SIGN AND RANK COVARIANCE MATRICES WITH APPLICATIONS TO MULTIVARIATE ANALYSIS

ESA OLLILA

July 29, 2002

Abstract

This dissertation consists of five original publications. The papers consider the statistical properties (consistency, limiting distribution, limiting efficiencies, robustness, computation, estimation of accuracy) of the affine equivariant sign and rank covariance matrix introduced by Visuri, Koivunen and Oja (2000) and their use in multivariate analysis. In particular, new estimates for principal component analysis and for the multivariate linear regression are proposed and their statistical properties are derived. The concepts of multivariate sign and rank are based on Oja's criterion function.

Key words: affine equivariance; asymptotic efficiency; multivariate sign and rank; multivariate analysis; robustness;

Acknowledgements

I wish to express my sincere gratitude to my supervisor, Professor Hannu Oja, for his expert guidance and constant support. Special thanks go also to Professor Visa Koivunen, Professor Thomas P. Hettmansperger and Professor Christophe Croux for the pleasant and fruitful collaboration on our joint papers. I also wish to thank my thesis reviewers Professor Marc Hallin and Professor Probal Chaudhuri for their constructive comments and careful reading of the thesis.

The research work for this thesis was carried out at the Department of Mathematics and Statistics, University of Jyväskylä and the Department of Statistics, Pennsylvania State University during the years 1999-2002. This research was funded by the Comas graduate school. Additional financial support for my doctoral studies was provided by the Finnish Cultural Foundation and Ellen and Artturi Nyyssönen foundation. These organizations and foundations are greatly acknowledged for making this work possible.

I would like to acknowledge my thankfulness to all my friends, colleagues and people who have been in this work in spirit. Special thanks go to Tom and Ann Hettmansperger for making my visit to State College, Pennsylvania, so memorable.

Finally, I would like to give my warmest thanks to my parents, Elvi and Antti, for all their support and for everything, basically. My brother Mikko and my sister Hanna and their families also deserve my warm thanks.

Espoo,
July 29, 2002

Esa Ollila

List of Original Publications

This thesis consists of an introductory part and five publications listed below. They will be referred to by Paper I, Paper II etc.

I Ollila, E., Oja, H and Croux, C. (2001), "The Affine Equivariant Sign Covariance Matrix: Asymptotic Behavior and Efficiencies". Conditionally accepted in *Journal of Multivariate Analysis.*

II Ollila, E., Hettmansperger, T. P. and Oja, H. (2002), "Estimates of Regression Coefficients Based on the Sign Covariance Matrix". *Journal of the Royal Statistical Society, Series B*, **64**, 447-466.

III Ollila, E., Hettmansperger, T. P. and Oja, H. (2001), "Affine Equivariant Multivariate Sign Methods". Submitted.

IV Ollila, E., Croux, C. and Oja, H. (2002), "Influence Function and Asymptotic Efficiency of the Affine Equivariant Rank Covariance Matrix". Conditionally accepted in *Statistica Sinica.*

V Ollila, E., Koivunen, V. and Oja, H. (2002), " Estimates of Regression Coefficient Based on Lift Rank Covariance Matrix". To appear in *Journal of the American Statistical Association.*

Contents

Chapter 1

Introduction

In the univariate case, signs and ranks are used to obtain procedures which are valid under wide nonparametric models. The estimates and tests based on signs and ranks are then typically also robust, that is, they are not sensitive to outlying observations. For multivariate observations, it is not so obvious anymore how the sign and the rank of an observation should be defined. In this dissertation the affine equivariant multivariate extensions of the concepts of sign and rank proposed by Brown and Hettmansperger (1987, 1989) are considered. The concepts of sign and rank are based on the Oja (1983) median and have been used in the construction of estimates and tests in a series of papers. Oja (1999) reviewed multivariate one-sample and two-sample location cases and Hettmansperger and Oja (1994) and Hettmansperger, Möttönen and Oja (1998) discuss multi-sample designs. Later, Visuri, Koivunen and Oja (2000) introduced the associated concepts of affine equivariant sign and rank covariance matrices and proposed their use in covariance matrix estimation. More recently, Visuri, Ollila, Koivunen, Möttönen and Oja (2001) illustrated the use of the rank covariance matrix in classical multivariate analysis problems.

In this dissertation, the statistical properties (consistency, limiting distribution, limiting efficiencies, robustness, computation, estimation of accuracy) of the affine equivariant sign and rank covariance matrix are considered and their use in classical multivariate analysis is examined. In particular, new estimates of the eigenvectors and eigenvalues of the covariance matrix and the parameters of the multiresponse linear regression model are proposed and their statistical properties are derived. It turns out that the proposed estimates of eigenvectors and eigenvalues and of regression parameters are highly efficient in the normal case and have a superior performance for heavy tailed distributions as compared to the classical estimates based on the sample covariance matrix. The proposed estimates are not robust in the strict sense as they do not have a positive breakdown point or a bounded influence function. However, they are not as sensitive to outliers as the classical estimates.

The outline of this paper is as follows. In Chapter 2 we discuss the estimation of multivariate location and scatter. Different tools to compare estimators, such as the breakdown point, the influence function and asymptotic efficiency, are discussed and some classes of robust estimators of location and scatter, namely the M-estimators and S-estimators, are presented and discussed. In Chapter 3, we review the notion of sign and rank in the univariate setting and then introduce different extensions to multivariate case. Among the extensions considered are the marginal, spatial and affine equivariant multivariate sign and rank. Also, the related sign and rank covariance matrices are introduced. The emphasis is then put on the affine equivariant signs and ranks and covariance matrices based on them. Their performance in estimation of the elements of the covariance matrix, and the eigenvalues and eigenvectors of the covariance matrix, are then compared with the classical estimators and their robust alternatives.

Chapter 2

Estimation of Multivariate Location and Scatter

2.1 Motivation and Notations

Let $\mathrm{PDS}(k)$ denote the set of all positive definite symmetric $k \times k$ matrices and let I_k denote the $k \times k$ identity matrix. Furthermore, let $\boldsymbol{x}$ be a k-dimensional random vector with finite second-order moments, mean vector $\boldsymbol{\mu} = E(\boldsymbol{x})$ and covariance matrix $\Sigma = \mathrm{Cov}(\boldsymbol{x}) \in \mathrm{PDS}(k)$. The *spectral decomposition* of the covariance matrix is given by $\Sigma = P\,\Lambda\,P^T$ where P is the matrix with the eigenvectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k$ of Σ in its columns and Λ is a diagonal matrix with the descending eigenvalues $\lambda_1 \geq \ldots \geq \lambda_k$ as diagonal elements. We may also state the eigenvalue decomposition in the form

$$\Sigma = \lambda\ P\ \Lambda^*\ P^T = \lambda\ \Sigma^*, \tag{2.1}$$

where $\lambda = (\lambda_1 \cdots \lambda_k)^{1/k}$ is the geometrical mean of the eigenvalues and $\Lambda = \lambda\Lambda^*$. The matrix Λ^* is then a diagonal matrix of *standardized eigenvalues*

$$\lambda_j^* = \frac{\lambda_j}{(\lambda_1 \cdots \lambda_k)^{1/k}}. \tag{2.2}$$

We will often refer to elliptically symmetric distributions. A k-dimensional random vector $\boldsymbol{x}$ has cdf $F_{\boldsymbol{\mu},\Omega}$ if its density is of the form

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \Omega) = \{\det(\Omega)\}^{-1/2} g\{(\boldsymbol{x} - \boldsymbol{\mu})^T \Omega^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\} \tag{2.3}$$

for some $\Omega \in \mathrm{PDS}(k)$ and nonnegative function g, which is independent of $\boldsymbol{\mu}$ and Ω. The parameter $\boldsymbol{\mu}$ is the symmetry centre of the distribution of $\boldsymbol{x}$ (or the expected value $E(\boldsymbol{x})$ when it exists) whereas the parameter Ω is proportional to the covariance matrix

Σ of $\boldsymbol{x}$ when it exists. Often the parameter Ω is called the scatter matrix or the pseudo-covariance matrix. If $\Omega = cI_k$ for some $c > 0$ then $\boldsymbol{x}$ is said to be spherically symmetric around $\boldsymbol{\mu}$. For notational convenience, we simply denote $F = F_{\boldsymbol{\mu},\Omega}$ and $F_0 = F_{\mathbf{0},I_k}$. Note that if $\boldsymbol{x} \sim F$ then $\boldsymbol{z} = \Omega^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}) \sim F_0$ and $r = \|\boldsymbol{z}\|$ and $\boldsymbol{u} = \boldsymbol{z}/r$ are independent with $\boldsymbol{u}$ being uniformly distributed on the unit sphere and with r^2 having the density function

$$f_{r^2}(t) = \{\pi^{k/2}/\Gamma(k/2)\}t^{k/2-1}g(t). \tag{2.4}$$

If the covariance matrix of $\boldsymbol{x}$ exist then $\Sigma = k^{-1}E(r^2)\Omega$.

The elliptical distributions are often used for studying the robustness of multivariate statistics. For this purpose the k-variate t-distribution with ν degrees of freedom, $t_{\nu,k}$, is particularly interesting as it yields distributions with varying heaviness of tails. To be more specific, if $\boldsymbol{x} \sim t_{\nu,k}$, then g in equation (2.3) and (2.4) is of the form

$$g(t) = c_{\nu,k}(1 + t/\nu)^{-(k+\nu)/2}, \tag{2.5}$$

where $c_{\nu,k} = (\pi\nu)^{-k/2}\Gamma\{(k+\nu)/2\}/\Gamma(\nu/2)$. The value $\nu = 1$ gives the k-variate Cauchy distribution whereas the multivariate standard normal density

$$g(t) = (2\pi)^{-k/2}\exp(-t/2) \tag{2.6}$$

is obtained as $\nu \to \infty$. For $\nu > 2$ the covariance matrix of the k-variate t-distribution is $\Sigma = \{\nu/(\nu-2)\}\Omega$. In the normal distribution case, $\Sigma = \Omega$. The information about the scale of Σ is confounded in g. However, for a specific distribution, the functional form g can be replaced by $g_0(t) = c^{-k/2}g(t/c)$ with $c = E(r^2)/k$. We call this the *standardized spherical distribution.* This results in a change in the scatter matrix parameter Ω to $c\Omega = \Sigma$. If the second moments of $\boldsymbol{x}$ exists, we use the standardized spherical distribution and consequently $\Omega = \Sigma$.

A starting point for most multivariate analysis procedures such as the Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), multiresponse regression analysis and discriminant analysis, is indeed the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ. In the PCA, for example, the statistics of interest are the eigenvectors $\boldsymbol{p}_i$ and standardized eigenvalues λ_i^* of Σ. These are often used to reduce the dimensionality of the data set by considering the first m ($< k$) principal components only (obtained by projecting the data on to the direction of the first m eigenvectors). The standardized eigenvalues then measure the relative amount of information explained by the principal components. In the CCA, one considers the partitioned mean vector and covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

of a random variable $(\boldsymbol{x}^T, \boldsymbol{y}^T)^T$ and the statistics of interest are the eigenvalues and eigenvectors of matrices

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \qquad \text{and} \qquad \Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

which are used to describe the dependence between the two sets of variables. Specifically, if $\boldsymbol{x}$ gives the explanatory variables and $\boldsymbol{y}$ the response variables in the multivariate multiple regression model, then the regression coefficient matrix is given by

$$\mathcal{B} = \begin{pmatrix} \boldsymbol{\mu}_y^T - \boldsymbol{\mu}_x^T\Sigma_{xx}^{-1}\Sigma_{xy} \\ \Sigma_{xx}^{-1}\Sigma_{xy} \end{pmatrix}.$$

Furthermore, in the linear discriminant analysis, the classification rules are defined using the mean vectors and the common covariance matrix of the populations from which the data come. Note that these applications require the covariance matrix up to a constant scalar factor, and hence a *shape matrix* (a function of the covariance matrix which is invariant under multiplication of the covariance statistic, such as Σ^* in (2.1)) could be used as well. Properties of the shape matrices and the multivariate procedures based on them are discussed in Paper III.

