf{n}(1-p^*)\Pr\left(\hat{b}\left(\boldsymbol{X}^{obs}\right)=i|\mathsf{n}\right)}\right) & :\mathrm{C/T(S=D)}, \\ \approx\tau_i^{*obs}\left(1-\frac{1}{\mathsf{n}(1-p^*)\Pr\left(\hat{b}\left((Y^{obs},\boldsymbol{X}^{obs})^T\right)=i|\mathsf{n}\right)}\right) & :\mathrm{CJ/TJ(S=D)}, \\ \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathsf{n}]\approx\frac{\sum_{l=1}^{n_c}h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]\tau_l^{*obs}}{\sum_{l=1}^{n_c}h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]} & :\mathrm{T(S=R^s)},\ \text{and} \\ \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathsf{n}]\approx\frac{\sum_{l=1}^{n_c}h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]\mathbb{E}[\hat{\tau}_l|\mathcal{Q}_1]}{\sum_{l=1}^{n_c}h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]} & :\mathrm{TJ(S=R^s)}.\end{cases}$$

in which

$$\hat{\tau}_l=\frac{1}{\mathsf{N}_l^{obs}}\sum_{j=1}^{\mathsf{N}_l^{obs}}(Y_{j,l}^{obs}-\hat{\mu}_l^s)^2,$$

where $Y_{j,l}^{obs}$ *is j:th (random) observation of* Y^{obs} *in the l:th cell.*

For TS-SOM joint $(Y,\boldsymbol{X})$ clustering methods classifier in quantities $\hat{b}(\boldsymbol{X}^{mis})$ and $b(\boldsymbol{X}^{mis})$ is replaced by randomized classifiers $\hat{b}^{\epsilon}$ and b^{ϵ}.

The bias of the second moment is dependent on the distribution of cells, on smoothing, on the classifier of incomplete observations, and on imputation strategy. The derivation of the result is based on the computation of

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\tau}^{comp}|\mathsf{n}] &= \mathbb{E}[\hat{\tau}^{comp}|\mathsf{n}] - \tau^{*} \\
&= \mathbb{E}[(1 - \frac{\mathsf{N}^{mis}}{\mathsf{n}-1})\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} + \frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs} - \hat{\mu}^{imp})^{2}|\mathsf{n}] - \tau^{*},
\end{aligned}$$

where the result is computed by applying the first order Taylor approximation around expected value of all random variables. We can summarize the behavior of the bias using Approximation 6.5.

Approximation 6.5 $\mathbb{B}\text{ias}[\hat{\tau}^{comp}|n]$.
The bias of $\hat{\tau}^{comp}$ *for* n *observations may be approximated as*

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\tau}^{comp}|n] \approx\ & \underbrace{p^{*}(\sum_{i=1}^{n_c} p_i^{mis}(\mu_i^{*imp} - \sum_{l=1}^{n_c} p_l^{mis}\mu_l^{*imp})^{2} + B - \tau^{*mis})}_{\text{difference between variance of imputed and missing Y values}} \\
& \underbrace{+\,p^{*}(1-p^{*})\Big((\mu^{*obs} - \sum_{l=1}^{n_c} p_l^{mis}\mu_l^{*imp})^{2} - (\mu^{*obs} - \mu^{*mis})^{2}\Big)}_{\text{difference between mean of imputed and missing Y values}} \\
& + \underbrace{\qquad O(n^{-1}) \qquad}_{\text{approximation error}},
\end{aligned}$$

where the term $p_l^{mis} = \Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = l|n\Big)$, *and terms* μ_l^{*imp} *depend on the cell method and strategy as follows:*

$$\mu_l^{*imp} = \begin{cases} \mu_l^{*obs} & : \mathrm{T/TJ(S = M/R/D)} \quad , \\ \mu_l^{*s} & : \mathrm{T/TJ(S = M^s/R^s)} \quad , \end{cases}$$

and term B *is due to noise modelling, and depends on the cell method and on imputation strategy* $\hat{\epsilon}^{S}$ *as follows:*

$$B = \begin{cases} 0 & : \mathrm{S = M/M^s} \quad , \\ & \\ \sum_{l=1}^{n_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{obs}|n] & : \mathrm{CJ/TJ/C/T(S = R)} \quad , \\ \sum_{l=1}^{n_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{obs}|n](1 - \frac{1}{\mathbb{E}[\mathsf{N}_l^{obs}|n]}) & : \mathrm{CJ/TJ/C/T(S = D)} \quad , \\ & \\ \sum_{l=1}^{n_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{T,R^s}|n] & : \mathrm{T(S = R^s)} \quad , \text{ and} \\ \sum_{l=1}^{n_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{TJ,R^s}|n] & : \mathrm{TJ(S = R^s)} \quad . \end{cases}$$

Quantity $\hat{b}(\cdot)$ is replaced by $\hat{b}^{\epsilon}(\cdot)$ in p_l^{mis} for TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods.

There are two possible sources for bias. First, the variance of the imputed and missing Y values may differ. Secondly, the means of the imputed and missing Y values may differ. Typically imputation leads to underestimation of the second moment (for finite sample size n), even if noise is modelled. Joint $Y, \boldsymbol{X}$ clustering with TS-SOM or K-Means methods are exceptions to this. Namely, they may yield a positively biased (overestimated) second moment. The reason for this is the misclassification of incomplete observations. It is possible that on average more incomplete observations are classified to cells with high variances than should be.

As always these results are simplified when $\mathsf{n} \to \infty$. The asymptotics is given in the form of Consequence 6.6.

Consequence 6.6 *Approximation of asymptotic behaviour $\mathsf{n} \to \infty$ of the first two moment estimators.*
Asymptotically we have the following approximations

i)

$$\begin{aligned}\lim_{\mathsf{n}\to\infty} \mathbb{B}\mathrm{ias}[\hat{\mu}^{comp}|\mathsf{n}] &\approx p^*(\sum_{i=1}^{\mathsf{n}_c} p_i^{mis}\mu_i^{*imp} - \mu^{*mis}) \\ \lim_{\mathsf{n}\to\infty} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathsf{n}] &\approx 0,\end{aligned}$$

where

$$\mu_i^{*imp} = \begin{cases} \mu_i^{*obs} & : \mathrm{C/CJ/T/TJ(S = M/R/D)} \quad , \\ \mu_i^{*s} \approx \frac{\sum_l h_{i,l}(p_l^{mis}+p_l^{obs})\mu_l^{*obs}}{\sum_l h_{i,l}(p_l^{mis}+p_l^{obs})} & : \mathrm{T/TJ(S = M^s/R^s)} \quad . \end{cases}$$

ii)

$$\begin{aligned}&\lim_{\mathsf{n}\to\infty} \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp}|\mathsf{n}] \\ \approx\ & \underbrace{p^*(\sum_{i=1}^{\mathsf{n}_c} p_i^{mis}(\mu_i^{*imp} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 + C - \tau^{*mis})}_{\text{difference between variance of imputed and missing Y values}} \\ & \underbrace{+ p^*(1-p^*)\Big((\mu^{*obs} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 - (\mu^{*obs} - \mu^{*mis})^2\Big)}_{\text{difference between the mean of imputed and missing Y values}},\end{aligned}$$

where $p_i^{mis} = \lim_{\mathsf{n}\to\infty} \Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)$, $p_i^{obs} = \lim_{\mathsf{n}\to\infty} \Pr\Big(\hat{b}\big((Y^{obs}, \boldsymbol{X}^{obs})^T\big) = i|\mathsf{n}\Big)$, and C in ii) is due to noise modelling and depends on the cell method and on imputation strategy $\hat{\epsilon}^S$ as follows:

$$C = \begin{cases} 0 & : \mathrm{(S = M/M^s)} \quad , \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\tau_l^{*obs} & : \mathrm{C/CJ/T/TJ(S = R/D)} \quad , \\ & \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\tau}_l^{T,R^s}|\mathsf{n}] & : \mathrm{T(S = R^s)} \quad , \text{ and} \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\tau}_l^{TJ,R^s}|\mathsf{n}] & : \mathrm{TJ(S = R^s)} \quad . \end{cases}$$

From Consequence 6.6 one sees that the approximative variance of the first moment estimators go to zero as sample size grows to infinity. This is likely to hold exactly (unless imputation variance goes to infinity which is not realistic). Therefore these estimators are consistent. The estimator of the second moment may be asymptotically biased. The main difference between the finite and limiting results is that the limiting results are clearer, because there are no finite sample estimation errors. Justifications for all approximations and the above consequence are given in Appendix A6.

Next we give some insight to covariance-variance matrix $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$ which appears in the approximative variances of the mean estimator. Further, some details of $\Pr(b^{\epsilon}(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2)$ are given also.

6.4.1 Distribution assumptions for the number of missing data values in the cells

In order to use Approximations 6.2, 6.4 and 6.5 in practice one should be able to estimate the values of $\Pr(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n})$, $\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}]$, $\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]$, $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$, $\Pr(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2)$, and $\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]$. Therefore our purpose here is to obtain estimates for some of these. Most importantly we are interested about variance-covariance matrix that describes the number of missing data values within cells: $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$. In addition, we need to know the classification probabilities $\Pr(b(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2)$. Recall that Approximation 6.3 about variance of mean estimator $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2]$ contains term $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$. Our approach is to define the distribution of $\boldsymbol{N}^{mis}$ given $\mathcal{Q}_2$ (or $\mathcal{Q}_3$) for standard cell methods with $\boldsymbol{X}$ clustering and for TS-SOM joint $(Y, \boldsymbol{X})$ clustering under some assumptions. Note that the "interpretable" distribution assumption is difficult to set at conditionalization $\mathcal{Q}_1 = \{\mathsf{n}\}$, because the imputation model (cells) is not identifiable. We do not know the number of incomplete observations, and the number of complete observations within the cells is random. Thus conditionalization $\mathcal{Q}_1$ is not considered here.

The assumptions are characterized using cell specific quantities p_i and π_i, which are representing missingness and the prior of the cells, as explained later. The exact values of p_i and π_i for each of the cells $i = 1, \dots, \mathsf{n}_c$ depend on the data and on the clustering method. Thus an application of the results of this chapter requires that one obtains the distributions of p_i and π_i via experiments, for example.

Standard clustering of $\boldsymbol{X}$-space

First note that at conditionalization $\mathcal{Q}_3 = \{\mathbf{d}^{train}, \mathbf{d}^{test}, \mathbf{w}_{\{i\}}\}$ the number of complete and incomplete observations within cells are fixed. Therefore covariance-variance matrix $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_3]$, where $\boldsymbol{N}^{mis} = (\mathsf{N}_1^{mis}, \dots, \mathsf{N}_{\mathsf{n}_c}^{mis})$, is zero.

Thus, the real interest is to study conditionalization $\mathcal{Q}_2$ where the test data $\mathbf{D}^{test}$ is random. We shall do this under the assumptions

i) Missingness between the cells is of MAR type where,

$\Pr(\text{"}Y \text{ is missing"}|\mathbf{x}, i, \mathcal{Q}_2) \approx p_i,\ i = 1, \ldots, \mathsf{n}_c.$

ii) Missingness inside the cells is MCAR.

iii) The distribution of $\boldsymbol{N}^{mis}|\mathcal{Q}_2$ (for incomplete observations counts) is assumed to follow multinomial distribution

$$\boldsymbol{N}^{mis}|\mathsf{n}, \mathbf{d}^{train}, \mathbf{w}_{\{i\}} \sim Multin\left(\mathsf{n}^{mis}; \frac{\pi_1 p_1}{p^*}, \ldots, \frac{\pi_{\mathsf{n}_c} p_{\mathsf{n}_c}}{p^*}\right), \tag{6.9}$$

where n_c is the number of cells, n^{mis} is the number of incomplete observations, $\pi_i = \Pr(i|\mathcal{Q}_2)$ is the prior of the i:th cell.

Under these assumptions the variance of $\boldsymbol{N}^{mis}|\mathcal{Q}_2$ can be written in terms of approximation 6.7.

Approximation 6.7 *Approximation to second order moments of* $\boldsymbol{N}^{mis}|\mathcal{Q}_2$.
Given assumption (6.9) the second order moments of $\boldsymbol{N}^{mis}$ *can be approximated as:*

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\mathsf{N}_i^{mis}|\mathcal{Q}_2] &\approx n^{mis}\frac{\pi_i p_i}{p^*}(1 - \frac{\pi_i p_i}{p^*}) \\ \mathbb{C}\mathrm{ov}[\mathsf{N}_i^{mis}, \mathsf{N}_l^{mis}|\mathcal{Q}_2] &\approx -n^{mis}\frac{\pi_i p_i \pi_l p_l}{(p^*)^2},\ i \neq l. \end{aligned} \tag{6.10}$$

The result follows immediately from the assumptions when using the basic properties of multinomial distribution (see Appendix A3.1.2 for details). Note that the covariances between the counts are negative because the counts must add to n^{mis}. The result can be simplified if the missingness between the cells is MCAR. Then Consequence 6.8 holds.

Consequence 6.8 *Simplification under MCAR mechanism.*
Under MCAR second order moments simplify to

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\mathsf{N}_i^{mis}|\mathcal{Q}_2] &\approx n^{mis}\pi_i(1 - \pi_i) \\ \mathbb{C}\mathrm{ov}[\mathsf{N}_i^{mis}, \mathsf{N}_l^{mis}|\mathcal{Q}_2] &\approx -n^{mis}\pi_i\pi_l,\ i \neq l. \end{aligned} \tag{6.11}$$

The result follows immediately when we notice that under MCAR $p_i = p^*, i = 1, \ldots, n_c$.

Joint $(Y, \boldsymbol{X})$-clustering with TS-SOM

The main difference between our standard clustering with $\boldsymbol{X}$-space and joint clustering with $(Y, \boldsymbol{X})$-space using TS-SOM is the way in which data are associated to cells. In standard clustering, classification of $(y^{mis}, \mathbf{x}^{mis})$ is done deterministically by finding the closest cell to $\mathbf{x}^{mis}$. In joint $(y, \mathbf{x})$ clustering, where decisions on "correct" cell depend on incomplete data Y, the selection of the best cell is randomized according to cell posterior probabilities $\Pr(i|\mathbf{x}^{mis})$. The benefit is that TJ type of imputation is able to handle multimodal distributions of $f_{Y^{mis}|\boldsymbol{X}^{mis}}(y^{mis}|\mathbf{x})$. The problem is that $\mathbf{x}^{obs}$ may not be associated to a "correct" cell, which makes the analysis of counts N_i^{mis} more difficult.

In our approach we need to decompose N_i^{mis} to "correct" and "incorrect" counts as below. Let $\mathsf{N}_{i,l}^{mis}$ be a random variable denoting how many observations belonging to the i:th cell were classified to the j:th cell, thus $i = 1, \ldots, \mathsf{n}_c$ and $l = 1, \ldots, \mathsf{n}_c$. The number of incomplete observations in the i:th cell is the sum of the number of correctly classified incomplete observations and misclassifications from other cells, formally

$$\mathsf{N}_i^{mis} = \underbrace{\mathsf{N}_{i,i}^{mis}}_{\text{correct classifications}} + \underbrace{\sum_{l\neq i} \mathsf{N}_{l,i}^{mis}}_{\text{misclassifications from other cells}} = \sum_{l=1}^{\mathsf{n}_c} \mathsf{N}_{l,i}^{mis}.$$

We assume that misclassifications of incomplete observations belonging to a cell are uniformly spread to other cells. Distributions of $\mathsf{N}_{i,l}^{mis}$ are supposed to be

$$\begin{aligned}
&\mathsf{N}_{1,1}^{mis}, \ldots, \mathsf{N}_{1,\mathsf{n}_c}^{mis} | \mathcal{Q}_3 \sim Multin\Big(\mathsf{n}_1^{mis,c}; q_1, \frac{1}{\mathsf{n}_c - 1}(1-q_1), \ldots, \frac{1}{\mathsf{n}_c - 1}(1-q_1)\Big),\\
&\vdots\\
&\mathsf{N}_{\mathsf{n}_c,1}^{mis}, \ldots, \mathsf{N}_{\mathsf{n}_c,\mathsf{n}_c}^{mis} | \mathcal{Q}_3 \sim Multin\Big(\mathsf{n}_{\mathsf{n}_c}^{mis,c}; \frac{1}{\mathsf{n}_c - 1}(1-q_{\mathsf{n}_c}), \ldots, \frac{1}{\mathsf{n}_c - 1}(1-q_{\mathsf{n}_c}), q_{\mathsf{n}_c}\Big),
\end{aligned}$$

where $\mathsf{n}_i^{mis,c}$ is the correct number of incomplete observations belonging to the i:th cell and q_i is the expected success ratio in the classification of incomplete observations belonging to the i:th cell given $\mathcal{Q}_3$. Value one of q_i means that all incomplete observations belonging to cell i are classified to it on expectation. However, in practice the value of q_i is below one because the Y part is missing. If the cells are well separated in the $\boldsymbol{X}$ space then the quantities q_i can be high, however if the cells are overlapping then q_is are expected to be somewhat low.

At conditionalization Q_2 the number of complete observations remains fixed due to the conditionalization of a complete part of the true sample. However, the true number of incomplete observations belonging to cells becomes random. Provided the MCAR within cells approximation holds then the distribution is assumed to be

$$\mathsf{N}_1^{mis,c}, \ldots, \mathsf{N}_{\mathsf{n}_c}^{mis,c} | \mathcal{Q}_2 \sim Multin\Big(\mathsf{n}^{mis}; \frac{\pi_1 p_1}{\sum_i \pi_i p_i}, \ldots, \frac{\pi_{\mathsf{n}_c} p_{\mathsf{n}_c}}{\sum_i \pi_i p_i}\Big).$$

The distribution of classification counts is supposed to be the following:

$$\begin{aligned}
&\mathsf{N}_{1,1}^{mis}, \mathsf{N}_{1,2}^{mis}, \ldots, \mathsf{N}_{1,\mathsf{n}_c}^{mis}, \ldots, \mathsf{N}_{\mathsf{n}_c,1}^{mis}, \mathsf{N}_{\mathsf{n}_c,2}^{mis}, \ldots, \mathsf{N}_{\mathsf{n}_c,\mathsf{n}_c-1}^{mis}, \mathsf{N}_{\mathsf{n}_c,\mathsf{n}_c}^{mis} | \mathcal{Q}_2 \sim\\
&Multin\Big(\mathsf{n}^{mis};\\
&\frac{1}{z}\pi_1 p_1 \mathbb{E}[\hat{q}_1], \frac{1}{z}\frac{1}{\mathsf{n}_c - 1}\pi_1 p_1 (1 - \mathbb{E}[\hat{q}_1]), \ldots, \frac{1}{z}\frac{1}{\mathsf{n}_c - 1}\pi_1 p_1 (1 - \mathbb{E}[\hat{q}_1]), \qquad (6.12)\\
&\vdots\\
&\frac{1}{z}\frac{1}{\mathsf{n}_c - 1}\pi_{\mathsf{n}_c} p_{\mathsf{n}_c} (1 - \mathbb{E}[\hat{q}_{\mathsf{n}_c}]), \ldots, \frac{1}{z}\frac{1}{\mathsf{n}_c - 1}\pi_{\mathsf{n}_c} p_{\mathsf{n}_c} (1 - \mathbb{E}[\hat{q}_{\mathsf{n}_c}]), \pi_{\mathsf{n}_c} p_{\mathsf{n}_c} \mathbb{E}[\hat{q}_{\mathsf{n}_c}]\Big),
\end{aligned}$$

where $z = \sum_{i=1}^{\mathsf{n}_c} \pi_i p_i$ is the normalization constant and $\mathbb{E}[\hat{q}_i|\mathcal{Q}_2]$ is the expected success ratio in classification of incomplete observations belonging to the i:th cell given the conditionalizers of the second level.

Given the above assumptions we can now i) compute the expected number of the missing data values within cells, and ii) compute the covariance-variance structure for the number of missing data values. The former allows one to better understand quantity $\Pr(b^\epsilon(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2)$ and the latter gives detailed information about $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$. The result is summarized in Approximation 6.9.

Approximation 6.9 $\Pr(b^\epsilon(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2)$ *with randomized classification.*
Under assumption of MCAR within cells and distribution (6.12), the randomized classification probability can be approximated as:

$$\Pr(b^\epsilon(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2) \approx \underbrace{\frac{1}{z}\pi_l p_l \mathbb{E}[\hat{q}_l|\mathcal{Q}_2]}_{\text{probability for correct classication}} + \underbrace{\sum_{j\neq l}^{\mathsf{n}_c} \frac{1}{z}\frac{1}{\mathsf{n}_c - 1}\pi_j p_j (1 - \mathbb{E}[\hat{q}_j|\mathcal{Q}_2]) \qquad ,}_{\text{probability for misclassication from other cells}}$$

which holds because

$$\begin{aligned} \mathsf{n}^{mis}\Pr(b^\epsilon(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2) &= \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2] = \mathbb{E}[\sum_{j=1}^{\mathsf{n}_c} \mathsf{N}_{j,l}^{mis}|\mathcal{Q}_2] \\ &\approx \underbrace{\mathsf{n}^{mis}\frac{1}{z}\pi_l p_l \mathbb{E}[\hat{q}_j|\mathcal{Q}_l]}_{\text{expected number of correct classifications}} \\ &\quad \underbrace{+\,\mathsf{n}^{mis}\sum_{j\neq l}^{\mathsf{n}_c}\frac{1}{z}\frac{1}{\mathsf{n}_c - 1}\pi_j p_j (1 - \mathbb{E}[\hat{q}_j|\mathcal{Q}_2])\,.}_{\text{expected number of misclassifications}} \end{aligned}$$

Under the MCAR missingness the randomized classification probability simplifies, as showin in consequence 6.10.

Consequence 6.10 *Simplification to* $\Pr(b^\epsilon(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2)$ *under MCAR.*
Under MCAR:

$$\Pr\big(b^\epsilon(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2\big) \approx \pi_l \mathbb{E}[\hat{q}_l|\mathcal{Q}_2] + \sum_{j\neq l}^{\mathsf{n}_c} \frac{1}{\mathsf{n}_c - 1}\pi_j (1 - \mathbb{E}[\hat{q}_j|\mathcal{Q}_2]),$$

because

$$\mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2] \approx \mathsf{n}^{mis}\pi_l \mathbb{E}[\hat{q}_l|\mathcal{Q}_2] + \mathsf{n}^{mis}\sum_{j\neq l}^{\mathsf{n}_c} \frac{1}{\mathsf{n}_c - 1}\pi_j (1 - \mathbb{E}[\hat{q}_j|\mathcal{Q}_2]).$$

Computing covariance-variance matrix $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$ is a bit more complicated. See Appendix A6.3 for details on derivation of variance and covariance results, which are given in Approximation 6.11.

Approximation 6.11 *Second order moments of $\boldsymbol{N}^{mis}|\mathcal{Q}_2$, with randomized classification.*

Under the assumption of MCAR within cells and distribution (6.12) one may approximate:

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}[\mathsf{N}_i^{mis}|\mathcal{Q}_2]\\
\approx\;& \underbrace{\mathsf{n}^{mis}\frac{1}{z}\pi_i p_i\mathbb{E}[\hat{q}_i](1-\frac{1}{z}\pi_i p_i\mathbb{E}[\hat{q}_i])}_{\text{variance due correct classifications}}\\
&+\underbrace{\mathsf{n}^{mis}(\mathsf{n}_c-1)\frac{1}{z}\pi_i p_i(1-\mathbb{E}[\hat{q}_i])\big(1-\frac{1}{z}\pi_i p_i(1-\mathbb{E}[\hat{q}_i])\big)}_{\text{variance due to misclassifications}}\\
&\underbrace{-\,\mathsf{n}^{mis}\sum_{l\neq i}\frac{1}{z}\pi_i p_i\mathbb{E}[\hat{q}_i]\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_l p_l(1-\mathbb{E}[\hat{q}_l])}_{\text{covariance between correct classifications to cell i and misclassifications from other cells}}\\
&\underbrace{-\,\mathsf{n}^{mis}\sum_{j\neq i}\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_j p_j(1-\mathbb{E}[\hat{q}_j])\frac{1}{z}\pi_i p_i\mathbb{E}[\hat{q}_i]}_{\text{covariance between misclassifications from other cells and correct classifications to cell i}}\\
&\underbrace{-\mathsf{n}^{mis}\sum_{l\neq j,j\neq i,l\neq i}\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_j p_j(1-\mathbb{E}[\hat{q}_j])\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_l p_l(1-\mathbb{E}[\hat{q}_l])}_{\text{covariance between misclassifications}}\,.
\end{aligned}$$

For $i\neq l$ covariance term is approximated as

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}[\mathsf{N}_i^{mis},\mathsf{N}_l^{mis}|\mathcal{Q}_2]\\
\approx\;& -\mathsf{n}^{mis}\sum_{j\neq i}\sum_{u\neq l}\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_j p_j(1-\mathbb{E}[\hat{q}_j])\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_u p_u(1-\mathbb{E}[\hat{q}_u])\\
&-\mathsf{n}^{mis}\sum_{j\neq i}\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_j p_j(1-\mathbb{E}[\hat{q}_j])\frac{1}{z}\pi_l p_l\mathbb{E}[\hat{q}_l]\\
&-\mathsf{n}^{mis}\sum_{u\neq l}\frac{1}{z}\pi_i p_i\mathbb{E}[\hat{q}_i]\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_u p_u(1-\mathbb{E}[\hat{q}_u])\\
&-\mathsf{n}^{mis}\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_i p_i(1-\mathbb{E}[\hat{q}_i])\frac{1}{z}\frac{1}{\mathsf{n}_c-1}\pi_l p_l(1-\mathbb{E}[\hat{q}_l]).
\end{aligned}$$

In the above variance and covariance formulas conditionalisation $\mathcal{Q}_2$ has been omitted from expectations $\mathbb{E}[\hat{q}_i]$ to make the formulas easier to read.

The second order statistics of the number of missing data values simplify under MCAR, as shown in Consequence 6.12.

Consequence 6.12 *Simplification to distribution of $\boldsymbol{N}^{mis}|\mathcal{Q}_2$ under MCAR. Under MCAR mechanism $p_i = p^*, i = 1, \ldots, \mathsf{n}_c$ and $z = p^*$, thus:*

$$
\begin{aligned}
& \mathbb{V}\mathrm{ar}[\mathsf{N}_i^{mis}|\mathcal{Q}_2] \\
\approx\ & \underbrace{\mathsf{n}^{mis}\pi_i\mathbb{E}[\hat{q}_i](1-\pi_i\mathbb{E}[\hat{q}_i])}_{\text{variance due correct classifications}} \\
& + \underbrace{\mathsf{n}^{mis}(\mathsf{n}_c-1)\pi_i(1-\mathbb{E}[\hat{q}_i])(1-\pi_i\big(1-\mathbb{E}[\hat{q}_i]\big))}_{\text{variance due to misclassifications}} \\
& \underbrace{\qquad\qquad - \mathsf{n}^{mis}\sum_{l\neq i}\pi_i\mathbb{E}[\hat{q}_i]\frac{1}{\mathsf{n}_c-1}\pi_l(1-\mathbb{E}[\hat{q}_l])\qquad\qquad}_{\text{covariance between correct classifications to cell i and misclassifications from other cells}} \\
& \underbrace{\qquad\qquad - \mathsf{n}^{mis}\sum_{j\neq i}\frac{1}{\mathsf{n}_c-1}\pi_j(1-\mathbb{E}[\hat{q}_j])\pi_i\mathbb{E}[\hat{q}_i]\qquad\qquad}_{\text{covariance between misclassifications from other cells and correct classifications to cell i}} \\
& \underbrace{-\mathsf{n}^{mis}\sum_{l\neq j, j\neq i, l\neq i}\frac{1}{\mathsf{n}_c-1}\pi_j(1-\mathbb{E}[\hat{q}_j])\frac{1}{\mathsf{n}_c-1}\pi_l(1-\mathbb{E}[\hat{q}_l])}_{\text{covariance between misclassifications}}\,.
\end{aligned}
$$

Covariance term for $i \neq l$ is approximated as

$$
\begin{aligned}
& \mathbb{C}\mathrm{ov}[\mathsf{N}_i^{mis}, \mathsf{N}_l^{mis}|\mathcal{Q}_2] \\
\approx\ & -\mathsf{n}^{mis}\Big(\sum_{j\neq i}\sum_{u\neq l}\frac{1}{\mathsf{n}_c-1}\pi_j(1-\mathbb{E}[\hat{q}_j])\frac{1}{\mathsf{n}_c-1}\pi_u(1-\mathbb{E}[\hat{q}_u]) \\
& +\sum_{j\neq i}\frac{1}{\mathsf{n}_c-1}p_j(1-\mathbb{E}[\hat{q}_j])\pi_l\mathbb{E}[\hat{q}_l] + \sum_{u\neq l}\pi_i\mathbb{E}[\hat{q}_i]\frac{1}{\mathsf{n}_c-1}p_u(1-\mathbb{E}[\hat{q}_u]) \\
& +\frac{1}{\mathsf{n}_c-1}\pi_i(1-\mathbb{E}[\hat{q}_i])\frac{1}{\mathsf{n}_c-1}\pi_l(1-\mathbb{E}[\hat{q}_l])\Big).
\end{aligned}
$$

Covariance-variance matrix $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$ is simple under the above assumptions for standard clustering methods. Variance $\mathbb{V}\mathrm{ar}[\mathsf{N}_i^{mis}|\mathcal{Q}_2]$ (the i:th diagonal element in the matrix) depends on the number of observations, on the prior of the i:th cell, on the missingness probability in the i:th cell, and on the cells. Covariance $\mathbb{C}\mathrm{ov}[\mathsf{N}_i^{mis}, \mathsf{N}_l^{mis}|\mathcal{Q}_2], i \neq l$ (off-diagonal element in the matrix) is of negative sign and depends on the number of observations, priors and missingness probabilities in the i:th and l:th cells. Under MCAR, missingness quantities simplify as missingness probabilities vanish.

Covariance-variance matrix $\mathbb{V}\mathrm{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$ is quite complicated for joint $(Y, \boldsymbol{X})$-clustering with TS-SOM due to the randomized classifier (which causes misclassifications). Variance $\mathbb{V}\mathrm{ar}[\mathsf{N}_i^{mis}|\mathcal{Q}_2]$ depends on the variance of correct classifications, variance of misclassifications, and on covariance terms. Covariance $\mathbb{C}\mathrm{ov}[\mathsf{N}_i^{mis}, \mathsf{N}_l^{mis}|\mathcal{Q}_2], i \neq l$ depends on four covariance terms, and is difficult to interpret. Classification probabilities $\Pr(b^{\epsilon}(\boldsymbol{X}^{mis}) = l|\mathcal{Q}_2)$ were decomposed into a probability for correct classification and a probability for misclassification from other cells. Decomposition terms for $\mathbb{V}\text{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$, covariance $\mathbb{C}\text{ov}[\mathsf{N}_i^{mis}, \mathsf{N}_l^{mis}|\mathcal{Q}_2]$, and classification probabilities depend on cell priors, missingness within cells, and classification success ratios. Under MCAR, dependency on missingness probabilities vanishes. Note that in our approximation it was assumed that misclassifications spread uniformly to other cells. This is a slightly pessimistic assumption. In practice if Y and $\boldsymbol{X}$ are dependent enough then it is likely that misclassifications spread close to the "correct cell". However, in such case the distribution of $\boldsymbol{N}^{mis}$ is likely to be more complicated.

6.4.2 How to apply results of Section 6.4 in practice

We demonstrate the applicability of previous results by the construction of a practical algorithm to use Approximation 6.4. Similar algorithms can be written for other approximations as well when the actual usage of these approximations is known. The current usage is demonstrated in the following example in Section 6.4.3.

The basic idea of the algorithm is simple: we just plug-in the assumptions given in Section 6.4.1 into Approximation 6.4.

First the approximation has to be rewritten in an explicit form for a given single data set. For this we assume that a first order Taylor approximation is sufficient (first row of the approximation). Next, for simplicity, it is assumed that the variance term (second row of the approximation) is only due to N^{mis} and N^{obs} and that covariance between $\mathsf{N}^{mis}/\mathsf{n}$ and $\mathsf{N}^{obs}/\mathsf{n}$ is neglible with respect to other terms. Now Approximation 6.4 may be simplified as follows

$$\begin{aligned}\mathbb{V}\text{ar}[\hat{\mu}^{comp}|\mathsf{n}] \quad \approx \quad & \underbrace{\frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\text{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]}_{=A} \qquad (6.13)\\ & \underbrace{+\frac{\mathbb{V}\text{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^2}\Big[\mathbb{E}[\hat{\mu}^{obs}]^2 + (\sum_{i=1}^{\mathsf{n}_c}\Pr(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n})\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}])^2\Big]}_{=B}\\ & \underbrace{+\frac{p^*}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\Pr(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n})\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]}_{=C}\,.\end{aligned}$$

Note that the variance of the number of missing data values $\mathbb{V}\text{ar}[\mathsf{N}^{mis}]$ cannot be estimated from a single data set. As a consequence one has to assume a parametric distribution form for N^{mis} to compute the variance. As an example, if $\mathsf{N}^{mis} \sim Bin(\mathsf{n}, p^*)$ then $\mathbb{V}\text{ar}[\mathsf{N}^{mis}] = \mathsf{n}p^*(1-p^*)$.

Secondly, one has to estimate all the quantities in Equation (6.13). The structure of covariance-variance matrix $\mathbb{V}\text{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$ depends on the clustering algorithm. For standard $\boldsymbol{X}$-clustering the entries of the matrix are computed using the results of the Approximation 6.7, whereas for joint $(Y, \boldsymbol{X})$ clustering one uses Approximation 6.11. For joint clustering one has to set by hand or by other means the classification success ratios $\mathbb{E}[\hat{q}_i]$, which cannot be estimated from a completed data set in practice, because the missing Y values are not known. If there is no other information then value 0.5 is a 'good guess'. Finally all values are assigned into Equation (6.13).

Assuming that we have applied standard TS-SOM clustering for a data set of n observations and obtained a clustering with n_c cells having the properties

cell means	$[\mu_1^{obs},\dots,\mu_{\mathsf{n}_c}^{obs}]^T = \boldsymbol{\mu}^{obs} \approx \mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}\vert\mathcal{Q}_2]$
cell variances	$[\tau_1^{obs},\dots,\tau_{\mathsf{n}_c}^{obs}]^T \approx \Big[\mathbb{E}[\hat{\tau}_1^{imp}\vert\mathcal{Q}_2],\dots,\mathbb{E}[\hat{\tau}_{\mathsf{n}_c}^{imp}\vert\mathcal{Q}_2]\Big]^T$
cell sizes	$[\mathsf{n}_1,\dots,\mathsf{n}_{\mathsf{n}_c}]^T$
missingness	$[\mathsf{n}_1^{mis},\dots,\mathsf{n}_{\mathsf{n}_c}^{mis}]^T$,

and let the mean estimate of the observed data values be $\mu^{obs} \approx \mathbb{E}[\hat{\mu}^{obs}]$. Then the practical algorithm can be written as follows

Algorithm 1: Application of 6.4 for standard TS-SOM clustering

Step 1:

$$A = \frac{1}{\mathsf{n}^2}(\boldsymbol{\mu}^{obs})^T\mathbf{V}\boldsymbol{\mu}^{obs},$$

where covariance-variance matrix $\mathbf{V} = \{v_{il}\}_{i=1,\dots,\mathsf{n}_c,l=1,\dots\mathsf{n}_c}$ is defined as

$$v_{il} = \begin{cases} \mathsf{n}^{mis}\frac{(\frac{\mathsf{n}_i}{\mathsf{n}})*(\frac{\mathsf{n}_i^{mis}}{\mathsf{n}_i})}{(\frac{\mathsf{n}^{mis}}{\mathsf{n}})}\left(1-\frac{(\frac{\mathsf{n}_i}{\mathsf{n}})*(\frac{\mathsf{n}_i^{mis}}{\mathsf{n}_i})}{(\frac{\mathsf{n}^{mis}}{\mathsf{n}})}\right), & \text{if } i=l \\ -\mathsf{n}^{mis}\frac{(\frac{n_i}{\mathsf{n}})(\frac{\mathsf{n}_i^{mis}}{\mathsf{n}_i})(\frac{n_l}{\mathsf{n}})(\frac{\mathsf{n}_l^{mis}}{\mathsf{n}_l})}{(\frac{\mathsf{n}^{mis}}{\mathsf{n}})^2}, & \text{if } i \neq l \end{cases}$$

Step 2:

$$B = \frac{\mathsf{n}(\frac{\mathsf{n}^{mis}}{\mathsf{n}})(1-\frac{\mathsf{n}^{mis}}{\mathsf{n}})}{\mathsf{n}^2}\left[[\mu^{obs}]^2 + (\sum_{i=1}^{\mathsf{n}_c}\frac{\mathsf{n}_i^{mis}}{\mathsf{n}^{mis}}\tau_i^{obs})^2\right],$$

where $\mathsf{n}^{mis} = \sum_{i=1}^{\mathsf{n}_c}\mathsf{n}_i^{mis}$.

Step 3:

$$C = \frac{(\frac{\mathsf{n}^{mis}}{\mathsf{n}})}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\frac{\mathsf{n}_i^{mis}}{\mathsf{n}^{mis}}\tau_i^{obs}$$

Step 4: variance estimates for imputation strategies are

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,T,M}\vert\mathsf{n}] &\approx A+B \\ \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,T,S}\vert\mathsf{n}] &\approx A+B+C, \end{aligned}$$

where strategy $S \in \{R,D\}$.

Note that in Step 2 of the algorithm we have assumed that $\mathsf{N}^{mis} \sim Bin(\mathsf{n},p^*)$. If this does not hold then nominator $\mathsf{n}(\frac{\mathsf{n}^{mis}}{\mathsf{n}})(1-\frac{\mathsf{n}^{mis}}{\mathsf{n}})$ of the term B should be replaced by a suitable estimate for $\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}\vert\mathsf{n}]$. Further, if the sample size is small (or the number of cells n_c is relative large) then one may encounter problems in Step 1 of the algorithm. Namely, there can be division by a zero in the estimation formulas for entries of the matrix $\mathbf{V}$. In such case one may either i) set the corresponding 'problematic' v_{il} entries to zero or ii) decrease the number of cells and redo clustering.

6.4.3 An example: the role of conditionalisation levels

In this example we demonstrate the role of conditionalisation levels. We study the joint $(Y, \boldsymbol{X})$ clustering method with simulated random imputation strategy. Further, we assume that imputed data $Y^{imp}|\boldsymbol{X}^{mis}$ is multimodal. This requirement excludes many imputation methods which would yield lower error quantities than the studied method. For simplicity, we investigate the aggregate level properties of $\hat{\mu}^{comp,TJ,R^s}$. More complicated error measures are utilized in the next chapter, where different imputation methods are compared.

It turns out that the bias of $\hat{\mu}^{comp,TJ,R^s}$ is neglible, and therefore we can focus to variance. The finite sample error of the first moment of $\hat{\mu}^{comp,TJ,R^s}$ contains various variance components, as seen from Approximation 6.4, but the roles of the components are difficult to interpret. To ease interpretation we utilize conditionalizations $\mathcal{Q}_1$, $\mathcal{Q}_2$, and $\mathcal{Q}_3$, as was done in Example 2 in Chapter 4. The studied quantities are

$$\begin{aligned} Err_1 &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,TJ,R^s}|\mathcal{Q}_1], \text{ where } \mathcal{Q}_1 = \{\mathsf{n}\} \\ Err_2 &= \mathbb{E}\Big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,TJ,R^s}|\mathcal{Q}_2]\Big], \text{ where } \mathcal{Q}_2 = \{\mathsf{n}, \mathbf{d}^{train}_{\mathsf{n}^{obs}}, \mathbf{w}_{\{i\}}\}, \text{ and} \\ Err_3 &= \mathbb{E}\Big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,TJ,R^s}|\mathcal{Q}_3]\Big], \text{ where } \mathcal{Q}_3 = \{\mathbf{d}^{train}_{\mathsf{n}^{obs}}, \mathbf{d}^{test}_{\mathsf{n}^{mis}}, \mathbf{w}_{\{i\}}\}. \end{aligned}$$

Data generator and the setup of the experiment

Our data generator is a four component mixture of gaussians (see Figure 6.4). The joint distribution of Y, X is

$$f_{Y,X}(y,x) = \sum_{i=1}^{4} \Pr(i) f_{Y,X|i}(y,x),$$

which parameters are given in Table 6.1. The distributions of Y^{obs}, X^{obs} and Y^{mis}, X^{mis} are:

$$f_{Y^{obs},X^{obs}}(y,x) = \sum_{i=1}^{4} \frac{1}{4} f_{Y,X|i}(y,x),$$

$$f_{Y^{mis},X^{mis}}(y,x) = \frac{1}{2} f_{Y,X|i=2}(y,x) + \frac{1}{2} f_{Y,X|i=3}(y,x),$$

where $f_{Y,X|i}, i = 1, \ldots, 4$ are bivariate gaussian distributions as described in Table 6.1. Figure 6.5 shows the marginal distributions of X^{obs} and X^{mis}. It should be noted that the missing data mechanism is NMAR. This becomes apparent as we notice that conditional expectations $\mathbb{E}[Y^{obs}|x]$ and $\mathbb{E}[Y^{mis}|x]$ are different, as shown in Figure 6.6.

We shall take a random sample of n observations, where a random number of N^{mis} observations are missing, using approximately $\mathsf{N}^{mis} \sim Bin(\mathsf{n}, 0.25)$ observations from $Y^{mis}, X^{mis} \sim f_{Y^{mis},X^{mis}}(y,x)$ and $\mathsf{N}^{obs} = \mathsf{n} - \mathsf{N}^{mis}$ observations from $Y^{obs}, X^{obs} \sim f_{Y^{obs},X^{obs}}(y,x)$. The sample sizes are $\mathsf{n} = \lfloor 25 * 1.6^k \rfloor$, where $k = 0, \ldots, 9$. A two-dimensional TS-SOM with four layers is fitted to data. This corresponds to modelling with 64 cells. Now we have everything that is required for the analysis.

Gaussian component i	Prior $\Pr(i)$	$\mathbb{E}[Y, X\|i]$	$\mathbb{V}\text{ar}[Y, X\|i]$	$\Pr(\text{"}missing\text{"}\|i)$
1	3/16	$(1, -1)$	$\text{diag}(0.15, 0.4)$	0
2	5/16	$(1, 0)$	$\text{diag}(0.15, 0.4)$	0.4
3	5/16	$(-1, 0)$	$\text{diag}(0.15, 0.4)$	0.4
4	3/16	$(-1, 1)$	$\text{diag}(0.15, 0.4)$	0

Table 6.1: Parametrization for data generator.

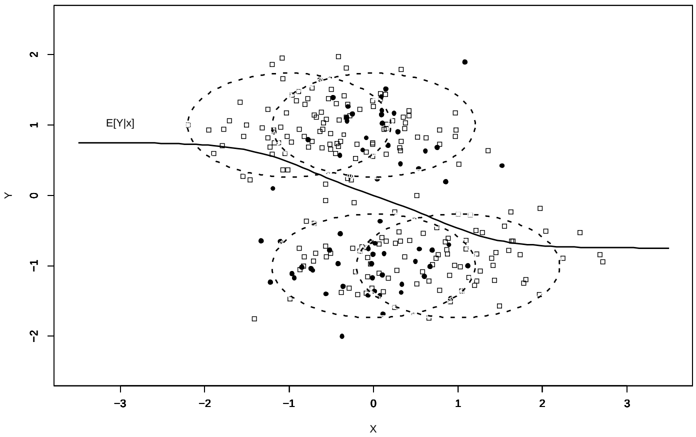


Figure 6.4: Random sample of size 262 from superpopulation and conditional expectation $\mathbb{E}[Y|x]$. The squares denote training data and the draws from missing population are denoted by black dots. 90% confidence ellipses (dashed lines) are drawn for each Gaussian component.

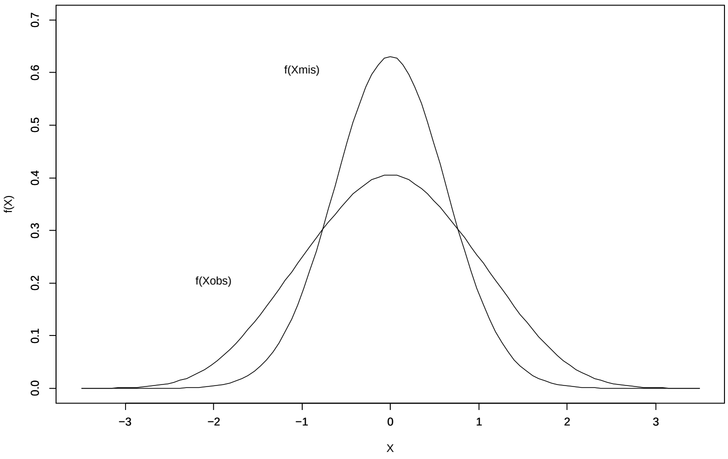


Figure 6.5: Distributions of X^{obs} and X^{mis}.

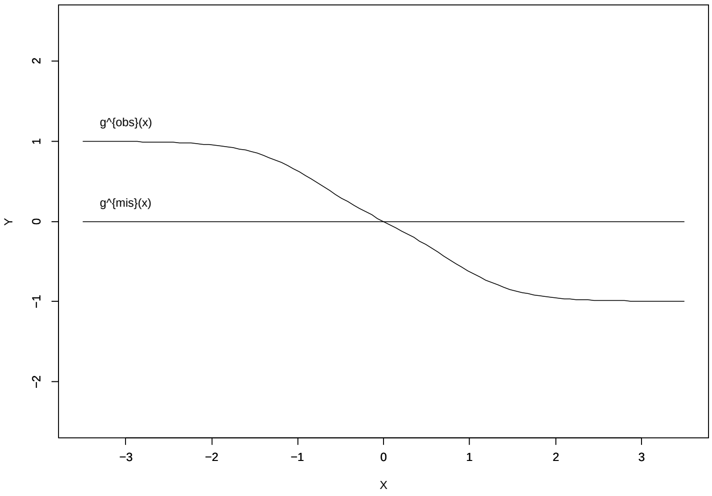


Figure 6.6: Conditional expectations $\mathbb{E}[Y^{obs}|x]$ and $\mathbb{E}[Y^{mis}|x]$ of observed and missing data.

Theoretical insight

Approximation 6.4 is applied to compute Err_1 from a single completed data set of size $\mathsf{n} = 1073$ using Algorithm 1 (page 122). We need to modify the algorithm for TS-SOM joint $(Y, \boldsymbol{X})$ clustering. Unsmoothed quantities μ_i^{obs} and τ_i^{obs} are replaced by smoothed ones. Further, estimation formulas for the entries of the covariance-variance matrix $\mathbf{V}$ are updated using Approximation 6.11. There are $\mathsf{n}^{mis} = 262$ imputed values in the data set. A suitably large completed data set is required in this example because many parameters need to be estimated (5 parameters for each of the 64 cells). In Phase 4 of the algorithm we assume that the number of missing data values is distributed as $\mathsf{N}^{mis} \sim Bin(\mathsf{n}, p^*)$. Therefore the variance of N^{mis} is

$$\mathbb{V}\text{ar}[\mathsf{N}^{mis}] = \mathsf{n}p^*(1-p^*) \approx \mathsf{n}\frac{262}{\mathsf{n}} * (1 - \frac{262}{\mathsf{n}}).$$

Even though we do know the value of p^* and the missing Y values (in this example) we do not use them to compute $\mathbb{V}\text{ar}[\mathsf{N}^{mis}]$ or classification success ratios, because in practice this information is not available. However, we try three values of classification success ratios: 75%, 50%, and 25%. This affects only analytical quantity Err_1(theor), and in practice we may approximate the real success ratio with some experimentations.

Error component Err_3(theor) is computed using the term C of Equation 6.13 (page 121), thus

$$Err_3(\text{theor}) \approx \frac{262}{\mathsf{n}} \sum_{i=1}^{\mathsf{n}_c} \Pr(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n})\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}],$$

where the probability terms and expectation of the second moments of imputed data are computed by applying Algorithm 6.4.2.

Results

The simulation results are shown in Figure 6.7 and Table 6.2 which contains also the theoretical values of Err_1 and Err_3. For comparison with Err_1, the variance of the mean of observed data (same as $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathsf{n}]$) is shown in Figure 6.8 and in Table 6.2. The results show how the conditionalization level affects the variance of the mean estimator. The B,M imputation method is not able to model the modes of marginal distribution of Y properly, as expected. The bias of $\hat{\mu}^{comp,TJ,R^s}$ is approximately zero, and is thus omitted. Variance at conditionalization $\mathcal{Q}_3$ is small. One can notice that the variance increase from $\mathcal{Q}_2$ to $\mathcal{Q}_1$ is the largest, as it was in Example 2 of Chapter 4. The increase in variance is due to additional variance sources: the number of missing data values, training data, and imputation model.

The following observations can be made from the theoretical results:

- Unfortunately all Err_1 values are larger than $\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}]$, which indicates that imputation leads to loss of efficiency. We note however that TJ,R^s is among the most random of all the strategies used in this thesis. We shall see in Section 9.4 that in practice it is better to use lower simulated noise than what is the estimated variance in the cells. Therefore a possible reason for large Err_1 values may be simulated noise level. To verify whether this is the reason we computed Err_1 for the TJ,M^s method using simulations and using an analytical approximation formula (for classification success ratio 0.75). The results were almost identical to the corresponding results for the TJ,R^s method for each sample size. This implies that the estimated variance in the cells (for TJ,R^s) is close to zero. Hence loss of efficiency might be reduced by decreasing the number of cells and simulated noise level.
- Theoretical formula for Err_1 overestimates the variance for classification success ratio 0.5, which is likely to mean that the assumed success ratio of only 0.5 is pessimistic. Classification success ratio 0.75 yields quite accurate variance results even for small sample size n.
- Theoretical formula for Err_1 shows that if the classification success ratio decreases then the variance of $\hat{\mu}^{comp,TJ}$ grows. Note that if the ratio decreases then incomplete observations spread to wider area.
- Theoretical formula for variance increase due to estimated noise, term Err_3(theor), is quite accurate from sample size 64 onwards.

Note that to verify that the theoretical results are not due to randomness we computed the analytical results using two other samples of size 1073. Similar results were observed.

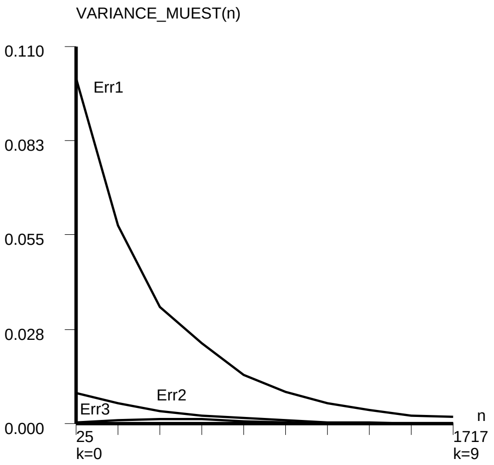


Figure 6.7: Simulation studies of measures $Err_i = \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,TJ,R^s}|\mathcal{Q}_i]]$ as a function of sample size n using TS-SOM.

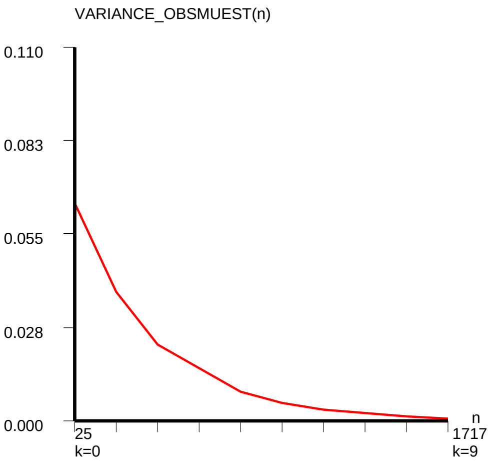


Figure 6.8: Simulated variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}]$ (same as $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathsf{n}]$) at $\mathcal{Q}_1$ (it is zero at other two conditionalisations).

n	25 (k=0)	40	64	102	163	262 (k=5)	419	671	1073	1717 (k=9)
Err_1(simulated)	0.1008	0.0578	0.0340	0.0235	0.0143	0.0092	0.0060	0.0040	0.0025	0.0019
Err_1(theor/0.75)	0.0978	0.0611	0.0382	0.0240	0.0150	0.0093	0.0058	0.0036	0.0023	0.0014
Err_1(theor/0.5)	0.1850	0.1156	0.0723	0.0453	0.0284	0.0177	0.0110	0.0069	0.0043	0.0027
Err_1(theor/0.25)	0.2708	0.1693	0.1058	0.0664	0.0415	0.0258	0.0162	0.0101	0.0063	0.0039
Err_2(simul)	0.0089	0.0060	0.0036	0.0025	0.0017	0.0009	0.0005	0.0003	0.0002	0.0001
Err_3(simul)	0.0005	0.0011	0.0013	0.0012	0.0006	0.0003	0.0002	0.0001	0.0001	0.0000
Err_3(theor)	0.0036	0.0022	0.0014	0.0009	0.0005	0.0003	0.0002	0.0001	0.0001	0.0000
$\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}]$(simul)	0.0640	0.0379	0.0226	0.0155	0.0087	0.0053	0.0034	0.0024	0.0013	0.0008

Table 6.2: Error components $Err_i = \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,TJ,R^s}|\mathcal{Q}_i]]$ as functions of the sample size $\mathsf{n} = \lfloor 25 * 1.6^k \rfloor$. Error components Err_1 and Err_3 are computed also theoretically. Theoretical value of Err_1 for varying classification success ratio is shown (values 0.75, 0.5, and 0.25 in parenthesis). Simulated variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}]$ is included for comparison with Err_1.

In order to interpret the causes of imputation errors Err_1, Err_2, and Err_3 we utilize the properties of data generator. Namely, marginal distributions of Y^{obs} and Y^{mis} have two modes. As a consequence, the mean estimator may be decomposed for a better interpretation as:

$$\hat{\mu}^{comp} = \frac{1}{\mathsf{n}}(N_1^{obs}\hat{\mu}_1^{obs} + N_2^{obs}\hat{\mu}_2^{obs} + N_1^{mis}\hat{\mu}_1^{imp} + N_2^{mis}\hat{\mu}_2^{imp}),$$

where the partitions for observed and imputed Y correspond to two modes. Observations of Y (observed or imputed) are classified to the upper mode cell if their values are greater or equal to 0.0, otherwise they are classified to the lower mode cell.

Bias and variance may be written as

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\mu}^{comp}] &= \mathbb{E}[\frac{1}{\mathsf{n}}(\mathsf{N}_1^{obs}\hat{\mu}_1^{obs} + \mathsf{N}_2^{obs}\hat{\mu}_2^{obs} + \mathsf{N}_1^{mis}\hat{\mu}_1^{imp} + \mathsf{N}_2^{mis}\hat{\mu}_2^{imp}] - \mu^* \\
&= \underbrace{\mathbb{E}[\frac{\mathsf{N}_1^{obs}}{\mathsf{n}}\hat{\mu}_1^{obs}] - (1-p^*)\frac{1}{2}\mu_1^{*obs}}_{\text{Bias due to observed Y in upper cell}} + \underbrace{\mathbb{E}[\frac{\mathsf{N}_2^{obs}}{\mathsf{n}}\hat{\mu}_2^{obs}] - (1-p^*)\frac{1}{2}\mu_2^{*obs}}_{\text{Bias due to observed Y in lower cell}} \\
&\quad + \underbrace{\mathbb{E}[\frac{\mathsf{N}_1^{mis}}{\mathsf{n}}\hat{\mu}_1^{imp}] - (p^*)\frac{1}{2}\mu_1^{*mis}}_{\text{Bias due to imputed Y in upper cell}} + \underbrace{\mathbb{E}[\frac{\mathsf{N}_2^{mis}}{\mathsf{n}}\hat{\mu}_2^{imp}] - (p^*)\frac{1}{2}\mu_2^{*mis}}_{\text{Bias due to imputed Y in lower cell}},
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,R^s}] &= \mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,M^s}] + \underbrace{\mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,R^s}] - \mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,M^s}]}_{\text{due to noise modelling, denoted by V}_{\text{noise}}} \\
&= \mathbb{V}\text{ar}[\frac{1}{\mathsf{n}}(\mathsf{N}_1^{obs}\hat{\mu}_1^{obs} + \mathsf{N}_2^{obs}\hat{\mu}_2^{obs} + \mathsf{N}_1^{mis}\hat{\mu}_1^{imp} + \mathsf{N}_2^{mis}\hat{\mu}_2^{imp})] + V_{noise} \\
&= \underbrace{\mathbb{V}\text{ar}[\frac{\mathsf{N}_1^{obs}}{\mathsf{n}}\hat{\mu}_1^{obs}]}_{\text{Variance due to observed Y in upper cell}} + \underbrace{\mathbb{V}\text{ar}[\frac{\mathsf{N}_2^{obs}}{\mathsf{n}}\hat{\mu}_2^{obs}]}_{\text{Variance due to observed Y in lower cell}} \\
&\quad + \underbrace{\mathbb{V}\text{ar}[\frac{\mathsf{N}_1^{mis}}{\mathsf{n}}\hat{\mu}_1^{imp}]}_{\text{Variance due to imputed Y (upper)}} + \underbrace{\mathbb{V}\text{ar}[\frac{\mathsf{N}_2^{mis}}{\mathsf{n}}\hat{\mu}_2^{imp}]}_{\text{Variance due to imputed Y (lower)}} \\
&\quad \underbrace{+ 2\mathbb{C}\text{ov}[\frac{\mathsf{N}_1^{obs}}{\mathsf{n}}\hat{\mu}_1^{obs}, \frac{\mathsf{N}_2^{obs}}{\mathsf{n}}\hat{\mu}_2^{obs}] + 2\mathbb{C}\text{ov}[\frac{\mathsf{N}_1^{mis}}{\mathsf{n}}\hat{\mu}_1^{imp}, \frac{\mathsf{N}_2^{mis}}{\mathsf{n}}\hat{\mu}_2^{imp}]}_{\text{Covariance terms}} \\
&\quad \underbrace{+2\mathbb{C}\text{ov}[\frac{\mathsf{N}_1^{obs}}{\mathsf{n}}\hat{\mu}_1^{obs} + \frac{\mathsf{N}_2^{obs}}{\mathsf{n}}\hat{\mu}_2^{obs}, \frac{\mathsf{N}_1^{mis}}{\mathsf{n}}\hat{\mu}_1^{imp} + \frac{\mathsf{N}_2^{mis}}{\mathsf{n}}\hat{\mu}_2^{imp}]}_{\text{Covariance terms cont.}} \\
&\quad +V_{noise},
\end{aligned}$$

where $\mu_1^{*obs} = 1, \mu_2^{*obs} = -1, \mu_1^{*mis} = 1, \mu_2^{*mis} = -1$, and $V_{noise} = \mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,R^s}] - \mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,M^s}]$.

According to our simulations, the joint $(Y, \boldsymbol{X})$ clustering cell method is able to model mixture distribution quite well. Namely, the last two terms in the above bias decomposition are roughly zero. As a consequence the bias of $\hat{\mu}^{comp,TJ,R^s}$ is approximately zero. Therefore we can concentrate on the analysis of imputation variance.

One is interested in how much of variability of $\hat{\mu}^{comp,TJ,R^s}$ is due to the components of variance decomposition. These are measured by computing the ratios between the terms in the previous equation for $\mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,R^s}|\mathsf{n}]$ and $\mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,M^s}|\mathsf{n}]$. We compute these quantities at conditionalization $\mathcal{Q}_1 = \{\mathsf{n}\}$ (for Err_1 measure). Table 6.3 contains ratios for sample size 1717. The values for other sample sizes are similar.

From Table 6.3 one can conclude that variances due to observed parts (first two rows) are equal, the same holds also for imputed parts (rows 4-5). The reason for this is due to the "symmetry" of data generator. Further, the cell model seems to be roughly symmetric in imputation Y values belonging to the upper and the lower mode cells. The covariance terms between the observed means and imputed means are positive. The distributions of $\hat{\mu}_1^{obs}$ and $\hat{\mu}_2^{obs}$ are equal up to the sign of expectation. Therefore the two quantities are negatively correlated. However, number of observations N_1^{obs} and N_2^{obs} are also negatively correlated, because they have sum to N^{obs}. As a consequence covariances between the products of the mean estimators and the observation count quantities are positively correlated. A similar reasoning holds also for the covariance term with $\hat{\mu}^{imp}$ quantities. The value of the last covariance term shows that the dependency between imputed and observed data contributes to approximately 20% of total variance. Finally, the ratio of variances of $\hat{\mu}^{comp,TJ,R^s}$ and $\hat{\mu}^{comp,TJ,M^s}$ shows that the random imputation strategy yields approximately 1% higher variance (at sample size 1717).

Quantity	Quantity/$\mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,M^s}\vert\mathsf{n}]$
$\mathbb{V}\text{ar}[\frac{\mathsf{N}_1^{obs}}{\mathsf{n}}\hat{\mu}_1^{obs}]$	0.090
$\mathbb{V}\text{ar}[\frac{\mathsf{N}_2^{obs}}{\mathsf{n}}\hat{\mu}_2^{obs}]$	0.090
$2\mathbb{C}\text{ov}[\frac{\mathsf{N}_1^{obs}}{\mathsf{n}}\hat{\mu}_1^{obs}, \frac{\mathsf{N}_2^{obs}}{\mathsf{n}}\hat{\mu}_2^{obs}]$	0.086
$\mathbb{V}\text{ar}[\frac{\mathsf{N}_1^{mis}}{\mathsf{n}}\hat{\mu}_1^{imp}]$	0.229
$\mathbb{V}\text{ar}[\frac{\mathsf{N}_2^{mis}}{\mathsf{n}}\hat{\mu}_2^{imp}]$	0.218
$2\mathbb{C}\text{ov}[\frac{\mathsf{N}_1^{mis}}{\mathsf{n}}\hat{\mu}_1^{imp}, \frac{\mathsf{N}_2^{mis}}{\mathsf{n}}\hat{\mu}_2^{imp}]$	0.090
$2\mathbb{C}\text{ov}[\frac{\mathsf{N}_1^{obs}}{\mathsf{n}}\hat{\mu}_1^{obs} + \frac{\mathsf{N}_2^{obs}}{\mathsf{n}}\hat{\mu}_2^{obs}, \frac{\mathsf{N}_1^{mis}}{\mathsf{n}}\hat{\mu}_1^{imp} + \frac{\mathsf{N}_2^{mis}}{\mathsf{n}}\hat{\mu}_2^{imp}]$	0.197
V_{noise}	0.012
$\mathbb{V}\text{ar}[\hat{\mu}^{comp,TJ,R^s}\vert\mathsf{n}]$	1.012

Table 6.3: Ratios at sample size 1717.

6.5 Unit level prediction errors

Cell methods are more flexible than our baseline methods and linear regression methods. For example, the smoothed version of our standard cell imputation T,M^s can be seen as a discrete approximation of kernel imputation. Unfortunately, flexibility leads to mathematical complexity of the analysis of the imputation properties of cell methods. Due to this difficulty the following results are given in a partially implicit form. As a consequence, the differences between the standard $\boldsymbol{X}$ and joint $(Y, \boldsymbol{X})$ clustering methods are not fully visible from the formulas.

As a new notation, a vector of the estimates of the means of the missing Y values within the cells is denoted by $\mu_{\{u\}}^{imp} = (\mu_1^{imp}, \ldots, \mu_{\mathsf{n}_c}^{imp})^T$.

Results are derived under the following assumptions

- Predictions based on crisp classifiers (maximum posterior and randomized) are approximated using a soft classifier: the estimator of Y^{mis} at $\mathbf{x}^{mis}$ given conditionalisation $\mathcal{Q}_3$ is

$$Y^{imp}|\mathbf{x}^{mis}, \mathcal{Q}_3 \;\approx\; \underbrace{\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\mu^{imp}_{i,\mathsf{n}^{obs}} +}_{=\overline{\mathrm{Y}}^{\mathrm{imp}}_{\mathbf{x}^{\mathrm{mis}}},\ \text{mean prediction}} \; \underbrace{\hat{\epsilon}(\mathbf{x}^{mis})}_{\text{imputation noise}} \quad , \tag{6.14}$$

where $u = 1, \ldots, \mathsf{n}_c$, and $\mathbb{E}[\hat{\epsilon}(\mathbf{x}^{mis})|\mathcal{Q}_3] = 0$ for any $\mathbf{x}^{mis}$. Note that in classification of incomplete observations only $\boldsymbol{X}$ part $\mathbf{w}_{\boldsymbol{X},\{u\}}$ of centroids $\mathbf{w}_{\{u\}}$ is used.

- Posterior probabilities $g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}}), i = 1, \ldots, \mathsf{n}_c$ are continuous and have first derivative with respect to $\mathbf{x}^{mis}$ and $\mathbf{w}_{\boldsymbol{X},\{u\}}$.
- Covariance between $\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}$ and $\hat{\boldsymbol{\mu}}^{imp}$ is neglible.

In addition, the first order Taylor approximation is used.

Next, variances of imputation noise at point $\boldsymbol{X}^{mis} = \mathbf{x}^{mis}$ are defined given $\mathcal{Q}_3$. The results for mean squared error at point $\mathbf{x}^{mis}$ are derived from $\mathcal{Q}_3$ conditionalisation by integrating over distribution of $\mathbf{D}^{train}_{\mathsf{n}^{obs}}, \hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}, \mathbf{D}^{mis}_{Y,\mathsf{n}^{mis}}|\mathbf{d}^{mis}_{\boldsymbol{X},\mathsf{n}^{mis}}$.

The difference between the imputations done using a crisp and a randomized classifier (which are both approximated using a soft classifier) is that the distribution of $\hat{\epsilon}(\mathbf{x}^{mis})$ is unimodal for the crisp classifier, whereas it is multimodal for the randomized classifier. As a consequence the variance of $\hat{\epsilon}(\mathbf{x}^{mis})$ consists of between (and possibly) within components for TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods. For other methods variance is the weighted sum of within cell variance estimators. The variances of $\hat{\epsilon}(\mathbf{x}^{mis})$ are

$$\mathbb{V}\mathrm{ar}[\hat{\epsilon}(\mathbf{x}^{mis})|\mathcal{Q}_3] = \begin{cases} 0 & : \mathrm{C/T/CJ(S = M)/} \\ & \;\;\mathrm{T(S = M^s)}, \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{obs} & : \mathrm{C/T/CJ(S = R)} \quad , \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{T,R^s} & : \mathrm{T(S = R^s)} \quad , \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{obs}(1 - \frac{1}{n_i^{obs}}) & : \mathrm{C/T/CJ(S = D)} \quad , \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})(\mu_i^{imp} - \overline{Y}^{imp}_{\mathbf{x}^{mis}})^2 & : \mathrm{TJ(S = M/M^s)} \quad , \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})(\mu_i^{imp} - \overline{Y}^{imp}_{\mathbf{x}^{mis}})^2 & : \mathrm{TJ(S = R)} \quad , \\ + \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{obs} & \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})(\mu_i^{imp} - \overline{Y}^{imp}_{\mathbf{x}^{mis}})^2 & : \mathrm{TJ(S = R^s)} \\ + \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{TJ,R^s} & \quad , \end{cases}$$

where $g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})$ is an estimate of the posterior probability of the i:th cell at $\boldsymbol{X}^{mis} = \mathbf{x}^{mis}$.

Note that the results of this section for joint $(Y, \boldsymbol{X})$ clustering with TS-SOM and mean strategy have a non-zero noise variance, as was shown in the above variance formulas. This is due to the fact that the impact of random selection of the best cell is formalized as multimodal imputation noise distribution.

Approximation for $Y^{imp}|\mathbf{x}^{mis}, \mathcal{Q}_3$, which was given in Equation (6.14) holds best for TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods. Further, a soft classifier with a donor strategy would yield a multimodal imputation noise distribution. To simplify things we analyse donor strategy as a random strategy (with soft classifier), but with smaller within cells variance estimates. The reason for selecting this approach is that it is closer to our practical implementation.

Mean squared error results at a given point $\mathbf{x}^{mis}$ are summarized in Approximation 6.13:

Approximation 6.13 *Approximation to mean squared error at a given point* $\boldsymbol{X}^{mis} = \mathbf{x}^{mis}$*.*

Mean squared error $\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, n^{mis}, n)$ *can be approximated as*

$$
\begin{aligned}
&\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, n^{mis}, n)\\
\approx\ &\underbrace{\left(\sum_{i=1}^{n_c} g_i(\mathbf{x}^{mis}|\mathcal{Q})\mu^{*imp}_{i,n^{obs}} - \mathbb{E}[Y^{mis}|\mathbf{x}^{mis}]\right)^2}_{\text{bias}}\\
&+\underbrace{\sum_{i=1}^{n_c} \mu^{*imp}_{i,n^{obs}} g_i'(\mathbf{x}^{mis}|\mathcal{Q})^T \mathbb{V}\mathrm{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})|\mathbf{d}_{\boldsymbol{X}}^{mis}, n^{mis}, n] \sum_{i=1}^{n_c} \mu^{*imp}_{i,n^{obs}} g_i'(\mathbf{x}^{mis}|\mathcal{Q})}_{\text{sampling variance due to estimation of classifier parameters}}\\
&+\underbrace{\big(g_1(\mathbf{x}^{mis}|\mathcal{Q})\dots g_{n_c}(\mathbf{x}^{mis}|\mathcal{Q})\big)\mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}_{\{u\}}|\mathbf{d}_{\boldsymbol{X}}^{mis}, n^{mis}, n]\big(g_1(\mathbf{x}^{mis}|\mathcal{Q})\dots g_{n_c}(\mathbf{x}^{mis}|\mathcal{Q})\big)^T}_{\text{sampling variance due to estimation of imputation model parameters}}\\
&+\underbrace{\sum_{i=1}^{n_c} g_i(\mathbf{x}^{mis}|\mathcal{Q})\mathbb{E}[\hat{\tau}_i^{imp}(\mathbf{x}^{mis})|n^{mis}, n]}_{\text{imputation variance}} + \underbrace{v^{*mis}(\mathbf{x}^{mis})}_{\text{target variance}} \qquad\qquad u = 1, \dots, n_c,
\end{aligned}
$$

the expected prediction in the i:th cell is

$$
\mu^{*imp}_{i,n^{obs}} = \begin{cases} \mu^{*obs}_{i,n^{obs}} \approx \mu^{*obs}_i + O\big((n^{obs})^{-1}\big) & : \mathrm{C/T/CJ/TJ(S = M/R/D)},\\[2ex] \mu^{*s}_{i,n^{obs}} \approx \frac{\sum_l h_{i,l}\mathbb{E}[N_l|n^{mis}, n]\mu^{*obs}_l}{\sum_l h_{i,l}\mathbb{E}[N_l|n^{mis}, n]} + O\big((n^{obs})^{-1}\big) & : \mathrm{T/TJ(S = M^s/R^s)}, \end{cases}
$$

$\mathcal{Q} = \mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbf{d}_{\boldsymbol{X}}^{mis}, n^{mis}, n]$*, quantity* $g_i'(\cdot)$ *is derivative of* $g_i(\cdot)$ *with respect to* $vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})$ *which is evaluated at* $\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbf{d}_{\boldsymbol{X}}^{mis}, n^{mis}, n]$*, and term* $\mathbb{E}[\hat{\tau}_i^{imp}(\mathbf{x}^{mis})|n^{mis}, n]$

depends on imputation strategy S:

$$\mathbb{E}[\hat{\tau}_i^{imp}(\mathbf{x}^{mis})|\mathsf{n}^{mis},\mathsf{n}] = \begin{cases} 0 & : \mathrm{C/T/CJ(S=M)/,} \\ & \ \ \mathrm{T(S=M^s),} \\ \approx \tau_i^{*obs} & : \mathrm{C/T/CJ(S=R),} \\ \approx \tau_i^{*obs}(1-\frac{1}{\mathbb{E}[\mathsf{N}_i^{obs}|\mathsf{n}^{mis},\mathsf{n}]}) & : \mathrm{C/T/CJ(S=D),} \\ & \\ \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathsf{n}^{mis},\mathsf{n}] \approx \frac{\sum_l h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}^{mis},\mathsf{n}]\mathbb{E}[\hat{\tau}_l^s|\mathsf{n}^{mis},\mathsf{n}]}{\sum_l h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}^{mis},\mathsf{n}]} & : \mathrm{T(S=R^s),} \\ \mathbb{E}[(\hat{\mu}_{i,\mathsf{n}^{obs}}^{imp} - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp})^2|\mathsf{n}^{mis},\mathsf{n}] & : \mathrm{TJ(S=M/M^s),} \\ \mathbb{E}[\hat{\tau}_i^{obs}|\mathsf{n}^{mis},\mathsf{n}] + \mathbb{E}[(\hat{\mu}_i^{obs} - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp})^2|\mathsf{n}^{mis},\mathsf{n}] & : \mathrm{TJ(S=R),} \\ \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathsf{n}^{mis},\mathsf{n}] + \mathbb{E}[(\hat{\mu}_i^{s} - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp})^2|\mathsf{n}^{mis},\mathsf{n}] & : \mathrm{TJ(S=R^s),} \end{cases}$$

in which quantities $\hat{\mu}_{i,\mathsf{n}^{obs}}^{imp} - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp}$ *are location shifts from mean prediction* $\hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp}$ *to the modes of multimodal imputation noise distribution.*

A justification of Approximation 6.13 is given in Appendix A6.2. In Approximation 6.13 one sees the decomposition of the mean squared error into squared bias, sampling variance, imputation variance, and target variance.

Approximative expected prediction, which is a part of the prediction bias, is based on the first order Taylor approximation. The first term of the approximation equals the approximate posterior probability of classifying to the i:th cell multiplied by the expectation of the corresponding mean estimator. Squared bias is expected to be low for large sample size and sufficient number of cells under MCAR. Large-sample performance is also good under MAR if the support of $\boldsymbol{X}^{obs}$ covers the support of $\boldsymbol{X}^{mis}$. As in kernel regression it is likely that one needs to decrease smoothing as the sample size grows to reduce squared bias. Under NMAR there is an irreducible squared bias if $\mathbb{E}[Y^{obs}|\mathbf{x}^{mis}]$ and $\mathbb{E}[Y^{mis}|\mathbf{x}^{mis}]$ are different.

Sampling variance is derived using the first order Taylor approximation. The variance consists of two parts: variance due to classifier parameter estimation and variance due to estimation of imputation model parameter estimation. Both parts are of quadratic form. The variance due to the estimation of classifier parameters is smaller the closer all expected predictions are to zero point or the larger the sample size is. The variance due to the estimation of imputation model parameters gets smaller as the sample size is increased.

The increase in the mean squared error due to noise modelling is measured by the imputation variance term. Mean strategy yields a lower variance than random strategy. Further, donor strategy yields a lower variance than random strategy. Note that the noise variance is non-zero for joint $(Y,\boldsymbol{X})$ clustering with TS-SOM and mean strategy because of random selection of node (however noise variance within any cell is zero). One can notice that for joint $(Y,\boldsymbol{X})$ clustering with TS-SOM and mean strategy there is between and within variance components. This is because noise distribution is multimodal.

Variance of target, an irreducible term, cannot be affected by the imputation method. If the signal to noise ratio is low, or in other words variability of the conditional mean is low relative to target variance, then the irreducible term may have a significant impact on the mean squared error. Note also that an increase of sample size does not affect the irreducible error.

Expectation of $\hat{mse}(Y^{comp})$ with n observations is computed by integrating mse at $\mathbf{x}^{mis}$ over distribution of the number of missing values N^{mis} and $\mathbf{D}^{mis}_{\mathbf{X}}$. The result is given in the form of Approximation 6.14:

Approximation 6.14 *The expected mean squared prediction error for* n *observations.*

Expectation of $\hat{mse}(Y^{comp})$ *with* n *observations can be approximated as*

$$
\begin{aligned}
&\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] \\
\approx\ & \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\mathbf{D}^{mis}_{\boldsymbol{X}}|\mathsf{n}}\left[\sum_{i=1}^{\mathsf{n}_c} g_i(\boldsymbol{X}^{mis}|\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathsf{n}^{mis},\mathsf{n}])\mu^{*imp}_{i,\mathsf{N}^{obs}}\right]}_{\text{variability of conditional mean estimate}} + (\underbrace{\mu^{*imp}_{\mathsf{n}} - \mu^{*mis}}_{\text{global bias}})^2 \\
& + \underbrace{\mathbb{V}\mathrm{ar}\big[\mathbb{E}[Y^{mis}|\boldsymbol{X}^{mis}]\big]}_{\text{variability of true model}} \\
& + \underbrace{2\mathbb{E}_{\mathsf{N}^{mis},\mathbf{D}^{mis}_{\boldsymbol{X}}|\mathsf{n}}\Bigg[\Bigg(\sum_{i=1}^{\mathsf{n}_c} g_i(\boldsymbol{X}^{mis}|\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathsf{n}^{mis},\mathsf{n}])\mu^{*imp}_{i,\mathsf{N}^{obs}} - \mu^{*imp}_{\mathsf{n}}\Bigg)}_{\text{cross term}} \\
& + \underbrace{\Bigg(\mu^{*imp}_{\mathsf{n}} - \mathbb{E}[Y^{mis}|\boldsymbol{X}^{mis}]\Bigg)\Bigg]}_{\text{cross term (cont.)}} \\
& + \underbrace{\sum_{i=1}^{\mathsf{n}_c}\mu_i^{*imp}g_i'(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})^T\mathbb{V}\mathrm{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})|\mathbb{E}[\mathbf{D}^{mis}_{\boldsymbol{X}},\mathsf{N}^{mis}|\mathsf{n}]]\sum_{i=1}^{\mathsf{n}_c}\mu_i^{*imp}g_i'(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})}_{\text{expected sampling variance due to estimation of classifier parameters}} \\
& + \underbrace{\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})^T\mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}_{\{u\}}|\mathbb{E}[\mathbf{D}^{mis}_{\boldsymbol{X}},\mathsf{N}^{mis}|\mathsf{n}]]\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})}_{\text{expected sampling variance due to estimation of imputation model parameters}} \\
& + \underbrace{\sum_{i=1}^{\mathsf{n}_c} g_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})\mathbb{E}[\hat{\tau}_i^{imp}(\boldsymbol{X}^{mis})|\mathsf{n}]}_{\text{expected imputation variance}} + \underbrace{v^{*mis}}_{\text{expected target variance}} \quad u = 1,\ldots,\mathsf{n}_c,
\end{aligned}
$$

where

$$
\mu^{*imp}_{\mathsf{n}} = \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}] \approx \begin{cases} \sum_{i=1}^{\mathsf{n}_c} g_i(\overline{X}^{*mis}|\mathcal{Q})\mu_i^{*obs} & : \mathrm{C/CJ/T/TJ(S = M/R/D)}, \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\overline{X}^{*mis}|\mathcal{Q})\mu_i^{*s} & : \mathrm{T/TJ(S = M^s/R^s)}, \end{cases}
$$

$\mathcal{Q} = \mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbb{E}[\mathbf{D}^{mis}_{\boldsymbol{X}}],\mathbb{E}[\mathsf{N}^{mis}],\mathsf{n}]$, $\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}) = [g_1(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}),\ldots,g_{\mathsf{n}_c}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})]^T$ *and constant* $\mathbb{E}[\hat{\tau}_i^{imp}(\boldsymbol{X}^{mis})|\mathsf{n}]$ *depends on the imputation method and strategy* S *as* *follows:*

$$\mathbb{E}[\hat{\tau}_i^{imp}(\boldsymbol{X}^{mis})|\mathsf{n}] = \begin{cases} 0 & : \mathrm{C/T/CJ(S = M)/} \quad , \\ & \;\; \mathrm{T(S = M^s)} \\ \approx \tau_i^{*obs} & : \mathrm{C/T/CJ(S = R)}, \\ \approx \tau_i^{*obs}(1 - \frac{1}{\mathbb{E}[\mathsf{N}_i^{obs}|\mathsf{n}]}) & : \mathrm{C/T/CJ(S = D)}, \\ & \\ \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathsf{n}] & : \mathrm{T(S = R^s)}, \\ \mathbb{E}[(\hat{\mu}_i^{imp} - \hat{\overline{Y}}{}^{imp}_{\overline{\boldsymbol{X}}^{*mis}})^2|\mathsf{n}] & : \mathrm{TJ(S = M/M^s)} \quad , \\ \mathbb{E}[\hat{\tau}_i^{obs}|\mathsf{n}] + \mathbb{E}[(\hat{\mu}_i^{obs} - \hat{\overline{Y}}{}^{imp}_{\overline{\boldsymbol{X}}^{*mis}})^2|\mathsf{n}] & : \mathrm{TJ(S = R)}, \\ \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathsf{n}] + \mathbb{E}[(\hat{\mu}_i^{s} - \hat{\overline{Y}}{}^{imp}_{\overline{\boldsymbol{X}}^{*mis}})^2|\mathsf{n}] & : \mathrm{TJ(S = R^s)}, \end{cases}$$

Expectation of $\widehat{mse}(Y^{comp})$ with n observations is quite complicated. In Approximation 6.14 we have decomposed it into more interpretable terms, similarly as was done in Approximation 6.13. Mean squared error consists of expected squared bias, expected sampling variance, imputation variance, and target variance. The expected squared bias is decomposed into variability of conditional mean estimate, squared global bias, variability of true model (which cannot be affected by the imputation method), and cross term.

Note that the variance of imputed values equals to variability of conditional mean estimate plus expected imputation variance. As a consequence it is possible (at least under MAR with suitable assumptions) to recover the first two moments of Y in multiple ways. One may use a stiff model with a large imputation variance, or a flexible model with a lower variance. With cell methods the stiff model equals to a low number of cells or high amount of smoothing. The flexible model equals to a large number of cells with no or relative low amount of smoothing.

Justifications to the above approximations and consequences can be found in Appendix A6. Next an example which demonstrates the differences between cell methods is given.

6.6 Example: comparison of cell methods

The purpose of this example is to demonstrate the differences between the proposed cell imputation methods. All six cell methods that were introduced in Chapter 6.1 are evaluated under the mean imputation strategy. The reason why the mean strategy is used is that clustering properties and imputation performances are easier to show when there is no added noise in the imputation results.

In our example we use a simple generator of Y, X data. The data generator is

$$Y \;\; = \;\; X + \epsilon, \;\; X \sim N(0,1), \;\; \epsilon \sim N(0,1),$$

as shown in Figure 6.9. It is easy to see that the correlation between Y and X is $1/\sqrt{2} \approx 0.71$ (strong linear dependency). The missing-data mechanism is

MCAR with 25% expected missingness. We shall take an iid sample of size n. As a consequence the number of missing observations is $\mathsf{N}^{mis} \sim Bin(\mathsf{n}, 0.25)$, and $\mathsf{N}^{obs} = \mathsf{n} - \mathsf{N}^{mis}$. In the experiments the sample size n is varied as: 32, 44, 62, 87 and 122.

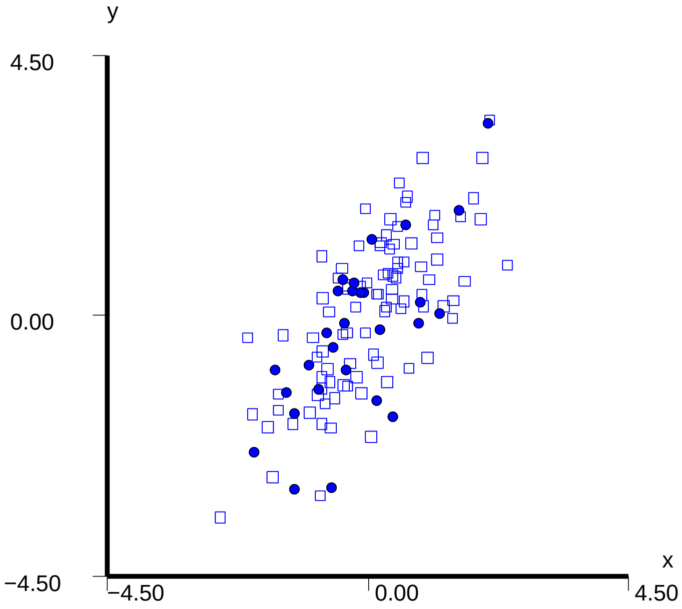


Figure 6.9: Random sample of size 122 from population. The squares are training data and the dots denote draws from the missing population.

Characteristics of clustering can be measured using decomposition of total-scatter matrix for training data. The total-scatter matrix is

$$\hat{\mathbf{S}}^{total} = \sum_{i=1}^{\mathsf{N}^{obs}} (\boldsymbol{Z}_i - \overline{Z})(\boldsymbol{Z}_i - \overline{Z})^T,$$

where $\boldsymbol{Z}_i = (Y_i^{obs}, X_i^{obs})^T$ and $\overline{Z} = (\hat{\mu}^{obs}, \hat{\overline{X}})^T$. Elements of $\hat{\mathbf{S}}^{total}$, which are used in later quantities, are denoted as

$$\hat{\mathbf{S}}^{total} = \begin{bmatrix} \hat{S}^{\text{total}}_{YY} & \hat{S}^{\text{total}}_{YX} \\ \hat{S}^{\text{total}}_{XY} & \hat{S}^{\text{total}}_{XX}. \end{bmatrix}. \tag{6.15}$$

Note that $\hat{\tau}^{obs} = \frac{1}{\mathsf{N}^{obs}-1}\hat{S}^{\text{total}}_{YY}$ and $\mathbb{E}[\frac{1}{\mathsf{n}^{obs}}\hat{\mathbf{S}}^{total}|\mathsf{n}^{obs}] \approx \mathbb{V}\text{ar}[(Y^{obs}, X^{obs})^T]$.

Decomposition for total-scatter matrix is

$$\hat{\mathbf{S}}^{\text{total}} = \underbrace{\hat{\mathbf{S}}^{\text{between}}}_{\text{between}-\text{cells}-\text{scatter matrix}} + \underbrace{\hat{\mathbf{S}}^{\text{within}}}_{\text{within}-\text{cells}-\text{scatter matrix}},$$

where $\hat{\mathbf{S}}^{between} = \sum_{i=1}^{\mathsf{n}_c}(\overline{Z}_j - \overline{Z})(\overline{Z}_j - \overline{Z})^T$, in which $\overline{Z}_j$ is the j:th cell position in the Y, X space. The between-cells scatter matrix characterizes statistics of the cells, and the within-cells scatter matrix can be considered as a residual structure.

Because the scatter matrices are symmetric, it is sufficient to compute three quantities:

$$T_1 = \mathbb{E}\Big[\frac{\hat{S}^{\text{between}}_{YY}}{\hat{S}^{\text{total}}_{YY}}|\mathsf{n}\Big],\ T_2 = \mathbb{E}\Big[\frac{\hat{S}^{\text{between}}_{XX}}{\hat{S}^{\text{total}}_{XX}}|\mathsf{n}\Big],\ T_3 = \mathbb{E}\Big[\frac{\hat{S}^{\text{between}}_{YX}}{\hat{S}^{\text{total}}_{YX}}|\mathsf{n}\Big].$$

Quantity T_1 measures how well the variance of Y^{obs} is preserved. Preservation of variance of X^{obs} is measured by T_2, and quantity T_3 can be used to evaluate how well the covariance between Y^{obs} and X^{obs} is preserved. If quantities $T_1 - T_3$ are all close to one then the clustering method is able to explain the variability of Y^{obs}, X^{obs}, and the covariance between Y^{obs} and X^{obs}. Note that $1 - T_1 = \mathbb{E}[\frac{\hat{S}^{\text{within}}_{YY}}{\hat{S}^{\text{total}}_{YY}}|\mathsf{n}]$ (and similarly for $1 - T_2$ and $1 - T_3$).

Imputation performance is measured using imputed data values. Mean squared error, and its decomposition, of mean estimator is computed as

$$\text{MSE}[\hat{\mu}^{imp}|\mathsf{n}] = \underbrace{(\mathbb{E}[\hat{\mu}^{imp} - \mu^{*mis}|\mathsf{n}])^2}_{\text{bias}} + \underbrace{\mathbb{V}\text{ar}[\hat{\mu}^{imp}|\mathsf{n}]}_{\text{variance}},$$

Further, the mean squared error of the second moment estimator is

$$\text{MSE}[\hat{\tau}^{imp}|\mathsf{n}] = \underbrace{(\mathbb{E}[\hat{\tau}^{imp} - \tau^{*mis}|\mathsf{n}])^2}_{\text{bias}} + \underbrace{\mathbb{V}\text{ar}[\hat{\tau}^{imp}|\mathsf{n}]}_{\text{variance}},$$

Note that in this example $\mu^{*mis} = 0$ and $\tau^{*mis} = 2$. The purpose of decomposition is to demonstrate the differences between biases and variances. This is useful when showing differences between unsmoothed and smoothed imputation methods.

In our experiments the number of cells is $\mathsf{n}_c = 16$ for all methods. One dimensional TS-SOM is used, implying lattice topology.

6.6.1 Summary of results

The results of clustering characteristics for all the methods are summarized in Tables 6.4-6.6. The variances of mean estimators are also shown in Figure 6.10. Squared biases are neglible, and thus the relative efficiencies of estimators can be given as

$$\text{Eff}[\hat{\mu}^{imp}|\mathsf{n}] = \frac{\mathbb{V}\text{ar}[\hat{\mu}^{imp,CJ,M}|\mathsf{n}]}{\mathbb{V}\text{ar}[\hat{\mu}^{imp}|\mathsf{n}]},$$

where $\hat{\mu}^{imp,CJ,M}$ is used as a reference estimator because it performs the worst for all sample sizes. Interpretation of the relative efficiency is simple. As an example, if efficiency is two then estimator $\hat{\mu}^{imp,CJ,M}$ requires twice as many observations to reach the same performance as estimator $\hat{\mu}^{imp}$. Efficiency results are shown in Table 6.7, and the results for the second moment estimators are shown in Figures 6.11-6.13.

n Method	32	44	62	87	122
T,M	0.79	0.72	0.66	0.61	0.58
T,M^s	0.55	0.53	0.51	0.49	0.48
C,M	0.84	0.73	0.66	0.61	0.57
TJ,M	0.98	0.98	0.97	0.97	0.96
TJ,M^s	0.89	0.88	0.87	0.87	0.86
CJ,M	0.99	0.98	0.97	0.96	0.96

Table 6.4: Quantity T_1 (preservation of $\mathbb{V}\text{ar}[Y^{obs}]$) as a function of the sample size.

n Method	32	44	62	87	122
T,M	0.96	0.95	0.94	0.93	0.93
T,M^s	0.96	0.95	0.94	0.93	0.93
C,M	1.00	0.99	0.99	0.99	0.99
TJ,M	0.96	0.94	0.91	0.90	0.88
TJ,M^s	0.85	0.85	0.84	0.84	0.83
CJ,M	0.98	0.96	0.94	0.93	0.92

Table 6.5: Quantity T_2 (preservation of $\mathbb{V}\text{ar}[X^{obs}]$) as a function of the sample size.

n Method	32	44	62	87	122
T,M	0.98	0.97	0.96	0.96	0.96
T,M^s	0.94	0.93	0.93	0.92	0.92
C,M	1.00	0.99	0.99	0.99	0.99
TJ,M	1.06	1.07	1.08	1.08	1.09
TJ,M^s	1.09	1.09	1.10	1.10	1.10
CJ,M	1.00	1.00	1.00	1.00	1.00

Table 6.6: Quantity T_3 (preservation of $\mathbb{C}\text{ov}[Y^{obs}, X^{obs}]$) as a function of the sample size.

n Method	32	44	62	87	122
T,M	1.28	1.37	1.55	1.91	2.28
T,M^s	1.74	1.92	2.11	2.42	2.72
C,M	1.19	1.39	1.63	2.00	2.33
TJ,M	1.14	1.20	1.25	1.41	1.52
TJ,M^s	1.20	1.27	1.35	1.51	1.65
CJ,M	1.00	1.00	1.00	1.00	1.00

Table 6.7: Efficiencies as functions of the sample size. Method CJ,M is used as reference.

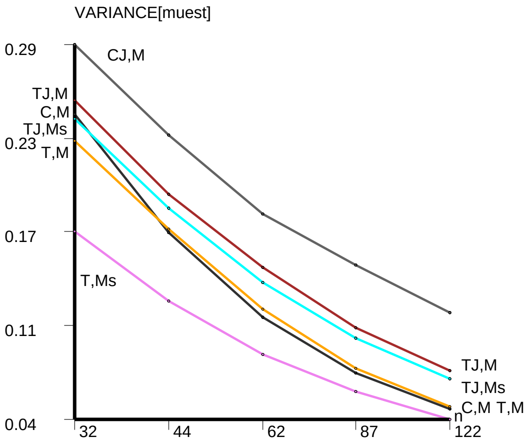


Figure 6.10: Variance $\mathbb{V}\text{ar}[\hat{\mu}^{imp}|\mathsf{n}]$ as a function of sample size.

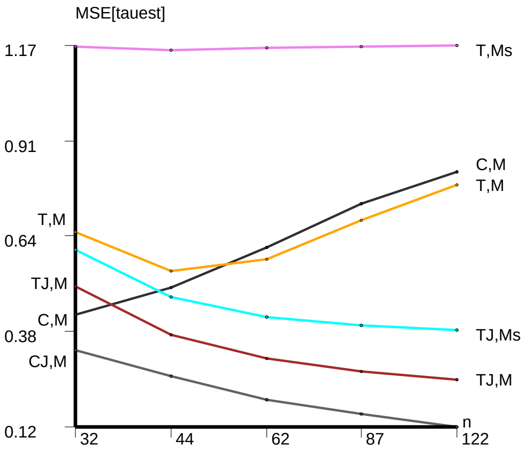


Figure 6.11: Mean squared error of $\hat{\tau}^{imp}$ as a function of sample size.

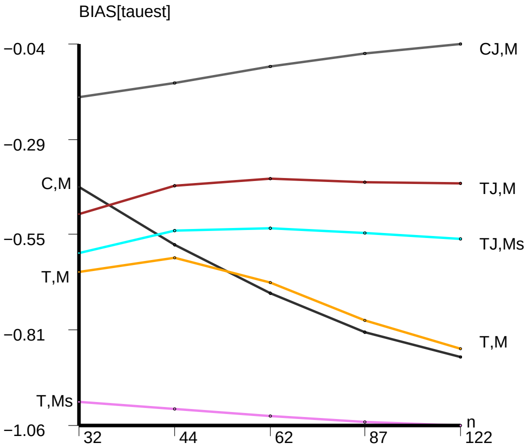


Figure 6.12: Bias of $\hat{\tau}^{imp}$ as a function of sample size.

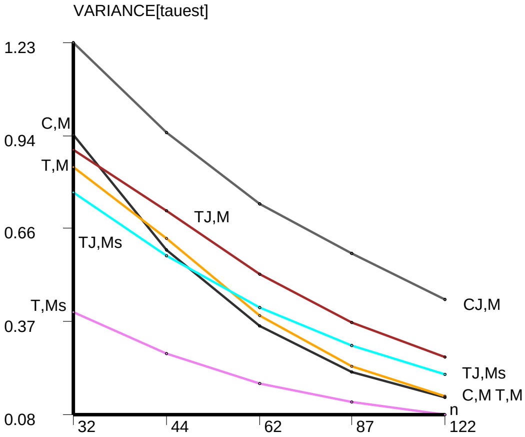


Figure 6.13: Variance $\mathbb{V}\text{ar}[\hat{\tau}^{imp}|\mathsf{n}]$ as a function of sample size.

From the results one can notice that joint cell methods CJ and TJ without smoothing preserve the variability of Y^{obs} best. The variability of X^{obs} is preserved best by C, T, and CJ. The covariance between Y^{obs}, X^{obs} is best preserved by the K-Means cell methods C and CJ. Smoothing used by TS-SOM leads to underestimation of the second moment, and it affects the covariance structure as well. However, smoothing is able to improve the mean estimator as discussed next.

Strength borrowing, which is due to neighborhood smoothing in the TS-SOM methods T,M^s and TJ,M^s, brings the cells closer to each other. This leads to underestimation of variability. On the other hand, smoothing reduces the variance of the first moment estimator without introducing bias. Variance reduction is visible in Figure 6.10 (compare T,M/T,M^s or TJ,M/TJ,M^s pairs). As a consequence smoothed mean estimators $\hat{\mu}^{imp,T,M^s}$ and $\hat{\mu}^{imp,TJ,M^s}$ have the lowest mean squared errors. Further, efficiencies for them are the best, especially for the smoothed covariate method.

A difference between X-clustering and joint (Y,X) clustering methods is that the latter ones are able to model the variability of Y^{obs} considerably better. This is expected, because covariate cell methods do not utilize the target variable.

Dependency between the characteristics of clustering and imputation performance is visible in the bias of the second moment estimators. For all methods, except the joint TS-SOM methods, the bias of the second moment is approximately

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\tau}^{imp}|\mathsf{n}] &\approx (T_1 - 1)\mathbb{V}\text{ar}[Y^{mis}] \\ &= -\mathbb{E}\Big[\frac{\hat{S}_{YY}^{\text{within}}}{\hat{S}_{YY}^{\text{total}}}|\mathsf{n}\Big]\mathbb{V}\text{ar}[Y^{mis}],\end{aligned}$$

where $\mathbb{V}\text{ar}[Y^{mis}] = \tau^{*mis} = 2$. Therefore the underestimation of the second moment depends on how large the residual variance of Y (related to the component $\hat{S}_{YY}^{\text{within}}$ in the within-cells scatter matrix) is. For joint TS-SOM methods dependency is more complicated because of probabilistic classification of incomplete observations.

Further, if the missingness were other than MCAR then the dependency would be even more complicated.

We note that all the results are sensitive to the number of cells, amount of smoothing used, and the number of observations. The second point is that TS-SOM cell methods may need different amounts of smoothing for mean estimators and variance estimators to better preserve the first two moments simultaneously. A change in the imputation strategy to random or donor is likely to yield a less biased second moment estimator.

