PAULIINA ILMONEN

Invariant Coordinate Selection and New Approaches for Independent Component Analysis

ACADEMIC DISSERTATION
To be presented, with the permission of
the board of the School of Health Sciences
of the University of Tampere,
for public discussion in the Auditorium of
School of Health Sciences, Medisiinarinkatu 3,
Tampere, on October 14th, 2011, at 12 o'clock.

UNIVERSITY OF TAMPERE

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere, School of Health Sciences
The Finnish Doctoral Programme in Stochastics and Statistics (FDPSS)
Finland

Supervised by
Professor Hannu Oja
University of Tampere
Finland
Professor Jaakko Nevalainen
University of Turku
Finland

Reviewed by
Professor Christophe Croux
Catholic University of Leuven
Belgium
Professor Ana Pires
Technical University of Lisbon
Portugal

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

Tel. +358 40 190 9800
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Mikko Reinikka

Acta Universitatis Tamperensis 1655
ISBN 978-951-44-8564-0 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1116
ISBN 978-951-44-8565-7 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2011

Acknowledgements

Writing my thesis has been a great journey. During the journey I have met many inspiring people, and the work itself has been exciting.

First of all, I wish to thank my supervisor, Academy Professor Hannu Oja. Hannu truly is an exceptional person. His intelligence, creativity, and enthusiasm is a never-ending source of inspiration. Hannu has not only provided excellent supervision - he has also allowed me to grow and to take some tiny steps on my own. Professionally, Hannu has given me strong roots, and he has given me wings as well. He has been very patient with me, and he has guided me gently but with certainty. Hannu is very open-minded in his work, and he is also very open-minded in general. I am certain that being my supervisor has been frustrating and tiring sometimes. However, Hannu has always accepted me as I am, and I am very grateful for that. Hannu's unique sense of humor is charming, and it has been fun to belong to his research group. One of the most important things that Hannu has taught me is that research has to be fun - the most important thing is to enjoy working. Hannu is my idol. He is my hero. It has been a great pleasure, and a great honor, to be his student. The journey will continue. We have plenty of ideas for new papers.

I also wish to thank my other supervisor, Professor Jaakko Nevalainen. Jaakko too is a role model for me. Jaakko is very straight and honest, when it is needed, and I greatly appreciate that. Professionally, I have learned a lot from Jaakko. He has also been an excellent emotional support.

I have been very fortunate to have two great unofficial supervisor, Professor Robert Serfling and Professor Davy Paindaveine. One of the first things, when I started my PhD studies, was to read the book *Approximation Theorems of Mathematical Statistics*. That is when I started wishing that I would someday meet the author of the book, Professor Robert Serfling. I just wanted to say hi to him, and tell him that I was impressed by his book. Now I have written a paper with Bob! It is difficult for me to try to find big enough words to express how much I respect Bob. Meeting Bob, and working with him, has been one of the greatest things in my professional life. Bob is also one the warmest and friendliest people I have ever met. I am waiting for my next trip to Dallas with great enthusiasm! I also wish to warmly thank my other unofficial supervisor, Professor Davy Paindaveine. When I met Davy, I was immediately impressed by his brightness and intelligence, and by his charming personality. Davy has no wires. Everything is clear to him right away. He is, however, still able to understand us that do have wires. Davy has been very patient with me, even when I have asked

the same questions one hundred times. It has been a great joy to work with him. As I have told Davy, my plan is to continue working with him until I am 99 years old. Then I will retire. Part of my thesis work was done in Brussels with Davy. During that half an year in Brussels I learned plenty of new things under Davy's guidance (including eating sushi). Davy is also a very precious friend of mine. He is adorable. I am very much looking forward to my upcoming post doc time in Brussels!

I wish to thank Dr. Klaus Nordhausen for being an excellent co-author. Klaus has also helped me with my constant problems with computers and programming. Computers do hate me, and it has been an honor to be guided by a true computer and programming genius! Klaus has also been a dear friend and a great emotional support for me, and I am very thankful for that too. I also wish to express my sincere gratitude to my other co-authors, Dr. Abhijit Mandal and Dr. Esa Ollila. It has been a great pleasure to work with them, and I have learned a lot from them. I would also like to thank Professor Anneli Ignatius for providing an interesting data set for the real data example section of my thesis.

I would like to thank my referees, Professor Ana Pires and Professor Christophe Croux, for their careful reading and very insightful comments.

The work was carried out when I was a researcher at the School of Health Sciences, University of Tampere, and it was financially supported by the Academy of Finland and by the Finnish Doctoral Programme in Stochastics and Statistics (FDPSS).

I would like to thank all the wonderful people studying and working at the School of Health Sciences. My special thanks go to the friends and colleagues from the Nonparametric and Robust Multivariate Methods research group, the Friday Seminar group, and the very unofficial Coffee Room group. I would like to thank the people from the Department of Mathematics and Statistics as well. That is the place where I learned to love science. I especially wish to thank Docent Pentti Haukkanen for guiding me when I took my first steps as researcher. Pentti taught me a lot about science, about life, and about friendship. With Pentti I wrote my very first scientific papers, but before anything else, Pentti is a friend. He has given me a lot of emotional support when I was writing my thesis, and I thank him for that too.

I have been blessed with many wonderful friends, and I wish to thank all my friends in Finland, and all my friends around the world, for their support and encouragement. I certainly have not had enough time for you all. I try to do better from now on...

When I was just a little girl I admired my mother Marja-Leena. She was something. I still admire my mother, because - well - she really is something. She is the most creative and the most intelligent person I have ever met. My mother always believed in me and encouraged me, also in my darkest moments, and I want to thank her for that. My mother is very strong. I wish I would be as strong as she is. Even when my mother is not literally by my side, she is still always there. My husband often says warmly that

even though we see Marja-Leena very rarely, she is constantly present in our everyday life. My husband knows that if I justify something by saying that 'but my mother always did it that way' or by 'but my mother has said that it is so', then it is completely useless to argue. My mother is the highest authority in my life. When I was a child I thought that my father Markku is the coolest father ever. That has not changed. My father always was, and he still is, full of new ideas. My father is a dreamer, but he is the type of dreamer that fulfills his dreams too. My father is very brave and I have always tried to be as brave as he is. My father has taught me to dream, and he has helped me when I have struggled. I thank him for that. Both of my parents always pushed me to reach higher and higher and higher and higher. I did my best in writing this thesis.

I also wish to thank my brother Jukka-Pekka and my sister Johanna. My sister is the best sister in the whole wide world, and my brother is the best brother in the whole wide world! That makes me a lucky one. You two have always supported me. Thank you for that.

I want to warmly thank Henna Laine and Liisa Kallioinen, and all the others that carried me through the darkness until I was able walk again. Without you I would not be here.

I would like to thank my husband's father, Kari-pappa, for helping us in so many ways. One of the most important things has been that Kari-pappa has looked after our children whenever we have needed help in that.

Finally, I wish to thank my husband Jukka, and our adorable children Elli and Osmo. When things go well at work, I celebrate with my family and we make pancakes. When I am stressed, Osmo wants to read his favorite book with me, and I am not stressed anymore. When I am down and blue, Elli wants to sing a song, and I am cheerful again. When I cry, Jukka kisses my tears away. With my family I forget everything else. When I met Jukka, my life changed completely. I found someone that I truly love from the bottom of my heart. Jukka knows very well that I am a workaholic, he knows very well that I always stress too much, he knows very very well that I am unstable and difficult. He still loves me the way I am. Jukka has been the greatest support for me, in everything. Jukka has stayed at home with the kids when they were babies, because I wanted to work. Jukka followed me to Brussels because my work took me there, and he will now follow me to Brussels again. Jukka always puts his family, me and the children, first. I love you Jukka, I love you and the kids more than anything. You three are my world.

Tampere, September 2011

Pauliina Ilmonen

Abstract

The aim of this doctoral thesis was to explore (asymptotical) characteristics of invariant coordinate system functionals and to introduce new approaches for independent component analysis.

Equivariance and invariance issues arise in multivariate statistical analysis. Sometimes statistical procedures have to be modified to obtain an affine equivariant or invariant version. This can be done by preprocessing the data, e.g., by standardizing the multivariate data or by transforming the data to an invariant coordinate system.

Two of the original articles deal with invariant coordinate selection and invariant coordinate system (ICS) functionals. Standardization of multivariate distributions, and characteristics of ICS functionals and statistics are examined. Also invariances up to some groups of transformations are discussed. Constructions of ICS functionals are addressed and asymptotical properties are explored. Also functionals and estimates of multivariate skewness and kurtosis are addressed. Application areas of ICS transformations are discussed. One important example of such application areas is independent component analysis.

Independent component analysis is a very timely research area with a wide field of applications. In the independent component model the elements of a p-variate random vector are assumed to be linear combinations of the elements of an unobservable p-variate vector with mutually independent components. In the independent component analysis the aim is to recover the independent components by estimating an unmixing matrix that transforms the observed p-variate vector to the independent components. New approaches for independent component analysis are provided in three of the original articles.

Deflation-based FastICA, where independent components are extracted one-by-one, is among the most popular methods for estimating an unmixing matrix in the independent component model. In the literature, it is often seen rather as an algorithm than an estimator related to a certain objective function, and only recently its statistical properties have been derived. One of the recent findings is that the order, in which the independent components are extracted in practice, has a strong effect on the performance of the estimator. A new reloaded procedure, to ensure that the independent components are extracted in an optimal order, is proposed in one of the articles.

In one of the original articles, new optimal (in Le Cam sense) inference procedures are developed under symmetry assumption of the independent components. The inference procedures are based on signed ranks. Hypothesis tests, estimators and confidence regions are provided, and asymptotical properties are examined.

The independent component model can be formulated in several ways: If the elements of a vector of independent components are permuted or multiplied by nonzero scalars, the vector still has independent components. The comparison of the performances of different unmixing matrix estimates is then difficult as the estimates are for different population quantities. A new natural performance index is suggested in one of the articles. The index is proven to possess several nice properties compared to previously presented indices, and it is easy and fast to compute. Also limiting behavior of the index, as the sample size approaches infinity, is explored.

To demonstrate the use of the new methods in practise, a data example is provided in the last chapter of this thesis.

KEY WORDS: multivariate analysis; invariant coordinate selection; invariant coordinate system functionals; independent component analysis; asymptotic normality; multivariate kurtosis; multivariate skewness; scatter and location functionals; semiparametric methods; sign and rank based methods; performance indices

Contents

Acknowledgements **3**

Abstract **7**

Abbreviations **11**

List of Original Publications **13**

1 Introduction **15**

2 Location and Scatter Functionals **16**
2.1 Definitions 16
2.2 Examples 17
2.3 Independence Property 18

3 Semiparametric Model **20**
3.1 Multivariate Location Scatter Model 20
3.2 Multivariate Normal Model 20
3.3 Multivariate Elliptical Model 20
3.4 Independent Component Model 21
3.5 Semiparametric Model 21

4 Invariant Coordinate Selection **22**
4.1 Invariance and Equivariance 22
4.2 Standardization of data 22
4.3 Invariant coordinate system (ICS) 24
4.4 Construction of ICS functionals 25
4.5 Applications of ICS 27

5 Independent Component Analysis **28**
5.1 Independent Component (IC) Model 28
5.2 IC Functionals 29
5.3 IC Functionals Based on the Use of Two Scatter Matrices . . 30
5.4 Deflation Based FastICA 31
5.4.1 Deflation Based FastICA reloaded 34
5.5 Inference Based on Signed Ranks in Symmetric IC model . . 34
5.5.1 ULAN 35
5.5.2 ULAN for symmetric IC models 36
5.5.3 Optimal signed-rank inference in symmetric IC models 37
5.6 Performance Indices 39

6 Data Example **42**

Summaries of Original Publications **57**

References **61**

Abbreviations

i.i.d.	independent and identically distributed
$\sim$	is distributed as
$\otimes$	Kronecker product
$\odot$	Hadamard (entrywise) product
e_i	ith vector of the canonical basis of $\mathbb{R}^p$
I, I_p	identity matrix
1_p	p variate vector of ones
P	permutation matrix
J	sign change matrix
D	scaling matrix
$C = PJD$	a matrix with exactly one non-zero element in each row and column
O	orthogonal matrix
$\operatorname{vec} A$	vector formed of the column vectors of a matrix A
$\operatorname{vecd}^\circ A$	vector obtained by removing the diagonal entries of A from $\operatorname{vec} A$
A^T	transpose of a vector or a matrix A
$\|\|A\|\|$	l_2 norm (Frobenius norm) of a vector or a matrix A
$Cov(\cdot)$	covariance matrix of $(\cdot)$
$E(\cdot)$	expected value of $(\cdot)$
$\inf(\cdot)$	infimum of $(\cdot)$
g'	derivative of a function g

List of Original Publications

I. P. Ilmonen, H. Oja and R. Serfling, On invariant coordinate system (ICS) functionals, *Submitted*, (2011).

II. P. Ilmonen, J. Nevalainen and H. Oja, Characteristics of multivariate distributions and the invariant coordinate system, *Statistics and Probability Letters*, Vol. 80(23-24) (2010), p. 1844-1853.

III. K. Nordhausen, P. Ilmonen, A. Mandal, H. Oja and E. Ollila, Deflation based fastICA reloaded, *Proceedings of EUSIPCO*, (2011), p. 1854–1858.

IV. P. Ilmonen and D. Paindaveine, Semiparametrically efficient inference based on signed ranks in symmetric independent component models, *The Annals of Statistics, to appear*, (2011).

V. P. Ilmonen, K. Nordhausen, H. Oja and E. Ollila, A new performance index for ICA: properties, computation and asymptotic analysis, *Proceedings of 9th International Conference on Latent Variable Analysis and Signal Separation*, (2010), p. 229–236.

1 Introduction

Various types of data sets are collected and stored into different databases these days. But how are all these data sets processed? We have loads of information. How can we extract and use all that information? What can we learn from it? There are endless different forms of data; signals, time series, images, functional data, etc. Sometimes a data set consists of very complicated elements. Often the tools to analyze new types of data sets are missing. If we are not able to analyze the data we have, we can not learn anything from it. The classical multivariate methods rely on the assumption of multivariate normality and i.i.d. observations. If these assumptions are not met, as often is the case, then traditional methods are misleading and inefficient.

Wider models, than multivariate normal model, are considered in this work, and new tools for analyzing multivariate data are developed.

Traditional location and scatter functionals and new competitors for them are discussed in Chapter 2. Different multivariate location-scatter models are considered in Chapter 3. In Chapter 4, invariant coordinate selection and invariant coordinate system functionals are explored. Independent component analysis is considered in Chapter 5. A data example is presented in Chapter 6.

2 Location and Scatter Functionals

Classical multivariate statistical inference methods (multivariate analysis of variance, principal component analysis, factor analysis, multivariate multiple regression, canonical correlation analysis, discriminant analysis, cluster analysis, etc) are typically based on the regular sample mean vector and covariance matrix. However, there exists a large number of competitors for those classical measures of location and scatter.

Various multivariate location and scatter functionals are discussed in this section.

2.1 Definitions

Let x denote a p-variate random vector with a cumulative distribution function F_x and let $X = [x_1....x_n]$, where $x_1, ..., x_n$ is a random sample from the distribution F_x.

Definition 1. A $p \times 1$ vector-valued functional $T(F_x)$, which is affine equivariant in the sense that

$$T(F_{Ax+b}) = AT(F_x) + b$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b, is called a *location functional*.

Definition 2. A $p \times p$ matrix-valued functional $S(F_x)$ which is positive definite and affine equivariant in the sense that

$$S(F_{Ax+b}) = AS(F_x)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b, is called a *scatter functional*.

The corresponding sample statistics are obtained if the functionals are applied to the empirical cumulative distribution F_n based on a sample $x_1, x_2, \ldots, x_n$. Notation $T(F_n)$ and $S(F_n)$ or $T(X)$ and $S(X)$ is used for the sample statistics. The location and scatter sample statistics then also satisfy

$$T(AX + b1_n^T) = AT(X) + b$$

and

$$S(AX + b1_n^T) = AS(X)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b.

Scatter matrix functionals are usually standardized such that in the case of standard multivariate normal distribution $S(F_x) = I$.

Definition 3. If a positive definite $p \times p$ matrix-valued functional $S(F_x)$ satisfies that $S(F_{Ax+b})$ is proportional to $AS(F_x)A^T$ for all nonsingular $p \times p$ matrices A and for all p-vectors b, then $S(F_x)$ is called a *shape functional*.

Note that clearly every scatter matrix functional is also a shape functional.

2.2 Examples

The first examples of location and scatter functionals are the mean vector and the regular covariance matrix:

$$T_1(F_x) = E(x) \quad \text{and} \quad S_1(F_x) = Cov(F_x) = E\left((x - E(x))(x - E(x))^T\right).$$

Location and scatter functionals can be based on the third and fourth moments as well. A location functional based on third moments is

$$T_2(F_x) = \frac{1}{p} E\left((x - E(x))^T Cov(F_x)^{-1}(x - E(x))x\right)$$

and a scatter matrix functional based on fourth moments is

$$S_2(F_x) = \frac{1}{p+2} E\left((x - E(x))(x - E(x))^T Cov(F_x)^{-1}(x - E(x))(x - E(x))^T\right).$$

These functionals, $T_2(F_x)$ and $S_2(F_x)$, together with $T_1(F_x)$ and $S_1(F_x)$, can be used to construct measures of multivariate skewness and kurtosis, respectively. In the case of standard multivariate normal distribution $T_2(F_x) = 0_p$ and $S_2(F_x) = I_p$.

There are several other location and scatter functionals, even families of them, having different desirable properties (robustness, efficiency, limiting multivariate normality, fast computations, etc). See for example Lopuhaä (1989); Maronna, Mardin and Yohai (2006); Davies (1987); Kent and Tyler (1996).

M-functionals of location and scatter are commonly used. They are defined as solutions of the two equations

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

and

$$S(F_x) = E(w_2(r)(x - T(F_x))(x - T(F_x))^T),$$

where $w_1(r)$ and $w_2(r)$ are nonnegative continuous functions of the Mahalanobis distance $r = ||S(F_x)^{-1/2}(x - T(F_x))||$. (The $||\cdot||$ here denotes the

l_2 norm of $\cdot$.) M-functionals were introduced by Maronna (1976). The mean vector and the regular covariance matrix are M-functionals with $w_1(r) = w_2(r) = 1$, and as an other example, the Hettmansperger-Randles functionals (Hettmansperger and Randles, 2002) have weight functions

$$w_1(r) = \frac{1}{r} \quad \text{and} \quad w_2(r) = \frac{p}{r^2}.$$

Several other weight functions have been proposed in the literature, see for example Huber (1964); Kent and Tyler (1991).

Another important family of location and scatter functionals is the family of one step M-functionals. Given a pair of location and scatter functionals (T_1, S_1), the one step M-functionals are defined to be

$$T_2(F_x) = E(w_1(r_1))^{-1} E(w_1(r_1)x)$$

and

$$S_2(F_x) = E(w_2(r_1)(x - T_1(F_x))(x - T_1(F_x))^T),$$

where $w_1(r)$ and $w_2(r)$ are again nonnegative continuous weight functions and $r_1 = ||S_1(F_x)^{-1/2}(x - T_1(F_x))||$. The location functional based on third moments and the scatter functional based on fourth moments are obtained with choices $T_1(F_x) = E(x)$, $S_1(F_x) = E\left((x - E(x))(x - E(x))^T\right)$, $w_1(r) = r^2/p$ and $w_2(r) = r^2/(p + 2)$. Tyler's shape matrix functional is obtained with $w_2(r) = p/r^2$ and it is calculated with respect to some given location functional (Hettmansperger and Randles, 2002). The symmetrized version (see Section 2.3) of Tyler's shape matrix is called Dümbgen's shape matrix (Dümbgen, 1998).

The Hallin-Paindaveine shape matrix functional $S_{HP}(F_x)$ is defined as

$$S_{HP}(F_x) = S_{HR}(F_x)^{1/2} E(\psi_p^{-1}(F_{||z||}(||z||)) \frac{zz^T}{||z||}) S_{HR}(F_x)^{1/2},$$

where ψ_p denotes the cdf of a chi-square distribution with p degrees of freedom, $z = S_{HR}(F_x)^{-1/2}(x - T_{HR}(x))$, $S_{HR}(F_x)$ and $T_{HR}(F_x)$ denote the Hettmansperger-Randles functionals and $S_{HR}(F_x)^{1/2}$ is the symmetric square root of $S_{HR}(F_x)$ (Hallin and Paindaveine, 2006).

Later, in a data example in Chapter 6, we use the regular covariance matrix, the matrix based on fourth moments, Dümbgen's shape matrix and the Hallin-Paindaveine shape matrix.

2.3 Independence Property

Let $S(F_x)$ denote any shape or scatter functional.

Definition 4. If $S(F_x)$ is a diagonal matrix for all x having independent components, it is said to posses the *independence property*.

The regular covariance matrix is a scatter matrix with the independence property. Another example of a scatter matrix with the independence property is the matrix based on fourth moments.

Most scatter functionals do posses the independence property only if all the components (or all the components except for one) are symmetric. However, every scatter/shape matrix functional $S(F_x)$ can be symmetrized by setting

$$S_{sym}(F_x) = S(F_{x_1 - x_2}),$$

where x_1 and x_2 are independent random vectors having the same cumulative distribution function F_x. The resulting symmetrized scatter matrix does always have the independence property (Sirkiä, Taskinen and Oja, 2007; Oja, Sirkiä and Eriksson, 2006). For a similar approach in the context of so called S-estimators, see Roelandt, Van Aelst and Croux (2009).

3 Semiparametric Model

The classical multivariate methods rely on the assumption of multivariate normality. There exist several more general models.

3.1 Multivariate Location Scatter Model

Different multivariate *location-scatter models* are obtained if one assumes that a p-variate random vector x can be written as

$$x = \Omega z + \mu$$

where z is a "standardized" p-variate latent vector, μ is a *location vector* and Ω is a full-rank $p \times p$ matrix, termed *mixing matrix* or *transformation matrix*. The inverse of Ω, $\Gamma = \Omega^{-1}$, is called an *unmixing matrix* or *retransformation matrix*, and $\Sigma = \Omega\Omega^T$ is the *scatter matrix*. Posing various assumptions on the distribution of z yields different parametric or semiparametric multivariate models with parameters μ and Σ, or μ and Ω. An excellent overview of different parametric and semiparametric location-scatter models is given in Nordhausen, Oja and Ollila (2011b).

3.2 Multivariate Normal Model

The classical multivariate methods rely on the assumption of multivariate normality, that is, $z \sim N_p(0, I_p)$. The location parameter μ is the mean vector and the scatter parameter Σ is the covariance matrix. As $Oz \sim N_p(0, I_p)$ for all orthogonal matrices O, the mixing matrix Ω or the unmixing matrix Γ are defined only up to an orthogonal transformation in the *multivariate normal model*. The mean vector and the sample covariance matrix are sufficient statistics for μ and Σ under the normality assumption, but they are extremely sensitive to outlying observations and have poor efficiency in models with heavy tailed distributions. One possibility to avoid these problems is to weaken the assumptions and to develop valid and efficient procedures in wider models than the multivariate normal model.

3.3 Multivariate Elliptical Model

In the *multivariate elliptical model* it is assumed that $z \sim Oz$ for all orthogonal matrices O. To fix Σ it is often assumed that $E(\|z\|^2) = p$ or that

$Med(\|z\|^2) = \chi^2_{p,1/2}$. (The first configuration naturally requires that finite second moments exist, but the second allows to avoid any moment assumptions.) As in the multivariate normal model, Ω and Γ are again defined only up to an orthogonal transformation. Elliptical distributions are thus symmetric in the sense that $z \sim Oz$ for all O, but they may vary in their kurtosis properties. The model permits for heavier (or lighter) tails than the multivariate normal model, and therefore elliptical models are commonly seen as a more realistic alternative to the multivariate normal model. Robust testing and estimation procedures considered in the literature, for example, often assume ellipticity.

3.4 Independent Component Model

In the *independent component (IC) model* it is assumed that z is a p-variate vector with mutually independent components. The IC model can be formulated in several ways: If the independent components are permuted or multiplied by nonzero scalars they still remain independent.

A *semiparametric independent component (IC) model* is obtained by either standardizing the marginal distributions of z (see Section 3.5) or by normalizing the mixing matrix Ω (Chapter 5, Section 5.1).

A *parametric independent component (IC) model* is obtained if the vector z is assumed to have independent and standardized components and the density function $f(z) = \prod_{j=1}^{p} f_j(z_j)$ with some known standardized marginal densities $f_1, ..., f_p$.

Unlike the multivariate normal or the multivariate elliptical model, the independent component model allows also asymmetric distribution.

3.5 Semiparametric Model

A general *semiparametric location-scatter-skewness-kurtosis model*, shortly *semiparametric model*, is standardized using two location functionals T_1 and T_2 and two scatter functionals S_1 and S_2. In the semiparametric model it is assumed that $T_1(F_z) = 0$, $S_1(F_z) = I_p$, $T_2(F_z) = \delta$ and $S_2(F_z) = \Lambda$, where δ is a p-vector with all components $\delta_i \geq 0$, $i = 1, ..., p$, and Λ is a diagonal matrix with diagonal elements $\lambda_1 \geq ... \geq \lambda_p > 0$. The parameters of the semiparametric model are the mean vector μ, the scatter matrix $\Sigma = \Omega\Omega^T$, the skewness vector δ, and the kurtosis matrix Λ. The mixing and unmixing matrices, Ω and $\Gamma = \Omega^{-1}$, are uniquely defined if $\delta_i > 0$, $i = 1, ..., p$, and $\lambda_1 > ... > \lambda_p > 0$. When the model parameters are fixed in this way, the unmixing matrix can be used to transform the random vector to an invariant coordinate system (ICS) (Tyler, Critchley, Dümbgen and Oja, 2009). See Chapter 4, Section 4.4. If the used scatter functionals do posses the independence property and if the components of z are independent, then the model is called *semiparametric independent component (IC) model* and the unmixing matrix Γ is a solution in the independent component analysis (ICA). See Chapter 5, Section 5.3.

4 Invariant Coordinate Selection

Equivariance and invariance issues often arise in multivariate statistical analysis. Multivariate data are often standardized someway or transformed to an invariant coordinate system in order to obtain affine equivariant or invariant versions of statistical procedures. Invariance or equivariance of a statistical procedure is essential for ensuring that the obtained results are not affected by the used coordinate system. Standardization of multivariate data and invariance and equivariance issues were discussed in Ilmonen, Oja and Serfling (2011b).

4.1 Invariance and Equivariance

It is required that multivariate location and scatter statistics are affine equivariant. Multivariate testing and estimation procedures in general are hoped to be affine invariant and affine equivariant, respectively.

Let x denote a p-variate random vector with a cumulative distribution function F_x and let $X = [x_1....x_n]$, where $x_1, ..., x_n$ is a random sample from the distribution F_x. Let $\mathcal{M}$ denote the set of all full-rank $p \times p$ matrices. Affine invariance and maximal (affine) invariance are defined as follows.

Definition 5. A statistic $Q(X)$ is *affine invariant* if

$$Q(AX + b1_n^T) = Q(X)$$

for all $A \in \mathcal{M}$ and $b \in \mathbb{R}^p$, and a statistic $Q(X)$ is *maximal invariant* under the group of affine transformations if it is affine invariant and if

$$Q(Y) = Q(X) \quad \Rightarrow \quad Y = AX + b1_n^T, \quad \text{for some } A \in \mathcal{M} \text{ and } b \in \mathbb{R}^p.$$

4.2 Standardization of data

Location and scatter functionals are often used to center and standardize distributions. Let Σ denote a positive definite $p \times p$ matrix. For the standardization we need the following definition of a matrix $\Sigma^{-1/2}$.

Definition 6. A matrix $\Sigma^{-1/2}$ denotes any matrix G, which satisfies

$$G\Sigma G^T = I.$$

For a scatter functional $S(F_x)$, let $S^{-1/2}(F_x)$ denote any functional $G(F_x)$ which satisfies

$$G(F_x)S(F_x)G(F_x)^T = I.$$

Note that $\Sigma^{-1/2}$ is defined only up to an orthogonal transformation: if $G\Sigma G^T = I$, then also $(VG)\Sigma(VG)^T = I$, for any orthogonal matrix V. One can always define Σ as $\Sigma = U\Lambda U^T$, where U is a unique orthogonal matrix and Λ is a unique diagonal matrix. (Here Λ is a diagonal matrix having the eigenvalues of Σ as its diagonal elements and the column vectors of U are the corresponding eigenvectors.) Now one can choose $\Sigma^{-1/2}$ in a unique way by requiring, for example, that

1. $\Sigma^{-1/2}$ is unique lower diagonal (the inverse of the lower diagonal matrix in the Cholesky decomposition of Σ),
2. $\Sigma^{-1/2}$ is unique upper diagonal (formed by permuting the rows of the inverse of the lower diagonal matrix in the Cholesky decomposition of Σ),
3. $\Sigma^{-1/2} = U\Lambda^{-1/2}U^T$, where $\Lambda^{-1/2} = diag(1/\lambda_1, \cdots, 1/\lambda_p)$ (symmetric version), or
4. $\Sigma^{-1/2} = \Lambda^{-1/2}U^T$ (rows are rescaled eigenvectors).

In general, $\Sigma^{-1/2}$ is any choice in the set of matrices

$$\{V\Lambda^{-1/2}U^T \mid V \text{ orthogonal and } \Sigma = U\Lambda U^T\}.$$

For a scatter functional $S(F_x)$, the corresponding eigenvector and eigenvalue functionals are defined implicitly by

$$S(F_x) = U(F_x)\Lambda(F_x)U(F_x)^T.$$

Now

$$S^{-1/2}(F_x) = V(F_x)\Lambda(F_x)^{-1/2}U(F_x)^T,$$

where $V(F_x)$ is an orthogonal matrix functional.

To fix functional $S^{-1/2}(F_x)$ uniquely, one thus has to fix the functional $V(F_x)$. Possible choices for $V(F_x)$ are for example $U(F_x)$ or I or a matrix $V(F_x)$ that makes $S^{-1/2}(F_x)$ unique upper or lower triangular. However, none of the choices above guarantees the invariance of $S^{-1/2}(F_x)x$. In fact, if $S(F_x)$ is a scatter functional then $S^{-1/2}(F_{Ax+b}) = US^{-1/2}(F_x)A^{-1}$ for some orthogonal $U = U(F_x, A)$. Thus $S^{-1/2}(F_x)x$ is not necessarily invariant under group of transformations

$$\{h \mid h(x) = Ax, A \in \mathcal{M}\},$$

and

$$S^{-1/2}(F_x)(x - T(F_x))$$

is not necessarily invariant under group of transformations

$$\{g \mid g(x) = Ax + b, A \in \mathcal{M}, b \in \mathbb{R}^p\}.$$

Note that the functional $S^{-1/2}(F_x)$ can be made affine equivariant with suitable choice of $V(F_x)$. That will be discussed in Section 4.4.

4.3 Invariant coordinate system (ICS)

A definition of an invariant coordinate system functional is given next.

Definition 7. An *invariant coordinate system (ICS) functional* is a nonsingular $p \times p$ matrix-valued functional $G(F_x)$ satisfying

$$G(F_{Ax+b}) = G(F_x)A^{-1},$$

for all $A \in \mathcal{M}$ and $b \in \mathbb{R}^p$, and an *invariant coordinate system (ICS) statistic* is a $p \times p$ matrix-valued sample statistic $G(X)$ satisfying

$$G(AX + b1_n^T) = G(X)A^{-1},$$

for all $A \in \mathcal{M}$ and $b \in \mathbb{R}^p$

For the following result for the ICS statistic $G(X)$, see Ilmonen et al. (2011b).

Theorem 1. *(i) If $G(X)$ satisfies $G(AX) = G(X)A^{-1}$ for all nonsingular $p \times p$ matrices A, then $G(X)X$ is maximal invariant under the transformations in $\{h \mid h(x) = Ax, A \in \mathcal{M}\}$.*
(ii) If $G(X)$ satisfies $G(AX + b1_n^T) = G(X)A^{-1}$ for all nonsingular $p \times p$ matrices A and for all p-vectors b and if $T(X)$ is a location statistic, then $G(X)(X - T(X)1_n^T)$ is maximal invariant under the transformations in $\{g \mid g(x) = Ax + b, A \in \mathcal{M}, b \in \mathbb{R}^p\}$.

In practical problems full invariance is not always needed. Weaker concepts of invariance are obtained if one only requires invariance up to some groups of transformations. Let $\mathcal{M}_s$ denote some particular subgroup of nonsingular $p \times p$ matrices.

Definition 8. A $p \times p$ matrix-valued functional $G(F_x)$ is an *invariant coordinate system (ICS) functional up to a group of transformations $\mathcal{M}_s$* if, for any $A \in \mathcal{M}$ and $b \in \mathbb{R}^p$,

$$G(F_{Ax+b}) = MG(F_x)A^{-1},$$

for some $M \in \mathcal{M}_s$, and a $p \times p$ matrix-valued sample statistic $G(X)$ is an *invariant coordinate system statistic up to a group of transformations $\mathcal{M}_s$* if, for any $A \in \mathcal{M}$ and $b \in \mathbb{R}^p$,

$$G(AX + b1_n^T) = MG(X)A^{-1}$$

for some $M \in \mathcal{M}_s$.

The next result then follows (Ilmonen et al., 2011b).

Theorem 2. *Let $Q(X)$ be invariant under transformations in $\mathcal{M}_s$, that is, $Q(MX) = Q(X)$, for all $M \in \mathcal{M}_s$. If $G(X)$ is an ICS statistic up to $\mathcal{M}_s$, then $Q(G(X)X)$ is affine invariant, that is, $Q(G(AX)AX) = Q(G(X)X)$ for any $A \in \mathcal{M}$.*

In the literature of multivariate nonparametric statistics one often has invariance under the following groups of transformations.

1. $\mathcal{D}_0 = \{dI \mid d > 0\}$ (homogeneous rescaling),
2. $\mathcal{D} = \{\text{diag}(d_1, ..., d_p) \mid d_i > 0,\ i = 1, ..., p\}$ (heterogeneous rescaling),
3. $\mathcal{J} = \{\text{diag}(c_1, ..., c_p) \mid c_i = \pm 1,\ i = 1, ..., p\}$ (heterogeneous sign changes),
4. $\mathcal{P} = \{P \mid P \text{ is a permutation matrix }\}$ (permuting the components),
5. $\mathcal{U} = \{U \mid U \text{ is orthogonal}\}$ (rotation and reflection), and
6. $\mathcal{C} = \{PJD \mid P \in \mathcal{P}, D \in \mathcal{D}, \text{and } J \in \mathcal{J}\ \}$ (permuting, rescaling and sign changes).

Using these definitions we can now also say that, if $S(F_x)$ is a scatter matrix functional, then $S(F_x)^{-1/2}$ is an ICS functional up to $\mathcal{U}$.

ICS functionals could be used to preprocess the data to obtain affine invariant or equivariant statistical procedures. Theorem 2 then shows what is needed for full invariance. For tests based on spatial signs and ranks, for example, we need ICS functionals only up to transformations in $\mathcal{U}$, see Oja (2010). For tests based on marginal signs and ranks it is sufficient to have ICS functionals up to a group of transformations $\mathcal{C}$, see Puri and Sen (1971); Nordhausen, Oja and Tyler (2006).