The true location parameter $\boldsymbol{\mu}$ and the covariance matrix Σ are unknown and need to be estimated from the data. In classical multivariate analysis one assumes that the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ come from a multivariate normal distribution (so the density g is as in equation (2.6)). Then the maximum likelihood estimators of $\boldsymbol{\mu}$ and Σ are the sample mean mean vector and the sample covariance matrix

$$\bar{\boldsymbol{x}} = \text{ave}\{\boldsymbol{x}_i\} \qquad \text{and} \qquad S = \text{ave}\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\},$$

respectively, where "ave" stands for arithmetic average over cases $i = 1, \ldots, n$. These estimators are easy to calculate and work with and their limiting distributions are reasonably simple. This is mostly true also for the procedures based on them. It is well known however that the sample mean and covariance matrix and the related procedures are highly sensitive to *outliers* or *gross errors* (atypical observations that deviate from the bulk of the data) or slight deviations from assumptions (e.g. if the true underlying distribution has tails slightly heavier that that of the normal distribution). Therefore, a fundamental problem in multivariate statistics is to develop robust estimators of $\boldsymbol{\mu}$ and Σ as alternatives to sample mean and covariance matrix. The robust estimators of the mean vector and the covariance matrix can then be used to construct robust multivariate analysis procedures such as robust estimators of eigenvectors and eigenvalues of Σ or robust estimators of the regression coefficient matrix $\mathcal{B}$.

2.2 Tools to Compare Estimators

In this section we introduce some tools to compare different location and scatter matrix estimators. (The words scatter matrix and covariance matrix will be abusively used as synonyms throughout this paper.) The tools introduced here are the breakdown point, influence function and asymptotic efficiency.

2.2.1 Estimators and Statistical Functionals

We define here what we mean by estimators and statistical functionals of the multivariate location and scatter. Let $\hat{\boldsymbol{\mu}}$ and $\hat{C}$ denote the estimates of $\boldsymbol{\mu}$ and Σ based on the data set $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in $\mathbb{R}^k$ which is a random sample from unknown F. Assume that $F \in \mathcal{F}$ where $\mathcal{F}$ denotes a set of all distributions on $\mathbb{R}^k$ (or a large subset of it). By "large" we mean one that contains plausible models F for the unknown population as well as the empirical distribution F_n associated with the data set X_n. Then, a map $M : \mathcal{F} \to \mathbb{R}^k$ is *a statistical functional corresponding to* $\hat{\boldsymbol{\mu}}$ if $\hat{\boldsymbol{\mu}} = M(F_n)$. Furthermore, a map $C : \mathcal{F} \to \mathrm{PDS}(k)$ is a *statistical functional corresponding to* $\hat{C}$ whenever $\hat{C} = C(F_n)$. For example, let F be a member in $\mathcal{F}$ containing all distributions on $\mathbb{R}^k$ with finite second-order moments. Then the population mean vector and covariance matrix

$$M(F) = E_F(\boldsymbol{x}) \qquad \text{and} \qquad C(F) = E_F[\{\boldsymbol{x} - E_F(\boldsymbol{x})\}\{\boldsymbol{x} - E_F(\boldsymbol{x})\}^T]$$

are statistical functionals corresponding to the sample mean and sample covariance matrix as $M(F_n) = \bar{\boldsymbol{x}}$ and $C(F_n) = S$. The estimates and the corresponding statistical functionals are supposed to satisfy the following two conditions.

Condition A. *Affine equivariance.* The estimates $\hat{\boldsymbol{\mu}}$ and $\hat{C}$ are *affine equivariant* in the sense that if $(\hat{\boldsymbol{\mu}}, \hat{C})$ are estimates calculated from the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, then $(A\hat{\boldsymbol{\mu}} + \boldsymbol{b}, A\hat{C}A^T)$ are estimates from the transformed data $A\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, A\boldsymbol{x}_n + \boldsymbol{b}$, for any non-singular $k \times k$-matrix A and $\boldsymbol{b} \in \mathbb{R}^k$. The corresponding statistical functionals $(M(F), C(F))$ are affine equivariant in an analogous fashion, that is, if F_y and F_x represents the distributions of the k-dimensional random vectors $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{b}$ and $\boldsymbol{x}$ respectively, then

$$(\, M(F_y), C(F_y)\,) = (\, AM(F_x) + \boldsymbol{b}, AC(F_x)A^T\,).$$

The sample mean and covariance matrix and their statistical functionals are naturally affine equivariant. The well-known estimator of the multivariate location, the co-ordinatewise sample median, for example, is not affine equivariant.

Condition B. *Concistency.* The functionals $M(F)$ and $C(F)$ are supposed to be *Fisher-consistent* to the true parameter values $\boldsymbol{\mu}$ and Σ at F respectively, i.e. $M(F) = \boldsymbol{\mu}$

and $C(F) = \Sigma$. Moreover, the estimates are supposed to be consistent in the regular sense that $\hat{\boldsymbol{\mu}}$ and $\hat{C}$ converge in probability to $M(F)$ and $C(F)$, respectively.

The spectral decomposition of $\hat{C}$ will be denoted by $\hat{C} = \hat{P}\hat{\Lambda}\hat{P}^T$, so $\hat{P} = [\hat{\boldsymbol{p}}_1 \cdots \hat{\boldsymbol{p}}_k]$ is an orthogonal matrix containing the eigenvectors as columns and $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ contains the descending eigenvalues as diagonal elements. The statistical functionals of $\hat{P}$ and $\hat{\Lambda}$ are naturally obtained from the spectral decomposition of the scatter functional: $C(F) = P_C(F)\Lambda_C(F)P_C(F)^T$, where again $P_C(F)$ is an orthogonal matrix with eigenvector functionals $\boldsymbol{p}_{C,j}(F)$, $j = 1, \ldots, k$ as columns and $\Lambda_C(F)$ is a diagonal matrix with eigenvalue functionals $\lambda_{C,j}(F)$, $j = 1, \ldots, k$ as diagonal elements. Condition B immediately implies that the functionals $P_C(F)$ and $\Lambda_C(F)$ are Fisher-consistent for P and Λ at F, so $P_C(F) = P$ and $\Lambda_C(F) = \Lambda$, and $\hat{P}$ and $\hat{\Lambda}$ converge to P and Λ in probability. Naturally, at the empirical distribution, $P_C(F_n) = \hat{P}$ and $\Lambda_C(F_n) = \hat{\Lambda}$.

We will use $T(F)$ to denote an statistical functional corresponding to an arbitrary estimator say $\hat{\theta} = T(F_n)$ obtaining values on some set Θ.

2.2.2 The Breakdown Point

The breakdown point (BP) of an estimator measures *quantitative robustness* of an estimator. Loosely speaking it is the percentage of contamination that the estimator (or statistic) can tolerate before it "breaks down" and becomes totally unreliable. We now define the notion of breakdown point for estimates (finite sample case) and for functionals (asymptotic, or, the population case).

The Finite Sample Breakdown Point

We consider two slightly different definitions of breakdown point for finite samples, namely the *replacement breakdown point* (RBP) and *addition breakdown point* (ABP), introduced by Donoho (1982) and Donoho and Huber (1983). The breakdown point of an estimate or statistic $\hat{\theta} \in \Theta$ can be defined in several ways but these two ways are the most popular in the literature. Let us first consider the RBP, which is, the smallest fraction $\varepsilon = m/n$ of observations in the data set X_n that need to be replaced to take the estimate over all bounds. To be more specific, let ε-replacement neighborhood of the data set X_n be defined as

$$B_\varepsilon(X_n) = \{\, Y \subset \mathbb{R}^k \; : \#\{Y\} = n \text{ and } \#\{X_n \cap Y\} = n - m \,\}.$$

That is, the corrupted sample Y is obtained by replacing the fraction $\varepsilon = m/n \in \{1/n, \ldots, (n-1)/n, 1\}$ of observations of X_n with arbitrary values. Let δ be a measure of dissimilarity between two elements of Θ. Then the *maximum bias* of the estimate $\hat{\theta}$ over ε-replacement neighborhood is defined by

$$\operatorname{bias}(\varepsilon, X_n; \hat{\theta}) = \sup\{\, \delta(\, \hat{\theta}(X_n), \hat{\theta}(Y)\,) \,:\, Y \in B_\varepsilon(X_n)\,\}.$$

The RBP of $\hat{\theta}$ is then given by

$$\varepsilon_n^* = \inf\{\, \varepsilon \,:\, \operatorname{bias}(\varepsilon, X_n; \hat{\theta}) = \infty \,\}.$$

Note that the RBP is dependent on the number of observations n and also on the particular data at hand. Usually some conditions are imposed on the data set X_n such as that $n \geq k+1$ and the data is in *general position*, that is, no $k+1$ points are contained in any hyperplane of dimension smaller than k; see Lopuhaä and Rousseeuw (1991) and Davies (1987).

In the case of location estimates $\hat{\boldsymbol{\mu}}$, $\Theta = \mathbb{R}^k$, and one could define $\delta(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$, where $\|\cdot\|$ is the Euclidean norm. In the case of scatter matrix estimates $\hat{C}$, $\Theta = \mathrm{PDS}(k)$, and one may define

$$\delta(A, B) = \max\{\, |\lambda_1(A) - \lambda_1(B)|, |\lambda_k^{-1}(A) - \lambda_k^{-1}(B)|\,\},$$

with $\lambda_1(A) \geq \ldots \geq \lambda_k(A)$ being the ordered eigenvalues of the matrix $A \in \mathrm{PDS}(k)$. Hence, for this choice of δ, the RBP of the scatter estimate is the smallest replacement fraction of outliers that can either take the largest eigenvalue over all bounds or the smallest eigenvalue arbitrarily close to zero. These two particular functions δ are used by Lopuhaä and Rousseeuw (1991). There are naturally several other possibilities in defining δ function for the location and scatter estimates. In the case of shape matrices, δ is often defined so that it involves ratios of eigenvalues. See Maronna and Yohai (1995).

Instead of considering the breakdown of location and scatter estimates separately, one could construct definitions of *simultaneous* RBP of location and scatter by combining the definitions appropriately as is done in Davies (1987) and Tyler (1994) and Kent and Tyler (1996). In other words, the simultaneous RBP is then just the smallest fraction $\varepsilon = m/n$ of replacement outliers which causes the breakdown of at least one of the estimates. The finite sample simultaneous RBP of any affine equivariant location and scatter estimate $\hat{\boldsymbol{\mu}}$ and $\hat{C}$ will always be at most $\lfloor \frac{n-k+1}{2} \rfloor / n$ if X_n is in general position and $n \geq k+1$ (Davies, 1987). Naturally affine equivariant location estimators can be defined without any reference to a companion scatter estimate (and vice versa), and the RBP may then exceed the above upper bound. For a detailed discussion, see Lopuhaä and Rousseeuw (1991).

