KLAUS NORDHAUSEN

On Invariant Coordinate Selection and Nonparametric Analysis of Multivariate Data

■

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine of the University of Tampere,
for public discussion in the Auditorium of
Tampere School of Public Health, Medisiinarinkatu 3,
Tampere, on December 12th, 2008, at 12 o'clock.

UNIVERSITY OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere, School of Public Health
Finland

Supervised by
Professor Hannu Oja
University of Tampere
Finland
Docent Tapio Nummi
University of Tampere
Finland

Reviewed by
Professor Ronald H. Randles
University of Florida
USA
Professor Anne Ruiz-Gazen
University Toulouse 1
France

Distribution
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 3 3551 6055
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Juha Siro

Acta Universitatis Tamperensis 1370
ISBN 978-951-44-7538-2 (print)
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 792
ISBN 978-951-44-7539-9 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2008

Acknowledgements

First of all I wish to warmly thank Professor Hannu Oja for supervising me during the past years. I would especially like to thank him for introducing me to this interesting topic, his expert guidance, constant support and also infinite patience when answering my many questions. I also wish to express my sincere gratitude to my second supervisor Professor Tapio Nummi for his guidance and continuous support as well as the many winter swimming sessions together.

I have been very fortunate to have excellent co-authors for all of my papers. It was a great experience and pleasure to work together with Professor David E. Tyler, Professor Davy Paindaveine and Dr. Esa Ollila as well as a great chance to learn from all of them. My special thanks go to Professor David E. Tyler for his great hospitality when I visited Rutgers in May 2008.

Furthermore, I would like to thank Professor Ilkka Pörsti for providing this interesting data set which brought the real world into this project.

I wish to thank Professor Uwe Ligges who was a great help when making the R packages and Jarmo Niemelä who helped me with any LaTeX problem I had.

I would like to express my gratitude to my referees Professor Anne Ruiz-Gazen and Professor Ronald H. Randles for their careful reading of the thesis and their constructive comments.

The work was financially supported by the Academy of Finland and the Tampere Graduate School of Information Science and Engineering (TISE) and was carried out while I was a researcher in the biometry group of the Tampere School of Public Health and at the Department of Mathematics and Statistics of the University Tampere. I especially wish to thank here Catarina Ståhle-Nieminen for her constant help with all my practical problems. Furthermore, all my colleagues made this time a memorable experience and it was especially great to be a member of the "Friday Book Seminar" and the "Nonparametric & Robust Multivariate Methods Research Group".

Finally I wish to warmly thank all my friends and my family for always being there for me and their constant support. Especially I want to thank Elina for her constant encouragement and support while still making sure that I do also other things in my free time.

Tampere, November 2008

Klaus Nordhausen

Abstract

The aim of this doctoral thesis was to investigate further properties and applications of the recently introduced two scatter matrices transformation of Oja, Sirkiä and Eriksson (2006) and Tyler, Critchley, Dümbgen and Oja (2008). We consider this transformation in the framework of multivariate model selection, robust independent component analysis and multivariate nonparametric location tests. Especially the last one leads to robust affine invariant location tests which are highly efficient for appropriately chosen score functions. The transformation was implemented in R with a large choice for scatter functionals and made publicly available. Together with the other R packages resulting from this work, all methods discussed can be easily applied. For a practical demonstration a hemodynamic data set is analyzed using the methods discussed here.

Key words: affine equivariance, affine invariance, multivariate distributions, marginal signs and ranks, transformation retransformation.

Contents

Abbreviations

$\sim$	distributed as
$.^T$	transpose of a vector $\cdot$ or a matrix $\cdot$
$\mathbf{E}(\cdot)$	multivariate expectation of $(\cdot)$
$\mathbf{E}_3(\cdot)$	location functional based on third moments of $(\cdot)$
$\mathbf{COV}(\cdot)$	variance-covariance matrix of $(\cdot)$
$\mathbf{COV}_4(\cdot)$	scatter matrix of fourth moments of $(\cdot)$
$\mathbf{e}_i$	p-variate vector that has at its ith position a 1 and otherwise zeros
$\mathbf{1}_p$	p-variate vector of ones
$\mathbf{I}_p$	p-variate identity matrix
$\mathbf{D}$	diagonal matrix
$diag(\mathbf{a})$	diagonal matrix with diagonal elements given in $\mathbf{a}$
$diag(\mathbf{A})$	vector of the diagonal elements of matrix $\mathbf{A}$
$\mathbf{P}$	permutation matrix (obtained by permuting the rows or columns of $\mathbf{I}_p$)
$\mathbf{J}$	sign change matrix (diagonal matrix with entries ± 1)
$\mathbf{O}$	orthogonal matrix
$\|\| \cdot \|\|$	vector norm of $\cdot$
$\|\| \cdot \|\|_k$	L_k norm of $\cdot$
$sgn(\cdot)$	sign of $(\cdot)$
$\bigodot$	Hadamard product (entrywise)
ICA	independent component analysis
ICS	invariant coordinate system
PCA	principal component analysis
df	degrees of freedom
i.i.d.	independent and identically distributed
$A \leftarrow B$	A is replaced by B
$Multin(n, \boldsymbol{\pi})$	Multinominal distribution with parameters n and $\boldsymbol{\pi}$

List of original publications

I. Nordhausen, K., Oja, H. and Ollila, E. (2008). "Multivariate models and the first four moments", (submitted to the Festschrift for Thomas P. Hettmansperger, edited by Hunter, D.R., Rosenberger, J.L. and Richards, D.).

II. Nordhausen, K., Oja, H. and Tyler, D.E. (2008). "Tools for exploring multivariate data: The package ICS", *Journal of Statistical Software*, 28, 1–31.

III. Nordhausen, K., Oja, H. and Ollila, E. (2008). "Robust independent component analysis based on two scatter matrices", *Austrian Journal of Statistics*, 37, 91–100.

IV. Nordhausen, K., Oja, H. and Tyler, D.E. (2006). "On the efficiency of invariant multivariate sign and rank tests". In Liski, E.P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G.P.H. (editors), "Festschrift for Tarmo Pukkila on his 60th birthday", 217–231, University of Tampere, Tampere, Finland.

V. Nordhausen, K., Oja, H. and Paindaveine, D. (2008). "Signed-rank tests for location in the symmetric independent component model", *Journal of Multivariate Analysis* (accepted).

1 Introduction

In multivariate data analysis several variables are observed for each experimental unit and the dependence structure between the different variables is considered relevant. Usually the analysis starts with some exploratory methods and a description of the data, like scatter plots and computing measures of central tendency and dispersion. Further steps consist often in making inference for example about the location or shape of the data, model building using regression methods or trying to find groups using classification (supervised or unsupervised). Especially for these further steps model assumptions must be made. In the classical multivariate analysis this means that one assumes normality of the residuals. This assumption has the advantage that the optimal methods under this assumption are tractable and relatively easy to apply and it is often justified by the central limit theorem which states that under general assumptions the distribution of the sum of random vectors converges to a multivariate normal distribution as the number of observations increases (see for example Morrison, 1998b). In reality, however, the assumption of normality is seldom met, not all observations are well behaving and the number of observations is often small.

In practise the analyst often tries to enforce normality by transforming the variables, for example by taking logarithms or so. This might make the interpretation of the model parameters more difficult. In research on the other side that led to investigations on how robust actually the classical methods are. Robustness in this context means that methods should not be sensitive to violations of the assumptions. An overview of such findings can for example be found in Krzanowski (1998) which shows for instance that the nominal level of the one sample Hotelling's T^2 tests suffers more from the skewness than from the kurtosis of the background distribution. However the normal model does not offer the possibility to incorporate different skewness or kurtosis values. The normal distribution is fully specified by its first two moments. As solutions to this dilemma robust methods are developed which often are based on replacing the mean vector and covariance matrix, the main tools of the classical methods, by more robust measures of location and scatter. Another approach extends the normal model to a more general semiparametric model or nonparametric model and develops methods which are valid under these more general assumptions.

It is often hoped that the results of a multivariate data analysis do not depend on the chosen coordinate system. A change of the coordinate system can be expressed as an affine transformation of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b},$$

where $\mathbf{x}$ is the point in the original coordinate system, $\mathbf{A}$ is a full rank $p \times p$ transformation matrix, $\mathbf{b}$ is a p-variate vector and $\mathbf{y}$ is the point given in the new

coordinate system. In this transformation $\mathbf{b}$ shifts the origin and $\mathbf{A}$ transforms the axes. The transformation of the axes can be better re-enacted using the singular value decomposition

$$\mathbf{A} = \mathbf{O}^*\mathbf{D}\mathbf{O}^T,$$

where $\mathbf{O}^*$ and $\mathbf{O}$ are orthogonal matrices and $\mathbf{D}$ is a diagonal matrix. When changing now the axes, $\mathbf{x}$ is first rotated / reflected using $\mathbf{O}^T$, then componentwise rescaled using $\mathbf{D}$ and finally rotated / reflected by $\mathbf{O}^*$. For example rescaling and shifting the origin is needed when converting temperatures measured in Fahrenheit into Centigrade. Estimates that follow a change in the coordinate system in the appropriate way are called affine equivariant and tests that do not depend on the coordinate system are called affine invariant.

The different approaches to generalize and/or robustify the classical methods have led to a huge body of literature (to name only a few textbooks, see for example Fang and Zhang (1990), Genton (2004), Hampel, Ronchetti, Rousseeuw and Stahel (1986), Hettmansperger and McKean (1998) Huber (1980) or Maronna, Martin and Yohai (2006)) with different families of location and dispersion measures and several alternative semiparametric and nonparametric models. The data analyst is now left with the decision to choose an appropriate model and appropriate analysis tools for the data at hand.

The structure of this thesis is as follows. In the next chapter different extensions of the classical normal model are discussed. In Chapter 3, location and scatter measures are defined in a more formal way and it is discussed when they refer to the same population quantities. Furthermore it is described how location and scatter functionals are used for data transformations. Usually only one location and one scatter measure are used at the time. Chapter 4 shows that a simultaneous usage of two location and two scatter measures can be very informative and may help to distinguish between the different models. The next two chapters consider in detail two applications of the joint usage of two scatter matrices - first in the context of independence component analysis and then in the nonparametric location problem. In the last chapter the theory is illustrated by an analysis of a hemodynamic data set.

2 Multivariate models

All models discussed in this thesis can be derived from the location scatter model

$$\mathbf{x} = \boldsymbol{\Omega}\boldsymbol{\epsilon} + \boldsymbol{\mu},$$

where $\mathbf{x} = (x_1, \ldots, x_p)^T$ is a p-variate random vector, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^T$ is a p-variate random vector standardized in a way explained later, $\boldsymbol{\Omega}$ a full rank $p \times p$ mixing matrix and $\boldsymbol{\mu}$ a p-variate location vector. The quantity $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}^T$ is the scatter matrix parameter. For further analysis, as mentioned above, the assumptions imposed on $\boldsymbol{\epsilon}$ are crucial. Note however, that the standardized vector $\boldsymbol{\epsilon}$ is actually not observed, only $\mathbf{x}$ is directly measureable. The vector $\boldsymbol{\epsilon}$ is rather a mental construction than something with a physical meaning. Yet, in some cases, $\boldsymbol{\epsilon}$ can have an interpretation and it might even be the goal of the analysis to recover it when only $\mathbf{x}$ is observed. Either way, one of the first challenges one faces in practical data analysis is to evaluate which assumptions on $\boldsymbol{\epsilon}$ can be justified for the data at hand.

The following eight models will be considered in more detail and they all differ by their assumptions on $\boldsymbol{\epsilon}$.

A1: Multivariate normal model. $\boldsymbol{\epsilon}$ has a standard multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_p)$.

A2: Elliptic model. $\boldsymbol{\epsilon}$ has a spherical distribution around the origin, i.e. $\mathbf{O}\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$ for all orthogonal $p \times p$ matrices $\mathbf{O}$.

A3: Exchangeable sign-symmetric model. In this model $\boldsymbol{\epsilon}$ is symmetric around the origin in the sense that $\mathbf{P}\mathbf{J}\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$ for all permutation matrices $\mathbf{P}$ and sign change matrices $\mathbf{J}$.

A4: Sign-symmetric model. $\boldsymbol{\epsilon}$ is symmetric around the origin in the sense that $\mathbf{J}\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$ for all sign change matrices $\mathbf{J}$.

A5: Central symmetric model. $\boldsymbol{\epsilon}$ is symmetric around the origin in the sense that $-\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$.

B1: Finite mixtures of elliptical distributions with proportional scatter matrices. For a fixed k is $\boldsymbol{\epsilon} = \sum_{i=1}^{k} p_i(\boldsymbol{\epsilon}_i + \boldsymbol{\mu}_i)$, where $\mathbf{p} = (p_1, \ldots, p_k)$ is $Multin(1, \boldsymbol{\pi})$ distributed with $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{k} \pi_i = 1$ and $\boldsymbol{\epsilon}_i$'s are all independent and follow **A2**.

B2: Skew-elliptical model. $\boldsymbol{\epsilon} = sgn(\epsilon^*_{p+1} - \alpha - \beta\epsilon^*_p)\boldsymbol{\epsilon}^*$, where $({\boldsymbol{\epsilon}^*}^T, \epsilon^*_{p+1})^T$ satisfies **A2**, but with dimension $p+1$, and $\alpha, \beta \in \mathbb{R}$ are constants.

B3: Independent component model. The components of $\boldsymbol{\epsilon}$ are independent with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = 1$.

The models $\mathbf{A1} - \mathbf{A5}$ are models that are symmetric whereas the distributions in the models $\mathbf{B1} - \mathbf{B3}$ may be asymmetric. The symmetric models $\mathbf{A1} - \mathbf{A5}$ are ranked from the strongest symmetry assumption to the weakest one, which means

$$\mathbf{A1} \subset \mathbf{A2} \subset \mathbf{A3} \subset \mathbf{A4} \subset \mathbf{A5}.$$

It is also easy to see that the symmetric models can be seen as border cases of the asymmetric models and we have the following relationships:

- $\mathbf{A1} = \mathbf{A2} \cap \mathbf{B3}$
- $\mathbf{A2} \subset \mathbf{B1}$ and $\mathbf{A2} \subset \mathbf{B2}$

Note that the different symmetry concepts of the models **A1**-**A5** do not cover all concepts of multivariate symmetry found in Serfling (2006), for example. For instance angular symmetry, which is defined as

$$\frac{\boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||_2} \sim -\frac{\boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||_2},$$

is not included in this list of models. Likewise also directional elliptical symmetry (Randles, 1989) is not included. This symmetry concept assumes

$$\frac{\boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||_2} \sim \frac{\mathbf{O}\boldsymbol{\epsilon}}{||\boldsymbol{\epsilon}||_2}.$$

The model with directional elliptic symmetry can be seen as an extension of the model **A2** and the models with angular symmetry yields an extension of model **A5**.

In the following we describe the models **A1**-**A5** and **B1**-**B3** in a bit more detail.

2.1 Multivariate normal model

The multivariate normal model is the classical model in multivariate analysis. It is the only fully parametric model in our list and has the density

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}||\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})||_2^2\}.$$

The $N(\mathbf{0}, \mathbf{I}_p)$ distribution is the only spherical distribution having independent components. The multivariate normal model is symmetric around $\boldsymbol{\mu}$. All marginal distributions are univariate normal and therefore have the same kurtosis value, the classical kurtosis measure is

$$\beta_2(x_i) = \frac{E((x_i - E(x_i))^4)}{(Var(x_i))^2} = 3.$$

2.2 Elliptical model

The density f of an elliptically distributed vector $\mathbf{x}$ is given by

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\rho(||\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})||_2^2)\},$$

where $\rho(\cdot)$ is a function independent of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

It is obvious that the multivariate normal distribution is a member of this model with $\rho(t) = 1/2t^2 + p/2\log(2\pi)$. Other prominent distributions in the elliptical model are the multivariate t-distribution (e.g. Kotz and Nadarajah, 2004) and the power-exponential distribution (Gomez, Gomez-Villegas and Marin, 1998). This model extends the normal model by allowing also lighter or heavier tails while still requiring that all marginal distributions are similar in shape. The center of symmetry is $\boldsymbol{\mu}$. Given the first two moments exist,

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = c_\rho \boldsymbol{\Sigma},$$

where c_ρ is a constant depending on ρ. In the multivariate normal case, for example, $c_\rho = 1$ and in the t_ν case $c_\rho = \nu/(\nu - 2)$.

This model is in practice the most common extension of the multivariate normal model and the standard multivariate methods have been extended to this model (see for example Fang and Zhang, 1990). Also robust estimation techniques often assume this model.

2.3 Exchangeable sign-symmetric model

This model is similar to the elliptic model but a broader range of shapes is possible. In this model the margins of $\boldsymbol{\epsilon}$ are exchangeable and symmetric around $\mathbf{0}$. For densities of $\boldsymbol{\epsilon}$ it holds that

$$f(\boldsymbol{\epsilon}) = f(\mathbf{JP}\boldsymbol{\epsilon}).$$

An example for densities with this property is the class

$$f(\boldsymbol{\epsilon}) = c_\rho \exp(-\rho(||\boldsymbol{\epsilon}||)),$$

where c_ρ is a normalizing constant depending on the radial function ρ. Any norm $||\cdot||$ which fulfills the condition $||\boldsymbol{\epsilon}|| = ||\mathbf{PJ}\boldsymbol{\epsilon}||$ can be used, as for example any L_p-norm.

Given the first two moments exist

$$\mathbf{E}(\boldsymbol{\epsilon}) = \mathbf{0} \quad \text{and} \quad \mathbf{COV}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_p,$$

and therefore

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = \sigma^2 \boldsymbol{\Sigma}.$$

The margins of $\boldsymbol{\epsilon}$ are uncorrelated but may be dependent and have all the same scale.

2.4 Sign-symmetric model

In the sign-symmetric model the components of $\boldsymbol{\epsilon}$ are again uncorrelated. They can have different scales however. This is the main difference compared with the previous model. Again $\boldsymbol{\mu}$ is the center of symmetry of the distribution of $\mathbf{x}$, and given the existence of the first two moments

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = \boldsymbol{\Omega}\ diag(\sigma_1^2, \ldots, \sigma_p^2)\ \boldsymbol{\Omega}^T,$$

where σ_i^2 is the variance of the i-th component of $\boldsymbol{\epsilon}$.

2.5 Central symmetric model

This kind of symmetry is often also called "reflective", "diagonal", "simple" or "antipodal" symmetry (see for example Serfling, 2006). For the density of $\mathbf{x}$ in this case it holds that

$$f(\mathbf{x} - \boldsymbol{\mu}) = f(\boldsymbol{\mu} - \mathbf{x}).$$

The central symmetry is the most direct analog of the univariate concept of symmetry (Serfling, 2006). There are no restrictions on the covariance structure of $\boldsymbol{\epsilon}$ and given the first two moments exist

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = \boldsymbol{\Omega}\ \mathbf{COV}(\boldsymbol{\epsilon})\ \boldsymbol{\Omega}^T.$$

2.6 Finite mixtures of elliptical distributions

In this model $\boldsymbol{\epsilon}$ is a mixture of k spherical populations which can have different symmetry centers and different scales. Assuming their existence, the first two moments of the different mixture populations are

$$\mathbf{E}(\boldsymbol{\epsilon}_i) = \boldsymbol{\mu}_i \quad \text{and} \quad \mathbf{COV}(\boldsymbol{\epsilon}_i) = \tau_i^2 \mathbf{I}_p,$$

$i = 1, \ldots, k$. Therefore

$$\mathbf{E}(\boldsymbol{\epsilon}) = \sum_{i=1}^{k} \pi_i \boldsymbol{\mu}_i$$

and

$$\mathbf{COV}(\boldsymbol{\epsilon}) = \left(\sum_{i=1}^{k} \pi_i \tau_i^2 \right) \mathbf{I}_p + \sum_{i=1}^{k} \pi_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T - \sum_{i=1}^{k} \sum_{j=1}^{k} \pi_i \pi_j \boldsymbol{\mu}_i \boldsymbol{\mu}_j^T.$$

The location $\boldsymbol{\mu}$ will not any longer be the center of symmetry of $\mathbf{x}$ and

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\Omega} \sum_{i=1}^{k} \pi_i \boldsymbol{\mu}_i + \boldsymbol{\mu} \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = \boldsymbol{\Omega}\mathbf{COV}(\boldsymbol{\epsilon})\boldsymbol{\Omega}^T.$$

The distributions in this model can be symmetric or skew depending on the means $\boldsymbol{\mu}_i$, the proportions π_i and the distributions of the $\boldsymbol{\epsilon}_i$'s. For example, if $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k$ then $\mathbf{x}$ is elliptically symmetric.

2.7 Skew-elliptical model

The skew-elliptical model was proposed quite recently and, in this approach, there are alternative ways to introduce skewness into the elliptical model. For a recent overview see Genton (2004).

The interpretation closest to our definition comes from a hidden truncation model. The "true" population distribution is an elliptic distribution but the sampling procedure is selective so that an observation is sampled only if its value of an hidden variable exceeds a threshold value. Conscripts, for example, do not represent the whole male population of their age group but only those considered fit enough to bear arms.

If $\alpha = 0$ and the population distribution is multivariate normal, then one obtains the canonical form of the skew-normal distribution of Azzalini and Capitanio (1999). In most skew-elliptical model definitions it is assumed that $\alpha = 0$ because that simplifies the models considerably (the sign change probability is then 0.5).

The moments of $\boldsymbol{\epsilon}$,if they exist, depend on the elliptical density, α and β. The expressions are quite complex. The first two moments of $\boldsymbol{\epsilon}$ for the skew-normal distributions with $\alpha = 0$ are, for example,

$$\mathbf{E}(\boldsymbol{\epsilon}) = \left(0, \ldots, 0, \sqrt{\frac{2}{\pi}} \frac{-\beta}{\sqrt{1+\beta^2}}\right)^T$$

and

$$\mathbf{COV}(\boldsymbol{\epsilon}) = diag\left(1, \ldots, 1, 1 - \frac{2}{\pi}\frac{\beta^2}{1+\beta^2}\right).$$

The corresponding moments of $\mathbf{x}$ are

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\Omega}\mathbf{E}(\boldsymbol{\epsilon}) + \boldsymbol{\mu} \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = \boldsymbol{\Omega}\mathbf{COV}(\boldsymbol{\epsilon})\boldsymbol{\Omega}^T.$$

A natural extension of this model would be given by allowing truncations in several different directions. This will however not be considered in this paper. For a general discussion about skew models related to a hidden truncation, see Arnorld and Beaver (2002, 2004).

2.8 Independent component model

The independent component model is often used in signal processing or in image analysis applications. It is a rather flexible model with possibly asymmetric distributions. The model is however ill specified, since for any permutation matrix $\mathbf{P}$ and any sign change matrix $\mathbf{J}$

$$\mathbf{x} = (\boldsymbol{\Omega}\mathbf{P}\mathbf{J})(\mathbf{J}\mathbf{P}^{-1}\boldsymbol{\epsilon}) = \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\epsilon}},$$

which means that the independent components do not have fixed signs and that their order is also arbitrary.

By definition

$$\mathbf{E}(\boldsymbol{\epsilon}) = \mathbf{0} \quad \text{and} \quad \mathbf{COV}(\boldsymbol{\epsilon}) = \mathbf{I}_p$$

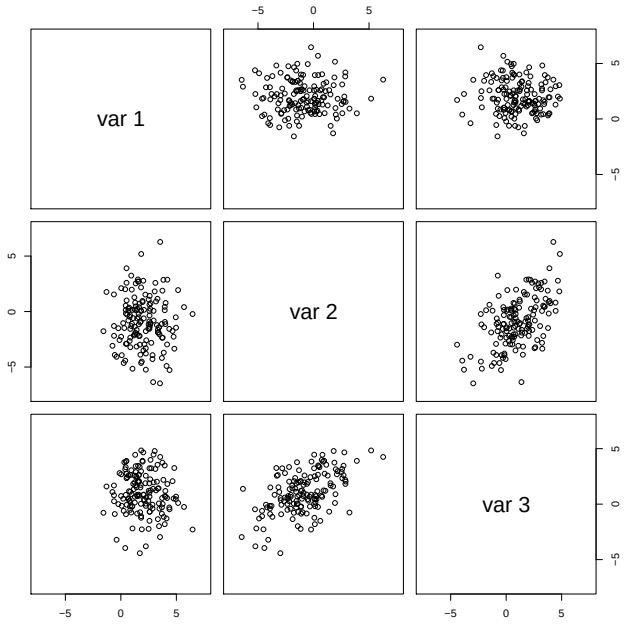

Figure 2.1: Pairwise scatter plots for a sample of size 150 following model **A1**.

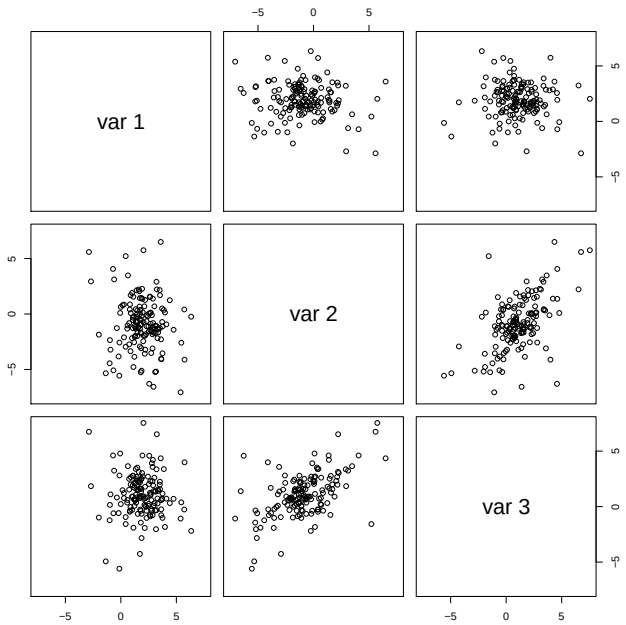

Figure 2.2: Pairwise scatter plots for a sample of size 150 following model **A2**.

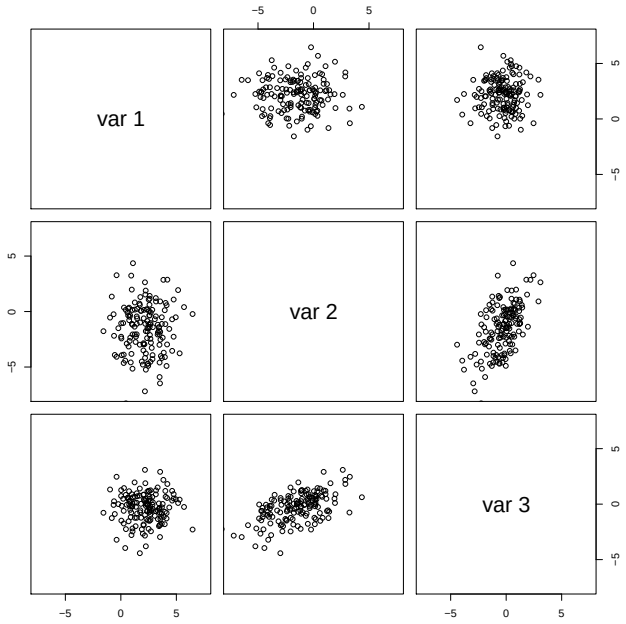

Figure 2.3: Pairwise scatter plots for a sample of size 150 following model **B2**.

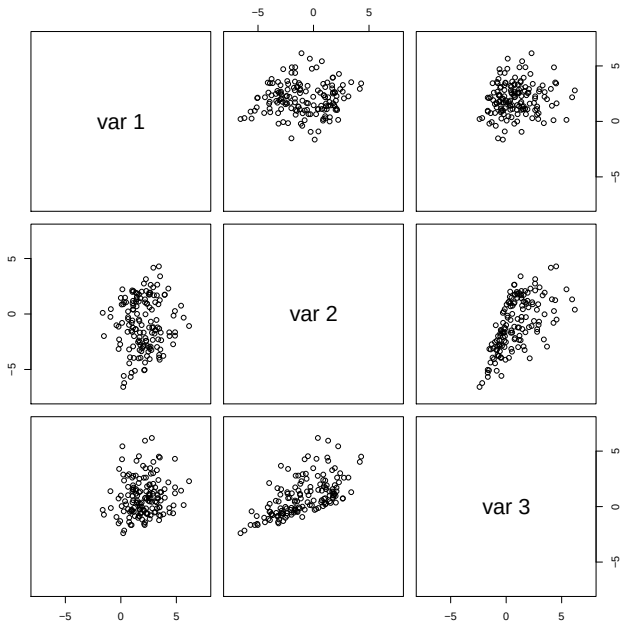

Figure 2.4: Pairwise scatter plots for a sample of size 150 following model **B3**.

and therefore the corresponding moments of $\mathbf{x}$ are

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\mu} \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = \boldsymbol{\Omega}\boldsymbol{\Omega}^T.$$

For a recent overview of this model see Hyvärinen, Karhunen and Oja (2001).

To demonstrate how much the shape of the data can vary just by changing the distribution of $\boldsymbol{\epsilon}$ four 3-variate samples of size 150 are shown in Figures 2.1 - 2.4. In the first case $\boldsymbol{\epsilon}$ is $N(\mathbf{0}, \mathbf{I}_3)$ distributed (model **A1**), in the second case $\boldsymbol{\epsilon}$ follows a t_5 distribution rescaled so that $\mathbf{COV}(\boldsymbol{\epsilon}) = \mathbf{I}_3$ (model **A2**), the third case is a skew-normal distribution with $\alpha = 0$ and $\beta = 3$ (model **B2**) and the last case is an independent component model where the three marginals have standardized normal, uniform and exponential distributions (model **B3**). In all cases the location parameter and the mixing matrix are

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega} = \begin{pmatrix} 1.5 & -0.5 & -0.3 \\ 1.0 & 2.0 & 0.5 \\ 0.5 & 0.5 & 1.7 \end{pmatrix}.$$

Note that the samples in Figure 2.1, Figure 2.2 and Figure 2.4 have the same first two theoretic moments. Yet, their scatter plots differ considerably. The data coming from the t_5 distribution have visibly much heavier tails than the data coming from the normal model. Both samples though have similar elliptically symmetric contours.

3 Location, scatter and their usage

The multivariate normal distribution is fully specified by its first two moments. It is then sufficient to base the analysis of the data only on the mean vector and the covariance matrix. Although the other models discussed in Chapter 2 offer the incorporation of more features like skewness or kurtosis, the inference in multivariate analysis is usually still based on location and scatter statistics.

In the following we define location and scatter functionals more carefully.

3.1 Location and scatter functionals

Let $\mathbf{x}$ be a p-variate random variable with cdf F. A vector valued functional $\mathbf{T}(F)$ or $\mathbf{T}(\mathbf{x})$ is a location functional if it is affine equivariant in the sense that

$$\mathbf{T}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\mathbf{T}(\mathbf{x}) + \mathbf{b}$$

for all full rank $p \times p$ matrices $\mathbf{A}$ and all p-variate vectors $\mathbf{b}$.

A matrix valued functional $\mathbf{S}(F)$ or $\mathbf{S}(\mathbf{x})$ is a scatter matrix if it is affine equivariant in the sense that

$$\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\mathbf{S}(\mathbf{x})\mathbf{A}^T$$

with $\mathbf{A}$ and $\mathbf{b}$ as defined above.

Location and scatter functionals are thus defined in such a way that they change in a logical way when the coordinate system is altered.

Location and scatter statistics, the finite sample versions of location and scatter functionals, must fulfill similar affine equivariance conditions. They will be denoted accordingly $\mathbf{T}(F_n)$ or $\mathbf{T}(\mathbf{X})$, respectively $\mathbf{S}(F_n)$ or $\mathbf{S}(\mathbf{X})$, where F_n is the empirical cdf of the $p \times n$ data matrix $\mathbf{X}$. Note that in the remainder of this chapter everything will be discussed only at the population level.

The classical location functional is the mean vector $\mathbf{E}(\mathbf{x})$ and the classical scatter functional the covariance matrix

$$\mathbf{COV}(\mathbf{x}) = \mathbf{E}((\mathbf{x} - \mathbf{E}(\mathbf{x}))(\mathbf{x} - \mathbf{E}(\mathbf{x}))^T).$$

There exist however a large number of different general techniques to construct other location and scatter functionals, such like M-estimates (Maronna, 1976), S-estimates (Davies, 1987) CM-estimates (Kent and Tyler, 1996), τ-estimates (Lopuhaä, 1991) and many more. For a recent overview see for example Maronna et al. (2006).

In this thesis only M-estimates will be discussed further. However, prior to that, some variations of scatter functionals need to be defined.

3.1.1 Special scatter matrices

Several scatter functionals actually do not achieve affine equivariance in the sense of

$$\mathbf{S}(\mathbf{A}\mathbf{x}+\mathbf{b}) = \mathbf{A}\mathbf{S}(\mathbf{x})\mathbf{A}^T,$$

but only in the sense of

$$\mathbf{S}(\mathbf{A}\mathbf{x}+\mathbf{b}) \propto \mathbf{A}\mathbf{S}(\mathbf{x})\mathbf{A}^T.$$

In that case they are usually called shape matrices. In many applications this form of equivariance is sufficient. To make different shape matrices comparable they are often normalized so that, for example, $tr(\mathbf{S}(\mathbf{x})) = p$ or $|\mathbf{S}(\mathbf{x})| = 1$. For further details about normalization methods of shape matrices, see Paindaveine (2008).

An important class of scatter functionals consists of those functionals that have the so called independence property (Oja et al., 2006). The independence property states that if $\mathbf{x}$ has independent components, then $\mathbf{S}(\mathbf{x})$ will be a diagonal matrix.

Most scatter functionals found in the literature do not have the independence property. For any scatter functional $\mathbf{S}$, a symmetrized version can be constructed as

$$\mathbf{S}_{sym}(\mathbf{x}) := \mathbf{S}(\mathbf{x}_1 - \mathbf{x}_2),$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent copies of $\mathbf{x}$. All symmetrized scatter functionals have the desired independence property (Oja et al., 2006).

3.1.2 M-estimators of location and scatter

The family of multivariate M-estimates was introduced by Maronna (1976) and the location and scatter functionals are usually estimated jointly. They are given as the simultaneous solution of the following two implicit equations

$$\mathbf{T}(\mathbf{x}) = E(w_1(r))^{-1}\mathbf{E}(w_1(r)\mathbf{x})$$

and

$$\mathbf{S}(\mathbf{x}) = \mathbf{E}(w_2(r)(\mathbf{x}-\mathbf{T}(\mathbf{x}))(\mathbf{x}-\mathbf{T}(\mathbf{x}))^T),$$

where $w_1(r)$ and $w_2(r)$ are nonnegative and continuous functions of the Mahalanobis distance $r = ||\mathbf{S}(\mathbf{x})^{-1/2}(\mathbf{x}-\mathbf{T}(x))||_2$. To obtain robust functionals usually the weight functions are chosen to be also nonincreasing.

The mean vector and the regular covariance matrix are M-estimators with $w_1(r) = w_2(r) = 1$. Other prominent members in this class are Huber's M-estimators (Huber, 1964) with the weight functions

$$w_1(r) = \begin{cases} 1 & r \le c \\ c/r & r > c \end{cases} \quad \text{and} \quad w_2(r) = \begin{cases} 1/\sigma^2 & r \le c \\ c/(r^2\sigma^2) & r > c \end{cases},$$

where c is a tuning constant chosen to satisfy $q = Pr(\chi^2_p \le c^2)$ and σ^2 is a scaling factor such that $E(\chi^2_p w_2(\chi_p)) = p$. Tyler's shape matrix (Tyler, 1987), which has $w_2(r) = p/r^2$ and is computed with respect to a given location functional $\mathbf{T}(\mathbf{x})$. A joint estimation of the affine equivariant spatial median with Tyler's shape matrix is obtained by using $w_1(r) = 1/r$ and $w_2(r) = p/r^2$ (Hettmansperger and Randles, 2002). This pair of estimates is denoted by $(\mathbf{T}_{HR}, \mathbf{S}_{HR})$. The weights

$w_1(r) = w_2(r) = (p+\nu)/(r^2+\nu)$ correspond to a M-estimator derived as the maximum likelihood solution from a t-distribution with $\nu \geq 1$ degrees of freedom and is described in Kent and Tyler (1991).

An important family of scatter functionals consists of the so called one-step M-functionals. Given a pair of location and scatter functionals $(\mathbf{T}_1, \mathbf{S}_1)$ and two weight functions $w_1(r)$ and $w_2(r)$ the one-step functionals are

$$\mathbf{T}_2(\mathbf{x}) = E(w_1(r_1))^{-1}\mathbf{E}(w_1(r_1)\mathbf{x})$$

and

$$\mathbf{S}_2(\mathbf{x}) = \mathbf{E}(w_2(r_1)(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))^T),$$

where $r_1 = ||\mathbf{S}_1(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))||_2$.

An interesting special case starts with the pair $(\mathbf{E}, \mathbf{COV})$ and uses the weight functions $w_1(r) = r^2$ and $w_2(r) = r^2/(p+2)$. The resulting estimates $\mathbf{T}_2$ and $\mathbf{S}_2$ are consistent at the multivariate normal model. They are called the vector of third moments $\mathbf{E}_3$ and the matrix of fourth moments $\mathbf{COV}_4$ and are given by

$$\mathbf{E}_3(\mathbf{x}) = \frac{1}{p}\mathbf{E}(r^2\mathbf{x}) \quad \text{and} \quad \mathbf{COV}_4(\mathbf{x}) = \frac{1}{p+2}\mathbf{E}(r^2(\mathbf{x} - \mathbf{E}(\mathbf{x}))(\mathbf{x} - \mathbf{E}(\mathbf{x}))^T).$$

A shape matrix used later in this thesis is the one-step normal score signed-rank scatter matrix of Hallin and Paindaveine (2006) which has the form

$$\mathbf{S}_{HP}(\mathbf{x}) = \mathbf{S}_{HR}^{1/2}(\mathbf{x})\mathbf{E}\left(\psi_p^{-1}(F_{||\mathbf{z}||_2}(||\mathbf{z}||_2))\frac{\mathbf{z}\mathbf{z}^T}{||\mathbf{z}||}\right)\mathbf{S}_{HR}^{1/2}(\mathbf{x}),$$

where the starting pair $(\mathbf{T}_{HR}, \mathbf{S}_{HR})$ is the Hettmansperger and Randles (2002) estimate, $\mathbf{z} = \mathbf{S}_{HR}^{-1/2}(\mathbf{x} - \mathbf{T}_{HR}(\mathbf{x}))$ and ψ_p denotes the cdf of a chi-square distribution with p degrees of freedom. This functional needs no moment assumptions and has a strong nonparametric nature.

In general M-estimator scatter functionals do not posses the independence property. Among the M-estimators mentioned so far only the regular covariance matrix and the matrix of fourth moments do. Symmetrized M-estimators of scatter with the independence property are described in Sirkiä, Taskinen and Oja (2007). The symmetrized version of Tyler's shape matrix is known also as Dümbgen's shape matrix (Dümbgen, 1998).

3.2 Location and scatter for data transformation

Besides just describing the central tendency or the dispersion of the data location and scatter functionals can also be used to transform the data. The goal then usually is to obtain a coordinate system that has nice mathematical properties or that highlights the features of interest. The transformation is traditionally based on the mean vector and the covariance matrix.

3.2.1 Whitening

The whitening transformation is a basic data transformation. It subtracts the mean vector to move the location center to the origin, rotates the data to jointly

uncorrelate the marginal variables and finally rescales the marginal variables to have unit variances. Formally this transformation can be described as

$$\mathbf{y} = \mathbf{COV}^{-\frac{1}{2}}(\mathbf{x})(\mathbf{x} - \mathbf{E}(\mathbf{x})),$$

where then

$$\mathbf{E}(\mathbf{y}) = \mathbf{0} \quad \text{and} \quad \mathbf{COV}(\mathbf{y}) = \mathbf{I}_p.$$

$\mathbf{COV}^{-\frac{1}{2}}$ denotes here the matrix square root of $\mathbf{COV}^{-1}$. In the new coordinate system of $\mathbf{y}$ no direction is more interesting than any other with respect to the variation. In the normal model (**A1**) whitening makes the coordinates independent whereas in the other models it only uncorrelates them. In the elliptic model (**A2**) $\mathbf{y}$ is spherical. Note that this transformation does not usually recover $\boldsymbol{\epsilon}$. This is related to the fact that the coordinate system obtained by whitening is not affine invariant and it only holds that

$$\mathbf{COV}^{-\frac{1}{2}}(\mathbf{Ax} + \mathbf{b})(\mathbf{Ax} + \mathbf{b} - \mathbf{E}(\mathbf{Ax} + \mathbf{b})) = \mathbf{O}\ \mathbf{COV}^{-\frac{1}{2}}(\mathbf{x})(\mathbf{x} - \mathbf{E}(\mathbf{x})),$$

for some orthogonal matrix $\mathbf{O}$. This means they might differ by a rotation which depends on the matrix $\mathbf{A}$, on the distribution of $\mathbf{x}$ and on the matrix square root of $\mathbf{COV}$ used. In this thesis matrix square roots are taken to be symmetric.

3.2.2 Principal component analysis

The principal component analysis also creates a coordinate system where the different coordinates are uncorrelated. The difference to whitening is however, that in PCA the variables in the different directions are not chosen to have unit variances. The aim is to create the new variables successively in such a way that they are the linear combinations that maximize the variation under the constraint of being orthogonal to the previous variables. This is obtained also by using the eigenvector-eigenvalue decomposition of the covariance matrix

$$\mathbf{COV}(\mathbf{x}) = \mathbf{ODO}^T,$$

where $\mathbf{D}$ is a diagonal matrix containing the ordered eigenvalues of $\mathbf{COV}(\mathbf{x})$ and $\mathbf{O}$ is orthogonal and the columns contain the corresponding eigenvectors. The new coordinates, also called the principal components, are then obtained as

$$\mathbf{y} = \mathbf{O}^T\mathbf{x},$$

and now $\mathbf{COV}(\mathbf{y}) = \mathbf{D}$.

There are often two transformations preliminary to the PCA. The first one is that $\mathbf{x}$ is centered, i.e. $\mathbf{x} \leftarrow \mathbf{x} - \mathbf{E}(\mathbf{x})$, which means that in the analysis $\mathbf{x}$ is replaced by $\mathbf{x} - \mathbf{E}(\mathbf{x})$. The other transformation scales the components, i.e. $x_i \leftarrow x_i/\sigma_i$, where σ_i^2 is the variance of x_i. This corresponds to perform the principal component analysis using the correlation matrix instead of the covariance matrix. Latter transformation is recommended when the components of $\mathbf{x}$ have completely different scales to give each component the same weight in the analysis. It is obvious that PCA, even with theses pretransformations, does not give an affine invariant coordinate system.

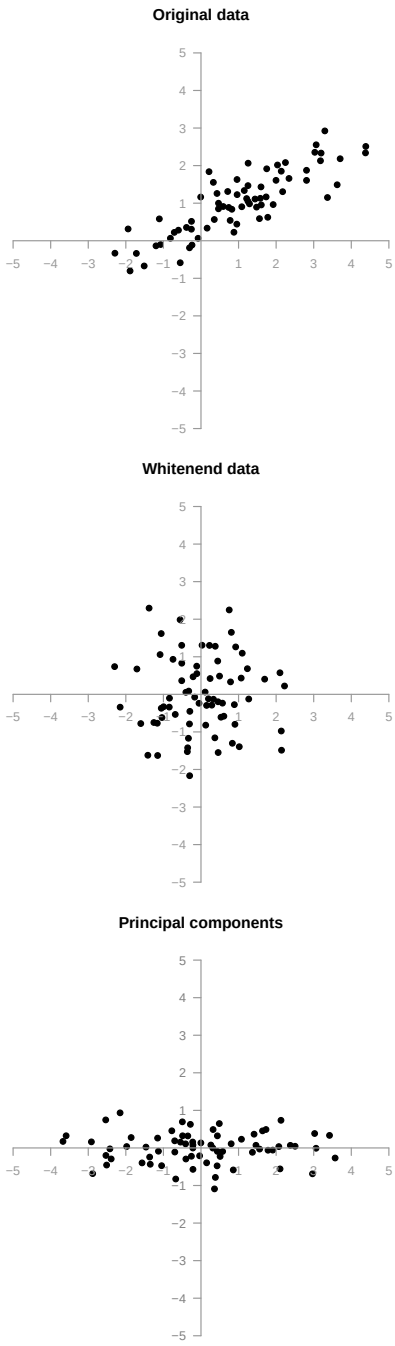

Figure 3.1: The top figure shows 70 observations from a bivariate normal distribution, the figure in the middle the whitened data and the bottom figure the principal components for the centered data.

The PC transformation has several nice geometrical properties as for example described in Jolliffe (2002). In general PCA does not make any distributional assumptions (except the existence of second moments) but has further interpretation possibilities in the multivariate normal case (Jolliffe, 2002).

The difference between whitening and PCA is demonstrated in Figure 3.1, where the two transformations are applied to a sample of 70 bivariate normal observations. The whitened data have along both axes the same variation whereas the first axis after the principal component transformation has a much larger variation than the second one.

3.2.3 Factor analysis

Factor analysis can be seen as a model based transformation where the underlying model is given by

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\zeta} + \boldsymbol{\upsilon} + \boldsymbol{\mu}$$

where the components of the observed p-variate random vector $\mathbf{x}$ are linear combinations of the latent $m(\leq p)$ components of $\boldsymbol{\zeta}$ added by some random noise $\boldsymbol{\upsilon}$. The location center is given by the p-vector $\boldsymbol{\mu}$. It is assumed that $\mathbf{E}(\boldsymbol{\zeta}) = \mathbf{0}$, $\mathbf{COV}(\boldsymbol{\zeta}) = \mathbf{I}_m$, $\mathbf{E}(\boldsymbol{\upsilon}) = \mathbf{0}$ and $\mathbf{COV}(\boldsymbol{\upsilon}) = \mathbf{D}$, and that $\boldsymbol{\zeta}$ and $\boldsymbol{\upsilon}$ are independent. The parameters in this model are however not well-defined and the $p \times m$ matrix $\mathbf{\Lambda}$ and $\boldsymbol{\zeta}$ are only defined up to a rotation. The aim of this model-based transformation is to find m latent variables that explain the dependence between the original variables.

There are several ways to estimate the new coordinate system and they all are based on the decomposition

$$\mathbf{COV}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{D}$$

given an value for m. For an overview see for example Anderson (2003). In some of the approaches multinormality of $\boldsymbol{\zeta}$ and $\boldsymbol{\upsilon}$ is assumed.

3.2.4 Canonical correlations

In the canonical correlation analysis the correlations between two subvectors $\mathbf{x}_1$ and $\mathbf{x}_2$ ($\mathbf{x} = (\mathbf{x}_1^T\ \mathbf{x}_2^T)^T$) are described in a simple way. One finds a new coordinate system that consists of two subsystems for $\mathbf{x}_1$ and $\mathbf{x}_2$ which explains the correlations between the two vectors in a canonic way. For the analysis, one uses a partition of $\mathbf{COV}(\mathbf{x}) = \mathbf{\Sigma}$ corresponding to $\mathbf{x}_1$ and $\mathbf{x}_2$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

Consider the vectors $\mathbf{a}$ and $\mathbf{b}$ satisfying

$$\mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a} = 1 \quad \text{and} \quad \mathbf{b}^T\mathbf{\Sigma}_{22}\mathbf{b} = 1$$

and then maximize $\mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b}$. Given the first solution one searches a second component that is uncorrelated to the previous one and so on. The solutions for $\mathbf{a}$ and $\mathbf{b}$ are the eigenvectors of

$$\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$$

and

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

respectively. For details, see for example Shubhabrata and Sen (1998).

3.2.5 Robust transformations

All the transformations considered so far are based on the regular covariance matrix which has the well-known disadvantages. In robust approaches the regular covariance matrix is replaced by some robust scatter functional. Robust PCA is described in Croux and Ruiz-Gazen (2005), robust factor analysis in Pison, Rousseeuw, Filzmoser and Croux (2003) and robust canonical correlation analysis in Taskinen, Croux, Kankainen, Ollila and Oja (2006). This replacement is basically not a problem if the scatter measures estimate the same population quantity; this is true in models **A1**-**A3** (see Tyler et al., 2008, who even conjecture that **A3** is the largest class with this property). In other models the interpretation of the results is difficult. It should be also noted, that the regular covariance matrix can not always just be replaced by another scatter functional (as we will see for example in Chapter 5), since sometimes some special properties of the regular covariance matrix are of importance. And as with the covariance matrix the same is also true with the mean vector - also the mean is often replaced without a question by a robust location functional. All location functionals yield the same population quantity, the center of symmetry, in the models **A1**-**A5**. The symmetry assumption which is very common in multivariate analysis can also be seen as an attempt to give the location a clear interpretation. In asymmetric models different location measures estimate different population quantities. A comparison of the values of two location measures and two scatter statistics is useful when describing skewness and kurtosis of multivariate data.

4 Simultaneous use of two location and/or two scatter functionals

In this chapter the simultaneous usage of two location and / or two scatter functionals is discussed. Note that in this thesis the interest lies only on applying the functionals to the same population, i.e., it is not of interest to compare the same functional computed for two different populations which has a long tradition in statistical analysis.

4.1 Early simultaneous usage of two different functionals

The simultaneous use of two location functionals $(\mathbf{T}_1, \mathbf{T}_2)$ and/or two scatter functionals $(\mathbf{S}_1, \mathbf{S}_2)$ does not have a long tradition in multivariate analysis. Whereas in univariate analysis already Karl Pearson saw the potential of the approach and measured the skewness of the data with the standardized difference of two location measures. Oja (1981) for example points out that the kurtosis of a random variable could be measured by the ratio of two scale measures. Also the classical univariate measures of skewness and kurtosis of a random variable x, namely

$$\beta_1(x) = \frac{E((x - E(x))^3)}{(Var(x))^{3/2}} \quad \text{and} \quad \beta_2(x) = \frac{E((x - E(x))^4)}{(Var(x))^2}$$

can be expressed in such ways. Setting

$$E_3(x) = E\left(\left(\frac{x - E(x)}{\sqrt{Var(x)}}\right)^2 x\right)$$

and

$$Var_4(x) = E\left(\left(\frac{x - E(x)}{\sqrt{Var(x)}}\right)^2 (x - E(x))^2\right),$$

then β_1 and β_2 can be rewritten as

$$\beta_1(x) = \frac{E_3(x) - E(x)}{\sqrt{Var(x)}} \quad \text{and} \quad \beta_2(x) = \frac{Var_4(x)}{Var(x)}.$$

However as Kotz, Balakrishnan and Johnson (2000) point out, so far using two functionals for measures of skewness and kurtosis has got less attention in the

multivariate case than in the univariate case. In the multivariate case Isogai (1982) uses the mean vector $\mathbf{T}_1$ and the multivariate mode $\mathbf{T}_2$ (not affine equivariant) to get a measure of skewness

$$Sk_{Isogai}(\mathbf{x}) = (\mathbf{T}_2(\mathbf{x}) - \mathbf{T}_1(\mathbf{x}))^T (w(\mathbf{COV}(\mathbf{x})))^{-1} (\mathbf{T}_2(\mathbf{x}) - \mathbf{T}_1(\mathbf{x})),$$

where $w(\cdot)$ is a matrix valued function for the covariance matrix, for example just the identity function. Oja (1983) suggests

$$Sk_{Oja}(\mathbf{x}) = (\mathbf{T}_2(\mathbf{x}) - \mathbf{T}_1(\mathbf{x}))^T \mathbf{COV}(\mathbf{x})^{-1} (\mathbf{T}_2(\mathbf{x}) - \mathbf{T}_1(\mathbf{x})),$$

where $\mathbf{T}_1$ is also the regular mean vector but $\mathbf{T}_2$ is the Oja median. In the same paper Oja defines also a kurtosis measure

$$Kurt_{Oja}(\mathbf{x}) = \frac{\mathbf{E}(\Delta(\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{E}(\mathbf{x})))^4}{(\mathbf{E}(\Delta(\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{E}(\mathbf{x})))^2)^2},$$

where $\mathbf{x}_i = (x_{i1} \ \ldots \ x_{ip})^T, \ i = 1, \ldots, p+1$, are independent copies of $\mathbf{x}$ and

$$\Delta(\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}) = abs \left(\begin{array}{c|cccc} \frac{1}{p!} & 1 & 1 & \cdots & 1 \\ & x_{11} & x_{11} & \cdots & x_{(p+1)1} \\ & x_{12} & x_{22} & \cdots & x_{(p+1)2} \\ & \vdots & \vdots & \vdots & \vdots \\ & x_{1p} & x_{2p} & \cdots & x_{(p+1)p} \end{array} \right)$$

is the volume of the simplex with vertices $\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}$. The joint usage of two scatter functionals can be found also in other areas of multivariate analysis. In the early applications only specific combinations of $(\mathbf{S}_1, \mathbf{S}_2)$ were considered. In cluster analysis and outlier detection, for example, two scatter functionals were used by Art, Gnanadesikan and Kettenring (1982), Yenyukov (1988), Caussinus and Ruiz (1990), Caussinus and Ruiz-Gazen (1993, 1995, 2006), Caussinus, Fekri, Hakam and Ruiz-Gazen (2003) or Critchley, Pires and Amado (2008). The approach of Caussinus and Ruiz-Gazen is known as the generalized principal component analysis (GPCA) and the approach of Critchley, Pires and Amado as the principal axis analysis (PAA). Furthermore in the independent component analysis (ICA) the so-called FOBI algorithm (Cardoso, 1989) can be seen as the joint usage of the regular covariance matrix ($\mathbf{COV}$) and the scatter matrix of 4th moments ($\mathbf{COV}_4$).