4.4 Construction of ICS functionals

The first example of ICS functional in the literature was introduced by Chaudhuri and Sengupta (1993) in the context of a location model $F(x-\mu)$ for testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Since $A\mu = 0$ for all $A \in \mathcal{M}$ if and only if $\mu = 0$, Chaudhuri and Sengupta (1993) suggest using a test function Q satisfying $Q(AX) = Q(X)$ for all $A \in \mathcal{M}$.

Here the focus is on an approach based on the use of two scatter matrix functionals. Construction of ICS functionals based on the use of two scatter matrix functionals was presented by Tyler et al. (2009). Let $S_1(F_x)$ and $S_2(F_x)$ denote two different scatter matrix functionals, and consider the set of distributions

$$\mathcal{F} = \{F_x \mid S_1^{-1}(F_x)S_2(F_x) \text{ has distinct eigenvalues}\}$$

In this model of distributions one can define ICS functionals in the following ways.

1. Find a transformation matrix functional $G(F_x)$ and a diagonal matrix valued functional $L(F_x)$ as a solution of the eigenvector and eigenvalue problem
$$S_1^{-1}(F_x)S_2(F_x)G(F_x)^T = G(F_x)^T L(F_x).$$
As the lengths, signs, and order of the eigenvectors are not fixed, G is an ICS functional in $\mathcal{F}$ up to $\mathcal{C}$. See Tyler et al. (2009).

2. Find a transformation matrix functional $G(F_x)$ and diagonal matrix valued functional $L(F_x)$ which solve the above eigenvector and eigenvalue problem and satisfy
$$G(F_x)S_1(F_x)G(F_x)^T = I \text{ and } G(F_x)S_2(F_x)G(F_x)^T = L(F_x)$$
where the eigenvalues in $L(F_x)$ are in decreasing order. Note that $G(F_x)$ is now chosen to be a certain version of $S^{-1/2}(F_x)$. With these restrictions, $G(F_x)$ is an ICS functional up to $\mathcal{J}$.

3. Let $T_1(F_x)$ and $T_2(F_x)$ denote two different location functionals. Find a transformation matrix functional $G(F_x)$ and diagonal matrix valued functional $L(F_x)$ which solve the above eigenvector and eigenvalue problem and satisfy
$$G(F_x)S_1(F_x)G(F_x)^T = I, \quad G(F_x)S_2(F_x)G(F_x)^T = L(F_x)$$
and
$$d(F_x) = G(F_x)(T_1(F_x) - T_2(F_x)) \geq 0,$$
where the eigenvalues in $L(F_x)$ are in a decreasing order. If
$$\mathcal{F} = \{F_x \mid L(F_x)_{11} > \cdots > L(F_x)_{pp} > 0, G(F_x)(T_1(F_x) - T_2(F_x)) > 0\},$$
then the functional $G(F_x)$ is an ICS functional in $\mathcal{F}$ and functionals $d(F_x)$ and $L(F_x)$ can be seen as multivariate measures of skewness and kurtosis. See Ilmonen, Nevalainen and Oja (2010a); Nordhausen et al. (2011b).

If $x = Az + b$ for some $A \in \mathcal{M}$ and $b \in \mathbb{R}^p$, where $JPz \sim z$ for all $J \in \mathcal{J}$ and $P \in \mathcal{P}$, then $S_1(F_x)$ and $S_2(F_x)$ are proportional and $F_x \notin \mathcal{F}$. Thus the ICS functional based on two scatter matrices is not uniquely defined for example for the distributions in the elliptic model, or for the distributions where the components of z are i.i.d. However, if X is a random sample from a continuous p-variate distribution, then $F_n \in \mathcal{F}$ with probability one and $G(X) = G(F_n)$ is an ICS statistic.

Sample statistics $G(X)$ and $L(X)$ as defined in point 3 above are affine equivariant and invariant in the sense that
$$G(AX + b1_n^T) = G(X)A^{-1} \text{ and } L(AX + b1_n^T) = L(X)$$
for all $A \in \mathcal{M}$ and $b \in \mathbb{R}^p$. For the asymptotics, it is therefore not a restriction to assume that X is a random sample from a distribution F_x with $S(F_x) = I$ and $S_2(F_x) = \Lambda$, where the diagonal elements of Λ are $\lambda_1 \geq ... \geq \lambda_p > 0$. For the following result, see Ilmonen et al. (2010a).

Theorem 3. *Assume that*

$$\sqrt{n}(S_1(X) - I) = O_p(1) \quad \textit{and} \quad \sqrt{n}(S_2(X) - \Lambda) = O_p(1),$$

with $\lambda_1 > ... > \lambda_p > 0$, *and assume that the diagonal elements of* $G(X)$ *are set to be positive. Then*

$$\begin{aligned}
\sqrt{n}(G(X)_{ii} - 1) &= -\frac{1}{2}\sqrt{n}(S_1(X)_{ii} - 1) + o_p(1), \\
(\lambda_i - \lambda_j)\sqrt{n}G(X)_{ij} &= \sqrt{n}S_2(X)_{ij} - \lambda_i\sqrt{n}S_1(X)_{ij} + o_p(1), \ i \neq j, \quad \textit{and} \\
\sqrt{n}(L(X)_{ii} - \lambda_i) &= \sqrt{n}(S_2(X)_{ii} - \lambda_i) - \lambda_i\sqrt{n}(S_1(X)_{ii} - 1) + o_p(1).
\end{aligned}$$

It is interesting to note that the asymptotic behavior of the diagonal elements of $G(X)$ does not depend on $S_2(X)$ at all.

The three equations in Theorem 3 above are in fact true if λ_i is distinct from all the other eigenvalues λ_j, $j \neq i$. The limiting joint distributions of the sample eigenvectors and sample eigenvalues for a subset with distinct population eigenvalues can then be derived from the limiting distributions of $S_1(X)$ and $S_2(X)$. (Ilmonen et al., 2011b)

4.5 Applications of ICS

There are several applications for ICS functionals based on the use of two scatter matrix functionals and two location functionals. For example, finding an unmixing matrix in the independent component analysis, see Chapter 5, Section 5.3; and deriving multivariate skewness and kurtosis measures, see Kankainen, Taskinen and Oja (2007); Nordhausen et al. (2011b); Ilmonen et al. (2010a). Also sliced inverse regression (Li, 1991) can be seen as an ICS functional application based on two scatter matrices, see Liski, Nordhausen and Oja (2011).

For other ICS-functionals and their applications, see Critchley, Pires and Amado (2006); Chakraborty and Chaudhuri (1998, 1996); Caussinus and Ruiz-Gazen (1993)

5 Independent Component Analysis

Independent component analysis (ICA) is an important and timely research area. The field of applications of ICA is wide and constantly expanding, varying from biomedical image data applications to signal processing, (Hyvärinen, Karhunen and Oja, 2001). ICA is also an interesting example of the use of the ICS functionals.

5.1 Independent Component (IC) Model

In the *independent component (IC) model* it is assumed that the p-variate vector

$$x = \Omega z, \tag{5.1}$$

where Ω is a full-rank $p \times p$ *mixing matrix* and z is a p-variate vector with mutually independent components.

In the *independent component analysis (ICA)* the aim is to find an estimate for an *unmixing matrix* Γ such that Γx has independent components. Naturally $\Gamma = \Omega^{-1}$ is one possible unmixing matrix. The IC model can be formulated in several ways: If the independent components are permuted or multiplied by nonzero scalars they still remain independent. Then the ICA problem reduces to estimating an unmixing matrix Ω^{-1} only up to the order, signs and scales of the row vectors.

Under the assumption that z has at most one Gaussian marginal, permutations (P), sign changes (J) and scale transformations (D) of the independent components are the only sources of unidentifiability for Ω, see, e.g., Theis (2004). Solving this identifiability problem requires either standardizing the marginal distributions of z (Ilmonen et al., 2010a) or normalizing the mixing matrix Ω (Ilmonen and Paindaveine, 2011). The marginal distributions of z can be standardized using two different location functionals T_1 and T_2 and two different scatter functionals S_1 and S_2 by setting

$$\begin{aligned} T_1(F_z) &= 0,\ S_1(F_z) = I_p, \\ T_2(F_z) &= \delta \text{ and } S_2(F_z) = D, \end{aligned}$$

where δ is a p-vector with all components $\delta_i \geq 0$, $i = 1, ..., p$, and D is a diagonal matrix with diagonal elements $d_1 \geq ... \geq d_p > 0$. If now $\delta_i > 0$,

$i = 1, ..., p$, and if the diagonal elements of D are distinct, then the mixing matrix Ω is uniquely defined. See also Chapter 3, Section 3.5.

The IC model 5.1 can also be standardized by standardizing the mixing matrix using a mapping

$$\Omega \mapsto L = \Omega D_1^+ P D_2,$$

where D_1^+ is the positive definite diagonal matrix that makes each column of ΩD_1^+ have Euclidean norm one, P is the permutation matrix for which the matrix $B = (b_{ij}) = \Omega D_1^+ P$ satisfies $|b_{ii}| > |b_{ij}|$ for all $i < j$, and D_2 is the diagonal matrix that makes all the diagonal entries of $L = \Omega D_1^+ P D_2$ to be equal to one. Ties may be taken care of e.g., by basing the ordering on subsequent rows of B above, but they may prevent the mapping to be continuous. Thus it is often convenient to restrict to the collection of mixing matrices Ω for which no ties occur in the permutation step.

Both standardization methods presented above enable to fix Model 5.1 uniquely.

There is a large number of estimates and algorithms for the ICA problem in the literature, and most popular algorithms proceed as follows.

1. In model (5.1) fix z such that $Cov(z) = I_p$. Then, for a prewhitened version of x, it holds that

   $$y = Cov(x)^{-1/2}(x - E(x)) = U(z - E(z))$$

   for some orthogonal matrix U.

2. Using y, find an orthogonal matrix $V = (v_1, ..., v_p)$ with the columns v_i, $i = 1, ..., p$, chosen to maximize (or minimize) a criterion function, say $|E[G(v_i^T y)]|$. The optimization may be conducted one by one or simultaneously. Measures of marginal nongaussianity (negentropy, kurtosis measures) or likelihood functions with parametric marginal distributions are often used.

3. The final IC estimate is $V^T Cov(x)^{-1/2}$. (Note that $V^T = PJU$.)

The fastICA estimate (Hyvärinen and Oja, 1997) uses an algorithm of such type where the columns of V are found by maximizing a negentropy criterion. In deflation based fastICA, the columns of V are found one by one and in symmetric fastICA algorithm, the optimizations are conducted simultaneously. For more details about fastICA and several similar estimates and algorithms, see Cichocki and Amari (2006); Hyvärinen et al. (2001). For other type of estimates, see Chen and Bickel (2005, 2006).

5.2 IC Functionals

The following formal mathematical definition of an IC functional was given in Ilmonen, Nordhausen, Oja and Ollila (2010b).

Let $\mathcal{M}$ denote the set of all full-rank $p \times p$ matrices. (Then naturally all unmixing matrices $\Gamma \in \mathcal{M}$.) Let P denote a permutation matrix, J a sign-change matrix, and D a scaling matrix. Let

$$\mathcal{C} = \{C \in \mathcal{M} \mid C = PJD \text{ for some } P, J, \text{ and } D\}.$$

Now two matrices Γ_1 and Γ_2 are said to be equivalent if $\Gamma_1 = C\Gamma_2$ for some $C \in \mathcal{C}$. We then write $\Gamma_1 \sim \Gamma_2$.

Definition 9. A functional $\Gamma(F_x) \in \mathcal{M}$ is an *IC functional* in the IC model (5.1) if

$$\Gamma(F_x)\Omega \sim I_p,$$

and if it is affine equivariant in the sense that

$$\Gamma(F_{Ax}) = \Gamma(F_x)A^{-1}$$

for all $A \in \mathcal{M}$.

If z has independent components, then so has Cz for all $C \in \mathcal{C}$. Then, for any $C \in \mathcal{C}$, the IC model can be reformulated as

$$x \;=\; (\Omega C^{-1})(Cz) = \Omega^* z^*$$

where Ω^* is a new mixing matrix and z^* is a new (transformed) vector of independent components. (Matrix C is used in the transformation.) Note that

$$\Gamma(F_x)\Omega \sim \Gamma(F_x)\Omega^*.$$

The functional $C(F_x) = \Gamma(F_x)\Omega$, with values in $\mathcal{C}$, depends on the distribution of z but not on the value of Ω. If the model is fixed by choosing $z^* = C(F_x)z$, and $x = \Omega^* z^*$, then $\Omega^* = \Gamma(F_x)^{-1}$. This formulation of the model is then most natural (canonical) for functional $\Gamma(F_x)$.

5.3 IC Functionals Based on the Use of Two Scatter Matrices

Let $S_1(F_x)$ and $S_2(F_x)$ denote two different scatter functionals with the independence property. The IC functional $\Gamma(F_x)$ based on the scatter matrix functionals $S_1(F_x)$ and $S_2(F_x)$ is defined as a solution of the equations

$$\Gamma S_1(F_x)\Gamma^T = I_p \;\text{ and }\; \Gamma S_2(F_x)\Gamma^T = \Lambda,$$

where $\lambda = \Lambda(F_x)$ is a diagonal matrix with diagonal elements $\lambda_1 \geq ... \geq \lambda_p > 0$. (See also Chapter 4 Section 4.4.) One of the first solutions for the ICA problem, the fourth order blind identification (FOBI) functional (Cardoso, 1989) is obtained if the scatter functionals $S_1(F_x)$ and $S_2(F_x)$ are the scatter matrices based on the second and fourth moments, respectively. The use of two scatter matrices in the ICA has been studied in

Nordhausen, Oja and Ollila (2008); Oja et al. (2006) (real data) and in Ollila, Oja and Koivunen (2008) (complex data).

For the asymptotical behavior of the corresponding estimates $\Gamma(F_n) = \Gamma(X)$ and $\Lambda(F_n) = \Lambda(X)$, see Theorem 3 in Chapter 4. Note however that in Theorem 3, the scatter functionals do not necessarily posses the independence property, but in the context of independent component analysis, the independence property is crucial. In the ICA one may also use shape functionals, instead of scatter functionals, but asymptotical analysis is then not as straightforward as when using scatter functionals. Note also that Ilmonen et al. (2010a) considered the limiting distribution of the FOBI estimate (with limiting covariance matrix) in more details.

Similar approaches like JADE (Cardoso and Souloumiac, 1993) or the matrix-pencil approach (Yeredor, 2009) (approximately) diagonalize jointly two or more data matrices (not necessarily scatter matrices). The estimates are typically not affine equivariant and their asymptotic behavior is still unknown.

5.4 Deflation Based FastICA

Another important family of IC functionals is given by the deflation-based fastICA algorithm. FastICA is one of the most popular and widespread ICA algorithms. Detailed examination of fastICA functionals are provided for example in Hyvärinen and Oja (1997) and Ollila (2010).

Assume that $x = \Omega z$ as in model (5.1) with finite first and second moments $E(x) = \mu$ and $Cov(x) = \Sigma$. In deflation based fastICA, a criterion function $|E(G(\gamma^T(x-\mu))|$ is first maximized under the constraint $\gamma^T\Sigma\gamma = 1$. Then, after finding $\gamma_1, ..., \gamma_{k-1}$, the kth source maximizes $|E(G(\gamma^T(x-\mu))|$ under the constraint

$$\gamma_k^T\Sigma\gamma_k = 1 \quad \text{and} \quad \gamma_j^T\Sigma\gamma_k = 0, \quad j = 1, ..., k-1.$$

If G satisfies the condition

$$|E(G(\alpha_1 z_1 + \alpha_2 z_2)| \leq \max(|E(G(z_1)|, |E(G(z_2)|)$$

for all independent z_1 and z_2 such that $E(z_1) = E(z_2) = 0$ and $E(z_1^2) = E(z_2^2) = 1$ and for all α_1 and α_2 such that $\alpha_1^2 + \alpha_2^2 = 1$, then the independent components are found using the above strategy. It is easy to check that the condition is true for $G(z) = z^4 - 3$, for example. See Bugrien (2005).

Let $T(F_x)$ denote the mean vector (functional) and $S(F_x)$ the covariance matrix (functional). Then the kth fastICA functional $\gamma_k(F_x)$ optimizes the Lagrangian function

$$|E[G(\gamma_k^T(x - T(F_x)))]| - \frac{\lambda_{kk}}{2}(\gamma_k^T S(F_x)\gamma_k - 1) - \sum_{j=1}^{k-1}\lambda_{jk}\gamma_j^T S(F_x)\gamma_k,$$

where $\lambda_{1k}, ..., \lambda_{kk}$ are the Lagrangian multipliers. If $g = G'$, then, under general assumptions, the functional $\Gamma(F_x) = \Gamma = (\gamma_1, ..., \gamma_p)^T$ satisfies the p estimating equations

$$E[g(\gamma_k^T(x-T(F_x))(x-T(F_x))] = S(F_x)\sum_{j=1}^{k}\gamma_j\gamma_j^T E[g(\gamma_k^T(x-T(F_x))(x-T(F_x))],$$

$k = 1, ..., p$. If $z = \Gamma x$ has independent components then Γ solves the above estimating equations. Note that the estimating equations do not fix the order of sources $\gamma_1, ..., \gamma_p$ anymore. See Ilmonen, Nordhausen, Oja and Ollila (2011a)

Popular choices of g for practical calculations are *pow3*: $g(z) = z^3$, *tanh*: $g(z) = tanh(z)$, and *gaus*: $g(z) = ze^{-z^2/2}$.

The limiting behavior of the deflation based fastICA estimate was examined in Ilmonen et al. (2011a) including proving asymptotic normality under some general conditions. The covariance structure of the row vectors of deflation based fastICA estimate was given (in closed form) in Ollila (2010), but the asymptotic normality remained unproven there.

If $E(x) = 0$ then the fastICA algorithm for γ_k uses the iteration steps

1. $\gamma_k \leftarrow \Sigma^{-1}E[g(\gamma_k^T x)x] - E[g'(\gamma_k^T x)]\gamma_k$
2. $\gamma_k \leftarrow \gamma_k - \sum_{j=1}^{k}(\gamma_k^T\Sigma\gamma_j)\gamma_j$
3. $\gamma_k \leftarrow \gamma_k/\sqrt{\gamma_k^T\Sigma\gamma_k}$

The sample version is naturally obtained if the expected values are replaced by the averages in the above formula. It is important to note that it is not known in which order the components are found in the above algorithm. The order depends strongly on the initial value in the iteration.

The limiting behavior of the sample statistic $\hat{\Gamma}$, based on a random sample $x_1, ..., x_n$, was examined in Ilmonen et al. (2011a). Assume that $E(x_i) = 0$ and $Cov(x_i) = I_p$ and that the true value $\Gamma = I_p = (e_1, ..., e_p)^T$. Let $T(F_n)$ denote the sample mean vector and $S(F_n)$ the sample covariance matrix. If the fourth moments exist, then $\sqrt{n}\text{vec}(T(F_n), S(F_n) - I_p)$ has a joint limiting multivariate normal distribution (CLT). Write $\hat{\Gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_p)^T$ for the fastICA estimate of Γ. Write also

$$\mu_k = E[g(e_k^T x_i)], \quad \lambda_k = E[g(e_k^T x_i)e_k^T x_i]$$

and

$$\tau_k = E[g'(e_k^T x_i)e_k^T x_i], \quad \delta_k = E[g'(e_k^T x_i)],$$

$k = 1, ..., p$. The assumption that $\lambda_k \neq \delta_k$, $k = 1, ..., p-1$ is needed later. (If $g(z) = z^3$, for example, this assumption rules that only the component that is found last, may be normally distributed.) For the sample statistics

$$T_k = \frac{1}{n}\sum_{i=1}^{n}(g(e_k^T x_i) - \mu_k)x_i \text{ and } \hat{T}_k = \frac{1}{n}\sum_{i=1}^{n} g(\hat{\gamma}_k^T(x_i - T(F_n)))(x_i - T(F_n))$$

it is needed that

$$(5.2)\quad \sqrt{n}(\hat{T}_k - \lambda_k e_k) = \sqrt{n}T_k - \tau_k e_k e_k^T \sqrt{n}T(F_n) + \Delta_k \sqrt{n}(\hat{\gamma}_k - e_k) + o_P(1)$$

where $\Delta_k = E[g'(e_k^T x_i)x_i x_i^T]$, $k = 1, ..., p$. Again, if $g(z) = z^3$ and the sixth moments exist, then (5.2) is true and $\sqrt{n}(\hat{T}_k - \lambda_k e_k)$ has a limiting multinormal distribution. The estimating equations for the fastICA solution $\hat{\Gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_p)'$ are then given by

$$(5.3)\qquad \hat{T}_k = S(F_n)[\hat{\gamma}_1\hat{\gamma}_1^T + ... + \hat{\gamma}_k\hat{\gamma}_k^T]\hat{T}_k, \quad k = 1, ..., p.$$

If (5.2) is true and $U_k = \sum_{j=1}^k e_j e_j^T$ then

$$\begin{aligned}(I_p - U_k)\sqrt{n}(\hat{T}_k - \lambda_k e_k) = \lambda_k[\sqrt{n}(S(F_n) - I_p)e_k &+ \sum_{j=1}^k e_j e_k^T \sqrt{n}(\hat{\gamma}_j - e_j) \\ &+ \sqrt{n}(\hat{\gamma}_k - e_k)] + o_P(1)\end{aligned}$$

and the next result follows (Ilmonen et al., 2011a).

Theorem 4. *Let $x_1, ..., x_n$ be a random sample from the model (5.1) with $\Omega = I_p$, $E(x_i) = 0$, and $Cov(x_i) = I_p$. Let $\hat{\Gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_p)$ be the solution for estimating equations in (5.3), and let the algorithm be chosen such that $\hat{\Gamma} \to_P I_p$. Then, under the general assumptions,*

$$\begin{aligned}\sqrt{n}\hat{\gamma}_{kl} &= \frac{1}{\lambda_k - \delta_k}\left[e_l^T\sqrt{n}T_k - \lambda_k\sqrt{n}S(F_n)_{kl}\right] + o_P(1), \qquad \textit{for } l > k \\ \sqrt{n}(\hat{\gamma}_{kk} - 1) &= -\frac{1}{2}\sqrt{n}(S(F_n)_{kk} - 1) + o_P(1), \qquad \textit{and} \\ \sqrt{n}\hat{\gamma}_{kl} &= \sqrt{n}\hat{\gamma}_{lk} - \sqrt{n}S(F_n)_{kl} + o_P(1) \qquad \textit{for } l < k\end{aligned}$$

Theorem 4 implies that, if $\sqrt{n}(T_k - \lambda_k e_k)$, $k = 1, ..., p$, and $\sqrt{n}\mathrm{vec}(S(F_n) - I_p)$ have a joint limiting multivariate distribution, then also the limiting distribution of $\sqrt{n}\mathrm{vec}(\hat{\Gamma} - I_p)$ is multivariate normal. Interestingly enough, the limiting distribution of the estimated sources $\hat{\gamma}_1, ..., \hat{\gamma}_p$ depends on the order in which they are found. The limiting behavior of the diagonal elements of $\hat{\Gamma}$ does not depend on the choice of the function $g(z)$. The initial value for $\hat{\Gamma}$ in the fastICA algorithm fixes the asymptotic order of the sources.

Remark 1. Let $\kappa_k = (E[x_{ik}^4] - 1)/4$, $\sigma_k^2 = Var[g(e_k^T x_i)]$ and let

$$\alpha_k = \frac{\sigma_k^2 - \lambda_k^2}{(\lambda_k - \delta_k)^2}.$$

Now it follows from Theorem 4 that

$$ASV(\hat{\gamma}_k) = \sum_{j=1}^{k-1}(\alpha_j + 1)e_j e_j^T + \kappa_k e_k e_k^T + \alpha_k \sum_{l=k+1}^{p} e_l e_l^T.$$

5.4.1 Deflation Based FastICA reloaded

Surprisingly, the order in which the independent components are found has an effect on the performance of the fastICA algorithm. Nordhausen, Ilmonen, Mandal, Oja and Ollila (2011a) presented an improved algorithm to ensure that the components are found in an optimal order. The minimum distance index, see Chapter 5, Section 5.6, suggests that for $\Gamma = I_p$, the optimal performance is achieved when the sum of the variances of the off-diagonal elements of the estimator are minimized. It follows from Remark 1 that

$$\sum_{i \neq j} ASV(\hat{\gamma}_{ij}) = 2 \sum_{j=1}^{p} (p-i)(\alpha_i) e_j e_j^T + \frac{p(p-1)}{2},$$

and it is minimized when the α_i :s are in increasing order.

Nordhausen et al. (2011a) suggest using first any equivariant and consistent estimate $\hat{\Gamma}_0$ such that $S(\hat{\Gamma}_0 X) = I_p$. After that the estimates $\hat{z}_i = (\hat{\Gamma}_0(x_i - \bar{x}))$ are used to calculate the estimates $\hat{\alpha}_k$ (expected values are replaced by averages). Then a permutation matrix $\hat{P}$ is found such that for the permuted estimated sources, the $\hat{\alpha}_k$ are in increasing order. Matrix $\hat{P}\hat{\Gamma}_0 S(F_n)^{-1/2}$ is used as a new initial value of the algorithm. The performance of this new reloaded fastICA algorithm is better than the performance of the original fastICA algorithm and the new improved algorithm works well also in preventing algorithm failures. For details, see Nordhausen et al. (2011a).

5.5 Inference Based on Signed Ranks in Symmetric IC model

The idea of using rank-based test statistics for point estimators and confidence regions in the context of one sample and two sample location models was presented by Hodges and Lehmann (1963). Since that, ranks and signs and signed ranks have been used in several sophisticated multivariate test and estimation statistics. Signs and ranks are known to be very robust, and many popular robust methods are based on using them. For an excellent overview, see Oja (2010).

In *symmetric IC model* it is assumed that the p-variate vector

$$x = \Omega z + \mu \tag{5.4}$$

where Ω is a full-rank $p \times p$ mixing matrix, μ is a location vector and z is a p-variate vector with mutually independent and symmetrically distributed components. Nordhausen, Oja and Paindaveine (2009) presented tests for one sample location problem, $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$, in symmetric IC model. Assuming that $x = (x_1^T, x_2^T)^T$, Oja, Paindaveine and Taskinen (2011) considered the problem of testing whether the multivariate subvectors x_1 and x_2 are independent. Ilmonen and Paindaveine (2011) considered tests and estimates for a mixing matrix Ω. The tests presented in

Nordhausen et al. (2009); Oja et al. (2011); Ilmonen and Paindaveine (2011) are based on signed ranks of the estimated independent components. All the three papers rely heavily on the uniform local asymptotic normality (ULAN) property of symmetric IC models and the tests and estimation procedures are locally and asymptotically optimal in the Le Cam sense (Le Cam, 1986) at given densities.

5.5.1 ULAN

A sequence of statistical models $P_f^{(n)} = \{\mathrm{P}_{\vartheta,f}^{(n)} \mid \vartheta \in \theta \in \Theta \subseteq \mathbb{R}^k, f \in \mathcal{F}\}$ is *uniformly locally asymptotically normal (ULAN)* if for any $\vartheta_n = \vartheta + O(n^{-1/2})$ and any bounded sequence (τ_n), there exists a symmetric positive definite matrix $G_{\vartheta,f}$ such that, under $\mathrm{P}_{\vartheta,f}^{(n)}$ as $n \to \infty$,

$$\log(d\mathrm{P}_{\vartheta_n+n^{-1/2}\tau_n,f}^{(n)} / d\mathrm{P}_{\vartheta_n,f}^{(n)}) = \tau_n^T \Delta_{\vartheta_n,f}^{(n)} - \frac{1}{2}\tau_n^T G_{\vartheta,f}\tau_n + o_{\mathrm{P}}(1),$$

and that, still under $\mathrm{P}_{\vartheta,f}^{(n)}$, $\Delta_{\vartheta_n,f}^{(n)}$ is asymptotically normal with mean zero and covariance matrix $G_{\vartheta,f}$.

Such ULAN property allows to derive parametric efficiency bounds at f and to construct the corresponding parametrically optimal inference procedures for ϑ, see Le Cam (1986). When testing $\mathcal{H}_0 : \vartheta = \vartheta_0$ against $\mathcal{H}_a : \vartheta \neq \vartheta_0$, parametrically optimal tests reject the null at asymptotic level α whenever

$$\Delta_{\vartheta_0,f}^{(n)T} G_{\vartheta_0,f}^{-1} \Delta_{\vartheta_0,f}^{(n)} > \chi^2_{k,1-\alpha},$$

where $\chi^2_{k,1-\alpha}$ denotes the α-upper quantile of the χ^2_k distribution. Under sequences of alternatives of the form $\mathrm{P}_{\vartheta_0+n^{-1/2}\tau,f}^{(n)}$, these tests have the asymptotic power $\Psi_k(\chi^2_{k,1-\alpha}; \tau^T G_{\vartheta_0,f}\tau)$, where $\Psi_k(\,\cdot\,;\delta)$ stands for the cumulative distribution function of the non-central χ^2_k distribution with non-centrality parameter δ. This settles the parametrically optimal (at f) performance for hypothesis testing. As for point estimation, an estimator $\hat{\vartheta}$ is parametrically efficient at f iff

$$\sqrt{n}\,(\hat{\vartheta} - \vartheta) \xrightarrow{d} \mathcal{N}_r\big(0, G_{\vartheta,f}^{-1}\big).$$

The underlying density f is often unspecified in practice, which leads to considering the semiparametric model $\mathcal{P}^{(n)} = \cup_h \cup_{\vartheta\in\Theta} \{\mathrm{P}_{\vartheta,h}^{(n)}\}$. In $\mathcal{P}^{(n)}$, semiparametrically optimal (still at f) inference procedures are based on the efficient central sequence $\Delta_{\vartheta,f}^{*(n)}$ resulting from the original central sequence $\Delta_{\vartheta,f}^{(n)}$ by performing adequate tangent space projections; see Bickel, Klaassen, Ritov and Wellner (1993). Under $\mathrm{P}_{\vartheta,f}^{(n)}$, the efficient central sequence $\Delta_{\vartheta,f}^{*(n)}$ typically is still asymptotically normal with mean zero, but now with covariance matrix $G_{\vartheta,f}^{*}$ (the efficient information matrix at f). Semiparametrically optimal tests (at f) reject the null at asymptotic level α whenever

$$\Delta_{\vartheta_0,f}^{*(n)T} (G_{\vartheta_0,f}^{*})^{-1} \Delta_{\vartheta_0,f}^{*(n)} > \chi^2_{k,1-\alpha}.$$

They have asymptotic powers $\Psi_k(\chi^2_{k,1-\alpha}; \tau^T(G^*_{\vartheta_0,f})\tau)$ under the sequences of alternatives considered above. An estimator $\hat{\vartheta}$ is semiparametrically efficient at f if and only if

$$\sqrt{n}\,(\hat{\vartheta} - \vartheta) \xrightarrow{d} \mathcal{N}_r\big(0, (G^*_{\vartheta,f})^{-1}\big).$$

5.5.2 ULAN for symmetric IC models

Let $\mathcal{M}_t$ denote the set of mixing matrices Ω for which no ties occur in the permutation step of the mapping $\Omega \mapsto L = \Omega D_1^+ P D_2$ described in Section 5.1 and let $\mathcal{M}_1$ denote the corresponding set of matrices L. The parametrization based on standardizing the mixing matrix now leads to considering the model associated with

$$x = Lz + \mu, \tag{5.5}$$

where $\mu \in \mathbb{R}^p$, $L \in \mathcal{M}_1$, and z has independent and symmetrically distributed marginals (among which at most one is normally distributed) with common median zero. The resulting collection of densities (of the form $h(z) = \prod_{r=1}^p h_r(z_r)$, where h_r is the symmetric density of z_r) will be denoted as $\mathcal{F}$.

The hypothesis under which n mutually independent observations x_i, $i = 1, \ldots, n$ are obtained from (5.5), where z has density h, will be denoted as $\mathrm{P}^{(n)}_{\vartheta,h}$, with $\vartheta = (\mu^T, (\mathrm{vecd}^\circ L)^T)^T \in \Theta = \mathbb{R}^p \times \mathrm{vecd}^\circ(\mathcal{M}_1)$, or alternatively, as $\mathrm{P}^{(n)}_{\mu,L,h}$. This leads to the semiparametric model

$$\mathcal{P}^{(n)} = \cup_h \mathcal{P}^{(n)}_h = \cup_h \cup_{\vartheta\in\Theta} \{\mathrm{P}^{(n)}_{\vartheta,h}\}.$$

As usual, ULAN at some specific $h = f$ requires further technical assumptions: it is needed that f belongs to the collection $\mathcal{F}_{\mathrm{ulan}}$ of densities in $\mathcal{F}$ for which each f_r, $r = 1, \ldots, p$, is absolutely continuous, with a derivative f_r' that satisfies (below we let $\varphi_{f_r} = -f_r'/f_r$)

$$\sigma^2_{f_r} = \int_{-\infty}^{\infty} y^2 f_r(y)\,dy < \infty, \quad \mathcal{I}_{f_r} = \int_{-\infty}^{\infty} \varphi^2_{f_r}(y) f_r(y)\,dy < \infty,$$

and

$$\mathcal{J}_{f_r} = \int_{-\infty}^{\infty} y^2 \varphi^2_{f_r}(y) f_r(y)\,dy < \infty.$$

For any $f \in \mathcal{F}_{\mathrm{ulan}}$, let $\gamma_{rs}(f) = \mathcal{I}_{f_r}\sigma^2_{f_s}$, define the optimal p-variate location score function $\varphi_f : \mathbb{R}^p \to \mathbb{R}^p$ through $z = (z_1, \ldots, z_p)^T \mapsto \varphi_f(z) = (\varphi_{f_1}(z_1), \ldots, \varphi_{f_p}(z_p))^T$, and denote by $\mathcal{I}_f$ the diagonal matrix with diagonal entries $\mathcal{I}_{f_r}$, $r = 1, \ldots, p$. Further, define

$$C = \sum_{r=1}^{p} \sum_{s=1}^{p-1} (e_r e_r^T \otimes u_s e^T_{s+\delta_{s\geq r}}),$$

where e_r and u_r stand for the rth vectors of the canonical basis of $\mathbb{R}^p$ and $\mathbb{R}^{p-1}$, respectively, and $\delta_{s\geq r}$ is equal to one if $s \geq r$ and to zero otherwise.