Let us now consider the finite sample addition breakdown point, which is, the smallest fraction $\varepsilon = m/(n+m)$ of arbitrary observations that need to be added to the data set X_n

to take the estimate over all bounds. In this case we consider ε-addition neighborhoods

$$B^*_\varepsilon(X_n) = \{\, Y = X_n \cup Z : \, Z \subset \mathbb{R}^k \text{ and } \#\{Z\} = m \,\}.$$

Thus the corrupted sample Y has size $n+m$ and contains a fraction $\varepsilon = m/(n+m)$ of bad observations with $m \in \mathbb{N}^+$. The definition of ABP is then as that of RBP but with $B_\varepsilon(X_n)$ replaced by $B^*_\varepsilon(X_n)$. Thus the two definitions, ABP and RBP, result in slightly different values of finite sample breakdown point. For example, the RBP and ABP of the sample mean vector is $1/n$ and $1/(n+1)$ respectively. Huber (1984) and Donoho and Gasko (1992) used the ABP concept whereas Davies (1987), Lopuhaä and Rousseeuw (1991), Tyler (1994), Kent and Tyler (1996) preferred the RBP. The relationship between the two versions of the finite sample breakdown point is established in a recent paper by Zuo (2001), thereby allowing one to obtain ABP directly from the RBP and vice versa.

The Asymptotic Breakdown Point

The asymptotic breakdown point considers the breakdown of statistical functionals. There are several ways to define the concept of the asymptotic breakdown point of the functional $T(F)$ corresponding to estimate $\hat{\theta} = T(F_n)$. Here we consider the ε-contamination breakdown point, which is analogous to the finite sample BP introduced above. Let the ε-contamination neighborhood of the distribution F be defined as

$$B_\varepsilon(F) = \{\, F_\varepsilon \, : \, F_\varepsilon = (1-\varepsilon)F + \varepsilon G \text{ with } \varepsilon \in [0,1) \text{ and } \; G \in \mathcal{F} \,\}. \tag{2.7}$$

The *maximum asymptotic bias* of the functional $T(F)$ over the ε-contamination neighborhood is then

$$\text{bias}(\varepsilon, F; T) = \sup\{\, \delta(\, T(F), T(F_\varepsilon)\,) \, : \, F_\varepsilon \in B_\varepsilon(F) \,\}.$$

Then the *asymptotic contamination BP* of T is

$$\varepsilon^* = \inf\{\, \varepsilon \, : \, \text{bias}(\varepsilon, F; T) = \infty \,\}.$$

Often $\mathcal{F} = \{\Delta_{\boldsymbol{x}} : \boldsymbol{x} \in \mathbb{R}^k\}$ where $\Delta_{\boldsymbol{x}}$ puts all the probability mass at $\boldsymbol{x}$. Then ε^* is often called the *asymptotic point-mass contamination BP*. In the case of location and scatter matrix functionals $M(F)$ and $C(F)$ one can define δ as earlier in the finite sample case. Sometimes, asymptotic breakdown point is defined using the finite sample breakdown point as $\varepsilon^* = \lim_{n\to\infty} \varepsilon^*_n$ (if the limit exist).

2.2.3 The Influence Function

The influence function (IF) measures *qualitative robustness* of an estimator. The influence function of a functional T is given by

$$\mathrm{IF}(\boldsymbol{x}; T, F) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_\varepsilon)\big|_{\varepsilon=0}, \tag{2.8}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon \Delta_{\boldsymbol{x}}$ and $\Delta_{\boldsymbol{x}}$ is the cdf of a distribution putting all its mass at $\boldsymbol{x}$. The above function, considered as a function of $\boldsymbol{x}$, has been introduced by Hampel (1968, 1974) under the name *influence curve* or *influence function* and it is essentially the first derivative of the statistical functional T. One may interpret the IF as describing the effect of an *infinitesimal* contamination at point $\boldsymbol{x}$ on the estimate, standardized by the mass of the contamination.

A robust estimator should have a bounded and continuous IF. Loosely speaking, the boundedness implies that a small amount of contamination at any point does not have an arbitrarily large influence on the estimate whereas the continuity implies that the small changes in the data set cause only small changes in the estimate. The IF is an asymptotic concept and therefore one should be careful in relating the form of the IF to the finite sample behavior of the estimate. Still the heuristic interpretation of describing the effect of additional observation $\boldsymbol{x}$ on the estimate is appealing. In this spirit, the finite sample version of the influence function at empirical F_n is obtained by suppressing the limit in (2.8) and choosing $\varepsilon = 1/(n+1)$. This yields the *empirical influence function* (EIF), or the *sensitivity curve*

$$\mathrm{EIF}(\boldsymbol{x}; \hat{\theta}, X_n) = (n+1)[\,\hat{\theta}(X_n \cup \{\boldsymbol{x}\}) - \hat{\theta}(X_n)\,].$$

The EIF thus calculates the standardized effect of an additional observation at $\boldsymbol{x}$ on the estimate.

If the functional T is sufficiently regular and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is a random sample from F, one has that (cfr. Huber, 1981; Hampel, 1986)

$$\sqrt{n}\{\hat{\theta} - T(F)\} = \sqrt{n}\,\mathrm{ave}\{\mathrm{IF}(\boldsymbol{x}_i; T, F)\} + o_p(1). \tag{2.9}$$

It turns out that $E[\mathrm{IF}(\boldsymbol{x}; T, F)] = \mathbf{0}$ and, by the central limit theorem, $\sqrt{n}\{\hat{\theta} - T(F)\}$ is asymptotically normal with mean vector zero and asymptotic variance-covariance matrix

$$\mathrm{ASV}(\hat{\theta}; F) = E_F[\mathrm{IF}(\boldsymbol{x}; T, F)\mathrm{IF}(\boldsymbol{x}; T, F)^T]. \tag{2.10}$$

Although equation (2.9) is often true, a rigorous proof may be difficult. However, given the form of the influence function, the equation (2.10) is often used to calculate an expression for the asymptotic variance-covariance matrix of the estimate in a heuristic way (as an educated guess). The influence function is also used to derive other measures of robustness. One such measure of particular importance is the *gross error sensitivity* defined as $\gamma^* = \sup_{\boldsymbol{x}\in\mathbb{R}^k} |\mathrm{IF}(\boldsymbol{x}; T, F)|$ for some norm $|\cdot|$. Loosely speaking, it describes the maximal influence of an individual observation.

The influence functions of the location and scatter matrix functionals $M(F)$ and $C(F)$ at elliptical $F = F_{\boldsymbol{\mu},\Sigma}$ have simple expressions. Their influence functions are of the forms

$$\mathrm{IF}(\boldsymbol{x}; M, F) = \gamma_M(r; F_0)\Sigma^{1/2}\boldsymbol{u}, \tag{2.11}$$

$$\mathrm{IF}(\boldsymbol{x}; C, F) = \alpha_C(r; F_0)\Sigma^{1/2}\boldsymbol{u}\boldsymbol{u}^T\Sigma^{1/2} - \beta_C(r; F_0)\Sigma, \tag{2.12}$$

respectively, where $r^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis-distance of $\boldsymbol{x}$, $\boldsymbol{u} = \Sigma^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu})/r$ is the Mahalanobis-angle of $\boldsymbol{x}$ and γ_M, α_C and β_C are real valued functions determined by the functionals M and C and the underlying spherical distribution $F_0 = F_{\mathbf{0},I_k}$. See Hampel *et al.* (1986) and Croux and Haesbroeck (2000). We may interpret γ_M, α_C and β_C as weight functions and for robust estimates the weight functions should be continuous and bounded. In conclusion, for influence function comparisons between different estimators of location and scatter, we only need to compare the functions γ_M, α_C and β_C. When the IF of the scatter matrix functional C is known, the IF of the eigenvalue and eigenvector functionals $\lambda_{C,j}$ and $\boldsymbol{p}_{C,j}$ of C at elliptical F are easily derived from

$$\mathrm{IF}(\boldsymbol{x}; \lambda_{C,j}, F) = \lambda_j\{\alpha_C(r; F_0)u_j^2 - \beta_C(r; F_0)\}, \tag{2.13}$$

$$\mathrm{IF}(\boldsymbol{x}; \boldsymbol{p}_{C,j}, F) = \alpha_C(r; F_0)\sum_{i\neq j}^{k} \frac{(\lambda_i\lambda_j)^{1/2}u_iu_j}{\lambda_i - \lambda_j}\boldsymbol{p}_i, \tag{2.14}$$

for distinct eigenvalues. See Croux and Haesbroeck (2000).

2.2.4 Asymptotic Efficiency

To calculate the asymptotic relative efficiency (ARE) of the location and scatter estimates $\hat{\boldsymbol{\mu}}$ and $\hat{C}$, we need the following additional restriction. Here we use "vec" as an operator working on matrices: $\mathrm{vec}(A)$ represents k^2-dimensional vector formed by stacking the columns of $k \times k$-matrix A on top of each other.

Condition C: *Asymptotic normality.* Assume that X_n is a random sample from F and that $n^{1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) \to \boldsymbol{n}$ and $n^{1/2}(\hat{C}-\Sigma) \to N$ in distribution, where $\boldsymbol{n}$ is a multivariate normal vector with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathrm{ASV}(\hat{\boldsymbol{\mu}}; F)$ and N is a multivariate normal matrix such that $\mathrm{vec}(N)$ has mean vector $\mathbf{0}$ and variance-covariance matrix $\mathrm{ASV}(\hat{C}; F)$.

Asymptotic efficiency of the location and scatter estimates are easily calculated in the elliptical models (Tyler, 1982). For that purpose, we need some notations. A commutation matrix $I_{k,k}$, is a $k^2 \times k^2$ block matrix with (i,j)-block being equal to a $k \times k$ matrix that has 1 at entry (j,i) and zero elsewhere. The Kronecker product of $k \times k$ matrices A and B, denoted by $A \otimes B$, is a $k^2 \times k^2$-block matrix with $k \times k$-blocks, the (i,j)-block equal to $a_{ij}B$. For relations of Kronecker products, commutation matrices and vec-operator, see Magnus and Neudecker (1988).

Let X_n be a random sample from an elliptical distribution $F = F_{\boldsymbol{\mu},\Sigma}$. The asymptotic variance-covariance matrices of $\hat{\boldsymbol{\mu}}$ and $\mathrm{vec}(\hat{C})$ are then given by

$$\mathrm{ASV}(\hat{\boldsymbol{\mu}}; F) = \tau_1 \Sigma \tag{2.15}$$

and

$$\mathrm{ASV}(\hat{C}; F) = \tau_2 (I_{k^2} + I_{k,k})(\Sigma \otimes \Sigma) + \tau_3 \mathrm{vec}(\Sigma)\mathrm{vec}(\Sigma)^T \tag{2.16}$$

respectively. Here $\tau_1 = \mathrm{ASV}(\hat{\mu}_1; F_0)$ is the asymptotic variance of an element of $\hat{\boldsymbol{\mu}}$, $\tau_2 = \mathrm{ASV}(\hat{C}_{12}; F_0)$ is the asymptotic variance of an off-diagonal element of $\hat{C}$ and $\tau_3 = \mathrm{ASC}(\hat{C}_{11}, \hat{C}_{22}; F_0)$ is the asymptotic covariance between two distinct on-diagonal elements of $\hat{C}$ for the corresponding spherical distribution $F_0 = F_{\mathbf{0},I_k}$. Tyler (1982) derived the expression (2.16) and the expression (2.15) follows at once using similar invariance in distribution arguments.