Finally, it is interesting to see what the decision boundaries of the cells for the evaluated methods are. Figures 6.14-6.19 depict the decision boundaries for all the methods, which were trained with a sample of size 122. From Figure 6.17 one can notice the 1-D continuum that is constructed by the smoothed joint (Y,X) clustering TS-SOM method. The impact of smoothing on cell positions is visible in Figures 6.15 and 6.17. Smoothed cells deviate less from diagonal line $y = x$. Linear trend, which is a data generator, is best preserved by the smoothed TS-SOM cell methods. The joint K-Means method does not form any kind of continuum, the cells are spread along the data space.

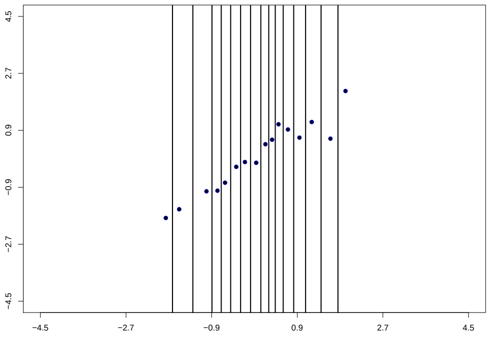


Figure 6.14: Decision boundaries for X-clustering method T,M (TS-SOM), dots denote cells and the sample size is 122.

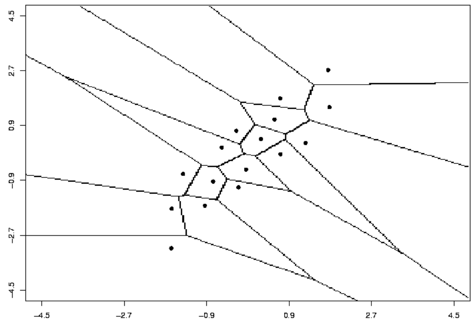


Figure 6.15: Decision boundaries for joint (Y,X) clustering method TJ,M (TS-SOM).

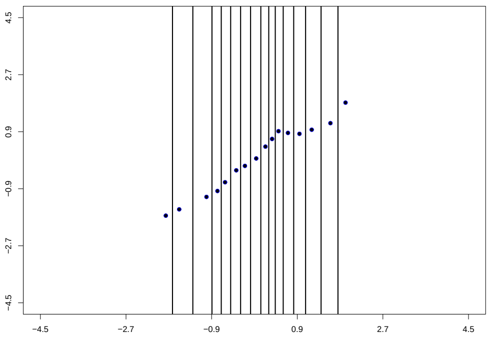


Figure 6.16: Decision boundaries for X-clustering method T,M^s (TS-SOM).

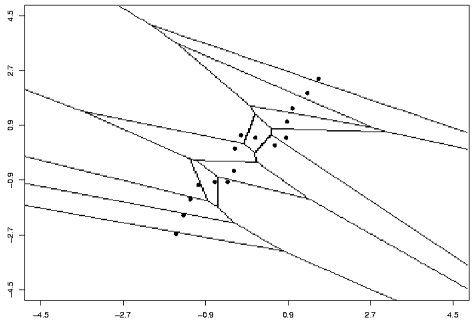


Figure 6.17: Decision boundaries for joint (Y,X) clustering method TJ,M^s (TS-SOM).

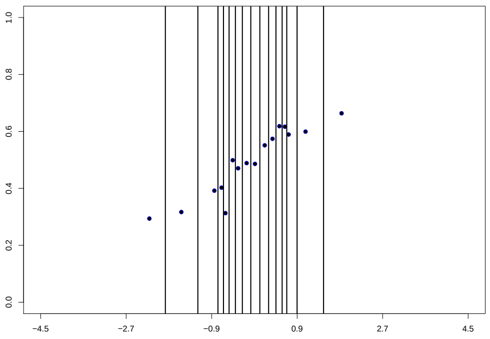


Figure 6.18: Decision boundaries for X-clustering method C,M (K-Means).

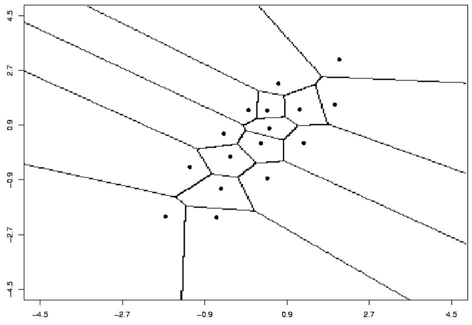


Figure 6.19: Decision boundaries for joint (Y,X) clustering method CJ,M (K-Means).

6.7 Summary

Imputation methods based on standard (K-Means type) clustering using completely observed $\boldsymbol{X}$-covariates, as well as clustering of joint distribution $f_{Y,\boldsymbol{X}}(y, \mathbf{x})$, and smoothed imputation (TS-SOM methods) approaches in imputation were analysed. When considering the results, there are two differences between the approaches: i) how incomplete observations are classified to cells and ii) how smoothing is used. If incomplete observations are classified in a probabilistic manner (joint $(Y, \boldsymbol{X})$ clustering using TS-SOM) then one is likely to produce misclassifications. However, the benefit is that one is able to model and impute multimodal distributions. Smoothing aims to improve estimates by borrowing strength from the neighborhood. The hope is that smoothing will reduce estimation variance without significantly increasing the estimation bias.

Variance of mean estimator given a model depends on the covariance structure of the missing data. This quantity is somewhat abstract. To give some insight to this structure, an assumption on the distributions of a number of missing data values within cells were given, under a MAR assumption, for a given model. A simulation example showed that a derived analytical formula may be used to estimate variance of mean estimator.

A view about differences between cell methods was given by a simulation example. Characteristics of clustering algorithms were evaluated by total scatter matrix and its decomposition into a between-cells scatter matrix and within-cells scatter matrix. A major difference between the K-Means and TS-SOM algorithms is smoothing used by TS-SOM. It leads to underestimation of variance but on the other hand it is able to improve the mean square error of an estimator (mean estimator in the example) sometimes considerably. A better performance, when considering simultaneously the mean square errors of the first two moment estimators, may be reached for the TS-SOM methods if neighborhood smoothing is used when computing mean and variance estimators. Further, a rough connection between the characteristics of clustering and imputation performance was observed. For all the cell methods with mean strategy, except the joint TS-SOM methods, underestimation of bias of the second moment is proportional to residual variance of Y (related to an element in the within-cells scatter matrix). A probabilistic classifier of incomplete observations makes things more complicated for joint $(Y, \boldsymbol{X})$ clustering with TS-SOM.

Chapter 7

Evaluation of imputation using simulated data sets

There are three ways to evaluate imputation methodology: theoretical studies, simulations, and real-world experiments. As we can see from the previous chapters, theoretical studies are often difficult to interpret in practical terms. Real-world experiments, on the other hand, may not tell us why some method is better than another. The role of simulations is to fill this cap.

Good simulations can show direct relations between certain variations in data and variations in the performance of the studied methodology. Ideally, simulated experiments can be linked to theoretical properties of the methods, and thus simulations could be the results of theoretical analyses. We have seen this kind of studies for individual methods already in Chapter 4. In the current chapter we extend simulations to cover all the methods simultaneously. The purpose is to see what the differences between the methods are.

Our carefully designed simulation experiments are introduced to evaluate a total of 24 combinations of methods and imputation strategies, as shown in Table 7.1. The three simulation cases are:

1) The role of MAR type of missingness, where data is varied from simple MCAR to strongly MAR.

2) Imputation of NMAR multimodal data, where $f(Y^{mis}|X)$ is varied from single to strongly multimodal distribution.

3a) The effect of dimensionality, which is studied for a simple classification problem with increasing number of covariates.

3b) Computational performance of the methods as a function of data size and dimension.

The methods are used more or less as described in Chapters 4, 5, and 6. The role of baseline methods is to give an insight to "easily" achieveable imputation performance. Thus we should do better. The nearest neighbor method is used with one neighbor only ($k = 1$). For kernel regression, a symmetric Gaussian kernel is used, with bandwidth λ, which is varied as a function of sample size as proposed by Mack in [72].

Abbreviation	**Method**
B,M	Baseline with mean imputation strategy.
B,R	Baseline with random strategy.
B,D	Baseline with donor strategy.
L,M	Linear regression with mean imputation strategy.
L,R	Linear regression with random strategy.
N,M	k-nearest neighbour with mean imputation strategy ($k = 1$ always in this chapter).
K,M	Kernel regression with mean imputation strategy.
K,R	Kernel regression with random strategy.
T,M	TS-SOM covariate $\boldsymbol{X}$ clustering with mean strategy.
T,R	TS-SOM covariate $\boldsymbol{X}$ clustering with random strategy.
T,D	TS-SOM covariate $\boldsymbol{X}$ clustering with donor strategy.
T,M^s	TS-SOM covariate $\boldsymbol{X}$ clustering with mean strategy and smoothing.
T,R^s	TS-SOM covariate $\boldsymbol{X}$ clustering with random strategy and smoothing.
C,M	K-Means covariate $\boldsymbol{X}$ clustering with mean strategy.
C,R	K-Means covariate $\boldsymbol{X}$ clustering with random strategy.
C,D	K-Means covariate $\boldsymbol{X}$ clustering with donor strategy.
TJ,M	TS-SOM joint $(Y, \boldsymbol{X})$ clustering with mean strategy.
TJ,R	TS-SOM joint $(Y, \boldsymbol{X})$ clustering with random strategy.
TJ,D	TS-SOM joint $(Y, \boldsymbol{X})$ clustering with donor strategy.
TJ,M^s	TS-SOM joint $(Y, \boldsymbol{X})$ clustering with mean strategy and smoothing.
TJ,R^s	TS-SOM joint $(Y, \boldsymbol{X})$ clustering with random strategy and smoothing.
CJ,M	K-Means joint $(Y, \boldsymbol{X})$ clustering with mean strategy.
CJ,R	K-Means joint $(Y, \boldsymbol{X})$ clustering with random strategy.
CJ,D	K-Means joint $(Y, \boldsymbol{X})$ clustering with donor strategy.

Table 7.1: Abbreviations for compared methods.

Simulations are done under random repetitions, where a random sample is taken from a specified data generator, and then all the methods are applied to impute the missing part of data. This process is repeated until the desired measures of imputation performance are stable enough. Thus we try to eliminate the role of simulation variation by giving empirical variants of expectation of simulated imputation measures. The measures used in the simulation cases 1, 2, and 3a are described briefly next.

First, two moments of mean estimator are estimated in **Case 1**. The estimates are

$$
\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp}] &\approx \frac{1}{\mathsf{n}_{sim}} \sum_{sim=1}^{\mathsf{n}_{sim}} \widehat{\mathbb{B}\mathrm{ias}}\Big[[\mu^{comp}]^{sim}\Big] \\
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}] &\approx \frac{1}{\mathsf{n}_{sim}-1} \sum_{sim=1}^{\mathsf{n}_{sim}} \Big([\mu^{comp}]^{sim} - \overline{\mu}^{comp}\Big)^2,
\end{aligned}
$$

where $\widehat{\mathbb{B}\mathrm{ias}}[[\mu^{comp}]^{sim}] = [\mu^{comp}]^{sim} - \mu^*$, $[\mu^{comp}]^{sim}$ is the value of $\hat{\mu}^{comp}$ in the sim:th repetition, $\overline{\mu}^{comp} = \frac{1}{\mathsf{n}_{sim}-1}\sum_{sim=1}^{\mathsf{n}_{sim}}[\mu^{comp}]^{sim}$, sim denotes one simulation run and n_{sim} is the number of repetitions. As one can imagine most of the repetitions are required to estimate imputation variances of higher moments.

In **Case 2** we estimate conditional and marginal biases of the first two moment estimators. In addition, marginal and conditional Kolmogorov-Smirnov distances are estimated. The quantities are computed as

$$\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{\mu}^{mis}|x=3.5] &\approx \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}} \widehat{\mathbb{B}\mathrm{ias}}\Big[[\mu^{imp}|x=3.5]^{sim}\Big] \\
\mathbb{B}\mathrm{ias}[\hat{\tau}^{mis}|x=3.5] &\approx \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}} \widehat{\mathbb{B}\mathrm{ias}}\Big[[\tau^{imp}|x=3.5]^{sim}\Big] \\
\mathbb{B}\mathrm{ias}[\hat{\mu}^{mis}] &\approx \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}} \widehat{\mathbb{B}\mathrm{ias}}\Big[[\mu^{imp}]^{sim}\Big] \\
\mathbb{B}\mathrm{ias}[\hat{\tau}^{mis}] &\approx \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}} \widehat{\mathbb{B}\mathrm{ias}}\Big[[\tau^{imp}]^{sim}\Big],
\end{aligned}$$

in which $[\mu^{imp}|x=3.5]$ and $[\tau^{imp}|x=3.5]$ are mean and variance estimates computed from imputations done at $x = 3.5$. Kolmogorov-Smirnov distances are estimated as

$$\begin{aligned}
\mathrm{KSc} &= \mathbb{E}\Big[\sup_y |F_{Y^{mis}|x}(y|x=3.5) - \hat{F}_{Y^{imp}|X}(y|x=3.5)|\Big|\mathsf{n}\Big] \approx \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}}[ksc]^{sim} \\
&= \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}} \max_{y\in\mathbb{R}}\Bigg|F_{Y^{mis}|x}(y|x=3.5) - \underbrace{\frac{1}{\mathsf{n}^{mis}}\sum_{j=1}^{\mathsf{n}^{mis}} I([y_j^{comp}]^{sim} \le y|x=3.5)}_{\text{estimate of } F_{Y^{imp}}(y)|x=3.5}\Bigg| \\
\mathrm{KS} &= \mathbb{E}\Big[\sup_y |F_{Y^{mis}}(y) - \hat{F}_{Y^{imp}}(y)|\Big|\mathsf{n}\Big] \approx \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}}[ks]^{sim} \\
&= \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}} \max_{y\in\mathbb{R}}\Bigg|F_{Y^{mis}}(y) - \underbrace{\frac{1}{\mathsf{n}^{mis}}\sum_{j=1}^{\mathsf{n}^{mis}} I([y_j^{comp}]^{sim} \le y)}_{\text{estimate of } F_{Y^{imp}}(y)}\Bigg| \\
\mathrm{KS}_2 &= \mathbb{E}\Big[\sup_y |F_{Y^{mis}|x}(y|x\in\Delta_x) - \hat{F}_{Y^{imp}|X}(y|x\in\Delta_x)\Big|\mathsf{n}\Big] \approx \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}}[ksc]^{sim} \\
&= \frac{1}{\mathsf{n}_{sim}}\sum_{sim=1}^{\mathsf{n}_{sim}} \max_{y\in\mathbb{R}}\Bigg|F_{Y^{mis}|x}(y|x\in\Delta_x) - \underbrace{\frac{1}{\mathsf{n}^{mis}}\sum_{j=1}^{\mathsf{n}^{mis}} I([y_j^{comp}]^{sim} \le y|x\in\Delta_x)}_{\text{estimate of } F_{Y^{imp}}(y|x\in\Delta_x)}\Bigg|,
\end{aligned}$$

where Δ_x is the interval of imputation positions. Note that distributions $F_{Y^{mis}}(y)$, $F_{Y^{mis}|x=3.5}(y)$, and $F_{Y^{mis}|x\in\Delta_x}(y)$ are also estimated (with a fixed sample), what is described in the Case 3a.

In **Case 3a** we estimate the integrated mean squared error

$$\text{MISE}(\hat{Y}^{imp}) = \mathbb{E}_{\boldsymbol{Z}}\mathbb{E}_{\boldsymbol{X}}[(\hat{Y}^{imp}_{|\mathbf{x},\mathbf{z}} - Y^{mis}_{|\mathbf{x},\mathbf{z}})^2|\mathbf{z}],$$

where $\mathbf{z}$ is a parameter which is described in the case. The MISE is decomposed to expected squared bias and variance terms (and expected variability of target, an irreducible term, which is omitted here)

$$\begin{aligned}
\mathbb{B}\text{ias}^2[\hat{Y}^{imp}] &= \mathbb{E}_{\boldsymbol{Z}}\mathbb{E}_{\boldsymbol{X}}\Big[\big(\mathbb{E}[\hat{Y}^{imp}|\mathbf{x},\mathbf{z}] - \mathbb{E}[Y^{mis}|\mathbf{x},\mathbf{z}]\big)^2|\mathbf{z}\Big] \\
&\approx \frac{1}{10}\sum_{l=1}^{10}\left[\frac{1}{1600}\sum_{j=1}^{1600}\big(\hat{\mathbb{E}}[\hat{Y}^{imp}|\mathbf{x}_{j,l},\mathbf{z}_l] - \mathbb{E}[Y^{mis}|\mathbf{x}_{j,l},\mathbf{z}_l]\big)^2\right] \\
&= \frac{1}{10}\sum_{l=1}^{10}\left[\frac{1}{1600}\sum_{j=1}^{1600}\big(\frac{1}{\mathsf{n}_{sim}}\sum_{i=1}^{\mathsf{n}_{sim}} y^{imp}_{j,l,i} - \mathbb{E}[Y^{mis}|\mathbf{x}_{j,l},\mathbf{z}_l]\big)^2\right],
\end{aligned}$$

where 1600 integration positions $\mathbf{x}_{j,l}$ are drawn once for each parameter $\mathbf{z}_l$. Expected variance is computed as

$$\begin{aligned}
\mathbb{V}\text{ar}[\hat{Y}^{imp}] &= \mathbb{E}_{\boldsymbol{Z}}\mathbb{E}_{\boldsymbol{X}}\Big[\mathbb{V}\text{ar}[\hat{Y}^{imp}|\mathbf{x},\mathbf{z}]|\mathbf{z}\Big] \\
&\approx \frac{1}{10}\sum_{l=1}^{10}\left[\frac{1}{1600}\sum_{j=1}^{1600}\Big[\hat{\mathbb{V}\text{ar}}[\hat{Y}^{imp}|\mathbf{x}_{j,l},\mathbf{z}_l]\Big]\right] \\
&= \frac{1}{10}\sum_{l=1}^{10}\left[\frac{1}{1600}\sum_{j=1}^{1600}\Big[\frac{1}{\mathsf{n}_{sim}-1}\sum_{i=1}^{\mathsf{n}_{sim}}(y^{imp}_{j,l,i} - \hat{\mathbb{E}}[\hat{Y}^{imp}|\mathbf{x}_{j,l},\mathbf{z}_l])^2\Big]\right].
\end{aligned}$$

7.1 Case 1: the role of missing-data mechanism

The purpose of this study is to evaluate the effect of a missing-data mechanism, missingness is varied from MCAR to MAR. We do this by increasing the difference between $\mathbb{E}[Y]$ and $\mathbb{E}[Y^{mis}]$. For simplicity, the results are evaluated in terms of the first moment bias and variance for a given number of observations.

7.1.1 Data generator

Our data follows a simple model

$$Y = X^I + \epsilon, \text{ where } \epsilon \sim N(0,3),\ I \sim Bernoulli(0.5)$$

and X^I is generated from a different Normal distribution for missing and observed parts, namely

$$X^I = \begin{cases} X^{obs} & \sim N(0,5), \quad \text{if } I = 0 \\ X^{mis} & \sim N(a_t,5), \quad \text{if } I = 1. \end{cases}$$

The mean a_t of missing data covariate X^{mis} is varied in ten experiments $t = 1,\dots,10$ such that $a_t = \frac{2}{9}(t-1)$, giving $a_t \in [0,2]$. Thus data varies from MCAR to MAR such that in the final experiments the data is as seen in Figure 7.1.

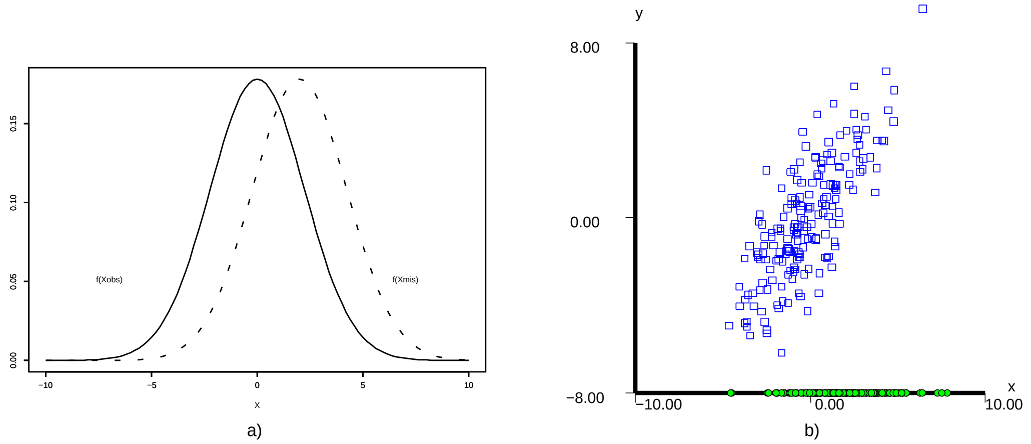


Figure 7.1: a) distributions of X^{obs} and X^{mis} at $t = 10$, b) random sample from superpopulation at $t = 10$.

We shall take an iid random sample of $\mathsf{n} = 1000$ observations. As a consequence the number of missing observations is $\mathsf{N}^{mis} \sim Bin(1000, 0.5)$, and $\mathsf{N}^{obs} = \mathsf{n} - \mathsf{N}^{mis}$.

The methods are used in a rather standard way. The bandwidth for kernel regression is set to $\lambda = 2(\mathsf{N}^{obs})^{-1/5}$. TS-SOM with a one-dimensional latent lattice topology and 32 cells is used, and the same number of cells is used in K-Means clustering too.

7.1.2 Theoretical considerations

From Chapters 4, 5 and 6 we may try to predict what the outcome of these experiments is. Clearly, our methods can be divided between predictive and nonpredictive ones. Baseline methods that cannot utilize covariate X will do much worse than all the other algorithms. In fact we may predict what the error is because

$$\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,B}] = p^*(\mu^{*obs} - \mu^{*mis}).$$

Applying the numbers we get

$$\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,B}] = 0.5\Big(0 - \frac{2}{9}(t-1)\Big),$$

which implies that $\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,B}]$ moves from 0 to -1 as $t = 1, \ldots, 10$. The variances for baseline methods are

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}] &\approx \tau^{*obs}\left(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\right) \\
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}] + \tau^{*obs}\frac{p^*}{\mathsf{n}} \\
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}] + \tau^{*obs}\frac{p^*}{\mathsf{n}}\Big(1 - \frac{1}{\mathsf{n}(1-p^*)}\Big)
\end{aligned}$$

Applying numbers gives

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}] &\approx 8\left(\frac{1}{1000(1-0.5)} + \frac{1000*0.5(1-0.5)}{1000^3(1-0.5)^3}\right) = 0.016 \\
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}] + 8\frac{0.5}{1000} = 0.02 \\
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}] + 8\frac{0.5}{1000}\Big(1 - \frac{1}{1000(1-0.5)}\Big) \approx 0.02.
\end{aligned}$$

For linear regression we have

$$\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,L}] \approx p^*(\mathbb{E}[\hat{\beta}^{obs}_{-0}|\mathsf{n}]^T\overline{X}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}] - \mu^{*mis})$$

Now $\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}] \approx \boldsymbol{\beta}^{*obs} = (1\ 0)^T$. Further, $\mu^{*mis} = \beta^{*mis}_{-0}\overline{X}^{*mis} + \beta^{*mis}_0 = \overline{X}^{*mis}$. Applying numbers we get

$$\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,L}] \approx 0.5(\overline{X}^{*mis} - \overline{X}^{*mis}) = 0,$$

which implies that linear regression is approximately unbiased for all experiments $t = 1, \ldots, 10$.

Approximate bias for kernel regression and 1-nearest neighbour regression is given by Approximation 5.1. Note that the bias is a function of the density of X^{obs} and conditional mean $\mathbb{E}[Y^{obs}|x] = g^{*obs}(x)$ and their first or second order derivatives. Thus a bit more work than earlier is required to solve the bias. To apply Approximation 5.1 one needs to notice that the NMAR bias is zero and the bias due to noise estimation is zero (donor strategy is not used). Computation of approximate bias is simplified by ignoring the approximation term and applying the first order Taylor approximation to the estimation bias term. Thus bias is computed as

$$\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] &\approx \mathbb{E}_{\mathsf{N}^{mis}}[\frac{\mathsf{N}^{mis}}{\mathsf{n}}C] + O(\mathsf{n}^{-1}) \approx \mathbb{E}_{\mathsf{N}^{mis}}[\frac{\mathsf{N}^{mis}}{\mathsf{n}}C] \\
&\overset{Taylor}{\approx} \mathbb{E}_{\mathsf{N}^{mis}}[\frac{\mathsf{N}^{mis}}{\mathsf{n}}]\mathbb{E}_{\mathsf{N}^{mis}}[C] \approx p^*\mathbb{E}_{\mathsf{N}^{mis}}[C] = 0.5\mathbb{E}_{\mathsf{N}^{mis}}[C].
\end{aligned}$$

Estimation bias term C for both of the methods depends on the following quantity

$$\begin{aligned}
(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis}) &= 2f'_{X^{obs}}(\overline{X}^{*mis}) \\
&= 2f'_{X^{obs}}(\frac{2}{9}(t-1)).
\end{aligned}$$

The expectation of C for kernel regression is approximately

$$\begin{aligned}
\mathbb{E}[C] &\overset{Taylor}{\approx} \frac{2f'_{X^{obs}}(\frac{2}{9}(t-1))}{2f_{X^{obs}}(\overline{X}^{*mis})}\int \xi^2K(\xi)d\xi\lambda^2\big(\mathsf{n}(1-p^*)\big) \\
&= \frac{f'_{X^{obs}}(\frac{2}{9}(t-1))}{f_{X^{obs}}(\frac{2}{9}(t-1))}\lambda^2\big(0.5\mathsf{n}\big) = \frac{f'_{X^{obs}}(\frac{2}{9}(t-1))}{f_{X^{obs}}(\frac{2}{9}(t-1))}4\big(0.5\mathsf{n}\big)^{-2/5}.
\end{aligned}$$

For 1-nearest neighbour the expectation of C is

$$\begin{aligned}\mathbb{E}[C] &\overset{Taylor}{\approx} \frac{2f'_{X^{obs}}(\frac{2}{9}(t-1))}{24f^3_{X^{obs}}(\overline{X}^{*mis})}\big(1/\mathbb{E}[\mathsf{N}^{obs}]\big)^2\\ &= \frac{f'_{X^{obs}}(\frac{2}{9}(t-1))}{12f^3_{X^{obs}}(\frac{2}{9}(t-1))}\big(1/(0.5\mathsf{n})\big)^2 = \frac{f'_{X^{obs}}(\frac{2}{9}(t-1))}{12*500^2 f^3_{X^{obs}}(\frac{2}{9}(t-1))}.\end{aligned}$$

Now we need to compute $f'_{X^{obs}}(x)$,

$$\begin{aligned}f'_{X^{obs}}(x) &= \frac{\partial}{\partial x}\frac{1}{\sqrt{2\pi*5}}\exp\Big(-\frac{x^2}{2*5}\Big)\\ &= \frac{1}{\sqrt{10\pi}}\frac{\partial}{\partial x}\exp\Big(-\frac{x^2}{10}\Big) = -\frac{2x}{10\sqrt{10\pi}}\exp(-\frac{x^2}{10}).\end{aligned}$$

The approximate bias for 1-nearest neighbour is zero at $t = 1$. The bias moves from $-2.4*10^{-7}$ to $-4.7*10^{-6}$ as $t = 2,\ldots,10$. Therefore the bias for the nearest neigbour is expected to be almost zero. However, for kernel regression the situation is different. Approximation to bias $\mathbb{B}\text{ias}[\hat{\mu}^{comp,K}|\mathsf{n}]$ is shown in Table 7.2. The bias is linear for these values of a_t and roughly follows equation $\mathbb{B}\text{ias}[\hat{\mu}^{comp,K}|\mathsf{n}] = -p^* \cdot 0.0666 \cdot a_t = -0.0333a_t$.

a_t	0 (t=1)	2/9	4/9	6/9	8/9	10/9 (t=6)	12/9	14/9	16/9	2 (t=10)
$\mathbb{B}\text{ias}[\hat{\mu}^{comp,K}\|\mathsf{n}]$ (theoretical)	0.000	-0.007	-0.015	-0.022	-0.030	-0.037	-0.044	-0.052	-0.059	-0.067

Table 7.2: Theoretical bias for kernel regression as a function of $a_t = \frac{2}{9}(t-1),\ t = 1,\ldots,10$.

7.1.3 Simulation results for Case 1

In simulations we used $\mathsf{n}_{sim} = 1000$ repetitions. The results are shown in Figures 7.2 and 7.3. The main effect of the simulated experiments is quite expected; the baseline methods do not preserve the first moment, as shown in Figure 7.2. There also seems to be a small overestimation of $\hat{\mu}^{comp}$ with TS-SOM joint clustering (Y, X) methods, namely TJ and TJ^s. As expected kernel regression methods underestimate the first moment as our MAR parameter a_t is increased from zero. All the other methods yield a bias that is close to zero.

Imputation variance $\mathbb{V}\text{ar}[\hat{\mu}^{comp}|\mathsf{n}]$ is not dependent on the parameter a_t for baseline methods as expected. The variances for baseline methods are as expected. The variances for other methods are not strongly dependent on the parameter, but there are large differences between the methods. The imputation variance of the joint (Y, X) K-Means clustering is largest, which relates directly to the estimation variance of 2-D mean vector $(\overline{y}_i, \overline{x}_i)$ from data sets of size $0.5*1000/32 \approx 15$ samples (recall that on average 50% of samples are missing). For TS-SOM this variance is considerably smaller, because of neighbour smoothing (strength borrowing). As t is increased it seems that variance grows for most of the imputation methods. For nonparametric regression methods and cell methods this is probably due to the fact that density of X^{obs} is decreasing at imputation positions. A partial reason for linear regression is that pointwise prediction variance grows quadratically as a function of covariate.

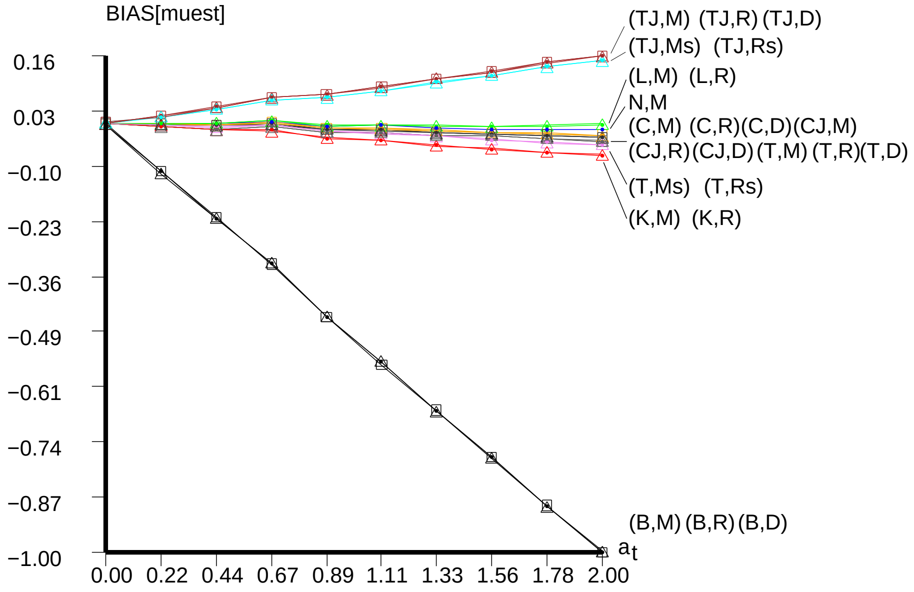


Figure 7.2: Bias of $\hat{\mu}^{comp}$ as function of a_t.

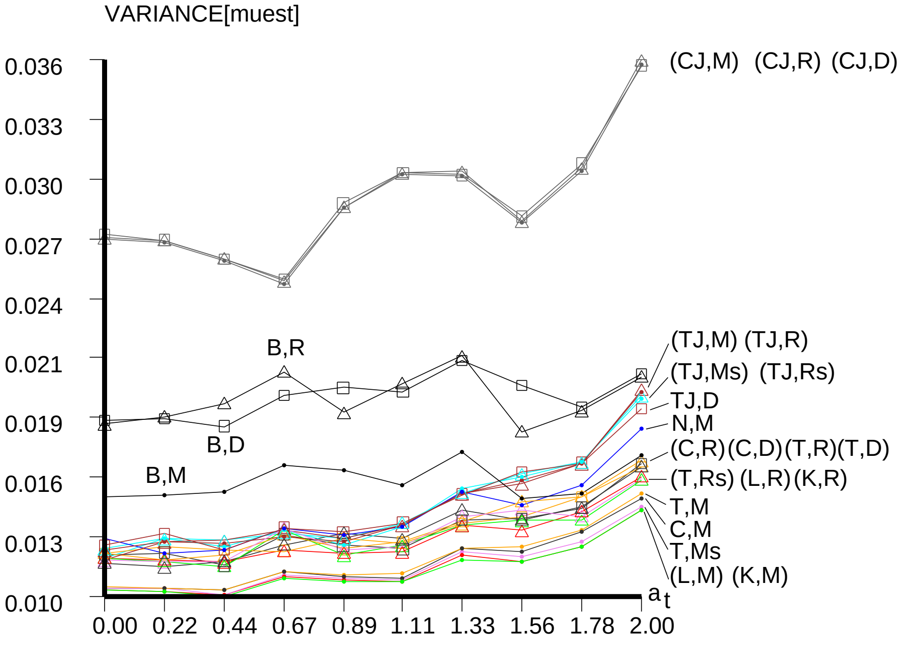


Figure 7.3: Variance of $\hat{\mu}^{comp}$ as function of a_t.

7.2 Case 2: imputation of multimodal $f_{Y|X}(y|x)$

In this example we shall study imputation in a case where covariate $\boldsymbol{X}$ does not fully explain the distribution of Y, but where missingness depends on Y itself. This is an example of NMAR type of missingness. The study was conducted by generating three gaussian components in (X,Y) space as depicted in Figure 7.4, such that the conditional distribution at $x = 3.5$ $f_{Y|X}(y|x = 3.5)$ is multimodal, because Gaussians B and C are centered at the same X-position.

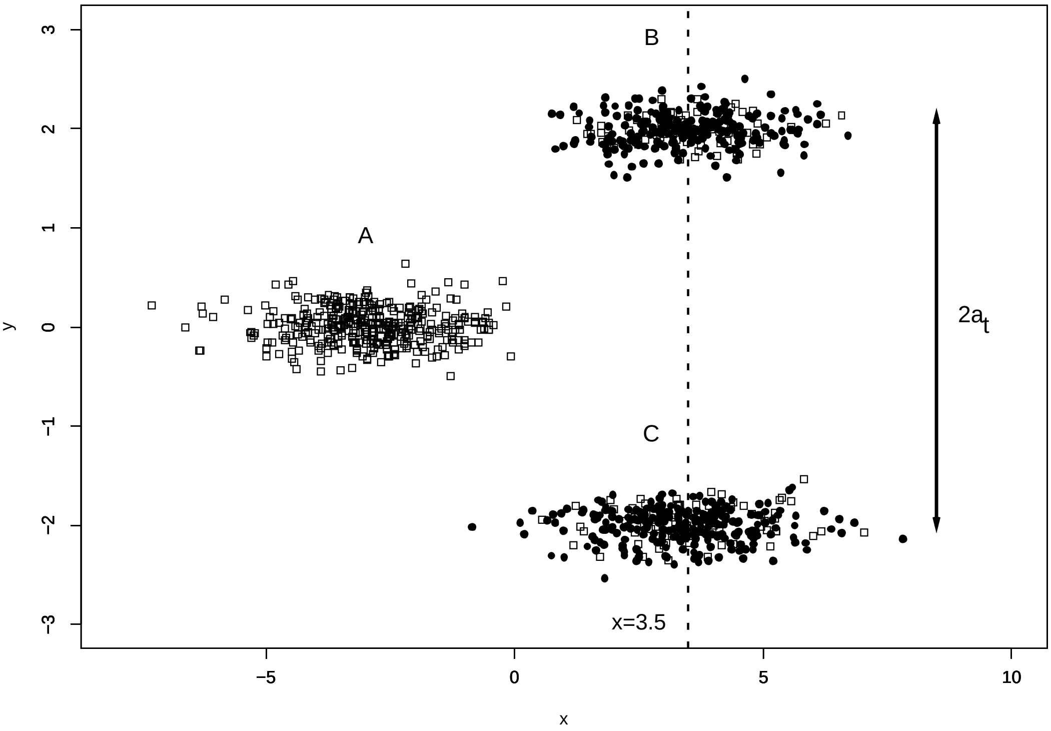


Figure 7.4: The data set for Case 2 ($a_t = 2$). The squares denote training data, and the draws from a missing distribution are shown as black dots.

In the experiments we study the imputation performance of the methods, in a setup where the distance $2a_t$ between Gaussian B and C at $x = 3.5$ is increased, and missingness takes place equally in components B and C. We shall use mainly conditional measures of imputation performance at $x = 3.5$, such as conditional Kolmogorov-Smirnov distance KSc($X = 3.5$), $\mathbb{B}\text{ias}[\hat{\mu}^{imp}|X = 3.5]$ and $\mathbb{B}\text{ias}[\hat{\tau}^{imp}|X = 3.5]$, but for the sake of completeness also marginal measures $\mathbb{B}\text{ias}[\hat{\mu}^{mis}]$, $\mathbb{B}\text{ias}[\hat{\tau}^{mis}]$, and KS are computed.

All models are built using incomplete finite sample as explained above. After training, the same data is used for the evaluation of measures related to statistical moments. However, the study of conditional measures at $X = 3.5$ is done separately by generating a sufficiently large test sample of observations at $X = 3.5$.

All the methods are used as usual. For TS-SOM a 1D latent structure is used with 32 cells. The same number of cells is used for K-Means. In the case of joint

(Y,X) clustering, a ring topology where the ends of 1D SOM-lattice are connected to form a "ring" of cells is used for TS-SOM. This seems to get organised better than a 1D-lattice. For kernel regression the bandwidth is selected to be $\lambda = 2(N^{obs})^{-1/5}$.

7.2.1 Data generator and setup of the experiment

The generator of data can be written as a Gaussian mixture

$$f_{X,Y} = \pi_A \cdot N(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) + \pi_B \cdot N(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B) + \pi_C \cdot N(\boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C),$$

where priors are $\pi_A = \frac{10}{28}, \pi_B = \frac{9}{28}, \pi_C = \frac{9}{28}$. The locations of Gaussian are at

$$\boldsymbol{\mu}_A = (-3, 0)^T, \boldsymbol{\mu}_B = (3.5, a_t), \text{ and } \boldsymbol{\mu}_C = (3.5, -a_t)$$

and their shapes are defined by

$$\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}_C = \text{diag}(1.5, 0.03).$$

MCAR missingness is applied with equal probability 7/9 to components B and C only, as depicted in Figure 7.4. Thus the marginal densities of the covariate are

$$\begin{aligned} f_{X^{obs}}(x^{obs}) &= \frac{5}{7}N(-3, 1.5) + \frac{2}{7}N(3.5, 1.5), \text{ and} \\ f_{X^{mis}}(x^{mis}) &= N(3.5, 1.5). \end{aligned}$$

The densities are illustrated in Figure 7.5.

In the experiments iid samples of $\mathsf{n} = 1000$ observations are taken. As a consequence the number of missing observations is $\mathsf{N}^{mis} \sim Bin(1000, 0.5)$ and $\mathsf{N}^{obs} = \mathsf{n} - \mathsf{N}^{mis}$.

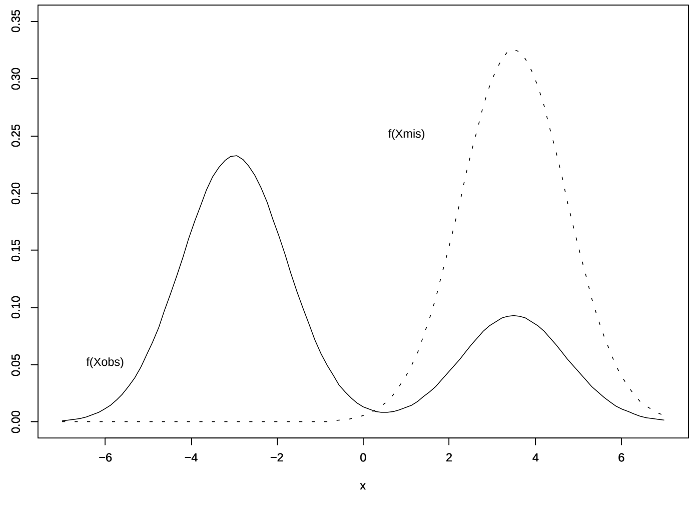


Figure 7.5: Marginal densities $f_X(x)$ for missing and observed datas.

7.2.2 Theoretical predictions of KSc at $X = 3.5$

Here we try to predict the outcome of simulation runs using theoretical insight that were developed in the previous chapters. Yet, We do not follow the previous results directly, but rather we use them as a guideline for a more detailed analysis under the current case. This analysis includes Kolmogorov-Smirnov measures (K-S), which we believe are the best descriptors of imputation performance under multimodal missingness.

The Kolmogorov-Smirnov measures for $f_{Y|X}(y|x = 3.5)$

Here we shall study the conditional measure

$$\mathrm{KSc} = \mathbb{E}\left[\sup_y |F_{Y^{mis}|x}(y|x = 3.5) - \hat{F}_{Y^{imp}|X}(y|x = 3.5)| \middle| \mathsf{n} \right].$$

When $\mathsf{n} \to \infty$ it is relatively easy to see that there are 10 types of behaviours among the combinations of imputation methods and strategies that are studied in this thesis. The behaviours are labelled A1-A3, B1-B5, C1-C2 and they can be characterized as in Table 7.3. They are illustrated in Figure 7.6.

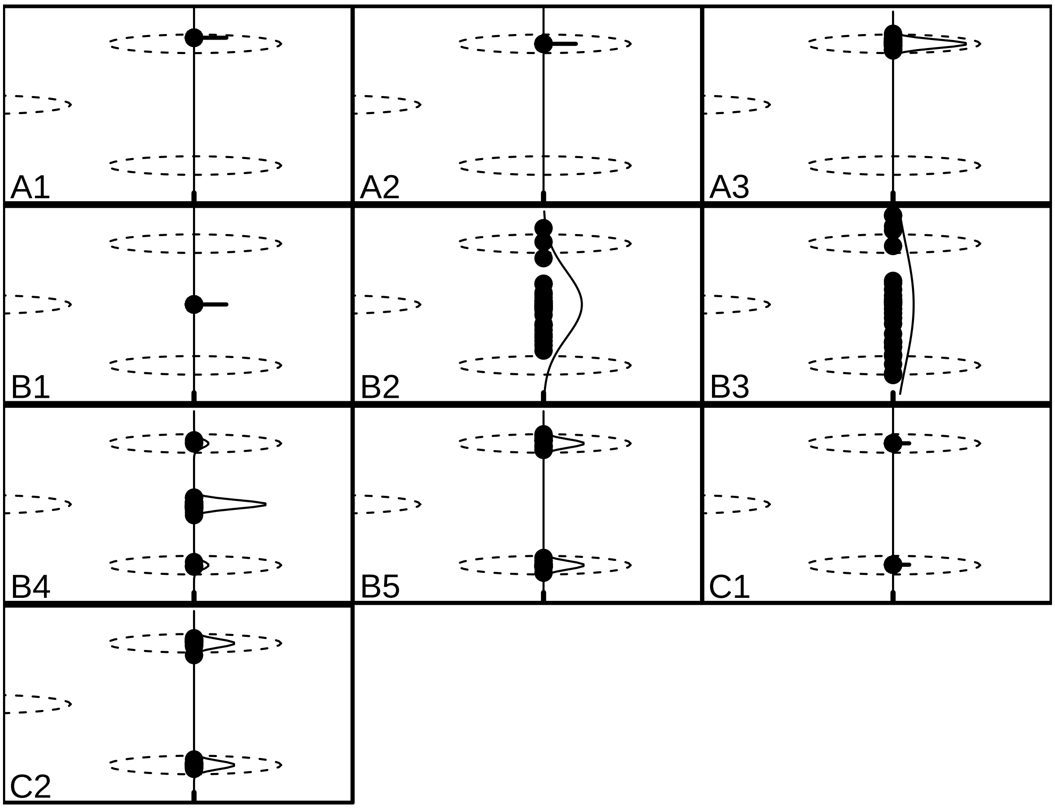


Figure 7.6: An illustration of the behaviour of ten different types of imputation models for Case 2. The dashed ellipses denote the 90% confidence regions for the Gaussian components A-C, and the dots illustrate realizations of imputations. Distributions $f_{Y^{imp}|x=3.5}(y)$ have also been drawn.

Class	**Methods**	**Description**
A1	N,M	Flexible model that goes through one observation at $x = 3.5$.
A2	CJ,M	Stiff model that goes through either mode of component B or C at $x = 3.5$.
A3	CJ,R CJ,D	Same as A2 but added noise in the realization of imputations.
B1	B,M L,M K,M (C,M),(T,M^s),(T,M)	Stiff model that goes through the center of data (between two modes).
B2	B,R L,R K,R	Same as B1 but added noise of Y^{obs} in the realization of imputation (mean + noise without covariate information).
B3	T,R^s T,R C,R	Same as B1 and B2 but added noise is estimated at $X = 3.5$ (mean + noise at $X = 3.5$).
B4	B,D	Random donor without covariate information (yields three modes).
B5	C,D T,D	Conditionalized random donor at $X = 3.5$.
C1	TJ,M^s TJ,M	Multimodal algorithm that randomly selects either mode for the realization of imputation.
C2	TJ,D TJ,R TJ,R^s	Methods can, in principle, model any conditional distribution $f_{Y^{mis}\|X}(y\|x)$ provided distributions $f_{Y^{mis}\|X}$ and $f_{Y^{obs}\|X}$ are same.

Table 7.3: Characteristical differences between methods in the multimodal case.

With the above idealizations we can assume that imputed data at $x = 3.5$ follows distribution

$$Y^{imp} = \begin{cases} Y^{obs}(j|x=3.5), & \text{if class A1} \\ \hat{\mu}_i, & \text{if class A2} \\ Y^{obs}(j|x=3.5) + \hat{\epsilon}_I, & \text{if class A3} \\ \hat{\mu}^{obs}, & \text{if class B1} \\ \hat{\mu}^{obs}_{|x=3.5} + \hat{\epsilon}, & \text{if class B2} \\ \hat{\mu}^{obs}_{|x=3.5} + \hat{\epsilon}_{|x=3.5}, & \text{if class B3} \\ \hat{\mu} + \hat{\epsilon}^D, & \text{if class B4} \\ \hat{\mu} + \hat{\epsilon}^D_{|x=3.5}, & \text{if class B5} \\ \hat{\mu}_I, & \text{if class C1} \\ \hat{\mu}_I + \hat{\epsilon}_I, & \text{if class C2,} \end{cases} \tag{7.1}$$

where

$Y^{obs}(j|x = 3.5)$ denotes some observation close to $x = 3.5$. Thus this model follows randomly selected observation.

$\hat{\mu}_i$ is the mean of mode (B or C) that is closest to $x = 3.5$.

$\hat{\epsilon}_I$ is randomness that is selected from either the component B or C. Thus $\hat{\epsilon}_I = N(0, 0.03)$.

$\hat{\mu}^{obs}$ is the mean of observed data: $\hat{\mu}^{obs} = \frac{1}{\mathsf{n}^{obs}} \sum_j Y_j^{obs}$.

$\hat{\mu}^{obs}_{|x=3.5}$ is the mean at $x = 3.5$.

$\hat{\epsilon}$ is randomness that is drawn from a zero mean Gaussian distribution with variance $\hat{\tau}^{obs} = \frac{1}{\mathsf{n}^{obs}-1} \sum_j (Y_j^{obs} - \hat{\mu}^{obs})^2$.

$\hat{\epsilon}^D$ is randomness that is drawn from residuals $\{Y_j^{obs} - \hat{\mu}^{obs}\}_j$.

$\hat{\epsilon}^D_{|x=3.5}$ is randomness that is drawn from residuals $\{Y_j^{obs} - \hat{\mu}^{obs}\}_j$ for which $\{X_j^{obs}\}$ are close to $x = 3.5$.

$\hat{\mu}_I$ is the mean of mode (B or C) which is selected randomly.

Using Equation (7.1) it is possible to compute the expected predictions of the K-S measure at $X = 3.5$. For some groups this is quite easy. For example we may assume that

$$\text{KSc} \approx \begin{cases} 0.75, & \text{for group A1,A2} \\ 0.5, & \text{for groups B1,A3} \\ 0, & \text{for group C2/B5.} \end{cases}$$

In group A2 predictions are in either mode (B or C). If they are in mode B then the cumulative probability for $Y^{mis}|x = 3.5$ is 0.75 ”epsilon” below mode B, and 0 for predictions. In case that predictions are in mode C cumulative probability of $Y^{mis}|x = 3.5$ is 0.25, whereas for predictions it is 1.00. In either case the KSc value is 0.75. If the variance within modes B and C is ignored then A1 behaves as A2. Of course this does not hold, but the approximation is still somewhat justifiable the because variance within modes B and C is small relative to the variance between the two modes (randomly selected closest observation is close to either of the modes).

The methods in group B1 yield predictions that are at the center of data (between the two modes). Thus the supremum of absolute differences between conditional cumulative distribution functions is reached at this position because $f_{Y^{mis}|x=3.5}$ is symmetric. The supremum value, KSc, is 1.0-0.5=0.5 .

In group A3 all predictions are in either component (B or C). Thus ”component prior is doubled”. Therefore supremum value is reached at the center of data due to the symmetry of $f_{Y^{mis}|x=3.5}$. If all predictions are in component B then the cumulative mass of predictions between the modes is zero, whereas it is 0.50 for missing values. On the other hand, if the predictions are in component C then mass is 1.0 for predictions and 0.50 for missing values. In either case the supremum value is $|1.0 - 0.5| = 0.5$.

Imputation distribution is exactly the same as the distribution of missing values at $x = 3.5$ in groups C2 and B5. Therefore supremum value is reached in infinitely many positions, and it equals zero in all the positions.

Further when $a_t = 0$ we may expect that KSc is 0.5 for group C1. The reason for this is that C1 behaves as group B1 (modes B and C are same). When modes B and C are "well separated" then the supremum value is reached at two positions. Namely at modes B and C. The absolute mass differences in these modes are $|0.25 - 0.50| = 0.25$ and $|0.75 - 0.50| = 0.25$. Thus the value of KSc is 0.25.

For the rest of the groups, including group C1, the estimate is achieved using numerical integration. The outcome of these predictions is illustrated in Figure 7.7.

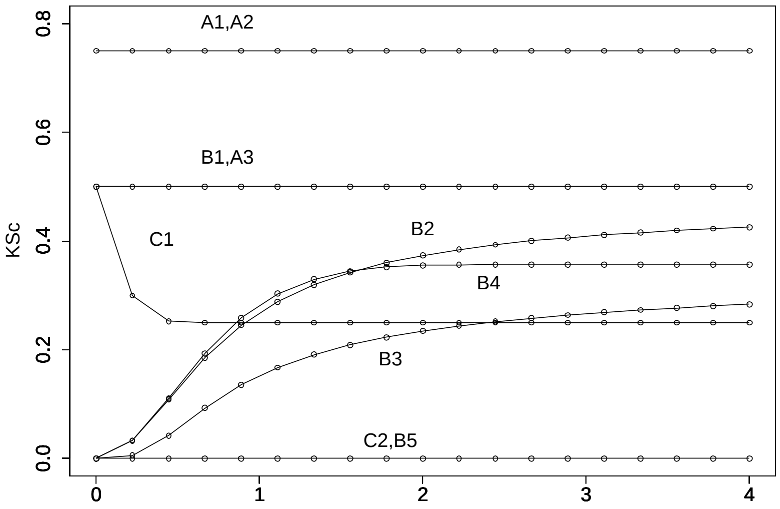


Figure 7.7: Theoretically predicted conditional Kolmogorov-Smirnov statistic as function of $2a_t$.

7.2.3 Simulation results at $x = 3.5$

We used $n_{sim} = 200$ repetitions in the simulations. Imputations were done 500 times at point $x = 3.5$ for each repetition. In the computation of the Kolmogorov-Smirnov distance the cumulative distribution function (cdf) of $Y^{mis}|x = 3.5$ was approximated by the empirical cdf, which was constructed by drawing a sample of size 2500 from $f(Y^{mis}|x = 3.5)$ once for each value of a_t.

The main difference between theoretical predictions and the simulated results is caused by an estimation error when using a finite sample of n observations. We can expect that this must be dominant in nonparametric methods like kernel regression and for some of the cell methods.

In addition two biases computed at $x = 3.5$ via simulations are

$$\mathbb{B}\text{ias}[\hat{\mu}^{imp}|x = 3.5] \;=\; \mathbb{E}[\hat{\mu}^{imp}|x = 3.5] - \mathbb{E}[Y^{mis}|x = 3.5]$$

and

$$\mathbb{B}\text{ias}[\hat{\tau}^{imp}|x = 3.5] \;=\; \mathbb{E}[\hat{\tau}^{imp}|x = 3.5] - \mathbb{V}\text{ar}[Y^{mis}|x = 3.5].$$

From a theoretical insight we may assume that $\mathbb{B}\text{ias}[\hat{\mu}^{imp}|x = 3.5]$ is close to zero, while $\mathbb{B}\text{ias}[\hat{\tau}^{imp}|x = 3.5]$ is close to zero only for the methods in classes B3, B5, C1, and C2. Note that C1 performs well, because the within modes (components B and C) variance of Y^{mis} at $x = 3.5$ is neglible (0.03) relative to the variance between the modes. These classes contain methods which are able to yield an approximately unbiased variance estimate at point $x = 3.5$. Conditional variance estimate (if any) for the other methods is worse than in the four mentioned classes.

The results of conditional simulations are summarized in Figures 7.8-7.11. The numerical values for KSc can be found from Table 7.4. For the sake of clarity the results of the KSc measures are divided into two Figures 7.8 and 7.9. As we can see, the results correspond quite well to our predictions in Figure 7.7.

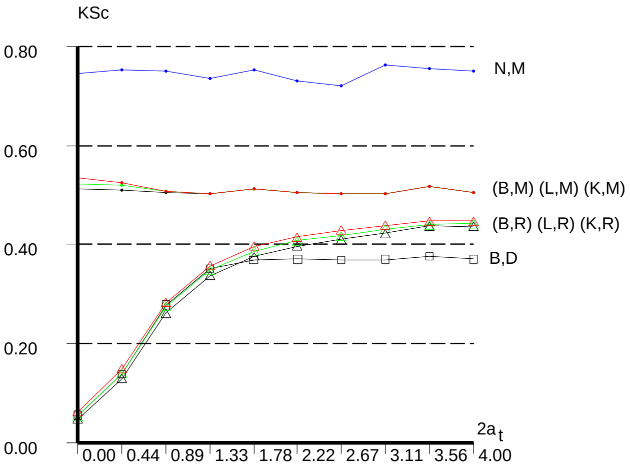


Figure 7.8: Simulated conditional K-S as function of $2a_t$ for baseline and regression methods. The dashed lines are included for an easier comparison with Figure 7.9.

The most notable differences to our predictions are found in classes C1 and C2. Methods in class C1 (TJ,M and TJ,M^s) yield a KSc value of close to 0.4, whereas we predicted the value to be 0.25. The reasons for this difference may be i) estimation errors due to finite sample size, and ii) the assumed imputation distribution differs from the actual one. In imputation, four best matching cells are searched at $x = 3.5$. If these four cells are not uniformly spread to components B and C (two cells in both components) then ”prior bias” may occur. Recall that in data generator components B and C are equally probable at $x = 3.5$. Methods in class C2 and B5 yield KSc values of 0.15-0.20 that are clearly higher than the expected value that is 0. A partial reason for this may be found in estimation errors due to small sample size, as for class C1. Further, ”prior bias” which was mentioned earlier is a possible reason for class C2 (which contains methods TJ,R, TJ,D and TJ,R^s).

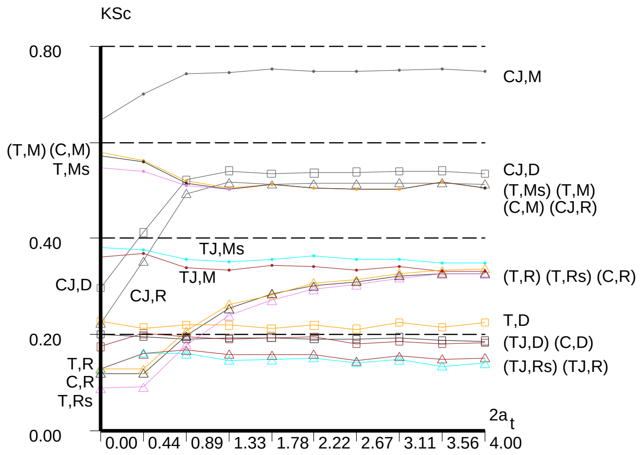


Figure 7.9: Simulated conditional K-S as function of $2a_t$ for cell methods.

The two biases $\mathbb{B}\text{ias}[\hat{\mu}^{imp}|x = 3.5]$ and $\mathbb{B}\text{ias}[\hat{\tau}^{imp}|x = 3.5]$ are shown in Figures 7.10 and 7.11.

Again the results are quite expected. The methods in categories B3, B5, C1 and C2 can estimate the variance of data quite well, while the methods in other classes are considerably worse.

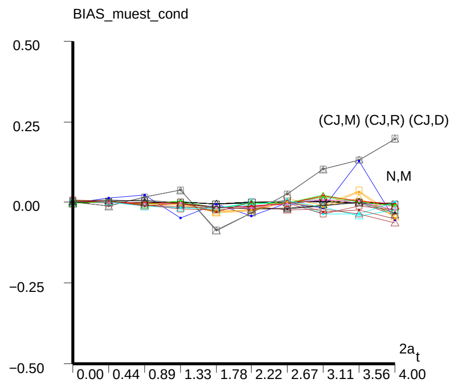


Figure 7.10: Conditional bias of $\hat{\mu}^{imp}|X = 3.5$ wrt. $\mathbb{E}[Y^{mis}|X = 3.5]$.

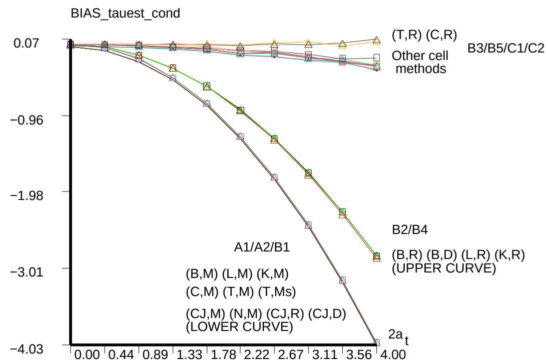


Figure 7.11: Conditional bias of $\hat{\tau}^{imp}|X = 3.5$ wrt. $\mathbb{V}\text{ar}[Y^{mis}|X = 3.5]$.

7.2.4 Preservation of marginals

As earlier, $n_{sim} = 200$ repetitions were used in the simulations. In the computation of the Kolmogorov-Smirnov distance the cumulative distribution function (cdf) of Y^{mis} was approximated by empirical cdf, which was constructed by drawing a sample of size 2500 from $f(Y^{mis})$ once for each a_t.

As we know from our basic decompositions in Chapter 3 the marginal distribution of Y^{imp} can be achieved via several ways. We can have for example a perfect model with optimal simulated randomness or a flexible (noisy) model with zero added noise. We may therefore assume that some models that behave badly at some conditional point x, can actually yield good results on a marginal level. This is especially true for very flexible models like 1-nearest neighbour imputation.

The results of the marginal measures for KS are shown in Figures 7.12 and 7.13. The numerical results are shown in table 7.4.

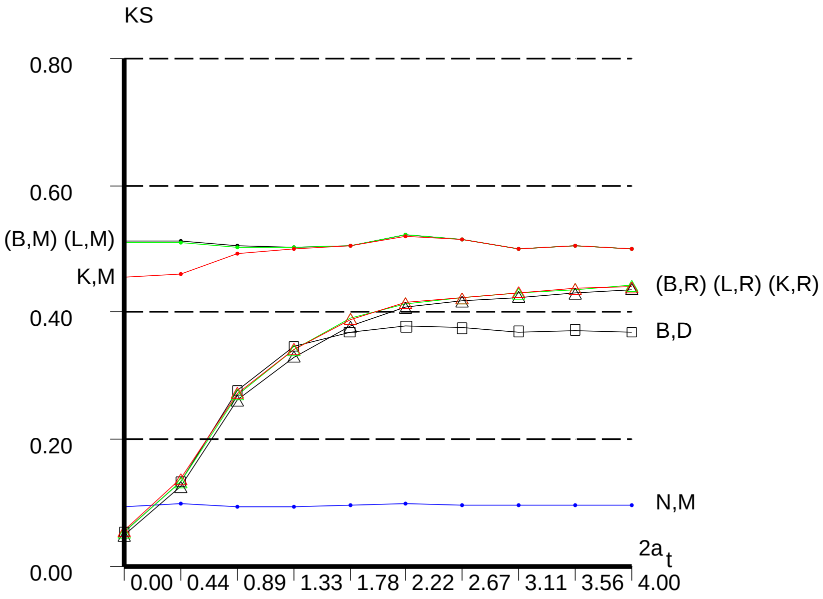


Figure 7.12: Simulated K-S as function of $2a_t$ for baseline and regression methods.

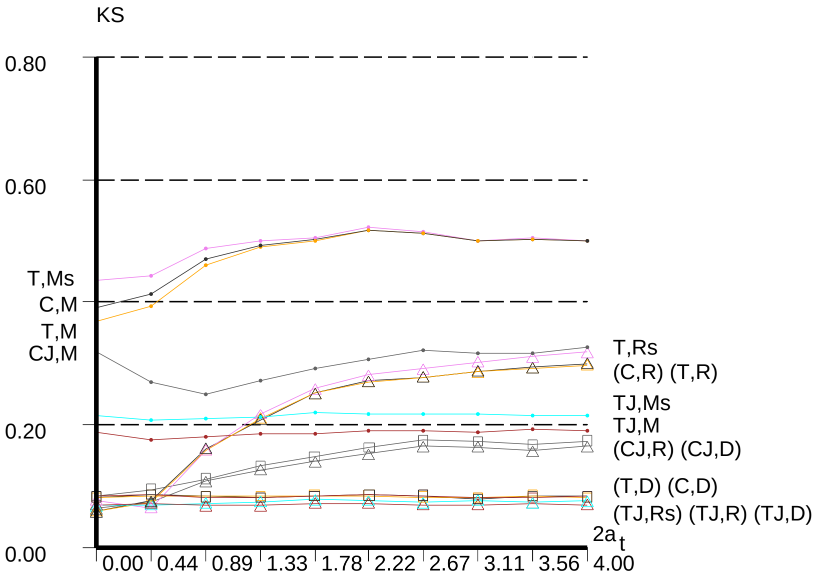


Figure 7.13: Simulated K-S as function of $2a_t$ for cell methods.

Method	**KS** $(2a_t = 0)$	**KS** $(2a_t = 20/9)$	**KS** $(2a_t = 4)$	**KSc** $(2a_t = 0)$	**KSc** $(2a_t = 20/9)$	**KSc** $(2a_t = 4)$
B,M	0.51	0.52	0.50	0.51	0.50	0.51
B,R	**0.05**	0.41	0.44	**0.05**	0.40	0.44
B,D	**0.05**	0.38	0.37	**0.05**	0.37	0.37
L,M	0.51	0.52	0.50	0.52	0.50	0.51
L,R	**0.05**	0.41	0.44	**0.05**	0.41	0.44
N,M	0.09	0.10	0.10	0.75	0.73	0.75
K,M	0.45	0.52	0.50	0.53	0.50	0.51
K,R	0.06	0.42	0.44	0.06	0.42	0.45
T,M	0.37	0.52	0.50	0.58	0.51	0.51
T,R	**0.06**	0.27	0.30	0.13	0.31	0.34
T,D	0.08	**0.08**	**0.08**	0.23	0.22	0.23
T,M^s	0.43	0.52	0.50	0.55	0.50	0.51
T,R^s	0.08	0.28	0.32	0.09	0.29	0.33
C,M	0.39	0.52	0.50	0.57	0.50	0.51
C,R	**0.06**	0.27	0.30	0.12	0.30	0.33
C,D	0.08	0.09	0.08	0.20	0.19	0.19
TJ,M	0.19	0.19	0.19	0.36	0.34	0.33
TJ,R	0.07	**0.07**	**0.07**	0.13	**0.16**	**0.15**
TJ,D	0.08	0.09	**0.08**	0.17	0.20	0.18
TJ,M^s	0.22	0.22	0.22	0.38	0.36	0.35
TJ,R^s	0.07	**0.08**	**0.08**	0.13	**0.15**	**0.14**
CJ,M	0.32	0.31	0.33	0.65	0.75	0.75
CJ,R	**0.06**	0.15	0.17	0.22	0.51	0.51
CJ,D	0.08	0.16	0.17	0.30	0.54	0.54

Table 7.4: Marginal (KS) and conditional (KSc) Kolmogorov-Smirnov statistics for $2a_t \in \{0, 20/9, 4\}$. The best results for each statistic are underlined, see Section 3.6.3 for details.

As predicted, the main difference in values of conditional and marginal K-S statistics is achieved for 1-nearest neighbour (N,M) and joint cell based on K-Means with the mean imputation strategy. The nearest neighbour method preserves marginal distribution well. The reason for this is that now predictions are done in various X positions, and given data both upper and lower mode distributions are roughly preserved.

Figures 7.14-7.15 depict biases of estimators for the first two moments of marginal distribution of Y. One can notice that estimators $\hat{\mu}^{comp}$ are roughly unbiased for all imputation methods. As in the conditional case some of the second moment estimators are clearly biased.

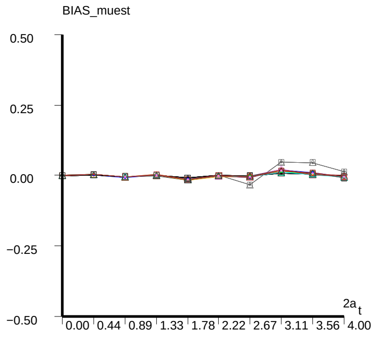


Figure 7.14: Marginal bias of $\hat{\mu}^{comp}$ with respect to $\mathbb{E}[Y]$ as function of $2a_t$.

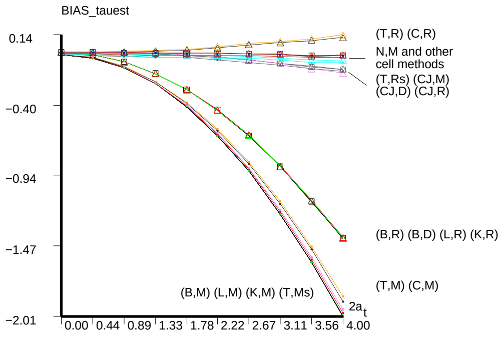


Figure 7.15: Marginal bias of $\hat{\tau}^{comp}$ wrt. $\mathbb{V}\mathrm{ar}[Y]$.

The methods from groups A3, B1, B2, and B4 are expected to perform poorly when considering marginal bias statistic $\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp}]$. The reason is that the variance of Y^{mis} is underestimated. In group B1 imputation variance is zero. Underestimation of the variance of Y^{mis} occurs also in groups A3, B2 and B4. As in the conditional biases case two quadratic curves corresponding to biased estimators may be seen in Figure 7.15.

7.2.5 About the role of conditionalization

We use $\mathsf{n}_{sim} = 200$ repetitions in our simulation. Further, the impact of conditionalisation is evaluated using 500 points, the distribution of which is described later.

The differences between the conditional results at $x = 3.5$ and marginal (integral) results over the full support of X bring out a question about the role of conditionalization. We like to know what happens to measures when conditionalizing a full marginal to one point at $x = 3.5$. To test this we simulate a multimodal case with $2a_t = 4$ (the largest separation between modes B and C) as a function of conditionalization level as described below.

The idea is to draw a picture that illustrates how conditionalization affects the Kolmogorov-Smirnov measure over all methods. For this purpose we define a new type of measure as

$$\mathrm{KS}_2 = \mathbb{E}\left[\sup_y \left|F_{Y^{mis}|x}(y|3.5-\alpha_t \le x \le 3.5+\alpha_t) - \hat{F}_{Y^{imp}}(y|3.5-\alpha_t \le x \le 3.5+\alpha_t)\right| \Big| \mathsf{n}\right],$$

where the range of X (imputation positions) is defined to be $x \in [3.5 - \alpha_t, 3.5 + \alpha_t]$. Imputation positions are uniformly distributed in this range, whereas distribution of data is as earlier. In our experiments we define

$$\alpha_t = \begin{cases} 1.24/2^{t-1} & \text{when } t = 1, \ldots, 9 \\ 0 & \text{when } t = 10. \end{cases}$$

In other words, when $t = 1$ KS_2 equals to the marginal Kolmogorov-Smirnov measure and when $t = 10$ KS_2 is the same as the conditional KS measure at $x = 3.5$.

Note that marginalisation is done with respect to uniform distribution (whereas it was done with respect to gaussian distribution earlier). By increasing t from 1 to 10 we move gradually from marginal K-S to conditional one.

The results are shown in Figure 7.16. As we can see, some methods are very sensitive to the level of conditionalisation. Especially sensitive are the most flexible models N,M and CJ,M. Also CJ,D and CJ,R are sensitive to conditionalisation. This is quite expected because as the range of imputation positions gets smaller it is likely that a given data set and model imputations using joint (Y, X) clustering with K-Means are done in either component B or C. For 1-nearest neighbour distribution of imputed values becomes more discrete as the range of imputation positions is decreased. Eventually distribution collapses to one point (Y value of the observation nearest to the position $x = 3.5$).

It is also apparent that those methods that perform well on conditional measures also perform well on marginal measures, but the opposite is not true. Thus we may conclude that conditional measures are better for the evaluation of overall performance, and that joint $(Y, \boldsymbol{X})$ clustering with TS-SOM seems to be optimal for this case.

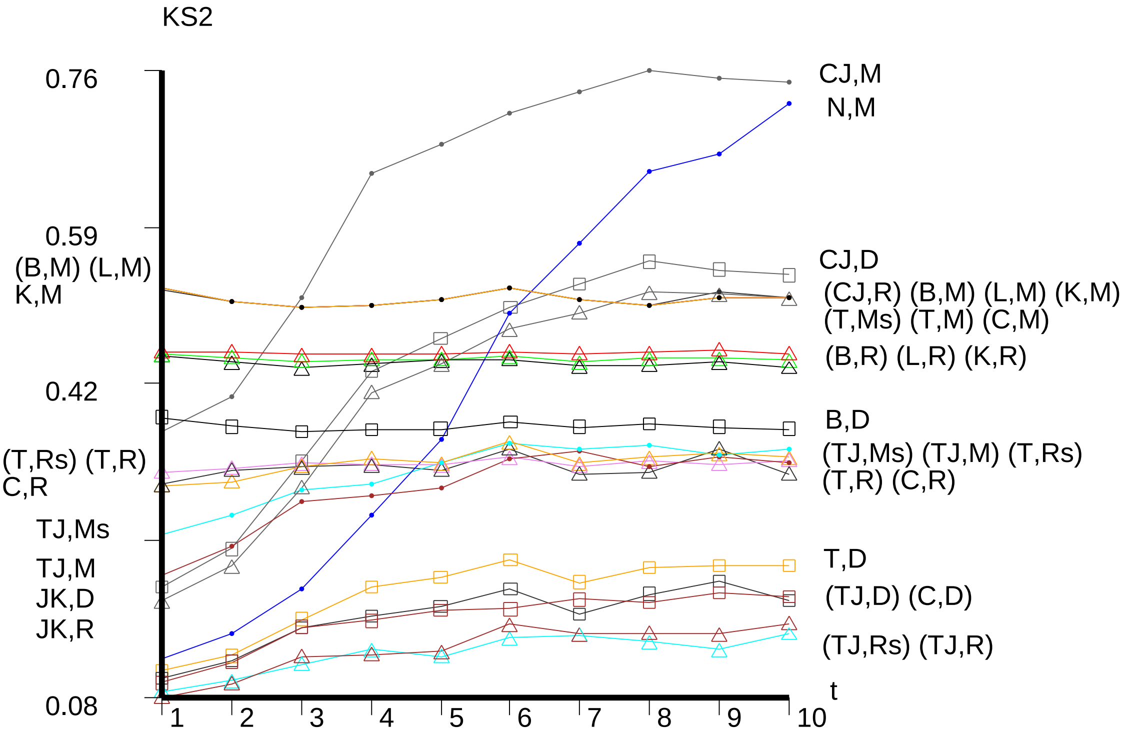


Figure 7.16: The role of increasing conditionalisation from marginal KS to conditional KSc at 3.5

7.3 Case 3: classification with multivariate $\boldsymbol{X}$

Until now we have studied cases with one response variable Y and one covariate X. This supports easy interpretations and allows some theoretical analyses to be coupled with the work. But in the real-world, data is seldom as easy as it is in our simulations. As a step towards more practical studies in Chapters 8 and 9, we shall now consider a case with an increasing dimension of the covariate. This complicates the problem considerably. Therefore we shall limit the study to a simple linear classification problem under MCAR missingness. Thus we know that linear regression should be the best method for the task.

Because we want to test the imputation performance of our methods, we complicate the problem slightly by adding Gaussian noise to class information $Y_{|\mathbf{x}}$, which otherwise is coded using two crips values: one and zero. Thus our optimal classifier returns correct class in $\mathbf{x} \in \mathbb{R}^p$ space and adds just the right type of noise to predicted class identifier $\hat{Y}$.

7.3.1 Data generator

Our problem is to predict noisy class information Y using a multivariate Gaussian covariate $\mathbf{x} \in \mathbb{R}^p$. The classes of Y are defined by a threshold equation

$$Y = H(\mathbf{z}^T \boldsymbol{X}) + \epsilon,$$

where $H(a)$ is the Heaviside step function

$$H(\mathbf{z}^T \boldsymbol{X}) = \begin{cases} 1 & \text{if } \mathbf{z}^T \boldsymbol{X} > 0 \\ 0 & \text{otherwise} \end{cases}$$

and the noise follows Gaussian distribution

$$\epsilon \sim N(0, 0.07).$$

We further assume that all $X_i, i = 1, \ldots, p$ are identical $X_i \sim N(0,1)$ implying that $\boldsymbol{X} \sim N\big([0, \ldots, 0]^T, \mathrm{diag}(1, \ldots, 1)\big)$. Density of X_i is illustrated in Figure 7.17.

The parameter vector $\mathbf{z}$ defines a half-space where $\mathbb{E}[Y|\mathbf{x}]$=1, while on the other half $\mathbb{E}[Y|\mathbf{x}] = 0$. As all inputs are identically distributed $X_i \sim N(0,1)$, the class boundary goes through the origin of the p-dimensional Gaussian covariate space $\boldsymbol{X}$. In the direction $\mathbf{z}$ the problem becomes essentially one-dimensional, as depicted in Figure 7.18.

For a given data set the direction vector $\mathbf{z}$ is fixed, but in our repeated simulations $\mathbf{z}$ is taken randomly for each experiment from $f_{\boldsymbol{X}}(\mathbf{x})$. By this we try to ensure that no direction $\mathbf{z}$ is "accidentially" favoured by some of the imputation methods. For 2+1-dimensional case example data is depicted in Figure 7.19. Two-dimensional projections are included for clarity.

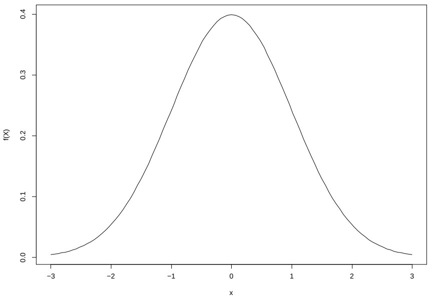


Figure 7.17: Distribution of X_i. This is also the distribution of Z_i.

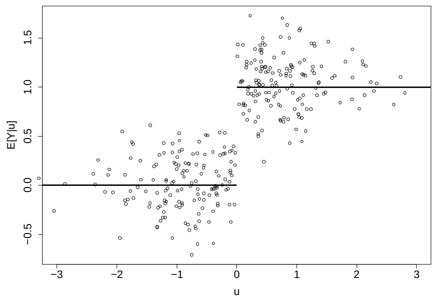


Figure 7.18: Conditional mean and a random sample of Y as function of projection $u = \mathbf{x}^T\mathbf{z}$.

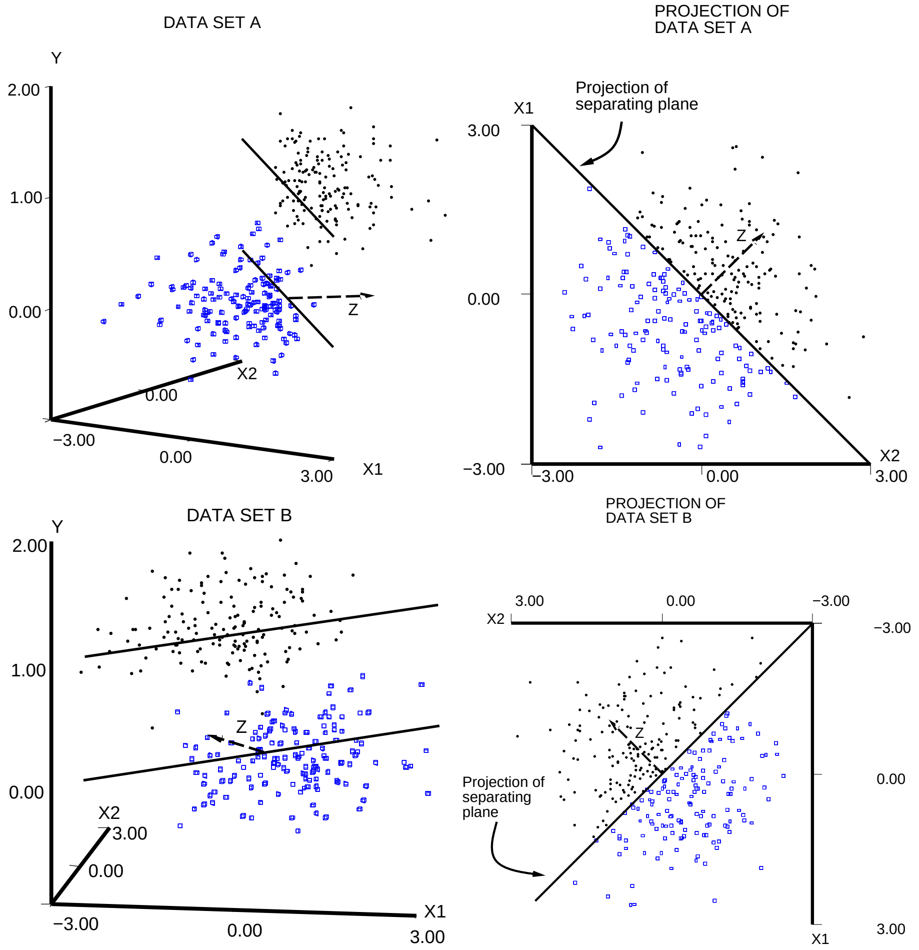


Figure 7.19: Two randomly selected data sets, and their 2D projections, with two-dimensional Gaussian covariate $\boldsymbol{X}$. For simplicity, only training observations are shown. Black dots denote class y=1, wheras class y=0 is marked by cubes/squares.

In the experiments the dimension p of the covariate $\mathbf{x} \in \mathbb{R}^p$ is varied from 1 to 11, while the number of observations n is kept constant $\mathsf{n} = 600$. The missingness is selected with Bernoulli probability $\frac{1}{2}$, implying that $\mathsf{N}^{mis} \sim Bin(600, 0.5)$ and $\mathsf{N}^{obs} = \mathsf{n} - \mathsf{N}^{mis}$.

The evaluation of imputation performance is done in terms of integrated mean squared error

$$\text{MISE}(\hat{Y}^{imp}) = \mathbb{E}_{\boldsymbol{Z}}\mathbb{E}[(\hat{Y}^{imp}_{|\mathbf{x},\mathbf{z}} - Y^{mis}_{|\mathbf{x},\mathbf{z}})^2|\mathbf{z}],$$

which can further be written in terms of bias and variance (and expected variability of target, irreducible term, which is constant 0.07 and is omitted here)

$$\begin{aligned}\mathbb{B}\text{ias}^2[\hat{Y}^{imp}] &= \mathbb{E}_{\boldsymbol{Z}}\mathbb{E}\Big[\big(\mathbb{E}[\hat{Y}^{imp}|\mathbf{x},\mathbf{z}] - H(\mathbf{x}^T\mathbf{z})\big)^2|\mathbf{z}\Big], \text{ and} \\ \mathbb{V}\text{ar}[\hat{Y}^{imp}] &= \mathbb{E}_{\boldsymbol{Z}}\mathbb{E}\Big[\mathbb{V}\text{ar}[\hat{Y}^{imp}|\mathbf{x},\mathbf{z}]|\mathbf{z}\Big].\end{aligned}$$

All the methods are used in a rather straightforward way; we try to predict Y^{mis} using observed covariate $\mathbf{x}^{mis}$. For some methods special tuning is required: for kernel regression the bandwidth was selected as $\lambda = 2(\mathsf{N}^{mis})^{-1/(4+p)}$, for TS-SOM a two-dimensional latent neighborhood was used with 64 cells, and 32 cells were used with K-Means clustering.

7.3.2 Theoretical insight

Due to the simplicity of the problem we can do some obvious predictions about the performance of the imputation methods. This is summarized in the following list

Baseline methods are strongly biased, making an expected squared bias

$$\mathbb{B}\text{ias}^2[\hat{Y}^{imp,B}] = \mathbb{E}_{\boldsymbol{Z}}\Big[(\frac{1}{2})^2\Big] = 0.25.$$

The variance is not affected by the dimension of $\boldsymbol{X}$ and therefore it is close to zero. Formally

$$\mathbb{V}\text{ar}[\hat{Y}^{imp,B}] = \mathbb{V}\text{ar}[\hat{\mu}^{obs}] \overset{Taylor}{\approx} \frac{\tau^{*obs}}{\mathsf{n}(1-p^*)} = 0.32/300 = 0.001.$$

and

$$\mathbb{V}\text{ar}[\hat{Y}^{imp,B,S}] = \mathbb{V}\text{ar}[\hat{Y}^{imp,B}] + \mathbb{E}[\hat{\tau}^{obs}] \approx 0.321,$$

where strategy S $\in$ { R, D }.

Linear regression is expected to lead a considerably lower expected squared bias than the baseline methods. Prediction variance grows quadratically as the distance from zero point is increased.

Cell methods are the most sensitive for increasing $\boldsymbol{X}$-dimension, because they try to spread cells over all data. Thus with fixed n and increasing dimension data more spread ensues. On average, there will be a larger X-space, for each cell, which leads to a larger estimation variance.

7.3.3 Simulation results

In simulation we do 10 draws of direction vector $\boldsymbol{Z}$ for each dimension of covariate $\boldsymbol{X}$. Given direction $\mathbf{z}$, 25 repetitions of data samplings and imputations are done in 1600 $\mathbf{x}$ positions. These positions are drawn once for each $\mathbf{z}$ and the dimension of $\boldsymbol{X}$. Thus over the repetitions the positions are fixed. The final results are computed by averaging the results for 10 directions of $\boldsymbol{Z}$.

The results are depicted in Figures 7.20-7.22 for expected squared bias, variance, and mean squared error. Irreducible error, which is constant 0.07 for all the methods, is excluded from the expected mean squared error curves. A minor simulation inaccuracy is visible in the squared bias results for random strategies. As an example, the results for baseline mean and random strategies should be same. This is caused by the higher variability of bias estimate for random strategies. Simulation inaccuracy is not major issue here, because it is low, roughly 0.01 squared bias units.

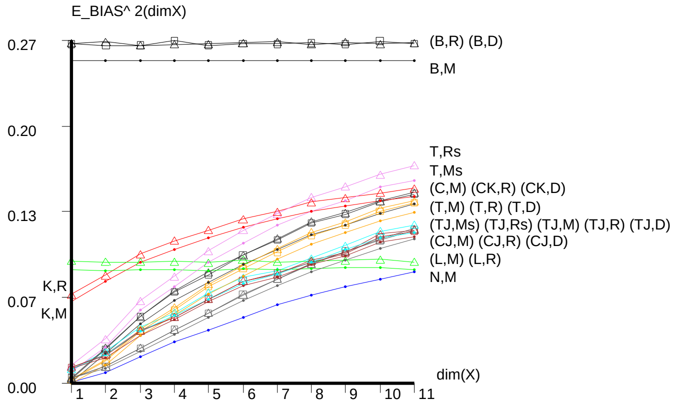


Figure 7.20: Expected squared bias as function of the dimension of $\boldsymbol{X}$.

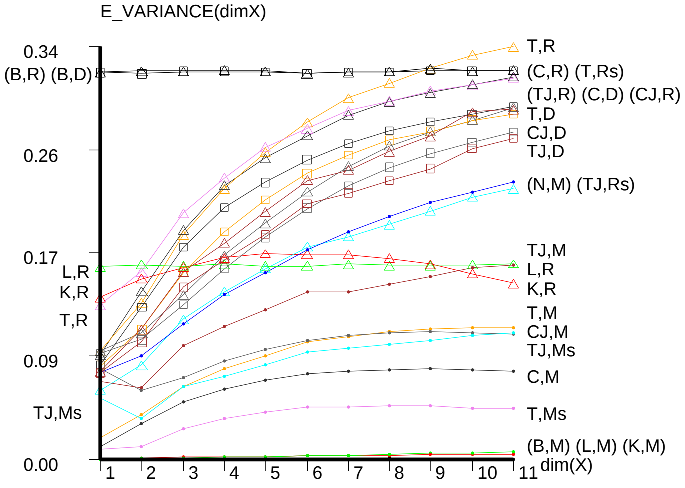


Figure 7.21: Expected prediction variance.

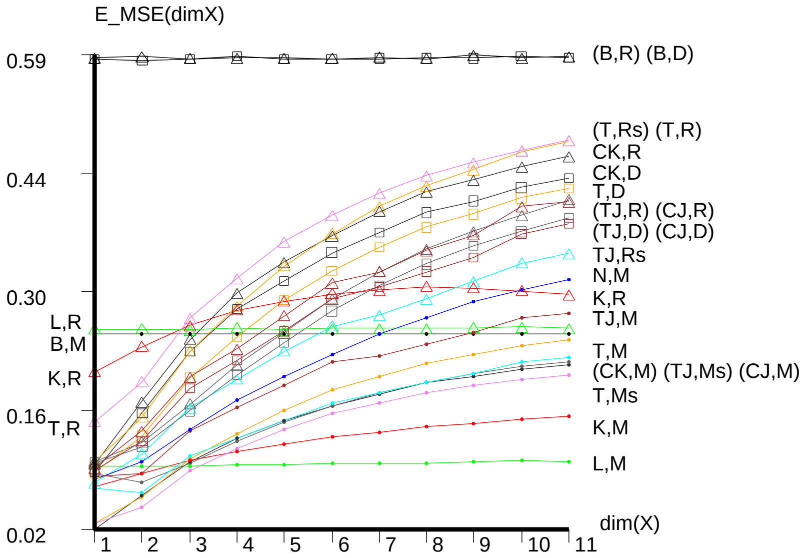


Figure 7.22: Expected mean squared error as function of the dimension of $\boldsymbol{X}$. Irreducible error, which is 0.07 for all methods, is excluded from the curves.

The following conclusions can be made from the results of the simulations:

Baseline methods perform the worst with respect to expectation of squared bias. Further, the baseline with random or donor strategy yields the highest variance at least up to dimensions 8-9 of $\boldsymbol{X}$. Squared bias is roughly 0.25 for all the baseline methods. Further, the mean imputation strategy yields an approximately expected variance of 0.001, whereas the random and donor strategies yield a variance of 0.32. Thus, theoretical results are verified.

Linear regression methods yield a low variance and squared bias. Further, both of the quantities are not affected, at least significantly, by the dimension of $\boldsymbol{X}$.

Nearest neighbour is able to yield the lowest squared bias which grows as the dimension is increased. However, the method is penalized by its high variability. Expected variance also increases as a function of the dimension of covariate.

Cell methods yield a smaller squared bias than linear regression methods at least up to dimension 4 of $\boldsymbol{X}$. Joint $(Y, \boldsymbol{X})$ clustering cell methods are able to do this to dimensions 7-8 of $\boldsymbol{X}$. Among cell methods, variance grows the lowest for the covariate TS-SOM method with smoothing.

The results for methods other than baseline and linear regression are impacted by the dimension of covariate.

7.4 A study of computational properties under Case 3

Here we shall study the computational time complexity of imputation methods. The effect of sample size, dimension of covariate $\boldsymbol{X}$, and the amount of missingness on computational times are analysed. We measure the average model training time, imputation time, and total time. For simplicity, we repeat the experiments of Case 3 with some exceptions. Parametrizations for imputation methods are the same, except that now 64 cells are used with K-Means based methods. The processor used in these simulations is Intel Pentium 4 Prescott, which runs at 3.0 GHz, with 512 megabytes of DDR RAM (with a frequency of 200 MHz and in a dual-channel mode). The operating system is Linux.

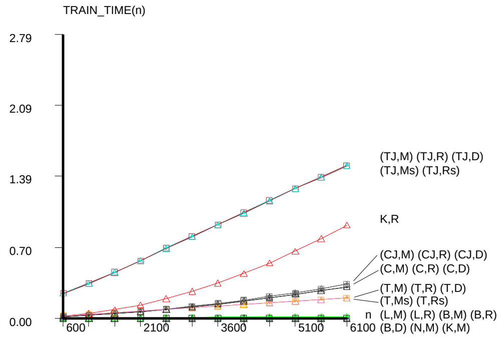
Three simulation studies were done to investigate the effects. In the first study the sample size is increased from 600 to 6100, while the dimension of $\boldsymbol{X}$ and the missingness probability are kept fixed. In the second study the dimension of $\boldsymbol{X}$ is increased from 1 to 25, while the sample size and missingness probability are fixed. In the third study the missingness probability is increased from 5% to 85%, and the sample size and the dimension of $\boldsymbol{X}$ are fixed.

7.4.1 Impact of the increase of sample size

In the experiment sample size is linearly increased from 600 to 6100 in 11 steps. The dimension of $\boldsymbol{X}$ is fixed to 11, and the missingness probability is 50%. In the simulation we use $\mathsf{n}_{sim} = 15$ repetitions. The curves for imputation model training, imputation, and total time are depicted in Figures 7.23-7.25. For clarity, the methods for which the imputation time results are between the results for baseline and kernel regression methods have not been marked.

From the results we can see that the training time is the highest for the joint (Y, X) clustering methods based on TS-SOM and for kernel regression with random imputation strategy (for which the estimation of residual variance is slow). Imputation model training time is zero for 1-nearest neighbour and kernel regression with mean imputation strategy (N,M and K,M). Imputation is slowest for non-parametric regression methods. Further, the corresponding curves are nonlinear and they grow fast.

The total computational cost curves are roughly linear for all the methods except for the nearest neighbour and kernel regression methods. Total computational cost is highest for the joint $Y, \boldsymbol{X}$ clustering cell methods for sample size less than 3100. The shape of the cost curves is linear for the cell methods. The nonparametric kernel and nearest neighbour regression methods have nonlinear cost curves. From the results one sees that from sample size 3600 onwards the cost for nearest neighbour is higher than for the fastest $Y, \boldsymbol{X}$ cell methods (TS-SOM cell method with mean or random strategy and smoothing).



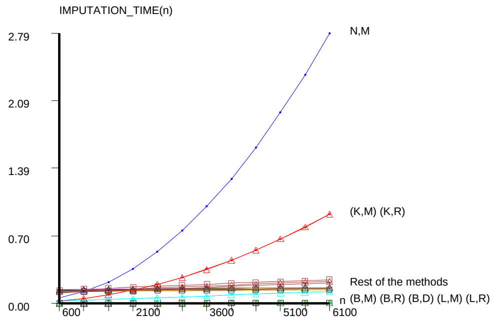


Figure 7.23: Training time as function of sample size. Time is measured in seconds.

Figure 7.24: Imputation time (seconds).

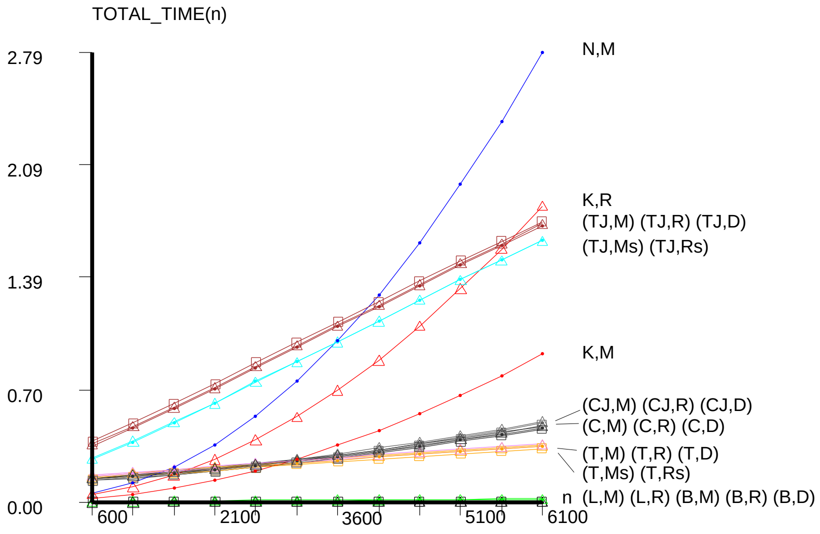


Figure 7.25: Total time (seconds).

7.4.2 Impact of increase of dimension of covariate

In this example the dimension of $\boldsymbol{X}$ is linearly increased from 1 to 25 in 8 steps. Sample size is fixed to 1100, and missingness probability is 50%. The number of repetitions is $\mathsf{n}_{sim} = 15$ in this simulation. Figures 7.26-7.28 depict computational times as a function of the dimension of $\boldsymbol{X}$. The baseline methods have the lowest total computational times, and applying linear methods is the second fastest. From Figure 7.28 one can notice that the total times for the $Y, \boldsymbol{X}$ TS-SOM cell methods are the highest. This is due to the slowness of model training. The increase for cell methods is only linear with respect to the dimension of $\boldsymbol{X}$. Further, note that the joint $Y, \boldsymbol{X}$ clustering TS-SOM cell methods utilize all available observed data, whereas the other methods utilize only complete observations. Therefore the TS-SOM joint cell methods use twice as much covariate data as do other methods on expectation.

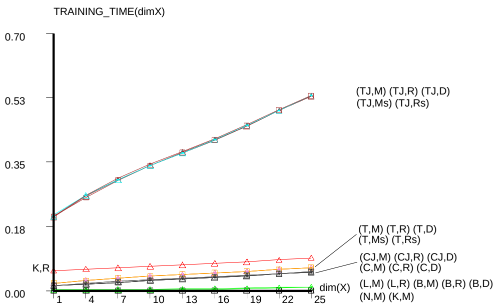


Figure 7.26: Training time as function of the dimension of $\boldsymbol{X}$. Time is measured in seconds.

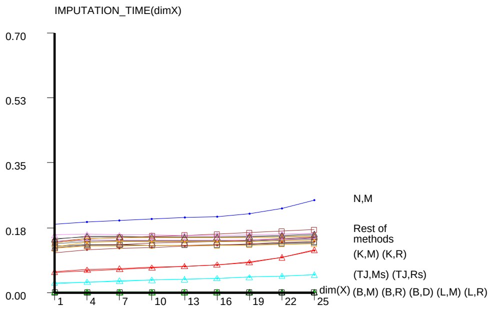


Figure 7.27: Imputation time (seconds).

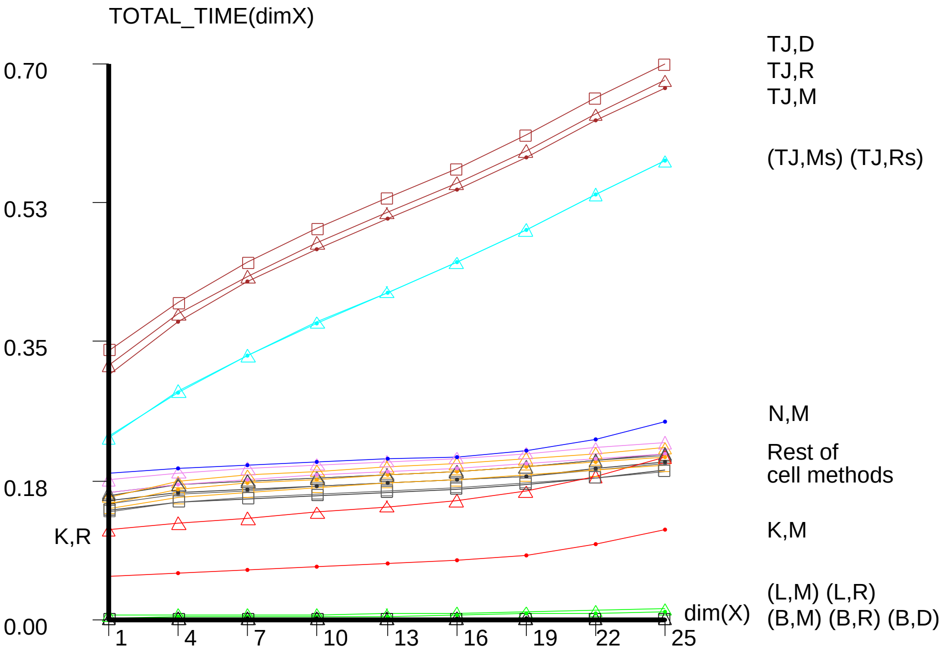


Figure 7.28: Total time (seconds).