Then the parametric model $\mathcal{P}_f^{(n)}$ is ULAN for any fixed $f \in \mathcal{F}_{\text{ulan}}$ (Oja et al., 2011; Ilmonen and Paindaveine, 2011), with central sequence

$$\Delta_{\vartheta,f}^{(n)} = \begin{pmatrix} \Delta_{\vartheta,f;1}^{(n)} \\ \Delta_{\vartheta,f;2}^{(n)} \end{pmatrix} = \begin{pmatrix} n^{-1/2}(L^{-1})^T \sum_{i=1}^n \varphi_f(z_i) \\ n^{-1/2} C(I_p \otimes L^{-1})^T \sum_{i=1}^n \text{vec}(\varphi_f(z_i) z_i^T - I_p) \end{pmatrix},$$

where $z_i = z_i(\vartheta) = L^{-1}(x_i - \mu)$, and full-rank information matrix

$$G_{L,f} = \begin{pmatrix} G_{L,f;1} & 0 \\ 0 & G_{L,f;2} \end{pmatrix},$$

where $G_{L,f;1} = (L^{-1})^T \mathcal{I}_f L^{-1}$ and

$$\begin{aligned} G_{L,f;2} &= C(I_p \otimes L^{-1})^T \Bigg[\sum_{r=1}^p (\mathcal{J}_{f_r} - 1)(e_r e_r^T \otimes e_r e_r^T) \\ &+ \sum_{r,s=1, r\neq s}^p \big(\gamma_{sr}(f)(e_r e_r^T \otimes e_s e_s^T) + (e_r e_s^T \otimes e_s e_r^T)\big) \Bigg] (I_p \otimes L^{-1}) C^T. \end{aligned}$$

5.5.3 Optimal signed-rank inference in symmetric IC models

Ilmonen and Paindaveine (2011) considered the problem of testing $\mathcal{H}_0 : L = L_0$ against $\mathcal{H}_a : L \neq L_0$, where $L_0 \in \mathcal{M}_1$ is fixed. As already mentioned, semiparametrically optimal procedures are based on the efficient central sequence $\Delta_{\vartheta,f}^*$. Classically, $\Delta_{\vartheta,f}^*$ is obtained by performing tangent space computations. When, however, the semiparametric model at hand enjoys a strong invariance structure, the efficient central sequence $\Delta_{\vartheta,f}^*$ can alternatively be obtained by conditioning the original central sequence $\Delta_{\vartheta,f}$ with respect to the corresponding maximal invariant; see Hallin and Werker (2003).

In the context of symmetric IC models, this maximal invariant is given by

$$(S_1(\vartheta), \ldots, S_n(\vartheta), R_1^+(\vartheta), \ldots, R_n^+(\vartheta)),$$

with $S_i(\vartheta) = (S_{i1}(\vartheta), \ldots, S_{ip}(\vartheta))^T$ and $R_i^+(\vartheta) = (R_{i1}^+(\vartheta), \ldots, R_{ip}^+(\vartheta))^T$, where $S_{ir}(\vartheta)$ is the sign of $z_{ir}(\vartheta) = (L^{-1}(x_i - \mu))_r$ and $R_{ir}^+(\vartheta)$ is the rank of $|z_{ir}(\vartheta)|$ among $|z_{1r}(\vartheta)|, \ldots, |z_{nr}(\vartheta)|$. This is what leads to considering signed-rank procedures when performing inference on L in the present context.

Let $\hat{\vartheta}_{0\#} = (\hat{\mu}_\#^T, (\text{vecd}^\circ L_0)^T)^T$ denote a root-n consistent (under the null) and locally asymptotically discrete sequence of estimators $\hat{\mu}_\#$ for μ. (An estimate $(\check{\vartheta}_\#^n)$ is said to be locally asymptotically discrete if the number of possible values of $\check{\vartheta}_\#^n$ in balls with $O(n^{-1/2})$ radius centered at ϑ is bounded as $n \to \infty$. For examples of such estimates in this present setup,

see Ilmonen and Paindaveine (2011).) The nonparametric counterpart of the parametrically optimal (at f) test statistic is given by (Ilmonen and Paindaveine, 2011)

$$\underline{Q}_f = (\underline{\Delta}^*_{\hat{\vartheta}_{0\#},f;2})^T \, (G^*_{L_0,f;2})^{-1} \underline{\Delta}^*_{\hat{\vartheta}_{0\#},f;2},$$

where

$$\begin{aligned}\underline{\Delta}^*_{\vartheta,f;2} &= C(I_p \otimes L^{-1})^T \mathrm{vec}\left[\mathrm{odiag}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(S_i(\vartheta) \odot \varphi_f\Big(F_+^{-1}\Big(\frac{R_i^+(\vartheta)}{n+1}\Big)\Big)\Big)\right. \\ &\times \left.\Big(S_i(\vartheta) \odot F_+^{-1}\Big(\frac{R_i^+(\vartheta)}{n+1}\Big)\Big)^T\Big)\right]\end{aligned}$$

(here odiag is the operator that replaces diagonal entries with zeroes) and

$$\begin{aligned}G^*_{L,f;2} &= C(I_p \otimes L^{-1})^T\left[\sum_{r,s=1,r\neq s}^{p}\big(\gamma_{sr}(f)(e_r e_r^T \otimes e_s e_s^T)\right. \\ &+ \left.(e_r e_s^T \otimes e_s e_r^T)\big)\right](I_p \otimes L^{-1})C^T.\end{aligned}$$

The resulting signed-rank tests, that reject $\mathcal{H}_0 : L = L_0$ at asymptotic level α whenever $\underline{Q}_f > \chi^2_{p(p-1),1-\alpha}$, are semiparametrically optimal (most stringent, see Le Cam (1986)) at f. Since they are signed-rank tests, they, however, remain valid in the sense that they meet asymptotically the level constraint, under a very broad class of densities h.

Ilmonen and Paindaveine (2011) also provided a one step point estimator $\hat{L}$. Let $\tilde{\vartheta}_\# = (\tilde{\mu}^T_\#, (\mathrm{vecd}^\circ \tilde{L}_\#)^T)^T$ denote a root-n consistent and locally asymptotically discrete preliminary estimator. (Several such practical estimators exist, see Ilmonen and Paindaveine (2011).) Let

$$\begin{aligned}G^*_{L,f,h;2} &= C(I_p \otimes L^{-1})^T\left[\sum_{r,s=1,r\neq s}^{p}\big(\gamma_{sr}(f,h)(e_r e_r^T \otimes e_s e_s^T)\right. \\ &+ \left.\rho_{rs}(f,h)(e_r e_s^T \otimes e_s e_r^T)\big)\right](I_p \otimes L^{-1})C^T,\end{aligned}$$

where

$$\gamma_{rs}(f,h) = \int_0^1 \varphi_{f_r}(F_r^{-1}(u))\,\varphi_{h_r}(H_r^{-1}(u))\,du \times \int_0^1 F_s^{-1}(u)\,H_s^{-1}(u)\,du$$

and

$$\rho_{rs}(f,h) = \int_0^1 F_r^{-1}(u)\,\varphi_{h_r}(H_r^{-1}(u))\,du \times \int_0^1 \varphi_{f_s}(F_s^{-1}(u))\,H_s^{-1}(u)\,du$$

and let $\hat{G}^*_{\tilde{L}_\#,f;2}$ denote an estimate of $G^*_{L,f,h;2}$ formed by plugging in preliminary a estimator $\tilde{\vartheta}_\#$ and estimators $\hat{\gamma}_{rs\#}(f)$ and $\hat{\rho}_{rs\#}(f)$ that (i) are locally asymptotically discrete and (ii) satisfy $\hat{\gamma}_{rs\#}(f) = \gamma_{rs}(f,h) + o_{\mathrm{P}}(1)$ and $\hat{\rho}_{rs\#}(f) = \rho_{rs}(f,h) + o_{\mathrm{P}}(1)$ as $n \to \infty$, under $\cup_{\vartheta\in\Theta} \cup_{h\in\mathcal{F}_{\mathrm{ulan}}} \{\mathrm{P}^{(n)}_{\vartheta,h}\}$.

Let

$$\mathrm{vecd}^\circ \underline{\hat{L}}_{f\#} = (\mathrm{vecd}^\circ \tilde{L}_\#) + n^{-1/2} (\hat{G}^*_{\tilde{L}_\#,f;2})^{-1} \underline{\Delta}^*_{\tilde{\vartheta}_\#,f;2},$$

where $\hat{G}^*_{\tilde{L}_\#,f;2}$ is the consistent estimate of $G^*_{L,f,h;2}$ just defined. Ilmonen and Paindaveine (2011) showed that

$$\sqrt{n}\ \mathrm{vecd}^\circ(\underline{\hat{L}}_{f\#} - L) \xrightarrow{\mathcal{L}} \mathcal{N}_{p(p-1)}\big(0, (\Gamma^*_{L,f;2})^{-1}\big)$$

as $n \to \infty$, under $\cup_{\mu\in\mathbb{R}^p} \{\mathrm{P}^{(n)}_{\mu,L,f}\}$.

5.6 Performance Indices

Due to the vast amount of different ICA estimates and algorithms, asymptotic as well as finite sample criteria are needed for their comparisons. While asymptotic results (convergence, asymptotic normality, etc.) are often missing, several finite-sample performance indices have been proposed in the literature to compare different estimates in simulation studies. First, one can compare the true sources z (which are of course known in the simulations) and the estimated sources $\hat{z} = \hat{\Gamma}x$. Second, one can measure the closeness of the true unmixing matrix Ω^{-1} (used in the simulations) and the estimated unmixing matrix $\hat{\Gamma}$. In both cases the problem is that the order, signs and scales of the rows of the estimated unmixing matrix may not match as $\hat{\Gamma}$ is typically not an estimate of Ω^{-1}. For a good estimate, the gain matrix $\hat{G} = \hat{\Gamma}\Omega$ is close to a matrix PJD, where P is a permutation matrix, J is a sign-change matrix, and D is a scaling matrix.

Normalized versions of $\hat{G}$ are used in the performance index constructions. One of the most popular indices, the Amari index (Amari, Cichocki and Yang, 1996) for example uses

$$\frac{1}{p}\left[\sum_{i=1}^{p} \frac{\sum_{j=1}^{p} |\hat{G}_{ij}|}{\max_j |\hat{G}_{ij}|} + \sum_{j=1}^{p} \frac{\sum_{i=1}^{p} |\hat{G}_{ij}|}{\max_i |\hat{G}_{ij}|}\right] - 2.$$

The smaller is the value of the index, the better is the estimate, and value 0 corresponds to perfect separation. Also the intersymbol interference (ISI) (Moreau and Macchi, 1994) is based on both row-wise and column-wise standardizations. The inference-to-signal ratio (ISR) (Ollila, 2010) and interchannel inference (ICI) (Douglas, 2007) use only row-wise standardization and

$$\sum_{i=1}^{p} \left(\sum_{j=1}^{p} \frac{\sum_{j=1}^{p} \hat{G}_{ij}^2}{\max_j \hat{G}_{ij}^2} - 1 \right).$$

Theis, Lang and Puntonet (2004) proposed an index called the generalized crosstalking error which is the shortest distance between the mixing matrix Ω and the set of estimates equivalent to $\hat{\Gamma}^{-1}$. Chen and Bickel (2006) compute the norm $||\hat{\Gamma}\Omega - I_p||$ after suitable rescaling and permutation of $\hat{\Gamma}$ and Ω.

Ilmonen et al. (2010b) introduced a new performance index based on the use of $\hat{G} = \hat{\Gamma}\Omega$. The index finds the shortest distance (using Frobenius norm) between the identity matrix and the set of matrices equivalent to the gain matrix $\hat{\Gamma}\Omega$.

Let A denote a $p \times p$ matrix. The shortest squared distance (divided by $p-1$) between the set $\{CA \mid C \in \mathcal{C}\}$ of equivalent matrices (to A) and I_p is given by

$$D^2(A) = \frac{1}{p-1} \inf_{C \in \mathcal{C}} \|CA - I_p\|^2$$

where $\|\cdot\|$ is the matrix (Frobenius) norm. For the following result, see Ilmonen et al. (2011a).

Theorem 5. *Let A be any $p \times p$ matrix having at least one nonzero element in each row. The shortest squared distance $D^2(A)$ fulfils the following four conditions:*

1. $1 \geq D^2(A) \geq 0$,
2. $D^2(A) = 0$ *if and only if* $A \sim I_p$,
3. $D^2(A) = 1$ *if and only if* $A \sim 1_p a^T$ *for some p-vector a, and*
4. *the function* $c \to D^2(I_p + c\ \mathrm{odiag(A)})$ *is increasing in* $c \in [0,1]$ *for all matrices A such that* $A_{ij}^2 \leq 1$, $i \neq j$.

Let $X = [x_1 ....x_n]$, where $x_1, ..., x_n$ is a random sample from a distribution F_x, where x obeys the IC model (5.1) with a mixing matrix Ω. Let $\Gamma(F)$ be an IC functional. Then clearly $D^2(\Gamma(F_x)\Omega) = 0$. If F_n is the empirical cumulative distribution function based on X then

$$\hat{\Gamma} = \hat{\Gamma}(X) = \Gamma(F_n)$$

is the unmixing matrix estimate based on the functional $\Gamma(F_x)$.

The shortest distance between the identity matrix and the set of matrices $\{C\hat{\Gamma}\Omega : C \in \mathcal{C}\}$ equivalent to the gain matrix $\hat{G} = \hat{\Gamma}\Omega$ is as given in the following definition.

Definition 10. The *minimum distance index* for $\hat{\Gamma}$ is

$$\hat{D} = D(\hat{\Gamma}\Omega) = \frac{1}{\sqrt{p-1}} \inf_{C \in \mathcal{C}} \|C\hat{\Gamma}\Omega - I_p\|.$$

It follows directly from Theorem 5, that $1 \geq \hat{D} \geq 0$, and $\hat{D} = 0$ if and only if $\hat{\Gamma} \sim \Omega^{-1}$. The worst case with $\hat{D} = 1$ is obtained if all the row vectors of $\hat{\Gamma}\Omega$ point to the same direction. Thus the value of the minimum distance index is easy to interpret. Note that $D(\hat{\Gamma}\Omega) = D(C\hat{\Gamma}\Omega)$ for all $C \in \mathcal{C}$. Also, if

$$x_i = \Omega z_i \ \text{ and } \ x_i^* = (A\Omega)z_i = \Omega^* z_i,$$

and $\hat{\Gamma}^*$ is calculated from $X^* = [x_1^*, ..., x_n^*]$, then $D(\hat{\Gamma}^*\Omega^*) = D(\hat{\Gamma}\Omega)$. Thus the minimum distance index provides a fair comparison for different IC functionals. Note also the nice and natural behavior described in Theorem 5, condition 4.

Note that the generalized crosstalking error in Theis et al. (2004) is defined as

$$E(\Omega, \hat{\Gamma}) = \inf_{C \in \mathcal{C}} \|\Omega - \hat{\Gamma}^{-1}C\|$$

where $\|\cdot\|$ denotes any matrix norm. Clearly, $E(\Omega, \hat{\Gamma}) = E(\Omega, C\hat{\Gamma})$ for all $C \in \mathcal{C}$, but $E(\Omega^*, \hat{\Gamma}^*) = E(\Omega, \hat{\Gamma})$ is not necessarily true. If the Frobenius norm is used, the new index may be seen as a standardized version of the generalized crosstalking error as

$$\hat{D} = \inf_{C \in \mathcal{C}} \|C^{-1}\hat{\Gamma}\left(\Omega - \hat{\Gamma}^{-1}C\right)\|.$$

Note that, unlike the minimum distance index, the values of the Amari index for $\hat{\Gamma}\Omega$ and $D\hat{\Gamma}\Omega$ (with a diagonal matrix D) may differ.

The limiting behavior of the value of the minimum distance index depends on the limiting behavior of the used estimate. The following theorem was given in Ilmonen et al. (2011a).

Theorem 6. *Assume that the model is fixed such that* $\Gamma(F_x) = \Omega = I_p$ *and that* $\sqrt{n}\ \mathrm{vec}(\hat{\Gamma} - I_p) \to_d N_{p^2}(0, \Sigma)$. *Then*

$$n\hat{D}^2 = \frac{n}{p-1}\|\mathrm{odiag}(\hat{\Gamma})\|^2 + \mathrm{o_P}(1)$$

and the limiting distribution of $n\hat{D}^2$ *is that of* $(p-1)^{-1}\sum_{i=1}^k \delta_i \chi_i^2$ *where* $\chi_1^2, ..., \chi_k^2$ *are independent chi squared variables with one degree of freedom, and* $\delta_1, ..., \delta_k$ *are the* k *nonzero eigenvalues (including all algebraic multiplicities) of*

$$ASCOV(\sqrt{n}\ vec(\mathrm{odiag}(\hat{\Gamma}))) = (\mathrm{I_{p^2}} - \mathrm{D_{p,p}})\Sigma(\mathrm{I_{p^2}} - \mathrm{D_{p,p}}),$$

with $D_{p,p} = \sum_{i=1}^p (e_i e_i^T) \otimes (e_i e_i^T)$.

It is also important to note that similar simple asymptotical results for the Amari index cannot be found since it is based on the use of l_1 norms.

6 Data Example

To demonstrate the use of ICS transformation in practise, we performed an ICS transformation to a real data. The data set we used was part of a large data set of height measurements that were collected retrospectively from health centers and schools for construction of the Finnish growth charts. The used data set comprised 525 boys and 571 girls, fullterm, healthy singletons, followed until approximately age 19, with measurements from three to 44 occasions. The original data set is explained in detail in Pere (2000).

We used the original observations to estimate each individual growth curve from birth to age 19 by fitting splines. We excluded the individuals that did not have enough measurements for fitting the splines. After that we had 829 (481 boys and 348 girls) estimated height curves. In our analysis, we used measurements (based on estimated curves) at ages 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18 years. Thus we had 13 dimensional sample with 829 observations.

Before going to ICS transformations, we first used PCA for dimension reduction. (That is often done also in ICA, see Hyvärinen et al. (2001).) The first principal component explained 77 %, the second 17 % and the third 4 % of the variance of the data. Thus the first, second and third principal component together already explained 98 % of the variance and we reduced the dimension of the data to three. We calculated the mean curve and the first three principal component curves (i.e. the three first column vectors of the estimated mixing matrix/the inverse of the loadings matrix). Mean curve of the estimated data points and the three first principal component curves are presented in Figure 6.1.

After transformation to the principal components and dimension reduction, we performed FOBI-transformation i.e. ICS transformation based on the use of the covariance matrix and the scatter matrix based on fourth moments. As in PCA, we also calculated the three FOBI component curves. Mean curve of the estimated data points and the three FOBI component curves (the three first column vectors of the mixing matrix estimate) are presented in Figure 6.2. To compare with robust ICS transformation, we also performed ICS transformation based on the use of Dümbgen shape matrix and Hallin-Paindaveine shape matrix and calculated the corresponding curves. Mean curve of the estimated data points and the three ICS component curves based on Dümbgen shape matrix and Hallin-Paindaveine shape matrix are presented in Figure 6.3.

The shapes of the three first principal component curves, the three FOBI component curves and the three robust ICS component curves represent different growth profiles. Some of the curves put more emphasis on early

growth and some on late growth. The first principal component curve puts emphasis on overall growth (shape of the curve is similar to the mean curve), the second on late growth, and the third on growth around age 14. The first FOBI component curve puts emphasis on growth peak around age 14, the second on overall growth, and the third on late growth. The robust ICS component curves are similar to the FOBI component curves.

To illustrate the usage of the PCA, FOBI and robust ICS component curves on individual level, we randomly picked one boy and one girl and presented their estimated height growth curves as sums of their principal component curves, FOBI component curves and robust ICS component curves (based on Dümbgen shape matrix and Hallin-Paindaveine shape matrix). The estimated growth curve of one randomly chosen boy in terms of principal components, FOBI components and robust ICS components is presented in Figures 6.4, 6.5 and 6.6 respectively and the estimated growth curve of one randomly chosen girl in terms of principal components, FOBI components and robust ICS components is presented in Figures 6.7, 6.8 and 6.9. All these three methods seem to work very well also on individual level. In these examples, the curves based on three FOBI components and three robust ICS components are very close to the curves estimated using the splines, whereas with principal components only two components are needed for being very close to the curve based on splines.

We also examined how well PCA, FOBI and robust ICS work in separating sexes. Scatter plot after PCA is presented in Figure 6.10, scatter plot after FOBI transformation is presented in Figure 6.11 and scatter plot after robust ICS transformation is presented in Figure 6.12. In separating sexes, robust ICS transformation works better than non-robust FOBI-transformation, but both of these ICS transformations separate sexes better than PCA. This is not surprising since ICS transformations are often used to find hidden structures of the data, even when PCA fails to discover them. The first FOBI component (and the first robust ICS component) is a mixture of two distributions with the same location, but different scales (high kurtosis). This component measuring the 'spurt' around the age 14 is doing well in separating the sexes. The second component is similar for boys and girls, and separates tall and short individuals. The third component is a mixture of two distributions with the same scale but different location (low kurtosis) and separates boys and girls. Boys grow later than girls!

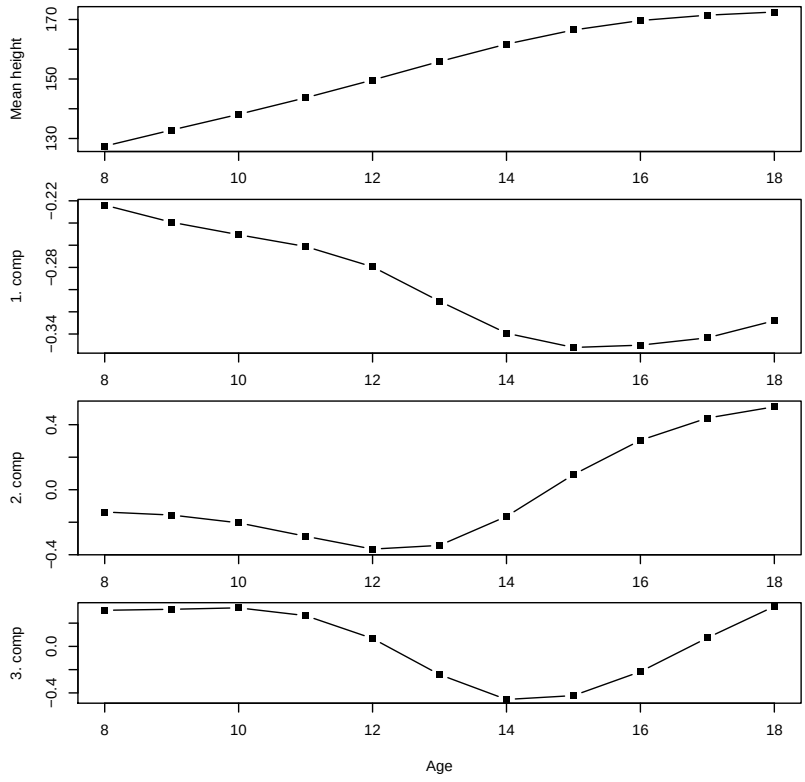

Figure 6.1: Mean curve of the estimated data points and the three first principal component curves.

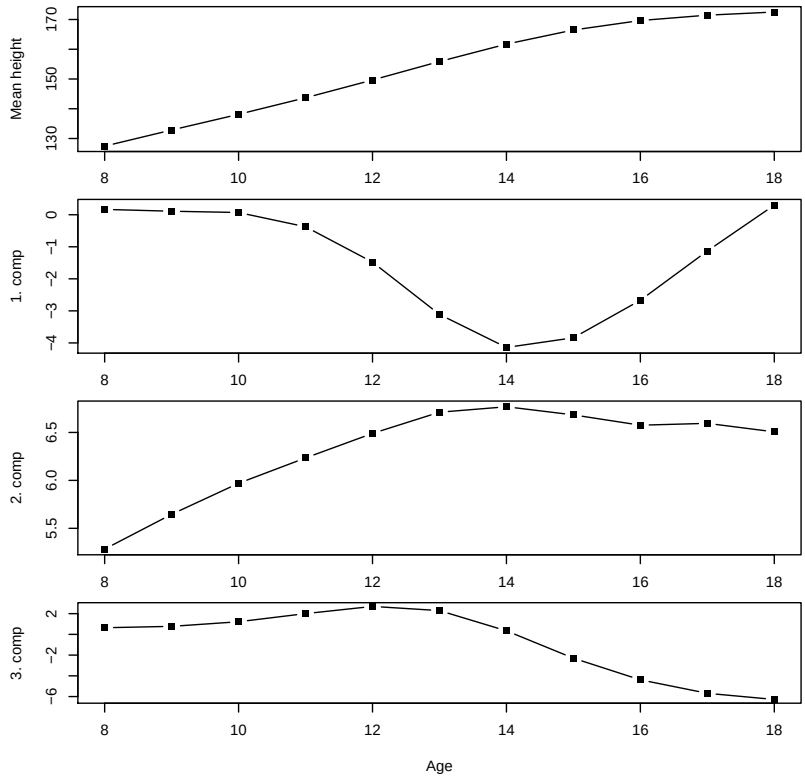

Figure 6.2: Mean curve of the estimated data points and the three first FOBI component curves.

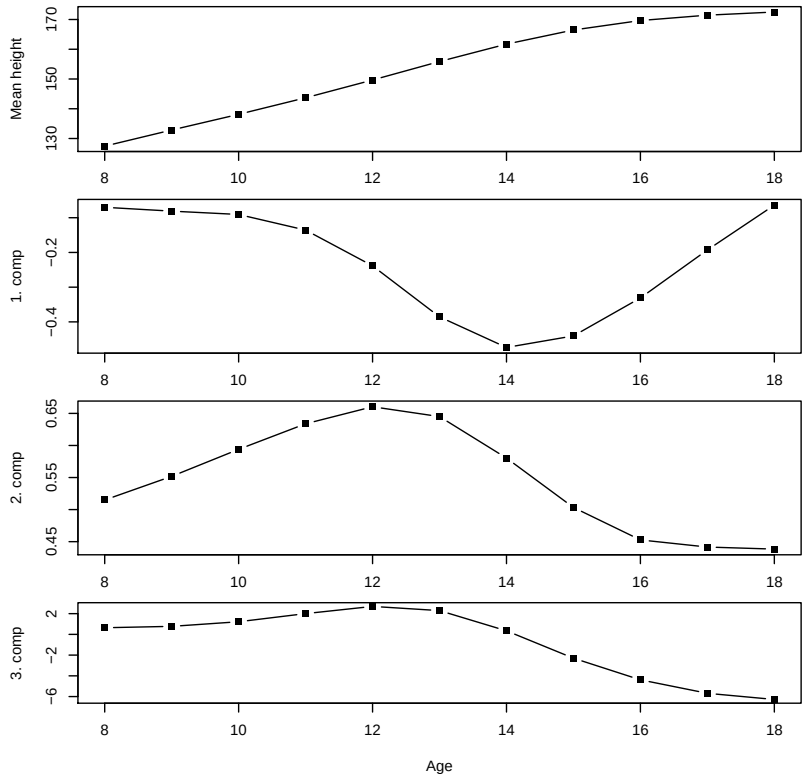

Figure 6.3: Mean curve of the estimated data points and the three first robust ICS component curves.

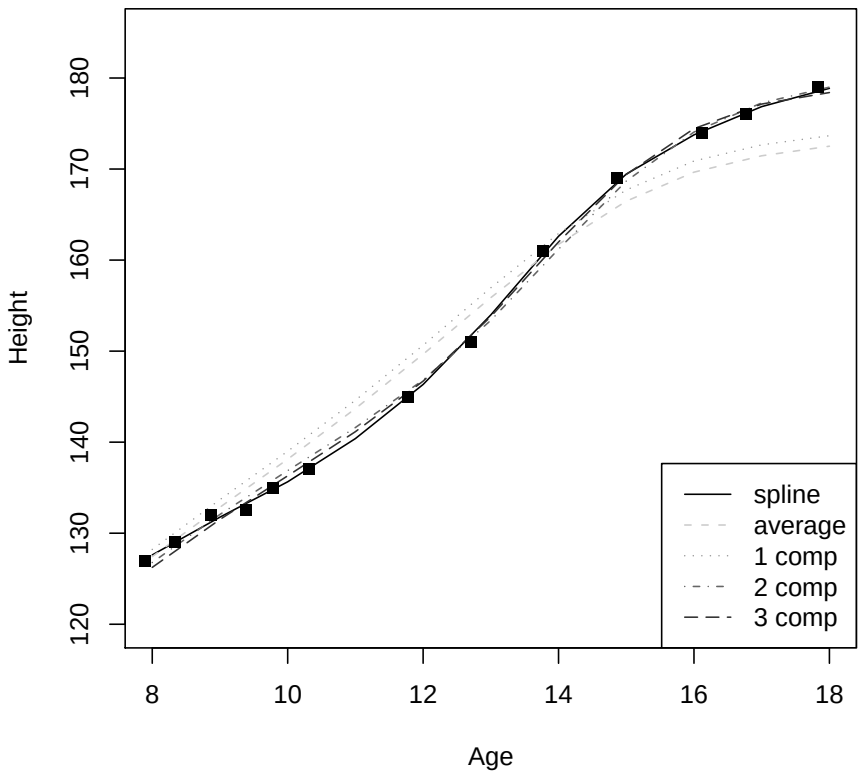

Figure 6.4: Estimated growth curve of one randomly chosen boy. Black squares are the original measurements and black solid line is the estimated growth curve based on fitting splines. Grey curves are the average height curve, the estimated curve based on the first principal component, the estimated curve based on the first and the second principal component, and the estimated curve based on the first, the second and the third principal component.

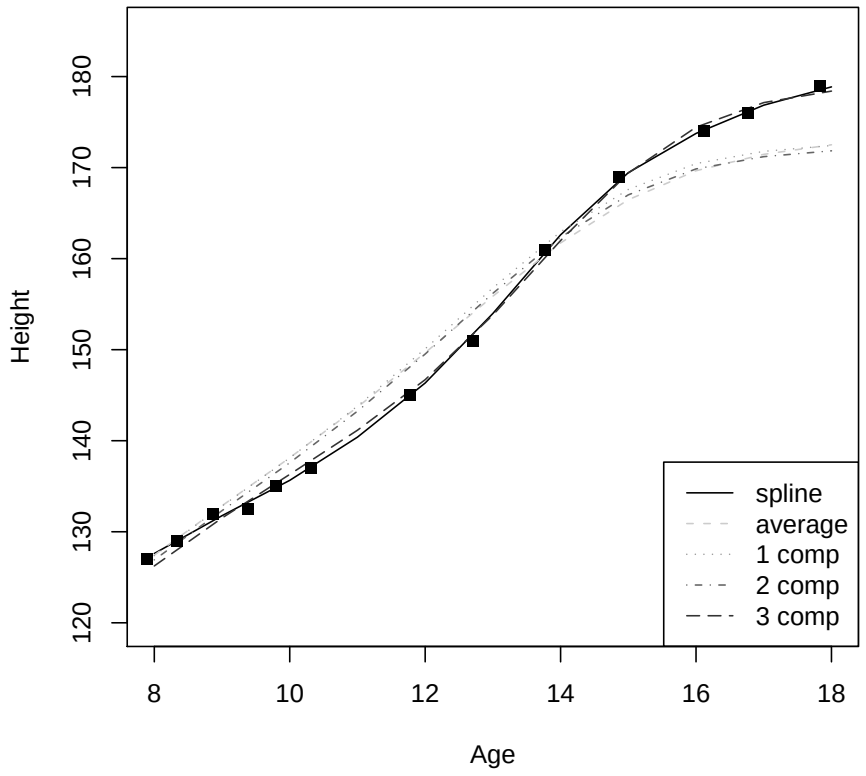

Figure 6.5: Estimated growth curve of one randomly chosen boy. Black squares are the original measurements and black solid line is the estimated growth curve based on fitting splines. Grey curves are the average height curve, the estimated curve based on the first FOBI component, the estimated curve based on the first and the second FOBI component, and the estimated curve based on the first, the second and the third FOBI component.

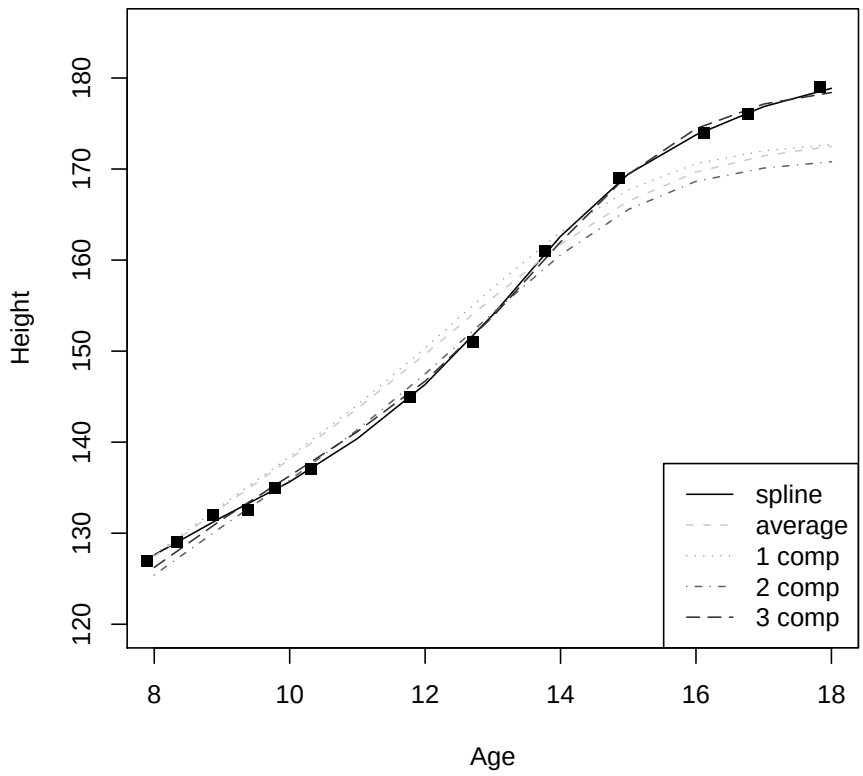

Figure 6.6: Estimated growth curve of one randomly chosen boy. Black squares are the original measurements and black solid line is the estimated growth curve based on fitting splines. Grey curves are the average height curve, the estimated curve based on the first robust ICS component, the estimated curve based on the first and the second robust ICS component, and the estimated curve based on the first, the second and the third robust ICS component.

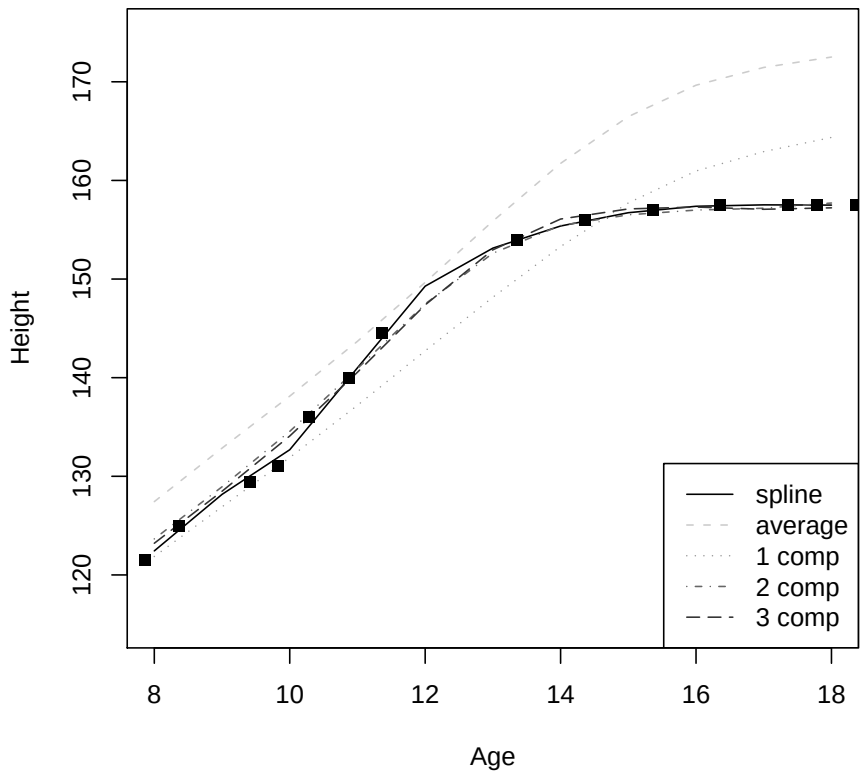

Figure 6.7: Estimated growth curve of one randomly chosen girl. Black squares are the original measurements and black solid line is the estimated growth curve based on fitting splines. Grey curves are the average height curve, the estimated curve based on the first principal component, the estimated curve based on the first and the second principal component, and the estimated curve based on the first, the second and the third principal component.