A general theory of the joint usage of location and / or scatter functionals has been developed quite recently. First Oja et al. (2006) investigated the general usage of two scatter matrices for real-valued ICA problems. Ollila, Oja and Koivunen (2008) extended then these results also to the complex data ICA problem. Kankainen, Taskinen and Oja (2007) developed a theory for testing multivariate normality using two location or two scatter functionals. The most general theory about the joint usage of two scatter functionals as a tool in multivariate analysis was given in Tyler et al. (2008). The approach was then the called invariant coordinate selection (ICS) and reconciled all the previously mentioned goals as well as found also further applications. The results in the following section are mainly due to Tyler et al. (2008).

4.2 Two scatter matrices and ICS

The main idea of ICS is to compare two scatter functionals $\mathbf{S}_1(\mathbf{x})$ and $\mathbf{S}_2(\mathbf{x})$ by solving $\mathbf{B}(\mathbf{x})$ and $\mathbf{D}(\mathbf{x})$ in the eigenvalue-eigenvector problem

$$\mathbf{S}_1^{-1}(\mathbf{x})\mathbf{S}_2(\mathbf{x})\mathbf{B}^T(\mathbf{x}) = \mathbf{B}^T(\mathbf{x})\mathbf{D}(\mathbf{x}).$$

$\mathbf{D}(\mathbf{x})$ is a diagonal matrix containing the p eigenvalues of $\mathbf{S}_1^{-1}(\mathbf{x})\mathbf{S}_2(\mathbf{x})$ and the rows of $\mathbf{B}(\mathbf{x})$ are the corresponding eigenvectors. For brevity, in the following denote $\mathbf{S}_1(\mathbf{x}) = \mathbf{S}_1$, $\mathbf{S}_2(\mathbf{x}) = \mathbf{S}_2$, $\mathbf{D}(\mathbf{x}) = \mathbf{D}$ and $\mathbf{B}(\mathbf{x}) = \mathbf{B}$.

It follows from this derivation of $\mathbf{B}$ that $\mathbf{B}$ jointly diagonalizes $\mathbf{S}_1$ and $\mathbf{S}_2$, i.e.,

$$\mathbf{B}\mathbf{S}_1\mathbf{B}^T = \mathbf{D}_1 \quad \text{and} \quad \mathbf{B}\mathbf{S}_2\mathbf{B}^T = \mathbf{D}_2,$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices with $\mathbf{D}_1^{-1}\mathbf{D}_2 = \mathbf{D}$.

Since the order, sign and length of the eigenvectors are not uniquely defined some conventions are needed. The following two conventions seem to be relevant in practise:

1. (a) Fixing the order of $\mathbf{D} \leftarrow \mathbf{P}\mathbf{D}$ and $\mathbf{B} \leftarrow \mathbf{P}\mathbf{B}$ so that $d_{11} \geq \ldots \geq d_{pp}$.
   (b) Rescaling $\mathbf{B} \leftarrow \mathbf{D}^*\mathbf{B}$, where $\mathbf{D}^*$ is a diagonal matrix such that $\mathbf{B}\mathbf{S}_1\mathbf{B}^T = \mathbf{I}_p$.
   (c) Fixing the sign of $\mathbf{B} \leftarrow \mathbf{J}\mathbf{B}$ so that $\mathbf{T}_1(\mathbf{B}\mathbf{x}) \geq \mathbf{T}_2(\mathbf{B}\mathbf{x})$ for two location functionals $\mathbf{T}_1$ and $\mathbf{T}_2$.
2. (a) Fixing the order of $\mathbf{D} \leftarrow \mathbf{P}\mathbf{D}$ and $\mathbf{B} \leftarrow \mathbf{P}\mathbf{B}$ so that $d_{11} \geq \ldots \geq d_{pp}$.
   (b) Rescaling $\mathbf{B} \leftarrow \mathbf{D}^*\mathbf{B}$, where $\mathbf{D}^*$ is a diagonal matrix such that $diag(\mathbf{B}^T\mathbf{B}) = \mathbf{I}_p$.
   (c) Fixing the sign of $\mathbf{B} \leftarrow \mathbf{J}\mathbf{B}$ so that, for $i = 1, \ldots, p$, $\max(b_{i1}, \ldots, b_{ip}) = \max(|b_{i1}|, \ldots, |b_{ip}|)$.

The first convention therefore fixes the length and sign of the eigenvectors based on the pairs $(\mathbf{T}_1, \mathbf{S}_1)$ and $(\mathbf{T}_2, \mathbf{S}_2)$ whereas in the second convention they do not depend on the functionals used. The second convention is more natural in the framework of ICA where different estimates of $\mathbf{B}$ might be compared. The convention used in the analysis does not change the interpretation of the eigenvectors as they only may have different scales and signs.

In the following we will assume, if not mentioned otherwise, that $\mathbf{B}$ is standardized according to the first method. The only ambiguities remaining are when some components of $\mathbf{B}\mathbf{x}$ are symmetric, and/or when $\mathbf{S}_1^{-1}\mathbf{S}_2$ has less than p distinct eigenvalues. For the reminding part of this section we will assume however that the eigenvalues are all distinct. For details about the case of non-distinct eigenvalues see Tyler et al. (2008).

This comparison of the two scatter matrices can also be interpreted as a combination of the two ways to uncorrelate a vector. In this case $\mathbf{x}$ is first uncorrelated using $\mathbf{S}_1^{-1/2}$, i.e. $\mathbf{y} = \mathbf{S}_1^{-1/2}(\mathbf{x})\mathbf{x}$ and then a principal component analysis is performed on $\mathbf{y}$ using $\mathbf{S}_2(\mathbf{y})$. Actually, the feature from which ICS got its name is then that this double decorrelation transformation yields an invariant coordinate system in the sense that

$$\mathbf{B}(\mathbf{x})\mathbf{x} = \mathbf{J}\mathbf{B}(\mathbf{A}\mathbf{x})(\mathbf{A}\mathbf{x}),$$

which means that the components differ at most by signs. The new coordinates $\mathbf{z} = \mathbf{B}(\mathbf{x})\mathbf{x}$ are therefore referred to as the invariant coordinates and $\mathbf{B}$ is the transformation matrix to the invariant coordinates.

Recalling the earlier statement that in the models **A1**-**A3** the scatter functionals $\mathbf{S}_1$ and $\mathbf{S}_2$ measure the same population quantity, it is obvious that in these models the PCA step with respect to $\mathbf{S}_2$ cannot find any directions with alternating variation. This can be seen as a way to explore whether $\mathbf{S}_1$ carries any additional information of the distribution in addition to $\mathbf{S}_2$.

A nice feature when comparing two scatter matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ in this way is that $\mathbf{S}_1^{-1}(\mathbf{x})\mathbf{S}_2(\mathbf{x})$ can be seen as a ratio of two scatter functionals and therefore as a measure of kurtosis. The connection to kurtosis becomes obvious when considering the univariate variable $y = \mathbf{a}^T\mathbf{x}$ with $\mathbf{a} \in \mathbb{R}^p$. Now $\mathbf{a}^T\mathbf{S}_1(\mathbf{x})\mathbf{a}$ is a measure of variation of y, or that of $\mathbf{x}$ in the direction $\mathbf{a}$ and similarly for $\mathbf{a}^T\mathbf{S}_2(\mathbf{x})\mathbf{a}$. Thus

$$\kappa(\mathbf{a}) = \frac{\mathbf{a}^T\mathbf{S}_2(\mathbf{x})\mathbf{a}}{\mathbf{a}^T\mathbf{S}_1(\mathbf{x})\mathbf{a}}$$

is the ratio of two different measures of variation in the direction of $\mathbf{a}$. Recall here the discussion in Chapter 4.1.

Tyler et al. (2008) point out, the maximum of $\kappa(\mathbf{a})$ is obtained if $\mathbf{a} = \mathbf{b}_1$ and the minimum is obtained if $\mathbf{a} = \mathbf{b}_p$, where $\mathbf{b}_i$ is the eigenvector corresponding to the eigenvalue d_{ii}, that is, the ith row of $\mathbf{B}$. Therefore the components of the invariant coordinates $\mathbf{z}$ are ordered according to their kurtosis values measured by $\mathbf{S}_1$ and $\mathbf{S}_2$. The values of these kurtosis measures correspond to the diagonal entries of $\mathbf{D}$. We refer to $\mathbf{D}$ as the generalized $\mathbf{S}_1 - \mathbf{S}_2$-kurtosis. The matrix $\mathbf{D}$ depends on the scaling of the scatter functionals. If both functionals are standardized under the multivariate normal model, $\mathbf{D}(\mathbf{x}) = \mathbf{I}_p$ if $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. However if they are standardized differently a direct comparison is difficult. Sometimes $\mathbf{D}$ is standardized so that $\prod_{i=1}^p d_{ii} = 1$.

Since two scatter functionals $\mathbf{S}_1$ and $\mathbf{S}_2$ typically measure different population quantities different choices of the pair $\mathbf{S}_1$ and $\mathbf{S}_2$ will yield different coordinate systems. The problem on how to choose $\mathbf{S}_1$ and $\mathbf{S}_2$ needs still to be investigated. It seems that there is no general best combination and the choice should depend on the purpose of the invariant coordinates and that the scatter functionals should have properties needed in further analysis (e.g. root-n consistency, asymptotic normality, the independence property, etc.).

Two rough guidelines that can be provided so far:

- Tyler et al. (2008) recommend to use two scatter matrices with a low breakdown point (like $\mathbf{COV}$ and $\mathbf{COV}_4$) only if there are no outlying observations or if the objective is to find outliers. Two scatter matrices with a high breakdown point (close to 1/2) on the other hand represent only the inner part of the data cloud. A good choice in their opinion is to use one medium robust scatter matrix (e.g. M-estimators with breakdown point $1/(p+1)$) and one scatter matrix with high breakdown point or to use two medium robust scatter matrices.

- Given a pair of scatter functionals, the roles of $\mathbf{S}_1$ and $\mathbf{S}_2$ can be interchanged since
$$\mathbf{S}_1^{-1}\mathbf{S}_2\mathbf{B}^T = \mathbf{B}^T\mathbf{D} \Leftrightarrow \mathbf{S}_2^{-1}\mathbf{S}_1\mathbf{B}^T = \mathbf{B}^T\mathbf{D}^{-1}.$$

This means the order has no effect on the transformation matrix $\mathbf{B}$, only the elements of $\mathbf{D}$ will be inverted and reversed.

But what is an ICS actually good for?

There are basically four uses of invariant coordinates considered so far : (i) Descriptive statistics and model selection; This is motivated by the idea that $\mathbf{D}$ provides measures of kurtosis and the two functionals $\mathbf{T}_1$ and $\mathbf{T}_2$ that were used to fix the signs of the rows of $\mathbf{B}$ provide a measure of skewness. These descriptive measures can then be used for model selection in the spirit of Pearson's system. This will be still discussed further in Section 4.3. (ii) ICS can be used in the independent component analysis (ICA) where the goal is to estimate $\mathbf{\Omega}$ assuming that the components of $\boldsymbol{\epsilon}$ are independent (model **B3**). In this case ICS is a generalization of Cardoso's FOBI algorithm (Cardoso, 1989). This will be discussed further in Chapter 5. Application (iii) is classification, outlier detection and dimension reduction; the idea behind this is that clusters or outliers might be found in the invariant coordinates with high or low kurtosis and for example $k < p$ invariant components might be chosen to be used in further analysis. This area of application is not carefully discussed in this thesis, but Paper II contains some examples. For further details, see Tyler et al. (2008). As pointed out earlier, affine equivariance is an important property for functionals. The invariance properties of tests under affine transformations is important as well, a test decision (the p-value) should not depend on the underlying coordinate system. Yet, there are for example nonparametrical tests and estimates which do not have this property. A solution is then to use invariant coordinates for the estimation and testing using the invariant coordinates - the estimates naturally have to be retransformed to the original coordinate system. This is our last area (iv) of application and will be discussed in Chapter 6.

We end this section with an example to compare the original coordinate system, the whitened coordinates, the principal components and the invariant coordinates. For this purpose 500 observations were sampled from model **B1** which in this case is a finite mixture of three 5-variate spherical normal distributions with different means and different variances. Figures 4.1 - 4.4 show the four different coordinates systems. Only the invariant coordinates find the underlying structure which is revealed in the first and last coordinate.

4.3 Multiple location and scatter functionals for descriptive data analysis and model selection

As mentioned above the diagonal elements of $\mathbf{D}$ can be seen as generalized kurtosis measures and together with two location functionals also measures of skewness are available. In this section we modify the general ICS by centering the components with respect to a scatter functional $\mathbf{T}_1$. We choose two pairs of functionals $(\mathbf{T}_1, \mathbf{S}_1)$ and $(\mathbf{T}_2, \mathbf{S}_2)$ and define the invariant coordinates as

$$\mathbf{z} = \mathbf{B}(\mathbf{x} - \mathbf{T}_1(\mathbf{x})).$$

The matrix $\mathbf{B}$ is now chosen to satisfy

$$\mathbf{T}_1(\mathbf{z}) = \mathbf{0}, \quad \mathbf{T}_2(\mathbf{z}) = \mathbf{s} \geq 0, \quad \mathbf{S}_1(\mathbf{z}) = \mathbf{I}_p \quad \text{and} \quad \mathbf{S}_2(\mathbf{z}) = \mathbf{D}.$$

The coordinate system is unique if $\mathbf{s} > \mathbf{0}$ and if the diagonal entries of $\mathbf{D}$ are distinct.

Based on these two pairs of functionals the distribution of $\mathbf{x}$ can be summarized having the following four descriptive functionals,

$$\begin{array}{ll} \textbf{Location: } \mathbf{T}_1(\mathbf{x}) & \textbf{Skewness: } \mathbf{T}_2(\mathbf{z}) \\ \textbf{Scatter: } \mathbf{S}_1(\mathbf{x}) & \textbf{Kurtosis: } \mathbf{S}_2(\mathbf{z}). \end{array}$$

The four descriptive measures given above can be used to get an impression of the data but furthermore also help to decide about a model that might fit the data best. Recall that in the well-known univariate Pearson system of distributions (see for example Ord, 1986) the choice of the model is based on the first four moments. This system is derived from a differential equation and β_1 and β_2 are used to choose an appropriate distribution for a sample at hand. Kotz (1975) noted however, that Pearson's system seems difficult to transfer to the multivariate case from a differential equations point of view. Paper I suggests to distinguish the eight models discussed in Chapter 2 using the the pairs $(\mathbf{T}_1, \mathbf{S}_1)$ and $(\mathbf{T}_2, \mathbf{S}_2)$.

Again any pairs of functionals could be used. The suggestion of paper I is to use the pairs

$$(\mathbf{E}, \mathbf{COV}) \quad \text{and} \quad (\mathbf{E}_3, \mathbf{COV}_4).$$

The main motivation of this choice is that the four functionals are multivariate versions of the classical univariate location, scale, skewness and kurtosis functionals based on the first four moments. This combination is not very robust, however, and the existence of the first four moments are assumed.

The number of distinct elements of $\mathbf{D}$ can be used to distinguish between the models. Let $\mathcal{D}(k)$ be the set of all positive definite diagonal matrices with at most k distinct diagonal entries. A first overview about the possible values of $\mathbf{s}$ and $\mathbf{D}$ in the 8 models of Chapter 2 is given in Table 4.1. The results of this table will now be a bit more elaborated.

Model	Skewness $\mathbf{s}$	Kurtosis $\mathbf{D}$
A1	$\mathbf{0}$	$\mathbf{I}_p$
A2	$\mathbf{0}$	$\mathcal{D}(1)$
A3	$\mathbf{0}$	$\mathcal{D}(1)$
A4	$\mathbf{0}$	$\mathcal{D}(p)$
A5	$\mathbf{0}$	$\mathcal{D}(p)$
B1	$\geq \mathbf{0}$	$\mathcal{D}(k)$
B2	$\propto \mathbf{e}_p$ or $\propto \mathbf{e}_1$	$\mathcal{D}(2)$
B3	$\geq \mathbf{0}$	$\mathcal{D}(p)$

Table 4.1: Possible values of $\mathbf{s}$ and $\mathbf{D}$ in the eight models under consideration.

In the symmetric models the skewness measure $\mathbf{s}$ is $\mathbf{0}$ and therefore only the kurtosis can be used to separate between **A1**-**A5**. Since both scatter functionals are normalized under the normal model, the kurtosis measure must be the identity matrix $\mathbf{I}_p$ in the normal model. Under **A2**, **COV** and $\mathbf{COV}_4$ measure the same population quantity and both are proportional to $\mathbf{\Sigma} = \mathbf{\Omega}\mathbf{\Omega}^T$. Therefore $\mathbf{D}$ must be a diagonal matrix and all diagonal entries are the same. For a t_ν distribution, for example, $\mathbf{D} = (\nu - 2)/(\nu - 4)\mathbf{I}_p$. However also in **A3**, $\mathbf{D}$ is a diagonal matrix

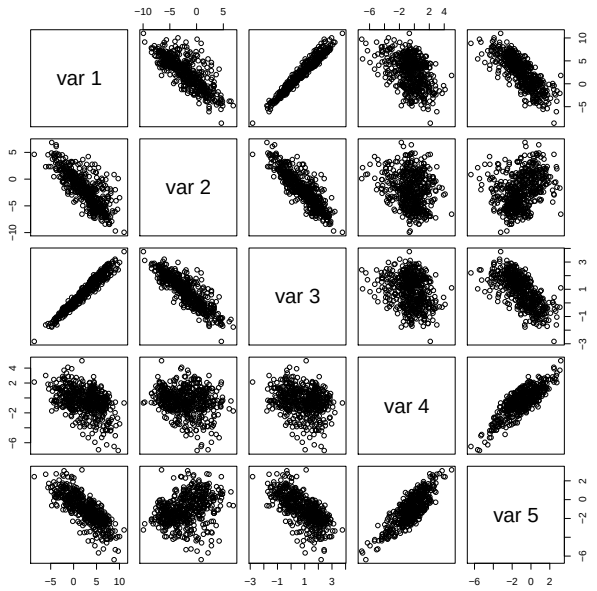

Figure 4.1: Scatter plot matrix of a sample of size 500 following model **B1**.

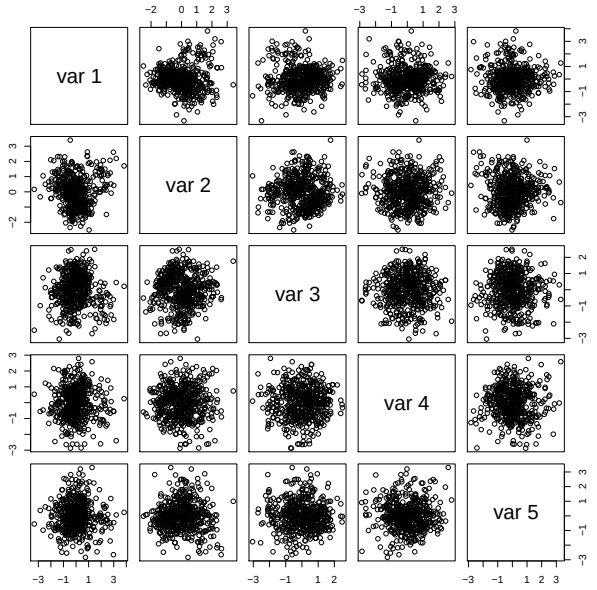

Figure 4.2: Scatter plot matrix of a sample of size 500 following model **B1** after whitening.

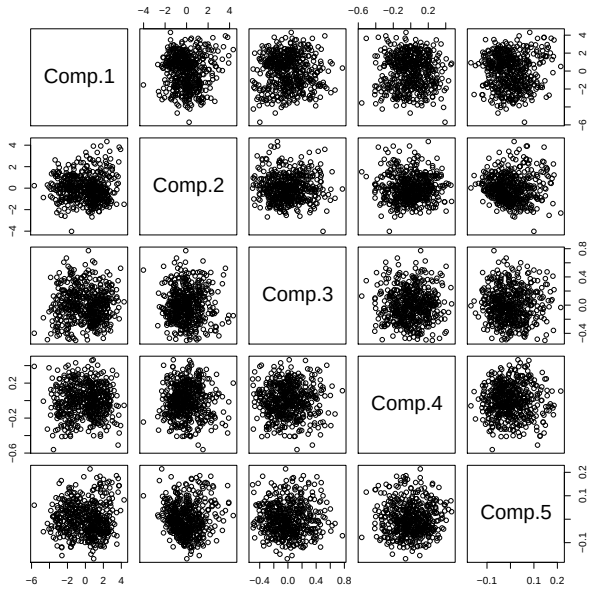

Figure 4.3: Scatter plot matrix of the principle components from a sample of size 500 following model **B1**.

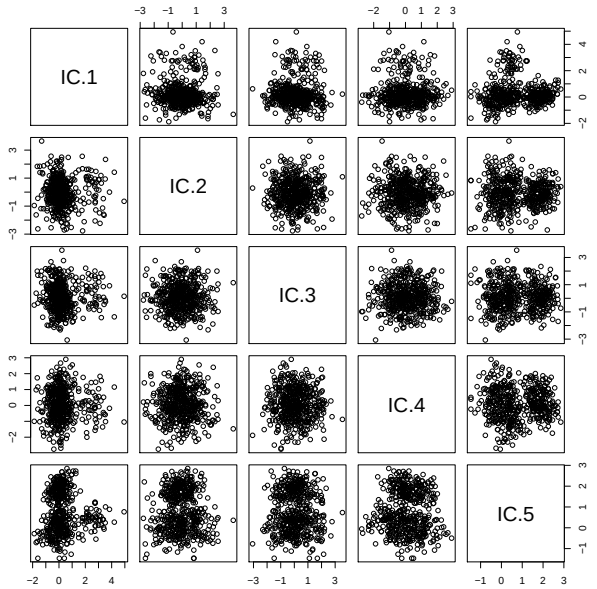

Figure 4.4: Scatter plot matrix of the invariant coordinates from a sample of size 500 following model **B1**.

with identical values on the diagonal. Therefore it is not possible to distinguish between **A2** and **A3**. In the sign-symmetric model **A4** however $\mathbf{D}$ can have distinct diagonal entries. This is the case for the central symmetry **A5** model as well. Therefore models **A1**-**A3** can be separated from the models **A4**-**A5**.

If however $\mathbf{s} \geq \mathbf{0}$, with some elements larger than zero, a skew model must be chosen. Tyler et al. (2008) considered **B1** in more detail and showed that there $\mathbf{D}$ will have at most k distinct diagonal values and the corresponding vectors of $\mathbf{B}$ span (heuristically spoken) Fisher's linear discriminant subspace. Therefore $\mathbf{D}$ in this case gives an idea about the number of mixture components and suggests the coordinates to separate them. In the skew-elliptical model Azzalini and Capitanio (1999) defined the skew-normal distribution in the canonical form where all skewness is absorbed in one component and the remaining $p-1$ components are $N(\mathbf{0}, \mathbf{I}_{p-1})$ distributed. Basically for any skew-elliptical distribution ICS will find the canonical from, it collects all skewness in one component whereas the remaining components are elliptic. All symmetric components have in this canonical form the same kurtosis value and therefore due to the convention of the order of the components the skew component will be the first or the last one. Hence, $\mathbf{s}$ and $\mathbf{D}$ must have in this model one of the two forms given in Table 4.1. In the IC model it is important that both, $\mathbf{S}_1$ and $\mathbf{S}_2$, have the independence property. If both scatter functionals have the independence property and if the components have different kurtosis measures with respect to $\mathbf{S}_1$ and $\mathbf{S}_2$ then, as Oja et al. (2006) show, $\mathbf{z}$ gives $\boldsymbol{\epsilon}$ up to sign, scale and order. The elements of $\mathbf{D}$ give the classical kurtosis measures of the components in the IC model **B3** when the pair $(\mathbf{COV}, \mathbf{COV}_4)$ is used.

This model will actually be considered in more detail in the next chapter.

Before that however still some concluding remarks. The method proposed here can be used to distinguish between a wide range of models and but should be combined with graphical displays. Tests to separate between different models are not yet available. In most models the transformation matrix $\boldsymbol{\Omega}$ can not be recovered, only under assumptions **A4** and **B3** this is often possible.

5 Independent components analysis and robustness

In this chapter we will have a closer look at the independent component model, where often the main goal is to estimate $\mathbf{\Omega}$. Focusing on this problem, we assume that $\mathbf{x}$ is centered. Therefore in this chapter we consider only the model

$$\mathbf{x} = \mathbf{\Omega}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ has independent components.

5.1 Main ICA estimation techniques

Since the model is ill defined, $\mathbf{\Omega}$ and $\boldsymbol{\epsilon}$ can not be recovered as such. Theis (2004) proves that $\mathbf{\Omega}^{-1}$ can be estimated up to signs, scales and permutations of its rows if there is at most one gaussian component. The signs and the order of the independent components are similarly arbitrary but the scales are fixed in model (**B3**) by $\mathbf{COV}(\boldsymbol{\epsilon}) = \mathbf{I}_p$. Actually this definition of the model requires finite second moments for the components of $\boldsymbol{\epsilon}$. This moment assumption is made in most ICA algorithms. Most algorithms usually start with whitening the observed data using the covariance matrix and then continue as in PCA to search for directions in which a measure of non-gaussianity is maximized. From this point of view these ICA algorithms can be seen also as projection pursuit methods.

The non-gaussianity criterion is heuristically motivated by the central limit theorem that suggests that the observed sums of random variables tend to be closer to a gaussian random variable than the latent ones.

The most popular algorithm of this type is the so called fastICA (Hyvärinen and Oja, 1997) which maximizes negentropy NE. Negentropy is a normalized version of the entropy EN and is defined for a standardized random variable x as

$$NE(x) = EN(z) - EN(x),$$

where $z \sim N(0,1)$. Then $NE(x) \geq 0$ and the equality holds only if x is gaussian. It is however difficult to apply directly negentropy, respectively entropy, in practise since it requires the knowledge of the density of x. Different expansion methods lead to different approximations and the following three are often used in practise.

A cumulant based approximation is

$$\frac{1}{12}E(y^3)^2 + \frac{1}{48}(E(y^4) - 3)^2.$$

The two parts of this approximation are based on the classical univariate skewness and kurtosis measures. This approximation is however not robust.

More robust approximations are given by

$$(E(G(x)) - E(G(z)))^2,$$

with the choices

$$G(y) = \frac{1}{\alpha}\log(\cosh(\alpha y)) \quad \text{or} \quad G(y) = -\exp\left(-\frac{1}{2}y^2\right),$$

with $1 \leq \alpha \leq 2$ as a tuning constant.

Brys, Hubert and Rousseeuw (2005) investigate possibilities to robustify ICA algorithms further, especially in the context of fastICA. They consider for example replacing the covariance matrix in the whitening step with the MCD scatter functional, a high-breakdown S-functional. They report however that this approach leads to convergence problems of the fastICA algorithm. This may be because the MCD scatter functional does not have the independence property, which plays an important role in this model. In the final approach of Brys et al. (2005) outlying observations are removed before whitening the data. They however point out the difficulties of this approach as the skewness and the tailweight are important features at the second step of the separation.

5.2 ICA based on two scatter matrices

FOBI was one of the first ICA algorithms and it is a special case of ICS with non-robust $\mathbf{S}_1 = \mathbf{COV}$ and $\mathbf{S}_2 = \mathbf{COV}_4$. Oja et al. (2006) show that any pair of scatter functionals can be used for ICA if both functionals have the independence property and $\mathbf{D}$ has distinct eigenvalues. In the case of FOBI this means that the components cannot have the same classical kurtosis values. The independence property requirement can actually be neglected (Tyler et al., 2008) if $p-1$ components are symmetric. Robust choices of $\mathbf{S}_1$ and $\mathbf{S}_2$ guarantee the robustness of the procedure. Also, moment assumptions can be avoided with suitable choices.

It is important that the $\mathbf{S}_1 - \mathbf{S}_2$-kurtoses of the components differ. In cases where FOBI fails, one may find another pair $(\mathbf{S}_1, \mathbf{S}_2)$ for which $\mathbf{D}$ has distinct elements. An example for such a situation is for example the bivariate independent component model where one component has a χ^2_4 distribution and the other a Laplace distribution. Then both components have the same classical kurtosis value $\beta_2 = 6$ but differ in their higher moments. Using the regular covariance matrix and a scatter functional based on sixth moments however separates these two components. One could conjecture therefore that the only case that never can be resolved with this method is the case of components with identical distributions, which is a restriction compared to the other type of algorithms which can also handle that case given the identical margins are not gaussian.

The two scatter matrices approach in ICA is similar to the main approach in the sense that (i) $\mathbf{S}_1$ is first used for the whitening and (ii) then $\mathbf{S}_2$ for rotation. But the maximization criterion $\kappa(\mathbf{a})$ of the second step depends on the distribution of the whole vector $\mathbf{x}$ and not only on the distribution of $\mathbf{a}^T\mathbf{x}$. Therefore this method is not a projection pursuit method.

The differences between the procedures make comparisons difficult. It is important that the criterion that is used to measure the performance of an algorithm

is scale invariant besides being permutation and sign change invariant since in the two scatter matrices approach $\mathbf{\Omega}$ is scaled so that $\mathbf{S}_1(\mathbf{\Omega}^{-1}\mathbf{x}) = \mathbf{I}_p$.

A common criterion is Amari's performance index (Amari, Cichocki and Yang, 1996) PI which uses the true mixing matrix $\mathbf{\Omega}$ and an estimated unmixing matrix $\hat{\mathbf{\Omega}}^{-1}$. The criterion is based on the product $\mathbf{G} = \hat{\mathbf{\Omega}}^{-1}\mathbf{\Omega}$. If $\hat{\mathbf{\Omega}}^{-1}$ is a good estimate then $\mathbf{G}$ should be a permuted diagonal matrix, which is measured by PI as

$$PI(\mathbf{G}) = \frac{1}{2p(p-1)} \left[\sum_{i=1}^{p} (\sum_{j=1}^{p} \frac{|g_{ij}|}{\max_h |g_{ih}|} - 1) + \sum_{j=1}^{p} (\sum_{i=1}^{p} \frac{|g_{ij}|}{\max_h |g_{hj}|} - 1) \right].$$

Now clearly $PI(\mathbf{PG}) = PI(\mathbf{G})$ but $PI(\mathbf{DG}) = PI(\mathbf{G})$ is not necessarily true. Therefore, for a fair comparison, $\hat{\mathbf{\Omega}}^{-1}$ should be standardized always in the same way. The second standardization method mentioned for ICS in Chapter 4.2 is for example a good way to fix the scales and signs, i.e. the rows of $\hat{\mathbf{\Omega}}^{-1} = (\boldsymbol{\omega}_1 \dots \boldsymbol{\omega}_p)'$ (i.e. the rows of $\mathbf{B}$ in ICS) so that

(i) $||\boldsymbol{\omega}_i|| = 1$, $i = 1, ..., p$,

(ii) $\max(\omega_{i1}, ..., \omega_{ip}) = \max(|\omega_{i1}|, ..., |\omega_{ip}|)$, $i = 1, ..., p$.

The performance index $PI(\mathbf{G})$ can take values in $[0, 1]$ and small values mean a good estimate $\hat{\mathbf{\Omega}}^{-1}$.

Paper III compares in a simulation study several combinations of scatter functionals and fastICA, using the two functions for G as defined above. The study shows that two robust scatter matrices seem a better choice than the FOBI combination and that in case of outliers such combinations have also a clear advantage compared to fastICA.

6 Inference on location based on marginal signs

So far the difference of two location functionals has been used as a measure of skewness. For asymmetric distributions different location functionals measure the location in different ways. Therefore the null hypotheses

$$H_0 : \mathbf{T}_{HR}(\mathbf{x}) = \mathbf{0} \quad \text{and} \quad H_0 : \mathbf{E}(\mathbf{x}) = \mathbf{0},$$

for example, represent different features of the distribution. Here recall that $\mathbf{T}_{HR}(\mathbf{x})$ is the affine equivariant spatial median as defined in Chapter 3.1.2. Hence, when testing or estimating location in asymmetric models one should first consider what "location" one has actually in mind. To avoid this confusion a common assumption is to require some form of symmetry. In that case different tests and estimates refer to the same population quantity. Some of the tests of location are also valid in more general models but this is not of interest here and we consider only symmetric models.

In this chapter let $\mathbf{X} = (\mathbf{x}_1 \ \dots \ \mathbf{x}_n)$ be the $p \times n$ data matrix of n i.i.d. p-variate observations where $\mathbf{x}_i = (x_{i1} \ \dots \ x_{ip})^T$, $i = 1, \dots, n$. For simplicity we will test whether the center of symmetry is the origin, i.e.

$$H_0 : \ \boldsymbol{\mu} = \mathbf{0} \quad \text{versus } H_1 : \ \boldsymbol{\mu} \neq \mathbf{0}.$$

Naturally other hypotheses $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ can be tested by shifting each observation using the null value $\mathbf{x}_i \leftarrow \mathbf{x}_i - \boldsymbol{\mu}_0$.

We call a test statistic Q_n with critical value $c_{n,\alpha}$ valid if

$$P_{H_0}(Q_n > c_{n,\alpha}) = \alpha.$$

The test sequence $(Q_n, c_{n,\alpha})$ is asymptotically valid if it holds that

$$\lim_{n\to\infty} P_{H_0}(Q_n > c_{n,\alpha}) = \alpha.$$

The following list contains some tests considered in the literature for the one sample location problem:

Hotelling's T^2: This test is the multivariate analogue of the t-test and can be seen as the classical test for multivariate location. The test can be derived from different points of view (see for example Morrison, 1998a). T^2 is a monotonic function of the likelihood ratio test in the multivariate normal model. This test is the uniformly most powerful affine invariant test in model **A1**. It is also asymptotically valid in models **A2**-**A5** given the first two moments exist. It is however well known that Hotelling's T^2 is not very efficient when the distribution has heavy tails. The test is also not very robust against outliers.

Signed-rank score tests by Hallin and Paindaveine: These tests combine the ranks of pseudo-Mahalanobis distances between the data points and the origin either with Randles' interdirections (Hallin and Paindaveine, 2002a) or with the so-called standardized spatial signs (Hallin and Paindaveine, 2002b). These tests need no moment assumptions, are affine invariant, robust and with a good choice of the score function also highly efficient. The tests are optimal in the Le Cam sense at correctly specified densities. The test using van der Warden scores is asymptotically at least as good as Hotelling's T^2 in model **A2** when considering asymptotic relative efficiencies. Oja and Paindaveine (2005) combined Randles' interdirections with lift-interdirections and constructed a hyperplane-based version of the original tests of Hallin and Paindaveine. All the tests however target only model **A2** and are in general not valid in models **A3**-**A5**.

Marginal sign and signed-rank tests: These tests are, for example, described in Puri and Sen (1971). The tests combine marginal signed-rank score tests and are asymptotically valid in the models **A1**-**A5**. No moment assumptions are required but the tests are not invariant under affine transformations. The efficiency of the tests suffers if the margins are dependent. We will discuss these type of tests later in more detail and describe also possibilities to make them affine invariant.

Spatial sign and signed-rank tests: These test have been for example reviewed by Möttönen and Oja (1995) and are based on spatial signs and signed-ranks. They are asymptotically valid in the models **A1**-**A5** and, as the previous marginal tests, also lack invariance under affine transformations. For spherical distributions they are however more efficient than the ones based on marginal signs and signed-ranks. These tests can be made invariant by pretransforming the data using any scatter matrix. Randles' spatial sign test (Randles, 2000) uses Tyler's shape matrix for this purpose and is strictly distribution-free in model **A2** (actually even in the larger directional elliptical symmetric model).

Tests based on Oja signs and signed-ranks: Hettmansperger, Nyblom and Oja (1994) and Hettmansperger, Möttönen and Oja (1997) use the affine equivariant Oja signs and signed-ranks and obtain invariant tests which are asymptotically valid in **A1**-**A5**. In the spherical case these tests and the tests based on spatial signs and signed-ranks are asymptotically equivalent. The tests are however difficult to compute.

6.1 Marginal sign and signed-rank tests

The tests based on marginal signs and signed-ranks are of special interest in this thesis. For this purpose let $\mathbf{w}_i = sgn(\mathbf{x}_i) = (sgn(x_{i1}) \ \dots \ sgn(x_{ip}))^T$ denote the vector of the marginal signs of the ith observation and $\mathbf{r}_i = (r_{i1} \ \dots \ r_{ip})^T$, where r_{ij} denotes the rank of $|x_{ij}|$ among all $|x_{1j}|, \dots, |x_{nj}|$. Furthermore let $\mathbf{K}(\mathbf{u}) = (K_1(u_1) \ \dots \ K_p(u_p))^T$ be a p-variate vector of score functions with K_i being (i) continuous, (ii) satisfying $\int_0^1 (K_i(u_i))^2 du_i < \infty$ and (iii) expressable as the difference of two monotone functions.

The test statistic is then given by

$$Q_{\mathbf{K}} = n\bar{\mathbf{m}}_{\mathbf{K}}^T \mathbf{V}_{\mathbf{K}}^{-1} \bar{\mathbf{m}}_{\mathbf{K}},$$

where $\bar{\mathbf{m}}_{\mathbf{K}} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_i \bigodot \mathbf{K}\left(\frac{\mathbf{r}_i}{n+1}\right)$ is the average of the signed-ranks and $\mathbf{V}_{\mathbf{K}} = \{v_{ij}\}$ is the sample covariance matrix of the signed-ranks with elements

$$v_{ij} = \frac{1}{n}\sum_{k=1}^n w_{ki} w_{kj} K_i\left(\frac{r_{ki}}{n+1}\right) K_j\left(\frac{r_{kj}}{n+1}\right).$$

Under the null hypothesis $Q_{\mathbf{K}}$ is asymptotically chi-square distributed with p degrees of freedom.

In practise all p score functions K_i are often chosen to be the same. The following three versions are the most prominent.

Marginal sign test: $K_1(u) = 1, \quad i = 1, \ldots, p.$

Marginal Wilcoxon signed-rank test: $K_i(u) = u, \quad i = 1, \ldots, p.$

Marginal van der Waerden signed-rank test:
$K_i(u) = \Phi_+^{-1}(u) = \Phi^{-1}(\frac{u+1}{2})$, $i = 1, \ldots, p$ where Φ is the cdf of the standard normal distribution.

6.1.1 Marginal signed-rank tests and affine invariance

The marginal signed-rank tests are not affine invariant, even not rotationally invariant. This is of course a huge drawback in practise since then the decision depends on the coordinate system. The larger the dependence between the different coordinates, the more the efficiency suffers. See Table 6.2.4 in Hettmansperger and McKean (1998) for the efficiencies of marginal Wilcoxon signed-rank test compared to Hotelling's T^2 for bivariate data.

Chakraborty and Chaudhuri (1999) applied their data-driven transformation (Chakraborty and Chaudhuri, 1996) on marginal tests and showed that this so-called transformation retransformation technique can be used to obtain invariant tests. For the one sample location problem the transformation retransformation technique selects then $p+1$ data points with indices listed in $I = (i_0, i_1, \ldots, i_p)$. The transformation matrix based on the selected points is then $\mathbf{B}_I = (\mathbf{x}_{i_1} - \mathbf{x}_{i_0} \ \ldots \ \mathbf{x}_{i_p} - \mathbf{x}_{i_0})^{-1}$ and the test is performed using the transformed observations $\mathbf{Z}_I = \mathbf{B}_I\mathbf{X}$. The resulting test is affine invariant.

The obvious question here is how to select the $p+1$ observations? In general any points can be selected to achieve affine invariance, but in order to make the tests as efficient as possible they should be chosen so that the coordinates are as uncorrelated as possible. Chakraborty and Chaudhuri (1999) suggest a choice such that $\mathbf{B}_I\mathbf{\Sigma}\mathbf{B}_I^T$ becomes as close to a diagonal matrix as possible. The scatter parameter $\mathbf{\Sigma}$ is of course unknown and should be replaced in practise with an affine equivariant estimate of a scatter matrix, e.g. with $\mathbf{COV}(\mathbf{X})$. An adaptive procedure for such a selection is described in Chakraborty, Chaudhuri and Oja (1998). The asymptotic theory of Chakraborty and Chaudhuri (1999) seems to be developed for the elliptical model **A2** only.

As shown earlier, ICS yields an invariant coordinate system that is easier to compute then the adaptive data-driven coordinate system of Chakraborty and Chaudhuri (1996). Paper IV investigates the performance of the marginal signed-rank tests in an invariant coordinate system. When constructing an invariant coordinate system for a location testing problem, one should use scatter functionals with respect to the null value. The scatter functionals should also be invariant under permutations of the observations. We thus require that

$$\mathbf{S}_k(\mathbf{AXPJ}) = \mathbf{A}\mathbf{S}_k(\mathbf{X})\mathbf{A}^T, \quad k = 1, 2.$$

An important difference between the original observations and the observations in the invariant coordinate system is that latter are not independent anymore. Under the null hypothesis they are exchangeable however.

An obvious idea when applying tests in an invariant coordinate system is trying to make use of the kurtosis ordering of the invariant coordinates. The components with extreme kurtosis values can be expected to give the directions of the location shift. Applying a univariate signed-rank test to any single component yields a distribution-free affine invariant multivariate test. Unfortunately however it seems that using only subsets of components for testing and combining them as done in the multivariate marginal signed-rank tests seems to have not as much power as using all of them.

6.2 Marginal signed-rank tests in the symmetric independent component model

The marginal signed-rank tests are asymptotically valid in the models **A1**-**A5**. The dependence between the components is taken care of with the estimated covariance matrix $\mathbf{V_K}$. Naturally if the components are independent $\mathbf{V_K}$ is converging to a diagonal matrix. For the independence we need stronger assumptions than **A5**. We assume an independent component model and symmetry, that is, $\mathbf{A5} \cap \mathbf{B3}$. This model and the elliptic model are different extensions of the multivariate normal model. The Maxwell-Hershell Theorem (see for instance Bilodeau and Brenner, 1999, pp. 51) states that the multivariate standard normal distribution is the only spherical distribution with independent margins.

The main idea when constructing the test in this model is first to recover the underlying independent components and then apply univariate signed-rank tests to the estimated independent components $\hat{\mathbf{Z}} = \hat{\mathbf{\Omega}}^{-1}\mathbf{X}$. Here $\hat{\mathbf{\Omega}}^{-1}$ is an estimate of the unmixing matrix $\mathbf{\Omega}^{-1}$. The estimated unmixing matrix $\hat{\mathbf{\Omega}}^{-1}$ must be (i) affine invariant, (ii) invariant under individual sign changes of the observations, (iii) invariant under permutation of the observations and (iv) root-n consistent. Secondly, the condition (ii) $\int_0^1 (K_i(u_i))^2 du_i < \infty$ used for the score functions K_i must be replaced by the stronger assumption (iib) $\int_0^1 (K_i(u_i))^{2+\delta} du_i < \infty$ for some $\delta > 0$.

The gain in this approach is that the matrix $\mathbf{V_K}$ can now be replaced by its probability limit

$$\mathbf{V_K} = diag(E(K_1(u)^2), \ldots, E(K_p(u)^2)),$$

where u is uniformly distributed on $[0, 1]$ and that this test is affine invariant.

In paper V the asymptotic relative efficiencies (ARE) of $Q_{\mathbf{K}}$ are computed with respect to Hotelling's T^2. The AREs can be expressed in terms of univariate AREs of the marginal signed-rank test with respect to the t-test.

If all score functions used correspond to the underlying densities of the independent components, then the test is optimal in the Le Cam sense.

In this thesis $\hat{\mathbf{\Omega}}^{-1}$ is based on two scatter matrices (the ICA problem). From a fully nonparametric point of view the two scatter functionals used should naturally avoid any moment assumptions and paper V recommends therefore to use the pair $(\mathbf{S}_{Tyl}, \mathbf{S}_{HP})$ where both are taken with respect to the origin. The only restriction is then that the underlying components must have distinct $\mathbf{S}_1 - \mathbf{S}_2$-kurtoses.

These tests are in detail explained in paper V which also contains finite sample efficiencies coming from a Monte Carlo study and some robustness considerations.

7 Example

To demonstrate the methods described in this thesis a medical data set is analyzed using R 2.7.0 (R Development Core Team, 2008) and the R packages ICS (Nordhausen, Oja and Tyler, 2008b), ICSNP (Nordhausen, Sirkiä, Oja and Tyler, 2007), JADE (Nordhausen, Cardoso, Oja and Ollila, 2008a), lattice (Sarkar, 2008) and zoo (Zeileis and Grothendieck, 2005).

In our example we give the estimates of the multivariate skewness and kurtosis. The estimates of the (asymptotical) variances and covariances of these estimates are not known yet and therefore no inference tools are available so far. These tools should be developed in future research.

7.1 The data

The data analyzed here are the hemodynamic data collected as a part of the Young Finns Study using whole-body impedance cardiography and plethysmographic blood pressure recordings from fingers. For these data, in 2003 and 2004, 243 healthy subjects between 25 and 42 years of age took part in the recording of the hemodynamic variables both in a supine position and during a passive head-up tilt on a motorized table. During that experiment the subject spent the first ten minutes in a supine position, then was tilted for five minutes to a head-up tilt position (60 Degrees) and for the last five minutes the table was returned to the supine position. Continuously during the experiment several hemodynamic variables were measured while the subject was supposed to be silent and not to move. For a more detailed description, see Päivä, Kähönen, Lehtimäki, Raitakari, Jula, Viikari, Alfthan, Juonala, Laaksonen and Hutri-Kähönen (2008). For this analysis only four one minute averages are available. These are the averages of the last minute before the upwards tilt, the first minute after the head-up tilt, the last minute before the downwards tilt and the fifth minute after the downwards tilt. The four periods will be denoted from now on as recording phases 1, 2, 3 and 4. However only for 235 subjects successful measurements for all recording phases are available. The profiles of those subjects are presented in Figure 7.1 for the three key variables heart rate, cardiac output and vascular resistance index (SVRI).

Different questions arise from the data.

1. Do the key variables return to the pretilt levels after the downwards tilt?

2. Do the subjects react in the same way to the tilt? (Hypothesis is that there are two ways to deal with the tilt.)

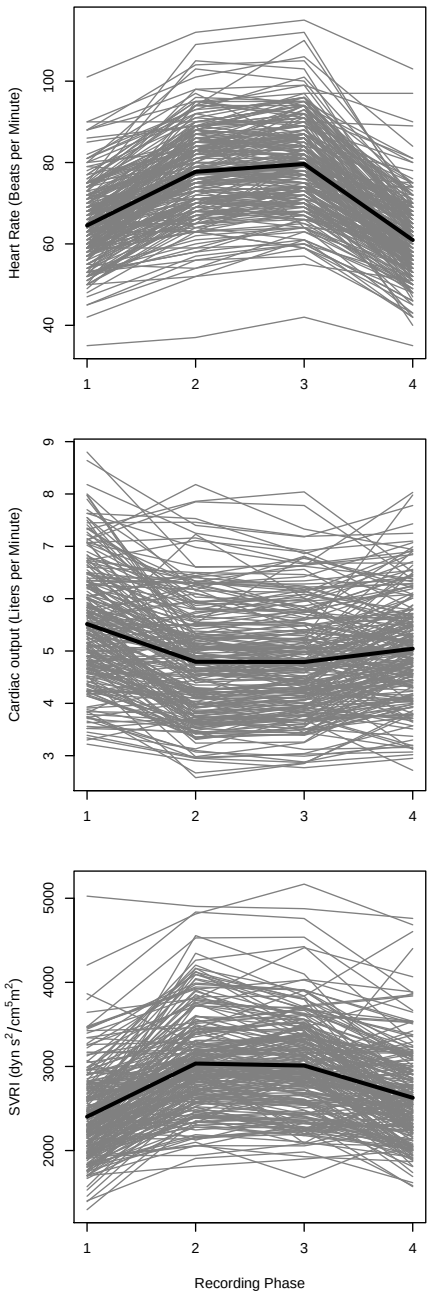

Figure 7.1: Profile of three key variables from the hemodynamic experiment in the Young Finns Study. The thick black line denotes the mean profile.

In the following we will try to answer these questions using the methods described in this thesis.

Naturally it would be also interesting to see the actual continuous signals from such an experiment. Such data was made available from a similar experiment by Ilkka Pörsti at the Tampere University Hospital. The measurements here shown later are from a 62 year old male subject and were recorded 2007. The number of observations available corresponds to the number of heart beats of the subject during the experiment.

7.2 The analysis

We will start by considering question 1.

7.2.1 Location tests

This question can be reformulated as a paired data problem with the hypothesis that the "expected difference" between the first and the last recording phase is zero. As pointed out earlier, the test one should use depends on the underlying distribution of the data.

Therefore, the data is first inspected using the sample versions of $\mathbf{E}$, $\mathbf{E}_3$, $\mathbf{COV}$ and $\mathbf{COV}_4$ to get a feeling for the data. The following values are obtained (ignoring the scatter):

Location: $\hat{\boldsymbol{\mu}} = (3.5872\ 0.4719\ -225.2213)^T$

Skewness: $\hat{\mathbf{s}} = (0.8383\ 0.3519\ 0.1783)^T$

Kurtosis: $diag(\hat{\mathbf{D}}) = (4.5584\ 2.2514\ 1.0062)$

These estimates suggest some skewness and also heavy tails in two directions. However, when looking at the scatter plot of the invariant coordinates as shown in Figure 7.2 one could suspect that the skewness and large kurtosis of the first two components is mainly due to the outliers found in those directions, which are clearly visible in the figure.

Thus, this suggests replacing $\mathbf{E}$, $\mathbf{E}_3$, $\mathbf{COV}$ and $\mathbf{COV}_4$ by robust measures. For example when using $\mathbf{T}_1 = \mathbf{T}_{HR}$, the affine equivariant spatial median, $\mathbf{T}_2=$ transformation retransformation componentwise median, $\mathbf{S}_1 = \mathbf{S}_{HR}$ Tyler's shape matrix jointly estimated with $\mathbf{T}_1$ and $\mathbf{S}_2 = \mathbf{S}_{HP}$, the one step M-estimator scatter matrix based on ranks the following results are obtained:

Location: $\hat{\boldsymbol{T}}_1 = (3.4934\ 0.4269\ -202.5677)^T$

Skewness: $\hat{\mathbf{s}} = (0.2114\ 0.0621\ 0.2200)^T$

Kurtosis: $diag(\hat{\mathbf{D}}) = (1.1777\ 1.0064\ 0.8437)$

The interpretation of the kurtosis values is partly lost since the scatter matrices used are shape matrices and the kurtosis values are standardized so that their product is one. It is also difficult to draw any conclusions without estimated standard errors. Believing however in symmetry, this suggests that models **A4** or **A5** might be good candidates for further analysis.

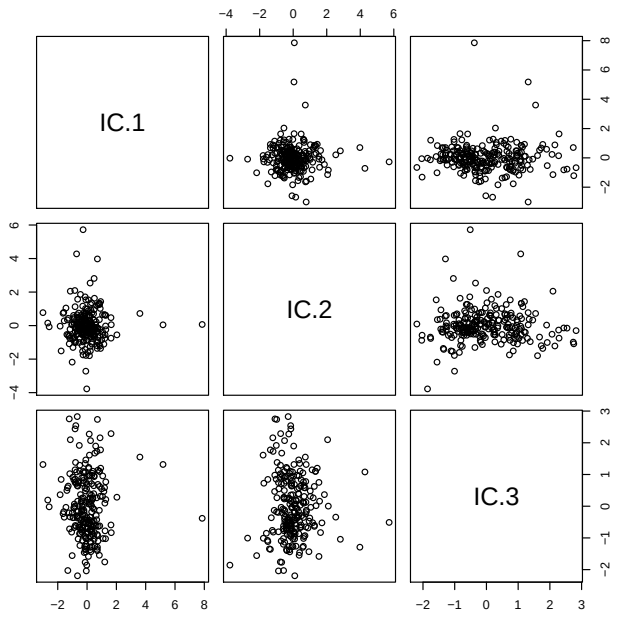

Figure 7.2: Scatter plot of the invariant coordinates of the pairwise differences of the three key variables of the Young Finns Study.