The practical implication of equations (2.15) and (2.16) then is that the limiting distribution of the location estimate is completely described by a single scalar τ_1 and the limiting distribution of the scatter estimate is completely described by the two scalars τ_2 and τ_3. The ARE's of the affine equivariant location estimates can then be calculated by simply comparing the scalars τ_1 given in (2.15). In case of scatter matrix estimates, the asymptotic variances of the on-diagonal and off-diagonal elements of the scatter matrices need to be compared separately. For results on shape matrix estimates, see Tyler (1983) and Paper III.

Consider again the PCA problem and calculate the asymptotic variance $\mathrm{ASV}(\hat{\lambda}_j; F)$ of the eigenvalue estimate $\hat{\lambda}_j$ and the asymptotic variance-covariance matrix $\mathrm{ASV}(\hat{\boldsymbol{p}}_j; F)$ of the eigenvector estimate $\hat{\boldsymbol{p}}_j$ based on $\hat{C}$ using the expression (2.13) and (2.14) of their IF respectively and equation (2.10). Then, for elliptical $F = F_{\boldsymbol{\mu},\Sigma}$ with distinct eigenvalues

$$\mathrm{ASV}(\hat{\lambda}_j; F) = \tau_4 \lambda_j^2, \tag{2.17}$$

$$\mathrm{ASV}(\hat{\boldsymbol{p}}_j; F) = \tau_2 \sum_{i \neq j} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)} \boldsymbol{p}_i \boldsymbol{p}_i^T, \tag{2.18}$$

(Croux and Haesbroeck, 2000). Here $\tau_4 = \mathrm{ASV}(\hat{C}_{11}; F_0)$ is the asymptotic variance of the on-diagonal element of $\hat{C}$ at spherical distribution $F_0 = F_{\mathbf{0},I_k}$. Note that τ_4 can be obtained from τ_2 and τ_3 from the identity

$$\tau_4 = \tau_3 + 2\tau_2.$$

See Paper I and IV. For the asymptotic distribution of the estimates of standardized eigenvalues based on $\hat{C}$, see Paper I. Equations (2.17) and (2.18) imply that the asymptotic efficiencies of the eigenvalue and eigenvector estimators can be obtained by comparing the corresponding values of τ_4 and τ_2 respectively.

2.3 Classes of Estimators

Several affine equivariant estimators of location and scatter have been introduced in the literature. Besides the sample mean and covariance matrix, examples include the M-estimators (Maronna, 1976; Huber, 1981), the Stahel-Donoho (Stahel, 1981; Donoho, 1982) estimator, the Minimum Volume Ellipsoid (MVE)-estimator and the Minimum Covariance Determinant (MCD)-estimator (Rousseeuw, 1985), the S-estimators (Rousseeuw and Leroy, 1987; Davies, 1987), the τ-estimators (Lopuhaä, 1991), *CM*-estimators (Kent and Tyler, 1996) and *MM*-estimators (Tatsuoka and Tyler, 2001) among others. We now briefly review two well-known location and covariance matrix estimator classes, namely the M-estimators and S-estimators. Their properties have been extensively studied in the literature. Later we compare the properties of M- and S-estimators to those based on multivariate affine equivariant signs and ranks.

2.3.1 M-estimators

M-estimators for multivariate location and covariance were first introduced by Maronna (1976). Later, Huber (1981, Chapter 8) extended Maronna's definition. In his definition, simultaneous M-estimators of location and scatter $\hat{\boldsymbol{\mu}} \in \mathbb{R}^k$ and $\hat{C} \in \mathrm{PDS}(k)$, solve

$$\mathrm{ave}\{\Psi(\boldsymbol{x}_i; \boldsymbol{\mu}, C)\} = 0, \tag{2.19}$$

where the average is over $i = 1, \ldots, n$ and $\Psi = (\Psi_1, \Psi_2)$ is the function

$$\begin{aligned}
\Psi_1(\boldsymbol{x}_i; \boldsymbol{\mu}, C) &= v_1(d_i)(\boldsymbol{x}_i - \boldsymbol{\mu}), \\
\Psi_2(\boldsymbol{x}_i; \boldsymbol{\mu}, C) &= v_2(d_i)(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T - v_3(d_i)C,
\end{aligned}$$

with

$$d_i = d(\boldsymbol{x}_i; \boldsymbol{\mu}, C) = \{(\boldsymbol{x}_i - \boldsymbol{\mu})^T C^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})\}^{1/2}, \tag{2.20}$$

and v_1, v_2 and v_3 are real-valued functions on $[0, \infty)$. The functionals $M(F)$ and $C(F)$ corresponding to estimates $\hat{\boldsymbol{\mu}}$ and $\hat{C}$, respectively, are then defined as the solution of $E_F[\Psi(\boldsymbol{x}; M, C)] = 0$. Note that if F_n is the empirical distribution then $\hat{\boldsymbol{\mu}} = M(F_n)$ and $\hat{C} = C(F_n)$. Under the specified underlying elliptical distribution $F = F_{\boldsymbol{\mu},\Sigma}$, i.e. under the specified density g, the weight functions v_2 and v_3 should be scaled so that they satisfy

$$E[r^2 v_2(r)] = k E[v_3(r)],$$

where r^2 has the density (2.4). This choice of weight functions yields the required Fisher-consistency of $C(F)$ to Σ at F, that is, $C(F) = \Sigma$.

Existence and uniqueness, robustness properties, limiting normality and asymptotic variance and covariances have been derived by Maronna (1976) and Huber (1981) under general assumptions. For results on the existence and uniqueness for an important subclass of M-estimates, see also Kent and Tyler (1991). In Maronna (1976) it is argued that the asymptotic point-mass contamination BP ε^* of M-functionals of location and scatter is at most $1/(k+1)$, when the solution is unique. If the solution is not unique there will always exist a solution with an asymptotic breakdown point of at most $1/k$ (Huber, 1981). The influence functions of the location and scatter M-functionals $M(F)$ and $C(F)$ at elliptical F are derived in Huber (1981; Section 8.7) which by rewriting gives the influence function expressions (2.11) and (2.12) with the weight functions

$$\gamma_M(r; F_0) = \frac{v_1(r) r}{\eta_1}, \quad \alpha_C(r; F_0) = \frac{v_2(r) r^2}{\eta_2}, \quad \beta_C(r; F_0) = \frac{\alpha_C(r; F_0)}{k} - \kappa_C(r; F_0), \tag{2.21}$$

where

$$\kappa_C(r; F_0) = \frac{2\{v_2(r) r^2 - k v_3(r)\}}{\eta_3}.$$

Here

$$\begin{aligned}
\eta_1 &= k^{-1} E[k v_1(r) + r v_1'(r)], \\
\eta_2 &= \{k(k+2)\}^{-1} E[(k+2) v_2(r) r^2 + v_2'(r) r^3], \\
\eta_3 &= E[v_2'(r) r^3 + 2 v_2(r) r^2 - k v_3'(r) r],
\end{aligned}$$

where r^2 has the density (2.4). Thus the influence function of the location functional $M(F)$ is continuous and bounded if $v_1(r)r$ is continuous and bounded. Furthermore, the influence function of the scatter functional $C(F)$ is continuous and bounded if both $v_2(r)r^2$ and $v_3(r)$ are continuous and bounded. The asymptotic variance-covariance matrices of the M-estimates of location and scatter can be derived from the influence functions using (2.10) (Huber 1981, Section 8.8). The scalars τ_1, τ_2 and τ_3 defining the asymptotic variance-covariance matrices (2.15) and (2.16) of the estimates $\hat{\boldsymbol{\mu}}$ and $\mathrm{vec}(\hat{C})$ at elliptical F are then given by

$$\tau_1 = \frac{E[v_1^2(r)r^2]}{k\eta_1^2}, \quad \tau_2 = \frac{E[v_2^2(r)r^4]}{\eta_2^2 k(k+2)} \quad \text{and} \quad \tau_3 = E[\kappa_C^2(r;F_0)] - \frac{2}{k}\tau_2, \tag{2.22}$$

where r^2 has the density (2.4). In conclusion: given the weight functions v_1, v_2 and v_3, the influence functions and asymptotic variance-covariance matrices of the M-estimators at the model distribution are easily calculated by (2.21) and (2.22) respectively.

An Example: *Huber's M-estimator.* Huber's M-estimator, denoted by $M(q,k)$ or $M(q)$, has weight functions

$$v_1(d) = \begin{cases} 1 & d \le c \\ c/d & d > c \end{cases}, \quad \sigma^2 v_2(d) = \begin{cases} 1 & d^2 \le c^2 \\ c^2/d^2 & d^2 > c^2 \end{cases}, \qquad v_3(d) = 1,$$

where c is a tuning constant defined so that $q = Pr(\chi_k^2 \le c^2)$ and the scaling factor σ^2 is defined so that $E[tv_2(t)] = k$, where t is a random variable from χ_k^2-distribution. Then $C(F) = \Sigma$ whenever the elliptical population is multivariate normal. Note also that Huber's weights are decreasing to zero which means that the outlying observations receive smaller weights. According to Tyler (1986) and Lopuhaä (1989) the asymptotic breakdown point of $M(q,k)$ is $\varepsilon^* = \min\{1/c^2, 1-k/c^2\}$ for $c^2 > k$. As an example, the choice $q = 0.9$ yields the breakdown points $\varepsilon = 0.217, 0.160, 0.129, 0.108$ for dimensions $k = 2, 3, 4, 5$.

2.3.2 S-estimators

The multivariate S-estimators were introduced by Rousseeuw and Leroy (1987) and Davies (1987). Let $\rho(d)$ be a bounded, nondecreasing, non-negative function on $\mathbb{R}^+$ obtaining its maximum at c_0 and satisfying $\rho(d) = \rho(c_0)$ for $d \ge c_0$. The S-estimators of multivariate location and scatter is the pair $\hat{\boldsymbol{\mu}} \in \mathbb{R}^k$ and $\hat{C} \in \mathrm{PDS}(k)$ which minimizes $\det(C)$ subject to the constraint

$$\mathrm{ave}\{\rho(d_i)\} = b_0 \tag{2.23}$$

where $d_i = d(\boldsymbol{x}_i; \boldsymbol{\mu}, C)$ is given in equation (2.20) and $0 < b_0 < \rho(c_0)$. The S-functionals of multivariate location and scatter is the pair $T(F) \in \mathbb{R}^k$ and $C(F) \in \text{PDS}(k)$ which minimizes $\det(C)$ subject to the constraint $E_F[\rho(d)] = b_0$, where $d = d(\boldsymbol{x}; \boldsymbol{\mu}, C)$ and $\boldsymbol{x}$ is a k-dimensional random vector having distribution F.

For elliptical distributions, the existence and uniqueness, robustness properties, limiting normality and asymptotic variance and covariances of the estimators have been solved by Davies (1987) and Lopuhaä (1989) under general assumptions. The constant b_0 is chosen in agreement with an assumed underlying distribution as $E_{F_0}[\rho(\|\boldsymbol{x}\|)]$ to guarantee consistency of the scatter matrix estimate to Σ. In that case the constant c_0 can be chosen such that

$$b_0 = \varepsilon \rho(c_0) \tag{2.24}$$

to attain $100\varepsilon\%$ asymptotic breakdown point. The highest finite sample RBP of S-estimators of location and scatter is reached with the choice of $\varepsilon = (n-k)/(2n)$ yielding $\varepsilon_n^* = \lfloor \frac{n-k+1}{2} \rfloor / n$ (See Lopuhaä and Rousseeuw 1991, Theorem 3.2). The equation (2.24) explicit's the relation with ε and the constant c_0. Given the value of ε, the value for c_0 is determined and vice versa, given the value of c_0, the value of ε is determined.