7.4.3 Impact of increase of missingness probability

Here the missingness probability is linearly increased from 5% to 85% in 16 steps. The dimension of $\boldsymbol{X}$ is fixed to 11 and sample size is 1100. Fifteen, $\mathsf{n}_{sim} = 15$, repetitions are used in this simulation. It is expected that the training time decreases as the missingness probability is increased (because the size of training data is smaller on average). However, joint $(Y, \boldsymbol{X})$ clustering with TS-SOM is an exception to this, as an incomplete training data algorithm is used. For all TJ methods the training time is expected to grow as the missingness probability is increased. Figures 7.29-7.31 depict computational time curves as a function of missingness probability.

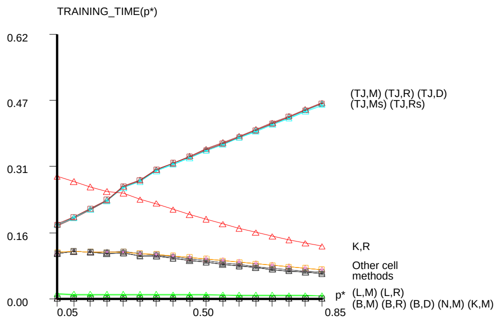


Figure 7.29: Training time as function of missingness probability p^*. Time is measured in seconds.

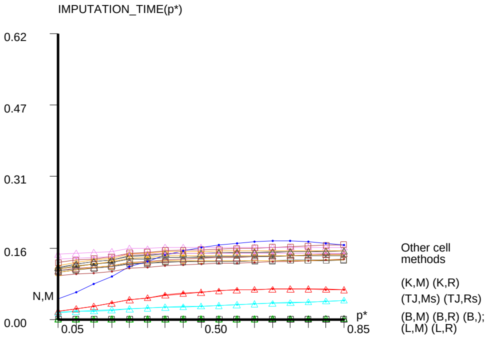


Figure 7.30: Imputation time (seconds).

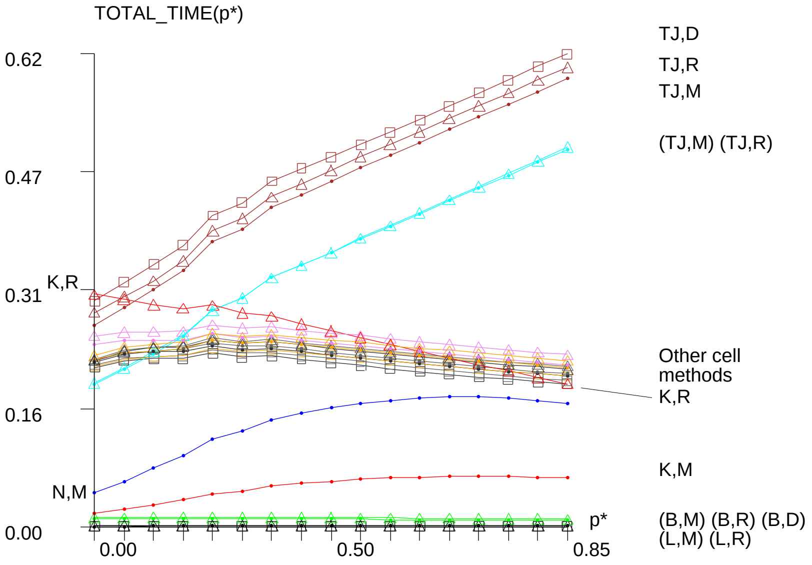


Figure 7.31: Total time (seconds).

The following observations are made from the results:

TJ methods: total computational costs for the joint $Y, \boldsymbol{X}$ clustering cell methods based on the TS-SOM algorithm grow linearly as the missingness probability is increased. The reason for the increase is that the corresponding training algorithm is costlier for incomplete observations than for complete observations.

Other cell methods: all other than the TJ cell methods get faster as missingness probability is increased. The reason for this is that the training data set, which is a fully observed part of the whole data, gets smaller on expectation, while training of models becomes faster.

Kernel regression with random strategy: estimation of residual variance becomes faster as missingness probability is increased, because the size of training data is on average smaller.

Finally, we would like to mention the following notes concerning all the three previous experiments:

Nearest neighbour: our implementation of k-nearest neighbour could be optimized for the value of k=1. Therefore the imputation times could be reduced close to the times for kernel regression method K,M.

K-Means clustering: no repetitions of model training, to prevent bad local solutions, is done. Thus in practical applications, with a data set of similar dimensions, the training times are larger by some factor (say 10-25 for example).

Non-smoothed joint $(Y, \boldsymbol{X})$ clustering TS-SOM cell methods: could be made as fast as the smoothed TJ methods.

7.5 Summary

In this Chapter we have studied distributional and unit level properties of imputation methods, as well as the impact of the dimension of covariates to imputation results. Finally, the methods were evaluated with respect to computational properties also. Next we summarize our findings.

It was shown that at a distributional level many methods are able to preserve the first moment under transition from MCAR to MAR mechanism. In case of multimodal distribution and NMAR mechanism some of the methods were able to preserve marginal distribution, but only a few of them were able to preserve conditional distribution. Of a particular note are the joint $Y, \boldsymbol{X}$ clustering methods based on TS-SOM, which were able to deliver a good performance in the multimodal case and in the preservation of conditional distribution.

At a unit level linear regression, kernel regression, and cell methods with mean imputation strategy performed quite well. Nearest neighbour regression was penalized by its high variance. The methods with random and donor strategies were, as expected, inferior in most of our experiments. Linear regression showed best robustness against the increase of data dimensionality with a low expected mean squared error. This is expected as our data model is well approximated by linear regression. The nearest neighbour method is able to produce quite good results at distributional and unit level studies. However, multimodal conditional distribution is not well preserved.

Computational properties included training time, imputation time, and total time. Training times increase roughly linearly for the joint $(Y, \boldsymbol{X})$ clustering cell methods, and for the covariate cell methods based on K-Means, as a function of data dimensions. The nearest neighbour and kernel regression methods showed rapid increase of imputation and total computational times as a function of sample size. High computational requirements may render the nearest neighbour or kernel regression methods to be unusable in practice. The use of cell methods is to be preferred with large data sets. Nonparametric regression methods may be modified for faster performance using a some kind of discretization, i.e., binned kernel regression (see for example [41]). One may except this modification to increase squared bias at least.

Chapter 8

Simplified case study: UK survey of small and medium-sized enterprises

This chapter describes a case study that is similar but not exactly the same as the annual business survey (ABI)[1] that was used in the Euredit project [11]. The motivation of the current study is to test our refined methodology using well-known data, and to overcome some shortcomings of the previous study. The main differences between the current case and the Euredit experiments are the following:

- We have a well-defined MAR type missingness generator, while the missingness in EurEdit project was due to MCAR.
- We compute "expected" results using repeated samples from some data generator, while only one sample was imputed in Euredit. Thus we can avoid "accidentially" good or bad results.
- We now have more experience about the studied survey and imputation. During Euredit we were still novices with the problem, which contributed to many of our shortcomings in the use of our technology.
- The current study is not as "objective" as Euredit, because we do know what the true values are. But, of course, we try to be as objective as we can, and no knowledge about the true values is used to "boost" the performance of the proposed methodology.
- In this experiment, only a small "clean" subset of data is used, while in Euredit the data was in a "real-world" format including special values and other nuisances. An experiment that is more close to a real-world case will be given in Chapter 9.

[1] See www.statistics.gov.uk/abi/ for general information on UK ABI data sets, referenced 02.05.2007

The data set in this case is based on United Kingdom Survey of Small and Medium-sized Enterprises' Finances 2004, produced by Fraser[2], sponsored by Bank of England, and supplied by the UK Data Archive. The data are copyright of Bank of England and University of Warwick. Here the first edition of data set [26], which is dated 1st February 2004, is used. The following statement is required for the use of this data set:

The original data creators, depositors or copyright holders, the funders of the Data Collections and the UK Data Archive bear no responsibility for further analysis or interpretation done in this thesis.

Log transformed turnover for ending accounting year is a variable of interest. Multiple covariates are used for predicting turnover. A description of the variables used is given a bit later. As explained later, a subset of data is used in the experiments as a clean "true data set". In the experiments we repeat samples from "true" data and generate missingness using a MAR type mechanism. For each incomplete sample the missing values are imputed and the result is evaluated using evaluation statistics. The process is repeated as long as it takes to eliminate biases due to the setup of the experiment.

8.1 Description of the dataset

There are 699 variables and 2500 observations in the dataset. Most of the variables are answers to survey questions. Different questions were asked from different kinds of enterprises (i.e., start-ups versus non start-ups), and many observations are incomplete. Due to the nature of our experiment we shall not impute "naturally" missing values. Instead we pick a complete subset of data and apply our own missingness generator to it. In this example we shall also omit all special values such as "don't know" and "not asked". In addition, zero values are omitted. See Appendix A8.2 for details on the construction of the subset. As a result we are left with a very small clean data set of 678 observations and eight variables, including sampling weights.

The variables of our data are summarized in Table 8.1. Missingness is applied to variable O2 (turnover) and six other variables are used as covariates. The covariates include information about turnover, number of employees and age of business, balance sheet, and accounts.

[2] University of Warwick. Warwick Business School. Centre for Small and Medium-sized Enterprises

Variable	Description
Income and profits	
O2	Turnover for ending year (variable of interest) [in pounds].
Screening for eligibility	
S4_SIZE	Number of employees [from 2 to 240].
AGE	Age of business, derived as 2004-S11 for non-special values, where S11 is the year when business was established.
Balance sheet information	
P1	Total amount of assets held [in pounds].
P2	Total amount of liabilities owned [in pounds].
Use of current accounts	
E3D	Approximate amount of money in current business or personal account at present.
E5	Total monthly or quarterly bank charges on the account [charges for banking services eg including writing or paying in cheques, making BACS (Bankers Automated Clearing System) payments but not interest or charges for any loans or overdrafts].
Sampling information	
PWEIGHT	Probability weights [approximate range is from 1.15 to 6288.81].

Table 8.1: List of the variables used and their descriptions.

The variables of the original data are preprocessed for our purposes as follows. In the original data there is variable S11 that contains information about the age of the business. We have changed it into a more direct form AGE=2004-S11. Then variables O2 (turnover), S4_SIZE, AGE, P1, P2, E3D and E5 are log transformed using formula

$$x' = \ln(x),$$

where $\ln(\cdot)$ denotes natural logarithm.

Relationships between the log-transformed variables are seen from a scatter matrix which is depicted in Figure 8.1. From the scatter plot matrix one can notice that turnover and sampling weights are negatively correlated. The scatter plot also reveals that turnover for the ending year (O2) and number of employees (S4_SIZE) are highly linearly dependent. There are also dependencies between turnover and other covariates.

Log-transformed turnover with a Gaussian fit is shown in Figure 8.2, indicating that the log-transformed turnover is close to Gaussian distributed. The distribution of sampling weights is visualized in Figure 8.3. As we can see, the number of small enterprises is underpresented in the sample. In this experiment sampling weights are used in the computation of evaluation statistics so that the computed imputation biases and variances can be interpreted in terms of real-world importance. The use of sampling weights in imputation methodology depends on the method. In our experience, the best practice is to use sampling weights with baseline methods, while methods that utilize covariates often do better without weighting. More details about this are given in Section 8.2.

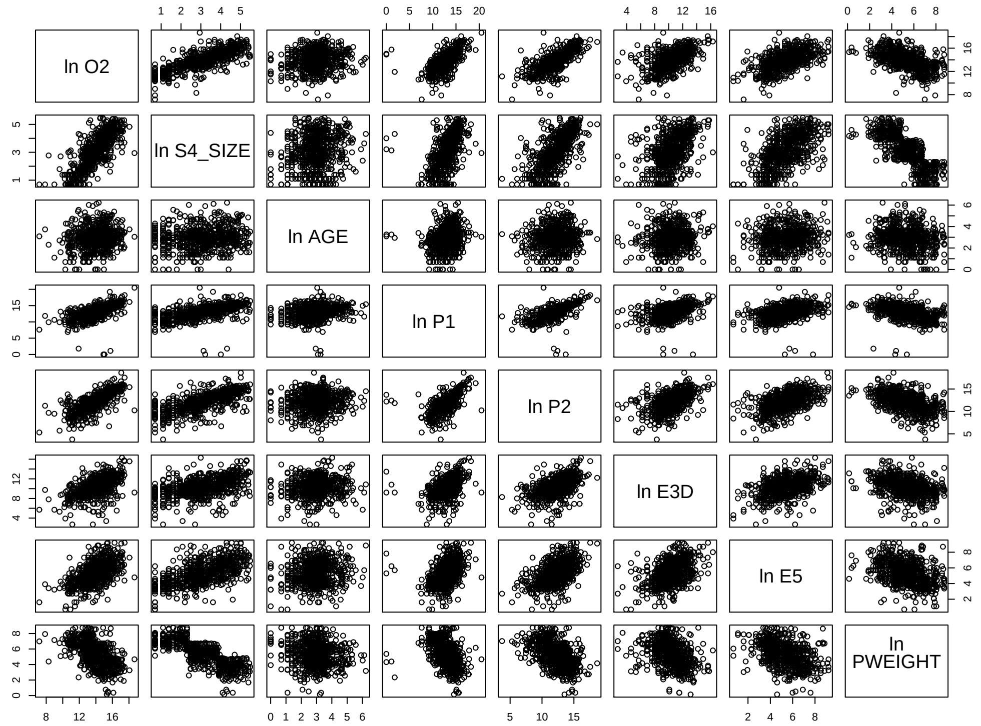


Figure 8.1: Scatter matrix of unweighted log-transformed variables. As an exception, weights are also in log scale here.

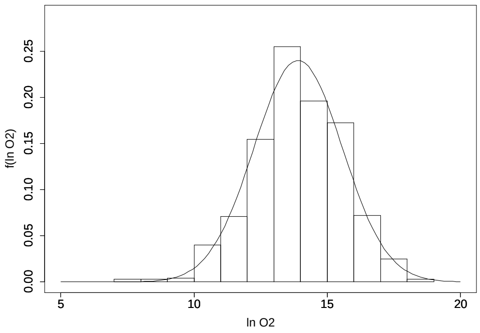


Figure 8.2: Histogram and Gaussian fit for logarithm of unweighted turnover (variable O2).

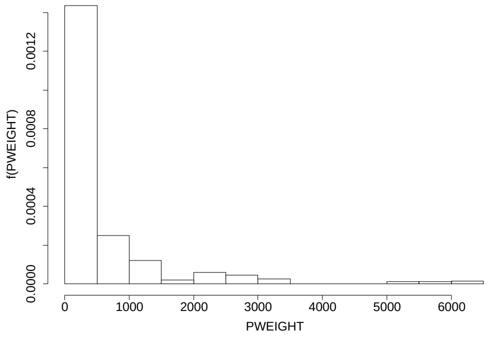


Figure 8.3: Histogram of probability weights (variable PWEIGHT).

8.1.1 Data generator and simulation technique

Our data generator is based on a previously constructed finite population and on synthetic missingness. Nonresponse mechanism of population units is defined as

$$\Pr(\text{"Nonresponse for O2"}|\text{S4\_SIZE} = x) = \frac{1}{\sqrt{x}},$$

where variable S4_SIZE is in the original scale. This is a missing at random (MAR) mechanism. Figure 8.4 depicts the curve for a probability that turnover (O2) is missing as the function of the number of employees (S4_SIZE). The distribution of the number of employees is shown in Figure 8.5. The reason for the use of MAR mechanism is that it is more realistic than MCAR. In business inquiries it is quite typical that smaller enterprises are likely to answer less often than larger enterprises, due to lack of resources.

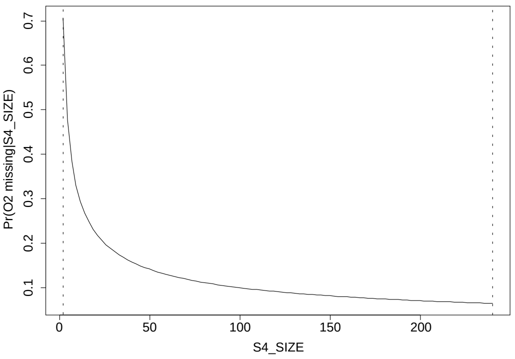


Figure 8.4: Probability of missingness of turnover (O2) as a function of employee (S4_SIZE).

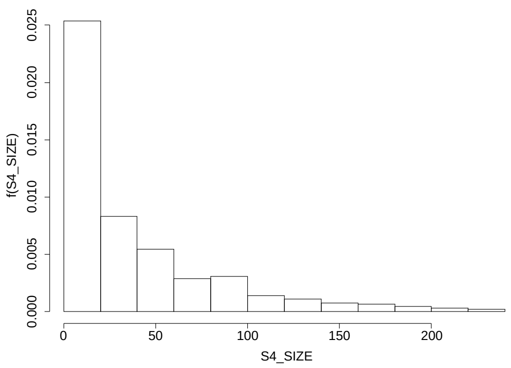


Figure 8.5: Distribution of unweighted number of employees (S4_SIZE) before log transform.

Simple random sampling of size $\mathsf{n} = 600$ without replacements is used as the simulation technique. This reflects sampling variability. A sampling fraction is approximately 0.88. Probability weights are not used when drawing a sample. However, weights are included in each drawn sample. A simulation technique with the above defined nonresponse probability function yields an expected missingness percentage of approximately 25% in random samples.

The use of sampling without replacements is the only possibility, because we do not want copies of true values in the simulation sample for which missingness is generated afterwards. However, we must be aware of possible problems due to the small size of our original data.

8.1.2 Analysis of incomplete sample data

In this section an incomplete random sample from our data generator is studied similarly as we would do with any real-world data. Analyses are done for log transformed variables so that the assumptions of imputation models hold better. First, correlation coefficients are estimated to reveal linear dependencies. Dependencies within the sme (small or medium) enterprises group and the large enterprises group are also briefly analysed. This is followed by the visualizations using the self-organizing map.

Correlation coefficients between turnover and covariates are computed from the complete part of the random sample. We divide the enterprises rather arbitrarily into two classes: enterprises with turnover (O2) less or equal to 4.7 million pounds belong to the small and medium (sme) class, while the rest of enterprises belong to the large enterprise class. In the example sample the complete part includes 445 enterprises of which 344 belong to the sme class and 101 to the large enterprise class. Estimated correlations between response and covariates are depicted in Table 8.2. Sampling weights have not been used. The correlation coefficients reveal that especially for small and medium enterprises there seems to be a somewhat strong positive correlation between turnover and number of employees (S4_SIZE) and turnover and liabilities (P2). There are also quite strong linear dependencies between turnover and other covariates excluding age of enterprise. For larger enterprises these correlations are weaker.

	ln S4_SIZE	ln AGE	ln P1	ln P2	ln E3D	ln E5
ln O2	0.75	0.12	0.65	0.72	0.52	0.56
$\ln O2 \mid O2 \leq 4.7 * 10^6$	0.70	0.13	0.50	0.64	0.46	0.51
$\ln O2 \mid O2 > 4.7 * 10^6$	0.19	0.13	0.42	0.19	0.27	0.00

Table 8.2: Estimated correlations between log-transformed turnover (O2) and covariates.

Data modelling using a two-dimensional self-organizing map (SOM) is done with 64 cells. A model is visualized in data space, and data local statistics are visualized in latent space. Figure 8.6 depicts a 2-D SOM in a data space of log transformed and min-max [0,1] equalized turnover (O2), number of employees (S4_SIZE), and assets (P1). The same model is shown in Figure 8.7 in a 2-D latent space, in which local averages of the three log-transformed variables are shown as bars. From the two figures linear dependency between number of employees and turnover is visible. One can also conclude that SOM is able to model dependencies between variables.

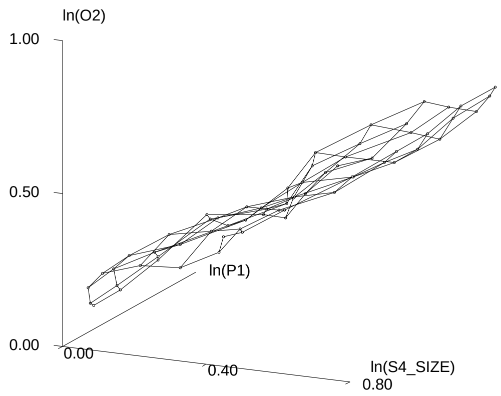


Figure 8.6: 2-D SOM model in data space of log transformed and min-max [0,1] equalized turnover, number of employees, and assets.

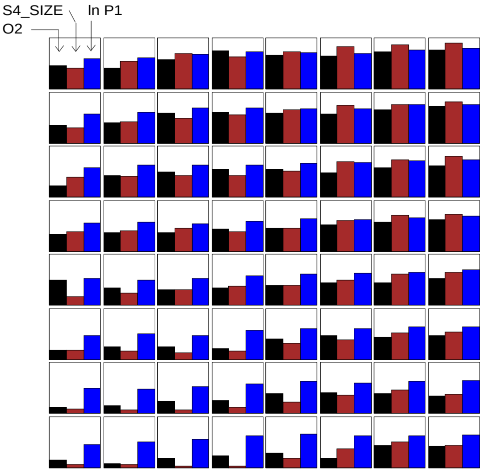


Figure 8.7: Visualization of local averages of log transformed variables (turnover O2, number of employees S4_SIZE, assets P1) in 2-D latent space.

8.2 Imputation procedures

A total of 24 combinations of imputation methods and strategies were used in this case, including the cell methods that were introduced in the previous chapters. As mentioned above, all the variables were log-transformed. Then the ranges of the variables were equalized using min-max rule

$$x' = \frac{x - x_{min}}{x_{\max} - x_{\min}},$$

where maximum and minimum values are computed from the complete part of data. As an exception for joint $(Y, \boldsymbol{X})$ clustering with TS-SOM the values are computed from all the available values of variable. Other equalizations, such as variance based equalization, were also tested but min-max gave the best performance.

After preprocessing, imputation methods and strategies were used to create completed data. As usual there were some methodology specific settings, which are described below:

i) Sampling weights are used only with baseline methods. Thus the weighted baseline imputation procedures are as follows

Mean strategy: weighted prediction is

$$Y^{imp,B,M} = \hat{\mu}^{obs,w} = \frac{1}{S} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{train}} W_j Y_j^{obs},$$

where $S = \sum_{j|\boldsymbol{X}_j \in \mathbf{D}^{train}} W_j$ and W_j is the j:th sampling weight[3].

[3] Weight equals the inverse of inclusion probability.

Random strategy: imputation distribution is

$$Y^{imp,B,R} \sim N(\hat{\mu}^{obs,w}, \hat{\tau}^{obs,w}),$$

where the weighted variance estimator is

$$\hat{\tau}^{obs,w} = \frac{N^{obs}}{(N^{obs}-1)S} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{train}} W_j (Y_j^{obs} - \hat{\mu}^{obs,w})^2.$$

Donor strategy: $Y^{imp,B,D}$ is drawn from observed data values $\{Y_j^{obs}\}_{j=1}^{N^{obs}}$ with replacements and using weights $\{W_j\}_{j=1}^{N^{obs}}$.

ii) Nearest neighbour is used with smoothing parameter $k = 1$.

iii) Kernel regression utilizes a spherical Gaussian kernel with a single smoothing parameter. The bandwidth that was used is loosely based on the optimality criteria for a random design, which is introduced for example in Mack's paper[72]. However, the estimation of true optimal bandwidth is somewhat complicated as it depends on the values of covariates and requires density estimation, among other things. To simplify things, the assumption made here is that smoothing is constant over all prediction positions (for a given number of observations and covariates). Formally, let covariate vector be $\boldsymbol{X}$. Smoothing bandwidth is now $0.2 * N^{obs^{-1/\lambda}}$ where N^{obs} is the number of complete observations and $\lambda = 1/(\dim(\boldsymbol{X})+4)$. The bandwidth 'scale' value of 0.2 was chosen, because it seemed to work quite well. Some other values were tested too, but the chosen alternative gave the best results.

iv) K-Means clustering is done with 32 cells.

v) Two-dimensional TS-SOM with lattice topology and 64 cells is used. Smoothing parameter $h = 0.5$ is used with covariate cell methods based on TS-SOM to form smoothed predictions of Y.

After imputation some simple post-processing operations were done to ensure that all the values are "realistic". Namely, the imputed values were thresholded to be less than or equal to $\ln(500 * 10^6)$ in logaritmic scale, which for example removes outliers caused by simulated randomness from the Gaussian model.

8.3 Evaluation statistics

For evaluation purposes the data set is divided between "large" and "sme" (small and medium) sized enterprises, as was done in the data analysis. Company with turnover Y_j^{orig} (before log-transform) obeys

$$enterprise = \begin{cases} \text{"large" if turnover } > 4.7 \text{ million pounds} \\ \text{"sme" if turnover } \leq 4.7 \text{ million pounds.} \end{cases}$$

Using this division we shall evaluate the imputation performance for three classes:

a) all data.

b) sme, which corresponds roughly to 90.7% of unweighted observations and 89.8% of weighted missing values.

c) large enterprises.

After some consideration, it was decided that all evaluation results will be given in a sampling weight corrected form under logaritmic scale. This is not the only possibility, but it is, we believe, the most illustrative way of presenting the results.

Let W_j denote the sampling weight for observation j. The computed measures of performance for the class of all enterprises are Kolmogorov-Smirnov distance, MSE, and biases of $\hat{\mu}^{imp}$ and $\hat{\tau}^{imp}$ as follows

$$\begin{aligned}
\text{KS} &= \mathbb{E}\left[\sup_y |\hat{F}_{Y^{mis}}(y) - \hat{F}_{Y^{imp}}(y)| \,\Big|\mathsf{n}\right], && (8.1)\\
\text{MSE} &= \mathbb{E}\left[\frac{1}{S_w} \sum_{j:Y_j \in \mathbf{D}_Y^{mis}} W_j (Y_j - Y_j^{comp})^2 \,\Big|\mathsf{n}\right], \\
\mathbb{B}\text{ias}[\hat{\mu}^{imp}|\mathsf{n}] &= \mathbb{E}[\hat{\mu}^{imp} - \hat{\mu}^{mis}|\mathsf{n}], \\
\mathbb{B}\text{ias}[\hat{\tau}^{imp}|\mathsf{n}] &= \mathbb{E}[\hat{\tau}^{imp} - \hat{\tau}^{mis}|\mathsf{n}],
\end{aligned}$$

where Y is log-transformed turnover variable, $S_w = \sum_{j:Y_j \in \mathbf{D}_Y^{mis}} W_j$, Y_j is the j:th missing (random) observation, Y_j^{comp} is the j:th imputed (random) observation, and $\hat{F}$ is an empirical cumulative distribution function.

In addition we shall compute the biases of imputed values at five quantiles of the distribution of Y as follows

$$\mathbb{B}\text{ias}[\hat{\xi}_q^{imp}|\mathsf{n}] = \mathbb{E}[\hat{\xi}_q^{imp} - \hat{\xi}_q^{mis}|\mathsf{n}] \text{ for q } \in \{0.05, 0.25, 0.50, 0.75, 0.95\},$$

where $\hat{\xi}_q^{imp}$ and $\hat{\xi}_q^{mis}$ are sample q-quantiles which are computed from N^{mis} missing and imputed data values.

The quantities for which evaluation statistics (8.1) is computed are weighted imputation results as follows

$$\begin{aligned}
\hat{F}_{Y^{imp}}(y) &= \frac{1}{S_w} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{mis}} W_j I(Y_j^{comp} \leq y), && (8.2)\\
\hat{\mu}^{imp} &= \frac{1}{S_w} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{mis}} W_j Y_j^{comp}, \text{ and} \\
\hat{\tau}^{imp} &= \frac{\mathsf{N}^{mis}}{(\mathsf{N}^{mis} - 1) S_w} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{mis}} W_j (Y_j^{comp} - \hat{\mu}^{imp})^2,
\end{aligned}$$

where $I(\cdot)$ denotes indicator function. Quantities which are based on missing data are defined similarly except that Y_j^{comp} is replaced by Y_j in the three formulas in (8.2).

The statistics for "sme" and "large" enterprises are the same as above, expect that the values are computed over conditionalized data:

$$
\begin{aligned}
\text{KSs} &= \mathbb{E}\left[\sup_y \left|\hat{F}_{Y^{mis}|Y^{mis}\leq \ln(4.7*10^6)}(y) - \hat{F}_{Y^{imp}|Y^{mis}\leq \ln(4.7*10^6)}(y)\right| \mid \mathsf{n}\right],\\
\text{KSl} &= \mathbb{E}\left[\sup_y \left|\hat{F}_{Y^{mis}|Y^{mis}> \ln(4.7*10^6)}(y) - \hat{F}_{Y^{imp}|Y^{mis}> \ln(4.7*10^6)}(y)\right| \mid \mathsf{n}\right],\\
\text{MSEs} &= \mathbb{E}\left[\frac{1}{\#\{j : Y_j \leq \ln(4.7*10^6)\}} \sum_{j:Y_j\leq \ln(4.7*10^6)} W_j(Y_j - Y_j^{comp})^2 \mid \mathsf{n}\right], \text{ and}\\
\text{MSEl} &= \mathbb{E}\left[\frac{1}{\#\{j : Y_j > \ln(4.7*10^6)\}} \sum_{j:Y_j> \ln(4.7*10^6)} W_j(Y_j - Y_j^{comp})^2 \mid \mathsf{n}\right],
\end{aligned}
$$

where $\#(A)$ is the cardinality of set A. Thus we have three versions for KS and MSE: all data, sme data, and large enterprise data.

Before describing the results a decomposition of mean squared error estimator is given. This decomposition provides us with a good way to compare between unit level performance of the methods.

Let $\hat{mse} = \sum_{j:\boldsymbol{X}_j\in\mathbf{D}^{mis}} \overline{W}_j\Big(Y_j - Y_j^{comp}\Big)^2$ and $\hat{\mu}^{mis} = \frac{1}{S_w}\sum_{j:\boldsymbol{X}_j\in\mathbf{D}^{mis}} \overline{W}_j Y_j$. Expectation of $\hat{mse}$ may be decomposed as (see Appendix A8.1 for details)

$$
\begin{aligned}
\mathbb{E}[\hat{mse}] \;=\; & \underbrace{\mathbb{E}[\sum_{j:Y_j\in\mathbf{D}_y^{mis}} \overline{W}_j(Y_j - \hat{\mu}^{mis})^2 \qquad]}_{\text{A: weighted variance of missing } \mathrm{Y_j^{mis}}} \\
+ & \underbrace{\mathbb{E}[\sum_{j:\boldsymbol{X}_j\in\mathbf{D}^{mis}} \overline{W}_j(Y_j^{comp} - \hat{\mu}^{imp})^2]}_{\text{B: weighted variance of imputed values}} \\
& \underbrace{-\,2\mathbb{E}[\sum_{j:Y_j\in\mathbf{D}_y^{mis}} \overline{W}_j(Y_j - \hat{\mu}^{mis})(Y_j^{comp} - \hat{\mu}^{imp})]}_{\text{C: }-2*\text{weighted covariance between missing and imputed}} \\
+ & \underbrace{(\mathbb{E}[(\hat{\mu}^{mis} - \hat{\mu}^{imp})^2]}_{\text{D: global estimation squared bias}},
\end{aligned}
\tag{8.3}
$$

where $\overline{W}_j = W_j/S_w$. Expectation of squared global bias estimator (term D) may be further decomposed as:

$$
\begin{aligned}
\mathbb{E}\Big[(\hat{\mu}^{mis} - \hat{\mu}^{imp})^2\Big] \;=\; & \underbrace{(\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{mis}])^2}_{d_1\text{: expected squared bias}} \\
& + \underbrace{\mathbb{V}\mathrm{ar}[\hat{\mu}^{mis}]}_{d_1\text{: variance of } \hat{\mu}^{\mathrm{mis}}} \\
& + \underbrace{\mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}]}_{d2\text{: sample variance of } \hat{\mu}^{\mathrm{imp}}} \\
& \underbrace{-\,2\mathbb{E}\Big[(\hat{\mu}^{mis} - \mathbb{E}[\hat{\mu}^{mis}])(\hat{\mu}^{imp} - \mathbb{E}[\hat{\mu}^{imp}])\Big]}_{d_4\text{: }-2*\text{covariance between } \hat{\mu}^{\mathrm{mis}} \text{ and } \hat{\mu}^{\mathrm{imp}}}\;.
\end{aligned}
\tag{8.4}
$$

8.4 Evaluation of results

The evaluation of imputation performance is a somewhat subjective task. From a single viewpoint one method may be better than another, but when several evaluation measures are used simultaneously, it can be difficult to pick one clear winner.

In our setup of the general problem, where the future use of imputed data is unknown, the best we can do is to try to preserve many evaluation measures simultaneously. In simple terms we like to simultaneously preserve distributional measures and minimize errors in unit level imputations. To understand which of the methods is best in this, the following evaluations are made

i) Preservation of moments in terms of $\mathbb{B}\text{ias}[\hat{\mu}^{imp}]$ and $\mathbb{B}\text{ias}[\hat{\tau}^{imp}]$.

ii) Preservation of five quantiles.

iii) Qualitative analysis of the preservation of distribution for selected methods.

iv) A detailed study of unit level errors in terms of MSE.

v) A comparison between distributional Kolmogorov-Smirnov measures vs. unit level MSE performance.

The number of repetitions is 10000 except in the evaluation iii) in which the analysis is based on a single repetition only.

As mentioned earlier, all these evaluations are done in log scale. Log scale is used due to a problem with conversion to normal scale. Translation of results from log-space to the original scale is not trivial as one might think, because "truthful" conversion from log scale to a normal scale must handle random noise terms in a sensible way. To demonstrate the problem let $Z = \exp(Y)$ be turnover in the original space and Y be the logarithmic one. First the two moments of Y^{mis} are denoted as μ^{*mis} and τ^{*mis}, and we assume that Y^{mis} follows Gaussian distribution. A basic result from statistics (see p.257 of [67]) states that $Z^{mis} = \exp(Y^{mis})$ follows log-normal distribution with the expectation:

$$\mathbb{E}[Z^{mis}] \;=\; \exp(\mu^{*mis} + \tau^{*mis}/2).$$

Provided that moment estimation is ignored, then the distribution of Y^{imp} is roughly $N(\mu^{*obs}, \tau^{*obs})$ when the baseline method is used with the random strategy, where μ^{*obs} and τ^{*obs} are the first two moments of Y^{obs}. Expectation of the mean estimator for the imputed values in the original scale is then computed as

$$\begin{aligned}\mathbb{E}[\hat{\mu}_z^{imp,B,R}|\mathsf{n}^{mis},\mathsf{n}] &= \frac{1}{\mathsf{n}^{mis}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \mathbb{E}[\exp(Y_j^{comp})|\mathsf{n}^{mis},\mathsf{n}] \approx \exp(\mu^{*obs} + \tau^{*obs}/2)\\ \Rightarrow \mathbb{E}[\hat{\mu}_z^{imp,B,R}|\mathsf{n}] &\approx \exp(\mu^{*obs} + \tau^{*obs}/2).\end{aligned}$$

Therefore the approximate biases for the two baseline methods are

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\mu}_z^{imp,B,M}|\mathsf{n}] &\approx \exp(\mu^{*obs}) - \underbrace{\exp(\mu^{*mis} + \tau^{*mis}/2)}_{\mathbb{E}[Z^{mis}]}, \text{ and}\\ \mathbb{B}\text{ias}[\hat{\mu}_z^{imp,B,R}|\mathsf{n}] &\approx \exp(\mu^{*obs} + \tau^{*obs}/2) - \exp(\mu^{*mis} + \tau^{*mis}/2).\end{aligned}$$

From the above bias results one can notice that the first two moments of Y^{imp} (log-transformed turnover) affect the bias of the first moment estimator in the original space.

8.4.1 Preservation of moments

It is quite well known that in log scale many enterprise variables like the turnover can be well explained using linear models. We may therefore expect that a linear model should preserve the two first moments of turnover rather well. To evaluate this together with other models, we shall study both $\mathbb{B}\text{ias}[\hat{\mu}^{imp}]$ and $\mathbb{B}\text{ias}[\hat{\tau}^{imp}]$.

To make interpretations easier, we include also relative performances in terms of percentages as follows. The error percentage for $\hat{\mu}^{imp}$ is computed as

$$\text{Err\%}(\hat{\mu}^{imp}) = \mathbb{E}\Big[\Big|\frac{\hat{\mu}^{imp} - \hat{\mu}^{mis}}{\hat{\mu}^{mis}}\Big| * 100\%|\mathsf{n}\Big], \text{ and}$$

$$\text{Err\%}(\hat{\tau}^{imp}) = \mathbb{E}\Big[\Big|\frac{\hat{\tau}^{imp} - \hat{\tau}^{mis}}{\hat{\tau}^{mis}}\Big| * 100\%|\mathsf{n}\Big].$$

The results are summarized in Table 8.3. As we can see from the table the first moment μ^{*mis} is best preserved by linear regression, nearest neighbour imputation, joint TS-SOM clustering with smoothing and by joint cell K-means clustering, but basically all covariate methods perform equally. As we can see, joint TS-SOM tends to underestimate μ^{*mis} while all the other methods overestimate it. This can be explained by the tendency of TS-SOM to move all clusters close to each other, which leads to underestimation of large values (large turnover is downward biased). Note also the differences between $\mathbb{B}\text{ias}[\hat{\mu}^{imp}]$ and $\text{Err\%}(\hat{\mu}^{imp})$. This is explained by fact that error percentage is an expectation of the ratio of two random quantities. Therefore for finite sample size n the variances of random quantities are mixed in the result.

As expected the preservation of the second moment is more difficult. At their best, methods can reach about 24% relative errors for the preservation of τ^{*mis}. In most cases the best performances are due to the simulated random imputation strategy, which fails only with the standard $\boldsymbol{X}$ clustering under smoothing. All the methods tend to underestimate the variance.

	$\mu^{*mis} \approx 12.48$		$\tau^{*mis} \approx 2.24$	
Method	$\mathbb{B}$ias$[\hat{\mu}^{imp}]$	Err%$(\hat{\mu}^{imp})$	$\mathbb{B}$ias$[\hat{\tau}^{imp}]$	Err%$(\hat{\tau}^{imp})$
B,M	0.697(0.002)	5.611(0.019)	-2.235(0.003)	100.000(0.000)
B,R	0.698(0.003)	5.629(0.023)	**-0.181**(0.006)	**24.441**(0.184)
B,D	0.697(0.003)	5.624(0.024)	**-0.182**(0.007)	25.792(0.206)
L,M	**0.040**(0.001)	$\underline{\mathbf{0.904}}$(0.007)	-1.236(0.003)	54.430(0.078)
L,R	**0.039**(0.002)	**1.137**(0.009)	-0.524(0.004)	**24.457**(0.131)
N,M	**0.033**(0.002)	**1.087**(0.008)	-0.446(0.005)	25.371(0.147)
K,M	0.168(0.001)	1.477(0.009)	-1.270(0.004)	55.813(0.108)
K,R	0.167(0.002)	1.550(0.010)	-0.861(0.004)	37.446(0.144)
T,M	0.048(0.002)	**1.100**(0.008)	-1.116(0.004)	48.700(0.132)
T,R	0.046(0.002)	1.332(0.010)	**-0.248**(0.006)	**24.538**(0.181)
T,D	0.048(0.002)	1.315(0.010)	-0.366(0.006)	25.983(0.171)
T,M^s	0.158(0.002)	1.486(0.010)	-1.427(0.004)	62.934(0.090)
T,R^s	0.158(0.003)	2.032(0.015)	1.375(0.009)	67.987(0.497)
C,M	0.104(0.002)	1.261(0.009)	-1.173(0.004)	51.208(0.130)
C,R	0.104(0.002)	1.476(0.011)	-0.275(0.006)	$\underline{\mathbf{23.893}}$(0.166)
C,D	0.103(0.002)	1.474(0.011)	-0.356(0.006)	25.731(0.172)
TJ,M	-0.078(0.002)	1.254(0.010)	-0.584(0.005)	29.454(0.158)
TJ,R	-0.077(0.002)	1.363(0.011)	$\underline{\mathbf{-0.143}}$(0.007)	**24.842**(0.198)
TJ,D	-0.077(0.002)	1.337(0.010)	-0.256(0.006)	25.100(0.179)
TJ,M^s	**-0.041**(0.002)	**1.130**(0.009)	-0.924(0.005)	40.517(0.153)
TJ,R^s	**-0.039**(0.002)	1.232(0.009)	-0.558(0.005)	28.479(0.157)
CJ,M	**0.012**(0.002)	**1.088**(0.008)	-0.892(0.005)	39.770(0.155)
CJ,R	**0.012**(0.002)	1.262(0.010)	**-0.251**(0.006)	**24.553**(0.183)
CJ,D	$\underline{\mathbf{0.011}}$(0.002)	1.261(0.010)	-0.303(0.006)	25.955(0.184)

Table 8.3: First two moments of missing turnover and biases of estimators $\hat{\mu}^{imp}$ and $\hat{\tau}^{imp}$. Standard deviations of simulation estimates are shown in parentheses. See Section 3.6.3 for details.

8.4.2 Preservation of quantiles

Five quantiles were computed from both the original data and the imputed data set as follows:

$$\begin{aligned}\hat{\xi}_q^{mis} &= \hat{F}^{-1}_{Y^{mis}}(q),\ q = (0.05, 0.25, 0.5, 0.75, 0.95)\\ \hat{\xi}_q^{imp} &= \hat{F}^{-1}_{Y^{imp}}(q),\end{aligned}$$

where $\hat{F}^{-1}_{Y^{mis}}(q)$ gives point y_q in which $\hat{F}_{Y^{mis}}(y_q) = q$, and similarly for imputed values.

The bias in quantities $\hat{\xi}_q^{imp}$ is simply $\mathbb{B}\text{ias}[\hat{\xi}_q^{imp}|\mathsf{n}] = \mathbb{E}[\hat{\xi}_q^{imp} - \hat{\xi}_q^{mis}|\mathsf{n}]$. These results are given in Table 8.4.

As expected, the largest biases can be found from quantiles $q = 0.05$ and $q = 0.95$. A deeper study of the results reveals that in other quantiles the cell methods dominate in this measure of performance. Out of 27 best results 22 are obtained by cell imputation. It is also notable that joint clustering using TS-SOM seems to be the best alternative, especially when the simulated random strategy is used. For small enterprises ($q = 0.05$) the best alternative seems to be nearest neighbour, but also linear regression with random strategy works quite well. The same comment applies also to the largest enterprises (quantile $q = 0.95$). Therefore for this type of data one can recommend the joint $(Y, \boldsymbol{X})$ clustering TS-SOM methods for mass imputation of small and medium enterprises.

	$q = 0.05$	$q = 0.25$	$q = 0.50$	$q = 0.75$	$q = 0.95$
$\xi_q^{*mis} \approx$	10.54	12.01	13.12	14.22	15.87
Method	$\mathbb{B}\text{ias}(\hat{\xi}_{0.05}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.25}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.50}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.75}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.95}^{imp})$
B,M	2.633(0.004)	1.167(0.004)	0.053(0.003)	-1.042(0.004)	-2.693(0.004)
B,R	0.275(0.004)	0.193(0.003)	0.054(0.002)	-0.065(0.003)	-0.342(0.004)
B,D	0.293(0.005)	0.331(0.004)	0.051(0.002)	-0.207(0.003)	-0.307(0.004)
L,M	0.603(0.003)	0.208(0.002)	**0.011**(0.001)	-0.151(0.002)	-0.344(0.003)
L,R	0.147(0.003)	0.080(0.002)	0.016(0.002)	0.032(0.002)	**0.008**(0.003)
N,M	**0.024**(0.006)	0.136(0.003)	**-0.013**(0.001)	-0.080(0.002)	-0.040(0.003)
K,M	0.766(0.004)	0.303(0.002)	0.145(0.001)	-0.233(0.002)	-0.401(0.003)
K,R	0.449(0.004)	0.248(0.002)	0.091(0.001)	-0.045(0.002)	-0.198(0.003)
T,M	0.641(0.005)	0.066(0.003)	0.093(0.002)	-0.141(0.002)	-0.285(0.003)
T,R	**0.003**(0.005)	0.095(0.003)	0.065(0.002)	0.065(0.002)	0.070(0.003)
T,D	0.117(0.005)	0.167(0.003)	0.047(0.002)	**0.017**(0.002)	**-0.001**(0.003)
T,M^s	1.029(0.005)	0.220(0.003)	0.123(0.002)	-0.155(0.002)	-0.417(0.003)
T,R^s	-0.622(0.005)	-0.164(0.003)	0.121(0.002)	0.471(0.003)	0.863(0.005)
C,M	0.767(0.005)	0.122(0.004)	0.146(0.002)	-0.175(0.002)	-0.320(0.003)
C,R	0.082(0.005)	0.116(0.003)	0.096(0.002)	0.108(0.002)	0.073(0.003)
C,D	0.147(0.005)	0.225(0.003)	0.079(0.002)	0.040(0.002)	**0.025**(0.003)
TJ,M	**0.024**(0.006)	**0.037**(0.003)	0.046(0.002)	-0.075(0.002)	-0.181(0.003)
TJ,R	-0.138(0.006)	**-0.002**(0.003)	**-0.002**(0.002)	**0.021**(0.002)	0.037(0.003)
TJ,D	**-0.047**(0.006)	**0.024**(0.003)	**-0.010**(0.002)	**-0.009**(0.002)	**-0.015**(0.003)
TJ,M^s	0.322(0.005)	**0.039**(0.003)	0.064(0.002)	-0.110(0.002)	-0.296(0.003)
TJ,R^s	0.113(0.005)	**0.026**(0.002)	**-0.001**(0.002)	**-0.006**(0.002)	-0.070(0.003)
CJ,M	0.421(0.006)	**0.034**(0.004)	0.073(0.002)	-0.155(0.002)	-0.188(0.003)
CJ,R	-0.087(0.006)	0.067(0.003)	0.039(0.002)	0.029(0.002)	0.043(0.003)
CJ,D	**0.017**(0.006)	0.126(0.003)	0.024(0.002)	**-0.010**(0.002)	**-0.005**(0.003)

Table 8.4: Quantiles and biases in quantiles (ξ_q^{*mis}) for imputation methods. Standard simulation deviations of error estimates are shown in parentheses.

8.4.3 Qualitative analysis of the preservation of distribution

In addition to quantitative (numerical) results it is always a good idea to do some qualitative studies as well. In this case we shall investigate three different approaches using data visualization:

a) B,R baseline with simulated randomness,

b) L,M linear regression with mean strategy, and

c) TJ,R^s joint (Y,X) TS-SOM clustering with simulated randomness.

The results of imputation are depicted in log scale using three types of plots

- Graphical confusion (density) tables,
- Density plots, and
- QQ-plots.

In all the plots we try to compare the distribution of imputed values against the distribution of true values of missing data.

In confusion plots (Figure 8.8, Figure 8.10, and Figure 8.12) the joint distribution between Y^{mis} and Y^{imp} is visualized using a contour display. Ideally the plot is concentrated close to the diagonal line, which indicates that Y^{imp} is close to Y^{mis}.

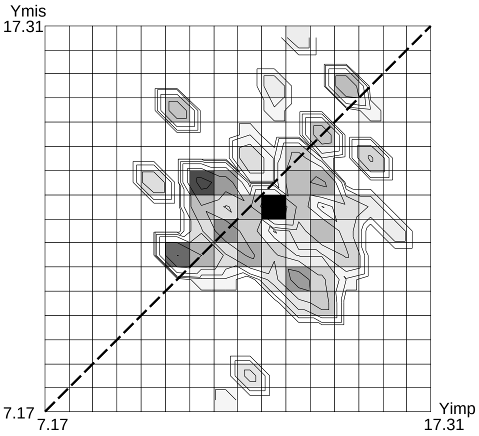


Figure 8.8: Confusion table (pdf) between $Y^{imp,B,R}$ and Y^{mis}.

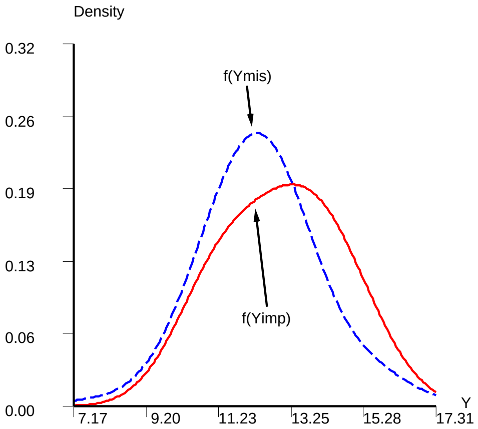


Figure 8.9: Density $f(Y^{imp,B,R})$ versus $f(Y^{mis})$.

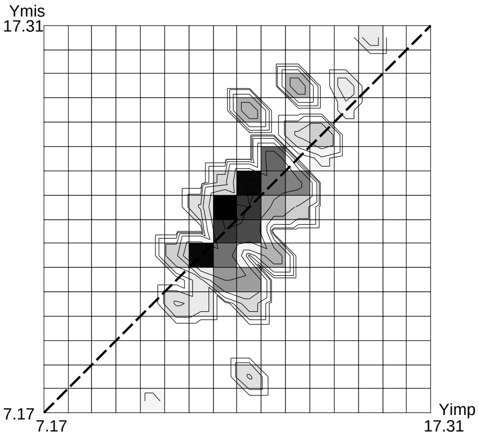


Figure 8.10: Confusion table (pdf) between $Y^{imp,L,M}$ and Y^{mis}.

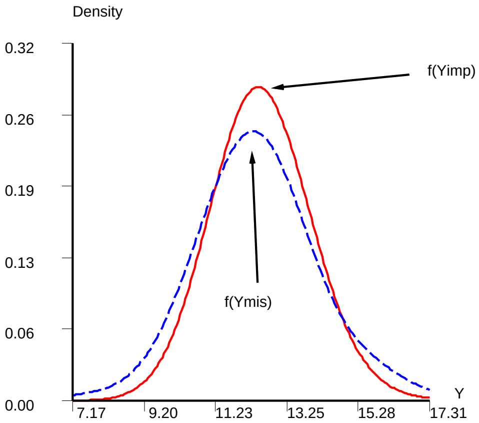


Figure 8.11: Density $f(Y^{imp,L,M})$ versus $f(Y^{mis})$.

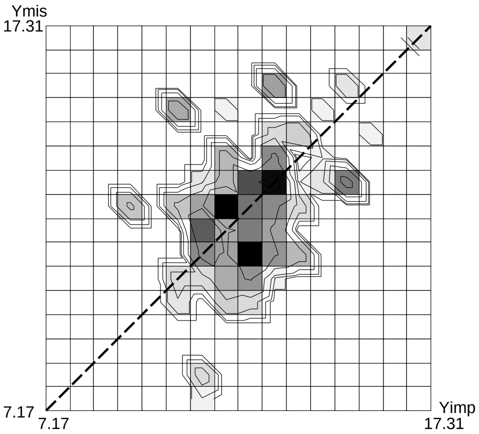


Figure 8.12: Confusion table (pdf) between Y^{imp,TJ,R^s} and Y^{mis}.

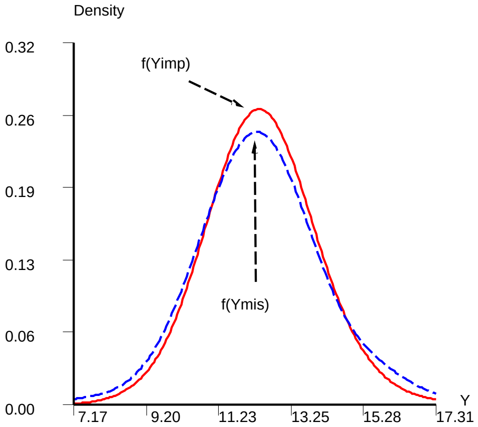


Figure 8.13: Density $f(Y^{imp,TJ,R^s})$ versus $f(Y^{mis})$.

In Figures 8.9, 8.11, and 8.13 the corresponding marginal distributions of confusion plots are shown. From the plots one can conclude that baseline B,R behaves much worse than the two other methods. It seems that the linear method is slightly better on a unit level as its confusion plot seems more diagonalized than that of TS-SOM, but on the level of marginal distributions TS-SOM seems to be a little better.

The conclusions are verified easily using QQ plots as shown in Figure 8.14.

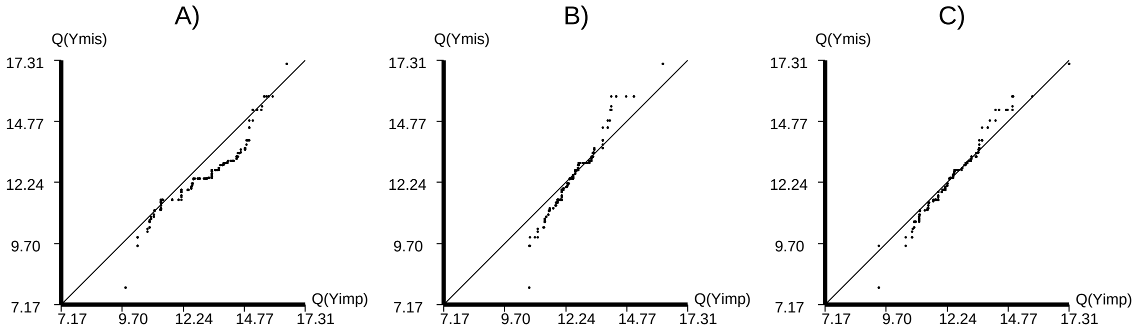


Figure 8.14: A) QQ plot for B,R method, B) QQ for L,M method, and C) QQ for TJ,R method.

It seems that the linear method and the baseline have larger deviations from diagonal than the TS-SOM. A notable difference between the baseline and other methods is that the baseline does best on small and big enterprises, while TS-SOM is best for the central part of the distribution.

8.4.4 Preservation of weighted unit level MSE

For the analysis of unit level performance we recall, from Section 8.3, the decomposition of expected (integrated) MSE

$$\mathbb{E}[\hat{mse}] = \mathbb{E}[\sum_{j:\mathbf{X}_j \in \mathbf{D}^{mis}} \overline{W}_j (Y_j^{imp} - Y_j^{true})^2] = A + B + C + \underbrace{d_1 + d_2 + d_3 + d_4}_{D},$$

where

$$\begin{aligned}
A &= \mathbb{E}[\sum_j \overline{W}_j (Y_j^{mis} - \hat{\mu}^{mis})^2], \text{ weighted variance of missing } \mathrm{Y_j^{mis}} \\
B &= \mathbb{E}[\sum_j \overline{W}_j (Y_j^{imp} - \hat{\mu}^{imp})^2], \text{ weighted variance of imputed values} \\
C &= -2\mathbb{E}\Big[\hat{\mathbb{C}\text{ov}}[Y^{imp}, Y^{mis}]\Big], \text{ weighted covariance between missing and imputed values} \\
D &= \mathbb{E}[(\hat{\mu}^{mis} - \hat{\mu}^{imp})^2], \text{ global estimation bias,}
\end{aligned}$$

where

$$\begin{aligned}
d_1 &= (\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{imp}])^2, \text{ expected bias} \\
d_2 &= \mathbb{V}\text{ar}[\hat{\mu}^{imp}], \text{ variance of } \hat{\mu}^{imp} \\
d_3 &= \mathbb{V}\text{ar}[\hat{\mu}^{mis}], \text{ sample variance of } \hat{\mu}^{mis} \\
d_4 &= -2\mathbb{C}\text{ov}[\hat{\mu}^{imp}, \hat{\mu}^{mis}], \text{ covariance between } \hat{\mu}^{imp} \text{ and } \hat{\mu}^{mis}.
\end{aligned}$$

Terms $A = \mathbb{V}\text{ar}[Y_j^{mis}]$ and $d_2 = \mathbb{V}\text{ar}[\hat{\mu}^{mis}]$ are caused by our data generator and have values $A \approx 2.22, d_2 \approx 0.021$. All the other terms are affected by the chosen methodology and thus they tell us about the differences of the methods in terms of unit level prediction accuracy. In general terms, we do expect that the chosen method is able to explain at least a part of $\mathbb{V}\text{ar}[Y_j^{mis}]$. Therefore we can expect MSE values that are less than 2.22.

The imputation results are summarized in Table 8.5. Most of the approaches can, indeed, reduce MSE from 2.22, and the best value is close to 1. In general terms, all nonparametric regression methods, covariate clustering cell methods with mean strategy, and joint clustering cell methods with mean strategy do quite well. The prediction ability is reflected by term $C = -2\mathbb{C}\text{ov}[Y^{imp}, Y^{mis}]$, but for clarity we have also computed the correlation coefficient between the imputed and missing values as follows:

$$Correlation = \begin{cases} -0.5 * C/\sqrt{A * B}, & \text{if } C > 0, \\ 0, & \text{if } C = 0. \end{cases}$$

The worst performance of T,R^s is most likely due to overestimation of simulated random noise, which contributes to term $B = \mathbb{V}\text{ar}[Y^{imp}]$. As we can expect, mean imputation always outperforms strategies where randomness is added to predicted values.

						d_1	d_3	d_4
Method	**MSE**	B	C	D	Correlation	Global bias^2	$\mathbb{V}\text{ar}[\hat{\mu}^{imp}]$	$-2\mathbb{C}\text{ov}[\hat{\mu}^{mis}, \hat{\mu}^{imp}]$
B,M	2.757(0.004)	0.000(0.000)	0.000(0.000)	0.537(0.003)	0.000(0.000)	0.486	0.009	0.019
B,R	4.828(0.008)	2.041(0.004)	-0.005(0.005)	0.572(0.004)	0.001(0.001)	0.487	0.045	0.019
B,D	4.830(0.008)	2.039(0.005)	-0.002(0.005)	0.572(0.004)	0.001(0.001)	0.485	0.046	0.020
L,M	**1.024**(0.002)	0.993(0.001)	-2.209(0.003)	0.019(0.000)	**0.745**(0.000)	**0.002**	0.011	-0.015
L,R	1.747(0.003)	1.700(0.003)	-2.205(0.004)	0.031(0.000)	0.569(0.001)	**0.002**	0.023	-0.015
N,M	1.473(0.003)	1.777(0.004)	-2.554(0.005)	0.029(0.000)	0.646(0.001)	**0.001**	0.025	-0.018
K,M	$\underline{\mathbf{1.019}}$(0.002)	0.958(0.002)	-2.207(0.003)	0.047(0.001)	$\underline{\mathbf{0.761}}$(0.000)	0.028	0.012	-0.014
K,R	1.431(0.003)	1.365(0.003)	-2.208(0.004)	0.054(0.001)	0.637(0.001)	0.028	0.019	-0.015
T,M	**1.234**(0.003)	1.112(0.002)	-2.128(0.003)	0.029(0.000)	**0.682**(0.001)	0.002	0.017	-0.012
T,R	2.107(0.005)	1.974(0.005)	-2.130(0.005)	0.043(0.001)	0.513(0.001)	**0.002**	0.032	-0.012
T,D	1.989(0.005)	1.857(0.005)	-2.131(0.005)	0.042(0.001)	0.531(0.001)	**0.002**	0.030	-0.012
T,M^s	**1.213**(0.003)	0.803(0.001)	-1.861(0.003)	0.049(0.001)	**0.701**(0.001)	0.025	0.012	-0.009
T,R^s	4.042(0.008)	3.586(0.008)	-1.863(0.007)	0.098(0.001)	0.332(0.001)	0.025	0.061	-0.009
C,M	**1.269**(0.003)	1.055(0.002)	-2.044(0.003)	0.037(0.000)	**0.672**(0.001)	0.011	0.016	-0.011
C,R	2.173(0.004)	1.948(0.004)	-2.047(0.005)	0.052(0.001)	0.496(0.001)	0.011	0.031	-0.011
C,D	2.100(0.005)	1.866(0.004)	-2.039(0.005)	0.052(0.001)	0.506(0.001)	0.011	0.030	-0.010
TJ,M	1.509(0.004)	1.641(0.004)	-2.391(0.004)	0.039(0.001)	0.632(0.001)	0.006	0.027	-0.015
TJ,R	1.950(0.005)	2.079(0.005)	-2.396(0.005)	0.047(0.001)	0.564(0.001)	0.006	0.034	-0.015
TJ,D	1.836(0.004)	1.966(0.005)	-2.395(0.005)	0.045(0.001)	0.579(0.001)	0.006	0.032	-0.015
TJ,M^s	**1.265**(0.003)	1.302(0.003)	-2.290(0.004)	0.032(0.000)	**0.679**(0.001)	**0.002**	0.023	-0.014
TJ,R^s	1.631(0.003)	1.666(0.004)	-2.294(0.005)	0.038(0.001)	0.601(0.001)	**0.002**	0.029	-0.014
CJ,M	**1.342**(0.003)	1.335(0.003)	-2.242(0.004)	0.029(0.000)	0.658(0.001)	$\underline{\mathbf{0.000}}$	0.021	-0.014
CJ,R	1.989(0.004)	1.972(0.005)	-2.242(0.005)	0.040(0.001)	0.541(0.001)	$\underline{\mathbf{0.000}}$	0.032	-0.014
CJ,D	1.939(0.005)	1.920(0.005)	-2.240(0.005)	0.039(0.001)	0.549(0.001)	$\underline{\mathbf{0.000}}$	0.031	-0.014

Table 8.5: Decomposition terms for mean squared error. Remark that: $A \approx 2.22(0.003)$, $d_2 = \mathbb{V}\text{ar}[\hat{\mu}^{mis}] \approx 0.021$, and $D = d_1 + d_2 + d_3 + d_4$.

8.4.5 Comparison between distribution level and unit level performance

The previous sections have given a mixed picture about the performances of different methods. Some methods demonstrate good performance on a distribution level while others do better on a unit level. This is quite expected as there is a well-demonstrated trade-off between the two performance measures. Especially, it becomes clear that random imputation strategies favor distributional measures while mean strategies provide a better unit level accurancy.

The choice of the "best" method is obviously a multicriteria decision (or optimization) problem [98], where the optimality before decision making is defined by a set of Pareto optimal solutions (see [48] for a good introduction). In the current example we try to approximate the Pareto optimal set in terms of mean squared error (MSE) and Kolmogorov-Smirnov distance (KS). The optimal methods are those that are on the Pareto front in two dimension MSE vs. KS plot, as given in Figure 8.15.

From Figure 8.15 we can observe that our estimate for the Pareto optimal set of methods consists of four methods

K,M - kernel regression with mean strategy

L,M - linear regression with mean strategy

N,M - 1-nearest neighbour imputation

TJ,D - joint $(Y, \boldsymbol{X})$ clustering using TS-SOM with donor strategy.

The first two of the methods should be used if the objective is to minimize unit level errors. The last two do better on distribution level. We also note that methods (TJ,R^s), (L,R), (TJ,R), (T,D), (CJ,D) and (CJ,R) are quite close to the Pareto front, which makes them good alternatives as well.

Numerical values for all the results are given in Table 8.6, and include separate results for sme and large enterprises.

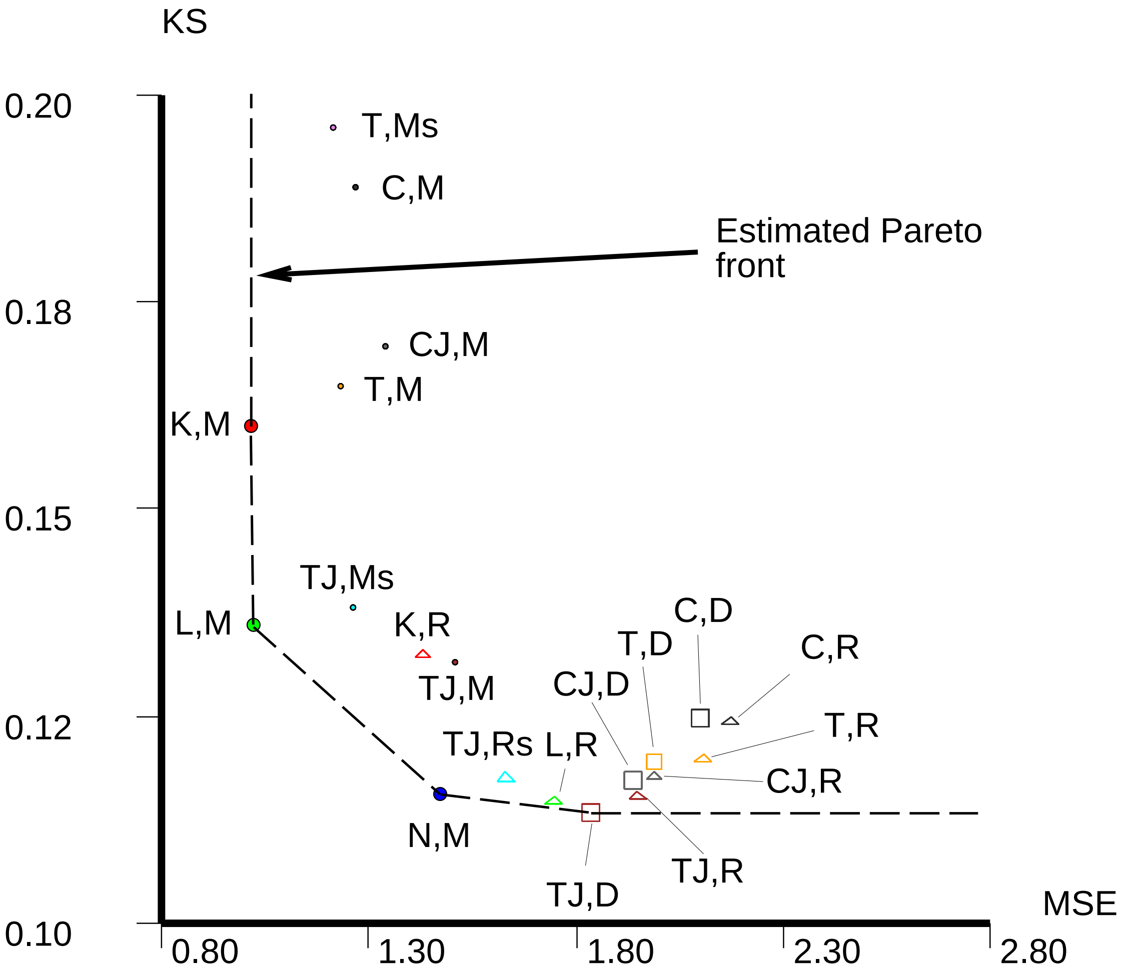


Figure 8.15: Mean squared error versus expected Kolmogorov-Smirnov statistic plot. The results for the baseline methods and the T,R^s method are available in Table 8.6.

Method	KS	KSs	KSl	MSE	MSEs	MSEl
B,M	0.561	0.587	1.000(0.0000)	2.757	2.582	7.689(0.012)
B,R	0.145	0.174	0.962(0.0007)	4.828	4.657	9.693(0.049)
B,D	0.163	0.186	0.956(0.0008)	4.830	4.657	9.731(0.051)
L,M	0.136	0.144	0.720(0.0016)	**1.024**	**0.947**	**2.804**(0.011)
L,R	**0.114**	**0.126**	**0.702**(0.0017)	1.747	1.670	3.527(0.022)
N,M	**0.116**	0.124	0.679(0.0018)	1.473	1.423	2.780(0.017)
K,M	0.160	0.172	0.751(0.0015)	1.019	0.934	**2.930**(0.012)
K,R	0.132	0.143	0.725(0.0017)	1.431	1.345	3.357(0.019)
T,M	0.165	0.179	0.726(0.0016)	**1.233**	**1.151**	**3.046**(0.015)
T,R	0.119	0.135	0.712(0.0017)	2.107	2.036	3.670(0.023)
T,D	0.120	0.133	0.713(0.0017)	1.989	1.917	3.559(0.023)
T,M^s	0.196	0.215	0.758(0.0015)	**1.213**	**1.133**	**2.982**(0.013)
T,R^s	0.152	0.189	**0.707**(0.0017)	4.042	3.967	5.704(0.045)
C,M	0.189	0.205	0.740(0.0016)	**1.269**	**1.181**	**3.175**(0.015)
C,R	0.124	0.141	0.710(0.0017)	2.173	2.096	3.832(0.024)
C,D	0.125	0.140	0.713(0.0017)	2.100	2.020	3.822(0.025)
TJ,M	0.132	0.141	0.722(0.0016)	1.509	1.423	3.315(0.021)
TJ,R	**0.115**	**0.127**	**0.708**(0.0017)	1.950	1.863	3.782(0.028)
TJ,D	0.113	**0.125**	**0.708**(0.0017)	1.836	1.749	3.697(0.026)
TJ,M^s	0.138	0.147	0.737(0.0016)	**1.265**	**1.176**	**3.191**(0.018)
TJ,R^s	**0.117**	**0.127**	0.719(0.0017)	1.631	1.544	3.531(0.022)
CJ,M	0.170	0.183	0.711(0.0017)	1.342	1.254	3.240(0.018)
CJ,R	**0.117**	**0.130**	**0.697**(0.0017)	1.989	1.907	3.732(0.024)
CJ,D	**0.117**	**0.129**	**0.701**(0.0017)	1.939	1.857	3.694(0.024)

Table 8.6: Kolmogorov-Smirnov and mean squared error results. Standard deviations are below 0.0007 for KS and KSs and below 0.009 for MSE and MSEs and thus have been removed to compress the table. The best results for each error measure have been marked using a bold font.

Using the division between sme and large enterprises, the Pareto optimal fronts are as depicted in Figures 8.16 and 8.17. Note that the baseline methods have been excluded from the figures as "outliers".

An investigation of Figure 8.16 reveals that for sme enterprises our estimate for the Pareto optimal front consists of three methods (K,M), (L,M), and (N,M). Further methods (TJ,R^s), (L,R), (TJ,D), and (TJ,R) are close to the front. Ordering of the results is very similar to that of in the plot for all enterprises. Large enterprises are preserved worst, as shown in Figure 8.17. Nearest neighbour imputation performs best, and it forms the Pareto front estimate. However, its distributional level performance is still bad. Linear regression with mean strategy is close to the front. There is a linear tradeoff between K-S and MSE measures among the other methods (between mean and random or donor strategies).

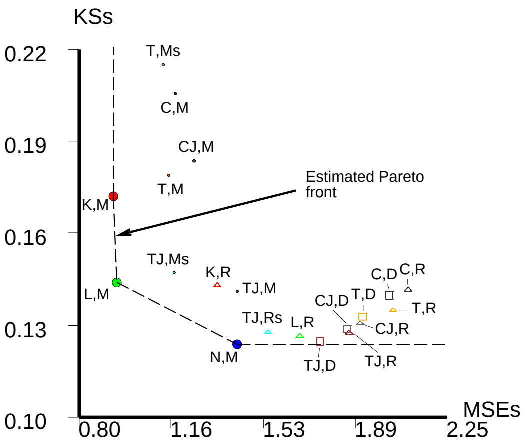


Figure 8.16: Pareto-optimal front for sme class of enterprises with turnover ≤ 4.7 million pounds.

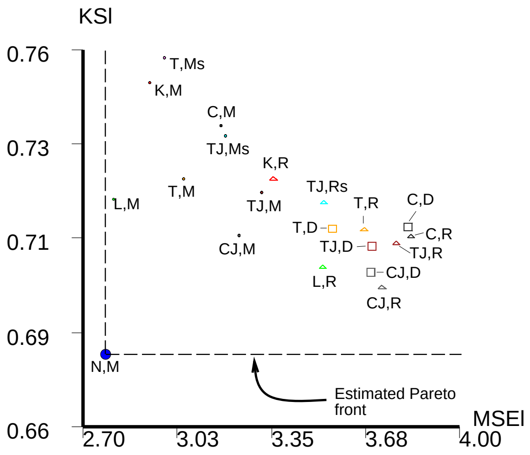


Figure 8.17: Pareto-optimal front for large enterprises with turnover > 4.7 million pounds.

8.5 Summary

In this chapter we demonstrated how different imputation methods perform in the imputation of turnover for enterprises under MAR missingness. Three different classes were evaluated: all data, small and medium enterprises, and large enterprises.

When simultaneous preservation of both unit level and distributional level for all data was considered, four methods came out as winners: kernel regression with mean strategy, linear regression with mean strategy, 1-nearest neighbour imputation, and joint $(Y, \boldsymbol{X})$ clustering using TS-SOM with donor strategy. The first two of these are good for preserving unit level. Distributional level is better preserved by the latter two methods.

Preservation of small and medium (sme) enterprises class is done best by kernel regression with mean strategy, linear regression with mean strategy, and 1-nearest neighbour. Linear regression with random strategy, and TS-SOM joint clustering with donor strategy or random (smoothed) strategy are also close to the estimated Pareto front.

Distribution level and unit level of large enterprises are not well preserved. Nearest neighbour imputation performs best, however, its performance is bad at distribution level. Information in covariates seems not to be adequate to predict large enterprises well.

To conclude, the proposed joint $(Y, \boldsymbol{X})$ clustering cell methods based on TS-SOM perform well in mass imputation of small and medium sized enterprises. The smallest and the largest enterprises are best preserved by 1-nearest neighbour and linear regression.

Chapter 9

Case study: Labour Force survey

The purpose of this chapter is to give a realistic example for the evaluation of imputation methods. This study uses Quarterly Labour Force Survey Household Dataset (LFH), April - June, 2006, produced by the Office for National Statistics (ONS)[1], sponsored by ONS and Northern Ireland[2], and supplied by the UK Data Archive. The data are Crown copyright. The first edition of the data [76], which is dated 16th October 2006, is used in the experiments. The following statement is required for the use of this data set:

The original data creators, depositors or copyright holders, the funders of the Data Collections and the UK Data Archive bear no responsibility for further analysis or interpretation done in this thesis.

The data is similar, but not the same as the SARS (UK Census 1991, Sample of Anonymised Records) data set that was used in the EurEdit project. Our previous experiments with this kind of data were not promising, but we suspect that this was due to the immaturity of our methodology and our inexperience in imputation. In the current study we hope to get a deeper understanding about the strength and weaknesses of cell methods for this kind of data.

By deeper evaluation we mean more careful setting of experiments and more detailed evaluation of the results and causes behind them. In comparison to Euredit this is done as follows

- We limit our study to two variables AGE of a person (in years) and SEX, which is a categorial variable.
- We have carefully designed a MAR type missingness mechanism.
- Results are computed over several repetitions including data sampling and missingness generation.

Like in the previous chapter, our evaluation is not completely objective because the missingness generator is known to us. Yet, we try to be as objective as we can.

[1] Social and Vital Statistics Division

[2] Department of Enterprise, Trade and Investment

Compared to the previous chapter, the current dataset is not simplified, which implies that we must be able to handle special values and other nuisances in data.

As before we shall compare the relative performances of imputation methods using both distributional and unit level measures. For variable AGE we use Kolmogorov-Smirnov (KS) measures and mean squared error (MSE) type of unit level measures. Two measures for categorial imputations are introduced for variable SEX. In addition several special studies are made as explained below:

- We shall investigate conditional performance on quantiles of the imputation of AGE variable under five categories defined by MARSTA (person's marital status).
- As in the previous chapter we use QQ plots and confusion plots to get a better insight on the imputation performance of AGE variable.
- A special study of the role of model flexibility and imputation strategy is done using SEX variable.
- Imputation performance of the categorial SEX variable are also reported in terms of the operating characteristics of the classifier.
- Experiments with the SEX variable are computed with different sample sizes in order to evaluate the relative efficiences of different methods.

9.1 Description of data

This is a hierarchical dataset with observations on household, family and individual levels. There are a total 124106 observations and 774 variables, including many categorial and multiresponse variables. We have omitted all derived variables, and coded both categorial and multiresponse variables as new indicator variables. In addition some special values like -8 (no answer) or -9 (not applicable) are coded with new indicator (class) variable. In the case of continuous variables, the special value is also marked as missing, and its handling depends on the imputation model as explained later. See Appendix A9.6 for a list of variables which are treated in the experiments as continuous.

Details of the data preparation including some discussion about practical problems of the data are given in Appendix A9. The result of the preparation gives us a total of 5628 variables, where there are 5588 indicator variables (of which 61 are indicators for special values of continuous variables), 40 continuous variables, and one sampling weight (HHWT03) that refers to proportional sampling of households. The sampling weight varies between 225-750 and it is distributed as shown in Figure 9.1, indicating that a typical observation corresponds roughly to 500 households in the real world. More details of the dataset can be found from the documentation of the dataset which is available from the UK Data Archive internet site[3].

[3] www.data-archive.ac.uk/findingData/snDescription.asp?sn=5500 (referenced 10.05.2007)

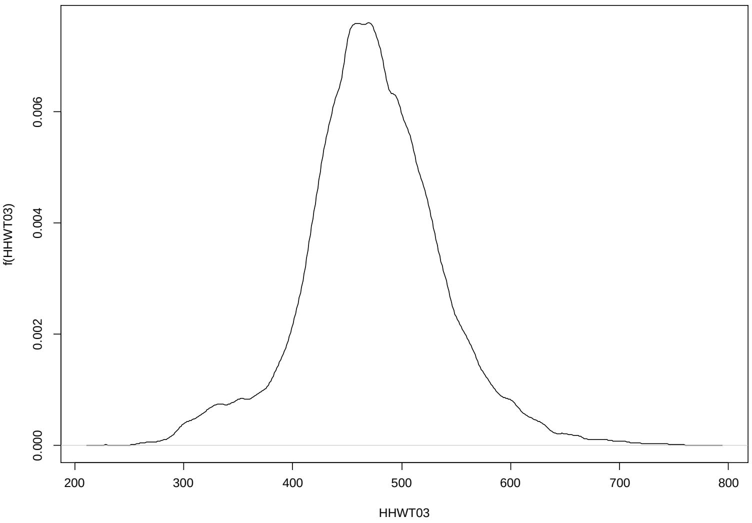


Figure 9.1: Density of household weights (variable HHWT03).

9.1.1 Data generator and missingness

Except of the variable selections for the imputation of AGE and one experiment with the SEX variable, we have used a random sample of n = 4000 observations in the tests. The sample is drawn without replacements under the specified incompleteness mechanism, and typically 500 sample-imputation repetions are taken in the AGE experiment and 110 in the SEX experiment to estimate the expected evaluation measures reliable. Sampling weights were ignored while sampling, but they were included in the sample for the computation of the weighted evaluation results.

The imputation experiments for the AGE and SEX variables were done separately using a different missingness generator, but in both cases about 39%-40% of observations were marked as missing on average.

In the case of the AGE variable, missingness was generated using an unweighted MARSTA (marital status) variable. Let Y=AGE and X=MARSTA. Then the missingness can be described as shown in Table 9.1. Note that these missingness values have been selected for the purpose of methodological testing. In the real world such high missigness might not be realistic for this particular dataset.

As we can see, the missingness probabilities differ for the values of marital status X, implying that this is a MAR type of missingness. When the sampling weights are applied, the distributions of observed and missing parts of the AGE variables are as shown in Figure 9.2.

Description of x	$\Pr(x)$	$\Pr(\text{"}Y\text{ is missing"}\mid x)$
1 (single)	0.44193	0.55
2 (married and living with)	0.41702	0.2
3 (civil partnership)	0.02066	0.4
4 (married and separated)	0.06135	0.5
5 (divorced)	0.05857	0.75
6 (widowed)	0.00042	0.6
7 (civil partnership and separated)	0.00005	0.45

Table 9.1: Missingness generator of Y=AGE with respect to covariate X=MARSTA.

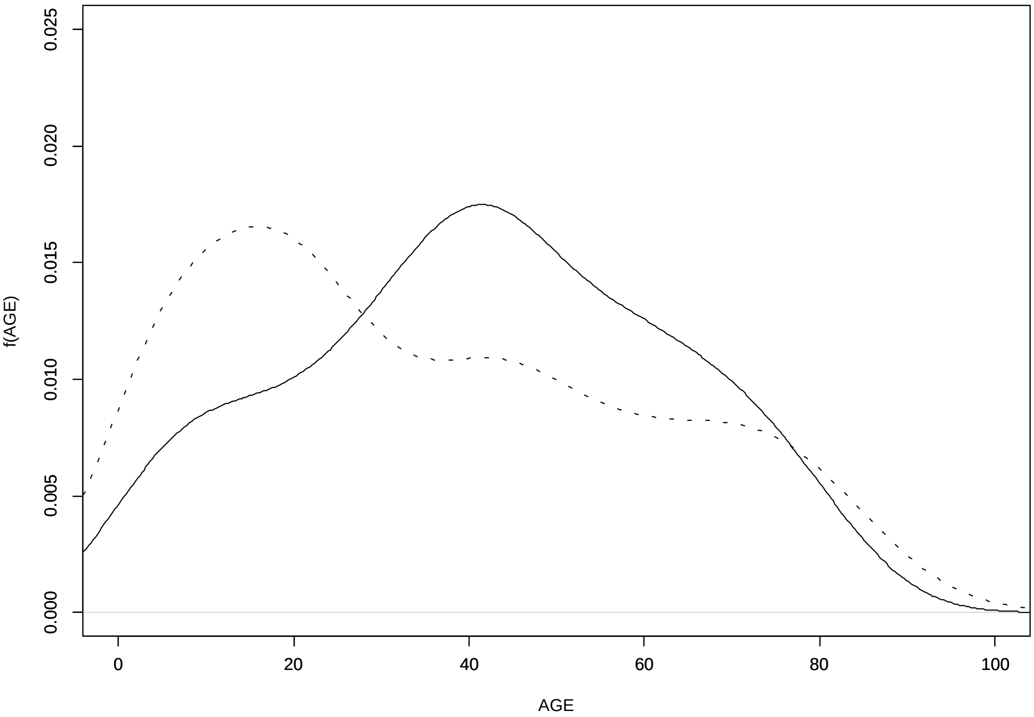


Figure 9.2: Distributions of weighted observed (solid line) and missing (dashed line) age.

In the experiments with the SEX variable, missingness is generated using the job status variable (FTPTWK) and the state benefits variable (BENFTS) as explained in Appendix A9.4. The application of missingness was selected such that the prior probabilities of SEX classes differ clearly from incomplete cases. This is clearly seen in Table 9.2 that shows the unweighted probabilities of the two classes of SEX for complete and incomplete cases.

The use of MAR type of missingness for both the AGE and SEX variables was chosen because we want to evaluate the role of predictions by different models. For example, a simple random donor imputation follows the distribution of observed data, which in our case differs from the distribution of missing ones. Thus a better model must be able to predict the missing data in terms of covariates.

9.2 About imputation models

Evaluation of many imputation methods with real-world data is a challenging task. Different methods require different setups and one should try to use the best setup for each of the methods. To summarize, we must select for each of the methods

i) the best way of presenting the selected variables. This is done with preprocessing.

ii) the best variables, because different methods can utilize different kinds of imputation.

iii) the best parametrizations for smoothing, number of clusters, etc.

iv) settings related to special values and missing data.

v) postprocessing of results, because different methods may have different mechanisms in how they present the results.

Clearly this is an overwhelming task, but we can only try to do our best. A semiautomatic variable selection method was used to find the best set of variables for each of the methods. Some manual work was also done to find the most effective ways of preprocessing and coding of variables. Finally special experiments were conducted to investigate the role of parametrizations.

9.2.1 Coding and preprocessing

Coding of variables and preprocessing must be done before using any of the methods. Thus it is done also before variable selection and the setting of model parameters. We first describe the coding of categorial variables and special values, because coding has to be done before the preprocessing phase (it introduces the missing data values in continuous covariates).

The role of coding is to present data in a form that is most suitable for the methods. While real (or integer) valued variables require only preprocessing (as described later), the situation is not the same for categorial and multiresponse variables. In addition, special values of certain variables are coded separately. The following codings were used in the experiments:

i) Dummy coding of categorial and multiresponse variables, where class variable

$$C \subset \{c_1, \ldots, c_m\},$$

	MALE	FEMALE
Pr(SEX|observed)	0.41	0.59
Pr(SEX|missing)	0.60	0.40

Table 9.2: Unweighted probability of the two classes of SEX for complete and incomplete cases.

where C contains more than one element only when C is a multiresponse variable. This is replaced with dummy indicator vector

$$\mathbf{x}' = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, \text{ where } x_i = \begin{cases} 1, & \text{if } c_i \in C \\ 0, & \text{otherwise,} \end{cases}$$

ii) Coding of special values is straightforward for categorial variables as they are handled as ”ordinary” classes, by coding them as dummy indicators. For continuous variables special values are coded as dummy indicators. This is followed by replacing the special values in a continuous variable with missing data values.

The codings have been applied as described in Table 9.3.

AGE experiments *Continuous variables*: dummy coding of the following special values: - values -8/-9 for all continuous variables - values 96/97 for EDAGE variable - value 99 for variables TOTUS1, USUHR, POTHR, UOTHR, TOTUS2, TOTAC1, ACTHR, ACTPOT, ACTUOT, TOTAC2, and OVHRS. *Categorial variables*: dummy coding.
SEX experiments Coding as in AGE experiments with the exception that variable SEX is coded as one variable (with values -1=MALE, 1=FEMALE).

Table 9.3: Codings for AGE and SEX experiments.

In our experiments the following preprocessing operations were used:

min-max equalization to hypercube $[a, b]^p$, thus the components of $\mathbf{x}$ are equalized as

$$\tilde{x}_i = (b - a) * \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} + a, \text{ where } i = 1, \ldots, p$$

where minimum and maximum are computed either from the complete part of data or from all the available observations of x_i (this is described later).

centering mean is subtracted as follows

$$\tilde{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$$

normalization scales the length of data vectors in such a way that they are on the surface of a unit hyper-sphere,

$$\tilde{\mathbf{x}} = \mathbf{x}/||\mathbf{x}||. \tag{9.1}$$

Normalization is straightforward for complete observations. However, the following phases are done for incomplete observations

1. replace all missing data values by zero values, forming observation $\mathbf{x}_{\text{zero}}$
2. normalize observation $\mathbf{x}_{\text{zero}}$ using Equation (9.1).
3. if the original observation had a missing data value in the target (which is either AGE or SEX depending on the experiment) then replace the zero value of the target in the preprocessed $\tilde{\mathbf{x}}_{\text{zero}}$ by the missing data value.

The preprocessing phases differ in AGE and SEX experiments. Further, different preprocessings are applied for different methods in the AGE experiments.

In AGE experiments, preprocessing, which is done after coding of variables, depends on the imputation method. The preprocessing groups are: linear regression methods (L,M and L,R), joint $(Y, \boldsymbol{X})$ clustering using TS-SOM, and other methods. Table 9.4 lists the preprocessings applied for these groups.

Method group	**Preprocessing phases**
Linear regression	*Categorial variables*: none *Continuous variables*: 1. min-max equalization to $[0, 1]$ $(a = 0, b = 1)$ using complete data statistics. 2. centering using observed data statistics.
Joint $(Y, \boldsymbol{X})$ clustering using TS-SOM	*Categorial variables*: none *Continuous variables*: min-max equalization to $[0, 1]$ by computing min-max statistics from all available observations.
Other methods	*Categorial variables*: none *Continuous variables*: min-max equalization to $[0, 1]$ using complete data statistics.

Table 9.4: Preprocessing phases for AGE experiments.

In the experiments for the SEX variable only one chain of preprocessing is used. Table 9.5 contains the preprocessings for the SEX experiments.

Preprocessing phases
1. *Categorial variables and special value indicators*: replacement of values 0 and 1 of dummy indicators with values -1 and 1.
2. *Continuous variables*: min-max equalization to $[-1, 1]$ by computing min-max statistics from all available observations.
3. *All observations*: normalization.

Table 9.5: Preprocessing phases for SEX experiments.

9.2.2 Variable selection, when imputing AGE variable

We made an attempt to use automatic variable selection to find the best predictive models for the AGE variable. This was not as easy as one might think. Since the goal was to find the best models under the best parameter setting we had a dilemma regarding whether to optimize first the selection of variables, their codings, or model parameters. In addition, we needed to decide what criteria to use to measure the ”optimality” of variables. To overcome these difficulties the following simplified forward selection procedure was used.

1. Fix parametrization of model as described in Section 9.2.4 (with the exception that the K-Means methods use only 32 cells) and fix data coding and preprocessing as described in Section 9.2.1 for the AGE experiments.
2. Set the maximum number of covariates $i_{\max}$ and clear covariate set $\Omega_{\boldsymbol{X}} = \emptyset$.
3. Loop through all the variables (excluding AGE and the sampling weights HHWT03) which are not in the set $\Omega_{\boldsymbol{X}}$ and for each of these variables do:

   3.1 Set $l = 0$ and set the maximum number of repetitions $l_{\max}$ to 3 if the imputation model is non-identifiable (cell methods) or if the imputation strategy is random or donor, otherwise set $l_{\max} = 1$.

   3.2 Train the model from a test data set.

   3.3 Impute the missing data values in both the training set and the test set using the covariates in $\Omega_{\boldsymbol{X}}$ and the current covariate.

   3.4 Compute the error criteria (over imputed data values) for the training data set and the test data set, store the results to Err_l^{train} and Err_l^{test}.

   3.5 Increase l by 1 if $l < l_{\max}$ GOTO 3.2

   3.6 Compute the averaged values of error criteria:

$$\mathrm{Err}^{train} = \frac{1}{l_{\max}} \sum_{i=1}^{l_{\max}} \mathrm{Err}_i^{train}, \ \mathrm{Err}^{test} = \frac{1}{l_{\max}} \sum_{i=1}^{l_{\max}} \mathrm{Err}_i^{test}.$$

4. Add variable producing minimum averaged training data set error criteria to $\Omega_{\boldsymbol{X}}$. Note that the definition of minimum depends on the error criteria as described below.
5. If $\#(\Omega_{\boldsymbol{X}}) < i_{\max}$ GOTO 3.
6. Select a covariate set for which the averaged test data set error criteria is minimum.

The simplified variable selection procedure was run with 523 original variables and 2000 observations of which 1000 belonged to the training data set and 1000 to the test data set. Note that both the training and the test data sets contain missing data values in AGE. Further, the data sets were generated by drawing a simple random sample of size 2000 without replacements from 124106 observations by splitting the data set into half. Therefore there are no duplicate observations in the training and the test data sets.

The procedure was run twice for each imputation method. The first run was made a using single objective: minimization of Kolmogorov-Smirnov (K-S) error criteria with 20 iterations. A minimum in Phases 4 and 6 is defined by the lowest value of the corresponding K-S error criteria. In the second run a combined K-S and mean squared error (MSE) rank error criteria was used with 10 iterations. Thus the K-S criteria was computed as in the first run, and in addition the mean squared error criteria was computed. In Phase 4 the rank values for K-S and MSE results were computed by sorting the error criterias from the lowest values to the highest values. A minimum was defined by a variable set for which the sum of the KS rank and the MSE rank was the lowest, this applied also to Phase 6.

Variable selection proved to be computationally expensive, as it took approximately one week (when measured in terms of total computing time used by two AMD Athlon 2400+ MP processors/2GB RAM, AMD Athlon64 3000+ processor/2GB RAM and Intel Pentium 4 Prescott 3 GHz processor/512MB RAM) to find ”optimal” sets for each of the models.

Unfortunately, the results of our forward selection were not fully satisfactory. This may be partially due to small sample size and too few imputation repetitions (parameter $l_{\max}$). We realized that for some methods the chosen variables were better, while for some methods the simplified procedure picked inferior predictors. Therefore the final selection of variables was done in part manually. The variable sets were evaluated (with sample of size $\mathsf{n} = 4000$) using the K-S and MSE evaluation criteria for imputation. Then the following steps were taken to create the final sets for the imputation experiment

i) A set for a method with random or donor strategy yielding bad imputations was, if possible, replaced by a set for the corresponding method with mean strategy, and vice versa.

ii) A better set from the two variable selection runs (KS criteria vs rank criteria) was chosen manually based on the evaluation results.

iii) The final sets for nearest neighbour imputation and joint $(Y, \boldsymbol{X})$ clustering using TS-SOM with smoothing performed badly. Therefore these final sets were created manually by merging the best set for linear regression with mean strategy (L,M) and the best set of kernel regression with random strategy (K,R). Further, variable HNPEN (number of people in a household who are of pensionable age) was added to the constructed sets.

The final selection of variables for each of the methods is given in Tables 9.6 and 9.7. See Appendix A9.5 for the descriptions of variables. From the selected variable sets one can notice that special value -9 (not applicable/NA) indicators are used by some methods. Therefore one may expect that the distributions of age for the observations with NA value in a variable and observations with non-NA value in the same variable are different. As an example, variables containing information on health problems and qualifications are used by some methods.

Variable	**Methods**																					#USE
	LM	LR	NM	KM	KR	TM	TR	TD	TM^s	TR^s	CM	CR	CD	TJM	TJR	TJD	TJM^s	TJR^s	CJM	CJR	CJD	
ACTWKDY2_4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
ACTWKDY2_7	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
ATTEND_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	1
CCTC5_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
CMBDEG01_16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
CMBDEG01_8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
CMBDEG05_-9	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
CMBMAIN_10	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
CMBMAIN_7	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	1
CRY01_59	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
EDAGE	-	-	-	-	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	4
ENROLL_-9	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4
EVERWK_-9	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4
EVERWK_1	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	1
FAMLY031_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
FUTUR13_2	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	1
GCSEFUL1_-9	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
HEAL02_1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
HEAL03_2	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	1
HEAL04_10	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
HEALPB01_3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	1
HEALYR_2	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
HLDCMP6_2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
HLDCMP6_7	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	1
HLDCMP6_9	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	1
HNFTIME	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
HNPEN	+	-	+	-	-	-	-	-	-	+	+	+	-	+	-	-	+	+	-	-	-	**8**
HNWKAGE	-	-	+	-	-	+	+	-	+	-	-	-	-	+	-	+	+	+	-	-	-	**8**
HNWOTH05_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	1
HOME_-9	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	2
HOME_4	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
HSNGNI_-9	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	1
ICOD92_331	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
ICOD92_382	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
ICOD92_423	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
JBAWAY_2	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
JSADUR_8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
LEFTYR_-9	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	2
LEVQUAL6_1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
LIMITA_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
LIMITA_1	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	1
LIVWTH_-9	-	-	-	+	-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	3
LIVWTH_1	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
LKTIMB_3	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
LLORD_6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
LNGLIM_1	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	1
LOOK4_2	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
LOOKM2_6	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
M3CRYO_52	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	1
MAINDRV_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1

Table 9.6: Variables used by the imputation methods (continues on the next page). Column #USE is a count for how many times each variable is used.

Variable	**Methods**																					#USE
	LM	LR	NM	KM	KR	TM	TR	TD	TM^s	TR^s	CM	CR	CD	TJM	TJR	TJD	TJM^s	TJR^s	CJM	CJR	CJD	
MAINDRV2_5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
MANAGLR_-9	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	1
MANAGLR_1	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
MARCHK_1	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4
MARSTA_1	+	+	+	+	+	-	-	-	-	-	-	-	+	-	+	-	+	+	+	+	+	**12**
METHAL02_2	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	1
METHMP04_7	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	1
NATO_84	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
NOLWM_-9	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
NOLWM_3	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	1
NUMAS_1	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	1
NVQSVQ_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
NVQUN_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	1
OYCIRC_-9	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	1
OYCIRC_10	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	1
OYCRY_1	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
OYSOLO_1	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	1
QGCSE41_1	-	-	-	-	-	+	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	3
QGNVQ_-9	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	1
QUALCH53_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
QUALS601_16	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	1
QUALS602_18	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	1
QUALS602_21	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	1
QUALS603_17	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	2
QUALS604_8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
QULHI4_-8	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	1
RELBUS_2	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4
RELH06_0	-	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	2
RELH06_3	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	2
RELHRP6_3	+	-	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	-	-	**15**
RELIG_1	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4
RESTME_6	+	-	+	+	+	+	+	-	+	+	-	-	-	-	-	+	+	+	-	-	-	**11**
SCHM04_66	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
SECJOB_2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	1
SEX_1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
SNGDEG_10040303	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
SNGDEG_18010100	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	1
SNGDEG_6010201	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
SNGDEG_6040200	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	2
SNGDEG_7020402	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
STAT_1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
SUBCOD1_21,1	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
TEACH41_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
TECLEC4_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
TEN1_2	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
TOTUS1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
TPBEN31_-9	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4
TPBEN31_4	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	1
TPBEN32_3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	3
TRSITE_9	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
TYPVCL3_1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	1
UNDABL_2	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	1
UNDEMP_-9	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	1
USEVCL_2	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4
XR01_-9	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	+	5
XR01_3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	1
XR02_-9	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	4

Table 9.7: Variables used by imputation methods.

9.2.3 Variable selection for the imputation of SEX variable

After the complex and rather disappointing results of automatic variable selection, much simpler procedure was used to select covariates to predict the categorial SEX variable. Now the same variables were used for all methods.

The selection was done by taking 16 variables that had the best Pearson correlation with SEX=MALE, and the obtained variables are listed in Table 9.8. It should be noted that this might be an inadequate way to do variable selection, but it was used because of its simplicity.

There are some immediate concerns about this variable selection. First, it does not eliminate collinearity between the covariates, and it favours linear models.

Therefore, some other studies were done to look for suitable variables. Clustering of variables was tried with the TS-SOM and K-Means algorithms with a different number of cells. This can be done by transposing a data set and clustering it. Then the clusters in which MALE and FEMALE indicators were classified were inspected. From the constructed clusterings the ones yielding a suitable number of covariates in the MALE and FEMALE clusters were tried with some of the cell imputation methods. However, the imputation results turned out to be bad typically both at unit and at distribution level. Secondly, the indicator variables for categorial classes or special values, with the highest mutual information between the SEX=MALE indicator, were searched. This study verified that the indicators among the 16 selected variables have a fairly high amount of mutual information with the indicator for the MALE class. This may indicate that the relationship between 'best available information' and the MALE class is reasonably linear. Better results may be obtained by doing proper variable selections for each imputation method.

