JYVÄSKYLÄN YLIOPISTO
MATEMATIIKAN JA TILASTO-
TIETEEN LAITOS

Master's Thesis in Statistics and Data Science

Estimating prediction error variances of a plant breeding hybrid model using Monte Carlo sampling

Antero Heikkilä

20.9.2024

UNIVERSITY OF JYVÄSKYLÄ
Department of Mathematics and Statistics
Antero Heikkilä: Estimating prediction error variances of a plant breeding hybrid model using Monte Carlo sampling, Master's thesis in Statistics and Data Science, 49 pages + 42 appendix pages
20.9.2024

Abstract

Genomic best linear unbiased prediction (GBLUP) is a method widely used in animal and plant breeding. It uses individuals' genomic information to estimate breeding values. Breeding values are an essential part of animal and plant breeding, and they tell the genetic merit of an individual compared to the others. Using estimated breeding values (EBVs), breeders can select the best individuals to be the ancestors of the next generation. To estimate breeding values accurately, relationship information from the breeding population should be used. A relationship matrix is constructed using either pedigree or genetic information. In GBLUP, the relationships of a population are presented in a genomic relationship matrix, which is constructed using the individuals' genetic information. The genomic information is usually based on single nucleotide polymorphisms (SNPs), which tell the variant of a gene an individual carries.

A linear mixed model is a typical choice for estimating breeding values. Individual breeding values are treated as random effects in the linear mixed model. Using Henderson's mixed model equations (MMEs) makes it possible to obtain the estimates for the fixed and random effects simultaneously. A hybrid model in plant breeding is a linear mixed model in which phenotypic observations are explained by both maternal and paternal effects separately and a cross effect. A cross is a plant that emerges when two plants reproduce. This thesis shows how a hybrid model is fitted using a GBLUP model.

When the number of individuals is large, the use of exact solving methods becomes computationally infeasible, making the use of iterative solving methods for solving the MME and approximate methods for obtaining prediction error variances (PEVs) necessary. The behaviour of four methods for approximating PEVs was studied using a hybrid model. The methods are called PEV1, PEV2, PEV3, and NF2, and they are widely used methods to approximate the exact PEV of a model. PEV measures the accuracy of an EBV. These methods, which are based on Monte Carlo (MC) sampling of the model, were compared across different genetic groups and situations. The results indicate that the methods PEV3 and NF2 work better than the methods PEV1 and PEV2. Especially the method PEV2 behaved poorly when the distribution of the exact PEV values was narrow. Overall, the thesis demonstrates that all the methods work in a hybrid model framework when the MC sample size is large enough.

Keywords: plant breeding, prediction error variance, estimated breeding value, GBLUP, linear mixed model, Monte Carlo sampling

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Antero Heikkilä: Kasvinjalostuksessa käytettävän hybridimallin ennustevirhevarianssien estimointi hyödyntäen Monte Carlo -simulointia, Pro gradu -tutkielma, tilastotiede ja datatiede, 49 sivua + 42 liitesivua
20.9.2024

Tiivistelmä

Jalostusarvoilla ilmaistaan yksilön geneettistä hyvyyttä jalostettavan ominaisuuden suhteen verrattuna muihin yksilöihin jalostettavassa populaatiossa. Seuraavan sukupolven vanhemmiksi valitaan tyypillisesti yksilöt, joiden jalostusarvojen ennusteet ovat suurimmat toivoen, että heidän jälkeläisilläänkin olisi hyvät ominaisuudet jalostettavan ominaisuuden suhteen. G-BLUP (genomic best linear unbiased prediction) -menetelmä on laajasti käytössä eläin- ja kasvinjalostuksessa. Siinä jalostusarvojen ennustamiseen käytetään yksilöiltä kerättyä geneettistä tietoa. Jotta jalostusarvojen ennusteet olisivat mahdollisimman hyviä ja tarkkoja, on tärkeää, että populaation sukulaisuussuhteet tiedetään. Erityisesti eläinpopulaatioissa on tavallisesti tiedossa populaation sukupuu, jonka avulla jalostusarvojen ennustamiseen käytetyissä menetelmissä, kuten BLUP- ja G-BLUP -menetelmässä, voidaan muodostaa niissä tarvittava sukulaisuusmatriisi. Nykyisin, kun yksilöiden genotyypittämisen hinta on laskenut, on entistä yleisempää muodostaa sukulaisuusmatriisi hyödyntäen yksilöiltä kerättyä tietoa snipeistä (SNP), eli yhden nukleotidin polymorfismeista. Snipit ovat edustava otos genomia, ja kuvaavat siinä olevaa geneettistä vaihtelua.

Tilastollisena mallina jalostuksessa käytetään tavallisesti lineaariseen sekamalliin pohjautuvaa mallia. Siinä yksilöiden fenotyyppisiä havaintoja selitetään joukolla kiinteitä tekijöitä, kuten ikää, sukupuolta ja painoa, ja satunnaistekijöitä. Satunnaistekijöinä mallissa ovat erityisesti yksilöiden jalostusarvot, joten ratkaisemalla satunnaistekijöiden ennusteet saadaan ennusteet jalostusarvoille. Kasvinjalostuksessa käytettävässä hybridimallissa satunnaistekijöitä on usein kolme: risteytyksen molempien vanhempien sekä itse risteytyksen satunnaisvaikutus fenotyyppiseen havaintoon. Tässä tutkielmassa hybridimalli sovitetaan käyttäen G-BLUP -menetelmää.

Tutkielman varsinaisena tavoitteena oli selvittää, miten ennustevirhevariansseja (PEV) approksimoivat menetelmät toimivat hybridimallin kanssa. Ennustevirhevarianssilla mitataan sitä, kuinka lähellä jalostusarvon ennuste on todellista jalostusarvoa. Approksimoivat menetelmät perustuvat mallin simuloimiseen Monte Carlo -menetelmällä. Approksimoivien menetelmien toimivuutta tutkittiin kolmen geneettisen ryhmän välillä, jotka olivat risteytyksen vanhempaiskasvit ja risteytys itse, jonka lisäksi tutkittiin, miten menetelmät toimivat tilanteessa, joissa geneettisiä varriansseja ja jäännösvarianssia muutettiin, ja tilanteessa, jossa analyysiin otettiin mukaan vain puolet havainnoista. Tutkielmaan otettiin mukaan neljä tunnettua menetelmää, joita kutsutaan nimillä PEV1, PEV2, PEV3 ja NF2. Menetelmät perustuvat mallin simuloimiseen ja niissä verrataan simuloidun jalostusarvon ja simuloidun datan perusteella saadun jalostusarvon estimaatin välistä eroa. Tämä tutkielma osoitti, että kaikki (tutkittavat) ennustevirhevarianssia approksimoivat menetelmät toimivat asymptoottisesti Monte Carlo -näytteiden määrän kasvaessa myös hybridimallin kanssa. Tutkielmassa kuitenkin selvisi, että menetelmien välillä on myös eroja. Parhaimmiksi havaittiin menetelmät PEV3 ja NF2. Sen sijaan erityisesti menetelmä PEV2 toimi huonosti tilanteessa, jossa ennustevirhevarianssin vaihteluväli oli pieni.

Avainsanat: kasvinjalostus, ennustevirhevarianssi, jalostusarvon ennuste, G-BLUP, lineaarinen sekamalli, Monte Carlo -simulointi

Contents

Acronyms

BLUP Best linear unbiased prediction.

EBV Estimated breeding value.

GBLUP Genomic best linear unbiased prediction.

GCA General combining ability.

GEBV Genomic estimated breeding value.

LHS Left-hand side of the MME.

MAD Maximum absolute difference.

MAF Minor allele frequency.

MC Monte Carlo.

MME Mixed model equation.

MVN Multivariate normal distribution.

NF2 Approximated prediction error variance using Hickey's New Formulation 2.

PEV Prediction error variance.

PEV1 Approximated prediction error variance using García-Cortés' method 1.

PEV2 Approximated prediction error variance using García-Cortés' method 2.

PEV3 Approximated prediction error variance using García-Cortés' method 3.

RHS Right-hand side of the MME.

RMSE Root mean square error.

SCA Specific combining ability.

SNP Single nucleotide polymorphism.

1 Introduction

The increasing number of genotyped individuals in plant and animal breeding models causes computational challenges when using exact methods to calculate estimated breeding values (EBVs) and their reliabilities. The reason for this is the size of the matrices required to compute the breeding value estimates. Inverting large matrices takes time and memory; even relatively small matrices cannot be inverted in a feasible amount of time. The time complexity, even for the best algorithms for matrix inversion and multiplication, is clearly over $\mathcal{O}(n^2)$, where n is the dimension of the matrix (Ambainis, Filmus, and Le Gall, 2015). Since the size of the matrices is directly related to the number of animals and plants, the limit of direct matrix inversion is reached quickly. Therefore, there is a need to develop both algorithms for estimating the breeding values, such as preconditioned conjugate gradient (Strandén and Lidauer, 1999), and for approximating the reliabilities of the breeding values, such as Monte Carlo (MC).

In the first part of this thesis, we will introduce the basics of animal and plant breeding models so that statisticians unfamiliar with the field of breeding will understand the basic concepts. The basics of animal and plant breeding are discussed in Section 2. As gene technology in animal and plant breeding models has increased rapidly, we will explain how genetic information can be included in the models. The main focus of this thesis is the accuracy and reliability of the EBVs, especially the concept of prediction error variance (PEV). We will show with examples of how we can express the relationships of animals and plants using either pedigree or genetic information of the individuals. A useful reference for animal breeding is provided by Juga et al. (1999).

Charles Roy Henderson (1911–1989) invented Henderson's mixed model equations (MMEs) in the mid-20th century. Using Henderson's MMEs, it is possible to simultaneously obtain the estimates for the fixed and random effects of a linear mixed model. After inventing Henderson's MMEs, Henderson developed many other important and useful methods and practices, which animal and plant breeders have used since then. Even though it sounds like a cliché, it can be said that Henderson is the father of animal breeding (Van Vleck, 1998).

Relationship matrices play an essential role in the models extensively used in animal and plant breeding. In Section 3, we will explore the construction of relationship matrices using two alternative methods. Most animal and plant breeding models are based on best linear unbiased prediction (BLUP), which is discussed in more detail in Section 4. We will also examine the extensions of this basic breeding model, namely genomic best linear unbiased prediction (GBLUP) and SNP-BLUP, where SNP stands for single nucleotide polymorphism. The different models used in animal breeding are discussed thoroughly by Mrode and Thompson (2014).

After the first part of the thesis, we will apply these models in practice using exact methods on data supplied by Natural Resources Institute Finland. The data are simulated and try to mimic real-world data as well as possible. A GBLUP model is applied to the data in Section 5. When using simulated data, we know and can decide the parameters, such as genetic variances. As implied, this thesis aims to evaluate how well the approximate methods for obtaining the accuracies of the EBVs work and how many MC samples are needed to achieve sufficient results. The MC simulation is presented in Section 6.

One of the first methods for estimating PEVs using resampling was presented by García-Cortés et al. (1995), and some new formulations were introduced by Hickey et al. (2009). Hickey et al. (2009) compared ten different MC-based methods for computing the approximated PEVs, and this thesis will have a similar perspective when comparing these methods. Although the methods have been tested and used for tens of years, they have not been used with a hybrid model in plant breeding. This thesis will show how these methods work with a hybrid model. The results are presented in Section 7. On the whole, this thesis provides an introduction to statistical models in animal and plant breeding and creates a foundation for further research on this topic. In Section 8, we will summarize the thesis and consider what could be done in the future.

This thesis was conducted in collaboration with Natural Resources Institute Finland. The author expresses gratitude to research professors Ismo Strandén and Martin Lidauer, as well as research scientist Anna-Kaisa Ylitalo, for providing the topic and data and guiding the project. The author also wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources.

The analysis was done with the statistical software R (R Core Team, 2023). Grammarly was used to improve the grammar of this thesis (Grammarly, 2024). ChatGPT was used to edit the drawing code for the correlation plots and scatterplots drawn with the function *ggplot2* in R (OpenAI, 2024; Wickham, 2016).

2 Basics of animal and plant breeding models

The main goal in animal and plant breeding is to improve a population by selecting the best individuals to be the ancestors of the next generation. The selection is usually based on EBVs, which are calculated for each individual. EBVs measure the genetic potential of an individual relative to the population to which it belongs. Breeding values in a statistical context are discussed in Isik, Holland, and Maltecca (2017) and Mrode and Thompson (2014). We will introduce two crucial terms to understand breeding: phenotype and genotype. A phenotype is "an organism's appearance or observable traits" (Campbell et al., 2018, pp. 319-326). Examples of phenotypes are the colour of a flower, the height of a human, the milk production of cows, or basically any physical characteristic. In practice, a phenotype is a characteristic that can be observed and measured. Phenotypes are usually affected by both genetic and environmental factors. For example, the weight of an animal depends on both its genes and environmental factors, such as the amount and quality of food available. Conversely, a genotype is an individual's "genetic makeup" (Campbell et al., 2018). It refers to an individual's genome, which is the inherited genetic information, including genes and DNA. The phenotypes of two individuals can be the same, even if their genotypes differ. For example, two flowers can have the same colour while having different genotypes. More precisely, these two flowers have different alleles, which are alternative versions of a gene. Therefore, different genotypes do not always result in different phenotypes. Conversely, the same genotype can result in a different phenotype due to environmental factors.

Breeders can base their decisions on one or more phenotypes or traits they want to breed in their population. In animal and plant breeding, breeders can use single or multiple-trait models. The difference between these two is that there is only one response variable in a single-trait model, while in a multiple-trait model, there are two or more response variables. A single-trait model is less complicated to fit, but with multiple-trait models, it is possible to get more accurate results.

In general, we want to determine how the variation in a phenotype is divided between genotypic and environmental variations. In other words, we want to know how much genes and genotype affect the phenotype compared to how much environmental factors, such as food, sunlight and maternal effects, affect the phenotype. The more the phenotype is affected by the genotype, the easier it is for breeders to estimate breeding values of the next generation's parents based on genetic information. On the other hand, if the phenotype is affected more by environmental factors, then the selection of the parents for the next generation based on genotypes is not as effective, and breeders should instead pay attention to a favourable environment.

In the next subsections, we will examine breeding values and genetic models more closely; the subsections are mainly based on Mrode and Thompson (2014).

2.1 Breeding value

A breeding value is an individual's genetic merit for a trait, which can be weight, milk production, fertility, or any other physical characteristic. The breeding value can be specific to only one trait or a combined value for many different traits. The breeding value indicates how good or bad an individual is compared to other individuals in the population. Breeding values cannot be measured directly from an animal or a plant but can be estimated. We will denote the true breeding value as a and the EBV as $\hat{a}$. Usually, breeders select individuals with the highest EBVs to reproduce with other individuals with high EBVs, hoping the offspring will also have high breeding values for the trait(s) of interest.

EBVs are presented in the same unit as the measurable trait. For example, if the trait is weight in kilograms, the breeding values are also in kilograms. Breeding values are usually presented as deviations from the population mean. Therefore, positive EBVs indicate that the individual has a better chance of producing offspring with good characteristics for the desired trait. There is a connection between an individual's breeding value and the breeding values of its sire and dam. Since an individual inherits half of its genes from the sire and half from the dam, on average, the breeding value of an individual is approximately the average of its parents' breeding values. However, since genetic variation exists between the genes that parents pass to their offspring, an individual's breeding value is not exactly the average of its parents' breeding values. The difference is called Mendelian sampling.

Thus,

$$a_i = \frac{1}{2}a_{i,s} + \frac{1}{2}a_{i,d} + m_i, \tag{1}$$

where a_i is the breeding value of individual i and $a_{i,s}$, and $a_{i,d}$ are the breeding values of its sire and dam, respectively. The last term, m_i, is the Mendelian sampling term, which is assumed to follow a normal distribution with mean zero.

For example, if the trait of interest is milk production per 305 days, and a cow has a breeding value of 200 litres of milk per 305 days, then the expected breeding value of its offspring would be 100 litres of milk per 305 days. This means that, on average, offspring of that cow would produce 100 litres more milk per 305 days compared to the offspring of a cow with a breeding value of 0.

2.2 Basic genetic model

As said in the previous section, genetic and environmental effects affect an individual's phenotype. Nevertheless, as usual in statistics and life, there is always an element of chance and unpredictability that cannot be attributed to genetic or environmental factors. It is something that our model can not explain, and we usually refer to it as a residual, which is the difference between our estimate and the true value. Mrode and Thompson (2014) define the basic genetic model as follows:

$$\begin{gathered}\text{Phenotypic observation} = \text{ Systematic environmental effects } + \\ \text{Genetic effects } + \text{ Random environmental effects,}\end{gathered}$$

or more mathematically

$$y_{ij} = \mu_i + g_i + \epsilon_{ij},$$

where y_{ij} is the jth record for the individual i, μ_i represents the environmental fixed effects for the individual i, like location or year of birth, g_i is the sum of the additive ($g_{a,i}$), dominance ($g_{d,i}$), and epistatic ($g_{e,i}$) genetic effects of the genotype of individual i, and ϵ_{ij} is the sum of random environmental effects. The additive genetic effect is the average additive effect of genes the individual inherits from its parents, so $g_{a,i}$ is the breeding value presented in (1) (Mrode and Thompson, 2014). The dominance means that if a gene has a dominant allele, the phenotype is determined directly regardless of the other allele (Campbell et al., 2018). For example, a locus that determines the colour of a flower might have two alleles, let us say P and p. Locus (plural loci) is the position of a gene on a chromosome (Alberts et al., 2002). It is the "address" of the gene. Then we have four options for the locus: PP, Pp, pP, and pp. If P is dominant over p, the three first options will yield the same colour, while the pp will yield a different colour. In summary, epistasis means that two or more genes at different loci affect together to a phenotype (Campbell et al., 2018). In other words, the alleles of two genes interact, which will affect the phenotype.

The effects of dominance and epistasis are often ignored because they make the calculations harder, and usually, their effect on the phenotype is relatively small. Also, in this thesis, we assume that their effect is zero if it is not otherwise mentioned. Alternatively, it is possible to assume that the effect of dominance and epistasis are included in ϵ_{ij}. The connection between the breeding values and the genetic model (when the dominance and epistasis are ignored) is:

$$y_{ij} = \mu_i + g_{a,i} + \epsilon_{ij} = \mu_i + a_i + \epsilon_{ij},$$

where a_i is the breeding value of the individual. The above genetic model can be presented also in a matrix form:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{a} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y}, \boldsymbol{\mu}, \mathbf{a},$ and $\boldsymbol{\epsilon}$ are all n_{obs}-dimensional vectors. The number of observations is denoted as n_{obs}. In the context of animal and plant breeding, it is usually assumed that the random vectors $\mathbf{a}$ and $\boldsymbol{\epsilon}$ follow multivariate normal distributions (MVN): $\mathbf{a} \sim \text{MVN}(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \mathbf{R})$. The actual structure of a breeding model is discussed in Section 4.

The genetic model presented in (2) is a starting point for a linear mixed model typically used in breeding. We are especially interested to solve the predictions for $\mathbf{a}$ in (2) since $\mathbf{a}$ holds the breeding values. The linear mixed model and BLUP are discussed in more detail in Section 4.

In the framework of a genetic model, we assume additionally that all the variance components, $\text{Var}(\mathbf{y})$, $\text{Var}(\mathbf{a})$, and $\text{Var}(\boldsymbol{\epsilon})$, are known. It is usually assumed that there is no correlation between environmental and genetic effects, nor a correlation between the individuals. However, there are some exceptions when there is a correlation between genotype and environment. Falconer and Mackay (1996) give an example related to dairy cattle: if the cows are fed according to their yield, the cows with better milk yield get more food. Even though there are some situations with a correlation between genotype and environment, the magnitude of the covariance between them is almost always unknown.

Assuming no correlation between environment and genetic effects, we get that $\mathrm{Cov}(\mathbf{a}, \boldsymbol{\mu}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{a}, \boldsymbol{\epsilon}) = \mathbf{0}$. The phenotypic variance is then:

$$\begin{aligned}\mathrm{Var}(\mathbf{y}) &= \mathrm{Var}(\boldsymbol{\mu} + \mathbf{a} + \boldsymbol{\epsilon}) \\ &= \mathrm{Var}(\boldsymbol{\mu}) + \mathrm{Var}(\mathbf{a}) + \mathrm{Var}(\boldsymbol{\epsilon}) \\ &= \mathrm{Var}(\mathbf{a}) + \mathrm{Var}(\boldsymbol{\epsilon}).\end{aligned}$$

As we can see, we can partition the phenotypic variance into different sources of variation and estimate how much of the variation in a phenotype is due to genetics and how much it is due to random environmental factors. With these variance components, an estimate for heritability can be calculated, which will be discussed in the next section.

2.3 Heritability

Breeders are interested in how much the differences between individuals' phenotypic observations are affected by genetic differences between them. Heritability of a trait expresses the proportion of the phenotypic variation $\mathrm{Var}(\mathbf{y})$ explained by the genotype (Juga et al., 1999). It is defined as a ratio:

$$h^2 = \frac{\mathrm{Var}(\mathbf{a})}{\mathrm{Var}(\mathbf{y})},$$

where $\mathrm{Var}(\mathbf{a})$ is the additive genetic variance of the population and $\mathrm{Var}(\mathbf{y})$ is the phenotypic variance. It is (clearly) a value between 0 and 1. The higher the heritability is, the easier it is for the breeders to make the selection since the variation is mainly in genes and not in things that breeders could not maybe affect so easily. Breeders hope to see as high heritability values as possible. According to Juga et al. (1999), irrespective of animal species, traits related to fertility and vitality have low heritabilities $h^2 \approx 0.0 - 0.1$, most of the structural traits of animals have high heritabilities: h^2 over 0.40, and most of the production traits, such as growth rate and milk production, have $h^2 \approx 0.15 - 0.30$. The heritability of human height is about 0.8, which is high value (McEvoy and Visscher, 2009). This means that genes control 80% of the variation in height between humans.

As we can see from the definition of heritability, if the random environmental effects affect a lot to the phenotypic observation, meaning that $\mathrm{Var}(\boldsymbol{\epsilon})$ is relatively large, then the heritability is low. Correspondingly, if the effect of random environmental effect is low, meaning that $\mathrm{Var}(\boldsymbol{\epsilon})$ is low, then the heritability is high (Juga et al., 1999).

2.4 Reliability and accuracy of an estimated breeding value

The accuracy of an EBV is defined as the correlation between the true breeding value and the estimated one. Thus, it can be calculated as follows:

$$r = \mathrm{Corr}(\hat{a}, a) = \frac{\mathrm{Cov}(\hat{a}, a)}{\sigma_{\hat{a}}\sigma_a}, \tag{3}$$

where a is the true breeding value, and $\hat{a}$ is the EBV. In (3), the components $\mathrm{Cov}(\hat{a}, a)$ and $\sigma_{\hat{a}}$ are estimated with breeding models, such as BLUP, which is discussed in Section 4. Additive genetic variance $\sigma_a^2 = \mathrm{Var}(\mathbf{a})$ can be estimated different time as the other components. Reliability r^2 is the square of accuracy.

PEV measures the precision of an EBV:

$$\mathrm{PEV} = \mathrm{Var}(a - \hat{a}). \tag{4}$$

PEV can also be estimated with BLUP. The main goal in this thesis is how we can approximate PEV using MC sampling. We will follow up by calculating the reliabilities, and especially the PEVs in Section 4.3.

3 Relationship matrices

One of the most important parts of animal and plant breeding models is to have information about the relationships among individuals in the population. The relationship information can be presented in two alternative ways. One way is to construct a pedigree-based relationship matrix. It is constructed using information about the parents of each individual. When constructing a pedigree-based relationship matrix, we must have the population's pedigree available. The pedigree-based approach is extensively used in animal breeding, where the pedigree is usually available. Nowadays, it is more common to construct a relationship matrix using the genotype information; in this approach, the pedigree is not needed. The genomic relationship matrix requires the genotype information of all individuals to be collected. When genomic information is known only from some of the individuals, then a relationship matrix combining genomic and pedigree information can be constructed.

3.1 Pedigree-based relationship matrix

The pedigree-based relationship matrix, also known as a numerator relationship matrix, **A** is a square and symmetric $n \times n$ matrix, where n is the number of individuals, and which, in the simplest case, has ones on the diagonal (Mrode and Thompson, 2014). However, if the inbreeding is considered, the diagonal values are $1 + F_i$, where F_i is the inbreeding coefficient of the animal i. Inbreeding can (accidentally) occur in small populations, where, for example, an animal's father and grandfather could be the same animal. In other words, inbreeding means that the individuals are very closely related. According to Mrode and Thompson (2014), the exact interpretation of the diagonal value of **A** is "twice the probability that two gametes taken at random from animal i will carry identical alleles by descent." The off-diagonal values of **A** express the degree of the relatedness between two individuals. They are called the coefficients of relationship (Henderson, 1976). A good rule of thumb is that the off-diagonal values lie substantially in the range of 0 to 1. Values close to one indicate that the animals are more related, and values close to zero indicate that the animals are less related. Multiplying the matrix **A** with the additive genetic variance σ_a^2, we will get the variance-covariance matrix for the population. There are many ways of constructing the pedigree-based matrix **A** and directly constructing its inverse; the inverse of the relationship matrix is needed in most animal and plant breeding models.

Next, we will have a simple example of constructing **A** matrix using a so-called recursive method (Henderson, 1976). Let us have the following pedigree as presented in Table 1. When constructing the relationship matrix **A**, the parents must precede the descendants in the corresponding pedigree. Next, we will introduce the rules to form the elements of **A**. In this example, we have inbreeding in the population because the sire and grandsire of animal number six is the animal number one.

Table 1: An example of a pedigree for six animals.