Knowing that there are outliers, robust tests are needed to test the hypothesis of no difference. We used the invariant version of the Puri and Sen test with Wilcoxon scores. If model **A5** is assumed then one has to estimate $\mathbf{V_K}$. Under the symmetric IC model, $\mathbf{B3} \cap \mathbf{A5}$, one can use the known diagonal probability limit of $\mathbf{V_K}$. In both cases the invariant coordinate system was based on Tyler's shape matrix and on the one step M-estimator scatter matrix based on ranks, both with respect to the origin. The results seem very similar as can be seen in in Table 7.1.

Test assumes	Test statistic	p-value
symmetric IC model	151.8241	< 0.001
symmetric NP model	151.3460	< 0.001

Table 7.1: Results of two locations test to test the hypothesis of no difference between recording phase 1 and 4.

Therefore both tests reject clearly and one has to reject the hypothesis that after the tilt those key variables return immediately to their pretilt levels. The affine equivariant version of the marginal Hodges-Lehmann estimator described in Paper IV can then be used to estimate the mean difference in pretilt and aftertilt levels.

7.2.2 Clustering

To answer the second question the clustering should find different profile shapes and therefore using a clustering method based on the four recording phases is not advisable. The shape of the profile is found more likely by considering the differences between recording phase 1 and recording phase 2 and between recording phase 1 and recording phase 3. This gives 6 variables which could be used easily for clustering since the dimension is still relatively small.

However to illustrate our approach, we would like to reduce the dimension. We first create an invariant coordinate system using Tyler's shape matrix and the one step M-estimator scatter matrix based on ranks. Furthermore, since the hypothesis is that there are two clusters, we use the invariant coordinates with the smallest kurtosis measure and with the largest kurtosis measure, since these are the most obvious candidates to show the clusters. The actual clustering is then done with the kmeans algorithm. The resulting groups are plotted in Figure 7.3.

This figure shows that subjects in cluster 1 have basically a constant cardiac output and only a little change in their vascular resistance, whereas in cluster 2 the vascular resistance changes more and the output of the heart is reduced during the head-up tilt. The view that these cardiovascular phenotypes are clinically significant is supported by the finding, that the two groups differ in their pulse wave velocities (PWV) at rest. Cluster 1 has a median pulse wave velocity of 9.40 (95% CI $(8.96, 9.84)$) and cluster 2 of 8.75 (95% CI $(8.51, 8.99)$) and the approximate Wilcoxon signed-rank test yields a p-value of 0.0018. Such an increased pulse wave velocity in cluster 1 suggests that subjects in that cluster have stiffer larger arteries than the subjects in the other cluster.

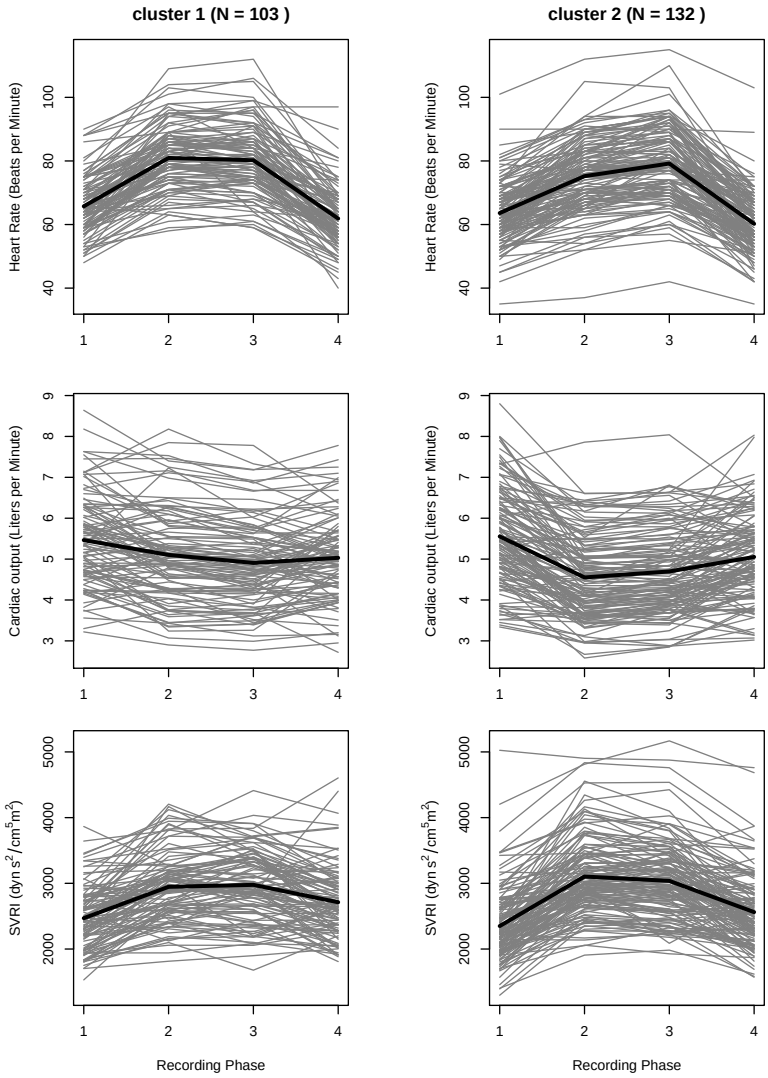

Figure 7.3: Profiles of three key variables from the hemodynamic experiment in the Young Finns Study for the two clusters obtained using kmeans on two invariant coordinates. The thick black lines denote the mean profiles.

7.2.3 Continuous signals

The subject which is now under a closer investigation had 1237 heartbeats in the 20 minutes lasting experiment. Ten recorded variables are shown in Figure 7.4. The grey vertical lines correspond to a change in position. (Note that the the first line actually is not a change is position but of the first ten minutes in the supine position, the first five minutes are considered as a phase where the subject should get used to the experimental environment.) As can be seen in the in Figure 7.4, there are several single atypical observations which can be considered artifacts due to small movements. For a more detailed description of such data and its analysis see Tahvanainen, Koskela, Tikkakoski, Lahtela, Leskinen, Kähönen, Nieminen, Kööbi, Mustonen and Pörsti (2008).

We will investigate these signals now by applying four ICA algorithms (33 measurements had to be removed because of missing measurements). The first three algorithms are based on two scatter matrices methods and the following choices were made

ICA1: $\mathbf{S}_1 = \mathbf{COV}$ and $\mathbf{S}_2 = \mathbf{COV}_4$. Corresponds to the FOBI algorithm. Not robust and requires the components to have different kurtosis measures.

ICA2: $\mathbf{S}_1$ = M-estimator of shape using Cauchy weights and $\mathbf{S}_2$ = M-estimator of shape using t_2 distribution weights. A robust solution but requires that at least nine of the components are symmetric and all components have different kurtosis measures.

ICA3: $\mathbf{S}_1$ = Dümbgen's shape estimator and $\mathbf{S}_2$ = Symmetrized Huber M-estimator. Robust and no assumptions about skewness, however requires also unequal kurtosis measures.

For the comparison we applied also the JADE algorithm which can deal also with components that have equal kurtosis values. It is however also not very robust.

The first thing one notices when comparing the four ICA algorithm results is, that the non-robust algorithms use three components to collect the outliers. This means that if structures would be in those directions they might be lost. Compared to that the outliers can be seen in almost all components of the robust algorithms and no components are used to collect them. This feature makes however also a further visual comparison difficult since especially the outliers have a great effect on the scaling of the components which differs anyway due to the different algorithms. A reasonable thing to do might be to use now some trimming method and trim for example 5 percent of each component - this might make them more comparable. Here we refrain however from doing so and just want to point out that all algorithms seem to have one component which has a clear level shift during the tilt and a component that shows higher variation during the tilt. Furthermore all solutions have a component that shows a level change after the tilt, this component for example might mainly describe the feature of the variable SV.

These findings are summarized in Table 7.2 and for example for further analysis the components with these phenomena (excluding the outliers) might be sufficient. In that case ICA would have been a tool reduce the number of dimensions.

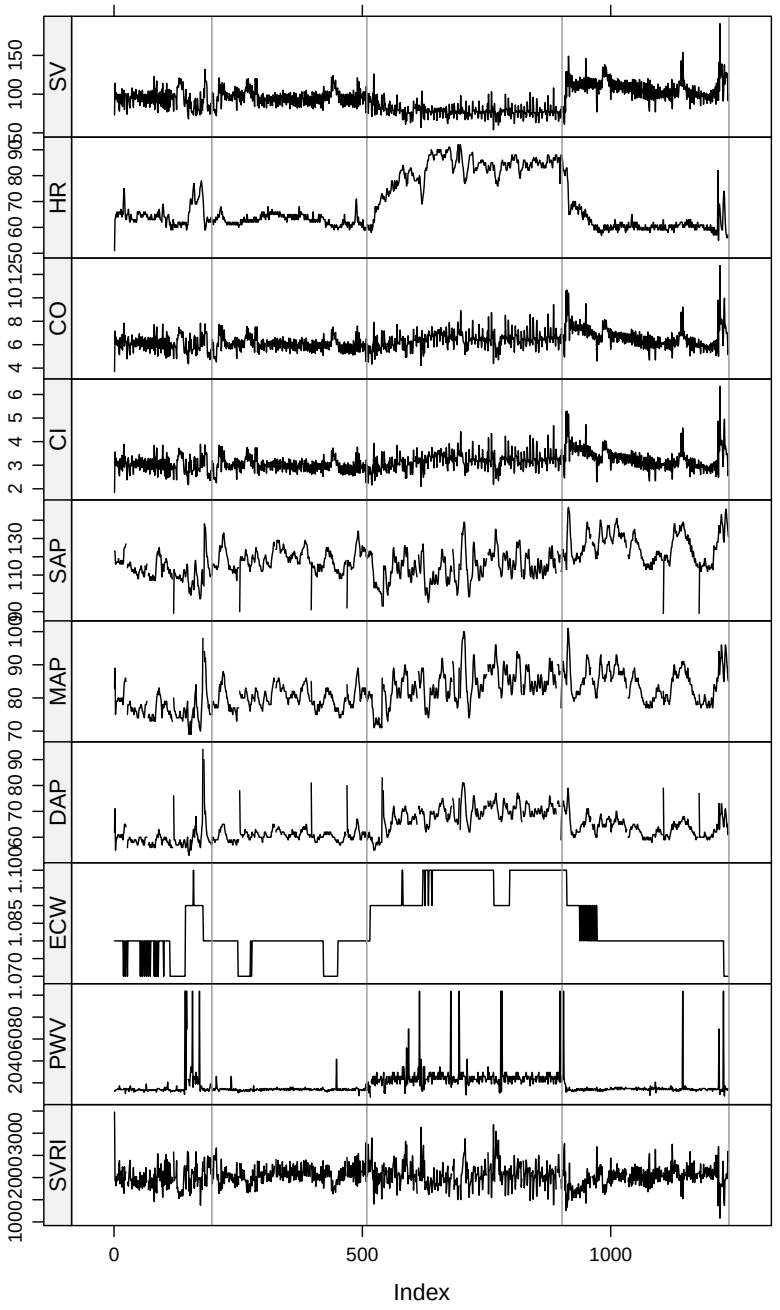

Figure 7.4: Hemodynamic measurements of a single subject. The grey vertical lines correspond to a change in position.

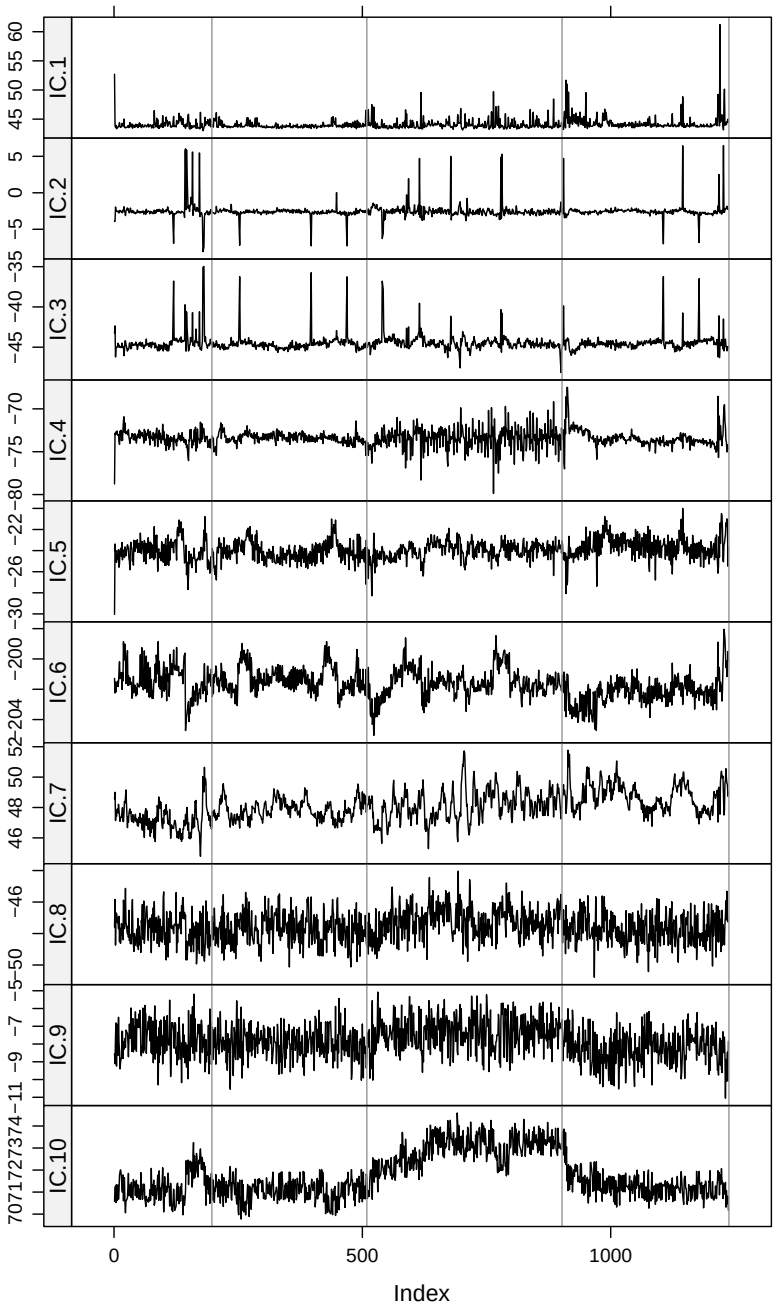

Figure 7.5: ICA signals using ICS with $\mathbf{S}_1 = \mathbf{COV}$ and $\mathbf{S}_2 = \mathbf{COV}_4$.

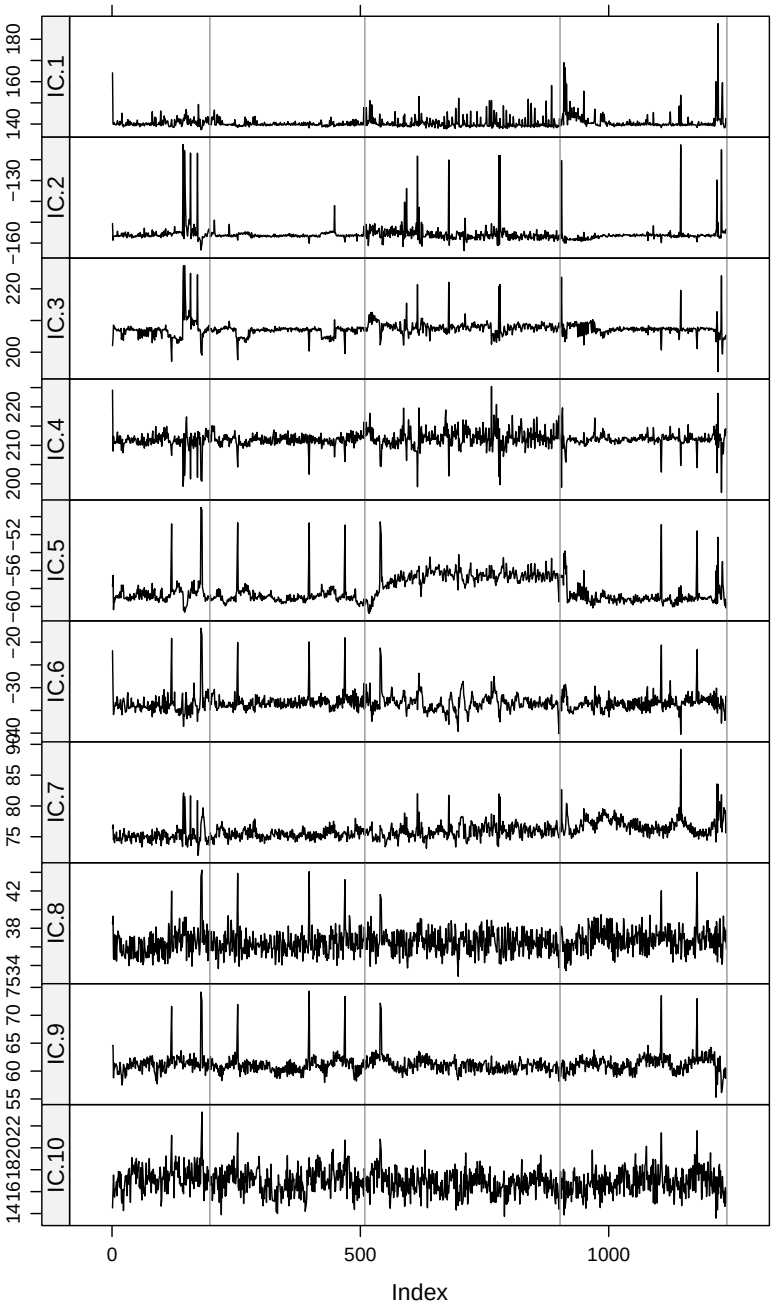

Figure 7.6: ICA signals using ICS with $\mathbf{S}_1$ = M-estimator of shape using Cauchy weights and $\mathbf{S}_2$ = M-estimator of shape using t_2 distribution weights.

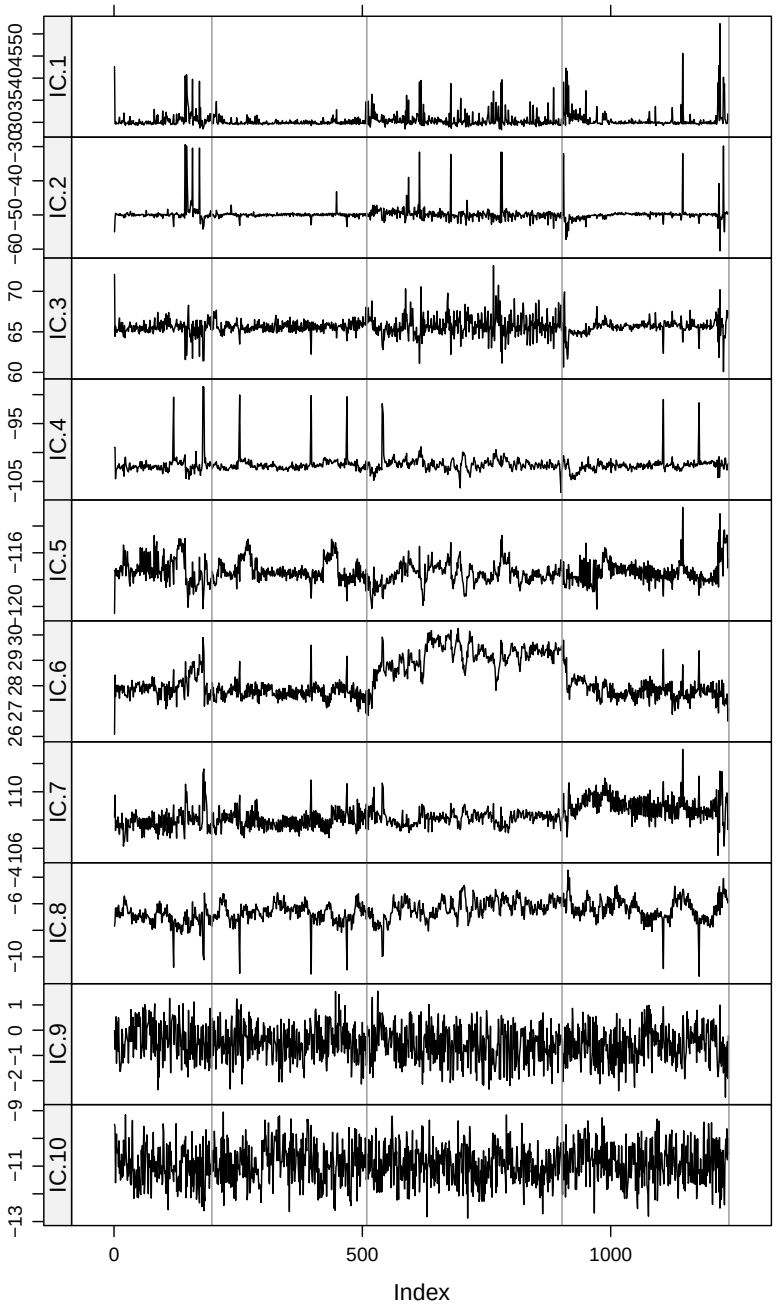

Figure 7.7: ICA signals using ICS with $\mathbf{S}_1$ = Dümbgen's shape estimator and $\mathbf{S}_2$ = Symmetrized Huber M-estimator.

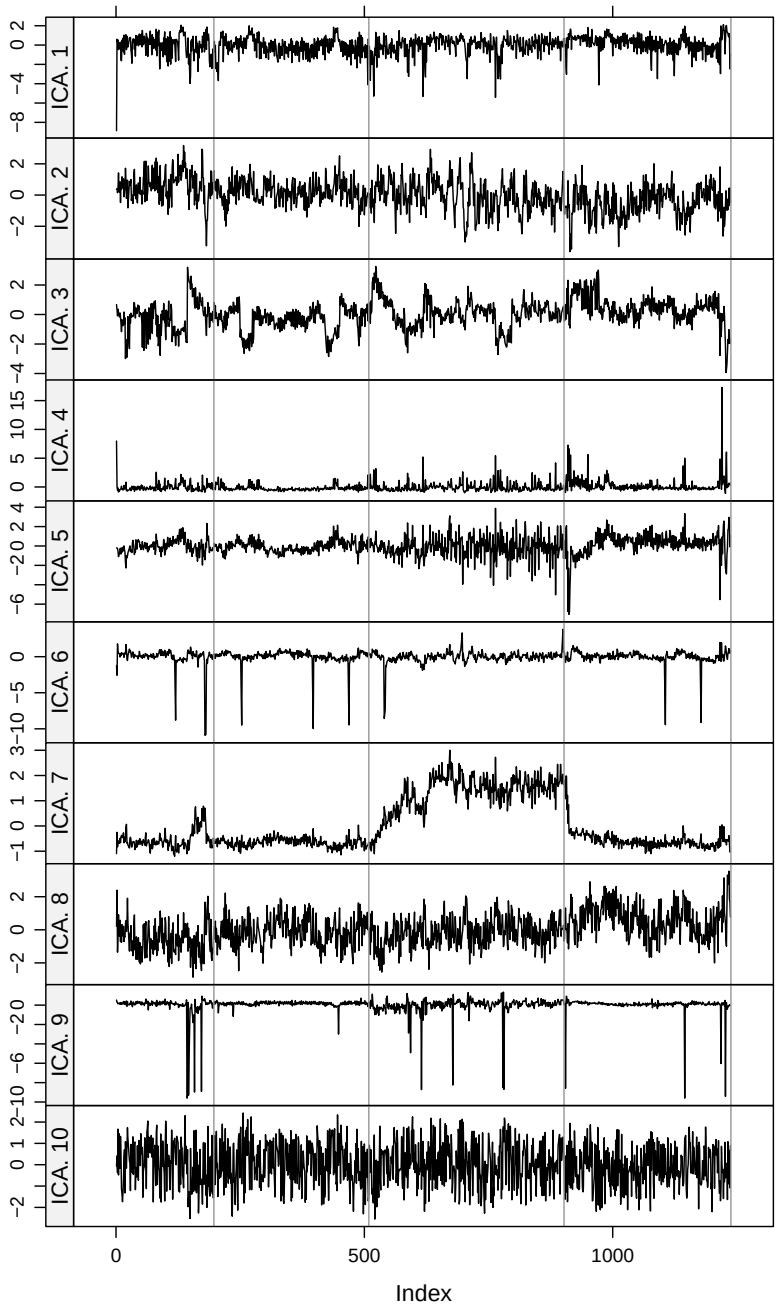

Figure 7.8: ICA signals using the JADE algorithm.

Phenomenon	ICA1	ICA2	ICA3	JADE
Outliers	1, 2, 3	-	-	4, 6, 9
Shift change during tilt	10	5	6	7
Higher variation during tilt	4	4	3	5
Shift change after tilt	9	8	7	8

Table 7.2: Description of the independent components produced by different ICA algorithms.

These sections showed that the methods presented, which are also more or less completely implemented in R, obtain reasonable results in practice. Naturally the analysis here is not sufficient but demonstrates the use of the introduced methods.

Summary of original publications

I. **Nordhausen, K., Oja, H. and Ollila, E. (2008). "Multivariate models and the first four moments", (submitted to the Festschrift for Thomas P. Hettmansperger, edited by Hunter, D.R., Rosenberger, J.L. and Richards, D.).**

The multivariate normal model has been generalized in several ways which all have shown to be useful in practical data analysis. This paper considers seven such extensions including the elliptical model, the skew-elliptical model, the independent component model, finite mixtures of elliptical distributions and so on. The paper suggests to use multivariate measures of skewness and kurtosis to separate the different models. The multivariate measures of skewness and kurtosis applied here are based on the simultaneous use of two location functionals and two scatter functionals. This approach should provide the practical data analyst with a tool to decide which model and methods might be most appropriate for a data set at hand. The decision making is demonstrated using simulated and real data examples.

II. **Nordhausen, K., Oja, H. and Tyler, D.E. (2008). "Tools for exploring multivariate data: The package ICS", *Journal of Statistical Software*, 28, 1–31.**

This paper reviews the invariant coordinate selection (ICS) due to Tyler et al. (2008) and illustrates how it can be applied for descriptive statistics, outlier identification, clustering, independent component analysis and in the context of multivariate nonparametrics using a wide range of examples. All applications are analyzed using the introduced R-package ICS which implements besides a function for ICS also several scatter functionals and two tests for multivariate normality.

III. **Nordhausen, K., Oja, H. and Ollila, E. (2008). "Robust independent component analysis based on two scatter matrices", *Austrian Journal of Statistics*, 37, 91–100.**

In independent component analysis (ICA) most algorithms start by whitening the data using the mean vector and the covariance matrix. The whitening step assumes the existence of second moments and is very sensitive to outliers. Oja et al. (2006) generalized the FOBI algorithm of Cardoso (1989) and showed how and when to use any two scatter matrices to estimate the mixing matrix and the independent components in ICA. A simulation study compares several combinations of scatter functionals for this purpose using the fastICA algorithm of Hyvärinen and Oja (1997) as a reference. The findings of this study show that two robust scatter functionals produce robust estimates of the target quantities.

IV. **Nordhausen, K., Oja, H. and Tyler, D.E. (2006). "On the efficiency of invariant multivariate sign and rank tests". In Liski, E.P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G.P.H. (editors), "Festschrift for Tarmo Pukkila on his 60th birthday", 217–231, University of Tampere, Tampere, Finland.**

Multivariate estimates and tests based on marginal signs and ranks are not affine equivariant and affine invariant, respectively. This paper considers the one and two sample location problem when performing the tests using invariant coordinates obtained from ICS. It is shown that these tests are then affine invariant. It is furthermore investigated whether any advantage can be taken from the ordering of the invariant coordinates. The finite sample efficiencies of the tests, which utilize different number of invariant coordinates, are compared to Hotelling's T^2 test.

V. **Nordhausen, K., Oja, H. and Paindaveine, D. (2008). "Signed-rank tests for location in the symmetric independent component model", *Journal of Multivariate Analysis* (accepted).**

New tests for the one sample location problem are introduced in the symmetric independent component model. The tests are based on marginal signed-ranks and are affine invariant when the estimate of the mixing matrix is affine equivariant. The tests do not require any moment assumptions and are for appropriate chosen score functions locally and asymptotically optimal in the Le Cam sense at given densities. Local powers and asymptotic relative efficiencies with respect to Hotelling's T^2 test are derived. These show that when using van der Waerden scores the asymptotic relative efficiency of the test is always greater than or equal to one when compared to Hotelling's T^2 test. Finite sample efficiencies and the robustness of the tests are investigated in a simulation study.

References

Amari, S. I., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge MA.

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Hobocken, US, 3rd edition.

Arnorld, B. and Beaver, R. (2002). Skewed multivariate models related to hidden truncation and / or selective reporting. *Test*, 11:7–54.

Arnorld, B. and Beaver, R. (2004). Elliptical models subject to hidden truncation or selective sampling. In Genton, M. G., editor, *Skew-Elliptical Distributions and Their Applications*, pages 101–112. Chapman & Hall/CRC, Boca Raton, USA.

Art, D., Gnanadesikan, R., and Kettenring, J. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica*, 21:75–99.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society Series B*, 61:579–602.

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics*. Springer, New York, US.

Brys, G., Hubert, M., and Rousseeuw, J. (2005). A robustification of independent component analysis. *Journal of Chemometrics*, 19:364–375.

Cardoso, J. F. (1989). Source separation using higher order moments. In *Proceedings of IEEE International Conference on Accoustics, Speech and Signal Processing*, pages 2109–2112. Glasgow, UK.

Caussinus, H., Fekri, M., Hakam, S., and Ruiz-Gazen, A. (2003). A monitoring display of multivariate outliers. *Computational Statistics & Data Analysis*, 44:237–252.

Caussinus, H. and Ruiz, A. (1990). Interesting projections of multidimensional data by means of generalized principal component analysis. In Momirovic, K. and Mildner, V., editors, *Proceedings of COMPSTAT 90*, pages 121–126. Physica Verlag, Heidelberg, Germany.

Caussinus, H. and Ruiz-Gazen, A. (1993). Projection pursuit and generalized principal component analysis. In Morgenthaler, S., Ronchetti, E., and Stahel, W. A., editors, *New Directions in Statistical Data Analysis and Robustness*, pages 35–46. Birkhäuser Verlag, Basel.

Caussinus, H. and Ruiz-Gazen, A. (1995). Metrics for finding typical structures by means of principal component analysis. In Escoufier, Y. and Hayashi, C., editors, *Data Science and Its Applications*, pages 177–192. Academic Press, Tokyo, Japan.

Caussinus, H. and Ruiz-Gazen, A. (2006). Principal component analysis, generalized. In Kotz, S., Balakrishnan, N., Read, C., Vidakovic, B., and Johnson, N., editors, *Encyclopedia of Statistical Science*, volume 10, page 6400. John Wiley & Sons.

Chakraborty, B. and Chaudhuri, P. (1996). On a transformation and retransformation technique for constructing affine equivariant multivariate median. In *Proceedings of the American Mathematical Society*, volume 124, pages 2539–2547.

Chakraborty, B. and Chaudhuri, P. (1999). On affine invariant sign and rank tests in one and two sample multivariate problems. In Ghosh, S., editor, *Multivariate Analysis, Design of Experiments, and Survey Sampling*, pages 499–522. Dekker, New York, US.

Chakraborty, B., Chaudhuri, P., and Oja, H. (1998). Operating transformation retransformation on the spatial median and angle test. *Statistica Sinica*, 8:767–784.

Critchley, F., Pires, A., and Amado, C. (2008). Principal axis analysis. *Manuscript*.

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection pursuit approach revisited. *Journal of Multivariate Analysis*, 95:206–226.

Davies, P. L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15:1269–1292.

Dümbgen, L. (1998). On Tyler's M-functional of scatter in high dimension. *Annals of the Institute of Statistical Mathematics*, 50:471–491.

Fang, K.-T. and Zhang, Y. T. (1990). *Generalized Multivariate Analysis*. Springer, New York, US.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman & Hall / CRC, Boca Raton, USA.

Gomez, E., Gomez-Villegas, M., and Marin, J. (1998). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods*, 27:589–600.

Hallin, M. and Paindaveine, D. (2002a). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30:1103–1133.

Hallin, M. and Paindaveine, D. (2002b). Randles' interdirections or Tyler's angles? In Dodge, Y., editor, *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 271–282. Birkhäuser, Basel, Switzerland.

Hallin, M. and Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *The Annals of Statistics*, 34:2707–2756.

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

Hettmansperger, T. P. and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. Arnold, London, UK.

Hettmansperger, T. P., Möttönen, J., and Oja, H. (1997). Affine invariant multivariate one-sample signed-rank test. *Journal of the American Statistical Association*, 92:1591–1600.

Hettmansperger, T. P., Nyblom, J., and Oja, H. (1994). Affine invariant multivariate one-sample sign test. *Journal of the Royal Statistical Society*, B 56:221–234.

Hettmansperger, T. P. and Randles, R. H. (2002). A practical affine equivariant multivariate median. *Biometrika*, 89:851–860.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1980). *Robust Statistics*. John Wiley & Sons, New York, USA.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, New York, USA.

Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analyis. *Neural Computation*, 9:1483–1492.

Isogai, T. (1982). On a measure of multivariate skewness and a test for multivariate normality. *Annals of the Institute of Mathematical Statistics*, 34:531–541.

Jolliffe, I. (2002). *Prinicpal Component Analysis*. Springer, New York, US, 2nd edition.

Kankainen, A., Taskinen, S., and Oja, H. (2007). Tests of multinormality based on location vectors and scatter matrices. *Statistical Methods & Applications*, 16:357–379.

Kent, J. T. and Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *The Annals of Statistics*, 19:2102–2119.

Kent, J. T. and Tyler, D. E. (1996). Constrained M-estimation of multivariate location and scatter. *The Annals of Statistics*, 24:1346–1370.

Kotz, S. (1975). Multivariate distributions at a cross road. In Patil, G., Kotz, S., and Ord, J., editors, *Statistical Distributions in Scientific Work*, volume 1, pages 247–270. D. Reidel Publishing, Dordrecht, Holland.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions: Models and Applications*. John Wiley & Sons, New York, US.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge, UK.

Krzanowski, W. J. (1998). Robustness of multivariate techniques. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*, volume 5, pages 3869–3873. John Wiley & Sons, Chichister, UK.

Lopuhaä, H. P. (1991). Multivariate τ-estimators for location and scatter. *Canadian Journal of Statistics*, 19:307–321.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics - Theory and Methods*. John Wiley & Sons, Chichester, UK.

Morrison, D. F. (1998a). Hotelling's T^2. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*, volume 3, pages 1956–1959. John Wiley & Sons, Chichister, UK.

Morrison, D. F. (1998b). Multivariate analysis, overview. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*, volume 4, pages 2832–2843. John Wiley & Sons, Chichister, UK.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Nonparametric Statistics*, 5:201–213.

Nordhausen, K., Cardoso, J.-F., Oja, H., and Ollila, E. (2008a). *JADE: JADE and ICA performance criteria*. R package version 1.0-1.

Nordhausen, K., Oja, H., and Tyler, D. E. (2008b). *ICS: Tools for Exploring Multivariate Data via ICS/ICA*. R package version 1.1-2.

Nordhausen, K., Sirkiä, S., Oja, H., and Tyler, D. E. (2007). *ICSNP: Tools for Multivariate Nonparametrics*. R package version 1.0-2.

Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics*, 8:154–168.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1:327–332.

Oja, H. and Paindaveine, D. (2005). Optimal signed-rank tests based on hyperplanes. *Journal of Statistical Planning and Inference*, 135:300–323.

Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 19:175–189.

Ollila, E., Oja, H., and Koivunen, V. (2008). Complex-valued ICA based on a pair of generalized covariance matrices. *Computational Statistics & Data Analysis*, 52:3789–3805.

Ord, J. (1986). Pearson system of distributions. In *Enclyclopedia of Statistical Science*, volume 6, pages 655–659. Wiley & Sons, New York, US.

Paindaveine, D. (2008). A canonical definition of shape. *Statistics & Probability*, 78:2240–2247.

Päivä, H., Kähönen, M., Lehtimäki, T., Raitakari, O. T., Jula, A., Viikari, J., Alfthan, G., Juonala, M., Laaksonen, R., and Hutri-Kähönen, N. (2008). Asymmetric dimethylarginine (adma) has a role in regulating systematic vascular tone in young healthy subjects: The cardiovascular risk in young finns study. *American Journal of Hypertension*, 21:873–878.

Pison, G., Rousseeuw, P., Filzmoser, P., and Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, 84:145–172.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, New York, USA.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Randles, R. H. (1989). A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, 84:1045–1050.

Randles, R. H. (2000). A simpler, affine-invariant, multivariate, distribution-free sign test. *Journal of the American Statistical Association*, 95:1263–1268.

Sarkar, D. (2008). *lattice: Lattice Graphics*. R package version 0.17-8.

Serfling, R. J. (2006). Multivariate symmetry and asymmetry. In Kotz, S., Balakrishnan, N., Read, C. B., Vidakovic, B., and Johnson, N. L., editors, *The Encyclopedia of Statistical Sciences*, pages 5338–5345. Wiley & Sons, Hoboken, US, 2nd edition.

Shubhabrata, D. and Sen, P. (1998). Canonical correlation. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*, volume 1, pages 468–482. John Wiley & Sons, Chichister, UK.

Sirkiä, S., Taskinen, S., and Oja, H. (2007). Symmetrised M-estimators of scatter. *Journal of Multivariate Analysis*, 98:1611–1629.

Tahvanainen, A., Koskela, J., Tikkakoski, A., Lahtela, J., Leskinen, M., Kähönen, M., Nieminen, T., Kööbi, T., Mustonen, J., and Pörsti, I. (2008). Analysis of cardiovascular responses to passive head-up tilt using continuous pulse wave analysis and impedance cardiography. *Scandinavian Journal of Clinical and Laboratory Investigation*. In press.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., and Oja, H. (2006). Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. *Journal of Multivariate Analysis*, 97:359–384.

Theis, F. J. (2004). A new concept for separability problems in blind source separation. *Neural Computation*, 16:1827–1850.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, 15:234–251.

Tyler, D. E., Critchley, F., Dümbgen, L., and Oja, H. (2008). Exploring multivariate data via multiple scatter matrices. *Journal of the Royal Statistical Society*. Accepted.

Yenyukov, I. S. (1988). Detecting structures by means of projection pursuit. In Edwards, D. and Raun, N. E., editors, *COMPSTAT 1988*, pages 47–58. Physica-Verlag, Heidelberg, Germany.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14:1–27.

Multivariate Models and the First Four Moments

Klaus Nordhausen[1], Hannu Oja[1], and Esa Ollila[2]

Version: October 26, 2008

[1]Tampere School of Public Health, University of Tampere, Finland
[2]Department of Mathematical Science, University of Oulu, Finland

Abstract

Several extensions of the multivariate normal model have been shown to be useful in practical data analysis. Therefore tools to identify which model might be appropriate for the analysis of a real data set are needed. This paper suggests the simultaneous use of two location and two scatter functionals to obtain multivariate descriptive measures for multivariate location, scatter, skewness and kurtosis, and shows how these measures can be used to distinguish between a wide range of models that extend the multivariate normal model. The method is demonstrated with examples on simulated and real data.

1 Introduction

Recently several extensions of the multivariate normal model with nice tractable features have been introduced and have appeared to be useful in multivariate data analysis. The extensions we have here in mind are the family of elliptical distributions shown to be useful by Fang and Zhang (1990), the independent component model frequently used in signal processing and medical image analysis (see e.g. Hyvärinen et al., 2001) or the family of skew-elliptical distributions described in Genton (2004). Naturally in the practical data analysis it is important to decide which model is the most appropriate one for the data and problem at hand. One can then use statistical inference tools, tests and estimates, tailored for that model. In this paper we propose some guidelines for this task: We use two old ideas from Karl Pearson but transfer them to the multivariate case.

Karl Pearson was perhaps the first scientist to see the importance of skewness and kurtosis in the model selection. In the univariate case, besides seeing these two properties as the properties measured by the standardized third and fourth moments, he also suggested to measure skewness with standardized differences between two location measures such as

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}} \quad \text{or} \quad \frac{\text{mean} - \text{median}}{\text{standard deviation}}.$$

In a similar way, kurtosis may be seen as a ratio of two univariate scale measures (see for example Oja, 1981).

The classical measures of skewness and kurtosis for a univariate random variable x are thus the standardized third and fourth moments,

$$\beta_1(x) = \frac{E\left[(x - E(x))^3\right]}{\left[Var(x)\right]^{3/2}} \quad \text{and} \quad \beta_2(x) = \frac{E\left[(x - E(x))^4\right]}{\left[Var(x)\right]^2}.$$

Also β_1 can be expressed as a standardized difference of two location measures and β_2 is a squared ratio of two scale measures, respectively. To see that, first write

$$E_3(x) = E\left[\left(\frac{x - E(x)}{\sqrt{Var(x)}}\right)^2 x\right] \quad \text{and} \quad Var_4(x) = E\left[\left(\frac{x - E(x)}{\sqrt{Var(x)}}\right)^2 (x - E(x))^2\right].$$

Now $E_3(x)$ is an affine equivariant location measure (based on three first moments) and the square root of $Var_4(x)$ yields an affine equivariant scale measure (based on four first moments). Then

$$\beta_1(x) = \frac{E_3(x) - E(x)}{\sqrt{Var(x)}} \quad \text{and} \quad \beta_2(x) = \frac{Var_4(x)}{Var(x)}.$$

Note also that, for a standardized variable $z = (x - E(x))/\sqrt{Var(x)}$, $\beta_1(x) = E_3(z)$ and $\beta_2(x) = Var_4(z)$.

The other idea of Karl Pearson we want to pick up again is the Pearson's system of frequency curves (see for example Ord, 1986). This flexible system of distributions has been an important tool in identifying the unknown distribution of the observations or of a sample statistic. The procedure is then to use β_1 and β_2 to decide which distribution might fit best to the data. The subfamilies of probability densities in this system were originally given as solutions of a simple differential equation. The extensions to the multivariate case have proved difficult (Kotz, 1975).

The structure of this paper is as follows. The next section describes how two location functionals and two scatter functionals can be jointly used to obtain descriptive statistics for multivariate data. Section 3 investigates the behavior of descriptive measures for multivariate skewness and kurtosis in a wide range of models. In Section 4 the model selection is demonstrated using simulated and real data. Some technical results are collected in the appendix.

2 Moments and multivariate descriptive statistics

Let now $\mathbf{x}$ be a p-variate random vector with cdf F. Then a vector valued functional $\mathbf{T} = \mathbf{T}(F) = \mathbf{T}(\mathbf{x})$ is a location functional if it is affine equivariant in the sense that $\mathbf{T}(\mathbf{Ax} + \mathbf{b}) = \mathbf{AT}(\mathbf{x}) + \mathbf{b}$ for any full rank $p \times p$ matrix $\mathbf{A}$ and any p-vector $\mathbf{b}$. A $p \times p$-matrix valued functional $\mathbf{S} = \mathbf{S}(F) = \mathbf{S}(\mathbf{x})$ is a scatter functional if it is affine equivariant in the sense that $\mathbf{S}(\mathbf{Ax} + \mathbf{b}) = \mathbf{AS}(\mathbf{x})\mathbf{A}'$, again for all $\mathbf{A}$ and $\mathbf{b}$ as defined above.

The most familiar location and scatter functionals are naturally the mean vector and covariance matrix,

$$\mathbf{E}(\mathbf{x}) \quad \text{and} \quad \mathbf{COV}(\mathbf{x}) = \mathbf{E}((\mathbf{x} - \mathbf{E}(\mathbf{x}))(\mathbf{x} - \mathbf{E}(\mathbf{x}))').$$

There are, however, several general families of location and scatter functionals with different properties. For a recent overview with references see Chapter 6 of Maronna et al. (2006). The mean vector and the covariance matrix are functionals based on the first two moments. As in the univariate case, we can define a multivariate location functional and a multivariate scatter functional which use first three moments and first four moments, respectively. These functionals are

$$\mathbf{E}_3(\mathbf{x}) = \frac{1}{p}\mathbf{E}(r^2\mathbf{x}) \quad \text{and} \quad \mathbf{C}\text{OV}_4(\mathbf{x}) = \frac{1}{p+2}\mathbf{E}[r^2(\mathbf{x}-\mathbf{E}(\mathbf{x}))(\mathbf{x}-\mathbf{E}(\mathbf{x}))'].$$

where $r^2 = (\mathbf{x} - E(\mathbf{x}))'\mathbf{C}\text{OV}(\mathbf{x})^{-1}(\mathbf{x} - E(\mathbf{x}))$. See Oja et al. (2006), for example. These two estimates can be seen as one step M-estimates and are the natural multivariate extensions of E_3 and Var_4 given in the introduction. It is remarkable that $\mathbf{C}\text{OV}(\mathbf{x})$ and $\mathbf{C}\text{OV}_4(\mathbf{x})$ have the so called independence property: If $\mathbf{x}$ has independent components, then both $\mathbf{C}\text{OV}(\mathbf{x})$ and $\mathbf{C}\text{OV}_4(\mathbf{x})$ are diagonal matrices but of course with possibly different diagonal elements. (If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then, however, $\mathbf{C}\text{OV}(\mathbf{x}) = \mathbf{C}\text{OV}_4(\mathbf{x}) = \boldsymbol{\Sigma}$. This is due to the appropriate scaling of the second matrix.) The idea in the following is to try to identify the multivariate model with two pairs of functionals, $(\mathbf{E}, \mathbf{C}\text{OV})$ and $(\mathbf{E}_3, \mathbf{C}\text{OV}_4)$.

The usage of two scatter functionals, say $\mathbf{S}_1$ and $\mathbf{S}_2$, has become quite popular recently in multivariate data analysis. Cardoso (1989) used in the first ICA algorithm FOBI the regular covariance matrix ($\mathbf{S}_1 = \mathbf{C}\text{OV}$) and the matrix of fourth moments ($\mathbf{S}_2 = \mathbf{C}\text{OV}_4$), Caussinus and Ruiz-Gazen (1994) used two different scatter matrices (Dispersion Matrix of Gnanadesikan and Kettenring and the regular covariance matrix) for projection pursuit and Critchley et al. (2008) use a one step Tyler matrix and the regular covariance matrix in their principal axis analysis. For a general theory for the comparison of different scatter matrices, see Tyler et al. (2008) and the references therein. In the following we list some of the results given in Oja et al. (2006) and Tyler et al. (2008).

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the values of two different scatter functionals at the distribution of a p-variate random variable $\mathbf{x}$. Let then a $p \times p$ matrix $\mathbf{B}$ and a $p \times p$ diagonal matrix $\mathbf{D}$ solve the eigenvector and eigenvalue problem

$$\mathbf{S}_1^{-1}\mathbf{S}_2\mathbf{B}' = \mathbf{B}'\mathbf{D}.$$

The column elements of $\mathbf{D}$ are then the eigenvalues and the rows of $\mathbf{B}$ are the eigenvectors of matrix $\mathbf{S}_1^{-1}\mathbf{S}_2$. The directions of the eigenvectors (up to sign) are well defined if the eigenvalues of $\mathbf{S}_1^{-1}\mathbf{S}_2$ are distinct. The eigenvectors are then unique up to a multiplication by non-zero constants. Then, for all choices of $\mathbf{B}$, the transformed observations $\mathbf{z} = \mathbf{B}\mathbf{x}$, which are called invariant coordinates, yield $\mathbf{S}_1(\mathbf{z}) = \mathbf{D}_1$ and $\mathbf{S}_2(\mathbf{z}) = \mathbf{D}_2$, where $\mathbf{D}_1$ and $\mathbf{D}_2$ are two diagonal matrices such that $\mathbf{D} = \mathbf{D}_1^{-1}\mathbf{D}_2$.

Besides two scatter matrices $\mathbf{S}_1$ and $\mathbf{S}_2$, two location functionals $\mathbf{T}_1$ and $\mathbf{T}_2$ may sometimes be used to standardize the random variable in a unique way. We then choose the matrix $\mathbf{B}$ in such a way that, if $\mathbf{z} = \mathbf{B}(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))$ then

$$\mathbf{T}_1(\mathbf{z}) = \mathbf{0}, \quad \mathbf{T}_2(\mathbf{z}) \geq \mathbf{0}, \quad \mathbf{S}_1(\mathbf{z}) = \mathbf{I}_p, \quad \text{and} \quad \mathbf{S}_2(\mathbf{z}) = \mathbf{D}.$$

The standardized vector is uniquely defined if $\mathbf{T}_2(\mathbf{z}) > \mathbf{0}$ and if the diagonal elements of $\mathbf{D}$ are distinct. This then suggests multivariate descriptive measures for location, scatter, skewness and kurtosis as

- *Location:* $\mathbf{T}_1(\mathbf{x})$
- *Scatter:* $\mathbf{S}_1(\mathbf{x})$
- *Skewness:* $\mathbf{T}_2(\mathbf{z})$
- *Kurtosis:* $\mathbf{S}_2(\mathbf{z})$

These measures are moment based if one chooses

$$\mathbf{T}_1 = \mathbf{E}, \quad \mathbf{S}_1 = \mathbf{C}\mathrm{OV}, \quad \mathbf{T}_2 = \mathbf{E}_3 \quad \text{and} \quad \mathbf{S}_2 = \mathbf{C}\mathrm{OV}_4.$$

In the univariate case we then get the classical measures given in the introduction.

3 The moments in some multivariate statistical models

3.1 General structures for models

We consider multivariate location-scatter models where the observations $\mathbf{x}$ are thought to be generated by

$$\mathbf{x} = \boldsymbol{\Omega}\boldsymbol{\epsilon} + \boldsymbol{\mu}$$

with a random vector $\boldsymbol{\epsilon}$ standardized in some way. Random vector $\boldsymbol{\epsilon}$ is often used just to formulate the model but sometimes it can be thought to be a real latent variable of its own interest. The models are then distinguished by different assumptions on the vector $\boldsymbol{\epsilon}$. The p-variate vector $\boldsymbol{\mu}$ is the location vector (parameter), and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$ is the scatter matrix (parameter). Matrix $\boldsymbol{\Omega}$ is called the transformation matrix.

In the following we often need some matrix notation: $\mathbf{J}$ is a sign change matrix, that is, a diagonal matrix with diagonal entries ± 1, $\mathbf{P}$ is a permutation matrix obtained from an identity matrix by successively permuting its rows and/or columns. Finally, $\mathbf{O}$ is an orthogonal matrix, that is, $\mathbf{O}\mathbf{O}' = \mathbf{O}'\mathbf{O} = \mathbf{I}$. Clearly, $\mathbf{J}$ and $\mathbf{P}$ are orthogonal.

A structure of symmetrical models is obtained with assumptions

(A1) $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \mathbf{I})$.

(A2) $\boldsymbol{\epsilon} \sim \mathbf{O}\boldsymbol{\epsilon}$ for all orthogonal $\mathbf{O}$.

(A3) $\mathbf{P}\mathbf{J}\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$ for all permutation matrices $\mathbf{P}$ and sign change matrices $\mathbf{J}$.

(A4) $\mathbf{J}\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$ for all sign change matrices $\mathbf{J}$.

(A5) $-\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$.

Skew distributions are obtained, for example, if we assume that

(B1) $\boldsymbol{\epsilon} = \sum_{i=1}^{k} p_i(\boldsymbol{\epsilon}_i + \boldsymbol{\mu}_i)$ where

$$(p_1, \ldots, p_k)' \sim Multin(1; (\pi_1, ..., \pi_k))$$

and $\boldsymbol{\epsilon}_i$ are independent and satisfy (A2).

(B2) $\boldsymbol{\epsilon} = sign(\epsilon^*_{p+1} - \alpha - \beta\epsilon^*_p)\boldsymbol{\epsilon}^*$ where $(\boldsymbol{\epsilon}^{*\prime}, \epsilon^*_{p+1})'$ satisfies (A2).