Variable	**Description**
RELH06_1	Relationship to the head of household [Head of household]
RELH06_0	Relationship to the head of household [Spouse]
XR00_1	Relationship to person 0 [Spouse]
USUHR	Usual hours worked excluding overtime
TOTUS1	Total usual hours worked excluding lunch breaks (no overtime)
TOTUS2	Usual hours worked including overtime
TOTAC1	Total actual hours worked (no overtime)
XR00_-9	Relationship to person 0 [Not applicable]
RELHRP6_1	Relationship to household reference person [Spouse]
ACTHR	Actual hours worked excluding overtime
TOTAC2	Actual hours worked including paid and unpaid overtime
TPBEN31_6	Type of 1. benefit claimed [Child benefit]
TPBEN31_-9	Type of 1. benefit claimed [Not applicable]
DAYSPZ	Number of different days per week worked
FTPTWK_2	Whether full or part time in main job [Part-time]
BENFTS_1	Whether claiming any State Benefits/Tax credits [Yes]

Table 9.8: Variables that were used to impute SEX.

9.2.4 Parametrization of models

The selection of optimal training parameters for each of the methods is quite challenging. The main tasks are to

- select the level of smoothing for the nonparametric regression methods.
- select the number of clusters for the TS-SOM and K-Means methods.

Some of the settings were specific to experiments and are described later. The more generic settings that where shared by most of the experiments were as follows:

Simple methods (B,M), (B,R), and (B,D) utilize sampling weights. Recall the weighted imputation procedures from Section 8.2. Note that other imputation methods do not use sampling weights.

Nearest neighbour imputation uses smoothing parameter $k = 1$.

Kernel regression methods use smoothing $\lambda = (\mathsf{N}^{obs})^{-1/(4+dim(\boldsymbol{X}))}$, where $dim(\boldsymbol{X})$ is the number of covariates.

Cell methods use $\mathsf{n}_c = 64$ cells.

TS-SOM cell methods use 2-D lattice topology.

9.2.5 Handling missing data in the training phase

Coding of special values imposes missing data values in covariates. For the SEX experiments this is not a "problem", because the normalization phase in preprocessing produces a data set which has no missing data values in covariates. However, with the AGE experiments some modifications to model trainings are required. The following modifications are done:

Linear regression: after the coding and preprocessing phases the missing data values in covariates are replaced by zero values.

Nonparametric regression: distances are computed using jointly observed parts of observations. Further, scaling of distance is done to compensate the reduced number of covariates. The formula for computing squared distance between two (possibly incomplete) observations with indexes i and j is

$$sd(i,j) = \frac{\dim(\mathbf{z})}{\dim(\mathbf{z}_{i,j}^{obs})} * sd^{comp}(i,j), \tag{9.2}$$

where $\dim(\mathbf{z}_{i,j}^{obs})$ is the number of variables which are observed for both $\mathbf{z}_i$ and $\mathbf{z}_j$ and $sd^{comp}(i,j)$ is the sum of squared differences over jointly observed variables. As a special case, if there are no jointly observed variables, then $sd(i,j)$ is set to a very large value.

K-Means cell methods: positions of centroids are computed using all observed data values for each variable. Classification of observation to a cell is done by computing the distance between the observed components of observation and the centroid using Equation (9.2).

Standard clustering using TS-SOM: the incomplete data training algorithm is used by the standard $\boldsymbol{X}$ clustering TS-SOM methods to build a clustering model. However, an incomplete observation is deterministically classified to the closest cell. The distance between an observation and the centroid is computed using the jointly observed part of the incomplete observation and the cell centroid, by Equation (9.2). This is a less sophisticated strategy than the one used in joint $(Y, \boldsymbol{X})$ clustering using TS-SOM methods. However, the imputation results seem still to be quite good.

9.2.6 Postprocessing of results

Some postprocessing phases were required in the AGE and SEX experiments. The phases were applied to all methods (for some of the methods they do nothing).

The original AGE variable is in range [0,99] years and it is integer valued. However, imputed values may not be in the range. Further, they may be floating point values due to Gaussian noise terms used by the random strategies. Therefore the following postprocessing phases were done in the AGE experiments:

1. Truncating imputed values to range [0,99].

2. Rounding the truncated values down to the closest integer value.

In the SEX experiments the target is coded as a single variable (MALE coded with the value -1 and FEMALE with the value 1). However, most of the imputation methods yield values which are floating point values, and possibly even out of range [-1,1] due to random Gaussian noise. Therefore the imputed values of SEX were postprocessed as follows:

$$\text{CLASS} = \begin{cases} \text{MALE} & \text{if } Y^{imp} < 0 \\ \text{FEMALE} & \text{otherwise.} \end{cases}$$

9.3 Evaluation of the imputation results of the AGE variable

As stated before, the evaluation of imputation results is quite challenging. The order of performance between the methods depends on the evaluation criteria. Therefore we shall investigate the results from several viewpoints. The first set of results is computed for the (almost) real valued AGE variable. The second, set, which evaluates the imputation performance of the categorial SEX variable, will be described in Chapter 9.4.

As described earlier, a lot of work was done with the AGE variable to find the best possible covariables for the different imputation methods. The parameters for the methods were set as described in Section 9.2.4 with the exception that methods TJ,M^s and TJ,R^s use 256 cells. The variety in approaches complicates the evaluation also, making it difficult to point out the exact reason for why one method performs better than another. Thus the results should be understood as "what one obtains" when using the methods.

There are six classes of results, which will be presented in the following sections. The classes are

i) Preservation of moments of AGE.

ii) Preservation of quantiles.

iii) Preservation of conditionalized quantiles.

iv) Qualitative analysis of the preservation of distribution.

v) Preservation of the unit level MSE measure.

vi) Comparative evaluation between the distributional KS measure and MSE.

9.3.1 Preservation of moments of the AGE variable

As usual we shall evaluate the biases of the first two moments $\hat{\mu}^{imp}$ and $\hat{\tau}^{imp}$. These are computed using weighted estimators

$$\begin{aligned}\hat{\mu}^{imp} &= \frac{1}{S_w} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{mis}} W_j Y_j^{comp}, \text{ and} \\ \hat{\tau}^{imp} &= \frac{\mathsf{N}^{mis}}{(\mathsf{N}^{mis}-1)S_w} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{mis}} W_j (Y_j^{comp} - \hat{\mu}^{imp})^2,\end{aligned}$$

where $S_w = \sum_{j:Y_j \in \mathbf{D}_Y^{mis}} W_j$. And the biases are as usual

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\mu}^{imp}] &= \mathbb{E}[\hat{\mu}^{imp} - \hat{\mu}^{mis} | \mathsf{n}], \text{ and} \\ \mathbb{B}\text{ias}[\hat{\tau}^{imp}] &= \mathbb{E}[\hat{\tau}^{imp} - \hat{\tau}^{mis} | \mathsf{n}],\end{aligned}$$

where the missing data moments are defined as imputed data moments but the imputed values are replaced by missing values. In addition, another relative measure, Err% is computed to assist in the evaluation of results. These are defined as

$$\begin{aligned}\text{Err\%}(\hat{\mu}^{imp}) &= \mathbb{E}\Big[| \frac{\hat{\mu}^{imp} - \hat{\mu}^{mis}}{\hat{\mu}^{mis}} | * 100\% | \mathsf{n} \Big], \text{ and} \\ \text{Err\%}(\hat{\tau}^{imp}) &= \mathbb{E}\Big[| \frac{\hat{\tau}^{imp} - \hat{\tau}^{mis}}{\hat{\tau}^{mis}} | * 100\% | \mathsf{n} \Big].\end{aligned}$$

The actual results are, as described earlier, mean values over 500 repetitions of the sampling and imputation cycle. The results are summarized in Table 9.9.

From the results one can conclude that certain nonparametric methods yield best performances. The baseline methods fail quite badly, which is due to the MAR type missingness mechanism. The preservation of the first moment is the best with standard TS-SOM under simulated random strategy. However, when the second moment $\hat{\tau}^{imp}$ is included in the evaluation, the best method seems to be nearest neighbour.

It seems apparent that all the methods underestimate the second moment (variance of AGE), and all the regression methods underestimate the first moment (mean). This might explain why the "noisiest" model, nearest neighbour, is the best one for the second moment.

An overall impression about the imputation strategies is that simulated randomness (R) and donor (D) should be used to optimize the second moment of AGE.

	$\mu^{*mis} \approx 35.29$		$\tau^{*mis} \approx 585.80$	
Method	$\mathbb{B}$ias$[\hat{\mu}^{imp}]$	**Err**%$(\hat{\mu}^{imp})$	$\mathbb{B}$ias$[\hat{\tau}^{imp}]$	**Err**%$(\hat{\tau}^{imp})$
B,M	5.75(0.03)	16.3(0.11)	-585.80(0.74)	100.0(0.0)
B,R	5.94(0.04)	16.9(0.12)	-148.02(0.99)	25.2(0.15)
B,D	6.16(0.04)	17.5(0.13)	-125.42(1.08)	21.3(0.17)
L,M	-1.00(0.02)	2.8(0.04)	-121.23(0.62)	20.7(0.10)
L,R	-0.68(0.02)	2.0(0.05)	-48.54(0.83)	8.3(0.14)
N,M	-0.30(0.03)	1.7(0.05)	**-8.71**(1.55)	**4.9**(0.16)
K,M	-0.58(0.01)	1.6(0.03)	-130.31(0.49)	22.2(0.07)
K,R	-0.75(0.02)	2.1(0.04)	-71.57(0.59)	12.2(0.09)
T,M	**-0.16**(0.02)	**1.0**(0.03)	-127.24(0.76)	21.7(0.12)
T,R	**-0.05**(0.02)	**1.1**(0.04)	**-40.06**(0.84)	**6.9**(0.13)
T,D	0.29(0.02)	**1.2**(0.04)	**-34.85**(0.64)	**5.9**(0.11)
T,M^s	0.66(0.06)	2.6(0.15)	-202.57(2.32)	34.5(0.39)
T,R^s	**0.10**(0.02)	**1.1**(0.04)	-48.34(0.89)	8.2(0.15)
C,M	-1.11(0.02)	3.2(0.07)	-175.16(1.15)	29.9(0.19)
C,R	0.26(0.02)	**1.2**(0.04)	-55.97(0.82)	9.5(0.13)
C,D	-0.60(0.02)	1.9(0.05)	-42.55(0.77)	7.2(0.13)
TJ,M	**0.22**(0.04)	2.2(0.07)	-79.19(1.23)	13.5(0.21)
TJ,R	-0.57(0.04)	2.3(0.07)	-37.79(1.27)	**6.7**(0.19)
TJ,D	**0.20**(0.05)	2.4(0.08)	-42.50(1.36)	7.5(0.21)
TJ,M^s	-0.69(0.03)	2.4(0.07)	-91.14(1.01)	15.5(0.17)
TJ,R^s	-0.74(0.03)	2.4(0.07)	-50.10(1.00)	8.5(0.17)
CJ,M	-0.88(0.04)	3.0(0.09)	-122.05(2.07)	20.8(0.34)
CJ,R	-0.68(0.05)	2.9(0.10)	**-35.91**(2.12)	8.2(0.26)
CJ,D	-0.34(0.05)	2.4(0.08)	**-25.15**(1.95)	**6.7**(0.24)

Table 9.9: Biases of first two moments of imputed age. Standard deviations of estimates are shown in parentheses. See Section 3.6.3 for details.

9.3.2 Preservation of quantiles of the AGE variable

The preservation of the quantiles of the AGE variable gives us our first indication of the distributional performance of the various imputation methods. Also, when compared to moments, quantiles are more robust against outliers. We may, for example, compare the biases of the first moment $\hat{\mu}^{imp}$ and median $\hat{\xi}^{imp}_{0.5}$, as well as the biases of the second moment $\hat{\tau}^{imp}$ and quantiles $\hat{\xi}^{imp}_{0.25}$ and $\hat{\xi}^{mis}_{0.25}$. The weighted quantiles are defined as

$$\begin{aligned}\hat{\xi}^{mis}_q &= \hat{F}^{-1}_{Y^{mis}}(q)\\ \hat{\xi}^{imp}_q &= \hat{F}^{-1}_{Y^{imp}}(q), \ \text{q} = 0.05, 0.25, 0.50, 0.75, 0.95,\end{aligned}$$

where

$$\hat{F}_{Y^{imp}}(y) \;=\; \frac{1}{S_w} \sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{mis}} W_j I(Y_j^{comp} \leq y),$$

where $I(\cdot)$ denotes an indicator function, and $\hat{F}_{Y^{mis}}(y)$ is computed by replacing imputed values by missing values. Biases are defined as $\mathbb{B}\text{ias}[\hat{\xi}_q^{imp}|\mathsf{n}] = \mathbb{E}[\hat{\xi}_q^{imp} - \hat{\xi}_q^{mis}|\mathsf{n}]$.

The results of these biases are summarized in Table 9.10. As with moments, the baseline methods perform quite badly due to MAR missingess. As before, nearest neighbour seems to perform quite well. It also seems that cell methods with donor strategy are always among the best methods, as is the standard TS-SOM with simulated randomness.

A notable difference between the results of the preservation of moments and the current analysis of quantiles is that in the bias of quantiles there are as large differences as there were between the first and second moments. This reflects, of course, the sensitivity of second order measures, when compared to ”robust” quantiles.

	$q = 0.05$	$q = 0.25$	$q = 0.50$	$q = 0.75$	$q = 0.95$
$\xi_q^{*mis} \approx$	2.72	14.27	31.22	54.15	79.61
Method	$\mathbb{B}\text{ias}(\hat{\xi}_{0.05}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.25}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.50}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.75}^{imp})$	$\mathbb{B}\text{ias}(\hat{\xi}_{0.95}^{imp})$
B,M	38.32(0.03)	26.77(0.04)	9.82(0.06)	-13.11(0.06)	-38.57(0.05)
B,R	3.03(0.06)	12.15(0.05)	9.84(0.06)	1.49(0.06)	-3.25(0.07)
B,D	2.51(0.05)	11.14(0.06)	10.93(0.07)	3.49(0.07)	-3.38(0.06)
L,M	2.60(0.03)	-2.26(0.04)	**-0.18**(0.06)	-4.77(0.05)	-7.48(0.05)
L,R	-2.13(0.04)	1.50(0.04)	**-0.02**(0.05)	-1.98(0.05)	-3.20(0.06)
N,M	**0.30**(0.10)	-0.62(0.05)	**-0.42**(0.10)	**-1.03**(0.10)	**-0.67**(0.17)
K,M	4.42(0.03)	-2.28(0.06)	2.03(0.05)	-5.42(0.05)	-7.69(0.04)
K,R	-2.40(0.03)	0.84(0.03)	1.71(0.04)	-3.44(0.04)	-4.53(0.05)
T,M	3.50(0.04)	0.90(0.06)	6.95(0.08)	-2.87(0.10)	-6.58(0.06)
T,R	-0.85(0.04)	**0.03**(0.04)	2.10(0.05)	**-0.90**(0.05)	-3.94(0.05)
T,D	**0.11**(0.04)	0.37(0.03)	1.69(0.06)	**-0.40**(0.05)	**-2.44**(0.05)
T,M^s	6.60(0.13)	2.50(0.19)	7.67(0.07)	-4.54(0.08)	-9.28(0.14)
T,R^s	-0.57(0.04)	**0.04**(0.04)	3.29(0.05)	-1.51(0.05)	-3.83(0.06)
C,M	4.73(0.03)	-3.66(0.23)	4.73(0.08)	-9.47(0.10)	-11.00(0.14)
C,R	**-0.40**(0.04)	**-0.07**(0.05)	3.48(0.05)	-1.46(0.05)	-4.27(0.06)
C,D	**-0.03**(0.04)	**-0.22**(0.04)	**-0.03**(0.06)	**-0.81**(0.06)	-3.71(0.06)
TJ,M	**-0.39**(0.05)	**-0.21**(0.09)	2.62(0.12)	-3.12(0.14)	-3.98(0.15)
TJ,R	-0.65(0.05)	-0.27(0.06)	0.95(0.07)	**-1.34**(0.07)	-4.1(0.08)
TJ,D	**0.15**(0.04)	**0.09**(0.04)	2.56(0.10)	**0.11**(0.10)	-3.96(0.09)
TJ,M^s	1.40(0.05)	0.49(0.10)	0.77(0.09)	-5.20(0.07)	**-2.79**(0.07)
TJ,R^s	0.65(0.04)	-0.54(0.06)	**0.42**(0.07)	-3.21(0.06)	**-2.42**(0.06)
CJ,M	3.51(0.09)	-1.05(0.14)	1.15(0.07)	-5.58(0.08)	-6.26(0.22)
CJ,R	**-0.41**(0.05)	-0.57(0.07)	1.13(0.06)	-2.19(0.08)	-3.06(0.17)
CJ,D	**0.06**(0.06)	**-0.16**(0.06)	**0.07**(0.07)	**-0.52**(0.08)	**-2.72**(0.15)

Table 9.10: Quantiles (ξ_q^{*mis}) and biases $\mathbb{B}\text{ias}(\hat{\xi}_{0.05}^{imp})$ for the AGE variable. Standard deviations of error estimates are shown in parentheses.

9.3.3 Preservation of conditionalized quantiles of the AGE variable

The preservation of quantiles was computed over the marginal distribution of the imputed AGE variable. This gives a very limited viewpoint of the performance of the different methods. As we have seen many times in this thesis, there are two ways to preserve marginal distributions, one using a "correct" model and "correct" level of simulated randomn noise; and the other using flexible models, where the residual noise is mixed with the model.

Now we are interested in seeing whether this marginal performance on quantiles extends to joint distribution of AGE and other variables. As an example of this we use a 7-category variable MARSTA (marital status) that takes the values

$$\text{MARSTA} = \begin{cases} \text{1 (single)} & \text{44.193\% of observations} \\ \text{2 (married and living with)} & \text{41.702\%} \\ \text{3 (civil partnership)} & \text{2.066\%} \\ \text{4 (married and separated)} & \text{6.135\%} \\ \text{5 (divorced)} & \text{5.857\%} \\ \text{6 (widowed)} & \text{0.042\%} \\ \text{7 (civil partnership and separated)} & \text{0.005\%} \end{cases}$$

Using again the five quantiles (0.05, 0.25, 0.50, 0.75, 0.95), we are now interested in finding out how the distribution of AGE behaves when conditionalized with the five first MARSTA classes. The "true" behaviour is visualized in Figure 9.3.

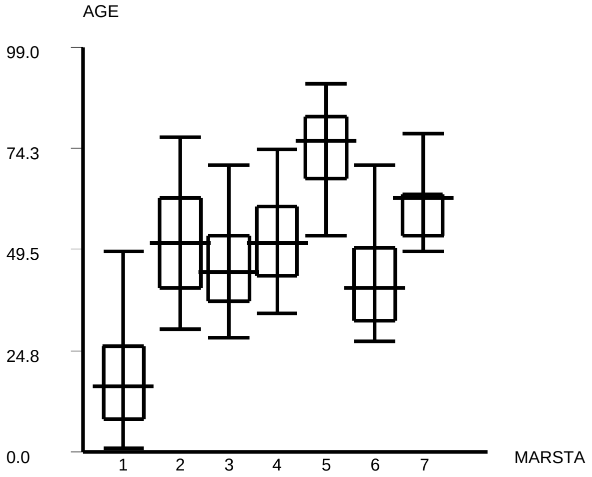


Figure 9.3: Visualization of weighted f(AGE|MARSTA) using box plots.

The focus is on the first five classes, which have 99% of all mass. Figures 9.4-9.9 illustrate the box plots of 0.05, 0.25, 0.50 (median), 0.75, and 0.95-quantiles for AGE^{imp}|MARSTA $= x$ where $x = 1, \ldots, 5$. The quantiles of AGE^{mis}|MARSTA have been plotted (dashed line) for comparison.

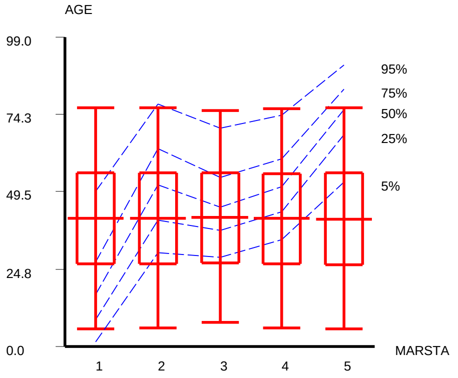


Figure 9.4: Preservation of $f(\text{AGE}^{mis}|\text{MARSTA})$ for B,R method.

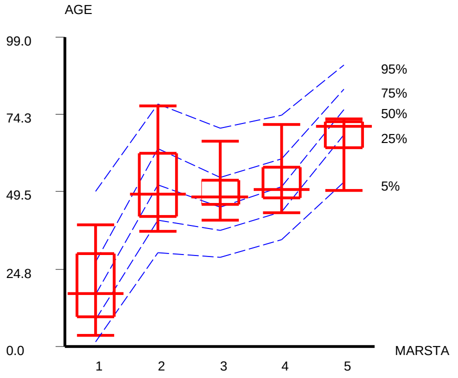


Figure 9.5: Preservation of $f(\text{AGE}^{mis}|\text{MARSTA})$ for L,M method.

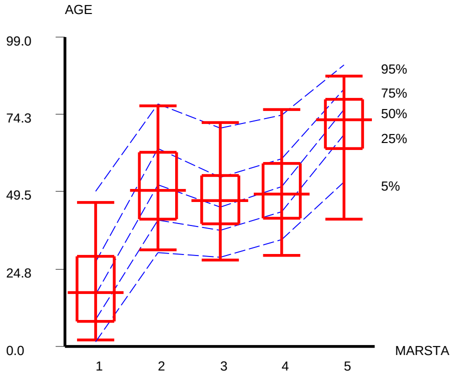


Figure 9.6: Preservation of $f(\text{AGE}^{mis}|\text{MARSTA})$ for TJ,R^s method.

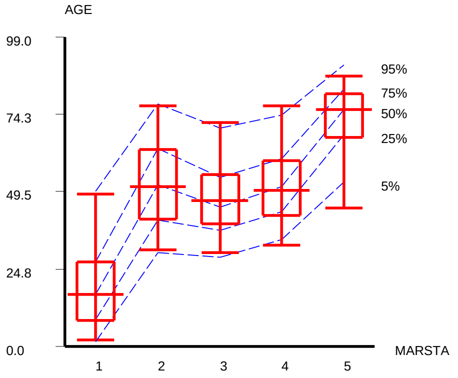


Figure 9.7: Preservation of $f(\text{AGE}^{mis}|\text{MARSTA})$ for N,M method.

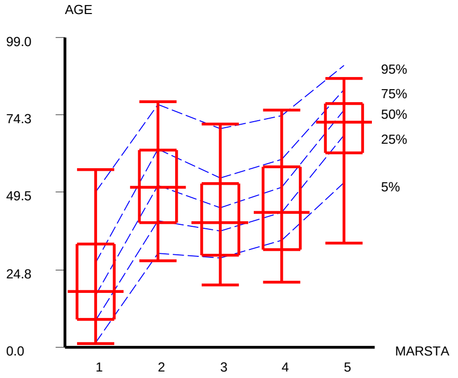


Figure 9.8: Preservation of $f(\text{AGE}^{mis}|\text{MARSTA})$ for T,D method.

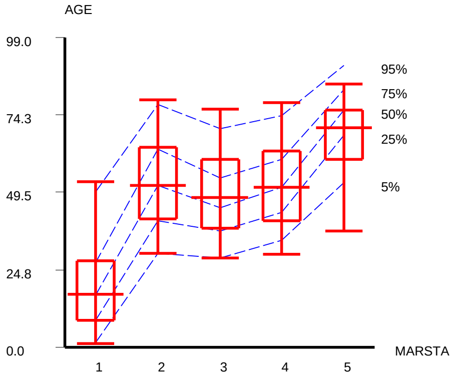


Figure 9.9: Preservation of $f(\text{AGE}^{mis}|\text{MARSTA})$ for CJ,D method.

The results indicate more clearly how some of the methods are able to preserve joint distributions while some are not.

From the results we can see that the results of baseline B,R are independent of covariates. Linear regression preserves the median quite well but fails on 5% and 95% quantiles. The nearest neighbour imputation, the joint $(Y, \boldsymbol{X})$ clustering version of TS-SOM with random strategy TJ,R^s and K-Means with donor strategy give the visually best results. The standard clustering TS-SOM method with donor strategy experiences problems for MARSTA values 3, 4, and 5. Namely, it results in underestimation of many conditional quantiles.

9.3.4 Qualitative analysis of the preservation of distribution of the AGE variable

We continue by doing an even deeper analysis of the chosen three methods: baseline (B,R), linear regression (L,M), and joint $(Y, \boldsymbol{X})$ TS-SOM clustering (TJ,R^s). The results are given in a form of confusion plots, density plots and QQ plots. Sampling weights have been applied in the estimation of the plots. As before, we try to compare the distribution of missing and imputed values.

The confusion plots in Figures 9.10, 9.12, and 9.14 reveal that the correlation between imputed and missing values for the baseline method is zero. This is expected as covariate information is not used. According to these plots, the performance for the linear regression method and the TS-SOM method is approximately the same.

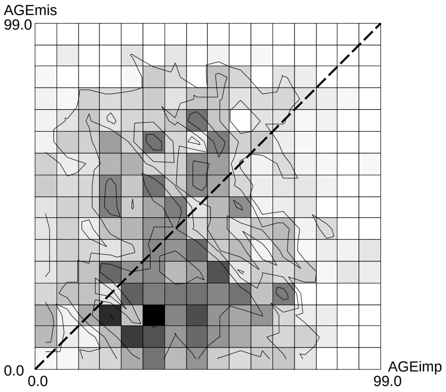


Figure 9.10: Confusion plot (pdf) between AGE^{imp,B,R} and AGE^{mis}.

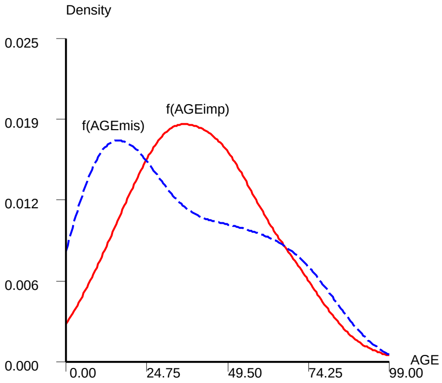


Figure 9.11: Density $f(\text{AGE}^{imp,B,R})$ versus $f(\text{AGE}^{mis})$.

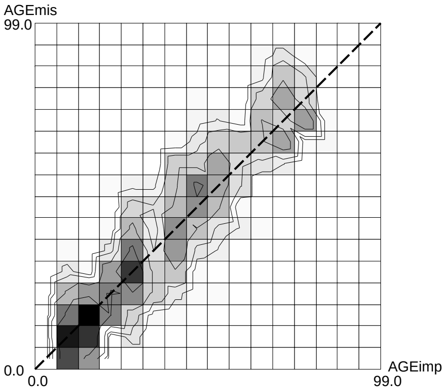


Figure 9.12: Confusion plot (pdf) between AGE^{imp,L,M} and AGE^{mis}.

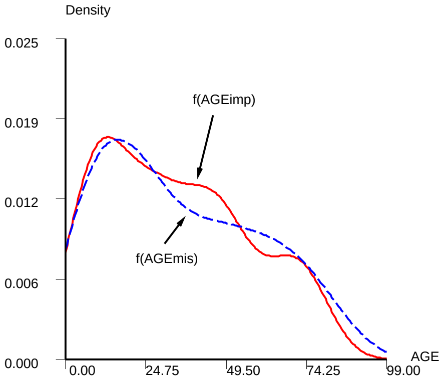


Figure 9.13: Density $f(\mathrm{AGE}^{imp,L,M})$ versus $f(\mathrm{AGE}^{mis})$.

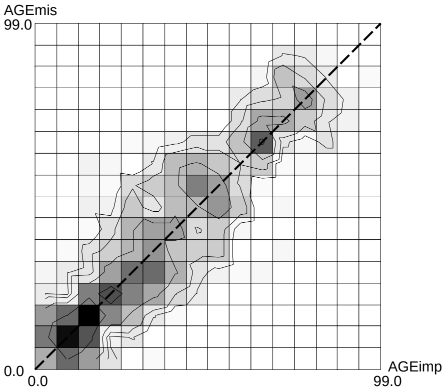


Figure 9.14: Confusion plot (pdf) between AGE$^{imp,TJ,R^s}$ and AGE^{mis}.

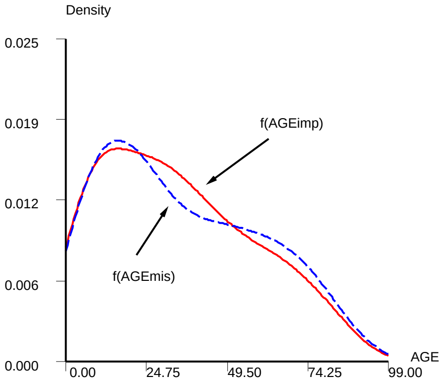


Figure 9.15: Density $f(\mathrm{AGE}^{imp,TJ,R^s})$ versus $f(\mathrm{AGE}^{mis})$.

A finer analysis of distribution preservation is easier via kernel density plots depicted in Figures 9.11, 9.13, and 9.15. From the density estimate for the baseline method it is obvious that distribution $f(\mathrm{AGE}^{mis})$ is not well preserved. Namely, the Gaussian assumption is wrong and the moments of imputation distribution are severely biased: expectation is too high and variance is underestimated. Linear regression with mean strategy recovers the underlying distribution considerably better than the baseline method. Linear regression with mean strategy as well as the joint $(Y, \boldsymbol{X})$ clustering version of TS-SOM with random strategy seem to perform equally well.

Figure 9.16 depicts QQ plots for the three methods. From the first plot one can conclude that the small quartiles for the imputed values of the baseline method are too high (deviation below diagonal line). This is obvious because the expectations of AGE^{obs} are higher than AGE^{mis}. The plot for linear regression with mean strategy seems to reveal that at least the right tail of AGE^{mis} is not well preserved. One reason for this may be that noise is not modelled, in other words, the variance is underestimated. The plot for joint $(Y, \boldsymbol{X})$ clustering with the TS-SOM method seems to be close to the diagonal line. This means that distribution is well preserved.

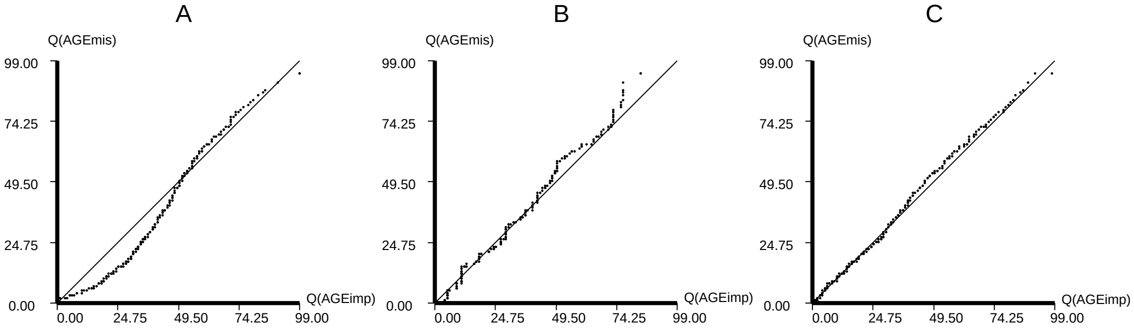


Figure 9.16: A: QQ plot for B,R method, B: QQ for L,M method, C: QQ for TJ,R^s method.

9.3.5 Preservation of weighted unit level MSE of the AGE variable

Recall from Section 8.4.4 the definition of expected MSE

$$\mathbb{E}[\hat{mse}] = \mathbb{E}[\sum_{j:\boldsymbol{X}_j \in \mathbf{D}^{mis}} \overline{W}_j (Y_j^{imp} - Y_j^{true})^2] = A + B + C + \underbrace{d_1 + d_2 + d_3 + d_4}_{D},$$

where

$$\begin{aligned} A &= \mathbb{E}[\sum_j \overline{W}_j (Y_j^{mis} - \hat{\mu}^{mis})^2], \text{ weighted variance of missing } \mathrm{Y_j^{mis}} \\ B &= \mathbb{E}[\sum_j \overline{W}_j (Y_j^{imp} - \hat{\mu}^{imp})^2], \text{ weighted variance of imputed values} \\ C &= -2\mathbb{E}\Big[\hat{\mathbb{C}\mathrm{ov}}[Y^{imp}, Y^{mis}]\Big], \text{ weighted covariance between missing and imputed values} \\ D &= \mathbb{E}[(\hat{\mu}^{mis} - \hat{\mu}^{imp})^2], \text{ global estimation bias,} \end{aligned}$$

where

$$\begin{aligned} d_1 &= (\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{imp}])^2, \text{ expected bias} \\ d_2 &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}], \text{ variance of } \hat{\mu}^{imp} \\ d_3 &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{mis}], \text{ sample variance of } \hat{\mu}^{mis} \\ d_4 &= -2\mathbb{C}\mathrm{ov}[\hat{\mu}^{imp}, \hat{\mu}^{mis}], \text{ covariance between } \hat{\mu}^{imp} \text{ and } \hat{\mu}^{mis}. \end{aligned}$$

Correlation is also computed as follows

$$Correlation = \begin{cases} -0.5 * C/\sqrt{A * B}, & \text{if } C > 0, \\ 0, & \text{if } C = 0. \end{cases}$$

The results are summarized in Table 9.11. Note that the expectation of weighted variance estimator for missing values (term A) is the same for all methods, thus it is omitted from the table. Its value is A = 585.4(0.7). Further, variance of mean estimator for the missing values is also the same for all methods, and its value is $\mathbb{V}\mathrm{ar}[\hat{\mu}^{mis}] = 0.36$.

The main results can be summarized as follows:

- the contribution of the squared global bias term to the mean squared error is neglible (except for baseline methods - however, for them its impact is about 5%)
- the methods with random or donor strategies correlate less between imputed and missing values than the methods with mean strategy,
- the variance of $\hat{\mu}^{imp}$ is bigger than the variance of $\hat{\mu}^{mis}$ for most methods.

						d_1	d_3	d_4
Method	**MSE**	B	C	D	Correlation	Global bias^2	$\mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}]$	$-2\mathbb{C}\mathrm{ov}[\hat{\mu}^{mis}, \hat{\mu}^{imp}]$
B,M	619.0(0.7)	0.0(0.0)	0.0(0.0)	33.6(0.4)	0.000(0.0000)	33.01	0.24	-0.02
B,R	1061.2(1.6)	437.5(0.7)	2.3(1.1)	36.0(0.5)	-0.002(0.0011)	35.25	0.40	-0.02
B,D	1082.2(1.6)	460.1(0.8)	-2.2(1.1)	38.9(0.5)	0.002(0.0011)	37.98	0.49	0.05
L,M	**81.1**(0.2)	464.3(0.7)	-969.7(1.3)	1.1(0.03)	**0.930**(0.0002)	1.00	0.33	-0.57
L,R	251.3(0.5)	536.9(0.8)	-871.8(1.4)	0.7(0.03)	0.777(0.0005)	0.46	0.39	-0.52
N,M	**87.3**(0.4)	576.7(1.6)	-1075.5(2.0)	0.5(0.03)	**0.926**(0.0003)	0.09	0.71	-0.62
K,M	**63.3**(0.1)	455.2(0.6)	-977.8(1.2)	0.4(0.02)	**0.947**(0.0001)	0.33	0.31	-0.59
K,R	141.2(0.3)	513.9(0.7)	-958.9(1.3)	0.7(0.03)	0.874(0.0003)	0.57	0.33	-0.56
T,M	122.2(0.5)	458.3(0.7)	-921.7(1.4)	0.2(0.01)	**0.890**(0.0004)	**0.02**	0.37	-0.57
T,R	214.8(0.8)	545.4(0.8)	-916.3(1.5)	0.2(0.01)	0.811(0.0007)	**0.00**	0.41	-0.53
T,D	186.6(0.7)	550.6(0.7)	-949.7(1.6)	0.3(0.01)	0.836(0.0007)	0.08	0.36	-0.55
T,M^s	154.7(2.1)	383.0(2.3)	-816.0(3.8)	2.2(0.37)	0.862(0.0022)	0.43	1.97	-0.53
T,R^s	208.4(0.6)	537.1(0.9)	-914.4(1.6)	0.3(0.03)	0.815(0.0006)	**0.01**	0.42	-0.54
C,M	116.9(0.6)	410.4(1.1)	-880.5(1.8)	1.5(0.06)	**0.898**(0.0005)	1.24	0.46	-0.54
C,R	203.1(0.8)	529.5(0.8)	-912.1(1.6)	0.3(0.02)	0.819(0.0008)	0.07	0.39	-0.55
C,D	221.7(1.0)	542.9(0.7)	-907.3(1.6)	0.6(0.03)	0.805(0.0009)	0.36	0.47	-0.57
TJ,M	190.0(1.1)	506.3(1.3)	-902.7(2.1)	0.9(0.07)	0.829(0.0011)	**0.05**	1.03	-0.52
TJ,R	324.3(1.1)	547.7(1.2)	-809.8(2.0)	1.0(0.06)	0.715(0.0011)	0.32	0.77	-0.47
TJ,D	247.7(0.9)	543.0(1.4)	-881.8(2.0)	1.1(0.08)	0.782(0.0008)	**0.04**	1.25	-0.53
TJ,M^s	**87.8**(0.6)	494.4(1.1)	-993.0(1.8)	1.0(0.05)	**0.923**(0.0006)	0.47	0.71	-0.57
TJ,R^s	126.9(0.5)	535.4(1.1)	-995.0(1.8)	1.0(0.06)	**0.889**(0.0005)	0.54	0.67	-0.58
CJ,M	156.4(1.2)	463.5(2.1)	-894.2(2.6)	1.7(0.09)	0.859(0.0011)	0.78	0.99	-0.47
CJ,R	246.5(1.5)	549.6(2.1)	-890.1(2.8)	1.6(0.11)	0.785(0.0013)	0.46	1.31	-0.54
CJ,D	248.3(1.3)	560.3(2.0)	-898.6(2.6)	1.1(0.08)	0.785(0.0012)	0.11	1.30	-0.64

Table 9.11: Decomposition terms for mean squared error. Remark that: $A = 585.4(0.7)$, $d_2 = \mathbb{V}\mathrm{ar}[\hat{\mu}^{mis}] = 0.36$.

The four lowest MSE results have been marked in the table. The methods that these results correspond to are: kernel regression with mean strategy, linear regression with mean strategy, 1-nearest neighbour, and TS-SOM joint $(Y, \boldsymbol{X})$ clustering with mean strategy and smoothing. All these methods yield a high correlation between missing and imputed values. There are some differences in the global bias which is lowest for the nearest neighbour method. The main difference between these methods is in the variability of imputations (term B). Linear regression and kernel regression yield a somewhat low variability among the imputed values. The variance of the missing values is approximately 584.4 and these two methods yield the variances 464.3 and 455.2. However, 1-nearest neighbour almost captures the true variability of the missing values. It yields a variance of 576.7. But this much of variance has been penalized in the mean squared error quantity. The joint $(Y, \boldsymbol{X})$ clustering with TS-SOM method is between these methods: yielding a variance of 535.4.

9.3.6 Comparison between the distributional KS measure and unit level MSE of the AGE variable

The final question about the imputation performance of the AGE variable concerns the tradeoff between the distributional and unit level performances. The MSE is computed as in the previous example in Section 9.3.5. To evaluate the distributional performance we use two variants of the Kolmogorov-Smirnov measure

a) The standard Kolmogorov-Smirnov measure

$$\mathrm{KS} = \mathbb{E}\Big[\sup_{y} |\hat{F}_{Y^{imp}}(y) - \hat{F}_{Y^{mis}}(y)| \big| \mathsf{n}\Big], \text{ and}$$

b) Integrated absolute error Kolmogorov-Smirnov which is defined as (this measure is KS$^{\alpha}$ for $\alpha = 1$ in [10])

$$\mathrm{KS(L1)} = \mathbb{E}\left[\frac{1}{T_{2\mathsf{N}^{mis}} - T_0} \sum_{j=1}^{2\mathsf{N}^{mis}} (T_j - T_{j-1}) \big|\hat{F}_{Y^{mis}}(T_j) - \hat{F}_{Y^{imp}}(T_j)\big| \; |\mathsf{n}\right],$$

where Y denotes age variable, $\hat{F}$ is the empirical weighted cumulative distribution function of age, $\{T_1, T_2, \ldots, T_{2\mathsf{N}^{mis}}\}$ is a sorted (from minimum to maximum) set of imputed and missing values, and T_0 is the biggest integer less or equal to T_1.

The results are given in Table 9.12, as well as in Plots 9.17 and 9.18. As before, the best methods are those that are close to the Pareto-optimal front. The results are more or less unaffected by the way how Kolmogorov-Smirnov measure is computed. In both Figures, 9.17 and 9.18, the best methods for under multivariate decision problem are

K,M kernel regression with mean strategy, which minimizes MSE

N,M nearest neighbour imputation

TJ,R^s Joint $(Y, \boldsymbol{X})$ clustering with TS-SOM under simulated random noise

T,D standard X-clustering TS-SOM with donor strategy, and

C,D standard K-Means clustering with donor strategy which minimizes the Kolmogorov-Smirnov measure.

In addition, methods (TJ,M^s), (T,R), (L,R), and (CJ,D) are close to the Pareto front.

The obvious conclusion is that nonparametric regression methods perform best if one tries to minimize unit level errors. For distributional performance it seems that cell methods are the best ones.

Method	**KS**	**KS(L1)**	**MSE**
B,M	0.61(0.0006)	0.23(0.0003)	619.0(0.7)
B,R	0.20(0.0007)	0.07(0.0003)	1061.2(1.6)
B,D	0.19(0.0009)	0.07(0.0004)	1082.2(1.6)
L,M	0.08(0.0004)	0.03(0.0001)	**81.1**(0.2)
L,R	**0.04**(0.0003)	0.02(0.0001)	251.3(0.5)
N,M	0.06(0.0007)	0.02(0.0002)	**87.3**(0.4)
K,M	0.12(0.0004)	0.04(0.0001)	**63.3**(0.1)
K,R	0.05(0.0003)	0.02(0.0001)	141.2(0.3)
T,M	0.11(0.0005)	0.04(0.0002)	122.2(0.5)
T,R	**0.04**(0.0003)	0.02(0.0001)	214.8(0.8)
T,D	**0.03**(0.0004)	**0.01**(0.0001)	186.6(0.7)
T,M^s	0.17(0.0022)	0.06(0.0006)	154.7(2.1)
T,R^s	0.05(0.0005)	0.02(0.0002)	208.4(0.6)
C,M	0.14(0.0006)	0.05(0.0002)	116.9(0.6)
C,R	0.05(0.0005)	0.02(0.0002)	203.1(0.8)
C,D	**0.03**(0.0003)	**0.01**(0.0001)	221.7(1.0)
TJ,M	0.10(0.0010)	0.03(0.0002)	190.0(1.1)
TJ,R	**0.04**(0.0006)	0.02(0.0002)	324.3(1.1)
TJ,D	0.05(0.0010)	0.02(0.0003)	247.7(0.9)
TJ,M^s	0.08(0.0007)	0.03(0.0002)	**87.8**(0.6)
TJ,R^s	0.05(0.0006)	0.02(0.0002)	126.9(0.5)
CJ,M	0.11(0.0008)	0.04(0.0003)	156.4(1.2)
CJ,R	0.05(0.0008)	0.02(0.0003)	246.5(1.5)
CJ,D	0.05(0.0008)	**0.01**(0.0002)	248.3(1.3)

Table 9.12: Kolmogorov-Smirnov and mean squared error results. Standard deviations of estimates are shown in parentheses.

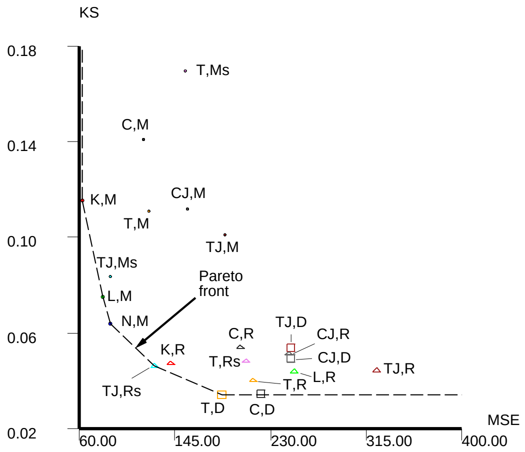


Figure 9.17: Mean squared error vs. Kolmogorov-Smirnov maximum statistic plot. Results for baseline methods are available in Table 9.12.

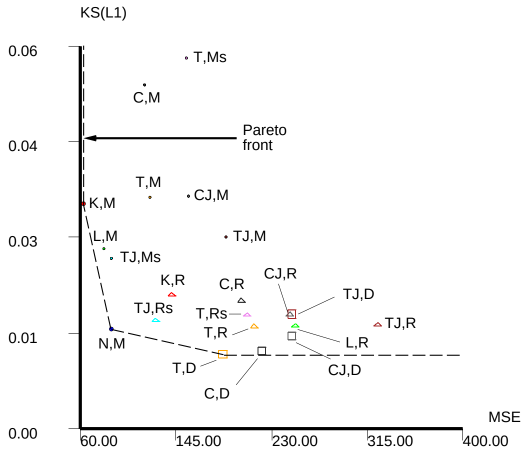


Figure 9.18: Mean squared error vs. Kolmogorov-Smirnov absolute statistic plot.

9.4 Evaluation of imputed SEX variable

In our final evaluation we shall investigate how well the different methods perform in the imputation of the categorial SEX variable. Since the missingness is directly explained by two covariates (FTPTWK and BENFTS) as explained earlier, the problem is solvable using a linear classifier. Therefore our interest is to see how efficient the other methods, especially cell imputation, are in this case. Typically, one would use more appropriate method, such as logistic regression, to impute a categorical variable. However, the results indicate that linear regression does well at unit level.

The evaluation consists of three different types of analyses

1) Evaluation of classification performance in terms of operating characteristics.
2) A comparison between unit level vs. distribution level measures.
3) A short review of computational time complexities of the methods.

In addition, a deeper study about the role of certain model parameters is studied in conjunction with the comparison between unit level and distribution level performances. In the study we try to understand the role of simulated noise and model flexibility. In addition the efficiency of different methods is studied in terms of a data set size n.

The parameters for the imputation methods were set as described in Section 9.2.4 with the exceptions that the TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods TJ* use 1024 cells in studies 1) and 2), and some parameters are different in the additional study of the efficiency as described later.

9.4.1 Classification performance of the SEX variable

In this study we investigate how well different methods can predict the correct class of variable SEX. For simplicity let $Y \in \{\mathrm{M}, \mathrm{F}\}$ be the SEX variable, where M denotes MALE and F denotes FEMALE. Then the task can be written in terms of soft classifier $g(\mathbf{x}) \in \mathbb{R}$ and thresholding parameter $\theta \in \mathbb{R}$ as follows

$$\hat{Y}_j^{imp} = \begin{cases} \mathrm{M} & \text{if } g(\mathbf{x}_j) + \epsilon < \theta \\ \mathrm{F} & \text{otherwise,} \end{cases}$$

where $\mathbf{x}_j$ is the covariate vector for the j^{th} observation. In the actual implementation, values are coded such that M$= -1$ and F$= 1$, which implies that $\hat{Y}_j^{imp} \in \{-1, 1\}$.

The classification performance can now be written in terms of posterior probabilities using a threshold θ as follows

$$\Pr(\hat{Y}^{imp} = \mathrm{M} | Y^{mis}, \theta) = 1 - \Pr(\hat{Y}^{imp} = \mathrm{F} | Y^{mis}, \theta).$$

The operating characteristics curve [20, 45] is computed in terms of the threshold θ and it explains how sensible the classifier is in terms of

$$\begin{aligned} v_1(\theta) &= \Pr(\hat{Y}^{imp} = \mathrm{M} | Y^{mis} = \mathrm{M}, \theta) \text{ vs.} \\ v_2(\theta) &= \Pr(\hat{Y}^{imp} = \mathrm{M} | Y^{mis} = \mathrm{F}, \theta). \end{aligned}$$

The result is then drawn in terms of curve $(v_1(\theta), v_2(\theta) | \theta),\ \theta : -\infty \rightarrow \infty$ as depicted in Figure 9.19. Another example is given, for example, in reference [22].

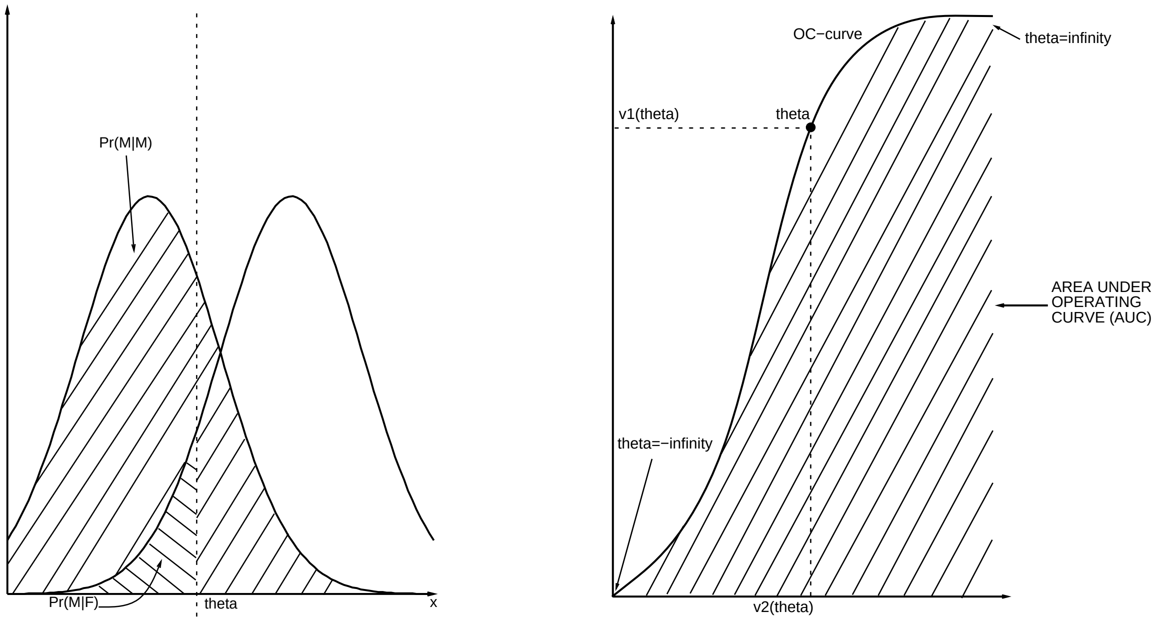


Figure 9.19: The idea behind operating characteristic curves.

It is obvious that the bigger the "area under OC-curve" (AUC) is, the better the classification performance is on a wide range of threshold values. Thus the results can be summarized using term AUC, which we compute using simple numerical integration over the observed values of the OC-curve.

Four examples of OC-curves are shown in Figures 9.20-9.23 for baseline mean, linear regression, standard $\boldsymbol{X}$-clustering TS-SOM, and joint $(Y, \boldsymbol{X})$ clustering TS-SOM. A box in the OC-curves depicts the value $\theta = 0$.

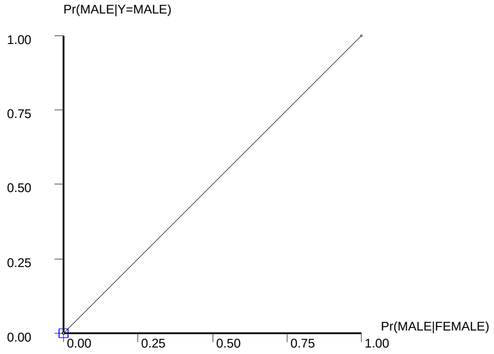


Figure 9.20: OC curve for baseline/mean strategy.

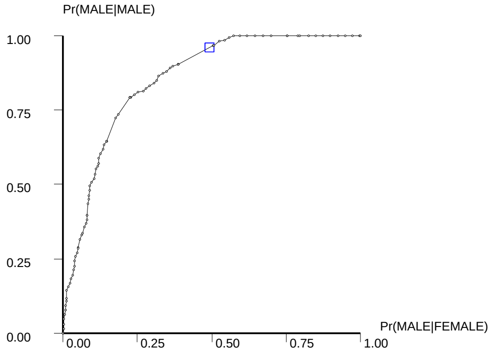


Figure 9.21: OC curve for linear regression/mean strategy.

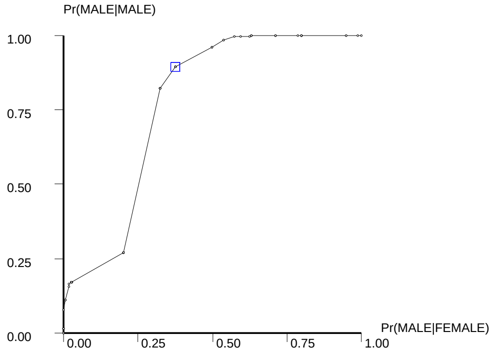


Figure 9.22: OC curve for TS-SOM standard $\boldsymbol{X}$-clustering/mean strategy (non-smoothed)

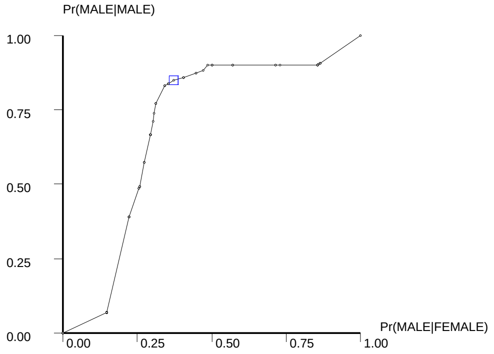


Figure 9.23: OC curve for TS-SOM joint $(Y, \boldsymbol{X})$ clustering/mean strategy (non-smoothed)

Clearly, for small values θ the classification performance of linear regression is the best, and it explains the highest AUC values. However, given $\theta = 0$, the differences between the methods are not as visible.

Since the actual implementation of imputation requires fixed θ, we have chosen $\theta = 0$ to compute the confusion matrices for all the methods. Thus we have computed the quantities

$$\Pr(\hat{Y}^{imp} = \mathrm{M} | Y^{mis} = \mathrm{M}, \theta = 0), \ \Pr(\hat{Y}^{imp} = \mathrm{F} | Y^{mis} = \mathrm{M}, \theta = 0)$$
$$\Pr(\hat{Y}^{imp} = \mathrm{M} | Y^{mis} = \mathrm{F}, \theta = 0), \ \Pr(\hat{Y}^{imp} = \mathrm{F} | Y^{mis} = \mathrm{F}, \theta = 0).$$

The results are summarized for all of the methods in Table 9.13.

The worst performer is the baseline method with mean strategy. Namely, it predicts all missing persons to be females (which is a mode class in the observed data). The best performer for male class is linear regression. This is expected as covariates with a higher linear dependency to male class are being used. The nonparametric regression methods and some cell methods perform also well for the male class predicting approximately 80%-90% of missing males correctly as male. For the female class, linear regression performs significantly worse than some other methods: it is able to predict only 55% of missing femals as female. Possible nonlinear dependencies between female class and some of the covariates may explain the bad performance of linear regression. The nonparametric regression methods and almost all the cell methods are able to predict 70% of missing females correctly.

The expected area under the operating characteristic curves is also summarized in Table 9.13. A good imputation method should have few misclassifications and thus a high expected area under the curve. Linear regression yields the highest area under curve for the male class. However, two standard $\boldsymbol{X}$-clustering cell methods yield also a high area. The best joint $(Y, \boldsymbol{X})$ clustering methods (TJ*,M^s and TJ*,R^s without incomplete training) yield an area of 0,74. One can notice that random or donor strategies yield lower areas than the corresponding method with mean strategy, however this is not the case for joint $(Y, \boldsymbol{X})$ clustering methods. For the joint $(Y, \boldsymbol{X})$ clustering methods, random or donor strategy yield at least as good a performance as mean strategy, often a bit better.

Method	Pr(M|M)	Pr(F|M)	Pr(M|F)	Pr(F|F)	**AUC**
B,M	0.00(0.000)	1.00(0.000)	**0.00**(0.000)	**1.00**(0.000)	0.500(0.000)
B,R	0.42(0.002)	0.58(0.002)	0.42(0.002)	0.58(0.002)	0.499(0.002)
B,D	0.41(0.002)	0.59(0.002)	0.41(0.002)	0.59(0.002)	0.500(0.001)
L,M	**0.95**(0.003)	**0.05**(0.003)	0.45(0.005)	0.55(0.005)	**0.857**(0.001)
L,R	0.55(0.002)	0.45(0.002)	0.46(0.002)	0.54(0.002)	0.563(0.001)
N,M	0.77(0.007)	0.23(0.007)	0.31(0.006)	0.69(0.006)	0.711(0.002)
K,M	0.79(0.010)	0.21(0.010)	**0.29**(0.012)	**0.71**(0.012)	**0.832**(0.001)
K,R	0.53(0.002)	0.47(0.002)	0.42(0.002)	0.58(0.002)	0.574(0.002)
T,M	**0.88**(0.006)	**0.12**(0.006)	0.37(0.009)	0.63(0.009)	**0.849**(0.002)
T,R	0.71(0.004)	0.29(0.004)	0.31(0.003)	0.69(0.003)	0.743(0.002)
T,D	0.72(0.004)	0.28(0.004)	0.31(0.003)	0.69(0.003)	0.699(0.002)
T,M^s	**0.89**(0.006)	**0.11**(0.006)	0.39(0.010)	0.61(0.010)	**0.846**(0.002)
T,R^s	0.63(0.004)	0.37(0.004)	0.31(0.002)	**0.69**(0.002)	0.700(0.002)
C,M	**0.90**(0.008)	**0.10**(0.008)	0.40(0.012)	0.60(0.012)	**0.850**(0.001)
C,R	0.64(0.005)	0.36(0.005)	**0.31**(0.002)	**0.69**(0.002)	0.710(0.002)
C,D	0.66(0.005)	0.34(0.005)	**0.31**(0.002)	**0.69**(0.002)	0.650(0.002)
TJ,M	0.72(0.018)	0.28(0.018)	0.32(0.006)	0.68(0.006)	0.640(0.008)
TJ,R	0.70(0.018)	0.30(0.018)	0.31(0.006)	0.69(0.006)	0.664(0.008)
TJ,D	0.71(0.018)	0.29(0.018)	0.32(0.007)	0.68(0.007)	0.680(0.008)
TJ,M^s	0.73(0.016)	0.27(0.016)	0.32(0.006)	0.68(0.006)	0.683(0.008)
TJ,R^s	0.73(0.016)	0.27(0.016)	0.32(0.005)	0.68(0.005)	0.703(0.006)
TJ*,M^s	**0.81**(0.009)	**0.19**(0.009)	0.33(0.009)	0.67(0.009)	0.735(0.005)
TJ*,R^s	**0.81**(0.008)	**0.19**(0.008)	0.33(0.009)	0.67(0.009)	0.737(0.004)
CJ,M	0.73(0.014)	0.27(0.014)	0.32(0.010)	0.68(0.010)	0.645(0.006)
CJ,R	0.68(0.017)	0.32(0.017)	**0.30**(0.011)	**0.70**(0.011)	0.675(0.006)
CJ,D	0.71(0.015)	0.29(0.015)	**0.30**(0.011)	**0.70**(0.011)	0.685(0.006)

Table 9.13: Conditional probabilities and areas under operating characteristic curves. Standard deviations of estimates are shown in parenthesis.

9.4.2 Comparison between unit level and distributional level imputation performances of the SEX variable

There are several alternatives to evaluate the distributional and unit level performance of categorial variables. In this context we have chosen new measures DIST and MR as defined next.

Distributional measure DIST is chosen as

$$\mathrm{DIST} = \mathbb{E}\Big[\sum_{y\in\{M,F\}} \Pr(Y^{mis}=y)|\Pr(Y^{imp}=y)-\Pr(Y^{mis}=y)|\;\Big|\mathsf{n}\Big],$$

which takes its maximum when there is the biggest difference between the proportion of imputed Y^{imp} and missing values Y^{mis} for males and females. The differences are weighted by class priors, which makes this somewhat similar to the Kullback-Leibler distance.

The unit level error is simply the weighted unit level error

$$\mathrm{MR} = \mathbb{E}\Big[\frac{1}{\sum_{j\in\Omega^{mis}} W_j} W_j I(Y_j^{mis} \neq Y_j^{imp})|\mathsf{n}\Big],$$

where Ω^{mis} is the index-set of missing values and $I(lexp)$ is the indicator function for logical expression $lexp$. As before, we are using the sampling weights.

The results are summarized in Table 9.14 and Figure 9.24.

Method	**DIST**	**MR**
B,M	0.610(0.001)	0.610(0.001)
B,R	0.186(0.002)	0.516(0.001)
B,D	0.200(0.002)	0.519(0.001)
L,M	0.145(0.004)	**0.209**(0.001)
L,R	0.097(0.002)	0.454(0.001)
N,M	**0.056**(0.004)	0.262(0.002)
K,M	0.105(0.004)	**0.245**(0.002)
K,R	0.125(0.002)	0.454(0.001)
T,M	0.089(0.005)	**0.219**(0.001)
T,R	**0.058**(0.003)	0.299(0.002)
T,D	**0.051**(0.003)	0.294(0.002)
T,M^s	0.099(0.006)	**0.221**(0.001)
T,R^s	0.105(0.003)	0.350(0.003)
C,M	0.124(0.006)	**0.218**(0.001)
C,R	0.100(0.003)	0.341(0.003)
C,D	0.087(0.003)	0.332(0.003)
TJ,M	0.097(0.010)	0.298(0.009)
TJ,R	0.105(0.010)	0.308(0.010)
TJ,D	0.104(0.010)	0.304(0.009)
TJ,M^s	0.083(0.009)	0.289(0.008)
TJ,R^s	0.082(0.009)	0.289(0.008)
TJ*,M^s	**0.070**(0.005)	**0.245**(0.004)
TJ*,R^s	**0.071**(0.004)	**0.244**(0.003)
CJ,M	0.096(0.008)	0.288(0.007)
CJ,R	0.125(0.010)	0.308(0.008)
CJ,D	0.104(0.008)	0.291(0.007)

Table 9.14: Distribution level and unit level results. The best five (or six) results have been marked using a bold font.

The baseline methods perform the worst at both unit level and distribution level, what is as expected when covariate information is not used.

The best unit level predictions are given by linear regression with mean strategy and the standard $\boldsymbol{X}$-clustering cell methods with smoothed or with non-smoothed mean strategy. A good performance for linear regression is expected, because variables with the highest linear dependency to male class were selected as covariates. The nonparametric regression methods with mean strategy perform also well. Joint $(Y, \boldsymbol{X})$ clustering with TS-SOM methods TJ*,M^s and TJ*,R^s perform better than nearest neighbour at unit level. Random or donor strategy increases the unit level error compared to mean strategy, as expected. On the other hand, inclusion of noise to predictions reduces the distributional level error in some cases. One may suspect noise distribution to be misspecified for those methods in which added noise increases the distributional level error.

From the results one can notice that the baseline methods with random imputation or donor strategy perform better at unit level than the baseline mean strategy. Typically one would not expect such behaviour. There is a simple explanation for this. Distributions of observed and missing SEX are different: in the observed part there are more females than males, and vice versa in the missing part. The baseline mean strategy predicts all the missing persons as female (an 'optimal baseline' unit level strategy would predict all of them as males). The donor and random imputation strategies predict more often male than the mean strategy does, and this produces a better, but still poor, unit level result.

The best distribution level results are obtained by nearest neighbour and TS-SOM standard $\boldsymbol{X}$-clustering with either random strategy (without smoothing) or with donor strategy. However, also the TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods which use smoothing yield a quite good distribution performance.

When considering simultaneously preservation of unit and distribution level, four methods pop up: nearest neighbour, the TS-SOM joint $(Y, \boldsymbol{X})$-clustering methods utilizing smoothing and using only complete observations in model training (abbreviations TJ*,M^s and TJ*,R^s), and standard $\boldsymbol{X}$-clustering TS-SOM with donor strategy.

Figure 9.24 confirms our conclusions. The best methods for multivariate decision problem are

L,M linear regression with mean strategy, which minimizes unit level error MR

T,M standard $\boldsymbol{X}$-clustering TS-SOM with mean strategy (T,M)

(TJ*,M^s),(TJ*,R^s) joint $(Y, \boldsymbol{X})$ clustering methods

N,M nearest neighbour, and

T,D donor version of $\boldsymbol{X}$-clustering using TS-SOM, which minimizes distribution level error DIST.

In addition methods (C,M), (K,M), (T,M^s), (T,R) are close to the Pareto front.

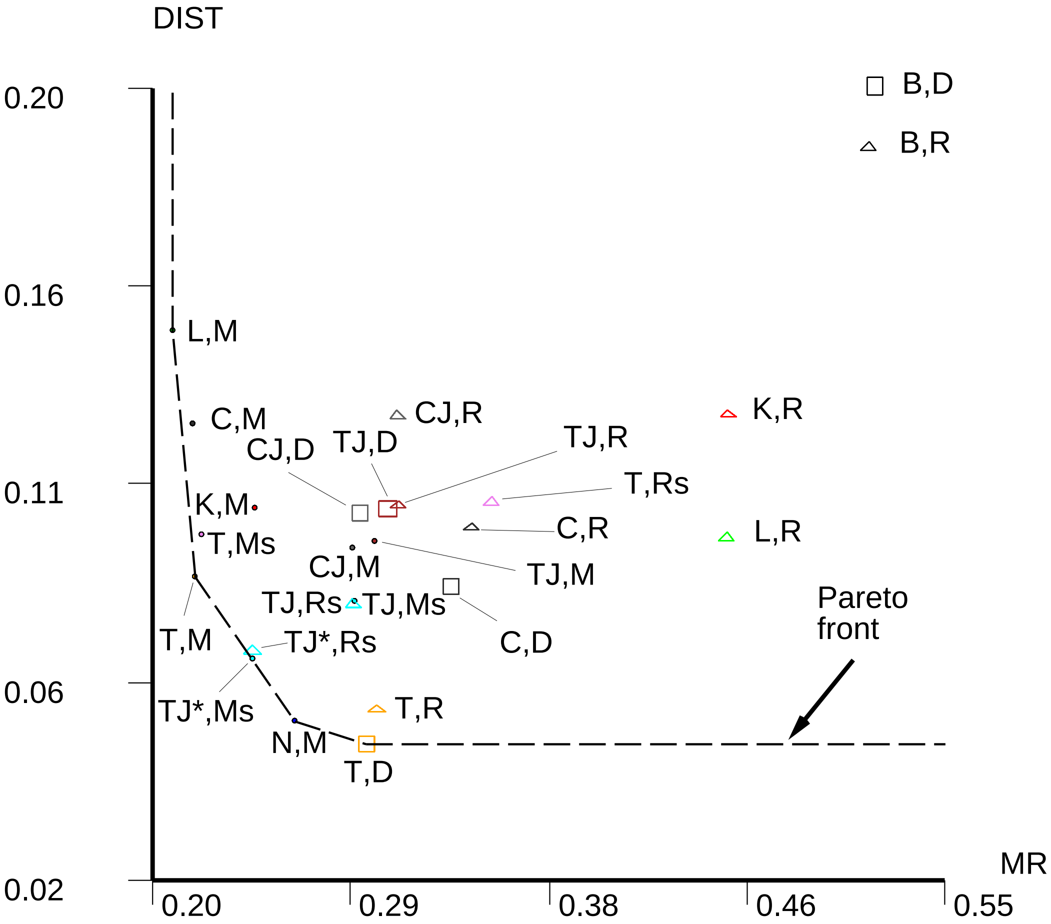


Figure 9.24: Unit level result versus distribution level result. Results for joint $(Y, \boldsymbol{X})$ clustering with TS-SOM methods which do not utilize observations with missing SEX are labeled by a star(*).

9.4.3 The roles of model flexibility and simulated randomness

Here we are interested to know how added simulated randomness changes the role of the method from unit level imputation to distributional level imputation. In addition, we shall investigate the role of model flexibility in the case of TS-SOM methods. This is done by testing the model performance with different number of clusters (cells).

In the case of simulated randomness the variation is done by changing the variance σ^2 in imputation using a soft clustering model, $g^{model}(\mathbf{x}_j)$, with a random term ϵ as

$$\hat{Y}_j^{imp} = \begin{cases} \text{M} & \text{if } g^{model}(\mathbf{x}_j) + \epsilon < 0, \ \epsilon \sim N(0, \sigma^2) \\ \text{F}, & \text{otherwise.} \end{cases}$$

In the experiments we see what happens when $\sigma^2 : 0 \rightarrow \infty$.

The results of linear regression vary as shown in Figure 9.25.

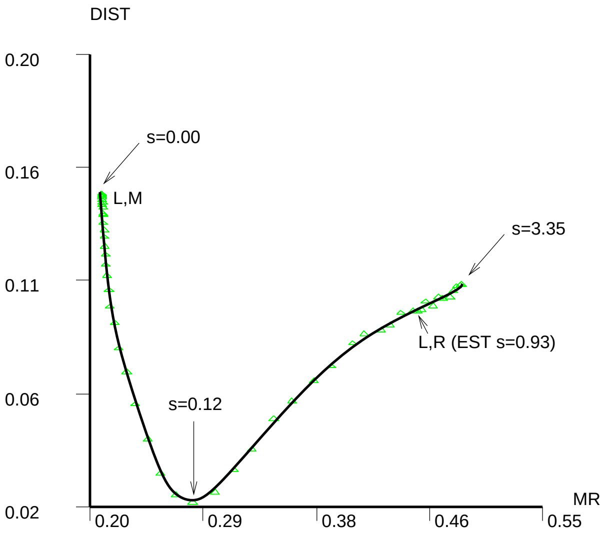


Figure 9.25: Performance of linear regression in terms of different values of simulated randomness σ. Notation "s" denotes standard deviation of simulated randomness (σ), and "EST s" denotes standard deviation of imputation noise estimated by the L, R method ($\hat{\sigma}$).

Clearly, at $\sigma = 0$ the unit level measure MR is minimized. Then as σ is increased to value 0.12, where the distributional measure DIST is minimized. Finally as $\sigma \to \infty$ the performance becomes worse on both terms of MR and DIST. Note also that the estimated residual standard deviation was $\hat{\sigma} = 0.93$, indicating that the optimal value of σ is $0.129 * \hat{\sigma}$ for this example.

Similar experiment was done with kernel regression as well, as shown in Figure 9.26.

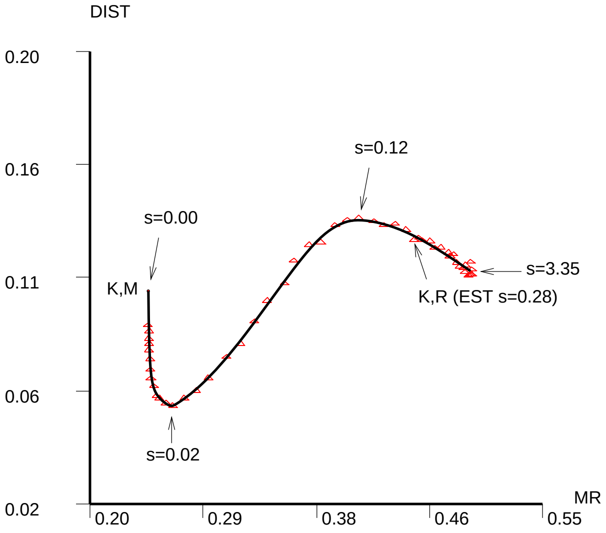


Figure 9.26: The role of simulated randomness in the results of kernel regression.

As we can see, the behaviour is rather similar to that of linear regression. Now

MR is optimum with $\sigma = 0$, DIST is optimal with $\sigma = 0.02$ and there is a local "worst" result when $\sigma = 0.12$. The results can be compared against the estimate of residual standard deviation $\hat{\sigma} = 0.28$.

With joint $(Y, \boldsymbol{X})$-clustering with TS-SOM also the complexity of the model was varied from 4 to 4096 clusters, denoted by $Ll = 2^l$, (L2=4,. . . ,L7=4096). The results are visualized in Figures 9.27 and 9.28, where the former is using incomplete data training and in the other the model is built using the fully observed part of data.

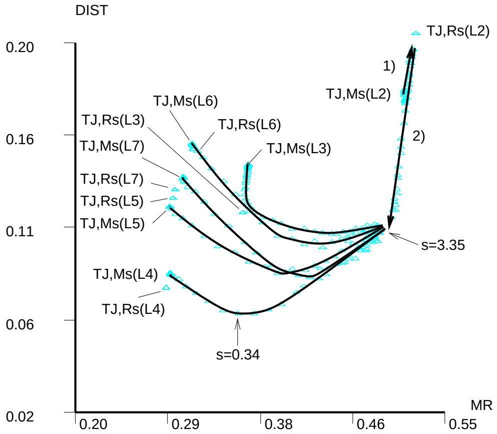


Figure 9.27: Results for TJ,M^s and TJ,R^s cell methods with a varied number of cells (L2=4 cells, L3=16, L4=64, L5=256, L6=1024, and L7=4096) and noise level.

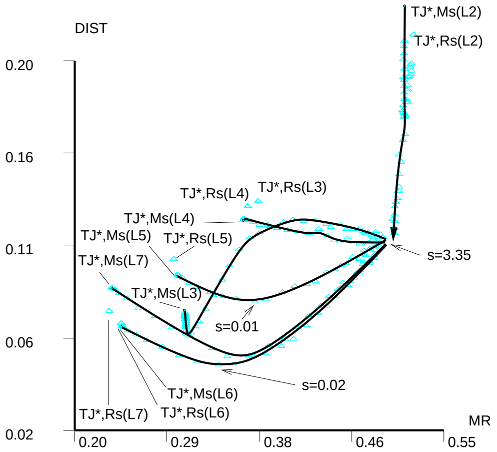


Figure 9.28: Results for TJ*,M^s and TJ*,R^s cell methods when observations with missing SEX are excluded from model training data.

In all cases the methods meet when $\sigma \to \infty$ as the model becomes closer to random imputation. Otherwise the MR measure is minimized (usually) when $\sigma = 0$. The best results correspond to the TJ*,R^s model using 1024 clusters (L6) with added randomness under variance $\sigma = 0.02$. Note that in the optimal case the estimated residual variance was $\hat{\sigma} = 0.15$, indicating that the optimal value of σ is $0.133 * \hat{\sigma}$ for this method and example.

9.4.4 The role of sample size in the imputation of the SEX variable

In this example we test the performance of different methods in terms of sample size. The aim is to compare relative efficiencies between the methods. Although the cause behind the differences in the efficiency is not fully studied here, it should be noted that there is a relation between our analytical results (approximations) and empirical results of this example. Unit level error MR, which is used here, equals to four times mean squared error (MSE). However, sampling weights complicate the situation.

As earlier, we use DIST and MR to measure distributional and unit level performances, respectively. In the experiments the sample size is increased from 4 to 65536 observations. However, the results for the nonparametric regression methods for sample sizes 32768 and 65536 have been omitted due to high computational requirements. Also, when the number of observations increases, the complexity of nonparametric models is allowed to grow. Thus, we try to maximize the role of data in imputation performance, rather than minimize the variance of estimators. For kernel regression this is done by decreasing the smoothing, and for the cell methods the number of cells is increased.

The exact value of smoothing bandwidth for kernel regression is taken as

$$\lambda = 0.5 * (N^{obs})^{-1/20},$$

which is based on Mack's recommendations [72]. However, note that approximately half of the covariates are categorials (Mack's formulas assume continuous covariates). Smoothing was halved from the earlier study, because it seemed to produce better results, especially at the smallest sample sizes.

Note that the imputation noise variances for the random imputation strategies were not set to optimal levels with respect to distribution level performance (recall the previous experiment). Instead we used an estimated amount of noise variance. As a consequence, the performance of some methods at a distribution level might be improved.

The number of cells for the clustering methods was selected after some experimentation. The obtained "best" performing cluster values are listed in Table 9.15. There are three different categories of the methods

i) SOM, which refers to all "normal" TS-SOM methods

ii) SOM*, which refers to new TJ* methods, where the incomplete part of data was omitted during the training

iii) K-means that refers to K-Means based methods.

Sample size n	Repetitions	SOM cells	SOM* cells	K-Means cells
4	500	4	4	1
8	500	4	4	1
16	400	4	4	2
32	400	4	4	4
64	300	16	16	6
128	300	16	16	8
256	200	16	16	16
512	200	16	16	24
1024	200	64	64	32
2048	100	64	256	48
4096	100	64	1024	64
8192	50	256	1024	96
16384	25	256	1024	128
32768	25	1024	1024	192
65536	25	1024	1024	256

Table 9.15: Sample size, the number of repeated experiments and the number of clusters used with the cell imputation methods.

As usual, the experiments were repeated several times in order to obtain reliable mean results. The number of repetitions was bigger when the sample size was small, as shown in Table 9.15.

The results are shown in Figures 9.29 and 9.30, as well as in Tables 9.16 and 9.17. The two figures contain relative results which are defined as

$$\begin{aligned}\Delta(\text{DIST}) &= \text{DIST} - \text{DIST}^{L,M}\\ \Delta(\text{MR}) &= \text{MR} - \text{MR}^{L,M},\end{aligned}$$

where linear regression with mean strategy has been used as the reference method. The two tables contain absolute results. Three best (or multiple if two best is not uniquely defined) results are marked for each sample size, and the best of them is underlined. However, there are situations in which the results for multiple methods differ only by 0.01 or 0.02 units, which amounts to an average or maximum standard deviation of the computed estimates. A full list of the estimates of standard deviations of the computed estimates is not included to compress the tables.

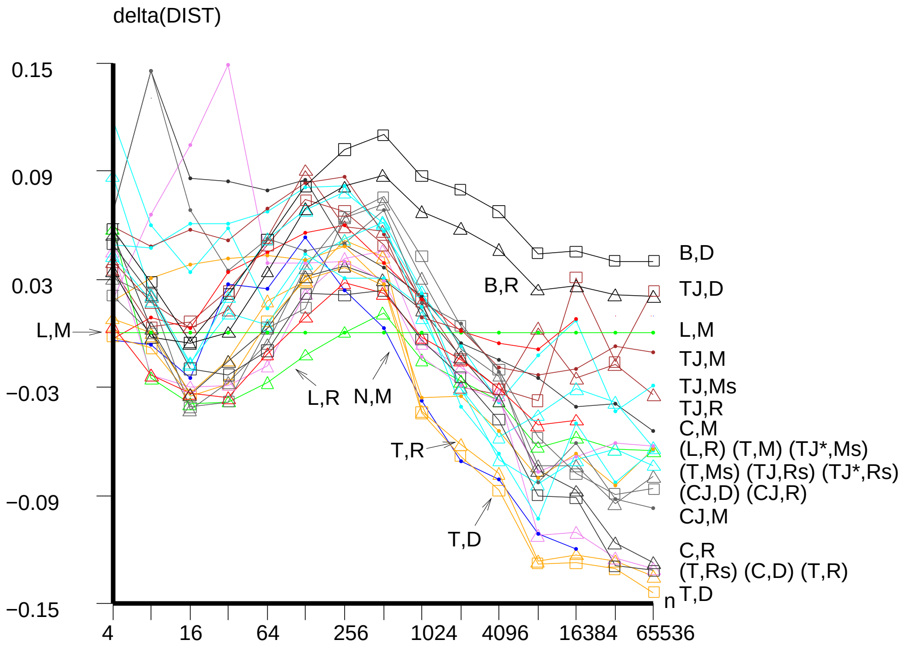


Figure 9.29: Relative distributional level results as functions of the sample size. The results for B,M have been omitted as 'outliers'. The results for the nonparametric regression methods have been omitted for n = 32768 and 65536 due to high computational requirements. The labels for kernel regression methods have been omitted for clarity. See Table 9.16 for all results.

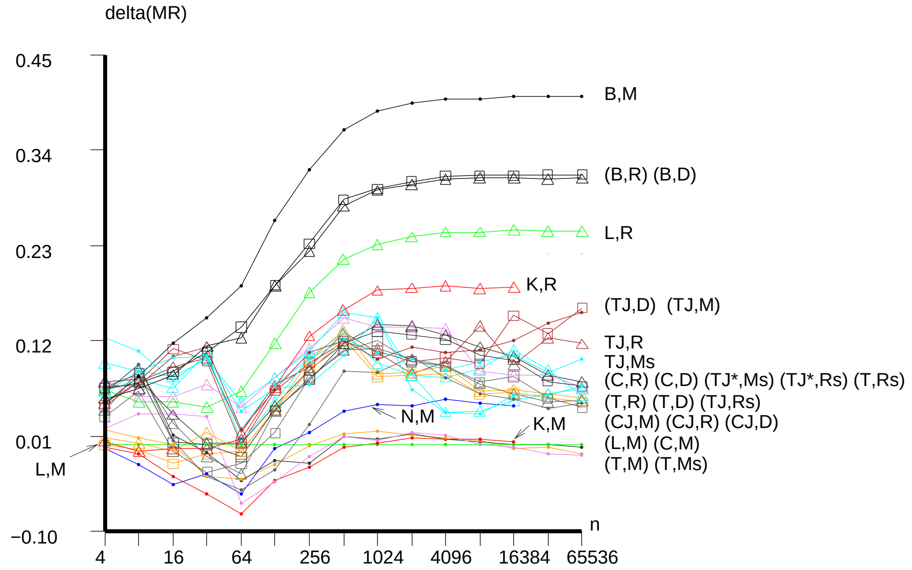


Figure 9.30: Relative unit level results as functions of the sample size. The results for nonparametric regression methods have been omitted for n = 32768 and 65536 due to high computational requirements.

n **Method**	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768	65536
B,M	0.49	0.51	0.56	0.56	0.58	0.60	0.60	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
B,R	0.47	0.38	0.32	0.24	0.20	0.19	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18
B,D	0.49	0.41	0.32	0.26	0.22	0.21	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
L,M	0.44	0.38	0.33	0.24	0.17	**0.12**	**0.10**	**0.09**	0.11	0.12	0.14	0.16	0.16	0.16	0.16
L,R	0.49	**0.35**	**0.29**	**0.20**	**0.14**	**0.11**	**0.10**	**0.10**	0.10	0.10	0.10	0.09	0.10	0.10	0.10
N,M	**0.43**	**0.37**	0.30	0.26	0.19	0.18	**0.12**	**0.10**	**0.08**	**0.05**	**0.06**	**0.05**	**0.04**	-	-
K,M	**0.43**	0.39	0.33	0.27	0.22	0.18	0.16	0.13	0.13	0.13	0.13	0.15	0.17	-	-
K,R	**0.44**	0.36	0.29	**0.20**	**0.16**	**0.13**	0.13	0.11	0.11	0.11	0.11	0.11	0.11	-	-
T,M	0.45	0.41	0.37	0.28	0.21	0.17	0.15	0.12	**0.08**	0.09	0.08	0.08	0.09	0.08	0.10
T,R	**0.44**	0.38	**0.29**	0.22	0.19	0.15	0.15	0.13	**0.07**	**0.06**	**0.06**	**0.03**	**0.03**	**0.03**	**0.03**
T,D	**0.43**	**0.37**	**0.29**	0.21	0.18	0.15	0.14	0.12	**0.07**	**0.06**	**0.05**	**0.03**	**0.03**	**0.03**	**0.02**
T,M^s	0.45	0.45	0.43	0.39	0.21	0.16	0.14	0.12	0.10	0.10	0.10	0.08	0.09	0.10	0.10
T,R^s	0.48	**0.36**	0.30	0.21	**0.15**	0.15	0.14	0.14	0.11	0.10	0.10	**0.05**	0.05	0.04	**0.03**
C,M	0.50	0.53	0.41	0.32	0.25	0.21	0.15	0.13	0.13	0.12	0.12	0.13	0.12	0.12	0.11
C,R	0.49	0.38	**0.29**	0.22	0.17	0.16	0.14	0.12	0.13	0.11	0.10	0.08	0.07	0.05	0.04
C,D	0.49	0.40	0.31	0.21	**0.16**	0.15	**0.12**	0.12	0.11	0.10	0.09	0.07	0.07	**0.03**	**0.03**
TJ,M	0.50	0.43	0.38	0.29	0.24	0.21	0.19	0.15	0.12	0.13	0.12	0.13	0.14	0.16	0.15
TJ,R	0.48	0.40	0.33	0.25	0.22	0.21	0.16	0.15	0.13	0.09	0.10	0.16	0.13	0.15	0.13
TJ,D	0.47	0.38	0.33	0.26	0.22	0.20	0.17	0.14	0.13	0.11	0.11	0.12	0.19	0.15	0.19
TJ,M^s	0.49	0.43	0.39	0.30	0.24	0.21	0.18	0.15	0.14	0.12	0.10	0.15	0.16	0.12	0.13
TJ,R^s	0.48	0.40	0.31	0.26	0.22	0.19	0.18	0.15	0.12	0.09	0.08	0.11	0.13	0.12	0.10
TJ*,M^s	0.55	0.44	0.36	0.30	0.18	0.17	0.13	0.12	0.13	0.08	0.07	**0.05**	0.11	0.08	0.10
TJ*,R^s	0.52	0.40	0.31	0.25	0.17	0.16	0.15	0.15	0.14	0.09	0.07	0.07	0.09	0.10	0.09
CJ,M	0.50	0.53	0.40	0.27	0.22	0.17	0.15	0.16	0.13	0.13	0.12	0.08	0.10	0.07	0.07
CJ,R	0.47	0.40	**0.28**	0.21	**0.16**	0.16	0.16	0.16	0.14	0.11	0.11	0.08	0.08	0.07	0.08
CJ,D	0.46	0.38	**0.29**	**0.20**	0.17	0.14	0.16	0.17	0.16	0.13	0.12	0.10	0.08	0.07	0.08
Deviation															
Max	0.019	0.015	0.014	0.012	0.011	0.010	0.010	0.009	0.010	0.013	0.012	0.017	0.031	0.028	0.027
Avg	0.019	0.014	0.012	0.009	0.008	0.007	0.007	0.006	0.005	0.007	0.006	0.007	0.010	0.009	0.009

Table 9.16: Distribution level results (DIST) as functions of the sample size. The maximum and average deviances of computed estimates are shown in bottom two rows. The best three methods for each sample size are marked and the best of them is underlined. The results for nonparametric regression methods have been omitted for sample sizes 32768 and 65536 due to high computational requirements.

n **Method**	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768	65536
B,M	0.49	0.51	0.56	0.56	0.58	0.60	0.60	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
B,R	0.50	0.50	0.52	0.53	0.52	0.52	0.51	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52
B,D	0.51	0.50	0.53	0.52	0.53	0.52	0.52	0.53	0.52	0.52	0.52	0.52	0.52	0.52	0.52
L,M	**0.44**	0.43	0.44	0.41	0.39	0.34	0.29	**0.24**	**0.22**	**0.22**	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**
L,R	0.51	0.48	0.49	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.45	0.46	0.46	0.45
N,M	**0.44**	**0.41**	**0.40**	**0.38**	**0.34**	0.33	0.30	0.28	0.27	0.26	0.26	0.26	0.25	-	-
K,M	**0.44**	**0.42**	**0.41**	**0.36**	**0.31**	**0.30**	**0.26**	**0.24**	**0.23**	**0.22**	**0.22**	0.22	**0.21**	-	-
K,R	0.45	**0.42**	0.44	0.41	0.40	0.40	0.41	0.40	0.40	0.40	0.39	0.39	0.39	-	-
T,M	0.46	0.44	0.44	**0.38**	0.36	0.32	0.29	0.26	0.24	**0.23**	**0.22**	**0.21**	**0.21**	**0.21**	**0.20**
T,R	0.45	0.43	0.44	0.43	0.39	0.38	0.39	0.38	0.31	0.30	0.30	0.27	0.27	0.27	0.26
T,D	**0.44**	0.43	**0.42**	0.40	0.39	0.38	0.37	0.36	0.30	0.30	0.29	0.27	0.27	0.26	0.26
T,M^s	0.46	0.47	0.48	0.45	**0.33**	**0.30**	**0.27**	**0.25**	**0.23**	**0.23**	**0.22**	**0.21**	**0.21**	**0.20**	**0.20**
T,R^s	0.52	0.49	0.50	0.48	0.44	0.42	0.40	0.39	0.36	0.35	0.34	0.29	0.29	0.29	0.27
C,M	0.50	0.53	0.45	0.41	0.35	0.32	**0.27**	**0.25**	**0.23**	**0.23**	**0.22**	**0.21**	**0.21**	**0.21**	**0.20**
C,R	0.52	0.51	0.48	0.42	0.40	0.38	0.37	0.36	0.36	0.35	0.34	0.32	0.31	0.29	0.28
C,D	0.51	0.51	0.45	0.42	0.38	0.38	0.36	0.36	0.35	0.34	0.33	0.31	0.31	0.28	0.28
TJ,M	0.51	0.50	0.55	0.53	0.41	0.40	0.39	0.36	0.32	0.33	0.32	0.32	0.33	0.35	0.36
TJ,R	0.49	0.50	0.53	0.52	0.41	0.41	0.38	0.37	0.33	0.30	0.31	0.35	0.30	0.33	0.32
TJ,D	0.49	0.51	0.55	0.51	0.40	0.40	0.38	0.36	0.33	0.31	0.31	0.30	0.36	0.34	0.37
TJ,M^s	0.50	0.52	0.54	0.52	0.41	0.41	0.39	0.36	0.33	0.31	0.30	0.30	0.31	0.29	0.31
TJ,R^s	0.51	0.51	0.51	0.52	0.40	0.39	0.39	0.37	0.32	0.30	0.29	0.30	0.32	0.29	0.27
TJ*,M^s	0.57	0.54	0.52	0.52	0.43	0.40	0.36	0.35	0.37	0.28	0.25	0.24	0.27	0.27	0.28
TJ*,R^s	0.54	0.52	0.52	0.51	0.44	0.42	0.40	0.40	0.37	0.30	0.25	0.25	0.26	0.27	0.27
CJ,M	0.50	0.53	0.45	**0.38**	**0.34**	**0.31**	0.31	0.33	0.31	0.32	0.29	0.27	0.26	0.25	0.26
CJ,R	0.48	0.52	0.47	0.40	0.36	0.38	0.36	0.35	0.34	0.31	0.30	0.28	0.27	0.26	0.26
CJ,D	0.48	0.49	0.44	**0.38**	0.37	0.35	0.37	0.35	0.34	0.31	0.30	0.28	0.29	0.26	0.25
Deviation															
Max	0.019	0.015	0.014	0.011	0.011	0.009	0.009	0.008	0.008	0.011	0.010	0.015	0.023	0.022	0.020
Avg	0.019	0.015	0.012	0.009	0.008	0.006	0.006	0.004	0.004	0.005	0.004	0.005	0.007	0.006	0.006

Table 9.17: Unit level results (MR) as functions of the sample size. The maximum and average deviances of computed estimates are shown in bottom two rows. The best three methods for each sample size are marked and the best of them is underlined. The results for nonparametric regression methods have been omitted for sample sizes 32768 and 65536 due to high computational requirements.

At distribution level most of the methods yield results better than those of the baseline methods from sample size 128 onwards. From sample size 8 to 256 linear regression with random strategy performs best. Nearest neighbour and the two TS-SOM cell methods seem to provide better results than the linear regression method from sample size 1024 onwards. A possible reason for this might be that the imputation distribution (for imputation noise) is misspecified: the Gaussian assumption may be incorrect.

At unit level all methods seem to yield better results than the baseline methods from sample size 64 onwards. From sample size 512 onwards the five best methods are: linear regression with mean strategy, the TS-SOM standard $\boldsymbol{X}$-clustering cell method with mean strategy and smoothing, the TS-SOM standard $\boldsymbol{X}$-clustering cell method with mean strategy without smoothing, the K-Means standard $\boldsymbol{X}$-clustering cell method with mean strategy, and kernel regression with mean strategy. The nearest neighbour method performs also quite well. The two introduced joint $(Y, \boldsymbol{X})$ clustering TS-SOM cell methods (abbreviated as TJ*,M^s and TJ*,R^s) perform quite similar to 1-nearest neighbour from sample size 4096 to 16384. This is because the number of cells has grown enough.

As a final conclusion we observe (these are best verified from result tables 9.16 and 9.17) that

i) Linear regression is likely to be the optimal method for this problem, which is due to our setup of the experiment. Note that at distribution level linear regression with random strategy is not among the best performing methods for the largest sample sizes. This is likely due to misspecified noise distribution.

ii) As the sample increases, some standard $\boldsymbol{X}$-clustering cell methods can reach the same unit level performance as linear regression with mean strategy. At distribution level many cell methods reach a better performance than linear regression with random strategy.

9.4.5 Computational requirements

In the final set of experiments we examine the computational requirements of various methods. To make the experiments comparable, the previous example with one incomplete variable (SEX) with a fixed set of 16 covariates are used. As we are only interested in computational times, there is no need to do repeated experiments. Thus, results are computed only once with increasing sample size n from 4 to 65536 observations. Simulations are done using an AMD Athlon64 3000+ processor, which is running at 1.8 GHz, with 2 gigabytes of RAM (DDR/400 MHz in dualchannel mode), under the Linux operating system.

The results are presented in log-scale, in Graphs 9.31, 9.32, and 9.33 because time requirements grow rapidly for nonparametric methods. The three graphs contain the results from sample sizes 256 to 65536. For a comparison, total times in the original scale, in seconds, are shown in Table 9.18 (for all sample sizes).

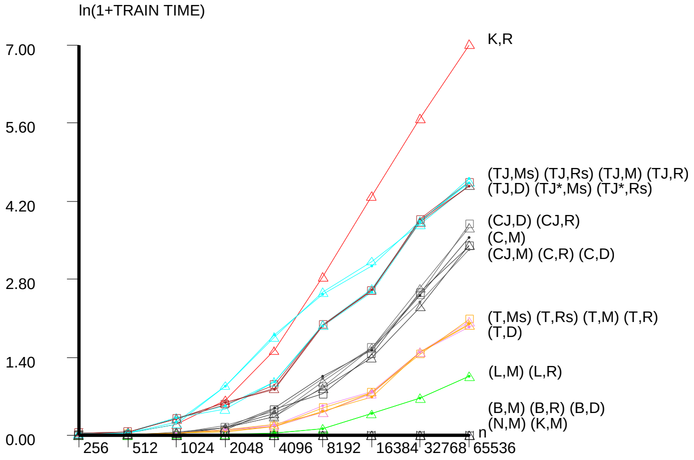


Figure 9.31: Logarithmic (natural) training times as functions of sample size. The Y-axis is seconds in logaritmic scale (natural base).

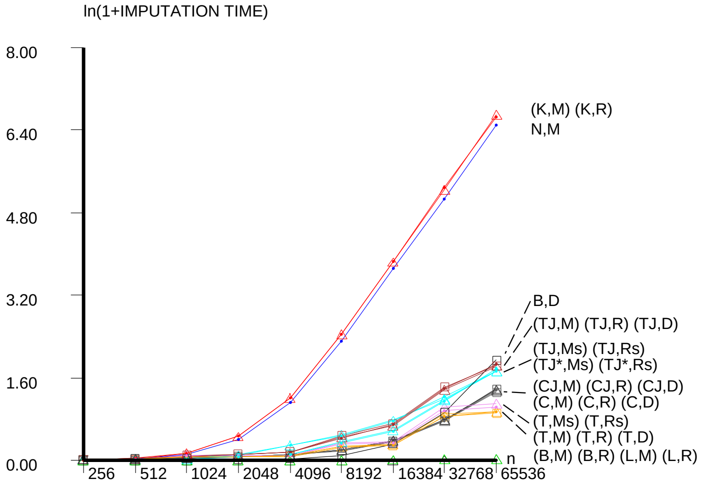


Figure 9.32: Logarithmic imputation times as functions of sample size. The Y-axis is seconds in logaritmic scale (natural base).