Animal	Sire of animal	Dam of animal
1	Unknown	Unknown
2	Unknown	Unknown
3	1	2
4	1	2
5	Unknown	4
6	1	4

The rules for constructing **A** follow Henderson (1976) and Mrode and Thompson (2014, pp. 22-33). We will denote the elements of **A** as q_{ij}, where i is the ith row of the matrix, and j is the jth column. Here, the letter s refers to the sire of the animal i, and the letter d refers to the dam of the animal i.

i) If both parents of the animal i are known, then the elements are

$$\begin{aligned} q_{ji} &= q_{ij} = \frac{1}{2} \cdot (q_{js} + q_{jd}), \text{when } i > j \quad \text{or} \\ q_{ii} &= 1 + \frac{1}{2} \cdot q_{sd}. \end{aligned}$$

ii) If only one parent of the animal i is known (sire or dam), then the elements are

$$\begin{aligned} q_{ji} &= q_{ij} = \frac{1}{2} \cdot q_{j(\text{s or d})}, \text{when } i > j \quad \text{or} \\ q_{ii} &= 1. \end{aligned}$$

iii) If both parents of the animal i are unknown, then the elements are

$$\begin{aligned} q_{ji} &= q_{ij} = 0, \text{when } i > j \quad \text{or} \\ q_{ii} &= 1. \end{aligned}$$

Using these three rules, we can build up **A** for our example using the pedigree given in Table 1. Below are some calculations of how these elements are calculated in our example. The resulting relationship matrix **A** is given in Table 2.

$q_{11} = 1$ (both parents of animal 1 are unknown, so we use the rule iii))

$q_{12} = q_{21} = 0$ (both parents of animal 2 are unknown, so we use the rule iii))

$q_{22} = 1$

$q_{13} = q_{31} = \frac{1}{2} \cdot (q_{11} + q_{12}) = \frac{1}{2} \cdot 1 = \frac{1}{2}$ (both parents of animal 3 are known, so we use the rule i))

$q_{23} = q_{32} = \frac{1}{2} \cdot (q_{21} + q_{22}) = \frac{1}{2}$ (both parents of animal 3 are known, so we use the rule i))

$q_{33} = 1$

$q_{14} = q_{41} = \frac{1}{2} \cdot (q_{11} + q_{12}) = \frac{1}{2}$ (both parents of animal 4 are known, so we use the rule i))

$\vdots$

$q_{66} = 1 + \frac{1}{2} \cdot q_{14} = 1 + \frac{1}{2} \cdot \frac{1}{2} = 1.25$

Table 2: The numerator relationship matrix **A** for six animals based on the pedigree given in Table 1.

Animal	1	2	3	4	5	6
1	1.000	0.000	0.500	0.500	0.250	0.750
2	0.000	1.000	0.500	0.500	0.250	0.250
3	0.500	0.500	1.000	0.500	0.250	0.500
4	0.500	0.500	0.500	1.000	0.500	0.750
5	0.250	0.250	0.250	0.500	1.000	0.375
6	0.750	0.250	0.500	0.750	0.375	1.250

In our example, most of the diagonal values of **A** are 1s, which means that those animals are not inbred. However, the inbreeding coefficient of animal number six is 0.25, which means the animal is inbred. The inverse of **A** is needed in animal models, so the next step is to invert the matrix. Here, we demonstrated one method to construct the pedigree-based relationship matrix. There are faster methods of constructing **A** and directly constructing its inverse. But we do not go deeper into those methods in this thesis.

More information about these methods is found, for example, in Mrode and Thompson (2014, pp. 22-33).

A numerator relationship matrix provides information about the expected relationships between individuals. The values in the matrix **A** indicate the expected relationships between the individuals (Wang, Misztal, and Legarra, 2014). The expected relationships are based on expectations for actual identity at individual loci (Hill and Weir, 2011). For instance, the expected relationship value for full-sibs is 0.5, and for half-sibs, it is 0.25. The expected genetic relationship between the parent and the offspring is 0.5, as the offspring inherits half of its genes from one parent. If two individuals are not connected by pedigree and do not share common ancestors, the expected relationship value is 0. Usually, it is assumed that the founders of the pedigree are unrelated. However, these assumptions might not hold in nature because there is a limited number of genes, and there might be linkage disequilibrium between the alleles of loci. This means that there is an association of alleles at two or more different loci, which is not random (Slatkin, 2008).

Next, we will introduce the alternative relationship matrix, the genomic relationship matrix. But before that, we will shortly tell what SNPs are, as genomic relationship matrices are based on SNP information of individuals.

3.2 Single nucleotide polymorphism

SNPs are specific locations in the genome sequence where the nucleotide differs within a population. Some fraction of the population has one nucleotide, and some fraction has another nucleotide (Alberts et al., 2002). Most of the SNPs do not affect phenotype, but a subset is responsible for the differences in a phenotype. For example, the DNA sequences of a human and a chimpanzee are 99 % identical, while the DNA sequences of two different humans are over 99.9 % identical (Alberts et al., 2002). This shows that even slight discrepancies can be noticeable.

3.3 Genomic relationship matrix

Instead of having a pedigree showing the relatedness of the individuals, we might have genetic information from the individuals. Including genetic information in animal breeding models began in the early 21st century when

Meuwissen, Hayes, and Goddard (2001) presented a method for estimating breeding values using genotypic data. A couple of years later, Schaeffer (2006) said that genotyping animals for thousands of SNPs could reduce the cost of proving bulls by over 90 %. Genomic estimated breeding value (GEBV) is a breeding value estimated using SNP data. Generally speaking, we have a marker matrix containing SNP information for all the individuals. The individuals can be genotyped for over 50,000 SNPs. Nowadays, when it is actually possible to measure the genotype of an animal or plant—with DNA marker technology—genotyping the individuals is on the up and up. The genetic similarity estimates obtained using SNP information are more accurate than the pedigree-based relationship estimates (Isik, Holland, and Maltecca, 2017). In this thesis, the genomic relationship matrix is denoted as $\mathbf{\Omega}$. As the pedigree-based relationship matrix, the genomic relationship matrix is also $n \times n$ matrix, where n is the number of individuals.

We will call the matrix that contains the SNP information for every individual as $\mathbf{M}$. It is an $n \times m$ -matrix, where n is the number of individuals and m is the number of markers. In the process of constructing the genomic relationship matrix $\mathbf{\Omega}$, there must not be any missing data in the $\mathbf{M}$ matrix, so either the missing cells must be deleted, or the values can be imputated with some method. In this context deleting means either deleting an individual having a lot of missing SNP information or deleting a marker having a lot of missing data. The marker matrix should be coded as 0s, 1s and 2s, such that homozygotes (AA or BB) are coded as 0s or 2s and heterozygotes (AB or BA) are coded as 1s. Homozygote means that an individual has a pair of identical alleles for a gene, and heterozygote means that an individual has different alleles for a gene (Campbell et al., 2018). For example, in Table 3, we see that plant 1 is homozygous for SNP1 and SNP3 and heterozygous for SNP2, SNP4, and SNP5. In our example, we will have the following marker matrix with three plants and five SNPs:

Table 3: An example of a marker matrix $\mathbf{M}$ for three plants. Plants are genotyped for five markers.

Plant	SNP1	SNP2	SNP3	SNP4	SNP5
1	0	1	0	1	1
2	2	2	1	0	1
3	0	0	1	2	0

First, we need to create a $\mathbf{P}$ matrix that contains twice the minor allele frequencies (MAFs) for each SNP. This means that each column of the matrix will represent twice the MAF for the corresponding SNP. To calculate the allele frequencies for each SNP, we can choose either allele A or B and determine the proportion of that allele in each column. In this case, we will calculate the allele frequencies with respect to allele B. The allele frequency of the first column is $\frac{1}{3}$, because plant 1 has (AA), plant 2 has (BB), and plant 3 has (AA). The allele frequency of B in the first column is:

$$\frac{\text{count of B}}{\text{count of A and B}} = \frac{2}{6} = \frac{1}{3}.$$

With the same logic, we can count the allele frequency of the second column, which is $\frac{1}{2}$, because 1 means (AB or BA), 2 means (BB), and 0 means (AA), so if we count the frequency of the letter B, we get $\frac{3}{6} = \frac{1}{2}$. This way we get all allele frequencies for SNP1, SNP2, SNP3, SNP4, and SNP5, which are $\frac{1}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{2}$, and $\frac{1}{3}$, respectively.

Remark 3.1. The allele frequencies can also be counted by simply counting the mean of the column and dividing it by two. The allele frequency for the jth SNP is $p_j = \frac{1}{2n}\sum_{i=1}^{n} m_{ij}$, where m_{ij} is the ith row and jth column of the marker matrix $\mathbf{M}$.

After counting the allele frequencies, we can continue constructing the genomic relationship matrix $\mathbf{\Omega}$ using VanRaden's first method (VanRaden, 2008). The calculations are done in a similar way as Putz (2018). As said, $\mathbf{P}$ contains twice the allele frequency of each SNP. In this case, our $\mathbf{P}$ is

$$\mathbf{P} = \begin{bmatrix} \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} \end{bmatrix}.$$

The next step is to center the marker matrix $\mathbf{M}$; the centered $\mathbf{M}$ is called $\mathbf{C}$ matrix. Using our $\mathbf{P}$ matrix, we can obtain the $\mathbf{C}$ matrix:

$$\begin{aligned} \mathbf{C} = \mathbf{M} - \mathbf{P} &= \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 2 & 2 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 & 0 \end{bmatrix} - \begin{bmatrix} \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{2}{3} & 0 & -\frac{2}{3} & 0 & \frac{1}{3} \\ \frac{4}{3} & 1 & \frac{1}{3} & -1 & \frac{1}{3} \\ -\frac{2}{3} & -1 & \frac{1}{3} & 1 & -\frac{2}{3} \end{bmatrix}. \end{aligned}$$

As VanRaden (2008) presented, a genomic relationship matrix $\boldsymbol{\Omega}$ has the following form, where p_j is the allele frequency of the jth SNP:

$$\boldsymbol{\Omega} = \frac{\mathbf{CC}'}{2\sum_{j=1}^{m} p_j(1-p_j)}. \tag{5}$$

Thus, applying (5) to our example, we finally have the genomic relationship matrix $\boldsymbol{\Omega}$, presented in Table 4.

Table 4: The genomic relationship matrix $\boldsymbol{\Omega}$ for three plants using genotype information from Table 3.

Plant	1	2	3
1	0.429	-0.429	0.000
2	-0.429	1.714	-1.286
3	0.000	-1.286	1.286

In R, genomic relationship matrices can be constructed using the function *G.matrix* in the R package *ASRgenomics* (R Core Team, 2023; Gezan et al., 2022).

The genomic inbreeding coefficient for individual i is $\Omega_{ii} - 1$, and the genomic relationship coefficient between individuals i and j is $\frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}}$ (VanRaden, 2008). "The genomic relationship matrix is an estimator of the actual proportion of genome that is identical by descent across individuals" (Hill and Weir, 2011). In theory, both the pedigree-based $\mathbf{A}$ and the genomic relationship matrix $\boldsymbol{\Omega}$ should reflect the actual relationships in the population. Additionally, these matrices should be quite similar when created for the same population. Significant differences between them indicate errors in the pedigree or in the genotyping (Hill and Weir, 2011).

3.4 Blending genomic relationship matrix

A common problem with $\boldsymbol{\Omega}$ matrices is that they are usually not invertible. That is a problem because, in GBLUP, introduced in the next section, we need to have the inverse of $\boldsymbol{\Omega}$. Fortunately, there are methods to overcome this problem. One alternative is to blend $\boldsymbol{\Omega}$ with the corresponding pedigree-based $\mathbf{A}$ matrix. As pedigree is not always available, this approach is not always possible.

Instead of blending $\boldsymbol{\Omega}$ with $\mathbf{A}$, we can blend it with a proportion of an identity matrix $\mathbf{I}_n$, which is multiplied by a small value (Hollifield et al., 2022; Gezan et al., 2022). If the original genomic relationship is $\boldsymbol{\Omega}$, then the blended one is:

$$\boldsymbol{\Omega}_{\text{blended}} = (1 - b) \cdot \boldsymbol{\Omega} + b \cdot \mathbf{I}_n,$$

where b is the desired proportion of an identity matrix. The proportion of b used changes according to the situation, but small proportions are usually favoured. For example, Himmelbauer, Schwarzenbacher, and Fuerst (2021) presented that $b = 0.01$ caused some bias to the modelling, whereas using $b = 0.001$ caused hardly any bias.

As Hollifield et al. (2022) discussed, the differences between blending $\boldsymbol{\Omega}$ with the pedigree-based matrix or with a weighted identity matrix were insignificant in terms of reliability and inflation of breeding values. In our example in Section 3.3, our $\boldsymbol{\Omega}$ is not invertible, but after blending it with a proportion of 1% of an identity matrix, inverting it is possible.

4 Best linear unbiased prediction

BLUP is a prevalent method in animal and plant breeding, and it is also the foundation for more complex models, such as GBLUP, which is discussed later. In the context of animal and plant breeding, the meaning of the words in BLUP is presented greatly and concisely by Mrode and Thompson (2014). In this context, *best* means that it maximizes the correlation between the true breeding value and the EBV. *Linear* means that "predictors are linear functions of observations" (Mrode and Thompson, 2014). *Unbiased* means that the estimates for fixed effects and random effects are unbiased. *Prediction* stands simply for the fact that we are estimating the true values, such as animal breeding values. Next, we will define a linear mixed model and see how the estimates for the fixed and random effects can be estimated simultaneously using so-called Henderson's MMEs.

4.1 Linear mixed model

Usually, in animal or plant models, we have both fixed and random effects in our models. Thus, the starting point is a simple linear mixed model presented, for example, by Isik, Holland, and Maltecca (2017):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{6}$$

where $\mathbf{y}$ is an n_{obs}-dimensional response vector, and n_{obs} is the number of observations. The incidence matrix $\mathbf{X}$ is an $n_{\text{obs}} \times k$ matrix, where k is the sum of the number of levels of categorical variables and continuous variables used as a fixed effect. The incidence matrix relates the fixed effects to the corresponding observation. The vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)'$ contains the unknown fixed effects and its size is $k \times 1$. The incidence matrix $\mathbf{Z}$ is an $n_{\text{obs}} \times n$ matrix, where n is the number of individuals with at least one observation, and it relates the random effects to the corresponding observation. The random vector $\mathbf{u}$ is an n-dimensional vector of unknown random effects. The residual vector $\mathbf{e}$ is an $n_{\text{obs}} \times 1$ vector containing the values of the residuals. The EBVs are obtained as predictions of $\mathbf{u}$. Assumptions for the random effects $\mathbf{u}$ and $\mathbf{e}$ depend on the situation. However, we will assume that they both follow a MVN. Thus:

$$\begin{aligned} \mathbf{u} &\sim \text{MVN}(\mathbf{0}, \mathbf{G}) \\ \mathbf{e} &\sim \text{MVN}(\mathbf{0}, \mathbf{R}), \end{aligned}$$

where $\mathbf{G}$ and $\mathbf{R}$ are the corresponding variance-covariance matrices.

Remark 4.1. The linear mixed model presented in (6) is presented without the intercept. If the intercept is included in the model, the incidence matrix $\mathbf{X}$ would be $n_{\text{obs}} \times (k+1)$ matrix having a column of ones as a first column. In this case $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$.

As $\mathbf{u}$ and $\mathbf{e}$ follow MVNs, it follows that $\mathbf{y}$ also follow a MVN. Usually, it is assumed that random effects and residuals are uncorrelated, meaning $\text{Cov}(\mathbf{u}, \mathbf{e}) = \mathbf{0}$. With these assumptions, it holds that (Isik, Holland, and Maltecca, 2017, pp. 74-77):

$$\begin{aligned} \mathbb{E}(\mathbf{y}) &= \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}) = \mathbf{X}\boldsymbol{\beta} \text{ and} \\ \text{Var}(\mathbf{y}) &= \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}) \\ &= \text{Var}(\mathbf{Z}\mathbf{u} + \mathbf{e}) \\ &= \mathbf{Z}\text{Var}(\mathbf{u})\mathbf{Z}' + \mathbf{R} + \mathbf{0} \\ &= \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}. \end{aligned}$$

The variance structure of the random effects depends on the situation. For example, if we assume that the observations are uncorrelated, the structure of $\mathbf{R}$ could be $\sigma_e^2\mathbf{I}_{n_{\text{obs}}}$, where σ_e^2 is the residual variance. In the context of animal and plant breeding, the variance-covariance matrix $\mathbf{G}$ is, usually, either a weighted pedigree-based relationship matrix $\sigma_u^2\mathbf{A}$ or a weighted genomic relationship matrix $\sigma_u^2\boldsymbol{\Omega}$, where σ_u^2 is additive genetic variance. Relationship matrices were discussed in Sections 3.1 and 3.3.

Standard notation is to mark $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$. Therefore, the model can be presented as follows (Isik, Holland, and Maltecca, 2017, pp. 74-77):

$$\begin{aligned}\mathbf{y} &\sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})\\ \mathbf{u} &\sim \text{MVN}(\mathbf{0}, \mathbf{G})\\ \mathbf{e} &\sim \text{MVN}(\mathbf{0}, \mathbf{R}).\end{aligned}$$

The connection to the genetic model, presented in Section 2.2, is that $\mathbf{u}$ presents the breeding values of the individuals: genetic effects. The fixed effects part $\mathbf{X}\boldsymbol{\beta}$ presents the environmental effects, and the residual part $\mathbf{e}$ refers to the random environmental effects. The linear mixed model presented in (6) is one way of modelling genetic effects.

Since the response variable $\mathbf{y}$ follows a MVN, the log-likelihood function for the model is

$$\log L(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where the constant part of the likelihood function has been ignored (Agresti, 2015). And if the variance-covariance matrix $\mathbf{V}$ is known, then maximizing the log-likelihood function $\log L(\boldsymbol{\beta}, \mathbf{V})$ gives the maximum likelihood estimate for the fixed effects:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Using the estimate for the fixed effects $\hat{\boldsymbol{\beta}}$, it is then possible to solve the predictors for the random effects $\hat{\mathbf{u}}$ as Henderson (1963) proved. The solution is

$$\hat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

If the variance-covariance matrix $\mathbf{V}$ is unknown, it has to be estimated. Estimation of variance parameters can be done, for example, by using Gibbs sampling. This thesis does not handle variance component estimation, as the variance parameters are assumed to be known. More information can be found from Mrode and Thompson (2014, pp. 251-297).

4.2 Henderson's mixed model equations

To obtain BLUP estimates for a linear mixed model, Henderson (1975) developed a method for simultaneously solving estimates for the fixed effects $\hat{\boldsymbol{\beta}}$ and the random effects $\hat{\mathbf{u}}$. One of the main advantages of Henderson's method is that there is no need to construct or invert the variance-covariance matrix $\mathbf{V}$. Henderson's MMEs are presented as follows:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{7}$$

$$[\text{LHS}] \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = [\text{RHS}].$$

The solutions for parameters are obtained by inverting the left-hand side of the equation. It is also a common and convenient practice to call the parts of (7) the left-hand side of the MME (LHS) and the right-hand side of the MME (RHS). This practice is beneficial when expanding models to be more complex. The predictor of random effects $\hat{\mathbf{u}}$ is the BLUP of $\mathbf{u}$ (Agresti, 2015), and the estimate for the fixed effects $\hat{\boldsymbol{\beta}}$ yields the same results as the generalized least squares for the $\boldsymbol{\beta}$, which means that they are equivalent.

The difference between BLUP and GBLUP is that in a BLUP model $\mathbf{G} = \sigma_u^2\mathbf{A}$, whereas in GBLUP $\mathbf{G} = \sigma_u^2\boldsymbol{\Omega}$. Thus, depending on the relationship matrix used, the model is called differently. The formulas for accuracies and PEVs, discussed in the next subsection, hold true for both approaches.

4.3 Accuracy of BLUP

As discussed in Section 2.4, the accuracy r of a prediction is the correlation between the true value and the predicted value. Sometimes, researchers want to calculate the reliability r^2 of the prediction, and that is simply the square of the accuracy r. We assume that the values of the variances σ_e^2 and σ_u^2 are known or estimated beforehand. If we define the LHS from (7) as

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$$

and the inverse of it as

$$\begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix},$$

then the PEV vector can be calculated as follows (Mrode and Thompson, 2014, pp. 44-45):

$$\mathbf{PEV} = \text{Var}(\mathbf{u} - \hat{\mathbf{u}}) = \mathbf{C}^{22}\sigma_e^2,$$

where σ_e^2 is the residual variance. For an individual i, the PEV is

$$\text{PEV}_i = \text{Var}(u_i - \hat{u_i}) = \mathbf{C}_{ii}^{22}\sigma_e^2.$$

PEV measures the magnitude of the additive genetic variance that is not explained by the model.

There is a deterministic connection between PEV and reliability. The reliability can be calculated as follows:

$$r^2 = 1 - \frac{\mathbf{PEV}}{\sigma_u^2}, \tag{8}$$

where σ_u^2 is the additive genetic variance. The standard error of prediction is SEP $= \sqrt{\text{Var}(\mathbf{u} - \hat{\mathbf{u}})} = \sqrt{\mathbf{C}^{22}}$.

4.4 SNP-BLUP

A method equivalent to GBLUP is SNP-BLUP (Mrode and Thompson, 2014), where we estimate the effect of each single SNP on the phenotype. That is the main difference from GBLUP, where we get the EBVs for the individuals, not for the SNPs, even though we use SNPs when constructing the genomic relationship matrix. Nevertheless, since the models are equivalent, we can indirectly calculate the EBVs using SNP-BLUP model's estimates.

We will use otherwise similar notation as Mrode and Thompson (2014), but we will have our own notation for the variance matrix of the markers. We assume that the markers follow a common normal distribution with mean zero and common variance σ_g^2. The marker variance is something that we do not usually know, but we must instead estimate it with the additive genetic variance and with a number of markers or with the allele frequencies. The marker variance can be estimated either as $\sigma_g^2 = \frac{\sigma_u^2}{m}$, where m is the number of SNPs or $\sigma_g^2 = \frac{\sigma_u^2}{2\sum_{j=1}^m p_j(1-p_j)}$, where p_j is the mean of the SNP j and σ_u^2 is the additive genetic variance.

The linear mixed model for estimating the SNP effects can be presented as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{g} + \mathbf{e},$$

where $\mathbf{y}$ is an n_{obs}-dimensional response vector, and n_{obs} is the number of observations. The incidence matrix $\mathbf{X}$ is an $n_{\text{obs}} \times k$ matrix, where k is the sum of the number of levels of categorical variables and continuous variables used as a fixed effect. The incidence matrix relates the fixed effects to the corresponding observation. The vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)'$ contains the unknown fixed effects and its size is $k \times 1$. $\mathbf{W}$ is an $n_{\text{obs}} \times m$ centralised marker matrix as discussed in Section 3.3, so basically $\mathbf{W} = \mathbf{M} - \mathbf{P}$, where m is the number of SNPs. The random vector $\mathbf{g}$ is an m-dimensional vector of unknown marker effects. The residual vector $\mathbf{e}$ is an $n_{\text{obs}} \times 1$ vector containing the values of the residuals. The marker effects are obtained as predictions of $\mathbf{g}$. Assumptions for the random effects $\mathbf{g}$ and $\mathbf{e}$ depend on the situation. However, we will assume that they both follow a MVN. Thus:

$$\begin{aligned} \mathbf{g} &\sim \text{MVN}(\mathbf{0}, \sigma_g^2 \mathbf{I}_m) \\ \mathbf{e} &\sim \text{MVN}(\mathbf{0}, \mathbf{R}), \end{aligned}$$

where $\mathbf{R}$ is the variance-covariance matrix for the residuals, usually having a diagonal form: $\mathbf{R} = \sigma_e^2 \mathbf{I}_{n_{\text{obs}}}$ (Mrode and Thompson, 2014).

Thus, in a SNP-BLUP model, the effects of SNPs are treated as random, and SNP effects get direct predictions using a SNP-BLUP model. However, the breeding values can also be estimated from the SNP-BLUP model using the following formula: $\hat{\mathbf{u}} = \mathbf{W}\hat{\mathbf{g}}$. Those estimated breeding values $\hat{\mathbf{u}}$ should be the same as obtained with GBLUP, thus, the results should be the same when applying SNP-BLUP and GBLUP for the same data. However, since there are a couple of different methods for constructing $\mathbf{G}$, the estimates for the breeding values might not be identical. Also, the blending of a genomic relationship matrix affects the results slightly. Similarly, as in BLUP and GBLUP, SNP-BLUP model can also be presented conveniently as a MME, where the results for the fixed and random effects can be obtained simultaneously. The MME for the SNP-BLUP is (Mrode and Thompson, 2014):

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix},$$

where the value of α is defined as $\alpha = \frac{\sigma_e^2}{\sigma_g^2}$.

One of the main advantages of the SNP-BLUP occurs when the number of genotyped individuals in the data exceeds the number of SNPs because the number of SNPs limits the number of random effects in the SNP-BLUP.

5 Application of a GBLUP model

Next, we will fit a GBLUP model to a simulated data set. We will use a subset of data from Natural Resources Institute Finland. The reason for only using the subset is to ensure that the computer's RAM memory is sufficient for conducting the exact modelling. The specifications of the computer used in this thesis are found in Appendix F. From now on, the data means the subset which we conduct the analysis. The data mimics a typical plant breeding scheme, where we have two heterotic groups, one male group and one female group. A male from the male group is crossed with a female from the female group to produce several genetically identical offspring (clones) that are tested on different locations. In the following, we name the offspring of the same parents as a cross, which has repeated phenotypic observations from different locations. The response variable in this application is the grain yield, and the observations are made on crosses. In this application, the parents of the crosses have been genotyped, which means that we can construct genomic relationship matrices for the parents. The crosses have not been genotyped.

Our interest in this application is to fit a single-trait GBLUP model for the data. The model is a linear mixed model with one fixed effect and four random effects. We will estimate the fixed effect of the location, where the crosses have been planted, and random effects, which are the general combining ability (GCA) of the male plants, the GCA of the female plants, and the specific combining ability (SCA) of the crosses. In addition to that, we have the residual effect, which is treated as random. In this case, GCA measures a parent's impact on the cross's phenotypic observation, which is the grain yield. Generally, GCAs predict parental breeding values (Isik, Holland, and Maltecca, 2017). SCA measures the effect of each independent cross on the phenotypic observation.

We have SNP information for both male and female plants in the data set: 50,000 SNPs are available from all the individual male and female plants. In this application, we have separate marker matrices for the male heterotic group and the female heterotic group because the genotypic variance differs in those heterotic groups. Since the data are simulated data set, there are no missing values in the marker matrices, so we do not have to imputate any missing SNPs. As said before, we will use a single-trait GBLUP model for the data; in other words, we will use a hybrid model with two genomic relationship matrices.