(B3) The components of $\boldsymbol{\epsilon}$ are independent with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = 1$.

The symmetry assumptions satisfy

$$(A1) \quad \Rightarrow \quad (A2) \quad \Rightarrow \quad (A3) \quad \Rightarrow \quad (A4) \quad \Rightarrow \quad (A5).$$

It is also easy to see that the symmetric models can be seen as border cases of the asymmetric models and we have the following relationships:

$$(A1) \Leftrightarrow (A2)\&(B3) \quad \text{and} \quad (A2) \Rightarrow (B1)\&(B2).$$

If the assumptions (A1)-(A5) or (B1)-(B3) are true, the resulting models are called models (A1)-(A5) or (B1)-(B3), respectively. In the following we describe these models in more detail.

3.2 Elliptical Model

Elliptical model is obtained under the assumption (A2). If assumption (A2) is true then the standardized random variable $\boldsymbol{\epsilon}$ has a spherical distribution around the origin and has a density of the form

$$f(\boldsymbol{\epsilon}) = \exp\{-\rho(||\boldsymbol{\epsilon}||)\},$$

with Euclidean distance $||\boldsymbol{\epsilon}|| = (\epsilon_1^2 + ... + \epsilon_p^2)^{1/2}$ and some function $\rho(\cdot)$. Random vector $\mathbf{x}$ is then elliptically symmetric. The scatter matrix $\boldsymbol{\Sigma}$ and ρ are confounded. Therefore one often assumes that $E(||\boldsymbol{\epsilon}||^2) = p$ (second moments exist) or $Med(||\mathbf{e}||^2) = \chi^2_{p,.5}$. Under this assumption, both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are uniquely defined, and $\boldsymbol{\Sigma}$ is the covariance matrix in the multivariate normal case. Transformation matrix $\boldsymbol{\Omega}$ is not well defined.

The multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given by assumption (A1) is naturally a member in the family of elliptical distributions with the choice $\rho(r) = \frac{p}{2}\log(2\pi) + \frac{1}{2}r^2$. The family of elliptically distributed random variables thus extends the normal model by allowing lighter as well as heavier tails while still maintaining symmetry around $\boldsymbol{\mu}$. Prominent distributions in the elliptical model are the multivariate t-distributions and the power-exponential distributions. This model is in practice the most popular extension of the multivariate model and standard multivariate gaussian methods have been extended to this wider model, see for example Fang and Zhang (1990). Also robust procedures often assume elliptical symmetry.

If we assume that first four moments exist then, in the elliptic model,

$$\mathbf{E}_3(\mathbf{z}) = \mathbf{0}, \quad \text{and} \quad \mathbf{C}\text{OV}_4(\mathbf{z}) = c_\rho \mathbf{I}_p.$$

Here, as before, $\mathbf{z} = \mathbf{B}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$. All marginal distributions have the same kurtosis which is a function of ρ. In the multivariate normal case $c_\rho = 1$ and in the p-variate t_ν case $c_\rho = (\nu - 2)/(\nu - 4)\mathbf{I}_p$.

3.3 Other symmetric models

First consider the model with assumption (A3). The model includes all elliptical distributions as well as the cases with i.i.d components $\epsilon_1, ..., \epsilon_p$. An interesting submodel with different shapes of density contours is obtained if the density of $\boldsymbol{\epsilon}$ is of the form

$$f(\boldsymbol{\epsilon}) = \exp(-\rho(||\boldsymbol{\epsilon}||)),$$

where the norm $||\cdot||$ is any norm that satisfies the condition that $||\mathbf{z}|| = ||\mathbf{PJz}||$ for all permutations $\mathbf{P}$ and sign changes $\mathbf{J}$. This is true for any L_p norm, for example. If we again assume that first four moments exist then also in this model

$$\mathbf{E}_3(\mathbf{z}) = \mathbf{0}, \quad \text{and} \quad \mathbf{COV}_4(\mathbf{z}) = c_f \mathbf{I}_p.$$

Remember that again $\mathbf{z} = \mathbf{B}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$. This means that the first four moments can not be used to distinguish model (A2) from model (A3). As in the elliptic model, both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, but not $\boldsymbol{\Omega}$, are uniquely defined.

A still wider symmetric model is obtained if we assume (A4). Assuming that first four moments exist then

$$\mathbf{E}_3(\mathbf{z}) = \mathbf{0}, \quad \text{and} \quad \mathbf{COV}_4(\mathbf{z}) = \mathbf{D}.$$

where $\mathbf{D}$ is a diagonal matrix. Thus model (A4) can be separated from model (A3) just by looking at the diagonal elements of $\mathbf{D}$: If the diagonal elements are not all the same, then the model (A3) is not correct any more. It is remarkable that, in this model (A4), both $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are uniquely defined, and $\boldsymbol{\Omega}^{-1}$ gives a transformation to a standardized latent variable, $\boldsymbol{\epsilon}$.

The widest symmetric model is obtained if one assumes (A5). Under the same assumptions as before,

$$\mathbf{E}_3(\mathbf{z}) = \mathbf{0}, \quad \text{and} \quad \mathbf{COV}_4(\mathbf{z}) = \mathbf{D}.$$

where $\mathbf{D}$ is again a diagonal matrix with possibly distinct diagonal elements. Thus it is not possible to make a distinction between models (A4) and (A5). Our conjecture is that the distinction can be made if one uses three different scatter matrices.

3.4 Models with skew distributions

Consider next the model of *finite mixtures of elliptical distributions*. This is given by assumption (B1). If the first four moments exist then in this model

$$\mathbf{E}_3(\mathbf{z}) = \mathbf{s}, \quad \text{and} \quad \mathbf{COV}_4(\mathbf{z}) = \mathbf{D}.$$

If $\boldsymbol{\mu}_1 = ... = \boldsymbol{\mu}_k$ we are back in the elliptic case. Also in the nonelliptic cases $\mathbf{s}$ may be zero under some special conditions. $\mathbf{D}$ has at most k distinct diagonal elements. Only $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, not $\boldsymbol{\Omega}$, are well defined parameters. Fisher's linear subspace to discriminate between the mixture populations corresponds to the subspace spanned by the components of $\mathbf{z} = \mathbf{B}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$ which correspond to the distinct values of $\mathbf{D}$ (without any knowledge on the population membership).

Skew-elliptical distributions are given by assumption (B2). Given the first four moments exist,

$$\mathbf{E}_3(\mathbf{z}) = \mathbf{s}, \quad \text{and} \quad \mathbf{COV}_4(\mathbf{z}) = \mathbf{D},$$

where $\mathbf{s}$ has at most one non-zero element and, in a similar way with the same division, the multiplicities of two possible values of the diagonal elements $\mathbf{D}$ are 1 and $p-1$. The corresponding $(p-1)$ subvector of $\mathbf{z} = \mathbf{B}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$ is spherically distributed, and one component of $\mathbf{z}$ absorbs all skewness.

Finally, the *independent components model* is given by (B3). If the first four moments exist, then

$$\mathbf{E}_3(\mathbf{z}) = \mathbf{s}, \quad \text{and} \quad \mathbf{COV}_4(\mathbf{z}) = \mathbf{D},$$

where the elements of $\mathbf{s}$ and of the diagonal of $\mathbf{D}$ can be related to the classical moment based univariate kurtosis and skewness measures if all diagonal elements of $\mathbf{D}$ are distinct: Then $\beta_1(z_i) = ps_i$ and $\beta_2(z_i) = (p+2)D_{ii} - p + 1$. As both $\mathbf{COV}$ and $\mathbf{COV}_4$ have the independence property, $\mathbf{z}$ is a latent vector of independent components. This procedure to find independent components, a solution for the ICA problem, is the well-known FOBI algorithm proposed by Cardoso (1989).

4 Examples

In this section we want to apply this approach on some simulated and real data. All computations are done in R 2.7.0 (R Development Core Team, 2008) by using the package ICS (Nordhausen et al., 2008b). Naturally all the population versions with expected values above will be replaced sample versions with sample means. However, general statistical inference tools (tests and estimates with confidence ellipsoids) based on $\mathbf{s}$ and $\mathbf{D}$ have not been developed so far.

4.1 Examples with simulated data

First we want to evaluate the procedure in four simulated 3-variate data sets, where the observations are obtained following the used model definition (sampling $\boldsymbol{\epsilon}$ and then transforming to $\mathbf{x} = \boldsymbol{\Omega}\boldsymbol{\epsilon} + \boldsymbol{\mu}$). We choose (i) a multivariate normal model, (ii) an elliptic t_{10} model, (iii) a skew-normal model with $\alpha = 0$ and $\beta = 4$ and (iv) an independent component model where the three independent components have a normal, t_{10} and uniform distribution, respectively. In all four cases, the sample sizes are $n = 1000$ and

$$\boldsymbol{\Omega} = \begin{pmatrix} 2 & -2 & -0.3 \\ 1 & 2 & 0.5 \\ 0.5 & 0.5 & 1.7 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}.$$

The estimates of $\mathbf{s}$ and $\mathbf{D}$ are then presented in Table 1.

True distribution	$\mathbf{s}$	$diag(\mathbf{D})$
Normal	$(0.0052\ 0.0557\ 0.0271)'$	$(1.0095\ 0.9766\ 0.9594)$
Elliptic	$(0.0247\ 0.0059\ 0.0133)'$	$(1.3060\ 1.2266\ 1.1218)$
Skew-normal	$(0.2304\ 0.0021\ 0.0021)'$	$(1.1753\ 1.0406\ 0.9437)$
IC (symmetric)	$(0.0810\ 0.0142\ 0.0069)'$	$(1.1845\ 0.9938\ 0.7478)$

Table 1: Skewness and Kurtosis measures for the four simulated data sets.

The estimates are based on higher moments and have therefore a large variation even with sample size $n = 1000$. Observed values of $\mathbf{s}$ and $\mathbf{D}$ calculated from the

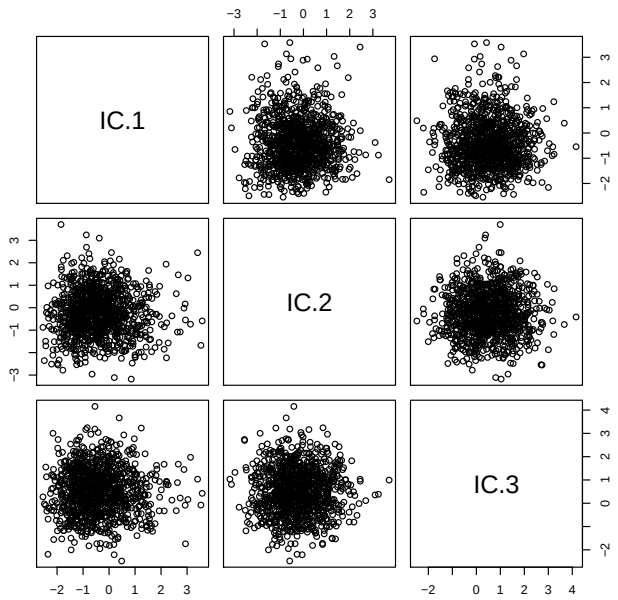

Figure 1: Scatter plot matrix of the invariant coordinates of the simulated skew-normal data set.

sample from a normal distribution clearly suggest model (A1) as $\mathbf{s}$ is close to zero and all the diagonal elements of $\mathbf{D}$ are close to 1. Observed values for the elliptical data only suggest (A2) or (A3) since the diagonal values of $\mathbf{D}$ are the same but distinct from 1. It is remarkable that, in the case of t_ν distribution, the value $D_{22} = 1.22$ suggests a value $\nu = 13$ whereas the true value is $\nu = 10$. The skew-normal sample offers the possibilities (B1)-(B3), however, since the skewness is concentrated in one component and since that component has a clear distinct kurtosis measure compared to the others a skew-elliptic model might be the most parsimonious solution. Furthermore, since the two remaining kurtosis measures are both more or less 1, the skew-normality assumption may seem realistic. Of course, instead of looking at the values of the estimates only, one should have a look at the bivariate scatter plots of the transformed variables in $\mathbf{z}$. See Figure 1 for the plots of the simulated skew-normal data. The last sample values suggest (A4) or (A5) or (B3) with symmetry as well. In general, and as suggested earlier, a look at a third scatter matrix might help to make the distinction. Similarly looking at the scatter plots can also help to distinguish between the models as will be shown in the next section.

4.2 Examples with real data sets

Now we turn from simulated data to real data. The first data set is the famous Fisher's Iris data set where there are four different measurements on 150 iris plants which come from three different species ($n = n_1 + n_2 + n_3$, $n_1 = n_2 = n_3 = 50$).

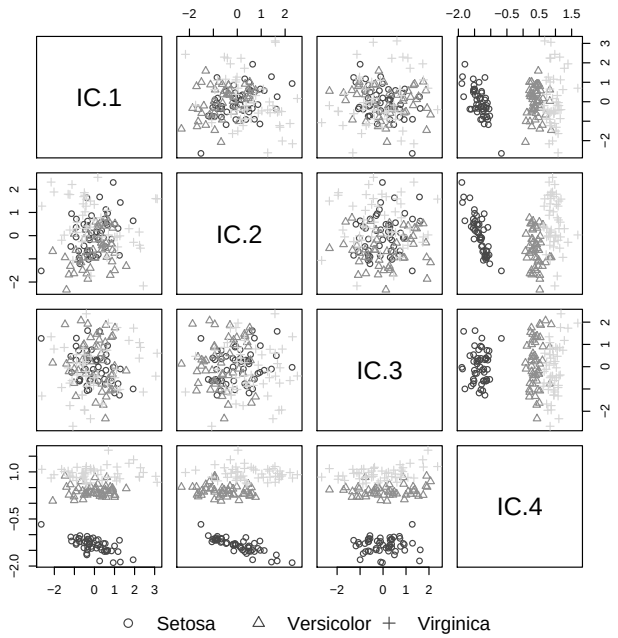

Figure 2: Scatter plot matrix of the invariant coordinates of the Iris data set.

The group membership information is ignored in the analysis. The second data set is the Australian Athletes data set (AIS) that was considered as an example of a skew-normal data in Azzalini and Capitanio (1999) (The four variables are Body Mass Index, Body Fat, Sum of Skin Folds and the Lean Body Mass). In this example, we consider separately the data set of all (men and women) $n = 202$ athletes and that of the $n = 100$ female athletes.

Data	$\mathbf{s}$	$diag(\mathbf{D})$
Iris	$(0.0613\ 0.1794\ 0.0226\ 0.1176)'$	(1.2074 1.0269 0.9292 0.7405)
AIS, all	$(0.4821\ 0.0928\ 0.1473\ 0.1687)'$	(1.7154 1.2788 0.9353 0.7659)
AIS, Female	$(0.3231\ 0.1611\ 0.1369\ 0.0654)'$	(1.3583 1.1320 1.0322 0.8330)

Table 2: Skewness and Kurtosis measures for the Iris and the Australian athletes data.

Table 2 gives the values of $\mathbf{s}$ and the diagonal elements of $\mathbf{D}$ for the above real data sets. It is obvious from these values that there are two still mildly skew components in the Iris data and that two components have a deviating kurtosis. Therefore there models (B1) and (B3) are possible candidates. Figure 2, the scatter plot of the invariant components reveals that the first component has probably a different kurtosis measure to catch some slightly outlying observations whereas the last component shows that the means of the three groups all lie on a line and that this component could be used for separation.

A skew-elliptical model for the Australian athletes data set sounds tempting since the individuals in the data set are certainly collected from a population

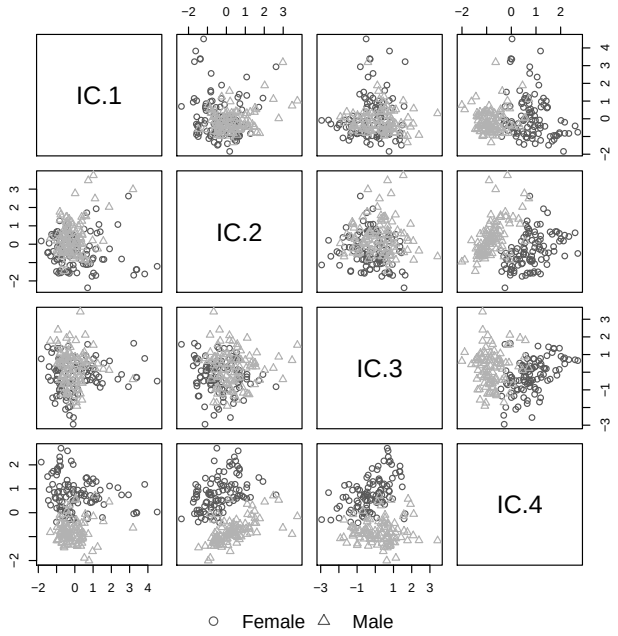

Figure 3: Scatter plot matrix of invariant coordinates of the Australian athletes data set.

using a kind of selective sampling. However the descriptive statistics for the full data set seems to contradict this. The reason becomes apparent when looking at Figure 3, where the last component shows the two clusters of men and women. Looking therefore only at the female athletes, a skew-normal model seems much more realistic; the first component is the one that absorbs all skewness.

5 Discussion

The approach proposed in the paper gives some simple tools to consider critically the model assumptions. Even if one is using multivariate nonparametrical tests as in Hallin and Paindaveine (2002) and Nordhausen et al. (2008a) , one can use tests and estimated tailored for certain target distribution, e.g. multivariate normal distribution. Naturally, the skewness and kurtosis cannot distinguish between all models, but using a third scatter matrix may help to decide about (A4) (A5) or (B3). It is also important to have a careful look at the scatter plots of the invariant coordinates. Also further extensions of the considered model (e.g. skewness in several directions in the skew-elliptic model) are still possible.

Finally note that the choice of the two location and two scatter functionals in this paper was motivated by their univariate classical counterparts $(\mu, \sigma^2, \beta_1, \beta_2)$. However this choice assumes the existence of first four moments and the sample statistics are highly nonrobust. It is obvious that the two location statistics and two scatter matrices used here can be replaced them by other, more robust,

functionals. The model selection follows then the same rough rules. The scatter functionals should then be rescaled so that the regular covariance matrix is obtained in the multivariate normal case. In the independent components model, the scatter functionals should have the independence property.

A Technical details

This section collects technical details which are valid when $\mathbf{S}_1$ and $\mathbf{S}_2$ are any two scatter matrices. Proofs are given only for new results.

Result 1. If $\boldsymbol{\epsilon}$ satisfies (A4) then any scatter functional gives a diagonal matrix.

Proof of Result 1. If $\boldsymbol{\epsilon} \sim \mathbf{Je}$, then $\mathbf{S}(\boldsymbol{\epsilon}) = \mathbf{S}(\mathbf{J}\boldsymbol{\epsilon}) = \mathbf{J}\mathbf{S}(\boldsymbol{\epsilon})\mathbf{J}$. Be $\mathbf{S}(\boldsymbol{\epsilon}) = \mathbf{A}$ where $\mathbf{A}$ is a symmetric psd matrix, and therefore must be $\mathbf{A} = \mathbf{JAJ}$ for any sign change matrix $\mathbf{J}$. Let $\mathbf{J}^i$ be a signchange matrix with

$$\mathbf{J}^i_{jj} = \begin{cases} 1, \text{ if } j \neq i \\ -1, \text{ if } j = i \end{cases}$$

Then $\mathbf{S}(\mathbf{J}^i\boldsymbol{\epsilon})$ changes the signs of the off-diagonal elements of the ith row and the ith column of $\mathbf{A}$. Therefore those elements must be 0 in order for $\mathbf{A} = \mathbf{J}^i\mathbf{A}\mathbf{J}^i$ to hold. Since this is true for all $i = 1, \ldots, p$, $\mathbf{A}$ has to be a diagonal matrix.

Result 2. If $\boldsymbol{\epsilon}$ follows (A4) then the two scatter transformation on $\mathbf{x} = \boldsymbol{\Omega}\boldsymbol{\epsilon} + \boldsymbol{\mu}$ estimates $\boldsymbol{\Omega}$ up to scale and permutation.

Proof of Result 2. Proof follows the lines of the Proof of Theorem 5.3 of Tyler et al. (2008) together with the Result 1 and the fact that in this model $\mathbf{x} = \boldsymbol{\Omega}\boldsymbol{\epsilon} + \boldsymbol{\mu} = (\boldsymbol{\Omega}\mathbf{D}^{-1}\mathbf{J})(\mathbf{JD}\boldsymbol{\epsilon}) + \boldsymbol{\mu} = \boldsymbol{\Omega}^*\boldsymbol{\epsilon}^* + \boldsymbol{\mu}$ where $\boldsymbol{\epsilon}^*$ is also part of (A4).

Result 3. Assume model (B1), $k = 2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ then

$$d_{11} = \ldots = d_{p-1,p-1} > d_{pp} \quad \text{or} \quad d_{11} > d_{22} = \ldots = d_{pp} \quad \text{or} \quad d_{11} = \ldots = d_{pp}.$$

In the first two cases, the subspace corresponding to the eigenvalue with multiplicity 1 is Fisher's discriminant subspace. The result for general k is stated in Theorem 5.2 of Tyler et al. (2008).

Result 4. If $\boldsymbol{\epsilon}$ satisfies (B2), then any scatter functional gives a diagonal matrix.

Proof of Result 4. If $(\boldsymbol{\epsilon}^{*\prime}\ \epsilon^*_{p+1})'$ satisfies (A2) then also $\boldsymbol{\epsilon}^*$ and any of its components satisfies (A2) but with different dimension. The $sign(\epsilon^*_{p+1} - \alpha - \beta\epsilon^*_p)$ is therefore for the first first $p-1$ components of $\boldsymbol{\epsilon}^*$ a random sign change. Clearly $\mathbf{J}^i\boldsymbol{\epsilon} \sim \boldsymbol{\epsilon}$ for $i = 1, \ldots, p-1$ and

$$\mathbf{S}(\boldsymbol{\epsilon}) = \mathbf{S}(\mathbf{J}^i\boldsymbol{\epsilon}) = \mathbf{J}^i S(\boldsymbol{\epsilon})\mathbf{J}^i.$$

Applying the same reasoning as in the proof of Result 1 shows again that $\mathbf{S}(\boldsymbol{\epsilon})$ must be a diagonal matrix.

Result 5. If $\boldsymbol{\epsilon}$ satisfies (B2) then

$$d_{11} = \ldots = d_{p-1,p-1} > d_{pp} \quad \text{or} \quad d_{11} > d_{22} = \ldots = d_{pp} \quad \text{or} \quad d_{11} = \ldots = d_{pp}.$$

Proof of Result 5. Using again the fact that the first $p-1$ components of $\boldsymbol{\epsilon}$ still follow (A2) one knows, that these components are exchangeable and therefore must hold

$$\mathbf{S}(\boldsymbol{\epsilon}) = \mathbf{S}(\mathbf{P}^*\boldsymbol{\epsilon}), \quad \text{where} \quad \mathbf{P}^* = \begin{pmatrix} P & \mathbf{0} \\ \mathbf{0}' & 1 \end{pmatrix}.$$

From which one can conclude that the first $p-1$ diagonal elements of $\mathbf{S}(\boldsymbol{\epsilon})$ must be the same. The rest follows by applying Theorem 4.2 of Tyler et al. (2008).

Result 6. Assume that $\boldsymbol{\epsilon}$ satisfies (B3) and $\mathbf{D}$ is based on any $\mathbf{S}_1$ and $\mathbf{S}_2$ having the independence property. If the diagonal elements of $\mathbf{D}$ are distinct then $\boldsymbol{\epsilon}$ and $\mathbf{B}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$ differ by at most a permutation and/or change in componentwise signs and scales. For the case of non-distinct eigenvalues, see Theorem 5.6 of Tyler et al. (2008).

References

Azzalini, A. and Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society, Series B, 61, 579–602.

Cardoso, J.F., 1989. Source separation using higher order moments. In Proceedings of IEEE International Conference on Acustics, Speech and Signal Processing, Glasgow, 2109–2112.

Caussinus, H., and Ruiz-Gazen, A., 1994. Projection pursuit and generalized principal component analysis. In: S. Morgenthaler, E. Ronchetti, and W. A. Stahel, (Eds.), New Directions in Statistical Data Analysis and Robustness, Birkhäuser Verlag Basel, 35–46.

Critchley, F., Pires, A., Amado, C., 2008, Principal axis analysis, Manuscript.

Fang, K.-T., and Zhang, Y. T., 1990. Generalized multivariate analysis, Springer, New York.

Genton, M. G., 2004. Skew-elliptical distributions and their applications: A journey beyond normality. Chapman & Hall / CRC, Boca Raton.

Hallin, M. and Paindaveine, D., 2002. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. The Annals of Statistics, 30, 1103–1133.

Hyvärinen, A., Karhunen, J. and Oja, E., 2001. Independent component analysis, Wiley & Sons, New York.

Kotz, S., 1975. Multivariate distributions at a crossroad. In: G.P. Patil, S. Kotz, J.K. Ord (Eds.), Statistical Distributions in Scientific Work, Vol.1, Reidel Publishing, Dordrecht, 247–270.

Maronna, R. A., Martin, D.M. and Yohai, V.J., 2006. Robust statistics. Theory and methods. Wiley & Sons, Chichester.

Nordhausen, K., Paindaveine, D. and Oja, H., 2008a. Signed-rank tests for location in the symmetric independent component model. Journal of Multivariate Analysis (accepted).

Nordhausen, K., Oja, H. and Tyler, D.E., 2008b. ICS: Tools for exploring multivariate data via ICS/ICA. R package version 1.1-2.

Oja, H., 1981. On location, scale, skewness and kurtosis of univariate distributions, Scandinavian Journal of Statistics, 8, 154–168.

Oja, H., Sirkiä, S. and Eriksson, J., 2006. Scatter matrices and independent component analysis. Austrian Journal of Statistics, 35, 175-189.

Ord, J.K., 1986. Pearson system of distributions. In: Enclyclopedia of Statistical Science, Vol. 6, Wiley & Sons, New York, 655–659.

R Development Core Team, R: A language and environment for statistical computing, ISBN 3-900051-07-0, http://www.R-project.org, Vienna, 2008.

Tyler, D.E., Critchley, F., Dümbgen, L. and Oja, H., Invariant coordinate selection, 2008. Journal of Royal Statistical Society, Series B, (accepted).

Journal of Statistical Software

November 2008, Volume 28, Issue 6. http://www.jstatsoft.org/

Tools for Exploring Multivariate Data: The Package ICS

Klaus Nordhausen
University of Tampere

Hannu Oja
University of Tampere

David E. Tyler
Rutgers, The State University of New Jersey

Abstract

Invariant coordinate selection (ICS) has recently been introduced as a method for exploring multivariate data. It includes as a special case a method for recovering the unmixing matrix in independent components analysis (ICA). It also serves as a basis for classes of multivariate nonparametric tests, and as a tool in cluster analysis or blind discrimination. The aim of this paper is to briefly explain the (ICS) method and to illustrate how various applications can be implemented using the R package **ICS**. Several examples are used to show how the ICS method and **ICS** package can be used in analyzing a multivariate data set.

Keywords: clustering, discriminant analysis, independent components analysis, invariant coordinate selection, R, transformation-retransformation method.

1. Introduction

Multivariate data normally arise by collecting p measurements on n individuals or experimental units. Such data can be displayed in tables with each row representing one individual and each column representing a particular measured variable. The resulting data matrix X is then $n \times p$, with the row vector $x_i \in \Re^p$ denoting the measurements taken on the ith individual or ith experimental unit. Hence $X^\top = [x_1^\top, \ldots, x_n^\top]$. To be consistent with the convention used in the programming language R, all vectors are understood to be row vectors.

We begin by introducing some concepts which are used throughout the paper. An affine transformation of a data vector x_i is a transformation of the form

$$x_i \to x_i A^\top + b, \qquad i = 1, \ldots, n,$$

or equivalently,

$$X \to XA^\top + 1_n^\top b,$$

where A is a nonsingular matrix of order p, $b \in \Re^p$, and $1_n \in \Re^n$ denotes a vector consisting of all ones. Besides affine transformations and linear transformations ($X \to XA^\top$ with A nonsingular), other important classes of transformations are orthogonal transformations ($X \to XU$ with $U^\top U = UU^\top = I_p$), sign-change transformations ($X \to XJ$ where J is a $p \times p$ diagonal matrix with diagonal elements ± 1), and permutations ($X \to XP$ where P is a $p \times p$ permutation matrix, i.e., one obtained by successively permuting rows or columns of I_p). These transformations can also be applied on the left-hand side of X, in which case A, U, J, and P are matrices of order n rather than of order p. Note, for example, that a right-sided sign-change transformation simply results in a change of the sign of the jth variable if the jth entry of J is -1, whereas a left-sided permutation transformation simply reorders the individual observations.

A fundamental multivariate data transformation method is the so-called 'whitening' or 'standardization' of the data. This is given by

$$X \to Z = (X - 1_n^\top \bar{x})\mathsf{COV}(X)^{-\frac{1}{2}},$$

where $\bar{x} = 1_n X/n = \sum_{i=1}^n x_i/n$ is the vector of the column means of X and $\mathsf{COV}(X) = (X - 1_n^\top \bar{x})^\top (X - 1_n^\top \bar{x})/(n-1)$ is the sample covariance matrix of the columns of X. The 'whitened' data matrix Z has its mean at the origin ($\bar{z} = 0$), with all the variables being standardized and uncorrelated with each other ($\mathsf{COV}(Z) = I_p$).

This transformation, however, has several drawbacks. First, it is not unique in the sense that it depends on the particular choice or definition of the square-root matrix $\mathsf{COV}(X)^{\frac{1}{2}}$. Recall that for a symmetric positive semi-definite matrix V, a square root of V is any matrix C such that $CC^\top = V$. Two common choices for the square-root C which are uniquely defined are the lower triangular square-root and the symmetric positive definite square-root. Second, even for a well defined square-root matrix, the 'whitening' transformation is not invariant under affine transformations of the data. Rather, one has

$$XA^\top [\mathsf{COV}(XA^\top)]^{-\frac{1}{2}} = X[\mathsf{COV}(X)]^{-\frac{1}{2}} U,$$

with $U = U(X, A)$ being an orthogonal matrix which is dependent on both X and A. In other words, a 'whitened' data matrix is only well defined up to post multiplication by an orthogonal matrix.

Also, this 'whitening' transformation is not very robust since both the sample mean vector $\bar{x}$ and the sample covariance matrix $\mathsf{COV}(X)$ are both highly non-robust statistics. In particular, just one 'bad' data point can greatly affect the standardization. An obvious remedy for the last problem is to simply replace those two statistics by more robust ones. This gives a more general framework for a whitening transformation, namely

$$X \to Z = [X - 1_n^\top T(X)]S(X)^{-\frac{1}{2}},$$

where the statistic $T(X)$ is a multivariate location statistic and $S(X)$ is a scatter matrix. Here, we say that $T(X)$ is a location statistic if it is affine equivariant, i.e., if

$$T(XA^\top + 1_n^\top b) = T(X)A^\top + b$$

for any nonsingular matrix A of order p, and any $b \in \Re^p$. The matrix $S(X)$ is said to be a scatter matrix if it is affine equivariate in the following sense:

$$S(XA^\top + 1b) = AS(X)A^\top$$

with A and b as before. Such a statistic is sometimes referred to as being affine 'covariant'. Choosing robust statistics for $T(X)$ and $S(X)$ then yields a robustly whitened coordinate system. This new coordinate system though is still not invariant under affine transformations of the original data matrix.

Besides whitening, there are other methods one can use to linearly transform a multivariate data set to a new coordinate system, such as those arising in principal components analysis (PCA), those arising in independent components analysis (ICA), and those arising in invariant coordinate selection (ICS). Principal components analysis has a long history and is perhaps one of the most common methods used in multivariate analysis, whereas independent components analysis is a fairly recent subject which is becoming increasing popular in areas such as computer science, engineering, meteorology and other applied areas where multivariate data arise. Invariant coordinate selection has recently been introduced as a very general method for exploring multivariate data, and is explained in more detail in the Section 3.1. These three methods respectively involve the following transformations of the data (here we ignore the centering part of the transformations, which if desired could be done after the transformation).

- Principal components analysis
  The principal components are obtained by rotating the data matrix, namely

  $$X \to Z = XU^\top,$$

  where $U^\top$ is an orthogonal matrix whose columns are the ordered eigenvectors of $\mathsf{COV}(X)$. This gives $\mathsf{COV}(Z) = D$, with D being a diagonal matrix whose diagonal elements are equal to the corresponding ordered eigenvalues of $\mathsf{COV}(X)$. The matrices U and D thus correspond to those in the spectral value decomposition $\mathsf{COV}(X) = U^\top DU$. PCA can also be viewed as a rotation of the data matrix arising from first finding the projection of maximal variance, and then finding subsequent projections of maximal variance subject to the constraint of being uncorrelated with the previously extracting projections.

- Independent components analysis
  Unlike principal components analysis, ICA transformations presume a model. The most common model is to presume that the p measured variables arise from a linear transformation of p independent variables. The goal of ICA is to recover the original independent variables. Most ICA algorithms first involve whitening the data and then rotating them in such a way as to make the resulting components as independent as possible. When the components are derived sequentially, this typically implies finding the 'most' nongaussian projection subject to being uncorrelated to the previously extracted projections. Such ICA transformations then have the form

  $$X \to Z = X\mathsf{COV}(X)^{-\frac{1}{2}}Q = XB^\top$$

where Q is an orthogonal matrix. The matrix B is typically called the unmixing matrix. For ways to choose the final rotation matrix Q, or more generally for a review of ICA, see Hyvärinen, Karhunen, and Oja (2001).

- Invariant coordinate selection
  The ICS transformation is based upon the use of two different scatter matrices. One scatter statistic is first used to 'whiten' the data, while a second scatter statistic, defined differently from the first, is used to find a rotation of the data obtained from a PCA of the 'whitened' data. Specifically, this gives the transformation

$$X \to Z = XS_1(X)^{-\frac{1}{2}}U_2^\top = XB^\top,$$

  where U_2 is given by the spectral value decomposition of $S_2(Z_1) = U_2^\top DU_2$ for $Z_1 = XS_1(X)^{-\frac{1}{2}}$. As described later in the paper, this new coordinate system is invariant up to a sign change under affine transformations of the original data matrix X.

The goal of this paper is to describe how the ICS method, as well as a certain class of ICA algorithms, can be implemented using the R package **ICS**. The structure of this paper is as follows. In the next section, a review of some scatter matrices to be used later within the paper is first given. Section 3 explains the ICS method in more detail and discusses its various applications. One such application involves the recovery of the unmixing matrix in the ICA problem. Section 4 describes the R package **ICS**, and finally Section 5 concludes the paper with several examples showing how the ICS method and package can be used in analyzing a multivariate data set.

2. Scatter matrices

Conceptually, the simplest alternative to the sample mean and sample covariance matrix is to use a weighted mean and covariance matrix respectively, with the weights dependent on the original Mahalanobis distances. This gives the location and scatter statistics

$$T(X) = \frac{\text{ave}[u_1(r_i)x_i]}{\text{ave}[u_1(r_i)]} \quad \text{and} \quad S(X) = \text{ave}[u_2(r_i)(x_i - \bar{x})^\top(x_i - \bar{x})],$$

where $r_i = ||x_i - \bar{x}||_{\mathsf{COV}(X)}$, and with $u_1(r)$ and $u_2(r)$ being non-negative weight functions. Here, we use the general notation

$$||y||_\Gamma^2 = y\Gamma^{-1}y^\top,$$

which defines a norm on $\Re^p$ whenever Γ is a symmetric positive definite matrix of order p. Since a single outlier can greatly affect the value of all of the Mahalanobis distances, the weighted mean and covariance statistics can be also highly sensitive to a single outlier.

Many classes of robust location and scatter statistics have been proposed. For our purposes, we briefly discuss only the multivariate M-estimates. For a detailed overview of the M-estimates and other robust estimates, we refer the reader to Maronna, Martin, and Yohai (2006).

The multivariate M-estimates of location and scatter may be viewed as adaptively weighted means and covariance matrices respectively. More specifically, they can be defined as solutions

to the M estimating equations

$$T(X) = \frac{\text{ave}[u_1(r_i)x_i]}{\text{ave}[u_1(r_i)]}, \quad \text{and} \quad S(X) = \text{ave}[u_2(r_i)(x_i - T(X))^\top (x_i - T(X))],$$

where now $r_i = ||x_i - T(X))||_{S(X)}$, and again $u_1(r)$ and $u_2(r)$ are non-negative weight functions. Note that these are implicit equations in $(T(X), S(X))$ since the weights on the right-hand side of the equations depend upon them.

The multivariate M-estimates of location and scatter were originally derived as a generalization of the maximum likelihood estimates for the parameters of an elliptically symmetric distribution. Hence these maximum likelihood estimates are special cases of the multivariate M-estimates. Of particular interest here are the maximum likelihood estimates associated with a p-dimensional elliptical t-distribution on ν degrees of freedom. These maximum likelihood estimates correspond to M-estimates with weight functions

$$u_1(r) = u_2(r) = \frac{p + \nu}{r^2 + \nu}.$$

An important property of these t M-estimates, for $\nu \geq 1$, is that they are one of the few M-estimates which are known to have a unique solution to its M-estimating equations and a proven convergent algorithm, see Kent and Tyler (1991).

A useful variation of a scatter matrix is a scatter matrix with respect to the origin. We defined this to be a statistic $S_o(X)$ which is invariant under sign changes of the individual observations and equivariant or 'covariant' under nonsingular linear transformations. That is,

$$S_o(JXA^\top) = AS_o(X)A^\top$$

for any nonsingular matrix of order p and any sign change matrix J of order n. An example of a scatter matrix about the origin is the matrix of second moments $M_2(X) = \text{ave}[x_i^\top x_i]$. Other examples are weighted second moment matrices and M-estimates of scatter about the origin. These are defined as

$$S_o(X) = \text{ave}[u_2(r_i)x_i^\top x_i],$$

with $r_i = ||x_i||_{M_2(X)}$ for the former and $r_i = ||x_i||_{S_o(X)}$ for the latter.

One important application of scatter matrices with respect to the origin is that they can be used to construct symmetrized scatter matrices. A symmetrized scatter matrix $S_s(X)$ is a scatter matrix defined by applying a scatter matrix with respect to the origin to pairwise differences of the data. More specifically, given a scatter functional with respect to the origin S_o, a symmetrized scatter matrix is then defined as

$$S_s(X) = S_o(X_s),$$

where X_s is $N = n(n-1)/2 \times p$ with row vectors $d_{i,j} = x_i - x_j$ for $i < j$. Note that a location statistic is not needed in defining a symmetrized scatter statistic. As explained later in Section 3.4, symmetrized scatter matrices play a crucial role when using ICS for independent components analysis.

Another scatter matrix which plays a role in independent components analysis involves the 4th central moments (Cardoso 1989). This is given by

$$\mathsf{COV}_4(X) = \frac{1}{p+2}\text{ave}[r_i^2(x_i - \bar{x})^\top (x_i - \bar{x})]$$

where $r_i = ||x_i - \bar{x}||_{\mathsf{COV}(X)}$. This scatter matrix is a special case of a weighted sample covariance matrix, namely one with weight function $u_2(r) = r^2/(p+2)$. A curious observation is that this weight function upweights rather than downweights outliers. The constant $1/(p+2)$ is used to make $\mathsf{COV}_4(X)$ consistent for the covariance matrix under random samples from a multivariate normal distribution.

Finally, a popular M estimates of scatter within the area of nonparametric multivariate statistics is Tyler's shape matrix (Tyler 1987). For a given location functional $T(X)$, this is defined as a solution to the implicit equation

$$S(X) = p \text{ ave} \left[\frac{(x_i - T(X))^\top (x_i - T(X))}{||x_i - T(X)||^2_{S(X)}} \right].$$

Tyler's shape matrix about the origin is obtained by simply setting $T(X) = 0$ in the above definition, which corresponds to an M estimate of scatter about the origin with weight function $u_2(r) = 1/r^2$. The symmetrized version of Tyler's shape matrix is known as Dümbgen's shape matrix (Dümbgen 1998). It is implicitly defined by

$$S_s(X) = p \text{ ave}_{i<j} \left[\frac{(x_i - x_j)^\top (x_i - x_j)}{||x_i - x_j||^2_{S_s(X)}} \right].$$

Tyler's shape matrix and Dümbgen's shape matrix are not well defined scatter matrices since they are defined only up to a constant. That is, if $S(X)$ and $S_s(X)$ satisfy the above definitions respectively, then so do $\lambda S(X)$ and $\lambda S_s(X)$ for any $\lambda > 0$. This however is the only indeterminancy in their definitions. Consequently, they possess the following equivariant property under affine transformations,

$$S(XA^\top + 1_n^\top b) \propto AS(X)A^\top,$$

for any nonsingular matrix A of order p and any $b \in \Re^p$. For the applications discussed in this paper this equivariant property is sufficient.

3. Multivariate data analysis using an ICS

3.1. Invariance of ICS

As noted in the introduction, using the sample mean and covariance matrix or some robust affine equivariate alternatives, say $(T(X), S(X))$, to 'whiten' a multivariate data set yields a new 'standardized' coordinate system in the sense that the 'new' data set has uncorrelated components with respect to S. This new coordinate system, however, is not invariant under affine transformations of the original data set X since for nonsingular A and $b \in \Re^p$, one obtains

$$[(XA^\top + 1_n^\top b) - 1_n^\top T(XA^\top + 1_n^\top b)]S(XA^\top)^{-\frac{1}{2}} = [X - 1_n^\top T(X)][S(X)]^{-\frac{1}{2}}U,$$

with U being an orthogonal matrix depending on X, A and S, and on the particular definition of the matrix square-root being used. Thus, 'standardizing' X does not necessarily give the same coordinate system as 'standardizing' $XA^\top + 1_n^\top b$.

Tyler, Critchley, Dümbgen, and Oja (2008) show however that an affine invariant 'whitening' of the data can be obtained by introducing a second scatter statistic. They call this transformation invariant coordinate selection ICS. The definition of ICS as given in the introduction can be seen as a two step transformation. First the data is 'standardized' with respect to one scatter statistic $S_1(X)$ and then a PCA transformation is performed on the 'standardized' data using a different scatter statistic $S_2(X)$. Note that if one applies the same scatter statistic $S_1(Z)$ to the 'standardized' data, then one simply obtains $S_1(Z) = I_p$, for which a PCA transformation is meaningless.

An alternative formulation of ICS, which makes some of its properties more transparent, is as follows. For two different scatter statistics $S_1(X)$ and $S_2(X)$, let $B(X)$ be the $p \times p$ matrix whose rows corresponds to the eigenvectors of $S_1(X)^{-1}S_2(X)$ and let $D(X)$ be the diagonal matrix consisting of the p corresponding eigenvalues. For brevity, denote $S_1 = S_1(X)$, $S_2 = S_2(X)$, $B = B(X)$ and $D = D(X)$, and so

$$S_1^{-1}S_2B^\top = B^\top D \quad \text{or} \quad S_2B^\top = S_1B^\top D$$

Note that any matrix B satisfying the above definition also jointly diagonalizes both S_1 and S_2. This gives

$$BS_1B^\top = D_1 \quad \text{and} \quad BS_2B^\top = D_2,$$

with D_1 and D_2 being diagonal matrices. Moreover, $D_1^{-1}D_2 = D$. If the roles of S_1 and S_2 are reversed, then the matrix of eigenvectors B is the unchanged, but $D \to D^{-1}$.

We hereafter use the convention of normalizing the eigenvectors to have length one relative to the scatter matrix S_1, i.e.,

$$BS_1B^\top = I_p,$$

and hence $D = D_2$. We also presume the eigenvalues, i.e., the diagonal elements of D, are ordered. The resulting transformation matrix B corresponds to that given in the introduction. The matrix B can be made unique by imposing some restrictions like the element in each row with largest absolute value must be positive.

The transformation defined by the matrix $B(X)$, i.e.,

$$X \to Z = XB(X)^\top$$

is invariant under nonsingular linear transformations in the following sense. Presuming the eigenvalues in $D(X)$ are all distinct, it follows that for any nonsingular matrix A

$$X_* = XA^\top \to Z_* = X_*B(X_*)^\top = (XA^\top)B(XA^\top)^\top = XB(X)^\top J = ZJ,$$

for some sign change matrix J. A similar statement can be made in the case of multiple eigenvalues, see Tyler *et al.* (2008) for details. Given an affine equivariant location statistic $T(Y)$, if either the variable X or the transformed data Z is center by subtracting $T(X)$ or $T(Z)$ respectively from each of the rows, then the resulting transformation is affine invariant up to a sign change matrix. Finally, we note that the eigenvalues are also affine invariant. Specifically,

$$D(XA^\top + 1_n^\top b) = D(X).$$

Thus, given two scatter statistics, one can easily generate an invariant coordinate system. For the most part, which scatter statistics are best to use is an open problem. Most likely it

depends on the particular application in mind. The following sections show some applications of using ICS in multivariate data analysis, and points out those situations where certain types of scatter matrices are needed.

3.2. Descriptive statistics

In this section, let $Z = XB(X)^\top$ be the invariant components obtained from ICS based on the scatter statistics $S_1(Y)$ and $S_2(Y)$. The components of Z are thus standardized with respect to S_1 and uncorrelated with respect to S_2, i.e.,

$$S_1(Z) = I \quad \text{and} \quad S_2(Z) = D,$$

where D is an ordered diagonal matrix. We hereafter refer to the diagonal elements of D as generalized kurtosis measures. In the univariate setting, the classical kurtosis measure can be viewed as a comparison of two different univariate dispersion measures, namely the square-root of the fourth central moment and the variance. The ratio of any two dispersion measures can be used to define a generalized univariate kurtosis measure. In the multivariate setting, one can consider the 'ratio' of two different scatter matrices $S_1(X)^{-1}S_2(X)$. The maximal invariants under nonsingular linear or under affine transformations can then be shown to be D, and thus we view D as a multivariate affine invariant generalized kurtosis measure, again see Tyler *et al.* (2008) for details. The individual elements of D represent a generalized kurtosis measure for the corresponding components of Z, with these components being ordered according to their generalized kurtosis.

Consequently, the two scatters and the ICS transformation along with two different location statistics (denoted correspondingly as T_1 and T_2) can be used to describe four of the most basic features of a data set:

- The location: $T_1(X)$
- The scatter: $S_1(X)$
- Measure of skewness: $T_2(Z) - T_1(Z)$
- Kurtosis measures: $S_2(Z)$

The last two measures can even be used to construct tests of multinormality or ellipticity. The usage of two different location and scatter statistics for such tests is described in more detail in Kankainen, Taskinen, and Oja (2007).

3.3. Diagnostic plots and dimension reduction

Perhaps the most common type of diagnostic plot for multivariate data is the classical Mahalanobis distance plots based upon the sample mean vector and sample covariance matrix. Such a plot, i.e., a plot of the index i versus the Mahalanobis distance $r_i = ||x_i - \bar{x}||_{\mathsf{COV}(X)}$, can be useful in detecting outliers in the data. Such plots though are known to suffer from the masking problem. To alleviate this problem, one can replace the sample mean and covariance by robust location and scatter statistics respectively, and then generate robust Mahalanobis distance plots, see e.g., Rousseeuw and van Zomeren (1990). Another type of diagnostic plot,

e.g., used in Rousseeuw and van Driessen (1999) to help uncover outliers or groups of outliers, is a plot of the classical Mahalanobis distances versus the robust Mahalanobis distances.

One feature of Mahalanobis distance plots is that they are invariant under affine transformations of the data. Given two location and scatter statistics, one can plot the corresponding Mahalanobis distances against each other. However, a more complete affine invariant view of the data is given by the pairwise plots of the invariant coordinates Z obtained from ICS. The ordering of the components of Z is with respect to their generalized kurtosis measures. Moreover, if we take

$$\kappa(b) = bS_2b^\top / bS_1b^\top$$

as a generalized kurtosis measure for the univariate linear combination $Xb^\top$, then $\kappa(b)$ achieves its maximum at the first component of Z and its minimum at the last component of Z. The other components of Z successively maximize or minimize $\kappa(b)$ subject to being 'uncorrelated' relative to S_1 or S_2, with the previously extracted components, e.g., $b_1S_1b_2^\top = b_1S_2b_2^\top = 0$. Extreme kurtosis measures can indicate non-normality of coordinates and hence indicate coordinates which may be of special interest for further examination. Thus, focusing on the 'extreme' ICS components yields a natural method for dimension reduction. This criterion for dimension reduction is demonstrated in Section 5.2.

The ICS transformation is also known to have other important properties which justifies its use as a data analytic method. For example, if the data arise as a location mixture of two multivariate normal distributions, or more general two possibly different elliptical distributions, with proportional population scatter matrices, then Fisher's linear discriminant function for discrimination between the two components of the mixture corresponds to one of the two extreme ICS components even though the classification of the data points are not known. For more details and generalizations to mixtures with more than two components, we again refer the reader to Tyler *et al.* (2008). Another important property of ICS is its relationship to independent components analysis, which is discussed in the next section.

Special cases of the ICS transformation have been proposed as diagnostic methods for detecting outliers or groups of outliers by Caussinus and Ruiz-Gazen (1994) and more recently by Critchley, Pires, and Amado (2008). The former consider the case when S_1 is taken to be the sample covariance matrix and S_2 is taken to be a weighted sample covariance matrix as defined in Section 2. They refer to their method as generalized principal components analysis (GPCA). Critchley *et al.* (2008) also consider the case when S_1 is taken to be the sample covariance matrix. For S_2, they use a weighted covariance matrix based upon the weight function $u_2(r_i) = 1/r_i^2$, and they refer to their method as principal axis analysis (PAA). Since these are special cases of ICS, the R package **ICS** can be used to implement GPCA or PAA.

3.4. Independent components analysis

So far no assumptions have been made as to how the data arises, other than the reference to mixture models in the previous section. In this section, we now assume the observations represent a random sample from a multivariate population, with the population representing a nonsingular linear transformation of a vector of independent components. More specifically, the observations

$$x_i = z_iA^\top, \quad i = 1, \ldots, n$$

where the mixing matrix A is a full rank $p \times p$ matrix A, and z_i is a p-variate latent vector with independent components. This is the independent components (IC) model in its simplest

form. The aim of independent components analysis (ICA) is to find an unmixing matrix B so that $x_i B^\top$ has independent components. Note that this model is not well defined since for any diagonal matrices D and permutation matrices P

$$X = Z^* A^{*\top} = (ZPD)(D^{-1}P^{-1}A^\top).$$

Therefore for any unmixing matrix B, $B^* = DPB$ is also a valid unmixing matrix. For a recent overview about ICA see Hyvärinen *et al.* (2001).

Oja, Sirkiä, and Eriksson (2006) show, under fairly general conditions, that the transformation matrix B defined in Section 3.1 is also an unmixing matrix for the IC model. One condition is that the population values of the generalized kurtosis values for the independent components of z_i have different values. Another condition is that the population version of the scatter matrices S_1 and S_2 posses the co-called 'independence property'. This independence property requires that if z_i has independent components, then the population version of $S(Z)$ is a diagonal matrix. In general, scatter matrices do not necessarily possess this property, but symmetrized scatter matrices do.

The regular covariance matrix COV and the matrix of 4th moments COV_4 also possess the aforementioned independence property since they can be represented as symmetrized scatter matrices. Consequently, the FOBI algorithm (Cardoso 1989), can be seen as a special case of the ICS based algorithm with $S_1 = COV$ and $S_2 = COV_4$. For this case, it turns out that that the generalized kurtosis measure D_{jj} can be transformed into an estimate of the classical kurtosis measure for the j^{th} independent components, specifically by taking $\widehat{\kappa}_j = (p+2)(D_{jj} - 1)$. Simulations given in Nordhausen, Oja, and Ollila (2008a) indicate the performance of the algorithm is better when more robust scatter functionals are used.

3.5. Multivariate nonparametrics

Multivariate extensions of univariate signs, ranks and the median can be easily obtained by applying signs, ranks and medians to the individual components of a multivariate dataset. Such componentwise or marginal signs, ranks and median, as well as spatial signs, spatial ranks and spatial median, are not invariant or equivariant under affine transformations of the data. This lack of invariance is partially responsible for the lack of power or efficiency when the data are highly correlated, see e.g. Bickel (1965) and Puri and Sen (1971).