S-estimators and M-estimators are closely related. Lopuhaä (1989) showed that the S-estimates of the location and scatter satisfy the M-estimation equation (2.19) with weight functions

$$v_1(d) = \rho'(d)/d, \qquad v_2(d) = k\rho'(d)/d \quad \text{and} \quad v_3(d) = \rho'(d)d - \rho(d) + b_0. \tag{2.25}$$

This implies that the influence functions and the limiting distribution of the S-estimators will be the same as those of the M-estimators. Hence the functions γ_C, α_C and β_C identifying the influence functions of $M(F)$ and $C(F)$ in (2.11) and (2.12) are given by equation (2.21) and the scalars τ_1, τ_2 and τ_3 for the asymptotic variance-covariance matrices are given by equation (2.22). The weight functions in (2.25) do not satisfy the requirements for the uniqueness of the solutions of M-estimation equations, so there may be multiple solutions and therefore one cannot just solve the M-estimation equation to find the S-estimates. Nevertheless, S-estimates are a solution with a high breakdown point but there will always be a solution to the M-estimation equation with an asymptotic breakdown point at most $1/k$. See Lopuhaä (1989) for a detailed discussion.

An Example: *The Biweight S-estimator.* A popular S-estimator, denoted by $S(\varepsilon, k)$ or $S(\varepsilon)$, uses Tukey's biweight ρ-function,

$$\rho(d) = \begin{cases} \frac{d}{2}^2 - \frac{d^4}{2c^2} + \frac{d^6}{6c^4}, & |d| \le c_0 \\ \frac{d^2}{6} & |d| \ge c_0. \end{cases} \tag{2.26}$$

Lopuhaä (1989) showed that, in the multinormal case, the asymptotic relative efficiency of the $S(\varepsilon, k)$-estimators with respect to the sample mean and covariance matrix goes to

1 as $c_0 \to \infty$. However, the breakdown point also depends on the value of c_0 and high breakdown point corresponds to low values of c_0. Therefore one has to make a tradeoff between high breakdown point and good asymptotic efficiency.

Chapter 3

Sign and Rank Covariance Matrices

3.1 Univariate Sign and Rank

The estimation and testing theory based on univariate signs and ranks is well studied and the estimates/tests are widely used in applied sciences. Typically, no strict assumptions about the underlying distributional form is needed, the estimates and test often have good robustness properties (high breakdown point and a bounded influence function) and they are highly efficient.

Let us recall the definition of the univariate sign and rank. Let $X_n = \{x_1, \ldots, x_n\}$ be a univariate data set. The univariate sign $S_i = S(x_i)$ and centered rank $R_i = R(x_i)$ of an observation $x_i \in \mathbb{R}$ can be defined implicitly through the following L_1-type objective functions

$$\text{ave}\{|x_i|\} = \text{ave}\{S_i\, x_i\}, \tag{3.1}$$

$$\text{ave}\{|x_i - x_j|\} = \text{ave}\{R_i\, x_i\}. \tag{3.2}$$

In the univariate case, signs and ranks have been extensively used to construct estimates and tests in the one sample, two-sample and several sample cases and also in univariate regression problems. For a review of the univariate sign and rank methods, see Hettmansperger (1984) and Lehmann (1998), for example.

3.2 Multivariate Signs and Ranks

The concepts of sign and rank can be extended to the multivariate case by considering extensions of the L_1-type objective functions in (3.1) and (3.2). Three possibilities found in the literature are described next. We let $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a k-variate data set in $\mathbb{R}^k$.

3.2.1 The Marginal Sign and Rank

For extensions of L_1-type objective equations (3.1) and (3.2) to the multivariate case one needs a distance or norm $|\cdot|$ in $\mathbb{R}^k$. If one uses the *Manhattan-distance*, $|\boldsymbol{x}| = |x_1| + \ldots + |x_k|$, then the multivariate sign $\boldsymbol{S}_i = \boldsymbol{S}(\boldsymbol{x}_i) \in \mathbb{R}^k$ and rank $\boldsymbol{R}_i = \boldsymbol{R}(\boldsymbol{x}_i) \in \mathbb{R}^k$ of an observation $\boldsymbol{x}_i$, defined through

$$\begin{aligned} \text{ave}\{|\boldsymbol{x}_i|\} &= \text{ave}\{\boldsymbol{S}_i^T \boldsymbol{x}_i\}, \\ \text{ave}\{|\boldsymbol{x}_i - \boldsymbol{x}_j|\} &= \text{ave}\{\boldsymbol{R}_i^T \boldsymbol{x}_i\} \end{aligned}$$

are just the vectors of marginal (componentwise) signs and ranks. The marginal signs and ranks have been used to construct multivariate analogues of the univariate sign and rank tests and estimates in the one sample, two-sample and multi-sample cases and also in the multivariate regression problems. See the text by Puri and Sen (1971). These test and estimates are scale invariant/equivariant but unfortunately not rotation equivariant/invariant. The lack of affine equivariance is undesirable as such procedures usually suffer a serious loss in efficiency when the marginal variables are substantially correlated. For a discussion on the relation of affine equivariance and efficiency, see e.g. Chakraborty and Chaudhuri (1998). Recently, affine equivariant extensions of the marginal sign and rank methods have been constructed by Chakraborty and Chaudhuri (1996, 1997, 1998) and Chakraborty (1998) using the so called transformation-retransformation technique.

3.2.2 The Spatial Sign and Rank

If the *Euclidean distance*, $\|\boldsymbol{x}\| = (x_1^2 + \ldots + x_k^2)^{1/2}$, is chosen as a norm in $\mathbb{R}^k$ then the associated spatial signs $\boldsymbol{S}_i = \boldsymbol{S}(\boldsymbol{x}_i) \in \mathbb{R}^k$ and ranks $\boldsymbol{R}_i = \boldsymbol{R}(\boldsymbol{x}_i) \in \mathbb{R}^k$ verify, analogous to (3.1) and (3.2),

$$\begin{aligned} \text{ave}\{\|\boldsymbol{x}_i\|\} &= \text{ave}\{\boldsymbol{S}_i^T \boldsymbol{x}_i\}, \\ \text{ave}\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|\} &= \text{ave}\{\boldsymbol{R}_i^T \boldsymbol{x}_i\}. \end{aligned}$$

The spatial signs and ranks have been used to construct multivariate analogues of the univariate sign and rank tests and estimates in the one-sample and multi-sample cases and also in the multivariate regression problems. See Brown (1983), Rao (1988), Bai, Chen, Miao and Rao (1990), Chaudhuri (1992), Möttönen and Oja (1995), Möttönen, Oja and Tienari (1997), Choi and Marden (1997) and Möttönen, Hüsler and Oja (2002). A drawback is that these testing/estimation procedures are not affine invariant/equivariant, but only rotation invariant/equivariant. Affine equivariant extensions of these methods have been constructed by Hössjer and Croux (1995), Chakraborty, Chaudhuri and Oja (1998) and Randles (2000).

3.2.3 The Affine Equivariant Sign and Rank

The affine equivariant signs $\boldsymbol{S}_i = \boldsymbol{S}(\boldsymbol{x}_i) \in \mathbb{R}^k$ and ranks $\boldsymbol{R}_i = \boldsymbol{R}(\boldsymbol{x}_i) \in \mathbb{R}^k$ are defined by

$$\text{ave}\{|\det\begin{pmatrix}\boldsymbol{x}_{i_1} & \cdots & \boldsymbol{x}_{i_{k-1}} & \boldsymbol{x}_i\end{pmatrix}|\} = \text{ave}\{\boldsymbol{S}_i^T\boldsymbol{x}_i\},$$
$$\text{ave}\{|\det\begin{pmatrix}\boldsymbol{x}_{i_1}-\boldsymbol{x}_i & \cdots & \boldsymbol{x}_{i_k}-\boldsymbol{x}_i\end{pmatrix}|\} = \text{ave}\{\boldsymbol{R}_i^T\boldsymbol{x}_i\},$$

where the first average is over $i = 1, \ldots, n$ and all $(k-1)$-subsets $1 \le i_1 < \ldots < i_{k-1} \le n$ whereas the second average is over $i = 1, \ldots, n$ and all k-subsets $1 \le i_1 < \ldots < i_k \le n$. The affine equivariant signs and ranks have been used to construct multivariate analogues of the univariate sign and rank tests and estimates in the one-sample and multi-sample cases and also in the multivariate regression problems. See Brown and Hettmansperger (1987, 1989), Brown, Hettmansperger, Nyblom and Oja (1992), Hettmansperger and Oja (1994), Hettmansperger, Nyblom and Oja (1994), Hettmansperger, Möttönen and Oja (1998) and Oja (1999). Unlike in the case of marginal and spatial sign and rank, the test/estimation procedures are now fully affine invariant/equivariant.

3.3 The Sign and Rank Covariance Matrices

Visuri, Koivunen and Oja (2000) defined the affine equivariant sign and rank covariance matrices as follows. For constructing the sign covariance matrix, first consider the centered data set

$$Y_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\} = \{\boldsymbol{x}_1 - \hat{\boldsymbol{\mu}}, \ldots, \boldsymbol{x}_n - \hat{\boldsymbol{\mu}}\}.$$

The location estimator $\hat{\boldsymbol{\mu}}$ is chosen so that the sum of sign vectors based on Y_n is a zero vector. When using marginal signs, $\hat{\boldsymbol{\mu}}$ becomes the *vector of marginal medians*. In case of spatial signs, $\hat{\boldsymbol{\mu}}$ is the *spatial median* or L_1-median (see e.g Brown, 1983), and the affine equivariant signs corresponds to the *Oja median* (Oja, 1983). The Oja median is a location estimate satisfying the conditions stated in Section 2.2.1. It is affine equivariant (Condition A) and consistent (Condition B) in a class of distributions with finite first order moments. The spatial median is not scale equivariant and the marginal median do not satisfy rotation equivariance.