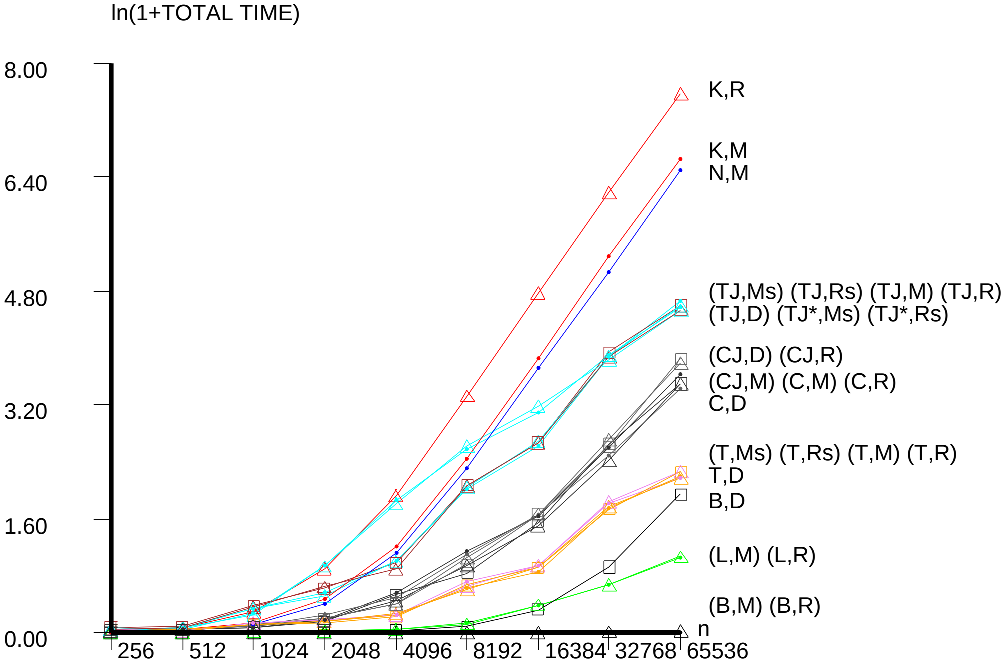


Figure 9.33: Logarithmic total times as functions of sample size. The Y-axis logarithmic values 1.6, 4.8 and 6.4 correspond to 4, 120 and 600 seconds. The maximum value (method K,R) is approximately 7.56 which corresponds to 1920 seconds. The Y-axis is seconds in logaritmic scale (natural base).

n	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768	65536
Method															
B,M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B,R	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02
B,D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.38	1.51	5.97
L,M	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.04	0.14	0.48	0.96	1.86
L,R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.13	0.48	0.97	1.91
N,M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.12	0.50	2.08	9.16	40.21	158.51	662.58
K,M	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.04	0.14	0.59	2.36	10.60	45.95	196.44	772.29
K,R	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.08	0.35	1.47	5.93	26.82	117.60	482.04	1928.57
T,M	0.01	0.01	0.01	0.01	0.02	0.03	0.03	0.04	0.12	0.16	0.27	0.86	1.33	4.77	7.98
T,R	0.00	0.01	0.01	0.00	0.02	0.02	0.02	0.03	0.13	0.18	0.31	0.83	1.53	4.86	7.85
T,D	0.01	0.01	0.00	0.01	0.02	0.02	0.03	0.04	0.11	0.14	0.25	0.94	1.49	4.63	8.52
T,M^s	0.01	0.01	0.01	0.00	0.03	0.02	0.03	0.03	0.14	0.20	0.28	1.06	1.57	5.11	7.74
T,R^s	0.01	0.01	0.01	0.01	0.03	0.03	0.04	0.03	0.14	0.20	0.28	0.93	1.56	5.26	8.67
C,M	0.01	0.01	0.00	0.01	0.01	0.02	0.02	0.05	0.07	0.21	0.74	2.14	4.09	12.54	36.66
C,R	0.01	0.00	0.00	0.02	0.01	0.02	0.02	0.05	0.09	0.21	0.58	1.55	3.49	10.19	32.21
C,D	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.04	0.08	0.16	0.70	1.32	3.73	13.11	31.97
TJ,M	0.01	0.01	0.01	0.01	0.02	0.03	0.05	0.10	0.42	0.92	1.48	6.60	13.59	49.87	92.54
TJ,R	0.01	0.01	0.01	0.01	0.02	0.03	0.05	0.08	0.43	0.90	1.47	6.86	13.43	47.52	92.29
TJ,D	0.02	0.01	0.01	0.01	0.02	0.04	0.07	0.09	0.45	0.87	1.67	6.92	13.40	50.49	97.87
TJ,M^s	0.01	0.00	0.00	0.01	0.01	0.02	0.04	0.08	0.40	0.75	1.75	6.47	12.84	46.80	104.03
TJ,R^s	0.01	0.01	0.01	0.00	0.01	0.02	0.05	0.07	0.41	0.67	1.74	6.64	13.73	47.67	97.83
TJ*,M^s	0.01	0.00	0.00	0.00	0.01	0.03	0.06	0.05	0.30	1.55	5.36	12.18	21.03	48.51	96.57
TJ*,R^s	0.00	0.00	0.00	0.01	0.02	0.02	0.03	0.06	0.27	1.54	5.16	12.77	22.96	44.99	92.02
CJ,M	0.01	0.00	0.00	0.01	0.01	0.02	0.01	0.04	0.09	0.27	0.65	1.97	4.26	11.03	29.84
CJ,R	0.01	0.00	0.00	0.01	0.00	0.02	0.03	0.04	0.11	0.22	0.51	1.84	4.30	14.17	43.20
CJ,D	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.04	0.09	0.22	0.50	1.61	4.25	12.62	45.98

Table 9.18: Total computational times in seconds as functions of the sample size.

From the total times one can notice that the nonparametric regression methods are the slowest, as expected. Further, computational requirements grow fastest for them (at least when considering imputation and total times). The second slowest are the TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods. Then the K-Means cell methods, the standard $\boldsymbol{X}$-clustering TS-SOM cell methods, baseline random donor, linear regression, and the rest of baseline methods follow. An unexpected result, the slowness of the baseline random donor method, is due to inefficient implementation of sampling with weights. With a better implementation the total time would likely to be close to the times of the other two baseline methods.

Recall our note about the optimization of performance of k-nearest neighbour for $k = 1$ from Section 7.4.3. In this study we have optimized our implementation of k-nearest neighbour for $k = 1$. As a consequence, imputation times are roughly the same as for the kernel regression method K,M. However, even with optimizations the

1-nearest neighbour method is slow. Next, a brief analysis for 1-nearest neighbour is given to illustrate why nonparametric regression methods are so slow. This is followed by a more detailed discussion on training and imputation time results.

Let an incomplete data set have $\mathsf{n}^{mis} \approx p^*\mathsf{n}$ missing values and $\mathsf{n}^{obs} \approx (1-p^*)\mathsf{n}$ observed data values. The number of operations required by the 1-nearest neighbour method is:

$$\begin{aligned}
\mathcal{T}^{N,M} &= \underbrace{\mathsf{n}^{mis}}_{\text{how many to impute}} * \Big(\underbrace{\mathsf{n}^{obs}}_{\text{distance computations}} \\
&\quad + \underbrace{\mathsf{n}^{obs}}_{\text{comparisons to search minimum distance}} \Big) \\
&\approx p^*\mathsf{n} * \left((1-p^*)\mathsf{n} + (1-p^*)\mathsf{n} \right) \\
&= \underbrace{p^*(1-p^*)\mathsf{n}^2}_{\text{distance computations}} + \underbrace{p^*(1-p^*)\mathsf{n}^2}_{\text{comparisons to search minimum distances}} ,
\end{aligned}$$

where $p^* \approx 0.39$ in this study. From the above formula one can notice that the number of operations equals the sample size squared. Methods requiring sample size squared operations are slow.

Training of the model is the most expensive for kernel regression with simulated random imputation. The reason is that estimation of residual variance is expensive, as one has to form n^{obs} smoothed predictions of observed Y values. The second slowest methods are the TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods. However, one must recall that they utilize approximately 64% more observations than do the other methods: incomplete observations are used too. Further, the number of cells is considerably larger for the TS-SOM methods than for the K-Means method which are the third lowest with respect to training time. The standard $\boldsymbol{X}$-clustering TS-SOM cell methods are quite fast, but both the linear regression methods, the three baseline methods, and the two nonparametric regression methods with mean strategy are even faster. The nonparametric regression methods with mean strategy require no training, because 'the observed data set is the model'.

In imputation the nonparametric methods are the slowest. The second slowest are the TS-SOM cell methods, which are followed by the K-Means methods. As earlier, the standard $\boldsymbol{X}$-clustering TS-SOM cell methods are fast. The baseline methods, excluding donor strategy, and the linear regression methods are the fastest in imputation. As the sample size grows, the time consumed by the baseline random donor method grows faster than one would except. However, as discussed earlier, this is due to inefficient implementation of sampling with weights.

To summarize, one can conclude that the nonparametric regression methods might be inusable if the sample size is too large. The computational times of all the other methods are clearly faster than those of the nonparametric ones. The standard $\boldsymbol{X}$-clustering TS-SOM cell methods are especially fast when compared to the other cell methods. In the study they were from 2.5 to 3 times faster than the corresponding K-Means methods. However, recall that the number of cells for the TS-SOM methods was much larger. Therefore in a 'fair comparison' (same number of cells) the TS-SOM standard $\boldsymbol{X}$-clustering cell methods are more than thrice as fast as the corresponding K-Means methods. Finally, it is fair to remark that nonparametric regression methods can be made faster. Discretization, binning, of data can be used to reduce the computational burden of nonparametric methods. Training data for nearest neighbour imputation may be replaced by K-Means centroids [40]. Further, training data can be binned for kernel regression [41]. However, these improvements are likely to yield an increase of biases in predictions.

9.5 Summary

In this chapter studies were conducted using a labour force data set. The imputation target was either the almost continuous variable AGE or categorial SEX. Earlier in this thesis we have not evaluated the imputation performance for categorial data which is common in practice. Therefore the experiments for variable SEX give valuable information.

In the AGE experiment, the nonparametric regression methods with mean strategy, the TS-SOM joint $(Y, \boldsymbol{X})$-clustering method with mean strategy and smoothing, and the standard $\boldsymbol{X}$-clustering methods with donor strategy performed the best at simultaneous preservation of unit level and distribution level. In the analysis of results for preservation of quantiles for age it was noticed that, typically, cell methods with donor strategy and TS-SOM standard $\boldsymbol{X}$-clustering with random strategy perform well. Most of the best results for the five analysed quantiles were achieved with cell methods. Preservation of the relationship between a target and a covariate was also evaluated by analysing imputation performance using conditional distribution of age given marital status. It was noticed that the TS-SOM joint $(Y, \boldsymbol{X})$-clustering cell method with random strategy and smoothing performs well. In addition, nearest neighbour imputation and K-Means joint $(Y, \boldsymbol{X})$ clustering with donor strategy perform well too.

In the SEX experiment the best unit level predictions were derived by linear regression with mean strategy and the standard $\boldsymbol{X}$-clustering cell methods with smoothed or non-smoothed mean strategy. Good performance for linear regression was expected because the variables with the highest linear dependency to male class were selected as covariates. The nonparametric regression methods with mean strategy performed also well. Some of the TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods performed better than nearest neighbour at unit level. The nearest neighbour method and the TS-SOM standard $\boldsymbol{X}$-clustering method with either random strategy (without smoothing) or with donor strategy yielded the best distribution level results for the largest sample sizes. Also the TS-SOM joint $(Y, \boldsymbol{X})$ clustering methods using smoothing had a good performance at distribution level. Unit and distribution level were simultaneously preserved best by the nearest neighbour imputation, the two TS-SOM joint $(Y, \boldsymbol{X})$-clustering cell methods utilizing smoothing and using only complete observations when training model, and the TS-SOM standard $\boldsymbol{X}$-clustering method with donor strategy.

The impact of sample size on imputations was measured with the SEX experiment. The nonparametric regression methods and some of the cell methods seemed to improve the results up to some level as sample size is increased (recall that the number of cells was also increased). This is good information as it means that one can sort of trust with this kind of data that these methods are able to yield a good performance for a suitably large sample size.

Computational time requirements were measured in the SEX experiment. It was found that the nonparametric regression methods are the slowest. Their requirements grow very fast, and may render them inusable in practice. All the other methods, including the cell methods, run much faster. The standard $\boldsymbol{X}$-clustering TS-SOM cell methods were quite fast. One must note that the nonparametric regression methods can be made faster by discretization of data as discussed earlier. However, this may be penalized by increases in predictions biases.

To conclude, the results of this chapter empirically justify that the proposed cell methods based on TS-SOM perform well with this kind of real-world data and MAR missingness. The nearest neighbour method is a good competitor for the proposed cell methods, provided sample size is not too large. Further, imputations using classical linear regression may be good enough. However, if the target and covariates are not linearly dependent and one is not able to linearize data (by suitable nonlinear transformations) then the linear regression method may not perform well.

Chapter 10

Conclusions and Future Work

The objective of this thesis was to analyse and evaluate cell imputation methods. This was done using both theoretical and empirical tools. For comparison, baseline, linear regression, and nonparametric regression were also included in the evaluation.

The claim was that cell methods provide a practical, multipurpose alternative to many imputation tasks. In addition, it was claimed that it is possible to build error estimates for cell imputation methods. This thesis contains all the necessary elements to prove these claims. Yet, due to time limitations the plug-in error estimate for cell imputation that was introduced in Chapter 6 was not used in the empirical studies in Chapters 7, 8 and 9. This was because the final details of the error estimates were completed after the empirical studies. In Chapters 4, 5 and 6 a theoretical framework to evaluate imputation errors was given. The practicality of error estimates was demonstrated in Algorithm 1 in page 122. The performance of cell imputation was evaluated against other methods in Chapters 7, 8, and 9. The results are clear: cell imputation provides a good alternative if one tries to preserve the distributional properties of data sets.

Carefully designed simulation studies were conducted in Chapter 7. Simulated studies allowed focusing on some specified phenomena (with real-world data sets different things get mixed and the interpretation of the results may be complicated). The first study showed that the proposed cell methods perform well under MCAR and MAR mechanisms. Preservation of multimodal distribution was analysed and evaluated in the second study. Approximate theoretical behaviour of the Kolmogorov-Smirnov (KS) statistic for the compared imputation methods were given. The numerical study results showed that the behaviour of KS for various methods was close to the expected behaviour. Further, the proposed TS-SOM based cell imputation was able to preserve multimodal marginal distribution and conditional distribution at a specified point under NMAR missingness. In the last study, the impact of data dimension to imputations was evaluated. Curse of dimensionality (in the form of increased imputation errors) appeared in most of the methods including the cell methods. Finally the cell methods showed good computational complexities, especially for large sample sizes, compared to the nonparametric regression methods, which can be considered to be the 'nearest' competitors.

Chapters 8 and 9 contained two empirical cases which demonstrated that the proposed cell methods perform well with real-world data sets. In Chapter 8 imputation of turnover of enterprise was evaluated using a small and medium-sized enterprises data set of the year 2004 from the UK. Further, a UK Labour Force data set from the year 2006 was used to analyse imputation of age and sex of respondents in Chapter 9. The proposed TS-SOM joint $(Y, \boldsymbol{X})$ clustering method came up among the best methods. The linear regression with added noise and 1-nearest neighbour methods performed also well. However, for large data sets 1-nearest neighbour (or kernel regression) has a high computational complexity that may render it inusable. Further, if the linearity assumption does not hold, and one is not able to linearize the data with suitable transformations then linear regression may not be usable. Cell methods on the other hand are nonparametric by nature and have clearly better computational properties than nonparametric regression methods. Therefore the proposed cell methods are applicable in practice.

10.1 Ideas for Future Work

Considering the future, researchers have at least the following ways to improve the work done in this thesis:

- make some of the formulas more accurate (for example Approximation 4.14 for variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}]$ of linear regression methods)
- simplify some analytical formulas (to make them more interpretable) (for example Approximation 5.3 for variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}]$ of nonparametric regression methods)
- test derived analytical results with real-world data sets. Especially more work is needed to compare theoretical error estimates $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,T}|\mathsf{n}]$ and $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,TJ}|\mathsf{n}]$ (Section 6.4.2, p.121) with real-world experiments.

Note that to present some of the formulas in a more accurate way may require one to apply higher order Taylor approximations. This would make the formulas even longer. Thus, simplification of those formulas may not be easy. Instead of Taylor approximations one might use some other techniques, such as variance upper bounds, which could yield accurate and simple (interpretable) results.

Bibliography

[1] Aldrich, J.: *R. A. Fisher and the Making of Maximum Likelihood 1912-1922*, Statistical Science, Vol. 12, No. 3, pp. 162-176, 1997

[2] Allison, P. D.: *Missing Data*, Series: Quantitative Applications in the Social Sciences (Series/Number 07-136), Sage University Press, 2002

[3] Anderson, T. W.: *Maximum Likelihood Estimates for a Multivariate Normal Distribution when some Observations are Missing*, Journal of the American Statistical Association, Vol. 52, No. 278, pp. 200-203, 1957

[4] Barnett, V.: *Elements Of Sampling Theory*, London : Hodder and Stoughton, 1982

[5] Bishop, C. M.: *Neural Networks for Pattern Recognition*, Oxford University Press, Inc., New York, 1995 Clarendon Press, Oxford, 1995

[6] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J.: *Classification and Regression Trees*, Chapman & Hall / CRC, 1984

[7] Buck, S. F.: *A Method of Estimation of Missing Values in Multivariate Data Suitable for use with an Electronic Computer*, Journal of the Royal Statistical Society Series B (Methodological), Vol 22., No. 2, pp. 302-306, 1960

[8] Celmins, A.: *The Method of Gauss in 1799*, Statistical Science, Vol. 13, No. 2, pp. 123-135, 1998

[9] Černý, V.: *A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm*, Journal of Optimization Theory and Applications, Vol. 45, No. 1, pp. 41-51, 1985

[10] Chambers, R.: *Evaluation Criteria for Statistical Editing and Imputation*, National Statistics Methodological Series No. 28, Her Majesty's Stationery Office (HMSO), 2001

[11] Charlton, J.: *Evaluation automatic edit and imputation methods, and the Euredit project*, Journal of the Royal Statistical Society: A, 167, Issue 2, pp. 199-207, 2004

[12] Chen, J.; Shao, J.: *Biases and variances of survey estimators based on nearest neighbour imputation*, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 365-370, 1997

[13] Hua Yun Chen; Little, R. J. A.: *A Test of Missing Completely at Random for Generalised Estimating Equations with Missing Data*, Biometrika, Vol. 86, No. 1, pp. 1-13, 1999

[14] Cheng, P. E.: *Nonparametric Estimation of Mean Functionals with Data Missing at Random*, Journal of the American Statistical Association, Vol. 89, No. 425, pp. 81-87, 1994

[15] Cottrell, M.; Letrémy, P.: *Missing values: processing with the Kohonen algorithm*, Proceedings of XIth International Symposium on Applied Stochastic Models and Data Analysis (ASMDA), pp. 489-496, 2005

[16] Young, J.; Lees, K.; Austin, J.: *Performance comparison of correlation matrix memory implementations*, Neural Information Processing, Proceedings, ICONIP '66, 6th International Coference on Neural Information Processing, Vol. 2, pp. 570-575, 1999

[17] Creel, D. V.; Krotki, K.: *Creating imputation classes using classification tree methodology*, Proceedings of the Survey Research Methods Section (ASA), pp. 2884-2887, 2006

[18] Dempster, A. P.; Laird, N. M.; Rubin, D. B.: *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1, pp. 1-38, 1977

[19] Diggle, P. J.: *Testing for Random Dropouts in Repeated Measurement Data*, Biometrics, Vol. 45, No. 4, pp. 1255-1258, 1989

[20] Duda, R. O.; Hart, P. E.; Stork, D. G.: *Pattern Classification*, 2nd edition, John Wiley & Sons, Inc., 2001

[21] Edgett, G. L.: *Multiple Regression with Missing Observations among the Independent Variables*, Journal of the American Statistical Association, Vol. 51, No. 273, pp. 122-131, 1956

[22] Everitt, B. S.: *The Cambridge Dictionary of Statistics*, 2nd edition, Cambridge University Press, 2002

[23] Fellegi, I. P.; Holt, D.: *Systematic Approach to Automatic Edit and Imputation*, Journal of the American Statistical Association, Vol. 71, No. 353, pp. 17-35, 1976

[24] Feller, W.: *On a General Class of "Contagious" Distributions*, The Annals of Mathematical Statistics, Vol. 14, No. 4., pp. 389-400, 1943

[25] Fessant, F.; Midenet, S.: *Self-Organising Map for Data Imputation and Correction in Surveys*, Neural Computing & Applications, pp. 300-310, 2002

[26] Fraser, S.: *United Kingdom Survey of Small- and Medium-sized Enterprises' Finances*, 2004 [computer file]. Colchester, Essex: UK Data Archive [distributor], February 2006. SN: 5326.

[27] Fuchs, C.: *Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data*, Journal of The American Statistical Association, Vol. 77, No. 378, pp. 270-278, 1982

[28] Gasser, T.; Müller, H.-G.: *Kernel Estimation of Regression Functions*, Smoothing Techniques for Curve Estimation, Ed. Gasser, T.; Rosenblatt, M., Lecture Notes in Mathematics 757, pp. 23-68, 1979

[29] Gasser, T.; Müller, H.-G.: *Estimating Regression Functions and Their Derivatives by the Kernel Method*, Scandinavian journal of statistics, Vol. 11, pp. 171-185, 1984

[30] Gelman, A; Carlin, J. B.; Stern, H. S.; Rubin, D. B.: *Bayesian Data Analysis*, Chapman & Hall/CRC, 2004

[31] Hartigan, J. A.: *Clustering Algorithms*, John Wiley & Sons, 1975

[32] Hartigan, J. A.; Wong, M. A.: *A K-Means Clustering Algorithm*, Applied Statistics, Vol. 28, No. 1, pp. 100-108, 1979

[33] Hartley, H. O.: *A Plan for Programming Analysis of Variance for General Purpose Computers*, Biometrics, Vol. 12, No. 2, pp. 110-122, 1956

[34] Hartley, H. O.: *Maximum likelihood estimation from Incomplete Data*, Biometrics, Vol. 14, No. 2, pp. 174-194, 1958

[35] Hastie, T.; Stuezle, W.: *Principal Curves*, JASA 406, pp. 502-516, 1989

[36] Hastie, T.; Tibshirani, R.; Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001

[37] Hess, K. T.: *Random Partitions of Samples*, Dresdner Schriften zur Versicherungsmathematik, Technische Universität Dresden, Institut für Mathematische Stochastik, 1/2000

[38] Hocking, R. R.; Smith, Wm. B.: *Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations*, Journal of the American Statistical Association, Vol. 63, No. 321, pp. 159-173, 1968

[39] Horvitz, D. G.; Thompson, D. J.: *A Generalization of sampling Without Replacement From a Finite Universe*, Journal of the American Statistical Association, Vol. 47, No. 260, pp. 663-685, 1952

[40] Hruschka, E. R.; Hruschka, E. R. Jr., Nelson, F. F. E.: *Towards Efficient Imputation by Nearest-Neighbors: A Clustering-Based Approach*, 17th Joint Australian Conference on Artificial Intelligence - AI'04, 2004, Cairns. Lecture Notes in Computer Science (LNAI 3339), Berlin: Springer-Verlag, Vol. 3339, pp. 513-525, 2004.

[41] Hoti, F.; Holmström, L.: *On the estimation error in binned local linear regression*, Journal of Nonparametric Statistics, Vol. 15, Issue 4-5, pp. 625-642, 2003.

[42] Häkkinen, E.: *Design, implementation and evaluation of the Neural Data Analysis environment*, PhD thesis, University of Jyväskylä, 2001

[43] Jaszi, G.: *National Income: Status and Prospects as Seen by an Estimator*, Journal of the American Statistical Association, Vol. 46, No. 255, pp. 345-357, 1951

[44] Jann-Huei Jinn: *The Effect of Different Imputation Methods on Analytical Statistics of Simple Linear Regression*, InterStat, 2000, available online **interstat.stat.vt.edu/InterStat/ARTICLES/2000/articles/O00002.pdf** referenced 31.12.2003.

[45] Johnson, D. H.; Dudgeon, D. E.: *Array Signal Processing - Concepts and Techniques*, Prentice Hall, Inc., 1993

[46] Kalton, G.; Kasprzyk, D.: *Imputing for Missing Survey Responses*, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 22-31, 1982

[47] Kalton, G.; Kish, L.: *Two Efficient Random Imputation Procedures*, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 146-151, 1981

[48] Keeney, R. L.; Raiffa H.: *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley & Sons, 1976

[49] G. M. P. van Kempen; L. J. van Vliet: *Mean and Variance of Ratio Estimators Used in Fluorescence Ratio Imaging*, Cytometry 39, pp. 300-305, 2000

[50] Kim, J. K.; Fuller, W.: *Fractional hot deck imputation*, Biometrika, 91, 3, pp. 559-578, 2004

[51] King, G.; Honaker, J.; Joseph, A.; Scheve, K.: *Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation*, American Political Science Review, Vol. 95, No. 1, pp. 49-69, 2001

[52] Kirkpatrick, S.; Gelatt, C. D.; Vecchi M. P.: *Optimization by Simulated Annealing*, Science, Vol. 220, Number 4598, pp. 671-680, 1983

[53] Knottnerus, P.: *Sample Survey Theory: Some Pythagorean Perspectives*, 2003

[54] Kohonen, T.: *Self-Organizing Maps*, Springer, Berlin, 1997

[55] Koikkalainen, P.; Oja, E.: *Self-Organizing Hierarchical Feature Maps*, In Proc. IJCNN-90: International Joint Conference on Neural Networks, IEEE Press, pp. 279-284, 1990

[56] Koikkalainen, P.: *Tree Structured Self-Organizing Maps*, In Oja, E. and Kaski, S., eds., Kohonen Maps, Elsevier, The Netherlands, pp. 121-130, 1999

[57] Koikkalainen, P.; Horppu, I.: *Handling Missing Data with the Tree-Structured Self-Organizing Map*, Proceedings of the 2007 International Joint Conference on Neural Networks (IJCNN 2007), to appear.

[58] Laaksonen, S.: *Regression-Based Nearest Neighbour Hot Decking*, ZUMA, Methoden und Analysen, No. 4., Nonresponse in Survey Research, Mannheim, pp. 285-298, 1998

[59] Lai, Y. L.: *Analysis of incomplete survey data with application to the construction of social indicators of Hong Kong*, M.Phil. Thesis, The University of Hong Kong, 1998

[60] LeBlanc, M.; Tibshirani, R.: *Adaptive Principal Surfaces*, JASA 425, pp. 53-64, 1994

[61] Little, R. J. A.: *A Test of Missing Completely at Random for Multivariate Data with Missing Values*, Journal of the American Statistical Association, Vol. 83, No. 404, pp. 1198-1202, 1988

[62] Little, R. J. A.; Rubin, D. B.: *Statistical Analysis With Missing Data*, 2nd edition, New York: Wiley, 2002

[63] Longford, N. T.: *Missing Data and Small-Area Estimation*, Springer, 2005

[64] Lord, F. M.: *Estimation of Parameters from Incomplete Data*, Journal of the American Statistical Association, Vol. 50, No. 271, pp. 870-876, 1955

[65] Lloyd, S. P.: *Least squares quantization in PCM*, IEEE Transactions on Information Theory, Vol. 28, pp. 129-137, 1982

[66] Magnus, R. M.; Neudecker, H.: *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley Series in Probability and Statistics, 2002

[67] Mandelbrot, B. B.: *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, Springer-Verlag, 1997

[68] Xiao-Li Meng; Rubin, D. B.: *Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm*, Journal of the American Statistical Association, Vol. 86, Issue 416, pp. 899-909, 1991

[69] Myrtveit, I.; Stensrud, E.; Olsson, U. H.: *Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods*, IEEE Transactions on Software Engineering, Vol. 27, No. 11, pp. 999-1013, 2001

[70] Nadaraya, É.: *On Non-Parametric Estimates of Density Functions and Regression Curves*, Theory of Probability & Its Applications, Vol. 10, Issue 1, pp. 186-190, 1965

[71] Carreira-Perpiñán, M. Á.: *Continuous latent variable models for dimensionality reduction and sequential data reconstruction*, PhD thesis, Department of Computer Science, University of Sheffield, UK, 2001

[72] Mack, Y. P.: *Local Properties of k-NN Regression Estimates*, SIAM Journal on Algebraic and Discrete Methods, Vol. 2, No. 3, pp. 311-323, 1981

[73] Mittinty, Murthy N.; Chacko, E.: *Imputation by Propensity Matching*, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 4022-4028, 2004

[74] Morrison, D. F.: *Expectations and Variances of Maximum Likelihood Estimates of the Multivariate Normal Distribution Parameters with Missing Data*, Journal of the American Statistical Association, Vol. 66, No. 335, pp. 602-604, 1971

[75] Nicholson, G. E.: *Estimation of Parameters From Incomplete Multivariate Samples*, Journal of the American Statistical Association, Vol. 52, No. 280, pp. 523-526, 1957

[76] Office for National Statistics. Social and Vital Statistics Division and Northern Ireland Statistics and Research Agency. Central Survey Unit: *Quarterly Labour Force Survey Household Dataset, April - June*, 2006 [computer file]. 1st Edition. Colchester, Essex: UK Data Archive [distributor], October 2006. SN: 5500.

[77] Phillips, H. S.: *United Kingdom Indices of Wholesale Prices*, Journal of the Royal Statistical Society. Series A (General), Vol. 199, No. 3, pp. 239-283, 1956

[78] Pollard, D.: *Strong Consistency of K-Means Clustering*, The Annals of Statistics, Vol. 9, No. 1 (Jan.), pp. 135-140, 1981

[79] Pollard, D.: *A Central Limit Theorem for k-Means Clustering*, The Annals of Probability, Vol. 10, No. 4 (Nov.), pp. 919-926, 1982

[80] Rallo, R.; Ferré-Giné, J.; Giralt, F.: *Best Feature Selection and Data Completion for the Design of Soft Neural Sensors*, Proceedings of AIChE 2003, 2nd Topical Conference on Sensors, San Francisco, November 2003

[81] Rallo, R.; Ferré-Giné, J.; Giralt, F.: *Design of soft sensors with multiple imputation of missing data by self-organizing map ensembles*, Proceedings of the 7th World Congress of Chemical Engineering (CDROM WCCE'05)

[82] Rao, J. N. K.: *Small Area Estimation*, John Wiley & Sons, 2003

[83] Rideout, M. S.; Diggle, P. J.: *Testing for Random Dropouts in Repeated Measurement Data*, Biometrics, Vol. 47, No. 4, pp. 1617-1621, 1991

[84] Rockwell, R. C.: *An Investigation of Imputation and Differential Quality of Data in the 1970 Census*, Journal of the American Statistical Association, Vol. 70, No. 349, pp. 39-42, 1975

[85] Rosenblatt, M.: *Conditional probability density and regression estimates*, Multivariate Analysis II, Ed. Krishnaiah, pp. 25-31, 1969

[86] Rubin, D. B.: *A Non-Iterative Algorithm for Least Squares Estimation of Missing Values in Any Analysis of Variance Design*, Applied Statistics, Vol. 21, No. 2, pp. 136-141, 1972

[87] Rubin, D. B.: *Multiple Imputation for nonresponse in surveys*, New York : Wiley, 1987

[88] Rubin, D. B.: *Formalizing subjective notions about the effect of nonrespondents in sample surveys*, Journal of the American Statistical Association, Vol. 72, No. 359, pp. 538-543, 1977

[89] Rubin, D. B.: *Multiple Imputation After 18+ Years*, Journal of the American Statistical Association, Vol. 91, No. 434, pp. 473-489, 1996

[90] Santos, R.: *Effects of imputation on complex statistics*, Survey Research Center, Institute for Social Research, University of Michigan, 1981

[91] Schafer, J. L.: *Analysis of Incomplete Multivariate Data*, Monographs on Statistics and Applied Probability 72, Chapman & Hall/CRC, 2000

[92] Serfling, R. J.: *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, 2001

[93] Shah, S. M.: *On Estimating the Parameter of a Doubly Truncated Binomial Distribution*, Journal of the American Statistical Association, Vol. 61, No. 313. (Mar.), pp. 259-263, 1966

[94] Shen, S. M.; Lai, Y. L.: *Handling Incomplete Quality-of-Life Data*, Social Indicators Research 55, pp. 121-166, 2001.

[95] Simon, G. A.; Simonoff, J. S.: *Diagnostic Plots for Missing Data in Least Squares Regression*, Journal of the American Statistical Association, Vol. 81, No. 394, pp.501-509, 1986

[96] Simonoff, J. S.: *Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression*, Technometrics, Vol. 30, No. 2, pp.205-214, 1988

[97] Skrivastava, R. C.: *Estimation of Probability Density Function Based on Random Number of Observations with Applications*, International Statistical Review / Revue Internationale de Statistique, Vol. 41, No. 1. (Apr.), pp. 77-86, 1973

[98] Steuer, R. E.: *Multiple criteria optimization : theory, computation, and application*, New York, Wiley, 1986

[99] Stigler: *Thomas Bayes's Bayesian Inference*, Journal of the Royal Statistical Society Seris A (General), Vol. 145, No. 2, pp. 250-258, 1982

[100] Tanner, M. A.; Wing Hung Wong: *The Calculation of Posterior Distributions by Data Augmentation*, Journal of the American Statistical Association, Vol. 82, No. 398, pp. 528-540, 1987

[101] Titterington, D. M.; Mill, G. M.: *Kernel-Based Density Estimates from Incomplete Data*, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 45, No. 2, pp. 258-266, 1983

[102] Toutenburg, H.; Srivastava, V. K.: *Estimation of Linear Regression Models with Missingness of Observations on Both the Explanatory and Study Variables-Part I: Theoretical Results*, SFB Discussion Paper 184, Universitat Munchen, Munchen, Germany, 2000

[103] Toutenburg, H.; Srivastava, V. K.; Fieger, A.: *Estimation of Parameters in Multiple Regression With Missing X-Observations using Modified First Order Regression*, SFB Discussion Paper 38, Universitat Munchen, Munchen, Germany, 1996

[104] Toutenburg, H.; Srivastava, V. K.: *A Revisit to the Application of Weighted Mixed Regression Estimation in Linear Regression Models with Missing Data*, SFB Discussion Paper 241, Universitat Munchen, Munchen, Germany, 2001

[105] Toutenburg, H.; Fieger A.: *Using diagnostic measures to detect non-MCAR processes in linear regression models with missing covariates*, SFB Discussion Paper 204, Universitat Munchen, Munchen, Germany, 2000

[106] Toutenburg, H.; Heumann, C.; Nittner, T.; Scheid, S.: *Parametric and Nonparametric Regression with Missing X's - A Review*, SFB Discussion Paper 286, Universitat Munchen, Munchen, Germany, 2002

[107] Toutenburg, H.; Shalabh: *Use of Prior Information in the form of Interval Constraints for the Improved Estimation of Linear Regression Models with Some Missing Responses*, SFB Discussion Paper 240, Universitat Munchen, Munchen, Germany, 2001

[108] Vapnik, V. N.: *Statistical learning theory*, Wiley, New York, 1998

[109] Watson, G. S.: *Smooth regression analysis*, Sankhya, Ser. A, Vol. 26, pp. 359-372, 1964.

[110] Wharton Jr, C. R.: *Processing Underdeveloped Data from an Underdeveloped Area*, Journal of the American Statistical Association, Vol. 55, No. 289, pp. 23-37, 1960

[111] Wilks, S. S.: *Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples*, The Annals of Mathematical Statistics, Vol. 3, No. 3(Aug.), pp. 163-195, 1932

[112] Wu, C. F. J.: *On the Convergence Properties of the EM Algorithm*, The Annals of Statistics, Vol. 11, No. 1, pp. 95-103, 1983

[113] Yamane, T.: *Elementary Sampling Theory*, Englewood Cliffs (NJ) : Prentice-Hall, cop. 1967

[114] Hujun Yin; Allinson, N. M.: *On the Distribution and Convergence of Feature Space in Self-Organizing Maps*, Neural Computation 7, pp. 1178-1187, 1995

[115] Zhao, X.: *Imputation by Neural Networks and Related Methods*, PhD Thesis, Department of Social Statistics, Faculty of Social Sciences, University of Southampton, 2002

[116] *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, 1st volume of the EurEdit scientific report, Eurostat, 2003, available online **www.cs.york.ac.uk/euredit/results/results.html** referenced 31.5.2007

Appendix for Chapter 1

In this appendix list of partners in the EurEdit project is given. In addition, list of operators and symbols, and lists of tables and figures are included.

List of partners in the EurEdit project

Partners in the EurEdit project were (followed by abbreviation in paranthesis) were

- Statistics Denmark (DST),
- University of York, UK, (YORK),
- Royal Holloway and Bedford New College - University of London, (RHUL),
- Statistics Finland (STATFI),
- Office for National Statistics, UK (ONS)
- University of Jyväskylä (JYU),
- Statistics Netherlands (CBS),
- University of Southampton (SOTON),
- Swiss Federal Statistical Office (SFSO),
- Statistics Italy (ISTAT),
- The Numerical Algorithms Group Ltd (NAG), and
- Qantaris.

As supplemental information we give here references to some of methods and techniques which were used by the partners. Multilayer perceptron network (MLP) was used by Statistics Denmark and Statistics Italy, see [5] for details on MLP. Implementation of imputation methods used by University of York were based on correlation matrix memory (CMM) technique [16]. Statistics Finland used for example regression based nearest neighbour imputation [58, 73]. Royal Holloway did imputations using support vector machines [108].

Operator and symbol list

Main operators and symbols which are used in this thesis are:

Operator/symbol	Description
argmax	Argument of the maximum.
sup	Supremum.
lim	Limit operator.
$\|\| \cdot \|\|_2^2$	Euclidean norm (L_2).
$\int$	Integral.
$\Pr(A)$	Probability of event A.
$\mathbb{B}\text{ias}[\hat{\theta}]$	Bias of $\hat{\theta}$.
$\mathbb{C}\text{ov}[X, Y]$	Covariance between X and Y.
$I(\cdot)$	Indicator function.
$\mathbf{I}$	Identity matrix.
$\text{tr}(\cdot)$	Trace operator: sum of diagonal elements of a given matrix.
$vec(\cdot)$	Vector operator: stacks columns of given matrix above each other (from first column to last).
$\mathbf{1}$	Vector of ones.
$Bin(\mathsf{n}, p^*)$	Binomial distribution with n trials and probability p^*.
$Multin(\mathsf{n}; w_1, \ldots, w_l)$	Multinomial distribution with n trials and probabilities $w_1, \ldots, w_l$
$cat(w_1, \ldots, w_l)$	Categorical distribution for random variable with outcomes $\{1, \ldots, l\}$ and probabilities $w_1, \ldots, w_l$.
μ^*	Expectation of target Y.
$\hat{\mu}$	Estimator of μ^*.
μ	Estimate of μ^*.
τ^*	Variance of target Y.
$\overline{\boldsymbol{X}}^*$	Expectation of covariate $\boldsymbol{X} = (X_1, \ldots, X_{p-1})^T$.
$p - 1$	Dimension of $\boldsymbol{X}$.
$\boldsymbol{\Sigma}_{\boldsymbol{X}}^*$	Variance of $\boldsymbol{X}$.
$g^*(\mathbf{x})$	Conditional expectation of Y given $\boldsymbol{X} = \mathbf{x}$.
$v^*(\mathbf{x})$	Conditional variance of Y given $\boldsymbol{X} = \mathbf{x}$.
v^*	Expectation of $v^*(\boldsymbol{X})$ over distribution of $\boldsymbol{X}$.
n	Sample size (fixed).
N^{mis}	Number of missing data values (random variable).
p^*	Probability for missingness (in target Y).
$p_{\mathbf{x}}^*$	Probability for missingness (in target Y) given $\boldsymbol{X} = \mathbf{x}$.
N^{obs}	Number of observed data values (r.v.).
Y_j	j:th random observation of Y.
$\boldsymbol{X}_j$	j:th random observation of $\boldsymbol{X}$.

Operator/symbol	Description
R_j	j:th random response indicator.
$\boldsymbol{\beta}^*$	Optimal (in least squares sense) linear regression coefficients, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^*_{-0}, \beta^*_0)^T$.
$\boldsymbol{\beta}^*_{-0}$	Slope terms.
β^*_0	Intercept term.
λ	Smoothing bandwidth (non-parametric regression)
$\lambda(\mathsf{n}^{obs})$	Emphasizes that smoothing bandwidth is varied as a function of n^{obs}.
k	Number of neighbours to use (in nearest neighbour imputation).
$k(\mathsf{n}^{obs})$	Emphasizes that number of neighbours is varied as a function of n^{obs}.
n_c	Number of cells (cell methods).
$\mathbf{w}_{\{i\}}$	Cell centroids: $\mathbf{w}_{\{i\}} = \mathbf{w}_{Y,\{i\}} \cup \mathbf{w}_{\boldsymbol{X},\{i\}}$, $i = 1, \ldots, \mathsf{n}_c$ where $\mathbf{w}_{Y,\{i\}}$ is Y part of centroids and $\mathbf{w}_{\boldsymbol{X},\{i\}}$ is the $\boldsymbol{X}$ part.
$h_{i,l}$	Smoothing parameter between cells i and l.
$b(\mathbf{x}')$	Hard classifier, where notation $\mathbf{x}'$ means that $\mathbf{x}'$ may consists of either $\mathbf{x}$ and $(y, \mathbf{x})$.
$b^\epsilon(\mathbf{x}')$	Probabilistic classifier.
$g_i(\mathbf{x}')$	Soft classifier which maps input $\mathbf{x}'$ to probability (or weight) in range $[0, 1]$ for cell i.
N_i	Number of observations in cell i (r.v.).
N_i^{obs}	Number of complete observations in cell i (r.v.).
N_i^{mis}	Number of incomplete observations in cell i (r.v.).
$\boldsymbol{N}^{mis}$	Number of incomplete observations in cells (r.v.): $\boldsymbol{\mathsf{N}}^{mis} = (\mathsf{N}_1^{mis}, \ldots, \mathsf{N}_{\mathsf{n}_c}^{mis})^T$
π_i	Cell i prior.
p_i	Missingness probability in cell i.
q_i	Correct classification probability of classifier for cell i.
$\mathbb{E}[\hat{q}_i]$	Expected classification success probability.
μ_i^*	Expectation of $Y\|i$, where i is cell index
τ_i^*	Variance of $Y\|i$

List of Tables

List of Figures

Appendix for Chapter 3

This appendix gives an overview of the basic statistical theory. Decompositions of second moments (for analytical quantity and estimator) are given. Finally limits for expectations of first moment estimator and second moment estimator are derived.

A3.1 Overview of the basic statistical theory

Analytical computations in this thesis utilize chain rules, some distributions, Taylor approximation, and magnitude of orders. Thus we describe them here briefly.

A3.1.1 Chain rules

Chain rules are useful for decomposing complicated integrations in easier tasks. In this thesis we utilize chain rules for expectation, variance, and covariance. Let X, Y, Z, W, be random variables.

Expectation of Y may be derived using chain rule as

$$\mathbb{E}[Y] = \mathbb{E}\Big[\mathbb{E}[Y|X]\Big],$$

where outer integration is with respect to distribution of X.

Variance of Y may be derived using chain rule as

$$\mathbb{V}\text{ar}[Y] = \mathbb{E}\Big[\mathbb{V}\text{ar}[Y|X]\Big] + \mathbb{V}\text{ar}\Big[\mathbb{E}[Y|X]\Big],$$

where the outer integrations are with respect to distribution of X. This decomposition is also useful as it allows one to interpret second moment of superpopulation. Namely it is sum of expected variance of noise and variability of conditional mean of Y given $\boldsymbol{X}$ (signal).

Covariance is defined as

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Chain rule for covariance is

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}\Big[\mathbb{C}\text{ov}[X, Y|Z]\Big] + \mathbb{C}\text{ov}\Big[\mathbb{E}[X|Z], \mathbb{E}[Y|Z]\Big],$$

where outer integrations are with respect to distribution of Z.

Conditional covariance is defined as

$$\mathbb{C}\mathrm{ov}[X,Y|Z] = \mathbb{E}\Big[(X-\mathbb{E}[X|Z])(Y-\mathbb{E}[Y|Z])\Big].$$

Chain rule for conditional covariance is

$$\mathbb{C}\mathrm{ov}[X,Y|Z] = \mathbb{E}\Big[\mathbb{C}\mathrm{ov}[X,Y|Z,W]\Big] + \mathbb{C}\mathrm{ov}\Big[\mathbb{E}[X|Z,W],\mathbb{E}[Y|Z,W]\Big],$$

where outer integrations are with respect to conditional distribution of W given Z.

A3.1.2 Distributions

In this thesis we have a lot of use for random variables which are Bernoulli, binomially, or multinomially distributed. We require statistical properties up to second order (expectation, variance, and covariance). Next we briefly describe these moments.

Random variable I which has value 1 with probability p and value 0 with probability $q = 1-p$ is Bernoulli distributed and is denoted as $X \sim \mathit{Bernoulli}(p)$, where

$$X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } q = 1-p \end{cases}$$

Basic result from statistics states that sum of n independently and identically distributed Bernoulli variables is binomially distributed with parameters n and p. Formally let $Z = X_1 + X_2 + \ldots + X_{\mathsf{n}}$, where X_j are iid sampled Bernoulli variables with success probability p. Probability for $Z = k$ in binomial distribution is $\binom{\mathsf{n}}{k}p^k q^{\mathsf{n}-k}$.

This can be generalized for multiple simultaneous outcomes $X_1, X_2, \ldots, X_l$ that sum to n. In other words if n items are divided between l classes such that probabilities for classes are $p_1, p_2, \ldots, p_l$ the distribution is called multinomial

$$Z_1, Z_2, \ldots, Z_l \sim \mathit{Multin}(\mathsf{n}; p_1, p_2, \ldots, p_l).$$

With average success $\alpha = \mathsf{n}p$ and sufficiently large number of trials with relatively small p binomial distribution becomes approximately Poisson distributed,

$$P(k) = \frac{\alpha^k}{k!}e^{-\alpha}$$

which again is approximately normal as $\mathsf{n} \to \infty$. Since we are mainly interested in properties with small n, it is the binomial and multinomial distributions that are most useful for us.

For the reader of this thesis the most important facts about multinomial distributions are that for $0 \le p_i \le 1, \sum p_i = 1$

$$Z_1, Z_2, \ldots, Z_l \sim \mathit{Multin}(\mathsf{n}; p_1, p_2, \ldots, p_l)$$

we know that,

$$\begin{aligned} &\mathbb{E}[Z_i] = \mathsf{n}p_i \\ &\mathbb{V}\mathrm{ar}[Z_i] = \mathsf{n}p_i(1-p_i) \\ &\mathbb{C}\mathrm{ov}[Z_i, Z_j] = -\mathsf{n}p_i p_j \text{ (when } i \neq j) \end{aligned}$$

First two above results follow from basic result that all marginal distributions Z_i are binomially distributed with n trials and probability parameter p_i, formally $Z_i \sim Bin(\mathsf{n}, p_i)$.

A3.1.3 Taylor approximation

One of the most important tools for us is the Taylor approximation. Continuous and differentiable function $g(\mathbf{x})$ may be approximated using Taylor series expansion. Expansion around $\mathbf{x}_0$ is written as

$$\begin{aligned} g(\mathbf{x}) &= g(\mathbf{x}_0) + \frac{1}{1!}(\mathbf{x}-\mathbf{x}_0)^T g'(\mathbf{x}_0) + \frac{1}{2!}(\mathbf{x}-\mathbf{x}_0)^T g''(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + R_{\mathsf{n}}(\mathbf{x}) \\ &= g(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T g'(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T g''(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + R_{\mathsf{n}}(\mathbf{x}) \end{aligned}$$

where g' denotes first derivative and g'' second, and $R_{\mathsf{n}}(\mathbf{x})$ is remainder (approximation error).

By using Taylor series expansion approximate expectation and variance of complicated non-linear functions parametrized by random variables may be computed. First two moments of (linear or non-linear) function of random variables is often needed in analytical derivations in this thesis.

Lets assume that $\boldsymbol{X}$ is a random vector variable with mean $\overline{\boldsymbol{X}}^*$ and variance $\boldsymbol{\Sigma}^*_{\boldsymbol{X}}$. Second order Taylor approximation around expectation of $\boldsymbol{X}$ yields to:

$$\begin{aligned} &\mathbb{E}[g(\boldsymbol{X})] \approx \mathbb{E}[g(\overline{\boldsymbol{X}}^*) + (\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T g'(\overline{\boldsymbol{X}}^*) + \frac{1}{2}(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T g''(\overline{\boldsymbol{X}}^*)(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)] \\ &= g(\overline{\boldsymbol{X}}^*) + \mathbb{E}[(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T] g'(\overline{\boldsymbol{X}}^*) + \frac{1}{2}\mathbb{E}[(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T g''(\overline{\boldsymbol{X}}^*)(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)] \\ &= g(\overline{\boldsymbol{X}}^*) + \frac{1}{2}\mathbb{E}[(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T g''(\overline{\boldsymbol{X}}^*)(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)] \\ &= g(\overline{\boldsymbol{X}}^*) + \frac{1}{2}\mathrm{tr}(\mathbb{E}[(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T g''(\overline{\boldsymbol{X}}^*)(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)]) \\ &= g(\overline{\boldsymbol{X}}^*) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^*_{\boldsymbol{X}} g''(\overline{\boldsymbol{X}}^*)), \end{aligned}$$

where g' and g'' denote first and second derivative of g.

In above derivation we have applied some supplemental results from mathematics (mainly from matrix algebra). First of all trace (tr) of scalar is scalar itself. Secondly trace is linear operator therefore trace of expectation equals to expectation of trace. We have also applied cyclic property of trace. Namely, $\mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{CAB})$ provided matrices $\mathbf{ABC}$ and $\mathbf{CAB}$ exist (are defined as is in above case).

First order Taylor approximation around expectation of $\boldsymbol{X}$ yields to:

$$\begin{aligned} \mathbb{V}\mathrm{ar}[g(\boldsymbol{X})] &\approx \mathbb{V}\mathrm{ar}[g(\overline{\boldsymbol{X}}^*) + (\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T g'(\overline{\boldsymbol{X}}^*)] \\ &= \mathbb{V}\mathrm{ar}[(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)^T g'(\overline{\boldsymbol{X}}^*)] = g'(\overline{\boldsymbol{X}}^*)^T \mathbb{V}\mathrm{ar}[(\boldsymbol{X}-\overline{\boldsymbol{X}}^*)] g'(\overline{\boldsymbol{X}}^*) \\ &= g'(\overline{\boldsymbol{X}}^*)^T \mathbb{V}\mathrm{ar}[\boldsymbol{X}] g'(\overline{\boldsymbol{X}}^*) = g'(\overline{\boldsymbol{X}}^*)^T \boldsymbol{\Sigma}^*_{\boldsymbol{X}} g'(\overline{\boldsymbol{X}}^*). \end{aligned}$$

For more details on Taylor approximation see [92] for example.

A3.1.4 Magnitude of orders

In this thesis we require two (deterministic) magnitude of order definitions which are defined next.

If function $h(\mathsf{n})$ is of order $O\big(g(\mathsf{n})\big)$ then

$$\Big|\lim_{\mathsf{n}\to\infty}\frac{h(\mathsf{n})}{g(\mathsf{n})}\Big| < \infty.$$

If function $h(\mathsf{n})$ is of order $o(g(\mathsf{n}))$ then

$$\lim_{\mathsf{n}\to\infty}\frac{h(\mathsf{n})}{g(\mathsf{n})} = 0.$$

Magnitude of order notations are useful to clean up formulas for example formula $\frac{1}{\mathsf{n}-1}$ (encountered for example in some formulas for second moment estimators) is of order $O(\mathsf{n}^{-1})$. Because

$$\lim_{\mathsf{n}\to\infty}\frac{\frac{1}{\mathsf{n}-1}}{\mathsf{n}^{-1}} = \lim_{\mathsf{n}\to\infty}\frac{\mathsf{n}}{\mathsf{n}-1} = \lim_{\mathsf{n}\to\infty}(1+\frac{1}{\mathsf{n}-1}) = 1.$$

A3.2 Decompositions for second moment

Here we give decompositions for $\hat{\tau}^{comp}$ and τ^{*}.

A3.2.1 Decomposition for $\hat{\tau}^{comp}$

Lemma: Decomposition for $\hat{\tau}^{comp}$.

$$\hat{\tau}^{comp} = \frac{\mathsf{N}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} + \frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2.$$

Proof:

$$\begin{aligned}
\hat{\tau}^{comp} &= \frac{1}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{n}}(Y_j^{comp}-\hat{\mu}^{comp})^2 \\
&= \frac{1}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{N}^{obs}}\Big(Y_j^{comp}-\hat{\mu}^{comp}\Big)^2 + \frac{1}{\mathsf{n}-1}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\Big(Y_j^{comp}-\hat{\mu}^{comp}\Big)^2,
\end{aligned}$$

in which first term can be written as

$$
\begin{aligned}
\frac{1}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{n}-\mathsf{N}^{mis}}\left(Y_j^{comp}-\hat{\mu}^{comp}\right)^2 &= \frac{1}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{N}^{obs}}\left(Y_j^{comp}-\frac{\mathsf{N}^{obs}}{\mathsf{n}}\hat{\mu}^{obs}-\frac{\mathsf{N}^{mis}}{\mathsf{n}}\hat{\mu}^{imp}\right)^2 \\
&= \frac{1}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{N}^{obs}}\left(Y_j^{comp}-\hat{\mu}^{obs}+\frac{\mathsf{N}^{mis}}{\mathsf{n}}(\hat{\mu}^{obs}-\hat{\mu}^{imp})\right)^2 \\
&= \frac{1}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{N}^{obs}}\left(Y_j^{comp}-\hat{\mu}^{obs}\right)^2 \\
&\quad +\frac{2\frac{\mathsf{N}^{mis}}{\mathsf{n}}}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{N}^{obs}}(Y_j^{comp}-\hat{\mu}^{obs})(\hat{\mu}^{obs}-\hat{\mu}^{imp}) \\
&\quad +\frac{(\frac{\mathsf{N}^{mis}}{\mathsf{n}})^2}{\mathsf{n}-1}\sum_{j=1}^{\mathsf{N}^{obs}}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2 \\
&= \frac{\mathsf{N}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+(\frac{\mathsf{N}^{mis}}{\mathsf{n}})^2\frac{\mathsf{N}^{obs}}{\mathsf{n}-1}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2,
\end{aligned}
$$

and second term equals to

$$
\begin{aligned}
\frac{1}{\mathsf{n}-1}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\left(Y_j^{comp}-\hat{\mu}^{comp}\right)^2 &= \frac{1}{\mathsf{n}-1}\sum_{j=\mathsf{n}-\mathsf{N}^{mis}+1}^{\mathsf{n}}\left(Y_j^{comp}-\frac{\mathsf{N}^{obs}}{\mathsf{n}}\hat{\mu}^{obs}-\frac{\mathsf{N}^{mis}}{\mathsf{n}}\hat{\mu}^{imp}\right)^2 \\
&= \frac{1}{\mathsf{n}-1}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\left(Y_j^{comp}-\hat{\mu}^{imp}+\frac{\mathsf{N}^{obs}}{\mathsf{n}}(\hat{\mu}^{imp}-\hat{\mu}^{obs})\right)^2 \\
&= \frac{1}{\mathsf{n}-1}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\left(Y_j^{comp}-\hat{\mu}^{imp}\right)^2 \\
&\quad +\frac{2\frac{\mathsf{N}^{obs}}{\mathsf{n}}}{\mathsf{n}-1}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}(Y_j^{imp}-\tilde{\mu}^{imp})(\hat{\mu}^{imp}-\hat{\mu}^{obs}) \\
&\quad +\frac{\mathsf{N}^{mis}}{\mathsf{n}-1}(\frac{\mathsf{N}^{obs}}{\mathsf{n}})^2(\hat{\mu}^{imp}-\hat{\mu}^{obs})^2 \\
&= \frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}+\frac{\mathsf{N}^{mis}}{\mathsf{n}-1}(\frac{\mathsf{N}^{obs}}{\mathsf{n}})^2(\hat{\mu}^{imp}-\hat{\mu}^{obs})^2.
\end{aligned}
$$

As a consequence sum of first and second term equals to:

$$\begin{aligned}
& \frac{1}{\mathsf{n}-1}\sum_{j=1}^{N^{obs}}\left(Y_j^{comp}-\hat{\mu}^{comp}\right)^2+\frac{1}{\mathsf{n}-1}\sum_{j=N^{obs}+1}^{\mathsf{n}}\left(Y_j^{comp}-\hat{\mu}^{comp}\right)^2 \\
= \; & \frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+(\frac{N^{mis}}{\mathsf{n}})^2\frac{N^{obs}}{\mathsf{n}-1}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2 \\
& +\frac{N^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}+\frac{N^{mis}}{\mathsf{n}-1}(\frac{N^{obs}}{\mathsf{n}})^2(\hat{\mu}^{imp}-\hat{\mu}^{obs})^2 \\
= \; & \frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+\frac{N^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}+(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2\Big((\frac{N^{mis}}{\mathsf{n}})^2\frac{N^{obs}}{\mathsf{n}-1}+\frac{N^{mis}}{\mathsf{n}-1}(\frac{N^{obs}}{\mathsf{n}})^2\Big) \\
= \; & \frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+\frac{N^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} \\
& +(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2\Big(\frac{(N^{mis})^2(\mathsf{n}-N^{mis})}{\mathsf{n}^2(\mathsf{n}-1)}+\frac{N^{mis}(\mathsf{n}^2-2\mathsf{n}N^{mis}+(N^{mis})^2)}{\mathsf{n}^2(\mathsf{n}-1)}\Big) \\
= \; & \frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+\frac{N^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}+(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2\frac{\mathsf{n}^2N^{mis}-\mathsf{n}(N^{mis})^2}{\mathsf{n}^2(\mathsf{n}-1)} \\
= \; & \frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+\frac{N^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}+(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2\frac{\mathsf{n}N^{mis}(\mathsf{n}-N^{mis})}{\mathsf{n}^2(\mathsf{n}-1)} \\
= \; & \frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+\frac{N^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}+\frac{N^{mis}N^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2 \qquad \square
\end{aligned}$$

A3.2.2 Decomposition for τ^*

Lemma: Decomposition for second moment of Y.

$$\tau^*=(1-p^*)\tau^{*obs}+p^*\tau^{*mis}+p^*(1-p^*)(\mu^{*mis}-\mu^{*obs})^2.$$

Proof:

$$\begin{aligned}
\tau^* &= \mathbb{V}\mathrm{ar}[Y] \\
&= \mathbb{E}[\mathbb{V}\mathrm{ar}[Y|R]]+\mathbb{V}\mathrm{ar}[\mathbb{E}[Y|R]] \\
&= \mathbb{E}[\mathbb{V}\mathrm{ar}[Y|R]]+\mathbb{E}[(\mathbb{E}[Y|R]-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[\mathbb{V}\mathrm{ar}[Y|R]]+\mathbb{E}[(\mathbb{E}[Y|R]-\mu^*)^2] \\
&= \Pr(R=1)\mathbb{V}\mathrm{ar}[Y|R=1]+\Pr(R=0)\mathbb{V}\mathrm{ar}[Y|R=0] \\
&\quad +\Pr(R=1)(\mathbb{E}[Y|R=1]-\mu^*)^2+\Pr(R=0)(\mathbb{E}[Y|R=0]-\mu^*)^2 \\
&= \Pr(R=1)\tau^{*obs}+\Pr(R=0)\tau^{*mis} \\
&\quad +\Pr(R=1)(\mu^{*obs}-\mu^*)^2+\Pr(R=0)(\mu^{*mis}-\mu^*)^2 \\
&= (1-p^*)\tau^{*obs}+p^*\tau^{*mis}+(1-p^*)(\mu^{*obs}-\mu^*)^2+p^*(\mu^{*mis}-\mu^*)^2 \\
&= (1-p^*)\tau^{*obs}+p^*\tau^{*mis}+p^*(1-p^*)(\mu^{*mis}-\mu^{*obs})^2 \qquad \square
\end{aligned}$$

A3.3 Limits for expectations of first moment estimator and second moment estimator

Limit for expectation of $\hat{\mu}^{comp}$ is derived using first order Taylor approximation as follows

$$\begin{aligned}
\lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\mu}^{comp}|\mathsf{n}] &= \lim \mathbb{E}[\underbrace{\frac{1}{\mathsf{n}}\Big(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathsf{N}^{mis}\hat{\mu}^{imp}\Big)}_{=A}|\mathsf{n}] \\
&\overset{Taylor}{=} \lim \mathbb{E}\left[\frac{1}{\mathsf{n}}\Big(\mathbb{E}[\mathsf{N}^{obs}|\mathsf{n}]\mathbb{E}[\hat{\mu}^{obs}|\mathsf{n}] + \mathbb{E}[\mathsf{N}^{mis}|\mathsf{n}]\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}]\Big) + R_{\mathsf{n}}\right] \\
&\approx \lim \left\{(1-p^*)\mu_{\mathsf{n}}^{*obs} + p^*\mu_{\mathsf{n}}^{*imp} + O(\mathsf{n}^{-1}) + \mathbb{E}[R_{\mathsf{n}}]\right\} \\
&= (1-p^*)\mu^{*obs} + p^*\mu^{*imp},
\end{aligned}$$

where Taylor approximation has been done for A around expectations of N^{obs}, $\hat{\mu}^{obs}$, N^{mis}, and $\hat{\mu}^{imp}$. Further, order term $O(\mathsf{n}^{-1})$ is due to fact that N^{mis} is ensured via technical assumptions to be such that all computed estimates do exist. To be precise, it is required that $\mathsf{N}^{mis} \in [2, \mathsf{n}-2]$. As a consequence it is assumed that $\mathbb{E}[\mathsf{N}^{mis}/\mathsf{n}, |\mathsf{n}, 2 \leq \mathsf{N}^{mis} \leq \mathsf{n}-2] = p^* + O(\mathsf{n}^{-1})$. We have assumed that limit for expectation of Taylor remainder R_{n} is zero.

Limit for $\hat{\tau}^{comp}$ is computed using first order Taylor approximation as follows

$$\begin{aligned}
\lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\tau}^{comp}|\mathsf{n}] &= \lim \mathbb{E}[\frac{\mathsf{N}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} + \frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2|\mathsf{n}] \\
&\overset{Taylor}{=} \lim \mathbb{E}\Bigg[\frac{\mathbb{E}[\mathsf{N}^{obs}|\mathsf{n}]-1}{\mathsf{n}-1}\mathbb{E}[\hat{\tau}^{obs}|\mathsf{n}] + \frac{\mathbb{E}[\mathsf{N}^{mis}]-1}{\mathsf{n}-1}\mathbb{E}[\hat{\tau}^{imp}|\mathsf{n}] \\
&\quad + \frac{\mathbb{E}[\mathsf{N}^{mis}]\mathbb{E}[\mathsf{N}^{obs}|\mathsf{n}]}{\mathsf{n}(\mathsf{n}-1)}(\mathbb{E}[\hat{\mu}^{obs}|\mathsf{n}] - \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2 + R_{\mathsf{n}}|\mathsf{n}\Bigg] \\
&= (1-p^*)\tau^{*obs} + p^*\tau^{*imp} + p^*(1-p^*)(\mu^{*imp}-\mu^{*obs})^2,
\end{aligned}$$

We have again assumed that limit for expectation of Taylor remainder R_{n} is zero. Note that remainder R_{n} is different than the remainder for the mean estimator.

A3.4 Decompositions

Equation (3.1) is shown to hold under pattern-mixture model for missingness and under selection model for missingness. Further, Equation (3.2) is shown to be a consequence of it. The Bayes theorem (see [63] for details on the theorem) is applied.

A3.4.1 Decomposition of the joint distribution $f(Y, \boldsymbol{X})$

Let Y be variable with missingness, $\boldsymbol{X}$ a fully observed covariate and R response indicator. Next it is shown that Equation (3.1),p.32 holds under pattern-mixture model for missingness and under selection model for missingness.

Pattern-mixture model for missingness

Under pattern-mixture model the joint distribution of $(R, Y, \boldsymbol{X})$ is factored as

$$f(R, Y, \boldsymbol{X}) = f(Y|R, \boldsymbol{X})f(R|\boldsymbol{X})f(\boldsymbol{X}).$$

Now the joint distribution of $Y, \boldsymbol{X}$ is computed as

$$\begin{aligned}
f(Y, \boldsymbol{X}) &= \int f(Y|R = r, X)f(R = r|\boldsymbol{X})f(\boldsymbol{X})dr \\
&\overset{Bayes\ theorem}{=} \int f(Y|R = r, \boldsymbol{X})\frac{f(R = r)f(\boldsymbol{X}|R = r)}{f(\boldsymbol{X})}f(\boldsymbol{X})dr \\
&= \int f(Y|R = r, \boldsymbol{X})f(R = r)f(\boldsymbol{X}|R = r)dr \\
&= \int f(Y, \boldsymbol{X}|R = r)f(R = r)dr \\
&= f(Y, \boldsymbol{X}|R = 1)f(R = 1) + f(Y, \boldsymbol{X}|R = 0)f(R = 0) \\
&= [1 - f(R = 0)]f(Y, \boldsymbol{X}|R = 1) + f(R = 0)f(Y, \boldsymbol{X}|R = 0) \\
&= (1 - p^*)f(Y, \boldsymbol{X}|R = 1) + p^* f(Y, \boldsymbol{X}|R = 0) \\
&= (1 - p^*)f_{Y^{obs}, \boldsymbol{X}^{obs}}(Y, \boldsymbol{X}) + p^* f_{Y^{mis}, \boldsymbol{X}^{mis}}(Y, \boldsymbol{X}).
\end{aligned}$$

Therefore joint distribution $f(Y, \boldsymbol{X})$ is mixture of observed population and missing distributions.

Selection model for missingness

Under selection model the joint distribution of $R, Y, \boldsymbol{X}$ is factored as

$$f(R, Y, \boldsymbol{X}) = f(R|Y, \boldsymbol{X})f(Y|\boldsymbol{X})f(\boldsymbol{X}).$$

Now

$$\begin{aligned}
f(Y, \boldsymbol{X}) &= \int f(R = r|Y, \boldsymbol{X})f(Y|\boldsymbol{X})f(\boldsymbol{X})dr \\
&\overset{Bayes\ theorem}{=} \int \frac{f(R = r)f(Y, \boldsymbol{X}|R = r)}{f(Y, \boldsymbol{X})}f(Y|\boldsymbol{X})f(\boldsymbol{X})dr \\
&= \int f(R = r)f(Y, \boldsymbol{X}|R = r)dr \\
&= (1 - p^*)f_{Y^{obs}, \boldsymbol{X}^{obs}}(Y, \boldsymbol{X}) + p^* f_{Y^{mis}, \boldsymbol{X}^{mis}}(Y, \boldsymbol{X}).
\end{aligned}$$

As with the pattern-mixture model the joint distribution $f(Y, \boldsymbol{X})$ is mixture of two distributions.

A3.4.2 Decomposition of the conditional distribution $f(Y|\boldsymbol{X})$

Here it is shown that if Equation (3.1),p.32 (see Section A3.4.1) holds then Equation (3.2),p.35 is consequence of it.

Let Y be variable with missingness, $\boldsymbol{X}$ a fully observed covariate and R response indicator. Decomposition (Equation 3.1) of the joint distribution of $Y, \boldsymbol{X}$ implies that conditional distribution $Y|\boldsymbol{X}$ may be decomposed as

$$\begin{aligned}
f(Y|\boldsymbol{X}) &= f(Y,\boldsymbol{X})/f(\boldsymbol{X}) \\
&\overset{Equation\ 3.1}{=} \frac{f(Y,\boldsymbol{X}|R=1)f(R=1)}{f(\boldsymbol{X})} + \frac{f(Y,\boldsymbol{X}|R=0)f(R=0)}{f(\boldsymbol{X})} \\
&= f(Y|\boldsymbol{X},R=1)\frac{f(\boldsymbol{X}|R=1)f(R=1)}{f(\boldsymbol{X})} \\
&\quad + f(Y|\boldsymbol{X},R=0)\frac{f(\boldsymbol{X}|R=0)f(R=0)}{f(\boldsymbol{X})} \\
&\overset{Bayes\ theorem}{=} f(Y|\boldsymbol{X},R=1)f(R=1|\boldsymbol{X}) + f(Y|\boldsymbol{X},R=0)f(R=0|\boldsymbol{X}) \\
&= f(Y|\boldsymbol{X},R=1)[1-f(R=0|\boldsymbol{X})] + f(Y|\boldsymbol{X},R=0)f(R=0|\boldsymbol{X}) \\
&= (1-p^*_{\mathbf{x}})f(Y|\boldsymbol{X},R=1) + p^*_{\mathbf{x}}f(Y|\boldsymbol{X},R=0).
\end{aligned}$$

Therefore conditional expectation $g^*(X) = \mathbb{E}[Y|\boldsymbol{X}]$ is

$$\begin{aligned}
g^*(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}] &= \int f(Y=y|\boldsymbol{X})dy \\
&= \int [(1-p^*_{\mathbf{x}})f(Y=y|\boldsymbol{X},R=1) + p^*_{\mathbf{x}}f(Y=y|\boldsymbol{X},R=0)]dy \\
&= (1-p^*_{\mathbf{x}})\int f(Y=y|\boldsymbol{X},R=1)dy + p^*_{\mathbf{x}}\int f(Y=y|\boldsymbol{X},R=0)]dy \\
&= (1-p^*_{\mathbf{x}})\mathbb{E}[Y|\boldsymbol{X},R=1] + p^*_{\mathbf{x}}\mathbb{E}[Y|\boldsymbol{X},R=0] \\
&= (1-p^*_{\mathbf{x}})g^{*obs}(\boldsymbol{X}) + p^*_{\mathbf{x}}g^{*mis}(\boldsymbol{X}).
\end{aligned}$$

Appendix for Chapter 4

In this appendix all justifications of approximations and consequences which were introduced in Chapter 4 are given. At first assumptions which allow comparison of results for different methods and prevent mathematical pathologies are described. Note that these assumptions are used also with non-parametric regression methods and cell imputation methods.

A4.1 Assumptions to prevent mathematical pathologies

In analytical analysis some assumptions are needed to prevent mathematical pathologies. Further, it is required that all results are comparable. The required assumptions are

1: there are at least two complete and incomplete observations in each data set,

2: ordinary least squares estimate of regression coefficients $\boldsymbol{\beta}^{obs}$ exists always, and

3: there are at least two incomplete and complete observations within each cell.

First assumption may have somewhat high impact on moments of N^{mis} and N^{obs} (for small sample sizes). The assumption is actually related to truncation of distribution. As an example, the number of missing data values N^{mis} may follow doubly truncated binomial distribution under assumption 1. Theory for it is available for example in [93].

Second assumption is less strong than the first one because typically probability that ordinary least squares estimate does not exist is zero (covariates are assumed to be continuous).

Finally, third assumption is best ensured by clustering algorithm because then data sets or their classifications to cells need not to be discarded. However, numbers of observations in cells are affected. To fulfill the third assumption observations are clustered and then the clustering is adjusted. The adjustment requires that sample size n is at least four times number of cells n_c.

A4.2 Baseline methods / moments

Theorem 4.1

The bias of $\hat{\mu}^{comp,B}$ for n observations is

$$\mathbb{B}\text{ias}[\hat{\mu}^{comp,B}|\mathsf{n}] = p^*(\mu^{*obs} - \mu^{*mis}).$$

Proof: for mean imputation holds

$$\hat{\mu}^{comp,B,M} = \frac{1}{\mathsf{n}}\big(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathsf{N}^{mis}\hat{\mu}^{imp}\big) = \frac{1}{\mathsf{n}}\big(\mathsf{N}^{obs}\hat{\mu}^{obs} + \sum_{j=1}^{\mathsf{N}^{mis}}\hat{\mu}^{obs}\big) = \hat{\mu}^{obs}.$$

Therefore

$$\begin{aligned}
\mathbb{E}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_3] &= \mu^{obs}\\
\mathbb{E}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_2] &= \mu^{obs}\\
\mathbb{E}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_1] &= \mu^{*obs}.
\end{aligned}$$

Bias result at $\mathcal{Q}_1 = \{\mathsf{n}\}$ follows immediately:

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\mu}^{comp,B,M}|\mathsf{n}] &= \mu^{*obs} - \mu^*\\
&= \mu^{*obs} - ((1-p^*)\mu^{*obs} + p^*\mu^{*mis})\\
&= p^*(\mu^{*obs} - \mu^{*mis}).
\end{aligned}$$

For random imputation strategy mean estimator may be decomposed as follows

$$\hat{\mu}^{comp,B,R} = \hat{\mu}^{comp,B,M} + \frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis},$$

where $\hat{\epsilon}_j^{mis}, j = \mathsf{N}^{obs}+1,\dots,\mathsf{n}$ is j:th imputation noise term. Expectation of imputation noise terms are zero, thus expectations of mean estimator are $\mathbb{E}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_i] = \mathbb{E}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_i],\ i = 1,2,3$. Therefore bias result follows immediately, and it is equal to bias for mean imputation.

As earlier, mean estimator may be written as

$$\hat{\mu}^{comp,B,D} = \hat{\mu}^{comp,B,M} + \frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{mis}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis},$$

where $\hat{\epsilon}_j^{mis}, j = \mathsf{N}^{obs}+1,\dots,\mathsf{n}$ is j:th randomly drawn with replacements and equal draw probabilities empirical residual which equals to observed Y value minus mean of observed Y values. Next expectation of each $\hat{\epsilon}_j^{mis}$ given $\mathcal{Q}_3$ is derived.

Basic sampling theory result is applied: expectation of simple random sample with replacement from population is mean of population (see for example Equation 2.10 in [53]). Here population is observed Y values which have been centered: $\{y_j^{obs} - \mu^{obs}\}_{j=1}^{\mathsf{n}^{obs}}$. Mean of population is zero. Therefore $\mathbb{E}[\hat{\epsilon}_j^{mis}|\mathcal{Q}_3] = 0, j = \mathsf{n}^{obs}+1,\dots,\mathsf{n}$. As a consequence $\mathbb{E}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_i] = \mathbb{E}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_i],\ i = 1,2,3$. Therefore bias result is same as for other baseline strategies. □

Approximation 4.2

The variance $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B}]$ with n observations is approximately

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B}] \approx \tau^{*obs}\left(\underbrace{\frac{1}{\mathsf{n}(1-p^*)}+\frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}}_{\text{due sampling}}+\underbrace{C}_{\text{due imputation strategy}}\right),$$

where term C depends on imputation strategy $\hat{\epsilon}^S$ as follows:

$$C=\begin{cases} 0 & \text{:S=M} \quad \text{(for mean imputation strategy),} \\ \frac{p^*}{\mathsf{n}} & \text{:S=R} \quad \text{(for simulated random imputation), and} \\ \frac{p^*}{\mathsf{n}}\left(1-\frac{1}{\mathsf{n}(1-p^*)}\right) & \text{:S=D} \quad \text{(for random donor).} \end{cases}$$

Justification: as in bias case result for mean strategy is derived first, which is followed by derivations for random and donor strategies.

Mean strategy: for mean strategy imputations are fixed at second and third conditionalization levels, thus

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_3] &= 0 \\ \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_2] &= 0.\end{aligned}$$

At first level mean of observed data values and number of missing data values are random over repetitions of sample drawings. Therefore variance of $\hat{\mu}^{comp}$ is non-zero. To derive the variance one needs to compute $\mathbb{E}[\frac{1}{\mathsf{n}-\mathsf{N}^{mis}}|\mathcal{Q}_1]$. Exact computation of the expectation is impossible because distribution of N^{mis} is not known. However, second order Taylor approximation (see Appendix A3.1.3 for details) can be applied using first two moments of N^{mis}. This yields to (conditionalizer $\mathcal{Q}_1$ is omitted for clarity):

Using Taylor approximation at $\mathbb{E}[\mathsf{N}^{obs}]$ and $\mathsf{N}^{obs}=\mathsf{n}-\mathsf{N}^{mis}$ one gets

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{\mathsf{N}^{obs}}\Big] &\approx \mathbb{E}\Big[\frac{1}{\mathbb{E}[\mathsf{N}^{obs}]}+\frac{1}{1!}\big(\frac{\partial}{\partial \mathsf{N}^{obs}}\frac{1}{\mathsf{N}^{obs}}\big)_{\mathsf{N}^{obs}=\mathbb{E}[\mathsf{N}^{obs}]}(\mathsf{N}^{obs}-\mathbb{E}[\mathsf{N}^{obs}]) \\ &\quad +\frac{1}{2!}\big(\frac{\partial^2}{\partial \mathsf{N}^{obs}\partial \mathsf{N}^{obs}}\frac{1}{\mathsf{N}^{obs}}\big)_{\mathsf{N}^{obs}=\mathbb{E}[\mathsf{N}^{obs}]}(\mathsf{N}^{obs}-\mathbb{E}[\mathsf{N}^{obs}])^2\Big] \\ &= \frac{1}{\mathbb{E}[\mathsf{N}^{obs}]}+\frac{1}{2}\big(2(\mathsf{N}^{obs})^{-3}\big)_{\mathsf{N}^{obs}=\mathbb{E}[\mathsf{N}^{obs}]}\mathbb{V}\mathrm{ar}[\mathsf{N}^{obs}|\mathsf{n}] \\ &\overset{\mathsf{N}^{obs}=\mathsf{n}-\mathsf{N}^{mis}}{\approx} \frac{1}{\mathsf{n}(1-p^*)}+\frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big),\end{aligned}$$

where $\frac{\partial}{\partial \mathsf{N}^{obs}}$ is ordinary derivative (not stochastic), and notation $\big(g(\mathsf{N}^{obs})\big)_{\mathsf{N}^{obs}=\mathbb{E}[\mathsf{N}^{obs}]}$ means that function $g(\mathsf{N}^{obs})$ is evaluated at position $\mathsf{N}^{obs}=$

$\mathbb{E}[\mathsf{N}^{obs}]$. Now variance is derived by applying chain rule of variance (see Appendix A3.1.1 for details) as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathcal{Q}_1] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathcal{Q}_1] \\
&= \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathcal{Q}_1, \mathsf{N}^{mis} = \mathsf{n}^{mis}]|\mathcal{Q}_1] \\
&\quad + \mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\mu}^{obs}|\mathcal{Q}_1, \mathsf{N}^{mis} = \mathsf{n}^{mis}]|\mathcal{Q}_1] \\
&= \mathbb{E}\Big[\tau^{*obs}\frac{1}{\mathsf{n} - \mathsf{N}^{mis}}|\mathcal{Q}_1\Big] + \mathbb{V}\mathrm{ar}\Big[\mu^{*obs}\Big] \\
&= \tau^{*obs}\mathbb{E}\Big[\frac{1}{\mathsf{n} - \mathsf{N}^{mis}}|\mathcal{Q}_1\Big] \\
&\overset{Taylor}{\approx} \tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big).
\end{aligned}$$

Random strategy: variance of the mean estimator for random imputation strategy at $\mathcal{Q}_3$ is

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\Big[\hat{\mu}^{comp,B,R}|\mathcal{Q}_3\Big] &= \mathbb{V}\mathrm{ar}\Big[\hat{\mu}^{comp,B,M} + \frac{1}{\mathsf{n}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{mis}|\mathcal{Q}_3\Big] \\
&= \mathbb{V}\mathrm{ar}\Big[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{mis}|\mathcal{Q}_3\Big] = \frac{1}{\mathsf{n}^2}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \mathbb{V}\mathrm{ar}\Big[\hat{\epsilon}_j^{mis}|\mathcal{Q}_3\Big] \\
&= \frac{1}{\mathsf{n}^2}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \tau^{obs} = \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\tau^{obs}.
\end{aligned}$$

At second conditionalization level randomness of $\mathbf{D}^{test}$ has no impact on variability of $\hat{\mu}^{comp}$ because imputations do not utilize covariate information. Therefore

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_2] = \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\tau^{obs}.$$

Result at first conditionalization level is derived using second order Taylor approximation and chain rule of variance. Required previously computed quantities are $\mathbb{E}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_2]$ and $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_2]$. Variance is computed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_1, \mathsf{N}^{mis} = \mathsf{n}^{mis}] &= \mathbb{V}\mathrm{ar}\Big[\hat{\mu}^{obs}|\mathcal{Q}_1, \mathsf{N}^{mis} = \mathsf{n}^{mis}\Big] \\
&\quad + \mathbb{E}\Big[\frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\hat{\tau}^{obs}|\mathcal{Q}_1, \mathsf{N}^{mis} = \mathsf{n}^{mis}\Big] \\
&= \frac{\tau^{*obs}}{\mathsf{n}-\mathsf{n}^{mis}} + \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\tau^{*obs} \\
&= \tau^{*obs}(\frac{1}{\mathsf{n}-\mathsf{n}^{mis}} + \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}). \\
\mathbb{E}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_1, \mathsf{N}^{mis} = \mathsf{n}^{mis}] &= \mu^{*obs}. \\
\Rightarrow \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,R}|\mathcal{Q}_1] &= \mathbb{V}\mathrm{ar}\Big[\mu^{*obs}|\mathcal{Q}_1\Big] \\
&\quad + \tau^{*obs}\mathbb{E}\Big[\frac{1}{\mathsf{n}-\mathsf{N}^{mis}} + \frac{\mathsf{N}^{mis}}{\mathsf{n}^2}|\mathcal{Q}_1\Big] \\
&= \tau^{*obs}\mathbb{E}\Big[\frac{1}{\mathsf{n}-\mathsf{N}^{mis}} + \frac{\mathsf{N}^{mis}}{\mathsf{n}^2}|\mathcal{Q}_1\Big] \\
&\approx \tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3} + \frac{p^*}{\mathsf{n}}\Big).
\end{aligned}$$

Donor strategy: for deriving variance of mean estimator one notes that

$$\hat{\mu}^{comp,B,D} = \hat{\mu}^{obs} + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\bar{\hat{\epsilon}}^{B,D},$$

where $\bar{\hat{\epsilon}}^{B,D}$ is mean of estimated noise terms. Now

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_3] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs} + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\bar{\hat{\epsilon}}^{B,D}|\mathcal{Q}_3] \\
&= \mathbb{V}\mathrm{ar}[\frac{\mathsf{n}^{mis}}{\mathsf{n}}\bar{\hat{\epsilon}}^{B,D}|\mathcal{Q}_3] = \frac{(\mathsf{n}^{mis})^2}{\mathsf{n}^2}\mathbb{V}\mathrm{ar}[\bar{\hat{\epsilon}}^{B,D}|\mathcal{Q}_3].
\end{aligned}$$

Another basic result from sampling theory to compute variance of mean of estimated noise terms is applied next. Namely, variance of mean estimator for simple random sample with replacement from population equals to population variance divided by sample size (see Equation 2.10 in [53]). Thus one needs to solve variance of our centered population, this is done as follows:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[Y^C|\mathcal{Q}_3] &= \sum_{j=1}^{\mathsf{n}-\mathsf{n}^{mis}} \Big(y_j^{obs} - \mu^{obs} - \mathbb{E}[Y^C]\Big)^2 \Pr(j) \\
&= \sum_{j=1}^{\mathsf{n}-\mathsf{n}^{mis}} \Big(y_j^{obs} - \mu^{obs}\Big)^2 \frac{1}{\mathsf{n}-\mathsf{n}^{mis}} \\
&= \frac{\mathsf{n}-\mathsf{n}^{mis}-1}{\mathsf{n}-\mathsf{n}^{mis}}\tau^{obs} = (1 - \frac{1}{\mathsf{n}-\mathsf{n}^{mis}})\tau^{obs},
\end{aligned}$$

where Y^C denotes centered random variable and $\Pr(j)$ sampling probability for population unit j. Now by applying sampling theory result (Equation 2.10

in [53]) one gets

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_3] &= \frac{(\mathsf{n}^{mis})^2}{\mathsf{n}^2}\mathbb{V}\mathrm{ar}[\bar{\hat{\epsilon}}^{B,D}|\mathcal{Q}_3] \\
&= \frac{(\mathsf{n}^{mis})^2}{\mathsf{n}^2}\frac{1}{\mathsf{n}^{mis}}(1-\frac{1}{\mathsf{n}-\mathsf{n}^{mis}})\tau^{obs} \\
&= \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}(1-\frac{1}{\mathsf{n}-\mathsf{n}^{mis}})\tau^{obs}.
\end{aligned}$$

Randomness of $\mathbf{D}^{test}$ has no impact on variability of $\hat{\mu}^{comp}$ at second level thus

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_2] = \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}(1-\frac{1}{\mathsf{n}-\mathsf{n}^{mis}})\tau^{obs}.$$

At first conditionalization level chain rule of variance and Taylor approximation are applied to derive variance. Required earlier results are $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_2]$ and $\mathbb{E}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_2]$. Variance is derived as:

$$\begin{aligned}
& \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_1] \\
= \quad & \mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_2]|\mathcal{Q}_1] + \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,D}|\mathcal{Q}_2]|\mathcal{Q}_1] \\
= \quad & \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathcal{Q}_1] + \mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}(1-\frac{1}{\mathsf{n}-\mathsf{N}^{mis}})\hat{\tau}^{obs}|\mathcal{Q}_1] \\
\overset{\text{2nd order Taylor}}{\approx} \quad & \tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \big(\mathsf{n}(1-p^*)\big)^{-3}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\Big) \\
& +\mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}(1-\frac{1}{\mathsf{n}-\mathsf{N}^{mis}})\hat{\tau}^{obs}|\mathcal{Q}_1] \\
\overset{\text{1st order Taylor}}{\approx} \quad & \tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \big(\mathsf{n}(1-p^*)\big)^{-3}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\Big) \\
& +\frac{p^*}{\mathsf{n}}(1-\frac{1}{\mathsf{n}(1-p^*)}) \\
= \quad & \tau^{*obs}\Bigg(\Big(\frac{1}{\mathsf{n}(1-p^*)} + \big(\mathsf{n}(1-p^*)\big)^{-3}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\Big) \\
& +\frac{p^*}{\mathsf{n}}(1-\frac{1}{\mathsf{n}(1-p^*)})\Bigg).
\end{aligned}$$

Approximation 4.3

The bias of $\hat{\tau}^{comp,B}$ for n observations is approximately

$$\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,B}|\mathsf{n}] \;\approx\; & p^*(\tau^{*imp}-\tau^{*mis}) - p^*(1-p^*)(\mu^{*mis}-\mu^{*obs})^2 \\
& +C+O(\mathsf{n}^{-1}),
\end{aligned}$$

where imputation variance τ^{*imp} and sampling error C depend on imputation strategy $\hat{\epsilon}^S$. Imputation variance τ^{*imp} and term C depend on imputation strategy $\hat{\epsilon}^S$ as follows:

$$\tau^{*imp} = \left\{ \begin{array}{lll} 0 & :\text{S=M} & \text{(for mean imputation strategy), and} \\ \tau^{*obs} & :\text{S=R and S=D} & \text{(random and donor imputation).} \end{array} \right.$$

and sample error C is

$$C = \begin{cases} 0 & \text{:S=M} \quad \text{(for mean imputation strategy),} \\ \frac{1-p^*}{\mathsf{n}}\tau^{*obs} & \text{:S=R} \quad \text{(for simulated random imputation), and} \\ \frac{\mathsf{n}(1-p^*)-1}{\mathsf{n}^2}\tau^{*obs} & \text{:S=D} \quad \text{(for random donor).} \end{cases}$$

Justification: derivation of results is based on decompositions of second moment given in Appendix A3.2.

Mean: for mean strategy it holds that

$$\begin{aligned} \mathbb{E}[\hat{\tau}^{comp,B,M}|\mathcal{Q}_3] &= \mathbb{E}[\frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{N^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} \\ &\quad + \frac{N^{mis}N^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2|\mathcal{Q}_3] \\ &= \mathbb{E}[\frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{N^{mis}-1}{\mathsf{n}-1}0 + \frac{N^{mis}N^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{obs})^2|\mathcal{Q}_3] \\ &= \mathbb{E}[\frac{N^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}|\mathcal{Q}_3] = \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs} = (1-\frac{\mathsf{n}^{mis}}{\mathsf{n}-1})\tau^{obs}. \end{aligned}$$

Imputation method does not use covariates, thus:

$$\mathbb{E}[\hat{\tau}^{comp,B,M}|\mathcal{Q}_2] = (1-\frac{\mathsf{n}^{mis}}{\mathsf{n}-1})\tau^{obs}.$$

Bias results at $\mathcal{Q}_2$ and $\mathcal{Q}_3$ follow directly. At first level:

$$\begin{aligned} \mathbb{E}[\hat{\tau}^{comp,B,M}|\mathcal{Q}_1] &= \mathbb{E}[(1-\frac{N^{mis}}{\mathsf{n}-1})\hat{\tau}^{obs}|\mathcal{Q}_1] \\ &\approx (1-p^*)\tau^{*obs} + O(\mathsf{n}^{-1}) \\ \Rightarrow \mathbb{B}\text{ias}[\hat{\tau}^{comp,B,M}|\mathcal{Q}_1] &\approx (1-p^*)\tau^{*obs} - \tau^* + O(\mathsf{n}^{-1}) \\ &= (1-p^*)\tau^{*obs} - ((1-p^*)\tau^{*obs} + p^*\tau^{*mis} \\ &\quad + p^*(1-p^*)(\mu^{*mis}-\mu^{*obs})^2) + O(\mathsf{n}^{-1}) \\ &= -p^*\tau^{*mis} - p^*(1-p^*)(\mu^{*mis}-\mu^{*obs})^2 + O(\mathsf{n}^{-1}). \end{aligned}$$

Random: quantity $\hat{\tau}^{imp}$ can be written as

$$\hat{\tau}^{imp} = \frac{1}{N^{mis}-1}\sum_{j=1}^{N^{mis}}\left(\hat{\epsilon}_j^{mis} - \bar{\hat{\epsilon}}^{mis}\right)^2,$$

where $\bar{\hat{\epsilon}}^{mis}$ is the mean of imputation noise terms. At conditionalization level three quantities $\hat{\epsilon}_j^{mis}$ are identically and independently distributed with expectation zero and variance τ^{obs}. Therefore expectation of $\hat{\tau}^{imp}$ is τ^{obs}. Now with $\mathcal{Q}_3 = \{\mathbf{d}^{train}, \mathbf{d}^{test}, g(\mathbf{x}|\boldsymbol{\theta})\}$

$$\begin{aligned} \mathbb{E}[\hat{\tau}^{B,R}|\mathcal{Q}_3] &= \mathbb{E}\Big[\frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2|\mathcal{Q}_3\Big] \\ &= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\tau^{obs} + \frac{\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\tau^{obs}. \end{aligned}$$

Expectation and bias are same in second level as in third level because covariate information is not used. First level expectation $\mathcal{Q}_1 = \{\mathsf{n}\}$ is now

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{B,R}|\mathcal{Q}_1] &= \mathbb{E}_{\mathsf{N}^{mis},\mathbf{D}^{train}|\mathsf{n}}\Big[\frac{\mathsf{N}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{obs} \\
&\quad + \frac{\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\hat{\tau}^{obs}\Big] \\
&\approx \frac{\mathsf{n}-\mathsf{n}p^*-1}{\mathsf{n}-1}\tau^{*obs} + \frac{\mathsf{n}p^*-1}{\mathsf{n}-1}\tau^{*obs} + \frac{\mathsf{n}-\mathsf{n}p^*}{\mathsf{n}(\mathsf{n}-1)}\tau^{*obs} \\
&= (1-p^*)\tau^{*obs} + p^*\tau^{*obs} + \frac{1-p^*}{\mathsf{n}}\tau^{*obs} + O(\mathsf{n}^{-1}) \\
\Rightarrow \mathbb{B}\text{ias}[\hat{\tau}^{comp,B,R}|\mathcal{Q}_1] &\approx p^*(\tau^{*obs} - \tau^{*mis}) - p^*(1-p^*)(\mu^{*mis} - \mu^{*obs})^2 \\
&\quad + \frac{1-p^*}{\mathsf{n}}\tau^{*obs} + O(\mathsf{n}^{-1}).
\end{aligned}$$

Donor: for deriving the bias for donor strategy one applies sampling theory results. Namely, basic sampling theory result states that expectation of sample variance, which is $\hat{\tau}^{imp}$ here, equals to population variance (for simple random sampling with replacement, see for example page 17 of [53]). Our population consists of centered Y values $\{y_j^{obs} - \mu^{obs}\}_{j=1}^{\mathsf{n}^{obs}}$. Variance of population is $(1 - \frac{1}{\mathsf{n}^{obs}})\tau^{obs}$. Thus:

$$\mathbb{E}[\hat{\tau}^{imp}|\mathcal{Q}_3] = (1 - \frac{1}{\mathsf{n}^{obs}})\tau^{obs}.$$

One needs also to solve

$$\begin{aligned}
\mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp})^2|\mathcal{Q}_3] &= \mathbb{E}\left[\Big(\hat{\mu}^{obs} - (\hat{\mu}^{obs} + \frac{1}{\mathsf{n}^{mis}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{B,D}\Big)^2|\mathcal{Q}_3\right] \\
&= \mathbb{E}\left[\Big(-\frac{1}{\mathsf{n}^{mis}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{B,D}\Big)^2|\mathcal{Q}_3\right] \\
&= \mathbb{E}\left[\Big(\frac{1}{\mathsf{n}^{mis}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{B,D}\Big)^2|\mathcal{Q}_3\right].
\end{aligned}$$

Now $\mathbb{E}[(\frac{1}{\mathsf{n}^{mis}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{B,D})^2|\mathcal{Q}_3] = \mathbb{V}\text{ar}[\hat{\mu}^{imp}|\mathcal{Q}_3]$. Basic sampling theory result states that variance of sample mean equals to population variance divided by sample size (for simple random sampling with replacement, see Equation 2.10 in [53]). Therefore,

$$\mathbb{V}\text{ar}[\hat{\mu}^{imp}|\mathcal{Q}_3] = \frac{1}{\mathsf{n}^{mis}}(1 - \frac{1}{\mathsf{n}^{obs}})\tau^{obs}.$$

As a consequence

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,B,D}|\mathcal{Q}_3] &= \mathbb{E}[\frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs}+\frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} \\
&\quad +\frac{\mathsf{n}^{mis}(\mathsf{n}^{obs})}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2|\mathcal{Q}_3] \\
&= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs}+\frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}(1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs} \\
&\quad +\frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\frac{1}{\mathsf{n}^{mis}}(1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs} \\
&= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs}+\frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}(1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs} \\
&\quad +\frac{\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs} \\
&= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs}+\frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}(1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs}+\frac{\mathsf{n}^{obs}-1}{\mathsf{n}(\mathsf{n}-1)}\tau^{obs}.
\end{aligned}$$

Expectation at second conditionalization level is equal to first level result, thus

$$\mathbb{E}[\hat{\tau}^{comp,B,D}|\mathcal{Q}_2] = \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs}+\frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}(1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs}+\frac{\mathsf{n}^{obs}-1}{\mathsf{n}(\mathsf{n}-1)}\tau^{obs}.$$

Approximate expectation at first level is computed using result for second level as

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,B,D}|\mathcal{Q}_1] &\approx (1-p^*)\tau^{*obs}+p^*\tau^{*obs}+(\frac{1-p^*}{\mathsf{n}}-\frac{1}{\mathsf{n}^2})\tau^{*obs} \\
&\quad +O(\mathsf{n}^{-1}) \\
\Rightarrow \mathbb{B}\text{ias}[\hat{\tau}^{comp,B,D}|\mathcal{Q}_1] &\approx p^*(\tau^{*obs}-\tau^{*mis})-p^*(1-p^*)(\mu^{*mis}-\mu^{*obs})^2 \\
&\quad +(\frac{1-p^*}{\mathsf{n}}-\frac{1}{\mathsf{n}^2})\tau^{*obs}+O(\mathsf{n}^{-1}).
\end{aligned}$$

Consequence 4.4

Asymptotically one has (approximately) the following

$$\begin{aligned}
\lim_{\mathsf{n}\to\infty}\mathbb{B}\text{ias}[\hat{\mu}^{comp,B}|\mathsf{n}] &= p^*(\mu^{*obs}-\mu^{*mis}) \\
\lim_{\mathsf{n}\to\infty}\mathbb{V}\text{ar}[\hat{\mu}^{comp,B}|\mathsf{n}] &\approx 0 \\
\lim_{\mathsf{n}\to\infty}\mathbb{B}\text{ias}[\hat{\tau}^{comp,B}|\mathsf{n}] &\approx p^*\Big[(\tau^{*imp}-\tau^{*obs})-(1-p^*)(\mu^{*mis}-\mu^{*obs})^2\Big],
\end{aligned}$$

where

$$\tau^{*imp}=\begin{cases} 0 & \text{:S=M} \quad \text{(for mean imputation strategy) and} \\ \tau^{*obs} & \text{:S=R and S=D} \quad \text{(random and donor imputation).}\end{cases}$$

Justification: Results are trivial to derive as they follow immediately by taking limits of results in theorem 4.1, approximation 4.2, and approximation 4.3.

A4.3 Baseline methods / unit level

Approximation 4.9

Expectation of $\hat{mse}(Y^{comp,B})$ with n observations is approximately

$$\begin{aligned}\mathbb{E}[\hat{mse}(Y^{comp,B})|\mathsf{n}] \;\approx\;& \underbrace{(\mu^{*obs} - \mu^{*mis})}_{\text{global bias}}{}^2 + \underbrace{\mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\ & + \underbrace{\tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big)}_{\text{expected sampling variance}} \\ & + \underbrace{C\tau^{*obs}}_{\text{expected imputation variance}} + \underbrace{v^{*mis}}_{\text{expected target variance}},\end{aligned}$$

where constant C depends on imputation strategy S:

$$C = \begin{cases} 0 & \text{:S=M} \quad \text{(for mean imputation strategy),} \\ 1 & \text{:S=R} \quad \text{(for simulated random imputation), and} \\ 1 - \frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3} & \text{:S=D} \quad \text{(for random donor).} \end{cases}$$

Justification: recall decomposition of mean squared error given in Equation 3.12 (Chapter 3). First term and first cross term in the equation are zero because $\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] = \mu^{*imp}_{\mathsf{n}} = \mu^{*obs}$. Last cross term is zero because $\hat{\mu}^{obs}$ and $\hat{\epsilon}_{\mathbf{x}^{mis}}$ are conditionally independent given training data and expectation of noise is zero. As a consequence

$$\begin{aligned}\mathbb{E}[\hat{mse}(Y^{comp,B})|\mathsf{n}] =& \underbrace{(\mu^{*obs} - \mu^{*mis})}_{\text{global bias}}{}^2 + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\ + & \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]\Bigg]}_{\text{expected variance of conditional mean estimate}} + \underbrace{v^{*imp}_{\mathsf{n}}}_{\text{expected imputation noise}} \\ + & \underbrace{v^{*mis}}_{\text{expected target noise}}.\end{aligned}$$

Expected variance of conditional mean estimate is computed as

$$\mathbb{E}_{\mathbf{x}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]\Bigg] \;=\; \frac{\tau^{*obs}}{\mathsf{n}^{obs}}.$$

Integration over distribution of response pattern N^{mis} is a bit complicated. Approximate result can be derived using second order Taylor approximation:

$$\begin{aligned}\mathbb{E}_{\mathsf{N}^{mis},\mathbf{x}^{mis}|\mathsf{n}}\Big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}\Big] \;&=\; \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Big[\frac{\tau^{*obs}}{\mathsf{n}^{obs}}\Big] \\ &\approx\; \tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\big(\mathsf{n}(1-p^*)\big)^3}\Big).\end{aligned}$$

Next variances for each strategy are computed.

Mean: because noise is not modelled it holds that $v^{*imp}_{\mathsf{n}} = 0$.