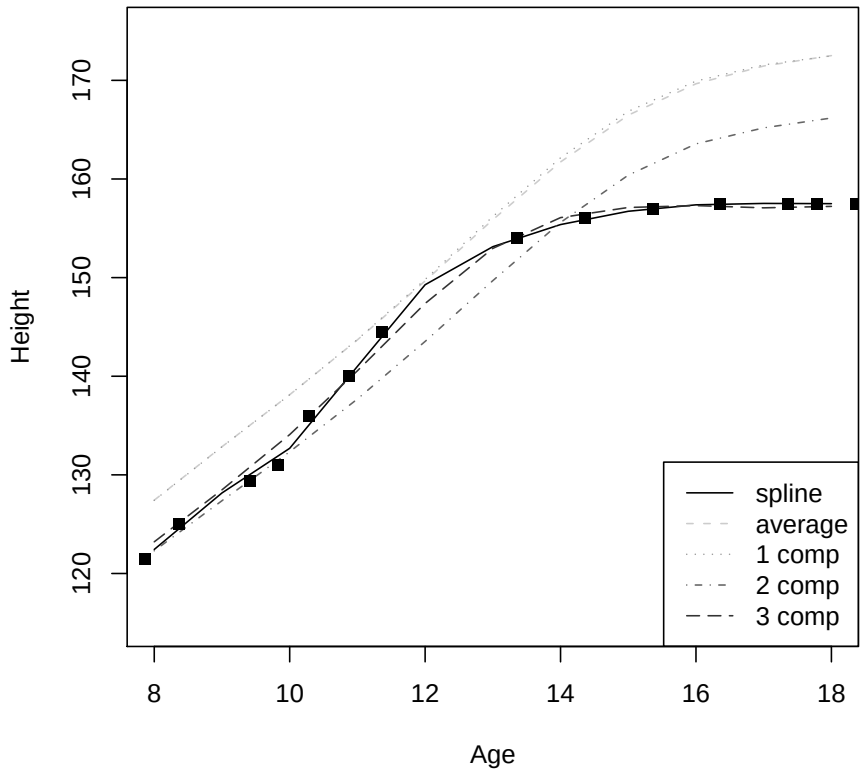

Figure 6.8: Estimated growth curve of one randomly chosen girl. Black squares are the original measurements and black solid line is the estimated growth curve based on fitting splines. Grey curves are the average height curve, the estimated curve based on the first FOBI component, the estimated curve based on the first and the second FOBI component, and the estimated curve based on the first, the second and the third FOBI component.

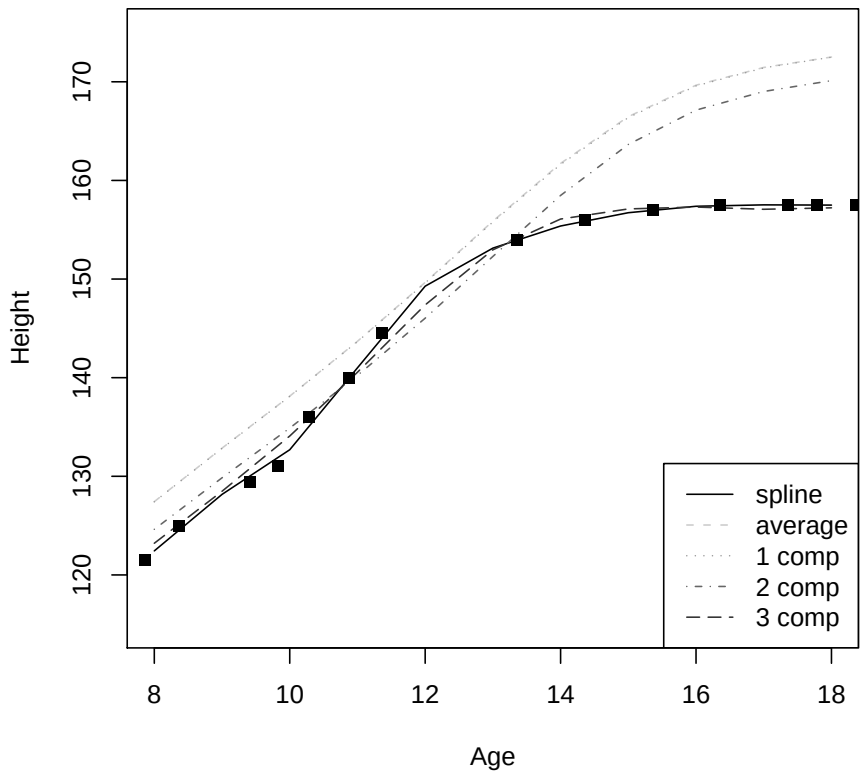

Figure 6.9: Estimated growth curve of one randomly chosen boy. Black squares are the original measurements and black solid line is the estimated growth curve based on fitting splines. Grey curves are the average height curve, the estimated curve based on the first robust ICS component, the estimated curve based on the first and the second robust ICS component, and the estimated curve based on the first, the second and the third robust ICS component.

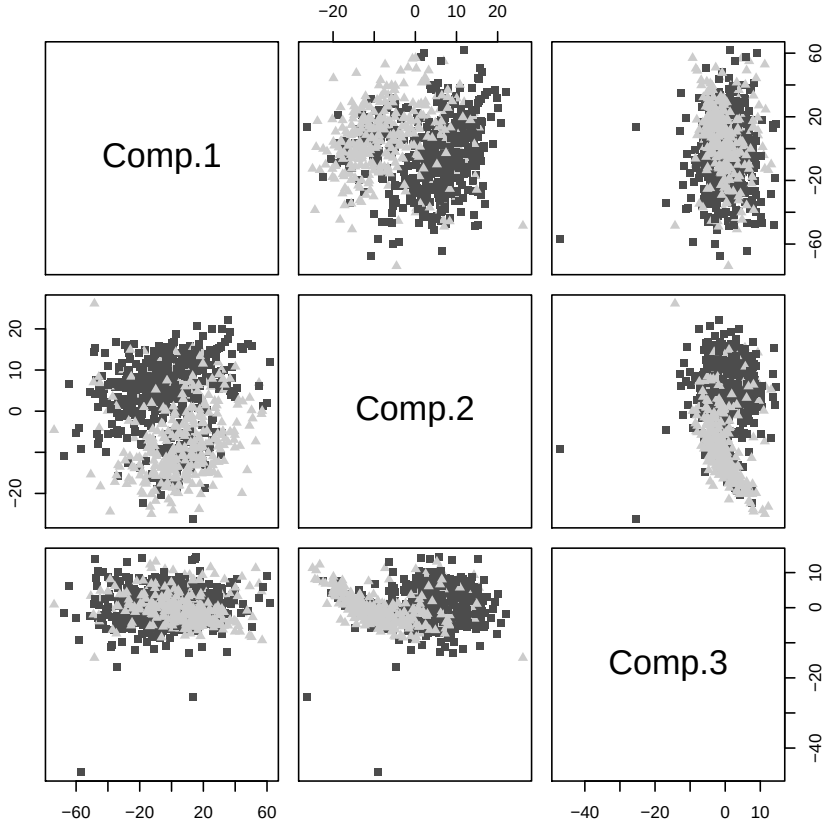

Figure 6.10: Scatter plot after PCA. Dark grey squares are used for the boys and light grey triangles for the girls.

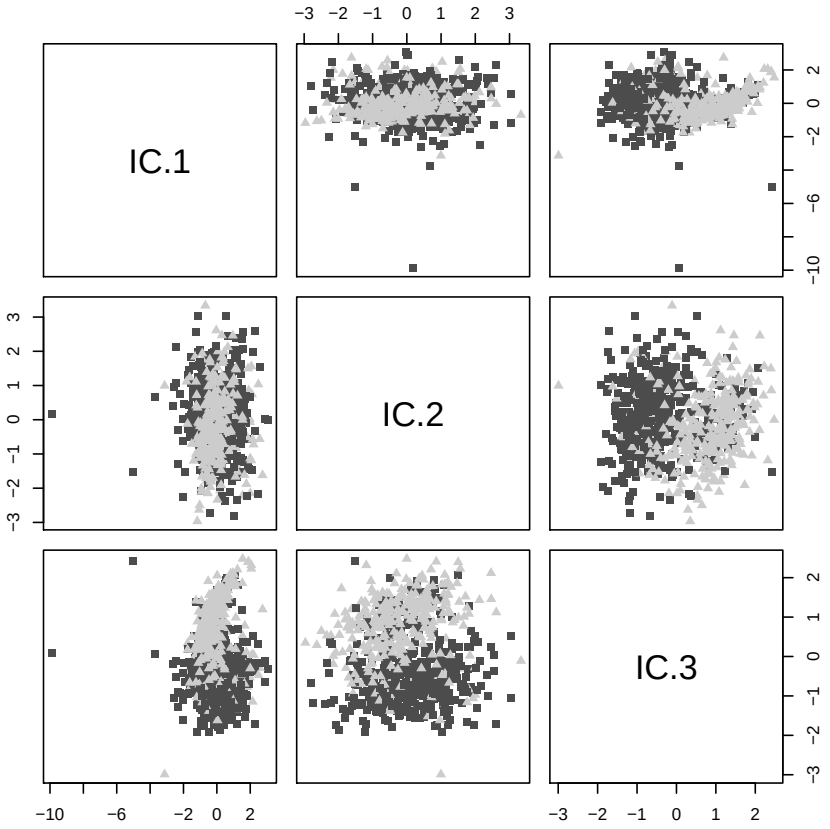

Figure 6.11: Scatter plot after FOBI transformation. Dark grey squares are used for the boys and light grey triangles for the girls.

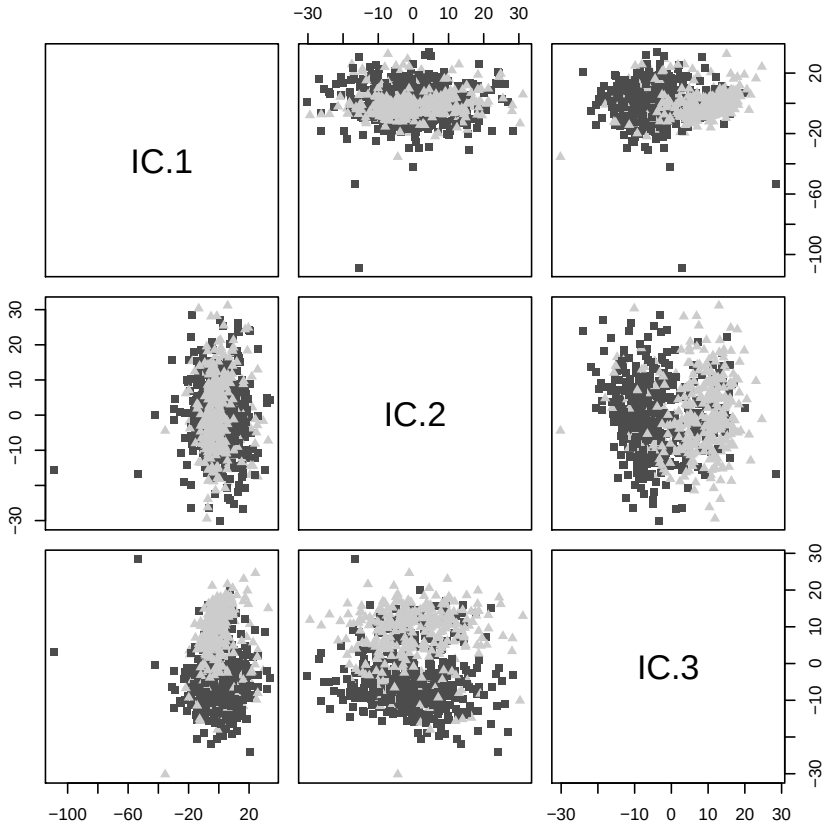

Figure 6.12: Scatter plot after robust ICS transformation. Dark grey squares are used for the boys and light grey triangles for the girls.

Summaries of Original Publications

I. Equivariance and invariance issues arise in multivariate statistical analysis. Often statistical procedures have to be modified to obtain an affine equivariant or invariant version. This is usually done by preprocessing the data, e.g., by standardizing the multivariate data or by transforming the data to an invariant coordinate system.

   In the article *On invariant coordinate system (ICS) functionals* (P. Ilmonen, H. Oja and R. Serfling), standardization of multivariate distributions, and characteristics of invariant coordinate system (ICS) functionals and statistics are examined. Also, invariances up to some groups of transformations are discussed. Constructions of different ICS functionals are addressed. In particular, the construction based on the use of two scatter matrix functionals presented by Tyler et al. (2009), and constructions based on the approach presented by Chaudhuri and Sengupta (1993) and related approaches, are examined. Several applications of ICS functionals are also discussed.

II. In the article *Characteristics of multivariate distributions and the invariant coordinate system* (P. Ilmonen, J. Nevalainen and H. Oja), a semiparametric multivariate location-scatter model, where p-variate vector

$$x = \Omega z + \mu,$$

   where μ is a location vector, Ω is a full rank $p \times p$ mixing matrix, and z is a 'standardized' p-variate vector, is considered. The model is fixed using simultaneously two location vectors and two scatter matrices. The approach using location and scatter functionals based on the first four moments serves as main example. The four functionals yield in a natural way the corresponding skewness, kurtosis and unmixing matrix functionals. Affine transformation based on the unmixing matrix transforms the variable to an invariant coordinate system. The limiting properties of the skewness, kurtosis, and unmixing matrix estimates are derived under general conditions. Related statistical inference problems, the role of the sample statistics in testing for normality and ellipticity, and connections to invariant coordinate selection and independent component analysis are discussed.

III. In the independent component (IC) model it is assumed that the p-

variate vector

$$x = \Omega z,$$

where Ω is a full rank $p \times p$ mixing matrix and z is a p-variate vector with mutually independent components.

In the independent component analysis (ICA) the aim is to find an estimate for an unmixing matrix Γ such that Γx has independent components.

Deflation-based FastICA, where independent components are extracted one-by-one, is among the most popular methods for estimating an unmixing matrix Γ. In the literature, it is often seen rather as an algorithm than an estimator related to a certain objective function, and only recently its statistical properties has been derived. One of the recent findings is that the order, in which the independent components are extracted in practice, has a strong effect on the performance of the estimator. In the article *Deflation-based fastICA reloaded* (K. Nordhausen, P. Ilmonen, A. Mandal, H. Oja and E. Ollila) these recent findings are reviewed, and a new reloaded procedure, to ensure that the independent components are extracted in an optimal order, is proposed. The reloaded algorithm improves the separation performance of the deflation-based FastICA estimator as amply illustrated by simulation studies. Reloading also seems to render the algorithm more stable.

IV. In symmetric independent component model it is assumed that the p-variate vector

$$x = \Omega z + \mu,$$

where μ is a location vector, Ω is a full rank $p \times p$ mixing matrix, and z is a p-variate vector with mutually independent and symmetrically distributed components.

In the article *Semiparametrically efficient inference based on signed ranks in symmetric independent component models* (P. Ilmonen and D. Paindaveine), optimal (in Le Cam sense) inference procedures are derived for a mixing matrix Ω in symmetric IC model. The inference procedures are based on the signed ranks of the residuals. Hypothesis tests, estimators and confidence zones are provided, and asymptotical properties are examined. In the article, optimality properties of the proposed inference procedures crucially rely on the uniform local asymptotic normality (ULAN) property of the model.

V. In the independent component (IC) model it is assumed that the p-variate vector

$$x = \Omega z,$$

where Ω is a full rank $p \times p$ mixing matrix and z is a p-variate vector with mutually independent components.

In the independent component analysis (ICA) the aim is to find an estimate for an unmixing matrix Γ such that Γx has independent components. Naturally $\Gamma = \Omega^{-1}$ is one possible unmixing matrix. The IC model can be formulated in several ways: If the independent components are permuted or multiplied by nonzero scalars they still remain independent. Thus the ICA problem reduces to estimating an unmixing matrix Ω^{-1} only up to the order, signs and scales of the row vectors. The comparison of the performances of different unmixing matrix estimates $\hat{\Gamma}$ is then difficult as the estimates are for different population quantities Γ. In the article *A new performance index for ICA: properties, computation and asymptotic analysis* (P. Ilmonen, K. Nordhausen, H. Oja and E. Ollila), a formal (mathematical) definition of the independent component (IC) functional $\Gamma(F)$ is given. The estimate is obtained when the functional is applied to the empirical cumulative distribution function.

A new natural performance index is suggested in the article. It finds the shortest distance (using Frobenius norm) between the identity matrix and the set of matrices equivalent to the gain matrix $\hat{\Gamma}\Omega$. The index is proven to possess several nice properties when compared to previously used indices, and it is easy and fast to compute. Limiting distribution of the index is provided when the limiting behavior of the estimate $\hat{\Gamma}$ is known. The theory is illustrated in a small simulation study.

References

Amari, S., Cichocki, A., and Yang, H. (1996). A new learning algorithm for blind source separation. *Advances in Neural Information Processing Systems*, 8:757–763.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Statistical Inference for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Bugrien, J. (2005). *Robust approaches to clustering based on density estimation and projection*. Ph.D. thesis, University of Leeds.

Cardoso, J. (1989). Source separation using higher moments. In *Proceedings of IEEE international conference on acoustics, speech and signal processing*, pages 2109–2112.

Cardoso, J. and Souloumiac, A. (1993). Blind beamforming for non gaussian signals. In *IEEE Proceedings-F*, volume 140, pages 362–370.

Caussinus, H. and Ruiz-Gazen, A. (1993). Projection pursuit and generalized principal component analysis. In *New Directions in Statistical Data Analysis and Robustness*, pages 34–46.

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. (1998). On an adaptive transformation and retransformation estimate of multivariate location. *Journal of the Royal Statistical Society, Series B*, 60:145–157.

Chaudhuri, P. and Sengupta, D. (1993). Sign tests in multidimension: Inference based on the geometry of the data cloud. *Journal of the American Statistical Association*, 88:1363–1370.

Chen, A. and Bickel, P. (2005). Consistent independent component anlysis and prewithening. In *IEEE Transactions on Signal Processing*, volume 53, pages 3625–3631.

Chen, A. and Bickel, P. (2006). Efficient independent component analysis. *The Annals of Statistics*, 34:2825–2855.

Cichocki, A. and Amari, S. (2006). *Adaptive Blind Signal and Image Processing*. John Wiley & Sons, Chichester.

Critchley, F., Pires, A., and Amado, C. (2006). Principal axis analysis. *Technical Report 06/14, The Open University, Milton Keynes*.

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

Douglas, S. (2007). Fixed-point algorithms for the blind separation of arbitrary complex-valued non-gaussian signal mixtures. *EURASIP Journal on Advances in Signal Processing*, 1:83–83.

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

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.

Hallin, M. and Werker, B. J. M. (2003). Semiparametric efficiency, distribution-freeness, and invariance. *Bernoulli*, 9:55–65.

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

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34:598–611.

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

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

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

Ilmonen, P., Nevalainen, J., and Oja, H. (2010a). Characteristics of multivariate distributions and the invariant coordinate system. *Statistics and Probability Letters*, 80(23–24):1844–1853.

Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2010b). A new performance index for ICA: properties, computation and asymptotic analysis. In *Proceedings of 9th International Conference on Latent Variable Analysis and Signal Separation*, pages 229–236.

Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2011a). Independent component (IC) functionals and a new performance index. *Submitted*.

Ilmonen, P., Oja, H., and Serfling, R. (2011b). On invariant coordinate system (ICS) functionals. *Submitted*.

Ilmonen, P. and Paindaveine, D. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *The Annals of Statistics*.

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.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–327.

Liski, E., Nordhausen, K., and Oja, H. (2011). Supervised invariant coordinate selection. *Submitted*.

Lopuhaä, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics*, 17:1662–1683.

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

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

Moreau, E. and Macchi, O. (1994). A one stage self-adaptive algorithm for source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 49–52.

Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., and Ollila, E. (2011a). Deflation based fastICA reloaded. In *Proceedings of EUSIPCO 2011*, pages 1854–1858.

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

Nordhausen, K., Oja, H., and Ollila, E. (2011b). Multivariate models and the first four moments. In *Festschrift for Thomas P. Hettmansperger*, pages 267–287.

Nordhausen, K., Oja, H., and Paindaveine, D. (2009). Signed-rank tests for location in the symmetric independent component model. *Journal of Multivariate Analysis*, 100:821–834.

Nordhausen, K., Oja, H., and Tyler, D. E. (2006). On the efficiency of invariant multivariate sign and rank tests. In *Festschrift for Tarmo Pukkila on his 60th Birthday*, pages 217–231.

Oja, H. (2010). *Multivariate Nonparametric Methods With R*. Springer-Verlag, New York, USA.

Oja, H., Paindaveine, D., and Taskinen, S. (2011). Parametric and nonparametric tests for multivariate independence in the independent component model. *Submitted*.

Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35:175–189.

Ollila, E. (2010). The deflation-based fastICA estimator: statistical analysis revisited. In *IEEE Transactions in Signal Processing*, volume 58, pages 1527–1541.

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.

Pere, A. (2000). Comparison of two methods for transforming height and weight to normality. *Annals of Human Biology*, 27:35–45.

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

Roelandt, E., Van Aelst, S., and Croux, C. (2009). *Journal of Multivariate Analysis*, 100:876–887.

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

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

Theis, F., Lang, E., and Puntonet, C. (2004). A geometric algorithm for overcomplete linear ICA. *Neurocomputing*, 56:381–398.

Tyler, D. E., Critchley, F., Dümbgen, L., and Oja, H. (2009). Invariant co-ordinate selection. *Journal of the Royal Statistical Society, Series B*, 71:549–592.

Yeredor, A. (2009). On optimal selection of correlation matrices for matrix-pencil-based separation. In *Lecture Notes in Computer Science (LNCS 5441): Independent Component analysis and Signal Separation*, volume 97, pages 1423–1426.

Submitted to the Annals of Statistics
arXiv: math.PR/0000000

SEMIPARAMETRICALLY EFFICIENT INFERENCE BASED ON SIGNED RANKS IN SYMMETRIC INDEPENDENT COMPONENT MODELS

BY PAULIINA ILMONEN[*,‡] AND DAVY PAINDAVEINE[†,§]

University of Tampere[‡] and Université Libre de Bruxelles[§]

We consider semiparametric location-scatter models for which the p-variate observation is obtained as $X = \Lambda Z + \mu$, where μ is a p-vector, Λ is a full-rank $p \times p$ matrix, and the (unobserved) random p-vector Z has marginals that are centered and mutually independent but are otherwise unspecified. As in blind source separation and independent component analysis (ICA), the parameter of interest throughout the paper is Λ. On the basis of n i.i.d. copies of X, we develop, under a symmetry assumption on Z, *signed-rank* one-sample testing and estimation procedures for Λ. We exploit the uniform local and asymptotic normality (ULAN) of the model to define signed-rank procedures that are semiparametrically efficient under correctly specified densities. Yet, as usual in rank-based inference, the proposed procedures remain valid (correct asymptotic size under the null, for hypothesis testing, and root-n consistency, for point estimation) under a very broad range of densities. We derive the asymptotic properties of the proposed procedures and investigate their finite-sample behavior through simulations.

1. Introduction. In multivariate statistics, concepts of location and scatter are usually defined through affine transformations of a noise vector. To be more specific, assume that the observation X is obtained through

$$X = \Lambda Z + \mu, \tag{1.1}$$

where μ is a p-vector, Λ is a full-rank $p \times p$ matrix, and Z is some *standardized* random vector. The exact nature of the resulting *location* parameter μ and *scatter* parameter $\Sigma = \Lambda\Lambda'$—or equivalently, *mixing matrix* parameter Λ, say—crucially depends on the standardization adopted.

[*]Supported by the Academy of Finland.

[†]Supported by an A.R.C. contract of the Communauté Française de Belgique. Davy Paindaveine is also member of ECORE, the association between CORE and ECARES.

AMS 2000 subject classifications: Primary 62G05, 62G10; secondary 62G20, 62H99

Keywords and phrases: Independent component analysis, Local asymptotic normality, Rank-based inference, Semiparametric efficiency, Signed ranks

The most classical assumption on Z specifies that Z is standard p-normal. Then μ and Σ simply coincide with the mean vector $\mathrm{E}[X]$ and variance-covariance matrix $\mathrm{Var}[X]$ of X, respectively. In robust statistics, it is often rather assumed that Z is spherically symmetric about the origin of $\mathbb{R}^p$—in the sense that the distribution of OZ does not depend on the orthogonal $p\times p$ matrix O. The resulting model in (1.1) is then called the *elliptical* model. If Z has finite second-order moments, then $\mu = \mathrm{E}[X]$ and $\Sigma = c\mathrm{Var}[X]$ for some $c > 0$, but this also defines μ and Σ in the absence of any moment assumption.

This paper focuses on an alternative standardization of Z, for which Z has mutually independent marginals with common median zero. The resulting model in (1.1)—the *independent component (IC) model*, say—is more flexible than the elliptical model, even if one restricts, as we will do, to vectors Z with symmetrically distributed marginals. The IC model indeed allows for heterogeneous marginal distributions for X, whereas, in contrast, marginals in the elliptical model all share—up to location and scale—the same distribution, hence also the same tail weight. This severely affects the relevance of elliptical models for practical applications, particularly so for moderate to large dimensions, since it is then very unlikely that all variables share, e.g., the same tail weight.

The IC model provides the most standard setup for *independent component analysis (ICA)*, in which the mixing matrix Λ is to be estimated on the basis of n independent copies $X_1, \ldots, X_n$ of X, the objective being to recover (up to a translation) the original unobservable independent signals $Z_1, \ldots, Z_n$—by premultiplying the X_i's with the resulting $\hat{\Lambda}^{-1}$. It is well-known in ICA, however, that Λ is severely unidentified : for any $p \times p$ permutation matrix P and any full-rank diagonal matrix D, one can always write

$$X = [\Lambda PD]\,[(PD)^{-1}Z] + \mu = \tilde{\Lambda}\,\tilde{Z} + \mu, \tag{1.2}$$

where $\tilde{Z}$ still has independent marginals with median zero. Provided that Z has at most one Gaussian marginal, two matrices Λ_1 and Λ_2 may lead to the same distribution for X in (1.1) if and only if they are equivalent (we will write $\Lambda_1 \sim \Lambda_2$) in the sense that $\Lambda_2 = \Lambda_1 PD$ for some matrices P and D as in (1.2); see, e.g., [25]. In other words, under the assumption that Z has at most one Gaussian marginal, permutations (P), sign changes and scale transformations (D) of the independent components are the only sources of unidentifiability for Λ.

This paper considers inference on the mixing matrix Λ. More precisely, because of the identifiability issues above, we rather consider a normalized

version L of Λ, where L is a well-defined representative of the class of mixing matrices that are equivalent to Λ. This parameter L is actually the parameter of interest in ICA : an estimate of L will indeed allow to recover the independent signals $Z_1, \ldots, Z_n$ equally well as an estimate of any other Λ with $\Lambda \sim L$. Interestingly, the situation is extremely similar when considering inference on Σ in the elliptical model. There, Σ is only identified up to a positive scalar factor, and it is often enough to focus on inference about the well-defined *shape* parameter $V = \Sigma/(\det \Sigma)^{1/p}$ (for instance, in PCA, principal directions, proportions of explained variance, etc. can be computed from V). Just as L is a normalized version of Λ in the IC model, V is a normalized version of Σ in the elliptical model, and in both classes of models, the normalized parameters actually are the natural parameters of interest in many inference problems. The similarities further extend to the semiparametric nature of both models : just as the density $g_{\|\cdot\|}$ of $\|Z\|$ in the elliptical model, the pdf g_r of the various independent components Z_r, $r = 1, \ldots, p$, in the IC model, can hardly be assumed to be known in practice.

These strong similarities motivate the approach we adopt in this paper : we plan to conduct inference on L (hypothesis testing and point estimation) in the IC model by adopting the methodology that proved extremely successful in [7, 8] for inference on V in the elliptical model. This methodology combines semiparametrically efficient inference and *invariance arguments*. In the IC model, the fixed-(μ, Λ) nonparametric submodels (indexed by $g_1, \ldots, g_p$) indeed enjoy a strong invariance structure that is parallel to the one of the corresponding elliptical submodels (indexed by $g_{\|\cdot\|}$). As in [7, 8], we exploit this invariance structure through a general result from [11] that allows to derive invariant versions of efficient central sequences, on the basis of which one can define semiparametrically efficient (at fixed target densities $g_r = f_r$, $r = 1, \ldots, p$) invariant procedures. As the maximal invariant associated with the invariance structure considered turns out to be the vector of marginal signed ranks of the residuals, the proposed procedures are of a signed-rank nature, and do not require to estimate densities. While they achieve semiparametric efficiency under correctly specified densities, they remain valid (correct asymptotic size under the null, for hypothesis testing, and root-n consistency, for point estimation) under misspecified densities.

We will consider the problem of estimating L and that of testing the null $\mathcal{H}_0 : L = L_0$ against the alternative $\mathcal{H}_1 : L \neq L_0$, for some fixed L_0. While point estimation is undoubtedly of primary importance for applications (e.g., in blind source separation), one might question the practical relevance of the testing problem considered, especially when L_0 is not the p-dimensional identity matrix. Solving this generic testing problem, how-

ever, is the main step in developing tests for any linear hypothesis on L, and we will explicitly describe the resulting tests in the sequel. An extensive study of these tests is beyond the scope of the present paper, though; we refer to [20] for an extension of our tests to the particular case of testing the (linear) hypothesis that L is block-diagonal, a problem that is obviously important in practice (non-rejection of the null would indeed allow practitioners to proceed with two separate, lower-dimensional, analyses). Testing linear hypotheses on L includes many other testing problems of high practical relevance, such as testing that a given column of L is equal to some fixed p-vector, and testing that a given entry of L is zero—the practical importance of these two testing problems, in relation, e.g., with functional magnetic resonance imaging (fMRI), is discussed in [22].

The paper is organized as follows. In Section 2, we fix the notation and describe the model (Section 2.1), state the corresponding *uniformly locally and asymptotically normal (ULAN)* property that allows to determine semiparametric efficiency bounds (Section 2.2), and then introduce, in relation with invariance arguments, *rank-based* efficient central sequences (Section 2.3). In Sections 3 and 4, we develop the resulting rank tests and estimators for the mixing matrix L, respectively. Our estimators actually require the delicate estimation of $2p(p-1)$ "cross-information coefficients", an issue we solve in Section 4.2 by generalizing the method recently developed in [5]. In Section 5, simulations are conducted both to compare the proposed estimators with some competitors and to investigate the validity of asymptotic results—simulation results for hypothesis testing are provided in the supplementary article [18]. Finally, the Appendix states some technical results (Appendix A) and reports proofs (Appendix B).

2. The model, the ULAN property, and invariance arguments.

2.1. *The model.* As already explained, the IC model above suffers from severe identifiability issues for Λ. To solve this, we map each Λ onto a unique representative $L = \Pi(\Lambda)$ of the collection of mixing matrices $\tilde{\Lambda}$ that satisfy $\tilde{\Lambda} \sim \Lambda$ (the equivalence class of Λ for $\sim$). We propose the mapping

$$\Lambda \mapsto \Pi(\Lambda) = \Lambda D_1^+ P D_2,$$

where D_1^+ is the positive definite diagonal matrix that makes each column of ΛD_1^+ have Euclidean norm one, P is the permutation matrix for which the matrix $B = (b_{ij}) = \Lambda D_1^+ P$ satisfies $|b_{ii}| > |b_{ij}|$ for all $i < j$, and D_2 is the diagonal matrix such that all diagonal entries of $\Pi(\Lambda) = \Lambda D_1^+ P D_2$ are equal to one.

If one restricts to the collection $\mathcal{M}_p$ of mixing matrices Λ for which no ties occur in the permutation step above, it can easily be shown that, for any $\Lambda_1, \Lambda_2 \in \mathcal{M}_p$, we have that $\Lambda_1 \sim \Lambda_2$ iff $\Pi(\Lambda_1) = \Pi(\Lambda_2)$, so that this mechanism succeeds in identifying a unique representative in each class of equivalence (this is ensured with the double scaling scheme above, which may seem a bit complicated at first). Besides, Π is then a continuously differentiable mapping from $\mathcal{M}_p$ onto $\mathcal{M}_{1p} := \Pi(\mathcal{M}_p)$. While ties may always be taken care of in some way (e.g., by basing the ordering on subsequent rows of the matrix B), they may prevent the mapping Π to be continuous, which would cause severe problems and would prevent us from using the Delta method in the sequel. It is clear, however, that the restriction to $\mathcal{M}_p$ only gets rid of a few particular mixing matrices, and will not have any implications in practice.

The parametrization of the IC model we consider is then associated with

$$X = LZ + \mu, \tag{2.1}$$

where $\mu \in \mathbb{R}^p$, $L \in \mathcal{M}_{1p}$, and Z has independent marginals with common median zero. Throughout, we further assume that Z admits a density with respect to the Lebesgue measure on $\mathbb{R}^p$, and that it has p symmetrically distributed marginals, among which at most one is Gaussian (as explained in the Introduction, this limitation on the number of Gaussian components is needed for L to be identifiable). We will denote by $\mathcal{F}$ the resulting collection of densities for Z. Of course, any $g \in \mathcal{F}$ naturally factorizes into $g(z) = \prod_{r=1}^p g_r(z_r)$, where g_r is the symmetric density of Z_r.

The hypothesis under which n mutually independent observations X_i, $i = 1, \ldots, n$ are obtained from (2.1), where Z has density $g \in \mathcal{F}$, will be denoted as $\mathrm{P}^{(n)}_{\vartheta,g}$, with $\vartheta = (\mu', (\mathrm{vecd}^\circ L)')' \in \Theta = \mathbb{R}^p \times \mathrm{vecd}^\circ(\mathcal{M}_{1p})$, or alternatively, as $\mathrm{P}^{(n)}_{\mu,L,g}$; for any $p \times p$ matrix A, we write $\mathrm{vecd}^\circ A$ for the $p(p-1)$-vector obtained by removing the p diagonal entries of A from its usual vectorized form $\mathrm{vec}\, A$ (diagonal entries of L are all equal to one, hence should not be included in the parameter).

The resulting semiparametric model is then

$$\mathcal{P}^{(n)} := \cup_{g\in\mathcal{F}} \mathcal{P}^{(n)}_g := \cup_{g\in\mathcal{F}} \cup_{\vartheta\in\Theta} \{\mathrm{P}^{(n)}_{\vartheta,g}\}. \tag{2.2}$$

Performing semiparametrically efficient inference on ϑ, at a fixed $f \in \mathcal{F}$, typically requires that the corresponding parametric submodel $\mathcal{P}^{(n)}_f$ satisfies the *uniformly locally and asymptotically normal (ULAN)* property.