There are 3,000 (n_{c}) unique crosses in the data set: each has eight observations, so we have 24,000 (n_{obs}) observations in total. Each of the eight observations has been planted in a different location, so we will see if the location has some effect. In the data, there are 1,492 (n_{m}) male plants and 1,510 (n_{f}) female plants. The main male plant has been crossed with all the other 1,491 female plants except the main female plant. Similarly, the main female plant was crossed with all the other 1,509 male plants except the main male plant. From each unique cross, eight clones were tested at eight locations out of 31 possible locations. The covariate used in this application, location, is a categorical variable. It has 31 levels, meaning there are 31 different locations where the crosses have been planted. The range of observations per location is 9–1,500, while the median is 755.

The linear mixed model, which we are going to use for this situation, is a single-trait GBLUP hybrid model:

$$\text{grain\_yield}_i = \text{location}_i + \text{GCA}_{\text{female},i} + \text{GCA}_{\text{male},i} + \text{SCA}_{\text{cross},i} + \epsilon_i. \tag{9}$$

The random effects are assumed to follow the following distributions:

$$\begin{aligned} \mathbf{GCA}_{\text{female}} &\sim \text{MVN}(\mathbf{0}, \sigma^2_{\text{female}}\boldsymbol{\Omega}_{\text{female}}) \\ \mathbf{GCA}_{\text{male}} &\sim \text{MVN}(\mathbf{0}, \sigma^2_{\text{male}}\boldsymbol{\Omega}_{\text{male}}) \\ \mathbf{SCA}_{\text{cross}} &\sim \text{MVN}(\mathbf{0}, \sigma^2_{\text{cross}}\mathbf{I}_{n_{\text{c}}}) \\ \boldsymbol{\epsilon} &\sim \text{MVN}(\mathbf{0}, \sigma^2_{\text{e}}\mathbf{I}_{n_{\text{obs}}}), \end{aligned} \tag{10}$$

where $\boldsymbol{\Omega}_{\text{female}}$ and $\boldsymbol{\Omega}_{\text{male}}$ are the genomic relationship matrices, with dimensions $n_{\text{f}} \times n_{\text{f}}$ and $n_{\text{m}} \times n_{\text{m}}$, constructed using VanRaden's method 1 introduced in Section 3.3 (VanRaden, 2008).

In plant breeding, it is usual to estimate the variance components at different times, as the EBVs. In this application, the variance values are $\sigma^2_{\text{female}} = 15$, $\sigma^2_{\text{male}} = 10$, $\sigma^2_{\text{cross}} = 6$, and $\sigma^2_{\text{e}} = 233$.

Let us formulate the model using matrix notations for linear mixed models as in Section 4.1. We have a similar situation as Luo et al. (2023), and we will use a similar matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{Z}_3\mathbf{u}_3 + \boldsymbol{\epsilon}, \tag{11}$$

where $\boldsymbol{\beta}$ is a 31-dimensional vector of fixed effects, and we denote $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ for the respective BLUPs of $\mathbf{GCA}_{\text{female}}$, $\mathbf{GCA}_{\text{male}}$, and $\mathbf{SCA}_{\text{cross}}$ effects, with dimensions of $n_{\text{f}} \times 1, n_{\text{m}} \times 1$, and $n_{\text{c}} \times 1$, respectively. These random vectors contain the breeding values. The incidence matrix for the fixed effect, $\mathbf{X}$, is an $n_{\text{obs}} \times 31$ matrix. The incidence matrices for the random effects $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ have dimensions of $n_{\text{obs}} \times n_{\text{f}}$, $n_{\text{obs}} \times n_{\text{m}}$, and $n_{\text{obs}} \times n_{\text{c}}$, respectively. The response vector $\mathbf{y}$ and the residual vector $\boldsymbol{\epsilon}$ are n_{obs}-dimensional vectors.

The variance-covariance matrix for the residuals $\mathbf{R} = \sigma^2_{\text{e}}\mathbf{I}_{n_{\text{obs}}}$. The variance-covariance matrices for the random effects $\mathbf{G}_1 = \sigma^2_{\text{female}}\boldsymbol{\Omega}_{\text{female}}$, $\mathbf{G}_2 = \sigma^2_{\text{male}}\boldsymbol{\Omega}_{\text{male}}$, and $\mathbf{G}_3 = \sigma^2_{\text{cross}}\mathbf{I}_{n_{\text{c}}}$.

As expressed in the previous section, we can construct the MME to solve the fixed and random effects simultaneously. In this case, the MME to be solved is the same as Luo et al. (2023) have:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{G}_1^{-1} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{G}_2^{-1} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{Z}_3 + \mathbf{G}_3^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \hat{\mathbf{u}}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}. \tag{12}$$

The estimates for the fixed and random effects are obtained by inverting the LHS. The MME is directly constructed using R (R Core Team, 2023). The genomic relationship matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ are blended with a function called *G.tuneup* from the R package *ASRgenomics.* The matrices were blended with one per cent of an identity matrix. After blending the matrices, they were invertible. Blending genomic relationship matrices were discussed in Section 3.4.

In this example, we do not do any marker selection because there is no missing marker information, and we do not specify any MAF threshold even though there are SNPs whose MAF is zero, which means that every plant has the same allele of that particular SNP.

5.1 Calculating the prediction error variances

As discussed in Section 4.3, PEVs of a model can be obtained using the diagonal values of the LHS. We denote the inverse of the left-hand side of

(12) as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{G}_1^{-1} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{G}_2^{-1} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_3 \\ \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{Z}_2 & \mathbf{Z}_3'\mathbf{R}^{-1}\mathbf{Z}_3 + \mathbf{G}_3^{-1} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} & \mathbf{C}^{13} & \mathbf{C}^{14} \\ \mathbf{C}^{21} & \mathbf{C}^{22} & \mathbf{C}^{23} & \mathbf{C}^{24} \\ \mathbf{C}^{31} & \mathbf{C}^{32} & \mathbf{C}^{33} & \mathbf{C}^{34} \\ \mathbf{C}^{41} & \mathbf{C}^{42} & \mathbf{C}^{43} & \mathbf{C}^{44} \end{bmatrix}.$$

The PEV vector of this model is:

$$\mathbf{PEV} = \operatorname{diag}\left(\begin{bmatrix} \mathbf{C}^{22} & \mathbf{C}^{23} & \mathbf{C}^{24} \\ \mathbf{C}^{32} & \mathbf{C}^{33} & \mathbf{C}^{34} \\ \mathbf{C}^{42} & \mathbf{C}^{43} & \mathbf{C}^{44} \end{bmatrix}\right). \tag{13}$$

PEV contains PEVs for all the predicted random effects: $\hat{\mathbf{u}}_1$, $\hat{\mathbf{u}}_2$, and $\hat{\mathbf{u}}_3$, so its dimension is $6,002 \times 1$. In this thesis, the components of **PEV** are called the exact PEV values, and they are used later when comparing the sampled PEV values to the exact PEV values.

The reliabilities from a GBLUP model can be obtained as follows. The reliability for an individual i can be calculated (Ben Zaabza, Van Tassell, et al., 2023):

$$r_i^2 = 1 - \frac{\mathbf{PEV}_i}{\mathbf{G}_{ii}}, \tag{14}$$

where $\mathbf{PEV}_i$ is the ith element of the **PEV** vector. In our case $\mathbf{G}_{ii}$ is the corresponding diagonal value of $\mathbf{G}_1$, $\mathbf{G}_2$, or $\mathbf{G}_3$, depending on the genetic group to which the individual i belongs. For example, if we are calculating the reliability for a female plant, we would use $\mathbf{G}_1$ since that is the variance-covariance matrix for the female plants.

If we would like to calculate the reliability of the third female plant in our data, $\mathbf{PEV}_3$ would be the third element from **PEV**. In the divisor, we would have the third diagonal element from the matrix $\mathbf{G}_1$. So, the reliability for the third female plant can now be calculated:

$$r_3^2 = 1 - \frac{\mathbf{PEV}_3}{\operatorname{diag}(\mathbf{G}_1)_3} = 1 - \frac{9.897}{30.682} \approx 0.677,$$

where $\operatorname{diag}(\mathbf{G}_1)_3$ is the third diagonal element from the matrix $\mathbf{G}_1$.

6 Approximating the prediction error variances using Monte Carlo sampling

Data sets involved in plant and animal breeding are usually very large. They can contain observations from millions of individuals. In such a scenario, exact modelling is not computationally possible. Because of this, we need to use approximate methods. According to Ben Zaabza, Mäntysaari, and Strandén (2020) and Ben Zaabza, Van Tassell, et al. (2023), inverting the genomic relationship matrix with one million genotyped individuals would take 70 days with 10 CPU threads. One important part of breeding is to determine how accurate EBVs are. As discussed before, we can measure that with reliabilities (r^2). Since there is a deterministic connection between PEVs and reliabilities, we will see how methods approximating PEVs work with a hybrid model.

PEVs obtained using (13) are called the exact PEV values. The sampled PEVs are values estimating the exact PEVs using MC samples. The sample size (n_{MC}) refers to the number of MC samples used to calculate the sampled PEV.

This thesis will compare different formulations to obtain the PEVs using MC sampling. We will use a similar perspective as Hickey et al. (2009). They compared ten different formulations to estimate the exact PEVs using MC sampling. In this thesis, we will compare four different formulations, of which three are presented by García-Cortés et al. (1995). We will see how they behave when the sample size increases and how they behave between the three different genetic groups: male plants, female plants, and crosses. In addition to comparing these three methods, we will study the behaviour of a new formulation, which was presented by Hickey et al. (2009). The methods from García-Cortés et al. (1995) were chosen because they are well-known, widely used, and easy to implement. The so-called New Formulation 2 (NF2) from Hickey et al. (2009) was chosen, as it worked well.

We will inspect the correlations between the exact PEVs and the sampled PEVs, calculate the root mean square errors (RMSEs), and fit linear regressions, where we explain the exact PEVs with the sampled PEVs; with linear regression, we can measure how biased the sampled PEV estimators are. We will also observe the maximum absolute differences (MADs) between the exact PEVs and the sampled PEVs. There is also a visualisation of how those indicators behave as the sample size increases.

Next, we will introduce how to use MC simulation with the hybrid model, and in Section 6.3, we will present the formulations for approximating PEVs.

6.1 Notations for Monte Carlo sampling

As MC simulation or sampling can mean many different things, we will state shortly what we mean by MC in this context. As presented in Section 5, we have assumed the distributions for the random variables as expressed in (10). We also assume that the variance components are known. The genomic relationship matrices $\boldsymbol{\Omega}_{\text{female}}$ and $\boldsymbol{\Omega}_{\text{male}}$ are also known since we have the marker matrices and can construct them with VanRaden's method. Thus, the variance values and the genomic relationship matrices do not change between simulation rounds. However, what does change are the random effects $\mathbf{GCA}_{\text{female}}$, $\mathbf{GCA}_{\text{male}}$, and $\mathbf{SCA}_{\text{cross}}$, which are denoted as $\mathbf{u}_1, \mathbf{u}_2$, and $\mathbf{u}_3$, respectively. Using these simulated random effects, we will generate new $\mathbf{y}$ using the equation (9). After that, we will solve the MME (12) and get estimates for the fixed effects and the random effects. We will call it MC sampling when repeating this process multiple times.

Before introducing the formulations, we must present some notations. The number of MC samples is denoted as n_{MC}. We will simulate n_{MC} independent MC samples for the random effects: $\mathbf{u}_{\text{female}}$, $\mathbf{u}_{\text{male}}$ and $\mathbf{u}_{\text{cross}}$, assuming their distributions, shown in (10). For clarity, we will denote $\mathbf{u}_{\text{female}} = \mathbf{u}_1$, $\mathbf{u}_{\text{male}} = \mathbf{u}_2$, and $\mathbf{u}_{\text{cross}} = \mathbf{u}_3$, introduced in (11). After the simulation, we will have n_{MC} samples for each individual and each cross. We will denote the simulated values as $\widetilde{\mathbf{u}}_{\text{female}}$, $\widetilde{\mathbf{u}}_{\text{male}}$ and $\widetilde{\mathbf{u}}_{\text{cross}}$, whose dimensions are $n_{\text{female}} \times n_{\text{MC}}$, $n_{\text{male}} \times n_{\text{MC}}$, and $n_{\text{cross}} \times n_{\text{MC}}$, respectively. We combine all of those simulated random effects into one matrix

$$\widetilde{\mathbf{u}} = \begin{bmatrix} \widetilde{\mathbf{u}}_{\text{female}} \\ \widetilde{\mathbf{u}}_{\text{male}} \\ \widetilde{\mathbf{u}}_{\text{cross}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{u}}^{[1]}_{\text{female}} & \widetilde{\mathbf{u}}^{[2]}_{\text{female}} & \cdots & \widetilde{\mathbf{u}}^{[n_{\text{MC}}]}_{\text{female}} \\ \widetilde{\mathbf{u}}^{[1]}_{\text{male}} & \widetilde{\mathbf{u}}^{[2]}_{\text{male}} & \cdots & \widetilde{\mathbf{u}}^{[n_{\text{MC}}]}_{\text{male}} \\ \widetilde{\mathbf{u}}^{[1]}_{\text{cross}} & \widetilde{\mathbf{u}}^{[2]}_{\text{cross}} & \cdots & \widetilde{\mathbf{u}}^{[n_{\text{MC}}]}_{\text{cross}} \end{bmatrix},$$

with dimension $n \times n_{\text{MC}}$. The number of female plants is denoted as n_{female}, n_{male} is the number of male plants, n_{cross} is the number of crosses, and $n = n_{\text{female}} + n_{\text{male}} + n_{\text{cross}}$. For clarity, we will denote the column j of $\widetilde{\mathbf{u}}$, where $j = 1, ..., n_{\text{MC}}$, as $\widetilde{\mathbf{u}}^{[j]}$.

Next, we will demonstrate how to simulate data from a breeding model.

6.2 Simulating data from a breeding model

Multivariate normal samples can be simulated using the following procedure (Rasmussen and Williams, 2006, pp. 200-201). To generate samples from $\mathbf{y} \sim \text{MVN}(\mathbf{m}, \mathbf{K})$, where $\mathbf{m}$ is a d-dimensional mean vector and $\mathbf{K}$ is a positive definite symmetric covariance matrix that can be presented as $\mathbf{K} = \mathbf{L}\mathbf{L}'$. The lower triangular matrix $\mathbf{L}$ is called the Cholesky decomposition of $\mathbf{K}$. Generating an auxiliary variable $\mathbf{x} \sim \text{MVN}(\mathbf{0}, \mathbf{I}_d)$. Then $\mathbf{y} = \mathbf{m} + \mathbf{L}\mathbf{x}$ has the desired distribution $\text{MVN}(\mathbf{m}, \mathbf{K})$.

In our case, we will simulate the random effects $\tilde{\mathbf{u}}_{\text{female}}$ and $\tilde{\mathbf{u}}_{\text{male}}$ with the procedure mentioned above:

$$\begin{aligned}
\mathbf{x}^{[j]}_{\text{female}} &\sim \text{MVN}(\mathbf{0}, \mathbf{I}_{n_{\text{female}}}) \\
\mathbf{x}^{[j]}_{\text{male}} &\sim \text{MVN}(\mathbf{0}, \mathbf{I}_{n_{\text{male}}}) \\
\tilde{\mathbf{u}}^{[j]}_{\text{female}} &= \mathbf{L}_{\text{female}}\mathbf{x}^{[j]}_{\text{female}} \\
\tilde{\mathbf{u}}^{[j]}_{\text{male}} &= \mathbf{L}_{\text{male}}\mathbf{x}^{[j]}_{\text{male}},
\end{aligned}$$

where $j = 1, \ldots, n_{\text{MC}}$ and $\mathbf{L}_{\text{female}}$ and $\mathbf{L}_{\text{male}}$ are the corresponding lower-triangular matrices. They can be obtained by solving the Cholesky decompositions of the blended genomic relationship matrices $\boldsymbol{\Omega}_{\text{female}}$ and $\boldsymbol{\Omega}_{\text{male}}$.

In practice, the auxiliary variables can be simulated from a univariate normal distribution such that $x_{\text{female}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{female}})$ and $x_{\text{male}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{male}})$. We will take n_{female}-sized and n_{male}-sized samples from those distributions and denote them as $\mathbf{x}^{[j]}_{\text{female}}$ and $\mathbf{x}^{[j]}_{\text{male}}$.

To simulate the random effects of the crosses $\tilde{\mathbf{u}}^{[j]}_{\text{cross}}$, we do not have to use Cholesky decomposition because the covariance structure is uncomplicated. They can be simulated such that $\tilde{\mathbf{u}}^{[j]}_{\text{cross}} \sim \text{MVN}(\mathbf{0}, \sigma^2_{\text{cross}}\mathbf{I}_{n_{\text{cross}}})$. And, in practice, we can draw n_{cross} samples from the corresponding distribution like this: $\tilde{\mathbf{u}}^{[j]}_{\text{cross}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{\text{cross}})$.

After we have simulated the random effects, we can generate the response variable $\mathbf{y}$, which we need to obtain the solutions of the MME (12). Since the relation between $\mathbf{y}$ and the random effects, as expressed in equations (9) and (11), is $y_i = \text{location}_i + \mathbf{u}_{\text{female},i} + \mathbf{u}_{\text{male},i} + \mathbf{u}_{\text{cross},i} + \epsilon_i$, we can construct the simulated observations such that $\tilde{\mathbf{y}}^{[j]} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\tilde{\mathbf{u}}^{[j]}_{\text{female}} + \mathbf{Z}_2\tilde{\mathbf{u}}^{[j]}_{\text{male}} + \mathbf{Z}_3\tilde{\mathbf{u}}^{[j]}_{\text{cross}} + \boldsymbol{\epsilon}$, where $\mathbf{Z}_1, \mathbf{Z}_2$, and $\mathbf{Z}_3$ are the same incidence matrices as presented in (11). However, in practice, we do not have to simulate any values for the fixed effects $\boldsymbol{\beta}$, because the expectation of $\mathbf{X}\boldsymbol{\beta}$ does not affect the distribution of random variables (García-Cortés et al., 1995). In other words, we can assume that $\mathbb{E}(\mathbf{X}\boldsymbol{\beta}) = 0$, because we are interested only in PEVs of the random effects. The generation of the simulated observations is made simply such that

$$\tilde{\mathbf{y}}^{[j]} = \mathbf{Z}_1\tilde{\mathbf{u}}^{[j]}_{\text{female}} + \mathbf{Z}_2\tilde{\mathbf{u}}^{[j]}_{\text{male}} + \mathbf{Z}_3\tilde{\mathbf{u}}^{[j]}_{\text{cross}} + \boldsymbol{\epsilon},$$

which yields

$$\tilde{\mathbf{y}} = \begin{bmatrix} \tilde{\mathbf{y}}^{[1]} & \tilde{\mathbf{y}}^{[2]} & \cdots & \tilde{\mathbf{y}}^{[n_{\text{MC}}]} \end{bmatrix}.$$

We will use a similar notation to the simulated observations $\tilde{\mathbf{y}}$, such that $\tilde{\mathbf{y}}^{[j]}$ means the jth column of $\tilde{\mathbf{y}}$. Using these simulated observations, we will solve the MME (12) n_{MC} times to obtain estimates with different values of random effects. We will denote the estimates that we get using the simulated observations $\tilde{\mathbf{y}}^{[j]}$ as $\hat{\mathbf{u}}^{[j]}_{\text{female}}, \hat{\mathbf{u}}^{[j]}_{\text{male}}$, and $\hat{\mathbf{u}}^{[j]}_{\text{cross}}$, combining them will result in

$$\hat{\mathbf{u}} = \begin{bmatrix} \hat{\mathbf{u}}_{\text{female}} \\ \hat{\mathbf{u}}_{\text{male}} \\ \hat{\mathbf{u}}_{\text{cross}} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{u}}^{[1]}_{\text{female}} & \hat{\mathbf{u}}^{[2]}_{\text{female}} & \cdots & \hat{\mathbf{u}}^{[n_{\text{MC}}]}_{\text{female}} \\ \hat{\mathbf{u}}^{[1]}_{\text{male}} & \hat{\mathbf{u}}^{[2]}_{\text{male}} & \cdots & \hat{\mathbf{u}}^{[n_{\text{MC}}]}_{\text{male}} \\ \hat{\mathbf{u}}^{[1]}_{\text{cross}} & \hat{\mathbf{u}}^{[2]}_{\text{cross}} & \cdots & \hat{\mathbf{u}}^{[n_{\text{MC}}]}_{\text{cross}} \end{bmatrix},$$

whose dimension is $n \times n_{\text{MC}}$. We will denote the column j of $\hat{\mathbf{u}}$, where $j = 1, ..., n_{\text{MC}}$, as $\hat{\mathbf{u}}^{[j]}$.

The formulations for approximating the exact PEV introduced in the next section rely on measuring how much the simulated random values $\tilde{\mathbf{u}}$ differ from the estimates $\hat{\mathbf{u}}$. The formulations also depend on the number of MC samples (n_{MC}) used. Comparing the results from different rounds of MC iterations will tell us how much variation there is in the random effects estimates between the iterations. Our primary interest is determining how the different methods of approximating the PEV behave as the sample size increases. We will also compare the accuracy and differences between these methods.

6.3 Methods for estimating prediction error variances

This section will present the four selected methods and their formulations to obtain the sampled PEV. They were selected as they are easy to implement and because they are well-known and widely used. The methods have different assumptions. The first three are presented by García-Cortés et al. (1995). The first three methods assume that $\text{Var}(\mathbf{u}) = \boldsymbol{\Omega}\sigma^2_u$. Here, $\boldsymbol{\Omega}$ is the genomic relationship matrix, and σ_u^2 is the corresponding additive genetic variance.

The first method from García-Cortés assumes additionally that $\text{Cov}(\mathbf{u}, \hat{\mathbf{u}}) = \text{Var}(\hat{\mathbf{u}})$. The formula for method PEV1 is:

$$\text{PEV1}(\hat{u}_i) = \text{Var}(u_i) - \text{Var}_{\text{MC}}(\hat{u}_i) = \boldsymbol{\Omega}_{ii}\sigma_u^2 - \frac{\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i}{n_{\text{MC}}}, \tag{15}$$

where $\hat{\mathbf{u}}_i$ is the ith row of $\hat{\mathbf{u}}$ and σ_u^2 is the corresponding additive genetic variance.

The second method from García-Cortés does not make any assumption of $\text{Cov}(\mathbf{u}, \hat{\mathbf{u}})$, so it handles the situations where $\text{Cov}(\mathbf{u}, \hat{\mathbf{u}}) \neq \text{Var}(\hat{\mathbf{u}})$. The formulation for the method PEV2 is:

$$\text{PEV2}(\hat{u}_i) = \text{Var}_{\text{MC}}(u_i - \hat{u}_i) = \frac{(\tilde{\mathbf{u}}_i - \hat{\mathbf{u}}_i)'(\tilde{\mathbf{u}}_i - \hat{\mathbf{u}}_i)}{n_{\text{MC}}}, \tag{16}$$

where $\tilde{\mathbf{u}}_i$ is the ith row of the simulated random effects.

The third method from García-Cortés et al. (1995) combines the PEV estimates from methods PEV1 and PEV2, and it uses information from both $\text{Var}_{\text{MC}}(\hat{u}_i)$ and $\text{Var}_{\text{MC}}(u_i - \hat{u}_i)$. As García-Cortés et al. (1995) said, we need to weight the pooled estimate with a covariance matrix when combining estimates of the same parameter. For the method PEV3 we assume that $\text{Cov}(\mathbf{u} - \hat{\mathbf{u}}, \hat{\mathbf{u}}) = \mathbf{0}$. We also need to know the asymptotic sampling variances of the methods PEV1 and PEV2.

$$\text{PEV3}(\hat{u}_i) = \frac{w_{1i}\text{PEV1}(\hat{u}_i) + w_{2i}\text{PEV2}(\hat{u}_i)}{w_{1i} + w_{2i}}, \tag{17}$$

$$\text{where } w_{1i} = \frac{1}{\text{Var}(\text{PEV1}(\hat{u}_i))} \text{ and } w_{2i} = \frac{1}{\text{Var}(\text{PEV2}(\hat{u}_i))}.$$

The asymptotic sampling variances can be approximated as follows:

$$\begin{aligned} \text{Var}(\text{PEV1}(\hat{u}_i)) &\approx \text{Var}_{\text{MC}}(\hat{\mathbf{u}}_i^2) \\ \text{Var}(\text{PEV2}(\hat{u}_i)) &\approx \text{Var}_{\text{MC}}((\tilde{\mathbf{u}}_i - \hat{\mathbf{u}}_i)^2). \end{aligned}$$

The fourth method we are using in the thesis is introduced by Hickey et al. (2009), and it is called NF2. It only assumes that $\text{Cov}(\mathbf{u} - \hat{\mathbf{u}}, \hat{\mathbf{u}}) = \mathbf{0}$, but does not make any assumption of $\text{Var}(\mathbf{u})$. It uses the PEV estimate obtained using method PEV2.

$$\text{PEV}_{\text{NF2}}(\hat{u}_i) = \frac{\text{PEV2}(\hat{u}_i)}{\text{PEV2}(\hat{u}_i) + \text{Var}_{\text{MC}}(\hat{\mathbf{u}}_i)}\boldsymbol{\Omega}_{ii}\sigma_u^2. \tag{18}$$

6.4 Algorithm for estimating prediction error variances of a hybrid model

This section will provide an algorithm to simulate random effects and solve the MME using the simulated values. The algorithm uses Cholesky decomposition to sample from a MVN. Algorithm 1 provides instructions on how a hybrid model can be simulated and how to produce MC samples from a hybrid model. It combines the information discussed in Section 6.

7 Comparing the methods approximating prediction error variances

In this section, we will implement the MC sampling to generate replicates of the data and solve the MME multiple times. We will compare the goodness of the four different methods approximating the exact PEV values, and we will also check how close those approximations are to the exact PEV values obtained via the direct method, presented in Section 5. We will also compare the accuracies between the three genetic groups and check for differences among them. The methods and principles used in the comparison follow the same perspective presented by Hickey et al. (2009). Next, we will present the tools we will use to determine the goodness of the methods.