Random: for random strategy

$$\begin{aligned} v^{*imp}_{\mathsf{n}} &= \mathbb{E}_{\boldsymbol{X}^{mis},\boldsymbol{\mathsf{N}}^{mis}|\mathsf{n}} \mathbb{V}\mathrm{ar}[\hat{\epsilon}^{B,R}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \\ &= \tau^{*obs}. \end{aligned}$$

Donor: for donor strategy it holds that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}} \mathbb{V}\mathrm{ar}[\hat{\epsilon}^{B,D}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] &\approx \mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}} \Big[(1 - \frac{1}{\mathsf{n}^{obs}})\tau^{*obs}\Big] \\ &= (1 - \frac{1}{\mathsf{n}^{obs}})\tau^{*obs} \\ \Rightarrow \mathbb{E}_{\boldsymbol{\mathsf{N}}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}} \mathbb{V}\mathrm{ar}[\hat{\epsilon}^{B,D}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] &\approx \tau^{*obs}(1 - \frac{1}{\mathsf{n}(1-p^*)} - \frac{\mathbb{V}\mathrm{ar}[\boldsymbol{\mathsf{N}}^{mis}]}{\big(\mathsf{n}(1-p^*)\big)^3}). \end{aligned}$$

Approximation 4.10

Mean squared error $\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n})$ can be approximated as:

$$\begin{aligned} \mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}) \approx\ & \underbrace{\big(\mu^{*obs} - g^{*mis}(\mathbf{x}^{mis})\big)^2}_{\text{squared bias}} + \underbrace{\frac{1}{\mathsf{n}^{obs}}\tau^{*obs}}_{\text{sampling variance}} \\ & + \underbrace{C\tau^{*obs}}_{\text{imputation variance}} + \underbrace{v^{*mis}(\mathbf{x}^{mis})}_{\text{target variance}}. \end{aligned}$$

where term C depends on imputation strategy S:

$$C = \begin{cases} 0 & \text{:S=M} \quad \text{(for mean imputation strategy),} \\ 1 & \text{:S=R} \quad \text{(for simulated random imputation), and} \\ 1 - \frac{1}{\mathsf{n}^{obs}} & \text{:S=D} \quad \text{(for random donor).} \end{cases}$$

Justification: recall from Chapter 3 that

$$\begin{aligned} \mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}) =\ & \Big(\underbrace{\mathbb{E}[\hat{g}(\mathbf{x}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - g^{*mis}(\mathbf{x}^{mis})}_{\text{imputation bias at } \mathrm{x}^{\mathrm{mis}}}\Big)^2 \\ & + \underbrace{\mathbb{V}\mathrm{ar}[Y^{imp}_{|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}}]}_{\text{imputation variance at } \mathrm{x}^{\mathrm{mis}}} \\ & + \underbrace{\mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}^{mis}}]}_{\mathrm{v}^{*\mathrm{mis}}(\mathbf{x}^{\mathrm{mis}}),\ \text{target noise at } \mathrm{x}^{\mathrm{mis}}}. \end{aligned}$$

Squared bias terms in $\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n})$ are same for all baseline methods because expected predictions are same. Squared bias term is

$$\Big(\mathbb{E}[Y^{imp,B}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - \mathbb{E}[Y^{mis}|\mathbf{x}^{mis}]\Big)^2 = \Big(\mu^{*obs} - g^{*mis}(\mathbf{x}^{mis})\Big)^2.$$

Prediction variances are derived as follows

Mean:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[Y^{imp,B,M}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\ &= \frac{\tau^{*obs}}{\mathsf{n}^{obs}}.\end{aligned}$$

Random: derivation of variance for random imputation strategies is a bit more complicated. By writing $Y^{imp,B,R} = Y^{imp,B,M} + \hat{\epsilon}^{B,R}$ one gets following result for simulated random strategy:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[Y^{imp,B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{V}\mathrm{ar}[Y^{imp,B,M}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\ &\quad +\mathbb{V}\mathrm{ar}[\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\ &\quad +2\mathbb{C}\mathrm{ov}[Y^{imp,B,M},\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}],\end{aligned}$$

where first term is already computed. Second and third terms are easy to derive by applying chain rules of variance and covariance (suitable conditionalizer is training data):

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{V}\mathrm{ar}\Big[\mathbb{E}[\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n},\mathbf{d}^{train}]\Big] \\ &\quad +\mathbb{E}\Big[\mathbb{V}\mathrm{ar}[\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n},\mathbf{D}^{train}]\Big] \\ &= \mathbb{V}\mathrm{ar}_{\mathbf{D}^{tr}|\mathsf{n}^{mis},\mathsf{n}}\Big[0\Big] + \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\Big[\hat{\tau}^{obs}\Big] \\ &= \tau^{*obs}.\end{aligned}$$

where outer integrations in first two rows are with respect to distribution $\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}$. For clarity of formulas, outer conditionalizers have been omitted in first two rows (they are same as in inner integrations except training data is excluded).

Derivation of covariance term is easy after noticing that $Y^{imp,B,M}$ and $\hat{\epsilon}^{B,R}$ are conditionally independent given training data $\mathbf{d}^{train}$. By applying chain rule of covariance one gets:

$$\begin{aligned}&\mathbb{C}\mathrm{ov}[Y^{imp,B,M},\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\ =\;& \mathbb{E}\Bigg[\mathbb{C}\mathrm{ov}[Y^{imp,B,M},\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n},\mathbf{d}^{train}\Bigg] \\ &+\mathbb{C}\mathrm{ov}\Bigg[\mathbb{E}[\hat{\mu}^{obs}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n},\mathbf{D}^{tr}],\mathbb{E}[\hat{\epsilon}^{B,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n},\mathbf{d}^{train}]\Bigg] \\ =\;& \mathbb{E}[0] + \mathbb{C}\mathrm{ov}\Bigg[\mathbb{E}[\hat{\mu}^{obs}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n},\mathbf{d}^{train}],0\Bigg] = 0.\end{aligned}$$

Again outer integrations in first two rows are with respect to distribution $\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}$. Outer conditionalizers have been omitted, for clarity, as in previous computation of variance term.

Donor: finally, one derives variance term for random donor method. Expectation of noise term given training data is zero. Variance result is derived using basic result from sampling theory. Link to sampling theory is following: provided training data is given then finite population is set of observed Y values, imputation of single Y value is equal to drawing a simple random sample, with replacement, of size one from the finite population. One can apply here classical result for variance of mean estimate which is computed from one observation. Result is available in multiple good sampling theory sources such as [53, 113, 4]. Here reference to Knottnerus [53] is made. Knottnerus' Equation (2.10) [53] states that the variance of mean estimate, in case of simple random sampling, equals to finite population variance divided by sample size. In our case variance of finite population is (see Section 2.1 in [53] for details) $\frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}(y_j-\mu^{obs})^2 = \frac{\mathsf{n}^{obs}-1}{\mathsf{n}^{obs}}\tau^{obs} = (1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs}$. This is also variance of individual random donor imputation because corresponding sample size is one, thus:

$$\mathbb{V}\text{ar}[\hat{\epsilon}^{B,D}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}, \mathbf{d}^{train}] \;=\; (1-\frac{1}{\mathsf{n}^{obs}})\tau^{obs}.$$

Carrying computation as in random simulated strategy (R) one gets following

$$\begin{aligned}\mathbb{V}\text{ar}[\hat{\epsilon}^{B,D}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] &= \mathbb{V}\text{ar}\Big[\mathbb{E}[\hat{\epsilon}^{B,D}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}, \mathbf{d}^{train}]\Big] \\ &\quad + \mathbb{E}\Big[\mathbb{V}\text{ar}[\hat{\epsilon}^{B,D}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}, \mathbf{d}^{train}]\Big] \\ &= \mathbb{V}\text{ar}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\Big[0\Big] + \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\Big[(1-\frac{1}{\mathsf{n}^{obs}})\hat{\tau}^{obs}\Big] \\ &= (1-\frac{1}{\mathsf{n}^{obs}})\tau^{*obs}.\end{aligned}$$

Consequence 4.11

Limit of expectation of $\hat{mse}(Y^{comp,B})$ is approximately

$$\begin{aligned}\lim_{\mathsf{n}\to\infty}\mathbb{E}[\hat{mse}(Y^{comp,B})|\mathsf{n}] \;\approx\; & \underbrace{(\mu^{*obs}-\mu^{*mis})^2}_{\text{global bias}} + \underbrace{\mathbb{V}\text{ar}_{\boldsymbol{X}^{mis}}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\ & + \underbrace{C\tau^{*obs}}_{\text{expected imputation variance}} + \underbrace{v^{*mis}}_{\text{expected target variance}},\end{aligned}$$

where constant C depends on imputation strategy S:

$$C = \begin{cases} 0 & \text{:S=M} \quad \text{(for mean imputation strategy), and} \\ 1 & \text{:S=R,S=D} \quad \text{(for random strategies).}\end{cases}$$

Justification: result follows directly by taking limit of approximation 4.9:

$$\begin{aligned}
\lim_{\mathsf{n}\to\infty} \mathbb{E}[m\hat{s}e(Y^{comp,B})|\mathsf{n}] &\approx (\mu^{*obs} - \mu^{*mis})^2 + \mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}}[g^{*mis}(\boldsymbol{X}^{mis})] \\
&\quad + \lim_{\mathsf{n}\to\infty} \tau^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big) \\
&\quad + \lim_{\mathsf{n}\to\infty} C'\tau^{*obs} \qquad + v^{*mis} \\
&= (\mu^{*obs} - \mu^{*mis})^2 + \mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}}[g^{*mis}(\boldsymbol{X}^{mis})] \\
&\quad + C\tau^{*obs} \qquad + v^{*mis},
\end{aligned}$$

where C' is C from approximation 4.9.

A4.4 Baseline methods / importance of higher order term (example)

Approximation 4.2 for $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathsf{n}]$ in Chapter 4 contains second order Taylor approximation term $\tau^{*obs}\frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}$. The second order Taylor term is shown to be important next. Consider example $\tau^* = 1$, MCAR missingness, where number of missing data values is roughly $\mathsf{N}^{mis} \sim Bin(\mathsf{n}, 0.5)$ (50% missingness). Let first order Taylor approximation be $A = \tau^{*obs}\frac{1}{\mathsf{n}(1-p^*)}$, B be the second order approximation term, and $V = \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B,M}|\mathsf{n}]$. Importance of second order Taylor approximation term is evaluated by computing relative error ratios $Err1 = \mathbb{E}[\frac{V-A}{V}|\mathsf{n}]$ and $Err2 = \mathbb{E}[\frac{V-(A+B)}{V}|\mathsf{n}]$. Sign of $Err1$ and $Err2$ tells to which direction approximation needs to be corrected.

Simulations of 5000 repetitions is done 50 times for each sample size $\mathsf{n} \in \{5, 9, 21, 37, 69, 101\}$ to compute expectation and deviation of error estimates. Results are (deviations of estimates are shown in paranthesis):

n	5	9	21	37	69	101
Err1	0.0419(0.0025)	0.0996(0.0030)	0.0501(0.0029)	0.0291(0.0024)	0.0175(0.0030)	0.0067(0.0025)
Err2	-0.1497(0.0030)	-0.0005(0.0034)	0.0049(0.0031)	0.0028(0.0024)	0.0032(0.0030)	-0.0031(0.0025)

From sample size 9 to 69 second order approximation is considerable better than first order approximation. Therefore second order Taylor term is important.

For smallest sample sizes first order approximation underestimates variance. Second order approximation overestimates variance for sample size 5, and absolute error is higher than for first order approximation. However, typically it is better to overestimate variance than underestimate (to not to get too narrow confidence intervals).

A4.5 Baseline methods / computations for simulation example

Lemma A4.5.1:
Let a and $c > 0$ be real numbers, then

i)

$$\begin{aligned}\int_{-\infty}^{\infty} x^3 * \exp(-(x-a)^2/c)dx &= \frac{1}{2}a\sqrt{c}(2a^2+3c)\sqrt{\pi} \\ &= K(a,c).\end{aligned}$$

ii)

$$\begin{aligned}\int_{-\infty}^{\infty} x^6 * \exp(-(x-a)^2/c)dx &= \frac{1}{8}\sqrt{c}(8a^6+60a^4c+90a^2c^2+15c^3)\sqrt{\pi} \\ &= K^2(a,c).\end{aligned}$$

Lemma A4.5.1 was derived using symbolical integration in Mathematica software. Computer softwares are known to have bugs (programming errors). Therefore numerical integration was used to empirically ”verify” validity of formulas (for set of values a and c required in this study).

Corollary A4.5.2:
The first two moments of Y^{obs} and Y^{mis} are

$$\begin{aligned}\mathbb{E}[Y^{obs}] &= -0.196 \\ \mathbb{V}\text{ar}[Y^{obs}] &= 0.39074 \\ \mathbb{E}[Y^{mis}] &= 0.196 \\ \mathbb{V}\text{ar}[Y^{mis}] &= 0.39074\end{aligned}$$

Proof: First moment of Y^{obs} is computed as

$$\begin{aligned}
\mathbb{E}[Y^{obs}] &= \mathbb{E}_{X^{obs}}\left[\mathbb{E}[Y^{obs}|X^{obs}=x]\right] = \mathbb{E}_{X^{obs}}\left[g(x)\right] \\
&= \mathbb{E}_{X^{obs}}\left[\frac{1}{500}x^3\right]\frac{1}{500}\int_{-\infty}^{\infty} x^3 f_{X^{obs}}(x)dx \\
&= \frac{1}{500}\int_{-\infty}^{\infty} x^3 \frac{1}{\sqrt{2\pi * 15}}\exp\Bigg(-\Big(\big(x-(-2)\big)^2/(2*15)\Big)\Bigg)dx \\
&= \frac{1}{500\sqrt{2\pi * 15}}\int_{-\infty}^{\infty} x^3 \exp\Bigg(-\Big(\big(x-(-2)\big)^2/(2*15)\Big)\Bigg)dx \\
&= \frac{1}{500\sqrt{2\pi * 15}}K(-2, 2*15) \\
&\overset{Lemma\ A4.5.1\ i)}{=} \frac{1}{500\sqrt{2\pi * 15}}\frac{1}{2}(-2)\sqrt{30}(2(-2)^2+3*30)\sqrt{\pi} \\
&= -\frac{1}{500\sqrt{30}}\sqrt{30}(2*4+3*30) = -\frac{1}{500}(8+90) \\
&= -\frac{98}{500} = -0.196.
\end{aligned}$$

Now because $K(a,c) = -K(-a,c)$ one gets that $\mathbb{E}[Y^{mis}] = -(-0,196) = 0.196$.

Remember that

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[Y^{obs}] &= \mathbb{V}\mathrm{ar}_{X^{obs}}[\mathbb{E}[Y|X^{obs}]] + \mathbb{E}_{X^{obs}}[\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs}]] \\
&= \mathbb{V}\mathrm{ar}_{X^{obs}}[g(X^{obs})] + 0.05 \\
&= \mathbb{E}_{X^{obs}}[(g(X^{obs}))^2] - \mathbb{E}_{X^{obs}}[g(X^{obs})]\mathbb{E}_{X^{obs}}[g(X^{obs})] + 0.15 \\
&= \mathbb{E}_{X^{obs}}\left[\frac{1}{500^2}(X^{obs})^6\right] - \mathbb{E}[Y^{obs}]\mathbb{E}[Y^{obs}] + 0.15.
\end{aligned}$$

Now one just needs to compute the first term which is done as

$$\begin{aligned}
& \mathbb{E}_{X^{obs}}\left[\frac{1}{500^2}(X^{obs})^6\right] \\
= \quad & \frac{1}{500^2}\int_{-\infty}^{\infty} x^6 f_{X^{obs}}(x)dx \\
= \quad & \frac{1}{500^2}\int_{-\infty}^{\infty} x^6 \frac{1}{\sqrt{2\pi * 15}} \exp\Bigg(-\Big(\big(x-(-2)\big)^2/(2*15)\Big)\Bigg)dx \\
= \quad & \frac{1}{500^2\sqrt{2\pi * 15}}\int_{-\infty}^{\infty} x^6 \exp\Bigg(-\Big(\big(x-(-2)\big)^2/(2*15)\Big)\Bigg)dx \\
= \quad & \frac{1}{500^2\sqrt{2\pi * 15}}K^2(-2, 2*15) \\
\overset{*}{=} \quad & \frac{1}{500^2\sqrt{2\pi * 15}}\frac{1}{8}\sqrt{30}(8(-2)^6 + 60(-2)^4 * 30 + 90(-2)^2 * 30^2 + 15 * 30^3)\sqrt{\pi} \\
= \quad & \frac{1}{8 * 500^2\sqrt{2\pi * 15}}\sqrt{30}(8*64 + 60*16*30 + 90*4*900 + 15*27000)\sqrt{\pi} \\
= \quad & \frac{1}{8 * 500^2\sqrt{30}}\sqrt{30}(8*64 + 60*16*30 + 90*4*900 + 15*27000) \\
= \quad & \frac{1}{2000000}758312 \\
= \quad & 0.379156,
\end{aligned}$$

where at * lemma A4.5.1 ii) was applied.

Now one can compute variance of Y^{obs} as

$$\mathbb{V}\text{ar}[Y^{obs}] = 0.379156 - (-0.196)^2 + 0.15 = 0.49074 \approx 0.49.$$

Now because $K^2(a, c) = K^2(-a, c)$ and residual variances are same for missing and observed Y values it immediately follows that $\mathbb{V}\text{ar}[Y^{mis}] = \mathbb{V}\text{ar}[Y^{obs}]$. □

A4.6 Linear regression / moments

In this section justifications for results of moments for linear regression methods are given.

Approximation 4.12

The approximation for the bias of $\hat{\mu}^{comp,L}$ for n observations is

$$\mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,L}|\mathsf{n}] \approx p^*(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}] - \mu^{*mis})$$

where $\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}]$ is the expected regression coefficients over training data and $\overline{\boldsymbol{X}}^{*mis} = \mathbb{E}[\boldsymbol{X}^{mis}]$ is the expected covariate vector over missing data.

Justification: result is derived as follows:

$$\begin{aligned}
\mathbb{E}[\hat{\mu}^{comp,L,M}|\mathsf{n}, \mathsf{N}^{mis} = \mathsf{n}^{mis}] &= \mathbb{E}\left[\frac{1}{\mathsf{n}}\Big(\mathsf{n}^{obs}\hat{\mu}^{obs} + \mathsf{n}^{mis}(\hat{\beta}^{obs}_0 + (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis})\Big)|\mathsf{n}, \mathsf{n}^{mis}\right] \\
&= \frac{1}{\mathsf{n}}\Bigg(\mathsf{n}^{obs}\mu^{*obs} + \mathsf{n}^{mis}(\mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}, \mathsf{n}^{mis}] \\
&\quad + \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}, \mathsf{N}^{mis} = \mathsf{n}^{mis}]^T\overline{\boldsymbol{X}}^{*mis})\Bigg) \\
\Rightarrow \mathbb{E}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_1] &\approx \frac{1}{\mathsf{n}}\Bigg(\mathbb{E}[\mathsf{N}^{obs}|\mathsf{n}]\mu^{*obs} \\
&\quad + \mathbb{E}[\mathsf{N}^{mis}|\mathsf{n}](\mathbb{E}[\hat{\beta}_0|\mathsf{n}] + \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis})\Bigg) \\
&= (1-p^*)\mu^{*obs} + p^*(\mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}] + \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}) \\
\Rightarrow \mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_1] &\approx (1-p^*)\mu^{*obs} + p^*(\mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}] + \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}) \\
&\quad -\Big((1-p^*)\mu^{*obs} + p^*\mu^{*mis}\Big) \\
&= p^*(\mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}] + \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis} - \mu^{*mis}),
\end{aligned}$$

Expectation and bias for random imputation strategy are same as for mean imputation because expectations of modelled noise terms are zero.

Approximation 4.13

The variance of first moment $\hat{\mu}^{comp,L}$ given $\mathcal{Q}_2$ is approximately

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathcal{Q}_2] \approx \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Big((\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0} + C\Big)$$

where $_{-0}$ subscript means that all other regression coefficients are picked except intercept term, and

$$C = \begin{cases} 0 & \text{:S=M} \quad \text{(mean imputation), and} \\ v^{obs,L,R} & \text{:S=R} \quad \text{(simulated random imputation).} \end{cases}$$

Justification: approximation 4.13 is straightforward to derive:

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2] &= \mathbb{V}\mathrm{ar}\left[\frac{1}{\mathsf{n}}\left(\mathsf{n}^{obs}\hat{\mu}^{obs} + \mathbf{1}^T_{\mathsf{n}^{mis}}\mathbf{D}^{mis}_{\mathbb{X}}\hat{\boldsymbol{\beta}}^{obs}\right)|\mathcal{Q}_2\right] \\ &= \mathbb{V}\mathrm{ar}\left[\frac{1}{\mathsf{n}}\mathbf{1}^T_{\mathsf{n}^{mis}}\mathbf{D}^{mis}_{\mathbb{X}}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2\right] = \frac{1}{\mathsf{n}^2}\mathbf{1}^T_{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}|\mathcal{Q}_2]\boldsymbol{\beta}^{obs}_{-0} \\ &= \frac{1}{\mathsf{n}^2}\mathbf{1}^T_{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0} = \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}(\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0}. \end{aligned}$$

Variance increases due to random imputation strategy. This increase is computed by first noticing that

$$\hat{\mu}^{comp,L,R} = \hat{\mu}^{comp,L,M} + \frac{1}{\mathsf{n}}\sum_{\hat{\epsilon}^{L,R}_j \in D^{imp}_{\hat{\epsilon},\mathsf{n}^{mis}}} \hat{\epsilon}^{L,R}_j.$$

Terms in decomposition for $\hat{\mu}^{comp,L,R}$ are conditionally independent given $\mathcal{Q}_2$, thus variance for random imputation strategy computed as:

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2] + \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{\hat{\epsilon}^{L,R}_j \in D^{imp}_{\hat{\epsilon},\mathsf{n}^{mis}}} \hat{\epsilon}^{L,R}_j|\mathcal{Q}_2] \\ &\quad +2\mathbb{C}\mathrm{ov}[\hat{\mu}^{comp,L,M}, \frac{1}{\mathsf{n}}\sum_{\hat{\epsilon}^{L,R}_j \in D^{imp}_{\hat{\epsilon},\mathsf{n}^{mis}}} \hat{\epsilon}^{L,R}_j|\mathcal{Q}_2] \\ &= \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}(\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0} + \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}v^{obs,L,R} + 0 \\ &= \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}((\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0} + v^{obs,L,R}). \end{aligned}$$

Approximation 4.14

When variance of $Y^{obs}|\mathbf{x}^{obs}$ is constant v^{*obs} (homoscedastic situation) an approximation for the variance of first moment $\hat{\mu}^{comp,L}$ given n observations is

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathsf{n}] &= \mathbb{E}_{\mathsf{N}^{mis},\mathbf{D}^{train}_{\mathsf{N}^{obs}},\hat{\boldsymbol{\beta}}^{obs}}\left[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L}|\mathcal{Q}_2]\right] \\ &\quad +\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\mathbf{D}^{train}_{\mathsf{N}^{obs}},\hat{\boldsymbol{\beta}}^{obs}}\left[\mathbb{E}[\hat{\mu}^{comp,L}|\mathcal{Q}_2]\right] \\ &\approx T_1 + T_2 + \underbrace{\qquad C \qquad}_{\text{variance due to noise modelling}} + \underbrace{O\left(\mathsf{n}^{-1}\right)}_{\text{approximation error}}, \end{aligned}$$

where $T_1 = \mathbb{E}_{\mathsf{N}^{mis}, \mathbf{D}^{train}_{\mathsf{N}^{obs}}, \hat{\boldsymbol{\beta}}^{obs}}[\mathbb{V}\text{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2]]$ thus

$$T_1 \;=\; \underbrace{\frac{v^{*obs}}{\mathsf{n}^2}\frac{p^*}{1-p^*}\text{tr}\bigg(\big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\big)^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\bigg)}_{\text{variance due to estimated coefficients}}$$
$$+\underbrace{\frac{p^*}{\mathsf{n}}\mathbb{V}\text{ar}\Big[\mathbb{E}_{\mathbf{D}^{train}_{\mathsf{N}^{obs}}, \hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\boldsymbol{X}^{mis}\Big]}_{\text{variability of expected model}}$$

and $T_2 = \mathbb{V}\text{ar}_{\mathsf{N}^{mis}, \mathbf{D}^{train}_{\mathsf{N}^{obs}}, \hat{\boldsymbol{\beta}}^{obs}}[\mathbb{E}[\hat{\mu}^{comp,L}|\mathcal{Q}_2]]$ hence

$$T_2 = \underbrace{\frac{1}{\mathsf{n}^2}\Big(\mathsf{n}(1-p^*)\tau^{*obs} + (\mu^{*obs})^2\mathbb{V}\text{ar}[\mathsf{N}^{obs}]\Big)}_{\text{sampling variance (due to } \hat{\mu}^{\text{obs}} \text{ and } \mathsf{N}^{\text{obs}})}$$
$$+ \quad \underbrace{\frac{1}{\mathsf{n}}v^{*obs}\frac{(p^*)^2}{1-p^*}(\overline{\boldsymbol{X}}^{*mis})^T\Big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\Big)^{-1}\overline{\boldsymbol{X}}^{*mis}}_{\text{imputation variance part 1}}$$
$$+ \quad \underbrace{\frac{1}{\mathsf{n}^2}\mathbb{V}\text{ar}[\mathsf{N}^{mis}]\Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\Big)^2}_{\text{imputation variance part 2}} + \underbrace{\frac{2}{\mathsf{n}^2}\Big[\mathsf{n}p^* + \mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\text{ar}[\mathsf{N}^{mis}]\Big]}_{\text{cross term (covariance)}},$$

in which

$$C = \begin{cases} 0 & \\ & \text{:S=M (mean)}, \\ \frac{p^*}{\mathsf{n}}\Bigg(v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0]\big)^2\Big] & \\ \quad +O\Big(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\text{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3}\Big)\Bigg) & \\ & \text{:S=R (random)} \end{cases}$$

Justification: first result is derived for mean strategy, which is followed by random imputation strategy.

Mean strategy

Variances at $\mathcal{Q}_3$ and $\mathcal{Q}_2$ are straighforward to derive:

$$\begin{aligned} \mathbb{V}\text{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_3] &= \mathbb{V}\text{ar}[\frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathbf{1}^T_{\mathsf{n}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})|\mathcal{Q}_3] = 0. \\ \mathbb{V}\text{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2] &= \mathbb{V}\text{ar}[\frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathbf{1}^T_{\mathsf{n}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})|\mathcal{Q}_2] \\ &= \mathbb{V}\text{ar}[\frac{1}{\mathsf{n}}\mathbf{1}^T_{\mathsf{n}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2] = \frac{1}{\mathsf{n}^2}\mathbf{1}^T_{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0} \\ &= \frac{1}{\mathsf{n}^2}\mathbf{1}^T_{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0} = \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}(\boldsymbol{\beta}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\boldsymbol{\beta}^{obs}_{-0}. \end{aligned}$$

Due to mathematical difficulty one approximates variance at first level using following regression coefficients estimates in predictions:

$$\begin{aligned}\hat{\beta}_0^{obs} &= 0\\ \hat{\boldsymbol{\beta}}_{-0}^{obs} &= \Big((\mathbf{D}_{\boldsymbol{X}}^{obs})^T\mathbf{D}_{\boldsymbol{X}}^{obs}\Big)^{-1}(\mathbf{D}_{\boldsymbol{X}}^{obs})^T\mathbf{D}_Y^{obs}.\end{aligned}$$

This approximation may underestimate variance because one parameter less is estimated. However, one compensates this by assuming that additional variance, including cross terms between intercept and slope terms, is $O(\mathsf{n}^{-1})$. As a remark, above slope terms are biased provided optimal intercept term is non-zero.

Variance given $\mathcal{Q}_1$ can be decomposed using chain rule as:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_1] &= \underbrace{\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\Big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2]\Big]}_{\text{first term}} \quad (1)\\ &\quad \underbrace{+\,\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\Big[\mathbb{E}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2]\Big]}_{\text{second term}},\end{aligned}$$

where

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2] &\approx \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Big((\boldsymbol{\beta}_{-0}^{obs})^T\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis}\boldsymbol{\beta}_{-0}^{obs}\Big)\\ \mathbb{E}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_2] &\approx \frac{1}{\mathsf{n}}\Big(\mathsf{n}^{obs}\mu^{obs}+\mathsf{n}^{mis}(\boldsymbol{\beta}_{-0}^{obs})^T\overline{\boldsymbol{X}}^{*mis}\Big),\end{aligned}$$

Here one assumes that variance for predictions using all regression coefficients can be approximated using predictions using slope terms estimates without intercept term.

One begins by computing the first term in Equation (1) conditionalized also by

N^{mis}, after which integration over distribution of N^{mis} is done.

$$
\begin{aligned}
& \mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}\left[\frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Big((\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{obs}_{-0}\Big)\right] \\
= \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathrm{tr}\Big((\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{obs}_{-0}\Big)\right] \\
= \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathrm{tr}\Big(\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Big)\right] \\
= \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\mathrm{tr}\left(\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}\Big[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Big]\right) \\
= \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\mathrm{tr}\left(\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}\Big[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\Big]\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\right) \\
= \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\mathrm{tr}\Bigg(\Big(\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}] \\
& +\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]\Big)\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Bigg) \\
= \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Bigg\{\mathrm{tr}\Big(\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Big) \\
& +\mathrm{tr}\Big(\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Big)\Bigg\} \\
= \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Bigg\{\mathrm{tr}\Big(\frac{v^{*obs}}{\mathsf{n}-\mathsf{N}^{mis}}\mathbb{E}_{\mathbf{X}|\boldsymbol{R},n}\big[\Big(\frac{1}{\mathsf{n}-\mathsf{N}^{mis}}(\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}}\Big)^{-1}\big]\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Big) \\
& +\mathbb{V}\mathrm{ar}[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\boldsymbol{X}^{mis}]\Bigg\} \\
\approx \;& \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Bigg\{\frac{v^{*obs}}{\mathsf{n}-\mathsf{N}^{mis}}\mathrm{tr}\Big(\mathbb{E}[\boldsymbol{X}^{obs}(\boldsymbol{X}^{obs})^T]^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\Big) \\
& +\mathbb{V}\mathrm{ar}[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\boldsymbol{X}^{mis}]\Bigg\},
\end{aligned}
$$

where one uses same (large sample) approximation for $\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\big[\Big(\frac{1}{\mathsf{n}-\mathsf{N}^{mis}}(\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}}\Big)^{-1}\big]$ as Hastie et al have done [36].

Integration over distribution of N^{mis} is done as

$$
\begin{aligned}
& \mathbb{E}_{\mathsf{N}^{mis}}\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}\left[\frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\Big((\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{obs}_{-0}\Big)\right] \qquad (2)\\
\approx\; & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\left\{\frac{v^{*obs}}{\mathsf{n}-\mathsf{N}^{mis}}\mathrm{tr}\left(\mathbb{E}[\boldsymbol{X}^{obs}(\boldsymbol{X}^{obs})^T]^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\right)\right.\right.\\
& \left.\left.+\mathbb{V}\mathrm{ar}[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\boldsymbol{X}^{mis}]\right\}\right]\\
=\; & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\frac{v^{*obs}}{\mathsf{n}-\mathsf{N}^{mis}}\right]\mathrm{tr}\left(\mathbb{E}[\boldsymbol{X}^{obs}(\boldsymbol{X}^{obs})^T]^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\right)\\
& +\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\mathbb{V}\mathrm{ar}[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\boldsymbol{X}^{mis}]\right]\\
\approx\; & \frac{v^{*obs}}{\mathsf{n}^2}\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}-\mathsf{N}^{mis}}\right]\mathrm{tr}\left(\mathbb{E}[\boldsymbol{X}^{obs}(\boldsymbol{X}^{obs})^T]^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\right)\\
& +\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\mathbb{V}\mathrm{ar}\left[\Big(\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T+O\big((\mathsf{n}^{obs})^{-1}\big)\Big)\boldsymbol{X}^{mis}\right]\right]\\
\approx\; & \frac{v^{*obs}}{\mathsf{n}^2}\frac{\mathsf{n}p^*}{\mathsf{n}-\mathsf{n}p^*}\mathrm{tr}\left(\mathbb{E}[\boldsymbol{X}^{obs}(\boldsymbol{X}^{obs})^T]^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\right)\\
& +\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\mathbb{V}\mathrm{ar}[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\boldsymbol{X}^{mis}]\right]\\
=\; & \frac{v^{*obs}}{\mathsf{n}^2}\frac{p^*}{1-p^*}\mathrm{tr}\left(\big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}}+\overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\big)^{-1}\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\right)\\
& +\frac{p^*}{\mathsf{n}}\mathbb{V}\mathrm{ar}\left[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\boldsymbol{X}^{mis}\right],
\end{aligned}
$$

where first order Taylor approximation has been applied to compute $\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}-\mathsf{N}^{mis}}\right]$. Impact due to coefficient approximation error $O((\mathsf{N}^{obs})^{-1})$ is assumed to be neglibe, and is thus ignored.

Now one needs to solve the second term in Equation (1).

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{1}{\mathsf{n}}\Big(\mathsf{N}^{obs}\hat{\mu}^{obs}+\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis})\Big)\right] \qquad (3)\\
= \; &\frac{1}{\mathsf{n}^2}\Bigg(\underbrace{\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\hat{\mu}^{obs}\right]}_{\text{first term}}\\
&+\underbrace{\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}\right]}_{\text{second term}}\\
&+\underbrace{2\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\hat{\mu}^{obs},\,\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}\right]}_{\text{third term}}\Bigg).
\end{aligned}$$

The first term within paranthesis is computed, using chain rule (conditionalizing by N^{mis}), as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\hat{\mu}^{obs}\right] &= \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\mathsf{N}^{obs}\hat{\mu}^{obs}]\right] \qquad (4)\\
&\quad+\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\mathsf{N}^{obs}\hat{\mu}^{obs}]\right]\\
&\approx \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[(\mathsf{N}^{obs})^2\frac{\tau^{*obs}}{\mathsf{N}^{obs}}\right]+\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\left[(\mathsf{N}^{obs})\mu^{*obs}\right]\\
&\approx \mathsf{n}(1-p^*)\tau^{*obs}+(\mu^{*obs})^2\mathbb{V}\mathrm{ar}[\mathsf{N}^{obs}].
\end{aligned}$$

Chain rule is applied similarly to compute second term within paranthesis. This yields to:

$$
\begin{aligned}
& \mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}\right] \\
= \; & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right] \\
& +\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right] \\
= \; & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}(\mathsf{N}^{mis})^2\left[\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right] \\
& +\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{mis}\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right] \\
= \; & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}(\mathsf{N}^{mis})^2\left[(\overline{\boldsymbol{X}}^{*mis})^T\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]\overline{\boldsymbol{X}}^{*mis}\right] \\
& +\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{mis}\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\overline{\boldsymbol{X}}^{*mis}\right] \\
\approx \; & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}(\mathsf{N}^{mis})^2\left[(\overline{\boldsymbol{X}}^{*mis})^T\frac{v^{*obs}}{\mathsf{n}-\mathsf{N}^{mis}}\mathbb{E}[\boldsymbol{X}^{obs}(\boldsymbol{X}^{obs})^T]^{-1}\overline{\boldsymbol{X}}^{*mis}\right] \\
& +\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{mis}\right]\big(\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\overline{\boldsymbol{X}}^{*mis}\big)^2 \\
\approx \; & v^{*obs}\frac{(\mathsf{n}p^*)^2}{\mathsf{n}-\mathsf{n}p^*}(\overline{\boldsymbol{X}}^{*mis})^T\mathbb{E}[\boldsymbol{X}^{obs}(\boldsymbol{X}^{obs})^T]^{-1}\overline{\boldsymbol{X}}^{*mis} \\
& +\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\big)^2 \\
= \; & v^{*obs}\frac{\mathsf{n}(p^*)^2}{1-p^*}(\overline{\boldsymbol{X}}^{*mis})^T\big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}}+\overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\big)^{-1}\overline{\boldsymbol{X}}^{*mis} \\
& +\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\big)^2,
\end{aligned}
\tag{5}
$$

where large sample approximation of Hastie et al [36] and first order Taylor expansion (for $\mathbb{E}[\frac{(\mathsf{N}^{mis})^2}{\mathsf{n}-\mathsf{N}^{mis}}]$) have been applied.

The covariance part in third term of Equation (3) is approximated as

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\hat{\mu}^{obs}, \mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}\right]\\
= \;& \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\mathsf{N}^{obs}\hat{\mu}^{obs}, \mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right]\\
& +\mathbb{C}\mathrm{ov}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}}[\mathsf{N}^{obs}\hat{\mu}^{obs}],\right.\\
& \left.\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right]\\
= \;& \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\mathsf{N}^{mis}\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\mu}^{obs}, (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right]\\
& +\mathbb{C}\mathrm{ov}_{\mathsf{N}^{mis}|\mathsf{n}}\left[(\mathsf{N}^{obs})\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\mu}^{obs}],\right.\\
& \left.\mathsf{N}^{mis}\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right]\\
= \;& \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\mathsf{N}^{mis}\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\mu}^{obs}, (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right]\\
& +\mathbb{C}\mathrm{ov}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\mu^{*obs}, \mathsf{N}^{mis}\mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\boldsymbol{\beta}}^{obs}_{-0}]^T\overline{\boldsymbol{X}}^{*mis}\right]\\
\approx \;& \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\mathsf{N}^{mis}\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\mu}^{obs}, (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right]\\
& +\mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{C}\mathrm{ov}_{\mathsf{N}^{mis}|\mathsf{n}}\left[(\mathsf{n}^{obs}), \mathsf{N}^{mis}\right]\\
= \;& \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\mathsf{N}^{mis}\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\mu}^{obs}, (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]\right]\\
& +\mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}].
\end{aligned}$$

Further

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\mu}^{obs}] &= O\big((\mathsf{n}^{obs})^{-1}\big)\\
\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}] &= O\big((\mathsf{n}^{obs})^{-1}\big).
\end{aligned}$$

Remembering that $|\mathbb{C}\mathrm{ov}[\hat{\mu}^{obs}, (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}|\mathsf{n}^{mis}, \mathsf{n}]| \leq \sqrt{\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}]\mathbb{V}\mathrm{ar}[(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}]}$ yields to:

$$|\mathbb{C}\mathrm{ov}[\hat{\mu}^{obs}, (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}|\mathsf{n}^{mis}, \mathsf{n}]| \leq \sqrt{O\big((\mathsf{n}^{obs})^{-1}\big)O\big((\mathsf{n}^{obs})^{-1}\big)}.$$

Next it is assumed, without rigorous proof, that following rough (order) approximation holds:

$$\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}[\hat{\mu}^{obs},(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}] \approx O\big((\mathsf{n}^{obs})^{-1}\big).$$

Therefore

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathsf{N}^{obs}\hat{\mu}^{obs},\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}\right] && (6)\\
\approx\;& \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[(\mathsf{N}^{mis}-(\mathsf{N}^{mis})^2)O\big((\mathsf{n}^{obs})^{-1}\big)\right]+\mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\\
\approx\;& \frac{\mathsf{n}^2p^*-\mathsf{n}^2(p^*)^2}{\mathsf{n}-\mathsf{n}p^*}+\mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\\
=\;& \frac{\mathsf{n}p^*(1-p^*)}{1-p^*}+\mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\\
=\;& \mathsf{n}p^*+\mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}],
\end{aligned}$$

where first order Taylor approximation has been applied to compute $\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\left[(\mathsf{N}^{mis}-(\mathsf{N}^{mis})^2)O\big((\mathsf{N}^{obs})^{-1}\big)\right]$.

Second term in Equation (1) is solved by plugging Equations (4), (5), (6) into Equation (3). This yields:

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\frac{1}{\mathsf{n}}\Big(\mathsf{N}^{obs}\hat{\mu}^{obs}+\mathsf{N}^{mis}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis})\Big)\right] && (7)\\
\approx\;& \frac{1}{\mathsf{n}^2}\Bigg(\mathsf{n}(1-p^*)\tau^{*obs}+(\mu^{*obs})^2\mathbb{V}\mathrm{ar}[\mathsf{N}^{obs}]\\
&+v^{*obs}\frac{\mathsf{n}(p^*)^2}{1-p^*}(\overline{\boldsymbol{X}}^{*mis})^T\big(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}}+\overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\big)^{-1}\overline{\boldsymbol{X}}^{*mis}\\
&+\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\big)^2\\
&+2(\mathsf{n}p^*+\mu^{*obs}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}])\Bigg).
\end{aligned}$$

Approximation for $\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathsf{n}]$ follows by summing results (2), (7), and order term $O(\mathsf{n}^{-1})$ which is to compensate ignored variance from intercept term.

Simulated random imputation

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_3] &= \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R}|\mathcal{Q}_3] \\
&= \frac{1}{\mathsf{n}^2}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \mathbb{V}\mathrm{ar}[\hat{\epsilon}_j^{L,R}|\mathcal{Q}_3] \\
&= \frac{1}{\mathsf{n}^2}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} v^{obs,L,R} = \frac{\mathsf{n}^{mis}}{\mathsf{n}^2} v^{obs,L,R}.
\end{aligned}$$

Variance at second conditionalization level is

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2] + \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R}|\mathcal{Q}_2] \\
&\quad +2\mathbb{C}\mathrm{ov}[\hat{\mu}^{comp,L,M}, \frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R}|\mathcal{Q}_2] \\
&= \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}(\boldsymbol{\beta}_{-0}^{obs})^T \boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis} \boldsymbol{\beta}_{-0}^{obs} + \frac{\mathsf{n}^{mis}}{\mathsf{n}^2} v^{obs,L,R} + 0 \\
&= \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}((\boldsymbol{\beta}_{-0}^{obs})^T \boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis} \boldsymbol{\beta}_{-0}^{obs} + v^{obs,L,R}).
\end{aligned}$$

Variance is derived at first conditionalization level next. Variance is decomposed as follows

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_1] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_1] + \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R}|\mathcal{Q}_1] \\
&\quad +2\mathbb{C}\mathrm{ov}[\hat{\mu}^{comp,L,M}, \frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R}|\mathcal{Q}_1],
\end{aligned}$$

where first term is already known. Therefore, last two terms need to be solved. Now

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}] &= \mathbb{E}[\mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_2]|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}] \\
&\quad +\mathbb{V}\mathrm{ar}[\mathbb{E}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_2]|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}] \\
&= \mathbb{E}[\mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_2]|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}] \\
&= \mathbb{E}[\frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\hat{v}^{obs,L,R}|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}] \\
&= \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\mathbb{E}[\hat{v}^{obs,L,R}|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}] \\
\Rightarrow \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_1] &= \mathbb{E}[\mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}]|\mathcal{Q}_1] \\
&\quad +\mathbb{V}\mathrm{ar}[\mathbb{E}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{mis}|\mathcal{Q}_1, \mathsf{N}^{mis}=\mathsf{n}^{mis}]|\mathcal{Q}_1] \\
&\approx \mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}|\mathcal{Q}_1]\mathbb{E}[\hat{v}^{obs,L,R}|\mathcal{Q}_1] \approx \frac{\mathsf{n}p^*}{\mathsf{n}^2}\mathbb{E}[\hat{v}^{obs,L,R}|\mathcal{Q}_1] \\
&= \frac{p^*}{\mathsf{n}}\mathbb{E}[\hat{v}^{obs,L,R}|\mathcal{Q}_1].
\end{aligned}$$

Further

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}[\hat{\mu}^{comp,L,R}, \frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{L,R}|\mathcal{Q}_1] \\
&= \mathbb{E}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{C}\mathrm{ov}[\hat{\mu}^{comp,L,R}, \frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{L,R}|\mathcal{Q}_2]\right] \\
&\quad +\mathbb{C}\mathrm{ov}_{\mathbf{D}^{train},\hat{\boldsymbol{\beta}}^{obs},\mathsf{N}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2], \mathbb{E}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{L,R}|\mathcal{Q}_2]\right] \\
&= \mathbb{E}[0] + \mathbb{C}\mathrm{ov}[\mathbb{E}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_2], 0] = 0.
\end{aligned}$$

As a consequence

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_1] \approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_1] + \frac{p^*}{\mathsf{n}}\mathbb{E}[\hat{v}^{obs,L,R}|\mathcal{Q}_1].$$

Using following approximation

$$\begin{aligned}
\mathbb{E}[\hat{v}^{obs,L,R}|\mathcal{Q}_1] = \mathbb{V}\mathrm{ar}[\hat{\epsilon}^{L,R}|\mathsf{n}] &\approx v^{*obs} \\
&\quad +\mathbb{E}_{\boldsymbol{X}^{obs}}\left[\left(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathsf{n}]^T\boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}_0^{obs}|\mathsf{n}]\right)^2\right] \\
&\quad +O\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big),
\end{aligned} \tag{8}$$

One gets

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,R}|\mathcal{Q}_1] &\approx \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,L,M}|\mathcal{Q}_1] \\ &+\frac{p^*}{\mathsf{n}}\Bigg(v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{X}^{obs} \\ &-\mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}]\big)^2\Big] + O\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big)\Bigg).\end{aligned}$$

Approximation 4.16

Bias of $\hat{\tau}^{comp,L}$ for n observations can be approximated with

$$\begin{aligned}&\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,L}|\mathsf{n}] \approx \underbrace{p^*\Big(\mathrm{tr}(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) + C - \tau^{*mis}\Big)}_{A} \\ &+\underbrace{p^*(1-p^*)\Big[(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 - (\mu^{*obs} - \mu^{*mis})^2\Big]}_{B} \\ &+O(\mathsf{n}^{-1}),\end{aligned}$$

where term A is due to difference between variance of imputed and missing Y values and B is due model missmatch. Term A varies for imputation strategies namely added imputation variance C is

$$C = \begin{cases} 0 & \text{:S=M (mean imputation)} \\ v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}]\big)^2\Big] \\ \quad +O\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big) & \text{:S=R (random imputation)}\end{cases}$$

Justification: in derivation of results for linear regression methods extended variable $\boldsymbol{X}^{mis,+} = [1\ (\boldsymbol{X}^{mis})^T]$ is used. This equals to variable $\boldsymbol{X}^{mis}$ which has been augmented by constant value one (in first component). Corresponding first two moments are $\mathbb{E}[\boldsymbol{X}^{mis,+}] = \overline{\boldsymbol{X}}^{*mis,+}$ and $\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}] = \boldsymbol{\Sigma}^{*mis,+}_{\boldsymbol{X}}$. Note that all final results are given with normal notation. Result for mean strategy is derived at first, what is followed by random strategy.

Mean strategy

Expectation of variance estimator given $\mathcal{Q}_3$ is

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_3] &= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\tau^{imp} + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\mu^{obs}-\mu^{imp})^2 \\
&= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\frac{1}{\mathsf{n}^{mis}-1}(\mathbb{X}^{mis}\boldsymbol{\beta}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\boldsymbol{\beta}^{obs} \\
&\quad + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\mu^{obs} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}\mathbb{X}^{mis}\boldsymbol{\beta}^{obs})^2,
\end{aligned}$$

where $\mathbf{H} = \mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}$. Therefore bias equals to

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_3] &= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\frac{1}{\mathsf{n}^{mis}-1}(\mathbb{X}^{mis}\boldsymbol{\beta}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\boldsymbol{\beta}^{obs} \\
&\quad + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\mu^{obs} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}\mathbb{X}^{mis}\boldsymbol{\beta}^{obs})^2 - \tau^*.
\end{aligned}$$

At second conditionalization level $\mathbf{D}^{mis}_{\boldsymbol{X}}$ is random, thus

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_2] &= \mathbb{E}[\frac{\mathsf{N}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\frac{1}{\mathsf{n}^{mis}-1}(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs} \\
&\quad + \frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs} - \frac{1}{\mathsf{N}^{mis}}\mathbf{1}^T\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2] \\
&= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\frac{1}{\mathsf{n}^{mis}-1}\mathbb{E}[(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2] \\
&\quad + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\mathbb{E}[(\hat{\mu}^{obs} - \frac{1}{\mathsf{N}^{mis}}\mathbf{1}^T\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2] \\
&= (1-\frac{1}{\mathsf{n}-1})\tau^{obs} + \frac{1}{\mathsf{n}-1}\mathbb{E}[(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2] \\
&\quad + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\mathbb{E}[(\hat{\mu}^{obs} - \frac{1}{\mathsf{N}^{mis}}\mathbf{1}^T\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2],
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}[(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2] &= \mathbb{E}[\mathrm{tr}((\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})|\mathcal{Q}_2] \\
&= \mathbb{E}[\mathrm{tr}(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T\mathbf{H})|\mathcal{Q}_2] \\
&= \mathbb{E}[\mathrm{tr}(\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T)|\mathcal{Q}_2] \\
&= \mathrm{tr}(\mathbb{E}[\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T|\mathcal{Q}_2]) \\
&= \mathrm{tr}(\mathbf{H}\mathbb{E}[\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T|\mathcal{Q}_2]). \\
\mathbb{E}[\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T|\mathcal{Q}_2] &= \Big(((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}} \\
&\quad + (\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}\mathbf{I}_{\mathsf{n}^{mis}}\Big),
\end{aligned}$$

thus

$$\begin{aligned}
& \mathbb{E}[(\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^T\mathbf{H}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2] \\
= \;& \mathrm{tr}(\mathbf{H}(((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}} + (\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}\mathbf{I}_{\mathsf{n}^{mis}})) \\
= \;& \mathrm{tr}(\mathbf{H}(((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})) + \mathrm{tr}(\mathbf{H}(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}\mathbf{I}_{\mathsf{n}^{mis}}) \\
= \;& \mathrm{tr}(\mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})((((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{mis,+})^2\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})) \\
& +\mathrm{tr}((\mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}\mathbf{I}_{\mathsf{n}^{mis}}) \\
= \;& ((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2\mathrm{tr}((\mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})(\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})) \\
& +(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}\mathrm{tr}((\mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})\mathbf{I}_{\mathsf{n}^{mis}}) \\
= \;& ((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2\mathrm{tr}((\mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}) \\
& +\boldsymbol{\beta}^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}\mathrm{tr}(\mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}) \\
= \;& ((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2\mathrm{tr}((\mathbf{I}_{\mathsf{n}^{mis}} - \frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}) \\
& +(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}(\mathsf{n}^{mis} - 1) \\
= \;& ((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2[\mathrm{tr}(\mathbf{I}_{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}) - \mathrm{tr}(\frac{1}{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})] \\
& +(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}(\mathsf{n}^{mis} - 1) \\
= \;& ((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2[\mathrm{tr}(\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}) - \frac{1}{\mathsf{n}^{mis}}\mathrm{tr}(\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}\mathbf{1}_{\mathsf{n}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}})] \\
& +(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}(\mathsf{n}^{mis} - 1) \\
= \;& ((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2[\mathsf{n}^{mis} - \frac{1}{\mathsf{n}^{mis}}(\mathsf{n}^{mis})^2] + (\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}(\mathsf{n}^{mis} - 1) \\
= \;& (\mathsf{n}^{mis} - 1)(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs} \\
= \;& (\mathsf{n}^{mis} - 1)(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0},
\end{aligned}$$

and

$$\begin{aligned}
& \mathbb{E}[(\hat{\mu}^{obs} - \frac{1}{\mathsf{N}^{mis}}\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2] \\
= \;& \mathbb{E}[(\hat{\mu}^{obs})^2 - 2\hat{\mu}^{obs}\frac{1}{\mathsf{N}^{mis}}\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs} + (\frac{1}{\mathsf{N}^{mis}}\mathbf{1}^T_{\mathsf{n}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2] \\
= \;& (\mu^{obs})^2 - 2\mu^{obs}\frac{1}{\mathsf{n}^{mis}}\mathbb{E}[\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2] + \mathbb{E}[\frac{1}{(\mathsf{N}^{mis})^2}(\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2] \\
= \;& (\mu^{obs})^2 - 2\mu^{obs}\frac{1}{\mathsf{n}^{mis}}\mathsf{n}^{mis}(\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+} + \mathbb{E}[\frac{1}{(\mathsf{N}^{mis})^2}(\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2] \\
= \;& (\mu^{obs})^2 - 2\mu^{obs}(\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+} + \frac{1}{(\mathsf{n}^{mis})^2}\mathbb{E}[\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs}|\mathcal{Q}_2]
\end{aligned}$$

Further

$$\begin{aligned}
& \mathbb{E}[(\hat{\mu}^{obs} - \frac{1}{\mathsf{N}^{mis}}\mathbf{1}^T_{\mathsf{N}^{mis}}\mathbb{X}^{mis}\hat{\boldsymbol{\beta}}^{obs})^2|\mathcal{Q}_2] \\
= \; & (\mu^{obs})^2 - 2\mu^{obs}(\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+} + \frac{1}{(\mathsf{n}^{mis})^2}((\mathsf{n}^{mis})^2((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{mis,+})^2 \\
& + \mathsf{n}^{mis}(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs}) \\
= \; & (\mu^{obs})^2 - 2\mu^{obs}(\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+} + ((\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{mis,+})^2 + \frac{1}{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs} \\
= \; & (\mu^{obs} - (\boldsymbol{\beta}^{obs})^T\overline{\boldsymbol{X}}^{*mis,+})^2 + \frac{1}{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis,+}]\boldsymbol{\beta}^{obs} \\
= \; & (\mu^{obs} - (\boldsymbol{\beta}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \beta^{obs}_0)^2 + \frac{1}{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0}.
\end{aligned}$$

As a consequence

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_2] &= \frac{\mathsf{n}^{obs}-1}{\mathsf{n}-1}\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}\frac{1}{\mathsf{n}^{mis}-1}(\mathsf{n}^{mis}-1)\boldsymbol{\beta}^T_{-0}\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0} \\
& + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\Big[(\mu^{obs} - (\boldsymbol{\beta}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \beta^{obs}_0)^2 \\
& + \frac{1}{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0}\Big] \\
&= (1 - \frac{\mathsf{n}^{mis}}{\mathsf{n}-1})\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0} \\
& + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\Big[(\mu^{obs} - (\boldsymbol{\beta}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \beta^{obs}_0)^2 \\
& + \frac{1}{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0}\Big].
\end{aligned}$$

To conclude bias given $\mathcal{Q}_2$ is

$$\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_2] &= (1 - \frac{\mathsf{n}^{mis}}{\mathsf{n}-1})\tau^{obs} + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0} \\
& + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\Big[(\mu^{obs} - (\boldsymbol{\beta}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \beta^{obs}_0)^2 \\
& + \frac{1}{\mathsf{n}^{mis}}(\boldsymbol{\beta}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\boldsymbol{\beta}^{obs}_{-0}\Big] - \tau^*.
\end{aligned}$$

Expectation at first level is derived as

$$
\begin{aligned}
& \mathbb{E}\Bigg[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1\Bigg] \\
= \quad & \mathbb{E}\Bigg[(1-\frac{\mathsf{N}^{mis}}{\mathsf{n}-1})\hat{\tau}^{obs}|\mathcal{Q}_1\Bigg] + \mathbb{E}\Bigg[\frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1\Bigg] \\
& +\mathbb{E}\Bigg[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}[(\hat{\mu}^{obs}-(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}-\hat{\beta}^{obs}_0)^2 \\
& +\frac{1}{\mathsf{N}^{mis}}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\hat{\boldsymbol{\beta}}^{obs}_{-0}]|\mathcal{Q}_1\Bigg] \\
\overset{Taylor}{\approx} \quad & (1-\frac{\mathsf{n}p^*}{\mathsf{n}-1})\tau^{*obs}+\frac{\mathsf{n}p^*-1}{\mathsf{n}-1}\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
& +\mathbb{E}\Bigg[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\Bigg[(\hat{\mu}^{obs}-(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis}-\hat{\beta}^{obs}_0)^2 \\
& +\frac{1}{\mathsf{N}^{mis}}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\hat{\boldsymbol{\beta}}^{obs}_{-0}\Bigg]|\mathcal{Q}_1\Bigg] \\
\approx \quad & (1-p^*)\tau^{*obs}+p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
& +\mathbb{E}\Bigg[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}^2}(\hat{\mu}^{obs}-(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis,}-\hat{\beta}^{obs}_0)^2|\mathcal{Q}_1\Bigg] \\
& +\mathbb{E}\Bigg[\frac{\mathsf{N}^{obs}}{\mathsf{n}^2}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1\Bigg].
\end{aligned}
$$

Thus

$$
\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1] \quad &\approx \quad (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}^2}(\hat{\mu}^{obs} - (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \hat{\beta}^{obs}_0)^2|\mathcal{Q}_1] \\
&\quad\quad +\mathbb{E}[\frac{\mathsf{N}^{obs}}{\mathsf{n}^2}|\mathcal{Q}_1]\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T)|\mathcal{Q}_1] \\
&\approx \quad (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}^2}(\hat{\mu}^{obs} - (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \hat{\beta}^{obs}_0)^2|\mathcal{Q}_1] \\
&\quad\quad +\frac{\mathsf{n}-\mathsf{n}p^*}{\mathsf{n}^2}\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\mathrm{tr}(\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T)|\mathcal{Q}_1] \\
&= \quad (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}^2}(\hat{\mu}^{obs} - (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \hat{\beta}^{obs}_0)^2|\mathcal{Q}_1] \\
&\quad\quad +\frac{1-p^*}{\mathsf{n}}\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&= \quad (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}^2}(\hat{\mu}^{obs} - (\hat{\boldsymbol{\beta}}^{obs}_{-0})^T\overline{\boldsymbol{X}}^{*mis} - \hat{\beta}^{obs}_0)^2|\mathcal{Q}_1] + O(\mathsf{n}^{-1}) \\
&\overset{Taylor}{\approx} \quad (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad\quad + \left[\begin{array}{l} \dfrac{\mathbb{E}[\mathsf{N}^{mis}|\mathcal{Q}_1]\mathsf{n} - \mathbb{E}[(\mathsf{N}^{mis})^2|\mathcal{Q}_1]}{\mathsf{n}^2} \\ \\ (\mathbb{E}[\hat{\mu}^{obs}|\mathcal{Q}_1] - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 \end{array}\right] + O(\mathsf{n}^{-1}) \\
&= \quad (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad\quad + \left[\begin{array}{l} \dfrac{\mathbb{E}[\mathsf{N}^{mis}|\mathcal{Q}_1]\mathsf{n} - \mathbb{E}[(\mathsf{N}^{mis})^2|\mathcal{Q}_1]}{\mathsf{n}^2} \\ \\ (\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 \end{array}\right] + O(\mathsf{n}^{-1}),
\end{aligned}
$$

where first order Taylor approximation has been used.

Further

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1] &\approx (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad + \left[\frac{\mathsf{n}p^*\mathsf{n} - (\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}|\mathcal{Q}_1] + \mathbb{E}[\mathsf{N}^{mis}|\mathcal{Q}_1]^2)}{\mathsf{n}^2}\right. \\
&\quad \left.(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2\right] + O(\mathsf{n}^{-1}) \\
&\approx (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad + \left[\frac{\mathsf{n}p^*\mathsf{n} - (\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}|\mathcal{Q}_1] + \mathsf{n}^2(p^*)^2)}{\mathsf{n}^2}\right. \\
&\quad \left.(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0]|\mathcal{Q}_1])^2\right] + O(\mathsf{n}^{-1}) \\
&= (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad + \left[\frac{\mathsf{n}p^*\mathsf{n} - \mathsf{n}^2(p^*)^2}{\mathsf{n}^2}\right. \\
&\quad \left.(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2\right] + O(\mathsf{n}^{-1}) \\
&= (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
&\quad + p^*(1-p^*)(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 + O(\mathsf{n}^{-1}).
\end{aligned}$$

As a consequence

$$\begin{aligned}
&\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1] \\
\approx\ & (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
& + p^*(1-p^*)(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 + O(\mathsf{n}^{-1}) - \tau^* \\
=\ & (1-p^*)\tau^{*obs} + p^*\mathrm{tr}(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) \\
& + p^*(1-p^*)(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 + O(\mathsf{n}^{-1}) \\
& - ((1-p^*)\tau^{*obs} + p^*\tau^{*mis} + p^*(1-p^*)(\mu^{*obs} - \mu^{*mis})^2) \\
=\ & p^*\left(\mathrm{tr}(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) - \tau^{*mis}\right) \\
& + p^*(1-p^*)\left[(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 - (\mu^{*obs} - \mu^{*mis})^2\right] \\
& + O(\mathsf{n}^{-1}).
\end{aligned}$$

Random imputation strategy

Note that

$$\hat{\tau}^{L,R} = \frac{\mathsf{N}^{obs} - 1}{\mathsf{n} - 1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis} - 1}{\mathsf{n} - 1}\hat{\tau}^{imp,L,R} + \frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n} - 1)}(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,R})^2,$$

where

$$
\begin{aligned}
&\hat{\tau}^{imp,L,R}\\
=\;& \frac{1}{N^{mis}-1}\sum_{j=N^{obs}+1}^{\mathsf{n}}\Bigg((\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{mis,+}+\hat{\epsilon}_j^{L,R}-\hat{\mu}^{imp,L,M}-\frac{1}{N^{mis}}\sum_{k=N^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_k^{L,R}\Bigg)^2\\
=\;& \frac{1}{N^{mis}-1}\sum_{j=N^{obs}+1}^{\mathsf{n}}\Bigg((\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{mis,+}-\hat{\mu}^{imp,L,M}+\hat{\epsilon}_j^{L,R}-\frac{1}{N^{mis}}\sum_{k=N^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_k^{L,R}\Bigg)^2\\
=\;& \frac{1}{N^{mis}-1}\sum_{j=N^{obs}+1}^{\mathsf{n}}\Bigg(\Big((\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{mis,+}-\hat{\mu}^{imp,L,M}\Big)^2\\
& +2\Big((\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{mis,+}-\hat{\mu}^{imp,L,M}\Big)\Big(\hat{\epsilon}_j^{L,R}-\frac{1}{N^{mis}}\sum_{k=1}^{N^{mis}}\hat{\epsilon}_k^{L,R}\Big)\\
& +\Big(\hat{\epsilon}_j^{L,R}-\frac{1}{\mathsf{n}^{mis}}\sum_{k=N^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_k^{L,R}\Big)^2\Bigg)\\
=\;& \hat{\tau}^{imp,L,M}\\
& +\frac{1}{N^{mis}-1}\sum_{j=N^{obs}+1}^{\mathsf{n}}\Bigg(2\Big((\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{mis,+}-\hat{\mu}^{imp,L,M}\Big)\\
& \Big(\hat{\epsilon}_j^{L,R}-\frac{1}{N^{mis}}\sum_{k=N^{obs}+1}^{N^{mis}}\hat{\epsilon}_k^{L,R}\Big)+\Big(\hat{\epsilon}_j^{L,R}-\frac{1}{N^{mis}}\sum_{k=N^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_k^{L,R}\Big)^2\Bigg),
\end{aligned}
$$

and

$$
\begin{aligned}
(\hat{\mu}^{obs}-\hat{\mu}^{imp,L,R})^2 &= (\hat{\mu}^{obs}-\hat{\mu}^{imp,L,M}-\frac{1}{N^{mis}}\sum_{j=N^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{L,R})^2\\
&= (\hat{\mu}^{obs}-\hat{\mu}^{imp,L,M})^2-2(\hat{\mu}^{obs}-\hat{\mu}^{imp,L,M})\frac{1}{N^{mis}}\sum_{j=N^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{L,R}\\
&\quad +(\frac{1}{N^{mis}}\sum_{j=N^{obs}+1}^{\mathsf{n}}\hat{\epsilon}_j^{L,R})^2.
\end{aligned}
$$

As a consequence

$$\begin{aligned}
& \mathbb{E}[\hat{\tau}^{imp,L,R}|\mathcal{Q}_3] - \mathbb{E}[\hat{\tau}^{imp,L,M}|\mathcal{Q}_3] \\
= \; & \mathbb{E}[\frac{1}{\mathsf{N}^{mis}-1} \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} (\hat{\epsilon}_j^{L,R} - \frac{1}{\mathsf{N}^{mis}} \sum_{k=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_k^{L,R})^2 |\mathcal{Q}_3] \\
= \; & \frac{1}{\mathsf{n}^{mis}-1} \sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \mathbb{E}[(\hat{\epsilon}_j^{L,R} - \frac{1}{\mathsf{n}^{mis}} \sum_{k=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_k^{L,R})^2 |\mathcal{Q}_3] \\
= \; & \frac{1}{\mathsf{n}^{mis}-1} \sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \Bigg(\mathbb{E}[(\hat{\epsilon}_j^{L,R})^2|\mathcal{Q}_3] - 2\mathbb{E}[\hat{\epsilon}_j^{L,R} \frac{1}{\mathsf{n}^{mis}} \sum_{k=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_k^{L,R}|\mathcal{Q}_3] \\
& + \mathbb{E}[(\frac{1}{\mathsf{n}^{mis}} \sum_{k=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_k^{L,R})^2|\mathcal{Q}_3] \Bigg) \\
= \; & \frac{1}{\mathsf{n}^{mis}-1} \sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \Bigg(v^{obs,L,R} - 2\frac{1}{\mathsf{n}^{mis}} \sum_{k=\mathsf{n}^{obs}+1}^{\mathsf{n}} \mathbb{E}[\hat{\epsilon}_j^{L,R}\hat{\epsilon}_k^{L,R}|\mathcal{Q}_3] \\
& + \frac{1}{(\mathsf{n}^{mis})^2} \sum_{k=\mathsf{n}^{obs}+1}^{\mathsf{n}} \sum_{l=\mathsf{n}^{obs}+1}^{\mathsf{n}} \mathbb{E}[\hat{\epsilon}_k^{L,R}\hat{\epsilon}_l^{L,R}|\mathcal{Q}_3] \Bigg) \\
= \; & \frac{1}{\mathsf{n}^{mis}-1} \sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} (v^{obs,L,R} - 2\frac{1}{\mathsf{n}^{mis}} v^{obs,L} + \frac{1}{(\mathsf{n}^{mis})^2} \mathsf{n}^{mis} v^{obs,L,R}) \\
= \; & \frac{1}{\mathsf{n}^{mis}-1} (\mathsf{n}^{mis} v^{obs,L,R} - 2v^{obs,L,R} + v^{obs,L}) = \frac{1}{\mathsf{n}^{mis}-1} (\mathsf{n}^{mis}-1) v^{obs,L,R} \\
= \; & v^{obs,L,R}.
\end{aligned}$$

Further

$$\begin{aligned}
\mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,R})^2|\mathcal{Q}_3] \; = \; & \mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,M})^2|\mathcal{Q}_3] \\
& -2\mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,M}) \frac{1}{\mathsf{N}^{mis}} \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R}|\mathcal{Q}_3] \\
& + \mathbb{E}[(\frac{1}{\mathsf{N}^{mis}} \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R})^2|\mathcal{Q}_3] \\
= \; & \mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,M})^2|\mathcal{Q}_3] + \mathbb{E}[(\frac{1}{\mathsf{N}^{mis}} \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{L,R})^2|\mathcal{Q}_3] \\
= \; & \mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,M})^2|\mathcal{Q}_3] \\
& + \frac{1}{(\mathsf{n}^{mis})^2} \sum_{j=\mathsf{n}^{obs}+1}^{n} \sum_{k=\mathsf{n}^{obs}+1}^{\mathsf{n}} \mathbb{E}[\hat{\epsilon}_j^{L,R}\hat{\epsilon}_k^{L,R}|\mathcal{Q}_3] \\
= \; & \mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,M})^2|\mathcal{Q}_3] + \frac{1}{(\mathsf{n}^{mis})^2} \mathsf{n}^{mis} v^{obs,L,R} \\
= \; & \mathbb{E}[(\hat{\mu}^{obs} - \hat{\mu}^{imp,L,M})^2|\mathcal{Q}_3] + \frac{1}{\mathsf{n}^{mis}} v^{obs,L,R}.
\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}[\hat{\tau}^{comp,L,R}|\mathcal{Q}_3] &= \mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_3] + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}v^{obs,L,R} + \frac{\mathsf{n}^{mis}\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}\frac{1}{\mathsf{n}^{mis}}v^{obs,L,R} \\ &= \mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_3] + \frac{\mathsf{n}^{mis}-1}{\mathsf{n}-1}v^{obs,L,R} + \frac{\mathsf{n}^{obs}}{\mathsf{n}(\mathsf{n}-1)}v^{obs,L,R} \\ &\approx \mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_3] + \frac{\mathsf{n}^{mis}}{\mathsf{n}}v^{obs,L,R} + \frac{\mathsf{n}^{obs}}{\mathsf{n}^2}v^{obs,L,R}.\end{aligned}$$

Expectation at second level is derived as

$$\mathbb{E}[\hat{\tau}^{comp,L,R}|\mathcal{Q}_2] = \mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_2] + \frac{\mathsf{n}^{mis}}{\mathsf{n}}v^{obs,L,R} + \frac{\mathsf{n}^{obs}}{\mathsf{n}^2}v^{obs,L,R}.$$

At first level expectation is computed as

$$\begin{aligned}\mathbb{E}[\hat{\tau}^{comp,L,R}|\mathcal{Q}_1] &= \mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1] + \mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\hat{v}^{obs,L}|\mathcal{Q}_1] + \mathbb{E}[\frac{\mathsf{N}^{obs}}{\mathsf{n}^2}\hat{v}^{obs,L,R}|\mathcal{Q}_1] \\ &\overset{Taylor}{\approx} \mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1] + \frac{\mathsf{n}p^*}{\mathsf{n}}\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}] + \frac{\mathsf{n}-\mathsf{n}p^*}{\mathsf{n}^2}\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}] \\ &= \mathbb{E}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1] + \underbrace{p^*\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]}_{\text{imputation noise variance}} \\ &\quad + \underbrace{\frac{(1-p^*)}{\mathsf{n}}\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]}_{\text{additional estimation variance}},\end{aligned}$$

where first order Taylor approximation has been applied.

As a consequence

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\tau}^{L,R}|\mathcal{Q}_1] &\approx \mathbb{B}\text{ias}[\hat{\tau}^{comp,L,M}|\mathcal{Q}_1] + p^*\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}] + O(\mathsf{n}^{-1}) \\ &= p^*\Big(\text{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis}\mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}(\hat{\boldsymbol{\beta}}_{-0}^{obs})^T|\mathcal{Q}_1]) + \mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}] - \tau^{*mis}\Big) \\ &\quad + p^*(1-p^*)[(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathcal{Q}_1]^T\overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}_0^{obs}|\mathcal{Q}_1])^2 \\ &\quad - (\mu^{*obs} - \mu^{*mis})^2] + O(\mathsf{n}^{-1}),\end{aligned}$$

where $\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]$ is approximated as:

$$\begin{aligned}\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}] \approx \Bigg(& v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\Big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathsf{n}]^T\boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}_0^{obs}|\mathsf{n}]\Big)^2\Big] \\ & + O\Big(\frac{1}{\mathsf{n}(1-p^*)} + \frac{\mathbb{V}\text{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big)\Bigg).\end{aligned}$$

Consequence 4.17

Asymptotically one has the following approximation

$$\lim_{\mathsf{n}\to\infty} \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,L}] \approx \underbrace{p^*\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \boldsymbol{\beta}^{*obs}_{-0} + Cv^{*obs,L,R} - \tau^{*mis}\Big)}_{A}$$

$$+ \underbrace{p^*(1-p^*)\Bigg\{\Big[\mu^{*obs} - (\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} - (\beta^{*obs}_{0})\Big]^2 - (\mu^{*obs} - \mu^{*mis})^2\Bigg\}}_{B},$$

where

$$C = \begin{cases} 0 & \text{:S=M} \quad \text{(mean imputation), and} \\ 1 & \text{:S=R} \quad \text{(simulated random imputation).} \end{cases}$$

Term A is bias due to difference between variance of imputed and missing Y values. Difference between mean of imputed and missing Y values is measured by bias term B. Further, $v^{*obs,L,R} = \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{v}^{obs,L,R}]$ is the optimal noise variance parameter over all possible training datas.

Justification: result is derived by taking limits of terms in approximation 4.16 as follows:

$$\lim_{\mathsf{n}\to\infty} \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp,L}|\mathsf{n}] \approx \lim_{\mathsf{n}\to\infty} p^*\Big(\mathrm{tr}(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) + C' - \tau^{*mis}\Big)$$
$$+ \lim_{\mathsf{n}\to\infty} p^*(1-p^*)\Big[(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T \overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 - (\mu^{*obs} - \mu^{*mis})^2\Big]$$
$$+ \lim_{\mathsf{n}\to\infty} O(\mathsf{n}^{-1})$$
$$= p^*\Big(\lim_{\mathsf{n}\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}(\hat{\boldsymbol{\beta}}^{obs}_{-0})^T|\mathcal{Q}_1]) + \lim_{\mathsf{n}\to\infty} C' - \tau^{*mis}\Big)$$
$$+ p^*(1-p^*)\Big[\lim_{\mathsf{n}\to\infty}(\mu^{*obs} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathcal{Q}_1]^T \overline{\boldsymbol{X}}^{*mis} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathcal{Q}_1])^2 - (\mu^{*obs} - \mu^{*mis})^2\Big]$$
$$= p^*\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \boldsymbol{\beta}^{*obs}_{-0} + Cv^{*obs,L,R} - \tau^{*mis}\Big)$$
$$+ p^*(1-p^*)\Bigg\{\Big[\mu^{*obs} - (\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} - (\beta^{*obs}_{0})\Big]^2 - (\mu^{*obs} - \mu^{*mis})^2\Bigg\},$$

where term C' is C from approximation 4.16.

A4.7 Linear regression / computations for simulation example

Mathematica is used to compute following theoretical quantities. The biased slope term $\beta_{-0}^{*obs,biased}$ is least squared (LS) optimal solution to minimization problem

$$\min_{\beta} \int_{-\infty}^{\infty} (\beta * x - \frac{1}{500} x^3)^2 f_{X^{obs}}(x) dx.$$

Optimal value for β is solved by computing zero point of derivative of the integral with respect to β. Assuming that order of differentiation and integration can be changed one gets

$$\begin{aligned}
\frac{\partial}{\partial \beta} \int_{-\infty}^{\infty} (\beta * x - \frac{1}{500} x^3)^2 f_{X^{obs}}(x) dx &= 0 \\
\iff \int_{-\infty}^{\infty} \frac{\partial}{\partial \beta} (\beta * x - \frac{1}{500} x^3)^2 f_{X^{obs}}(x) dx &= 0 \\
\iff \int_{-\infty}^{\infty} 2(\beta * x - \frac{1}{500} x^3) * x f_{X^{obs}}(x) dx &= 0 \\
\iff \beta &= \frac{\int_{-\infty}^{\infty} \frac{1}{500} x^4 f_{X^{obs}}(x) dx}{\int_{-\infty}^{\infty} x^2 f_{X^{obs}}(x) dx} \\
\overset{Mathematica}{\iff} \beta &= 69639/475000.
\end{aligned}$$

Term C in approximation 4.14 is computed as

$$\begin{aligned}
C \quad &= \quad v^{*obs} + \mathbb{E}_{X^{obs}} \Big[\big(g^{*obs}\big(X^{obs}\big) - \mathbb{E}[\hat{\beta}_{-0}^{obs} | \mathsf{n}] X^{obs} - \mathbb{E}[\hat{\beta}_{0}^{obs}] \big)^2 \Big] \\
&\approx \quad v^{*obs} + \mathbb{E}_{X^{obs}} \Big[\big(g^{*obs}\big(X^{obs}\big) - \beta_{-0}^{*obs,biased} X^{obs} \big)^2 \Big] \\
&= \quad v^{*obs} + \int_{-\infty}^{\infty} \Big[\big(\frac{1}{500} x^3 - \beta_{-0}^{*obs,biased} x \big)^2 \Big] \\
&\overset{Mathematica}{=} \quad 0.15 + 69639/475000,
\end{aligned}$$

thus to simplify computations optimal regression coefficients are replaced by zero intercept term and 'optimal' biased slope term $\beta_{-0}^{*obs,biased}$.

A4.8 Linear regression / unit level

Approximation 4.18

Provided variance of $Y^{obs}|\mathbf{x}^{obs}$ is constant v^{*obs} (homoscedastic situation) the expectation of $m\hat{s}e(Y^{comp,L})$ with n observations can be approximated as:

$$
\begin{aligned}
\mathbb{E}[m\hat{s}e(Y^{comp,L})|\mathsf{n}] \approx\; & \underbrace{\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]}_{\text{variability of approximative model}} \\
+ \quad & \underbrace{\Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}] - \mu^{*mis}\Big)^2}_{\text{global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
+ \quad & \underbrace{2\Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}]\Big)\Big(\mu^{*mis} - g^{*mis}(\overline{\boldsymbol{X}}^{*mis})\Big)}_{\text{cross term}} \\
+ \quad & \underbrace{v^{*obs}\Big(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3}\Big)\mathrm{tr}(\mathbf{A})}_{\text{expected variance of approximative model predictions}} \\
+ \quad & \underbrace{\quad C \quad}_{\text{expected imputation variance}} + \underbrace{\quad v^{*mis} \quad}_{\text{expected target variance}} + \underbrace{O\big(\mathsf{n}^{-1}\big)}_{\text{approximation error}} \quad ,
\end{aligned}
$$

where $\mathbf{A} = \left(\left(\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*mis}(\overline{\boldsymbol{X}}^{*mis})^T\right)\left(\boldsymbol{\Sigma}^{*obs}_{\boldsymbol{X}} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1}\right)$ and term C depends on imputation strategy S:

$$
C = \begin{cases}
0 \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{:S=M mean imputation} \\
v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\Big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}]\Big)^2\Big] \\
\;+O\Big(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3}\Big) \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{:S=R random imputation}
\end{cases}
$$

Justification: recall decomposition of mean squared error given in Equation 3.12 (Chapter 3):

$$\begin{aligned}
\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] &= \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{variability of conditional mean estimate}} \qquad (9)\\
+ &\quad \underbrace{(\mu_{\mathsf{n}}^{*imp}-\mu^{*mis})^2}_{\text{global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}}\\
+ &\quad \underbrace{2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]-\mu_n^{*imp}\right)\left(\mu_n^{*imp}-g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{cross term}}\\
+ &\quad \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\quad\right]}_{\text{expected variance of conditional mean estimate}} + \underbrace{v_{\mathsf{n}}^{*imp}}_{\text{expected imputation noise}}\\
+ &\quad \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[2\mathbb{C}\mathrm{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{second cross term}} + \underbrace{v^{*mis}}_{\text{expected target noise}}.
\end{aligned}$$

Variability of conditional mean estimate is computed by assuming that variance due to intercept term is of order $O(\mathsf{n}^{-1})$, hence

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right] &\approx \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{N}^{mis},\mathsf{n}]\boldsymbol{X}^{mis}\right]\\
&\quad +O(\mathsf{n}^{-1})\\
&\approx \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}}\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]+O(\mathsf{n}^{-1}).
\end{aligned}$$

Squared global bias is

$$(\mu_{\mathsf{n}}^{*imp}-\mu^{*mis})^2 = (\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}+\mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}]-\mu^{*mis})^2.$$

Let t denote cross term (divided by two), now

$$\begin{aligned}
t &= \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\mu_n^{*imp}\right]\\
&\quad -\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]g^{*mis}(\boldsymbol{X}^{mis})\right]\\
&\quad -\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[(\mu_n^{*imp})^2\right]\\
&\quad +\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mu_n^{*imp}g^{*mis}(\boldsymbol{X}^{mis})\right]\\
&= (\mu_n^{*imp})^2\\
&\quad -\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}]^T\boldsymbol{X}^{mis}+\mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}^{mis},\mathsf{n}])g^{*mis}(\boldsymbol{X}^{mis})\right]\\
&\quad -(\mu_n^{*imp})^2+\mu_n^{*imp}\mu^{*mis}.
\end{aligned}$$

Further, integration over N^{mis} and applying first order Taylor approximation gives

$$\begin{aligned}
t &= -\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Big[(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{X}^{mis}+\mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}])g^{*mis}(\boldsymbol{X}^{mis})\Big] \\
&\quad +\mu^{*imp}_{n}\mu^{*mis} \\
&\overset{Taylor}{\approx} -\Big[(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}+\mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}])g^{*mis}(\overline{\boldsymbol{X}}^{*mis})\Big] \\
&\quad +(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}+\mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}])\mu^{*mis} \\
&= \Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\overline{\boldsymbol{X}}^{*mis}+\mathbb{E}[\hat{\beta}^{obs}_{0}|\mathsf{n}]\Big)\Big(\mu^{*mis}-g^{*mis}(\overline{\boldsymbol{X}}^{*mis})\Big).
\end{aligned}$$

In computation of expected variance of conditional mean estimate one assumes that intercept term estimate is zero and slope terms are estimated as

$$\hat{\boldsymbol{\beta}}^{obs}_{-0}=\Big((\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}}\Big)^{-1}(\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{Y}.$$

Expected variance of conditional mean estimate is approximated as

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Bigg] \\
\approx\;&\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[(\boldsymbol{X}^{mis})^T\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}]\boldsymbol{X}^{mis}\Bigg]+O\big((\mathsf{n}^{obs})^{-1}\big) \\
=\;&\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\mathrm{tr}\Bigg((\boldsymbol{X}^{mis})^T\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}]\boldsymbol{X}^{mis}\Bigg)\Bigg]+O\big((\mathsf{n}^{obs})^{-1}\big) \\
=\;&\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\mathrm{tr}\Bigg(\boldsymbol{X}^{mis}(\boldsymbol{X}^{mis})^T\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}]\Bigg)\Bigg]+O\big((\mathsf{n}^{obs})^{-1}\big) \\
=\;&\mathrm{tr}\Bigg(\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\boldsymbol{X}^{mis}(\boldsymbol{X}^{mis})^T\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}]\Bigg]\Bigg)+O\big((\mathsf{n}^{obs})^{-1}\big) \\
=\;&\mathrm{tr}\Bigg(\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\boldsymbol{X}^{mis}(\boldsymbol{X}^{mis})^T\Bigg]\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}]\Bigg)+O\big((\mathsf{n}^{obs})^{-1}\big) \\
=\;&\mathrm{tr}\Bigg(\Bigg(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]+\overline{\boldsymbol{X}}^{*mis}(\overline{\boldsymbol{X}}^{*mis})^T\Bigg)\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis},\mathsf{n}]\Bigg)+O\big((\mathsf{n}^{obs})^{-1}\big) \\
\approx\;&\mathrm{tr}\Bigg(\Bigg(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]+\overline{\boldsymbol{X}}^{*mis}(\overline{\boldsymbol{X}}^{*mis})^T\Bigg)\frac{v^{*obs}}{\mathsf{n}^{obs}}\Bigg(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{obs}]+\overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\Bigg)^{-1}\Bigg) \\
&+O\big((\mathsf{n}^{obs})^{-1}\big) \\
=\;&\frac{v^{*obs}}{\mathsf{n}^{obs}}\mathrm{tr}\Bigg(\Bigg(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]+\overline{\boldsymbol{X}}^{*mis}(\overline{\boldsymbol{X}}^{*mis})^T\Bigg)\Bigg(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{obs}]+\overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\Bigg)^{-1}\Bigg) \\
&+O\big((\mathsf{n}^{obs})^{-1}\big).
\end{aligned}$$

Finally integration over N^{mis} using second order Taylor approximation yields to

$$\begin{aligned}
&\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]\\
\approx\ & v^{*obs}\Big(\frac{1}{\mathsf{n}(1-p^*)}+\frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big)\\
&\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis}+\overline{\boldsymbol{X}}^{*mis}(\overline{\boldsymbol{X}}^{*mis})^T\right)\left(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*obs}+\overline{\boldsymbol{X}}^{obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1}\right)\\
&+O\big(\mathsf{n}^{-1}+\mathsf{n}^{-2}\big).
\end{aligned}$$

The second cross term in Equation (9) is zero. For mean strategy this is clear because noise is not modelled. For random strategy reason is that conditional mean estimate and imputation noise are conditionally independent given training data, and expectation of imputation noise is zero.

Finally, expected imputation noise is computed in two integrals. Integration over distribution of $\boldsymbol{X}^{mis}$ is trivial as:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}^{mis}}[\mathbb{V}\mathrm{ar}[\hat{\epsilon}^{L,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]] \ \approx\ & v^{*obs}\\
&+\mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big)-\mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathsf{n}]\boldsymbol{X}^{obs}-\mathbb{E}[\hat{\beta}_0^{obs}|\mathsf{n}]\big)^2\Big]\\
&+O\big((\mathsf{n}^{obs})^{-1}\big).
\end{aligned}$$

Integration over distribution of response pattern is done as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\epsilon}^{L,R}|\mathsf{n}] \ \approx\ & v^{*obs}+\mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big)-(\mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{*obs}|\mathsf{n}])^T\boldsymbol{X}^{obs}-\mathbb{E}[\hat{\beta}_0^{*obs}|\mathsf{n}]\big)^2\Big]\\
&+O\Big(\frac{1}{\mathsf{n}(1-p^*)}+\frac{\mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]}{\mathsf{n}^3(1-p^*)^3}\Big).
\end{aligned}$$

Approximation 4.19

Provided variance of $Y^{obs}|\mathbf{x}$ is constant v^{*obs} for all $\mathbf{x}$ (homoscedastic situation) the mean squared error $\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n})$ can be approximated as:

$$\begin{aligned}
\mathrm{mse}(Y^{imp}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}) \ \approx\ & \underbrace{\Big(\mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathsf{n}^{mis},\mathsf{n}]^T\mathbf{x}^{mis}+\mathbb{E}[\hat{\beta}_0^{obs}|\mathsf{n}^{mis},\mathsf{n}]-g^{*mis}(\mathbf{x}^{mis})\Big)^2}_{\text{prediction bias}}\\
&+\underbrace{(\mathbf{x}^{mis})^T\frac{v^{*obs}}{\mathsf{n}^{obs}}\left(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*obs}+\overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1}\mathbf{x}^{mis}}_{\text{sampling variance (slopes)}}\\
&+\underbrace{\qquad C\qquad}_{\text{imputation variance}}+\underbrace{v^{*mis}(\mathbf{x}^{mis})}_{\text{target variance}}+\underbrace{O\big((\mathsf{n}^{obs})^{-1}\big)}_{\text{approximation error}}\ .
\end{aligned}$$

where constant C depends on imputation strategy S:

$$C = \begin{cases} 0 & \text{:S=M (mean imputation)}, \\ \underbrace{v^{*obs}}_{\text{expectation of variance of } \mathrm{Y}^{\mathrm{obs}}|\boldsymbol{X}^{\mathrm{obs}}} \\ + \underbrace{\mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T\boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}]\big)^2\Big]}_{\text{expected squared bias}} + \underbrace{O\big((\mathsf{n}^{obs})^{-1}\big)}_{\text{sampling variance}} & \text{:S=R (random imputation)}. \end{cases}$$

Justification: recall from Chapter 3 that

$$\begin{aligned} \mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}) &= \Big(\underbrace{\mathbb{E}[\hat{g}(\mathbf{x}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - g^{*mis}(\mathbf{x}^{mis})}_{\text{bias at } \mathrm{x}^{\mathrm{mis}}}\Big)^2 \\ &+ \underbrace{\mathbb{V}\mathrm{ar}[Y^{imp}_{|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}}]}_{\text{variance at } \mathrm{x}^{\mathrm{mis}}} \\ &+ \underbrace{\mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}^{mis}}]}_{\mathrm{v}^{*\mathrm{mis}}(\mathbf{x}^{\mathrm{mis}}),\text{ target noise at } \mathrm{x}^{\mathrm{mis}}}. \end{aligned}$$

Squared imputation bias is same for mean and random imputation strategies, thus:

$$\begin{aligned} \Big(\mathbb{E}[Y^{imp,L}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - \mathbb{E}[Y^{mis}|\mathbf{x}^{mis}]\Big)^2 &= \Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis}, \mathsf{n}]^T\mathbf{x}^{mis} + \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}^{mis}, \mathsf{n}] \\ &\quad - g^{*mis}(\mathbf{x}^{mis})\Big)^2. \end{aligned}$$

Mean strategy

In computation of variance one replaces estimator $\hat{\boldsymbol{\beta}}^{obs}$ by following simplification:

$$\begin{aligned} \hat{\beta}^{obs}_0 &= 0 \\ \hat{\boldsymbol{\beta}}^{obs}_{-0} &= ((\mathbf{D}^{obs}_{\boldsymbol{X}})^T\mathbf{D}^{obs}_{\boldsymbol{X}})^{-1}\mathbf{D}^T_{\boldsymbol{X}}\mathbf{D}^{obs}_Y. \end{aligned}$$

Above approximation may underestimate true variance of non-simplified $\hat{\boldsymbol{\beta}}^{obs}$. Therefore this is compensated by adding $O\big((\mathsf{n}^{obs})^{-1}\big)$ term. Thus

$$\begin{aligned} \mathbb{V}\mathrm{ar}[Y^{imp,L,M}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] &\approx (\mathbf{x}^{mis})^T\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis}, \mathsf{n}]\mathbf{x}^{mis} \\ &\quad + O\big((\mathsf{n}^{obs})^{-1}\big). \end{aligned} \tag{10}$$

where variance of regression coefficients is computed by applying chain rule of variance. This is done by first fixing covariates of training data, then integration over their distribution. Thus:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{V}\mathrm{ar}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\boldsymbol{X},\mathsf{n}^{mis},\mathsf{n}\right]\right] \\
&\quad +\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{V}\mathrm{ar}\left[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\boldsymbol{X},\mathsf{n}^{mis},\mathsf{n}\right]\right] \\
&= \mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\boldsymbol{X},\mathsf{n}^{mis},\mathsf{n}] + \mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[v^{*obs}\Big((\mathbf{D}_{\boldsymbol{X}}^{obs})^T\mathbf{D}_{\boldsymbol{X}}^{obs}\Big)^{-1}\right] \\
&\approx \mathbb{V}\mathrm{ar}[\boldsymbol{\beta}_{-0}^{*obs}] + \mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[v^{*obs}\Big((\mathbf{D}_{\boldsymbol{X}}^{obs})^T\mathbf{D}_{\boldsymbol{X}}^{obs}\Big)^{-1}\right] \\
&= v^{*obs}\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[\Big((\mathbf{D}_{\boldsymbol{X}}^{obs})^T\mathbf{D}_{\boldsymbol{X}}^{obs}\Big)^{-1}\right].
\end{aligned}$$

Next one applies same large sample approximation as Hastie et al. have done [36]. However, case here is little more complicated as expectation of $\boldsymbol{X}^{obs}$ is not zero. One gets

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}_{-0}^{obs}|\mathsf{n}^{mis},\mathsf{n}] &\approx v^{*obs}\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[\frac{1}{\mathsf{n}^{obs}}\Big(\frac{1}{\mathsf{n}^{obs}}(\mathbf{D}_{\boldsymbol{X}}^{obs})^T\mathbf{D}_{\boldsymbol{X}}^{obs}\Big)^{-1}\right] \qquad (11) \\
&= v^{*obs}\frac{1}{\mathsf{n}^{obs}}\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[\Big(\frac{1}{\mathsf{n}^{obs}}(\mathbf{D}_{\boldsymbol{X}}^{obs})^T\mathbf{D}_{\boldsymbol{X}}^{obs}\Big)^{-1}\right] \\
&\approx \frac{v^{*obs}}{\mathsf{n}^{obs}}\left(\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{obs}] + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1} \\
&= \frac{v^{*obs}}{\mathsf{n}^{obs}}\left(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*obs} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1}.
\end{aligned}$$

By plugging Equation (11) into Equation (10) one gets

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[Y^{imp,L,M}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] &\approx (\mathbf{x}^{mis})^T\frac{v^{*obs}}{\mathsf{n}^{obs}}\left(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{*obs} + \overline{\boldsymbol{X}}^{*obs}(\overline{\boldsymbol{X}}^{*obs})^T\right)^{-1}\mathbf{x}^{mis} \\
&\quad +O\big((\mathsf{n}^{obs})^{-1}\big).
\end{aligned}$$

Random imputation strategy

Variance for linear regression with noise term is computed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[Y^{imp,L,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{V}\mathrm{ar}[Y^{imp,L,M}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \qquad (12) \\
&\quad +\mathbb{V}\mathrm{ar}[\hat{\epsilon}^{L,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\
&\quad +2\mathbb{C}\mathrm{ov}[Y^{imp,L,M},\hat{\epsilon}^{L,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\
&= \mathbb{V}\mathrm{ar}[Y^{imp,L,M}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] + \mathbb{V}\mathrm{ar}[\hat{\epsilon}^{L,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}],
\end{aligned}$$

Note that $Y^{imp,L,M}$ and $\hat{\epsilon}^{L,R}$ are conditionally independent given training data. Further, expectation of $\hat{\epsilon}^{L,R}$ given training data is zero. As a consequence covariance term in Equation (12) is zero (recall chain rule of covariance).

Variance of imputation noise is computed as

$$
\begin{aligned}
& \mathbb{V}\text{ar}[\hat{\epsilon}^{L,R}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \\
= \;& \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{V}\text{ar}\left[\hat{\epsilon}^{L,R}|\mathbf{D}^{train}, \mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}\right]\right] \\
& + \mathbb{V}\text{ar}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{E}\left[\hat{\epsilon}^{L,R}|\mathbf{D}^{train}, \mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}\right]\right] \\
= \;& \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{V}\text{ar}\left[\hat{\epsilon}^{L,R}|\mathbf{D}^{train}, \mathsf{n}^{mis}, \mathsf{n}\right]\right] = \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}[\hat{v}^{obs,L,R}] \\
= \;& \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\left[\frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\Big(Y_j - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\Big)^2\right] \\
= \;& \frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\left[\Big(\epsilon\big(\boldsymbol{X}_j\big) + g^{*obs}(\boldsymbol{X}_j) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\Big)^2\right] \\
= \;& \frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{E}_{\mathbf{Y}|\mathbf{X},\mathsf{n}^{mis},\mathsf{n}}\left[\big(\epsilon\big(\boldsymbol{X}_j\big) + g^{*obs}\big(\boldsymbol{X}_j\big) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\big)^2\right]\right] \\
= \;& \frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{E}_{\mathbf{Y}|\mathbf{X},\mathsf{n}^{mis},\mathsf{n}}\left[\big(\epsilon\big(\boldsymbol{X}_j\big) + g^{*obs}\big(\boldsymbol{X}_j\big) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\big)^2\right]\right] \\
= \;& \frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\mathbb{E}_{\mathbf{Y}|\mathbf{X},\mathsf{n}^{mis},\mathsf{n}}\Big[\big(\epsilon\big(\boldsymbol{X}_j\big)^2 + \big(g^{*obs}\big(\boldsymbol{X}_j\big) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\big)^2 \\
& + 2\epsilon\big(\boldsymbol{X}_j\big)\big(g^{*obs}\big(\boldsymbol{X}_j\big) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\big)\Big]\Bigg] \\
= \;& \frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\mathbb{E}_{\mathbf{X}|\mathsf{n}^{mis},\mathsf{n}}\Bigg[\mathbb{V}\text{ar}[Y^{obs}|\boldsymbol{X}_j] + \mathbb{E}_{\mathbf{Y}|\mathbf{X},\mathsf{n}^{mis},\mathsf{n}}[\big(g^{*obs}\big(\boldsymbol{X}_j\big) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\big)^2] \\
& - 2\mathbb{E}_{\mathbf{Y}|\mathbf{X},\mathsf{n}^{mis},\mathsf{n}}[\epsilon\big(\boldsymbol{X}_j\big)(\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}]\Bigg] \\
= \;& \frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\Bigg[v^{*obs} + \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}[\big(g^{*obs}\big(\boldsymbol{X}_j\big) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\big)^2] \\
& - 2\mathbb{E}_{\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}^{mis},\mathsf{n}}\left[\mathbb{E}_{\mathbf{D}^{train}|\hat{\boldsymbol{\beta}}^{obs},\mathsf{n}^{mis},\mathsf{n}}[\epsilon\big(\boldsymbol{X}_j\big)(\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}]\right]\Bigg] \\
= \;& v^{*obs} + \frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\left[\mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}\left[\big(g^{*obs}\big(\boldsymbol{X}_j\big) - (\hat{\boldsymbol{\beta}}^{obs})^T\boldsymbol{X}_j^{+}\big)^2\right]\right],
\end{aligned}
$$

Without rigorous mathematical proof above quantity is assumed approximately to be following:

$$
\begin{aligned}
& \mathbb{V}\mathrm{ar}[\hat{\epsilon}^{L,R}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \\
\approx \quad & \underbrace{\qquad\qquad v^{*obs} \qquad\qquad}_{\text{expectation of variance of } \mathrm{Y}^{\mathrm{obs}}|\boldsymbol{X}^{\mathrm{obs}}} \\
& \underbrace{+\,\mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}^{mis}, \mathsf{n}]^T \boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}^{mis}, \mathsf{n}]\big)^2\Big]}_{\text{expected squared bias}} \\
& + \underbrace{O\big((\mathsf{n}^{obs})^{-1}\big)}_{\text{sampling variance}} \\
\approx \quad & v^{*obs} + \mathbb{E}_{\boldsymbol{X}^{obs}}\Big[\big(g^{*obs}\big(\boldsymbol{X}^{obs}\big) - \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \boldsymbol{X}^{obs} - \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}]\big)^2\Big] \\
& + O\big((\mathsf{n}^{obs})^{-1}\big).
\end{aligned}
$$

Consequence 4.21

Limit of expectation of $m\hat{s}e(Y^{comp,L})$ can be approximated as:

$$
\begin{aligned}
\lim_{\mathsf{n}\to\infty} \mathbb{E}[m\hat{s}e(Y^{comp,L})|\mathsf{n}] \approx \quad & \underbrace{(\boldsymbol{\beta}^{*obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \boldsymbol{\beta}^{*obs}_{-0}}_{\text{variability of limit of approximative model}} \\
+ \quad & \underbrace{\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} + \beta^{*obs}_0 - \mu^{*mis}\Big)^2}_{\text{asymptotic global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
+ \quad & \underbrace{2\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} + \beta^{*obs}_0\Big)\Big(\mu^{*mis} - g^{*mis}(\overline{\boldsymbol{X}}^{*mis})\Big)}_{\text{cross term}} \\
+ \quad & \underbrace{\qquad C \qquad}_{\text{optimal imputation variance}} + \underbrace{\qquad v^{*mis} \qquad}_{\text{expected target variance}},
\end{aligned}
$$

where term C depends on imputation strategy S:

$$
C = \begin{cases} 0 & \text{:S=M} \quad \text{(mean imputation)} \\ \\ v^{*obs,L,R} & \text{:S=R} \quad \text{(simulated random imputation),} \end{cases}
$$

in which $v^{*obs,L,R} = \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]$. The expectation $\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]$ is decomposed in approximation 4.16 (see term C for random strategy).