To construct invariant tests and estimates using such multivariate signs and ranks, Chakraborty and Chaudhuri (1996), Chakraborty and Chaudhuri (1998) and Chakraborty, Chaudhuri, and Oja (1998) introduced the 'transformation-retransformation' (TR) technique. The TR method first linearly transforms the data to a new invariant coordinate system, and then the marginal tests or estimates are constructed on the transformed coordinates. Finally, estimates can then be retransformed to the original coordinate system. The transformation used in the TR technique in one sample problems is based on the selection of p data points. The data is then linearly transformed so that these p data points are mapped into the Euclidean basis vector. A major difficulty with the TR procedure involves the selection of the 'best' p data vectors.

In this section we discuss how the ICS transformation can be used as a simpler alternative in the construction of invariant componentwise tests and estimates. We concentrate here on the signs, ranks, and medians for the one sample problem. For applications of ICS in the two sample problem see Nordhausen, Oja, and Tyler (2006).

Suppose X arises as a random sample from a p-variate continuous distribution. Further, assume its distribution is symmetric about some unknown location parameter μ. In other words, the distributions of $(x_i - \mu)$ and $-(x_i - \mu)$ are assumed to be the same. Consider first the problem of testing the null hypothesis $H_0 : \mu = 0$. To apply the ICS method to this testing problem, we require now that the two scatter matrices be scatter matrices with respect to the origin, as defined in Section 2, and to be invariant under permutations of the data points. Hence, for $k = 1, 2$, we require

$$S_k(PJXA^\top) = AS_k(X)A^\top$$

for any nonsingular A, permutation P and sign-change J, which then implies

$$B(PJX) = B(X).$$

Under the null hypothesis, the rows of X represent a random sample from a distribution symmetric about the origin. Hence, the distribution of X and PJX are the same for any permutation matrix P and any sign-change matrix J. Consequently, for such P and J,

$$Z(X) = XB(X)^\top \sim_d PJZ(X).$$

Note that the rows of Z, z_i for $i = i, \ldots, n$, do not represent a random sample since the transformation matrix $B(X)$ is data dependent. Nevertheless, under the null hypothesis, the n observations in Z have an exchangeable and symmetric distribution.

Consider now the jth component or column of Z which corresponds to $(z_{1j}, \ldots, z_{nj})^\top$. It then readily follows that under the null hypothesis

$$U_j = \sum_{i=1}^{n} I(z_{ij} > 0) \sim_d Bin(n, 0.5)$$

for each $j = 1, \ldots, p$. Hence, the sign test statistic U_j is distribution-free and invariant under any nonsingular linear transformation of the data. Likewise, if we denote R_{ij}^+ to be the rank of $|z_{ij}|$ among $|z_{1j}|, \ldots, |z_{nj}|$, then the Wilcoxon signed-rank statistic

$$W_j = \sum_{i=1}^{n} \text{sgn}(z_{ij}) R_{ij}^+$$

is also distribution-free under the null hypothesis, specifically it has the distribution of the univariate Wilcoxon sign-rank statistic, and is similarly invariant. Note that these test statistics are distribution-free under any symmetric model and not only under elliptically symmetric models. Of course, other score functions can be used in an analogous manner.

Unfortunately, $U_1, \ldots, U_p$ as well as $W_1, \ldots, W_p$ are not mutually independent and their corresponding joint distributions are not distribution-free under the null hypothesis. Exact finite sample distribution-free tests can be constructed though if one uses only one of the extreme ICS, specifically U_1 or U_p for the sign test or W_1 or W_p for the sign-rank test. Which extreme should be used depends on the choice of the scatter statistics used in the ICS transformation. Alternatively, conservative finite sample distribution-free tests can be constructed if one uses each of the test statistics, that is either $U_1, \ldots, U_p$ or $W_1, \ldots, W_p$, together with Bonferonni's method. Another alternative is to combine the individual test

statistics, either the two extremes or all p, to form approximate χ^2 statistics as described in Puri and Sen (1971).

Nordhausen *et al.* (2006) compare the efficiencies of the following three strategies: (i) using only one of the extreme components, (ii) using an approximate χ^2 statistic based on the first and last component, and (iii) using an approximate χ^2 statistic based on all the components. Although the exact as well as the asymptotic distribution of (ii) and (iii) are still open questions, the efficiency comparisons showed that a χ^2_p approximation works well for strategy (iii). Furthermore, strategy (iii) using the Wilcoxon signed-rank statistics appears to be the best test statistic among these, and is a serious competitor to Hotelling's T^2 test even at the multivariate normal model. These tests using signs and ranks in an ICS are not only distribution-free under elliptically symmetric models but rather under any symmetric model.

To obtain an affine equivariant location estimate in this setting, let $\hat{\mu}$ be either the vector of marginal medians or the vector or marginal Hodges-Lehmann estimators. These, by themselves, are not a true multivariate location statistics since they are not affine equivariant. However, they can be applied to the ICS transformed coordinates (where the scatter matrices now are not taken with respect to the origin), and then transformed back to the original coordinates. This gives

$$\tilde{\mu}(X) = \hat{\mu}(XB^\top)(B^{-1})^\top,$$

where $B = B(X)$ is the ICS transformation matrix. The resulting statistic $\tilde{\mu}(X)$ then corresponds to an affine equivariant multivariate median, or respectively Hodges-Lehmann estimator. Applying this method with any other univariate location statistics yields an affine equivariant multivariate version of the statistic. Note that if the univariate statistic is the sample mean, then the resulting multivariate statistic is the usual multivariate sample mean.

4. ICS and R

The package **ICS** is freely available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=ICS` and comes under the GNU General Public Licence (GPL) 2.0 or higher licence.

The main function of the package **ICS** is the function `ics`. This function computes for a given numeric data frame or matrix the unmixing matrix, the (generalized) kurtosis values and the invariant coordinates. The function is written in a flexible way so that the user can choose for their computations any two scatter functions desired. The user can either submit the name of two arbitrary functions that return a scatter matrix or submit two scatter matrices already computed in advance to the arguments `S1` and `S2`.

In principle after deciding on two scatter matrices which scatter matrix is chosen as S_1 and which as S_2 makes no difference. The effect of relabeling S_1 and S_2 is that the coordinate order is reversed and that the kurtosis values are inverted. The later is however only the case when S_1 and S_2 are both actual scatter matrices and none of them is a shape matrix. If one or both of S_1 and S_2 are shape matrices, the product of the kurtosis after reversing one of the vectors is no longer 1 anymore but only constant since in this case, the kurtosis measures are only relative.

To avoid arbitrary scales for the kurtosis values and in order to make them also more comparable, the logical argument `stdKurt` can be used to decide if one wants the absolute values of

the kurtosis measures or one rather wants them standardized in such a way, that the product of the kurtosis elements is 1.

The best choice for S_1 and S_2 for a given data set is still an open question, in most cases the choice seem not to have a very big effect, in some cases however as shown for example in Tyler *et al.* (2008) it can have a substantial effect. Also the choice can depend heavily on the application. When, for example, the estimation of the mixing matrix in an independent components analysis is the goal, then the simulation study given in Nordhausen *et al.* (2008a) shows that robust combinations always dominate non-robust combinations, even when there are no outliers present. Whereas in Nordhausen, Oja, and Paindaveine (2008b) the combination of scatter functionals had no impact on the efficiency of a test for location in the symmetric independent component model, where ICS was used to recover the independent components. In general, given the current knowledge of ICS, we recommend trying several combinations of scatter matrices for S_1 and S_2. Here, R offers many possibilities. The package **ICS** itself offers, for example, the matrix of fourth moments (`cov4`), the covariance matrix with respect to the origin (`covOrigin`), a one-step Tyler shape matrix (`covAxis`) or an M estimator based on the t distribution (`tM`). Other packages offer still more scatter matrices. The following list names a few functions from different packages. For details about the functions see the corresponding help pages.

- **covRobust** (Wang, Raftery, and Fraley 2003): `cov.nnve`.
- **ICSNP** (Nordhausen, Sirkiä, Oja, and Tyler 2007): `tyler.shape`, `duembgen.shape`, `HR.Mest`, `HP1.shape`.
- **MASS** (Venables and Ripley 2002): `cov.rob`, `cov.trob`.
- **robustbase** (Mächler, Rousseeuw, Croux, Todorov, Ruckstuhl, and Salibian-Barrera 2008): `covMcd`, `covOGK`.
- **rrcov** (Todorov 2008): `covMcd`, `covMest`, `covOgk`.

Naturally the user should ascertain that the scatter matrices he uses have all the different properties like affine equivariance or independence property and so on, needed for the application at hand. The application has also an impact on the preferred form of the unmixing matrix B, which as mentioned above is not unique. The function `ics` offers two options via the argument `stdB`. Setting this argument to `Z` standardizes the unmixing matrix B in such a way, that all invariant coordinates are right skewed. The criterion used to achieve this is to use the sign between the mean and median of each component. Whereas the option `stdB = "B"` standardizes the ummixing matrix such that each row has norm 1 and in each row the element with the largest absolute value has a positive sign. The later method is more natural in an independent component model framework.

A call to the function `ics` creates an object of the S4 class `ics` and the package offers several functions to work with such objects. The two most basic ones are the functions `show` (equivalent to `print`) for a minimal output and `summary` for a more detailed output. The generic function `plot` for an `ics` object returns a scatter plot matrix which shows by default when $p > 7$ only those components with the three smallest kurtosis measures and the three largest kurtosis measures, since often the main interest is on the components with 'extreme' kurtosis values. However using the `index` argument any component can be included

or excluded in the scatterplot. Another plotting function for an `ics` object is the generic `screeplot.ics` which works similar as R's function `screeplot` for principal components with the difference, that it plots the kurtosis values against the number of the component. The function `fitted` returns the original data but it can also be used in the ICA framework when some components may be suppressed. The invariant coordinates or independent components can be obtained by calling `ics.components`. The transformation matrix or unmixing matrix B can be extracted from an `ics` object by using `coef`.

Not mentioned so far is, that the package offers also two tests for multinormality. The function `mvnorm.skew.test` is based on the difference between the mean vector and the vector of third moments, implemented as `mean3`. And in the same spirit compares `mvnorm.kur.test` the regular covariance matrix and covariance matrix of fourth moments.

For further details on the functions see their help pages and the references therein.

5. Examples for multivariate data analysis using an ICS

In this section we will present how to use **ICS** for the different purposes previously discussed. For the examples we use for the output the option `options(digits = 4)` in R 2.7.1 (R Development Core Team 2008) together with the packages **ICS** 1.2-0, **ICSNP** 1.0-2 (Nordhausen *et al.* 2007), **MASS** 7.2-44 (Venables and Ripley 2002), **mvtnorm** 0.9-2 (Genz, Bretz, and Hothorn 2008), **pixmap** 0.4-9 (Bivand, Leisch, and Mächler 2008) and **robustbase** 0.4-3 (Mächler *et al.* 2008). Random seeds are provided for reproducibility of all examples.

5.1. Descriptive statistics

The first example will show how to obtain the four summary statistics from Section 3.2 using the the regular covariance matrix, the matrix of fourth moments, the mean vector and the location estimated based on third moments. At the beginning we will load the needed packages, create a random sample from a multivariate normal distribution, and create our ICS. Note that due to our interest in the kurtosis the absolute kurtosis values are needed.

```
R> library("ICS")
R> library("mvtnorm")
R> set.seed(2)
R> X <- rmvnorm(1000, c(0, 0, 1))
R> ics.X <- ics(X, stdKurt = FALSE)
R> Z <- ics.components(ics.X)
```

The first summary statistic is the vector of means:

```
R> colMeans(X)
```

```
[1] 0.06200 0.02102 1.06619
```

The second summary statistic is the covariance matrix:

```
R> cov(X)
```

```
          [,1]      [,2]      [,3]
[1,]  1.030117  0.033838 -0.001187
[2,]  0.033838  0.988104 -0.003239
[3,] -0.001187 -0.003239  1.037944
```

The skewness measures are:

```
R> mean3(Z) - colMeans(Z)

      IC.1       IC.2       IC.3
 0.0010350  0.0135414 -0.0002974
```

Finally, as noted in Section 3.4, for this special case of ICS we can estimate the excess kurtosis values from the generalized kurtosis measures of the `ics` object as follows:

```
R> (dim(X)[2] + 2) * (ics.X@gKurt - 1)

[1]  0.40294  0.02736 -0.22158
```

5.2. Diagnostic plots and dimension reduction

Exploratory data analysis is often used to get some understanding of the data at hand, with one important aspect being the possible occurrence of atypical observations. Sometimes the identification of these atypical observations is the goal of the data analysis, more often however they must be identified and dealt with in order to assure the validity of inferential methods. Most classical methods are not very robust when the multinormality assumption is violated, and in particular when outliers are present.

Mahalanobis distance plots are commonly used to identify outliers. As we will demonstrate now, outliers can also be identified using ICS. The example we use here is the modified wood gravity data set which is for example part of the **robustbase** package as the data set `wood`. This data set consists of 20 observations for six variables, with a few of the observations being known outliers inserted into the data. This is a common data set used to demonstrate the need for robust scatter matrices, and in particular high breakdown point scatter matrices, to identify the outliers.

To demonstrate this idea, we first compute the Mahalanobis distances based on the sample mean vector and the sample covariance matrix and then one based on the minimum volume ellipsoid (MVE) estimate as implemented by `cov.rob` in the **MASS** package. Points which have distances larger than $\sqrt{\chi^2_{p;0.975}}$ are usually viewed as potential outliers, and so we will label such points accordingly.

```
R> library("MASS")
R> library("ICS")
R> data("wood", package = "robustbase")
R> maha1.wood <- sqrt(mahalanobis(wood, colMeans(wood), cov(wood)))
R> set.seed(1)
R> covmve.wood <- cov.rob(wood)
```

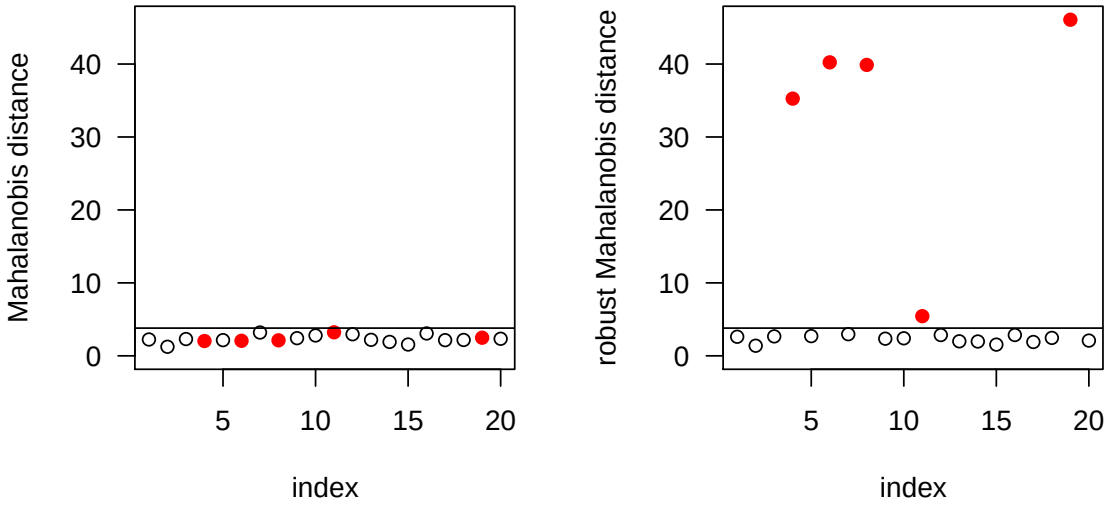

Figure 1: Mahalanobis distance plots for the `wood` data set. The red points are according to the robust Mahalanobis distances outliers.

```
R> maha2.wood <- sqrt(mahalanobis(wood, covmve.wood$center, covmve.wood$cov))
R> max.maha.wood <- max(c(maha1.wood, maha2.wood))
R> out.id <- ifelse(maha2.wood <= sqrt(qchisq(0.975, 6)), 0, 1)
```

It is worth noting that `cov.rob` in the **MASS** package does not actually give the raw MVE but rather a reweigthed scatter matrix which uses the location and scatter from the MVE as the initial statistics.

We next plot the distances against the observation number, include a horizontal line at the cutoff value and color the points that exceed the cutoff according to the robust distances.

```
R> par(mfrow = c(1, 2), las = 1)
R> plot(maha1.wood, xlab = "index" ,ylab = "Mahalanobis distance",
+    ylim = c(0, max.maha.wood), col = out.id + 1, pch = 15 * out.id + 1)
R> abline(h = sqrt(qchisq(0.975, 6)))
R> plot(maha2.wood, xlab = "index", ylab = "robust Mahalanobis distance",
+    ylim = c(0, max.maha.wood), col = out.id + 1, pch = 15 * out.id + 1)
R> abline(h = sqrt(qchisq(0.975, 6)))
R> par(mfrow = c(1, 1))
```

As can be seen from Figure 1, the classical Mahalanobis distances do not reveal any outlier whereas the robust distances classify 4 points as clear outliers and one borderline case.

The difference between the two Mahalanobis distances can also be observed in a distance versus distance plot, which ideally should have all points on the bisector. The results of the following code are given in Figure 2.

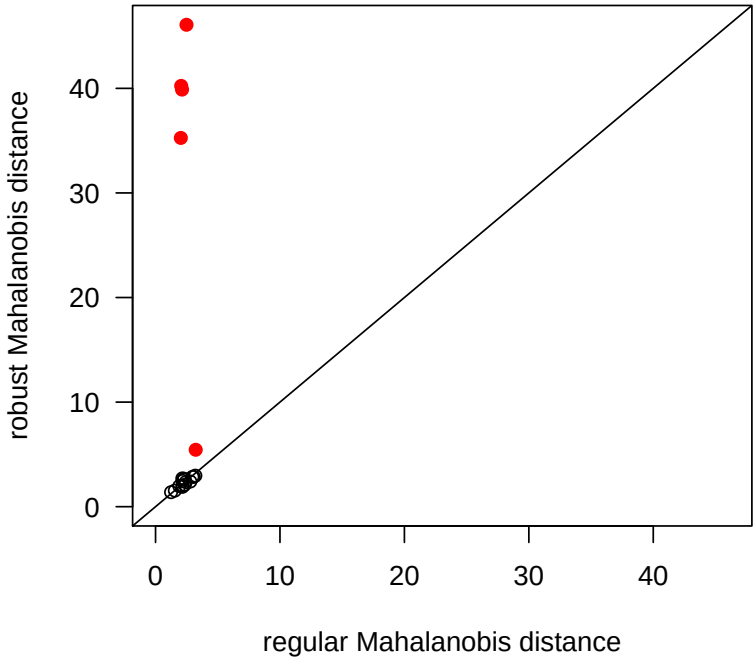

Figure 2: Distance distance plots for the `wood` data set. The red points are according to the robust Mahalanobis distances outliers.

```
R> plot(maha1.wood, maha2.wood, xlab = "regular Mahalanobis distance",
+    ylab = "robust Mahalanobis distance", ylim = c(0, max.maha.wood),
+    xlim = c(0, max.maha.wood), col = out.id + 1, pch = 15 * out.id + 1,
+    las = 1)
R> abline(0, 1)
```

For outlier identification, it is usually necessary to use Mahalanobis distances based on robust location and scatter statistics. Although, we still advise using robust scatter statistics for ICS, identifying atypical observations using ICS tends to be less dependent on the robustness properties of the scatter matrices being used. As an example, we fit here three different ICS systems based on three different combinations of scatter matrices for the wood data set, and observe that the choice of S_1 and S_2 does not seem to greatly affect the results.

```
R> library("ICSNP")
R> my.HR.Mest <- function(X,...) HR.Mest(X,...)$scatter
R> ics.default.wood <- ics(wood)
R> ics.2.wood <- ics(wood, tM(wood)$V, tM(wood, 2)$V)
R> ics.3.wood <- ics(wood, my.HR.Mest, HP1.shape)
R> par(mfrow=c(1, 3), las = 1, mar = c(5, 4, 1, 1) + 0.1)
```

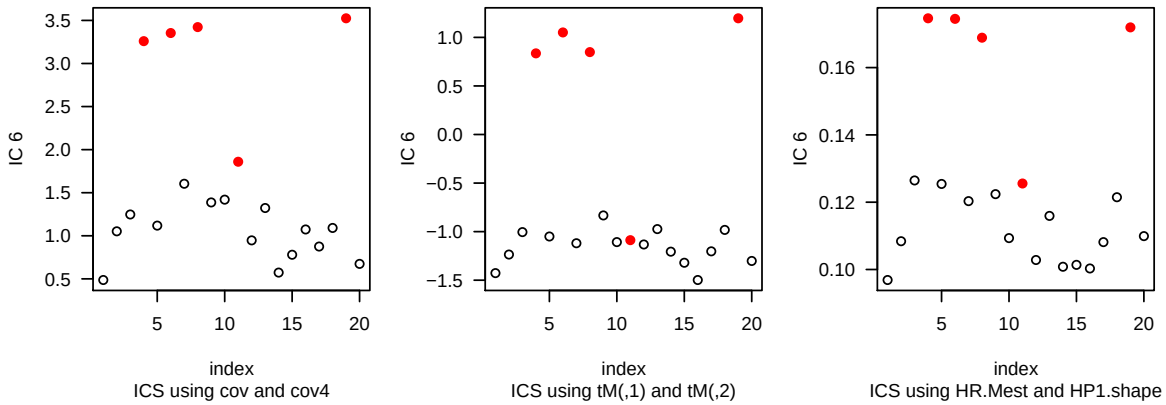

Figure 3: The last invariant coordinate from three different ICS's. The red points are according to the robust Mahalanobis distances outliers.

```
R> plot(ics.components(ics.default.wood)[,6], xlab = "index", ylab = "IC 6",
+    sub = "ICS using cov and cov4", col = out.id + 1, pch = 15 * out.id + 1)
R> plot(ics.components(ics.2.wood)[,6], xlab = "index", ylab = "IC 6",
+    sub = "ICS using tM(,1) and tM(,2)", col = out.id + 1,
+    pch = 15 * out.id + 1)
R> plot(ics.components(ics.3.wood)[,6], xlab = "index", ylab = "IC 6",
+    sub = "ICS using HR.Mest and HP1.shape", col = out.id + 1,
+    pch = 15 * out.id + 1)
R> par(mfrow = c(1, 1), las = 0)
```

From Figure 3, it can be noted that all three plots clearly display the four extreme points, even though the three pairs of scatter matrices are quite different. The first ICS uses two highly nonrobust scatter matrices, namely they have unbounded influence functions and zero breakdown points. The other two ICS have bounded influence functions, non-zero but not necessarily high breakdown points. The second ICS system presumes first moments, whereas the third does not presume any moments.

The last example also demonstrates the ease of use for the `ics` function. One can submit just two function names when the functions return only the scatter estimates, one can write without difficulties a wrapper around functions that return more than a scatter matrix, as done was done here for `HR.Mest`, or one can submit directly scatter matrices computed in advance, such as (`tM(wood)$V` and `tM(wood, 2)$V`).

In practice, one often encounters very high dimensional data sets, and so a common practice nowadays is to use PCA or other methods as a dimension reduction technique. The invariant coordinates, i.e., ICS, can also be used for this purpose. We will demonstrate this on Fisher's Iris data set (Fisher 1936).

We start by loading the needed packages and call for the 4 explanatory variables in the data set `ics`.

```
R> library("ICS")
R> library("MASS")
```

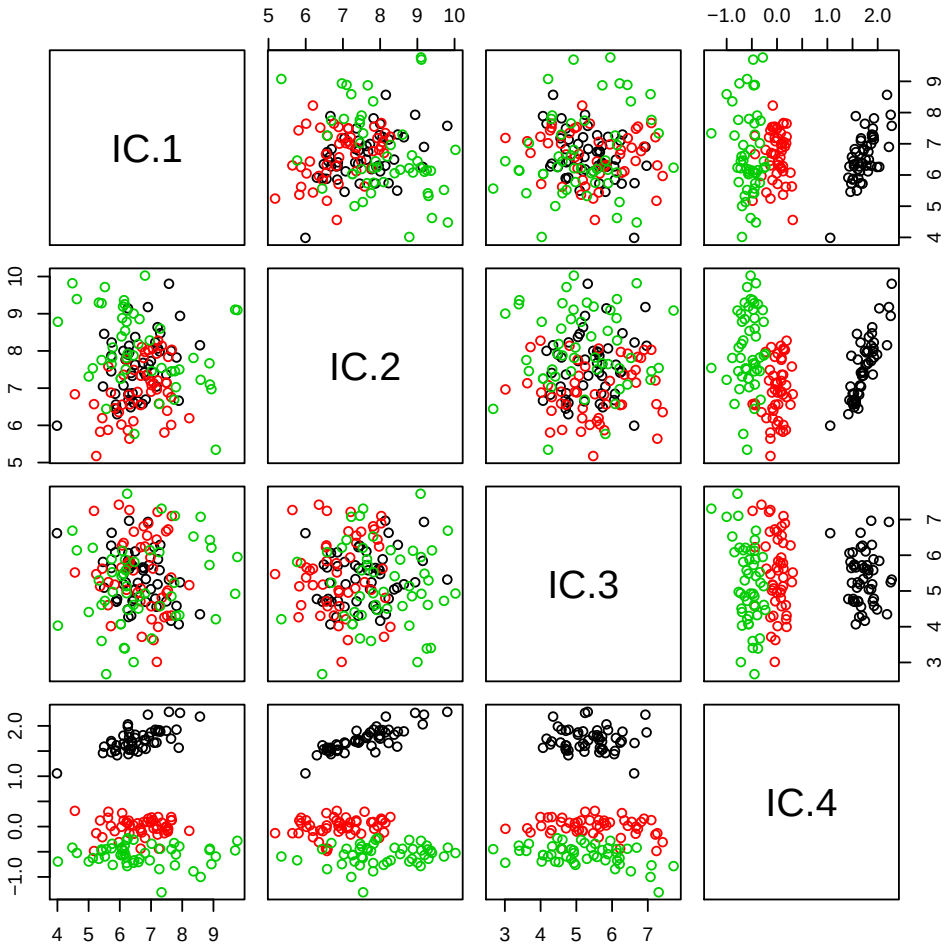

Figure 4: Scatter plot matrix for invariant coordinates of the Iris data set.

```
R> data("iris")
R> iris.ics <- ics(iris[,1:4])
R> plot(iris.ics, col = as.numeric(iris[,5]))
```

The invariant coordinates are then plotted with different colors for the different species in Figure 4. As can be seen in this figure, the coordinate with the lowest generalized kurtosis separates the three species very well, even though the species identification is not being taken into account in this analysis. Heuristically spoken one can say that the last coordinate corresponds to Fisher's linear discriminant subspace.

Since both ICS and PCA can serve as dimension reduction methods which helps identify clusters, we also plot for comparison purposes the principal component variables for the Iris data.

```
R> pairs(princomp(iris[,1:4])$scores, col = as.numeric(iris[,5]))
```

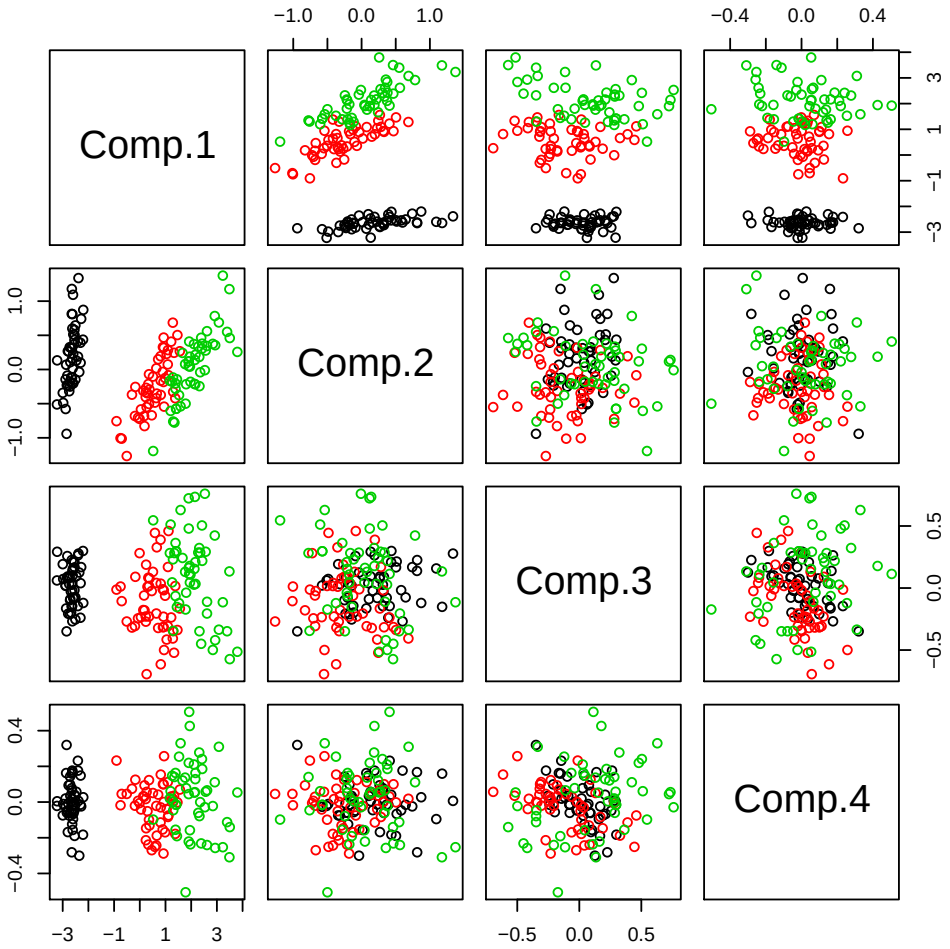

Figure 5: Scatter plot matrix for principal components of the Iris data set.

By comparing Figures 4 and 5, we note that both plots clearly separates one species from the other two, but the PCA plot is less successful than the ICS plot at distinguishing between the other two species.

Finally, we look at a so called discriminate coordinate plot, which unlike the two previous plots takes into account the group memberships. Such a plot can be done using `ics` by specifying as S_1 the regular covariance matrix and as S_2 the within group matrix, which we will call `cov.within`.

```
R> p <- dim(iris[, 1:4])[2]
R> n <- dim(iris[, 1:4])[1]
R> ngroup <- aggregate(iris$Species, list(iris$Species), length)$x
R> colMeans.iris <- colMeans(iris[, 1:4])
R> colMeans.iris.groups <- by(iris[, 1:4], iris$Species, colMeans)
```

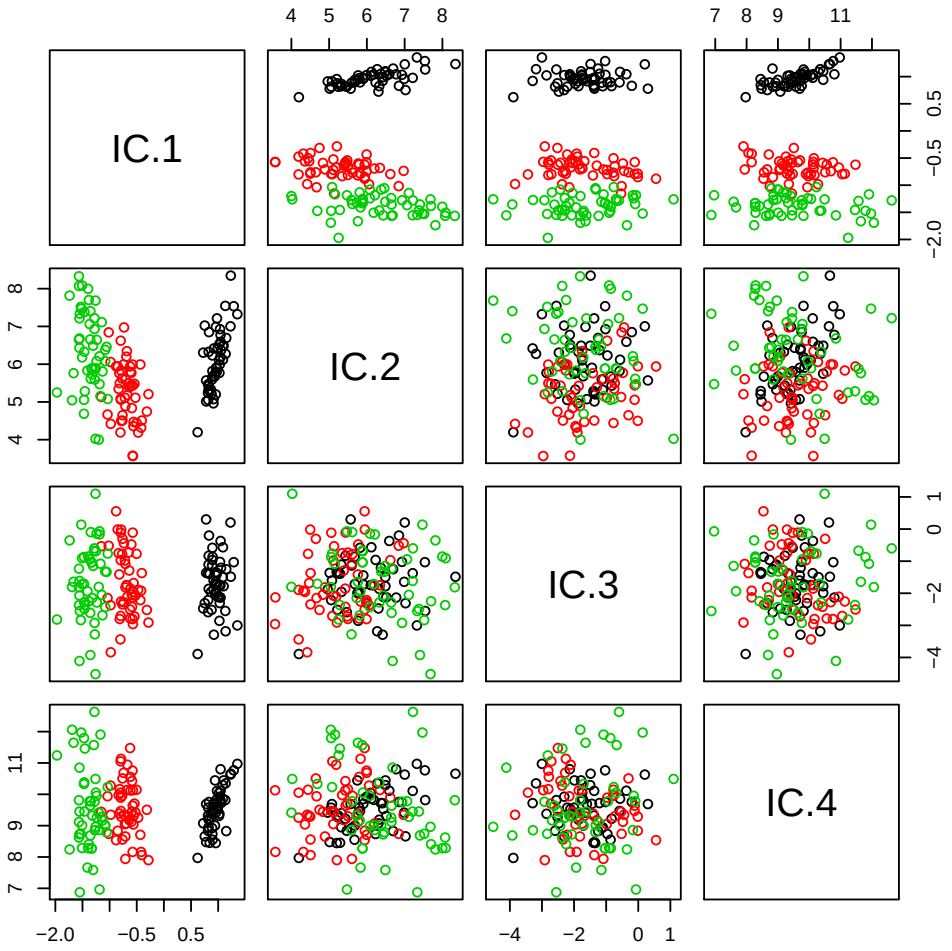

Figure 6: Discriminate coordinate plot for the Iris data set.

```
R> colMeans.iris.diffs <- sapply(colMeans.iris.groups,"-",
+  colMeans.iris, simplify = FALSE)
R> matrix.iris <- sapply(colMeans.iris.diffs, tcrossprod, simplify = FALSE)
R> freq <- rep(ngroup, each = p^2)
R> matrix.iris <- array(unlist(matrix.iris),
+    dim = c(p, p, nlevels(iris$Species)))
R> cov.within <- rowSums(matrix.iris * freq, dims = 2)/n
R> ics.iris.disc <- ics(iris[,1:4], cov(iris[,1:4]), cov.within)
R> plot(ics.iris.disc, col = as.numeric(iris$Species))
```

As can be seen from Figures 4 and 6, the fourth component of ICS a and the first component of the discriminate analysis are similar. As noted in Section 3.3, this is what is theoretically anticipated. We continue by taking a closer look at the 4th invariant coordinate of `iris.ics`.

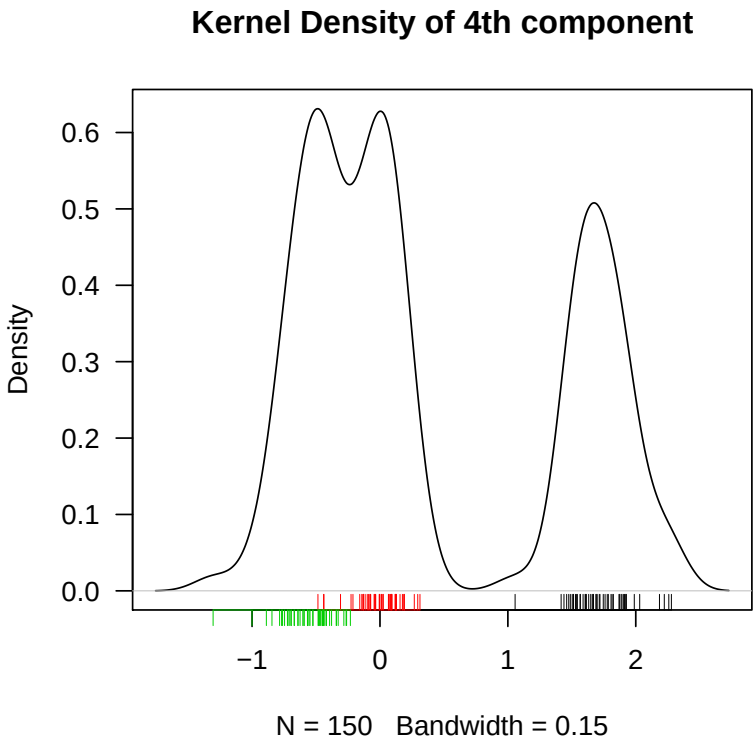

Figure 7: Kernel density estimate of the 4th invariant coordinate of the Iris data set with rugs for the different species. Bandwidth = 0.15.

Looking at a kernel density estimate of that component, with rugs representing the different species, confirms that this component serves very well for discriminating among the three species (see Figure 7).

```
R> iris.z <- ics.components(iris.ics)
R> plot(density(iris.z[,4], bw = 0.15), las = 1,
+  main = "Kernel Density of 4th component")
R> rug(iris.z[1:50, 4], col = 1)
R> rug(iris.z[51:100, 4], col = 2)
R> rug(iris.z[101:150, 4], col = 3, ticksize = -0.03)
```

This result agrees also with Bugrien (2005) who used ICA components for classification for the same data.

To demonstrate this we will randomly select 80% of the observations of the data set as the training set and use first the regular data to create a linear discrimination rule to classify the remaining 20% of the observations and afterwards we will use the training set to create an invariant coordinate system and use only the 4th component to create the discrimination rule and classify the test sample using this rule.

```
R> set.seed(4321)
```

```
R> train <- sample(1:150, 120)
R> lda.iris <- lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length +
+    Petal.Width, prior = c(1, 1, 1)/3, data = iris, subset = train)
R> table(iris[-train, 5], predict(lda.iris, iris[-train, ])$class)

             setosa versicolor virginica
  setosa         12          0         0
  versicolor      0         11         1
  virginica       0          1         5

R> ics.iris <- ics(as.matrix(iris[train, 1:4]))
R> iris.comp4 <- (ics.components(ics.iris))[,4]
R> lda.ics.iris <- lda(iris$Species[train] ~ iris.comp4, prior = c(1, 1, 1)/3)
R> iris.comp4.pred <- (as.matrix(iris[-train, 1:4]) %*% t(coef(ics.iris)))[,4]
R> table(iris[-train, 5], predict( lda.ics.iris,
+  data.frame(iris.comp4 = iris.comp4.pred))$class)

             setosa versicolor virginica
  setosa         12          0         0
  versicolor      0         12         0
  virginica       0          1         5
```

As the two tables show, both methods classify the species pretty well, however using an ICS we were able to reduce the number of explanatory variables from four to one.

In our analysis of the Iris data, the number of components considered for further analysis has been based only on graphical arguments. The values of the generalized kurtosis parameters can also be used to help decide which components may be of further interest. Within the framework of principal axis analysis (PAA) clear guidelines have been proposed. As pointed out in Section 3.3, PAA is a special case of ICS. Consequently, we demonstrate with the Iris data how PAA can be implemented using the function `ics`.

ICS yields PAA by calling `ics` using `cov` and `covAxis` for the centered data and requires the absolute values of the generalized kurtosis measures. Which in this case correspond to what is called the empirical alignment values in PAA.

```
R> iris.centered <- sweep(iris[,1:4], 2, colMeans(iris[,1:4]), "-")
R> iris.paa <- ics(iris.centered, cov, covAxis, stdKurt = FALSE)
```

In PAA, the generalized kurtosis measures are referred to as the empirical alignment values, which we now extract. The mean of the empirical alignment values always equals one.

```
R> emp.align <- iris.paa@gKurt
R> mean(emp.align)

[1] 1

R> emp.align
```

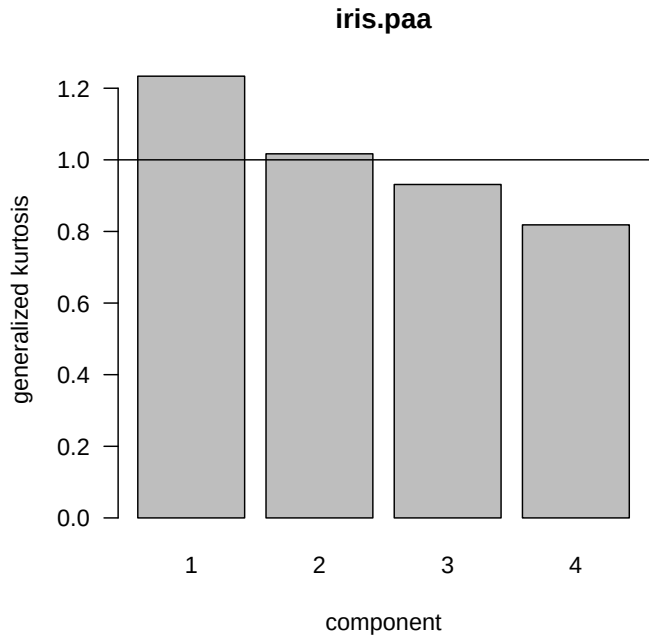

Figure 8: Screeplot for `iris.paa`. Components that exceed the vertical line are of interest.

```
[1] 1.2336 1.0168 0.9312 0.8184
```

The PAA guidelines given in Critchley *et al.* (2008) for deciding which components deserve to be considered for further analysis are those which have an empirical alignment greater than one. This can be visualized by using a screeplot and checking which components are indeed larger than one (see Figure 8).

```
R> screeplot(iris.paa, las = 1)
R> abline(h = 1)
```

So, in this example, we note that the first component is of clear interest whereas the second component may be of boarderline interest.

5.3. Independent components analysis

Independent components analysis has many applications as, for example, in signal processing or image separation. We will demonstrate here how the function `ics` can be used to restore three images which have been mixed by a random mixing matrix. The three images, which are displayed in the first row of Figure 9, are part of the package **ICS**. Each of them is on a greyscale and has 130×130 pixels. The figures are loaded as follows:

```
R> library("ICS")
R> library("pixmap")
```

```
R> fig1 <- read.pnm(system.file("pictures/cat.pgm", package = "ICS")[1])
R> fig2 <- read.pnm(system.file("pictures/road.pgm", package ="ICS")[1])
R> fig3 <- read.pnm(system.file("pictures/sheep.pgm", package = "ICS")[1])
```

For our analysis we have to vectorize the pixel matrices and combine them to form a data set.

```
R> p <- dim(fig1@grey)[2]
R> X <- cbind(as.vector(fig1@grey), as.vector(fig2@grey), as.vector(fig3@grey))
```

Next, we create a 3×3 mixing matrix A (the random seed is here set to ensure a proper mixing of the three pictures), mix the three pictures and use the FOBI algorithm via `ics` to recover the pictures.

```
R> set.seed(4321)
R> A <- matrix(rnorm(9), ncol = 3)
R> X.mixed <- X %*% t(A)
R> ICA.fig <- ics(X.mixed, stdB="B")
```

For a good comparison we plot into one figure in the first row the three original pictures, in the second row the three mixed pictures, and in the last row the recovered images.

```
R> par(mfrow = c(3, 3), omi = rep(0.1, 4), mai = rep(0.1, 4))
R> plot(fig1)
R> plot(fig2)
R> plot(fig3)
R> plot(pixmapGrey(X.mixed[,1], ncol = p))
R> plot(pixmapGrey(X.mixed[,2], ncol = p))
R> plot(pixmapGrey(X.mixed[,3], ncol = p))
R> plot(pixmapGrey(ics.components(ICA.fig)[,1], ncol = p))
R> plot(pixmapGrey(ics.components(ICA.fig)[,2], ncol = p))
R> plot(pixmapGrey(ics.components(ICA.fig)[,3], ncol = p))
```

As Figure 9 shows, we are able to recover the three images quite well. The new order of the images is related to their generalized kurtosis measures. Also, the cat is now a negative, since the signs of the components are not fixed. However the positive version of the cat could be easily obtained by multiplying the corresponding component by -1 before the plotting.

5.4. Multivariate nonparametrics

In this section, we demonstrate via examples, the use of an invariant coordinate system for estimation and testing problems. For the estimation example, we choose the componentwise Hodges-Lehmann estimator (Hettmansperger and McKean 1998). For the testing example, we use the one sample location test using marginal normal scores (Puri and Sen 1971).

We start the demo with loading the three packages needed.

```
R> library("ICS")
R> library("mvtnorm")
R> library("ICSNP")
```

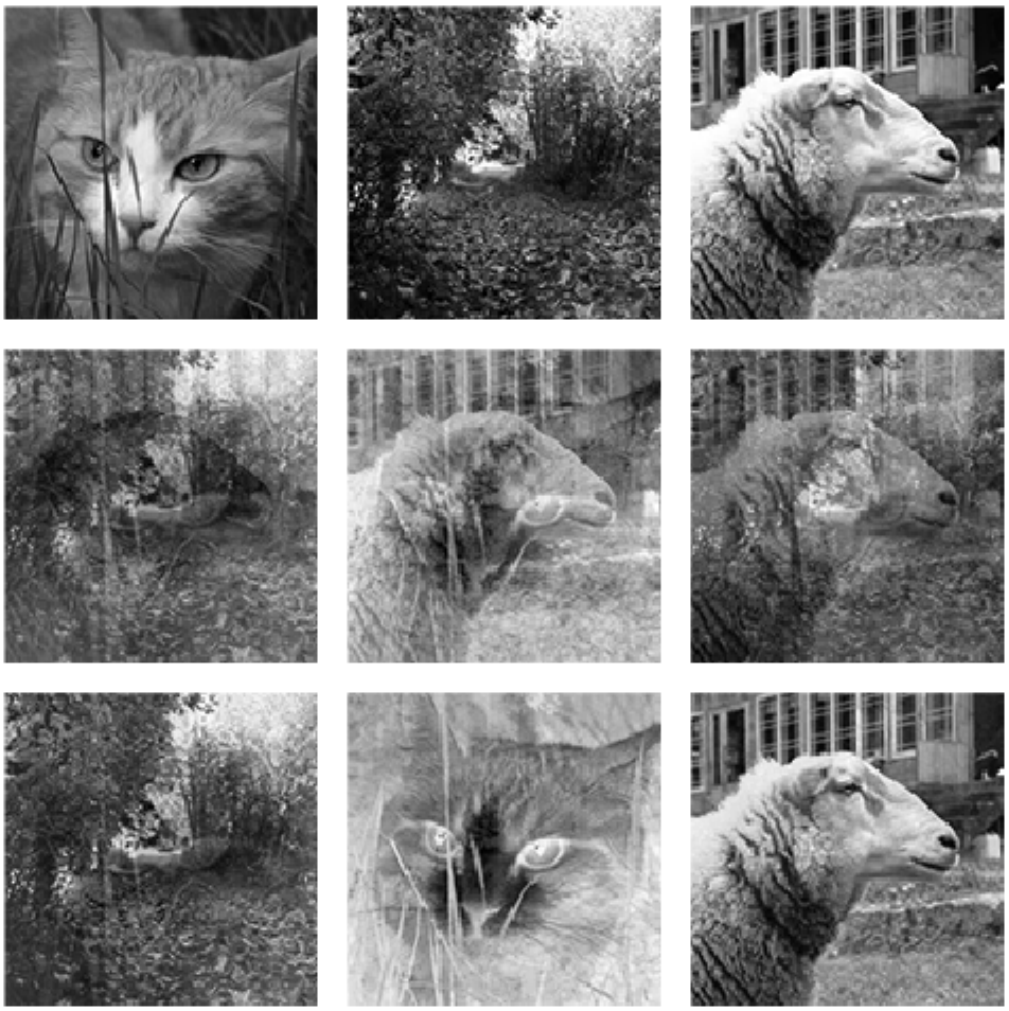

Figure 9: ICA for three pictures. First row are the original pictures, second row the mixed pictures and the last row the pictures recovered by ICA.

Now we will create a simulated data matrix X of 150 samples coming form a $N_3((1\ 2\ -1), I)$ distribution, a 3×3 transformation matrix A and a location shift vector $b = (1\ 1\ 1)$. The transformed data will be denoted X_{trans}. Also needed is the function `HL.estimator` in order to extract the Hodges-Lehmann estimator from the function `wilcox.test`.

```
R> set.seed(2000)
R> X <- rmvnorm(150, c(1, 2,-1))
R> A <- matrix(rnorm(9), ncol = 3)
R> b <- c(1, 1, 1)
R> X.trans <- sweep(X %*% t(A), 1, b, "+")
R> HL.estimator <- function(x){
+  wilcox.test(x, exact = TRUE, conf.int = TRUE)$estimate}
```

The results when applying then the Hodges-Lehmann estimator on X and transforming the estimate using A and b and applying the estimator directly on X_{trans} differ as the following lines show.

```
R> HLE.X <- apply(X, 2, HL.estimator)
R> as.vector(HLE.X %*% t(A) + b)

[1]  4.2068 -0.3024 -1.9104

R> apply(X.trans, 2, HL.estimator)

[1]  4.2025 -0.2617 -1.9073
```

since the Hodges-Lehmann estimator is not affine equivariant.

This can be avoided as explained in Section 3.5 by using an ICS. We therefore use the function `ics` and choose as S_1 the regular covariance matrix and as S_2 Tyler's shape matrix. First we will apply it only on X, estimate using the obtained coordinates the Hodges-Lehmann estimate and transform the estimate back into the original coordinates using the inverse of the transformation matrix B^{-1}, this estimate is denoted as `HL.ics.X`. Repeating the same procedure on the transformed data $XA^\top$ we can see that the corresponding estimate `HL.ics.AX` equals the transformed estimate of `HL.ics.X`.

```
R> ics.X <- ics(X, S1 = cov, S2 = tyler.shape)
R> HL.ics.Z1 <- apply(ics.components(ics.X), 2, HL.estimator)
R> HL.ics.X <- as.vector(HL.ics.Z1 %*% t(solve(coef(ics.X))))
R> ics.X.trans <- ics(X.trans, S1 = cov, S2 = tyler.shape)
R> HL.ics.Z2 <- apply(ics.components(ics.X.trans), 2, HL.estimator)
R> HL.ics.X.trans <- as.vector(HL.ics.Z2 %*% t(solve(coef(ics.X.trans))))
R> as.vector(HL.ics.X %*% t(A) +b)

[1]  4.2092 -0.3084 -1.9269

R> HL.ics.X.trans

[1]  4.2092 -0.3084 -1.9269
```

For the testing example we first generate a random sample of size 60 coming from a 4-variate t_6 distribution having mean (0 0 0 0.48). The 4×4 transformation matrix in this context is called $A2$.

```
R> set.seed(1979)
R> Y <- rmvt(60, diag(4), df = 6) + matrix(rep(c(0, 0.48), c(3*60, 60)),
+    ncol = 4)
R> A2 <- matrix(rnorm(16), ncol = 4)
```

We test the null hypothesis that the sample has the origin as its location on the original data Y first, and then for the transformed data $YA2^\top$. For invariant tests, the decisions are the same.

```
R> rank.ctest(Y, scores = "normal")

        Marginal One Sample Normal Scores Test

data:  Y
T = 9.653, df = 4, p-value = 0.04669
alternative hypothesis: true location is not equal to c(0,0,0,0)

R> rank.ctest((Y %*% t(A2)), scores = "normal")

        Marginal One Sample Normal Scores Test

data:  (Y %*% t(A2))
T = 9.387, df = 4, p-value = 0.05212
alternative hypothesis: true location is not equal to c(0,0,0,0)
```

As expected the decisions differ, they differ even that much, that assuming an α-level of 0.05 we would once reject and once fail to reject the null hypothesis.

Again, using an ICS avoids this problem. However, when testing a location parameter we have a hypothesis for it which should also be used in the computation of the scatter matrices. Therefore when creating our ICS we use scatter matrices with respect to the origin.

```
R> Z.Y <- as.matrix(ics.components(ics(Y,
+    S1 = covOrigin, S2 = cov4, S2args = list(location = "Origin"))))
R> rank.ctest(Z.Y, scores = "normal")

        Marginal One Sample Normal Scores Test

data:  Z.Y
T = 9.737, df = 4, p-value = 0.04511
alternative hypothesis: true location is not equal to c(0,0,0,0)

R> Z.Y.trans <- as.matrix(ics.components(ics(Y %*% t(A2),
+    S1 = covOrigin, S2 = cov4, S2args = list(location = "Origin"))))
R> rank.ctest(Z.Y.trans , scores = "normal")

        Marginal One Sample Normal Scores Test

data:  Z.Y.trans
T = 9.737, df = 4, p-value = 0.04511
alternative hypothesis: true location is not equal to c(0,0,0,0)
```

Acknowledgments

The authors are very grateful for Uwe Ligges' helpful comments that improved a lot the functionality of the package. The authors are also grateful for the comments of the editors,

associate editor and the two anonymous referees. The work of Klaus Nordhausen and Hannu Oja was supported by grants from the Academy of Finland. The work of David Tyler was supported by NSF Grant DMS-0604596.

References

Bickel PJ (1965). "On Some Asymptotically Nonparametric Competitors of Hotelling's T2." *Annals of Mathematical Statistics*, **36**, 160–173.

Bivand R, Leisch F, Mächler M (2008). ***pixmap**: Bitmap Images ("Pixel Maps")*. R package version 0.4-9, URL http://CRAN.R-project.org/package=pixmap.

Bugrien JB (2005). *Robust Approaches to Clustering Based on Density Estimation and Projection*. Ph.D. thesis, University of Leeds.