Let $\hat{\boldsymbol{S}}_i$, $i = 1, \ldots, n$, be the signs of the centered observations. The *sign covariance matrix* (SCM) is now the usual covariance matrix computed from the multivariate signs, that is,

$$\text{SCM} = \text{ave}\{\hat{\boldsymbol{S}}_i\hat{\boldsymbol{S}}_i^T\}. \tag{3.3}$$

In the same way, the *rank covariance matrix* (RCM) is defined as

$$\text{RCM} = \text{ave}\{\boldsymbol{R}_i\boldsymbol{R}_i^T\}, \tag{3.4}$$

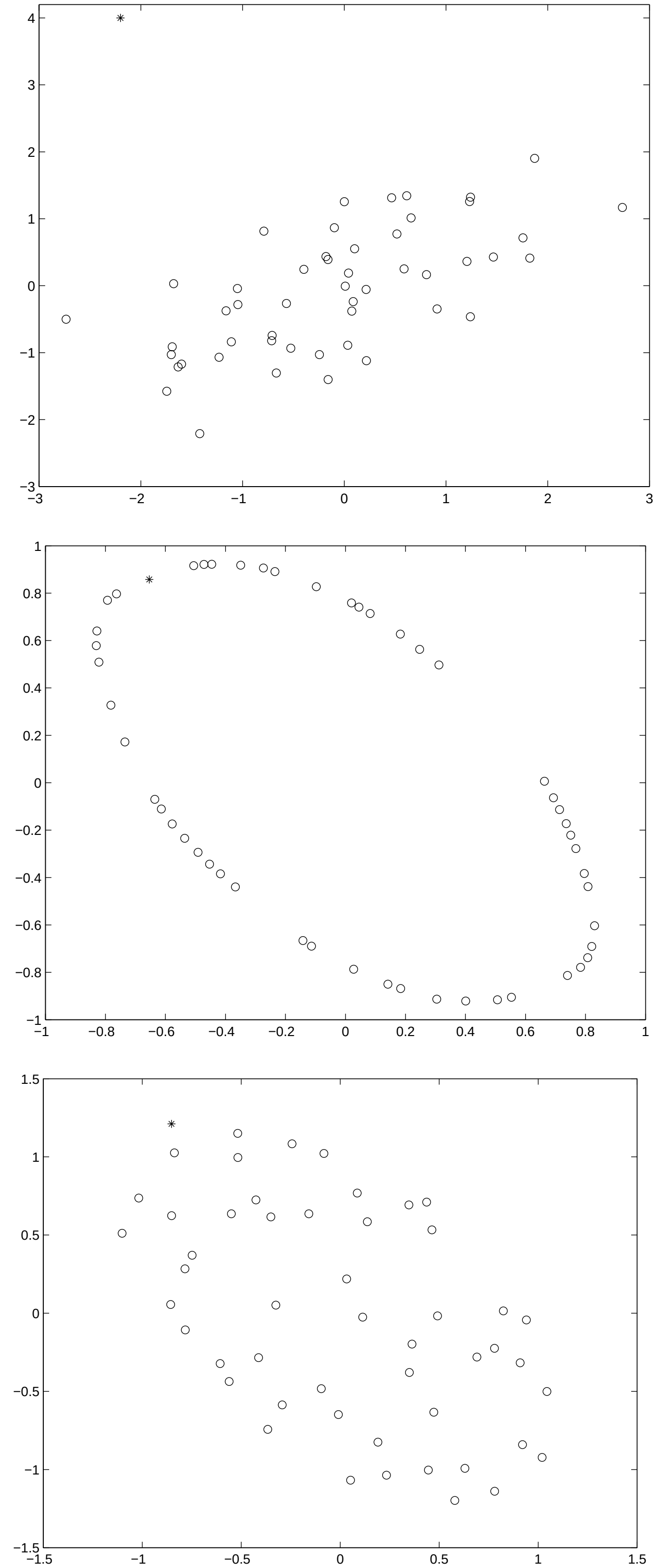
where the rank vectors are computed from the uncentered data.

Depending on which sign and rank concepts are used, covariance matrices with different equivariance, efficiency and robustness properties will be obtained. Statistical properties (consistency, limiting distribution and efficiencies, influence function, etc.) of the affine equivariant sign and rank covariance are established in Paper I and Paper IV and in Visuri, Ollila, Koivunen, Möttönen and Oja (1998). In Paper II and V, an estimator of the parameters of the multiresponse regression model based on the affine equivariant sign and rank covariance matrix are introduced and studied respectively. Properties of the spatial sign and rank covariance matrices have been studied in Marden (1998), Visuri, Koivunen and Oja (1998), Visuri (2001), Croux, Ollila and Oja (2002). Their asymptotic behavior (limiting covariance matrices and efficiencies) has not yet been fully solved, however. Results concerning the limiting variances and efficiencies of the eigenvectors of the spatial sign covariance matrix can be found from Croux, Ollila and Oja (2002).

In the following, affine equivariant sign and ranks and the related sign and rank covariance matrices will be considered. In Fig. 3.1, a data set generated from bivariate elliptical distribution is pictured together with the sign and rank vectors. We see that the sign transformation move the data points towards the periphery of an ellipse. The sign vector points in the direction of the corresponding observation vector, while its magnitude depends on the dispersion of the data in the space orthogonal to the vector. The form of this ellipse is therefore determined by the *inverse* of the covariance structure of the data. As can be seen from Fig. 3.1, this structure has not been influenced by the outlier (marked by $*$) present in the data cloud. Also notice that the rank vectors have the covariance structure of the inverse of the covariance matrix and the transformed data cloud, i.e. the ranks of the observations, adapts the shape of the ellipsoid. Again, this shape has not been influenced by the outlier.

3.4 The Affine Equivariant Sign and Rank Covariance Matrix

Contrary to the univariate case, the multivariate affine equivariant signs and ranks can be used to estimate the covariance structure also. See Paper I and Paper IV. We will now discuss and review the most important properties of the affine equivariant sign and rank covariance matrix and covariance matrix estimates based on them. We will make comparison in terms of efficiency and robustness to the sample covariance matrix, M-estimators and the S-estimators.



a)

b)

c)

Figure 3.1: A bivariate data cloud (Fig. a) together with the corresponding sign vectors (Fig. b) and rank vectors (Fig. c)

The SCM and RCM are affine equivariant in the sense that if SCM* and RCM* are calculated from the transformed observations $\boldsymbol{x}_i^* = A\boldsymbol{x}_i + \boldsymbol{b}$, $i = 1, \ldots, n$ with nonsingular A, then

$$\mathrm{SCM}^* = \det(A)^2 (A^{-1})^T\, \mathrm{SCM}\, A^{-1} \tag{3.5}$$

and

$$\mathrm{RCM}^* = \det(A)^2 (A^{-1})^T\, \mathrm{RCM}\, A^{-1}. \tag{3.6}$$

The Oja median $\hat{\boldsymbol{\mu}}$ and the associated functional $M(F)$ are affine equivariant. The sign and rank covariance matrices SCM and RCM are asymptotically equivalent with U-statistics and the population counterparts, the statistical functionals SCM(F) and RCM(F) are defined as the expected values of the related kernels. It also follows that the SCM and RCM are consistent under general conditions and have limiting multinormal distributions. For details, see Paper I and Paper IV and Visuri, Ollila, Koivunen, Möttönen and Oja (2002).

At elliptical $F = F_{\boldsymbol{\mu},\Sigma}$, the functionals SCM$(F)$ and RCM(F) satisfy

$$\mathrm{SCM}(F) = \lambda_S(F_0) \det(\Lambda) P \Lambda^{-1} P^T \tag{3.7}$$

and

$$\mathrm{RCM}(F) = \lambda_R(F_0) \det(\Lambda) P \Lambda^{-1} P^T, \tag{3.8}$$

where $\lambda_S(F_0)$ and $\lambda_R(F_0)$ are scalars depending only on the underlying spherical distribution $F_{\mathbf{0},I_k}$. The equations (3.7) and (3.8) also hold true in a more general location-scale model, see Paper I and Visuri, Ollila, Koivunen, Möttönen and Oja (2001). SCM(F) and RCM(F) are then proportional to the inverse of the covariance matrix ($\Sigma^{-1} = P\Lambda^{-1}P^T$). Therefore the eigenvectors $\boldsymbol{p}_i$ and the standardized eigenvalues λ_i^* of Σ can be estimated using the affine equivariant SCM and RCM. As many applications require the knowledge of the covariance matrix only up to constant scalar factor, applications of SCM and RCM are possible in multivariate multiple regression analysis, discriminant analysis, canonical correlations analysis (CCA), etc. See Papers I-V, Visuri, Koivunen and Oja (2000), and Visuri, Ollila, Koivunen, Möttönen and Oja (2002).

As constants $\lambda_S(F_0)$ and $\lambda_R(F_0)$ in equations (3.7) and (3.8) can be made explicit under a specified underlying elliptical model F, the covariance structure of F can be retrieved from the SCM and RCM. For elliptical distribution F, the following scatter matrix functionals based on the SCM and RCM,

$$C_S(F) = \left[\frac{\det\{\mathrm{SCM}(F)\}}{\lambda_S(F_0)}\right]^{1/(k-1)} \mathrm{SCM}(F)^{-1} \tag{3.9}$$

and

$$C_R(F) = \left[\frac{\det\{\mathrm{RCM}(F)\}}{\lambda_R(F_0)}\right]^{1/(k-1)} \mathrm{RCM}(F)^{-1} \tag{3.10}$$

and the corresponding scatter matrix estimates $\hat{C}_S$ and $\hat{C}_R$, are affine equivariant (Condition A) and consistent (Condition B) under general assumptions. See Paper I and Paper IV for details.

At elliptical F, the influence functions of the Oja median $M(F)$ and the associated scatter functional $C_S(F)$ based on the SCM are as in (2.11) and (2.12) respectively with weight functions

$$\gamma_M(r;F_0) = \frac{k}{(k-1)E_{F_0}(\|\boldsymbol{x}\|^{-1})}, \quad \alpha_{C_S}(r;F_0) = \frac{2kr}{E_{F_0}(\|\boldsymbol{x}\|)} - k, \quad \beta_{C_S}(r;F_0) = 1. \tag{3.11}$$

The γ_M function of the Oja median is constant in r; its influence function is thus continuous and bounded. Unfortunately, the scatter functional $C_S(F)$ is not robust as its influence function is unbounded and linear in r. For the scatter functional $C_R(F)$, the weight functions $\alpha_{C_R}(r;F_0)$ and $\beta_{C_R}(r;F_0)$ are somewhat complicated and can be found from Paper IV, Corollary 1. The function $\alpha_{C_R}(r;F_0)$ is approximately linear in r, so $C_R(F)$ has an unbounded influence function. Recall that the functions α_C and β_C also determine the influence functions of the corresponding eigenvalue and eigenvector functionals through equations (2.13) and (2.14). It is obvious that the asymptotic breakdown point of the SCM and RCM scatter estimators is zero. This is also the case for the Oja median whose finite sample ABP is $k/(n+k)$ (Niinimaa, Oja and Tableman, 1990). The scalars τ_1, τ_2 and τ_3 needed for the asymptotic variance-covariance matrices (2.15) and (2.16) of the Oja median $\hat{\boldsymbol{\mu}}$ and the scatter estimate $\mathrm{vec}(\hat{C}_S)$ at elliptical distribution F are given by

$$\tau_1 = \frac{k}{(k-1)^2E^2(r^{-1})}, \quad \tau_2 = \frac{k}{k+2}\left[\frac{4E(r^2)}{E^2(r)} - 3\right] \quad \text{and} \quad \tau_3 = \tau_2 - 1, \tag{3.12}$$

where r^2 has the density (2.4). For $\hat{C}_R$, the expressions for τ_2 and τ_3 are more complicated and can be found from Paper IV.

3.4.1 Comparisons to Existing Estimators

Let us first compare the influence functions of different location functionals. In Fig. 3.2, the function $\gamma_M(r;F_0)$ of different location functionals $M(F)$ are depicted in the bivariate ($k=2$) standard normal distribution ($F_0 = N(\mathbf{0}, I_2)$). The location functionals considered are the Oja median, the mean vector $M(F) = E_F(\boldsymbol{x})$, the $M(0.9,2)$- and $S(0.25,2)$-functionals. We see that all considered competitors of the mean vector are robust to outliers: their influence function is continuous and bounded. The γ_M-function of the S-functional is even redescending to zero so observations with large r receive zero weight.