Justification: applying approximation 4.18 and taking limit gives asymptotical approximation (sample size is suppressed from limit to compress formulas):

$$\begin{aligned}
& \lim_{\mathsf{n}\to\infty} \mathbb{E}[\widehat{mse}(Y^{comp,L})|\mathsf{n}] \approx \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}] \\
+ \quad & \lim_{\mathsf{n}\to\infty} \Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}] - \mu^{*mis} \Big)^2 + \mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})] \\
+ \quad & \lim_{\mathsf{n}\to\infty} 2 \Big(\mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}_{-0}|\mathsf{n}]^T \overline{\boldsymbol{X}}^{*mis} + \mathbb{E}[\hat{\beta}^{obs}_0|\mathsf{n}] \Big) \Big(\mu^{*mis} - g^{*mis}(\overline{\boldsymbol{X}}^{*mis}) \Big) \\
+ \quad & \lim_{\mathsf{n}\to\infty} v^{*obs} \Big(\mathsf{n}^{-1}(1-p^*)^{-1} + \mathbb{V}\mathrm{ar}[\mathsf{N}^{mis}]\mathsf{n}^{-3}(1-p^*)^{-3} \Big) \mathrm{tr}(\mathbf{A}) \\
+ \quad & \lim_{\mathsf{n}\to\infty} C' + \lim_{\mathsf{n}\to\infty} v^{*mis} + \lim_{\mathsf{n}\to\infty} O\big(\mathsf{n}^{-1}\big)
\end{aligned}$$

where C' is term C in approximation 4.18.

Remarking that i) $\lim \mathbb{E}[\hat{\boldsymbol{\beta}}^{obs}|\mathsf{n}] = \boldsymbol{\beta}^{*obs}$, ii) variability of true model and expected target noise do not depend on sample size, and iii) expected variance of approximative model predictions vanishes asymptotically one gets:

$$\begin{aligned}
& \lim_{\mathsf{n}\to\infty} \mathbb{E}[\widehat{mse}(Y^{comp,L)}|\mathsf{n}] \approx \underbrace{(\boldsymbol{\beta}^{*obs}_{-0})^T \boldsymbol{\Sigma}^{*mis}_{\boldsymbol{X}} \boldsymbol{\beta}^{*obs}_{-0}}_{\text{variability of limit of approximative model}} \\
+ \quad & \underbrace{\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} + \beta^{*obs}_0 - \mu^{*mis} \Big)^2}_{\text{asymptotic global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
+ \quad & \underbrace{2\Big((\boldsymbol{\beta}^{*obs}_{-0})^T \overline{\boldsymbol{X}}^{*mis} + \beta^{*obs}_0 \Big) \Big(\mu^{*mis} - g^{*mis}(\overline{\boldsymbol{X}}^{*mis}) \Big)}_{\text{limit of cross term}} \\
+ \quad & \underbrace{\qquad C \qquad}_{\text{asymptotic imputation variance}} + \underbrace{\qquad v^{*mis} \qquad}_{\text{expected target variance}},
\end{aligned}$$

where term C depends on imputation strategy S:

$$C = \begin{cases} 0 & \text{:S=M} \quad \text{(mean imputation)} \\ \\ v^{*obs,L,R} & \text{:S=R} \quad \text{(simulated random imputation),} \end{cases}$$

in which $v^{*obs,L,R} = \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]$. The expectation $\mathbb{E}[\hat{v}^{obs,L,R}|\mathsf{n}]$ is decomposed in proposition 4.16 (see term C for random strategy).

Appendix for Chapter 5

In this appendix all justifications of approximations and consequences which were introduced in Chapter 5 are given.

A5.1 Justifications for nonparametric regression / moments

Here results for moment estimators are derived.

Approximation 5.1

The bias of the first moment for kernel and k-nn can be approximated with

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] \;\approx\;& \underbrace{p^*\Big(\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis}) - g^{*mis}(\boldsymbol{X}^{mis})]\Big)}_{\text{NMAR bias}} \\ &+ \quad \underbrace{\mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}C\right]}_{\text{estimation bias wrt. } \mathrm{g}^{*\mathrm{obs}}(\mathbf{x}^{\mathrm{mis}})} \quad + \quad \underbrace{D}_{\text{bias due to noise estimation}} \\ &+ \quad \underbrace{O(\mathsf{n}^{-1})}_{\text{approximation term}} \quad ,\end{aligned}$$

where terms $C = \mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[\mathbb{B}\text{ias}[\hat{g}^{obs,K/N}(\mathbf{x}^{mis})|\mathsf{n}^{mis}, \mathsf{n}]\Bigg]$ (expected conditional mean estimation bias) and D vary according to kernel/k-nn and imputation strategy as

$$C = \begin{cases} \frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{2f_{X^{obs}}(\overline{X}^{*mis})} \int \xi^2 K(\xi)d\xi\lambda^2(\mathsf{n}^{obs}) & \\ +o\big(\lambda^2(\mathsf{n}^{obs})\big) + O\big((\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs}))^{-1}\big) & (\mathrm{Kernel, p = 2}) \\ & \\ \frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{24f^3_{X^{obs}}(\overline{X}^{*mis})} \big(k(\mathsf{n}^{obs})/\mathsf{n}^{obs}\big)^2 & \\ +o\big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^2\big) + O\big((k(\mathsf{n}^{obs}))^{-1}\big) & (\mathrm{K-nn, p = 2}) \\ & \\ \frac{Q(g^{*obs}f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis}) - g^{*obs}(\overline{\boldsymbol{X}}^{*mis})Q(f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis})}{2f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis})(v_{p-1}f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis}))^{2/(p-1)}} (\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)} & \\ +o\Big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)}\Big) + O\big((k(\mathsf{n}^{obs}))^{-1}\big) & (\mathrm{K-nn, p > 2}), \end{cases}$$

and

$$D = \begin{cases} 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad :\text{S=M/S=R} \ \ \text{(mean and random strategy)} \\ p^*\mu^{*obs} - \mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\frac{1}{\mathsf{N}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}} \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n},\mathsf{n}^{mis}}\left[\hat{g}^{obs,K/N}\big(\boldsymbol{X}_j\big)\right]\right] \\ \qquad\qquad\qquad\qquad\qquad\qquad :\text{S=D} \ \ \text{(random donor)} \end{cases}$$

where second derivative of function $h(x)$ is denoted as $h''(x)$, and product of two functions $g(x)$ and $f(x)$ is denoted as $(gf)(x) = g(x)f(x)$,

$$Q(h)(\mathbf{x}) = \sum_{i=1,l=1}^{p-1,p-1}\int_{\mathbb{R}^{p-1}} \xi_i\xi_l \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_l} h(\mathbf{x}) I(||\boldsymbol{\xi}|| < 1) * (1/v_{p-1}) d\boldsymbol{\xi}, \tag{13}$$

$p-1$ is dimension of $\boldsymbol{X}$, and v_{p-1} is volume of unit ball in $\boldsymbol{X}$ space which is $\mathbb{R}^{p-1}$.

Justification: expectations of added noise terms are zero for random imputation strategies. Thus following holds for strategies $S \in \{M, R\}$:

$$\begin{aligned} & \mathbb{E}[\hat{\mu}^{comp,K/N}|\mathsf{n}] = \mathbb{E}\left[\frac{1}{\mathsf{n}}\Big(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathsf{N}^{mis}\hat{\mu}^{imp}\Big)\right] \\ \approx \quad & (1-p^*)\mu^{*obs} \\ + \quad & \mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\mathbb{E}_{\boldsymbol{X}^{mis}}\left[\underbrace{\mathbb{E}_{\mathbf{D}^{train}|\boldsymbol{X}^{mis},\mathsf{N}^{mis}}\Big[\sum_{(\boldsymbol{Y},\boldsymbol{X})_j\in\mathbf{D}^{train}} \overline{K}\big(\boldsymbol{X}^{mis},\boldsymbol{X}_j\big)Y_j|\mathsf{N}^{mis}\Big]}_{g^{*obs}(\mathbf{x}^{mis})+\mathbb{B}\mathrm{ias}[\hat{g}^{obs,K/N}(\mathbf{x}^{mis})|\mathsf{n}^{mis},\mathsf{n}]}\right]\right] \\ + \quad & O(\mathsf{n}^{-1}) \\ \approx \quad & (1-p^*)\mu^{*obs} + p^*\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})] \\ + \quad & \mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\mathbb{E}_{\boldsymbol{X}^{mis}}\Big[\mathbb{B}\mathrm{ias}[\hat{g}^{obs,K/N}(\mathbf{x}^{mis})|\mathsf{n}^{mis},\mathsf{n}]\Big]\right] + O(\mathsf{n}^{-1}), \end{aligned}$$

where $\mathbb{B}\mathrm{ias}[\hat{g}^{obs,K/N}(\mathbf{x}^{mis})|\mathsf{n}^{mis},\mathsf{n}]$ is estimation bias with respect to $g^{*obs}(\mathbf{x}^{mis})$ at point $\mathbf{x}^{mis}$. Bias result is immediately derivable as follows

$$\begin{aligned} \mathbb{B}\mathrm{ias}[\hat{\mu}^{comp,K/N,S}|\mathsf{n}] \ \approx \ & \underbrace{p^*\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis}) - g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{NMAR bias}} \\ & \underbrace{+\mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\mathbb{E}_{\boldsymbol{X}^{mis}}\Big[\mathbb{B}\mathrm{ias}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big]\right]}_{\text{estimation bias}} \\ & + \ \underbrace{O(\mathsf{n}^{-1})}_{\text{technical term}} \ , \end{aligned}$$

For donor strategy additional bias due to noise terms has to be computed. Now

$$\mathbb{E}[\hat{\mu}^{comp,K/N,D}|\mathsf{n}] \ = \ \mathbb{E}[\hat{\mu}^{comp,K/N}|\mathsf{n}] + \frac{1}{\mathsf{n}}\mathbb{E}[\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{K/N,D}].$$

where $\hat{\epsilon}_j^{K/N,D}$ is random draw from finite population $\{y_l - \hat{g}^{K/N}\big(\mathbf{x}_l\big)\}_{l=1}^{\mathsf{n}^{obs}}$ at conditionalisation $\mathcal{Q}_3$. For applying sampling theory result one needs to compute $\mathbb{E}[Y^c]$, which is done as follows

$$\mathbb{E}[Y^c|\mathcal{Q}_3] \;=\; \frac{1}{\mathsf{n}^{obs}}\sum_{l=1}^{\mathsf{n}^{obs}}\Big[y_l - g^{K/N}\big(\mathbf{x}_l\big)\Big].$$

Therefore for donor strategy one has

$$\begin{aligned}
& \mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N,D}|\mathsf{n}] \\
= \;& \mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] + \frac{1}{\mathsf{n}}\mathbb{E}[\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{K/N,D}|\mathsf{n}] \\
= \;& \mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] + \frac{1}{\mathsf{n}}\mathbb{E}_{\mathbf{D}^{train},\mathbf{D}^{test},\mathsf{N}^{mis},\hat{g}(\mathbf{x})|\mathsf{n}}\Big[\mathbb{E}[\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{K/N,D}|Q_3]\Big] \\
= \;& \mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] + \frac{1}{\mathsf{n}}\mathbb{E}_{\mathsf{N}^{mis}}\left[\mathbb{E}_{\mathbf{D}^{train},\hat{g}(\mathbf{x})|\mathsf{n},\mathsf{n}^{mis}}\left[\mathsf{N}^{mis}\frac{1}{\mathsf{N}^{obs}}\sum_{l=1}^{\mathsf{N}^{obs}}\Big[Y_l - \hat{g}^{K/N}\big(\boldsymbol{X}_l\big)\Big]\right]\right] \\
= \;& \mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] + \frac{1}{\mathsf{n}}\mathbb{E}_{\mathsf{N}^{mis}}\left[\mathsf{N}^{mis}\mu^{*obs} - \mathsf{N}^{mis}\mathbb{E}_{\mathbf{D}^{train}|\mathsf{n},\mathsf{n}^{mis}}\left[\frac{1}{\mathsf{N}^{obs}}\sum_{l=1}^{\mathsf{N}^{obs}}\hat{g}^{K/N}\big(\boldsymbol{X}_l\big)\right]\right] \\
\approx \;& \mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] + p^*\mu^{*obs} - \mathbb{E}_{\mathsf{N}^{mis}}\left[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\frac{1}{\mathsf{N}^{obs}}\sum_{l=1}^{\mathsf{N}^{obs}}\mathbb{E}_{\mathbf{D}^{train}|\mathsf{n},\mathsf{n}^{mis}}\Big[\hat{g}^{K/N}\big(\boldsymbol{X}_l\big)\Big]\right] \\
& +O(\mathsf{n}^{-1}).
\end{aligned}$$

Consequence 5.2

Following bounds can be derived for kernel and nearest neighbour methods

$$\begin{aligned}
\lim_{\lambda\to\infty}\mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] \;&=\; \mathbb{B}\text{ias}[\hat{\mu}^{comp,B}|\mathsf{n}] \approx p^*(\mu^{*mis} - \mu^{*obs}) + O(\mathsf{n}^{-1}) \\
\lim_{\lambda\to 0,\mathsf{n}\to\infty}\mathbb{B}\text{ias}[\hat{\mu}^{comp,K/N}|\mathsf{n}] \;&=\; p^*\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis}) - g^{*mis}(\boldsymbol{X}^{mis})].
\end{aligned}$$

Justification: when $\lambda \to \infty$ then $\mu^{comp,K/N} \to \mu^{comp,B}$ for any realization of $\mathbf{D}^{train}$, thus first limit result follows.

The second limit result follows by noticing that when sample size grows and smoothing is decreased at suitable rate then estimation bias at any given point (perhaps not in a zero measure set of points) converges towards zero. As a consequence only NMAR bias, if there is such, remains and the limit result follows.

Approximation 5.3

The variance of the first moment for kernel and k-nn can be approximated with

$$
\begin{aligned}
& \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}] \approx \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[(\frac{\mathsf{N}^{obs}}{\mathsf{n}})^2 \underbrace{\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}^{mis}]}_{\text{sampling variance}} \\
+ \quad & (\frac{\mathsf{N}^{mis}}{\mathsf{n}})^2 \Big(\frac{1}{\mathsf{N}^{mis}} \Big(\underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}} \Big[\underbrace{\mathbb{V}\mathrm{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis}, \mathsf{N}^{mis}, \mathsf{n}]}_{A} \Big]}_{\text{variance due to conditional mean prediction}} \\
+ \quad & \underbrace{\mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) + \underbrace{\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis}, \mathsf{n}]}_{B} \Big] \Big)}_{\text{variance due to conditional mean prediction}} \\
+ \quad & \underbrace{O\Big((\mathsf{N}^{obs})^{-\frac{1}{2}}\Big)}_{\text{due to correlated predictions}} \quad \Big) + \quad \underbrace{2\frac{\mathsf{N}^{obs}\mathsf{N}^{mis}}{\mathsf{n}^2} O\Big((\mathsf{N}^{mis}\mathsf{N}^{obs})^{-\frac{1}{2}}\Big)}_{\text{approximation for cross term (covariance)}} \quad \Bigg] \\
+ \quad & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}} \Bigg[\underbrace{\Big(1 - \frac{\mathsf{N}^{mis}}{\mathsf{n}}\Big)\mu^{*obs} + \frac{\mathsf{N}^{mis}}{\mathsf{n}} \Big(\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})]}_{\mathbb{E}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}^{mis}]} \\
& \underbrace{+\mathbb{E}_{\boldsymbol{X}^{mis}} \mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis}, \mathsf{n}] \Big)}_{\mathbb{E}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}^{mis}]} \Bigg] + \underbrace{\qquad C \qquad}_{\text{imputation noise variance}},
\end{aligned}
$$

where terms A, B, and C depend on estimation method (kernel or k-nn) and on imputation strategy $\hat{\epsilon}^S$ as follows:

$$
A = \begin{cases}
\frac{\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs}=X^{mis}]}{f_{X^{obs}}(X^{mis})\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})} \int K^2(\xi)d\xi + o\Big(\frac{1}{\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\Big) & (\mathrm{Kernel, p = 2}), \\[2em]
\frac{v_{p-1}\mathbb{V}\mathrm{ar}[Y^{obs}|\mathbf{X}^{obs}=\mathbf{X}^{mis}]}{k(\mathsf{n}^{obs})} + o((k(\mathsf{n}^{obs}))^{-1}) & (\mathrm{K - nn, p \geq 2}),
\end{cases}
$$

$$
B = \begin{cases}
\frac{(g^{*obs}f_{X^{obs}})''(X^{mis}) - g^{*obs}(X^{mis})f''_{X^{obs}}(X^{mis})}{2f_{X^{obs}}(X^{mis})} \int \xi^2 K(\xi)d\xi\lambda^2(\mathsf{n}^{obs}) & \\
+o(\lambda^2(\mathsf{n}^{obs})) + O((\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs}))^{-1}) & (\mathrm{Kernel, p = 2}), \\[2em]
\frac{(g^{*obs}f_{X^{obs}})''(X^{mis}) - g^{*obs}(X^{mis})f''_{X^{obs}}(X^{mis})}{24f^3_{X^{obs}}(X^{mis})}(k(\mathsf{n}^{obs})/(\mathsf{n}^{obs}))^2 & \\
+o((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^2) + O((k(\mathsf{n}^{obs}))^{-1}) & (\mathrm{K - nn, p = 2}), \\[2em]
\frac{Q(g^{*obs}f_{\boldsymbol{X}^{obs}})(\boldsymbol{X}^{mis}) - g^{*obs}(\boldsymbol{X}^{mis})Q(f_{\boldsymbol{X}^{obs}})(\boldsymbol{X}^{mis})}{2f_{\boldsymbol{X}^{obs}}(\boldsymbol{X}^{mis})(v_{p-1}f_{\boldsymbol{X}^{obs}}(\boldsymbol{X}^{mis}))^{2/(p-1)}} \Big(\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}}\Big)^{2/(p-1)} & \\
+o\Big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)}\Big) + O((k(\mathsf{n}^{obs}))^{-1}) & (\mathrm{K - nn, p > 2}),
\end{cases}
$$

where $Q(h)(x)$ is defined in Equation (13) and

$$C = \begin{cases} 0 \\ \qquad\qquad\qquad\qquad\qquad :\text{S=M} \;\; \text{(mean imputation)} \\ \frac{p^{*}v^{*obs}}{\mathsf{n}} + \frac{1}{\mathsf{n}^{2}}\mathbb{E}_{\mathsf{N}^{mis}}\left[\mathsf{N}^{mis}\mathbb{E}_{\boldsymbol{X}^{obs}}\left[\left(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis}}[\hat{g}^{obs,K/N}\big(\boldsymbol{X}^{obs})]\right)^{2}\right]\right] \\ \qquad\qquad\qquad\qquad\qquad :\text{S=R,S=D} \;\; \text{(random and donor strategies)} \end{cases}$$

Justification: variance at $\mathcal{Q}_1 = \{\mathsf{n}\}$ can be decomposed as

$$\begin{aligned} \mathbb{V}\text{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}] &= \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Big[\mathbb{V}\text{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n},\mathsf{n}^{mis}]\Big] \\ &\quad + \mathbb{V}\text{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\Big[\mathbb{E}[\hat{\mu}^{comp,K/N}|\mathsf{n},\mathsf{n}^{mis}]\Big]. \end{aligned}$$

Now $\mathbb{V}\text{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n},\mathsf{n}^{mis}]$ is computed as

$$\begin{aligned} \mathbb{V}\text{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n},\mathsf{n}^{mis}] &= \mathbb{V}\text{ar}\left[\frac{1}{\mathsf{n}}\Big(\mathsf{n}^{obs}\hat{\mu}^{obs} + \mathsf{n}^{mis}\hat{\mu}^{imp}\Big)\right] \\ &\approx (\frac{\mathsf{n}^{obs}}{\mathsf{n}})^{2}\mathbb{V}\text{ar}[\hat{\mu}^{obs}|\mathsf{n}^{mis}] + (\frac{\mathsf{n}^{mis}}{\mathsf{n}})^{2}\mathbb{V}\text{ar}[\hat{\mu}^{imp,K/N,M}|\mathsf{n}^{mis}] \\ &\quad + 2\frac{\mathsf{n}^{obs}\mathsf{n}^{mis}}{\mathsf{n}^{2}}O\big((\mathsf{n}^{mis}\mathsf{n}^{obs})^{-\frac{1}{2}}\big), \end{aligned}$$

where it is assumed that covariance term is approximately of order $\big(\mathsf{n}^{mis}\mathsf{n}^{obs}\big)^{-\frac{1}{2}}$. Now $\mathbb{V}\text{ar}[\hat{\mu}^{obs}|\mathsf{n}^{mis}] = \frac{\tau^{*obs}}{\mathsf{n}^{obs}} + O\big((\mathsf{n}^{obs})^{-1}\big)$, thus one needs to solve $\mathbb{V}\text{ar}[\hat{\mu}^{imp,K/N,M}|\mathsf{n}^{mis}]$. This is done as

$$\begin{aligned} \mathbb{V}\text{ar}[\hat{\mu}^{imp,K/N,M}|\mathsf{n}^{mis}] &= \mathbb{V}\text{ar}\Big[\frac{1}{\mathsf{n}^{mis}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}}\hat{g}^{obs,K/N}\big(\boldsymbol{X}_j\big)|\mathsf{n}^{mis}\Big] \\ &\approx \frac{1}{\mathsf{n}^{mis}}\mathbb{V}\text{ar}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis}}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathsf{n}^{mis}] + O\big((\mathsf{n}^{obs})^{-\frac{1}{2}}\big), \end{aligned}$$

where it has been assumed that covariance terms are approximately of order $O((\mathsf{n}^{obs})^{-\frac{1}{2}})$.

Now

$$\begin{aligned} \mathbb{V}\text{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Big[\mathbb{V}\text{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis},\mathsf{N}^{mis},\mathsf{n}]\Big] \\ &\quad + \mathbb{V}\text{ar}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Big[g^{*obs}(\boldsymbol{X}^{mis}) + \mathbb{B}\text{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big], \end{aligned}$$

where $\mathbb{V}\text{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathsf{n}^{mis},\mathsf{n}]$ and $\mathbb{B}\text{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]$ terms are pointwise prediction bias (with respect to $g^{*obs}(\mathbf{x}^{mis})$) and variance. Terms are available from mean squared error results for kernel and k-nn methods.

Term $\mathbb{E}[\hat{\mu}^{comp,K/N}|\mathsf{n}^{mis},\mathsf{n}]$ is computed as

$$\begin{aligned}
\mathbb{E}[\hat{\mu}^{comp,K/N}|\mathsf{n}^{mis},\mathsf{n}] &= (1-\frac{\mathsf{n}^{mis}}{\mathsf{n}})\mathbb{E}[\hat{\mu}^{obs}|\mathsf{n}^{mis},\mathsf{n}] + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}^{mis},\mathsf{n}] \\
&= (1-\frac{\mathsf{n}^{mis}}{\mathsf{n}})\mu^{*obs} + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}^{mis},\mathsf{n}] + O\big((\mathsf{n}^{obs})^{-1}\big) \\
&= (1-\frac{\mathsf{n}^{mis}}{\mathsf{n}})\mu^{*obs} + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mathbb{E}_{\boldsymbol{X}^{mis}}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathsf{n}^{mis},\mathsf{n}] \\
&\quad +O\big((\mathsf{n}^{obs})^{-1}\big) \\
&= (1-\frac{\mathsf{n}^{mis}}{\mathsf{n}})\mu^{*obs} + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\Big(\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})] \\
&\quad +\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big) + O\big((\mathsf{n}^{obs})^{-1}\big),
\end{aligned}$$

where $\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]$ is estimation bias with respect to $g^{*obs}(\boldsymbol{X}^{mis})$ and it is available from mean squared error results for kernel and k-nn.

Putting above results together yields to

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}] &= \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Big[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n},\mathsf{n}^{mis}]\Big] + \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\Big[\mathbb{E}[\hat{\mu}^{comp,K/N}|\mathsf{n},\mathsf{n}^{mis}]\Big] \\
&\approx \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Big[(\frac{\mathsf{N}^{obs}}{\mathsf{n}})^2\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}^{mis}] + (\frac{\mathsf{N}^{mis}}{\mathsf{n}})^2\mathbb{V}\mathrm{ar}[\hat{\mu}^{imp,K/N,M}|\mathsf{n}^{mis}] \\
&\quad +2\frac{\mathsf{N}^{obs}\mathsf{N}^{mis}}{\mathsf{n}^2}O\big(\mathsf{N}^{mis}\mathsf{N}^{obs})^{-\frac{1}{2}}\big)\Big] \\
&\quad +\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[(1-\frac{\mathsf{N}^{mis}}{\mathsf{n}})\mu^{*obs} \\
&\quad +\frac{\mathsf{N}^{mis}}{\mathsf{n}}\Big(\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})] + \mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big)\Bigg] \\
&\approx \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Big[(\frac{\mathsf{N}^{obs}}{\mathsf{n}})^2\mathbb{V}\mathrm{ar}[\hat{\mu}^{obs}|\mathsf{n}^{mis}] \\
&\quad +(\frac{\mathsf{N}^{mis}}{\mathsf{n}})^2(\frac{1}{\mathsf{N}^{mis}}(\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Big[\mathbb{V}\mathrm{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis},\mathsf{N}^{mis},\mathsf{n}]\Big] \\
&\quad +\mathbb{V}\mathrm{ar}_{\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}}\Big[g^{*obs}(\boldsymbol{X}^{mis}) + \mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big]) \\
&\quad +O\big((\mathsf{N}^{obs})^{-\frac{1}{2}}\big)) \\
&\quad +2\frac{\mathsf{N}^{obs}\mathsf{N}^{mis}}{\mathsf{n}^2}O\big((\mathsf{N}^{mis}\mathsf{N}^{obs})^{-\frac{1}{2}}\big)\Big] \\
&\quad +\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[(1-\frac{\mathsf{N}^{mis}}{\mathsf{n}})\mu^{*obs} + \frac{\mathsf{N}^{mis}}{\mathsf{n}}\Big(\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})] \\
&\quad +\mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big)\Bigg].
\end{aligned}$$

For strategies $S \in \{R, D\}$ variance of mean estimator increases, compared to mean strategy, due to modelled noise terms. For random imputation strategy one has

$$
\begin{aligned}
& \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N,R}|\mathsf{n}] \\
= \; & \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}] + \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{K/N,R}] \\
= \; & \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}] + \mathbb{E}_{\mathsf{N}^{mis}}[\mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{j=\mathsf{n}^{obs}+1}^{\mathsf{n}} \hat{\epsilon}_j^{K/N,R}|\mathsf{N}^{mis}]] \\
= \; & \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}] + \frac{1}{\mathsf{n}^2}\mathbb{E}_{\mathsf{N}^{mis}}[\mathsf{N}^{mis}\hat{v}^{obs,K/N,R}] \\
\approx \; & \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}] + \frac{1}{\mathsf{n}^2}\mathbb{E}_{\mathsf{N}^{mis}}\Bigg[\mathsf{N}^{mis}\bigg(v^{*obs} + \\
& +\mathbb{E}_{\boldsymbol{X}^{obs}}\bigg[\Big(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs})|\boldsymbol{X}^{obs}, \mathsf{n}^{mis}, \mathsf{n}]\Big)^2\bigg] + O\big((\mathsf{N}^{obs})^{-1}\big)\bigg)\Bigg] \\
\approx \; & \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N,M}|\mathsf{n}] + \frac{p^* v^{*obs}}{\mathsf{n}} + \\
& +\frac{1}{\mathsf{n}^2}\mathbb{E}_{\mathsf{N}^{mis}}\Bigg[\mathsf{N}^{mis}\mathbb{E}_{\boldsymbol{X}^{obs}}\bigg[\Big(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs})|\boldsymbol{X}^{obs}, \mathsf{n}^{mis}, \mathsf{n}]\Big)^2\bigg]\bigg)\Bigg] \\
& +O(\mathsf{n}^{-1}),
\end{aligned}
$$

where covariances between imputation noise terms are zero.

We assume that variance increase for donor strategy is approximately the same as for random imputation strategy.

Consequence 5.4

$$
\begin{aligned}
\lim_{\lambda \to \infty} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}] &= \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,B}|\mathsf{n}] \\
\lim_{\mathsf{n} \to \infty} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,K/N}|\mathsf{n}] &\approx 0.
\end{aligned}
$$

Justification: first result follows by noting that $\mu^{comp,K/N}$ converges to $\mu^{comp,B}$ when $\lambda \to \infty$ for any realization of $\mathbf{D}^{train}$. Second result follows by noticing that all decomposition terms in approximation 5.3 are decreasing functions of sample size n.

Approximation 5.5

Approximate bias can be written as

$$
\begin{aligned}
\mathbb{B}\text{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}] \approx{}& p^*(\underbrace{\mathbb{V}\text{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big]}_{\text{variability of expected conditional mean estimate}} \\
+ \quad & \underbrace{\qquad C \qquad}_{\text{imputation noise variance}} - \tau^{*mis}) \\
+ \quad & p^*(1-p^*)\left[(\mu^{*obs} - \mathbb{E}[Y^{imp,K/N}|\mathsf{n}])^2 - (\mu^{*obs} - \mu^{*mis})^2\right] \\
+ \quad & \underbrace{p^*\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{V}\text{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big]}_{\text{expected sampling variance}} \\
+ \quad & \underbrace{\qquad\qquad\qquad\qquad O(\mathsf{n}^{-1}) \qquad\qquad\qquad\qquad}_{\text{sampling variance of } \hat{\mu}^{\text{imp}} \text{ and approximation error (finite sample vs asymptotic)}},
\end{aligned}
$$

where term C is

$$
C = \begin{cases}
0 & \text{:S=M (mean imputation)} \\
v^{*obs} + \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{obs}}\left[\Big(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs})|\boldsymbol{X}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big)^2\right]}_{\text{expected squared bias}} & \\
+O(\mathsf{n}^{-1}) & \text{:S=R,D (random strategies)}
\end{cases}
$$

and $\mathbb{E}[Y^{imp}|\mathsf{n}] = \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\left[\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis},\mathsf{N}^{mis},\mathsf{n}]\right]$ is

$$
\mathbb{E}[Y^{imp}|\mathsf{n}] \approx \begin{cases}
\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})] + \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\Big[\mathbb{B}\text{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big] \\
\qquad\qquad\qquad\qquad\qquad\qquad \text{:S=M,R (mean and random)} \\
\\
\mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})] + \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\Big[\mathbb{B}\text{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big] \\
+\mu^{*obs} - \mathbb{E}[\frac{1}{\mathsf{N}^{obs}}\sum_{j=1}^{\mathsf{N}^{obs}}\hat{g}^{obs,K/N}\big(\boldsymbol{X}_j\big)] \\
\qquad\qquad\qquad\qquad\qquad\qquad \text{:S=D (random donor)}
\end{cases}
$$

where $\mathbb{B}\text{ias}[\boldsymbol{X}^{mis}|\mathsf{N}^{mis},\mathsf{n}]$ is estimation bias with respect to $g^{*obs}(\boldsymbol{X}^{mis})$.

Justification: rough approximation for expectation at conditionalisation $\mathcal{Q}_1 = \{\mathsf{n}\}$ is computed as

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{comp,K/N}|\mathsf{n}] &= \mathbb{E}[\frac{\mathsf{N}^{obs}-1}{\mathsf{n}-1}\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp,K/N} \\
&\quad + \frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp,K/N})^2|\mathsf{n}] \\
&= (1-p^*)\tau^{*obs} + \mathbb{E}[\frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp,K/N}|\mathsf{n}] \\
&\quad + \mathbb{E}[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp,K/N})^2|\mathsf{n}] + O(\mathsf{n}^{-1}) \\
&\overset{Taylor}{\approx} (1-p^*)\tau^{*obs} + p^*\mathbb{E}[\hat{\tau}^{imp,K/N}|\mathsf{n}] \\
&\quad + p^*(1-p^*)(\mu^{*obs} - \mathbb{E}[\hat{\mu}^{imp,K/N}|\mathsf{n}])^2 + O(\mathsf{n}^{-1}),
\end{aligned}$$

where Taylor remainder has been assumed to be of order $O(\mathsf{n}^{-1})$. Remainder cannot be smaller because estimation variance of $\hat{\mu}^{imp}$ is included in it.

Now it is further assumed that

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{imp,K/N}|\mathsf{n}] &\approx \mathbb{V}\mathrm{ar}[Y^{imp}|\mathsf{n}], \text{ and} \\
\mathbb{E}[\hat{\mu}^{imp,K/N}|\mathsf{n}] &\approx \mathbb{E}[Y^{imp}|\mathsf{n}].
\end{aligned}$$

Variance of imputation is computed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[Y^{imp}|\mathsf{n}] &= \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{V}\mathrm{ar}[Y^{imp}|\boldsymbol{X}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big] \\
&\quad + \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{E}[Y^{imp}|\boldsymbol{X}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big] \\
&\approx \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{V}\mathrm{ar}[\hat{g}^{obs}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big]}_{\text{expected sampling variance}} \\
&\quad + \underbrace{v^{*obs}}_{\text{expected imputation noise variance}} \\
&\quad + \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\left[\Big(g^{*obs}(\boldsymbol{X}^{mis}) - \mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\boldsymbol{X}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big)^2\right]}_{\text{expected imputation noise variance (cont.)}} \\
&\quad + \underbrace{O(\mathsf{n}^{-1})}_{\text{expected imputation noise variance (cont.)}} \\
&\quad + \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{E}[\hat{g}^{obs}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Big]}_{\text{variability of conditional mean estimate}}.
\end{aligned}$$

Further, for mean and simulated random imputation strategies term $\mathbb{E}[Y^{imp,K/N}|\mathsf{n}]$ equals to

$$\begin{aligned}
\mathbb{E}[Y^{imp,K/N}|\mathsf{n}] &= \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}\left[\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs})|\boldsymbol{X}^{obs},\mathsf{N}^{mis},\mathsf{n}]\right] \\
&= \mathbb{E}_{\boldsymbol{X}^{mis}}[g^{*obs}(\boldsymbol{X}^{mis})] + \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}}[\mathbb{B}\mathrm{ias}^{K/N}[\mathbf{x}^{mis}|\mathsf{n}^{mis},\mathsf{n}]],
\end{aligned}$$

where $\mathbb{B}\text{ias}^{K/N}[\mathbf{x}^{mis}|\mathsf{n}^{mis},\mathsf{n}]$ is estimation bias with respect to $g^{*obs}(\mathbf{x}^{mis})$.

Approximate bias is

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}] \;\approx\;& p^*(\mathbb{V}\text{ar}[Y^{imp}|\mathsf{n}]-\tau^{*mis})\\ &+p^*(1-p^*)\left[(\mu^{*obs}-\mathbb{E}[Y^{imp,K/N}|\mathsf{n}])^2-(\mu^{*obs}-\mu^{*mis})^2\right]\\ &+O(\mathsf{n}^{-1}).\end{aligned}$$

Claimed result follows by isolating expected sampling variance term from $\mathbb{V}\text{ar}[Y^{imp}|\mathsf{n}]$.

Consequence 5.6

Bounds for $\mathbb{B}\text{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}]$ are:

$$\begin{aligned}\lim_{\lambda\to\infty}\mathbb{B}\text{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}] \;=\;& \mathbb{B}\text{ias}[\hat{\tau}^{comp,B}|\mathsf{n}]\\ \lim_{\lambda\to 0,\mathsf{n}\to\infty}\mathbb{B}\text{ias}[\hat{\tau}^{comp,K/N}|\mathsf{n}] \;\approx\;& p^*(\mathbb{V}\text{ar}[g^{*obs}(\boldsymbol{X}^{mis})]+C-\tau^{*mis})\\ &+p^*(1-p^*)[(\mu^{*obs}-\mathbb{E}[g^{*obs}(\boldsymbol{X}^{mis})]-D)^2\\ &-(\mu^{*obs}-\mu^{*mis})^2],\end{aligned}$$

where terms C and D depend on imputation strategy $\hat{\epsilon}^S$ as follows:

$$C=\begin{cases}0 & \text{:S=M (mean)},\\ v^{*obs}+\lim\mathbb{E}_{\mathsf{N}^{mis}}\mathbb{E}_{\boldsymbol{X}^{obs}}\left[\left(g^{*obs}(\boldsymbol{X}^{obs})-\mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis}}[\hat{g}^{obs,K/N}\big(\boldsymbol{X}^{obs}\big)]\right)^2\right] & \text{:S=R,D (random)},\end{cases}$$

and

$$D=\begin{cases}0 & \text{:S=M,R (for mean and random imputation)},\\ \mu^{*obs}-\lim\mathbb{E}_{\mathsf{N}^{mis},\mathbf{D}^{train}}\left[\frac{1}{\mathsf{N}^{obs}}\sum_{j=1}^{\mathsf{N}^{obs}}\hat{g}^{obs,K/N}\big(\mathbf{x}_j\big)\right] & \text{:S=D (for random donor)}.\end{cases}$$

Justification: first limit result follows because imputations y_j^{imp} converge to μ^{obs} as $\lambda\to\infty$ for any realization of $\mathbf{D}^{train}$. Second limit follows because estimation bias and sampling variance converge towards zero as sample size grows and smoothing is decreased at suitable rate.

A5.2 Justifications for nonparametric regression / unit level

Unit level, mean squared error, results are computed here.

Approximation 5.12

Over distribution of training data set with n^{obs} observations the mean squared error at point $\mathbf{x}_0$ can be approximated as follows:

$$
\begin{aligned}
mse^{K/N}(Y^{imp}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}) \approx \Big(&\underbrace{g^{*obs}(\mathbf{x}_0) - g^{*mis}(\mathbf{x}_0)}_{\text{NMAR bias}} + \underbrace{\mathbb{B}\text{ias}[\hat{g}^{obs,K/N}(\mathbf{x}_0)|\mathsf{n}^{mis}, \mathsf{n}]}_{\text{A: estimation bias wrt. } g^{*obs}(\mathbf{x}_0)} \\
+ \quad &\underbrace{\qquad\qquad C \qquad\qquad}_{\text{Bias due to noise modelling}}\Big)^2 + \underbrace{\mathbb{V}\text{ar}[\hat{g}^{obs,K/N}(\mathbf{x}_0)|\mathsf{n}^{mis}, \mathsf{n}]}_{\text{B: imputation model variance}} \\
+ \quad &\underbrace{\qquad\quad D \qquad\quad}_{\text{imputation noise variance}} + \underbrace{v^{*mis}(\mathbf{x}_0)}_{\text{target variance}} .
\end{aligned}
$$

where terms A-D depend on non-parametric estimate (kernel vs k-nn) and imputation strategy as follows

$$
A = \begin{cases}
\frac{(g^{*obs} f_{X^{obs}})''(x_0) - g^{*obs}(x_0) f''_{X^{obs}}(x_0)}{2 f_{X^{obs}}(x_0)} \int \xi^2 K(\xi) d\xi \lambda^2(\mathsf{n}^{obs}) & (\mathrm{Kernel, p = 2}), \\
+o(\lambda^2(\mathsf{n}^{obs})) + O\big((\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs}))^{-1}\big) & \\
& \\
\frac{(g^{*obs} f_{X^{obs}})''(x_0) - g^{*obs}(x_0) f''_{X^{obs}}(x_0)}{24 f^3_{X^{obs}}(x_0)} \big(k(\mathsf{n}^{obs})/\mathsf{n}^{obs}\big)^2 & (\mathrm{K-nn, p = 2}), \\
+o((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^2) + O\big((k(\mathsf{n}^{obs}))^{-1}\big) & \\
& \\
\frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0) - g^{*obs}(\mathbf{x}_0) Q(f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0)}{2 f_{\boldsymbol{X}^{obs}}(\mathbf{x}_0)(v_{p-1} f(\mathbf{x}_0))^{2/(p-1)}} (\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)} & (\mathrm{K-nn, p > 2}), \\
+o\Big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)}\Big) + O\Big((k(\mathsf{n}^{obs}))^{-1}\Big) &
\end{cases}
$$

$$
B = \begin{cases}
\frac{\mathbb{V}\text{ar}[Y^{obs}|X^{obs}=x_0]}{f_{X^{obs}}(x_0)\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})} \int K^2(\xi) d\xi + o\big(\frac{1}{\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\big) & (\mathrm{Kernel, p = 2}), \\
& \\
\frac{v_{p-1}\mathbb{V}\text{ar}[Y^{obs}|\boldsymbol{X}^{obs}=\mathbf{x}_0]}{k(\mathsf{n}^{obs})} + o\Big((k(\mathsf{n}^{obs}))^{-1}\Big) & (\mathrm{K-nn, p \geq 2}),
\end{cases}
$$

and

$$
C = \begin{cases}
0 & \\
& \text{:S=M,R (mean and simulated random)}, \\
& \\
\mu^{*obs} - \mathbb{E}\Big[\frac{1}{\mathsf{n}^{obs}} \sum_{j=1}^{\mathsf{n}^{obs}} \hat{g}^{obs,K/N}(X_j)|\mathsf{n}^{mis}, \mathsf{n}\Big] + O(\mathsf{n}^{-1}) & \\
& \text{:S=D (random donor)},
\end{cases}
$$

and

$$D=\begin{cases} 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad :\text{S=M (mean)}, \\ v^{*obs}+\mathbb{E}_{X^{obs}}\left[(g^{*obs}(X^{obs})-\mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}[\hat{g}^{obs,K/N}(X^{obs}|\mathsf{n},\mathsf{n}^{mis}])^2\right] \\ \qquad\qquad\qquad\qquad\qquad\qquad :\text{S=R (simulated random)}, \\ v^{*obs}+\mathbb{E}_{\boldsymbol{X}^{obs}}\left[(g^{*obs}(\boldsymbol{X}^{obs})-\mathbb{E}_{\mathbf{D}^{train}_{\mathsf{n}^{obs}}|\mathsf{n}^{mis},\mathsf{n}}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{obs}|\mathsf{n},\mathsf{n}^{mis}])^2\right] \\ +\mathbb{E}_{\mathbf{D}^{train}_{\mathsf{n}^{obs}}|\mathsf{n}^{mis},\mathsf{n}}\left[\left(\frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\left(Y_j-\hat{g}^{obs,K/N}(\boldsymbol{X}_j)\right)\right)^2\right] \\ \qquad\qquad\qquad\qquad\qquad\qquad :\text{S=D (random donor)}. \end{cases}$$

Justification:

Kernel regression

Kernel regression results for univariate covariate X are immediately derived by applying theorem 5.8 with sample size n^{obs} as follows

$$\begin{aligned}\mathrm{mse}^{K,M}(Y^{imp}|x_0,\mathsf{n}^{mis},\mathsf{n}) &= \Big(g^{*obs}(x_0)-g^{*mis}(x_0) \\ &\quad +\frac{(g^{*obs}f_{X^{obs}})''(x_0)-g^{*obs}(x_0)f''_{X^{obs}}(x_0)}{2f_{X^{obs}}(x_0)}\int\xi^2K(\xi)d\xi\lambda^2(\mathsf{n}^{obs}) \\ &\quad +o(\lambda^2(\mathsf{n}^{obs}))+O((\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs}))^{-1})\Big)^2 \\ &\quad +\frac{\mathbb{V}\mathrm{ar}[Y^{obs}|X^{obs}=x_0]}{f_{X^{obs}}(x_0)\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\int K^2(\xi)d\xi+o(\frac{1}{\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}) \\ &\quad +v^{*mis}(x_0).\end{aligned}$$

Imputation variance increases for random strategies. For simulated random imputation one gets

$$\begin{aligned}\mathrm{mse}^{K,R}(Y^{imp}|x_0,\mathsf{n}^{mis},\mathsf{n}) &= \mathrm{mse}^{K,R}(Y^{imp}|x_0,\mathsf{n}^{mis},\mathsf{n})+\mathbb{V}\mathrm{ar}[\hat{\epsilon}^{K,R}|\mathsf{n}^{mis},\mathsf{n}] \\ &\quad +2\underbrace{\mathbb{C}\mathrm{ov}[\hat{g}^K(x_0),\hat{\epsilon}^{K,R}(x_0)|\mathsf{n},\mathsf{n}^{mis}]}_{=0} \\ &= \mathrm{mse}^{K,R}(x_0|\mathsf{n}^{mis},\mathsf{n})+\mathbb{E}\left[\frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\left(Y_j-\hat{g}^{obs,K}(X_j)\right)^2\right] \\ &\approx \mathrm{mse}^{K,R}(x_0|\mathsf{n}^{mis},\mathsf{n})+v^{*obs} \\ &\quad +\mathbb{E}_{X^{obs}}\left[(g^{*obs}(X^{obs})-\mathbb{E}_{\mathbf{D}^{train}_{\mathsf{n}^{obs}}|\mathsf{n}^{mis},\mathsf{n}}[\hat{g}^{obs,K}(X^{obs}|\mathsf{n},\mathsf{n}^{mis}])^2\right].\end{aligned}$$

For random donor strategy both imputation bias and variance are more complicated. Change in bias and variance are

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{Y}^{K,D}_{x_0}|\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{B}\text{ias}[\hat{g}^K_{x_0}|\mathsf{n}^{mis},\mathsf{n}] + \mathbb{E}[\hat{\epsilon}^{K,D}|\mathsf{n}^{mis},\mathsf{n}]\\
&= \mathbb{B}\text{ias}[\hat{g}^K_{x_0}|\mathsf{n}^{mis},\mathsf{n}] + \mathbb{E}\left[\frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\Big(Y_j - \hat{g}^K\big(X_j\big)\Big)|\mathsf{n}^{mis},\mathsf{n}\right]\\
&= \mathbb{B}\text{ias}[\hat{g}^K_{x_0}|\mathsf{n}^{mis},\mathsf{n}] + \mu^{*obs} - \mathbb{E}\left[\frac{1}{\mathsf{n}^{obs}}\sum_{j=1}^{\mathsf{n}^{obs}}\hat{g}^K\big(X_j\big)|\mathsf{n}^{mis},\mathsf{n}\right]\\
&\quad +O\big((\mathsf{n}^{obs})^{-1}\big)\\
\mathbb{V}\text{ar}[\hat{Y}^{K,D}_{x_0}|\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{V}\text{ar}[\hat{g}^K_{x_0}|\mathsf{n}^{mis},\mathsf{n}] + \mathbb{V}\text{ar}[\hat{\epsilon}^{K,D}|\mathsf{n}^{mis},\mathsf{n}] + 2\mathbb{C}\text{ov}[\hat{g}^K(x_0),\hat{\epsilon}^{K,D}|\mathsf{n}^{mis},\mathsf{n}]\\
&\approx \mathbb{V}\text{ar}[\hat{g}^K_{x_0}|\mathsf{n}^{mis},\mathsf{n}] + v^{*obs}\\
&\quad +\mathbb{E}_{X^{obs}}\left[(g^{*obs}(X^{obs}) - \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}[\hat{g}^K(X^{obs}|\mathsf{n}^{mis},\mathsf{n}])^2\right]\\
&\quad +2\mathbb{C}\text{ov}[\hat{g}^K(x_0),\hat{\epsilon}^{K,D}|\mathsf{n}^{mis},\mathsf{n}],
\end{aligned}$$

where for computation of variance it has been assumed that expectation of modelled noise terms is approximate zero (this is not assumed in the bias term). Remark that the covariance term is not generally zero.

K-nearest neighbour

In case of univariate covariate X and by applying corollary 5.9 with sample size n^{obs} one gets

$$\begin{aligned}
\text{mse}^{N,M}(Y^{imp}|\mathbf{x}_0,\mathsf{n}^{mis},\mathsf{n}) &= \Bigg(g^{*obs}(x_0) - g^{*mis}(x_0)\\
&\quad +\frac{(g^{*obs}f_{X^{obs}})''(x_0) - g^{*obs}(x_0)f''_{X^{obs}}(x_0)}{24f^3_{X^{obs}}(x_0)}\big(k(\mathsf{n}^{obs})/\mathsf{n}^{obs}\big)^2\\
&\quad +o\big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^2\big) + O\big(k(\mathsf{n}^{obs})^{-1}\big)\Bigg)^2\\
&\quad +\frac{v_{p-1}v^{*obs}(x_0)}{k(\mathsf{n}^{obs})} + o\big(k(\mathsf{n}^{obs})^{-1}\big) + v^{*mis}(x_0),
\end{aligned}$$

Additional terms for random strategies are same as in kernel regression except that $\hat{g}^K$ is replaced by $\hat{g}^N$.

For multivariate covariate $\boldsymbol{X}$ application of theorem 5.7 with sample size n^{obs}

yields to

$$\begin{aligned}
\mathrm{mse}^{N,M}(Y^{imp}|\mathbf{x}_0, \mathsf{n}^{mis}, \mathsf{n}) \; = \; & \Bigg(g^{*obs}(\mathbf{x}_0) - g^{*mis}(\mathbf{x}_0) \\
& + \frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0) - g^{*obs}(\mathbf{x}_0) Q(f_{\boldsymbol{X}^{obs}})(\mathbf{x}_0)}{2 f_{\boldsymbol{X}^{obs}}(\mathbf{x}_0)(v_{p-1} f(\mathbf{x}_0))^{2/(p-1)}} (\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)} \\
& + o\Big((\frac{k(\mathsf{n}^{obs})}{\mathsf{n}^{obs}})^{2/(p-1)} \Big) + O\big(k(\mathsf{n}^{obs})^{-1}\big) \Bigg)^2 \\
& + \frac{v_{p-1} v^{*obs}(\mathbf{x}_0)}{k(\mathsf{n}^{obs})} + o\big(k(\mathsf{n}^{obs})^{-1}\big) + v^{*mis}(\mathbf{x}_0),
\end{aligned}$$

Additional terms for random strategies are same as in kernel regression except that $\hat{g}^K$ is replaced by $\hat{g}^N$.

Approximation 5.13

Expectation of mean square error can be approximated as

$$\begin{aligned}
& \mathbb{E}[\hat{mse}(Y^{comp,K/N})|\mathsf{n}] \approx \underbrace{\qquad A \qquad}_{\text{expected squared imputation bias}} \\
+ \quad & \underbrace{\mathbb{E}_{\mathsf{N}^{mis}, \boldsymbol{X}^{mis}|\mathsf{n}} \Big[\mathbb{V}\mathrm{ar}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \Big]}_{\text{B: expected variance of conditional mean estimate}} + \underbrace{\quad v_{\mathsf{n}}^{*obs,K/N} \quad}_{\text{C: expected imputation noise}} \\
+ \quad & \underbrace{\mathbb{E}_{\mathsf{N}^{mis}, \boldsymbol{X}^{mis}|\mathsf{n}} \Big[2\mathbb{C}\mathrm{ov}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis}), \hat{\epsilon}_{\mathbf{x}^{mis}} | \mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \Big]}_{\text{D: cross term}} + \underbrace{\quad v^{*mis} \quad}_{\text{expected target noise}} \\
+ \quad & \underbrace{O\big(\mathsf{n}^{-1}\big)}_{\text{technical term}} \;,
\end{aligned}$$

where terms are as follows

$$\begin{aligned}
A \; = \; & \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis}, \boldsymbol{X}^{mis}|\mathsf{n}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) + \mathbb{B}\mathrm{ias}^{K/N}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis}, \mathsf{n}] \Big]}_{\text{variability of expected conditional mean estimate}} \\
& + (\underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) \Big] + E - \mu^{*mis})^2}_{\text{global bias}} + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
& + \underbrace{2\mathbb{E}_{\mathsf{N}^{mis}, \boldsymbol{X}^{mis}|\mathsf{n}} \Bigg[\Bigg(\mathbb{E}[\hat{g}^{obs,K/N}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - \mathbb{E}_{\boldsymbol{X}^{mis}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) \Big] - E \Bigg)}_{\text{cross term}} \\
& \underbrace{\Bigg(\mathbb{E}_{\boldsymbol{X}^{mis}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) \Big] + E - g^{*mis}(\boldsymbol{X}^{mis}) \Bigg) \Bigg]}_{\text{cross term}},
\end{aligned}$$

in which term E is

$$E = \begin{cases} \begin{array}{l} \Big[\frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis}) - g^{*obs}(\overline{\boldsymbol{X}}^{*mis}) Q(f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis})}{2 f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis})(v_{p-1} f(\boldsymbol{X}^{mis}))^{2/(p-1)}} \Big(\frac{\mathbb{E}\Big[k(\mathsf{N}^{obs})\Big]}{\mathsf{n}(1-p^*)}\Big)^{2/(p-1)} \\ +o\Big(\big(\frac{\mathbb{E}[k(\mathsf{N}^{obs})]}{\mathsf{n}(1-p^*)}\big)^{2/(p-1)}\Big)\Big] + O\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) \qquad (\mathrm{k-nn, p > 2}), \end{array} \\ \\ \begin{array}{l} \Bigg[\frac{(g^{*obs} f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis}) f''_{X^{obs}}(\overline{X}^{*mis})}{24 f^3_{X^{obs}}(\overline{X}^{*mis})} \Big(\mathbb{E}[k(\mathsf{N}^{obs})]/\mathsf{n}(1-p^*)\Big)^2 \\ +o\Big(\big(\frac{\mathbb{E}[k(\mathsf{N}^{obs})]}{\mathsf{n}(1-p^*)}\big)^2\Big) + O\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big)\Bigg] \qquad (\mathrm{k-nn, p = 2}), \end{array} \\ \\ \begin{array}{l} \frac{(g^{*obs} f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis}) f''_{X^{obs}}(\overline{X}^{*mis})}{2 f_{X^{obs}}(\overline{X}^{*mis})} \int \xi^2 K(\xi) d\xi \lambda^2\big(\mathsf{n}(1-p^*)\big) \\ +o\Big(\lambda^2\big(\mathsf{n}(1-p^*)\big)\Big) + O\Big(\big(\mathsf{n}(1-p^*)\lambda(\mathsf{n}(1-p^*))\big)^{-1}\Big) \\ \qquad (\mathrm{kernel, p = 2}). \end{array} \end{cases}$$

Terms B-D are following:

$$B = \begin{cases} \begin{array}{l} \frac{v_{p-1}}{\mathbb{E}[k(\mathsf{N}^{obs})]} \Bigg(v^{*obs}(\overline{\boldsymbol{X}}^{*mis}) + \frac{1}{2}\mathrm{tr}\Big(\mathbf{H}_{v^{*obs}} \mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\Big)\Bigg) + o\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) \\ \qquad (\mathrm{k-nn, p > 2}), \end{array} \\ \begin{array}{l} \frac{2}{\mathbb{E}[k(\mathsf{N}^{obs})]} \Bigg(v^{*obs}(\overline{X}^{*mis}) + \frac{1}{2}\Big(\frac{\partial^2}{\partial x^{mis} \partial x^{mis}} v^{*obs}(x^{mis})\Big)_{x^{mis} = \overline{X}^{*mis}} \mathbb{V}\mathrm{ar}[X^{mis}]\Bigg) \\ +o\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) \qquad (\mathrm{k-nn, p = 2}) \end{array} \\ \\ \begin{array}{l} \frac{v^{*obs}(\overline{X}^{*mis})}{f_{X^{obs}}(\overline{X}^{*mis})(\mathsf{n}(1-p^*))\lambda(\mathsf{n}(1-p^*))} \int K^2(\xi) d\xi + o\Big(\frac{1}{(\mathsf{n}(1-p^*))\lambda(\mathsf{n}(1-p^*))}\Big) \\ \qquad (\mathrm{kernel, p = 2}), \end{array} \end{cases}$$

where $\mathbf{H}_{v^{*obs}}$ is Hessian of $\mathbb{V}\mathrm{ar}[Y^{obs}|\boldsymbol{X} = \mathbf{x}]$ and

$$C = \begin{cases} 0 \qquad \text{:S=M (mean),} \\ v^{*obs} + \mathbb{E}_{\mathsf{N}^{mis}, X^{obs}}\Bigg[(g^{*obs}(X^{obs}) - \mathbb{E}_{\mathbf{D}^{train}_{\mathsf{n}^{obs}} | \mathsf{n}^{mis}, \mathsf{n}}[\hat{g}^{obs, K/N}(X^{obs}|\mathsf{n}, \mathsf{n}^{mis}])^2\Bigg] \\ \qquad \text{:S=R,D (random)} \end{cases}$$

and

$$D = \begin{cases} 0 & \text{:S=M/S=R} \quad \text{(mean and simulated random),} \\ O(\mathsf{n}^{-1}) & \text{:S=D} \quad \text{(random donor).} \end{cases}$$

Justification: Result for kernel regression is justified at first, after which results for nearest neighbour imputation (with univariate covariate and multivariate covariate) are derived.

Kernel regression

Recall decomposition for $\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}]$ from Equation (3.12). Approximation for expected variance for conditional mean estimate (term B) is computed as follows:

$$\begin{aligned}
& \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{V}\mathrm{ar}[\hat{g}^{K}(\boldsymbol{X}^{mis})|x^{mis},\mathsf{n}^{mis},\mathsf{n}\Bigg] \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[\mathbb{E}_{X^{mis}}\bigg[\frac{\mathbb{V}\mathrm{ar}[Y^{obs}|X^{mis}]}{f_{X^{obs}}(X^{mis})\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\int K^2(\xi)d\xi + o\Big(\frac{1}{\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\Big)\bigg]\Bigg] \\
\overset{Taylor}{\approx} \quad & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\bigg[\frac{v^{*obs}(\overline{X}^{*mis})}{f_{X^{obs}}(\overline{X}^{*mis})\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\int K^2(\xi)d\xi + o\Big(\frac{1}{\mathsf{n}^{obs}\lambda(\mathsf{n}^{obs})}\Big)\bigg] + O(\mathsf{n}^{-1}) \\
\overset{Taylor}{\approx} \quad & \frac{v^{*obs}(\overline{X}^{*mis})}{f_{X^{obs}}(\overline{X}^{*mis})(\mathsf{n}(1-p^*))\lambda(\mathsf{n}(1-p^*))}\int K^2(\xi)d\xi \\
& + o\Big(\frac{1}{(\mathsf{n}(1-p^*))\lambda(\mathsf{n}(1-p^*))}\Big) + O(\mathsf{n}^{-1}),
\end{aligned}$$

where first order Taylor approximation has been applied twice. Therefore result is quite rough at least for small sample size.

Expected imputation noise (term C) and cross term (D) are zero for mean imputation strategy. Therefore one just needs to compute expected squared imputation bias (term A). Thus variability of conditional mean estimate and global bias terms have to be computed. Due to mathematical difficulty variability of conditional mean estimate is written in implicit form as follows

$$\begin{aligned}
& \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{E}[\hat{g}^{obs,K}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Bigg] \\
& = \mathbb{V}\mathrm{ar}\Big[g^{*obs}(\boldsymbol{X}^{mis}) + \mathbb{B}\mathrm{ias}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Big].
\end{aligned}$$

Global bias is solved by computing μ_{n}^{*imp} which is done as

$$
\begin{aligned}
\mu_{\mathsf{n}}^{*K,imp} \quad &= \quad \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\mathbb{E}[\hat{g}^{obs,K}(\boldsymbol{X}^{mis})|\mathsf{n}^{mis},\mathbf{x}^{mis},\mathsf{n}] \\
&= \quad \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[g^{*obs}(\boldsymbol{X}^{mis}) + \mathbb{B}\mathrm{ias}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Bigg] \\
&= \quad \mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[g^{*obs}(\boldsymbol{X}^{mis})\Bigg] + \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[\mathbb{B}\mathrm{ias}[\boldsymbol{X}^{mis}|\mathsf{n}^{mis},\mathsf{n}]\Bigg] \\
&= \quad \mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[g^{*obs}(\boldsymbol{X}^{mis})\Bigg] \\
&\quad\quad +\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[\mathbb{E}_{\boldsymbol{X}^{mis}}\Big[\frac{(g^{*obs}f_{X^{obs}})''(X^{mis}) - g^{*obs}(X^{mis})f''_{X^{obs}}(X^{mis})}{2f_{X^{obs}}(X^{mis})} \\
&\quad\quad \int \xi^2 K(\xi)d\xi\lambda^2(\mathsf{N}^{obs}) + o(\lambda^2(\mathsf{N}^{obs})) + O((\mathsf{N}^{obs}\lambda(\mathsf{N}^{obs}))^{-1})\Big]\Bigg] \\
&\overset{Taylor}{\approx} \quad \mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[g^{*obs}(\boldsymbol{X}^{mis})\Bigg] \\
&\quad\quad +\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[\frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{2f_{X^{obs}}(\overline{X}^{*mis})} \\
&\quad\quad \int \xi^2 K(\xi)d\xi\lambda^2(\mathsf{N}^{obs}) + o\Big(\lambda^2(\mathsf{N}^{obs})\Big) + O\Bigg(\Big(\mathsf{N}^{obs}\lambda(\mathsf{N}^{obs})\Big)^{-1}\Bigg)\Bigg] \\
&\overset{Taylor}{\approx} \quad \mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[g^{*obs}(\boldsymbol{X}^{mis})\Bigg] \\
&\quad\quad +\frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{2f_{X^{obs}}(\overline{X}^{*mis})} \\
&\quad\quad \int \xi^2 K(\xi)d\xi\lambda^2\big(\mathsf{n}(1-p^*)\big) \\
&\quad\quad +o\Big(\lambda^2\big(\mathsf{n}(1-p^*)\big)\Big) + O\Bigg(\Big(\mathsf{n}(1-p^*)\lambda\big(\mathsf{n}(1-p^*)\big)\Big)^{-1}\Bigg),
\end{aligned}
$$

where first order Taylor approximation has been applied twice.

Approximation for additional terms for random imputation strategy are given next. For simulated random imputation expected noise variance is approximated as

$$
\begin{aligned}
v_{\mathsf{n}}^{*obs} \quad &= \quad \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}[\mathbb{V}\mathrm{ar}[\hat{\epsilon}_{\mathbf{x}^{mis}}^{K,R}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]] \\
&\approx \quad v^{*obs} + \mathbb{E}_{\mathsf{N}^{mis},\mathsf{N}^{obs},X^{obs}}\Bigg[(g^{*obs}(X^{obs}) - \mathbb{E}_{\mathbf{D}_{\mathsf{n}^{obs}}^{train}|\mathsf{n}^{mis},\mathsf{n}}[\hat{g}^{obs,K}(X^{obs}|\mathsf{n},\mathsf{n}^{mis}])^2\Bigg].
\end{aligned}
$$

Donor strategy is assumed to behave similarly as random strategy.

K-nearest neighbour (K-nn)

Recall decomposition for $\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}]$. According to Equation (3.12)

$$\begin{aligned}
\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] \;=\; & \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\right)^2\right]}_{\text{expected squared imputation bias=ESIB}} \\
& + \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{expected variance of conditional mean estimate}} \\
& + \underbrace{v_{\mathsf{n}}^{*imp}}_{\text{expected imputation noise}} \\
& + \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[2\mathbb{C}\mathrm{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{cross term}} \\
& + \underbrace{v^{*mis}}_{\text{expected target noise}}.
\end{aligned}$$

where $v_{\mathsf{n}}^{*imp} = \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}[\mathbb{V}\mathrm{ar}[\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]]$. Expected squared imputation bias, ESIB, may be decomposed as (see Equation 3.14)

$$\begin{aligned}
\text{ESIB} \;=\; & \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]}_{\text{variability of conditional mean estimate}} \\
& + (\underbrace{\mu_n^{*imp} - \mu^{*mis}}_{\text{global bias}})^2 + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\
& + \underbrace{2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] - \mu_{\mathsf{n}}^{*imp}\right)\left(\mu_n^{*imp} - g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{cross term}},
\end{aligned}$$

where $\mu_{\mathsf{n}}^{*imp} = \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]$.

K-nn / multivariate covariate
Expected variance for conditional mean estimate is computed as

$$\begin{aligned}
& \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{V}\mathrm{ar}[\hat{g}^N(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}\Bigg] \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\frac{v_{p-1}\mathbb{V}\mathrm{ar}[Y^{obs}|\boldsymbol{X}^{obs}=\mathbf{x}^{mis}]}{k(\mathsf{N}^{obs})}+o\big(k(\mathsf{N}^{obs})^{-1}\big)\Bigg] \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\frac{v_{p-1}\mathbb{V}\mathrm{ar}[Y^{obs}|\boldsymbol{X}^{obs}=\mathbf{x}^{mis}]}{k(\mathsf{N}^{obs})}\Bigg]+o\big(k(\mathsf{N}^{obs})^{-1}\big) \\
= \quad & \frac{v_{p-1}}{\mathbb{E}[k(\mathsf{N}^{obs})]}\mathbb{E}_{\boldsymbol{X}^{mis}}\Bigg[v^{*obs}(\boldsymbol{X}^{mis})\Bigg]+o\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) \\
\overset{Taylor}{\approx} \quad & \frac{v_{p-1}}{\mathbb{E}[k(\mathsf{N}^{obs})]}\Bigg(v^{*obs}(\overline{\boldsymbol{X}}^{*mis})+\frac{1}{2}\mathrm{tr}\Bigg(\Big(\frac{\partial^2}{\partial\mathbf{x}^{mis}\partial\mathbf{x}^{mis}}v^{*obs}(\mathbf{x}^{mis})\Big)_{\mathbf{x}^{mis}=\overline{\boldsymbol{X}}^{*mis}}\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\Bigg)\Bigg) \\
& +o\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big) \\
= \quad & \frac{v_{p-1}}{\mathbb{E}[k(\mathsf{N}^{obs})]}\Bigg(v^{*obs}(\overline{\boldsymbol{X}}^{*mis})+\frac{1}{2}\mathrm{tr}\Bigg(\mathbf{H}_{v^{*obs}}\mathbb{V}\mathrm{ar}[\boldsymbol{X}^{mis}]\Bigg)\Bigg)+o\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big).
\end{aligned}$$

Expected imputation noise and cross term are zero for mean imputation strategy. Therefore one just needs to compute expected squared imputation bias.

As earlier variability of conditional mean estimate is written in implicit form as follows

$$\begin{aligned}
& \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\mathbb{E}[\hat{g}^N(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]\Bigg] \\
= \quad & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[g^{*obs}(\boldsymbol{X}^{mis})+\mathbb{B}\mathrm{ias}[\hat{g}^N(\boldsymbol{X}^{mis})|\mathsf{n}^{mis},\mathsf{n}]\Bigg].
\end{aligned}$$

Global bias is solved by computing μ_{n}^{*imp} which is done as

$$
\begin{aligned}
& \mu_{\mathsf{n}}^{*imp,N} \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis}, \boldsymbol{X}^{mis}|\mathsf{n}} \mathbb{E}[\hat{g}^N(\boldsymbol{X}^{mis})|\mathsf{n}^{mis}, \mathbf{x}^{mis}, \mathsf{n}] \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}} \mathbb{E}_{\boldsymbol{X}^{mis}} \mathbb{E}[\hat{g}^N(\boldsymbol{X}^{mis})|\mathsf{n}^{mis}, \mathbf{x}^{mis}, \mathsf{n}] \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}} \mathbb{E}_{\boldsymbol{X}^{mis}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) + \mathbb{B}\mathrm{ias}[\hat{g}^N(\boldsymbol{X}^{mis})|\mathsf{n}^{mis}, \mathsf{n}] \Big] \\
= \quad & \mathbb{E}_{\boldsymbol{X}^{mis}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) \Big] + \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}} \mathbb{E}_{\boldsymbol{X}^{mis}} \Big[\mathbb{B}\mathrm{ias}[\hat{g}^N(\boldsymbol{X}^{mis})|\mathsf{n}^{mis}, \mathsf{n}] \Big] \\
= \quad & \mathbb{E}_{\boldsymbol{X}^{mis}} \Big[g^{*obs}(\boldsymbol{X}^{mis}) \Big] \\
& + \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}} \mathbb{E}_{\boldsymbol{X}^{mis}} \Bigg[\frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\boldsymbol{X}^{mis}) - g^{*obs}(\boldsymbol{X}^{mis}) Q(f_{\boldsymbol{X}^{obs}})(\boldsymbol{X}^{mis})}{2 f_{\boldsymbol{X}^{obs}}(\boldsymbol{X}^{mis}) \Big(v_{p-1} f(\boldsymbol{X}^{mis}) \Big)^{2/(p-1)}} \\
& * \Big(\frac{k(\mathsf{N}^{obs})}{\mathsf{N}^{obs}} \Big)^{2/(p-1)} + o\bigg(\Big(\frac{k(\mathsf{N}^{obs})}{\mathsf{N}^{obs}} \Big)^{2/(p-1)} \bigg) + O\big(k(\mathsf{N}^{obs})^{-1} \big) \Bigg] \\
\overset{Taylor}{\approx} \quad & g^{*obs}(\overline{\boldsymbol{X}}^{*mis}) + \frac{1}{2} \mathrm{tr}(\mathbf{H}_{g^{*obs}} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis}) \\
& + \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}} \Bigg[\frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis}) - g^{*obs}(\overline{\boldsymbol{X}}^{*mis}) Q(f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis})}{2 f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis}) (v_{p-1} f(\overline{\boldsymbol{X}}^{*mis}))^{2/(p-1)}} \\
& * \Big(\frac{k(\mathsf{N}^{obs})}{\mathsf{N}^{obs}} \Big)^{2/(p-1)} + o\bigg(\Big(\frac{k(\mathsf{N}^{obs})}{\mathsf{N}^{obs}} \Big)^{2/(p-1)} \bigg) + O\big(k(\mathsf{N}^{obs})^{-1} \big) \Bigg] + O(\mathsf{n}^{-1}) \\
\overset{Taylor}{\approx} \quad & g^{*obs}(\overline{\boldsymbol{X}}^{*mis}) + \frac{1}{2} \mathrm{tr}(\mathbf{H}_{g^{*obs}} \boldsymbol{\Sigma}_{\boldsymbol{X}}^{*mis}) \\
& + \Bigg[\frac{Q(g^{*obs} f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis}) - g^{*obs}(\overline{\boldsymbol{X}}^{*mis}) Q(f_{\boldsymbol{X}^{obs}})(\overline{\boldsymbol{X}}^{*mis})}{2 f_{\boldsymbol{X}^{obs}}(\overline{\boldsymbol{X}}^{*mis}) (v_{p-1} f(\overline{\boldsymbol{X}}^{*mis}))^{2/(p-1)}} \Big(\frac{\mathbb{E}[k(\mathsf{N}^{obs})]}{\mathsf{n}(1-p^*)} \Big)^{2/(p-1)} \\
& + o\bigg(\Big(\frac{\mathbb{E}[k(\mathsf{N}^{obs})]}{\mathsf{n}(1-p^*)} \Big)^{2/(p-1)} \bigg) + O\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1} \big) \Bigg] + O(\mathsf{n}^{-1}),
\end{aligned}
$$

where $\mathbf{H}_{g^{*obs}}$ is Hessian of $g^{*obs}(\mathbf{x}^{mis})$ evaluated at $\overline{\boldsymbol{X}}^{*mis}$. Remark that first order Taylor approximation has been applied twice. Therefore result may be quite rough.

Additional terms for random strategies are derived next. For simulated random imputation expected noise variance is approximated as

$$
v_{\mathsf{n}}^{*imp} \approx v^{*obs} + \mathbb{E}_{\mathsf{N}^{mis}, \boldsymbol{X}^{obs}} \Bigg[(g^{*obs}(\boldsymbol{X}^{obs}) - \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis}, \mathsf{n}}[\hat{g}^N(\boldsymbol{X}^{obs}|\mathsf{n}^{mis}, \mathsf{n}])^2 \Bigg].
$$

Further cross term is zero (expectation of noise terms are zero and conditional mean estimate and noise term are conditionally independent given training data).

Expected noise term is assumed to be close to zero for random donor strategy. Therefore additional terms due to noise modelling are roughly same as for simulated random imputation strategy.

K-nn / univariate covariate
Expected variance for conditional mean estimate is computed as:

$$
\begin{aligned}
&\mathbb{E}_{\mathsf{N}^{mis},X^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}^N(X^{mis})|x^{mis},\mathsf{n}^{mis},\mathsf{n}\right]\\
=\;&\mathbb{E}_{\mathsf{N}^{mis},X^{mis}|\mathsf{n}}\left[\frac{2v^{*obs}(x^{mis})}{k(\mathsf{N}^{obs})}+o\big(k(\mathsf{N}^{obs})^{-1}\big)\right]\\
\approx\;&\frac{v_{p-1}}{k(\mathsf{n}(1-p^*))}\left(v^{*obs}(\overline{X}^{*mis})+\frac{1}{2}\Big(\frac{\partial^2}{\partial x^{mis}\partial x^{mis}}v^{*obs}(x^{mis})\Big)_{x^{mis}=\overline{X}^{*mis}}\mathbb{V}\mathrm{ar}[X^{mis}]\right),
\end{aligned}
$$

where $p-1=1$.

Expected imputation noise variance (term C) and cross term (D) are zero for mean strategy.

For expected squared bias one needs to compute variability of conditional mean estimate and global bias. Variability of conditional mean estimate is written in implicit form as

$$
\begin{aligned}
&\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},X^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}^N(X^{mis})|x^{mis},\mathsf{n}^{mis},\mathsf{n}]\right]\\
&=\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},X^{mis}|\mathsf{n}}\Big[g^{*obs}(X^{mis})+\mathbb{B}\mathrm{ias}[\hat{g}^N(X^{mis})|\mathsf{n}^{mis},\mathsf{n}]\Big].
\end{aligned}
$$

Further, global bias is solved by computing μ_{n}^{*imp} which is done as

$$
\begin{aligned}
& \mu_{\mathsf{n}}^{*imp,N} \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis},X^{mis}|\mathsf{n}}\mathbb{E}[\hat{g}^{N}(X^{mis})|\mathsf{n}^{mis},\mathbf{x}^{mis},\mathsf{n}] \\
= \quad & \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\mathbb{E}_{X^{mis}}\Bigg[g^{*obs}(X^{mis}) + \mathbb{B}\mathrm{ias}[\hat{g}^{N}(X^{mis})|\mathsf{n}^{mis},\mathsf{n}]\Bigg] \\
= \quad & \mathbb{E}_{X^{mis}}\Bigg[g^{*obs}(X^{mis})\Bigg] + \mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\mathbb{E}_{X^{mis}}\Bigg[\mathbb{B}\mathrm{ias}[\hat{g}^{N}(X^{mis})|\mathsf{n}^{mis},\mathsf{n}]\Bigg] \\
= \quad & \mathbb{E}_{X^{mis}}\Bigg[g^{*obs}(X^{mis})\Bigg] \\
& +\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[\mathbb{E}_{X^{mis}}\Bigg[\frac{(g^{*obs}f_{X^{obs}})''(X^{mis}) - g^{*obs}(X^{mis})f''_{X^{obs}}(X^{mis})}{24f^{3}_{X^{obs}}(X^{mis})}\big(k(\mathsf{N}^{obs})/(\mathsf{N}^{obs})\big)^{2} \\
& +o\Big((\frac{k(\mathsf{N}^{obs})}{\mathsf{N}^{obs}})^{2}\Big) + O\big(k(\mathsf{N}^{obs})^{-1}\big)\Bigg]\Bigg] \\
\overset{Taylor}{\approx} \quad & \mathbb{E}_{X^{mis}}\Bigg[g^{*obs}(X^{mis})\Bigg] \\
& +\mathbb{E}_{\mathsf{N}^{mis}|\mathsf{n}}\Bigg[\frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{24f^{3}_{X^{obs}}(\overline{X}^{*mis})}\big(k(\mathsf{N}^{obs})/\mathsf{N}^{obs}\big)^{2} \\
& +o\Bigg(\Big(\frac{k(\mathsf{N}^{obs})}{\mathsf{N}^{obs}}\Big)^{2}\Bigg) + O\big(k(\mathsf{N}^{obs})^{-1}\big)\Bigg] \\
\overset{Taylor}{\approx} \quad & \mathbb{E}_{X^{mis}}\Bigg[g^{*obs}(X^{mis})\Bigg] \\
& +\Bigg[\frac{(g^{*obs}f_{X^{obs}})''(\overline{X}^{*mis}) - g^{*obs}(\overline{X}^{*mis})f''_{X^{obs}}(\overline{X}^{*mis})}{24f^{3}_{X^{obs}}(\overline{X}^{*mis})}(\mathbb{E}[k(\mathsf{N}^{obs})]/(\mathsf{n}(1-p^{*})))^{2} \\
& +o\Bigg(\Big(\frac{\mathbb{E}[k(\mathsf{N}^{obs})]}{\mathsf{n}(1-p^{*})}\Big)^{2}\Bigg) + O\big(\mathbb{E}[k(\mathsf{N}^{obs})]^{-1}\big)\Bigg],
\end{aligned}
$$

where first order Taylor approximation has been applied twice and impact of technical conditionalizers has been assumed to be neglible.

Additional terms for random strategies are derived next. For simulated random imputation approximation for expected noise variance is

$$
v_{\mathsf{n}}^{*imp} \approx v^{*obs} + \mathbb{E}_{\mathsf{N}^{mis},X^{obs}}\Bigg[(g^{*obs}(X^{obs}) - \mathbb{E}_{\mathbf{D}^{train}|\mathsf{n}^{mis},\mathsf{n}}[\hat{g}^{N}(X^{obs}|\mathsf{n}^{mis},\mathsf{n}])^{2}\Bigg].
$$

Further, cross term is zero (expectation of noise terms are zero and conditional mean estimate and noise term are conditionally independent given training data).

Expected noise term is assumed to be close to zero for random donor strategy as earlier. Therefore additional terms due to noise modelling are roughly same as for simulated random imputation strategy.

A5.3 MSE re-derivation for nonlinear simulation example

To simplify numerical computation of terms for decomposition of mean squared error $\mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}]$ order of integration is changed and one integration level is 'reduced' (by combining two integrations together).

First, recall that

$$\begin{aligned}
\mathrm{mse}(\hat{Y}|\mathbf{x}_0, \mathsf{n}) &= \mathbb{E}_{\hat{Y},Y|\mathbf{x}_0,\mathsf{n}}[\hat{Y}^2_{|\mathbf{x}_0} - 2\hat{Y}_{|\mathbf{x}_0}Y_{|\mathbf{x}} + Y^2_{|\mathbf{x}_0}] \\
&= \Big(\underbrace{\mathbb{E}[\hat{g}(\mathbf{x}_0)|\mathbf{x}_0, \mathsf{n}] - g^{*mis}(\mathbf{x}_0)}_{\text{imputation bias at } \mathrm{x}_0}\Big)^2 + \underbrace{\qquad \mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0,n}] \qquad}_{\text{imputation variance at } \mathrm{x}_0} \\
&\quad + \underbrace{\qquad \mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}_0}] \qquad}_{\mathrm{v}^{*\mathrm{mis}}(\mathbf{x}_0),\ \text{target noise at } \mathrm{x}_0},
\end{aligned}$$

Further, imputation variance can be decomposed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{Y}_{|\mathbf{x}_0,n}] &= \mathbb{V}\mathrm{ar}[\hat{g}(\mathbf{x}_0) + \hat{\epsilon}_{\mathbf{x}_0}|\mathbf{x}_0, \mathsf{n}] \\
&= \underbrace{\qquad \mathbb{V}\mathrm{ar}[\hat{g}(\mathbf{x}_0)|\mathbf{x}_0, \mathsf{n}] \qquad}_{\text{variance of conditional mean estimate}} + \underbrace{\mathbb{V}\mathrm{ar}[\hat{\epsilon}_{\mathbf{x}_0}|\mathbf{x}_0, \mathsf{n}]}_{\text{imputation noise, } \hat{\mathrm{v}}(\mathbf{x}_0)} \\
&\quad + \underbrace{2\mathbb{C}\mathrm{ov}[\hat{g}(\mathbf{x}_0), \hat{\epsilon}_{\mathbf{x}_0}|\mathbf{x}_0, \mathsf{n}]}_{\text{cross term}}.
\end{aligned}$$

One should note that the cross term above is zero for k-nn and random imputation strategy which is used in this example.

Decomposition at population level can be written as

$$
\begin{aligned}
\mathbb{E}[m\hat{s}e(Y^{comp})|\mathsf{n}] &= \mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\text{mse}(\boldsymbol{X}^{mis}|\mathsf{n})\Big]\\
&= \mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\Big(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\Big)^2\Bigg]\\
&\quad +\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{V}\text{ar}\big[\hat{Y}_{|\mathbf{x}^{mis},\mathsf{n}}\big]\Big] + \mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Big[v^{*mis}(\boldsymbol{X}^{mis})\Big]\\
&= \mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\Big(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\Big)^2\Bigg]\\
&\quad +\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{V}\text{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}] + \mathbb{V}\text{ar}[\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}]\\
&\quad +2\mathbb{C}\text{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}]\Big]\\
&\quad +v^{*mis}\\
&= \underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Bigg[\Big(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}] - g^{*mis}(\boldsymbol{X}^{mis})\Big)^2\Bigg]}_{\text{expected squared imputation bias (ESIB)}}\\
&\quad + \underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\mathbb{V}\text{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis},\mathsf{n}]\Big]}_{\text{expected variance of conditional mean estimate}} + \underbrace{v_{\mathsf{n}}^{*imp}}_{\text{expected imputation noise}}\\
&\quad + \underbrace{\mathbb{E}_{\boldsymbol{X}^{mis}|\mathsf{n}}\Big[2\mathbb{C}\text{ov}[\hat{g}(\boldsymbol{X}^{mis}),\hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis},\mathsf{n}]\Big]}_{\text{cross term}}\\
&\quad + \underbrace{v^{*mis}}_{\text{expected target noise}}.
\end{aligned}
$$

First of all, one should note that the cross term is zero in this example. Most interesting terms are expected squared bias, expected variance of conditional mean estimate, expected imputation noise variance and expected target variance.

A5.4 Verification of Mack's conditions for ordinary k-nn regression

Mack's paper [72] contains bias and variance results for generalized k-nearest neighbour regression. Ordinary k-nearest neighbour regression estimate may be written using Mack's generalized form by using weight function

$$
w(\mathbf{v}) = \begin{cases} 1/v_p, & \text{when } ||\mathbf{v}|| < 1,\\ 0, & \text{otherwise,} \end{cases}
$$

where $v_p = \pi^{p/2}/\Gamma((p+2)/2)$ equals to volume of unit ball in $\mathbb{R}^p$. Next weight function requirements of Mack's theorems 1 and 2 are verified.

Theorem 1 requires following conditions for $w(\mathbf{v})$:

1a) $\int w(\mathbf{v})d\mathbf{v} = 1$, $w(\mathbf{v}) = 0$ for $||\mathbf{v}|| \geq 1$

1b) $\int ||\mathbf{v}||^2 |w(\mathbf{v})| d\mathbf{v} < \infty$

1c) $\int \mathbf{v}_\alpha w(\mathbf{v}) d\mathbf{v} = 0$ for $\alpha = 1, \ldots, p$, where $\mathbf{v}_\alpha$ denotes α:th component of $\mathbf{v}$.

Now

$$\int w(\mathbf{v})d\mathbf{v} = \int_{||\mathbf{v}||<1} 1/v_p d\mathbf{v} = 1,$$

and by definition of $w(\mathbf{v})$ it holds that $w(\mathbf{v}) = 0$ for $||\mathbf{v}|| \geq 1$. Therefore 1a) holds. Further

$$\begin{aligned}\int ||\mathbf{v}||^2 |w(\mathbf{v})| d\mathbf{v} &= \int_{||\mathbf{v}||<1} ||\mathbf{v}||^2 (1/v_p) d\mathbf{v} \\ &< \int_{||\mathbf{v}||<1} 1^2 (1/v_p) d\mathbf{v} = (1/v_p) \int_{||\mathbf{v}||<1} d\mathbf{v} = 1 < \infty,\end{aligned}$$

Thus 1b) also holds. Finally

$$\begin{aligned}\int \mathbf{v}_\alpha w(\mathbf{v}) d\mathbf{v} &= \int_{||\mathbf{v}||<1} \mathbf{v}_\alpha (1/v_p) d\mathbf{v} \\ &= (1/v_p) \int_{||\mathbf{v}||<1} \mathbf{v}_\alpha d\mathbf{v} = 0.\end{aligned}$$

Theorem 2 requires following additional condition for $w(\mathbf{v})$:

2a) $\int |\mathbf{v}_\alpha| |w(\mathbf{v})| d\mathbf{v} < \infty$.

Now

$$\begin{aligned}\int |\mathbf{v}_\alpha| |w(\mathbf{v})| d\mathbf{v} &= \int_{||\mathbf{v}||<1} |\mathbf{v}_\alpha| (1/v_p) d\mathbf{v} \\ &< \int_{||\mathbf{v}||<1} |1| (1/v_p) d\mathbf{v} = (1/v_p) \int_{||\mathbf{v}||<1} 1 d\mathbf{v} = 1 < \infty.\end{aligned}$$

Therefore requirement 2a) is satisfied.

To summarize, all weight function requirements of Mack's theorems 1 and 2 are satisfied by ordinary k-nn weight function $w(\mathbf{v})$. Now one can utilize Mack's bias and variance results. It is somewhat difficult to simplify bias result but variance result may be simplified as follows:

$$\begin{aligned}\int w^2(v) dv &= \int_{||v||<1} (1/v_p)^2 dv \\ &= (1/v_p)^2 \int_{||v||<1} 1 dv \\ &= (1/v_p)^2 v_p \\ &= 1/v_p.\end{aligned}$$

Substituting above result into Mack's Equation (12) (by replacing $v_p = c$) gives claimed variance.

Appendix for Chapter 6

In this appendix justifications of approximations and consequences which were introduced in Chapter 6 are given.

A6.1 Cell methods / moments

Here results for moment estimators based on cell imputation are derived. The following decomposition is used for mean estimator:

$$\begin{aligned}\hat{\mu}^{comp} &= \frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathsf{N}^{mis}\hat{\mu}^{imp}) \\ &= \frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \sum_{i=1}^{\mathsf{n}_c} \mathsf{N}_i^{mis}\hat{\mu}_i^{imp}).\end{aligned}$$

Approximation 6.1

The bias of $\hat{\mu}^{comp}$ for n observations, with fixed training data, and fixed imputation model may be approximated as

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}_2] \approx\; & \underbrace{\frac{1}{\mathsf{n}}(\sum_i \mathbb{E}[\mathsf{N}_i^{mis}|\mathcal{Q}_2]\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] - \mathsf{n}^{mis}\mu^{*mis})}_{\text{bias due to imputation method}} \\ & \underbrace{+\frac{1}{\mathsf{n}}(\mathsf{n}^{obs}\mu^{obs} + \mathsf{n}^{mis}\mu^{*mis}) - \mu^*}_{\text{finite sample estimation error}},\end{aligned}$$

where $\mathbb{E}[\mathsf{N}_i^{mis}|\mathcal{Q}_2] = \Pr\Big(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big)\mathsf{n}^{mis}$ and

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] = \begin{cases} \mu_i^{obs} & : \text{C/CJ/T/TJ(S = M/R/D)} \\ \mathbb{E}[\hat{\mu}_i^s|\mathcal{Q}_2] \approx \frac{\sum_l h_{i,l}(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])\mu_l^{obs}}{\sum_l h_{i,l}(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])} & : \text{T/TJ(S = M}^{\text{s}}\text{/R}^{\text{s}}\text{)},\end{cases}$$

Justification: Bias is approximated using decomposition of $\hat{\mu}^{comp}$ and first order Taylor approximation as

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathcal{Q}_2] &= \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_2] + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mu^{*mis} - \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mu^{*mis} - \mu^* \\
&= \mathbb{E}[\frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \sum_{i=1}^{\mathsf{n}_c} \mathsf{N}_i^{mis}\hat{\mu}_i^{imp})|\mathcal{Q}_2] + \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mu^{*mis} - \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mu^{*mis} - \mu^* \\
&= \frac{1}{\mathsf{n}}(\sum_{i=1}^{\mathsf{n}_c}\mathbb{E}[\mathsf{N}_i^{mis}\hat{\mu}_i^{imp}|\mathcal{Q}_2] - \mathsf{n}^{mis}\mu^{*mis}) \\
&\quad +\frac{1}{\mathsf{n}}(\mathsf{n}^{obs}\mu^{obs} + \mathsf{n}^{mis}\mu^{*mis}) - \mu^* \\
&\overset{Taylor}{\approx} \underbrace{\frac{1}{\mathsf{n}}(\sum_{i=1}^{\mathsf{n}_c}\mathbb{E}[\mathsf{N}_i^{mis}|\mathcal{Q}_2]\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] - \mathsf{n}^{mis}\mu^{*mis})}_{\text{bias due to imputation method}} \\
&\quad \underbrace{+\frac{1}{\mathsf{n}}(\mathsf{n}^{obs}\mu^{obs} + \mathsf{n}^{mis}\mu^{*mis}) - \mu^*}_{\text{finite sample estimation error}},
\end{aligned}$$

where term $\mathbb{E}[\frac{1}{\mathsf{n}}\mathsf{N}^{mis}\hat{\mu}^{mis}|\mathcal{Q}_2] = \frac{\mathsf{n}^{mis}}{\mathsf{n}}\mu^{*mis}$ has been added and subtracted in order to separate imputation and sample estimation errors from each other.

Now one needs to solve term $\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2]$, the expectation of mean of imputed values within cell i.

For non-smoothed imputation method with any strategy (mean, random imputation, or donor) the expectation equals to

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] = \mu_i^{obs}.$$

For mean strategy this is trivial as $\hat{\mu}_i^{imp} = \hat{\mu}_i^{obs}$. For random strategy

$$\hat{\mu}_i^{imp} = \hat{\mu}_i^{obs} + \frac{1}{\mathsf{N}_i^{obs}}\sum_{j=1}^{\mathsf{N}_i^{obs}}\hat{\epsilon}_{j,i},$$

where $\hat{\epsilon}_{j,i}$ is j:th noise term in cell i. Note that expectation of modelled noise term within any cell is zero. Therefore

$$\mathbb{E}[\hat{\mu}_i^{imp,R}|\mathcal{Q}_2] = \mu_i^{obs}.$$

Further, for donor strategy expectation of prediction of missing data value in any cell equals to mean of observed Y values within the cell. This result follows by applying basic sampling theory result ('finite population' is centered observations within the cell).

For smoothed imputation method with mean or random strategy the expectation is computed using first order Taylor approximation as

$$\begin{aligned}
\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] &= \mathbb{E}[\hat{\mu}_i^{s}|\mathcal{Q}_2] = \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}N_l\hat{\mu}^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}N_l}|\mathcal{Q}_2] \\
&\overset{Taylor}{\approx} \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]\mathbb{E}[\hat{\mu}_l^{obs}|\mathcal{Q}_2]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]} = \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])\mu_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])}.
\end{aligned}$$

Approximation 6.2

The bias of first moment $\hat{\mu}^{comp}$ given n observations can be approximated as

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathsf{n}] &\approx p^*(\underbrace{\sum_{i=1}^{\mathsf{n}_c}\Pr\Big(\hat{b}(\boldsymbol{X}^{mis})=i|\mathsf{n}\Big)\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}]-\mu^{*mis}}_{\text{(weighted) difference between mean of imputed and missing Y values}})\\ &\quad+\underbrace{O(\mathsf{n}^{-1})}_{\text{approximation error}},\end{aligned}$$

where

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}]=\begin{cases}\mathbb{E}[\hat{\mu}_i^{obs}|\mathsf{n}]=\mu_i^{*obs} & :\text{C/CJ/T/TJ(S = M/R/D)},\\[2ex] \mathbb{E}[\hat{\mu}_i^{s}|\mathsf{n}]\approx\frac{\sum_{l=1}^{\mathsf{n}_c}h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}]\mu_l^{*obs}}{\sum_{l=1}^{\mathsf{n}_c}h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}]} & :\text{T/TJ(S} = \text{M}^{\text{s}}/\text{R}^{\text{s}})\end{cases}$$

where μ_l^{*obs} is expectation of observed Y values in l:th cell and in which

$$\begin{aligned}\mathbb{E}\Big[\mathsf{N}_l|\mathsf{n}\Big] &= \mathbb{E}\Big[\mathsf{N}_l^{obs}|\mathsf{n}\Big]+\mathbb{E}\Big[\mathsf{N}_l^{mis}|\mathsf{n}\Big]\\ &\approx \underbrace{\mathsf{n}(1-p^*)\Pr\Big(\hat{b}\big((Y^{obs},\boldsymbol{X}^{obs})^T\big)=l|\mathsf{n}\Big)}_{\text{expected number of complete observations}}\\ &\quad+\underbrace{\mathsf{n}p^*\Pr\Big(\hat{b}(\boldsymbol{X}^{mis})=l|\mathsf{n}\Big)}_{\text{expected number of incomplete observations}}.\end{aligned}$$

Justification: applying decomposition of $\hat{\mu}^{comp}$, decomposition of μ^*, and first order Taylor approximation yield to

$$\begin{aligned}\mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathsf{n}] &= \mathbb{E}[\frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs}+\sum_{i=1}^{\mathsf{n}_c}\mathsf{N}_i^{mis}\hat{\mu}_i^{imp})|\mathsf{n}]-\mu^*\\ &= \mathbb{E}[\frac{\mathsf{N}^{obs}}{\mathsf{n}}\hat{\mu}^{obs}]-(1-p^*)\mu^{*obs}+\mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\frac{\mathsf{N}_i^{mis}}{\mathsf{N}^{mis}}\hat{\mu}_i^{imp}]-p^*\mu^{*imp}\\ &\overset{Taylor}{\approx} p^*\sum_{i=1}^{\mathsf{n}_c}\Pr\Big(\hat{b}(\boldsymbol{X}^{mis})=i|\mathsf{n}\Big)\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}]-p^*\mu^{*imp}+O(\mathsf{n}^{-1})\\ &= p^*(\sum_{i=1}^{\mathsf{n}_c}\Pr\Big(\hat{b}(\boldsymbol{X}^{mis})=i|\mathsf{n}\Big)\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}]-\mu^{*imp})+O(\mathsf{n}^{-1}).\end{aligned}$$

Approximation of expectation $\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}]$ is quite straightforward. For non-smoothed imputation method with any strategy (mean, random imputation, or donor) the expectation equals to

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}]=\mu_i^{*obs}.$$

This result follows directly from the fact that $\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] = \mu_i^{obs}$ for all three imputation strategies. See justification for approximation 6.1 for details. For smoothed imputation method with mean or random strategy the expectation is computed using first order Taylor approximation as

$$\begin{aligned}
\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] &= \mathbb{E}[\hat{\mu}_i^{s}|\mathsf{n}] = \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l \hat{\mu}^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l}|\mathsf{n}] \\
&\overset{Taylor}{\approx} \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathsf{n}] \mathbb{E}[\hat{\mu}_l^{obs}|\mathsf{n}]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathsf{n}]} = \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathsf{n}] \mu_l^{*obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathsf{n}]}.
\end{aligned}$$

Approximation 6.3

The variance of $\hat{\mu}^{comp}$ for n observations, fixed training data, and fixed imputation model can be approximated as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2] \approx\; & \underbrace{\frac{1}{\mathsf{n}^2} \mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big] \mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]}_{\text{due to randomness of number of missing Y values within cells}} \\
& + \underbrace{\frac{\mathsf{n}^{mis}}{\mathsf{n}^2} \sum_{i=1}^{\mathsf{n}_c} \Pr\Big(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big) \mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2]}_{\text{A: due to modelled noise}} \;,
\end{aligned}$$

where $\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] = (\mathbb{E}[\hat{\mu}_1^{imp}|\mathcal{Q}_2], \ldots, \mathbb{E}[\hat{\mu}_{\mathsf{n}_c}^{imp}|\mathcal{Q}_2])^T$ in which

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] = \begin{cases} \mathbb{E}[\hat{\mu}_i^{obs}|\mathcal{Q}_2] = \mu_i^{obs} & : \mathrm{C/T/CJ/TJ(S = M/R/D)} \quad , \\ \\ \mathbb{E}[\hat{\mu}_i^{s}|\mathcal{Q}_2] \approx \frac{\sum_l h_{i,l}(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])\mu_l^{obs}}{\sum_l h_{i,l}(\mathsf{n}_l^{obs} + \mathbb{E}[\mathsf{N}_l^{mis}|\mathcal{Q}_2])} & : \mathrm{T/TJ(S = M^s/R^s)} \quad . \end{cases}$$

Term $\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2]$ depends on cell method and on imputation strategy $\hat{\epsilon}^S$ as follows:

$$\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2] = \begin{cases} 0 & : \mathrm{S = M/M^s} \quad , \\ \\ \mathbb{E}[\hat{\tau}_i^{obs}|\mathcal{Q}_2] = \tau_i^{obs} & : \mathrm{C/CJ/T/TJ(S = R)} \quad , \\ \tau_i^{obs}(1 - \frac{1}{\mathsf{n}_i^{obs}}) & : \mathrm{C/CJ/T/TJ(S = D)} \quad , \\ \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathcal{Q}_2] = \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l \hat{\tau}_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l}|\mathcal{Q}_2] & : \mathrm{T(S = R^s)} \quad , \\ \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2] \tau_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2]} & \\ \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathcal{Q}_2] = \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l \hat{\tau}_l^{w}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l}|\mathcal{Q}_2] & : \mathrm{TJ(S = R^s)} \quad , \\ \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2] \mathbb{E}[\hat{\tau}_l^{w}|\mathcal{Q}_2]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} \mathbb{E}[N_l|\mathcal{Q}_2]} & \end{cases}$$

in which

$$\begin{aligned}
\hat{\tau}_i^{obs} &= \frac{1}{\mathsf{N}_i^{obs}} \sum_{j=1}^{\mathsf{N}_i^{obs}} (Y_{j,i}^{obs} - \hat{\mu}_i^{obs})^2 \\
\mathbb{E}[\hat{\tau}_l^{w}|\mathcal{Q}_2] &= \mathbb{E}[\frac{1}{\mathsf{N}_l^{obs}} \sum_{j=1}^{\mathsf{N}_l^{obs}} (Y_{j,l}^{obs} - \hat{\mu}_l^{s})^2|\mathcal{Q}_2] \approx \frac{1}{\mathsf{n}_l^{obs}} \sum_{j=1}^{\mathsf{n}_l^{obs}} (y_{j,l}^{obs} - \mathbb{E}[\hat{\mu}_l^{s}|\mathcal{Q}_2])^2,
\end{aligned}$$

where $Y^{obs}_{j,i}$ is the j:th random observation of Y^{obs} in i:th cell and

$$\mathbb{E}[\hat{\mu}^s_i|\mathcal{Q}_2] \quad \approx \quad \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\big(\mathsf{n}^{obs}_l + \mathbb{E}[\mathsf{N}^{mis}_l|\mathcal{Q}_2]\big)\mu^{obs}}{\sum_l h_{i,l}\big(\mathsf{n}^{obs}_l + \mathbb{E}[\mathsf{N}^{mis}_l|\mathcal{Q}_2]\big)}.$$

Justification: let $\hat{\mu}^{imp,M}_i$ be mean estimator for imputations of mean strategy (unsmoothed or smoothed) in cell i. Approximation of variance for random strategy (unsmoothed or smoothed) is derived as

$$\begin{aligned}
& \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp,R}|\mathcal{Q}_2] \\
= \quad & \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \sum_{i=1}^{\mathsf{n}_c} \mathsf{N}^{mis}_i \hat{\mu}^{imp,M}_i + \sum_{i=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}^{mis}_i} \hat{\epsilon}_{j,i})|\mathcal{Q}_2] \\
= \quad & \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c} \mathsf{N}^{mis}_i \hat{\mu}^{imp,M}_i|\mathcal{Q}_2] + \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}^{mis}_i} \hat{\epsilon}_{j,i}|\mathcal{Q}_2] \\
\approx \quad & \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c} \mathsf{N}^{mis}_i \mathbb{E}[\hat{\mu}^{imp,M}_i|\mathcal{Q}_2]|\mathcal{Q}_2] + \mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}^{mis}_i} \hat{\epsilon}_{j,i}|\mathcal{Q}_2] \\
= \quad & \frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] \\
& + \mathbb{E}[\mathbb{V}\mathrm{ar}[\frac{1}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}^{mis}_i} \hat{\epsilon}_{j,i}|\mathcal{Q}_2, \mathsf{N}^{mis}_1, \ldots, \mathsf{N}^{mis}_{\mathsf{n}_c}]|\mathcal{Q}_2] \\
& \underbrace{+\, \mathbb{V}\mathrm{ar}[\mathbb{E}[\frac{1}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}^{mis}_i} \hat{\epsilon}_{j,i}|\mathcal{Q}_2, \mathsf{N}^{mis}_1, \ldots, \mathsf{N}^{mis}_{\mathsf{n}_c}]|\mathcal{Q}_2]}_{=0} \\
= \quad & \frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] \\
& + \mathbb{E}[\frac{1}{\mathsf{n}^2}\sum_{i=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}^{mis}_i} \mathbb{V}\mathrm{ar}[\hat{\epsilon}_{j,i}|\mathcal{Q}_2, \mathsf{N}^{mis}_1, \ldots, \mathsf{N}^{mis}_{\mathsf{n}_c}]|\mathcal{Q}_2] \\
= \quad & \frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] + \frac{1}{\mathsf{n}^2}\sum_{i=1}^{\mathsf{n}_c}\mathbb{E}[\mathsf{N}^{mis}_i \hat{\tau}^{imp}_i|\mathcal{Q}_2] \\
\overset{Taylor}{\approx} \quad & \frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] + \frac{1}{\mathsf{n}^2}\sum_{i=1}^{\mathsf{n}_c}\mathbb{E}[\mathsf{N}^{mis}_i|\mathcal{Q}_2]\mathbb{E}[\hat{\tau}^{imp}_i|\mathcal{Q}_2] \\
= \quad & \frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] \\
& + \frac{\mathsf{n}^{mis}}{\mathsf{n}^2}\sum_{i=1}^{\mathsf{n}_c}\mathrm{Pr}\Big(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big)\mathbb{E}[\hat{\tau}^{imp}_i|\mathcal{Q}_2].
\end{aligned}$$

Elements of $\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] = (\mathbb{E}[\hat{\mu}^{imp}_1|\mathcal{Q}_2], \ldots, \mathbb{E}[\hat{\mu}^{imp}_{\mathsf{n}_c}|\mathcal{Q}_2])^T$ are already derived in justification for approximation 6.1.