7.1 Techniques for comparing formulations

First, to check for the bias of the approximated PEV, we will fit a basic linear regression, where the response variable is the exact PEV, and the explanatory variable is the sampled PEV:

$$\mathrm{PEV}_{\mathrm{exact},i} = \beta_0 + \beta_1 \mathrm{PEV}_{\mathrm{sampled},i} + \epsilon_i, \tag{19}$$

where

$$\epsilon_i \sim N(0, \sigma^2).$$

Estimates for the slope and intercept tell us how unbiased our approximated PEVs are. The closer the slope is to one, and the closer the intercept is to zero, the better our sampled PEV values are. We assume and hope that the approximated PEV values are as identical as possible compared to the

Algorithm 1 An algorithm for approximating PEVs of a GBLUP hybrid model.

Data: A data set with male and female plants and their crosses. Phenotypic observations are made on crosses. Male and female plants have been genotyped.

We will denote the three genetic groups as 1, 2, and 3, such that females are 1, males are 2, and crosses are 3.

Assumptions: Values for variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$, and σ_4^2 are assumed to be known. Incidence matrices $\mathbf{Z}_1, \mathbf{Z}_2$, and $\mathbf{Z}_3$ for random effects are assumed to be known. Genomic relationship matrices $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ are assumed to be known.

Set n_{MC} then do:

```
Solve Cholesky decompositions of the genomic relationship matrices $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$. Denote them as $\mathbf{L}_1$ and $\mathbf{L}_2$.
for $j = 1$ to $n_{\text{MC}}$ do
    Simulate $\mathbf{x}^{[i]}_{\text{female}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_1^2)$, $i = 1, ..., n_{\text{female}}$.
    Calculate $\tilde{\mathbf{u}}^{[j]}_{\text{female}} = \mathbf{L}_1 \mathbf{x}^{[i]}_{\text{female}}$.
    Simulate $\mathbf{x}^{[i]}_{\text{male}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_2^2)$, $i = 1, ..., n_{\text{male}}$.
    Calculate $\tilde{\mathbf{u}}^{[j]}_{\text{male}} = \mathbf{L}_{\text{male}} \mathbf{x}^{[i]}_{\text{male}}$.
    Simulate $\mathbf{x}^{[i]}_{\text{cross}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_3^2)$, $i = 1, ..., n_{\text{cross}}$.
    Set $\tilde{\mathbf{u}}^{[j]}_{\text{cross}} = \mathbf{x}^{[i]}_{\text{cross}}$.
    Generate simulated data $\tilde{\mathbf{y}}$. $\tilde{\mathbf{y}}^{[j]} = \mathbf{Z}_1 \tilde{\mathbf{u}}^{[j]}_{\text{female}} + \mathbf{Z}_2 \tilde{\mathbf{u}}^{[j]}_{\text{male}} + \mathbf{Z}_3 \tilde{\mathbf{u}}^{[j]}_{\text{cross}} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_4^2)$ and $i = 1, ..., n$.
end for
Construct the MME (12).
Solve the MME and calculate the exact PEVs for the random effects $\mathbf{u}_{\text{female}}, \mathbf{u}_{\text{male}}$ and $\mathbf{u}_{\text{cross}}$ using (13).
for $j = 1$ to $n_{\text{MC}}$ do
    Solve the MME such that $\mathbf{y}$ is replaced with $\tilde{\mathbf{y}}^{[j]}$. Save the solutions of the random effects $\hat{\mathbf{u}}^{[j]}_{\text{female}}$, $\hat{\mathbf{u}}^{[j]}_{\text{male}}$ and $\hat{\mathbf{u}}^{[j]}_{\text{cross}}$.
end for
Calculate the sampled PEVs using formulas (15), (16), (17) and (18).
Compare the exact PEVs and the sampled PEVs.
```

exact PEVs, meaning that each unit of increase in the exact PEV would also increase the sampled PEV value by one. It is also meaningful to assume that if the exact PEV is zero, the sampled one is also zero.

When fitting a linear regression, we can also calculate the coefficient of determination, usually R-squared or R^2. It is a value between 0 and 1, and it is the proportion of the variability in the response variable that can be explained using the explanatory variable(s) (Gareth et al., 2021). If R^2 is close to 1, the linear regression explains almost all the variability in the response variable. Correspondingly, a value close to 0 indicates that the linear regression does not explain the variability in the response variable.

We will also calculate the RMSE, which tells us how much the approximated PEVs differ from the exact PEVs.

$$\mathrm{RMSE}_{\mathrm{PEV}} = \left(\frac{1}{n} \sum_{i=1}^{n} (\mathrm{PEV}_{\mathrm{exact},i} - \mathrm{PEV}_{\mathrm{sampled},i})^2 \right)^{\frac{1}{2}}.$$

We can also calculate the correlation coefficient q between the exact PEV and the sampled PEV. In simple linear regression, where there is one response variable and one explanatory variable, there is a connection between q and R^2 because, in this simple case, it holds that $R^2 = q^2$ (Gareth et al., 2021). In a more general setting, the correlation coefficient can be calculated as

$$\begin{aligned}&\mathrm{Cor}(\mathrm{PEV}_{\mathrm{exact}}, \mathrm{PEV}_{\mathrm{sampled}}) = \\ &\frac{\sum_{i=1}^{n} (\mathrm{PEV}_{\mathrm{exact},i} - \overline{\mathrm{PEV}}_{\mathrm{exact}})(\mathrm{PEV}_{\mathrm{sampled},i} - \overline{\mathrm{PEV}}_{\mathrm{sampled}})}{\left(\sum_{i=1}^{n} (\mathrm{PEV}_{\mathrm{exact},i} - \overline{\mathrm{PEV}}_{\mathrm{exact}})^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} (\mathrm{PEV}_{\mathrm{sampled},i} - \overline{\mathrm{PEV}}_{\mathrm{sampled}})^2\right)^{\frac{1}{2}}},\end{aligned}$$

where $\overline{\mathrm{PEV}}$ is the mean of PEV values.

In addition to the above-mentioned techniques, we will also calculate the MAD between the exact PEV and the sampled PEV. MAD is simply the largest deviation between those values.

Even though there are many possible statistics to calculate, we will also draw scatter plots to show how the sampled PEVs and the exact PEVs differ. We will also see plots showing how the correlation behaves as the sample size increases. All the plots are drawn using the R-package *ggplot2*.

7.2 Main results

In the previous subsection, we discussed various statistics for evaluating four different formulas that approximate the exact PEVs. One approach is to compare the absolute difference between the sampled PEV and the exact PEV. Another method is to calculate the RMSE. Our preferred method is to calculate the correlation between the sampled PEV and the exact PEV. However, it is worth noting that the correlation coefficient may not be as reliable as the RMSE (García-Cortés et al., 1995). We can also measure the bias of the estimation by fitting a basic linear model, where the response variable is the exact PEV and the explanatory variable is the sampled PEV. Ideally, we want to see the intercept close to zero and the regression coefficient close to one. Appendix A demonstrates how these measurements improve as the sample size increases. These statistics are calculated using four sample sizes (n_{MC}): 100, 1,000, 10,000, and 200,000. The sample size of 200,000 mainly checks that everything works and that the calculations are implemented correctly. The statistics are found in Appendix A, and all the plots related to the analysis of this original design are found in Appendix D.

Using the direct method introduced in Section 5, we can obtain the exact PEVs, which we can use to compare with the sampled PEVs. The distribution of the exact PEVs for females was in the range of 3.89–12.09, with a mean of 7.75 and a standard deviation of 0.96. For males, the same statistics were 4.90–8.73 and (6.44, 0.48). For crosses, they were 5.09–5.33 and (5.18, 0.03). As we can see, the range of cross' PEVs is very narrow compared to the male and female plants.

It was surprising, when comparing to Hickey et al., 2009, how slowly the correlation between the sampled PEV values and the exact PEV values converged to 1. However, this is happening almost surely, as we see from the correlation plots in Figures 1, 2, and 3. Maximum number of MC samples used in this thesis was 200,000. With this number of samples, most of the correlations were over 0.99. In general, there were differences between the methods and between the genetic groups. PEV1 converges most slowly with the males and females, as seen in Figures 1 and 2, as the blue line yields regularly lower correlations. The method NF2 yields slightly higher correlations than the methods PEV2 and PEV3. However, what is highly satisfying is that all the methods seem to work: some converge slower and some faster.

The convergence is faster for the female plants than for the male plants. Nevertheless, there is a major difference when comparing the convergence of the cross group to the male and female groups: the convergence is much slower with all the methods, as seen in Figure 3. In addition, PEV2 is working very poorly with the crosses. The correlation after 200,000 MC samples was still less than 0.9 when using the method PEV2. Overall, the same story is true for the cross group; the method NF2 works well. With crosses, the method PEV3 works even better than the method NF2.

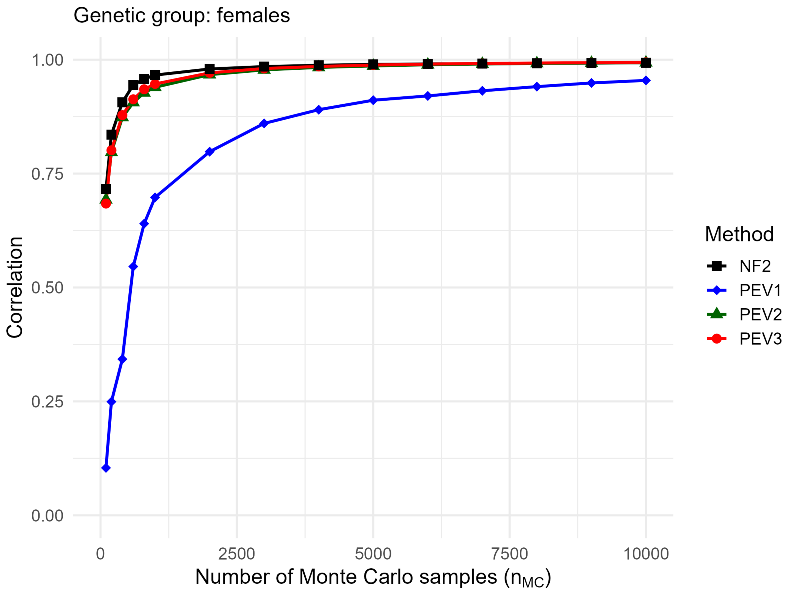

Figure 1: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. The correlation is calculated using different numbers of MC samples. Genetic group: females. Design: original.

More or less, it was the same phenomenon as seen in Figure 4, which happened at different paces using different methods and with different genetic groups. As the sample size increases, the sampled PEV values tend to be closer to the exact ones as presumed. The convergence was slowest with crosses. The reason for that is not self-evident. One reason might be that the range for the PEV values for crosses is small compared to the male and female groups. It is only from 5.09 to 5.33. It might be reasonable to think that it might take relatively many samples to estimate these PEV values precisely; however, the RMSEs and the MADs are small. The reason why PEV2 behaves so poorly might be related to this relatively small range of these values. Nonetheless, all methods seem to be working when the number of MC samples is large enough, as demonstrated using 200,000 MC samples.

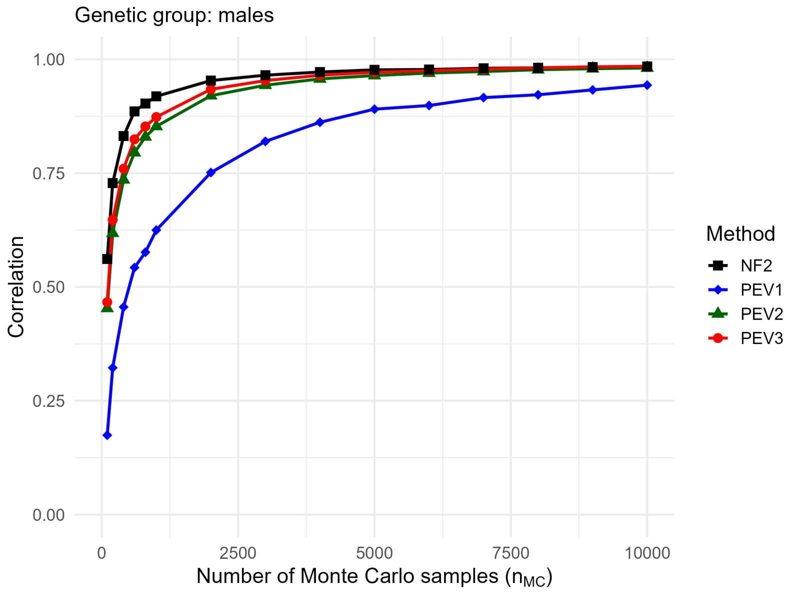

Figure 2: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. The correlation is calculated using different numbers of MC samples. Genetic group: males. Design: original.

7.3 Estimating reliabilities

As discussed in Section 4.3, there is a deterministic connection between PEV and the reliability of BLUP. As the breeders are usually more interested in the reliabilities of breeding values rather than PEVs, it is easy to present these same results in the reliability scale. As the main interest in this thesis was PEVs, we will not thoroughly analyse the accuracy and goodness of the prediction of the reliabilities. However, because reliability is always a value between 0 and 1, and PEVs can take any positive value, it is fruitful to investigate these reliabilities as well. We will use the same statistics presented in Section 7.1. Table 5 shows the same statistics for reliabilities calculated using 10,000 MC samples. Instead of comparing the exact PEV and the sampled PEV, we will compare the exact and sampled reliability. Thus, we will use the equation (14) with the corresponding diagonal values from $\mathbf{G}_{\text{ii}}$.

As Table 5 shows, the MADs are very low with almost all the methods. However, the recommended methods for approximating the reliabilities are PEV3 and NF2. They seem to work with all genetic groups.

Table 5: Comparison of different methods approximating the exact PEV and their estimates of the reliabilities with different statistics using 10,000 MC samples. Design: original.

Statistic	**Genetic group**	**PEV1**	**PEV2**	**PEV3**	**NF2**
Correlation	female	0.92	0.99	0.99	0.99
	male	0.89	0.97	0.98	0.98
	cross	0.93	0.37	0.93	0.90
RMSE	female	0.011	0.004	0.004	0.004
	male	0.010	0.005	0.004	0.004
	cross	0.000	0.002	0.000	0.000
Slope	female	0.84	0.98	0.98	0.98
	male	0.81	0.95	0.96	0.96
	cross	0.86	0.14	0.86	0.80
Intercept	female	0.12	0.01	0.01	0.02
	male	0.13	0.04	0.03	0.03
	cross	0.12	0.74	0.12	0.17
MAD	female	0.04	0.01	0.01	0.01
	male	0.03	0.02	0.01	0.01
	cross	0.00	0.00	0.00	0.00

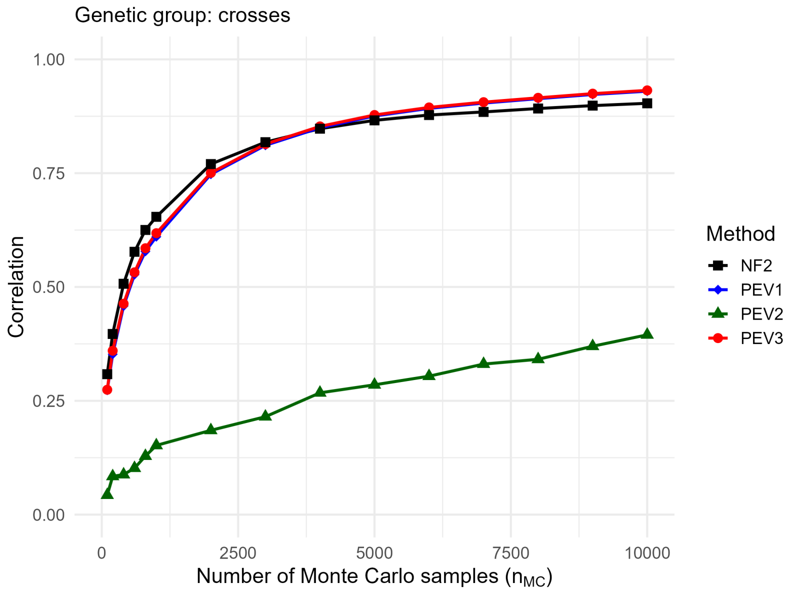

Figure 3: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. The correlation is calculated using different numbers of MC samples. Genetic group: crosses. Design: original.

7.4 Analysis with adjusted variances and with different number of observations

The behaviour of four different methods approximating the exact PEVs was also studied with changed variance components and a different number of observations. We changed the variances such that we doubled all the additive genetic variances while we halved our residual variance. Thus, the variance values in this design were: $\sigma^2_{\text{female}} = 30, \sigma^2_{\text{male}} = 20, \sigma^2_{\text{cross}} = 12,$ and $\sigma^2_{\text{e}} = 116.5$. The same analysis was done, and no significant differences were noticed, according to Appendix B. The distribution of the exact PEV with the changed variances for females was in the range of 4.98–14.25, with a mean of 9.39 and a standard deviation of 1.11. For males, the same statistics were 6.06–13.24 and (8.16, 0.61). For crosses, they were 7.60–9.49 and (8.37, 0.22).

We also conducted the analysis in a scenario where we used only half of the observations, such that we randomly took only four observations from each cross into the analysis. In this design, where only half of the observations were used, the exact PEV values for females were in the range of 5.01–15.46,

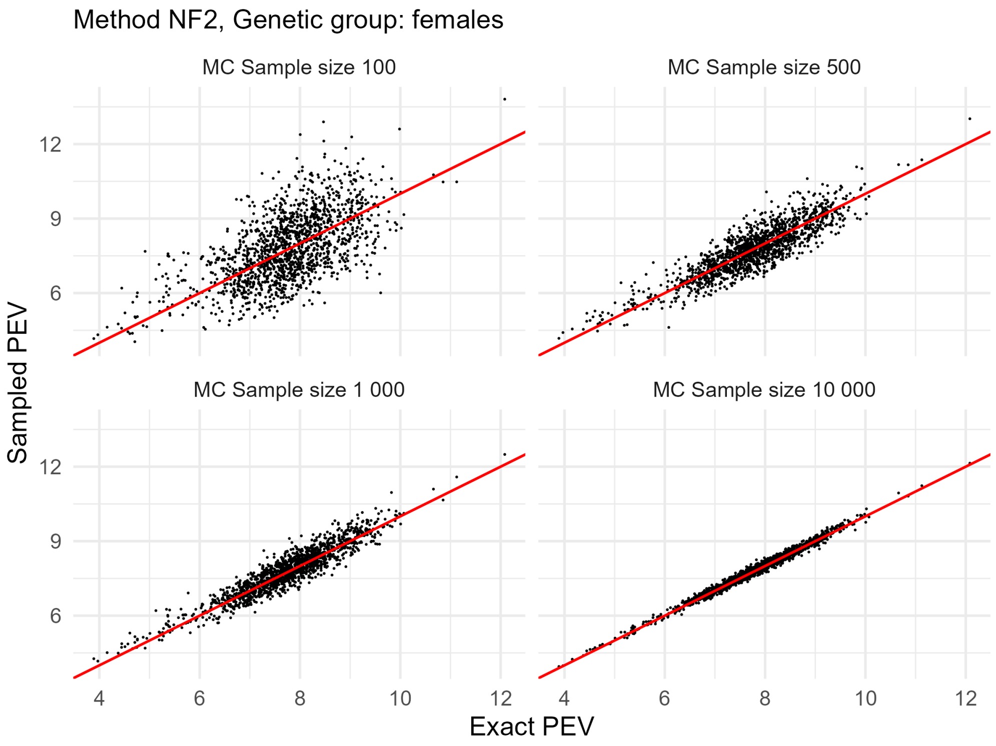

Figure 4: Scatterplots of the exact PEVs and the sampled PEVs using the method NF2. The scatterplots are drawn using four different numbers of MC samples. The red line represents the straight $y = x$. Genetic group: female plants. Design: original.

with a mean of 10.07 and a standard deviation of 1.25. These same statistics for males were 6.26–10.11 and (8.14, 0.59), and for crosses, 5.48–5.59 and (5.52, 0.01).

We will call the design, where we use all observations and original variances as the **original design**. The original design was presented in detail in Section 5. The design where we changed the variance components is called the **changed variances -design**. The design where we randomly took only half of the observations into the analysis is called the **smaller data -design**. Captions related to the figures and tables mention which design is in question. In addition, at the beginning of all appendices is a brief text indicating to which design pictures in that appendix are related.

Figures related to the changed variances -design are found in Appendix B. Figures related to the smaller data -design are found in Appendix C. All figures related to the original design introduced in Section 5 are found in Appendix D.

We will compare statistics between these three different designs: original design, changed variances -design, and smaller data -design using 10,000 MC samples. The results are found in Tables 6, 7, and 12. When comparing the correlations, we see that when there are fewer observations, the correlation of the exact PEV2 and the sampled PEV2 is only 0.14 for crosses. For other genetic groups, there were no significant differences in the correlations. The comparison of the correlation estimates between the designs is presented in Table 8. In Table 7, the correlation for PEV2 is 0.88 for crosses, so it seems that when there is more variance, the method PEV2 works better. RMSEs and MADs are not directly comparable between the designs since the absolute values of the exact PEVs are not the same due to a change in the designs. Overall, the method PEV2 is not recommended if the distribution of the exact PEV values is assumed to be very narrow. Also, the method PEV1 converges slower than the three other methods, so it is recommended to use the methods PEV3 and NF2: they work best in every situation covered in this thesis.

7.5 Computing times

The simulation (and analysis) was done using a computer, whose specifications are found in Appendix F. The simulation for 200,000 MC samples was

Table 6: Comparison of different methods approximating the exact PEV and their estimates with different statistics using 10,000 MC samples. Design: smaller data.

Statistic	**Genetic group**	**PEV1**	**PEV2**	**PEV3**	**NF2**
Correlation	female	0.98	0.99	0.99	0.99
	male	0.96	0.98	0.99	0.99
	cross	0.87	0.14	0.87	0.81
RMSE	female	0.27	0.15	0.13	0.13
	male	0.17	0.12	0.10	0.10
	cross	0.01	0.08	0.01	0.01
Slope	female	0.97	0.99	0.99	0.99
	male	0.92	0.96	0.97	0.97
	cross	0.75	0.02	0.76	0.65
Intercept	female	0.18	0.14	0.09	0.05
	male	0.67	0.29	0.23	0.24
	cross	1.36	5.40	1.35	1.92
MAD	female	1.02	0.62	0.43	0.50
	male	0.58	0.43	0.35	0.34
	cross	0.02	0.25	0.02	0.03

Table 7: Comparison of different methods approximating the exact PEV and their estimates with different statistics using 10,000 MC samples. Design: changed variances.

Statistic	**Genetic group**	**PEV1**	**PEV2**	**PEV3**	**NF2**
Correlation	female	0.83	0.99	0.99	0.99
	male	0.81	0.98	0.98	0.98
	cross	0.97	0.88	0.98	0.97
RMSE	female	0.71	0.13	0.13	0.16
	male	0.45	0.11	0.11	0.13
	cross	0.05	0.11	0.05	0.05
Slope	female	0.72	0.98	0.98	0.98
	male	0.64	0.97	0.97	0.95
	cross	0.95	0.77	0.96	0.95
Intercept	female	2.66	0.18	0.16	0.17
	male	2.95	0.25	0.25	0.39
	cross	0.42	1.90	0.36	0.42
MAD	female	2.61	0.56	0.55	0.62
	male	1.60	0.40	0.40	0.43
	cross	0.18	0.45	0.17	0.19

Table 8: Comparison of different methods approximating the exact PEV and their estimate of the **correlation** between the exact and sampled PEV using 10,000 MC samples. Design A refers to the original design. Design B refers to the changed variances -design. Design C refers to the smaller data -design.

Method	**Genetic group**	**Design**		
		A	**B**	**C**
PEV1	female	0.95	0.83	0.98
	male	0.94	0.81	0.96
	cross	0.93	0.97	0.87
PEV2	female	0.99	0.99	0.99
	male	0.98	0.98	0.98
	cross	0.37	0.88	0.14
PEV3	female	0.99	0.99	0.99
	male	0.99	0.98	0.99
	cross	0.93	0.98	0.87
NF2	female	0.99	0.99	0.99
	male	0.99	0.98	0.99
	cross	0.90	0.97	0.81

done using Puhti super computer provided by CSC - IT Center for Science, Finland (2024). Computing times for some MC sample sizes (n_{MC}) are presented in Table 9. In this context, computing time presents the combined time for simulating the random effects and solving the MME (12). The reported computing time is the median of five measurements. Computing times are measured in R using the function *system.time* (R Core Team, 2023).

The computing time measures both our functions `generate_simulated_u_and_y` and `solve_mme_s_times`. The code for the functions is found in Appendix E, where the R-code for this thesis is provided. As seen from Table 9, the time complexity for simulating and solving seems linear, as expected. It means that by doubling the MC sample size, the computing time also doubles. Thus, producing more MC samples is not a problem.

Table 9: Computing times in seconds for simulating the random effects and solving the MME (12) using different numbers of MC samples (n_{MC}). Computing time is the median of five measurements.

Sample size (n_{MC})	**Time (s)**
1,000	119
2,000	178
4,000	302
8,000	552
16,000	1,046
32,000	2,054

8 Conclusion

As we have seen, the basic methods for estimating PEVs of breeding values using GBLUP work, even with a hybrid model. The convergence for the PEVs of the crosses was slow, at least when using measures such as correlation, slope, and intercept. This may be due to the small range of the exact PEV values discussed in the previous section. Since the effects of the cross effect in the model are relatively small, one might argue that the cross effect should be left completely out of the model. However, when we changed the genetic variances, the range of the exact PEV values was wider for crosses, and the convergence was faster for crosses as well.

Based on the scenarios covered in this thesis, we recommend using the method

NF2 to approximate the exact PEVs of a single-trait hybrid model. The method performed best in almost every scenario handled in this thesis. However, for the crosses the method PEV3 worked even better.

The convergence of the correlation was slower than in Hickey et al. (2009), where the correlations exceeded 0.9 even with only 50 MC samples. However, the genetic and residual variances were much smaller there: genetic variance was 1.0, and the residual variance was 3.0; this might affect the correlation rate. Also, the values of the PEVs were all in the range of 0 to 1. Hickey et al. (2009) did not present any calculations for the reliabilities, so the comparison is not straightforward here. They also used the pedigree-based relationship matrix, not the genomic relationship matrix. Moreover, they did not have any genetic groups.