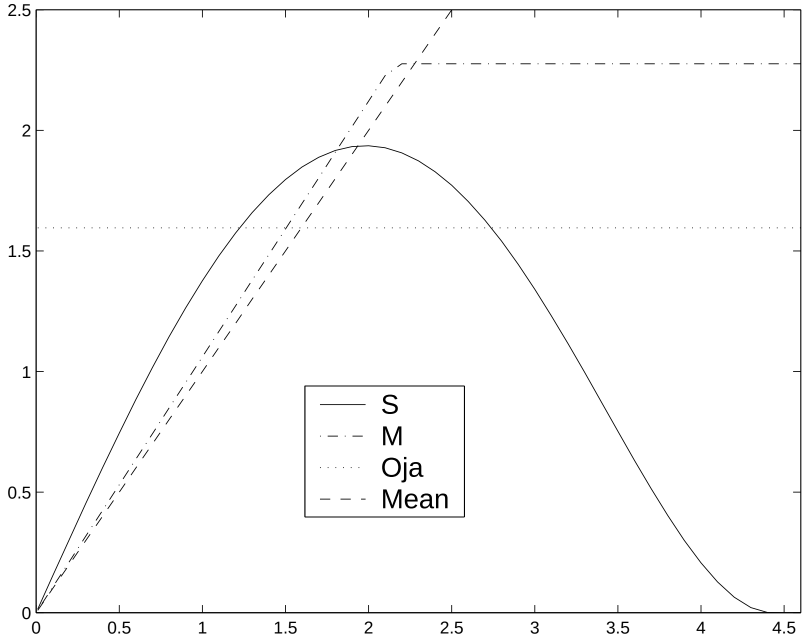
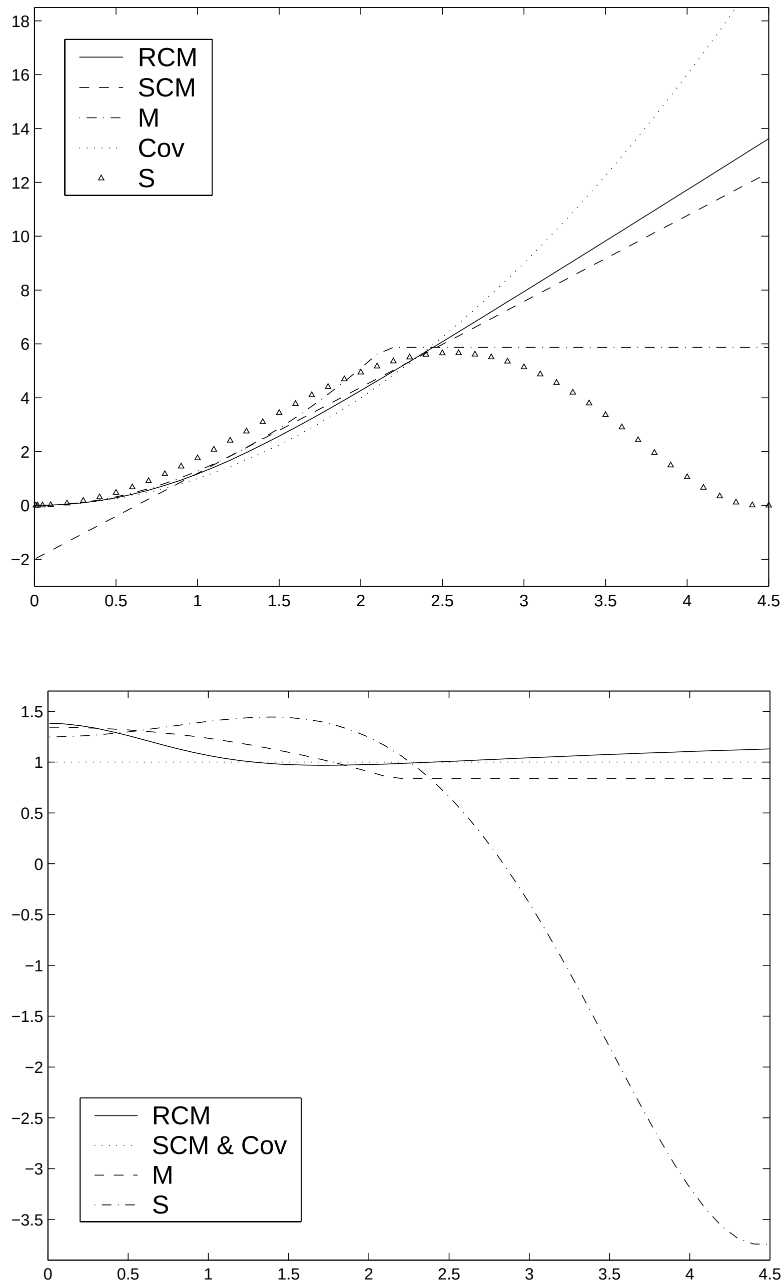


Figure 3.2: Function $\gamma_M(r; F_0)$ for the Oja median, the mean vector and the $M(0.9, k)$- and $S(0.25, k)$-functionals at the bivariate ($k = 2$) standard normal distribution ($F_0 = N(\mathbf{0}, I_2)$).

Similarly, in Fig. 3.3, the functions $\alpha_C(r; F_0)$ and $\beta_C(r; F_0)$ for different scatter functionals $C(F)$ are illustrated again at the bivariate ($k = 2$) standard normal distribution. The scatter functionals considered are the scatter functionals $C_S(F)$ and $C_R(F)$ based on the SCM and RCM respectively, the covariance matrix functional $C(F) = \mathrm{Cov}_F(\boldsymbol{x})$, the $M(0.9, 2)$- and the $S(0.25, 2)$-functional of scatter. We see that the α_C-function of the covariance matrix is quadratic, whereas that of $C_S(F)$ and $C_R(F)$ is linear for large r. Thus the scatter estimators based on the SCM and RCM give more protection to outliers than the covariance matrix. Their influence function remains unbounded however so the SCM and RCM scatter estimators are not robust in the strict sense. Again, we see that α_C-function of the S-functional is even redescending to zero so observations with large r receive zero weight. All β_C-functions are seen to be bounded but the β_C-function of the S-estimator differs from that of other estimators giving a large negative weight for observations with large r.

Let us now compare the asymptotic performance of the location estimates and the scatter estimates under the multinormal model. We first calculated the asymptotic variance $\tau_1 = \mathrm{ASV}(\hat{\mu}_1; F_0)$ for different location estimators $\hat{\boldsymbol{\mu}}$ at the k-variate standard normal distribution $F_0 = N(\mathbf{0}, I_k)$ with various choices of k. The location estimators considered were the sample mean, the Oja median, the $M(0.9, k)$-, $S(0.25, k)$- and $S(0.5, k)$-estimators of location. The variances are reported in Table 3.1. First



a)

b)

Figure 3.3: Functions $\alpha_C(r; F_0)$ (Fig. a) and $\beta_C(r; F_0)$ (Fig. b) for the SCM and RCM scatter functionals, the covariance matrix functional, and the $M(0.9, k)$- and $S(0.25, k)$-functionals of scatter at the bivariate ($k = 2$) standard normal distribution ($F_0 = N(\mathbf{0}, I_2)$).

	$k=2$	$k=3$	$k=5$	$k=8$	$k=10$
Mean	1.000	1.000	1.000	1.000	1.000
Oja	1.273	1.178	1.105	1.064	1.051
$M(0.9)$	1.012	1.008	1.005	1.003	1.002
$S(0.25)$	1.097	1.051	1.025	1.013	1.010
$S(0.50)$	1.725	1.384	1.182	1.096	1.072

Table 3.1: The asymptotic variances τ_1 of the sample mean, the Oja median, the $M(0.9, k)$-estimator of location and the $S(0.25, k)$- and $S(0.5, k)$-estimators of location at the k-variate standard normal distribution.

recall that the ARE's of the estimators are obtained as ratios of corresponding values of τ_1. We see that the sample mean has the smallest variance as it should since it is optimal estimator under the multinormal model. The $M(0.9, k)$-estimator and the 25% breakdown S-estimator $S(0.25, k)$ are also performing well. Indeed the sample mean is only marginally better than the Huber's M-estimator especially in high dimension. The Oja median is also efficient except for small dimensions $k = 2$ and $k = 3$, the efficiency loss being less than 10% for $k \geq 5$. We see that there is a serious loss in efficiency for choosing 50% breakdown S-estimator $S(0.50, k)$ instead of the 25% breakdown estimator $S(0.25, k)$ for small dimensions. Finally note that for all the robust estimators, the variances τ_1 are decreasing in dimension and already for $k = 10$ the differences in performance are only marginal.

The asymptotic variances $\tau_2 = \mathrm{ASV}(\hat{C}_{12}; F_0)$ and $\tau_4 = \mathrm{ASV}(\hat{C}_{11}; F_0)$ and the asymptotic covariances $\tau_3 = \mathrm{ASC}(\hat{C}_{11}, \hat{C}_{22}; F_0)$ of different scatter estimators $\hat{C}$ at the k-variate standard normal distribution $F_0 = N(\mathbf{0}, I_k)$ with various choices of k were also calculated. The scatter estimators considered were the sample covariance matrix, the SCM and RCM scatter estimators, the $M(0.9, k)$-, $S(0.25, k)$- and $S(0.5, k)$-estimators of scatter. The numbers are reported in Table 3.2. By comparing the corresponding values of τ_2 and τ_4 of the estimators we attain their asymptotic performance in estimating the off-diagonal and on-diagonal elements of the covariance matrix respectively. Also recall that ARE's of the estimators of eigenvectors are also obtained by comparing the values of τ_2 and the ARE's of the estimators of eigenvalues by comparing the values of τ_4. From Table 3.2 we see that the sample covariance matrix naturally yields the smallest variances. The SCM and RCM estimators have the best performance among the alternatives, the sample covariance matrix being only marginally better. Also the $M(0.9, k)$-estimator is performing well. Note again the loss in efficiency when choosing $S(0.50, k)$ instead of $S(0.25, k)$ especially in low dimension. For the competitors of the sample covariance matrix, the variances τ_2 and τ_4 are decreasing with dimension and,

		$k=2$	$k=3$	$k=5$	$k=8$	$k=10$
Cov	τ_2	1.000	1.000	1.000	1.000	1.000
	τ_4	2.000	2.000	2.000	2.000	2.000
	τ_3	0	0	0	0	0
SCM	τ_2	1.047	1.027	1.013	1.006	1.004
	τ_4	2.139	2.082	2.038	2.018	2.012
	τ_3	0.047	0.027	0.013	0.006	0.004
RCM	τ_2	1.010	1.010	1.031	1.010	1.031
	τ_4	2.041	2.041	2.083	2.020	2.062
	τ_3	0.021	0.021	0.022	0	0
$M(0.9)$	τ_2	1.087	1.056	1.032	1.019	1.014
	τ_4	2.268	2.171	2.095	2.055	2.042
	τ_3	0.093	0.058	0.032	0.018	0.014
$S(0.25)$	τ_2	1.177	1.082	1.034	1.017	1.012
	τ_4	2.225	2.126	2.060	2.031	2.023
	τ_3	-0.123	-0.038	-0.009	-0.003	-0.001
$S(0.50)$	τ_2	0.377	0.579	0.778	0.884	0.915
	τ_4	3.980	3.090	2.491	2.241	2.175
	τ_3	-1.332	-0.362	-0.078	-0.021	-0.012

Table 3.2: The limiting variances of the off-diagonal (τ_2) and on-diagonal (τ_4) and the limiting covariances of the on-diagonal elements (τ_3) of the sample covariance matrix (labeled Cov), the SCM and RCM scatter estimators, the $M(0.9,k)$-, $S(0.25,k)$- and $S(0.50,k)$-estimators of scatter at the k-variate standard normal distribution.

already for $k = 10$, the differences in performance are quite narrow. A curious fact is that the asymptotic covariances τ_3 of S-estimators are negative. This may be related to the behavior of $\beta_C(r; F_0)$ function of the S-estimator. For the efficiencies of the SCM and RCM scatter estimator at t-distribution $t_{\nu,k}$ with various choices of degrees of freedom ν and dimension k, see Papers I and IV.