For mean strategy (with or without smoothing) expectation $\mathbb{E}[\hat{\tau}^{imp}_i|\mathcal{Q}_2]$ is zero because noise is modelled. For unsmoothed random strategy the expectation equals to $\mathbb{E}[\hat{\tau}_i^{imp,R}|\mathcal{Q}_2] = \mathbb{E}[\hat{\tau}_i^{obs}|\mathcal{Q}_2] = \tau_i^{obs}$. Basic result from sampling theory gives that for donor strategy (simple random sampling with replacements) holds $\mathbb{E}[\hat{\tau}_i^{imp,R}|\mathcal{Q}_2] = \tau_i^{obs}(1 - \frac{1}{\mathsf{n}_i^{obs}})$.

The expectation for standard $\boldsymbol{X}$-clustering with smoothed random strategy is derived using first order Taylor approximation as

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2] &= \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathcal{Q}_2] \\
&= \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l \hat{\tau}_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l}|\mathcal{Q}_2] \overset{Taylor}{\approx} \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]\tau_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]}.
\end{aligned}$$

For joint $(Y, \boldsymbol{X})$ clustering the smoothed random strategy yields to following expectation

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2] &= \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathcal{Q}_2] \\
&= \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l \hat{\tau}_l^{w}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l} N_l}|\mathcal{Q}_2] \overset{Taylor}{\approx} \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]\mathbb{E}[\hat{\tau}_l^{w}|\mathcal{Q}_2]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]} \\
&= \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]\mathbb{E}[\frac{1}{\mathsf{N}_l^{obs}}\sum_{j=1}^{\mathsf{N}_l^{obs}}(Y_{j,l}^{obs} - \hat{\mu}_l^s)^2|\mathcal{Q}_2]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]} \\
&\overset{Taylor}{\approx} \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]\frac{1}{\mathsf{n}_l^{obs}}\sum_{j=1}^{\mathsf{n}_l^{obs}}(y_{j,l}^{obs} - \mathbb{E}[\hat{\mu}_l^s|\mathcal{Q}_2])^2}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]}.
\end{aligned}$$

Approximation 6.4

The variance of first moment $\hat{\mu}^{comp}$ given n observations can be approximated as

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathsf{n}] \\
\approx\quad & \underbrace{\mathbb{E}\left[\frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T \mathbb{V}\mathrm{ar}\left[\boldsymbol{N}^{mis}|\mathcal{Q}_2\right]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]\middle|\mathsf{n}\right]}_{\text{due to randomness of test data and classification of incomplete observations}} \\
+\quad & \underbrace{\mathbb{V}\mathrm{ar}\left[\frac{1}{\mathsf{n}}\left(\mathsf{N}^{obs}\hat{\mu}^{obs} + \mathsf{N}^{mis}\sum_{i=1}^{\mathsf{n}_c}\Pr(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2)\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2]\right)\middle|\mathsf{n}\right]}_{\text{due to randomness of training data,imputation model, and number of missing Y values}} \\
+\quad & \underbrace{\frac{p^*}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\Pr(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n})\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]\quad}_{\text{variance due to modelled noise}},
\end{aligned}$$

where $\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2] = (\mathbb{E}[\hat{\mu}_1^{imp}|\mathcal{Q}_2], \ldots, \mathbb{E}[\hat{\mu}_{\mathsf{n}_c}^{imp}|\mathcal{Q}_2])^T$ in which

$$\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2] = \begin{cases} \mu_i^{*obs} & \\ & : \mathrm{C/CJ/T/TJ(S = M/R/D)}, \\ \mathbb{E}[\hat{\mu}_i^s|\mathcal{Q}_2] \approx \frac{1}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\Pr(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2)\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]\hat{\mu}_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_2]} & \\ & : \mathrm{T/TJ(S = M^s/R^s)}, \end{cases}$$

and terms $\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]$ depend on cell method and on imputation strategy $\hat{\epsilon}^S$ as follows:

$$\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}] = \begin{cases} 0 & : \mathrm{S} = \mathrm{M}/\mathrm{M}^{\mathrm{s}}, \\ \mathbb{E}[\hat{\tau}_i^{obs}|\mathsf{n}] \approx \tau_i^{*obs} & : \mathrm{C/CJ/T/TJ}(\mathrm{S} = \mathrm{R}), \\ \approx \tau_i^{*obs}\left(1 - \frac{1}{\mathsf{n}(1-p^*)\Pr\left(\hat{b}\left(\boldsymbol{X}^{obs}\right)=i|\mathsf{n}\right)}\right) & : \mathrm{C/T}(\mathrm{S} = \mathrm{D}), \\ \approx \tau_i^{*obs}\left(1 - \frac{1}{\mathsf{n}(1-p^*)\Pr\left(\hat{b}\left((Y^{obs},\boldsymbol{X}^{obs})^T\right)=i|\mathsf{n}\right)}\right) & : \mathrm{CJ/TJ}(\mathrm{S} = \mathrm{D}), \\ \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathsf{n}] \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]\tau_l^{*obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]} & : \mathrm{T}(\mathrm{S} = \mathrm{R}^{\mathrm{s}}),\ \text{and} \\ \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathsf{n}] \approx \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]\mathbb{E}[\hat{\tau}_l|\mathcal{Q}_1]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]} & : \mathrm{TJ}(\mathrm{S} = \mathrm{R}^{\mathrm{s}}). \end{cases}$$

in which

$$\hat{\tau}_l = \frac{1}{\mathsf{N}_l^{obs}} \sum_{j=1}^{\mathsf{N}_l^{obs}} (Y_{j,l}^{obs} - \hat{\mu}_l^s)^2,$$

where $Y_{j,l}^{obs}$ is j:th (random) observation of Y^{obs} in l:th cell.

Justification: recall chain rule of variance which states that

$$\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathsf{n}] \;=\; \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2]|\mathsf{n}] + \mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_2]|\mathsf{n}],$$

where the conditional variance in the first term has already been derived in the justification for approximation 6.3. Thus

$$\begin{aligned} \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathcal{Q}_2]|\mathsf{n}] &\overset{approximation\ 6.3}{\approx} \mathbb{E}[\frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T\mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]|\mathsf{n}] \\ &\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}^2}\sum_{i=1}^{\mathsf{n}_c}\Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big)\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2]|\mathsf{n}] \\ &\overset{Taylor}{\approx} \mathbb{E}[\frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T\mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]|\mathsf{n}] \\ &\quad +\frac{p^*}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\mathbb{E}[\Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2\Big)|\mathsf{n}]\mathbb{E}[\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2]|\mathsf{n}] \\ &= \mathbb{E}[\frac{1}{\mathsf{n}^2}\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]^T\mathbb{V}\mathrm{ar}\Big[\boldsymbol{N}^{mis}|\mathcal{Q}_2\Big]\mathbb{E}[\hat{\boldsymbol{\mu}}^{imp}|\mathcal{Q}_2]|\mathsf{n}] \\ &\quad +\frac{p^*}{\mathsf{n}}\sum_{i=1}^{\mathsf{n}_c}\Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}], \end{aligned}$$

where $\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]$ is computed almost like $\mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_2]$ in justification for 6.3. First order Taylor approximation is applied to compute results for donor and smoothed random strategies).

For mean (smoothed or unsmoothed) strategies the $\mathbb{E}[\hat{\tau}_i^{imp}|\mathsf{n}]$ is zero as noise is not modelled. For random strategy $\mathbb{E}[\hat{\tau}_i^{imp,R}|\mathsf{n}] = \mathbb{E}[\hat{\tau}_i^{obs}|\mathsf{n}] \approx \tau_i^{*obs}$, whereas for donor strategy

$$\begin{aligned} \mathbb{E}[\hat{\tau}_i^{imp,D}|\mathsf{n}] &= \mathbb{E}[\mathbb{E}[\hat{\tau}_i^{imp,D}|\mathcal{Q}_2]|\mathsf{n}] = \mathbb{E}[\hat{\tau}_i^{obs}(1 - \frac{1}{\mathsf{N}_i^{obs}})|\mathsf{n}] \\ &\overset{Taylor}{\approx} \tau_i^{*obs}(1 - \frac{1}{\mathbb{E}[\mathsf{N}_i^{obs}|\mathsf{n}]}), \end{aligned}$$

where

$$\mathbb{E}[\mathsf{N}_i^{obs}|\mathsf{n}] = \begin{cases} \mathsf{n}(1-p^*)\Pr\Big(\hat{b}(\boldsymbol{X}^{obs}) = i|\mathsf{n}\Big), & : \text{C/T} \\ \mathsf{n}(1-p^*)\Pr\Big(\hat{b}\big((Y^{obs},\boldsymbol{X}^{obs})^T\big) = i|\mathsf{n}\Big) & : \text{CJ/TJ.} \end{cases}$$

For smoothed random strategies first order Taylor approximation is applied. Approximation for the expectation for standard $\boldsymbol{X}$-clustering with smoothed random strategy is

$$\begin{aligned} \mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_1] &= \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathcal{Q}_1] \\ &= \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}N_l\hat{\tau}_l^{obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}N_l}|\mathcal{Q}_1] \overset{Taylor}{\approx} \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]\tau_l^{*obs}}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]}. \end{aligned}$$

Further, approximation for smoothed random strategy and joint $(Y, \boldsymbol{X})$ clustering is

$$\begin{aligned} \mathbb{E}[\hat{\tau}_i^{imp}|\mathcal{Q}_1] &= \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathcal{Q}_1] \\ &= \mathbb{E}[\frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}N_l\hat{\tau}_l}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}N_l}|\mathcal{Q}_1] \overset{Taylor}{\approx} \frac{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]\mathbb{E}[\hat{\tau}_l|\mathcal{Q}_1]}{\sum_{l=1}^{\mathsf{n}_c} h_{i,l}\mathbb{E}[N_l|\mathcal{Q}_1]}. \end{aligned}$$

The conditional expectation in the second term of the variance decomposition is computed as

$$\begin{aligned} \mathbb{E}[\hat{\mu}^{comp}|\mathcal{Q}_2] &= \mathbb{E}[\frac{1}{\mathsf{n}}(\mathsf{N}^{obs}\hat{\mu}^{obs} + \sum_{i=1}^{\mathsf{n}_c} \mathsf{N}_i^{mis}\hat{\mu}_i^{imp})|\mathcal{Q}_2] \\ &= \frac{1}{\mathsf{n}}(\mathsf{n}^{obs}\mu^{obs} + \sum_{i=1}^{\mathsf{n}_c} \mathbb{E}[\mathsf{N}_i^{mis}\hat{\mu}_i^{imp}|\mathcal{Q}_2]) \\ &\overset{Taylor}{\approx} \frac{1}{\mathsf{n}}(\mathsf{n}^{obs}\mu^{obs} + \sum_{i=1}^{\mathsf{n}_c} \mathbb{E}[\mathsf{N}_i^{mis}|\mathcal{Q}_2]\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2]) \\ &\approx \frac{1}{\mathsf{n}}(\mathsf{n}^{obs}\mu^{obs} + \mathsf{n}^{mis}\sum_{i=1}^{\mathsf{n}_c} \Pr(b(\boldsymbol{X}^{mis}) = i|\mathcal{Q}_2)\mathbb{E}[\hat{\mu}_i^{imp}|\mathcal{Q}_2]). \end{aligned}$$

A6.1.1 Subresult for justifying approximation 6.5

Before justifying approximation 6.5 a useful result is introduced.

A useful decompositions of $\hat{\tau}^{imp}$ for cell imputation methods is

$$\begin{aligned}
\hat{\tau}^{imp} &= \frac{1}{\mathsf{N}^{mis}-1}\sum_{j=1}^{\mathsf{N}^{mis}}\Big(Y_j^{imp}-\hat{\mu}^{imp}\Big)^2 \qquad (14)\\
&= \frac{1}{\mathsf{N}^{mis}-1}\sum_{l=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}_l^{mis}}\Big(Y_{j,l}^{imp}-\hat{\mu}^{imp}\Big)^2\\
&= \frac{1}{\mathsf{N}^{mis}-1}\sum_{l=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}_l^{mis}}\Big(Y_{j,l}^{imp}-\hat{\mu}_l^{imp}+\hat{\mu}_j^{imp}-\hat{\mu}^{imp}\Big)^2\\
&= \frac{1}{\mathsf{N}^{mis}-1}\sum_{l=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}_l^{mis}}\Big((Y_{j,l}^{imp}-\hat{\mu}_l^{imp})^2+2(Y_{j,l}^{imp}-\hat{\mu}_l^{imp})(\hat{\mu}_l^{imp}-\hat{\mu}^{imp})\\
&\quad +(\hat{\mu}_l^{imp}-\hat{\mu}^{imp})^2\Big)\\
&= \frac{1}{\mathsf{N}^{mis}-1}\sum_{l=1}^{\mathsf{n}_c}\sum_{j=1}^{\mathsf{N}_l^{mis}}\Big((Y_{j,l}^{imp}-\hat{\mu}_l^{imp})^2+(\hat{\mu}_j^{imp}-\hat{\mu}^{imp})^2\Big)\\
&= \frac{1}{\mathsf{N}^{mis}-1}\sum_{l=1}^{\mathsf{n}_c}\Big((\mathsf{N}_l^{mis}-1)\hat{\tau}_l^{imp}+\mathsf{N}_l^{mis}(\hat{\mu}_l^{imp}-\hat{\mu}^{imp})^2\Big).
\end{aligned}$$

Approximation 6.5

The bias of $\hat{\tau}^{comp}$ for n observations may be approximated as

$$\begin{aligned}
\mathbb{B}\text{ias}[\hat{\tau}^{comp}|\mathsf{n}] \approx\ & \underbrace{p^*(\sum_{i=1}^{\mathsf{n}_c}p_i^{mis}(\mu_i^{*imp}-\sum_{l=1}^{\mathsf{n}_c}p_l^{mis}\mu_l^{*imp})^2+B-\tau^{*mis})}_{\text{difference between variance of imputed and missing Y values}}\\
&+\underbrace{p^*(1-p^*)\Big((\mu^{*obs}-\sum_{l=1}^{\mathsf{n}_c}p_l^{mis}\mu_l^{*imp})^2-(\mu^{*obs}-\mu^{*mis})^2\Big)}_{\text{difference between mean of imputed and missing Y values}}\\
&+\underbrace{\qquad\qquad O(\mathsf{n}^{-1})\qquad\qquad}_{\text{approximation error}},
\end{aligned}$$

where term $p_l^{mis}=\Pr\Big(\hat{b}(\boldsymbol{X}^{mis})=l|\mathsf{n}\Big)$, and terms μ_l^{*imp} depend on cell method and strategy as follows:

$$\mu_l^{*imp}=\begin{cases}\mu_l^{*obs} & :\mathrm{T/TJ(S=M/R/D)}\quad,\\ \mu_l^{*s} & :\mathrm{T/TJ(S=M^s/R^s)}\quad,\end{cases}$$

and term B is due to noise modelling, and depends on cell method and on imputation strategy $\hat{\epsilon}^S$ as follows:

$$B = \left\{\begin{array}{lll} 0 & : \mathrm{S}=\mathrm{M}/\mathrm{M}^{\mathrm{s}} & , \\ & & \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{obs}|\mathsf{n}] & : \mathrm{CJ}/\mathrm{TJ}/\mathrm{C}/\mathrm{T}(\mathrm{S}=\mathrm{R}) & , \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{obs}|\mathsf{n}](1-\frac{1}{\mathbb{E}[\mathsf{N}_l^{obs}|\mathsf{n}]}) & : \mathrm{CJ}/\mathrm{TJ}/\mathrm{C}/\mathrm{T}(\mathrm{S}=\mathrm{D}) & , \\ & & \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{T,R^s}|\mathsf{n}] & : \mathrm{T}(\mathrm{S}=\mathrm{R}^{\mathrm{s}}) & , \text{ and} \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{TJ,R^s}|\mathsf{n}] & : \mathrm{TJ}(\mathrm{S}=\mathrm{R}^{\mathrm{s}}) & . \end{array}\right.$$

Justification: derivation of result is based on decompositions

$$\begin{aligned} & \mathbb{E}[\hat{\tau}^{comp}|\mathsf{n}] \\ = \ & \mathbb{E}[(1-\frac{\mathsf{N}^{mis}}{\mathsf{n}-1})\hat{\tau}^{obs} + \frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp} + \frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2|\mathsf{n}] - \tau^*, \end{aligned}$$

and

$$\tau^* = (1-p^*)\tau^{*obs} + p^*\tau^{*mis} + p^*(1-p^*)(\mu^{*obs}-\mu^{*mis})^2.$$

Expectations in decomposition of $\mathbb{E}[\hat{\tau}^{comp}|\mathsf{n}]$ are computed using first order Taylor approximation. Thus the bias is approximated as follows

$$\begin{aligned} \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp}|\mathsf{n}] &= \mathbb{E}[\hat{\tau}^{comp}|\mathsf{n}] - \tau^* \\ &= \mathbb{E}[(1-\frac{\mathsf{N}^{mis}}{\mathsf{n}-1})\hat{\tau}^{obs}|\mathsf{n}] - (1-p^*)\tau^{*obs} \\ &\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}\hat{\tau}^{imp}|\mathsf{n}] - p^*\tau^{*mis} \\ &\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}\mathsf{N}^{obs}}{\mathsf{n}(\mathsf{n}-1)}(\hat{\mu}^{obs}-\hat{\mu}^{imp})^2|\mathsf{n}] - p^*(1-p^*)(\mu^{*obs}-\mu^{*mis})^2 \\ &\overset{Taylor}{\approx} \mathbb{E}[(1-\frac{\mathsf{N}^{mis}}{\mathsf{n}-1})|\mathsf{n}]\mathbb{E}[\hat{\tau}^{obs}|\mathsf{n}] - (1-p^*)\tau^{*obs} \\ &\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}-1}{\mathsf{n}-1}|\mathsf{n}]\mathbb{E}[\hat{\tau}^{imp}|\mathsf{n}] - p^*\tau^{*mis} \\ &\quad +\mathbb{E}[\frac{\mathsf{N}^{mis}}{\mathsf{n}}|\mathsf{n}]\mathbb{E}[\frac{\mathsf{N}^{obs}}{\mathsf{n}-1}|\mathsf{n}](\mathbb{E}[\hat{\mu}^{obs}|\mathsf{n}]-\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2 \\ &\quad -p^*(1-p^*)(\mu^{*obs}-\mu^{*mis})^2 \\ &\approx (1-p^*)\tau^{*obs} - (1-p^*)\tau^{*obs} + p^*\mathbb{E}[\hat{\tau}^{imp}|\mathsf{n}] - p^*\tau^{*mis} \\ &\quad +p^*(1-p^*)(\mu^{*obs}-\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2 - p^*(1-p^*)(\mu^{*obs}-\mu^{*mis})^2 \\ &\quad +O(\mathsf{n}^{-1}) \\ &= p^*(\mathbb{E}[\hat{\tau}^{imp}|\mathsf{n}] - \tau^{*mis}) \\ &\quad +p^*(1-p^*)\Big((\mu^{*obs}-\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2 - (\mu^{*obs}-\mu^{*mis})^2\Big) \\ &\quad +O(\mathsf{n}^{-1}), \end{aligned} \tag{15}$$

where $\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}]$ is computed as

$$\begin{aligned}
\mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}] &= \mathbb{E}[\frac{1}{\mathsf{N}^{mis}}\sum_{i=1}^{\mathsf{n}_c}\mathsf{N}_i^{mis}\hat{\mu}_i^{imp}|\mathsf{n}] \\
&\overset{Taylor}{\approx} \frac{1}{\mathbb{E}[\mathsf{N}^{mis}|\mathsf{n}]}\sum_{i=1}^{\mathsf{n}_c}\mathbb{E}[\mathsf{N}_i^{mis}|\mathsf{n}]\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] \\
&\approx \frac{1}{\mathsf{n}p^*}\sum_{i=1}^{\mathsf{n}_c}\mathsf{n}p^*\mathrm{Pr}\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] \\
&\approx \sum_{i=1}^{\mathsf{n}_c}\mathrm{Pr}\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)\mu_i^{*imp} + O(\mathsf{n}^{-1}) \\
&= \sum_{i=1}^{\mathsf{n}_c}p_i^{mis}\mu_i^{*imp} + O(\mathsf{n}^{-1}),
\end{aligned}\tag{16}$$

where μ_i^{*imp} equals to μ_i^{*obs} for unsmoothed imputation methods and to μ_i^{*s} for smoothed methods.

Expectation $\mathbb{E}[\hat{\tau}^{imp}|\mathsf{n}]$ is computed by i) replacing $\mathsf{N}_l^{mis} - 1$ by N_l^{mis}, using Equation (14), and first order Taylor approximation as follows

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{imp}|\mathsf{n}] &\overset{eq.\ (14)}{=} \mathbb{E}\Bigg[\frac{1}{\mathsf{N}^{mis}-1}\sum_{l=1}^{\mathsf{n}_c}\Big((\mathsf{N}_l^{mis}-1)\hat{\tau}_l^{imp} + \mathsf{N}_l^{mis}(\hat{\mu}_l^{imp} - \hat{\mu}^{imp})^2\Big)|\mathsf{n}\Bigg] \\
&\overset{i)}{\approx} \mathbb{E}\Bigg[\sum_{l=1}^{\mathsf{n}_c}\frac{\mathsf{N}_l^{mis}}{\mathsf{N}^{mis}}\Big(\hat{\tau}_l^{imp} + (\hat{\mu}_l^{imp} - \hat{\mu}^{imp})^2\Big)|\mathsf{n}\Bigg] \\
&\overset{Taylor}{\approx} \sum_{l=1}^{\mathsf{n}_c}\frac{\mathbb{E}[\mathsf{N}_l^{mis}|\mathsf{n}]}{\mathbb{E}[\mathsf{N}^{mis}|\mathsf{n}]}\Big(\mathbb{E}[\hat{\tau}_l^{imp}|\mathsf{n}] + (\mathbb{E}[\hat{\mu}_l^{imp}|\mathsf{n}] - \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2\Big) \\
&\approx \sum_{l=1}^{\mathsf{n}_c}\frac{\mathsf{n}p^*\mathrm{Pr}\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)}{\mathsf{n}p^*}\Big(\mathbb{E}[\hat{\tau}_l^{imp}|\mathsf{n}] + (\mathbb{E}[\hat{\mu}_l^{imp}|\mathsf{n}] - \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2\Big) \\
&= \sum_{l=1}^{\mathsf{n}_c}p_l^{mis}\Big(\mathbb{E}[\hat{\tau}_l^{imp}|\mathsf{n}] + (\mathbb{E}[\hat{\mu}_l^{imp}|\mathsf{n}] - \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2\Big) \\
&= \sum_{l=1}^{\mathsf{n}_c}p_l^{mis}(\mathbb{E}[\hat{\mu}_l^{imp}|\mathsf{n}] - \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}])^2 + \sum_{l=1}^{\mathsf{n}_c}p_l^{mis}\mathbb{E}[\hat{\tau}_l^{imp}|\mathsf{n}] \\
&\approx \sum_{l=1}^{\mathsf{n}_c}p_l^{mis}(\mu_l^{*imp} - \sum_{i=1}^{\mathsf{n}_c}p_i^{mis}\mu_i^{*imp})^2 + \sum_{l=1}^{\mathsf{n}_c}p_l^{mis}\mathbb{E}[\hat{\tau}_l^{imp}|\mathsf{n}] + O(\mathsf{n}^{-1}).
\end{aligned}\tag{17}$$

By plugging Equations (16) and (17) into Equation (15) yields to

$$\begin{aligned}
\mathbb{B}\mathrm{ias}[\hat{\tau}^{comp}|\mathsf{n}] \approx\ & p^*(\sum_{l=1}^{\mathsf{n}_c}p_l^{mis}(\mu_l^{*imp} - \sum_{i=1}^{\mathsf{n}_c}p_i^{mis}\mu_i^{*imp})^2 + \sum_{l=1}^{\mathsf{n}_c}p_l^{mis}\mathbb{E}[\hat{\tau}_l^{imp}|\mathsf{n}] - \tau^{*mis}) \\
& + p^*(1-p^*)\Big((\mu^{*obs} - \sum_{i=1}^{\mathsf{n}_c}p_i^{mis}\mu_i^{*imp})^2 - (\mu^{*obs} - \mu^{*mis})^2\Big) \\
& + O(\mathsf{n}^{-1}),
\end{aligned}$$

where $\sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{imp}|\mathsf{n}]$ is left in implicit form as given in the approximation. However, for donor strategy first order Taylor approximation is used to distinguish the result from result for random strategy:

$$\begin{aligned}
\sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}[\hat{\tau}_l^{imp,D}|\mathsf{n}] &= \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}\Big[\mathbb{E}[\hat{\tau}_l^{imp,D}|\mathcal{Q}_2]|\mathsf{n}\Big] \\
&= \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}\Big[\hat{\tau}_l^{obs}(1-\frac{1}{\mathsf{N}_l^{obs}})|\mathsf{n}\Big] \\
&\overset{Taylor}{\approx} \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mathbb{E}\Big[\hat{\tau}_l^{obs}|\mathsf{n}\Big](1-\frac{1}{\mathbb{E}[\mathsf{N}_l^{obs}|\mathsf{n}]}).
\end{aligned}$$

Consequence 6.6

Asymptotically one has the following approximations

i)

$$\begin{aligned}
\lim_{\mathsf{n}\to\infty} \mathbb{B}\text{ias}[\hat{\mu}^{comp}|\mathsf{n}] &\approx p^*(\sum_{i=1}^{\mathsf{n}_c} p_i^{mis}\mu_i^{*imp} - \mu^{*mis}) \\
\lim_{\mathsf{n}\to\infty} \mathbb{V}\text{ar}[\hat{\mu}^{comp}|\mathsf{n}] &\approx 0,
\end{aligned}$$

where

$$\mu_i^{*imp} = \begin{cases} \mu_i^{*obs} & : \text{C/CJ/T/TJ(S = M/R/D)} \quad , \\ \mu_i^{*s} \approx \frac{\sum_l h_{i,l}(p_l^{mis}+p_l^{obs})\mu_l^{*obs}}{\sum_l h_{i,l}(p_l^{mis}+p_l^{obs})} & : \text{T/TJ(S = M}^{\text{s}}\text{/R}^{\text{s}}\text{)} \quad . \end{cases}$$

ii)

$$\begin{aligned}
&\lim_{\mathsf{n}\to\infty} \mathbb{B}\text{ias}[\hat{\tau}^{comp}|\mathsf{n}] \\
\approx\ &\underbrace{p^*(\sum_{i=1}^{\mathsf{n}_c} p_i^{mis}(\mu_i^{*imp} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 + C - \tau^{*mis})}_{\text{difference between variance of imputed and missing Y values}} \\
&+\underbrace{p^*(1-p^*)\Big((\mu^{*obs} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 - (\mu^{*obs}-\mu^{*mis})^2\Big)}_{\text{difference between mean of imputed and missing Y values}},
\end{aligned}$$

where

$$p_i^{mis} = \lim_{\mathsf{n}\to\infty} \Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big),$$

$$p_i^{obs} = \lim_{\mathsf{n}\to\infty} \Pr\Big(\hat{b}\big((Y^{obs}, \boldsymbol{X}^{obs})^T\big) = i|\mathsf{n}\Big),$$

and C in ii) is due to noise modelling and depends on cell method and on imputation strategy $\hat{\epsilon}^S$ as follows:

$$C = \begin{cases} 0 & : (\mathrm{S} = \mathrm{M}/\mathrm{M}^{\mathrm{s}}) \quad , \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \tau_l^{*obs} & : \mathrm{C/CJ/T/TJ}(\mathrm{S} = \mathrm{R/D}) \quad , \\ \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\tau}_l^{T,R^s}|\mathsf{n}] & : \mathrm{T}(\mathrm{S} = \mathrm{R}^{\mathrm{s}}) \quad , \text{ and} \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\tau}_l^{TJ,R^s}|\mathsf{n}] & : \mathrm{TJ}(\mathrm{S} = \mathrm{R}^{\mathrm{s}}) \quad . \end{cases}$$

Justification: i) approximative limit of bias of first moment estimator is computed using approximation 6.2 as

$$\begin{aligned} \lim_{\mathsf{n}\to\infty} \mathbb{B}\mathrm{ias}[\hat{\mu}^{comp}|\mathsf{n}] &\approx \lim_{\mathsf{n}\to\infty} p^*(\sum_{i=1}^{\mathsf{n}_c} \Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] - \mu^{*mis}) \\ &\quad + \lim_{\mathsf{n}\to\infty} O(\mathsf{n}^{-1}) \\ &= \lim_{\mathsf{n}\to\infty} p^*(\sum_{i=1}^{\mathsf{n}_c} \Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big)\mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] - \mu^{*mis}) \\ &\approx p^*(\sum_{i=1}^{\mathsf{n}_c} \lim_{\mathsf{n}\to\infty} \Pr\Big(\hat{b}(\boldsymbol{X}^{mis}) = i|\mathsf{n}\Big) \lim_{\mathsf{n}\to\infty} \mathbb{E}[\hat{\mu}_i^{imp}|\mathsf{n}] - \mu^{*mis}) \\ &= p^*(\sum_{i=1}^{\mathsf{n}_c} p_i^{mis} \mu_i^{*imp} - \mu^{*mis}). \end{aligned}$$

Note that all terms in approximate variance for $\hat{\mu}^{comp}$ given n (approximation 6.4) are of order $O(\mathsf{n}^{-1})$, thus

$$\lim_{\mathsf{n}\to\infty} \mathbb{V}\mathrm{ar}[\hat{\mu}^{comp}|\mathsf{n}] \approx 0.$$

ii) approximate limit of bias of second moment estimator is computed using approximation 6.5 as.

$$\begin{aligned} \lim_{\mathsf{n}\to\infty} \mathbb{B}\mathrm{ias}[\hat{\tau}^{comp}|\mathsf{n}] &\approx \lim_{\mathsf{n}\to\infty} p^*(\sum_{i=1}^{\mathsf{n}_c} p_i^{mis}(\mu_i^{*imp} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 + B - \tau^{*mis}) \\ &\quad + \lim_{\mathsf{n}\to\infty} p^*(1-p^*)\Big((\mu^{*obs} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 - (\mu^{*obs} - \mu^{*mis})^2\Big) \\ &\quad + \lim_{\mathsf{n}\to\infty} O(\mathsf{n}^{-1}) \\ &= p^*(\sum_{i=1}^{\mathsf{n}_c} p_i^{mis}(\mu_i^{*imp} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 + \lim_{\mathsf{n}\to\infty} B - \tau^{*mis}) \\ &\quad + p^*(1-p^*)\Big((\mu^{*obs} - \sum_{l=1}^{\mathsf{n}_c} p_l^{mis}\mu_l^{*imp})^2 - (\mu^{*obs} - \mu^{*mis})^2\Big). \end{aligned}$$

Now $C = \lim B$ is

$$C = \begin{cases} 0 & : \mathrm{S} = \mathrm{M}/\mathrm{M}^{\mathrm{s}} & , \\ \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \tau_l^{*obs} & : \mathrm{CJ}/\mathrm{TJ}/\mathrm{C}/\mathrm{T}(\mathrm{S} = \mathrm{R}/\mathrm{D}) & , \\ \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \lim \mathbb{E}[\hat{\tau}_l^{T,R^s} | \mathsf{n}] & : \mathrm{T}(\mathrm{S} = \mathrm{R}^{\mathrm{s}}) & , \text{ and} \\ \sum_{l=1}^{\mathsf{n}_c} p_l^{mis} \lim \mathbb{E}[\hat{\tau}_l^{TJ,R^s} | \mathsf{n}] & : \mathrm{TJ}(\mathrm{S} = \mathrm{R}^{\mathrm{s}}) & . \end{cases}$$

Note that limit within cells for unsmoothed random and donor strategies are same.

A6.2 Cell methods / unit level

Unit level results for cell imputation are derived here.

Approximation 6.13

Mean squared error $\mathrm{mse}(Y^{imp} | \mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n})$ can be approximated as

$$\begin{aligned} & \mathrm{mse}(Y^{imp} | \mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}) \\ \approx \; & \Big(\underbrace{\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis} | \mathcal{Q}) \mu_{i,\mathsf{n}^{obs}}^{*imp} - \mathbb{E}[Y^{mis} | \mathbf{x}^{mis}]}_{\text{bias}} \Big)^2 \\ & + \underbrace{\sum_{i=1}^{\mathsf{n}_c} \mu_{i,\mathsf{n}^{obs}}^{*imp} g_i'(\mathbf{x}^{mis} | \mathcal{Q})^T \mathbb{V}\mathrm{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}) | \mathbf{d}_{\boldsymbol{X}}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \sum_{i=1}^{\mathsf{n}_c} \mu_{i,\mathsf{n}^{obs}}^{*imp} g_i'(\mathbf{x}^{mis} | \mathcal{Q})}_{\text{sampling variance due to estimation of classifier parameters}} \\ & + \underbrace{\big(g_1(\mathbf{x}^{mis} | \mathcal{Q}) \ldots g_{\mathsf{n}_c}(\mathbf{x}^{mis} | \mathcal{Q}) \big) \mathbb{V}\mathrm{ar}[\hat{\mu}_{\{u\}}^{imp} | \mathbf{d}_{\boldsymbol{X}}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \big(g_1(\mathbf{x}^{mis} | \mathcal{Q}) \ldots g_{\mathsf{n}_c}(\mathbf{x}^{mis} | \mathcal{Q}) \big)^T}_{\text{sampling variance due to estimation of imputation model parameters}} \\ & + \underbrace{\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis} | \mathcal{Q}) \mathbb{E}[\hat{\tau}_i^{imp}(\mathbf{x}^{mis}) | \mathsf{n}^{mis}, \mathsf{n}]}_{\text{imputation variance}} + \underbrace{v^{*mis}(\mathbf{x}^{mis})}_{\text{target variance}} \qquad u = 1, \ldots, \mathsf{n}_c, \end{aligned}$$

expected prediction in i:th cell is

$$\mu_{i,\mathsf{n}^{obs}}^{*imp} = \begin{cases} \mu_{i,\mathsf{n}^{obs}}^{*obs} \approx \mu_i^{*obs} + O\big((\mathsf{n}^{obs})^{-1}\big) & : \mathrm{C}/\mathrm{T}/\mathrm{CJ}/\mathrm{TJ}(\mathrm{S} = \mathrm{M}/\mathrm{R}/\mathrm{D}), \\ \\ \mu_{i,\mathsf{n}^{obs}}^{*s} \approx \frac{\sum_l h_{i,l} \mathbb{E}[N_l | \mathsf{n}^{mis}, \mathsf{n}] \mu_l^{*obs}}{\sum_l h_{i,l} \mathbb{E}[N_l | \mathsf{n}^{mis}, \mathsf{n}]} + O\big((\mathsf{n}^{obs})^{-1}\big) & : \mathrm{T}/\mathrm{TJ}(\mathrm{S} = \mathrm{M}^{\mathrm{s}}/\mathrm{R}^{\mathrm{s}}), \end{cases}$$

$\mathcal{Q} = \mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}} | \mathbf{d}_{\boldsymbol{X}}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]$, quantity $g_i'(\cdot)$ is derivative of $g_i(\cdot)$ with respect to $vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})$ which is evaluated at $\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}} | \mathbf{d}_{\boldsymbol{X}}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]$, and term

$t = \mathbb{E}[\hat{\tau}_i^{imp}(\mathbf{x}^{mis})|\mathsf{n}^{mis}, \mathsf{n}]$ depends on imputation strategy S:

$$t = \begin{cases} 0 & : \mathrm{C/T/CJ(S = M)/} \\ & \ \mathrm{T(S = M^s)}, \\ \approx \tau_i^{*obs} & : \mathrm{C/T/CJ(S = R)}, \\ \approx \tau_i^{*obs}(1 - \frac{1}{\mathbb{E}[\mathsf{N}_i^{obs}|\mathsf{n}^{mis},\mathsf{n}]}) & : \mathrm{C/T/CJ(S = D)}, \\ & \\ \mathbb{E}[\hat{\tau}_i^{T,R^s}|\mathsf{n}^{mis}, \mathsf{n}] \approx \frac{\sum_l h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}^{mis},\mathsf{n}]\mathbb{E}[\hat{\tau}_l^s|\mathsf{n}^{mis},\mathsf{n}]}{\sum_l h_{i,l}\mathbb{E}[\mathsf{N}_l|\mathsf{n}^{mis},\mathsf{n}]} & : \mathrm{T(S = R^s)}, \\ \mathbb{E}[(\hat{\mu}_{i,\mathsf{n}^{obs}}^{imp} - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp})^2|\mathsf{n}^{mis}, \mathsf{n}] & : \mathrm{TJ(S = M/M^s)}, \\ \mathbb{E}[\hat{\tau}_i^{obs}|\mathsf{n}^{mis}, \mathsf{n}] + \mathbb{E}[(\hat{\mu}_i^{obs} - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp})^2|\mathsf{n}^{mis}, \mathsf{n}] & : \mathrm{TJ(S = R)}, \\ \mathbb{E}[\hat{\tau}_i^{TJ,R^s}|\mathsf{n}^{mis}, \mathsf{n}] + \mathbb{E}[(\hat{\mu}_i^s - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp})^2|\mathsf{n}^{mis}, \mathsf{n}] & : \mathrm{TJ(S = R^s)}, \end{cases}$$

in which quantities $\hat{\mu}_{i,\mathsf{n}^{obs}}^{imp} - \hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp}$ are location shifts from mean prediction $\hat{\overline{Y}}_{\mathbf{x}^{mis}}^{imp} = \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})\hat{\mu}_{i,\mathsf{n}^{obs}}^{imp}$ to modes of multimodal imputation noise distribution.

Justification: at first, new notation is described. Vector of estimates of means of missing Y values within cells is denoted as $\boldsymbol{\mu}_{\{u\}}^{imp} = (\mu_1^{imp}, \ldots, \mu_{\mathsf{n}_c}^{imp})^T$.

Recall the following decomposition of mean squared error (which is given in theorem 3.5, Chapter 3)

$$\begin{aligned} \mathrm{mse}(Y^{imp}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}) &= \Big(\underbrace{\mathbb{E}[\hat{g}(\mathbf{x}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - g^{*mis}(\mathbf{x}^{mis})}_{\text{imputation bias at } \mathbf{x}^{\text{mis}}}\Big)^2 \\ &+ \underbrace{\quad \mathbb{V}\mathrm{ar}[Y^{imp}_{|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}}] \quad}_{\text{imputation variance at } \mathbf{x}^{\text{mis}}} \\ &+ \underbrace{\qquad \mathbb{V}\mathrm{ar}[Y_{|\mathbf{x}^{mis}}] \qquad}_{\text{v}^{*\text{mis}}(\mathbf{x}^{\text{mis}}),\ \text{target noise at } \mathbf{x}^{\text{mis}}}. \end{aligned}$$

To compute the first two terms (the last term is not affected by imputation method) some approximations are applied. Following assumptions are used in derivation of results:

- Predictions based on crisp classifiers (maximum posterior and randomized) are approximated using soft classifier: estimator of Y^{mis} at $\mathbf{x}^{mis}$ given conditionalisation $\mathcal{Q}_3$ is

$$Y^{imp}|\mathbf{x}^{mis}, \mathcal{Q}_3 \approx \underbrace{\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\hat{\mu}_{i,\mathsf{n}^{obs}}^{imp}}_{=\overline{\text{Y}}^{\text{imp}}_{\mathbf{x}^{\text{mis}}},\ \text{mean prediction}} + \underbrace{\hat{\epsilon}(\mathbf{x}^{mis})}_{\text{imputation noise}}, \tag{18}$$

  where $u = 1, \ldots, \mathsf{n}_c$, and $\mathbb{E}[\hat{\epsilon}(\mathbf{x}^{mis})|\mathcal{Q}_3] = 0$ for any $\mathbf{x}^{mis}$. Note that in classification of incomplete observations only $\boldsymbol{X}$ part $\mathbf{w}_{\boldsymbol{X},\{u\}}$ of centroids $\mathbf{w}_{\{u\}}$ is used.

- Posterior probabilities $g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}}), i = 1, \ldots, \mathsf{n}_c$ are continuous and have first derivative with respect to $\mathbf{x}^{mis}$ and $\mathbf{w}_{\boldsymbol{X},\{u\}}$.

Results for mean squared error at point $\mathbf{x}^{mis}$ are derived from $\mathcal{Q}_3$ conditionalisation by integrating over distribution of $\mathbf{D}^{train}_{\mathsf{n}^{obs}}, \hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}, \mathbf{D}^{mis}_{Y,\mathsf{n}^{mis}}|\mathbf{d}^{mis}_{\boldsymbol{X},\mathsf{n}^{mis}}$.

Imputation bias term is computed using first order Taylor approximation as

$$\begin{aligned}
& \mathbb{E}[\hat{g}(\mathbf{x}^{mis})|\mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}] - g^{*mis}(\mathbf{x}^{mis}) \\
= \quad & \mathbb{E}[\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})\hat{\mu}^{imp}_{i,\mathsf{n}^{obs}}|\mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}] - \mathbb{E}[Y^{mis}|\mathbf{x}^{mis}] \\
\overset{Taylor}{\approx} \quad & \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}])\mathbb{E}[\hat{\mu}^{imp}_{i,\mathsf{n}^{obs}}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - \mathbb{E}[Y^{mis}|\mathbf{x}^{mis}] \\
\approx \quad & \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathcal{Q})\mu^{*imp}_{i,\mathsf{n}^{obs}} - \mathbb{E}[Y^{mis}|\mathbf{x}^{mis}],
\end{aligned}$$

where $\mathcal{Q} = \{\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}]\}$.

Imputation variance is derived using chain rule and approximation (18) as

$$\begin{aligned}
\mathbb{V}\text{ar}[Y^{imp}_{|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}}] \quad = \quad & \mathbb{V}\text{ar}[\mathbb{E}[Y^{imp}|\mathbf{x}^{mis}, \mathcal{Q}_3]|\mathbf{x}^{mis}, \mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}] \\
& +\mathbb{E}[\mathbb{V}\text{ar}[Y^{imp}|\mathbf{x}^{mis}, \mathcal{Q}_3]|\mathbf{x}^{mis}, \mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}] \\
\approx \quad & \mathbb{V}\text{ar}[\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})\hat{\mu}^{imp}_{i,\mathsf{n}^{obs}}|\mathbf{x}^{mis}, \mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}] \\
& +\mathbb{E}[\mathbb{V}\text{ar}[\hat{\epsilon}(\mathbf{x}^{mis})|\mathbf{x}^{mis}, \mathcal{Q}_3]|\mathbf{x}^{mis}, \mathbf{d}^{mis}_{\boldsymbol{X}}, \mathsf{n}^{mis}, \mathsf{n}],
\end{aligned} \tag{19}$$

where the first term in approximation is derived using first order Taylor approximation. For this i) covariance between $vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})$ and $\hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}}$ is assumed to be neglible. Let $h(vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}), \hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}})$ be a function which variance one is interested of. Taylor approximation yields

$$\begin{aligned}
& \mathbb{V}\text{ar}[h(vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}), \hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}})] \\
\overset{Taylor}{\approx} \quad & [\frac{\partial h(\cdot)}{\partial vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}), \hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}}}]^T \mathbb{V}\text{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}), \hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}}][\frac{\partial h(\cdot)}{\partial vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}), \hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}}}] \\
\overset{i)}{\approx} \quad & [\frac{\partial h(\cdot)}{\partial vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})}]^T \mathbb{V}\text{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})][\frac{\partial h(\cdot)}{\partial vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})}] \\
& +[\frac{\partial h(\cdot)}{\partial \hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}}}]^T \mathbb{V}\text{ar}[\hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}}][\frac{\partial h(\cdot)}{\partial \hat{\boldsymbol{\mu}}^{imp}_{\mathsf{n}^{obs}}}],
\end{aligned} \tag{20}$$

where brackets denote here evaluation of random variables within brackets at their expected value.

The first term in approximation given in Equation 19 is computed by applying approximation 20, what yields to

$$
\begin{aligned}
&\mathbb{V}\mathrm{ar}[\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})\hat{\mu}^{imp}_{i,\mathsf{n}^{obs}}|\mathbf{x}^{mis},\mathbf{d}^{mis}_{\boldsymbol{X}},\mathsf{n}^{mis},\mathsf{n}] \\
\approx\ & \underbrace{\sum_{i=1}^{\mathsf{n}_c}\mu^{*imp}_{i,\mathsf{n}^{obs}} g_i'(\mathbf{x}^{mis}|\mathcal{Q})^T\mathbb{V}\mathrm{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})|\mathbf{d}^{mis}_{\boldsymbol{X}},\mathsf{n}^{mis},\mathsf{n}]\sum_{i=1}^{\mathsf{n}_c}\mu^{*imp}_{i,\mathsf{n}^{obs}} g_i'(\mathbf{x}^{mis}|\mathcal{Q})}_{\text{sampling variance due to estimation of classifier parameters}} \\
+\ & \underbrace{\mathbf{g}(\mathbf{x}^{mis}|\mathcal{Q})^T\mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}_{\{u\}}|\mathbf{d}^{mis}_{\boldsymbol{X}},\mathsf{n}^{mis},\mathsf{n}]\mathbf{g}(\mathbf{x}^{mis}|\mathcal{Q})}_{\text{sampling variance due to estimation of imputation model parameters}} \quad , \quad u = 1,\ldots,\mathsf{n}_c
\end{aligned}
$$

where $\mathbf{g}(\mathbf{x}^{mis}|\mathcal{Q}) = (g_1(\mathbf{x}^{mis}|\mathcal{Q}),\ldots,g_{\mathsf{n}_c}(\mathbf{x}^{mis}|\mathcal{Q}))^T$.

Impact of modelled noise (the second term in approximation 19 for imputation variance) needs to be computed next. Variances of $\hat{\epsilon}(\mathbf{x}^{mis})$ at $\mathcal{Q}_3$ are

$$
\mathbb{V}\mathrm{ar}[\hat{\epsilon}(\mathbf{x}^{mis})|\mathcal{Q}_3] = \begin{cases}
0 & : \mathrm{C/T/CJ(S=M)/} \\
 & \ \mathrm{T(S=M^s)}, \\
\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{obs} & : \mathrm{C/T/CJ(S=R)} \quad , \\
\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{T,R^s} & : \mathrm{T(S=R^s)} \quad , \\
\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{obs}(1-\frac{1}{n_i^{obs}}) & : \mathrm{C/T/CJ(S=D)} \quad , \\
 & \\
\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})(\mu_i^{imp}-\overline{Y}^{imp}_{\mathbf{x}^{mis}})^2 & : \mathrm{TJ(S=M/M^s)} \quad , \\
\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})(\mu_i^{imp}-\overline{Y}^{imp}_{\mathbf{x}^{mis}})^2 & : \mathrm{TJ(S=R)} \quad , \\
+\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{obs} & \\
\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})(\mu_i^{imp}-\overline{Y}^{imp}_{\mathbf{x}^{mis}})^2 & : \mathrm{TJ(S=R^s)} \\
+\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})\tau_i^{TJ,R^s} & \quad ,
\end{cases}
$$

where $g_i(\mathbf{x}^{mis}|\mathbf{w}_{\boldsymbol{X},\{u\}})$ is estimate of posterior probability of i:th cell at $\boldsymbol{X}^{mis} = \mathbf{x}^{mis}$. Applying first order Taylor approximations yields to decomposition for imputation variance: $\sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathcal{Q})\mathbb{E}[\hat{\tau}_i^{imp}(\mathbf{x}^{mis})|\mathsf{n}^{mis},\mathsf{n}]$ and formulas for $\mathbb{E}[\hat{\tau}_i^{imp}(\mathbf{x}^{mis})|\mathsf{n}^{mis},\mathsf{n}]$. As an example, for donor strategy first order Taylor approximation yields to

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}[\hat{\epsilon}^D(\mathbf{x}^{mis})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] &= \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\epsilon}^D(\mathbf{x}^{mis})|\mathcal{Q}_3]|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\
&= \underbrace{\mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\epsilon}^D(\mathbf{x}^{mis})|\mathcal{Q}_3]|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}]}_{=0} \\
&= \sum_{i=1}^{\mathsf{n}_c}\mathbb{E}[g_i(\mathbf{x}^{mis}|\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})\hat{\tau}_i^{obs}(1-\frac{1}{\mathsf{N}_i^{obs}})|\mathbf{x}^{mis},\mathsf{n}^{mis},\mathsf{n}] \\
&\overset{Taylor}{\approx} \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathcal{Q})\mathbb{E}[\hat{\tau}_i^{obs}|\mathsf{n}^{mis},\mathsf{n}](1-\frac{1}{\mathbb{E}[\mathsf{N}_i^{obs}|\mathsf{n}^{mis},\mathsf{n}]}) \\
&\approx \sum_{i=1}^{\mathsf{n}_c} g_i(\mathbf{x}^{mis}|\mathcal{Q})\tau_i^{*obs}(1-\frac{1}{\mathbb{E}[\mathsf{N}_i^{obs}|\mathsf{n}^{mis},\mathsf{n}]}),
\end{aligned}
$$

where $\mathcal{Q} = \{\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbf{d}^{mis}_{\boldsymbol{X}},\mathsf{n}^{mis},\mathsf{n}]\}$.

Approximation 6.14

Expectation of $\widehat{mse}(Y^{comp})$ with n observations can be approximated as

$$
\begin{aligned}
&\mathbb{E}[\widehat{mse}(Y^{comp})|\mathsf{n}]\\
\approx\ & \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\mathbf{D}^{mis}_{\boldsymbol{X}}|\mathsf{n}}\left[\sum_{i=1}^{\mathsf{n}_c} g_i(\boldsymbol{X}^{mis}|\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathsf{n}^{mis},\mathsf{n}])\mu^{*imp}_{i,\mathsf{N}^{obs}}\right]}_{\text{variability of conditional mean estimate}} + (\underbrace{\mu^{*imp}_{\mathsf{n}} - \mu^{*mis}}_{\text{global bias}})^2\\
&+ \underbrace{\mathbb{V}\mathrm{ar}\big[\mathbb{E}[Y^{mis}|\boldsymbol{X}^{mis}]\big]}_{\text{variability of true model}}\\
&+ \underbrace{2\mathbb{E}_{\mathsf{N}^{mis},\mathbf{D}^{mis}_{\boldsymbol{X}}|\mathsf{n}}\left[\left(\sum_{i=1}^{\mathsf{n}_c} g_i(\boldsymbol{X}^{mis}|\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathsf{n}^{mis},\mathsf{n}])\mu^{*imp}_{i,\mathsf{N}^{obs}} - \mu^{*imp}_{\mathsf{n}}\right)\right.}_{\text{cross term}}\\
&\underbrace{\left.*\left(\mu^{*imp}_{\mathsf{n}} - \mathbb{E}[Y^{mis}|\boldsymbol{X}^{mis}]\right)\right]}_{\text{cross term (cont.)}}\\
&+ \underbrace{\sum_{i=1}^{\mathsf{n}_c} \mu^{*imp}_i g'_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})^T \mathbb{V}\mathrm{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})|\mathbb{E}[\mathbf{D}^{mis}_{\boldsymbol{X}},\mathsf{N}^{mis}|\mathsf{n}]] \sum_{i=1}^{\mathsf{n}_c} \mu^{*imp}_i g'_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})}_{\text{expected sampling variance due to estimation of classifier parameters}}\\
&+ \quad \underbrace{\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})^T \mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}_{\{u\}}|\mathbb{E}[\mathbf{D}^{mis}_{\boldsymbol{X}},\mathsf{N}^{mis}|\mathsf{n}]]\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})}_{\text{expected sampling variance due to estimation of imputation model parameters}}\\
&+ \quad \underbrace{\sum_{i=1}^{\mathsf{n}_c} g_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})\mathbb{E}[\hat{\tau}^{imp}_i(\boldsymbol{X}^{mis})|\mathsf{n}]}_{\text{expected imputation variance}}\\
&+ \underbrace{v^{*mis}}_{\text{expected target variance}} \qquad\qquad u = 1,\ldots,\mathsf{n}_c,
\end{aligned}
$$

where $\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}) = (g_1(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}),\ldots,g_{\mathsf{n}_c}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}))^T$ and

$$
\mu^{*imp}_{\mathsf{n}} = \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}] \approx \begin{cases} \sum_{i=1}^{\mathsf{n}_c} g_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})\mu^{*obs}_i & : \mathrm{C/CJ/T/TJ(S = M/R/D)},\\ \sum_{i=1}^{\mathsf{n}_c} g_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})\mu^{*s}_i & : \mathrm{T/TJ(S = M^s/R^s)}, \end{cases}
$$

$\mathcal{Q} = \{\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbb{E}[\mathbf{D}^{mis}_{\boldsymbol{X}}], \mathbb{E}[\mathsf{N}^{mis}], \mathsf{n}]\}$, and constant $\mathbb{E}[\hat{\tau}^{imp}_i(\boldsymbol{X}^{mis})|\mathsf{n}]$ depends on imputation method and strategy S as follows:

$$\mathbb{E}[\hat{\tau}^{imp}_i(\boldsymbol{X}^{mis})|\mathsf{n}] = \begin{cases} 0 & : \mathrm{C/T/CJ(S = M)/} \\ & \ \mathrm{T(S = M^s)} \\ \approx \tau^{*obs}_i & : \mathrm{C/T/CJ(S = R)}, \\ \approx \tau^{*obs}_i(1 - \frac{1}{\mathbb{E}[\mathsf{N}^{obs}_i|\mathsf{n}]}) & : \mathrm{C/T/CJ(S = D)}, \\ & \\ \mathbb{E}[\hat{\tau}^{T,R^s}_i|\mathsf{n}] & : \mathrm{T(S = R^s)}, \\ \mathbb{E}[(\hat{\mu}^{imp}_i - \hat{\overline{Y}}^{imp}_{\overline{\boldsymbol{X}}^{*mis}})^2|\mathsf{n}] & : \mathrm{TJ(S = M/M^s)} \\ \mathbb{E}[\hat{\tau}^{obs}_i|\mathsf{n}] + \mathbb{E}[(\hat{\mu}^{obs}_i - \hat{\overline{Y}}^{imp}_{\overline{\boldsymbol{X}}^{*mis}})^2|\mathsf{n}] & : \mathrm{TJ(S = R)}, \\ \mathbb{E}[\hat{\tau}^{TJ,R^s}_i|\mathsf{n}] + \mathbb{E}[(\hat{\mu}^{s}_i - \hat{\overline{Y}}^{imp}_{\overline{\boldsymbol{X}}^{*mis}})^2|\mathsf{n}] & : \mathrm{TJ(S = R^s)}, \end{cases} ,$$

Justification: recall the following decomposition of mean squared error at population level (which is given in theorem 3.6, Chapter 3)

$$\begin{aligned} \mathbb{E}[\hat{mse}(Y^{comp})|\mathsf{n}] &= \underbrace{\mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]\right]}_{\text{variability of conditional mean estimate}} \\ + &\quad \underbrace{(\mu^{*imp}_{\mathsf{n}} - \mu^{*mis}}_{\text{global bias}})^2 + \underbrace{\mathbb{V}\mathrm{ar}[g^{*mis}(\boldsymbol{X}^{mis})]}_{\text{variability of true model}} \\ + &\quad \underbrace{2\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\left(\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] - \mu^{*imp}_n\right)\left(\mu^{*imp}_n - g^{*mis}(\boldsymbol{X}^{mis})\right)\right]}_{\text{cross term}} \\ + &\quad \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]\right]}_{\text{expected variance of conditional mean estimate}} + \underbrace{v^{*imp}_{\mathsf{n}}}_{\text{expected imputation noise}} \\ + &\quad \underbrace{\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[2\mathbb{C}\mathrm{ov}[\hat{g}(\boldsymbol{X}^{mis}), \hat{\epsilon}_{\mathbf{x}^{mis}}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]\right]}_{\text{second cross term}} + \underbrace{v^{*mis}}_{\text{expected target noise}}, \end{aligned}$$

where variability of true model and expected target noise (variance) are not affected by imputation method. Further, second cross term is assumed to be zero. This holds strictly at least for all other methods than joint $(Y, \boldsymbol{X})$ clustering TS-SOM methods (which utilize covariates for missing Y values).

Variability of conditional mean estimate is derived as

$$\begin{aligned} & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]\right] \\ = \quad & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{E}[\sum_{i=1}^{\mathsf{n}_c} g_i(\boldsymbol{X}^{mis}|\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})\hat{\mu}^{imp}_{i,\mathsf{N}^{obs}}|\mathbf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]\right] \\ \overset{Taylor}{\approx} \quad & \mathbb{V}\mathrm{ar}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\sum_{i=1}^{\mathsf{n}_c} g_i(\boldsymbol{X}^{mis}|\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathsf{n}^{mis}, \mathsf{n}])\mu^{*imp}_{i,\mathsf{N}^{obs}}\right]. \end{aligned}$$

Global bias requires computation of μ_{n}^{*imp}, which is approximated as

$$\mu_{\mathsf{n}}^{*imp} = \mathbb{E}[\hat{\mu}^{imp}|\mathsf{n}] \approx \begin{cases} \sum_{i=1}^{\mathsf{n}_c} g_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})\mu_i^{*obs} & : \mathrm{C/CJ/T/TJ(S = M/R/D)}, \\ \sum_{i=1}^{\mathsf{n}_c} g_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})\mu_i^{*s} & : \mathrm{T/TJ(S = M^s/R^s)}. \end{cases}$$

Cross term in its implicit form follows by assigning approximation for expectation $\mathbb{E}[\hat{g}(\boldsymbol{X}^{mis})|\mathsf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}]$ (computed earlier in derivation of variability of conditional mean estimate).

Expected variance of conditional mean estimate is computed using approximation 6.13 and first order Taylor approximation, thus yielding to

$$\begin{aligned}
& \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbb{V}\mathrm{ar}[\hat{g}(\boldsymbol{X}^{mis})|\mathsf{x}^{mis}, \mathsf{n}^{mis}, \mathsf{n}] \qquad\right] \\
\approx\ & \mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\left[\mathbf{s}^T\mathbb{V}\mathrm{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})|\mathbf{D}_{\boldsymbol{X}}^{mis}, \mathsf{N}^{mis}, \mathsf{n}]\mathbf{s}\right] \\
& +\mathbb{E}_{\mathsf{N}^{mis},\boldsymbol{X}^{mis}|\mathsf{n}}\Big[\big(g_1(\boldsymbol{X}^{mis}|\mathcal{Z}) \dots g_{\mathsf{n}_c}(\boldsymbol{X}^{mis}|\mathcal{Z})\big)\mathbb{V}\mathrm{ar}[\hat{\mu}_{\{u\}}^{imp}|\mathbf{D}_{\boldsymbol{X}}^{mis}, \mathsf{N}^{mis}, \mathsf{n}] \\
& \big(g_1(\boldsymbol{X}^{mis}|\mathcal{Q}) \dots g_{\mathsf{n}_c}(\boldsymbol{X}^{mis}|\mathcal{Q})\big)^T\Big] \\
\approx\ & \underbrace{\sum_{i=1}^{\mathsf{n}_c} \mu_i^{*imp} g_i'(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})^T\mathbb{V}\mathrm{ar}[vec(\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}})|\mathbb{E}[\mathbf{D}_{\boldsymbol{X}}^{mis}, \mathsf{N}^{mis}|\mathsf{n}]] \sum_{i=1}^{\mathsf{n}_c} \mu_i^{*imp} g_i'(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})}_{\text{expected sampling variance due to estimation of classifier parameters}} \\
& + \quad \underbrace{\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})^T\mathbb{V}\mathrm{ar}[\hat{\mu}_{\{u\}}^{imp}|\mathbb{E}[\mathbf{D}_{\boldsymbol{X}}^{mis}, \mathsf{N}^{mis}|\mathsf{n}]]\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})}_{\text{expected sampling variance due to estimation of imputation model parameters}} \quad , \quad u = 1, \dots, \mathsf{n}_c
\end{aligned}$$

where $\mathbf{s} = \sum_{i=1}^{\mathsf{n}_c} \mu_{i,\mathsf{N}^{obs}}^{*imp} g_i'(\boldsymbol{X}^{mis}|\mathcal{Z})$, in which $\mathcal{Z}$ is conditionalizer $\mathcal{Q}$ from approximation 6.13, $\mathcal{Q} = \{\mathbb{E}[\hat{\boldsymbol{W}}_{\boldsymbol{X},\{u\}}|\mathbb{E}[\mathbf{D}_{\boldsymbol{X}}^{mis}], \mathbb{E}[\mathsf{N}^{mis}], \mathsf{n}]\}$, and
$\mathbf{g}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}) = (g_1(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}), \dots, g_{\mathsf{n}_c}(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q}))^T$.

Expected imputation noise (variance) is approximated as

$$\begin{aligned}
v_{\mathsf{n}}^{*imp} &= \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\epsilon}(\boldsymbol{X}^{mis})|\mathsf{n}] \\
&\approx \sum_{i=1}^{\mathsf{n}_c} g_i(\overline{\boldsymbol{X}}^{*mis}|\mathcal{Q})\mathbb{E}[\hat{\tau}_i^{imp}(\boldsymbol{X}^{mis})|\mathsf{n}].
\end{aligned}$$

A6.3 Second order moments for random number of missing data values within joint $(Y, \boldsymbol{X})$ cells

Here elements of covariance-variance matrix $\mathbb{V}\text{ar}[\boldsymbol{N}^{mis}|\mathcal{Q}_2]$ are computed, where $\boldsymbol{N}^{mis} = (\mathsf{N}_1^{mis}, \ldots, \mathsf{N}_{\mathsf{n}_c}^{mis})^T$. Recall that number of missing data values within cell i was specified as

$$\mathsf{N}_i^{mis} = \sum_{j=1}^{\mathsf{n}_c} \mathsf{N}_{j,i}^{mis},$$

where $\mathsf{N}_{j,i}^{mis}$ is number of observations belonging to cell j but which were classified to cell i.

Multinomial distribution for $\boldsymbol{N}^{mis}|\mathcal{Q}_2$ is given in Equation (6.12). In derivation of second order moments properties of multinomial distribution (variance and covariance) come in handy. See Appendix A3.1.2 for details on multinomial distribution.

Variance of number of incomplete observations in cell i is computed as

$$\begin{aligned}
\mathbb{V}\text{ar}[\mathsf{N}_i^{mis}|\mathcal{Q}_2] &= \mathbb{V}\text{ar}[\sum_{j=1}^{\mathsf{n}_c} \mathsf{N}_{j,i}^{mis}|\mathcal{Q}_2] \\
&= \sum_j \mathbb{V}\text{ar}[\mathsf{N}_{j,i}^{mis}|\mathcal{Q}_2] + \sum_j \sum_{l \neq j} \mathbb{C}\text{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{l,i}^{mis}|\mathcal{Q}_2] \\
&\approx \mathsf{n}^{mis}\frac{1}{z}\pi_i p_i \mathbb{E}[\hat{q}_i](1 - \frac{1}{z}\pi_i p_i \mathbb{E}[\hat{q}_i]) \\
&\quad + \sum_{j \neq i} \mathsf{n}^{mis}\frac{1}{z}\pi_i p_i (1 - \mathbb{E}[\hat{q}_i])(1 - \frac{1}{z}\pi_i p_i (1 - \mathbb{E}[\hat{q}_i])) \\
&\quad + \sum_j \sum_{l \neq j} \mathbb{C}\text{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{l,i}^{mis}|\mathcal{Q}_2] \\
&\approx \underbrace{\mathsf{n}^{mis}\frac{1}{z}\pi_i p_i \mathbb{E}[\hat{q}_i](1 - \frac{1}{z}\pi_i p_i \mathbb{E}[\hat{q}_i])}_{\text{variance due correct classifications}} \\
&\quad + \underbrace{\mathsf{n}^{mis}(\mathsf{n}_c - 1)\frac{1}{z}\pi_i p_i (1 - \mathbb{E}[\hat{q}_i])(1 - \frac{1}{z}\pi_i p_i (1 - \mathbb{E}[\hat{q}_i]))}_{\text{variance due misclassifications}} \\
&\quad + \underbrace{\sum_j \sum_{l \neq j} \mathbb{C}\text{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{l,i}^{mis}|\mathcal{Q}_2]}_{\text{covariance between correct classifications and misclassifications}} \quad ,
\end{aligned}$$

where the last covariance term is

$$\begin{aligned}
&\sum_j \sum_{l \neq j} \mathbb{C}\mathrm{ov}[\mathsf{N}^{mis}_{j,i}, \mathsf{N}^{mis}_{l,i} | \mathcal{Q}_2] \\
\approx\ & \underbrace{-\mathsf{n}^{mis} \sum_{l \neq j, j = i, l \neq i} \frac{1}{z} \pi_i p_i \mathbb{E}[\hat{q}_i] \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_l p_l (1 - \mathbb{E}[\hat{q}_l])}_{\text{covariance between correct classifications to cell i and misclassifications from other cells}} \\
& \underbrace{-\mathsf{n}^{mis} \sum_{l \neq j, l = i, j \neq i} \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_j p_j (1 - \mathbb{E}[\hat{q}_j]) \frac{1}{z} \pi_i p_i \mathbb{E}[\hat{q}_i]}_{\text{covariance between misclassifications from other cells and correct classifications to cell i}} \\
& \underbrace{-\mathsf{n}^{mis} \sum_{l \neq j, j \neq i, l \neq i} \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_j p_j (1 - \mathbb{E}[\hat{q}_j]) \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_l p_l (1 - \mathbb{E}[\hat{q}_l]),}_{\text{covariance between misclassifications}}
\end{aligned}$$

where indexing of the first two sums can be simplified yielding to:

$$\begin{aligned}
&\sum_j \sum_{l \neq j} \mathbb{C}\mathrm{ov}[\mathsf{N}^{mis}_{j,i}, \mathsf{N}^{mis}_{l,i} | \mathcal{Q}_2] \\
\approx\ & \underbrace{-\mathsf{n}^{mis} \sum_{l \neq i} \frac{1}{z} \pi_i p_i \mathbb{E}[\hat{q}_i] \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_l p_l (1 - \mathbb{E}[\hat{q}_l])}_{\text{covariance between correct classifications to cell i and misclassifications from other cells}} \\
& \underbrace{-\mathsf{n}^{mis} \sum_{j \neq i} \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_j p_j (1 - \mathbb{E}[\hat{q}_j]) \frac{1}{z} \pi_i p_i \mathbb{E}[\hat{q}_i]}_{\text{covariance between misclassifications from other cells and correct classifications to cell i}} \\
& \underbrace{-\mathsf{n}^{mis} \sum_{l \neq j, j \neq i, l \neq i} \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_j p_j (1 - \mathbb{E}[\hat{q}_j]) \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_l p_l (1 - \mathbb{E}[\hat{q}_l]) \,.}_{\text{covariance between misclassifications}}
\end{aligned}$$

For $i \neq l$ covariance term is computed as

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[\mathsf{N}_i^{mis}, \mathsf{N}_l^{mis}|\mathcal{Q}_2] &= \mathbb{C}\mathrm{ov}[\sum_j \mathsf{N}_{j,i}^{mis}, \sum_u \mathsf{N}_{u,l}^{mis}|\mathcal{Q}_2] \\
&= \sum_j \sum_u \mathbb{C}\mathrm{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{u,l}^{mis}|\mathcal{Q}_2] \\
&= \sum_{j\neq i} \sum_{u\neq l} \mathbb{C}\mathrm{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{u,l}^{mis}|\mathcal{Q}_2] \\
&\quad + \sum_{j\neq i} \sum_{u=l} \mathbb{C}\mathrm{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{u,l}^{mis}|\mathcal{Q}_2] \\
&\quad + \sum_{j=i} \sum_{u\neq l} \mathbb{C}\mathrm{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{u,l}^{mis}|\mathcal{Q}_2] \\
&\quad + \sum_{j=i} \sum_{u=l} \mathbb{C}\mathrm{ov}[\mathsf{N}_{j,i}^{mis}, \mathsf{N}_{u,l}^{mis}|\mathcal{Q}_2] \\
&\approx -\mathsf{n}^{mis} \sum_{j\neq i} \sum_{u\neq l} \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_j p_j (1 - \mathbb{E}[\hat{q}_j]) \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_u p_u (1 - \mathbb{E}[\hat{q}_u]) \\
&\quad -\mathsf{n}^{mis} \sum_{j\neq i} \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_j p_j (1 - \mathbb{E}[\hat{q}_j]) \frac{1}{z} \pi_l p_l \mathbb{E}[\hat{q}_l] \\
&\quad -\mathsf{n}^{mis} \sum_{u\neq l} \frac{1}{z} \pi_i p_i \mathbb{E}[\hat{q}_i] \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_u p_u (1 - \mathbb{E}[\hat{q}_u]) \\
&\quad -\mathsf{n}^{mis} \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_i p_i (1 - \mathbb{E}[\hat{q}_i]) \frac{1}{z} \frac{1}{\mathsf{n}_c - 1} \pi_l p_l (1 - \mathbb{E}[\hat{q}_l]).
\end{aligned}$$

Appendix for Chapter 8

In this appendix decomposition of weighted mean squared error estimator is given. In addition, it is described how data set used in Chapter 8 is constructed.

A8.1 Decomposition of weighted mean squared error estimator

Let weighted (by sampling weights) mean squared error estimator for N^{mis} imputed observations be

$$\hat{mse} = \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j \left(Y_j - Y_j^{comp}\right)^2,$$

where $\overline{W}_j = \frac{1}{\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} W_j} W_j$.

A possible (reasonable) decomposition of WMSE for situation in which true model is not known and there is no repetitions of imputations at some given covariate positions is given next.

Let weighted means of missing and imputed data be

$$\begin{aligned}\hat{\mu}^{mis} &= \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j Y_j \\ \hat{\mu}^{imp} &= \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j Y_j^{comp}.\end{aligned}$$

Now $\hat{mse}$ is decomposed as

$$\begin{aligned}
\hat{mse} &= \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j \Big(Y_j - Y_j^{comp}\Big)^2 \\
&= \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j \Big(Y_j - \hat{\mu}^{mis} + \hat{\mu}^{mis} - Y_j^{comp}\Big)^2 \\
&= \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j \left[(Y_j - \hat{\mu}^{mis})^2 + (\hat{\mu}^{mis} - Y_j^{comp})^2 + 2(Y_j - \hat{\mu}^{mis})(\hat{\mu}^{mis} - Y_j^{comp})\right] \\
&= \underbrace{\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})^2}_{\text{Weighted sample variance estimator for missing values}} + \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (\hat{\mu}^{mis} - Y_j^{comp})^2 \\
&\quad +2 \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})(\hat{\mu}^{mis} - Y_j^{comp}) \\
&= \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})^2 + \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (\hat{\mu}^{mis} - \hat{\mu}^{imp} + \hat{\mu}^{imp} - Y_j^{comp})^2 \\
&\quad +2 \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})(\hat{\mu}^{mis} - Y_j^{comp}) \\
&= \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})^2 + \underbrace{\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (\hat{\mu}^{mis} - \hat{\mu}^{imp})^2}_{=(\hat{\mu}^{mis} - \hat{\mu}^{imp})^2} \\
&\quad + \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (\hat{\mu}^{imp} - Y_j^{comp})^2 + \underbrace{2 \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (\hat{\mu}^{mis} - \hat{\mu}^{imp})(\hat{\mu}^{imp} - Y_j^{comp})}_{=0} \\
&\quad -2 \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})(Y_j^{comp} - \hat{\mu}^{mis})
\end{aligned}$$

where last sum term is decomposed for better interpretation as

$$\begin{aligned}
& \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})(Y_j^{comp} - \hat{\mu}^{mis}) \\
= & \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})(Y_j^{comp} - \hat{\mu}^{imp} + \hat{\mu}^{imp} - \hat{\mu}^{mis}) \\
= & \sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})(Y_j^{comp} - \hat{\mu}^{imp}) \\
& +(\hat{\mu}^{imp} - \hat{\mu}^{mis}) \underbrace{\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})}_{=0} .
\end{aligned}$$

Therefore terms in $\hat{mse}$ decomposition are

$$\begin{aligned}
\hat{mse} \;=\; & \underbrace{\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})^2}_{\text{=A: weighted variance estimator for missing Y values}} \\
+ & \underbrace{\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j^{comp} - \hat{\mu}^{imp})^2}_{\text{=B: weighted variance estimator for imputed Y values}} \\
-2 & \underbrace{\sum_{j=\mathsf{N}^{obs}+1}^{\mathsf{n}} \overline{W}_j (Y_j - \hat{\mu}^{mis})(Y_j^{comp} - \hat{\mu}^{imp})}_{\text{=C: weighted covariance estimator between missing Y and imputed Y}} \\
+(& \underbrace{\hat{\mu}^{mis} - \hat{\mu}^{imp}}_{\text{=D: squared global bias estimator}})^2.
\end{aligned}$$

When computing expectation of $\hat{mse}$ the squared global bias term (term D) may be decomposed as

$$\begin{aligned}
\mathbb{E}\Big[(\hat{\mu}^{mis} - \hat{\mu}^{imp})^2\Big] &= \mathbb{E}\Big[(\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{imp}] + \hat{\mu}^{mis} - \mathbb{E}[\hat{\mu}^{mis}] + \mathbb{E}[\hat{\mu}^{imp}] - \hat{\mu}^{imp})^2\Big] \\
&= (\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{mis}])^2 + \mathbb{E}\Big[(\hat{\mu}^{mis} - \mathbb{E}[\hat{\mu}^{mis}] + \mathbb{E}[\hat{\mu}^{imp}] - \hat{\mu}^{imp})^2\Big] \\
&\quad + \underbrace{2\Big(\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{imp}]\Big)\mathbb{E}\Big[\hat{\mu}^{mis} - \mathbb{E}[\hat{\mu}^{mis}] + \mathbb{E}[\hat{\mu}^{imp}] - \hat{\mu}^{imp}\Big]}_{=0} \\
&= (\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{mis}])^2 \\
&\quad +\mathbb{E}\Big[(\hat{\mu}^{mis} - \mathbb{E}[\hat{\mu}^{mis}])^2 + (\mathbb{E}[\hat{\mu}^{imp}] - \hat{\mu}^{imp})^2 \\
&\quad +2(\hat{\mu}^{mis} - \mathbb{E}[\hat{\mu}^{mis}])(\mathbb{E}[\hat{\mu}^{imp}] - \hat{\mu}^{imp})\Big] \\
&= (\mathbb{E}[\hat{\mu}^{mis}] - \mathbb{E}[\hat{\mu}^{mis}])^2 + \mathbb{V}\mathrm{ar}[\hat{\mu}^{mis}] + \mathbb{V}\mathrm{ar}[\hat{\mu}^{imp}] \\
&\quad -2\,\underbrace{\mathbb{E}\Big[(\hat{\mu}^{mis} - \mathbb{E}[\hat{\mu}^{mis}])(\hat{\mu}^{imp} - \mathbb{E}[\hat{\mu}^{imp}])\Big]}_{\mathbb{C}\mathrm{ov}[\hat{\mu}^{mis},\hat{\mu}^{imp}]} .
\end{aligned}$$

A8.2 Construction of data set

The data construction phases are:

1. Read variables O2, S4_SIZE, PWEIGHT, S11, P1, P2, E3D, E5 from tabulator separated file `uksmef2004.tab` (1st edition, dated February 1st / 2006)

2. Pick observations for which turnover (O2) is not missing, variable S4_SIZE is not equal to values -2, -1 or 1, and variables S11, P1, P2, E3D, E5 do not have values -2, -1 or 0.

Appendix for Chapter 9

This appendix contains some details for Chapter 9. At first, issues with documentation of data set which is used in the chapter are discussed about. Construction of data set which is used in experiments is described, as is the missingness generator for SEX variable. Brief descriptions of variables for age experiment are given. Finally, variables which are treated as continuous in the experiments are listed.

A9.1 Issues with documentation of the data set

Documentation supplied with the Quaterly Labour Force Survey Household 2006 data set was partially out-dated. Namely, details on LFS variables and derived variables were from year 2003. Fortunately, version of details of LFS variables for year 2006 was available from National Statistics[1]. Even with documentation for year 2006 there were some issues. Namely, variables WKAGG1664, LEVQUAL6 and ICOD92 were not documented. First of these variables was not used at all, whereas the other two variables were used. Fortunately, LEVQUAL5 which is level of highest qualification held (year 2005) is documented. Therefore author assumes that LEVQUAL6 is level of highest qualification held for year 2006, and that its coding is same as for LEVQUAL5. Variable ICOD92 is assumed to be industry of current or last job. Its values map to Standard Industrial Classification of economic activities 1992 (SIC92) classes. Documentation did not include mappings from ICOD92 values to SIC92 classes. However, such a mapping was available in internet[2].

A9.2 Construction of data set

In data construction we remove system variables, derived variables, and variables with zero variance (no information). Further variables WKAGG1664 (unknown variable), DVHRPNUM (counter variable), IOUTCOME (contains AGE information class), and FAMUNIT (counter variable) are removed. Main reason for removal of variables is to remove duplicate information about variables AGE and SEX, and to focus on information in answers to questionnaires. Undocumented variable WKAGG1664 is removed as it may contain information about AGE.

[1] www.statistics.gov.uk/downloads/theme_labour/LFSUG_Vol3.pdf (checked 06.03.2007)

[2] Do Google search using keyword `uk97delessification.htm` and use cached link to site: www.lisproject.org/les/uk/uk97delessification.htm (checked 06.03.2007)

Data set for age experiment is constructed using the following phases:

1. Read tabulator separated file `aj06hp.tab` (1st edition, dated October 16th / 2006)

2. Convert SNGDEG variable from string format to integer valued as described in A9.3

3. Remove system variables:

   > QUOTA, WEEK, W1YR, QRTR, WAVFND, HHLD, THISWV, REFDTE, REFWKD, REFWKM, REFWKY, NUMHHLD, NURSE, HOUT, PERSNO, RESPNO, RECNO, ADD, TYPINT, HALLRES, DOBM, DOBY

4. Remove derived variables (year 2003 documentation):

   FUSERIAL, HSERIAL, INECAC05, AOFL16, AOFL19, AOHL16, AOHL19, AYFL19, AYHL19, CAIND, EXTFU, FDPCH15, FDPCH16, FDPCH19, FDPCH2, FDPCH4, FDPCH9, FMDP, FMNDP, FMPLUS, HDPCH19, HOHID, RELHFU, SMSXFU, TOTFU, TOTNUM, TOTXFU, XFMDC, XFMNDC, AGEDFE, AGES, BACTHR, BUSHR, CLAIMS, CRYOX, CURED, DISCURR, DURUN, DURUN2, EMPLEN, EMPMON, ETHCEN15, ETHCEN6, FLED9D, FLEXW1, FLEXW2, FLEXW3, FLEXW4, FLEXW5, FLEXW6, FLEXW7, FLEXW8, FLEXW9, FTPT, FTPTW, GB, GOR3, GORONE, GOVTOF, GOVTOF2, GOVTOR, HRP, ILLFRI, ILLMON, ILLOFF, ILLSAT, ILLSUN, ILLTHU, ILLTUE, ILLWED, INDD92L, INDD92M, INDD92S, INDG92L, INDG92M, INDG92S, INDM92L, INDM92M, INDM92O, INDM92S, INDS92L, INDS92M, INDS92S, INDSECT, LKWFWM, NATIDB, NATIDE, NATIDI, NATIDO, NATIDS, NATIDW, NATOX, NSECM, NSECMMJ, PAIDHRA, PAIDHRU, PRXREL, REDCLS, REDUND, REG3, REGONE, SAMELAD, SC2KLMJ, SC2KLMN, SC2KMMJ, SC2KMMN, SC2KOMJ, SC2KOMN, SC2KSMJ, SC2KSMN, SIC80L, SIC80M, SIC80O, SOC2KAP, SOC2KL, SOC2KM, SOC2KO, SOC2KR, SOC2KS, STUCUR, SUMHRS, TOTHRS, TTACHR, TTUSHR, TYEMPS, URESMC, WCHFR, WCHMO, WCHSA, WCHSU, WCHTH, WCHTU, WCHWE, WKFRI, WKMON, WKSAT, WKSUN, WKTHU, WKTUE, WKWED, WNLEFT, WNLEFT2, WRKAGE, XDISDDA

5. Remove additional derived variables (year 2006 documentation):

   MARDY6, MARSEX6, HHTYPE6, HDPCH4, HDC515, HDPCH18, HEACOMB, HEAHEAD, HEAWIFE, HNFTSTUD, HNOTSTUD, FUTYPE6, ILODEFR, PUBLICR, MPNR02, RESTMR6, REDINDYR, REGWKR, GORWKR, DIFFHR6, SECJMBR, MPNSR02, REGWK2R, GORWK2R, MPNLR02, OYMPR02, HIQUAL5, HIQUAL5D, HITQUA5,

6. Remove variables: WKAGG1664, DVHRPNUM, FAMUNIT, IOUTCOME

7. Remove variables with zero variance (no information):

   XR13, XR14, XR15, PRIVCL7, PRIVCL8, REDMPN2, HOMED23, UNDY986, UNDY987, UNDY988, UNDY989, NOLOWA06, NOLOWA07, NOLOWA08, NOLOWA09, NOLOWA10, METHSE5, METHSE6, METHAL11, METHAL12, METHAL13, METHAL14, QLSTY606, QLSTY607, QLSTY608, QLSTY609, QLSTY610, QLSTY611, CMBDEG11, CMBDEG12, QGCSE44, QGCSE45, GCSEFUL4, GCSEFUL5, SUBCOD8, SCQUL, GNVQUL5, RSAQUL, HSTQUL, HEAL16, HEAL17, HEALPB08, HEALPB09, HEALPB10, SKDSBN36, SKDSBN37, PENBEN34, FAMLY032, FAMLY033, TPBEN37, TPBEN38, SUBNO8

8. Remove two observations having value -9 in variable AGE

A9.3 Coding of SNGDEG variable

Original values, excluding special values, for variable SNGDEG were given in following forms: V1, V1.V2, V1.V2.V3, and V1.V2.V3.V4, in which V1-V4 are two digit positive numerical values, and V1 is most significant level and V4 is least significant level. Due to technical reason (indicator variable coding routine could not handle non-numeric values) the non-numeric answers were coded as integer values. Coding to integer values is described next. All original values can be presented in hierarchical form as V1.V2.V3.V4 as follows:

ORIGINAL VALUE	**TRANSFORMED VALUE V1.V2.V3.V4**
X	X.0.0.0
X.Y	X.Y.0.0
X.Y.Z	X.Y.Z.0
X.Y.Z.W	X.Y.Z.W

Finally, SNGDEG is coded as:

$$\text{VALUE} = V4 + V3 * 100 + V2 * 10000 + V1 * 1000000.$$

A9.4 Missingness generator for SEX experiment

Missingness of SEX depends on FTPTWK (whether full or part time in main job) and BENFTS (whether claiming any State Benefits/Tax credits) variables.

Probability for nonresponse of SEX as function of FTPTWK and BENFTS is:

		FTPTWK=FULL-TIME	PART-TIME	-8/-9
BENFTS=	YES	0.7	0.05	0.05
	NO	0.9	0.05	0.15
	-8/-9	0.6	0.2	0.1

Remark that -8 (no answer) and -9 (not applicable) are special values. Note that missing-data mechanism is MCAR within (FTPTWK, BENFTS) cells. This is a MAR mechanism because response probabilities vary within cells.

A9.5 Descriptions of Labour Force Survey Household dataset 2006 variables for AGE experiment

Following five tables contain variables which were used in the AGE experiment. Short descriptions of the variables are given. Note that these are coded variables, which include individual categories and special values.

Variable	Description
AGE	Age of respondent (variable of interest).
ACTWKDY2_4	Days scheduled to work [Thursday]
ACTWKDY2_7	Days scheduled to work [Sunday]
ATTEND_-9	Whether still attending education course [Not applicable]
CCTC5_-9	Child care tax credit [Not applicable]
CMBDEG01_8	1. subject area of combined subject degree [Technology]
CMBDEG01_16	1. subject area of combined subject degree [Humanities]
CMBDEG05_-9	5. subject area of combined subject degree [Not applicable]
CMBMAIN_7	Main subject area studied in qualification [Engineering]
CMBMAIN_10	Main subject area studied in qualification [Social sciences]
CRY01_59	Country of birth [Other (than available options)]
EDAGE	Age when completed full time education
ENROLL_-9	Whether enrolled on education course [Not applicable]
EVERWK_1	Ever had a paid job or place on scheme [Yes]
EVERWK_-9	Ever had a paid job or place on scheme [Not applicable]
FAMLY031_-9	Type of family related 1. benefit claimed [Not applicable]
FUTUR13_2	Job related training or education in the last 13 weeks [No]
GCSEFUL1_-9	Type of 1. GCSE or equivalent held above grade C/1 [Not applicable]
HEAL02_1	What 2. health problem does respondent have [Problems or disabilities (including arthritis or rheumatism) connected with ...arms or hands]
HEAL03_2	What 3. health problem does respondent have [Problems or disabilities (including arthritis or rheumatism) connected with ...legs or feet]
HEAL04_10	What 4. health problem does the respondent have [Stomach, liver kidney or digestive problems]
HEALPB01_3	1. health problem/disability that affected respondent in the past [Problems or disabilities (including arthritis or rheumatism) connected with ...back or neck
HEALYR_2	Any other past health problems or disabilities that have lasted longer than a year [No]
HHWT03	Household weight (sampling information)

Variable	**Description**
HLDCMP6_2	Composition of household [1 female over pensionable age with no children]
HLDCMP6_7	Composition of household [Cohabiting couple both under pensionable age with no children]
HLDCMP6_9	Composition of household [Cohabiting couple one or more over pensionable age with no children]
HNFTIME	Number of people in household who are working full-time
HNPEN	Number of people in household who are of pensionable age
HNWKAGE	Number of people in household who are of working age
HNWOTH05_-9	Number of people in household who are inactive for other reasons but would like to work [Not applicable]
HOME_4	Whether working from home in main job [Somewhere quite separate from home]
HOME_-9	Whether working from home in main job [Not applicable]
HSNGNI_-9	Whether receiving rent or rate rebate (NI) [Not applicable]
ICOD92_331	Industry of last or current work [Standard Industrial Classification (SIC92) code 60.23: Other passenger land transport]
ICOD92_382	Industry of last or current work [72.50: Repair of office, computer eqt]
ICOD92_423	Industry of last or current work [85.12: Medical practise activities]
JBAWAY_2	Whether temporarily away from paid work [No]
JSADUR_8	Length of time claiming Job Seekers Allowance and/or NI Credits [3 years but less than 4 years]
LEFTYR_-9	Year left last job [Not applicable]
LEVQUAL6_1	Level of highest qualification held? [NVQ level 4 and above]
LIMITA_1	Whether health problem affects the amount of paid work that can be done [Yes]
LIMITA_-9	Whether health problem affects the amount of paid work that can be done [Not applicable]
LIVWTH_1	Whether living together as couple [Yes]
LIVWTH_-9	Whether living together as couple [Not applicable]
LKTIMB_3	How long looking for work [1 month but less than 3 months]

Variable	**Description**
LLORD_6	Landlord of accommodation [Individual employer]
LNGLIM_1	Whether health problem lasting more than 12 months [Yes]
LOOK4_2	Whether looking for paid work in last four weeks [No]
LOOKM2_6	2. reason for looking for different job [Respondent wants to work shorter hours than in present job]
M3CRYO_52	Country of residence 3 months ago [United States of America]
MAINDRV_-9	Driver with most mileage [Not applicable]
MAINDRV2_5	Driver with 2. most mileage [Person 5].
MANAGLR_1	Managerial status last job (reported) [Manager]
MANAGLR_-9	Managerial status last job (reported) [Not applicable]
MARCHK_1	Whether spouse is household member [Yes]
MARSTA_1	Marital status [Single, never married]
METHAL02_2	2. method of looking for work (no preference) [Visit a Jobclub]
METHMP04_7	Method of looking for work (employees or Government scheme) [Study situations vacant in newspapers or journals]
NATO_84	Nationality (other) [Portugal (inc. Azores & Madeira)]
NOLWM_3	Main reason not looking for work in last 4 weeks [Looking after the family/home]
NOLWM_-9	Main reason not looking for work in last 4 weeks [Not applicable]
NUMAS_1	Number of A-S levels [1 A-S level]
NVQSVQ_-9	Whether respondent has any full NVQs or SVQs [Not applicable]
NVQUN_-9	Whether respondent has any units towards NVQs or SVQs [Not applicable]
OYCIRC_-9	Circumstances twelve months ago [Not applicable]
OYCIRC_10	Circumstances twelve months ago [Retired from paid work]
OYCRY_1	Country of residence 12 months ago [UK]

Variable	Description
OYSOLO_1	On own or with employees 1 year ago [Alone or with partner(s) but not employees]
QGCSE41_1	Type of GCSE or equivalent held below grade C/1 [GCSE s below grade C]
QGNVQ_-9	Whether respondent has any GNVQs/GSVQs [Not applicable]
QUALCH53_-9	Holds 3. educational/training qualification from [Not applicable]
QUALS601_16	Type of 1. qualification already held [AS-level/Vocational AS-level or equivalent]
QUALS602_18	Type of 2. qualification already held [Access to HE]
QUALS602_21	Type of 2. qualification already held [GCSE/Vocational GCSE]
QUALS603_17	Type of 3. qualification already held [Certificate of sixth year studies (CSYS) or equivalent]
QUALS604_8	Type of 4. qualification already held [Nursing or other medical qualification not yet mentioned]
QULHI4_-8	What highest qualification current study towards [No answer]
RELBUS_2	Whether doing unpaid work for relative's business [No]
RELH06_0	Relationship to head of household [Head of household]
RELH06_3	Relationship to head of household [Child]
RELHRP6_3	Relationship to household reference person [Child]
RELIG_1	Religion [Christian]
RESTME_6	Length of time at this address [10 years or longer]
SCHM04_66	Government employment and training programme [None of available options]
SECJOB_2	Whether had second job in reference week [No]
SEX_1	Sex of respondent [Male]
SNGDEG_10040303	Subject of single subject degree [Social Policy / Education Policy]
SNGDEG_18010100	Subject of single subject degree [Creative Arts & Design / Drawing]
SNGDEG_6010201	Subject of single subject degree [Mathematical & Computer Sciences / Mechanics (Mathematical)]

Variable	**Description**
SNGDEG_6040200	Subject of single subject degree [Computer Science / Networks & Communications]
SNGDEG_7020402	Subject of single subject degree [Civil Engineering / Engineering Surveying]
STAT_1	Employment status [Employee]
SUBCOD1_21,1	Area of 1. study [Arts / Fine Arts].
TEACH41_-9	Type of 1. teaching qualification already held [Not applicable].
TECLEC4_-9	On scheme run by a TEC or LEC [Not applicable]
TEN1_2	Accommodation details [Being bought with mortgage or loan]
TOTUS1	Total usual hours worked excluding lunch breaks (no overtime)
TPBEN31_4	Type of 1. benefit claimed [State pension]
TPBEN31_-9	Type of 1. benefit claimed [Not applicable]
TPBEN32_3	Type of 2. benefit claimed [Sickness or disability (excluding tax credits)]
TRSITE_9	Main place of education or training in work [At home (OU, Open Tech, correspondence course)]
TYPVCL3_1	Type of 3. vehicle [Car]
UNDABL_2	Whether employer able to increase hours [No]
UNDEMP_-9	Whether would like to work longer hours, at current basic rate of pay, given the opportunity [Not applicable]
USEVCL_2	Own or use motor vehicle [No]
XR01_3	Relationship to person 1 [Natural son or daughter]
XR01_-9	Relationship to person 1 [Not applicable]
XR02_-9	Relationship to person 2 [Not applicable]

A9.6 Variables which are treated as continuous

Following variables were treated as continuous (possible special values coded as indicator variables) in experiments with the QLFS data set:

> LEFTYR, CAMEYR, CONMPY, CONSEY, TOTUS1, USUHR, POTHR, UOTHR, TOTUS2, TOTAC1, ACTHR, ACTPOT, ACTUOT, TOTAC2, ACTHR2, UNDHRS, OVHRS, TRHR93, TRONJB, EDAGE, YERQAL2, YERQAL3, TOTNUM, TOTFU, FMDP, FMNDP, FMPLUS, TOTXFU, XFMDC, XFMNDC, NFAMHH, NPERSFM, NPERSHH, HDPCH19, HDPCH4, HDC515, HDPCH18, HNWKAGE, HNPEN, HNDK, HNEMP, HNUNEMP, HNINAC05, HNINACT, HNFTSTUD, HNOTSTUD, HNFTIME, HNPTIME, HNIWSTU, HNIWSKD, HNIWDSC, HNIWFAM, HNWOTH05, HNIWOTH, HNNOWK05, HNINOWK, FDPCH2, FDPCH4, FDPCH9, FDPCH15, FDPCH16, FDPCH19, ONETEN, EMPMON, ILLOFF, SUMHRS, DAYSPZ, HOLS, BNKHOLF, TOTWRK, TRNDAY, LEISHRS, EDHRS, THRS, T4HRS, NUMILL