Cardoso JF (1989). "Source Separation Using Higher Order Moments." In "Proceedings of IEEE International Conference on Accoustics, Speech and Signal Processing," pp. 2109–2112. Glasgow.

Caussinus H, Ruiz-Gazen A (1994). "Projection Pursuit and Generalized Principal Component Analysis." In S Morgenthaler, E Ronchetti, WA Stahel (eds.), "New Directions in Statistical Data Analysis and Robustness," pp. 35–46. Birkhäuser Verlag, Basel.

Chakraborty B, Chaudhuri P (1996). "On a Transformation and Re-Transformation Technique for Constructing Affine Equivariant Multivariate Median." In "Proceedings of the American Mathematical Society," volume 124, pp. 2539–2547.

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

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

Critchley F, Pires A, Amado C (2008). "Principal Axis Analysis." Unpublished Manuscript.

Dümbgen L (1998). "On Tyler's M-Functional of Scatter in High Dimension." *Annals of the Institute of Statistical Mathematics*, **50**, 471–491.

Fisher RA (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, **7**, 179–188.

Genz A, Bretz F, Hothorn T (2008). ***mvtnorm**: Multivariate Normal and t Distribution*. R package version 0.9-2, URL http://CRAN.R-project.org/package=mvtnorm.

Hettmansperger TP, McKean JW (1998). *Robust Nonparametric Statistical Methods*. Arnold, London, UK.

Hyvärinen A, Karhunen J, Oja E (2001). *Independent Component Analysis*. John Wiley & Sons, New York, USA.

Kankainen A, Taskinen S, Oja H (2007). "Tests of Multinormality Based on Location Vectors and Scatter Matrices." *Statistical Methods & Applications*, **16**, 357–379.

Kent JT, Tyler DE (1991). "Redescending M-Estimates of Multivariate Location and Scatter." *The Annals of Statistics*, **19**, 2102–2119.

Mächler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M (2008). ***robustbase**: Basic Robust Statistics*. R package version 0.4-3, URL http://CRAN.R-project.org/package=robustbase.

Maronna RA, Martin RD, Yohai VJ (2006). *Robust Statistics - Theory and Methods*. John Wiley & Sons, Chichester, UK.

Nordhausen K, Oja H, Ollila E (2008a). "Robust Independent Component Analysis Based on Two Scatter Matrices." *Austrian Journal of Statistics*, **37**, 91–100.

Nordhausen K, Oja H, Paindaveine D (2008b). "Signed-Rank Tests for Location in the Symmetric Independent Component Model." *Journal of Multivariate Analysis*. doi:10.1016/j.jmva.2008.08.004.

Nordhausen K, Oja H, Tyler DE (2006). "On the Efficiency of Invariant Multivariate Sign and Rank Test." In EP Liski, J Isotalo, J Niemelä, S Puntanen, GPH Styan (eds.), "Festschrift for Tarmo Pukkila on his 60th Birthday," pp. 217–231. University of Tampere, Tampere, Finland.

Nordhausen K, Sirkiä S, Oja H, Tyler DE (2007). ***ICSNP**: Tools for Multivariate Nonparametrics*. R package version 1.0-2, URL http://CRAN.R-project.org/package=ICSNP.

Oja H, Sirkiä S, Eriksson J (2006). "Scatter Matrices and Independent Component Analysis." *Austrian Journal of Statistics*, **35**, 175–189.

Puri ML, Sen PK (1971). *Nonparametric Methods in Multivariate Analysis*. John Wiley & Sons, New York, USA.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rousseeuw PJ, van Driessen K (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics*, **41**, 212–223.

Rousseeuw PJ, van Zomeren BC (1990). "Unmasking Multivariate Outliers and Leverage Points." *Journal of the American Statistical Association*, **85**, 633–639.

Todorov V (2008). ***rrcov**: Scalable Robust Estimators with High Breakdown Point*. R package version 0.4-07, URL http://CRAN.R-project.org/package=rrcov.

Tyler DE (1987). "A Distribution-Free M-Estimator of Multivariate Scatter." *The Annals of Statistics*, **15**, 234–251.

Tyler DE, Critchley F, Dümbgen L, Oja H (2008). "Exploring Multivariate Data via Multiple Scatter Matrices." *Journal of the Royal Statistical Society B*. Forthcoming.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer-Verlag, New York.

Wang N, Raftery A, Fraley C (2003). ***covRobust**: Robust Covariance Estimation via Nearest Neighbor Cleaning*. R package version 1.0, URL http://CRAN.R-project.org/package=covRobust.

Affiliation:

Klaus Nordhausen
Tampere School of Public Health
University of Tampere
33014 University of Tampere, Finland
E-mail: klaus.nordhausen@uta.fi
URL: http://www.uta.fi/~klaus.nordhausen/

Hannu Oja
Tampere School of Public Health
University of Tampere
33014 University of Tampere, Finland
E-mail: hannu.oja@uta.fi
URL: http://www.uta.fi/~hannu.oja/

David E. Tyler
Department of Statistics
The State University of New Jersey
Piscataway NJ 08854, USA
E-mail: dtyler@rci.rutgers.edu
URL: http://www.rci.rutgers.edu/~dtyler/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 28, Issue 6 *Submitted:* 2008-02-19
November 2008 *Accepted:* 2008-09-03

AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 1, 91–100

Robust Independent Component Analysis Based on Two Scatter Matrices

Klaus Nordhausen[1], Hannu Oja[1] and Esa Ollila[2]

[1]University of Tampere, Finland
[2]Helsinki University of Technology, Finland

Oja, Sirkiä, and Eriksson (2006) and Ollila, Oja, and Koivunen (2007) showed that, under general assumptions, any two scatter matrices with the so called independent components property can be used to estimate the unmixing matrix for the independent component analysis (ICA). The method is a generalization of Cardoso's (Cardoso, 1989) FOBI estimate which uses the regular covariance matrix and a scatter matrix based on fourth moments. Different choices of the two scatter matrices are compared in a simulation study. Based on the study, we recommend always the use of two robust scatter matrices. For possible asymmetric independent components, symmetrized versions of the scatter matrix estimates should be used.

Keywords: Affine Equivariance, Kurtosis, Source Separation.

1 Introduction

Let $x_1, x_2, \ldots, x_n$ be a random sample from a p-variate distribution, and write

$$X = \left(x_1 \; x_2 \; \ldots \; x_n \right)$$

for the $p \times n$ data matrix. We assume that X is generated by

$$X = AZ,$$

where $Z = (z_1 z_2 \ldots z_n)$ and $z_1, \ldots, z_n$ are independent and identically distributed latent random vectors having independent components and A is a full-rank $p \times p$ *mixing matrix*. This model is called the *independent component (IC) model*. The model is not well defined in the sense that the model may also be written as

$$X = A^* Z^*$$

where

$$A^* = AP'D^{-1} \qquad \text{and} \qquad Z^* = DPZ$$

for any diagonal matrix D (with nonzero diagonal elements) and for any permutation matrix P. (A permutation matrix P is obtained from identity matrix I_p by permuting its rows.) If Z has independent components, then also the components of $Z^* = DPZ$ are independent. The problem in the so called *independent component analysis (ICA)* is to find an *unmixing matrix* B such that Bx_i has independent components. Based on the discussion above, the solution is then not unique: If B is an unmixing matrix, then so is DPB.

Most ICA algorithms then proceed as follows. (For a recent review of different approaches, see Hyvärinen, Karhunen, and Oja, 2001.)

1. To simplify the problem it is first commonly assumed that the x_i are *whitened* so that $\mathrm{E}(x_i) = 0$ and $\mathrm{cov}(x_i) = I_p$. Then

$$X = UZ^*$$

with an orthogonal matrix U and Z^* with (columns having) independent components such that $\mathrm{E}(z_i^*) = 0$ and $\mathrm{cov}(z_i^*) = I_p$

2. For the whitened data X, find a $p \times r$ matrix U with orthonormal columns ($r \leq p$) which maximizes (or minimizes) a chosen criterion function, say $g(U'X)$. Measures of marginal nongaussianity (negentropy, kurtosis measures) $g(u'X)$ and likelihood functions with different choices of marginal distributions are often used.

In the FastICA algorithm (Hyvärinen and Oja, 1997) for example in each iteration step (for stage 2) the columns of U are updated one by one and then orthogonalized. The criterion of the FastICA algorithm maximizes the negentropy which is approximated by

$$g(u'X) = [\text{ave}\{h(u'x_i)\} - \mathrm{E}[h(z)]]^2 \tag{1}$$

with $z \sim N(0,1)$ and with several possible choices for the function $h(\cdot)$.

A different solution to the ICA problem, called FOBI, was given by Cardoso (1989): After whitening the data as above (stage 1), an orthogonal matrix U is found as the matrix of eigenvectors of a kurtosis matrix (matrix of fourth moments; this will be discussed later). The data transformation consists of a joint diagonalization of the regular covariance matrix and of the scatter matrix based on fourth moments. FOBI was generalized in Oja et al. (2006) (real data) and Ollila et al. (2007) (complex data) where any two scatter matrices which have the so called independent components property can be used. An interesting question then naturally arises: How should one choose these two scatter matrices in a good or optimal way?

The paper is organized as follows. First, in Section 2 scatter matrices and their use in the estimation of an unmixing matrix is reviewed. In Section 3 we describe the results from simulation studies where new ICA estimates with several choices of scatter matrices are compared to classical FastICA and FOBI estimates. Also an image analysis example is given. The paper ends with some conclusions in Section 4.

2 Two Scatter Matrices and ICA

Let x be a p-variate random vector with cdf F_x. A functional $T(F)$ is a p-variate *location vector* if it is affine equivariant in the sense that $T(F_{Ax+b}) = AT(F_x) + b$ for all x, all full-rank $p \times p$ matrices A and all p-variate vectors b. Using the same notation, a matrix-valued $p \times p$ functional $S(F)$ is called a *scatter matrix* if it is positive definite, symmetric and affine equivariant in such way that $S(F_{Ax+b}) = AS(F_x)A'$ for all x, A and b. The regular mean vector $E(x)$ and covariance matrix $Cov(x)$ serve as first examples. There are numerous alternative techniques to construct location and scatter functionals, e.g. M-functionals, S-functionals, etc. See e.g. Maronna, Martin, and Yohai (2006).

A scatter matrix $S(F)$ is said to have the *independent components (IC-) property* if $S(F_z)$ is a diagonal matrix for all z having independent components. The covariance

matrix naturally has the IC-property. Other classical scatter functionals (M-functionals, S-functionals, etc.) developed for elliptical distributions do not generally possess the IC-property. However, if z has independent and symmetrically distributed components, then $S(F_z)$ is a diagonal matrix for all scatter functionals S. It is therefore possible to develop a symmetrized version of a scatter matrix $S(F)$, say $S_{sym}(F)$, which has the IC-property; just define

$$S_{sym}(F_x) = S(F_{x_1-x_2}),$$

where x_1 and x_2 are two independent copies of X. See Oja et al. (2006), Ollila et al. (2007) and Sirkiä, Taskinen, and Oja (2007).

An alternative approach to the ICA using two scatter matrices with IC-property (Oja et al., 2006, Ollila et al., 2007) has the following two steps:

1. The x_i are whitened using S_1 (instead of the covariance matrix) so that $S_1(F_{x_i}) = I_p$. Then
$$X = UZ^*$$
with an orthogonal matrix U and with Z^* with (columns having) independent components such that $S_1(z_i^*) = I_p$.
2. For the whitened data X, find an orthogonal matrix U as the matrix of eigenvectors of $S_2(F_{x_i})$.

The resulting data transformation $X \rightarrow \hat{B}X$ then jointly diagonalizes S_1 and S_2 ($S_1(\hat{B}X) = I_p$ and $S_2(\hat{B}X) = D$) and the unmixing matrix $\hat{B}$ solves

$$S_2^{-1}S_1B' = B'D^{-1}.$$

The matrix $\hat{B}$ is the matrix of eigenvectors and the diagonal matrix $\hat{D}$ is the matrix of eigenvalues of $S_2^{-1}S_1$. Note the similarity between our ICA procedure and the principal component analysis (PCA): The direction u of the first eigenvector of $S_2^{-1}S_1$ maximizes the criterion function $(u'S_1u)/(u'S_2u)$ which is a measure of kurtosis (ratio of two scale measures) rather than a measure of dispersion (as in PCA) in the direction u, etc. The independent components are then ordered according to this specific kurtosis measure. The solution is unique if the eigenvalues of $S_2^{-1}S_1$ are distinct.

Different choices of S_1 and S_2 naturally yield different estimates $\hat{B}$. First, the resulting independent components $\hat{B}X$ are rescaled by S_1 and they are given in an order determined by S_2. Also the statistical properties of the estimates $\hat{B}$ (convergence, limiting distributions, efficiency, robustness) naturally depend on the choices of S_1 and S_2.

3 Performance Study

3.1 The Estimates $\hat{\mathbf{B}}$ to be Compared

We now study the behavior of the new estimates $\hat{B}$ with different (robust and non-robust) choices for S_1 and S_2. The classical FastICA procedures which use

$$h_1(u'x_i) = \log(\cosh(u'x_i)) \qquad \text{or} \qquad h_2(u'x_i) = -\exp(-u'x_i)$$

in equation (1) serve as a reference. These algorithms will be denoted as *FastICA1* and as *FastICA2*, respectively. According to Hyvärinen and Oja (2000), these choices are more robust than the traditional negentropy estimate with criterion

$$g(u'X) = \frac{1}{12}\left[\text{ave}\left\{(u'x_i)^3\right\}\right]^2 + \frac{1}{48}\left[\text{ave}\left\{(u'x_i)^4\right\} - 3\right]^2.$$

The *FOBI* estimate by Cardoso (1989) assumes that the centering is done using the mean vector, and

$$S_1(F_x) = \text{cov}(x) \quad \text{and} \quad S_2(F_x) = \frac{1}{p+2}\text{E}\left[||S_1^{-1/2}(x - E(x))||^2(x - \text{E}(x))(x - \text{E}(x))'\right].$$

Then S_2 is a scatter matrix based on the fourth moments, both S_1 and S_2 possess the IC-property, and the independent components are ordered with respect to their classical kurtosis measure. The FOBI estimate is member in the new class of estimates but highly non-robust due to the choices of S_1 and S_2.

In our simulation study we consider scatter matrices which are (unsymmetrized and symmetrized) M-functionals. Simultaneous M-functionals for location and scatter corresponding to chosen weight functions $w_1(r)$ and $w_2(r)$ are functionals which satisfy implicit equations

$$T(F_x) = [\text{E}[w_1(r)]]^{-1}\text{E}[w_1(r)x] \qquad \text{and} \qquad S(F_x) = \text{E}[w_2(r)xx'],$$

where r is the Mahalanobis distance between x and $T(F_x)$, i.e.

$$r^2 = (x - T(F_x))'S(F_x)^{-1}(x - T(F_x)).$$

In this paper we consider Huber's M-estimator (Maronna et al., 2006) with

$$w_1(r) = \begin{cases} 1 & r \le c \\ c/r & r > c \end{cases} \qquad \text{and} \qquad w_2(r) = \begin{cases} 1/\sigma^2 & r \le c \\ c^2/\sigma^2 r^2 & r > c. \end{cases}$$

The tuning constant c is chosen to satisfy $q = \Pr(\chi_p^2 \le c^2)$ and the scaling factor σ^2 so that $\text{E}[\chi_p^2 w_2(\chi_p^2)] = p$. Tyler's shape matrix (Tyler, 1987) is often called the most robust M-estimator. Tyler's shape matrix and simultaneous spatial median estimate, see (Hettmansperger and Randles, 2002), have the weight functions

$$w_1(r) = \frac{1}{r} \qquad \text{and} \qquad w_2(r) = \frac{p}{r^2}.$$

Symmetrized versions of Huber's estimate and Tyler's estimate then possess the IC-property. The symmetrized version of Tyler's shape matrix is also know as Dümbgen's shape matrix (Dümbgen, 1998).

In this simulation study we compare

- FastICA1 and FastICA2 estimates
- E1: FOBI estimate
- E2: Estimate based on the covariance matrix and Tyler's shape matrix
- E3: Estimate based on Tyler's shape matrix and the covariance matrix

- E4: Estimate based on Tyler's shape matrix and Dümbgen's shape matrix
- E5: Estimate based on Tyler's shape matrix and Huber's M-estimator ($q = 0.9$)
- E6: Estimate based on Dümbgen's shape matrix and symmetrized Huber's M-estimator ($q = 0.9$).

All computations are done in R 2.4.0 (R Development Core Team, 2006); the package fastICA (Marchini, Heaton, and Ripley, 2006) was used for the FastICA solutions and the package ICS (Nordhausen, Oja, and Tyler, 2006) for the new method.

3.2 Simulation Designs

In this simulation study the independent components are all symmetrically distributed. Therefore all choices of S_1 and S_2 are acceptable. The designs were as follows:

- *Design I:* The $p = 4$ independent components were generated from (i) a normal distribution, (ii) a uniform distribution, (iii) a t_3 distribution, and (iv) a Laplace distribution, respectively (all distributions with unit variance.) The sample sizes ranged from $n = 50$ to $n = 2000$. For each sample size, we had 300 repetitions. For all samples, the elements of a mixing matrix A were generated from a $N(0, 1)$ distribution.
- *Design II:* As Design I but with outliers. The $\max(1, 0.01n)$ observations x_i with the largest L_2 norms were multiplied by $s_i u_i$ where s_i is $+1$ or -1 with probabilities $1/2$ and u_i has a Uniform$(1, 5)$ distribution. This was supposed to partially destroy the dependence structure.

3.3 Performance Index

Let A be the "true" mixing matrix in a simulation and $\hat{B}$ an estimate of an unmixing matrix. For any true unmixing matrix B, $BA = PD$ with a diagonal matrix D and a permutation matrix P. Write $G = (g_{ij}) = \hat{B}A$. The performance index (Amari, Cichocki, and Yang, 1996)

$$PI(G) = \frac{1}{2p(p-1)} \left[\sum_{i=1}^{p} \left(\sum_{j=1}^{p} \frac{|g_{ij}|}{\max_h |g_{ih}|} - 1 \right) + \sum_{j=1}^{p} \left(\sum_{i=1}^{p} \frac{|g_{ij}|}{\max_h |g_{hj}|} - 1 \right) \right]$$

is then often used in comparisons. Now clearly $PI(PG) = PI(G)$ but $PI(DG) = PI(G)$ is not necessarily true. Therefore, for a fair comparison, we standardize and reorder the rows of $B = (b_1 \dots b_p)'$ ($B \to PDB$) such that

- $||b_i|| = 1, i = 1, \dots, p$
- $\max(b_{i1}, \dots, b_{ip}) = \max(|b_{i1}|, \dots, |b_{ip}|), i = 1, \dots, p$
- $\max(b_{i1}, \dots, b_{ip}) \geq \max(b_{j1}, \dots, b_{jp}), 1 \leq i \leq j \leq p.$

For the comparison, also A^{-1} is standardized in a similar way.

The performance index $PI(G)$ can take values in $[0, 1]$; the smaller is $PI(\hat{B}A)$ the better is the estimate $\hat{B}$.

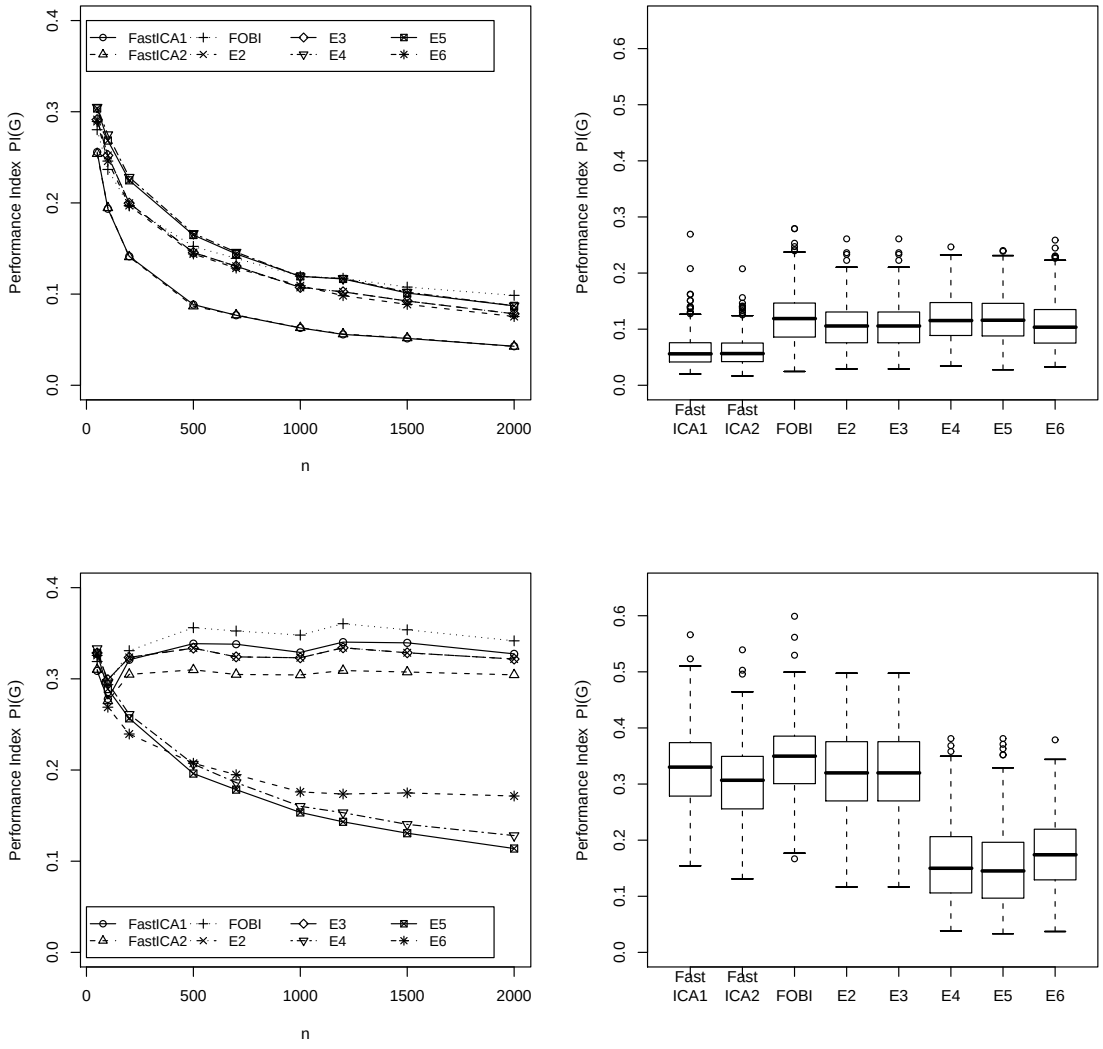

Figure 1: Results of the simulations. The top row shows the results for Design I and the bottom row for Design II. The left column shows the mean of $PI(\hat{B}A)$ for 300 repetitions and the right column boxplots of $PI(G)$ when $n = 1000$. The estimates based on two scatter matrices besides $FOBI$ are $E2$: covariance matrix & Tyler's shape matrix, $E3$: Tyler's shape matrix & covariance matrix, $E4$: Tyler's shape matrix & Dümbgen's shape matrix, $E5$: Tyler's shape matrix & Huber's M-estimator and $E6$: Dümbgen's shape matrix & Symmetrized Huber's M-estimator.

3.4 Simulation Results

The results of the simulations are summarized in Figure 1 and show, that in the non-contaminated case (Design I) the two versions of the fastICA algorithm dominate all estimates based on two scatter matrices. Surprisingly, in this case, the FOBI estimate seems to be the worst choice among all, whereas the best is estimate E6 which is based on two symmetrized scatter matrices. The differences are minor, however. The results change considerably when adding outliers (Design II). The procedures $E4$, $E5$ and $E6$

based on two robust scatter matrices are least affected by the outliers. Estimate $E6$ using robust symmetrized estimates presumably has a lowest breakdown point among the robust estimates which may explain its slightly worse behavior here. The order in which the two scatter matrices are used has no effect on the results; $E2$ and $E3$ have naturally the same performance in the simulations.

3.5 An Example

To demonstrate the effect of outliers in a real example we will attempt to unmix three mixed images. The original images which show a cat, a forest track and a sheep, are all in a greyscale having each 130×130 pixels and are part of the the R-package ICS. In the analysis of image data, the pixels are thought to be individuals ($n = 130 \times 130$), and each individual has three measurements corresponding to the three pictures ($p = 3$). The three pictures are first mixed with a random 3×3 matrix using the vector representation of the pictures. Contamination is added to the first mixed image by blackening 60 pixels in the right upper corner, which corresponds to less than 1 percent of outliers. The algorithms $E5$ and $FastICA2$ are then applied to recover the original images. To retransform the independent components to a reasonable greyscale, for all independent components, values smaller than the 2.5% quantile are replaced by the quantile and the same was done for values larger than the the 97.5% quantile. The result is shown in Figure 2.

As can be seen, some images are negatives of the original images. This is due to the arbitrary sign of the independent components. Nevertheless, it can be observed, that $E5$ performs better than $FastICA2$ even when the amount of contamination is so small. The algorithm $E5$ recovers the two images with the sheep and the cat well and only in the image of the forest track the head of the cat is slightly present. In the images recovered by $FastICA2$ however none could be called well separated. The picture with the cat has still the windows that belong to the picture with the sheep and in the picture of the sheep and of the forest track the head of the cat is still visible. The good performance of $E5$ is noteworthy here especially when considering that the images probably do not have underlying symmetric distributions. Using two robust scatter matrices having the IC-property like symmetrized scatter matrices might therefore even improve the result. However the dimension of this example with 16900 observations and three variates is currently too large to apply symmetrized scatter matrices since the resulting large number of pairwise differences is a too huge computational task and hence not feasible.

4 Conclusion

Based on the simulation results, we recommend the use of two robust scatter matrices in all cases. For possible asymmetric independent components, symmetrized versions of the scatter matrix estimates should be used. Symmetrized scatter matrices are however based on U-statistics and computationally expensive; $n = 1,000$ observations for example means almost $500,000$ pairwise differences. However, as the image example shows, ICA problems have easily several thousand observations and therefore this is not feasible yet. To relieve the computational burden, the original estimate may then be re-

Figure 2: ICA for three pictures. The first row shows the original pictures, the second row the mixed pictures including some contamination. The third row used two robust scatter matrices ($E5$) to recover the pictures and the fourth row the $FastICA2$ algorithm.

placed by an estimate which is based on an incomplete U-statistic. Further investigation is needed to examine the situations where the components are not symmetric. For asymmetric independent components, FastICA algorithms for example are known to have a poorer performance.

References

Amari, S., Cichocki, A., and Yang, H. (1996). A new learning algorithm for blind source separation. In *Advances in neural information processing systems 8* (p. 757-763). Cambridge, MA.: MIT Press.

Cardoso, J. (1989). Source separation using higher order moments. In *Proceedings of IEEE international conference on acustics, speech and signal processing* (p. 2109-2112). Glasgow.

Dümbgen, L. (1998). On Tyler's M-functional of scatter in high dimension. *Annals of Institute of Statistical Mathematics*, *50*, 471-491.

Hettmansperger, T. P., and Randles, R. H. (2002). A practical affine equivariant multivariate median. *Biometrika*, *89*, 851–860.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Hyvärinen, A., and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*, 1483-1492.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*, 411-430.

Marchini, J., Heaton, C., and Ripley, B. (2006). fastICA: FastICA algorithms to perform ICA and projection pursuit [Computer software manual]. (R package version 1.1-8)

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust statistics*. Chichester: Wiley.

Nordhausen, K., Oja, H., and Tyler, D. (2006). ICS: ICS / ICA computation based on two scatter matrices [Computer software manual]. (R package version 0.1-2)

Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, *35*, 175-189.

Ollila, E., Oja, H., and Koivunen, V. (2007). *Complex-valued ICA based on a pair of generalized covariance matrices.* (Conditionally accepted by Computational Statistics & Data Analysis)

R Development Core Team. (2006). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Sirkiä, S., Taskinen, S., and Oja, H. (2007). Symmetrized M-estimators of multivariate scatter. *Journal of Multivariate Analysis*, *98*, 1611-1629.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, *15*, 234-251.

Authors' Addresses:

Klaus Nordhausen
Tampere School of Public Health
FIN-33014 University of Tampere
Finland
E-mail: klaus.nordhausen@uta.fi

Hannu Oja
Tampere School of Public Health
FIN-33014 University of Tampere
Finland
E-mail: hannu.oja@uta.fi

Esa Ollila
Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000
FIN-02015 HUT
Finland
E-mail: esollila@wooster.hut.fi

Festschrift for Tarmo Pukkila on his 60th Birthday
Eds. E. P. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. P. H. Styan

Univ. of Tampere, 2006, ISBN 978-951-44-6620-5, pages 217–231

On the efficiency of invariant multivariate sign and rank tests

KLAUS NORDHAUSEN HANNU OJA DAVID E. TYLER

Abstract. Invariant coordinate selection (ICS) is proposed in Oja and Tyler (2006) for constructing invariant multivariate sign and rank tests. The multivariate data vectors are first transformed to invariant coordinates, and univariate sign and rank tests are then applied to the components of the transformed vectors. In this paper, the powers of different versions of the one sample and two samples location tests are compared via simulation studies.

2000 MSC codes: 62H12, 62G10, 62G05.

Key words and phrases: Hodges Lehmann estimate; Kurtosis; M-estimate; Multivariate median; Transformation and retransformation technique; Wilcoxon test.

1 Introduction

The classical L_1 type univariate sign and rank methods, estimates and tests, have been extended quite recently to the multivariate case. Multivariate extensions of the concepts of sign and rank based on (i) the vector of marginal medians, (ii) the so called spatial median or vector median, and (iii) the affine equivariant Oja median (Oja 1983) have been developed in a series of papers with natural analogues of one-sample, two-sample and multisample sign and rank tests. See e.g. Puri and Sen (1971), Möttönen and Oja (1995), Oja (1999), and Oja and Randles (2004) and references therein. These multivariate location estimates and tests are robust and nonparametric competitors of the classical MANOVA inference methods.

Unfortunately, the tests based on marginal signs and ranks and those based on spatial signs and ranks are not invariant under affine transformations of the observation vectors. Chakraborty and Chaudhuri (1996, 1998) and Chakraborty et al. (1998) introduced and discussed the so called transformation and retransformation technique to circumvent the problem: The data vectors are first linearly transformed back to a new, invariant coordinate system, the tests and estimates are constructed for these new vectors of variables, and, finally, the estimates are linearly retransformed to the original coordinate system. In the one sample and several samples p-variate location

problems, the transformation matrix was then based on p and $p+1$ original observation vectors, respectively.

Other nonparametric approaches for multivariate data analysis include the depth-based rank sum tests introduced by Liu and Singh (1993). The so called zonotopes and lift-zonotopes have been used to describe and investigate the properties of a multivariate distribution, see Mosler (2002). Randles (1989) developed an affine invariant sign test based on *interdirections*, and was followed by a series of papers introducing nonparametric sign and rank interdirection tests for multivariate one-sample and two-sample location problems. These tests are typically asymptotically equivalent with spatial sign and rank tests. Finally, in a series of papers, Hallin and Paindaveine constructed *optimal signed-rank tests* for the location and scatter problems in the elliptical model; see the seminal papers by Hallin and Paindaveine (2002, 2006).

In this paper, as proposed by Oja and Tyler (2006), two different scatter matrices are used to construct an invariant coordinate system. It is remarkable that, in the new coordinate system, the marginal variables are ordered according to their kurtosis. The multivariate variables are first transformed to invariant coordinates, and the univariate sign and rank tests are then applied to these transformed variables. Unlike most other invariant multivariate sign and rank methods, the resulting tests are distribution-free not only at elliptically symmetric models but rather at any symmetric model. The powers of different versions of the one sample and two samples location tests are compared via simulation studies.

Hence the structure of the paper is as follows. In Section 2 we introduce the basic notations and tools that are necessary to construct an invariant coordinate system and show its relationship with the kurtosis of the components. In Section 3 we point out different strategies to use univariate tests on the transformed data components to test the location problem in the one and two sample case. Section 4 gives results of a simulation study which compares the performance of the different strategies. The paper ends with a brief discussion in Section 5. For a complete discussion of this approach, see Oja and Tyler (2006).

2 Invariant coordinate selection (ICS)

2.1 Notations

Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ be independent p-variate observations and write

$$\boldsymbol{Y} = (\boldsymbol{y}_1\ \boldsymbol{y}_2\ \ldots\ \boldsymbol{y}_n)$$

for the corresponding $p \times n$ *data matrix* in the one sample case. In the several samples case, write

$$\boldsymbol{Y} = (\boldsymbol{Y}_1\ \ldots\ \boldsymbol{Y}_c)$$

where $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_c$ are independent random samples with sample sizes $n_1, \ldots, n_c$, $n = n_1 + \cdots + n_c$, from p-variate distributions. In this paper we consider the one sample and two samples multivariate location problems only.

It is often desirable to have statistical methods which are invariant or equivariant under *affine transformations* of the data matrix, i.e. under transformations of the form

$$\boldsymbol{y}_i \to \boldsymbol{A}\boldsymbol{y}_i + \boldsymbol{b}, \qquad i = 1, \ldots, n,$$

or equivalently

$$\boldsymbol{Y} \to \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}\boldsymbol{1}',$$

where $\boldsymbol{A}$ is a full-rank $p \times p$ matrix and $\boldsymbol{b}$ is a p-vector. The vector $\boldsymbol{1}$ is a n-vector full of ones. Some interesting transformations are *orthogonal transformations* ($\boldsymbol{Y} \to \boldsymbol{U}\boldsymbol{Y}$ with $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}$), *sign-change transformations* ($\boldsymbol{Y} \to \boldsymbol{J}\boldsymbol{Y}$ where $\boldsymbol{J}$ is a $p \times p$ diagonal matrix with diagonal elements ± 1), and *permutations* ($\boldsymbol{Y} \to \boldsymbol{P}\boldsymbol{Y}$ where $\boldsymbol{P}$ is a $p \times p$ permutation matrix obtained by successively permuting the rows and/or columns of $\boldsymbol{I}$). Note that transformation $\boldsymbol{Y} \to \boldsymbol{Y}\boldsymbol{P}$ with a $n \times n$ permutation matrix $\boldsymbol{P}$ permutes the observations.

2.2 Location vector and scatter matrices

We start by defining what we mean by a *location statistic*, a *scatter statistic*, and a *scatter statistic with respect to the origin*:

Definition. (i) A p-vector valued statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{Y})$ is called a *location statistic* if it is affine equivariant, that is,

$$\boldsymbol{T}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}\boldsymbol{1}') = \boldsymbol{A}\boldsymbol{T}(\boldsymbol{Y}) + \boldsymbol{b}$$

for all full-rank $p \times p$-matrices $\boldsymbol{A}$ and for all p-vectors $\boldsymbol{b}$.

(ii) Second, $p \times p$ matrix $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{Y}) \geq 0$ is a *scatter statistic* if it is affine equivariant in the sense that

$$\boldsymbol{S}(\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}\boldsymbol{1}') = \boldsymbol{A}\boldsymbol{S}(\boldsymbol{Y})\boldsymbol{A}'$$

for all full-rank $p \times p$-matrices $\boldsymbol{A}$ and for all p-vectors $\boldsymbol{b}$.

(iii) Third, a *scatter statistic with respect to the origin* is affine equivariant in the sense that

$$\boldsymbol{S}(\boldsymbol{A}\boldsymbol{Y}\boldsymbol{J}) = \boldsymbol{A}\boldsymbol{S}(\boldsymbol{Y})\boldsymbol{A}'$$

for all full-rank $p \times p$-matrices $\boldsymbol{A}$ and for all $n \times n$ sign change matrices $\boldsymbol{J}$.

If $\boldsymbol{Y}$ is a random sample, it is also natural to require that the statistics are invariant under permutations of the observations, that is,

$$\boldsymbol{T}(\boldsymbol{Y}\boldsymbol{P}) = \boldsymbol{T}(\boldsymbol{Y}) \quad \text{and} \quad \boldsymbol{S}(\boldsymbol{Y}\boldsymbol{P}) = \boldsymbol{S}(\boldsymbol{Y})$$

for all $n \times n$ permutation matrices $\boldsymbol{P}$.

In the semiparametric elliptic model, for example, the location statistic estimates the unknown center of symmetry $\boldsymbol{\mu}$ and the scatter statistic $\boldsymbol{S}(\boldsymbol{Y})$, possibly multiplied by a correction factor, is an estimate of the regular covariance matrix $\boldsymbol{\Sigma}$ if it exists. Different scatter statistics $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots$ then estimate the same population quantity but have different statistical properties (consistency, efficiency, robustness, computational convenience). In practice, one would choose the one that is most suitable for the problem at hand.

Different location and scatter statistics may also be used to construct skewness and kurtosis statistics; e.g. as in Kankainen et al. (2006),

$$\|\boldsymbol{T}_1 - \boldsymbol{T}_2\|_{\boldsymbol{S}}^2 \quad \text{and} \quad \|\boldsymbol{S}_1^{-1}\boldsymbol{S}_2 - \boldsymbol{I}\|^2$$

that is, the squared Mahalanobis distance between location statistics $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ and the squared matrix norm (Frobenius norm) of $\boldsymbol{S}_1^{-1}\boldsymbol{S}_2 - \boldsymbol{I}$ where $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ (again equipped with correction factors) are different consistent estimates of the regular covariance matrix at the normal model. In this paper we will use two different scatter statistics to transform the data to invariant coordinates. See Section 2.4.

2.3 M-estimates of location and scatter

One of the earliest robust estimates developed for multivariate data are the M-estimates of multivariate location and scatter (Maronna 1976). The pseudo maximum likelihood (ML) estimates, including the regular mean vector and covariance matrix among others, are members of this class. Many other classes of estimates, like the S-estimates, CM-estimates and MM-estimates may be seen as special cases of M-estimates with auxiliary scale (Tyler 2002). M-estimates of location and scatter (one version), $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{Y})$ and $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{Y})$, satisfy implicit equations

$$\boldsymbol{T} = [\text{ave}[w_1(r_i)]]^{-1} \text{ave}[w_1(r_i)\boldsymbol{y}_i]$$

and

$$\boldsymbol{S} = \text{ave}[w_2(r_i)(\boldsymbol{y}_i - \boldsymbol{T})(\boldsymbol{y}_i - \boldsymbol{T})']$$

for some suitably chosen weight functions $w_1(r)$ and $w_2(r)$. The scalar r_i is the Mahalanobis distance between $\boldsymbol{y}_i$ and $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{Y})$, that is, $r_i = \|\boldsymbol{y}_i - \boldsymbol{T}\|_{\boldsymbol{S}}$. Mean vector and covariance matrix are given by the choices $w_1(r) = w_2(r) = 1$.

If $\boldsymbol{T}_1 = \boldsymbol{T}_1(\boldsymbol{Y})$ and $\boldsymbol{S}_1 = \boldsymbol{S}_1(\boldsymbol{Y})$ are any affine equivariant location and scatter functionals then one-step M-functionals $\boldsymbol{T}_2 = \boldsymbol{T}_2(\boldsymbol{Y})$ and $\boldsymbol{S}_2 = \boldsymbol{S}_2(\boldsymbol{Y})$, starting from $\boldsymbol{T}_1$ and $\boldsymbol{S}_1$, are given by

$$\boldsymbol{T}_2 = [\text{ave}[w_1(r_i)]]^{-1} \text{ave}[w_1(r_i)\boldsymbol{y}_i]$$

and

$$\boldsymbol{S}_2 = \text{ave}[w_2(r_i)(\boldsymbol{y}_i - \boldsymbol{T}_1)(\boldsymbol{y}_i - \boldsymbol{T}_1)']$$

where now $r_i = \|\boldsymbol{y}_i - \boldsymbol{T}_1\|_{\boldsymbol{S}_1}$. It is easy to see that $\boldsymbol{T}_2$ and $\boldsymbol{S}_2$ are affine equivariant as well. Repeating this step until it converges yields a solution to the M-estimating equations with weight functions w_1 and w_2. If $\boldsymbol{T}_1$ is the mean vector and $\boldsymbol{S}_1$ is the covariance matrix, then

$$\boldsymbol{T}_2 = \frac{1}{p}\,\text{ave}[r_i^2 \boldsymbol{y}_i] \quad \text{and} \quad \boldsymbol{S}_2 = \frac{1}{p+2}\,\text{ave}[r_i^2(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})']$$

are one-step or reweighted M-estimates of location and scatter. Note that the scatter statistic $\boldsymbol{S}_2 = \boldsymbol{S}_2(\boldsymbol{Y})$ is a *scatter matrix estimate based on fourth moments*. It is consistent for the regular covariance matrix at the multinormal model.

2.4 Invariant coordinate selection

Scatter matrices are often used to standardize the data:

$$\boldsymbol{Y} \to \boldsymbol{Z} = [\boldsymbol{S}(\boldsymbol{Y})]^{-1/2}\boldsymbol{Y}.$$

Transformation matrix $[\boldsymbol{S}(\boldsymbol{Y})]^{-1/2}$ thus yields the new coordinate system with uncorrelated components (in the sense of $\boldsymbol{S}$). Unfortunately, this new coordinate system is not invariant under affine transformations; it is only true that

$$[\boldsymbol{S}(\boldsymbol{AY})]^{-1/2}(\boldsymbol{AY}) = \boldsymbol{U}[\boldsymbol{S}(\boldsymbol{Y})]^{-1/2}\boldsymbol{Y}$$

with an orthogonal matrix $\boldsymbol{U}$ depending on $\boldsymbol{Y}$, $\boldsymbol{A}$ and $\boldsymbol{S}$.

Two different scatter functionals $\boldsymbol{S}_1 = \boldsymbol{S}_1(\boldsymbol{Y})$ and $\boldsymbol{S}_2 = \boldsymbol{S}_2(\boldsymbol{Y})$ may be used to find an invariant coordinate system as follows. For a more detailed discussion of the *invariant coordinate selection (ICS)*, see Oja and Tyler (2006). Starting with $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, define a $p \times p$ transformation matrix $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{Y})$ and a diagonal matrix $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{Y})$ by

$$\boldsymbol{S}_2^{-1}\boldsymbol{S}_1\boldsymbol{B}' = \boldsymbol{B}'\boldsymbol{D}$$

that is, $\boldsymbol{B}$ gives the eigenvectors of $\boldsymbol{S}_2^{-1}\boldsymbol{S}_1$. The following result can then be shown to hold.

Result 1. The transformation $\boldsymbol{Y} \to \boldsymbol{Z} = \boldsymbol{B}(\boldsymbol{Y})\boldsymbol{Y}$ yields an *invariant coordinate system* in the sense that

$$\boldsymbol{B}(\boldsymbol{AY})(\boldsymbol{AY}) = \boldsymbol{J}\boldsymbol{B}(\boldsymbol{Y})\boldsymbol{Y}$$

for some $p \times p$ sign change matrix $\boldsymbol{J}$. Matrix $\boldsymbol{B}$ can be made unique by requiring that the element with largest absolute value in each row of $\boldsymbol{B}$ is positive.

2.5 Kurtosis and ICS

Let $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{Y})$ be the transformation matrix yielded by $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. Observe that the elements of $\boldsymbol{Z} = \boldsymbol{B}(\boldsymbol{Y})\boldsymbol{Y}$ are now standardized with respect to $\boldsymbol{S}_1$ and uncorrelated with respect to $\boldsymbol{S}_2$, that is,

$$\boldsymbol{S}_1(\boldsymbol{Z}) = \boldsymbol{I} \quad \text{and} \quad \boldsymbol{S}_2(\boldsymbol{Z}) = \boldsymbol{D}$$

where $\boldsymbol{D}$ is a diagonal matrix. The diagonal elements of $\boldsymbol{D}$ yield the kurtosis measures for the components. Therefore the components of $\boldsymbol{Z}$ are *ordered with respect to kurtosis*. Recall the discussion on kurtosis in Section 2.3.

In the simulations in this paper we use the invariant coordinate selection based on the regular covariance matrix $\boldsymbol{S}_1$ and the scatter matrix $\boldsymbol{S}_2$ based on the fourth moments. The jth diagonal element of matrix $\boldsymbol{D}$ is then

$$D_{jj} = \frac{1}{p+2}\,\mathrm{ave}_i\{z_{ij}^2(z_{i1}^2 + \cdots + z_{ip}^2)\}, \qquad j = 1, \ldots, p.$$

Consider the case having some special interest in our simulations: Assume that $\boldsymbol{Y} = \{\boldsymbol{y}_1 \ \ldots \ \boldsymbol{y}_n\}$ is a random sample from a distribution which is a mixture of two multivariate normal distribution differing only in location: $\boldsymbol{y}_i$ has a $N_p(\boldsymbol{0}, \boldsymbol{I})$-distribution with probability $1 - \varepsilon$ and a $N_p(\Delta \boldsymbol{e}_p, \boldsymbol{I})$-distribution with probability ε ($\varepsilon \leq 0.5$). (The last element in vector $\boldsymbol{e}_p$ is one, other elements are zero.) Then $\boldsymbol{S}_1(\boldsymbol{Y}) \to_p \boldsymbol{I}$ and $\boldsymbol{S}_2(\boldsymbol{Y}) \to_p \boldsymbol{D}$ where $\boldsymbol{D}$ is a diagonal matrix with $D_{11} = \cdots = D_{p-1,p-1} = 1$. The last diagonal element is $1 + b_2/(p+2)$ where b_2 is the *classical univariate kurtosis* measure for the last component. Note that the last component has the highest kurtosis for $\varepsilon < (3 + \sqrt{3})^{-1}$ and lowest kurtosis otherwise (compare Preston 1953). Also the amount of kurtosis strongly depends on the value of Δ; the greater Δ the larger is the absolute value of kurtosis. This behavior is visualized in Figures 1 and 2.

3 Invariant sign and rank tests

3.1 Marginal signs and ranks

Let $\boldsymbol{z}_i$, $i = 1, \ldots, n$, be the p-variate residuals in the multivariate location case, and consider the L_1 type criterion functions

$$\mathrm{ave}_i\{|z_{i1}| + \cdots + |z_{ip}|\} \quad \text{and} \quad \mathrm{ave}_{i,j}\{|z_{i1} - z_{j1}| + \cdots + |z_{ip} - z_{jp}|\}.$$

The resulting L_1 estimates are the vectors of marginal medians and marginal Hodges-Lehmann estimates. The corresponding score tests are based on the vectors of marginal (univariate) signs or marginal (univariate) ranks. See Puri and Sen (1971) for a complete discussion of this approach. The inference methods are invariant/equivariant under componentwise rescaling

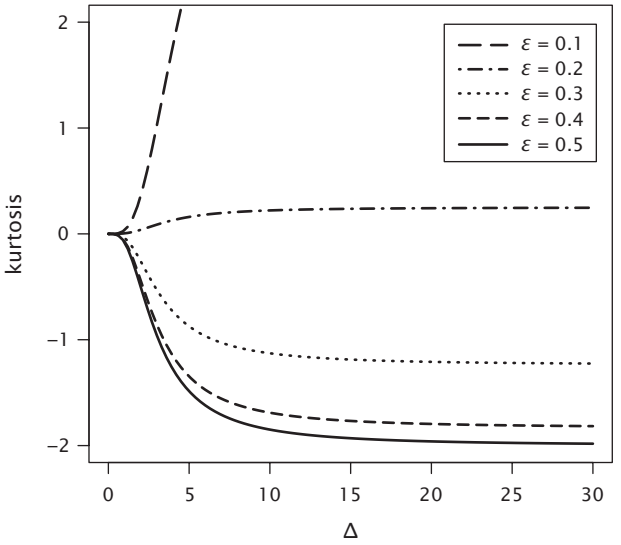

Figure 1. Kurtosis for a location mixture of normal distributions as a function of Δ for different ε.

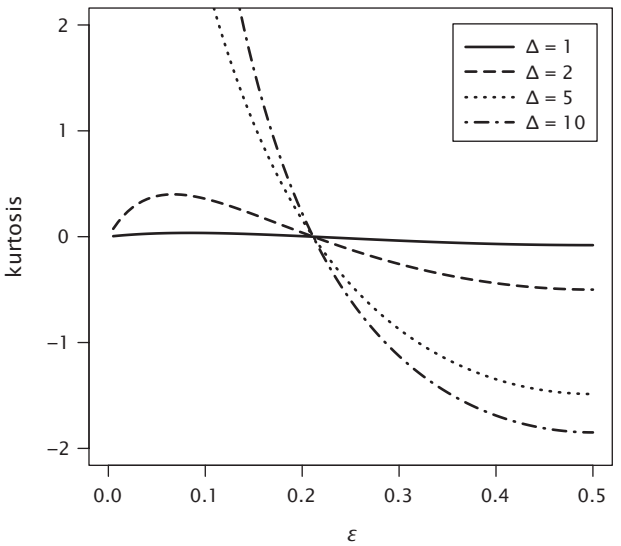

Figure 2. Kurtosis for a location mixture of normal distributions as a function of ε for different Δ.

but not orthogonally invariant/equivariant. The efficiencies do not exceed the univariate efficiencies and are quite low if the margins are highly correlated.

Invariant test versions can be obtained by first transforming the data to invariant coordinates. The use of the standardized data set $[\boldsymbol{S}(\boldsymbol{Y})]^{-1/2}\boldsymbol{Y}$ does not help as the standardization is not affine invariant. See Section 2.4. Chakraborty and Chaudhuri (1996, 1998) avoided the problem by using p observations with indices listed in $\alpha = (i_1, \ldots, i_p)$, $1 \le i_1 < \cdots < i_p \le n$, to construct, in the one-sample location case, a transformation matrix $\boldsymbol{B}(\alpha) = (\boldsymbol{y}_{i_1}\ \boldsymbol{y}_{i_2}\ \ldots\ \boldsymbol{y}_{i_p})^{-1}$. Now clearly $\boldsymbol{B}(\alpha)\boldsymbol{Y}$ is invariant under affine transformations $\boldsymbol{Y} \to \boldsymbol{AY}$ and the data set $\boldsymbol{B}(\alpha)\boldsymbol{Y}$ may then be used for invariant one-sample test construction. In the several sample case, they choose $\alpha = (i_1, \ldots, i_{p+1})$, $1 \le i_1 < \cdots < i_{p+1} \le n$ and $\boldsymbol{B}(\alpha) = (\boldsymbol{y}_{i_1} - \boldsymbol{y}_{i_{p+1}}\ \boldsymbol{y}_{i_2} - \boldsymbol{y}_{i_{p+1}}\ \ldots\ \boldsymbol{y}_{i_p} - \boldsymbol{y}_{i_{p+1}})^{-1}$. This technique is then called the *transformation and re-transformation* (*TR*) *technique*. The problem naturally is how to choose α, that is, the coordinate system in an optimal adaptive way. Techniques proposed for choosing α tend to be computationally intensive since they require optimizing some criterion over all possible subsets of size $p + 1$ from the sample. In the following we use the computationally simple invariant coordinate selection method based on two scatter matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.

3.2 One sample case

Let $\boldsymbol{Y} = (\boldsymbol{y}_1\ \ldots\ \boldsymbol{y}_n)$ be a random sample from a p-variate continuous distribution symmetric around unknown $\boldsymbol{\mu}$. We wish to test the null hypothesis H_0: $\boldsymbol{\mu} = \boldsymbol{0}$ and estimate the unknown $\boldsymbol{\mu}$. For the test, let $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ be two scatter matrices with respect to the origin. Assume also that they are invariant under permutations to the observations. Then, for $k = 1, 2$,

$$\boldsymbol{S}_k(\boldsymbol{AYPJ}) = \boldsymbol{A}\boldsymbol{S}_k(\boldsymbol{Y})\boldsymbol{A}', \qquad \forall\, \boldsymbol{A}, \boldsymbol{P}, \boldsymbol{J},$$

and therefore

$$\boldsymbol{B}(\boldsymbol{YJP}) = \boldsymbol{B}(\boldsymbol{Y})$$

As, under the null hypothesis, $\boldsymbol{Y}$ is a random sample from distribution symmetric around the origin, it is also true that

$$\boldsymbol{Z}(\boldsymbol{Y}) \sim \boldsymbol{Z}(\boldsymbol{Y})\boldsymbol{JP}, \qquad \forall\, \boldsymbol{J}, \boldsymbol{P}.$$

Clearly $\boldsymbol{Z} = (\boldsymbol{z}_1\ \ldots\ \boldsymbol{z}_n)$ is not a random sample any more. However, under the null hypothesis, the variables in $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ are exchangeable.