2.2. *The ULAN property.* As always, the ULAN property requires technical regularity conditions on f. In the present context, we need that each

corresponding univariate pdf f_r, $r = 1, \ldots, p$, is absolutely continuous, with a derivative f_r' that satisfies

$$\sigma_{f_r}^2 := \int_{-\infty}^{\infty} y^2 f_r(y)\, dy < \infty, \quad \mathcal{I}_{f_r} := \int_{-\infty}^{\infty} \varphi_{f_r}^2(y) f_r(y)\, dy < \infty,$$

and

$$\mathcal{J}_{f_r} := \int_{-\infty}^{\infty} y^2 \varphi_{f_r}^2(y) f_r(y)\, dy < \infty,$$

where we let $\varphi_{f_r} := -f_r'/f_r$. In the sequel, we denote by $\mathcal{F}_{\text{ulan}}$ the collection of pdfs $f \in \mathcal{F}$ meeting these conditions.

For any $f \in \mathcal{F}_{\text{ulan}}$, let $\gamma_{rs}(f) := \mathcal{I}_{f_r}\sigma_{f_s}^2$, define the optimal p-variate location score function $\varphi_f : \mathbb{R}^p \to \mathbb{R}^p$ through $z = (z_1, \ldots, z_p)' \mapsto \varphi_f(z) = (\varphi_{f_1}(z_1), \ldots, \varphi_{f_p}(z_p))'$, and denote by $\mathcal{I}_f$ the diagonal matrix with diagonal entries $\mathcal{I}_{f_r}$, $r = 1, \ldots, p$. Further write I_ℓ for the ℓ-dimensional identity matrix and define

$$C := \sum_{r=1}^{p} \sum_{s=1}^{p-1} (e_r e_r' \otimes u_s e_{s+\delta_{[s\geq r]}}'),$$

where $\otimes$ is the usual Kronecker product, e_r and u_r stand for the rth vectors of the canonical basis of $\mathbb{R}^p$ and $\mathbb{R}^{p-1}$, respectively, and $\delta_{[s\geq r]}$ is equal to one if $s \geq r$ and to zero otherwise. The following ULAN result then easily follows from Proposition 2.1 in [20] by using a simple chain rule argument.

PROPOSITION 2.1. *Fix $f \in \mathcal{F}_{\text{ulan}}$. Then the collection of probability distributions $\mathcal{P}_f^{(n)}$ is ULAN, with central sequence*

(2.3)
$$\Delta_{\vartheta,f} = \begin{pmatrix} \Delta_{\vartheta,f;1} \\ \Delta_{\vartheta,f;2} \end{pmatrix} = \begin{pmatrix} n^{-1/2}(L^{-1})' \sum_{i=1}^n \varphi_f(Z_i) \\ n^{-1/2} C (I_p \otimes L^{-1})' \sum_{i=1}^n \text{vec}(\varphi_f(Z_i) Z_i' - I_p) \end{pmatrix},$$

where $Z_i = Z_i(\vartheta) = L^{-1}(X_i - \mu)$, and full-rank information matrix

$$\Gamma_{L,f} = \begin{pmatrix} \Gamma_{L,f;1} & 0 \\ 0 & \Gamma_{L,f;2} \end{pmatrix},$$

where $\Gamma_{L,f;1} := (L^{-1})' \mathcal{I}_f L^{-1}$ and

$$\Gamma_{L,f;2} := C(I_p \otimes L^{-1})' \Bigg[\sum_{r=1}^{p} (\mathcal{J}_{f_r} - 1)(e_r e_r' \otimes e_r e_r') + \sum_{r,s=1, r\neq s}^{p} (\gamma_{sr}(f)(e_r e_r' \otimes e_s e_s') + (e_r e_s' \otimes e_s e_r')) \Bigg] (I_p \otimes L^{-1}) C'.$$

More precisely, for any $\vartheta_n = \vartheta + O(n^{-1/2})$ *(with* $\vartheta = (\mu', (\mathrm{vecd}^\circ L)')'$*) and any bounded sequence* (τ_n) *in* $\mathbb{R}^{p^2}$*, we have that, under* $\mathrm{P}^{(n)}_{\vartheta_n,f}$ *as* $n \to \infty$*,*

$$\log(d\mathrm{P}^{(n)}_{\vartheta_n+n^{-1/2}\tau_n,f}/d\mathrm{P}^{(n)}_{\vartheta_n,f}) = \tau_n'\Delta_{\vartheta_n,f} - \frac{1}{2}\tau_n'\Gamma_{L,f}\tau_n + o_{\mathrm{P}}(1),$$

and $\Delta_{\vartheta_n,f}$ *converges in distribution to a* p^2*-variate normal distribution with mean zero and covariance matrix* $\Gamma_{L,f}$*.*

Semiparametrically efficient (at f) inference procedures on L then may be based on the so-called *efficient central sequence* $\Delta^*_{\vartheta,f;2}$ resulting from $\Delta_{\vartheta,f;2}$ by performing adequate tangent space projections; see [3]. Under $\mathrm{P}^{(n)}_{\vartheta,f}$, $\Delta^*_{\vartheta,f;2}$ is still asymptotically normal with mean zero, but now with covariance matrix $\Gamma^*_{L,f;2}$ (the *efficient information matrix*). This matrix $\Gamma^*_{L,f;2}$ settles the semiparametric efficiency bound at f when performing inference on L. For instance, an estimator $\hat{L}$ is semiparametrically efficient at f if

$$\sqrt{n}\,\mathrm{vecd}^\circ(\hat{L} - L) \xrightarrow{\mathcal{L}} \mathcal{N}_{p(p-1)}(0, (\Gamma^*_{L,f;2})^{-1}). \tag{2.4}$$

The performance of semiparametrically efficient tests on L can similarly be characterized in terms of $\Gamma^*_{L,f;2}$: a test of $\mathcal{H}_0 : L = L_0$ is semiparametrically efficient at f (at asymptotic level α) if its asymptotic powers under local alternatives of the form $\mathcal{H}^{(n)}_1 : L = L_0 + n^{-1/2}H$, where H is an arbitrary $p \times p$ matrix with zero diagonal entries, are given by

$$1 - \Psi_{p(p-1)}(\chi^2_{p(p-1),1-\alpha}; (\mathrm{vecd}^\circ H)'\Gamma^*_{L_0,f;2}(\mathrm{vecd}^\circ H)), \tag{2.5}$$

where $\chi^2_{p(p-1),1-\alpha}$ stands for the α-upper quantile of the $\chi^2_{p(p-1)}$ distribution, and $\Psi_{p(p-1)}(\,\cdot\,;\delta)$ denotes the cumulative distribution function of the non-central $\chi^2_{p(p-1)}$ distribution with non-centrality parameter δ.

2.3. *Invariance arguments.* Instead of the classical tangent space projection approach to compute $\Delta^*_{\vartheta,f;2}$ (as in [6]), we adopt an approach—due to [11]—that rather exploits the invariance structure of the model considered. This will provide a version of the efficient central sequence (parallel to central sequences, efficient central sequences are defined up to $o_{\mathrm{P}}(1)$'s only) that is based on *signed ranks*. Here, signed ranks are defined as $S_i(\vartheta) = (S_{i1}(\vartheta), \ldots, S_{ip}(\vartheta))'$ and $R^+_i(\vartheta) = (R^+_{i1}(\vartheta), \ldots, R^+_{ip}(\vartheta))'$, where $S_{ir}(\vartheta)$ is the sign of $Z_{ir}(\vartheta) = (L^{-1}(X_i - \mu))_r$ and $R^+_{ir}(\vartheta)$ is the rank of $|Z_{ir}(\vartheta)|$ among $|Z_{1r}(\vartheta)|, \ldots, |Z_{nr}(\vartheta)|$. This signed-rank efficient central sequence—$\underline{\Delta}^*_{\vartheta,f;2}$, say—is given in Theorem 2.1 below (the asymptotic behavior of $\underline{\Delta}^*_{\vartheta,f;2}$ will be studied in Appendix A).

To be able to state Theorem 2.1, we need to introduce the following notation. Let $z \mapsto F_+(z) = (F_{+1}(z_1), \ldots, F_{+r}(z_p))'$, with $F_{+r}(t) := \mathrm{P}^{(n)}_{\vartheta,f}[|Z_r(\vartheta)| < t] = 2(\int_{-\infty}^t f_r(s)\,ds) - 1$, $t \geq 0$. Based on this, define $\underline{\Delta}^*_{\vartheta,f;2} := C(I_p \otimes L^{-1})'\mathrm{vec}\,\underline{T}_{\vartheta,f}$, with

$$\underline{T}_{\vartheta,f} := \mathrm{odiag}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(S_i(\vartheta) \odot \varphi_f\Big(F_+^{-1}\Big(\frac{R_i^+(\vartheta)}{n+1}\Big)\Big)\Big)\Big(S_i(\vartheta) \odot F_+^{-1}\Big(\frac{R_i^+(\vartheta)}{n+1}\Big)\Big)'\right],$$

where $\odot$ is the Hadamard (i.e., entrywise) product of two vectors and where $\mathrm{odiag}(A)$ denotes the matrix obtained from A by replacing all diagonal entries with zeros. Finally, let $\underline{\mathcal{F}}_{\mathrm{ulan}}$ be the collection of pdfs $f \in \mathcal{F}_{\mathrm{ulan}}$ for which each φ_{f_r}, $r = 1, \ldots, p$, is continuous and can be written as the difference of two monotone increasing functions. We then have the following result (see Appendix B for a proof).

THEOREM 2.1. *Fix $\vartheta = (\mu', (\mathrm{vecd}^\circ L)')' \in \Theta$ and $f \in \underline{\mathcal{F}}_{\mathrm{ulan}}$. Then, (i) denoting by $\mathrm{E}^{(n)}_{\vartheta,f}$ expectation under $\mathrm{P}^{(n)}_{\vartheta,f}$,*

$$\begin{aligned}\underline{\Delta}^*_{\vartheta,f;2} &:= C(I_p \otimes L^{-1})'\mathrm{vec}\,\underline{T}_{\vartheta,f} \\ &= \mathrm{E}^{(n)}_{\vartheta,f}[\Delta_{\vartheta,f;2} \,|\, S_1(\vartheta), \ldots, S_n(\vartheta), R_1^+(\vartheta), \ldots, R_n^+(\vartheta)] + o_{L^2}(1)\end{aligned}$$

as $n \to \infty$, under $\mathrm{P}^{(n)}_{\vartheta,f}$; (ii) the signed-rank quantity $\underline{\Delta}^*_{\vartheta,f;2}$ is a version of the efficient central sequence at f (that is, $\underline{\Delta}^*_{\vartheta,f;2} = \Delta^*_{\vartheta,f;2} + o_{L^2}(1)$ as $n \to \infty$, under $\mathrm{P}^{(n)}_{\vartheta,f}$).

Would the (nonparametric) fixed-ϑ submodels $\mathcal{P}^{(n)}_\vartheta := \cup_{g\in\mathcal{F}}\{\mathrm{P}^{(n)}_{\vartheta,g}\}$ of the semiparametric model $\cup_{\theta\in\Theta}\cup_{g\in\mathcal{F}}\{\mathrm{P}^{(n)}_{\theta,g}\}$ in (2.2) be invariant under a group of transformations $\mathcal{G}^\vartheta$ that generates $\mathcal{P}^{(n)}_\vartheta$, then the main result of [11] would show that the expectation of the original central sequence $\Delta_{\vartheta,f;2}$ conditional upon the corresponding maximal invariant—$\mathcal{I}^{(n)}_{\max}(\vartheta)$, say—is a version of the efficient central sequence $\Delta^*_{\vartheta,f;2}$ at f : as $n \to \infty$, under $\mathrm{P}^{(n)}_{\vartheta,f}$,

$$\Delta^*_{\vartheta,f;2} = \mathrm{E}^{(n)}_{\vartheta,f}[\Delta_{\vartheta,f;2} \,|\, \mathcal{I}^{(n)}_{\max}(\vartheta)] + o_{L^2}(1). \tag{2.6}$$

Such an invariance structure actually exists and the relevant group $\mathcal{G}^\vartheta$ collects all transformations

$$\begin{aligned} g^\vartheta_h : \mathbb{R}^p \times \ldots \times \mathbb{R}^p &\to \mathbb{R}^p \times \ldots \times \mathbb{R}^p \\ (x_1, \ldots, x_n) &\mapsto (Lh(z_1(\vartheta)) + \mu, \ldots, Lh(z_n(\vartheta)) + \mu),\end{aligned}$$

with $z_i(\vartheta) := L^{-1}(x_i - \mu)$ and $h((z_1, \ldots, z_p)') = (h_1(z_1), \ldots, h_p(z_p))'$, where each h_r, $r = 1, \ldots, p$, is continuous, odd, monotone increasing, and fixes $+\infty$. It is easy to check that $\mathcal{P}^{(n)}_{\vartheta}$ is invariant under (and is generated by) $\mathcal{G}^{\vartheta}$, and that the corresponding maximal invariant is the vector of signed ranks

$$\mathcal{I}^{(n)}_{\max}(\vartheta) = (S_1(\vartheta), \ldots, S_n(\vartheta), R^+_1(\vartheta), \ldots, R^+_n(\vartheta)); \tag{2.7}$$

Theorem 2.1(ii) then follows from (2.6) and Theorem 2.1(i).

Inference procedures based on $\underline{\Delta}^*_{\vartheta,f;2}$, unlike those (from [6]) based on the efficient central sequence $\Delta^*_{\vartheta,f;2}$ obtained through tangent space projections, are measurable with respect to signed ranks, hence enjoy all nice properties usually associated with rank methods : robustness, ease of computation, validity without density estimation (and, for hypothesis testing, even distribution-freeness), etc.

3. Hypothesis testing. We now consider the problem of testing the null hypothesis $\mathcal{H}_0 : L = L_0$ against the alternative $\mathcal{H}_1 : L \neq L_0$, with unspecified underlying density g. Beyond their intriseque interest, the resulting tests will play an important role in the construction of the R-estimators of Section 4 below, and they pave the way to testing linear hypotheses on L.

The objective here is to define a test that is semiparametrically efficient at some target density f, yet that remains valid—in the sense that it meets asymptotically the level constraint—under a very broad class of densities g. As we will show, this objective is achieved by the signed-rank test—$\underline{\phi}_f$, say—that rejects $\mathcal{H}_0$ at asymptotic level $\alpha \in (0, 1)$ whenever

$$\underline{Q}_f := (\underline{\Delta}^*_{\hat{\vartheta}_0,f;2})' \, (\Gamma^*_{L_0,f;2})^{-1} \underline{\Delta}^*_{\hat{\vartheta}_0,f;2} > \chi^2_{p(p-1),1-\alpha}, \tag{3.1}$$

where $\Gamma^*_{L,f;2}$ was introduced in Page 7 (an explicit expression is given below) and where $\hat{\vartheta}_0 = (\hat{\mu}', (\mathrm{vecd}^\circ L_0)')'$ is based on a sequence of estimators $\hat{\mu}$ that is locally asymptotically discrete (see Appendix A for a precise definition) and root-n consistent under the null.

Possible choices for $\hat{\mu}$ include (discretized versions of) the sample mean $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ or the transformation-retransformation componentwise median $\hat{\mu}_{\mathrm{Med}} := L_0 \, \mathrm{Med}[L_0^{-1}X_1, \ldots, L_0^{-1}X_n]$, where $\mathrm{Med}[\cdot]$ returns the vector of univariate medians. We favor the sign estimator $\hat{\mu}_{\mathrm{Med}}$, since it is very much in line with the signed-rank tests $\underline{\phi}_f$ and enjoys good robustness properties. However, we stress that Theorem 3.1 below, which states the asymptotic properties of the proposed signed-rank tests, implies that the choice of $\hat{\mu}$ does not affect the asymptotic properties of $\underline{\phi}_f$, at any $g \in \mathcal{F}_{\mathrm{ulan}}$.

In order to state this theorem, we need to define

$$\Gamma^*_{L,f,g;2} := C(I_p \otimes L^{-1})' G_{f,g} (I_p \otimes L^{-1}) C' \tag{3.2}$$
$$:= C(I_p \otimes L^{-1})'$$
$$\times \left[\sum_{r,s=1, r\neq s}^{p} (\gamma_{sr}(f,g)(e_r e_r' \otimes e_s e_s') + \rho_{rs}(f,g)(e_r e_s' \otimes e_s e_r')) \right] (I_p \otimes L^{-1}) C',$$

where we let

$$\gamma_{rs}(f,g) := \int_0^1 \varphi_{f_r}(F_r^{-1}(u))\, \varphi_{g_r}(G_r^{-1}(u))\, du \times \int_0^1 F_s^{-1}(u)\, G_s^{-1}(u)\, du \tag{3.3}$$

and

$$\rho_{rs}(f,g) := \int_0^1 F_r^{-1}(u)\, \varphi_{g_r}(G_r^{-1}(u))\, du \times \int_0^1 \varphi_{f_s}(F_s^{-1}(u))\, G_s^{-1}(u)\, du. \tag{3.4}$$

We also let $\Gamma^*_{L,f;2} := \Gamma^*_{L,f,f;2}$ and $G_f := G_{f,f}$, that involve $\gamma_{rs}(f,f) = \gamma_{rs}(f)$ (see Section 2.2) and $\rho_{rs}(f,f) = 1$. We then have the following result (see Appendix B for a proof).

THEOREM 3.1. *Fix $f \in \underline{\mathcal{F}}_{\text{ulan}}$. Then (i) under $\mathrm{P}^{(n)}_{\vartheta_0,g}$ and under $\mathrm{P}^{(n)}_{\vartheta_0+n^{-1/2}\tau,g}$, with $\vartheta_0 = (\mu', (\mathrm{vecd}^\circ L_0)')'$, $\tau = (\tau_1', \tau_2')' \in \mathbb{R}^p \times \mathbb{R}^{p(p-1)}$, and $g \in \mathcal{F}_{\text{ulan}}$,*

$$\underline{Q}_f \xrightarrow{\mathcal{L}} \chi^2_{p(p-1)} \quad \text{and} \quad \underline{Q}_f \xrightarrow{\mathcal{L}} \chi^2_{p(p-1)}(\tau_2'(\Gamma^*_{L_0,f,g;2})'(\Gamma^*_{L_0,f;2})^{-1}\Gamma^*_{L_0,f,g;2}\tau_2),$$

respectively, as $n \to \infty$. (ii) The sequence of tests $\underline{\phi}^{(n)}_f$ has asymptotic level α under $\cup_{\mu\in\mathbb{R}^p} \cup_{g\in\mathcal{F}_{\text{ulan}}} \{\mathrm{P}^{(n)}_{\mu,L_0,g}\}$. (iii) The sequence of tests $\underline{\phi}^{(n)}_f$ is semiparametrically efficient, still at asymptotic level α, when testing $\mathcal{H}_0 : L = L_0$ against $H_1^f : L \neq L_0$ with noise density f (i.e., when testing $\cup_{\mu\in\mathbb{R}^p} \cup_{g\in\mathcal{F}_{\text{ulan}}} \{\mathrm{P}^{(n)}_{\mu,L_0,g}\}$ against $\cup_{\mu\in\mathbb{R}^p} \cup_{L\in\mathcal{M}_{1p}\setminus\{L_0\}} \{\mathrm{P}^{(n)}_{\mu,L,f}\}$).

The test $\underline{\phi}_f$ achieves semiparametric efficiency at f (Theorem 3.1(iii)), and also at any f_σ, with $f_\sigma(z) := \prod_{r=1}^p \sigma_r^{-1} f_r(z_r/\sigma_r)$, where $\sigma_r > 0$ for all r—it can indeed be checked that $\underline{\phi}_{f_\sigma} = \underline{\phi}_f$. Most importantly, Theorem 3.1 shows also that $\underline{\phi}_f$ remains valid under any $g \in \mathcal{F}_{\text{ulan}}$. By proceeding as in Lemma 4.2 of [20], this can even be extended to any $g \in \mathcal{F}$, which allows to avoid any finite moment condition.

This is to be compared to the semiparametric approach of Chen and Bickel [6]—these authors focus on point estimation, but their methodology

also leads to tests that enjoy the same properties as their estimators. Their procedures achieve uniform (in g) semiparametric efficiency, while our methods achieve semiparametric efficiency at the target density f only—more precisely, at any corresponding f_σ. However, it turns out that the performances of our procedures do not depend much on the target density f, so that our procedures are close to achieving uniform (in g) semiparametric efficiency; see the simulations in the supplemental article [18]. As any uniformly semiparametrically efficient procedures (see [1]), Chen and Bickel's procedures require estimating g, hence choosing various smoothing parameters. In contrast, our procedures, by construction, are invariant (here, signed-rank) ones. As such, they do not require to estimate densities, and they are robust, easy to compute, etc.

One might still object that the choice of f is quite arbitrary. This choice should be based on the practitioner's prior belief on the underlying densities. If he/she has no such prior belief, a kernel estimate $\hat{f}$ of f could be used. The resulting test $\underline{\phi}_{\hat{f}}$ would then enjoy the same properties as any $\underline{\phi}_f$ in terms of validity, since kernel density estimators typically are measurable with respect to the order statistics of the $|Z_{ir}(\hat{\vartheta}_0)|$'s, that, asymptotically, are stochastically independent of the signed ranks $S_{ir}(\hat{\vartheta}_0), R^+_{ir}(\hat{\vartheta}_0)$ used in $\underline{\phi}_f$; see [11] for details. The test $\underline{\phi}_{\hat{f}}$ would further achieve uniform semiparametric efficiency.

Further results on the proposed tests are given in the supplemental article [18]. More precisely, a simple explicit expression of the test statistics, local asymptotic powers of the corresponding tests, and simulation results can be found there.

We finish this section by describing the extension of our signed-rank tests to the problem of testing a fixed (arbitrary) linear hypothesis on L, which includes many instances of high practical relevance (we mentioned a few in the Introduction). Denoting by $\mathcal{V}(\Omega)$ the vector space that is spanned by the columns of the $p(p-1)\times\ell$ matrix Ω (that is assumed to have full rank ℓ), we consider the testing problem

$$\begin{cases} \mathcal{H}_0(L_0,\Omega) : (\mathrm{vecd}^\circ L) \in (\mathrm{vecd}^\circ L_0) + \mathcal{V}(\Omega) \\ \mathcal{H}_1(L_0,\Omega) : (\mathrm{vecd}^\circ L) \notin (\mathrm{vecd}^\circ L_0) + \mathcal{V}(\Omega), \end{cases} \tag{3.5}$$

for some fixed $L_0 \in \mathcal{M}_{1p}$. If one forgets about the tacitly assumed constraint that $L \in \mathcal{M}_{1p}$ in (3.5), the null hypothesis above imposes a set of linear constraints on L. This clearly includes all testing problems mentioned in the Introduction : testing that a given column of L is equal to a fixed vector, testing that a given (off-diagonal) entry of L is zero, and testing block-diagonality of L.

Inspired by the tests from [15] (Section 10.9), the analog of our signed-rank test $\underline{\phi}_f$ above then rejects $\mathcal{H}_0(L_0, \Omega)$ for large values of

$$\underline{Q}_f(L_0, \Omega) := (\underline{\Delta}^*_{\hat{\vartheta},f;2})' P_\Omega \, \underline{\Delta}^*_{\hat{\vartheta},f;2},$$

with $P_\Omega := (\Gamma^*_{\hat{L},f;2})^- - \Omega(\Omega'\Gamma^*_{\hat{L},f;2}\Omega)^-\Omega'$, where B^- denotes the Moore-Penrose pseudoinverse of B, and where $\hat{\vartheta} = (\hat{\mu}', (\mathrm{vecd}^\circ \hat{L})')'$ is an estimator of ϑ that is locally and asymptotically discrete, root-n consistent under the null, and *constrained*—in the sense that $\hat{L}$ satisfies the linear constraints in $\mathcal{H}_0(L_0, \Omega)$.

It can be shown that this signed-rank test achieves semiparametric optimality at f (the relevant optimality concept here is *most stringency*; see, e.g., [20] for a discussion) and remains valid under any $g \in \underline{\mathcal{F}}_{\mathrm{ulan}}$. Its null asymptotic distribution is still chi-square, now with $r := \mathrm{Trace}[P_\Omega \Gamma^*_{L,f;2}]$ degrees of freedom (this directly follows from Theorem 9.2.1 in [24] and Theorem A.1); at asymptotic level α, the resulting asymptotic critical value (that actually does not depend on the true value L) therefore is $\chi^2_{r;1-\alpha}$. Just as for the tests $\underline{\phi}_f$, it is still possible to compute asymptotic powers under sequences of local alternatives. It is clear, however, that a thorough study of the properties of the tests above, for a general linear hypothesis, is beyond the scope of the present paper, hence is left for future research. In the important particular case of testing block-diagonality of L, a complete investigation of the signed-rank tests can be found in [20].

4. Point estimation. We turn to the problem of estimating L, which is of primary importance for applications. Denoting by $\underline{Q}_f = \underline{Q}_f(L_0)$ the signed-rank test statistic for $\mathcal{H}_0 : L = L_0$ in (3.1), a natural signed-rank estimator of L is obtained by "inverting the corresponding test" :

$$\underline{\hat{L}}_{f;\mathrm{argmin}} = \arg \min_{L \in \mathcal{M}_{1p}} \underline{Q}_f(L).$$

This estimator, however, is not satisfactory : as any signed-rank quantity, the objective function $L \mapsto \underline{Q}_f(L)$ is piecewise constant, hence discontinuous and non-convex, which makes it very difficult to derive the asymptotic properties of $\underline{\hat{L}}_{f;\mathrm{argmin}}$. It is also virtually impossible to compute $\underline{\hat{L}}_{f;\mathrm{argmin}}$ in practice, since this lack of smoothness and convexity essentially forces computing the estimator by simply running over a grid of possible values of the $p(p-1)$-dimensional parameter L—a strategy that cannot provide a reasonable approximation of $\underline{\hat{L}}_{f;\mathrm{argmin}}$, even for moderate values of p. Finally, there is no way to estimate the asymptotic covariance matrix of $\underline{\hat{L}}_{f;\mathrm{argmin}}$, which

rules out the possibility to derive confidence zones for L, hence drastically restricts the practical relevance of this estimator.

In order to avoid the aforementioned drawbacks, we propose adopting a one-step approach that was first used in [7] for the problem of estimating the shape of an elliptical distribution or in [9] in a more general context. The resulting one-step signed-rank estimators—in the sequel, we simply speak of *one-step rank estimators* or *one-step R-estimators*—can easily be computed in practice, their asymptotic properties can be derived explicitly, and their asymptotic covariance matrix can be estimated consistently.

4.1. *One-step R-estimators of L.* To initiate the one-step procedure, a preliminary estimator is needed. In the present context, we will assume that a root-n consistent and locally asymptotically discrete estimator $\tilde{\vartheta} = (\tilde{\mu}', (\mathrm{vecd}^\circ \tilde{L})')'$ is available. As we will show, the asymptotic properties of the proposed one-step R-estimators will not be affected by the choice of $\tilde{\vartheta}$. Practical choices will be provided in Section 5.

Describing our one-step R-estimators requires

ASSUMPTION (A). For all $r \neq s \in \{1, \ldots, p\}$, we dispose of sequences of estimators $\hat{\gamma}_{rs}(f)$ and $\hat{\rho}_{rs}(f)$ that (i) are locally asymptotically discrete and that (ii), for any $g \in \mathcal{F}_{\mathrm{ulan}}$, satisfy $\hat{\gamma}_{rs}(f) = \gamma_{rs}(f,g) + o_{\mathrm{P}}(1)$ and $\hat{\rho}_{rs}(f) = \rho_{rs}(f,g) + o_{\mathrm{P}}(1)$ as $n \to \infty$, under $\cup_{\vartheta \in \Theta} \{\mathrm{P}^{(n)}_{\vartheta,g}\}$.

Sequences of estimators fulfilling this assumption will be provided in Section 4.2 below. At this point, just note that plugging in (3.2) the estimators from Assumption (A) and the preliminary estimator $\tilde{L}$, defines a statistic—$\hat{\Gamma}^*_{\tilde{L},f;2}$, say—that consistently estimates $\Gamma^*_{L,f,g;2}$ under $\cup_{\vartheta \in \Theta} \{\mathrm{P}^{(n)}_{\vartheta,g}\}$.

For any target density f, we propose the one-step R-estimator $\underline{\hat{L}}_f$, with values in $\mathcal{M}_{1p}$, defined by

$$\mathrm{vecd}^\circ \underline{\hat{L}}_f := (\mathrm{vecd}^\circ \tilde{L}) + n^{-1/2} (\hat{\Gamma}^*_{\tilde{L},f;2})^{-1} \underline{\Delta}^*_{\tilde{\vartheta},f;2}. \tag{4.1}$$

The following result states the asymptotic properties of this estimator (see Appendix B for a proof).

THEOREM 4.1. *Let Assumption (A) hold and fix $f \in \underline{\mathcal{F}}_{\mathrm{ulan}}$. Then (i) under $\mathrm{P}^{(n)}_{\vartheta,g}$, with $\vartheta = (\mu', (\mathrm{vecd}^\circ L)')' \in \Theta$ and $g \in \mathcal{F}_{\mathrm{ulan}}$, we have that,*

$$\sqrt{n}\ \mathrm{vec}(\underline{\hat{L}}_f - L) = C'(\Gamma^*_{L,f,g;2})^{-1} \underline{\Delta}^*_{\vartheta,f;2} + o_{\mathrm{P}}(1) \tag{4.2}$$

$$= C'(\Gamma^*_{L,f,g;2})^{-1} \Delta^*_{\vartheta,f,g;2} + o_{\mathrm{P}}(1) \tag{4.3}$$

$$\xrightarrow{\mathcal{L}} \mathcal{N}_{p(p-1)}\big(0, C'(\Gamma^*_{L,f,g;2})^{-1} \Gamma^*_{L,f;2} (\Gamma^*_{L,f,g;2})^{-1\prime} C\big) \tag{4.4}$$

as $n \to \infty$, *where* $\Delta^*_{\vartheta,f,g;2}$ *is defined in Theorem A.1 (see Appendix A). (ii) The estimator* $\hat{\underline{L}}_f$ *is semiparametrically efficient at* f.

The result in (4.2) justifies calling $\hat{\underline{L}}_f$ an R-estimator since it shows that $n^{1/2}(\hat{\underline{L}}_f - L)$ is asymptotically equivalent to a random matrix that is measurable with respect to the signed ranks $S_i(\vartheta), R_i^+(\vartheta)$ in (2.7). The asymptotic equivalence in (4.3) gives a Bahadur-type representation result for $\hat{\underline{L}}_f$ with summands that are independent and identically distributed, hence leads trivially to the asymptotic normality result in (4.4). Recalling that $\hat{\Gamma}^*_{\tilde{L},f;2}$ consistently estimates $\Gamma^*_{L,f,g;2}$ under $\cup_{\vartheta\in\Theta}\{\mathrm{P}^{(n)}_{\vartheta,g}\}$, it is clear that asymptotic (signed-rank) confidence zones for L may easily be obtained from this asymptotic normality result.

For $r \neq s \in \{1,\ldots,p\}$, define $\hat{\alpha}_{rs}(f)$ and $\hat{\beta}_{rs}(f)$ as the statistics obtained by plugging the estimators $\hat{\gamma}_{rs}(f)$ and $\hat{\rho}_{rs}(f)$ from Assumption (A) in

$$\text{(4.5)} \qquad \begin{cases} \alpha_{rs}(f,g) := \dfrac{\gamma_{rs}(f,g)}{\gamma_{rs}(f,g)\gamma_{sr}(f,g) - \rho_{rs}(f,g)\rho_{sr}(f,g)} \\ \beta_{rs}(f,g) := \dfrac{-\rho_{rs}(f,g)}{\gamma_{rs}(f,g)\gamma_{sr}(f,g) - \rho_{rs}(f,g)\rho_{sr}(f,g)}, \end{cases}$$

and let $\hat{\alpha}_{rr}(f) := 0 =: \hat{\beta}_{rr}(f)$, $r = 1,\ldots,p$. The estimator $\hat{\underline{L}}_f$ then admits the following explicit expression (see Appendix B for a proof).

THEOREM 4.2. *Let Assumption (A) hold and fix* $f \in \underline{\mathcal{F}}_{\mathrm{ulan}}$. *Let* $\hat{N}_f := (\hat{\mathcal{A}}'_f \odot \underline{T}_{\tilde{\vartheta},f}) + (\hat{\mathcal{B}}'_f \odot \underline{T}'_{\tilde{\vartheta},f})$, *where we let* $\hat{\mathcal{A}}_f := (\hat{\alpha}_{rs}(f))$ *and* $\hat{\mathcal{B}}_f := (\hat{\beta}_{rs}(f))$. *Then the estimator* $\hat{\underline{L}}_f$ *rewrites*

$$\text{(4.6)} \qquad \hat{\underline{L}}_f = \tilde{L} + \frac{1}{\sqrt{n}}\tilde{L}[\hat{N}_f - \mathrm{diag}(\tilde{L}\hat{N}_f)],$$

where $\mathrm{diag}(A) = A - \mathrm{odiag}(A)$ *stands for the diagonal matrix with the same diagonal entries as* A.

It is straightforward to check that the role of the term $-\frac{1}{\sqrt{n}}\tilde{L}\,\mathrm{diag}(\tilde{L}\hat{N}_f)$ in the one-step correction $\frac{1}{\sqrt{n}}\tilde{L}[\hat{N}_f - \mathrm{diag}(\tilde{L}\hat{N}_f)]$ of $\tilde{L}$ is merely to ensure that the diagonal entries of $\hat{\underline{L}}_f$ remain equal to one, hence that $\hat{\underline{L}}_f$ takes values in $\mathcal{M}_{1p}$ (for n large enough).

As shown above, the estimator $\hat{\underline{L}}_f$ enjoys very nice properties : its asymptotic behavior is completely characterized, it is semiparametrically efficient

under correctly specified densities, yet remains root-n consistent and asymptotically normal under a broad range of densities g, its asymptotic covariance matrix can easily be estimated consistently, etc.

However, $\hat{\underline{L}}_f$ requires estimates $\hat{\gamma}_{rs}(f)$ and $\hat{\rho}_{rs}(f)$ that fulfill Assumption (A). We now provide such estimates.

4.2. *Estimation of cross-information coefficients.* Of course, it is always possible to estimate consistently the cross-information coefficients $\gamma_{rs}(f,g)$ and $\rho_{rs}(f,g)$ by replacing g in (3.3)-(3.4) with appropriate window or kernel density estimates—this can be achieved since the residuals $Z_{ir}(\tilde{\vartheta})$, $i = 1,\ldots,n$ typically are asymptotically i.i.d. with density g_r. Rank-based methods, however, intend to eliminate—through invariance arguments—the nuisance g without estimating it, so that density estimation methods simply are antinomic to the spirit of rank-based methods.

Therefore, we rather propose a solution that is based on ranks and avoids estimating the underlying nuisance g. The method, that relies on the asymptotic linearity—under g—of an appropriate rank-based statistic $\underline{S}_{\vartheta,f}$, was first used in [7], where there is only one cross-information coefficient $J(f,g)$ to be estimated. There, it is crucial that $J(f,g)$ is involved as a scalar factor in the asymptotic covariance matrix, under g, between the rank-based efficient central sequence $\underline{\Delta}^*_{\vartheta,f}$ and the parametric central sequence $\Delta_{\vartheta,g}$. In [5], the method was extended to allow for the estimation of a cross-information coefficient that appears as a scalar factor in the linear term of the asymptotic linearity, under g, of an arbitrary (possibly vector-valued) rank-based statistic $\underline{S}_{\vartheta,f}$.