According to this thesis, the sufficient sample size depends on how accurate estimates we want to have and how much time we can use to simulate. 10,000 MC samples may be enough, at least with the methods PEV3 and NF2. The poor behaviour of crosses must be kept in mind. The sufficient sample size depends, at least, on genetic variances and the number of observations. It still remains unclear how the number of individuals affects the sufficient sample size.

The next step would be to try these methods with more complicated models, such as multi-trait models, and with more covariates. Also, the use of a SNP-BLUP model could be interesting. Another aspect that could be explored further is how the different levels of genetic and residual variances affect the behaviour of the sampled PEVs. As discussed, the expected values of the fixed effects do not affect the estimation of random effects, so adding the covariates to the model should not affect these calculations. However, the variance values certainly affect the approximation of PEV values. The structure of the genetic covariance matrix for the crosses is still questionable. Is it too simple?

The questions we would like to answer in the future are: How do the methods approximating the exact PEVs work using multi-trait models? How do the values of the genetic and residual variances affect the sufficient number of MC samples? How does the more accurate structure of the genetic covariance matrix for the crosses affect these results? Nevertheless, this thesis offers a thorough picture of using GBLUP with a hybrid model, even for statisticians with little animal and plant breeding knowledge. The first part of the thesis provides examples of animal and plant breeding basics and genetic models.

This thesis is an introduction to the still unanswered questions that were presented. The thesis is a good foundation for further research on this topic.

References

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* Wiley Series in Probability and Statistics. Wiley. ISBN: 9781118730034.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter (2002). *Molecular Biology of the Cell.* 4th. New York: Garland Science. ISBN: 9780815332183.

Ambainis, A., Y. Filmus, and F. Le Gall (2015). "Fast Matrix Multiplication: Limitations of the Coppersmith-Winograd Method". *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing.* STOC '15. Portland, Oregon, USA: Association for Computing Machinery, pp. 585–593. DOI: 10.1145/2746539.2746554.

Ben Zaabza, H., E.A. Mäntysaari, and I. Strandén (2020). "Using Monte Carlo method to include polygenic effects in calculation of SNP-BLUP model reliability". *Journal of Dairy Science* 103.6, pp. 5170–5182. DOI: https://doi.org/10.3168/jds.2019-17255.

Ben Zaabza, H., C. P. Van Tassell, J. Vandenplas, P. VanRaden, Z. Liu, H. Eding, S. McKay, K. Haugaard, M. H. Lidauer, E. A. Mäntysaari, and I. Strandén (2023). "Invited review: Reliability computation from the animal model era to the single-step genomic model era". *Journal of Dairy Science* 106.3, pp. 1518–1532. DOI: https://doi.org/10.3168/jds.2022-22629.

Campbell, N. A., L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, and J. B. Reece (2018). *Biology: A Global Approach.* 11th, global edition. Pearson. ISBN: 9781292170442.

CSC - IT Center for Science, Finland (2024). *Suomen supertietokoneet: Mahti ja Puhti.* Accessed: 2024-08-26. URL: https://csc.fi/osaamisemme/suurteholaskenta/mahti-ja-puhti-supertietokoneet/.

Falconer, D. S. and T. F. C. Mackay (1996). *Introduction to Quantitative Genetics.* 4th. Harlow: Prentice Hall. ISBN: 0582243025.

García-Cortés, L. A., C. Moreno, L. Varona, and J. Altarriba (1995). "Estimation of prediction-error variances by resampling". *Journal of Animal Breeding and Genetics* 112.1-6, pp. 176–182. DOI: https://doi.org/10.1111/j.1439-0388.1995.tb00556.x.

Gareth, J., D. Witten, T. Hastie, and R. Tibshirani (2021). *An Introduction to Statistical Learning with Applications in R*. 2nd. Spinger. ISBN: 9781071614174.

Gezan, S. A., A. A. de Oliveira, G. Galli, and D. Murray (2022). *ASRgenomics: An R package with Complementary Genomic Functions*. Version 1.1.0. User's Manual for ASRgenomics v. 1.1.0. VSN International. Hemel Hempstead, United Kingdom. URL: https://CRAN.R-project.org/package=ASRgenomics.

Grammarly, Inc. (2024). *Grammarly*. Last used on: 2024-09-19. URL: https://www.grammarly.com.

Henderson, C. R. (1976). "A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values". *Biometrics* 32.1, pp. 69–83. DOI: https://doi.org/10.2307/2529339.

— (1975). "Best linear unbiased estimation and prediction under a selection model". *Biometrics*, pp. 423–447. DOI: https://doi.org/10.2307/2529430.

— (1963). "Selection Index and Expected Genetic Advance". *Statistical Genetics and Plant Breeding* 982, pp. 141–163.

Hickey, J. M., R. F. Veerkamp, M. P. Calus, H. A. Mulder, and R. Thompson (2009). "Estimation of prediction error variances via Monte Carlo sampling methods using different formulations of the prediction error variance". *Genetics Selection Evolution* 41. DOI: https://doi.org/10.1186/1297-9686-41-23.

Hill, W.G. and B.S. Weir (2011). "Variation in actual relationship as a consequence of Mendelian sampling and linkage". *Genetics Research* 93.1, pp. 47–64. DOI: https://10.1017/S0016672310000480.

Himmelbauer, J., H. Schwarzenbacher, and C. Fuerst (2021). "Implementation of single-step evaluations for fitness traits in the German and Austrian Fleckvieh and Brown Swiss populations". *Interbull Bulletin* 56, pp. 82–89. URL: https://journal.interbull.org/index.php/ib/article/view/79/79.

Hollifield, M. K., M. Bermann, D. Lourenco, and I. Misztal (2022). "Impact of blending the genomic relationship matrix with different levels of pedigree relationships or the identity matrix on genetic evaluations". *JDS Communications* 3.5, pp. 343–347. DOI: https://doi.org/10.3168/jdsc.2022-0229.

Isik, F., J. Holland, and C. Maltecca (2017). *Genetic Data Analysis for Plant and Animal Breeding*. Springer. ISBN: 9783319551777.

Juga, J., K. Maijala, A. Mäki-Tanila, E. Mäntysaari, M. Ojala, and J. Syväjärvi (1999). *Kotieläinjalostus*. suomi. Suomen Kotieläinjalostusosuuskunta. ISBN: 9519635653.

Luo, P., H. Wang, Z. Ni, R. Yang, F. Wang, H. Yong, L. Zhang, Z. Zhou, W. Song, M. Li, J. Yang, J. Weng, Z. Meng, D. Zhang, J. Han, Y. Chen, R. Zhang, L. Wang, M. Zhao, W. Gao, X. Chen, W. Li, Z. Hao, J. Fu, X. Zhang, and X. Li (2023). "Genomic prediction of yield performance among single-cross maize hybrids using a partial diallel cross design". *The Crop Journal* 11.6, pp. 1884–1892. DOI: `https://doi.org/10.1016/j.cj.2023.09.009`.

McEvoy, B. P. and P. M. Visscher (2009). "Genetics of human height". *Economics & Human Biology* 7.3, pp. 294–306. DOI: `https://doi.org/10.1016/j.ehb.2009.09.005`.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard (2001). "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps". *Genetics* 157.4, pp. 1819–1829. DOI: `10.1093/genetics/157.4.1819`.

Mrode, R.A. and R. Thompson (2014). *Linear Models for the Prediction of Animal Breeding Values*. 3rd. Cabi. ISBN: 9781780643908.

OpenAI (2024). *ChatGPT*. Last used on: 2024-08-16. URL: `https://www.openai.com/chatgpt`.

Putz, A. (2018). *AnS 562A - Introduction to GBLUP and single-step GBLUP*. URL: `https://rpubs.com/amputz/GBLUP_and_ssGBLUP` (visited on 05/20/2024).

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT press. ISBN: 9780262182539.

Schaeffer, L.R. (2006). "Strategy for applying genome-wide selection in dairy cattle". *Journal of Animal Breeding and Genetics* 123.4, pp. 218–223. DOI: `https://doi.org/10.1111/j.1439-0388.2006.00595.x`.

Slatkin, M. (2008). "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future". *Nature Reviews Genetics* 9.6, pp. 477–485. DOI: `https://doi.org/10.1038/nrg2361`.

Strandén, I. and M. Lidauer (1999). "Solving Large Mixed Linear Models Using Preconditioned Conjugate Gradient Iteration". *Journal of Dairy Science* 82.12, pp. 2779–2787. DOI: `https://doi.org/10.3168/jds.S0022-0302(99)75535-9`.

Van Vleck, L. D. (1998). "Charles Roy Henderson (1911–1989): A Biographical Memoir". *Biographical Memoirs*. URL: `https://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/henderson-charles.pdf`.

VanRaden, P.M. (2008). “Efficient Methods to Compute Genomic Predictions”. *Journal of Dairy Science* 91.11, pp. 4414–4423. DOI: https://doi.org/10.3168/jds.2007-0980.

Wang, H., I. Misztal, and A. Legarra (2014). “Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals”. *Journal of Animal Breeding and Genetics* 131.6, pp. 445–451. DOI: https://doi.org/10.1111/jbg.12109.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer. ISBN: 9783319242774.

Appendix A

In this appendix there is additional tables of statistics calculated using 100, 1,000, 10,000, and 200,000 MC samples (n_{MC}). They are calculated using the original design. The simulation using 200,000 MC samples were done using Puhti super computer, the results are in Table 13 (CSC - IT Center for Science, Finland, 2024).

Table 10: Comparison of different methods approximating the exact PEV and their estimates with different statistics using 100 MC samples. Design: original.

Statistic	**Genetic group**	**PEV1**	**PEV2**	**PEV3**	**NF2**
Correlation	female	0.27	0.68	0.70	0.66
	male	0.22	0.45	0.48	0.47
	cross	0.26	0.06	0.27	0.22
RMSE	female	3.10	1.08	1.04	1.14
	male	1.90	0.94	0.87	0.89
	cross	0.12	0.74	0.12	0.15
Slope	female	0.08	0.44	0.46	0.42
	male	0.06	0.21	0.23	0.23
	cross	0.06	0.00	0.06	0.04
Intercept	female	7.13	4.34	4.18	4.51
	male	6.07	5.13	4.94	4.98
	cross	4.86	5.17	4.85	4.96
MAD	female	11.34	4.79	3.79	3.99
	male	6.78	4.35	3.40	3.54
	cross	0.49	3.29	0.49	0.65

Table 11: Comparison of different methods approximating the exact PEV and their estimates with different statistics using 1,000 MC samples. Design: original.

Statistic	**Genetic group**	**PEV1**	**PEV2**	**PEV3**	**NF2**
Correlation	female	0.73	0.94	0.95	0.94
	male	0.61	0.85	0.88	0.86
	cross	0.63	0.14	0.63	0.56
RMSE	female	1.00	0.35	0.33	0.37
	male	0.64	0.29	0.26	0.28
	cross	0.04	0.24	0.04	0.04
Slope	female	0.48	0.89	0.89	0.86
	male	0.36	0.74	0.77	0.74
	cross	0.39	0.02	0.40	0.30
Intercept	female	4.04	0.82	0.82	1.09
	male	4.13	1.69	1.47	1.66
	cross	3.16	5.09	3.12	3.61
MAD	female	3.46	1.12	1.10	1.45
	male	2.32	0.99	0.83	0.95
	cross	0.12	0.82	0.13	0.17

Table 12: Comparison of different methods approximating the exact PEV and their estimates with different statistics using 10,000 MC samples. Design: original.

Statistic	**Genetic group**	**PEV1**	**PEV2**	**PEV3**	**NF2**
Correlation	female	0.95	0.99	0.99	0.99
	male	0.94	0.98	0.99	0.99
	cross	0.93	0.37	0.93	0.90
RMSE	female	0.32	0.11	0.11	0.12
	male	0.19	0.09	0.08	0.08
	cross	0.01	0.07	0.01	0.01
Slope	female	0.90	0.99	0.99	0.99
	male	0.83	0.97	0.97	0.96
	cross	0.86	0.14	0.86	0.80
Intercept	female	0.77	0.08	0.09	0.11
	male	1.10	0.18	0.20	0.27
	cross	0.74	4.48	0.73	1.05
MAD	female	1.42	0.40	0.37	0.42
	male	0.60	0.37	0.29	0.30
	cross	0.04	0.26	0.04	0.05

Table 13: Comparison of different methods approximating the exact PEV and their estimates with different statistics using 200,000 MC samples. Design: original.

Statistic	**Genetic group**	**PEV1**	**PEV2**	**PEV3**	**NF2**
Correlation	female	0.997	>0.999	0.999	>0.999
	male	0.996	0.999	0.999	0.999
	cross	0.996	0.868	0.970	0.994
RMSE	female	0.072	0.026	0.038	0.027
	male	0.045	0.021	0.025	0.021
	cross	0.003	0.017	0.007	0.003
Slope	female	0.989	0.999	0.995	0.997
	male	0.988	0.996	0.995	0.996
	cross	0.992	0.763	0.945	0.990
Intercept	female	0.101	0.009	0.043	0.024
	male	0.088	0.022	0.038	0.027
	cross	0.041	1.226	0.284	0.051
MAD	female	0.232	0.093	0.128	0.102
	male	0.175	0.078	0.091	0.075
	cross	0.010	0.078	0.034	0.012

Appendix B

In this appendix, there are all the figures related to the changed variances -design.

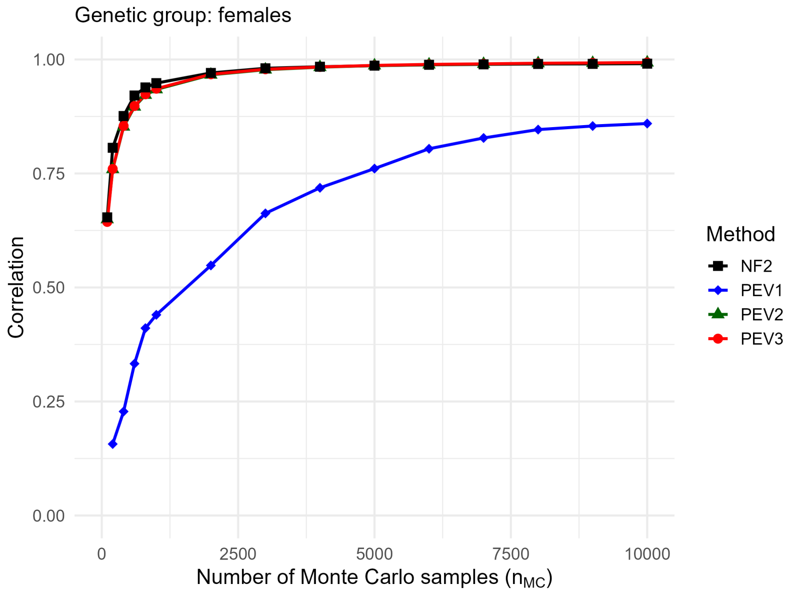

Figure 5: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. The correlation is calculated using different numbers of MC samples. Genetic group: females. Design: changed variances.

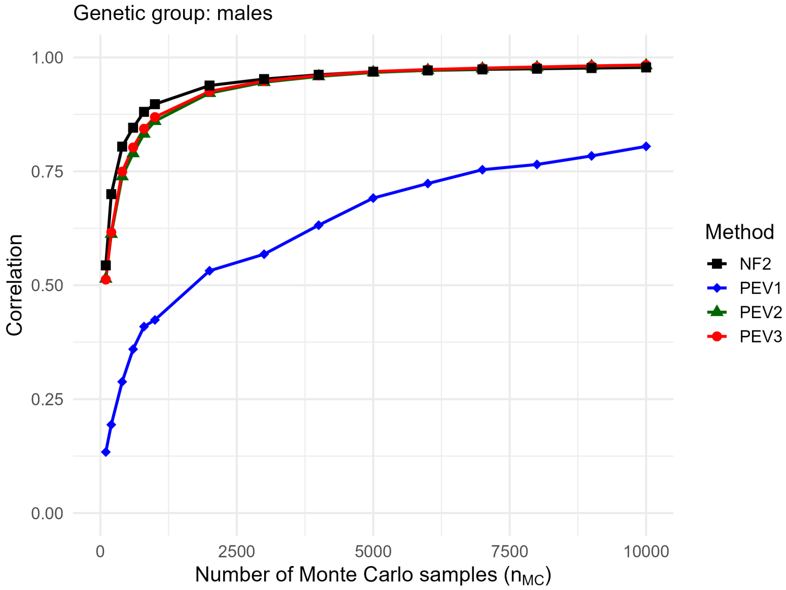

Figure 6: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. Genetic group: males. Design: changed variances.

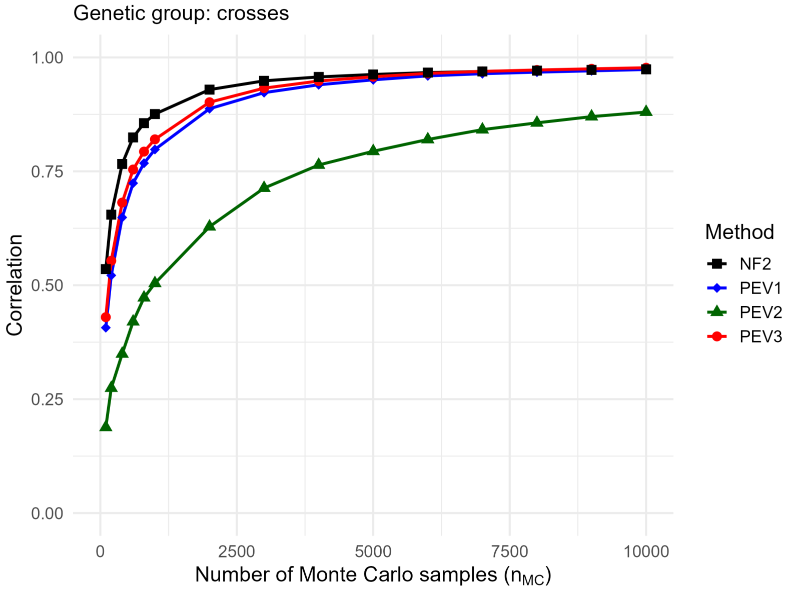

Figure 7: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. Genetic group: crosses. Design: changed variances.

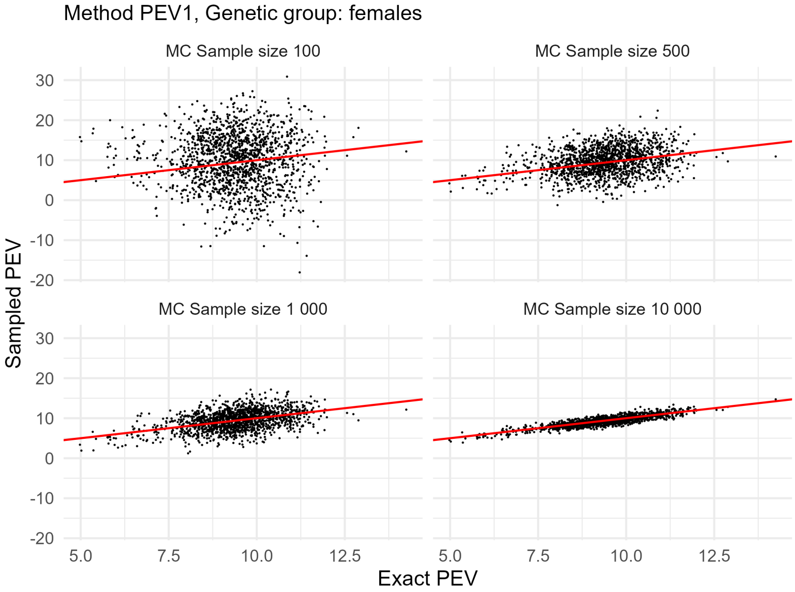

Figure 8: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: changed variances.

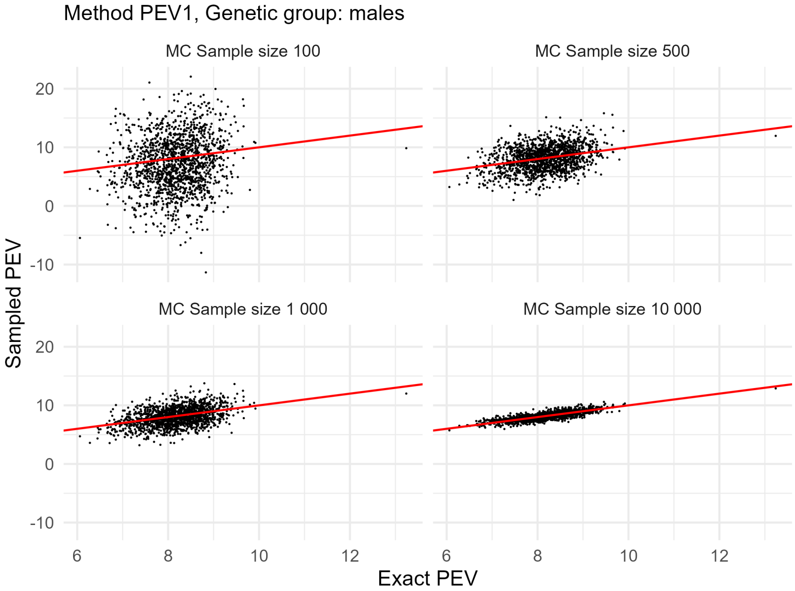

Figure 9: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: changed variances.

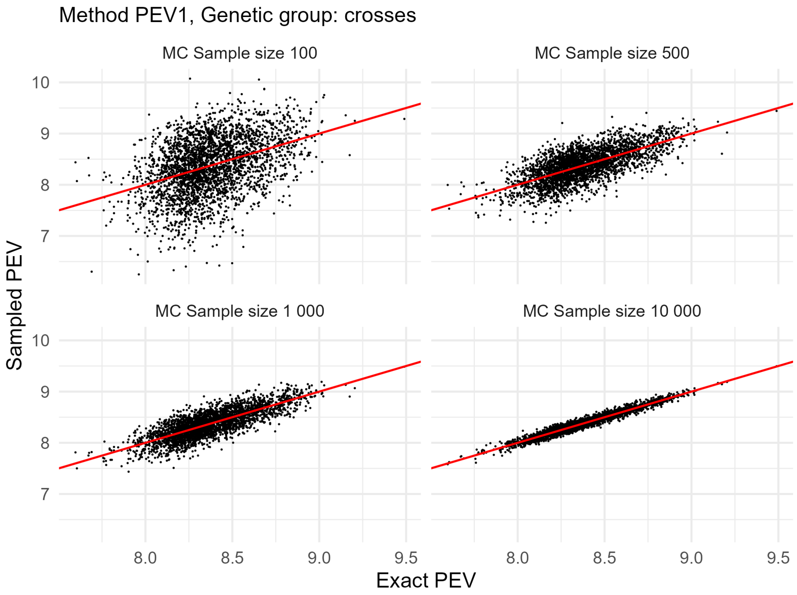

Figure 10: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: changed variances.

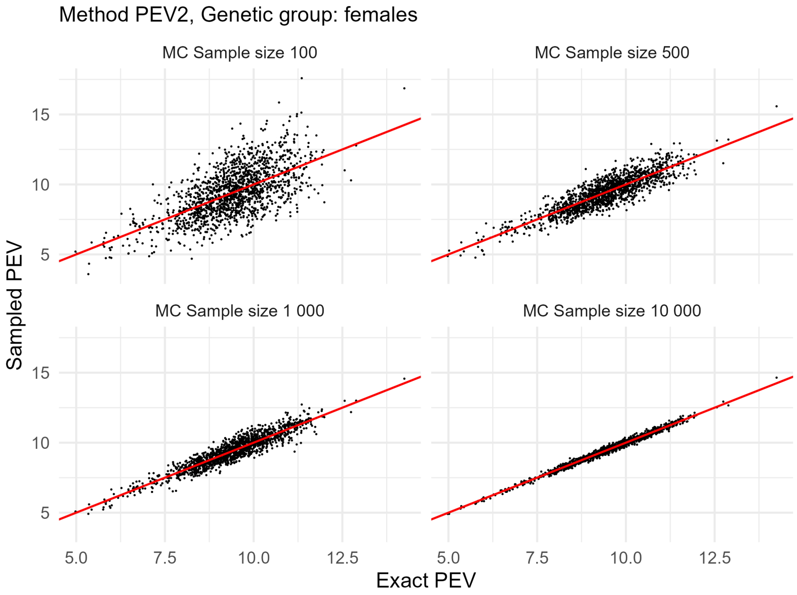

Figure 11: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: changed variances.

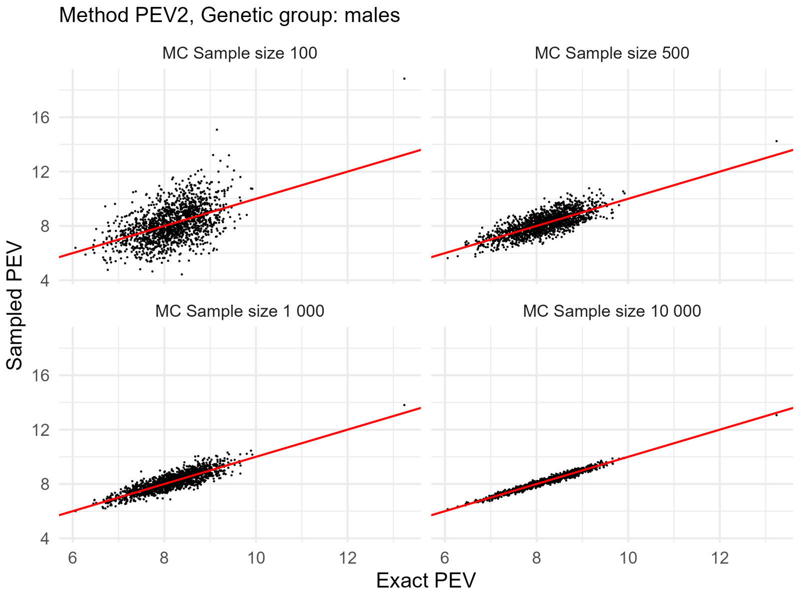

Figure 12: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: changed variances.