Consider next the jth component of the $\boldsymbol{z}_i$ vectors, that is, the observations $(z_{j1}, \ldots, z_{jn})$. Then, it is easy to see that

Result 2. Under the null hypothesis, the univariate *sign test statistic*

$$U_j = \sum_{i=1}^{n} I(z_{ji} > 0) \sim \mathrm{Bin}(n, 0.5).$$

Thus, for all $j = 1, \ldots, p$, U_j is an invariant distribution-free multivariate sign test statistic. Unfortunately, the p sign test statistics $U_1, \ldots, U_p$ are not mutually independent.

Let next R_{ji}^+ be the rank of $|z_{ji}|$ among $|z_{j1}|, \ldots, |z_{jn}|$. The univariate *Wilcoxon signed-rank test statistic*

$$W_j = \sum_{i=1}^{n} \operatorname{sgn}(z_{ji}) R_{ji}^+$$

is then distribution-free as well:

Result 3. Under the null hypothesis, the distribution of W_j is that of the one-sample Wilcoxon signed-rank test statistic.

The result easily follows from the facts that $\operatorname{sgn}(z_{j1}), \ldots, \operatorname{sgn}(z_{jn})$ are iid and independent of $(|z_{j1}|, \ldots, |z_{jn}|)$. Also, $|z_{j1}|, \ldots, |z_{jn}|$ are exchangeable.

All the test statistics $U_1, \ldots, U_p$ and $W_1, \ldots, W_p$ are thus distribution-free but dependent (the dependence structure depends on the background distribution). How then to choose U_j or W_j, or how to combine these statistics for the testing problem? One goal of the present paper then is to provide some insight into this rather complex question. As the components are ordered according to their kurtosis, and one expects to see a high absolute value of kurtosis in the direction of $\boldsymbol{\mu}$, often the last (or first) component is most powerful and contains the most information. This fact can be utilised when constructing the "overall" test statistic where one can choose between different strategies. For example one could use only the first or only the last component or those two components combined. One could also use a rule like use the $k \leq p$ components with the highest absolute value of kurtosis or one could simply use all components.

The corresponding *affine equivariant location estimates* are obtained as follows: Let $\boldsymbol{T}$ be the vector of marginal medians or the vector of marginal Hodges-Lehmann estimators. These estimates are not location statistics as they are not affine equivariant. Let $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{Y})$ be the transformation based on two scatter matrix estimates. Then multivariate affine equivariant *transformation-retransformation median* and *Hodges-Lehmann estimate* are obtained as

$$\tilde{\boldsymbol{T}}(\boldsymbol{Y}) = \boldsymbol{B}^{-1} \boldsymbol{T}(\boldsymbol{B}\boldsymbol{Y})$$

3.3 Two samples case

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1 \ \boldsymbol{Y}_2)$ where $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent random samples of sizes n_1 and n_2, $n = n_1 + n_2$, from p-variate continuous distributions with cumulative density functions $F(\boldsymbol{y})$ and $F(\boldsymbol{y} - \boldsymbol{\mu})$, respectively. We wish to test the null hypothesis H_0: $\boldsymbol{\mu} = \boldsymbol{0}$ and estimate the unknown location shift $\boldsymbol{\mu}$. Let $\boldsymbol{S}_1 = \boldsymbol{S}_1(\boldsymbol{Y})$ and $\boldsymbol{S}_2 = \boldsymbol{S}_2(\boldsymbol{Y})$ be two scatter matrices *calculated from*

the combined data set and invariant under permutations to the observations. This is to say that, for $k = 1, 2$,

$$\boldsymbol{S}_k\big((\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}\boldsymbol{1}')\boldsymbol{P}\big) = \boldsymbol{A}\boldsymbol{S}_k(\boldsymbol{Y})\boldsymbol{A}', \qquad \forall\, \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{P},$$

and $\boldsymbol{B}(\boldsymbol{Y}\boldsymbol{P}) = \boldsymbol{B}(\boldsymbol{Y})$. Under the null hypothesis, the combined sample $\boldsymbol{Y} = (\boldsymbol{Y}_1 \boldsymbol{Y}_2)$ is a random sample of size n, and

$$Z(\boldsymbol{Y}) \sim Z(\boldsymbol{Y})\boldsymbol{P}, \qquad \forall\, \boldsymbol{P}.$$

Again, $\boldsymbol{Z} = (\boldsymbol{z}_1 \ \ldots \ \boldsymbol{z}_n)$ is not a random sample but, under the null hypothesis, the variables in $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ are exchangeable.

Affine invariant distribution-free multivariate rank tests may be constructed as follows. Let now R_{ji} be the rank of z_{ji} among $z_{j1}, \ldots, z_{jn}$. As $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are exchangeable,

Result 4. Under the null hypothesis the distribution of the univariate *Wilcoxon rank test statistic*

$$W_j = \sum_{i=n_1+1}^{n} R_{ji}$$

is that of regular two samples Wilcoxon test statistic with sample sizes n_1 and n_2.

General rank score test statistics $\sum_{i=n_1+1}^{n} a(R_{ji})$ may be constructed as well. The *two samples sign test statistic* (*Mood's test statistic*) is given by the choice $a(i) = 1(0)$ for $i > (\leq)(n+1)/2$. All the test statistics $W_1, \ldots, W_p$ are thus distribution-free but unfortunately dependent (the dependence structure depends on the background distribution). The question of which of those test statistics to use for the decision making allows the same strategies as in the one sample case.

Corresponding *affine equivariant multivariate shift estimates* are obtained as follows: Let $\boldsymbol{T}$ be the vector of marginal difference of the medians (Mood's test) or the vector of marginal two-sample Hodges-Lehmann shift estimators (Wilcoxon test). These estimates are not affine equivariant. Let $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{Y})$ be the transformation based on two scatter matrix estimates. Then multivariate affine equivariant *transformation retransformation* estimates are again obtained as

$$\tilde{\boldsymbol{T}}(\boldsymbol{Y}) = \boldsymbol{B}^{-1}\boldsymbol{T}(\boldsymbol{B}\boldsymbol{Y})$$

4 Simulation results

As mentioned in Section 3.2 and 3.3, several strategies are available for the decision making. We performed a simulation study to compare the following strategies in the one and two sample case:

(i) Using a componentwise sign test and signed rank test as described in Puri and Sen (1971) based on all p components, denoted as $U[1:p]$, respectively as $W[1:p]$.

(ii) Using the same componentwise sign test and signed rank test as before but only to combine the first and last component, denoted as $U[1,p]$, respectively as $W[1,p]$.

(iii) Using an exact sign test respectively a Wilcoxon signed rank test for the last component only, denoted as $U[p]$, respectively as $W[p]$.

for different sample sizes and underlying distributions. As a reference test also Hotelling's T^2 for the original observations is included. We note that both the exact and asymptotic distributions for case (i) and (ii) are still open questions. To approximate their distributions we suggest using distributions analogous to the asymptotic distributions given by Puri and Sen (1971), and conjecture that these approximate distributions are asymptotically correct. A size simulation (not shown here) supports this conjecture.

All simulations are based on 5000 repetitions and were performed using R 2.2.0 (R Development Core Team 2005) at the level $\alpha = 0.05$. The critical values for the tests were based on the limiting null distributions.

Not shown in the following subsections are results for the strategy which uses only the component with the largest absolute value of the kurtosis since this strategy had in all settings in the one sample case always less power than strategy (iii) and in the two sample case it was less powerful or about equal when compared to strategy (iii).

4.1 One sample case

In this simulation we obtained the ICS with respect to the origin as described in Section 2.5 for data coming from a normal distribution and t_3 and t_{10} distributions for different dimensions and sample sizes.

A size simulation (not shown here) yielded for all tests the designated level except for $U[p]$ which was always smaller than 0.05 due to the discreteness of the test statistic and for Hotelling's T^2 for heavy tailed distributions and small sample sizes.

To compare the power of the different strategies the location parameter of the distributions were set to $\boldsymbol{\mu}_0 = (\Delta, 0, \ldots, 0)'$ and Δ in such a way chosen, that given the dimension p and the sample size n the power of Hotelling's T^2 is 0.5 under normality. This means

$$P[F(p, n-p, \delta) > F_\alpha(p, n-p)] = 0.5$$

where $F(p, n-p, \delta)$ is a random variable having a noncentral F distribution with degrees of freedom p and $n-p$ and noncentrality parameter $\delta = \frac{1}{n}\Delta^2$ and $F_\alpha(p, n-p)$ is the $1-\alpha$ quantile of $F(p, n-p) = F(p, n-p, 0)$. This gives in our case a range for Δ from 0.159 to 0.471.

The simulation results provided in Table 1 show that there is a lot of information in the last component, however the power of the strategies increases with the number of components they are based on and strategy (iii) can therefore not compete with strategy (i). Especially the signed rank test $W[1:p]$ can be seen as a serious competitor to Hotelling's T^2 since it is almost as efficient as Hotelling's T^2 under normality and more efficient for heavier tails.

Table 1. Simulated power in the one sample case in number of rejections per 1000 cases.

				sign tests			signed rank tests		
Dist.	p	n	T^2	$U[1:p]$	$U[1,p]$	$U[p]$	$W[1:p]$	$W[1,p]$	$W[p]$
normal	2	50	499	340	340	208	472	472	323
		200	502	333	333	220	479	479	327
	5	50	500	281	180	122	441	257	203
		200	508	319	197	137	472	283	213
	10	50	507	204	124	89	385	168	159
		200	503	288	140	104	458	195	152
t_{10}	2	50	415	317	317	194	417	417	309
		200	413	324	324	213	429	429	298
	5	50	405	256	180	138	387	235	211
		200	414	301	195	139	419	255	193
	10	50	427	191	124	92	334	166	158
		200	409	283	138	101	417	185	147
t_3	2	50	261	286	286	180	334	334	257
		200	221	281	281	193	334	334	249
	5	50	244	237	169	117	299	200	182
		200	215	267	177	129	315	205	168
	10	50	270	173	130	95	267	155	149
		200	213	246	131	105	313	153	135

4.2 Two samples case

The setup for the two sample simulations are of a similar fashion as in the one sample case. The size simulation (also not shown here) gave similar results as in the one sample case, namely that the size of $U[p]$ was always smaller than 0.05 and also Hotelling's T^2 was smaller for heavier tails when the sample size was small.

The difference of the population locations $\boldsymbol{\mu}_0 = (\Delta, 0 \ldots, 0)'$ was set also in such a way that under normality Hotelling's T^2 would achieve a power of

0.5. The corresponding value of Δ can then be computed via

$$P[F(p, n-p-1, \delta) > F_\alpha(p, n-p-1)] = 0.5$$

where the noncentrality parameter δ is given as $\delta = \frac{n_1 n_2}{n_1+n_2}\Delta^2$. This gives a range for Δ from 0.223 to 0.637.

Table 2 shows the results for the two sample power simulations where in two settings the two populations are of equal size and in one setting the mixture probability is $\varepsilon = 0.2$ (compare Section 2.5).

The same conclusions as for the one sample case apply basically also for the two sample case except one surprising occurrance for the rank test $W[1:p]$ where the power drops considerably when the dimension and the sample sizes of both populations are large.

Table 2. Simulated power in the two sample case in number of rejections per 1000 cases.

					sign tests			signed rank tests		
Dist.	p	n_1	n_2	T^2	$U[1:p]$	$U[1,p]$	$U[p]$	$W[1:p]$	$W[1,p]$	$W[p]$
normal	2	50	50	504	321	321	177	482	482	321
		200	50	494	326	326	205	477	477	332
		200	200	494	329	329	203	477	477	317
	5	50	50	504	307	201	117	475	278	210
		200	50	491	309	191	137	464	267	199
		200	200	507	316	203	130	482	292	211
	10	50	50	499	259	136	86	449	192	159
		200	50	484	282	144	99	443	198	154
		200	200	501	212	145	92	310	199	153
t_{10}	2	50	50	404	304	304	170	423	423	297
		200	50	405	310	310	214	418	418	307
		200	200	409	306	306	195	421	421	294
	5	50	50	402	277	182	114	409	252	207
		200	50	400	290	195	135	422	256	201
		200	200	393	290	191	121	405	255	194
	10	50	50	422	251	130	89	416	186	167
		200	50	414	293	142	97	421	179	146
		200	200	410	189	132	86	280	176	138
t_3	2	50	50	233	277	277	160	334	334	244
		200	50	219	278	278	179	330	330	239
		200	200	214	285	285	179	336	336	246
	5	50	50	233	249	177	102	320	221	175
		200	50	213	254	181	122	321	211	164
		200	200	194	268	174	110	318	210	152
	10	50	50	230	204	123	72	296	145	132
		200	50	209	241	136	99	306	152	126
		200	200	197	183	135	84	211	149	119

5 Final comments

This simulation study serves as an introduction to the use of two different scatter matrices to obtain an ICS where invariant sign and rank tests can be constructed. It is obvious that invariance of the test statistics is a worthwhile aim to pursuit and the ICS is a promising tool to achieve this goal and has for example compared to the TR technique the advantage that not p, respectively $p+1$, data points have to be singled out on which the transformation depends on. However for the ICS a choice of the two scatter matrices must be made and further research is necessary to compare the effect of different choices. For instance from a nonparametric point of view the assumption of fourth order moments as in this study is not fortunate. Also surprising for us was that contrary to the spatial sign test in the elliptical case for large n and p the efficiencies of the tests used here seem not to tend to 1 in the two sample case.

Another point to pursuit would be the efficiencies of the tests for different values of Δ which would occur for example if a larger power for Hotelling's T^2 would be required because then, as can be seen in Figure 1, the main direction of the data would become more distinct given in the two sample case that the mixing probability ε would be not too close to $1/(3+\sqrt{3})$.

Acknowledgements

The work of Dave Tyler was supported by the NSF Grant DMS-0305858. The work of Klaus Nordhausen and Hannu Oja was supported by grants from Academy of Finland.

References

Chakraborty, B. and Chaudhuri, P. (1996). On a transformation retransformation technique for constructing affine equivariant multivariate median. *Proceedings of American Mathematical Society*, 124, 1529–1537.

Chakraborty, B. and Chaudhuri, P. (1998). On an adaptive transformation retransformation affine equivariant estimate of multivariate location. *Journal of the Royal Statistical Society*, Series B, 60, 145–157.

Chakraborty, B., Chaudhuri, P., and Oja, H. (1998). Operating transformation retransformation on spatial median and angle test. *Statistica Sinica*, 8, 767–784.

Hallin, M. and Paindaveine, D. (2002). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Annals of Statistics*, 30, 1103–1133.

Hallin, M. and Paindaveine, D. (2006). Optimal rank-based tests for sphericity. *Annals of Statistics*, to appear.

Kankainen, A., Taskinen, S., and Oja, H. (2006). Tests of multinormality based on location vectors and scatter matrices. Submitted.

Liu, R.Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88, 252–260.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics*, 17, 1608–1630.

Mosler, K. (2002). *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach*. Lecture Notes in Statistics, Vol. 165. New York: Springer.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5, 201–213.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1, 327–332.

Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scandinavian Journal of Statistics*, 26, 319–343.

Oja, H. and Randles, R. (2004). Multivariate nonparametric tests. *Statistical Science*, 19, 598–605.

Oja, H. and Tyler, D. E. (2006). Invariant multivariate sign and rank tests. *Manuscript in preparation*.

Preston, E. J. (1953). A graphical method for the analysis of statistical distributions into two normal components. *Biometrika*, 40, 460–464.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. New York: Wiley & Sons.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Randles, R. H. (1989). A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, 84, 1045–1050.

Tyler, D. E. (2002). High breakdown point multivariate M-estimation. *Estadistica*, 52, 213–247.

Klaus Nordhausen
Tampere School of Public Health
FI-33014 University of Tampere, Finland
Klaus.Nordhausen@uta.fi
http://www.uta.fi/~klaus.nordhausen/

Hannu Oja
Tampere School of Public Health
FI-33014 University of Tampere, Finland
Hannu.Oja@uta.fi
http://www.uta.fi/~hannu.oja/

David E. Tyler
Department of Statistics
The State University of New Jersey
Piscataway NJ 08854, USA
dtyler@rci.rutgers.edu
http://www.rci.rutgers.edu/~dtyler/

ARTICLE IN PRESS

Journal of Multivariate Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Signed-rank tests for location in the symmetric independent component model

Klaus Nordhausen [a,*], Hannu Oja [a], Davy Paindaveine [b]

[a] Tampere School of Public Health, University of Tampere, 33014 University of Tampere, Finland

[b] E.C.A.R.E.S., Institut de Recherche en Statistique, and Département de Mathématique, Université Libre de Bruxelles, Campus de la Plaine CP 210, 1050 Bruxelles, Belgium

ARTICLE INFO

Article history:
Received 18 December 2007
Available online xxxx

AMS subject classifications:
primary 62G10
secondary 62H15

Keywords:
One-sample location problem
Rank tests
Independent component models
Elliptical symmetry
Hotelling's test
Local asymptotic normality

ABSTRACT

The so-called independent component (IC) model states that the observed p-vector X is generated via $X = \Lambda Z + \mu$, where μ is a p-vector, Λ is a full-rank matrix, and the centered random vector Z has independent marginals. We consider the problem of testing the null hypothesis $\mathcal{H}_0 : \mu = 0$ on the basis of i.i.d. observations $X_1, \ldots, X_n$ generated by the symmetric version of the IC model above (for which all ICs have a symmetric distribution about the origin). In the spirit of [M. Hallin, D. Paindaveine, Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks, Annals of Statistics, 30 (2002), 1103–1133], we develop nonparametric (signed-rank) tests, which are valid without any moment assumption and are, for adequately chosen scores, locally and asymptotically optimal (in the Le Cam sense) at given densities. Our tests are measurable with respect to the marginal signed ranks computed in the collection of null residuals $\hat{\Lambda}^{-1}X_i$, where $\hat{\Lambda}$ is a suitable estimate of Λ. Provided that $\hat{\Lambda}$ is affine-equivariant, the proposed tests, unlike the standard marginal signed-rank tests developed in [M.L. Puri, P.K. Sen, Nonparametric Methods in Multivariate Analysis, Wiley & Sons, New York, 1971] or any of their obvious generalizations, are affine-invariant. Local powers and asymptotic relative efficiencies (AREs) with respect to Hotelling's T^2 test are derived. Quite remarkably, when Gaussian scores are used, these AREs are always greater than or equal to one, with equality in the multinormal model only. Finite-sample efficiencies and robustness properties are investigated through a Monte Carlo study.

1. Introduction

Let $X_1, \ldots, X_n$ be a sample of p-variate random vectors generated by the location-scatter model

$$X_i = \Lambda Z_i + \mu, \quad i = 1, \ldots, n,$$

where the p-vector μ is the location center, the full-rank $p \times p$ matrix Λ is called the *mixing matrix*, and the Z_i's are i.i.d. *standardized* p-variate random vectors. We consider the multivariate one-sample location problem, that is, we wish to test $\mathcal{H}_0 : \mu = 0$ versus $\mathcal{H}_1 : \mu \neq 0$ (any other null value μ_0 can be tested by replacing X_i with $X_i - \mu_0$). Of course, different standardizations of the Z_i's lead to different location-scatter models—and to different definitions of μ and Λ. Such models include

- *The multinormal model:* Z_i has a standard multinormal distribution. This is a parametric model with mean vector μ and covariance matrix $\Sigma = \Lambda\Lambda'$.

* Corresponding author.
E-mail address: klaus.nordhausen@uta.fi (K. Nordhausen).

0047-259X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.08.004

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

- *The elliptic model:* Z_i has a spherical distribution around the origin ($OZ_i \stackrel{\mathcal{D}}{=} Z_i$ for any orthogonal $p \times p$ matrix O; throughout, $\stackrel{\mathcal{D}}{=}$ stands for equality in distribution) with $\text{Med}[\|Z_i\|^2] = \chi^2_{p,.5}$, where $\text{Med}[\cdot]$ denotes the population median and $\chi^2_{\ell,\alpha}$ denotes the α quantile of the χ^2_ℓ distribution. This is a semiparametric model with symmetry center μ and scatter matrix $\Sigma = \Lambda\Lambda'$ (in the multinormal submodel, Σ is the covariance matrix).
- *The symmetric independent component (IC) model:* the components of Z_i are independent and symmetric ($-Z_i^{(r)} \stackrel{\mathcal{D}}{=} Z_i^{(r)}$) with $\text{Med}[(Z_i^{(r)})^2] = \chi^2_{1,.5}$, $r = 1, \ldots, p$. This is a semiparametric model with symmetry center μ and mixing matrix Λ (again, in the multinormal submodel, $\Sigma = \Lambda\Lambda'$ is the covariance matrix). This model is used in the so-called independent component analysis (ICA), where the problem is to estimate Λ.
- *The symmetric nonparametric model:* Z_i has a distribution symmetric around the origin ($-Z_i \stackrel{\mathcal{D}}{=} Z_i$). Then, neither Λ nor Σ are uniquely defined.

Note that the semiparametric/nonparametric models above do not require any moment assumption, and that μ, irrespective of the model adopted, is properly identified as the center of symmetry of X_i. The assumption of symmetry is common in the one-sample location case. It is for example quite natural in the classical matched pairs design for the comparison of two treatments: if for pair i, $i = 1, \ldots, n$, the response variable is $X_{1i} = Y_i + \varepsilon_{1i} + \mu_1$ for treatment 1 and $X_{2i} = Y_i + \varepsilon_{2i} + \mu_2$ for treatment 2, with mutually independent Y_i, ε_{1i}, and ε_{2i} ($\stackrel{\mathcal{D}}{=} \varepsilon_{1i}$), then the difference used in the analysis, namely $X_i = X_{2i} - X_{1i}$, is symmetric about $\mu = \mu_2 - \mu_1$. The literature proposes a vast list of multivariate one-sample location tests. Some of the tests do not require symmetry; note however that only in the symmetric case the different tests are for the same population quantity. The tests include.

- *The Hotelling's T^2 test*, which is equivalent to the Gaussian likelihood ratio test (and actually is uniformly most powerful affine-invariant at the multinormal), is asymptotically valid (i.e., asymptotically meets the nominal level constraint) under any distribution with finite variances. However, its power is poor away from the multinormal (particularly so under heavy tails), and it is also very sensitive to outlying observations.
- *The optimal signed-rank scores tests by* Hallin and Paindaveine [1,2] are based on standardized spatial signs (or Randles' interdirections; see [3] for the corresponding sign test) and the ranks of Mahalanobis distances between the data points and the origin. They do not require any moment assumption and are optimal (in the Le Cam sense) at correctly specified (elliptical) densities. They are affine-invariant, robust, and highly efficient under a broad range of densities (AREs of their Gaussian-score version with respect to Hotelling's test are uniformly larger than or equal to one in the elliptic model). Later [4] showed that interdirections together with the so-called lift-interdirections allow for building hyperplane-based versions of these tests. All these tests however strictly require ellipticity.
- *The signed-rank scores tests by* Puri and Sen [5] combine marginal signed-rank scores tests in the widest symmetric nonparametric model. Unfortunately, these tests are not affine-invariant and may be poorly efficient for dependent margins. Invariant tests are obtained if the data points are first transformed to invariant coordinates; see [6,7].
- *The spatial sign and signed-rank tests* (see [8] for a review), which are based on spatial signs and signed ranks, are also valid in the symmetric nonparametric model. They improve over the Puri and Sen tests in terms of efficiency, but not in terms of affine-invariance. Again, affine-invariance can be achieved if the data is first transformed by using any scatter matrix (the spatial sign test based on Tyler's scatter matrix [9] is strictly distribution-free in the elliptic model and even in the wider directional elliptic model; see [10]).
- *The sign and signed-rank tests by* Hettmansperger et al. [11,12] are based on multivariate Oja signs and ranks. They can be used in all models above, are asymptotically equivalent to spatial sign and signed-rank tests in the spherical case, and are affine-invariant. However, at the elliptic model, their efficiency (as well as that of the spatial sign and signed-rank tests) may be poor when compared with the Hallin and Paindaveine tests.

Only the [1,2,4] tests combine robustness and affine-invariance with a locally optimal – and uniformly excellent – power behavior. The required ellipticity assumption, however, may not be appropriate in practice. This model assumption is often easily discarded just by a visual inspection of bivariate scatter plots or marginal density plots; equidensity contours should be elliptical, and the marginal densities should be similar in shape. The IC model which serves as an alternative extension of the multivariate normal model cannot be ruled out as easily in practice. Of course, more statistical tools should be developed for the important model choice problem.

This paper introduces signed-rank tests which enjoy the nice properties of the Hallin and Paindaveine ones (absence of moment assumption, robustness, affine-invariance, Le Cam optimality at prespecified densities, uniform dominance over Hotelling for Gaussian scores, etc.), but are valid in the *symmetric IC model*. The proposed tests are marginal signed-rank tests (with optimal scores) applied to the residuals $\hat{\Lambda}^{-1}X_i$, $i = 1, \ldots, n$, where $\hat{\Lambda}$ is a suitable (see Section 3) estimate of the mixing matrix Λ. Although they are based on marginal signed-rank statistics, our tests, unlike the marginal Puri and Sen signed-rank tests or any of their obvious generalizations, are affine-invariant.

The outline of the paper is as follows. Section 2 defines more carefully the IC models under consideration. Section 3 introduces the proposed tests and studies their asymptotic null behavior. In Section 4, we explain how to choose score functions to achieve Le Cam optimality at prespecified densities, derive the local powers of our tests under contiguous alternatives, and compute their AREs with respect to Hotelling's T^2 test. Section 5 discusses the practical implementation

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

of our tests and presents simulations that investigate their finite-sample efficiencies and robustness properties. Finally, the appendix collects proofs of technical results.

2. IC models and identifiability

In the absolutely continuous case, the IC model will be indexed by the location vector μ, mixing matrix Λ, and the pdf g of the standardized vectors. The location vector μ is a p-vector and Λ belongs to the collection $\mathcal{M}_p$ of invertible $p \times p$ matrices. As for g, it throughout belongs to the collection $\mathcal{F}$ of densities of absolutely continuous p-vectors $Z = (Z^{(1)}, \ldots, Z^{(p)})'$ whose marginals are (i) mutually independent, (ii) symmetric about the origin (i.e., $-Z^{(r)} \stackrel{\mathcal{D}}{=} Z^{(r)}$ for all r), and (iii) standardized so that $\mathrm{Med}[(Z^{(r)})^2] = \chi^2_{1,.5}$ for all $r = 1, \ldots, p$. Any $g \in \mathcal{F}$ of course decomposes into $z = (z^{(1)}, \ldots, z^{(p)})' \mapsto g(z) =: \prod_{r=1}^p g_r(z^{(r)})$. Denote then by $P^n_{\mu,\Lambda,g}$, $g \in \mathcal{F}$, the hypothesis under which the p-variate observations $X_1, \ldots, X_n$ are generated by the model $X_i = \Lambda Z_i + \mu$, $i = 1, \ldots, n$, where $Z_i = (Z_i^{(1)}, \ldots, Z_i^{(p)})'$, $i = 1, \ldots, n$ are i.i.d. with pdf g. Clearly, the likelihood, under $P^n_{\mu,\Lambda,g}$, is given by $L^n_{\mu,\Lambda,g} = |\det \Lambda|^{-n} \prod_{i=1}^n (\prod_{r=1}^p g_r(e_r' \Lambda^{-1}(X_i - \mu)))$, where e_r is the vector with a one in position r and zeros elsewhere.

In the symmetric IC model above, the location parameter μ is the unique center of symmetry of the common distribution of the X_i's and therefore is a well-defined parameter. In sharp contrast, the parameters Λ and g are not identifiable: letting P be any $p \times p$ permutation matrix and S be any $p \times p$ diagonal matrix with diagonal entries in $\{-1, 1\}$, one can write $X_i = (\Lambda PS)(SP^{-1}Z_i) + \mu =: \tilde{\Lambda}\tilde{Z}_i + \mu$, where $\tilde{Z}_i$ still satisfies (i), (ii) and (iii) above. If $\tilde{g}$ is the density of $\tilde{Z}_i$, then $P^n_{\mu,\Lambda,g} = P^n_{\mu,\tilde{\Lambda},\tilde{g}}$. This indeterminacy can be avoided by requiring, for instance, that marginal densities are given in a specified (e.g., kurtosis) order and that the entry having largest absolute value in each column of Λ is positive.

In the independent component analysis (ICA) one wishes to find an estimate of *any* Λ such that $\Lambda^{-1}X_i$ has independent components. If $\Lambda^{-1}X_i$ has independent components then so has $DSP\Lambda^{-1}X_i$, where D is any diagonal matrix with positive diagonal elements. This same identifiability problem is well recognized in the ICA literature, and it has been proven (see, e.g., [13] for a simple proof) that these three sources of non-identifiability are the only ones, *provided that not more than one IC is Gaussian*, an assumption that is therefore made throughout in the ICA literature. Note that the third source of non-identifiability D is avoided in our model building by fixing the scales of the marginals of Z_i in (iii) above. In the classical ICA the estimation of Λ is the main goal, whereas in our problem it is only a primary device to yield the components used for the testing. The sign-change or permutation of the components will not be a problem in our test construction. We naturally also would like to deal with distributions where there are more than one Gaussian IC. In particular, we do not want to rule out the multinormal case, for which all ICs are Gaussian! Quite fortunately, the resulting lack of identifiability will not affect the behavior of our tests (we discuss this further in Section 5).

3. The proposed tests

Define the (null) residual associated with observation X_i and value Λ of the mixing matrix as $Z_i(\Lambda) := \Lambda^{-1}X_i$. The signed ranks of these residuals are the quantities $S_i(\Lambda)R_i(\Lambda)$, with $S_i(\Lambda) := (S_i^{(1)}(\Lambda), \ldots, S_i^{(p)}(\Lambda))'$ and $R_i(\Lambda) := (R_i^{(1)}(\Lambda), \ldots, R_i^{(p)}(\Lambda))'$, $i = 1, \ldots, n$, where $S_i^{(r)}(\Lambda) := I_{[Z_i^{(r)}(\Lambda)>0]} - I_{[Z_i^{(r)}(\Lambda)<0]}$ is the sign of $Z_i^{(r)}(\Lambda)$ and $R_i^{(r)}(\Lambda)$ is the rank of $|Z_i^{(r)}(\Lambda)|$ among $|Z_1^{(r)}(\Lambda)|, \ldots, |Z_n^{(r)}(\Lambda)|$. Let $K^{(r)} : (0, 1) \to \mathbb{R}$, $r = 1, \ldots, p$ be score functions and consider the corresponding p-variate score function K defined by $u = (u^{(1)}, \ldots, u^{(p)})' \mapsto K(u) := (K^{(1)}(u^{(1)}), \ldots, K^{(p)}(u^{(p)}))'$. We throughout assume that the $K^{(r)}$'s are (i) continuous, (ii) satisfy $\int_0^1 (K^{(r)}(u))^{2+\delta} du < \infty$ for some $\delta > 0$, and (iii) can be expressed as the difference of two monotone increasing functions. These assumptions are required for Hájek's classical projection result for linear signed-rank statistics; see, e.g., [14], Chapter 3 (actually, Hájek's result requires square-integrability rather than the reinforcement of square-integrability in (ii); we will need the latter however to control the unspecification of Λ; see the proof of Lemma 3.3).

The (K-score version of the) test statistic we propose is then

$$Q_K(\Lambda) := (T_K(\Lambda))' \Gamma_K^{-1} T_K(\Lambda),$$

where $T_K(\Lambda) := n^{-1/2} \sum_{i=1}^n T_{K;i}(\Lambda) := n^{-1/2} \sum_{i=1}^n [S_i(\Lambda) \odot K(\frac{R_i(\Lambda)}{n+1})]$ and $\Gamma_K := \mathrm{diag}(\mathrm{E}[(K^{(1)}(U))^2], \ldots, \mathrm{E}[(K^{(p)}(U))^2])$; throughout, $\odot$ denotes the Hadamard (i.e., entrywise) product and U stands for a random variable that is uniformly distributed over $(0, 1)$.

The asymptotic behavior of $Q_K(\Lambda)$ can be investigated quite easily by using the *representation* result in Lemma 3.1 below. In order to state this result, we define $z = (z^{(1)}, \ldots, z^{(p)})' \mapsto G_+(z) := (G_+^{(1)}(z^{(1)}), \ldots, G_+^{(p)}(z^{(p)}))'$, where $G_+^{(r)}$ stands for the cdf of $|Z_1^{(r)}(\Lambda)|$ under $P^n_{0,\Lambda,g}$. Symmetry of g_r yields $G_+^{(r)}(t) = 2G^{(r)}(t) - 1$, where $t \mapsto G^{(r)}(t) = \int_{-\infty}^t g_r(s)ds$ is the cdf of $Z_1^{(r)}(\Lambda)$ under $P^n_{0,\Lambda,g}$.

Lemma 3.1. *Define* $T_{K;g}(\Lambda) := n^{-1/2} \sum_{i=1}^n T_{K;g;i}(\Lambda) := n^{-1/2} \sum_{i=1}^n [S_i(\Lambda) \odot K(G_+(|Z_i(\Lambda)|))]$, *where* $|Z_i(\Lambda)| := (|Z_i^{(1)}(\Lambda)|, \ldots, |Z_i^{(p)}(\Lambda)|)'$. *Then, for any* $\Lambda \in \mathcal{M}_p$ *and* $g \in \mathcal{F}$, $\mathrm{E}[\|T_K(\Lambda) - T_{K;g}(\Lambda)\|^2] = o(1)$ *as* $n \to \infty$, *under* $P^n_{0,\Lambda,g}$.

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

Lemma 3.1 implies that under the null – hence also under sequences of contiguous alternatives (see Section 4.2 for the form of those alternatives) – $T_K(\Lambda)$ is asymptotically equivalent to $T_{K;g}(\Lambda)$, where g is the "true" underlying noise density. Since $T_{K;g}(\Lambda)$ is a sum of i.i.d. terms, the asymptotic null distribution of $T_K(\Lambda)$ then follows from the multivariate CLT.

Lemma 3.2. *For any $\Lambda \in \mathcal{M}_p$, $T_K(\Lambda)$, under $\cup_{g\in\mathcal{F}}\{P^n_{0,\Lambda,g}\}$, is asymptotically multinormal with mean zero and covariance matrix Γ_K.*

It readily follows from Lemma 3.2 that $Q_K(\Lambda)$, under $\cup_{g\in\mathcal{F}}\{P^n_{0,\Lambda,g}\}$, is asymptotically chi-square with p degrees of freedom. The resulting test therefore consists in rejecting the null at asymptotic level α iff $Q_K(\Lambda) > \chi^2_{p,1-\alpha}$.

Of course, as already mentioned, Λ in practice is unspecified and should be replaced with some suitable estimate $\hat{\Lambda}$. The choice of this estimate is discussed in Section 5, but we will throughout assume that $\hat{\Lambda}$ is (i) root-n consistent, (ii) invariant under permutations of the observations, and (iii) invariant under individual reflections of the observations with respect to the origin (i.e., $\hat{\Lambda}(s_1X_1,\ldots,s_nX_n) = \hat{\Lambda}(X_1,\ldots,X_n)$ for all $s_1,\ldots,s_n \in \{-1,1\}$). The replacement of Λ with $\hat{\Lambda}$ in $Q_K(\Lambda)$ yields the genuine test statistic $\hat{Q}_K := Q_K(\hat{\Lambda})$. The following result establishes that this replacement has no effect on the asymptotic null behavior of the test (see the appendix for a proof).

Lemma 3.3. *For any $\Lambda \in \mathcal{M}_p$, $T_K(\hat{\Lambda}) = T_K(\Lambda) + o_P(1)$ (hence also $\hat{Q}_K = Q_K(\Lambda) + o_P(1)$) as $n\to\infty$, under $\cup_{g\in\mathcal{F}}\{P^n_{0,\Lambda,g}\}$.*

The following theorem, which is the main result of this section, is then a direct corollary of Lemmas 3.2 and 3.3.

Theorem 3.1. *Under $\cup_{\Lambda\in\mathcal{M}_p}\cup_{g\in\mathcal{F}}\{P^n_{0,\Lambda,g}\}$, $\hat{Q}_K$ is asymptotically χ^2_p, so that, still under $\cup_{\Lambda\in\mathcal{M}_p}\cup_{g\in\mathcal{F}}\{P^n_{0,\Lambda,g}\}$, the test ϕ_K that rejects the null as soon as $\hat{Q}_K > \chi^2_{p,1-\alpha}$ has asymptotic level α.*

The behavior of our tests under local alternatives will be studied in Section 4.

Let us finish this section with some particular cases of the proposed test statistics $\hat{Q}_K$. To this end, write $\hat{S}_i$ and $\hat{R}_i$ for the empirical signs $S_i(\hat{\Lambda})$ and ranks $R_i(\hat{\Lambda})$, respectively. Then (i) sign test statistics are obtained with constant score functions ($K^{(r)}(u) = 1$ for all r, say). The resulting test statistics are

$$\hat{Q}_S = \hat{T}_S'\hat{T}_S = \frac{1}{n}\sum_{i,j=1}^{n}\hat{S}_i'\hat{S}_j = \frac{1}{n}\sum_{i,j=1}^{n}\sum_{r=1}^{p}\hat{S}_i^{(r)}\hat{S}_j^{(r)}. \tag{1}$$

(ii) The Wilcoxon-type test statistics, associated with linear score functions ($K^{(r)}(u) = u$ for all r, say), take the form

$$\hat{Q}_W = 3\hat{T}_W'\hat{T}_W = \frac{3}{n(n+1)^2}\sum_{i,j=1}^{n}\sum_{r=1}^{p}\hat{S}_i^{(r)}\hat{S}_j^{(r)}\hat{R}_i^{(r)}\hat{R}_j^{(r)}. \tag{2}$$

(iii) Gaussian (or *van der Waerden*) scores are obtained with $K^{(r)}(u) = \Phi_+^{-1}(u) = \Phi^{-1}((u+1)/2)$, where Φ is the cdf of the standard normal distribution. The corresponding test statistics are

$$\hat{Q}_{\mathrm{vdW}} = \hat{T}_{\mathrm{vdW}}'\hat{T}_{\mathrm{vdW}} = \frac{1}{n}\sum_{i,j=1}^{n}\sum_{r=1}^{p}\hat{S}_i^{(r)}\hat{S}_j^{(r)}\Phi_+^{-1}\left(\frac{\hat{R}_i^{(r)}}{n+1}\right)\Phi_+^{-1}\left(\frac{\hat{R}_j^{(r)}}{n+1}\right). \tag{3}$$

As we show in the next section, this van der Waerden test is optimal in the Le Cam sense (more precisely, locally and asymptotically maximin) at the multinormal submodel.

4. Optimality, local powers, and AREs

In this section, we exploit Le Cam's theory of asymptotic experiments in order to define versions of our tests that achieve Le Cam optimality under correctly specified noise densities. We also study the behavior of our tests under sequences of local alternatives and compare their asymptotic performances with those of Hotelling's T^2 test in terms of asymptotic relative efficiencies (AREs).

4.1. Local asymptotic normality and optimal signed-rank tests

The main technical result here is the locally and asymptotically normal (LAN) structure of the IC model with respect to μ, for fixed values of Λ and g. Such LAN property requires more stringent assumptions on g. Define accordingly $\mathcal{F}_{\mathrm{LAN}}$ as

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

the collection of noise densities $g \in \mathcal{F}$ that (i) are absolutely continuous and (ii) have finite Fisher information for location, i.e., $\mathcal{I}_{g_r} := \int_{-\infty}^{\infty}(\varphi_{g_r}(z))^2 g_r(z)\mathrm{d}z < \infty$ for all r, where, denoting by g_r' the a.e.- derivative of g_r, we let $\varphi_{g_r} := -g_r'/g_r$. For $g \in \mathcal{F}_{\mathrm{LAN}}$, define the p-variate optimal location score function φ_g by $z = (z^{(1)}, \ldots, z^{(p)})' \mapsto \varphi_g(z) := (\varphi_{g_1}(z^{(1)}), \ldots, \varphi_{g_p}(z^{(p)}))'$. We then have the following LAN result, which is an immediate corollary of the more general result established in [15].

Proposition 4.1. *For any $\Lambda \in \mathcal{M}_p$ and $g \in \mathcal{F}_{\mathrm{LAN}}$, the family of distributions $\mathcal{P}^n_{\Lambda,g} := \{P^n_{\mu,\Lambda,g}, \mu \in \mathbb{R}^p\}$ is LAN. More precisely, for any p-vector μ and any bounded sequence of p-vectors (τ_n), we have that (letting $S_i(\mu, \Lambda)$ stand for the sign of $Z_i(\mu, \Lambda) := \Lambda^{-1}(X_i - \mu)$) (i) under $P^n_{\mu,\Lambda,g}$, as $n \to \infty$,*

$$\log\left(\mathrm{d}P^n_{\mu+n^{-1/2}\tau_n,\Lambda,g}/\mathrm{d}P^n_{\mu,\Lambda,g}\right) = \tau_n'\Delta^{(n)}_{\mu,\Lambda,g} - \frac{1}{2}\tau_n'\Gamma_{\Lambda,g}\tau_n + o_P(1),$$

with central sequence $\Delta^{(n)}_{\mu,\Lambda,g} := n^{-1/2}(\Lambda^{-1})'\sum_{i=1}^n \varphi_g(Z_i(\mu,\Lambda)) = n^{-1/2}(\Lambda^{-1})'\sum_{i=1}^n [S_i(\nu,\Lambda) \odot \varphi_g(|Z_i(\mu,\Lambda)|)]$ and information matrix $\Gamma_{\Lambda,g} := (\Lambda^{-1})'\mathcal{I}_g\Lambda^{-1} := (\Lambda^{-1})'\mathrm{diag}(\mathcal{I}_{g_1}, \ldots, \mathcal{I}_{g_p})\Lambda^{-1}$, and that (ii) still under $P^n_{\mu,\Lambda,g}$, $\Delta^{(n)}_{\mu,\Lambda,g}$ is asymptotically multinormal with mean zero and covariance matrix $\Gamma_{\Lambda,g}$.

Fix now some noise density $f \in \mathcal{F}_{\mathrm{LAN}}$. Le Cam's theory of asymptotic experiments (see, e.g., Chapter 11 of [16]) implies that an f-optimal (actually, locally and asymptotically maximin at f) test for $\mathcal{H}_0 : \mu = 0$ versus $\mathcal{H}_1 : \mu \neq 0$, under fixed $\Lambda \in \mathcal{M}_k$, consists, at asymptotic level α, in rejecting the null as soon as

$$Q_f(\Lambda) := \left(\Delta^{(n)}_{0,\Lambda,f}\right)' \Gamma^{-1}_{\Lambda,f}\Delta^{(n)}_{0,\Lambda,f} > \chi^2_{p,1-\alpha}.$$

Letting K_f be the p-variate score function defined by $K^{(r)} := \varphi_{f_r} \circ F^{-1}_{+r}$, $r = 1, \ldots, p$ (with the same notation as in Section 3), one straightforwardly checks that $Q_f(\Lambda) = (T_{K_f;f}(\Lambda))'\,\Gamma^{-1}_{K_f}\,T_{K_f;f}(\Lambda)$, which, by Lemmas 3.1 and 3.3 (provided that the score function K_f satisfies the assumptions of Section 3), is asymptotically equivalent to $\hat{Q}_{K_f}$ under $P^n_{0,\Lambda,f}$. Therefore, denoting by $\mathcal{F}^{\mathrm{opt}}_{\mathrm{LAN}}$ the collection of densities $f \in \mathcal{F}_{\mathrm{LAN}}$ for which the K_{f_r}'s (i) are continuous, (ii) satisfy $\int_0^1 (K_{f_r}(u))^{2+\delta}\mathrm{d}u < \infty$ for some $\delta > 0$, and (iii) can be expressed as the difference of two monotone increasing functions, we have proved the following.

Theorem 4.1. *For any $f \in \mathcal{F}^{\mathrm{opt}}_{\mathrm{LAN}}$, the test ϕ_{K_f} that rejects the null as soon as $\hat{Q}_{K_f} > \chi^2_{p,1-\alpha}$ (i) has asymptotic level α under $\cup_{\Lambda\in\mathcal{M}_p}\cup_{g\in\mathcal{F}}\{P^n_{0,\Lambda,g}\}$ and (ii) is locally and asymptotically maximin, at asymptotic level α, for $\cup_{\Lambda\in\mathcal{M}_p}\cup_{g\in\mathcal{F}}\{P^n_{0,\Lambda,g}\}$ against alternatives of the form $\cup_{\mu\neq 0}\cup_{\Lambda\in\mathcal{M}_p}\{P^n_{\mu,\Lambda,f}\}$.*

This justifies the claim (see the end of the previous section) stating that the van der Waerden version of the proposed signed-rank tests is optimal at the multinormal model. More generally, Theorem 4.1 indicates how to achieve Le Cam optimality at a fixed (smooth) noise density f.

4.2. Local powers and asymptotic relative efficiencies

Local powers of our signed-rank tests ϕ_K under local alternatives of the form $P^n_{n^{-1/2}\tau,\Lambda,g}$, $g \in \mathcal{F}_{\mathrm{LAN}}$ can be straightforwardly computed from the following result (the proof is given in the appendix).

Theorem 4.2. *Fix $g \in \mathcal{F}_{\mathrm{LAN}}$ and define $I_{K,g} := \mathrm{diag}(I_{K^{(1)},g_1}, \ldots, I_{K^{(p)},g_p})$, with $I_{K^{(r)},g_r} := \mathrm{E}[K^{(r)}(U)\varphi_{g_r}((G^{(r)}_+)^{-1}(U))]$, where U is uniformly distributed over $(0, 1)$. Then, $\hat{Q}_K$ is asymptotically $\chi^2_p(\tau'(\Lambda^{-1})'I_{K,g}\Gamma_K^{-1}I_{K,g}\Lambda^{-1}\tau)$ under $P^n_{n^{-1/2}\tau,\Lambda,g}$, where $\chi^2_\ell(c)$ stands for the noncentral chi-square distribution with ℓ degrees of freedom and noncentrality parameter c.*

This also allows for computing asymptotic relative efficiencies (AREs) with respect to our benchmark competitor, namely Hotelling's T^2 test. In the following result (see the appendix for a proof), we determine these AREs at any g belonging to the collection $\mathcal{F}^2_{\mathrm{LAN}}$ of noise densities in $\mathcal{F}_{\mathrm{LAN}}$ with finite variances. We want to stress however that our signed-rank tests ϕ_K, unlike Hotelling's test, remain valid without such moment assumption, so that, when the underlying density does not admit a finite variance, the ARE of any ϕ_K with respect to Hotelling's test actually can be considered as being infinite.

Theorem 4.3. *Fix $g \in \mathcal{F}^2_{\mathrm{LAN}}$. Then the asymptotic relative efficiency of ϕ_K with respect to Hotelling's T^2 test, when testing $\mathcal{H}_0 : \mu = 0$ against $\mathcal{H}_1(\tau) : \mu = n^{-1/2}\tau$, under mixing matrix $\Lambda \in \mathcal{M}_p$ and noise density g, is given by*

$$\mathrm{ARE}_{\Lambda,\tau,g}[\phi_K, T^2] = \frac{\tau'(\Lambda^{-1})'I_{K,g}\Gamma_K^{-1}I_{K,g}\Lambda^{-1}\tau}{\tau'(\Lambda^{-1})'\Sigma_g^{-1}\Lambda^{-1}\tau}, \tag{4}$$

where $\Sigma_g := \mathrm{diag}(\sigma^2_{g_1}, \ldots, \sigma^2_{g_p})$, with $\sigma^2_{g_r} := \int_{-\infty}^{z} z^2 g_r(z)\mathrm{d}z$.

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

Table 1
AREs of various univariate signed-rank tests (with sign, Wilcoxon, and van der Waerden scores, as well as scores achieving optimality under t_{12}, t_6, and t_3 densities) with respect to Student's test, under t (with 3, 6, 12 degrees of freedom), Gaussian, and power-exponential densities (with tail parameter $\eta = 2, 3, 5$)

		Underlying density						
		t_3	t_6	t_{12}	$\mathcal{N}$	e_2	e_3	e_5
Score	S	1.621	0.879	0.733	0.637	0.411	0.370	0.347
	W	1.900	1.164	1.033	0.955	0.873	0.881	0.907
	vdW	1.639	1.093	1.020	1.000	1.129	1.286	1.533
	t_{12}	1.816	1.151	1.040	0.981	0.973	1.024	1.102
	t_6	1.926	1.167	1.026	0.936	0.820	0.800	0.779
	t_3	2.000	1.124	0.944	0.820	0.569	0.479	0.385

For $p = 1$, ϕ_K (resp., T^2) boils down to the standard univariate location signed-rank test ϕ_K^{univ} based on the score function K (resp., to the one-sample Student test St), and the ARE in (4) reduces to the well-known result

$$\text{ARE}^{\text{univ}}_{\Lambda,\tau,g}[\phi_K^{\text{univ}}, St] = \frac{\sigma_g^2 I_{K,g}^2}{\text{E}[K^2(U)]}, \tag{5}$$

which does not depend on τ, nor on Λ. For $p \geq 2$, however, the ARE in (4) depends on τ and Λ. Letting $v = (v^{(1)}, \ldots, v^{(p)})' := \frac{\Sigma_g^{-1/2}\Lambda^{-1}\tau}{\|\Sigma_g^{-1/2}\Lambda^{-1}\tau\|}$, we can write

$$\text{ARE}_{\Lambda,\tau,g}[\phi_K, T^2] = \sum_{r=1}^{p}(v^{(r)})^2 \frac{\sigma_r^2 I_{K^{(r)},g_r}^2}{\text{E}[(K^{(r)}(U))^2]} = \sum_{r=1}^{p}(v^{(r)})^2 \text{ARE}^{\text{univ}}_{\Lambda=1,\tau^{(r)},g_r}[\phi_{K^{(r)}}^{\text{univ}}, St], \tag{6}$$

which shows that $\text{ARE}_{\Lambda,\tau,g}[\phi_K, T^2]$ can be seen as *a weighted mean of the corresponding univariate* AREs (those of the univariate signed-rank tests with respect to Student's). The weights depend on the shift τ through the "standardized" shift $\Lambda^{-1}\tau$; if the latter is in the direction of the rth coordinate axis, then $\text{ARE}_{\Lambda,\tau,g}[\phi_K, T^2] = \text{ARE}^{\text{univ}}_{\Lambda=1,\tau^{(r)},g_r}[\phi_{K^{(r)}}^{\text{univ}}, St]$. In all cases, irrespective of τ and Λ, $\text{ARE}_{\Lambda,\tau,g}[\phi_K, T^2]$ always lies between the smallest and the largest "univariate" AREs in $\{\text{ARE}^{\text{univ}}_{\Lambda=1,\tau^{(r)},g_r}[\phi_{K^{(r)}}^{\text{univ}}, T^2], r = 1, \ldots, p\}$.

This explains that it is sufficient to give numerical values for these univariate AREs. Such values are provided in Table 1, for various scores (sign, Wilcoxon, and van der Waerden scores, as well as scores achieving optimality at fixed t distributions) and various underlying densities (t, Gaussian, and power-exponential densities with lighter-than-normal tails). Power-exponential densities refer to densities of the form $g_\eta(r) = c_\eta \exp(-a_\eta r^{2\eta})$, where c_η is a normalization constant, $\eta > 0$ determines the tail weight, and $a_\eta > 0$ standardizes g_η in the same way as the marginal densities in $\mathcal{F}$ (see Section 2).

All numerical values for the van der Waerden signed-rank test ϕ_{vdW} in Table 1 are larger than one, except in the normal case, where it is equal to one. This is an empirical illustration of the [17] result showing that $\text{ARE}^{\text{univ}}_{\Lambda=1,\tau,g}[\phi_{vdW}^{\text{univ}}, St] \geq 1$ for all τ and g (with equality iff g is Gaussian). Hence, (6) entails that, in the IC model under consideration, the AREs of our p-variate van der Waerden test ϕ_{vdW}, with respect to Hotelling's, are always larger than or equal to one, with equality in the multinormal model only.

Coming back to the general expressions of our AREs in (4) and (6), it is clear (in view of (5)) that, in order to maximize the local powers/AREs above with respect to the score function K, one should maximize the cross-information quantities $I_{K^{(r)},g_r}$, $r = 1, \ldots, p$. The Cauchy–Schwarz inequality shows that $I_{K^{(r)},g_r}$ is maximal at $K^{(r)} = \varphi_{g_r} \circ (G_+^{(r)})^{-1}$, which confirms the rule for determining optimal score functions that was derived in Section 4.1.

5. Practical implementation and simulations

In this section, we first focus on the main issue for the practical implementation of our tests, namely the estimation of the mixing matrix Λ. Several approaches are possible, but the approach presented in [18] is chosen here. Then finite-sample efficiencies and robustness properties of our tests are investigated through Monte Carlo studies.