Chapter 4

Summary of Original Publications

Paper I considers influence functions and limiting distributions of the affine equivariant sign covariance matrix and its eigenvectors and eigenvalues. An estimate of the covariance and correlation matrix based on the sign covariance matrix is also constructed and their limiting efficiencies with respect to the sample covariance and correlation matrix are calculated in multivariate normal and t distribution cases. Finite sample efficiencies are examined by simulations.

In Paper II, a new affine equivariant estimator of the parameters of the multivariate multiple regression model is constructed using the affine equivariant sign covariance matrix and the Oja median. The influence function and limiting distribution of the proposed estimate are examined and asymptotic relative efficiencies with respect to the least squares estimate are calculated in the multivariate normal and t distribution cases. Various properties of the estimate are investigated by simulations. The theory is illustrated with an example.

Paper III presents a framework for constructing sign based location vector and shape matrix estimates. The used sign concepts are the spatial sign vector based on the spatial median and the affine equivariant sign vector based on the Oja median. The influence functions and limiting distributions of the resulting location vector and shape matrix estimates are derived in the elliptic case and their use in multivariate analysis problems are discussed.

In Paper IV, the influence functions and the limiting variances and covariances of the affine equivariant rank covariance matrix and the associated covariance matrix estimate based on the rank covariance matrix are derived in the multivariate elliptic case. Limiting efficiencies are calculated in multivariate normal and t distribution cases.

In Paper V, a new affine equivariant estimator of the parameters of the multivariate multiple regression model is constructed using the lift rank covariance matrix where the lift rank vectors are based on Oja's criterion function. The new estimate as well as the least squares estimate are shown to be weighted sums of the so called elemental regression estimates. The limiting distribution and influence function of the proposed estimate are examined and the theory is illustrated with an analysis of a real data set. Various properties of the estimate are investigated by simulations.

Bibliography

[1] Bai, Z. D., Chen, N. R., Miao, B. Q., and Rao, C. R. (1990), "Asymptotic Theory of Least Distances Estimate in Multivariate Linear Models," *Statistics*, **21**, 503–519.

[2] Brown, B. M. (1983), "Statistical Uses of the Spatial Median," *Journal of the Royal Statistical Society, Series B*, **45**, 25–30.

[3] Brown, B. M. and Hettmansperger, T. P. (1987), "Affine Invariant Rank Methods in the Bivariate Location Model," *Journal of the Royal Statistical Society, Series B*, **49**, 301–310.

[4] Brown, B. M. and Hettmansperger, T. P. (1989), "An Affine Invariant Bivariate Version of the Sign Test," *Journal of the Royal Statistical Society, Series B*, **51**, 117–125.

[5] Brown, B. M., Hettmansperger, T. P., Nyblom, J. and Oja, H. (1992), "On Certain Bivariate Sign Tests and Medians," *Journal of the American Statistical Association*, **87**, 127–135.

[6] Chakraborty, B. (1999), "On Multivariate Median Regression," *Bernoulli*, **5**, 683–703.

[7] Chakraborty, B. and Chaudhuri, P. (1996), "On a Transformation and Retransformation Technique for Constructing Affine Equivariant Multivariate Median," *Proceedings of the American Mathematical Society*, **124**, 2359–2547.

[8] Chakraborty, B. and Chaudhuri, P. (1997), "On Multivariate Rank Regression," In L_1*-Statistical Procedures and Related Topics* (ed. Y. Dodge), IMS Lecture Notes - Monograph Series, Volume 1. Hayward, CA. pp. 399–414.

[9] Chakraborty, B. and Chaudhuri, P. (1998), "On an Adaptive Transformation-retransformation Estimate of Multivariate Location," *Journal of the Royal Statistical Society, Series B*, **60**, 145–157.

[10] Chakraborty, B., Chaudhuri, P. and Oja, H. (1998), "Operating Transformation Retransformation on Spatial Median and Angle Test," *Statistica Sinica*, **8**, 767–784.

[11] Chaudhuri, P. (1992), "Multivariate Location Estimation Using Extension of R-estimates Through U-statistics Type Approach," *The Annals of Statistics*, **20**, 897–916.

[12] Choi, K. and Marden, J. (1997), "An Approach to Multivariate Rank Tests in Multivariate Analysis of Variance," *Journal of the American Statistical Association*, **92**, 1581–1590.

[13] Croux, C. and Haesbroeck, G. (2000), "Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies," *Biometrika*, **87**, 603–618.

[14] Croux, C., Ollila, E. and Oja, H. (2002), "Sign and Rank Covariance Matrices: Statistical Properties and Application to Principal Component Analysis," *Statistical data analysis based on L_1-norm and related methods* (ed. Y. Dodge). Birkhauser, Basel.

[15] Davies, P. L. (1987), "Asymptotic Behavior of S-estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, **15**, 1269–1292.

[16] Donoho, D. L. (1982), "Breakdown Properties of Multivariate Location Estimators," Ph.D. qualifying paper. Harvard University.

[17] Donoho, D. L. and Huber, P.J. (1983), "The Notion of Breakdown Point," In *Festschrift in Honor of Erich Lehmann* (eds. P. Bickel, K. Doksum and J. L. Hodges). Wadsworth, Belmont, CA.

[18] Donoho, D. L. and Gasko, M. (1992), "Breakdown Properties of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, **15**, 1269–1292.

[19] Hampel, F. R. (1968), "Contributions to the Theory of Robust Estimation", Ph. D. Thesis, University of California, Berkeley.

[20] Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation", *Journal of the American Statistical Association*, **69**, 383–393.

[21] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

[22] Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks.* Wiley, New York.

[23] Hettmansperger, T. P. and Oja, H. (1994), "Affine Invariant Multivariate Multisample Sign Tests," *Journal of the Royal Statistical Society, Series B*, **56**, 235–249.

[24] Hettmansperger, T. P., Nyblom, J. and Oja, H. (1994), "Affine Invariant Multivariate One-sample Sign Tests," *Journal of the Royal Statistical Society, Series B*, **56**, 221–234.

[25] Hettmansperger, T. P., Möttönen, J. and Oja, H. (1998), "Affine Invariant Multivariate Rank Tests for Several Samples," *Statistica Sinica*, **8**, 765–800.

[26] Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

[27] Huber, P. J. (1984), "Finite Sample Breakdown of M- and P-estimators," *The Annals of Statistics*, **12**, 119–126.

[28] Hössjer, O. and Croux, C. (1995), "Generalizing Univariate Signed Rank Statistics for Testing and Estimating a Multivariate Location Parameter," *Nonparametric Statistics*, **4**, 293–308.

[29] Kent, J. T. and Tyler, D. E. (1991), "Redescending M-estimates of Multivariate Location and Scatter," *The Annals of Statistics*, **19**, 2102–2119.

[30] Kent, J. T. and Tyler, D. E. (1996), "Constrained M-estimation for Multivariate Location and Scatter," *The Annals of Statistics*, **24**, 1346–1370.

[31] Lehmann, E. L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, Prentice-Hall, Upper Saddle River.

[32] Lopuhaä, H. P. (1989), "On the Relation Between S-estimators and M-estimators of Multivariate Location and Scatter," *The Annals of Statistics*, **17**, 1662–1683.

[33] Lopuhaä, H. P. (1991), "Multivariate τ-estimators for Location and Scatter," *Canadian Journal of Statistics*, **19**, 307–321.

[34] Lopuhaä, H. P. and Rousseeuw P. J. (1991), "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, **19**, 229–248.

[35] Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, New York.

[36] Marden, J. I. (1999), "Some Robust Estimates of Principal Components," *Statistics and Probability Letters*, **43**, 349–359.

[37] Maronna, R. A. (1976), "Robust M-estimators of Multivariate Location and Scatter," *The Annals of Statistics*, **4**, 51–67.

[38] Maronna, R. A. and Yohai, V. J. (1995), "The Behavior of the Stahel-Donoho Robust Multivariate Estimator", *Journal of the American Statistical Association*, **90**, 330–341.

[39] Möttönen, J. and Oja, H. (1995), "Multivariate Spatial Sign and Rank Methods," *Journal of Nonparametric Statistics*, **5**, 202–213.

[40] Möttönen, J., Oja, H. and Tienari, J. (1997), "On the Efficiency of Multivariate Spatial Sign and Rank Tests," *The Annals of Statistics*, **25**, 542–552.

[41] Möttönen, J., Hüsler, J. and Oja, H. (2002). Multivariate Nonparametric Tests in a Randomized Complete Block Design. *Journal of Statistical Planning and Inference*, To appear.

[42] Niinimaa, A., Oja, H. and Tableman, M. (1990), "The Finite Sample Breakdown Point of the Oja Bivariate Median," *Statistics and Probability Letters*, **10**, 325–328.

[43] Oja, H. (1983), "Descriptive Statistics for Multivariate Distributions," *Statistics and Probability Letters*, **1**, 327–332.

[44] Oja, H. (1999), "Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: a Review," *Scandinavian Journal of Statistics*, **26**, 319–343.

[45] Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[46] Randles, R. H. (2000), "A Simpler, Affine-invariant, Multivariate, Distribution-free Sign Test," *Journal of the American Statistical Association*, **95**, 1263–1268.

[47] Rao, C. R. (1988), "Methodology Based on L_1-norm in Statistical Inference," *Sankhya, Ser. A*, **50**, 289–313.

[48] Rousseeuw, P. J. (1985), "Multivariate Estimation With High Breakdown Point," In *Mathematical Statistics and Applications*, Vol. B (eds. W. Grossmann, G. Pflug, I Vincze and W. Wertz) 283–297. Reidel, Dordrecht.

[49] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

[50] Stahel, W. A. (1981), "Breakdown of Covariance Estimators," Research Report 31, Fachgruppe für Statistic, E.T.H. Zürich.

[51] Tatsuoka, K. S. and Tyler D. E. (2000), "On the Uniqueness of S-functionals Under Non-Elliptical Distributions", *The Annals of Statistics*, **28**, 1219-1243.

[52] Tyler, D. E. (1982), "Radial Estimates and the Test for Sphericity," *Biometrika*, **69**, 429–436.

[53] Tyler, D. E. (1983), "Robustness and Efficiency Properties of Scatter Matrices," *Biometrika*, **70**, 411–420.

[54] Tyler, D. E. (1986), "Breakdown Properties of the M-estimator of Multivariate Scatter," Report, Dept. Statistics, Rutgers Univ.

[55] Tyler, D. E. (1994), "Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics," *The Annals of Statistics*, **22**, 1024–1044.

[56] Visuri, S., Koivunen, V., and Oja, H. (2000), "Sign and Rank Covariance Matrices," *Journal of Statistical Planning and Inference*, **91**, 557–575.

[57] Visuri, S., Ollila, E., Koivunen, V., Möttönen, J. and Oja, H. (2001), "Affine Equivariant Multivariate Rank Methods," *Journal of Statistical Planning and Inference*, To appear.

[58] Visuri, S. (2001), "Array and Multichannel Signal Processing Using Nonparametric Statistics," Ph. D. Thesis, Helsinki University of Technology.

[59] Zuo, Y. (2001), "Some Quantitative Relationships Between Two Types of Finite Sample Breakdown Point," *Statistics and Probability Letters*, **51**, 369–375.