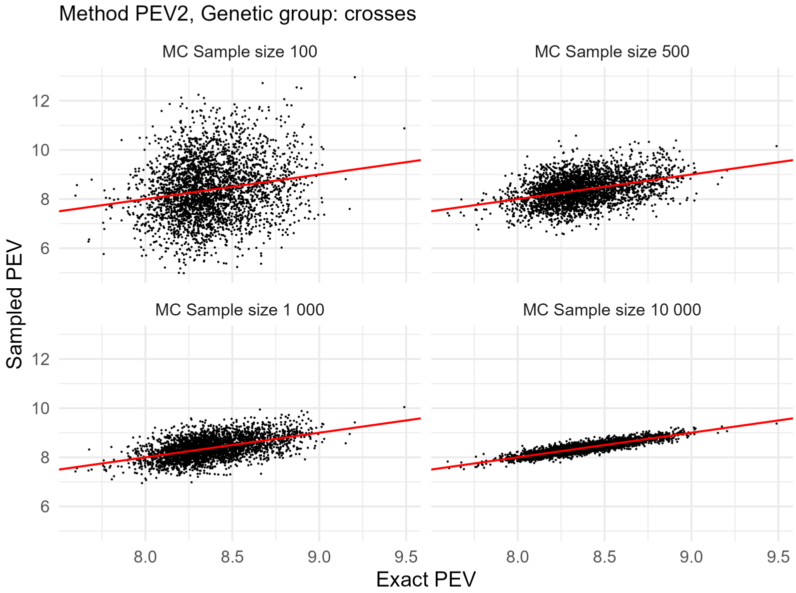

Figure 13: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: changed variances.

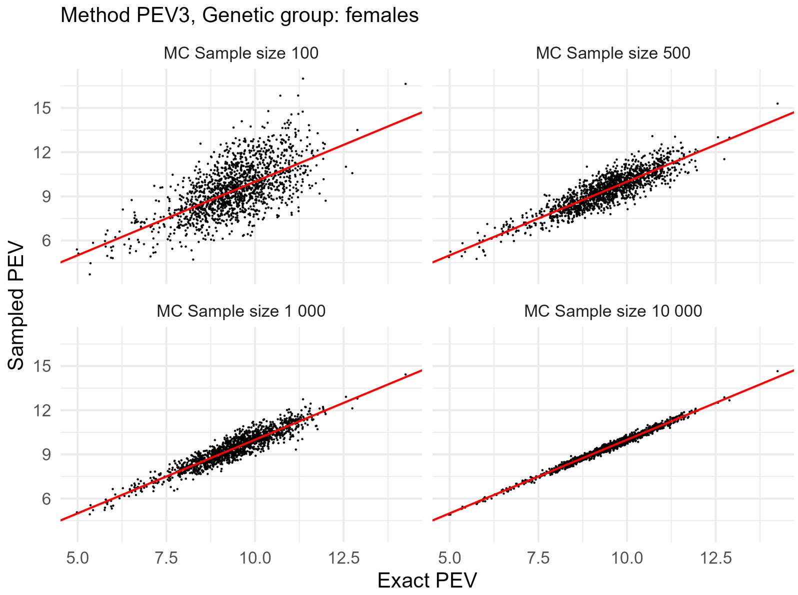

Figure 14: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: changed variances.

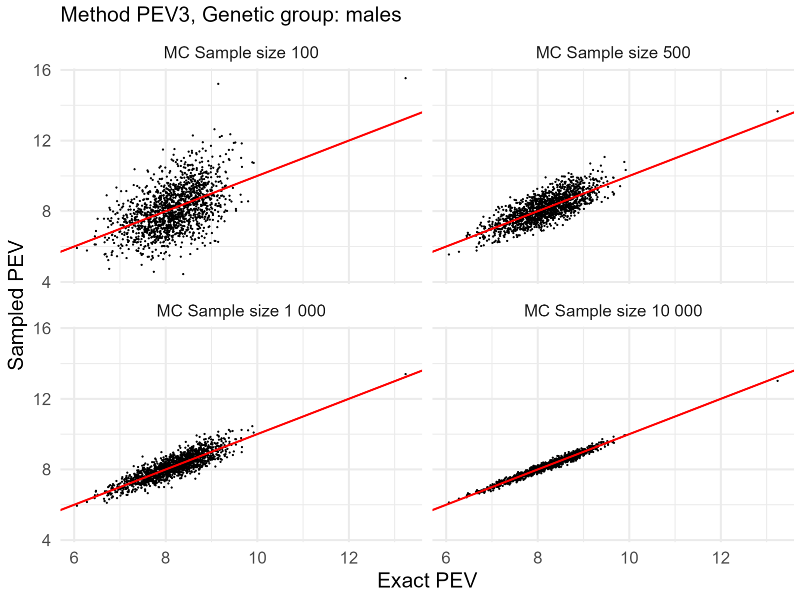

Figure 15: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: changed variances.

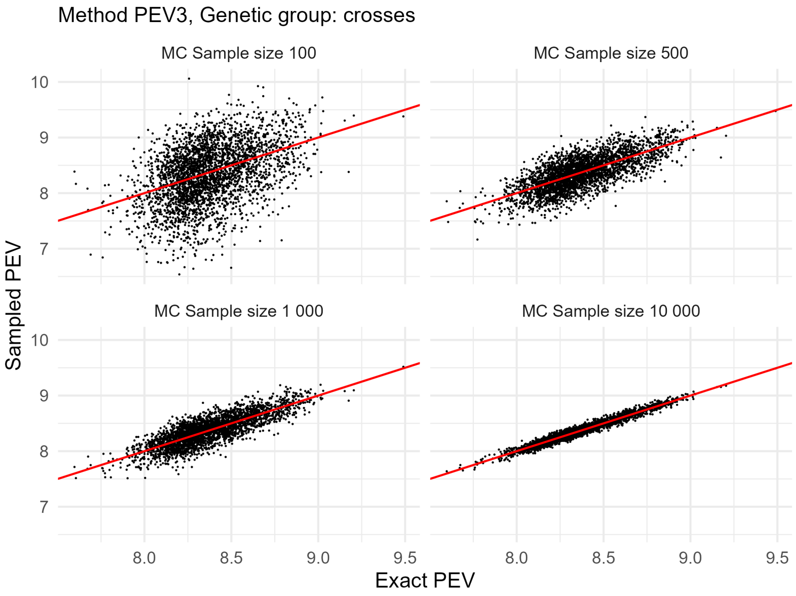

Figure 16: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: changed variances.

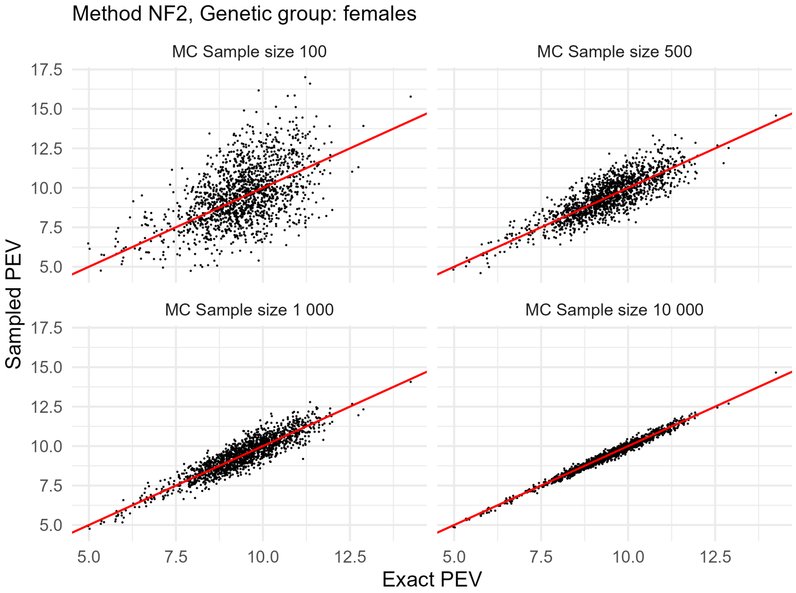

Figure 17: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: changed variances.

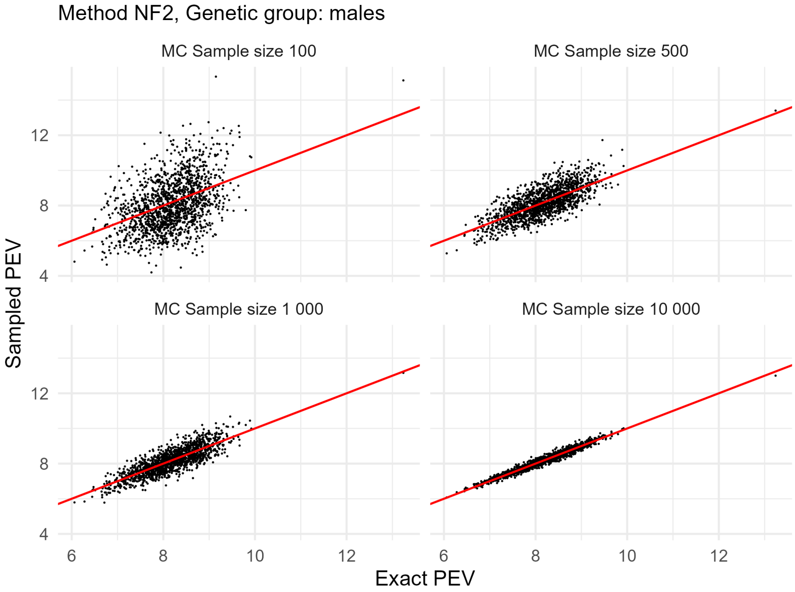

Figure 18: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: changed variances.

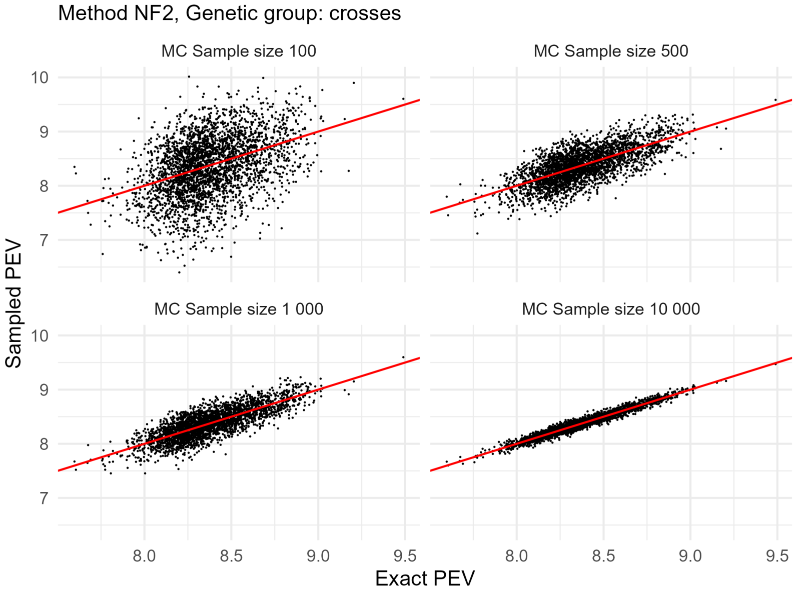

Figure 19: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: changed variances.

Appendix C

In this appendix, there are all the figures related to the smaller data -design.

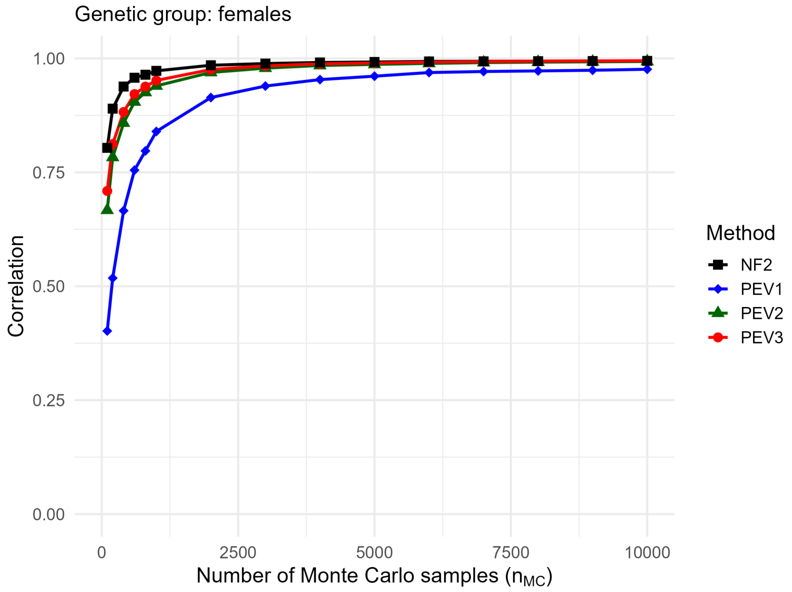

Figure 20: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. The correlation is calculated using different numbers of MC samples. Genetic group: females. Design: smaller data.

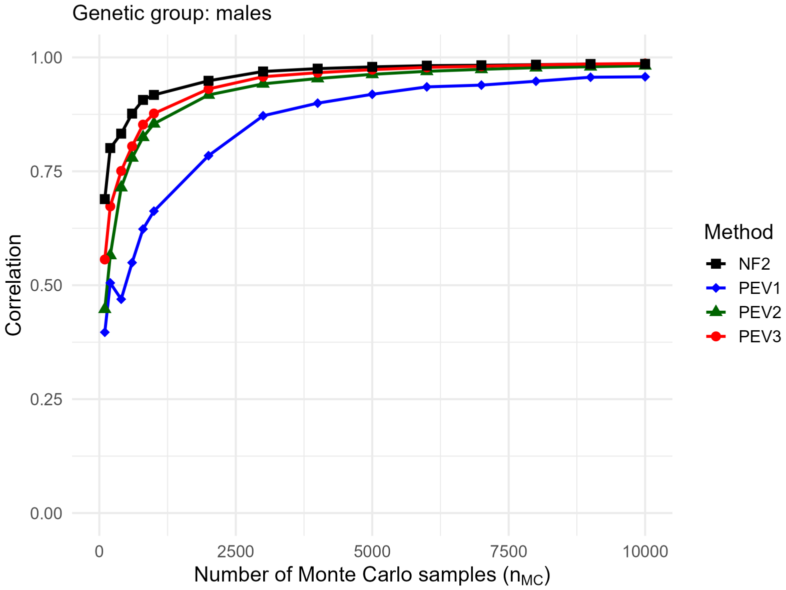

Figure 21: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. Genetic group: males. Design: smaller data.

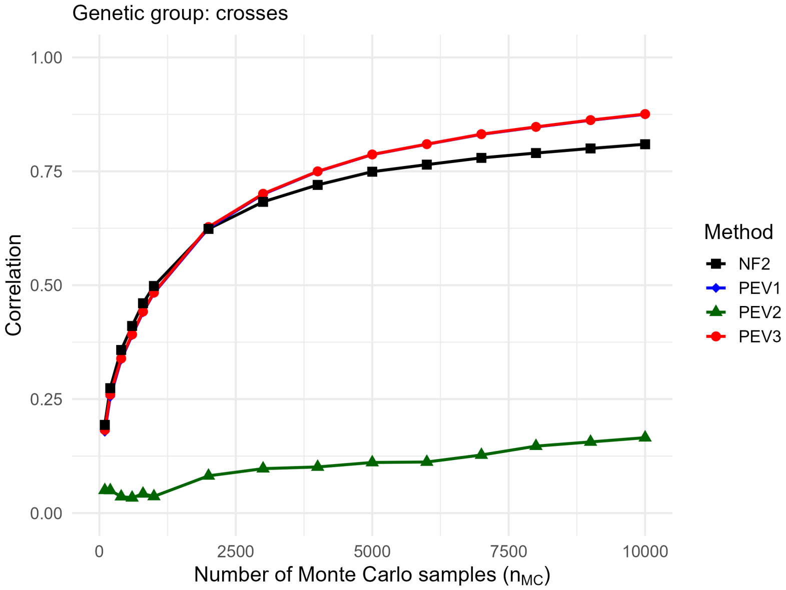

Figure 22: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. Genetic group: crosses. Design: smaller data.

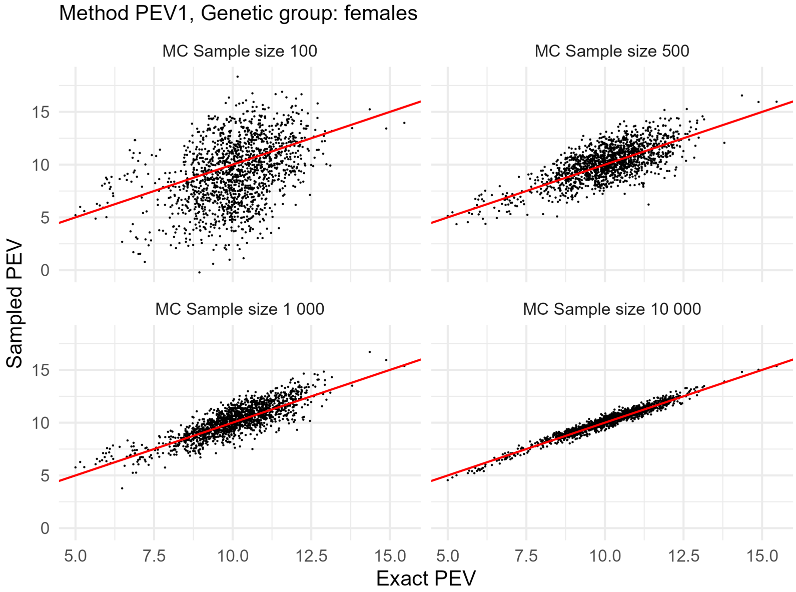

Figure 23: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: smaller data.

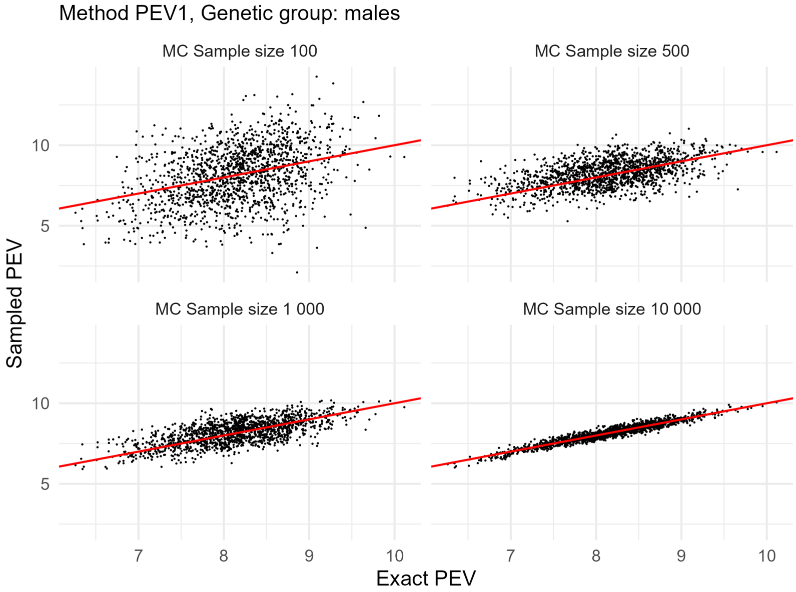

Figure 24: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: smaller data.

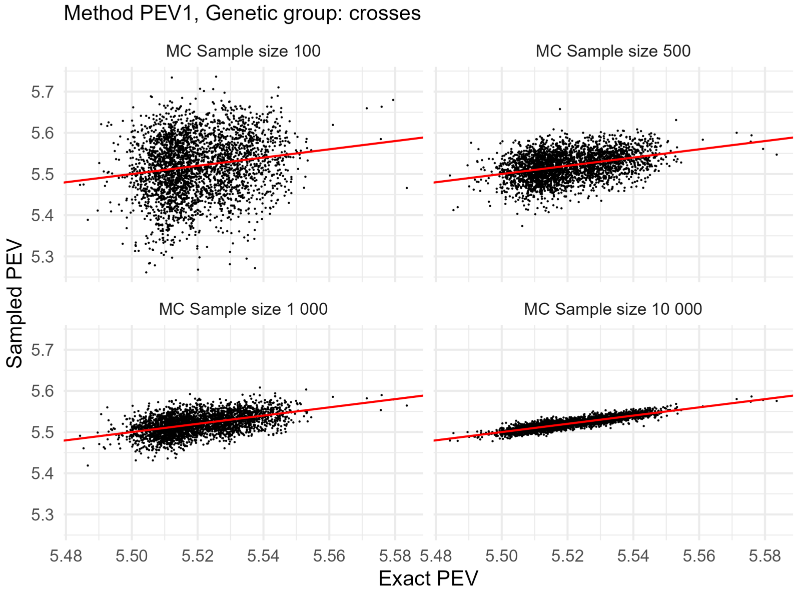

Figure 25: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: smaller data.

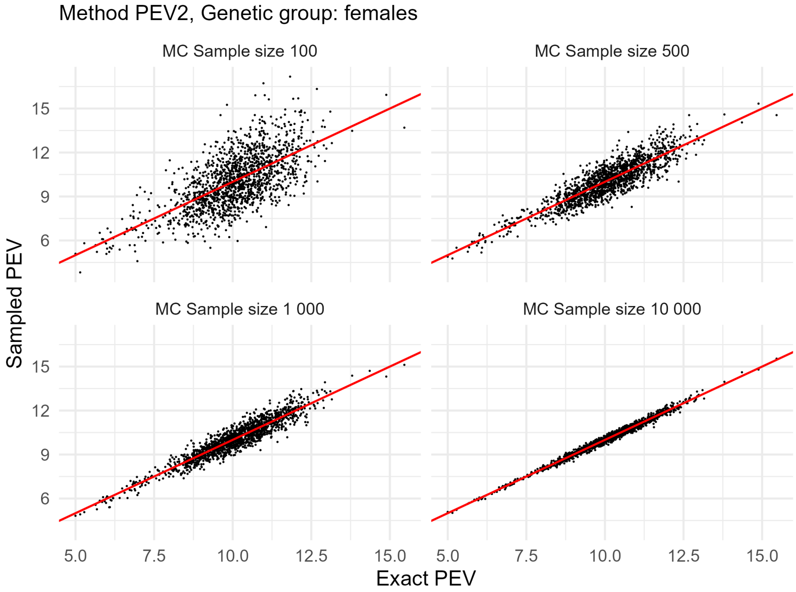

Figure 26: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: smaller data.

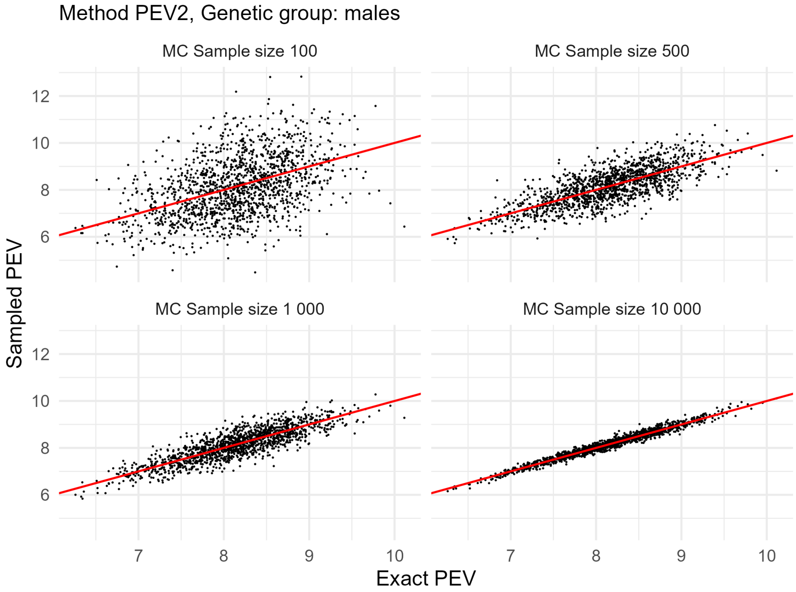

Figure 27: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: smaller data.

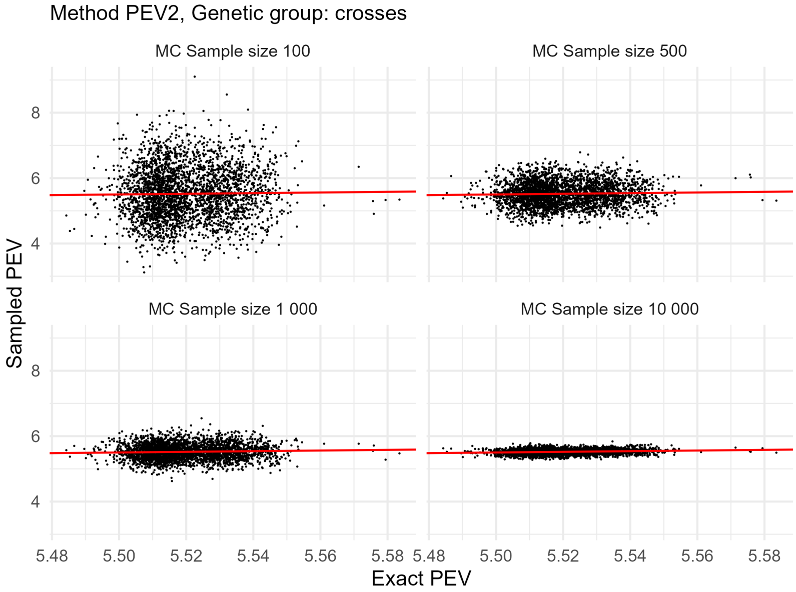

Figure 28: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: smaller data.