Computations were done using the statistical software package R 2.6.0 [19]. Note that the proposed method for estimating Λ is implemented in the R-package ICS [20], whereas the tests proposed in this paper are implemented in the R-package ICSNP [21]. Both packages are available on the CRAN website.

5.1. Estimation of Λ

An interesting way to obtain a root-n consistent estimate of Λ is to use two different root-n consistent scatter matrix estimates as in [18].

Let X be a p-variate random vector and denote its cdf by F_X. A *scatter matrix functional* S (with respect to the null value of the location center, namely the origin) is a $p \times p$ matrix-valued functional such that $S(F_X)$ is positive definite, symmetric, and

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

affine-equivariant in the sense that $S(F_{AX}) = AS(F_X)A'$, $\forall A \in \mathcal{M}_p$. Examples of scatter matrices are the covariance matrix $S_{\text{cov}}(F_X) := E[XX']$, the scatter matrix based on fourth-order moments $S_{\text{kurt}}(F_X) := E\left[(X'(S_{\text{cov}}(F_X))^{-1}X)XX'\right]$, and [9]'s scatter matrix S_{Tyl} defined implicitly by $S_{\text{Tyl}}(F_X) = E\left[(X'(S_{\text{Tyl}}(F_X))^{-1}X)^{-1}XX'\right]$.

As we now show, the mixing matrix Λ can be estimated by using a couple of different scatter matrices (S_1, S_2). Recall that our tests require a root-n consistent estimate of Λ *under the null*, that is, under $\mathcal{P}_0^n := \{P_{0,\Lambda,g}^n, \Lambda \in \mathcal{M}_p, g \in \mathcal{F}\}$. However, since Λ is not identifiable in $\mathcal{P}_0^n$ (see Section 2), estimation of Λ is an ill-posed problem. We therefore restrict to a submodel by using a couple of scatter matrices (S_1, S_2) as follows.

Define the model $\mathcal{P}_0^n(S_1, S_2)$ as the collection of probability distributions of $(X_1, \ldots, X_n)$ generated by $X_i = \Lambda Z_i$, $i = 1, \ldots, n$, where $Z_i = (Z_i^{(1)}, \ldots, Z_i^{(p)})'$, $i = 1, \ldots, n$ are i.i.d. from a distribution F_Z for which $S_1(F_Z) = I$ and $S_2(F_Z) = \Omega$, where $\Omega = (\Omega_{ij})$ is diagonal with $\Omega_{11} > \Omega_{22} > \cdots > \Omega_{pp} (> 0)$. Theorem 5.5 of [22] and our assumption that Z has independent and symmetric marginals imply that $S_\ell(F_Z)$, $\ell = 1, 2$ are diagonal matrices, so that this submodel actually only imposes that the quantities Ω_{rr}, $r = 1, \ldots, p$ are pairwise different. Before discussing the severity of this restriction, we note that $\mathcal{P}_0^n(S_1, S_2)$ takes care of the permutation (and scale) indetermination by merely assuming that the ICs are first standardized in terms of their "S_1-scales" and then ordered according to their "(S_1, S_2)-kurtoses". As for the signs of the ICs, they can be fixed by requiring, e.g., that the entry having largest absolute value in each column of Λ is positive (and similarly with $\hat{\Lambda}$); see Section 2.

Most importantly, the affine-equivariance of S_1 and S_2 then implies that

$$(S_2(F_X))^{-1}S_1(F_X)(\Lambda^{-1})' = (\Lambda^{-1})'\Omega^{-1} \tag{7}$$

(where X stands for a p-variate random vector with the same distribution as X_i, $i = 1, \ldots, n$), that is, Λ^{-1} and Ω^{-1} list the eigenvectors and eigenvalues of $(S_2(F_X))^{-1}S_1(F_X)$, respectively. Replacing $S_1(F_X)$ and $S_2(F_X)$ with their natural estimates $\hat{S}_1$ and $\hat{S}_2$ in (7) yields estimates $\hat{\Lambda}$ and $\hat{\Omega}$. Clearly, if $\hat{S}_1$ and $\hat{S}_2$ are root-n consistent, then $\hat{\Lambda}$ is root-n consistent as well. Since our tests are based on statistics that are invariant under heterogeneous rescaling and reordering of the ICs, their versions based on such a $\hat{\Lambda}$ will remain valid (i.e., will meet the asymptotic level constraint) independently of the particular signs, scales, and order of the ICs fixed above in $\mathcal{P}_0^n(S_1, S_2)$. Note that their optimality properties, however, require to order the scores K_{f_r}, $r = 1, \ldots, p$ according to the corresponding "(S_1, S_2)-kurtoses".

As we have seen above, the only restriction imposed by $\mathcal{P}_0^n(S_1, S_2)$ is that the "(S_1, S_2)-kurtoses" of the ICs are pairwise different, so that the ordering of the ICs is well defined. Note that this rules out cases for which two (or more) ICs would be identically distributed. More precisely, consider the case for which exactly k ($\geq$2) ICs are equally distributed and the distributions of the remaining $p - k$ ICs are pairwise different. Then the estimator $\hat{\Lambda}$ above allows for recovering the $p - k$ ICs with different distributions, but estimates the remaining k ones up to some random rotation. Note however that if those k ICs are Gaussian, the components of $\hat{\Lambda}^{-1}X$ – conditional on this random rotation – converge in distribution to Z (since – possibly rotated – uncorrelated Gaussian variables with a common scale are independent), so that the asymptotic null distribution of our test statistics is still χ_p^2 (also unconditionally, since this conditional asymptotic distribution does not depend on the value of the random rotation). As a conclusion, while our tests, when based on such $\hat{\Lambda}$, would fail being valid when several ICs share the same distribution, they are valid in the case where the only equally distributed ICs are Gaussian, which includes the important multinormal case.

If however one thinks that ruling out equally distributed non-Gaussian ICs is too much of a restriction, then he/she can still use a root-n consistent estimator of Λ that does not require this assumption. See for example [23] for an overview.

5.2. Finite-sample performances

We conducted a simulation study in the trivariate case ($p = 3$) in order to evaluate the finite-sample performances of our signed-rank tests.

We started by generating i.i.d. centered random vectors $Z_i = (Z_i^{(1)}, Z_i^{(2)}, Z_i^{(3)})'$, $i = 1, \ldots, n$ (we used $n = 50$ and $n = 200$) with marginals that are standardized so that $\text{Med}[(Z_1^{(r)})^2] = 1$, $r = 1, 2, 3$. We considered four settings with the following marginal distributions for $Z_1^{(1)}, Z_1^{(2)}$, and $Z_1^{(3)}$:

Setting I: t_9, Gaussian, and power-exponential with $\eta = 2$ (see Section 4.2) distributions
Setting II: t_3, t_6, and Gaussian distributions
Setting III: t_1, t_6, and Gaussian distributions
Setting IV: three Gaussian distributions (the multinormal case).

Denoting by I_ℓ the ℓ-dimensional identity matrix, samples were then obtained from the IC models $X_i = \Lambda Z_i + \mu$, $i = 1, \ldots, n$, with mixing matrix $\Lambda = I_3$ (this is without loss of generality, since all tests involved in this study are affine-invariant) and location values $\mu = 0$ (null case) and $\mu = n^{-1/2}\tau_\ell e_r$, $\ell = 1, 2, 3, 4$, $r = 1, 2, 3$, (cases in the alternative), where $\tau_1 = 2.147$, $\tau_2 = 3.145$, $\tau_3 = 3.966$, and $\tau_4 = 4.895$ were chosen so that the asymptotic powers of Hotelling's T^2 test, in Setting IV, are equal to .2, .4, .6, and .8, respectively.

First, we studied the sensitivity of our tests with respect to the choice of the estimator $\hat{\Lambda}$ in Setting I. To this end, we considered three estimators in the class of estimators introduced in Section 5.1:

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

(1) The estimator $\hat{\Lambda}_1$ is based on $S_1 = S_{\text{cov}}$ and $S_2 = S_{\text{kurt}}$; root-n consistency of $\hat{\Lambda}_1$ requires finite eighth-order moments.

(2) The estimate $\hat{\Lambda}_2$ is based on $S_1 = S_{\text{Tyl}}$ and $S_2 = S_{\text{Düm}}$, where $S_{\text{Düm}}$ stands for [24]'s scatter matrix (which is the symmetrized version of S_{Tyl}); although $\hat{\Lambda}_2$ is root-n consistent without any moment assumption, it does not fulfill the assumptions of Section 3, since $S_{\text{Düm}}$ (hence also $\hat{\Lambda}_2$) is not invariant under individual sign changes of observations.

(3) Finally, defining $S_{\text{rank}} = E[\Psi_p^{-1}(F_{\|S_{\text{Tyl}}^{-1/2}X\|}(\|S_{\text{Tyl}}^{-1/2}X\|))\frac{XX'}{X'S_{\text{Tyl}}^{-1}X}]$, where Ψ_p denotes the distribution function of a χ_p^2 random variable, the estimate $\hat{\Lambda}_3$, based on $S_1 = S_{\text{Tyl}}$ and $S_2 = S_{\text{rank}}$ fulfills all the assumptions of Section 3 and is root-n consistent without any moment conditions.

For the sake of comparison, we also considered the unrealistic case for which Λ is known. For brevity reasons we refrain from showing the results and only point out that the behavior of our tests does not depend much on the choice of the estimator for Λ. Actually even knowing the true value of Λ did not show to be of any clear advantage. However, it is crucial that the estimator $\hat{\Lambda}$ that is used is root-n consistent, which, in Setting I, is the case of $\hat{\Lambda}_i$, $i = 1, 2, 3$. In Settings I, II and III, the "(S_1, S_2)-kurtoses" from (1), (2) and (3) order the marginal distributions in the same way.

Second, we compared, in Settings I to IV, several versions of our tests with Hotelling's T^2 test. We considered the following signed-rank tests: the sign test based on $\hat{Q}_S$ in (1), the Wilcoxon test based on $\hat{Q}_W$ in (2), and the van der Waerden test based on $\hat{Q}_{\text{vdW}}$ in (3). In each setting, we also included the corresponding optimal signed-rank test (based on $\hat{Q}_{K_f}$ in Section 4.1); we denote by $\hat{Q}_{\text{opt}}^{\text{I}}$, $\hat{Q}_{\text{opt}}^{\text{II}}$, and $\hat{Q}_{\text{opt}}^{\text{III}}$ the statistics of these setting-dependent tests (the optimal test in Setting IV is the van der Waerden test based on $\hat{Q}_{\text{vdW}}$). Of course, these optimal tests use the unspecified underlying density, which is unrealistic, but this is done in order to check how much is gained, in each setting, by using optimal scores. Since the properties of the proposed tests are not very sensitive to the choice of $\hat{\Lambda}$, each signed-rank test was based on the estimator $\hat{\Lambda}_3$ (only the latter satisfies our assumptions on $\hat{\Lambda}$ in all settings). All tests were performed at asymptotic level 5%.

Figs. 1–4 report rejection proportions (based on 5000 replications) and asymptotic powers of the above tests in Settings I to IV, respectively. We should stress that preliminary simulations showed that, under the null in Setting I, the van der Waerden test and the test based on $\hat{Q}_{\text{opt}}^{\text{I}}$, when based on their asymptotic chi-square critical values, are conservative and significantly biased at small sample size $n = 50$. In order to remedy this, we rather used critical values based on the estimation of the (distribution-free) quantile of the test statistic under $\mu = 0$ and under known value $\Lambda = I_3$ of the mixing parameter. These estimations, just as the asymptotic chi-square quantile, are consistent approximations of the corresponding exact quantiles under the null, and were obtained, for the van der Waerden test and the test based on $\hat{Q}_{\text{opt}}^{\text{I}}$, as the empirical 0.05-upper quantiles $q_{.95}$ of the corresponding signed-rank test statistics in a collection of 10 000 simulated (standard) multinormal samples, yielding $q_{.95}^{\text{vdW}} = 7.239$ and $q_{.95}^{\text{opt,I}} = 6.859$, respectively. These bias-corrected critical values both are smaller than the asymptotic chi-square one $\chi_{3,.95}^2 = 7.815$, so that the resulting tests are uniformly less conservative than the original ones. Note that these critical values were always applied when any of those tests were used with $n = 50$ since in practice one does not know the underlying distribution. In all other cases (i.e., for all other tests at $n = 50$, and for all tests at $n = 200$), the asymptotic chi-square critical value $\chi_{3,.95}^2$ was used.

Based on the simulation studies we therefore recommend that for small sample sizes one should calculate the p-value based on simulations or just use a conditionally distribution-free test version. This is not a problem with the current speed of computers, and all three approaches have been implemented in the package ICSNP. Our simulations show that alternative ways to calculate p-values are needed especially when one of the score functions K_{f_r} used is associated with a light-tailed density f_r.

A glance at the rejection proportions under the null in Figs. 1–4 shows that all signed-rank tests appear to satisfy the 5% probability level constraint. In particular, for $n = 50$, the bias-corrected versions of the tests based on $\hat{Q}_{\text{vdW}}$ and on $\hat{Q}_{\text{opt}}^{\text{I}}$ are reasonably unbiased, whereas the asymptotic χ_3^2 approximation seems to work fine in all other cases. Note that Hotelling's T^2 test satisfies the 5% probability level constraint also in Setting III, which was unexpected since one of the marginals (the t_1 distributed one) has infinite second-order moments whereas in all other settings Hotelling's T^2 seems to reject too often.

As for the power properties, the proposed signed-rank tests behave uniformly well in all settings, unlike Hotelling's test, which, for instance, basically never detects the shift in the t_1 component of Setting III (still, it should be noticed that, in the same setting, Hotelling's test works pretty well if the shift is in another component; we will explain this unexpected behavior of Hotelling's test in Section 5.3). In Setting II (see Fig. 2), Hotelling's test competes reasonably well with our tests for small sample sizes, when the shift occurs in a heavy-tailed component. For larger sample sizes, however, our tests outperform Hotelling's and, except for $\hat{Q}_S$, behave essentially as Hotelling's test when the shift occurs in the Gaussian component (this is totally in line with the ARE values in Table 1). Note that when a light-tailed component is present as in Setting I (see Fig. 1), our tests perform as expected. Furthermore the proposed tests also work well in the multinormal model (Fig. 4), although $\hat{\Lambda}_3$ is there only a random rotation; see the comments at the end of Section 5.1. As a conclusion, our optimal tests exhibit very good finite-sample performances in IC models, both in terms of level and power.

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

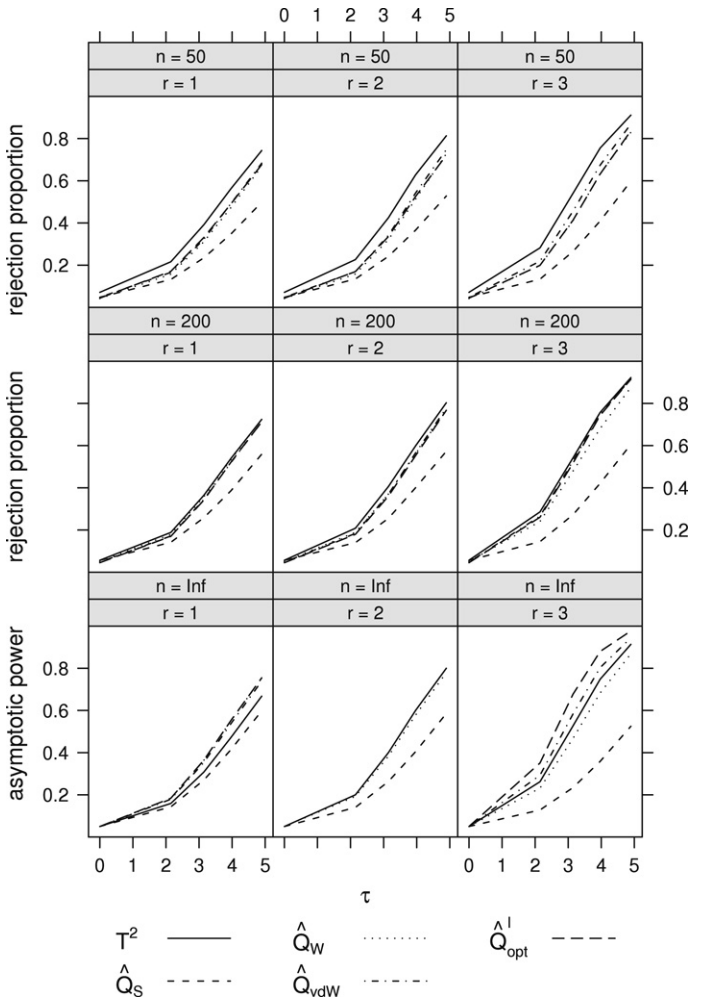

Fig. 1. Rejection proportions (for $n = 50$ and $n = 200$, based on 5000 replications) and asymptotic powers, in Setting I, of Hotelling's T^2 test and of the $\hat{\Lambda}_3$-based versions of the sign, Wilcoxon, van der Waerden, and Setting I optimal signed-rank tests. The integer r indicates in which coordinate the shift occurs.

5.3. Robustness evaluation

In this section, we investigate the robustness properties of the proposed signed-rank tests (in the bivariate case) by studying their power functions under contamination, and by comparing the results with Hotelling's test.

Starting with bivariate i.i.d. random vectors $Z_i = (Z_i^{(1)}, Z_i^{(2)})'$, $i = 1, \ldots, n$ (we used $n = 50$ in this section) with centered t_3 and Gaussian marginals in the first and second components, respectively (still standardized so that $\text{Med}[(Z_1^{(r)})^2] = 1$, $r = 1, 2$), we generated bivariate observations according to $X_i = \Lambda Z_i + \frac{\tau}{\sqrt{n}}(0, 1)'$, $i = 1, \ldots, n$, where $\Lambda = I_2$ and where $\tau = 3.301$ is so that the asymptotic power of Hotelling's test (at asymptotic level $\alpha = .05$) is .5. For any fixed $\delta = (\delta^{(1)}, \delta^{(2)})' \in \mathbb{R}^2$, denote then by $\mathbf{X}(\delta)$ the sample of size n obtained by replacing the first observation X_1 with $X_1 + \delta$.

Clearly, the value of a test statistic computed on $\mathbf{X}(\delta)$ – hence, also the power of the corresponding test – depends on δ. For any test ϕ rejecting $\mathcal{H}_0 : \mu = 0$ at asymptotic level α whenever $Q > \chi^2_{2,1-\alpha}$, we define the *power function* of ϕ as $\delta \mapsto \text{power}(\delta, Q) := P[Q(\mathbf{X}(\delta)) > \chi^2_{2,1-\alpha}]$. Of course, this function can be estimated by generating a large number of independent samples $\mathbf{X}(\delta)$ and by computing rejection frequencies.

We estimated the power functions over $\delta = (\pm 5i, \pm 5j)'$, with $i, j = 0, \ldots, 10$, of Hotelling's T^2 test and of two versions of the van der Waerden signed-rank tests based on (3): the first one (resp., the second one) is based on $\hat{\Lambda}_1$ (resp., on $\hat{\Lambda}_3$), where $\hat{\Lambda}_i$, $i = 1, 3$ are as in Section 5.2. To be in line with what we did there, all van der Waerden tests were based on an estimate (under the null) of the exact (at $n = 50$) distribution-free 95%-quantile of the known-Λ van der Waerden test statistic. In this bivariate case, this estimated quantile, based on 10 000 independent values of this statistic, took the value 5.354 ($<5.991 = \chi^2_{2,.95}$).

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

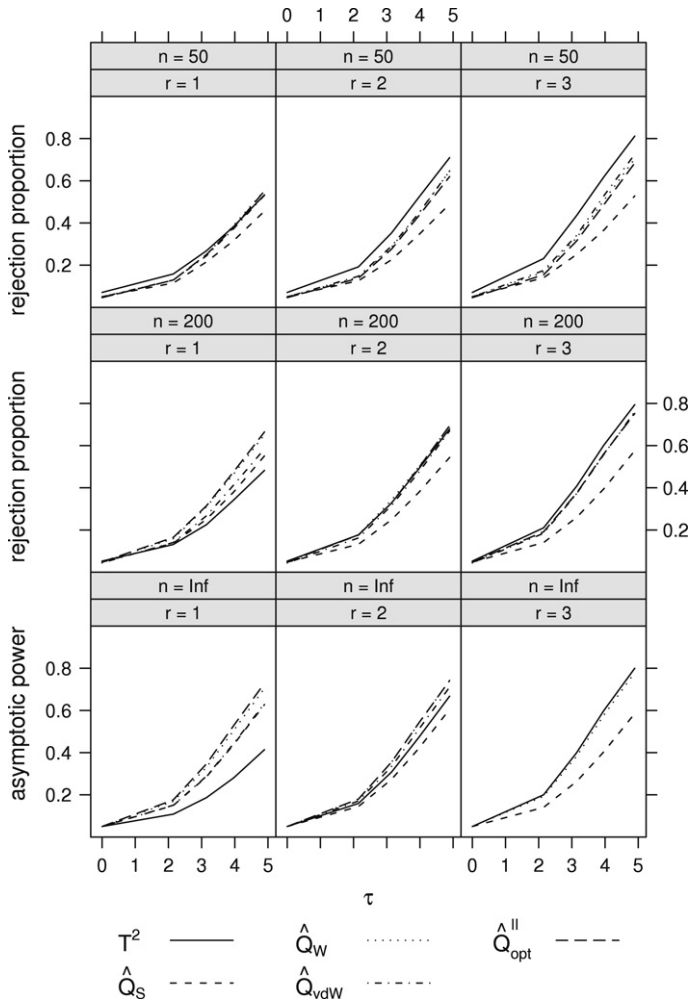

Fig. 2. Rejection proportions (for $n = 50$ and $n = 200$, based on 5000 replications) and asymptotic powers, in Setting II, of Hotelling's T^2 test and of the $\hat{\Lambda}_3$-based versions of the sign, Wilcoxon, van der Waerden, and Setting II optimal signed-rank tests. The integer r indicates in which coordinate the shift occurs.

Fig. 5 presents the estimated power functions (based on 1000 replications) of Hotelling's T^2 test and of the $\hat{\Lambda}_3$-based van der Waerden test. Results for the $\hat{\Lambda}_1$-based version of the latter are not shown since they are very similar to those of the $\hat{\Lambda}_3$-based one (which is actually surprising since one would guess that the lack of robustness of $\hat{\Lambda}_1$ would severely affect the test).

Quite unexpectedly, for $\delta^{(2)} = 0$, the power of Hotelling's test does not suffer under the value of $\delta^{(1)}$. It is even so that compared with the noncontaminated case $\delta = 0$, for which the power functional of Hotelling has the value .516, the functional shows higher power for $|\delta^{(1)}| < 10$ and $0 < \delta^{(2)} \leq 10$. However, if $|\delta^{(2)}|$ is large, the power drops quickly, especially so when there is no or little contamination in $\delta^{(1)}$. The power can then drop even below the size value of .05; e.g., at $\delta = (0, -20)'$, it is only .012.

The puzzling robustness of Hotelling's test with respect to an outlying observation in the first variate can be explained as follows. Let $\mathbf{X} = (X_1 X_2 \cdots X_n)$ be a sample of i.i.d. p-variate observations (whose common distribution admits finite second-order moments) and partition it into

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \end{pmatrix},$$

where the X_{i1}'s are random variables and the X_{i2}'s are $(p-1)$-random vectors. Now, by using (14) in [25], it can be shown that, if one replaces $X_1 = (X_{11}, X_{12}')'$ with $(X_{11}+\delta, X_{12}')'$ and lets $\delta \to \infty$, then, under the assumption (as in the setting above) that the X_{i2}'s are i.i.d. with mean $\tau/\sqrt{n}$ and covariance matrix Σ_{22}, $\lim_{\delta\to\infty} T^2(\mathbf{X}) = T^2(\mathbf{X}_2) + 1 + o_P(1) \xrightarrow{\mathcal{L}} \chi^2_{p-1}(\tau'\Sigma_{22}^{-1}\tau) + 1$, as $n \to \infty$, where $\xrightarrow{\mathcal{L}}$ denotes convergence in law. This is to be compared with the asymptotic $\chi^2_p(\tau'\Sigma_{22}^{-1}\tau)$ distribution of $T^2(\mathbf{X})$ under the assumption that the $X_i = (X_{i1}', X_{i2}')'$'s are i.i.d. with mean $(0, \tau')'$ and with an arbitrary covariance matrix

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

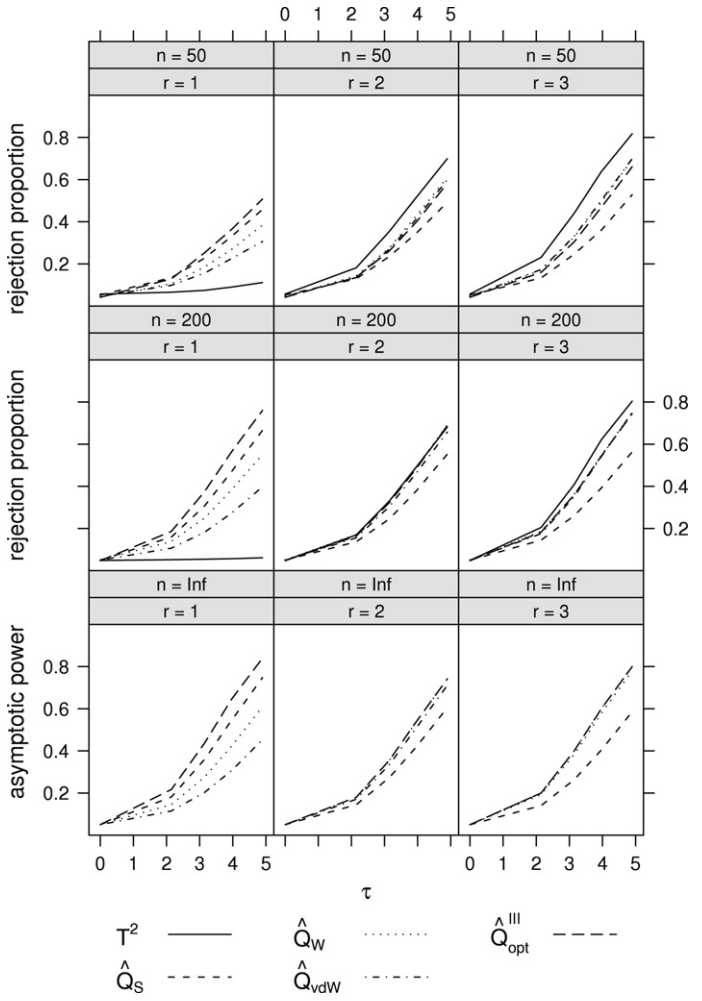

Fig. 3. Rejection proportions (for $n = 50$ and $n = 200$, based on 5000 replications) and asymptotic powers, in Setting III, of Hotelling's T^2 test and of the $\hat{\Lambda}_3$-based versions of the sign, Wilcoxon, van der Waerden, and Setting III optimal signed-rank tests. The integer r indicates in which coordinate the shift occurs.

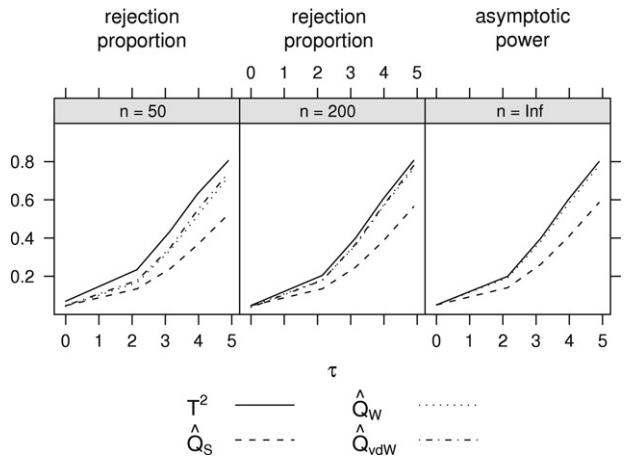

Fig. 4. Rejection proportions (for $n = 50$ and $n = 200$, based on 5000 replications) and asymptotic powers, in Setting IV, of Hotelling's T^2 test and of the $\hat{\Lambda}_3$-based versions of the sign, Wilcoxon, and van der Waerden (which is optimal in Setting IV) signed-rank tests. Without loss of generality (since the underlying distribution is spherically symmetric), the shift occurs in the first coordinate only.

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

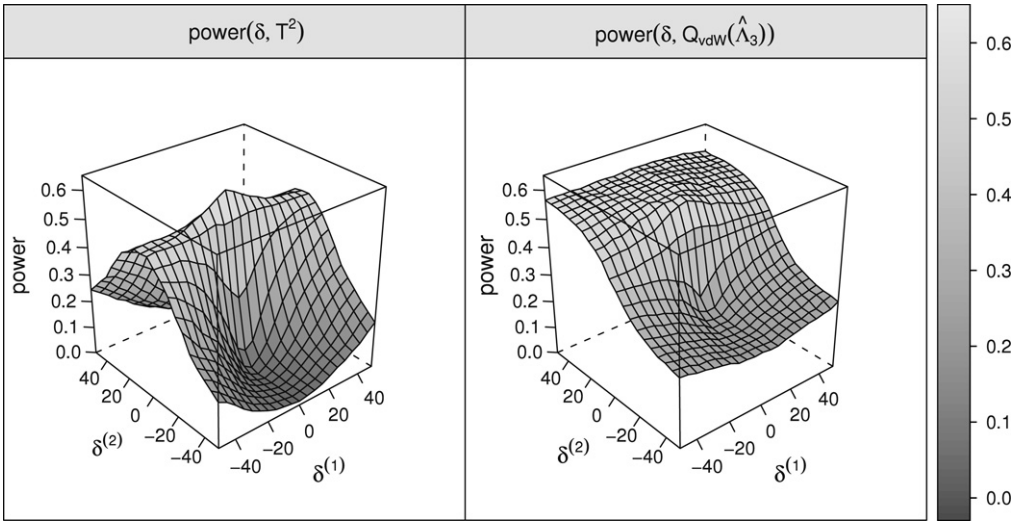

Fig. 5. Estimates of the power functions $power(\delta, T^2)$, $power(\delta, Q_{vdW}(\Lambda))$, $power(\delta, Q_{vdW}(\hat{\Lambda}_1))$, and $power(\delta, Q_{vdW}(\hat{\Lambda}_3))$. The sample size is $n = 50$ and the estimation is based on 1000 replications.

such that $\mathrm{Var}[X_{i2}] = \Sigma_{22}$. For small dimensions p, obtaining (by contaminating a single observation) a $\chi^2_{p-1}(\tau'\Sigma_{22}^{-1}\tau) + 1$ distribution rather than the expected $\chi^2_p(\tau'\Sigma_{22}^{-1}\tau)$ one can bias the results considerably.

Hence, one can say that an outlier in one variate (i) destroys all information about that variate and (ii) biases the result for the "remaining data". This also explains the unexpected behavior of Hotelling's test in Setting III of Section 5.2: the t_1-distributed variate can be seen as a completely contaminated variate which therefore basically contains no information; still, Hotelling's test can detect shifts in the remaining variates.

Fig. 5 shows that on the other hand the test based on $Q_{vdW}(\hat{\Lambda}_3)$ proves much more robust than Hotelling's and is hardly affected by the value of δ_1. Note that if the contamination δ_2 is negative (resp., positive), the power of this test slightly goes down (resp., up) as δ_1 goes through the Z_{i1} data cloud. This slight decrease (resp., increase) of the power function can be explained by the fact that, for any negative (resp., positive) value of δ_2, the contaminated observation – with the scale used in our setting – immediately gets the smallest (resp., largest) rank assigned. The range of the $Q_{vdW}(\hat{\Lambda}_3)$-power function in Fig. 5 goes from .193 to .582, which is comparable with those associated with $Q_{vdW}(\Lambda)$ (from .263 to .576) and with $Q_{vdW}(\hat{\Lambda}_1)$ (from .237 to .580).

Acknowledgments

The research work of Klaus Nordhausen and Hannu Oja was partially supported by research grants from the Academy of Finland and by the Tampere Graduate School in Information Science and Engineering. The research work of Davy Paindaveine was supported by a Mandat d'Impulsion Scientifique of the Fonds National de la Recherche Scientifique, Communauté française de Belgique. The authors would like to thank the editor, the associate editor, and two anonymous referees for their valuable comments that helped them to improve significantly the presentation and the quality of the paper.

Appendix A. Proofs of Lemmas 3.1–3.3

In this section, we will write, $T_K^{(r)}(\Lambda)$ (resp., $T_{K;g}^{(r)}(\Lambda)$) for the rth component of $T_K(\Lambda)$ (resp., of $T_{K;g}(\Lambda)$), $r = 1, \ldots, p$.

Proof of Lemma 3.1. Fix $\Lambda \in \mathcal{M}_p$, $g \in \mathcal{F}$, and $r \in \{1, \ldots, p\}$. Then, under $P^n_{0,\Lambda,g}$, the vector of signs $(S_1^{(r)}(\Lambda), \ldots, S_n^{(r)}(\Lambda))$ collects i.i.d. random variables with $P^n_{0,\Lambda,g}[S_i^{(r)}(\Lambda) = 1] = P^n_{0,\Lambda,g}[S_i^{(1)}(\Lambda) = -1] = 1/2$, (ii) the vector of ranks $(R_1^{(r)}(\Lambda), \ldots, R_n^{(r)}(\Lambda))$ is uniformly distributed over the set of all permutations of $\{1, 2, \ldots, n\}$, and (iii) the vector of signs is independent of the vector of ranks. Consequently, Hájek's classical projection result for signed-rank linear statistics (see, e.g.,[14], Chapter 3) yields that $\mathrm{E}[(T_K^{(r)}(\Lambda) - T_{K;g}^{(r)}(\Lambda))^2] = \mathrm{E}[(n^{-1/2}\sum_{i=1}^n S_i^{(r)}(\Lambda)[K^{(r)}(\frac{R_i^{(r)}(\Lambda)}{n+1}) - K^{(r)}(G_+^{(r)}(|Z_i^{(r)}(\Lambda)|))])^2]$ is $o(1)$ under $P^n_{0,\Lambda,g}$, as $n \to \infty$, which establishes the result. □

Note that this also shows that $\mathrm{E}[(K^{(r)}(R_1^{(r)}(\Lambda)/(n+1)) - K^{(r)}(G_+^{(r)}(|Z_1^{(r)}(\Lambda)|)))^2] = \mathrm{E}[(n^{-1/2}\sum_{i=1}^n S_i^{(r)}(\Lambda)[K^{(r)}(R_i^{(r)}(\Lambda)/(n+1)) - K^{(r)}(G_+^{(r)}(|Z_i^{(r)}(\Lambda)|))])^2]$ is $o(1)$ as $n \to \infty$, under $P^n_{0,\Lambda,g}$.

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004

Proof of Lemma 3.2. Fix $\Lambda \in \mathcal{M}_p$ and $g \in \mathcal{F}$. For any $r = 1, \ldots, p$, the CLT shows that $T^{(r)}_{K;g}(\Lambda)$ is, under $P^n_{0,\Lambda,g}$, asymptotically normal with mean zero and variance $\mathrm{E}[(K^{(r)}(U))^2]$. Therefore, the mutual independence (still under $P^n_{0,\Lambda,g}$) of $T^{(r)}_{K;g}(\Lambda)$, $r = 1, \ldots, p$ entails that $T_{K;g}(\Lambda)$ is asymptotically multinormal with mean zero and covariance matrix Γ_K. The result then follows from Lemma 3.1. □

It remains to prove Lemma 3.3. We do so by showing that, for any $\Lambda \in \mathcal{M}_p, g \in \mathcal{F}$, and $r \in \{1, \ldots, p\}$,

$$E[(T^{(r)}_K(\hat{\Lambda}) - T^{(r)}_K(\Lambda))^2] = o(1) \tag{A.1}$$

as $n \to \infty$, under $P^n_{0,\Lambda,g}$. In the rest of this section, we therefore fix such Λ, g, and r. All expectations and stochastic convergences will then be under $P^n_{0,\Lambda,g}$, and we will write $Z^{(r)}_i$, $S^{(r)}_i$, and $R^{(r)}_i$ for $Z^{(r)}_i(\Lambda)$, $S^{(r)}_i(\Lambda)$, and $R^{(r)}_i(\Lambda)$, respectively. Finally, we will denote the empirical counterparts of these quantities (based on $\hat{\Lambda}$) by $\hat{Z}^{(r)}_i$, $\hat{S}^{(r)}_i$, and $\hat{R}^{(r)}_i$.

We will need the following preliminary result.

Lemma A.1. *As $n \to \infty$,* (i) $\hat{Z}^{(r)}_1 - Z^{(r)}_1 = o_P(1)$,(ii) $\mathrm{E}[(K^{(r)}(\hat{R}^{(r)}_1/(n+1)) - K^{(r)}(G^{(r)}_+(|Z^{(r)}_1(\Lambda)|)))^2] = o(1)$ *and* (iii) $\mathrm{E}[|\hat{S}^{(r)}_1 - S^{(r)}_1|^a] = o(1)$ *for any $a > 0$.*

Proof of Lemma A.1. (i) Denoting by $\|A\|_{\mathcal{L}}$ the sup norm of the array A, we have $|\hat{Z}^{(r)}_1 - Z^{(r)}_1| \le \|\hat{Z}_1 - Z_1\| \le \|\hat{\Lambda}^{-1} - \Lambda^{-1}\|_{\mathcal{L}}\|X_1\|$. The claim therefore follows from the root-n consistency of $\hat{\Lambda}$.

(ii) Applying Lemma 2 of [26], with $\alpha = (\mathrm{vec}\,\Lambda)$ and $g(X, \alpha) = |e'_r[\Lambda^{-1}X]|$ yields that $(\hat{R}^{(r)}_1/(n+1)) - G^{(r)}_+(|Z^{(r)}_1|)$ is $o(1)$ as $n \to \infty$ (note that Conditions (a) and (b) of that lemma are fulfilled: (a) is our root-n consistency assumption on $\hat{\Lambda}$, whereas (b) can be checked exactly along the same lines as in [26], once it is noticed that $\|e'_r[(\Lambda + n^{-1/2}L)^{-1}X]| - |e'_r[\Lambda^{-1}X]\| \le \|[(\Lambda + n^{-1/2}L)^{-1} - \Lambda^{-1}]X\|$, for any fixed $p \times p$ matrix L).

Now, the continuity of $K^{(r)}$ entails that

$$K^{(r)}\left(\frac{\hat{R}^{(r)}_1}{n+1}\right) - K^{(r)}(G^{(r)}_+(|Z^{(r)}_1|)) \tag{A.2}$$

is $o_P(1)$ as $n \to \infty$. To prove that this convergence also holds in quadratic mean (which is precisely Part (ii) of the lemma), it is sufficient to show that (A.2) is uniformly integrable. The second term in (A.2) is of course uniformly integrable since the integrable random variable $K_r(G^{(r)}_+(|Z^{(1r)}_1|))$ does not depend on n. As for the first term in (A.2), recall that $K^{(r)}(R^{(r)}_1/(n+1)) - K^{(r)}(G^{(r)}_+(|Z^{(r)}_1|)) = o_{L^2}(1)$ as $n \to \infty$ (see the remark after the proof of Lemma 3.1), which implies that $K^{(r)}(\frac{R^{(r)}_1}{n+1})$ is uniformly integrable. Finally, the latter uniform integrability and the invariance of $\hat{\Lambda}$ under permutations of the observations in turn imply that $K^{(r)}(\frac{\hat{R}^{(r)}_1}{n+1})$ is uniformly integrable. We conclude that (A.2) is indeed uniformly integrable, and the result follows.

(iii) Since $\hat{S}^{(r)}_1 - S^{(r)}_1 = (|\hat{Z}^{(r)}_1|^{-1} - |Z^{(r)}_1|^{-1})\hat{Z}^{(r)}_1 + |Z^{(r)}_1|^{-1}(\hat{Z}^{(r)}_1 - Z^{(r)}_1)$, we have $|\hat{S}^{(r)}_1 - S^{(r)}_1| \le 2|\hat{Z}^{(r)}_1 - Z^{(r)}_1|/|Z^{(r)}_1| =: Y^{(r)}_1$. Now, fix some $\delta > 0$. Then, for all $\eta > 0$, $P[Y^{(r)}_1 > \delta] \le P[Y^{(r)}_1 I_{[|Z^{(r)}_1|<\eta]} > \delta/2] + P[Y^{(r)}_1 I_{[|Z^{(r)}_1|\ge\eta]} > \delta/2] \le P[|Z^{(r)}_1| < \eta] + P[Y^{(r)}_1 I_{[|Z^{(r)}_1|\ge\eta]} > \delta/2] =: p^{(n)}_1 + p^{(n)}_2$, say. For all $\varepsilon > 0$, there exists $\eta = \eta(\varepsilon)$ such that $p^{(n)}_1 < \varepsilon/2$. As for $p^{(n)}_2$, note that $Y^{(r)}_1 I_{[|Z^{(r)}_1|\ge\eta]} \le (2/\eta)|\hat{Z}^{(r)}_1 - Z^{(r)}_1|$, so that Part (i) of the lemma entails that $p^{(n)}_2 < \varepsilon/2$ for large n. We conclude that $|\hat{S}^{(r)}_1 - S^{(r)}_1| \le Y^{(r)}_1$ converges to zero in probability, which establishes the result (since $|\hat{S}^{(r)}_1 - S^{(r)}_1|$ is bounded). □

Proof of Lemma 3.3. We have to prove (A.1). Since the proof of Lemma 3.1 establishes that $E[(T^{(r)}_K(\Lambda) - T^{(r)}_{K;g}(\Lambda))^2] = o(1)$ as $n \to \infty$, it is sufficient to show that $E[(T^{(r)}_K(\hat{\Lambda}) - T^{(r)}_{K;g}(\Lambda))^2] = o(1)$ as $n \to \infty$. To do so, write $T^{(r)}_K(\hat{\Lambda}) - T^{(r)}_{K;g}(\Lambda) = H_1 + H_2$, with $H_1 := n^{-1/2}\sum_{i=1}^n \hat{S}^{(r)}_i(K^{(r)}(\hat{R}^{(r)}_i/(n+1)) - K^{(r)}(G^{(r)}_+(|Z^{(r)}_i|)))$ and $H_2 := n^{-1/2}\sum_{i=1}^n(\hat{S}^{(r)}_i - S^{(r)}_i)K^{(r)}(G^{(r)}_+(|Z^{(r)}_i|))$. Then, by using the invariance of $\hat{\Lambda}$ under individual reflections of the observations about the origin, we obtain

$$\begin{aligned}\mathrm{E}[(H_1)^2] &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}[(\hat{S}^{(r)}_i)^2(K^{(r)}(\hat{R}^{(r)}_i/(n+1)) - K^{(r)}(G^{(r)}_+(|Z^{(r)}_i|)))^2] \\ &= \mathrm{E}[(K^{(r)}(\hat{R}^{(r)}_1/(n+1)) - K^{(r)}(G^{(r)}_+(|Z^{(r)}_1|)))^2]\end{aligned}$$

and, by using Holder's inequality,

$$\begin{aligned}\mathrm{E}[(H_2)^2] &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}[(\hat{S}^{(r)}_i - S^{(r)}_i)^2(K^{(1)})^2(G^{(r)}_+(|Z^{(r)}_i|))] \\ &= \mathrm{E}[(\hat{S}^{(r)}_1 - S^{(r)}_1)^2(K^{(r)}(G^{(r)}_+(|Z^{(r)}_1|)))^2] \le (\mathrm{E}[(\hat{S}^{(r)}_1 - S^{(r)}_1)^{\frac{2d_\delta}{\delta}}])^{\frac{\delta}{d_\delta}}(\mathrm{E}[(K^{(r)}(U))^{d_\delta}])^{\frac{2}{d_\delta}},\end{aligned}$$

where $d_\delta := 2+\delta$, U is uniformly distributed over $(0,1)$, and $\delta > 0$ is the real number involved in our assumptions on $K^{(r)}$ (see the beginning of Section 3). By applying Lemma A.1(ii)–(iii), we then conclude that $E[(T_K^{(r)}(\hat{\Lambda}) - T_{K;g}^{(r)}(\Lambda))^2] \le 2(E[(H_1)^2] + E[(H_2)^2])$ is $o(1)$ as $n \to \infty$. □

Appendix B. Proofs of Theorems 4.2 and 4.3

Proof of Theorem 4.2. Fix $\Lambda \in \mathcal{M}_p$ and $g \in \mathcal{F}_{\mathrm{LAN}}$. Applying successively Lemmas 3.1 and 3.3 yields that, as $n \to \infty$, under $P^n_{0,\Lambda,g}$,

$$\hat{Q}_K = (T_{K;g}(\Lambda))' \Gamma_K^{-1} T_{K;g}(\Lambda) + o_P(1). \tag{B.1}$$

Recall that $T_{K;g}(\Lambda)$, under $P^n_{0,\Lambda,g}$, is asymptotically multinormal with mean zero and covariance matrix Γ_K; see the proof of Lemma 3.2. Now, it is easy to see that, under $P^n_{0,\Lambda,g}$, $T_{K;g}(\Lambda)$ and the local log-likelihood $\log(\mathrm{d}P^n_{n^{-1/2}\tau,\Lambda,g}/\mathrm{d}P^n_{0,\Lambda,g})$ asymptotically are jointly multinormal with covariance $I_{K,g}\Lambda^{-1}\tau$. Le Cam's third Lemma thus yields that $T_{K;g}(\Lambda)$, under $P^n_{n^{-1/2}\tau,\Lambda,g}$, is asymptotically multinormal with mean $I_{K,g}\Lambda^{-1}\tau$ and covariance matrix Γ_K. The result then follows from the fact that contiguity implying (B.1) holds also under $P^n_{n^{-1/2}\tau,\Lambda,g}$. □

Proof of Theorem 4.3. Fix $\Lambda \in \mathcal{M}_p$ and $g \in \mathcal{F}^2_{\mathrm{LAN}}$. In this proof, all expectations, variances, and covariances are under $P^n_{0,\Lambda,g}$.

Since $\mathrm{Var}[X_1] = \Lambda\Sigma_g\Lambda'$ (where Σ_g is defined in the statement of the theorem), we have that $S := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' = \Lambda\Sigma_g\Lambda' + o_P(1)$ as $n \to \infty$, under $P^n_{0,\Lambda,g}$. Consequently, letting $Z_i := Z_i(\Lambda) = \Lambda^{-1}X_i$ and $\bar{Z} := \frac{1}{n}\sum_{i=1}^n Z_i$, Hotelling's test statistic T^2 satisfies $T^2 = n\bar{X}'S^{-1}\bar{X} + o_P(1) = (\sqrt{n}\bar{Z})'\Sigma_g^{-1}(\sqrt{n}\bar{Z}) + o_P(1)$ as $n \to \infty$, under $P^n_{0,\Lambda,g}$, hence also under $P^n_{n^{-1/2}\tau,\Lambda,g}$ (from contiguity). Clearly, $\sqrt{n}\bar{Z}$ is asymptotically multinormal with mean zero and covariance matrix Σ_g under $P^n_{0,\Lambda,g}$. Proceeding as in the previous proof, one then shows that $\sqrt{n}\bar{Z}$ and the local log-likelihood $\log(\mathrm{d}P^n_{n^{-1/2}\tau,\Lambda,g}/\mathrm{d}P^n_{0,\Lambda,g})$ asymptotically are jointly multinormal under $P^n_{0,\Lambda,g}$, with asymptotic covariance $\Lambda^{-1}\tau$. Le Cam's third Lemma thus implies that $\sqrt{n}\bar{Z}$, under $P^n_{n^{-1/2}\tau,\Lambda,g}$, is asymptotically multinormal with mean $\Lambda^{-1}\tau$ and covariance matrix Σ_g. Therefore, T^2 is asymptotically $\chi^2_p(\tau'(\Lambda^{-1})'\Sigma_g^{-1}\Lambda^{-1}\tau)$ under $P^n_{n^{-1/2}\tau,\Lambda,g}$.

This establishes the result since the AREs of ϕ_K with respect to Hotelling's T^2 test are obtained by computing the ratios of the noncentrality parameters in their respective asymptotic distributions under local alternatives. □

References

[1] M. Hallin, D. Paindaveine, Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks, Annals of Statistics 30 (2002a) 1103–1133.
[2] M. Hallin, D. Paindaveine, Multivariate signed ranks: Randles' interdirections or Tyler's angles? in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L1-norm and Related Methods, Birkhauser, Basel, 2002b, pp. 271–282.
[3] R.H. Randles, A distribution-free multivariate sign test based on interdirections, Journal of the American Statistical Association 84 (1989) 1045–1050.
[4] H. Oja, D. Paindaveine, Optimal signed-rank tests based on hyperplanes, Journal of Statistical Planning and Inference 135 (2005) 300–323.
[5] M.L. Puri, P.K. Sen, Nonparametric Methods in Multivariate Analysis, Wiley & Sons, New York, 1971.
[6] B. Chakraborty, P. Chaudhuri, On affine invariant sign and rank tests in one sample and two sample multivariate problems, in: S. Ghosh (Ed.), Multivariate, Design and Sample Survey, Marcel-Dekker, New York, 1999, pp. 499–414.
[7] K. Nordhausen, H. Oja, D.E. Tyler, On the efficiency of invariant multivariate sign and rank test, in: E.P. Liski, J. Isotalo, J. Niemelä, S. Puntanen, G.P.H. Styan, Festschrift for Tarmo Pukkila on his 60th birthday, University of Tampere, Tampere, 2006, pp. 217–231.
[8] J. Möttönen, H. Oja, Multivariate spatial sign and rank methods, Journal of Nonparametric Statistics 5 (1995) 201–213.
[9] D.E. Tyler, A distribution-free M-estimator of multivariate scatter, Annals of Statistics 15 (1987) 234–251.
[10] R.H. Randles, A simpler, affine-invariant, multivariate, distribution-free sign test, Journal of the American Statistical Association 95 (2000) 1263–1268.
[11] T.P.J. Hettmansperger, J. Nyblom, H. Oja, Affine invariant multivariate one-sample sign test, Journal of the Royal Statistical Society, B 56 (1994) 221–234.
[12] T.P.J. Hettmansperger, J. Möttönen, H. Oja, Affine invariant multivariate one-sample signed-rank test, Journal of the American Statistical Association 92 (1997) 1591–1600.
[13] F.J. Theis, A new concept for separability problems in blind source separation, Neural Computation 16 (2004) 1827–1850.
[14] M.L. Puri, P.K. Sen, Nonparametric Methods in General Linear Models, J. Wiley, New York, 1985.
[15] H. Oja, D. Paindaveine, S. Taskinen, Parametric and nonparametric tests for multivariate independence in IC models (in preparation).
[16] L.M. Le Cam, Asymptotic Methods in Statistical Decision Theory, Springer-Verlag, New York, 1986.
[17] H. Chernoff, I.R. Savage, Asymptotic normality and efficiency of certain nonparametric tests, Annals of Institute of Mathematical Statistics 29 (1958) 972–994.
[18] H. Oja, S. Sirkiä, J. Eriksson, Scatter matrices and independent component analysis, Austrian Journal of Statistics 35 (2006) 175–189.
[19] R Development Core Team, R: A language and environment for statistical computing, ISBN 3-900051-07-0, Vienna, 2007. http://www.R-project.org.
[20] K. Nordhausen, H. Oja, D.E. Tyler, ICS: Tools for Exploring Multivariate Data via ICS/ICA, 2007, R package version 1.1-0.
[21] K. Nordhausen, S. Sirkiä, H. Oja, D.E. Tyler, ICSNP: Tools for Multivariate Nonparametrics, 2007, R package version 1.0-1.
[22] D.E. Tyler, F. Critchley, L. Dümbgen, H. Oja, Invariant coordinate selection, 2008. Conditionally accepted.
[23] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley & Sons, New York, 2001.
[24] L. Dümbgen, On Tyler's M-functional of scatter in high dimensions, Annals of Institute of Statistical Mathematics 50 (1998) 1269–1292.
[25] A.C. Rencher, The contribution of individual variables to Hotelling's T^2, Wilks' Λ, and R^2, Biometrics 49 (1993) 479–489.
[26] D. Peters, R.H. Randles, A multivariate signed-rank test for the one-sample location problem, Journal of the American Statistical Association 85 (1990) 552–557.

Please cite this article in press as: K. Nordhausen, et al., Signed-rank tests for location in the symmetric independent component model, Journal of Multivariate Analysis (2008), doi:10.1016/j.jmva.2008.08.004