In all cases, thus, this method was only used to estimate a *single* cross-information coefficient that appears as a *scalar factor* in some structural—typically, cross-information—matrix. In this respect, our problem, which requires to estimate $2p(p-1)$ cross-information quantities appearing in various entries of the cross-information matrix $\Gamma^*_{L,f,g;2}$, is much more complex. Yet, as we now show, it allows for a solution relying on the same basic idea of exploiting the asymptotic linearity, under g, of an appropriate f-score rank-based statistic.

Based on the preliminary estimator $\tilde{\vartheta} := (\tilde{\mu}', (\mathrm{vecd}^\circ \tilde{L})')'$ at hand, define $\tilde{\vartheta}^{\gamma_{rs}}_\lambda := (\tilde{\mu}', (\mathrm{vecd}^\circ \tilde{L}^{\gamma_{rs}}_\lambda)')'$, $\lambda \geq 0$, with

$$\tilde{L}^{\gamma_{rs}}_\lambda := \tilde{L} + n^{-1/2}\lambda(\underline{T}_{\tilde{\vartheta},f})_{rs}\tilde{L}(e_r e_s' - \mathrm{diag}(\tilde{L}e_r e_s')),$$

and $\tilde{\vartheta}^{\rho_{rs}}_\lambda := (\tilde{\mu}', (\mathrm{vecd}^\circ \tilde{L}^{\rho_{rs}}_\lambda)')'$, $\lambda \geq 0$, with

$$\tilde{L}^{\rho_{rs}}_\lambda := \tilde{L} + n^{-1/2}\lambda(\underline{T}_{\tilde{\vartheta},f})_{sr}\tilde{L}(e_r e_s' - \mathrm{diag}(\tilde{L}e_r e_s'));$$

note that, at $\lambda = 0$, $\tilde{\vartheta}_\lambda^{\gamma_{rs}} = \tilde{\vartheta}_\lambda^{\rho_{rs}} = \tilde{\vartheta}$ We then have the following result, that is crucial for the construction of the estimators $\hat{\gamma}_{rs}(f)$ and $\hat{\rho}_{rs}(f)$ (see Appendix B for a proof).

LEMMA 4.1. *Fix $\vartheta \in \Theta$, $f \in \underline{\mathcal{F}}_{\text{ulan}}$, $g \in \mathcal{F}_{\text{ulan}}$, and $r \neq s \in \{1, \ldots, p\}$. Then $h^{\gamma_{rs}}(\lambda) := (\underline{T}_{\tilde{\vartheta},f})_{rs}(\underline{T}_{\tilde{\vartheta}_\lambda^{\gamma_{rs}},f})_{rs} = (1-\lambda\gamma_{rs}(f,g))\,((\underline{T}_{\tilde{\vartheta},f})_{rs})^2 + o_{\mathrm{P}}(1)$ and $h^{\rho_{rs}}(\lambda) := (\underline{T}_{\tilde{\vartheta},f})_{sr}(\underline{T}_{\tilde{\vartheta}_\lambda^{\rho_{rs}},f})_{sr} = (1-\lambda\rho_{rs}(f,g))\,((\underline{T}_{\tilde{\vartheta},f})_{sr})^2 + o_{\mathrm{P}}(1)$ as $n \to \infty$, under $\mathrm{P}_{\vartheta,g}^{(n)}$.*

The mappings $\lambda \mapsto h^{\gamma_{rs}}(\lambda)$ and $\lambda \mapsto h^{\rho_{rs}}(\lambda)$ assume a positive value in $\lambda = 0$, and, as shown by Lemma 4.1, are—up to $o_{\mathrm{P}}(1)$'s as $n \to \infty$ under $\mathrm{P}_{\vartheta,g}^{(n)}$—monotone decreasing functions that become negative at $\lambda = (\gamma_{rs}(f,g))^{-1}$ and $\lambda = (\rho_{rs}(f,g))^{-1}$, respectively. Restricting to a grid of values of the form $\lambda_j = j/c$ for some large discretization constant c (which is needed to achieve the required discreteness), this naturally leads—via linear interpolation—to the estimators $\hat{\gamma}_{rs}(f)$ and $\hat{\rho}_{rs}(f)$ defined through

$$(\hat{\gamma}_{rs}(f))^{-1} := \lambda_{\gamma_{rs}} := \lambda^-_{\gamma_{rs}} + \frac{(\lambda^+_{\gamma_{rs}} - \lambda^-_{\gamma_{rs}})h^{\gamma_{rs}}(\lambda^-_{\gamma_{rs}})}{h^{\gamma_{rs}}(\lambda^-_{\gamma_{rs}}) - h^{\gamma_{rs}}(\lambda^+_{\gamma_{rs}})} = \lambda^-_{\gamma_{rs}} + \frac{c^{-1}\,h^{\gamma_{rs}}(\lambda^-_{\gamma_{rs}})}{h^{\gamma_{rs}}(\lambda^-_{\gamma_{rs}}) - h^{\gamma_{rs}}(\lambda^+_{\gamma_{rs}})}, \tag{4.7}$$

with $\lambda^-_{\gamma_{rs}} := \inf\{j \in \mathbb{N} : h^{\gamma_{rs}}(\lambda_{j+1}) < 0\}$ and $\lambda^+_{\gamma_{rs}} := \lambda^-_{\gamma_{rs}} + \frac{1}{c}$, and

$$(\hat{\rho}_{rs}(f))^{-1} := \lambda_{\rho_{rs}} := \lambda^-_{\rho_{rs}} + \frac{c^{-1}\,h^{\rho_{rs}}(\lambda^-_{\rho_{rs}})}{h^{\rho_{rs}}(\lambda^-_{\rho_{rs}}) - h^{\rho_{rs}}(\lambda^+_{\rho_{rs}})}, \tag{4.8}$$

with $\lambda^-_{\rho_{rs}} := \inf\{j \in \mathbb{N} : h^{\rho_{rs}}(\lambda_{j+1}) < 0\}$ and $\lambda^+_{\rho_{rs}} := \lambda^-_{\rho_{rs}} + \frac{1}{c}$. We have the following result (see the supplemental article [18] for a proof).

THEOREM 4.3. *Fix $\vartheta \in \Theta$ and $f, g \in \mathcal{F}_{\text{ulan}}$. Assume that $\tilde{\vartheta}$ is such that, for all $\varepsilon > 0$, there exist $\delta_\varepsilon > 0$ and an integer N_ε such that*

$$\mathrm{P}_{\vartheta,g}^{(n)}[(\underline{T}_{\tilde{\vartheta},f})_{rs} \geq \delta_\varepsilon] \geq 1 - \varepsilon, \tag{4.9}$$

for all $n \geq N_\varepsilon$, $r \neq s \in \{1, \ldots, p\}$. Then, for any such r, s, $\hat{\gamma}_{rs}(f) = \gamma_{rs}(f,g) + o_{\mathrm{P}}(1)$ and $\hat{\rho}_{rs}(f) = \rho_{rs}(f,g) + o_{\mathrm{P}}(1)$, as $n \to \infty$ under $\mathrm{P}_{\vartheta,g}^{(n)}$, hence $\hat{\gamma}_{rs}(f)$ and $\hat{\rho}_{rs}(f)$ satisfy Assumption (A).

We point out that the assumption in (4.9) is extremely mild, as it only requires that there is no couple (r,s), $r \neq s$, for which $(\underline{T}_{\tilde{\vartheta},f})_{rs}$ asymptotically has an atom in zero. It therefore rules out preliminary estimators $\tilde{L}$ defined through the (rank-based) f-likelihood equation $(\underline{T}_{\vartheta,f})_{rs} = 0$.

5. Simulations. Here we report simulation results for point estimation only—simulation results for hypothesis testing can be found in the supplemental article [18]. Our aim is to both compare the proposed estimators with some competitors and to investigate the validity of asymptotic results.

We used the following competitors : (i) FastICA from [12, 13], which is by far the most commonly used estimate in practice; we used here its deflation based version with the standard nonlinearity function pow3. (ii) FOBI from [4], which is one of the earliest solutions to the ICA problem and is often used as a benchmark estimate. (iii) The estimate based on two scatter matrices from [19]; here the two scatter matrices used are the regular empirical covariance matrix (COV) and the van der Waerden rank-based estimator (HOP) from [7] (actually, HOP is not a scatter matrix but rather a shape matrix, which is allowed in [19]). Root-n consistency of the resulting estimates $\hat{L}_{\mathrm{FICA}}$, $\hat{L}_{\mathrm{FOBI}}$, and $\hat{L}_{\mathrm{COV\_HOP}}$ of L requires finite sixth-, eighth-, and fourth-order moments, respectively, and follows from [16, 17] and [21].

We focused on the bivariate case $p = 2$, and we generated, for three different setups indexed by $d \in \{1, 2, 3\}$, $M = 2,000$ independent random samples $Z_i^{(d,m)} = (Z_{i1}^{(d,m)}, Z_{i2}^{(d,m)})'$, $i = 1, \ldots, n$, of size $n = 4,000$. Denoting by $g^{(d)}(z) = g_1^{(d)}(z_1)g_2^{(d)}(z_2)$ the common pdf of $Z_i^{(d,m)}$, $i = 1, \ldots, n$, $m = 1, \ldots, M$, the marginal densities $g_1^{(d)}$ and $g_2^{(d)}$ were chosen as follows.

(i) In Setup $d = 1$, $g_1^{(d)}$ is the pdf of the standard normal distribution ($\mathcal{N}$) and $g_2^{(d)}$ is the pdf of the Student distribution with 5 degrees of freedom (t_5);
(ii) In Setup $d = 2$, $g_1^{(d)}$ is the pdf of the logistic distribution with scale parameter one (log), and $g_2^{(d)}$ is t_5;
(iii) In Setup $d = 3$, $g_1^{(d)}$ is t_8 and $g_2^{(d)}$ is t_5.

We chose to use $L = I_2$ and $\mu = (0, 0)'$, so that the observations are given by $X_i^{(d,m)} = LZ_i^{(d,m)} + \mu = Z_i^{(d,m)}$ (other values of L and μ led to extremely similar results).

For each sample, we computed the competing estimates $\hat{L}_{\mathrm{FICA}}$, $\hat{L}_{\mathrm{FOBI}}$, and $\hat{L}_{\mathrm{COV\_HOP}}$ defined above. Each of these were also used as a preliminary estimator $\tilde{L}$ in the construction of three R-estimators : $\underline{\hat{L}}_{f^{(j)}}$, $j = 1, 2, 3$, with $f^{(j)} = g^{(j)}$ for all j. In the resulting nine R-estimators, we used the location estimate $\hat{\mu} = \tilde{L}\,\mathrm{Med}[\tilde{L}^{-1}X_1, \ldots, \tilde{L}^{-1}X_n]$, based on the preliminary estimate $\tilde{L}$ used to initiate the one-step procedure.

Figure 1 reports, for each setup d, a boxplot of the M squared errors

$$\|\hat{L}(X_1^{(d,m)},\ldots,X_n^{(d,m)})-L\|^2=\sum_{\substack{r,s=1\\ r\neq s}}^{p}\left(\hat{L}_{rs}(X_1^{(d,m)},\ldots,X_n^{(d,m)})-L_{rs}\right)^2 \tag{5.1}$$

for each of the twelve estimators $\hat{L}$ considered (the nine R-estimators and their three competitors).

The results show that, in each setup, all R-estimators dramatically improve over their competitors. The behavior of the R-estimators does not much depend on the preliminary estimator $\tilde{L}$ used. Optimality of $\hat{L}_{f^{(d)}}$ in Setup d is confirmed. Most importantly, as stated for hypothesis testing at the end of Section 3, the performances of the R-estimators do not depend much on the target density $f^{(j)}$ adopted, so that one should not worry much about the choice of the target density in practice. Quite surprisingly, R-estimators behave remarkably well even when based on preliminary estimators that, due to heavy tails, fail to be root-n consistent.

In order to investigate small-sample behavior of the estimates, we reran the exact same simulation with sample size $n = 800$; in ICA, where most applications involve sample sizes that are not in hundreds, but much larger, this sample size can indeed be considered small. Results are reported in Figure 2. They indicate that, in Setups 2 and 3, R-estimators still improve significantly over their competitors, and particularly over $\hat{L}_{\mathrm{FOBI}}$ and $\hat{L}_{\mathrm{COV\_HOP}}$. In Setup 1, there seem to be no improvement. Compared to results for $n = 4,000$, the behavior of one-step R-estimators here depends more on the preliminary estimator used. Performances of R-estimators again do not depend crucially on the target density, and optimality under correctly specified densities is preserved in most cases.

As a conclusion, for practical sample sizes, the proposed R-estimators outperform the standard competitors considered, and their behavior is very well in line with our asymptotic results.

Finally, we illustrate the proposed method for estimating cross-information coefficients. We consider again the first 50 replications of our simulation with $n = 4,000$, and focus on Setup 1 ($g = g^{(1)}$) and the target density $f = f^{(3)}(\neq g^{(1)})$. The cross-information coefficients to be estimated then are $\gamma_{12}(f,g) \approx 1.478$, $\gamma_{21}(f,g) \approx 0.862$, $\rho_{12}(f,g) \approx 1.149$, and $\rho_{21}(f,g) \approx 0.887$. The upper left picture in Figure 3 shows 150 graphs of the mapping $\lambda \mapsto h^{\gamma_{12}}(\lambda)$ (based on $f = f^{(3)}$), among which the 50 pink curves are based on $\tilde{L} = \hat{L}_{\mathrm{FICA}}$, the 50 green curves are based on $\tilde{L} = \hat{L}_{\mathrm{FOBI}}$, and the 50 blue ones are based on $\tilde{L} = \hat{L}_{\mathrm{COV\_HOP}}$. The upper right, bottom left, and bottom right pictures of the same figure provide the corresponding graphs for the

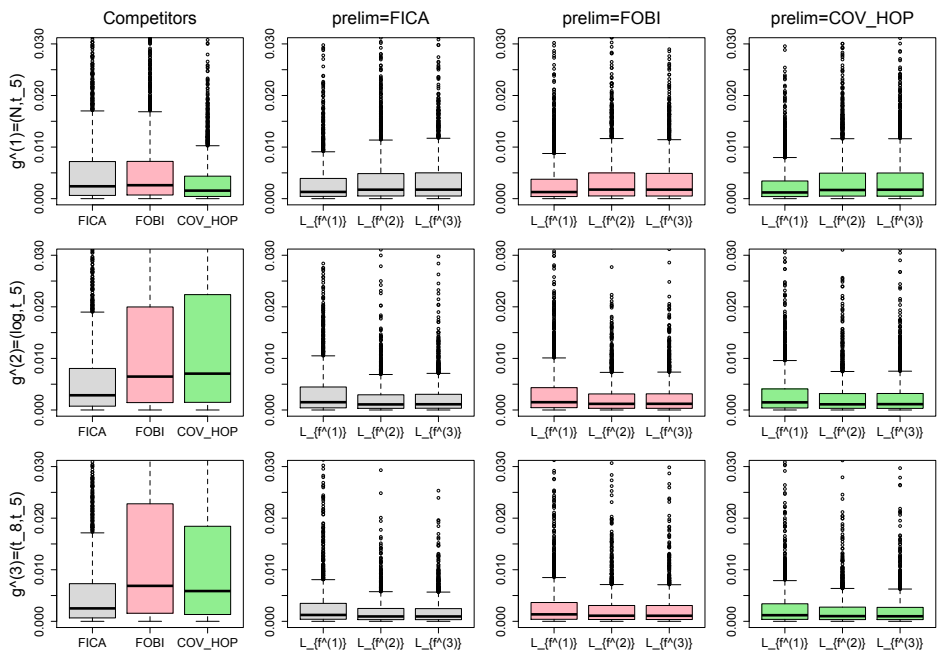

FIG 1. *Boxplots of the squared errors* $\|\hat{L} - L\|^2$ *(see (5.1)) obtained in* $M = 2,000$ *replications from setups* $d = 1, 2, 3$ *(associated with underlying distributions* $g^{(d)}$*,* $d = 1, 2, 3$*) for the competitors* $\hat{L}_{\text{FICA}}$*,* $\hat{L}_{\text{FOBI}}$*, and* $\hat{L}_{\text{COV\_HOP}}$*, and the nine R-estimators* $\hat{L}_f$ *resulting from all combinations of a target density* $f^{(j)} = g^{(j)}$*,* $j = 1, 2, 3$*, and one of the three preliminary estimators* $\hat{L}_{\text{FICA}}$*,* $\hat{L}_{\text{FOBI}}$*, and* $\hat{L}_{\text{COV\_HOP}}$*; see Section 5 for details. The sample size is* $n = 4,000$*.*

mappings $\lambda \mapsto h^{\gamma_{21}}(\lambda)$, $\lambda \mapsto h^{\rho_{12}}(\lambda)$, and $\lambda \mapsto h^{\rho_{21}}(\lambda)$, respectively. The value at which each graph crosses the λ-axis is the resulting estimate of the inverse of the associated cross-information coefficient. To be able to evaluate the results, we plotted, in each picture, a vertical black line at the corresponding theoretical value, namely at $1/\gamma_{12}(f, g)$, $1/\gamma_{21}(f, g)$, $1/\rho_{12}(f, g)$, and $1/\rho_{21}(f, g)$. Clearly, the results are excellent, and there does not seem to be much dependence on the preliminary estimator $\tilde{L}$ used.

APPENDIX A: RANK-BASED EFFICIENT CENTRAL SEQUENCES

In this first appendix, we study the asymptotic behavior of the rank-based efficient central sequences $\underline{\Delta}^*_{\vartheta,f;2}$. The main result is the following (see Appendix B for a proof).

THEOREM A.1. *Fix* $\vartheta = (\mu', (\text{vecd}^\circ L)')' \in \Theta$ *and* $f \in \underline{\mathcal{F}}_{\text{ulan}}$*. Then (i)*

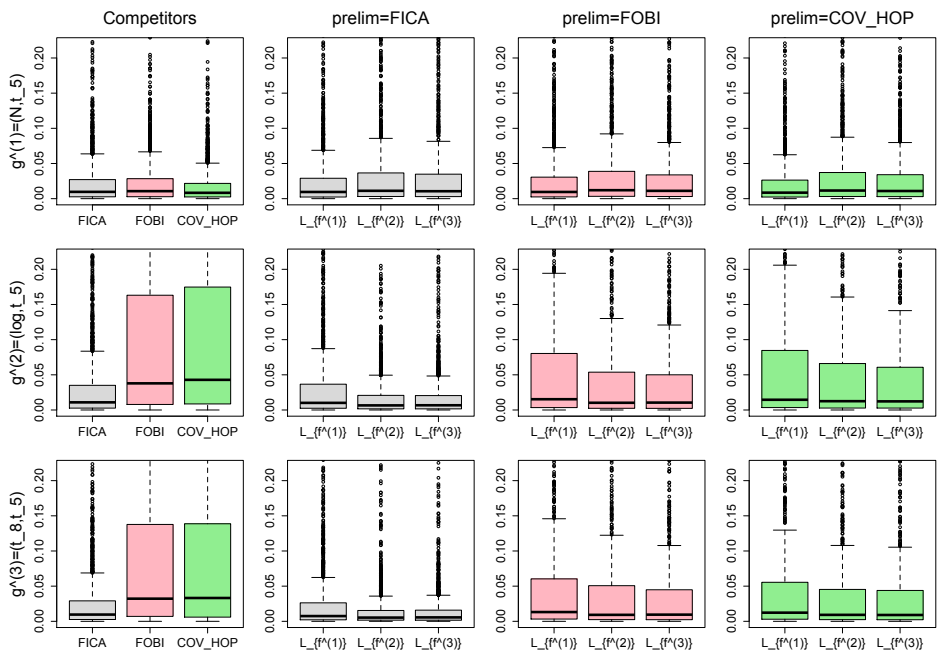

FIG 2. *The same boxplots as in Figure 1, but based on sample size* $n = 800$.

for any $g \in \mathcal{F}$,

$$\underline{\Delta}^*_{\vartheta,f;2} = \Delta^*_{\vartheta,f,g;2} + o_{L^2}(1)$$

as $n \to \infty$, *under* $\mathrm{P}^{(n)}_{\vartheta,g}$, *where* $\Delta^*_{\vartheta,f,g;2} := C(I_p \otimes L^{-1})'\mathrm{vec}\big[\mathrm{odiag}\big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (S_i \odot \varphi_f(F_+^{-1}(G_+(|Z_i|)))) (S_i \odot F_+^{-1}(G_+(|Z_i|)))'\big)\big]$. *(ii) Under* $\mathrm{P}^{(n)}_{\vartheta+n^{-1/2}\tau,g}$, *with* $\tau = (\tau_1', \tau_2')' \in \mathbb{R}^p \times \mathbb{R}^{p(p-1)}$ *and* $g \in \mathcal{F}_{\mathrm{ulan}}$,

$$\underline{\Delta}^*_{\vartheta,f;2} \xrightarrow{\mathcal{L}} \mathcal{N}_{p(p-1)}(\Gamma^*_{L,f,g;2}\tau_2, \Gamma^*_{L,f;2}),$$

as $n \to \infty$ *(for* $\tau = 0$, *the result only requires that* $g \in \mathcal{F}$*). (iii) Still with* $\tau = (\tau_1', \tau_2')' \in \mathbb{R}^p \times \mathbb{R}^{p(p-1)}$ *and* $g \in \mathcal{F}_{\mathrm{ulan}}$, $\underline{\Delta}^*_{\vartheta+n^{-1/2}\tau,f;2} - \underline{\Delta}^*_{\vartheta,f;2} = -\Gamma^*_{L,f,g;2}\tau_2 + o_{\mathrm{P}}(1)$ *as* $n \to \infty$, *under* $\mathrm{P}^{(n)}_{\vartheta,g}$.

Both for hypothesis testing and point estimation, we had to replace in $\underline{\Delta}^*_{\vartheta,f;2}$ the parameter ϑ with some estimator ($\check{\vartheta}^{(n)}$, say). The asymptotic behavior of the resulting (so-called *aligned*) rank-based efficient central sequence $\underline{\Delta}^*_{\check{\vartheta}^{(n)},f;2}$ is given in the following result.

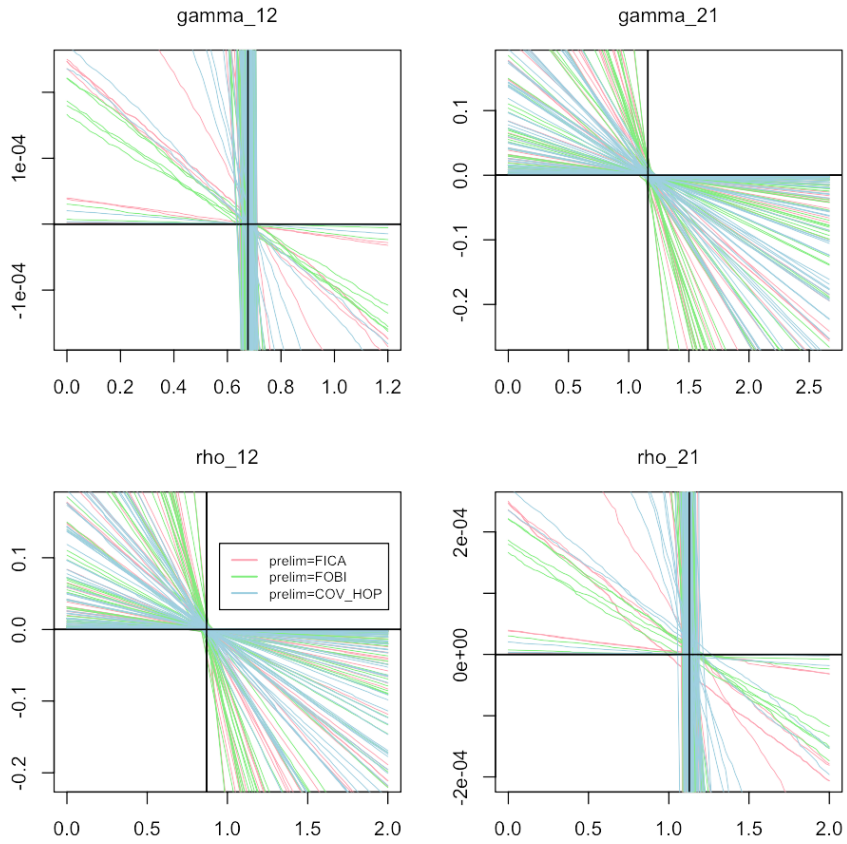

FIG 3. *Top left: 150 graphs of the mapping* $\lambda \mapsto h^{\gamma_{12}}(\lambda)$ *based on* $f = f^{(3)}$*, associated with the first 50 replications from Setup 1 (*$g = g^{(1)}$*) in Figure 1 (sample size is* $n = 4,000$*): the 50 curves in pink, green, and blue are based on the preliminary estimators* $\hat{L}_{\text{FICA}}$*,* $\hat{L}_{\text{FOBI}}$*, and* $\hat{L}_{\text{COV\_HOP}}$*, respectively. Top right, bottom left, and bottom right: the corresponding plots for the mappings* $\lambda \mapsto h^{\gamma_{21}}(\lambda)$*,* $\lambda \mapsto h^{\rho_{12}}(\lambda)$*, and* $\lambda \mapsto h^{\rho_{21}}(\lambda)$*, respectively.*

COROLLARY A.1. *Fix* $\vartheta = (\mu', (\text{vecd}^\circ L)')' \in \Theta$ *and* $f \in \underline{\mathcal{F}}_{\text{ulan}}$*, and* $g \in \mathcal{F}_{\text{ulan}}$*. Let* $\check{\vartheta} = \check{\vartheta}^{(n)} = (\check{\mu}', (\text{vecd}^\circ \check{L})')'$ *be a locally asymptotically discrete sequence of random vectors satisfying* $n^{1/2}(\check{\vartheta} - \vartheta) = O_{\text{P}}(1)$ *as* $n \to \infty$*, under* $\text{P}^{(n)}_{\vartheta,g}$*. Then* $\underline{\Delta}^*_{\check{\vartheta},f;2} - \underline{\Delta}^*_{\vartheta,f;2} = -\Gamma^*_{L,f,g;2} n^{1/2} \text{vecd}^\circ(\check{L} - L) + o_{\text{P}}(1)$*, still as* $n \to \infty$*, under* $\text{P}^{(n)}_{\vartheta,g}$*.*

Since the sequence of estimators $\check{\vartheta}^{(n)}$ is assumed to be locally asymptotically discrete (which means that the number of possible values of $\check{\vartheta}^{(n)}$ in balls with $O(n^{-1/2})$ radius centered at ϑ is bounded as $n \to \infty$), this result is a direct consequence of Theorem A.1(iii) and Lemma 4.4 from [14]. Local

asymptotic discreteness is a concept that goes back to Le Cam and is quite standard in one-step estimation; see, e.g., [2] or [14].

Of course, a sequence of estimators $\check{\vartheta}^{(n)}$ can always be discretized by replacing each component $(\check{\vartheta}^{(n)})_\ell$ with

$$(\check{\vartheta}^{(n)}_{\#})_\ell := (cn^{1/2})^{-1}\mathrm{sign}((\check{\vartheta}^{(n)})_\ell)\lceil cn^{1/2}|(\check{\vartheta}^{(n)})_\ell|\rceil, \quad \ell = 1,\ldots,p^2,$$

for some arbitrary constant $c > 0$. In practice, however, one can safely forget about such discretizations : irrespective of the accuracy of the computer used, the discretization constant c can always be chosen large enough to make discretization be irrelevant at the fixed sample size n_0 at hand—hence also at any $n > n_0$.

APPENDIX B: PROOFS

B.1. Proofs of Theorems 2.1 and A.1. The proofs of this section make use of the Hájek projection theorem for linear signed-rank statistics (see, e.g., [23], Chapter 3), which states that, if $Y_i = \mathrm{Sign}(Y_i)|Y_i|$, $i = 1,\ldots,n$ are i.i.d. with (absolutely continuous) cdf G and if $K : (0,1) \to \mathbb{R}$ is a continuous and square-integrable score function that can be written as the difference of two monotone increasing functions, then

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{Sign}(Y_i)K(G_+(|Y_i|)) \\
&\qquad = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{Sign}(Y_i)\, K\Big(\frac{R_i^+}{n+1}\Big) + o_{L^2}(1) \qquad \text{(B.1)} \\
&\qquad = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{Sign}(Y_i)\, \mathrm{E}\big[K(G_+(|Y_i|))\,|\,R_i^+\big] + o_{L^2}(1) \qquad \text{(B.2)}
\end{aligned}$$

as $n \to \infty$, where G_+ stands for the common cdf of the $|Y_i|$'s and R_i^+ denotes the rank of $|Y_i|$ among $|Y_1|,\ldots,|Y_n|$. The quantities in (B.1) and (B.2) are linear signed-rank quantities that are said to be based on *approximate* and *exact* scores, respectively.

In the rest of this section, we fix $\vartheta \in \Theta$, $f \in \mathcal{F}_{\mathrm{ulan}}$, and $g \in \mathcal{F}$. We write throughout Z_i, S_i, and R_i^+, for $Z_i(\vartheta)$, $S_i(\vartheta)$, and $R_i^+(\vartheta)$, respectively. We also write E_h instead of $\mathrm{E}^{(n)}_{\vartheta,h}$, with $h = f, g$. We then start with the proof of Theorem A.1(i).

Proof of Theorem A.1(i). Fix $r \neq s \in \{1,\ldots,p\}$ and two score functions $K_a, K_b : (0,1) \to \mathbb{R}$ with the same properties as K above. Then,

by using (i) $\mathrm{E}_g[S_{ir}] = 0$, (ii) the independence (under $\mathrm{P}^{(n)}_{\vartheta,g}$) between the S_{ir}'s and the $(R_{ir}, |Z_{ir}|)$'s, and (iii) the independence between the Z_{ir}'s and the Z_{is}'s, we obtain

$$\begin{aligned}
&\mathrm{E}_g\Big[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ir}S_{is}\Big(K_a(G_{+r}(|Z_{ir}|))K_b(G_{+s}(|Z_{is}|)) - K_a\Big(\frac{R^+_{ir}}{n+1}\Big)K_b\Big(\frac{R^+_{is}}{n+1}\Big)\Big)\Big)^2\Big] \\
&\quad = \frac{1}{n}\sum_{i=1}^n \mathrm{E}_g\Big[\Big(K_a(G_{+r}(|Z_{ir}|))K_b(G_{+s}(|Z_{is}|)) - K_a\Big(\frac{R^+_{ir}}{n+1}\Big)K_b\Big(\frac{R^+_{is}}{n+1}\Big)\Big)^2\Big] \\
&\quad \le 2\mathrm{E}_g\Big[\Big(K_a(G_{+r}(|Z_{ir}|)) - K_a\Big(\frac{R^+_{ir}}{n+1}\Big)\Big)^2\Big]\mathrm{E}_g\Big[K_b^2(G_{+s}(|Z_{is}|))\Big] \\
&\qquad +2\mathrm{E}_g\Big[K_a^2\Big(\frac{R^+_{ir}}{n+1}\Big)\Big]\mathrm{E}_g\Big[\Big(K_b(G_{+s}(|Z_{is}|)) - K_b\Big(\frac{R^+_{is}}{n+1}\Big)\Big)^2\Big].
\end{aligned}$$

Consequently, the square integrability of K_a, K_b, and the convergence to zero of both $\mathrm{E}_g[(K_a(G_{+r}(|Z_{ir}|)) - K_a(\frac{R^+_{ir}}{n+1}))^2]$ and $\mathrm{E}_g[(K_b(G_{+r}(|Z_{is}|)) - K_b(\frac{R^+_{is}}{n+1}))^2]$ (which directly follows from (B.1)) entail

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ir}S_{is}K_a(G_{+r}(|Z_{ir}|))K_b(G_{+s}(|Z_{is}|)) \\
&\qquad = \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ir}S_{is}K_a\Big(\frac{R^+_{ir}}{n+1}\Big)K_b\Big(\frac{R^+_{is}}{n+1}\Big) + o_{L^2}(1)
\end{aligned}$$

as $n \to \infty$, under $\mathrm{P}^{(n)}_{\vartheta,g}$. Theorem A.1(i) follows by taking $K_a = \varphi_{f_r} \circ F^{-1}_{+r}$ and $K_b = F^{-1}_{+s}$. □

We go on with the proof of Theorem 2.1, for which it is important to note that, by proceeding as in the proof of Theorem A.1(i) but with (B.2) instead of (B.1), we further obtain that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ir}S_{is}K_a(G_{+r}(|Z_{ir}|))K_b(G_{+s}(|Z_{is}|)) \\
&\qquad = \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ir}S_{is}K_a\Big(\frac{R^+_{ir}}{n+1}\Big)K_b\Big(\frac{R^+_{is}}{n+1}\Big) + o_{L^2}(1) \\
&\qquad = \frac{1}{\sqrt{n}}\sum_{i=1}^n S_{ir}S_{is}\mathrm{E}\big[K_a(G_{+r}(|Z_{ir}|)) \,|\, R_{+ir}\big] \\
&\qquad\qquad \times \mathrm{E}\big[K_b(G_{+s}(|Z_{is}|)) \,|\, R_{+is}\big] + o_{L^2}(1), \qquad \text{(B.3)}
\end{aligned}$$

still as $n \to \infty$ under $\mathrm{P}^{(n)}_{\vartheta,g}$.