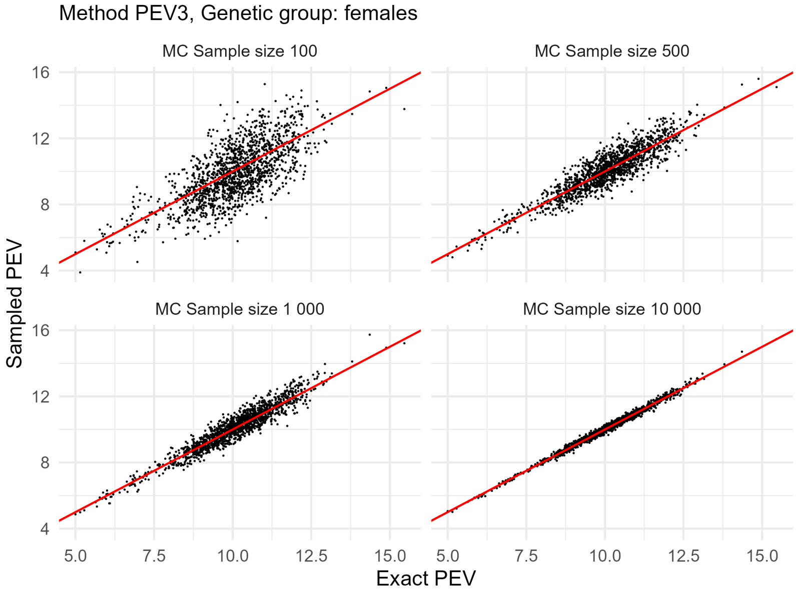

Figure 29: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: smaller data.

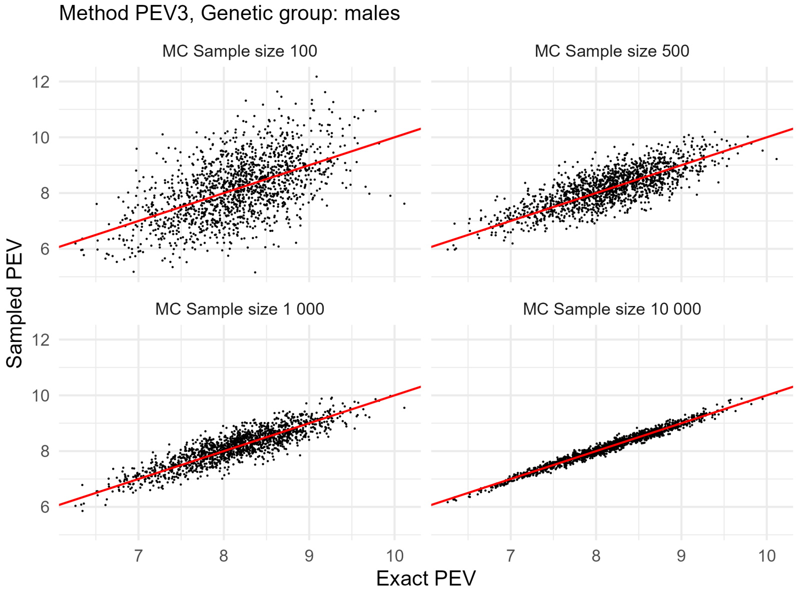

Figure 30: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: smaller data.

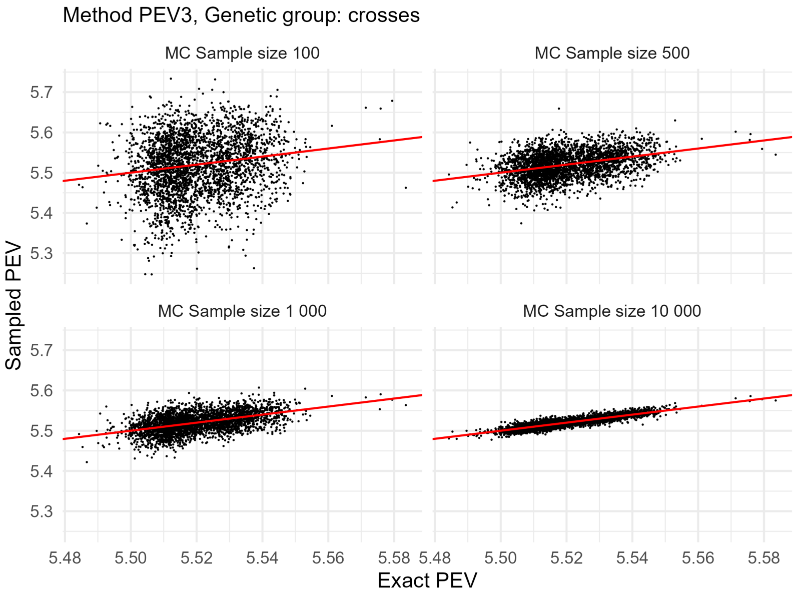

Figure 31: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: smaller data.

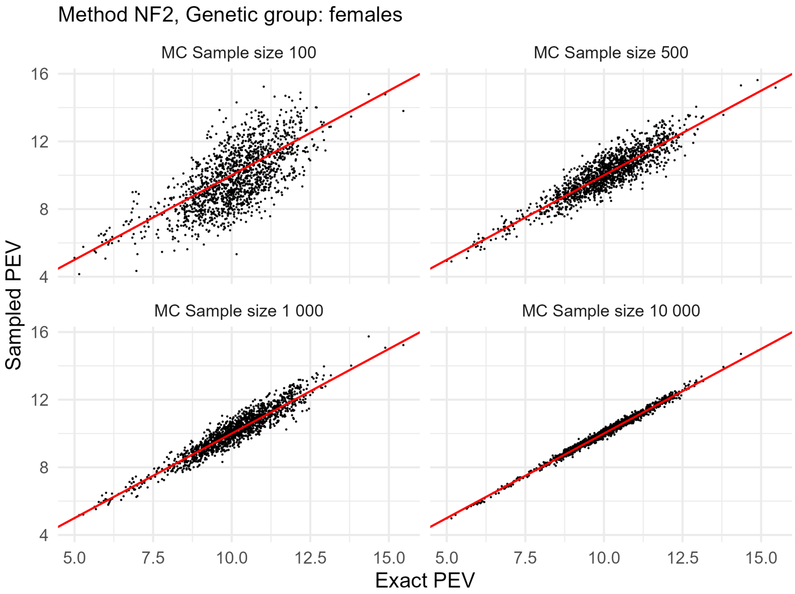

Figure 32: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: smaller data.

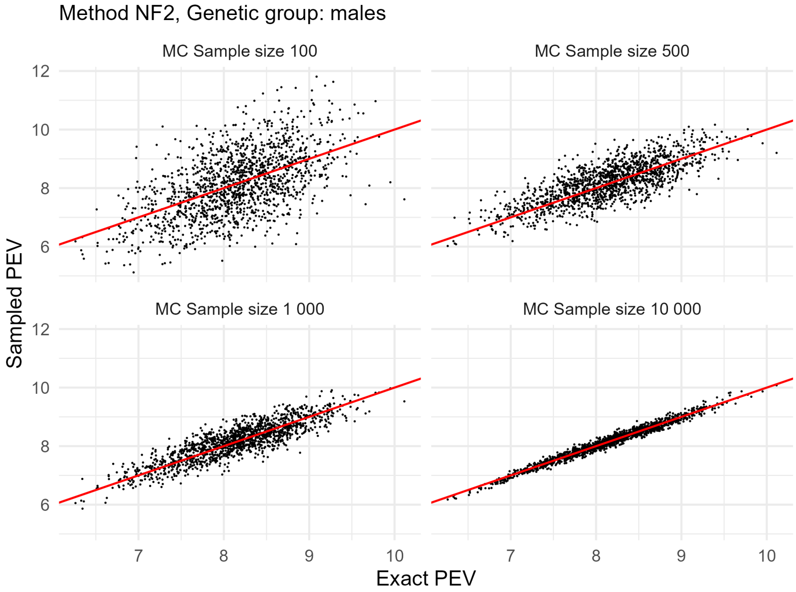

Figure 33: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: smaller data.

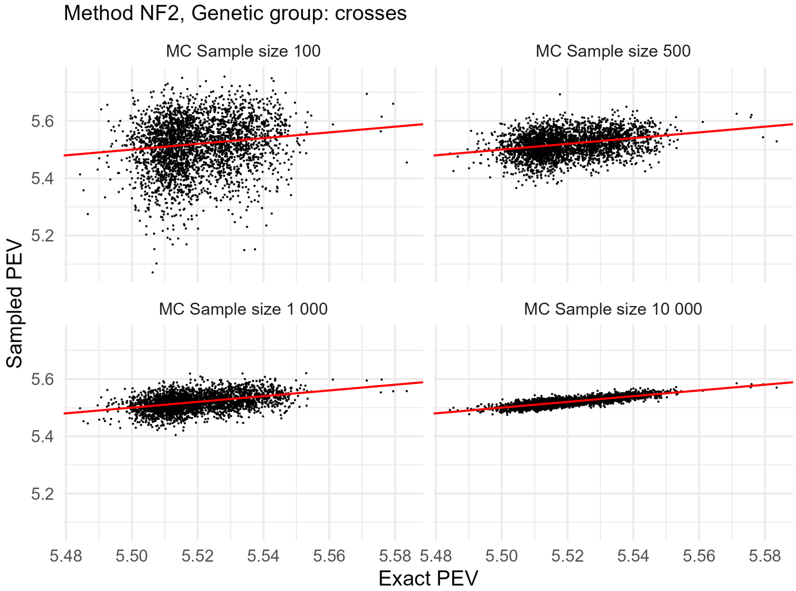

Figure 34: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different number of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: smaller data.

Appendix D

In this appendix, there are all the figures related to the original design.

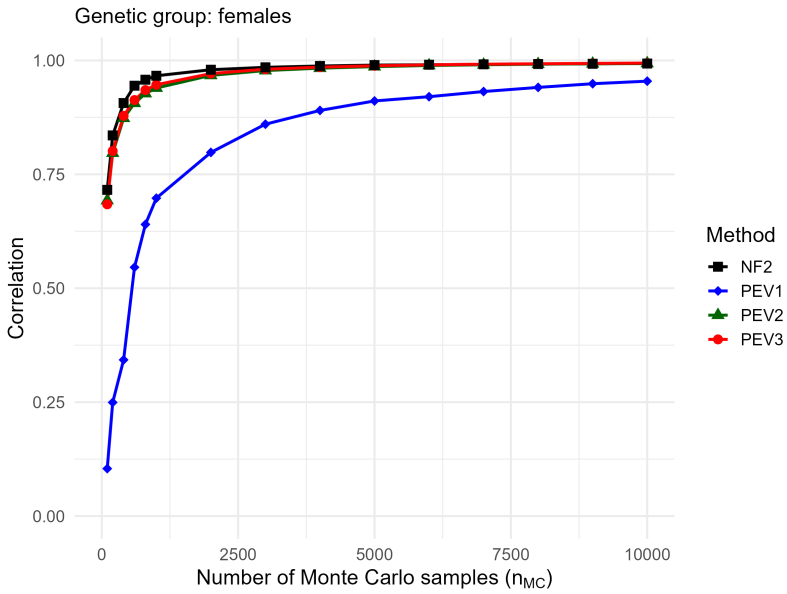

Figure 35: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. The correlation is calculated using different numbers of MC samples. Genetic group: females. Design: original.

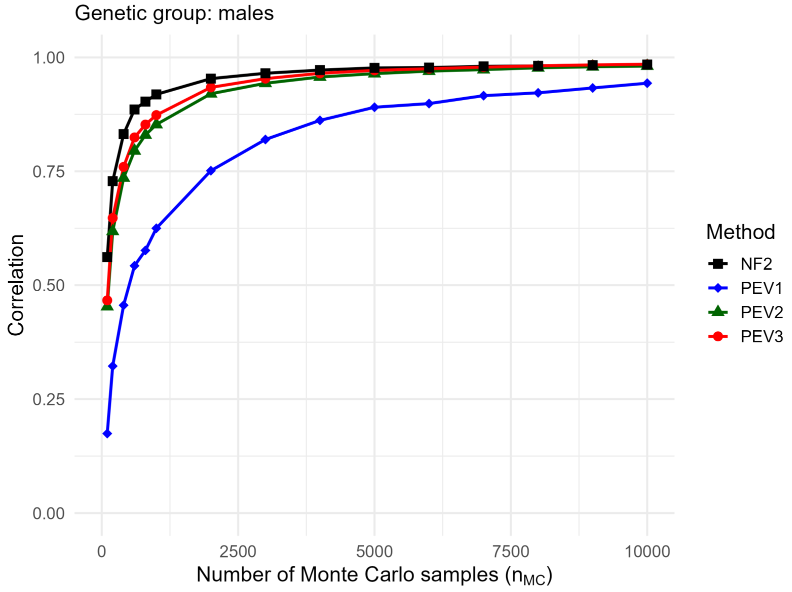

Figure 36: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. Genetic group: males. Design: original.

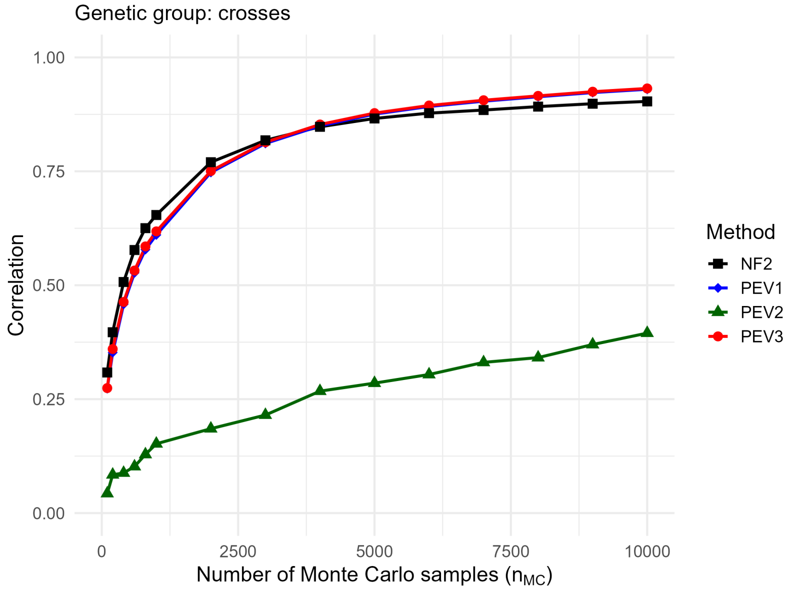

Figure 37: Correlation plots with the exact PEV and the sampled PEV using four different methods approximating the exact PEV. Genetic group: crosses. Design: original.

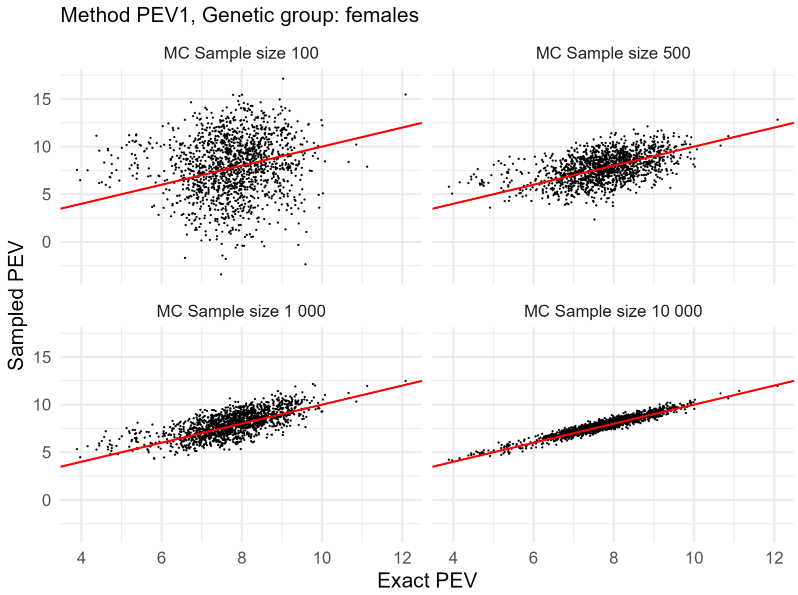

Figure 38: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: original.

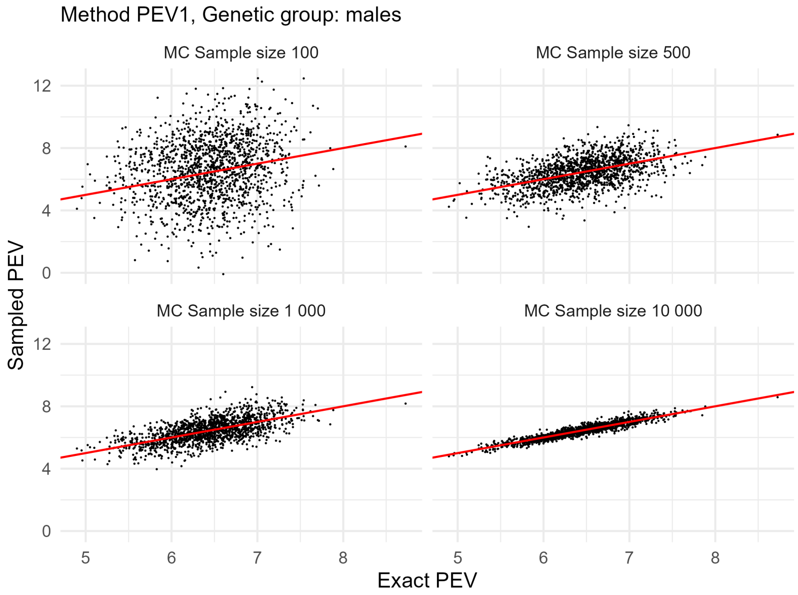

Figure 39: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: original.

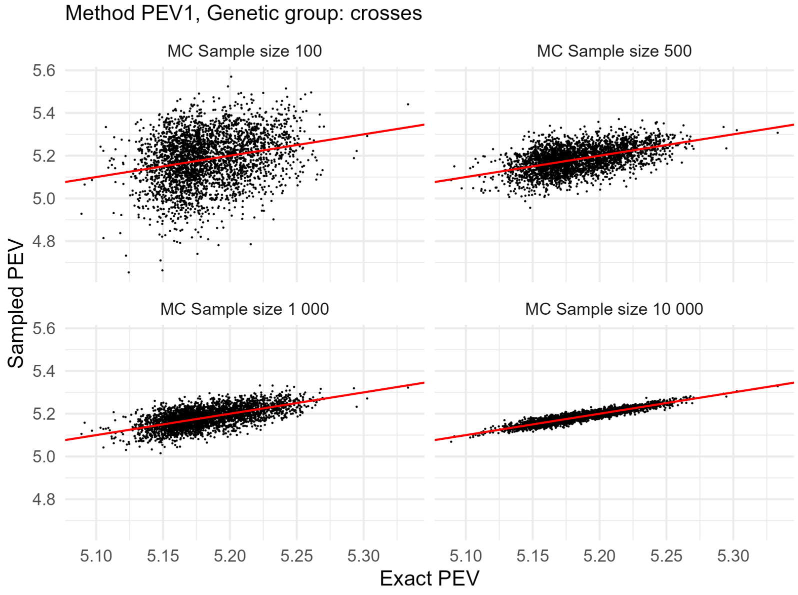

Figure 40: Scatterplots of the exact PEVs and the sampled PEVs using method PEV1 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: original.

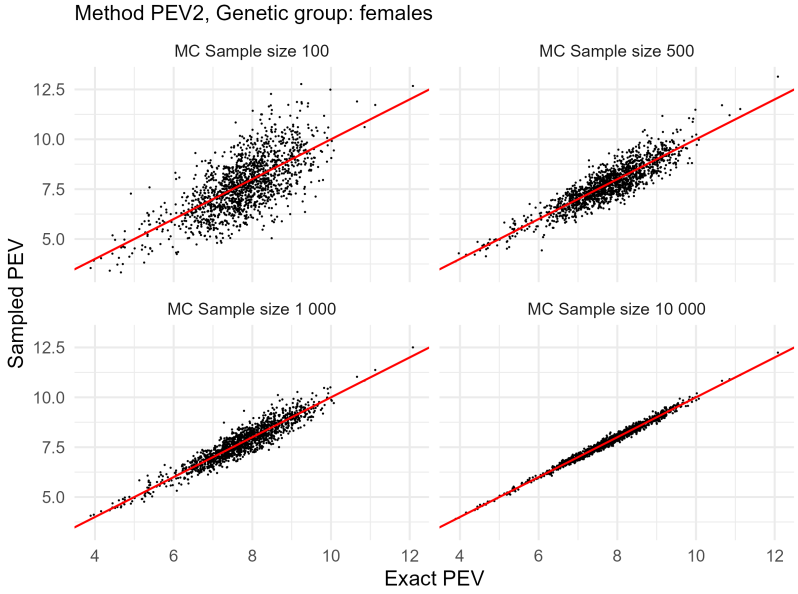

Figure 41: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: original.

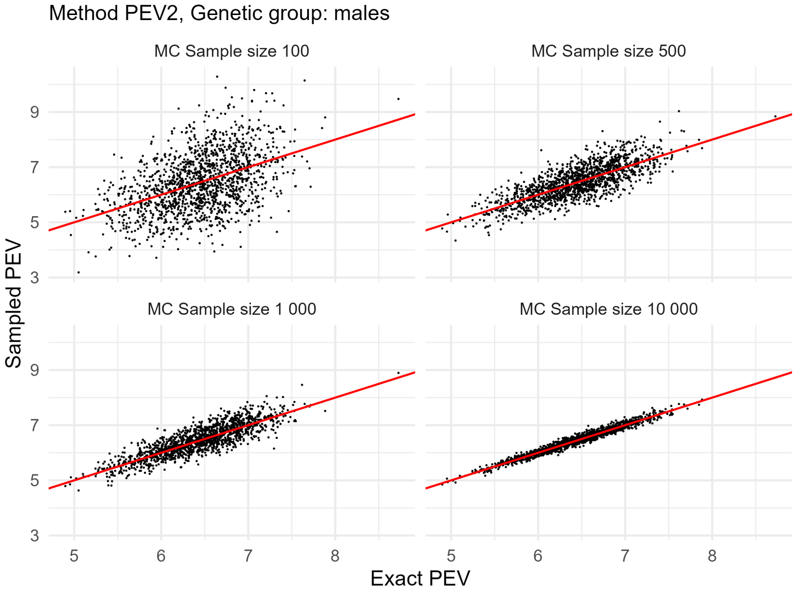

Figure 42: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: original.

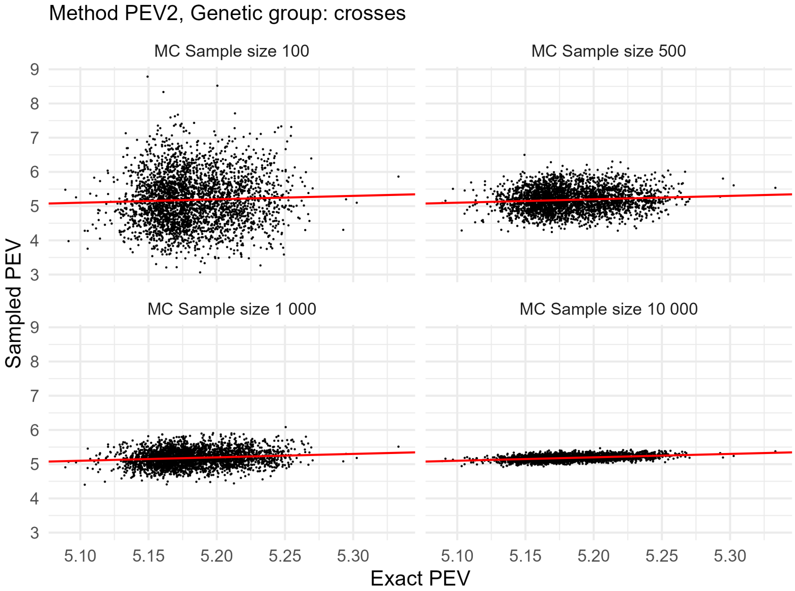

Figure 43: Scatterplots of the exact PEVs and the sampled PEVs using method PEV2 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: original.

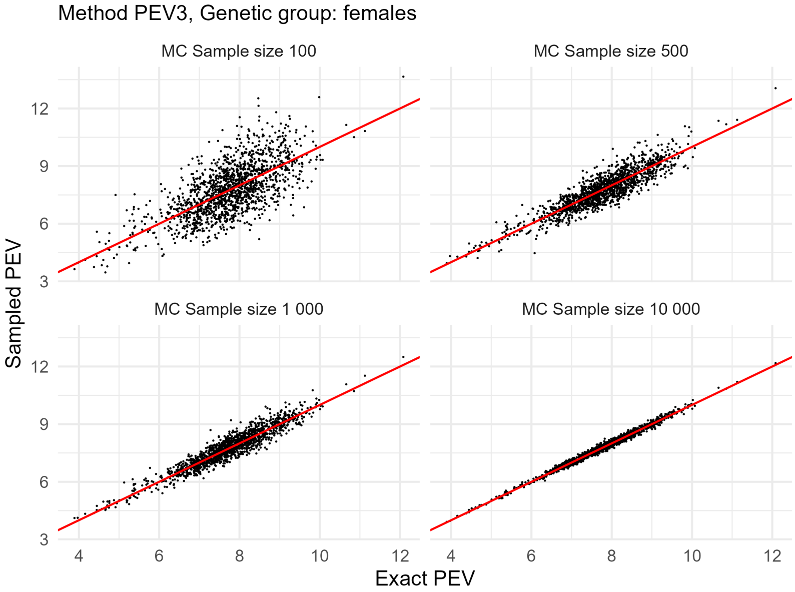

Figure 44: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: original.

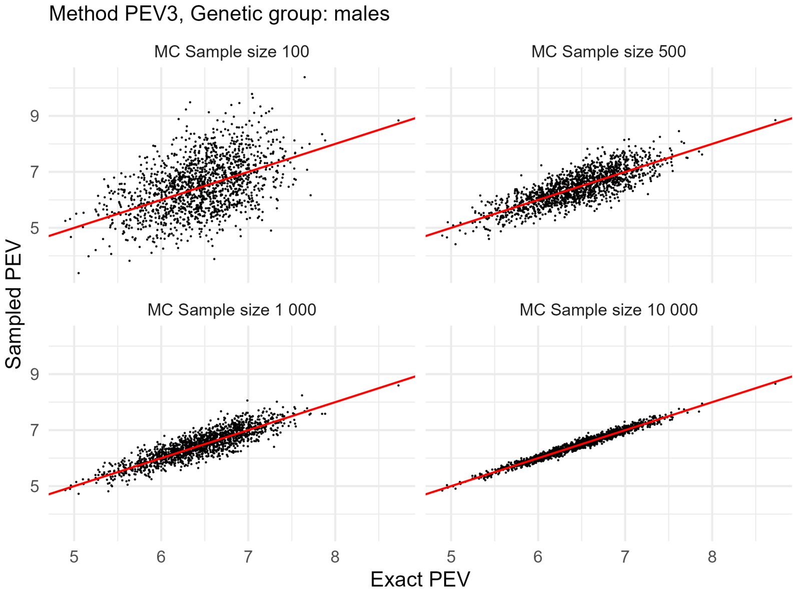

Figure 45: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: original.

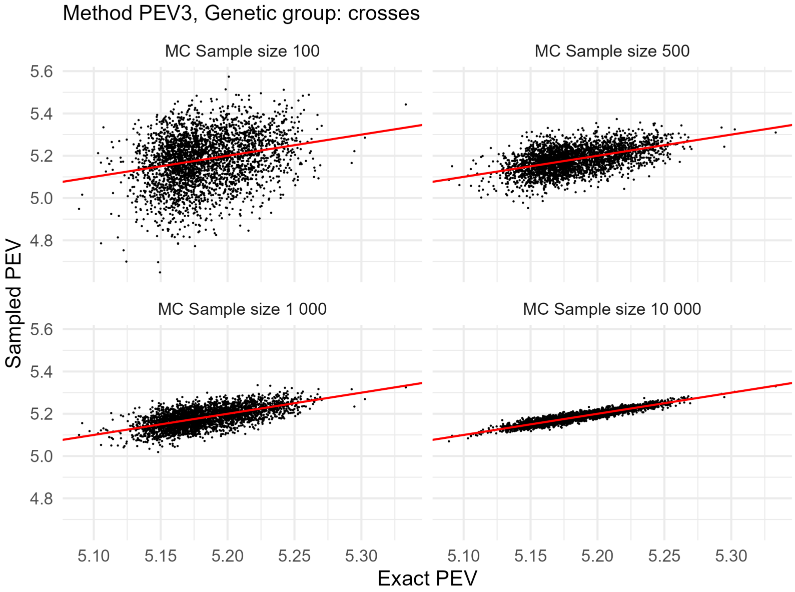

Figure 46: Scatterplots of the exact PEVs and the sampled PEVs using method PEV3 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: original.

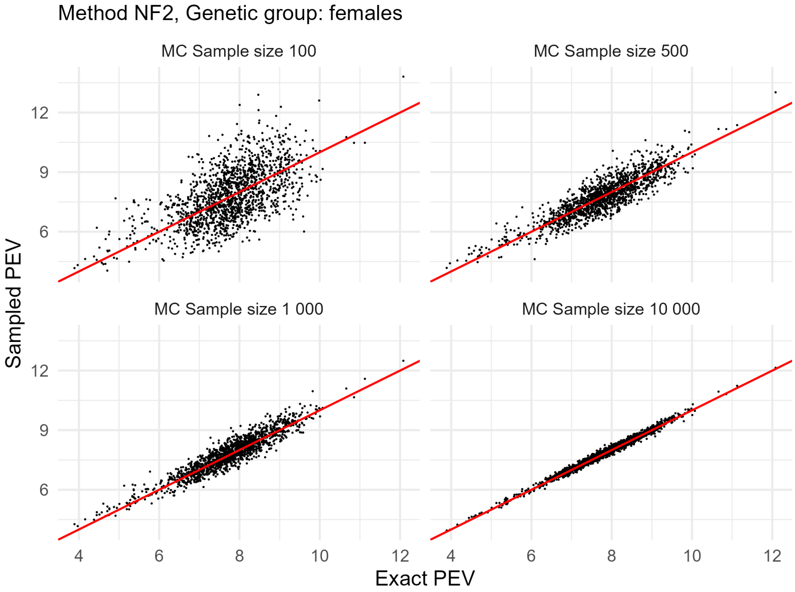

Figure 47: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: females. Design: original.

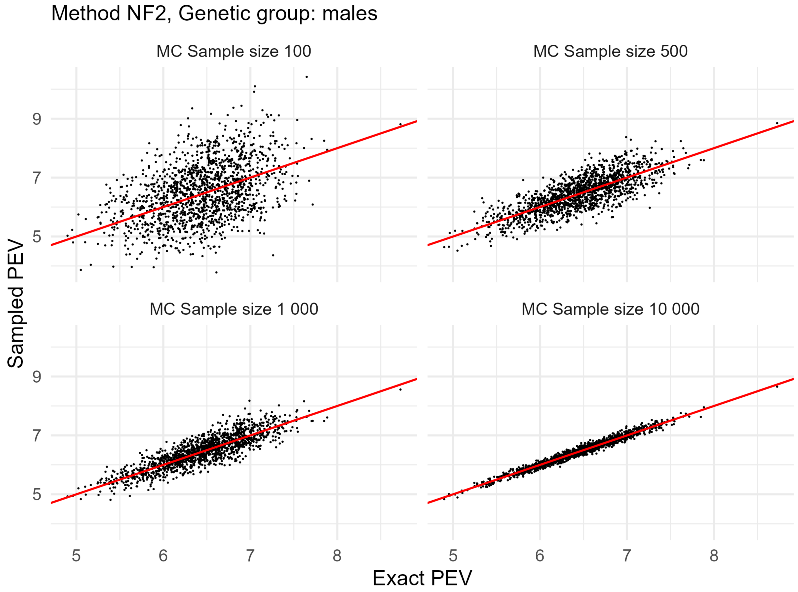

Figure 48: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: males. Design: original.

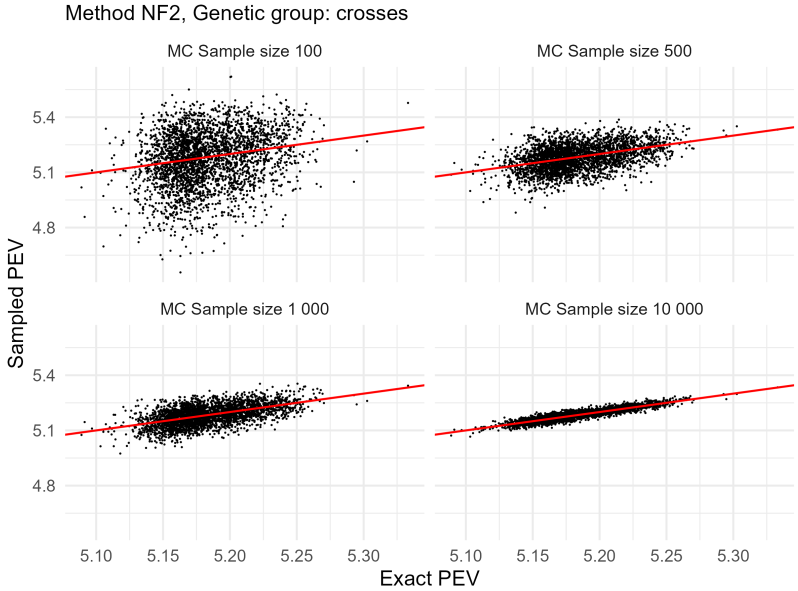

Figure 49: Scatterplots of the exact PEVs and the sampled PEVs using method NF2 with four different numbers of MC samples. The red line represents straight $y = x$. Genetic group: crosses. Design: original.

Appendix E

Part of the R-code used in the analysis and simulation is provided in this appendix.

```r
# Reading the data into R. The data is simulated by Natural
# Resources Institute Finland. The data is not publicly
# available.
MWY3000 <- read.table("MWY3000.data",
  quote = "\"",
  comment.char = ""
)
# Changing the names of the columns.
library(dplyr)
MWY3000 <- MWY3000 %>%
  rename(
    female_id = V1,
    male_id = V2,
    cross_id = V3,
    type_of_cross = V4,
    location = V5,
    loc_spec_cov1 = V7,
    loc_spec_cov2 = V8,
    loc_spec_cov3 = V9,
    moisture = V10,
    weigth = V11,
    grain_yield = V12
  )
# Loading the genotype information into R.
geno_male <- read.table("GenotypesM.txt", header = FALSE)
geno_female <- read.table("GenotypesF.txt", header = FALSE)
# Constructing the genetic relationship matrix by hand as
# an example.
m_female <- as.matrix(geno_female)
rownames(m_female) <- m_female[, 1]
m_female <- m_female[, -1]
# Counting the allele frequencies.
p_female <- apply(m_female, 2, mean) / 2
# Creating P matrix for females.
P_female <- matrix(rep(p_female * 2, nrow(m_female)),
  ncol = ncol(m_female), nrow = nrow(m_female), byrow = TRUE
)
rownames(P_female) <- rownames(m_female)
colnames(P_female) <- colnames(m_female)
# Creating Z matrix for females.
Z_female <- m_female - P_female
# Scaling for the final G matrix.
q_female <- 1 - p_female
sum2pq_female <- 2 * sum(p_female * q_female)
# Now we have G matrix for females
G_female <- (Z_female %*% t(Z_female)) / sum2pq_female
# Now we will try the package ASRgenomics to see if we get the
# same results.
library(ASRgenomics)
G_female_v2 <- qc.filtering(M = m_female, impute = FALSE)
Ghat_female <- G.matrix(
  M = G_female_v2$M.clean,
  method = "VanRaden", na.string = NA
)$G
# Here we used VanRaden's first method of
# constructing G matrix
# Even though we get the same results, these matrices are
# not invertible.

# Now we will construct the genetic relationship matrices.
# We must make sure that the individuals are in the same order
# in all of our matrices.
# We start with sorting the genotypic matrix of females.
unique_ids_female <- unique(MWY3000$female_id)
sorted_indices_female <- match(
  unique_ids_female,
  rownames(m_female)
)
sorted_m_female <- m_female[sorted_indices_female, ]
m_female <- sorted_m_female
# We will do the same for males.
m_male <- as.matrix(geno_male)
rownames(m_male) <- m_male[, 1]
m_male <- m_male[, -1]
unique_ids_male <- unique(MWY3000$male_id)
```

```r
sorted_indices_male <- match(unique_ids_male, rownames(m_male))
sorted_m_male <- m_male[sorted_indices_male, ]
m_male <- sorted_m_male
# Now we can construct G matrices, when the individuals are
# in the correct order.
Ghat_female <- G.matrix(
  M = m_female, method = "VanRaden",
  na.string = NA
)$G
Ghat_male <- G.matrix(
  M = m_male, method = "VanRaden",
  na.string = NA
)$G
# They are still not invertible, so we will blend them a
# little with an identity matrix.
G_female.blend <- G.tuneup(
  G = Ghat_female, blend = TRUE,
  pblend = 0.01
)$Gb
G_male.blend <- G.tuneup(
  G = Ghat_male, blend = TRUE,
  pblend = 0.01
)$Gb
# After that they are invertible.

# Now we can construct the variance-covariance matrices.
# Defining the variances.
var_female <- 15
var_male <- 10
var_sca <- 6
var_e <- 233
library(Matrix)
n <- nrow(MWY3000)
G_1 <- var_female * G_female.blend
G_2 <- var_male * G_male.blend
G_3 <- Matrix(var_sca * diag(length(unique(MWY3000$cross_id))))
# Variance-covariance matrix for residuals.
R <- Matrix(var_e * diag(n), sparse = TRUE)
# Inverting the matrices.
G_1_inv <- solve(G_1)
```

```r
G_2_inv <- solve(G_2)
G_3_inv <- solve(G_3)
inverse_R <- solve(R)
# Then we can construct incidence matrices Z_1, Z_2, Z_3 and X.
# Factorizing the location.
MWY3000$location <- as.factor(MWY3000$location)
# Incidence matrix for the fixed effects.
X <- model.matrix(~ 0 + location, MWY3000)
X <- Matrix(X, sparse = T)
# Incidence matrix for the female effect.
MWY3000$female_id <- as.factor(MWY3000$female_id)
Z_1 <- model.matrix(~ 0 + female_id, data = MWY3000)
# Removing the text from the column names.
colnames(Z_1) <- as.integer(gsub("[^0-9]", "", colnames(Z_1)))
# The order in Z matrices must be the same as in
# the corresponding G matrix.
sorted_indices <- match(rownames(G_1), colnames(Z_1))
Z_1_sorted <- Z_1[, sorted_indices]
Z_1 <- Z_1_sorted
Z_1 <- Matrix(Z_1, sparse = TRUE)
# Incidence matrix for the male effect.
MWY3000$male_id <- as.factor(MWY3000$male_id)
Z_2 <- model.matrix(~ 0 + male_id, data = MWY3000)
colnames(Z_2) <- as.integer(gsub("[^0-9]", "", colnames(Z_2)))
sorted_indices <- match(rownames(G_2), colnames(Z_2))
Z_2_sorted <- Z_2[, sorted_indices]
Z_2 <- Z_2_sorted
Z_2 <- Matrix(Z_2, sparse = TRUE)
# Incidence matrix for the cross effect.
MWY3000$cross_id <- as.factor(MWY3000$cross_id)
Z_3 <- model.matrix(~ 0 + cross_id, data = MWY3000)
Z_3 <- Matrix(Z_3, sparse = TRUE)
# Now we have all the elements for the MME.
# First we will make LHS of the MME.
col1_lhs <- rbind(
  t(X) %*% inverse_R %*% X,
  t(Z_1) %*% inverse_R %*% X, t(Z_2) %*% inverse_R %*% X,
  t(Z_3) %*% inverse_R %*% X
)
col2_lhs <- rbind(
  t(X) %*% inverse_R %*% Z_1,
  t(Z_1) %*% inverse_R %*% Z_1 + G_1_inv,
  t(Z_2) %*% inverse_R %*% Z_1, t(Z_3) %*% inverse_R %*% Z_1
)
col3_lhs <- rbind(
  t(X) %*% inverse_R %*% Z_2,
  t(Z_1) %*% inverse_R %*% Z_2, t(Z_2) %*% inverse_R %*% Z_2
    + G_2_inv, t(Z_3) %*% inverse_R %*% Z_2
)
col4_lhs <- rbind(
  t(X) %*% inverse_R %*% Z_3,
  t(Z_1) %*% inverse_R %*% Z_3, t(Z_2) %*% inverse_R %*% Z_3,
  t(Z_3) %*% inverse_R %*% Z_3 + G_3_inv
)
left_lhs <- cbind(col1_lhs, col2_lhs, col3_lhs, col4_lhs)
# Inverse of LHS.
inv_left_lhs <- solve(left_lhs)
# Response vector y.
y <- MWY3000$grain_yield
# RHS of the MME.
row1_rhs <- t(X) %*% inverse_R %*% y
row2_rhs <- t(Z_1) %*% inverse_R %*% y
row3_rhs <- t(Z_2) %*% inverse_R %*% y
row4_rhs <- t(Z_3) %*% inverse_R %*% y
rhs <- rbind(row1_rhs, row2_rhs, row3_rhs, row4_rhs)
results <- inv_left_lhs %*% rhs

# PEVs and reliabilities.
female_indices <- 32:1541
male_indices <- 1542:3033
cross_indices <- 3034:6033
pev_values <- diag(inv_left_lhs)
pev_female <- pev_values[female_indices]
pev_male <- pev_values[male_indices]
pev_cross <- pev_values[cross_indices]
G_female_diagonals <- diag(G_female.blend)
reliability_female <- 1 - pev_values[female_indices] /
  (var_female * G_female_diagonals)
G_male_diagonals <- diag(G_male.blend)
reliability_male <- 1 - pev_values[male_indices] /
```

```r
  (var_male * G_male_diagonals)
G_cross_diagonals <- diag(G_3)
reliability_cross <- 1 - pev_values[cross_indices] /
  (var_sca * G_cross_diagonals)
# Accuracies.
accuracies_female <- sqrt(reliability_female)
accuracies_male <- sqrt(reliability_male)
accuracies_cross <- sqrt(reliability_cross)

# Here is the function to simulate random effects and generate
# then the response y.
generate_simulated_u_and_y <- function(s) {
  # s is the number iterations.
  # Initializing the matrices, where we will store the
  # simulated random effects and simulated data.
  simulated_y <- matrix(ncol = s, nrow = nrow(MWY3000))
  ordered_samples_female <- matrix(
    ncol = s,
    nrow = nrow(MWY3000)
  )
  ordered_samples_male <- matrix(
    ncol = s,
    nrow = nrow(MWY3000)
  )
  ordered_samples_cross <- matrix(
    ncol = s,
    nrow = nrow(MWY3000)
  )
  unique_crosses <- unique(MWY3000$cross_id)
  # Solving the Cholesky decomposition of the
  # genomic relationship matrices.
  L_female <- chol(G_female.blend)
  L_male <- chol(G_male.blend)
  for (i in 1:s) {
    # Sampling the random effects assuming their distributions
    # and using the genomic relationship matrix and known
    # variance.
    # samples_female <- mvrnorm(n_samples, mu_female, G_1)
    # The above line is just for reference.
    x_a_female <- matrix(rnorm(ncol(Z_1),
      mean = 0,
      sd = sqrt(var_female)
    ), ncol = 1)
    samples_female <- t(L_female) %*% x_a_female
    # samples_male<- mvrnorm(n_samples, mu_male, G_2)
    # The above line is just for reference.
    x_a_male <- matrix(rnorm(ncol(Z_2),
      mean = 0,
      sd = sqrt(var_male)
    ), ncol = 1)
    samples_male <- t(L_male) %*% x_a_male
    # samples_cross<- mvrnorm(n_samples, mu_cross, G_3)
    # The above line is just for reference.
    x_a_cross <- matrix(rnorm(ncol(Z_3),
      mean = 0,
      sd = sqrt(var_sca)
    ), ncol = 1)
    samples_cross <- x_a_cross # t(L_cross) %*% x_a_cross
    # Here we make sure that the simulated data is in right
    # order: the same order as in MWY3000 data set.
    names(samples_cross) <- unique_crosses
    ordered_samples_female[, i] <-
      samples_female[match(
        MWY3000$female_id,
        row.names(samples_female)
      )]
    ordered_samples_male[, i] <-
      samples_male[match(
        MWY3000$male_id,
        row.names(samples_male)
      )]
    ordered_samples_cross[, i] <-
      samples_cross[match(
        MWY3000$cross_id,
        names(samples_cross)
      )]
    # Generating the y: simulated data.
    y_simulated <- ordered_samples_cross[, i] +
      ordered_samples_female[, i] + ordered_samples_male[, i] +
      rnorm(nrow(MWY3000), 0, sqrt(var_e))
    simulated_y[, i] <- y_simulated
  }
  list(
    simulated_y = simulated_y,
    ordered_samples_female = ordered_samples_female,
    ordered_samples_male = ordered_samples_male,
    ordered_samples_cross = ordered_samples_cross
  )
}

# Next we will make the function that will solve the MME for s
# times. It will use the output from the function
# "generate_simulated_u_and_y".
# It uses the simulated_y to solve the MME.
solve_mme_s_times <- function(s, simulated_y_df) {
  results_sim <- matrix(nrow = nrow(results), ncol = s)
  col_lhs_1_sim <- rbind(
    t(X) %*% inverse_R %*% X,
    t(Z_1) %*% inverse_R %*% X, t(Z_2) %*% inverse_R %*% X,
    t(Z_3) %*% inverse_R %*% X
  )
  col_lhs_2_sim <- rbind(
    t(X) %*% inverse_R %*% Z_1,
    t(Z_1) %*% inverse_R %*% Z_1 + G_1_inv, t(Z_2) %*%
      inverse_R %*% Z_1, t(Z_3) %*% inverse_R %*% Z_1
  )
  col_lhs_3_sim <- rbind(
    t(X) %*% inverse_R %*% Z_2,
    t(Z_1) %*% inverse_R %*% Z_2, t(Z_2) %*% inverse_R %*% Z_2
      + G_2_inv, t(Z_3) %*% inverse_R %*% Z_2
  )
  col_lhs_4_sim <- rbind(
    t(X) %*% inverse_R %*% Z_3,
    t(Z_1) %*% inverse_R %*% Z_3, t(Z_2) %*% inverse_R %*% Z_3,
    t(Z_3) %*% inverse_R %*% Z_3 + G_3_inv
  )
  left_lhs_sim <- cbind(
    col_lhs_1_sim, col_lhs_2_sim,
    col_lhs_3_sim, col_lhs_4_sim
  )
  inv_left_lhs_sim <- solve(left_lhs_sim)

  for (i in 1:s) {
    # RHS
    y_sim <- simulated_y_df[, i]
    row1_rhs_sim <- t(X) %*% inverse_R %*% y_sim
    row2_rhs_sim <- t(Z_1) %*% inverse_R %*% y_sim
    row3_rhs_sim <- t(Z_2) %*% inverse_R %*% y_sim
    row4_rhs_sim <- t(Z_3) %*% inverse_R %*% y_sim
    rhs_sim <- rbind(
      row1_rhs_sim, row2_rhs_sim, row3_rhs_sim,
      row4_rhs_sim
    )
    results_sim[, i] <-
      as.vector(inv_left_lhs_sim %*% right_lhs_sim)
  }
  results_sim
}

# Functions can be called as follows:
n_mc <- 100
simulated_data <- generate_simulated_u_and_y(n_mc)
simulated_results <- solve_mme_s_times(
  n_mc,
  simulated_data$simulated_y
)
# Before using the simulated data to approximate PEVs, we must
# only have one breeding value per individual.
row.names(simulated_data$ordered_samples_female) <-
  MWY3000$female_id
row.names(simulated_data$ordered_samples_male) <-
  MWY3000$male_id
row.names(simulated_data$ordered_samples_cross) <-
  MWY3000$cross_id
unique_samples_female <- simulated_data$ordered_samples_female[
  !duplicated(rownames(simulated_data$ordered_samples_female)),
]
unique_samples_male <- simulated_data$ordered_samples_male[
  !duplicated(rownames(simulated_data$ordered_samples_male)),
]
```

```r
unique_samples_cross <- simulated_data$ordered_samples_cross[
  !duplicated(rownames(simulated_data$ordered_samples_cross)),
]

# Method 1 Garcia-Gortes (PEV1):
G_female_diagonals <- diag(G_female.blend)
PEV1_female <- G_female_diagonals * var_female -
  rowMeans(simulated_results[female_indices, ]^2)
G_male_diagonals <- diag(G_male.blend)
PEV1_male <- G_male_diagonals * var_male -
  rowMeans(simulated_results[male_indices, ]^2)
PEV1_cross <- diag(diag(length(unique(MWY3000$cross_id)))) *
  var_sca - rowMeans(simulated_results[cross_indices, ]^2)
# Method 2 Garcia-Gortes (PEV2):
squared_diff_between_sim_true_female <- matrix(
  nrow = nrow(unique_samples_female),
  ncol = ncol(simulated_data$ordered_samples_female)
)
s <- ncol(simulated_results)
for (i in 1:s) {
  squared_diff_between_sim_true_female[, i] <-
    (unique_samples_female[, i] -
      simulated_results[female_indices, i])^2
}
PEV2_female <- rowMeans(squared_diff_between_sim_true_female)
squared_diff_between_sim_true_male <- matrix(
  nrow = nrow(unique_samples_male),
  ncol = ncol(simulated_data$ordered_samples_male)
)
for (i in 1:s) {
  squared_diff_between_sim_true_male[, i] <-
    (unique_samples_male[, i] -
      simulated_results[male_indices, i])^2
}
PEV2_male <- rowMeans(squared_diff_between_sim_true_male)
squared_diff_between_sim_true_cross <- matrix(
  nrow = nrow(unique_samples_cross),
  ncol = ncol(simulated_data$ordered_samples_cross)
)
for (i in 1:s) {
  squared_diff_between_sim_true_cross[, i] <-
    (unique_samples_cross[, i] -
      simulated_results[cross_indices, i])^2
}
PEV2_cross <- rowMeans(squared_diff_between_sim_true_cross)
# Method 3 Garcia-Gortes (PEV3):
var_PEV2_female <- apply(
  squared_diff_between_sim_true_female,
  1, var
)
var_PEV1_female <- apply(
  simulated_results[female_indices, ]^2,
  1, var
)
w1_female <- 1 / var_PEV1_female
w2_female <- 1 / var_PEV2_female
PEV3_female <- (w1_female * PEV1_female + w2_female *
  PEV2_female) /
  (w1_female + w2_female)
var_PEV2_male <- apply(
  squared_diff_between_sim_true_male,
  1, var
)
var_PEV1_male <- apply(
  simulated_results[male_indices, ]^2,
  1, var
)
w1_male <- 1 / var_PEV1_male
w2_male <- 1 / var_PEV2_male
PEV3_male <- (w1_male * PEV1_male + w2_male * PEV2_male) /
  (w1_male + w2_male)
var_PEV2_cross <- apply(
  squared_diff_between_sim_true_cross,
  1, var
)
var_PEV1_cross <- apply(
  simulated_results[cross_indices, ]^2,
  1, var
)
w1_cross <- 1 / var_PEV1_cross
```

```r
w2_cross <- 1 / var_PEV2_cross
PEV3_cross <- (w1_cross * PEV1_cross + w2_cross * PEV2_cross) /
  (w1_cross + w2_cross)
summary(PEV3_cross)
# Method NF2 (NF2):
PEV_NF2_female <- (PEV2_female / (PEV2_female +
  rowMeans(simulated_results[female_indices, ]^2))) *
  G_female_diagonals * var_female
PEV_NF2_male <- (PEV2_male / (PEV2_male
+ rowMeans(simulated_results[male_indices, ]^2))) *
  G_male_diagonals * var_male
PEV_NF2_cross <- (PEV2_cross / (PEV2_cross
+ rowMeans(simulated_results[cross_indices, ]^2))) *
  diag(diag(length(unique(MWY3000$cross_id)))) * var_sca

# Examples of how these estimates can be compared to the exact
# value of PEV:
cor(pev_cross, PEV_NF2_cross)
plot(pev_cross, PEV_NF2_cross,
  xlab = "Exact PEV",
  main = "Method NF2 estimates for cross"
)
model1 <- lm(pev_cross ~ PEV_NF2_cross)
summary(model1)
rmse_PEV_NF2_cross <- sqrt(mean((pev_cross - PEV_NF2_cross)^2))
```

Appendix F

Specifications of the computer used in this thesis:

Processor: Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz 2.90 GHz

Installed RAM: 64,0 GB (63,7 GB usable)

System type: 64-bit operating system, x64-based processor