PROOF OF THEOREM 2.1. It is sufficient to prove Theorem 2.1(i) only, since, as already mentioned at the end of Section 2.3, Theorem 2.1(ii) follows from (2.6) and Theorem 2.1(i). That is, we have to show that, for any $r, s \in \{1, \ldots, p\}$,

(B.4)
$$\mathrm{E}_f\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\varphi_f(Z_i)Z_i' - I_p\right)_{rs} \,|\, S_1, \ldots, S_n, R_1^+, \ldots, R_n^+\right] = (\underline{T}_{\vartheta,f})_{rs} + o_{L^2}(1)$$

as $n \to \infty$, under $\mathrm{P}_{\vartheta,f}^{(n)}$. Now, the left-hand side of (B.4) rewrites

$$\begin{aligned}
&\mathrm{E}_f\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\varphi_f(Z_i)Z_i' - I_p\right)_{rs} \,|\, S_1, \ldots, S_n, R_1^+, \ldots, R_n^+\right] \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{E}_f\left[S_{ir}S_{is}\varphi_f(|Z_{ir}|)|Z_{is}| - \delta_{rs} \,|\, S_1, \ldots, S_n, R_1^+, \ldots, R_n^+\right] \\
\text{(B.5)} &\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(S_{ir}S_{is}\mathrm{E}_f\left[\varphi_f(|Z_{ir}|)|Z_{is}| \,|\, R_{1r}^+, \ldots, R_{nr}^+, R_{1s}^+, \ldots, R_{ns}^+\right] - \delta_{rs}\right).
\end{aligned}$$

For $r \neq s$, this yields

$$\begin{aligned}
&\mathrm{E}_f\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\varphi_f(Z_i)Z_i' - I_p\right)_{rs} \,|\, S_1, \ldots, S_n, R_1^+, \ldots, R_n^+\right] \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{ir}S_{is}\mathrm{E}_f\left[\varphi_f(|Z_{ir}|) \,|\, R_{1r}^+, \ldots, R_{nr}^+\right]\mathrm{E}_f\left[|Z_{is}| \,|\, R_{1s}^+, \ldots, R_{ns}^+\right] \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{ir}S_{is}\varphi_{f_r}\Big(F_{+r}^{-1}\Big(\frac{R_{ir}^+}{n+1}\Big)\Big)F_{+r}^{-1}\Big(\frac{R_{is}^+}{n+1}\Big) + o_{L^2}(1) \\
&\quad = (\underline{T}_{\vartheta,f})_{rs} + o_{L^2}(1)
\end{aligned}$$

as $n \to \infty$, under $\mathrm{P}_{\vartheta,f}^{(n)}$, where we have used (B.3), still with $K_a = \varphi_{f_r} \circ F_{+r}^{-1}$ and $K_b = F_{+s}^{-1}$, but this time at $g = f$. This establishes (B.4) for $r \neq s$. As $r = s$, (B.5) now entails (writing $K_{ab}(u) := \varphi_f(F_{+r}^{-1}(u)) \times F_{+r}^{-1}(u)$ for

all u)

$$\begin{aligned}
&\mathrm{E}_f\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\varphi_f(Z_i)Z_i'-I_p\big)_{rs}\,|\,S_1,\ldots,S_n,R_1^+,\ldots,R_n^+\Big]\\
&\qquad=\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{E}_f\big[\varphi_f(|Z_{ir}|)|Z_{ir}|\,|\,R_{1r}^+,\ldots,R_{nr}^+\big]\Big)-\sqrt{n}\\
&\qquad=\mathrm{E}_f\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_{ab}(F_{+r}(|Z_{ir}|))\,|\,R_{1r}^+,\ldots,R_{nr}^+\Big]-\sqrt{n}\\
&\qquad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_{ab}\Big(\frac{R_i^+}{n+1}\Big)-\sqrt{n}+o_{L^2}(1) \qquad \text{(B.6)}\\
&\qquad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_{ab}\Big(\frac{i}{n+1}\Big)-\sqrt{n}+o_{L^2}(1)\\
&\qquad=\sqrt{n}\int_0^1 K_{ab}(u)\,du-\sqrt{n}+o_{L^2}(1) \qquad \text{(B.7)}\\
&\qquad=o_{L^2}(1), \qquad \text{(B.8)}
\end{aligned}$$

still as $n\to\infty$, under $\mathrm{P}_{\vartheta,f}^{(n)}$, where (B.6), (B.7), and (B.8) follow from the Hájek projection theorem for linear *rank* (not *signed-rank*) statistics (see, e.g., [23], Chapter 2), the square-integrability of $K_{ab}(\cdot)$ (see the proof of Proposition 3.2(i) in [10]), and integration by parts, respectively. This further proves (B.4) for $r=s$, hence also the result. □

Proof of Theorem A.1(ii)-(iii). (ii) In view of Theorem A.1(i), it is sufficient to show that both asymptotic normality results hold for $\Delta^*_{\vartheta,f,g;2}$. The result under $\mathrm{P}_{\vartheta,g}^{(n)}$ then straightforwardly follows from the multivariate CLT. As for the result under local alternatives (which, just as the result in Part (iii), requires that $g\in\mathcal{F}_{\mathrm{ulan}}$), it is obtained as usual, by establishing the joint normality under $\mathrm{P}_{\vartheta,g}^{(n)}$ of $\log(d\mathrm{P}_{\vartheta+n^{-1/2}\tau,f}^{(n)}/d\mathrm{P}_{\vartheta,g}^{(n)})$ and $\Delta^*_{\vartheta,f,g;2}$, then applying Le Cam's third Lemma; the required joint normality follows from a routine application of the classical Cramér-Wold device. (iii) The proof, that is long and tedious, is also a quite trivial adaptation of the proof of Proposition A.1. in [7]. We therefore omit it. □

B.2. Proof of Theorem 3.1.

Proof of Theorem 3.1. (i) Applying Corollary A.1, with $\check{\vartheta}:=\hat{\vartheta}_0=(\hat{\mu}',(\mathrm{vecd}\circ L_0)')'$ and $\vartheta:=\vartheta_0=(\mu',(\mathrm{vecd}\circ L_0)')'$, entails that $\underline{\Delta}^*_{\hat{\vartheta}_0,f;2}$

$= \underline{\Delta}^*_{\vartheta_0,f;2} + o_{\mathrm{P}}(1)$ as $n \to \infty$ under $\mathrm{P}^{(n)}_{\vartheta_0,g}$. Consequently, we have that

$$\underline{Q}_f = (\mathrm{vec}\,\underline{\Delta}^*_{\vartheta_0,f;2})'(\Gamma^*_{L_0,f;2})^{-1}(\mathrm{vec}\,\underline{\Delta}^*_{\vartheta_0,f;2}) + o_{\mathrm{P}}(1), \tag{B.9}$$

still as $n \to \infty$, under $\mathrm{P}^{(n)}_{\vartheta_0,g}$—hence also under $\mathrm{P}^{(n)}_{\vartheta_0+n^{-1/2}\tau,g}$ (from contiguity). The result then follows from Theorem A.1(ii). (ii) It directly follows from (i) that, under the sequence of local alternatives $\mathrm{P}^{(n)}_{\vartheta_0+n^{-1/2}\tau,f}$, $\underline{\phi}^{(n)}_f$ has asymptotic power $1 - \Psi_{p(p-1)}(\chi^2_{p(p-1),1-\alpha}; \tau_2'\Gamma^*_{L_0,f;2}\tau_2)$. This establishes the result, since these local powers coincide with the semiparametrically optimal (at f) powers in (2.5). □

B.3. Proofs of Lemma 4.1, Theorem 4.1, and Theorem 4.2.

Proof of Theorem 4.1. (i) Fix $\vartheta \in \Theta$ and $g \in \mathcal{F}_{\mathrm{ulan}}$. From (4.1), the fact that $\hat{\Gamma}^*_{\tilde{L},f;2} - \Gamma^*_{L,f,g;2} = o_{\mathrm{P}}(1)$ as $n \to \infty$ under $\mathrm{P}^{(n)}_{\vartheta,g}$, and Theorem A.1(iii), we obtain

$$\begin{aligned}
\sqrt{n}\ \mathrm{vecd}^\circ(\hat{L}_f - L) &= \sqrt{n}\ \mathrm{vecd}^\circ(\tilde{L} - L) + (\hat{\Gamma}^*_{\tilde{L},f;2})^{-1}\underline{\Delta}^*_{\vartheta,f;2} \\
&= \sqrt{n}\ \mathrm{vecd}^\circ(\tilde{L} - L) + (\Gamma^*_{L,f,g;2})^{-1}\underline{\Delta}^*_{\vartheta,f;2} + o_{\mathrm{P}}(1) \\
&= (\Gamma^*_{L,f,g;2})^{-1}\underline{\Delta}^*_{\vartheta,f;2} + o_{\mathrm{P}}(1) \qquad \text{(B.10)}
\end{aligned}$$

as $n \to \infty$ under $\mathrm{P}^{(n)}_{\vartheta,g}$. Consequently, Theorem A.1(i)-(ii) entails that, still as $n \to \infty$ under $\mathrm{P}^{(n)}_{\vartheta,g}$,

$$\begin{aligned}
\sqrt{n}\ \mathrm{vecd}^\circ(\hat{L}_f - L) &= (\Gamma^*_{L,f,g;2})^{-1}\Delta^*_{\vartheta,f,g;2} + o_{\mathrm{P}}(1) \qquad \text{(B.11)} \\
&\xrightarrow{\mathcal{L}} \mathcal{N}_{p(p-1)}(0, (\Gamma^*_{L,f,g;2})^{-1}\Gamma^*_{L,f;2}(\Gamma^*_{L,f,g;2})^{-1\prime}). \qquad \text{(B.12)}
\end{aligned}$$

Now, by using the fact that $C'(\mathrm{vecd}^\circ H) = (\mathrm{vec}\, H)$ for any $p \times p$ matrix H with only zero diagonal entries, we have that $\sqrt{n}\,\mathrm{vec}(\hat{L}_f - L) = \sqrt{n}\,C'\mathrm{vecd}^\circ(\hat{L}_f - L)$, so that (4.2), (4.3), and (4.4) follow from (B.10), (B.11), and (B.12), respectively.

(ii) The asymptotic covariance matrix of $\sqrt{n}\,\mathrm{vecd}^\circ(\hat{L}_f - L)$, under $\mathrm{P}^{(n)}_{\vartheta,f}$, reduces to $(\Gamma^*_{L,f;2})^{-1}$ (let $g = f$ in (B.12)), which establishes the result. □

To prove Theorem 4.2, we will need the following result.

LEMMA B.1. *Fix $\vartheta = (\mu', (\mathrm{vecd}^\circ L)')' \in \Theta$ and $f, g \in \mathcal{F}_{\mathrm{ulan}}$. Then*

$$(I_p \otimes L^{-1})C'\,(\Gamma^*_{L,f,g;2})^{-1}C(I_p \otimes L^{-1})' = \sum_{r,s=1,r\neq s}^{p} \Big\{\alpha_{rs}(f,g)(e_re_r' \otimes (L_{rs}^2 e_re_r' + e_se_s' - L_{rs}e_re_s' - L_{rs}e_se_r')) + \beta_{rs}(f,g)(e_re_s' \otimes (L_{rs}L_{sr}e_re_s' - L_{rs}e_re_r' - L_{sr}e_se_s' + e_se_r'))\Big\},$$

where L_{rs} denotes the entry (r,s) of L.

PROOF OF THEOREM 4.2. By using again the fact that $C'(\mathrm{vecd}^\circ H) = (\mathrm{vec}\, H)$ for any $p \times p$ matrix H with only zero diagonal entries, and then Lemma B.1, we obtain

$$\begin{aligned}
&\mathrm{vec}(\hat{L}_f - \tilde{L}) \\
&= C'\mathrm{vecd}^\circ(\hat{L}_f - \tilde{L}) = \frac{1}{\sqrt{n}}\, C'(\hat{\Gamma}^*_{\tilde{L},f;2})^{-1}C(I_p \otimes \tilde{L}^{-1})'\mathrm{vec}\,\underline{T}_{\tilde{\vartheta},f} \\
&= \frac{1}{\sqrt{n}}\,(I_p \otimes \tilde{L}) \\
&\times \Bigg[\sum_{r,s=1,r\neq s}^{p} \Big\{\hat{\alpha}_{rs}(f)(e_re_r' \otimes (\tilde{L}_{rs}^2 e_re_r' + e_se_s' - \tilde{L}_{rs}e_re_s' - \tilde{L}_{rs}e_se_r')) \\
&+ \hat{\beta}_{rs}(f)(e_re_s' \otimes (\tilde{L}_{rs}\tilde{L}_{sr}e_re_s' - \tilde{L}_{rs}e_re_r' - \tilde{L}_{sr}e_se_s' + e_se_r'))\Big\}\Bigg]\mathrm{vec}\,\underline{T}_{\tilde{\vartheta},f}.
\end{aligned}$$

Since all diagonal entries of $\underline{T}_{\tilde{\vartheta},f}$ are zeros, we have that

$$\begin{aligned}
\mathrm{vec}(\hat{L}_f - \tilde{L}) = \frac{1}{\sqrt{n}}\,(I_p \otimes \tilde{L})\Bigg[\sum_{r,s=1,r\neq s}^{p} \Big\{&\hat{\alpha}_{rs}(f)(e_re_r' \otimes (e_se_s' - \tilde{L}_{rs}e_re_s')) \\
&+ \hat{\beta}_{rs}(f)(e_re_s' \otimes (e_se_r' - \tilde{L}_{rs}e_re_r'))\Big\}\Bigg]\mathrm{vec}\,\underline{T}_{\tilde{\vartheta},f}. \qquad \text{(B.13)}
\end{aligned}$$

The identity $(C' \otimes A)(\mathrm{vec}\, B) = \mathrm{vec}(ABC)$ then yields

$$\mathrm{vec}(\hat{L}_f - \tilde{L}) = \frac{1}{\sqrt{n}}\,(I_p \otimes \tilde{L})\,\mathrm{vec}\Bigg[\sum_{r,s=1,r\neq s}^{p} (\hat{N}_f)_{sr}(e_se_r' - \tilde{L}_{rs}e_re_r')\Bigg].$$

Hence, we have

$$\begin{aligned}\hat{L}_f - \tilde{L} &= \frac{1}{\sqrt{n}}\tilde{L}\sum_{r,s=1,r\neq s}^{p}(\hat{N}_f)_{sr}(e_s e_r' - \tilde{L}_{rs}e_r e_r')\\ &= \frac{1}{\sqrt{n}}\tilde{L}\sum_{r,s=1}^{p}(\hat{N}_f)_{sr}(e_s e_r' - \tilde{L}_{rs}e_r e_r') = \frac{1}{\sqrt{n}}\tilde{L}\Big(N_f - \sum_{r,s=1}^{p}\tilde{L}_{rs}(\hat{N}_f)_{sr}e_r e_r'\Big)\\ &= \frac{1}{\sqrt{n}}\tilde{L}\Big(\hat{N}_f - \sum_{r=1}^{p}(\tilde{L}N_f)_{rr}e_r e_r'\Big) = \frac{1}{\sqrt{n}}\tilde{L}(\hat{N}_f - \operatorname{diag}(\tilde{L}N_f)),\end{aligned}$$

which proves the result. □

Proof of Lemma 4.1. In this proof, all stochastic convergences are as $n \to \infty$ under $\mathrm{P}^{(n)}_{\vartheta,g}$. First note that, if $\check{\vartheta} := (\check{\mu}', (\operatorname{vecd}^\circ \check{L})')'$ is an arbitrary locally asymptotically discrete root-n consistent estimator for $\vartheta = (\mu', (\operatorname{vecd}^\circ L)')'$, we then have that

$$\operatorname{vec}(\underline{T}_{\check{\vartheta},f} - \underline{T}_{\vartheta,f}) = -G_{f,g}(I_p \otimes \check{L}^{-1})C'\sqrt{n}\,\operatorname{vecd}^\circ(\check{L} - L) + o_{\mathrm{P}}(1) \tag{B.14}$$

(compare with Corollary A.1). Incidentally, note that (B.14) implies that $\operatorname{vec}\underline{T}_{\check{\vartheta},f}$ is $O_{\mathrm{P}}(1)$ (by proceeding exactly as in the proof of Theorem A.1(i)-(ii), we can indeed show that, under $\mathrm{P}^{(n)}_{\vartheta,g}$, $\operatorname{vec}\underline{T}_{\vartheta,f}$ is asymptotically multinormal, hence stochastically bounded).

Now, from (B.14), we obtain

$$\begin{aligned}&\operatorname{vec}(\underline{T}_{\tilde{\vartheta}^{\gamma_{rs}}_{\lambda},f} - \underline{T}_{\tilde{\vartheta},f})\\ &= -G_{f,g}(I_p \otimes \tilde{L}^{-1})C'\sqrt{n}\operatorname{vecd}^\circ(\tilde{L}^{\gamma_{rs}}_{\lambda} - \tilde{L}) + o_{\mathrm{P}}(1)\\ &= -\lambda(\underline{T}_{\tilde{\vartheta},f})_{rs}G_{f,g}(I_p \otimes \tilde{L}^{-1})C'\operatorname{vecd}^\circ(\tilde{L}e_r e_s' - \tilde{L}\operatorname{diag}(\tilde{L}e_r e_s')) + o_{\mathrm{P}}(1),\end{aligned}$$

which, by using the fact that $C'(\operatorname{vecd}^\circ H) = (\operatorname{vec} H)$ for any $p \times p$ matrix H with only zero diagonal entries, leads to

$$\begin{aligned}&\operatorname{vec}(\underline{T}_{\tilde{\vartheta}^{\gamma_{rs}}_{\lambda},f} - \underline{T}_{\tilde{\vartheta},f})\\ &\quad= -\lambda(\underline{T}_{\tilde{\vartheta},f})_{rs}G_{f,g}(I_p \otimes \tilde{L}^{-1})\operatorname{vec}(\tilde{L}e_r e_s' - \tilde{L}\operatorname{diag}(\tilde{L}e_r e_s')) + o_{\mathrm{P}}(1)\\ &\quad= -\lambda(\underline{T}_{\tilde{\vartheta},f})_{rs}G_{f,g}\operatorname{vec}(e_r e_s' - \operatorname{diag}(\tilde{L}e_r e_s')) + o_{\mathrm{P}}(1).\end{aligned}$$

This yields

$$\begin{aligned}\operatorname{vec}(\underline{T}_{\tilde{\vartheta}^{\gamma_{rs}}_{\lambda},f} - \underline{T}_{\tilde{\vartheta},f}) &= -\lambda(\underline{T}_{\tilde{\vartheta},f})_{rs}G_{f,g}\operatorname{vec}(e_r e_s') + o_{\mathrm{P}}(1)\\ &= -\lambda(\underline{T}_{\tilde{\vartheta},f})_{rs}(\gamma_{rs}(f,g)\operatorname{vec}(e_r e_s') + \rho_{rs}(f,g)\operatorname{vec}(e_s e_r')) + o_{\mathrm{P}}(1).\end{aligned}$$

Premultiplying by $(\underline{T}_{\tilde{\vartheta},f})_{rs}(e_s \otimes e_r)'$, we then obtain

$$(\underline{T}_{\tilde{\vartheta},f})_{rs}(\underline{T}_{\tilde{\vartheta}_\lambda^{\gamma rs},f})_{rs} - ((\underline{T}_{\tilde{\vartheta},f})_{rs})^2 = -\lambda((\underline{T}_{\tilde{\vartheta},f})_{rs})^2\gamma_{rs}(f,g) + o_{\mathrm{P}}(1)$$

(recall indeed that $\underline{T}_{\tilde{\vartheta},f} = O_{\mathrm{P}}(1)$), which establishes the γ-part of the lemma. The proof of the ρ-part follows along the exact same lines, but for the fact that the premultiplication is by $(\underline{T}_{\tilde{\vartheta},f})_{sr}(e_r \otimes e_s)'$. □

ACKNOWLEDGEMENTS

We would like to express our gratitude to the Co-Editor, Professor Peter Bühlmann, an Associate Editor and one referee. Their careful reading of a previous version of the paper and their comments and suggestions led to a considerable improvement of the present paper. We are also grateful to Klaus Nordhausen for sending to us the R code for FastICA authored by Abhijit Mandal.

SUPPLEMENTARY MATERIAL

Supplement: Further results on tests and a proof of Theorem 4.3 (http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). This supplement provides a simple explicit expression for the proposed test statistics, derives local asymptotic powers of the corresponding tests, and presents simulation results for hypothesis testing. It also gives a proof of Theorem 4.3.

REFERENCES

[1] Amari, S. (2002). Independent component analysis and method of estimating functions. *IEICE Trans. Fundamentals Electronics, Communications and Computer Sciences* **E85-A** 540–547.

[2] Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

[3] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Statistical Inference for Semiparametric Models*, Johns Hopkins University Press, Baltimore.

[4] Cardoso, J.-F. (1989). Source Separation Using Higher Moments. *Proceedings of IEEE international conference on acoustics, speech and signal processing* 2109–2112.

[5] Cassart, D., Hallin, M., and Paindaveine, D. (2010). On the estimation of cross-information quantities in R-estimation. In J. Antoch, M. Hušková and P.K. Sen, Editors: *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková*, I.M.S. Monographs-Lecture Notes, 35–45.

[6] Chen, A., and Bickel, P. J. (2006). Efficient independent component analysis. *Ann. Statist.* **34** 2825–2855.

[7] Hallin, M., Oja, H., and Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. II. Optimal R-Estimation of Shape. *Ann. Statist.* **34** 2757–2789.

[8] Hallin, M., and Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape I: Optimal rank-based tests for sphericity. *Ann. Statist.* **34** 2707–2756.
[9] Hallin, M., and Paindaveine, D. (2008). Semiparametrically efficient one-step R-Estimation. Unpublished manuscript.
[10] Hallin, M., Vermandele, C., and Werker, B. J. M. (2006). Serial and nonserial sign-and-rank statistics: asymptotic representation and asymptotic normality. *Ann. Statist.* **34** 254–289.
[11] Hallin, M., and Werker, B. J. M. (2003). Semiparametric efficiency, distribution-freeness, and invariance. *Bernoulli* **9** 55–65.
[12] Hyvärinen, A., and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9** 1483–1492.
[13] Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks* **10** 626–634.
[14] Kreiss, J-P. (1987). On adaptative estimation in stationary ARMA processes. *Ann. Statist.* **15** 112–133.
[15] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.
[16] Ilmonen, P., Nevalainen, J., and Oja, H. (2010). Characteristics of multivariate distributions and the invariant coordinate system. *Statist. Probab. Lett.* **80** 1844–1853.
[17] Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2011). Independent component (IC) functionals and a new performance index. Submitted.
[18] Ilmonen, P., and Paindaveine, D. (2011). Supplement to "Semiparametrically efficient inference based on signed ranks in symmetric independent component models."
[19] Oja, H., Sirkiä, S., and Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian J. Statist.* **35** 175–189.
[20] Oja, H., Paindaveine, D., and Taskinen, S. (2011). Parametric and nonparametric tests for multivariate independence in IC models. Submitted.
[21] Ollila, E. (2010). The deflation-based FastICA estimator: statistical analysis revisited. *IEEE Trans. Signal Processing* **58** 1527–1541.
[22] Ollila, E., and Kim, H.-J. (2011). On testing hypotheses of mixing vectors in the ICA model using FastICA. *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI'11)*, 325–328.
[23] Puri, M. L., and Sen, P. K. (1985). *Nonparametric Methods in General Linear Models.* J. Wiley, New York.
[24] Rao, C. R., and Mitra, S. K. (1971). *Generalized Inverses of Matrices and its Applications*, J. Wiley, New York.
[25] Theis, F. J. (2004). A new concept for separability problems in blind source separation. *Neural Comput.* **16** 1827–1850.

Pauliina Ilmonen
Tampere School of Health Sciences
University of Tampere
FIN-33014 University of Tampere
Finland
E-mail: Pauliina.Ilmonen@uta.fi

Davy Paindaveine
E.C.A.R.E.S.,
and Département de Mathématique
Université libre de Bruxelles
50, Avenue F.D. Roosevelt, CP114/04
B-1050 Brussels, Belgium
E-mail: dpaindav@ulb.ac.be
URL: http://homepages.ulb.ac.be/~dpaindav

SUPPLEMENTARY MATERIAL

Further results on tests and a proof of Theorem 4.3
(doi: ???http://lib.stat.cmu.edu/aos/???/???; .pdf). This supplement provides a simple explicit expression for the proposed test statistics (Section 1), derives local asymptotic powers of the corresponding tests (Section 2), and presents simulation results for hypothesis testing (Section 3). It also gives a proof of Theorem 4.3 (Section 4).

Below, (M-3.1), Page M.9, Section M.3, Lemma M-4.1, etc. refer to Expression (3.1), Page 9, Section 3, Lemma 4.1, etc. from [2]. Unless otherwise stated, other cross-references relate to this supplement itself.

1. Explicit expressions of the proposed test statistics. The following result provides a simple and explicit expression of the signed-rank test statistic $\underline{Q}_f$ in (M-3.1).

THEOREM 1.1. *Fix $f \in \underline{\mathcal{F}}_{\text{ulan}}$. Then the test statistic $\underline{Q}_f$ rewrites*

$$\begin{aligned} \underline{Q}_f &= (\operatorname{vec} \underline{T}_{\hat{\vartheta}_0,f})' M_f (\operatorname{vec} \underline{T}_{\hat{\vartheta}_0,f}) \\ &= \sum_{r,s=1,r\neq s}^{p} \big(\alpha_{rs}(f)(\underline{T}_{\hat{\vartheta}_0,f})^2_{sr} + \beta_{rs}(f)(\underline{T}_{\hat{\vartheta}_0,f})_{rs}(\underline{T}_{\hat{\vartheta}_0,f})_{sr}\big), \end{aligned} \tag{1.1}$$

where we let $\alpha_{rs}(f) = \alpha_{rs}(f,f)$, and $\beta_{rs}(f) = \beta_{rs}(f,f)$ (see (M-4.5)*) and $M_f := \sum_{r,s=1,r\neq s}^{p}(\alpha_{rs}(f)(e_r e_r' \otimes e_s e_s') + \beta_{rs}(f)(e_r e_s' \otimes e_s e_r')$.*

PROOF. Applying to $\underline{Q}_f$ Lemma M-B.1 with $g = f$, we obtain

$$\begin{aligned} \underline{Q}_f &= (\operatorname{vec} \underline{T}_{\hat{\vartheta}_0,f})' \\ &\times \Bigg[\sum_{r,s=1,r\neq s}^{p} \Big\{ \alpha_{rs}(f)\big(e_r e_r' \otimes \big(L_{0rs}^2 e_r e_r' + e_s e_s' - L_{0rs} e_r e_s' - L_{0rs} e_s e_r'\big)\big) \\ &+ \beta_{rs}(f)\big(e_r e_s' \otimes \big(L_{0rs} L_{0sr} e_r e_s' - L_{0rs} e_r e_r' - L_{0sr} e_s e_s' + e_s e_r'\big)\big) \Big\} \Bigg] (\operatorname{vec} \underline{T}_{\hat{\vartheta}_0,f}), \end{aligned}$$

which, as all diagonal entries of $\underline{T}_{\hat{\vartheta}_0,f}$ are equal to zero, indeed yields $\underline{Q}_f = (\operatorname{vec} \underline{T}_{\hat{\vartheta}_0,f})' M_f (\operatorname{vec} \underline{T}_{\hat{\vartheta}_0,f})$. The equality (1.1) then easily follows from the identity $(C' \otimes A)(\operatorname{vec} B) = \operatorname{vec}(ABC)$. □

2. Local asymptotic powers. Theorem M.3.1 allows to compute the asymptotic powers of $\underline{\phi}_f$ under sequences of local alternatives of the form $\mathrm{P}^{(n)}_{\mu,L_0+n^{-1/2}H,g}$, where H is an arbitrary $p \times p$ matrix with zero diagonal entries (only such a H provides a perturbed mixing matrix $L_0 + n^{-1/2}H$ that belongs—for n large enough—to the parameter space $\mathcal{M}_{1p}$). The corresponding asymptotic powers are given by

$$1 - \Psi_{p(p-1)}\big(\chi^2_{p(p-1),1-\alpha}; (\mathrm{vecd}^\circ H)'(\Gamma^*_{L_0,f,g;2})'(\Gamma^*_{L_0,f;2})^{-1}\Gamma^*_{L_0,f,g;2}(\mathrm{vecd}^\circ H)\big),$$

where $\Psi_{p(p-1)}(\,\cdot\,;\delta)$ and $\chi^2_{p(p-1),1-\alpha}$ were defined in Page M.7. By using the fact that $C'(\mathrm{vecd}^\circ H) = (\mathrm{vec}\, H)$ and then applying Lemma M-B.1, the noncentrality parameter above, after painful yet straightforward computations, simplifies to

$$\sum_{r,s=1,r\neq s}^{p} \big(\xi_{rs}(f,g)\,((L_0^{-1}H)_{sr})^2 + \eta_{rs}(f,g)\,(L_0^{-1}H)_{rs}(L_0^{-1}H)_{sr}\big), \tag{2.1}$$

with

$$\xi_{rs}(f,g) = \frac{\gamma_{rs}(f)\gamma^2_{sr}(f,g) + \rho^2_{rs}(f,g)\gamma_{sr}(f) - 2\rho_{rs}(f,g)\gamma_{sr}(f,g)}{\gamma_{rs}(f)\gamma_{sr}(f) - 1}$$

and

$$\eta_{rs}(f,g) = \frac{\rho_{sr}(f,g)(\gamma_{rs}(f)\gamma_{sr}(f,g) - \rho_{rs}(f,g))}{\gamma_{rs}(f)\gamma_{sr}(f) - 1} + \frac{\gamma_{rs}(f,g)(\gamma_{sr}(f)\rho_{rs}(f,g) - \gamma_{sr}(f,g))}{\gamma_{rs}(f)\gamma_{sr}(f) - 1}.$$

At $g = f$, this reduces to $\sum_{r,s=1,r\neq s}^{p}(\gamma_{sr}(f)\,((L_0^{-1}H)_{sr})^2 + (L_0^{-1}H)_{rs}(L_0^{-1}H)_{sr})$. In the simulations of the next section, we will compare the ranking of finite-sample rejection frequencies associated with various tests $\underline{\phi}_f$ with the corresponding theoretical ranking derived from (2.1).

3. Simulations for hypothesis testing. We considered the trivariate case $p = 3$ and concentrated on the particular case for which the null value of L is $L_0 = I_3$. For three trivariate densities of the form $z \mapsto g(z) = g^{(d)}(z) = \prod_{r=1}^{3} g_r^{(d)}(z_r)$, $d \in \{1,2,3\}$, we generated $M = 5,000$ independent random samples $Z_i^{(d,m)} = (Z_{i1}^{(d,m)}, Z_{i2}^{(d,m)}, Z_{i3}^{(d,m)})'$, $i = 1,\ldots,n$, $m = 1,\ldots,M$, of size $n = 500$. The pdfs $g^{(d)}$ have the following marginals:

(i) In Setup $d = 1$, $g_1^{(d)}$, $g_2^{(d)}$ and $g_3^{(d)}$ are the pdfs of the standard normal distribution ($\mathcal{N}$), the Student distribution with 6 degrees of freedom (t_6), and the beta distribution with parameters 3 and 3 ($\beta_{3,3}$), respectively;

(ii) In Setup $d = 2$, $g_1^{(d)}$ is t_6, $g_2^{(d)}$ is $\beta_{3,3}$, and $g_3^{(d)}$ is the pdf of the double-exponential distribution with scale parameter one (d-exp);

(iii) In Setup $d = 3$, $g_1^{(d)}$ is t_6, $g_2^{(d)}$ is d-exp, and $g_3^{(d)}$ is the pdf of the logistic distribution with scale parameter one (log).

We then generated samples of n observations $X_1, \ldots, X_n$ according to

$$X_i^{(d,m)} = (L_0 + a\,\kappa^{(d)} H)\, Z_i^{(d,m)} + \mu, \tag{3.1}$$

with $a = 0, 1, 2, 3, 4$,

$$\begin{pmatrix} \kappa^{(1)} \\ \kappa^{(2)} \\ \kappa^{(3)} \end{pmatrix} = \begin{pmatrix} .002 \\ .007 \\ .0025 \end{pmatrix}, \quad H = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 2 & 0 \end{pmatrix}, \quad \text{and} \quad \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Clearly, these samples correspond to the null hypothesis for $a = 0$ and to increasingly severe alternatives for $a = 1, 2, 3, 4$. The quantities $\kappa^{(d)}$ were chosen in such a way that the rejection frequencies obtained for $a = 4$ were approximately .95 for all d. All samples were subjected, at asymptotic level $\alpha = 5\%$, to the signed-rank tests $\underline{\phi}_{f^{(j)}}$, $j = 1, 2, 3, 4$, where $f^{(j)} = g^{(j)}$ for $j = 1, 2, 3$, and where $f^{(4)}$ uses a t_3 pdf for each marginal density. The first three tests therefore achieve asymptotic optimality in Setups 1 to 3, respectively. In all tests, the location estimate $\hat{\mu}$ used is the componentwise median defined in Page M-9.

Rejection frequencies are plotted against a in the first column of Figure 1. These rejection frequencies indicate that, when based on their asymptotic chi-square critical values, the signed-rank tests are conservative and significantly biased at the sample size considered. In order to remedy this, we also implemented versions of each of the signed-rank procedures based on estimations of the (distribution-free) quantile of the test statistic under known parameter values μ and L_0. These estimations, just as the asymptotic chi-square quantile, are consistent approximations of the corresponding exact quantiles under the null, and were obtained, for each of the four tests above, as the empirical 0.05-upper quantiles $q_{.95}^{(n)}$ of each signed-rank test statistic in a collection of 10^6 simulated multinormal samples, yielding $q_{.95}^{(n)} = 10.34$, 11.56, 10.88, and 9.74, respectively. These bias-corrected critical values are all smaller than the asymptotic chi-square one $\chi^2_{6;.95} = 12.60$, so that the resulting tests are uniformly less conservative than the original ones. The

resulting rejection frequencies are plotted in the second column of Figure 1, where it is readily seen that all tests now are roughly unbiased.

At the sample size $n = 500$, the asymptotic properties derived in Section M.3 do not show so clearly in the simulation results, not only because the signed-rank tests are biased, but also because the test $\underline{\phi}_{f^{(d)}}$ does not seem to be the most powerful one in Setup d. To question correctness of our asymptotic results, we reran the same simulation as above, but now with $n = 10,000$ and with $(\kappa^{(1)}, \kappa^{(2)}, \kappa^{(3)})'$ divided by $\sqrt{10,000/500}$. The resulting simulated critical values are given by $q^{(n)}_{.95} = 11.59$, 12.38, 11.83, and 11.46, respectively, and are all much closer to the asymptotic one $\chi^2_{6;.95} = 12.60$, so that the signed-rank tests, in their asymptotic versions, may only suffer a small bias for this large sample size. Consequently, it is justified to restrict to these asymptotic versions. The corresponding rejection frequencies are plotted in the last column of Figure 1 and confirm, under any $g^{(d)}$, $d = 1, 2, 3$, both the optimality of $\underline{\phi}_{f^{(d)}}$ and—more generally—the whole ranking of the local asymptotic powers of $\underline{\phi}_{f^{(j)}}$, $j = 1, 2, 3, 4$, which can be obtained from (2.1).

Finally, we point out that, for each fixed sample size, setup, and type of critical values considered, the various signed-rank tests exhibit very similar performances. This implies that, just as for point estimation, one should not worry too much about the choice of the target density f in hypothesis testing.

4. Proof of Theorem M-4.3. The proof follows the same scheme as that of Proposition 2.1 in [1]. We report the proof here for the sake of completeness.

PROOF. We fix $\vartheta \in \Theta$, $f \in \underline{\mathcal{F}}_{\text{ulan}}$, $g \in \mathcal{F}_{\text{ulan}}$, and $r \neq s \in \{1, \dots, p\}$, and concentrate on establishing that $\hat{\gamma}_{rs}(f) = \gamma_{rs}(f, g) + o_{\text{P}}(1)$, as $n \to \infty$ under $\text{P}^{(n)}_{\vartheta,g}$ (the proof of the ρ-result is entirely similar). In the sequel, we stress the dependence in n of the various statistics with superscripts $^{(n)}$.

Let us first show that, under $\text{P}^{(n)}_{\vartheta,g}$, $\lambda^{(n)-}_{\gamma_{rs}}$, hence also $\lambda^{(n)+}_{\gamma_{rs}}$, is $O_{\text{P}}(1)$ as $n \to \infty$. Assume therefore it is not: then, there exist $\epsilon > 0$ and a sequence $n_i \nearrow \infty$ such that, for all $\ell \in \mathbb{R}$ and i, $\text{P}^{(n_i)}_{\vartheta,g}[\lambda^{(n_i)-}_{\gamma_{rs}} > \ell] > \epsilon$. This implies, for arbitrarily large ℓ, that $\text{P}^{(n_i)}_{\vartheta,g}[h^{(n_i)\gamma_{rs}}(\ell) > 0] > \epsilon$, hence, in view of Lemma M-4.1,

$$\text{P}^{(n_i)}_{\vartheta,g}\left[(1 - \ell\gamma_{rs}(f, g))\, h^{(n_i)\gamma_{rs}}(0) + \zeta^{(n_i)} > 0\right] > \epsilon$$

for all i, where $\zeta^{(n)}$, $n \in \mathbb{N}$ is some $o_{\text{P}}(1)$ sequence. For $\ell > (\gamma_{rs}(f, g))^{-1}$,

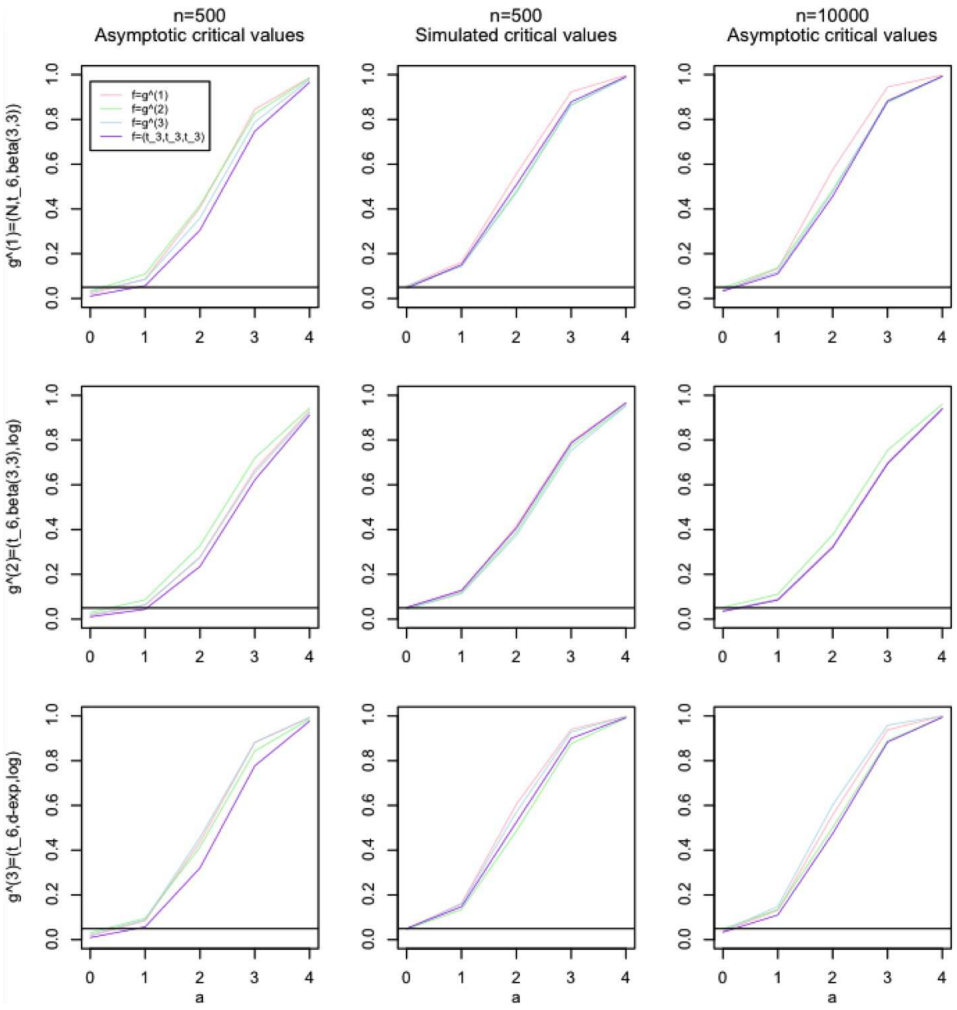

FIG 1. *Rejection frequencies (out of $M = 5,000$ replications), under the null ($a = 0$) and increasingly severe alternatives ($a = 1, 2, 3, 4$), of the signed-rank tests $\underline{\phi}_{f^{(j)}}$, $j = 1, 2, 3, 4$; see Section 3 for details. The sample size is $n = 500$ in both first columns and $n = 10,000$ in the third one. In the first and third columns, tests are based on their asymptotic null distribution, whereas the second column uses simulated critical values, obtained from 10^6 standard multinormal samples.*

this entails, for all i,

$$\mathrm{P}_{\vartheta,g}^{(n_i)}\big[0 < h^{(n_i)\gamma_{rs}}(0) < (\ell\gamma_{rs}(f,g)-1)^{-1}|\zeta^{(n_i)}|\big] > \epsilon,$$

which contradicts (M-4.9). It follows that $\lambda_{\gamma_{rs}}^{(n)-}$ is $O_{\mathrm{P}}(1)$ under $\mathrm{P}_{\vartheta,g}^{(n)}$.

By using again (M-4.9), there exist, for all $\eta > 0$, a positive real number δ_η and an integer N_η such that

$$\mathrm{P}_{\vartheta,g}^{(n)}\big[h^{(n)\gamma_{rs}}(0) \geq \delta_\eta\big] \geq 1 - \frac{\eta}{2}$$

for all $n \geq N_\eta$. Since $\lambda_{\gamma_{rs}}^{(n)-}$ and $\lambda_{\gamma_{rs}}^{(n)+}$ are $O_{\mathrm{P}}(1)$, Lemma M-4.1 implies that, for all $\eta > 0$ and $\varepsilon > 0$, there exists an integer $N_{\varepsilon,\delta} \geq N_\eta$ such that, for all $n \geq N_{\varepsilon,\delta}$ (with $\lambda_{\gamma_{rs}}^{(n)\pm}$ standing for either $\lambda_{\gamma_{rs}}^{(n)-}$ or $\lambda_{\gamma_{rs}}^{(n)+}$),

$$\mathrm{P}_{\vartheta,g}^{(n)}\big[(1-\lambda_{\gamma_{rs}}^{(n)\pm}\gamma_{rs}(f,g))h^{(n)\gamma_{rs}}(0) \in [h^{(n)\gamma_{rs}}(\lambda_{\gamma_{rs}}^{(n)\pm}) \pm \varepsilon]\big] \geq 1 - \frac{\eta}{2}.$$

It follows that for all $\eta > 0$, $\varepsilon > 0$ and $n \geq N_{\varepsilon,\delta}$, letting $\delta = \delta_\eta$,

$$\begin{aligned}\mathrm{P}_{\vartheta,g}^{(n)}\big[A_{\varepsilon,\delta}^{(n)}\big] := \mathrm{P}_{\vartheta,g}^{(n)}\Big[&(1-\lambda_{\gamma_{rs}}^{(n)\pm}\gamma_{rs}(f,g))h^{(n)\gamma_{rs}}(0) \in [h^{(n)\gamma_{rs}}(\lambda_{\gamma_{rs}}^{(n)\pm}) \pm \varepsilon] \\ &\qquad \text{and } h^{(n)\gamma_{rs}}(0) \geq \delta\Big] \\ &\geq 1-\eta.\end{aligned}$$

Next, denote by $\hat{D}^{(n)}$, $D^{(n)}$, and $D_\pm^{(n)}$ the graphs of the mappings

$$\begin{aligned}\lambda &\mapsto h^{(n)\gamma_{rs}}(\lambda_{\gamma_{rs}}^{(n)-}) - c(\lambda - \lambda_{\gamma_{rs}}^{(n)-})(h^{(n)\gamma_{rs}}(\lambda_{\gamma_{rs}}^{(n)-}) - h^{(n)\gamma_{rs}}(\lambda_{\gamma_{rs}}^{(n)+})) \\ \lambda &\mapsto (1-\lambda\gamma_{rs}(f,g))\, h^{(n)\gamma_{rs}}(0),\end{aligned}$$

and

$$\lambda \mapsto (1-\lambda\gamma_{rs}(f,g))\, h^{(n)\gamma_{rs}}(0) \pm \epsilon,$$

respectively. These graphs take the form of four random straight lines, intersecting the horizontal axis at $\lambda_{\gamma_{rs}}^{(n)}$ (our estimator of $(\gamma_{rs}(f,g))^{-1}$), $\lambda_0 := (\gamma_{rs}(f,g))^{-1}$, $\lambda_0^{(n)+}$ and $\lambda_0^{(n)-}$, respectively. Since $D_\pm^{(n)}$ and $D^{(n)}$ are parallel, with a negative slope, we have that $\lambda_0^{(n)-} \leq \lambda_0 \leq \lambda_0^{(n)+}$. Under $A_{\varepsilon,\delta}^{(n)}$, that common slope has absolute value at least $\delta\gamma_{rs}(f,g)$, which implies that $\lambda_0^{(n)+} - \lambda_0^{(n)-} \leq \frac{2\varepsilon}{\delta\gamma_{rs}(f,g)}$. Still under $A_{\varepsilon,\delta}^{(n)}$, for λ values between $\lambda_{\gamma_{rs}}^{(n)-}$

and $\lambda_{\gamma_{rs}}^{(n)+}$, $\hat{D}^{(n)}$ is lying between $D_-^{(n)}$ and $D_+^{(n)}$, which entails $\lambda_0^{(n)-} \leq \lambda_{\gamma_{rs}}^{(n)} \leq \lambda_0^{(n)+}$.

Summing up, for all $\eta > 0$ and $\varepsilon > 0$, there exist $\delta = \delta_\eta > 0$, and $N = N_{\varepsilon\gamma_{rs}(f,g)\delta/2,\delta}$ such that, for any $n \geq N$, with $\mathrm{P}_{\vartheta,g}^{(n)}$ probability larger than $1-\eta$, $|\lambda_{\gamma_{rs}}^{(n)} - \lambda_0| \leq \lambda_0^{(n)+} - \lambda_0^{(n)-} \leq \varepsilon$. □

References.

[1] Cassart, D., Hallin, M., and Paindaveine, D. (2010). On the estimation of cross-information quantities in R-estimation. In J. Antoch, M. Hušková and P.K. Sen, Editors: *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková*, I.M.S. Monographs-Lecture Notes, 35–45.

[2] Ilmonen, P., and Paindaveine, D. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models.

DEFLATION-BASED FASTICA RELOADED

Klaus Nordhausen[†], *Pauliina Ilmonen*[†], *Abhijit Mandal*[†], *Hannu Oja*[†] *and Esa Ollila*[‡,⋆]

[†]School of Health Sciences,
University of Tampere
FIN-33014 Tampere, Finland

[‡]Signal Processing and Acoustics
Aalto University
FIN-00076 Aalto, Finland

[⋆]Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA

ABSTRACT

Deflation-based FastICA, where independent components (IC's) are extracted one-by-one, is among the most popular methods for estimating an unmixing matrix in the independent component analysis (ICA) model. In the literature, it is often seen rather as an algorithm than an estimator related to a certain objective function, and only recently has its statistical properties been derived. One of the recent findings is that the order, in which the independent components are extracted in practice, has a strong effect on the performance of the estimator. In this paper we review these recent findings and propose a new "reloaded" procedure to ensure that the independent components are extracted in an optimal order. The reloaded algorithm improves the separation performance of the deflation-based FastICA estimator as amply illustrated by our simulation studies. Reloading also seems to render the algorithm more stable.

1. INTRODUCTION

The independent component (IC) model is a semiparametric model which has gained increasing interest in various fields of science and engineering during the recent years [6]. The basic IC model assumes that the observed p-variate random vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ is a linear mixture of the p mutually independent sources (IC's) $\mathbf{s} = (s_1, \ldots, s_p)^T$. Then

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{A}$ is assumed to be a full rank $p \times p$ unknown mixing matrix. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote a random sample from the IC model (1). The aim of the independent component analysis (ICA) is to find an estimate $\hat{\mathbf{W}}$ (using the random sample $\mathbf{X}$) of some $p \times p$ unmixing matrix $\mathbf{W}$ verifying $\mathbf{s} = \mathbf{W}\mathbf{x}$ up to permutation, sign and scale changes; see [6]. Naturally $\mathbf{W} = \mathbf{A}^{-1}$ is one possible solution.

In the following, $\mathbf{P}$ denotes a permutation matrix (obtained by permuting the rows or columns of $\mathbf{I}_p$), $\mathbf{J}$ denotes a sign-chance matrix (a $p \times p$ diagonal matrix with entries ± 1), and $\mathbf{D}$ denotes a $p \times p$ diagonal matrix with positive diagonal elements. Let $\mathscr{G}$ denote the set of all full-rank $p \times p$ matrices. Then the set of $p \times p$ matrices, defined as

$$\mathscr{C} = \{\mathbf{C} : \mathbf{C} = \mathbf{P}\mathbf{J}\mathbf{D} \text{ for some } \mathbf{P}, \mathbf{J} \text{ and } \mathbf{D}\},$$

is a subset of $\mathscr{G}$. If a matrix $\mathbf{W} \in \mathscr{G}$ is an unmixing matrix in the IC model (1), then so is $\mathbf{C}\mathbf{W}$ for any $\mathbf{C} \in \mathscr{C}$. We then say that two unmixing matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are (ICA) equivalent if $\mathbf{W}_1 = \mathbf{C}\mathbf{W}_2$ for some $\mathbf{C} \in \mathscr{C}$, and we write $\mathbf{W}_1 \sim \mathbf{W}_2$.

All reasonable estimates $\hat{\mathbf{W}}$ should naturally converge in probability to some population value $\mathbf{W}(F_{\mathbf{x}})$, that is, the value of an independent component (IC) functional $\mathbf{W}$ at $F_{\mathbf{x}}$, where $F_{\mathbf{x}}$ denotes the cumulative distribution function (cdf) of $\mathbf{x}$. A formal (model independent) definition [9] of an IC functional is given below.

Definition 1. *Let $F_{\mathbf{x}}$ denote the cdf of $\mathbf{x}$. The functional $\mathbf{W}(F_{\mathbf{x}}) \in \mathscr{G}$ is an IC functional in the IC model (1) if (i) $\mathbf{W}(F_{\mathbf{x}})\mathbf{A} \sim \mathbf{I}_p$ and (ii) it is affine equivariant in the sense that $\mathbf{W}(F_{\mathbf{B}\mathbf{x}}) = \mathbf{W}(F_{\mathbf{x}})\mathbf{B}^{-1}$ for all $\mathbf{B} \in \mathscr{G}$.*

Note that $\mathbf{W}(F_{\mathbf{B}\mathbf{x}})\mathbf{B}\mathbf{x} = \mathbf{W}(F_{\mathbf{x}})\mathbf{x}$, and therefore $\mathbf{W}(F_{\mathbf{x}})\mathbf{x}$ is invariant under invertible linear transformations of the observation vectors. A finite sample estimator corresponding to an IC functional is obtained if the functional is applied to the empirical distribution based on $\mathbf{X}$. We then write $\hat{\mathbf{W}} = \mathbf{W}(\mathbf{X})$ for the obtained estimator. The estimator is then also affine equivariant in the sense that $\hat{\mathbf{W}}(\mathbf{B}\mathbf{X}) = \hat{\mathbf{W}}(\mathbf{X})\mathbf{B}^{-1}$. Let us denote by $\mathbf{S}(F_{\mathbf{x}}) \equiv \mathrm{COV}(\mathbf{x})$ the covariance matrix (functional) of a random vector $\mathbf{x}$. We note that many IC functionals proposed in the literature are defined either implicitly or explicitly in such a way that the covariance matrix of the obtained source vector is equal to the identity matrix, *i.e.* $\mathrm{COV}(\mathbf{W}(F_{\mathbf{x}})\mathbf{x}) = \mathbf{I}_p$, in which case $\mathbf{W}(F_{\mathbf{x}}) = \mathbf{U}(F_{\mathbf{x}})\mathbf{S}^{-1/2}(F_{\mathbf{x}})$, where $\mathbf{U}(F_{\mathbf{x}})$ is an orthogonal matrix.

The estimator of interest in this paper is the deflation-based FastICA estimator [4, 5]. The paper is organized as follows. Sections 2 recalls the deflation-based FastICA algorithm and estimating equations, while statistical properties of the estimator are discussed in Sections 3. In Section 4, a new novel method is proposed, called the reloaded FastICA, to optimize the extraction order of the sources in succeeding FastICA deflation stages. A Simulation study in Section 5 illustrates the usefulness of our approach, whereas Section 6 presents our conclusions.

2. DEFLATION-BASED FASTICA

Deflation-based FastICA, hereafter FastICA for short, was introduced in [4] and further developed in [5]. Up to date it can be considered among one of the most popular methods to solve the ICA problem.

2.1 FastICA algorithm

Write $\mathbf{z} = \mathbf{S}^{-1/2}(F_{\mathbf{x}})(\mathbf{x} - \mathrm{E}(\mathbf{x}))$ for the whitened random variable, where the square root matrix is chosen to be symmetric. FastICA can be seen as a projection pursuit method, where the directions $\mathbf{u}_k$, maximizing a measure of non-Gaussianity $|\mathrm{E}\left(G(\mathbf{u}_k^T\mathbf{z})\right)|$, are found successively under the constraint that $\mathbf{u}_k$ is orthonormal with the previously found directions $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ (for $k = 1, \ldots, p-1$), where $G(\cdot)$

can be any twice continuously differentiable nonlinear and nonquadratic function with $G(0) = 0$. The unmixing matrix is then $\mathbf{W} = \mathbf{U}\mathbf{S}^{-1/2}$ where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)^T$. Note that the last vector $\mathbf{u}_p$ is set as a unit vector orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_{p-1}$. Let $g(\cdot)$ denote the derivative of $G(\cdot)$, called the nonlinearity. Commonly used nonlinearities are *pow3*: $g(u) = u^3$, *tanh*: $g(u) = \tanh(a_1 u)$, *gaus*: $g(u) = u\exp(-a_2 u^2/2)$ and *skew*: $g(u) = u^2$, where a_1 and a_2 are tuning parameters, usually chosen to be equal to 1.

Due to the whitening, the FastICA method is commonly formulated as an algorithm for finding an estimator $\hat{\mathbf{U}}$. The algorithm (and its slight variations) given below for the directions $\hat{\mathbf{u}}_k$, $k = 1, \ldots, p-1$, is generally accepted in the literature. In the algorithm, $\hat{\mathbf{u}}_j$, $j = 1 \ldots, k-1$, are the previously found directions and the sample mean vector and the sample covariance matrix are denoted by $\bar{\mathbf{x}}$ and $\hat{\mathbf{S}}$, respectively.

Algorithm 1 deflation-based FastICA algorithm for $\hat{\mathbf{u}}_k$

$\mathbf{x}_i \leftarrow \hat{\mathbf{S}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$ {Whiten the data}
$\mathbf{u}_{k,0} \leftarrow \mathbf{u}_{k,init}$ {Choose an initial value}
$\Delta = \infty$
while $\varepsilon < \Delta$ **do**
  $\mathbf{u}_{k,1} \leftarrow \text{ave}(\mathbf{x}_i g(\mathbf{u}_{k,0}^T \mathbf{x}_i)) - \text{ave}(g'(\mathbf{u}_{k,0}^T \mathbf{x}_i))\mathbf{u}_{k,0}$
  $\mathbf{u}_{k,1} \leftarrow \mathbf{u}_{k,1} - \sum_{j=1}^{k-1}(\mathbf{u}_{k,1}^T \hat{\mathbf{u}}_j)\hat{\mathbf{u}}_j$
  $\mathbf{u}_{k,1} \leftarrow \mathbf{u}_{k,1}/\|\mathbf{u}_{k,1}\|$
  $\Delta = \|\mathbf{u}_{k,1} - \mathbf{u}_{k,0}\|$
  $\mathbf{u}_{k,0} \leftarrow \mathbf{u}_{k,1}$
end while
RETURN $\hat{\mathbf{u}}_k = \mathbf{u}_{k,1}$

The FastICA estimator of the unmixing matrix is thus $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{S}}^{-1/2}$ with $\hat{\mathbf{U}}$ coming from the algorithm. The order in which the sources are found depends heavily on the initial value $\mathbf{U}_{init} = (\mathbf{u}_{1,init}, \ldots, \mathbf{u}_{p,init})^T$. Write next $\mathbf{W}(\mathbf{U}, \mathbf{X})$ for the estimate based on the data $\mathbf{X}$ and the initial value $\mathbf{U}_{init} = \mathbf{U}$. If $\mathbf{U}$ is random, then the estimate $\mathbf{W}(\mathbf{U}, \mathbf{X})$ may get $p!$ different values depending on random $\mathbf{U}$, and the different solutions may not be ICA equivalent.

Let $\mathbf{S}(\mathbf{X})$ be the covariance matrix computed from $\mathbf{X}$. It is well known that $\mathbf{S}(\mathbf{B}\mathbf{X})^{-1/2}(\mathbf{B}\mathbf{X}) = \mathbf{V}_\mathbf{B}\mathbf{S}(\mathbf{X})^{-1/2}\mathbf{X}$ where $\mathbf{V}_B$ is an orthogonal matrix depending on $\mathbf{B}$ (and $\mathbf{X}$). With a fixed choice $\mathbf{U}$, the estimate $\mathbf{W}(\mathbf{U}, \mathbf{X})$ is affine equivariant in the sense that $\mathbf{W}(\mathbf{U}, \mathbf{B}\mathbf{X}) = \mathbf{W}(\mathbf{U}, \mathbf{X})\mathbf{B}^{-1}$ if $\mathbf{W}(\mathbf{U}, \mathbf{X}) = \mathbf{W}(\mathbf{U}\mathbf{V}_\mathbf{B}, \mathbf{X})$, that is, if $\mathbf{W}(\mathbf{U}, \mathbf{X})$ and $\mathbf{W}(\mathbf{U}\mathbf{V}_\mathbf{B}, \mathbf{X})$ find the sources in the same order. (The equalities above are up to sign changes of the rows.) A natural question then is: Is there any choice $\mathbf{U}_{init} = \mathbf{U}(\mathbf{X})$ such that the "reloaded" fastICA estimate $\mathbf{W}(\mathbf{U}(\mathbf{X}), \mathbf{X})$ is fully affine equivariant. We answer this question in Section 4.

2.2 Estimating equations

To facilitate statistical analysis, it is appropriate to formulate the method as an estimator verifying a set of estimating equations. Furthemore, it is useful to formulate the estimator without the pre-whitening stage. Let $\mathbf{T}(F_\mathbf{x}) = \mathrm{E}(\mathbf{x})$ denote the mean vector (functional). The deflation-based FastICA functional $\mathbf{w}_k(F_\mathbf{x})$, $k = 1, \ldots, p-1$, may be seen [11, 12] as an optimizer of

$$|\,\mathrm{E}[G(\mathbf{w}_k^T(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))]\,|$$

under the constraints (i) $\mathbf{w}_k^T \mathbf{S}(F_\mathbf{x})\mathbf{w}_k = 1$ and (ii) $\mathbf{w}_j^T \mathbf{S}(F_\mathbf{x})\mathbf{w}_k = 0$ for $j = 1, \ldots, k-1$. (For $\mathbf{w}_1$, only the first constraint is needed.) Note that, for the definition of the functional $\mathbf{w}_k$, we need functionals $\mathbf{T}$, $\mathbf{S}$, and $\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$.

Using the Lagrange multiplier technique, one can easily show [9, 12] that (under general assumptions) the unmixing matrix functional $\mathbf{W}(F_\mathbf{x}) = (\mathbf{w}_1(F_\mathbf{x}), \ldots, \mathbf{w}_p(F_\mathbf{x}))^T$ satisfies the p estimating equations

$$\mathrm{E}\Big[g\big(\mathbf{w}_k^T(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\big)(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\Big]$$
$$= \mathbf{S}(F_\mathbf{x}) \sum_{j=1}^{k} \mathbf{w}_j \mathbf{w}_j^T \mathrm{E}\Big[g\big(\mathbf{w}_k^T(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\big)(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\Big],$$

$k = 1, \ldots, p$. Note that, if $\mathbf{s} = \mathbf{W}\mathbf{x}$ has independent components, then $\mathbf{W}$ solves the estimating equations. It is also important to note that, for all permutation matrices $\mathbf{P}$, also $\mathbf{P}\mathbf{W}$ then solves the estimating equations, and therefore the estimating equations do not fix the order of the unmixing vectors $\mathbf{w}_1, \ldots, \mathbf{w}_p$.

3. STATISTICAL PROPERTIES

Despite being such a popular tool, rigorous statistical analysis of the deflation-based FastICA estimator has not been given until quite recently in [9, 11–13]. In this section we discuss the limiting distribution and robustness properties of the deflation-based FastICA estimator. Without loss of generality we assume that $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$, $\mathrm{COV}(\mathbf{x}_i) = \mathbf{I}_p$, and the true mixing matrix is $\mathbf{A} = \mathbf{I}_p = (\mathbf{e}_1, \ldots, \mathbf{e}_p)^T$.

3.1 Limiting distribution

If the first four moments of $\mathbf{s}$ exist, then by the central limit theorem, the joint distribution of $\sqrt{n}\bar{\mathbf{x}}$ and $\sqrt{n}\mathrm{vec}(\hat{\mathbf{S}} - \mathbf{I}_p)$ is asymptotically normal. Furthermore, the existence of the expected values $\mu_{g,k} = \mathrm{E}[g(\mathbf{e}_k^T \mathbf{x}_i)]$,

$$\sigma_{g,k}^2 = Var[g(\mathbf{e}_k^T \mathbf{x}_i)], \quad \lambda_{g,k} = \mathrm{E}[g(\mathbf{e}_k^T \mathbf{x}_i)\mathbf{e}_k^T \mathbf{x}_i]$$

and

$$\delta_{g,k} = \mathrm{E}[g'(\mathbf{e}_k^T \mathbf{x}_i)], \quad \tau_{g,k} = \mathrm{E}[g'(\mathbf{e}_k^T \mathbf{x}_i)\mathbf{e}_k^T \mathbf{x}_i]$$

are required. We also need to assume that $\delta_{g,k} \neq \lambda_{g,k}$, $k = 1, \ldots, p-1$, and we write

$$\alpha_{g,k} = \frac{\sigma_{g,k}^2 - \lambda_{g,k}^2}{(\lambda_{g,k} - \delta_{g,k})^2}, \quad k = 1, \ldots, p. \tag{2}$$

Write $\mathbf{T}_k = \frac{1}{n}\sum_{i=1}^n (g(\mathbf{e}_k^T \mathbf{x}_i) - \mu_{g,k})\mathbf{x}_i$ and $\hat{\mathbf{T}}_k = \frac{1}{n}\sum_{i=1}^n g(\hat{\mathbf{w}}_k^T(\mathbf{x}_i - \bar{\mathbf{x}}))(\mathbf{x}_i - \bar{\mathbf{x}})$. Then, under general assumptions and using Taylor's expansion, we get

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{T}}_k - \lambda_{g,k}\mathbf{e}_k) &= \sqrt{n}\mathbf{T}_k - \tau_{g,k}\mathbf{e}_k\mathbf{e}_k^T\sqrt{n}\bar{\mathbf{x}} \\ &+ \Delta_{g,k}\sqrt{n}(\hat{\mathbf{w}}_k - \mathbf{e}_k) + o_P(1), \end{aligned} \tag{3}$$

where $\Delta_{g,k} = \mathrm{E}[g'(\mathbf{e}_k^T \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T]$.
Now recall that the FastICA unmixing matrix estimator $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_p)^T$ needs to verify the estimating equations

$$\hat{\mathbf{T}}_k = \hat{\mathbf{S}}[\hat{\mathbf{w}}_1\hat{\mathbf{w}}_1^T + ... + \hat{\mathbf{w}}_k\hat{\mathbf{w}}_k^T]\hat{\mathbf{T}}_k, \quad k = 1, \ldots, p. \tag{4}$$

But then

$$(\mathbf{I}_p - \mathbf{U}_k)\sqrt{n}(\hat{\mathbf{T}}_k - \lambda_{g,k}\mathbf{e}_k) = \lambda_{g,k}[\sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p)\mathbf{e}_k + \sum_{j=1}^{k} \mathbf{e}_j\mathbf{e}_k^T\sqrt{n}(\hat{\mathbf{w}}_j - \mathbf{e}_j) + \sqrt{n}(\hat{\mathbf{w}}_k - \mathbf{e}_k)] + o_P(1),$$

where $\mathbf{U}_k = \sum_{j=1}^{k} \mathbf{e}_j\mathbf{e}_j^T$, and, using (3), we get the following result.

Theorem 1. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from the IC model (1) with $\mathbf{A} = \mathbf{I}_p$, $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$, and $\mathrm{COV}(\mathbf{x}_i) = \mathbf{I}_p$. Let $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_p)^T$ be the solution for the estimating equations in (4) such that $\hat{\mathbf{W}} \to_P \mathbf{I}_p$. Then, under the general assumptions,*

$$\begin{aligned}\sqrt{n}\hat{\mathbf{w}}_{kl} &= \frac{1}{\lambda_{g,k} - \delta_{g,k}}\left[\mathbf{e}_l^T\sqrt{n}\mathbf{T}_k - \lambda_{g,k}\sqrt{n}\hat{\mathbf{S}}_{kl}\right] \\ &+ o_P(1) \quad \textit{for } l > k, \\ \sqrt{n}\hat{\mathbf{w}}_{kl} &= -\sqrt{n}\hat{\mathbf{w}}_{lk} - \sqrt{n}\hat{\mathbf{S}}_{kl} + o_P(1) \quad \textit{for } l < k,\end{aligned}$$

and

$$\sqrt{n}(\hat{\mathbf{w}}_{kk} - 1) = -\frac{1}{2}\sqrt{n}(\hat{\mathbf{S}}_{kk} - 1) + o_P(1).$$

Remark 1. *It follows from Theorem 1 that, for $\mathbf{A} = \mathbf{I}_p$, the asymptotic covariance matrix (ASV) of the k-th source $\hat{\mathbf{w}}_k$ is*

$$ASV(\hat{\mathbf{w}}_k) = \sum_{j=1}^{k-1}(\alpha_{g,j} + 1)\mathbf{e}_j\mathbf{e}_j^T + \kappa_k\mathbf{e}_k\mathbf{e}_k^T + \alpha_{g,k}\sum_{l=k+1}^{p}\mathbf{e}_l\mathbf{e}_l^T.$$

where $\kappa_k = (\mathrm{E}(x_{ik}^4) - 1)/4$ and $\alpha_{g,j}$ is defined in (2)*. We note that this result is in accordance with [12, Corollary 1]. Note that the asymptotic variances of the diagonal elements of $\hat{\mathbf{W}}$ do not depend on the choice of the function $g(\cdot)$, but only on the kurtosis of the corresponding source.*

Remark 2. *Theorem 1 implies that, if $\sqrt{n}\mathbf{T}_k$, $k = 1, ..., p$, and $\sqrt{n}vec(\hat{\mathbf{S}} - \mathbf{I}_p)$ have a joint limiting multivariate distribution, the limiting distribution of $\sqrt{n}vec(\hat{\mathbf{W}} - \mathbf{I}_p)$ is also multivariate normal. Interestingly, the limiting distributions of the estimated directions $\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_p$ depend on the order in which they are found; see [12] for details and illustrations. The initial value $\mathbf{U}_{init}$ in the FastICA algorithm mainly determines the order of the extracted sources in practice and hence plays a crucial role in the performance of the estimator.*

3.2 Robustness

Due to the different options for the nonlinearity function $g(\cdot)$, FastICA is often called robust when used with 'robust' nonlinearity functions, for example, *tanh* or *gaus* function. The influence function (IF) of the FastICA functional $\mathbf{w}_k$, $k = 1, \ldots, p$, in the IC model (1) is given in [12] as

$$\mathrm{IF}(\mathbf{z}; \mathbf{w}_k, F) = -p_k\sum_{j=1}^{k-1}(q_j + p_j)\mathbf{w}_j - \frac{p_k^2 - 1}{2}\mathbf{w}_k + q_k\sum_{l=k+1}^{p} p_l\mathbf{w}_l,$$

where $p_k = \mathbf{w}_k^T(\mathbf{z} - \mathrm{E}(\mathbf{x}))$ and

$$q_k = \frac{g(p_k) - \mu_{g,k} - \lambda_{g,k}p_k}{\lambda_{g,k} - \delta_{g,k}}.$$

Since the IF is a weighted sum of the sources $\mathbf{w}_1, \ldots, \mathbf{w}_p$, where the weights are unbounded functions of p_k, any large value of p_j, $j = 1, \ldots, p$ can have unbounded impact on $\mathbf{w}_k$ - irrelevant of the choice of the nonlinearity $g(\cdot)$. Thus, according to its IF, the deflation-based FastICA will never be robust - independently of the choice of $g(\cdot)$ (see [12] for details).

Note also that it is not straightforward to robustify deflation-based FastICA by replacing mean vector and covariance matrix with their more robust counterparts as reported in [1].

4. RELOADING FASTICA BY OPTIMIZING THE EXTRACTION ORDER

In this section, we first discuss the properties of the performance index *MD* for the ICA estimates, and show how it is connected to the asymptotic distribution of the estimate. We then suggest a two-step modified FastICA procedure which optimizes the extraction order.

4.1 Minimum distance performance criterion

Many different performance measures for the IC estimates have been suggested in the literature, see, for example, [10]. In this paper we use the so called minimum distance (*MD*) measure which was recently suggested in [8,9]. The measure is defined as

$$MD(\hat{\mathbf{W}}, \mathbf{A}) = \frac{1}{\sqrt{p-1}}\inf_{\mathbf{C}\in\mathscr{C}}\|\mathbf{C}\hat{\mathbf{W}}\mathbf{A} - \mathbf{I}_p\|.$$

This index is independent of the model specification and surprisingly easy to compute in practice (for details see [8, 9]). The asymptotic behavior of the index *MD* is as follows. If an equivariant estimator $\hat{\mathbf{W}}$ satisfies $\sqrt{n}\mathrm{vec}(\hat{\mathbf{W}} - \mathbf{I}_p) \to_d N_{p^2}(0, \Sigma)$, then

$$nMD^2(\hat{\mathbf{W}}, \mathbf{A}) = \frac{n}{p-1}\|\mathrm{off}(\hat{\mathbf{W}})\|^2 + o_P(1),$$

and the limiting distribution of $nMD^2(\hat{\mathbf{W}}, \mathbf{A})$ is that of a weighted sum of independent chi-square variables [9]. Also, the expected value $n(p-1)\mathrm{E}[MD^2(\hat{\mathbf{W}}, \mathbf{A})]$ converges to the sum of the limiting variances of the off-diagonal elements of $\hat{\mathbf{W}}$ as $n \to \infty$.

4.2 Reloaded FastICA

In order to achieve optimal performance in terms of the *MD* measure, we thus should minimize the sum of the variances of the off-diagonal elements of the FastICA estimator. Using Remark 1 it is easy to see that, for $\mathbf{A} = \mathbf{I}_p$,

$$\sum_{i\neq j} ASV(\hat{\mathbf{w}}_{ij}) = 2\sum_{i=1}^{p}(p-i)\alpha_{g,i} + \frac{p(p-1)}{2},$$

which is minimized if the $\alpha_{g,i}$'s are in the increasing order of magnitude.

To optimize the performance of the deflation-based FastICA, we therefore suggest the following simple procedure.

$g(\cdot)$	$\alpha_{g,E}$	$\alpha_{g,C}$	$\alpha_{g,L}$
pow3	5	15	6
tanh	3.14	32.13	2.01

Table 1: The theoretical values of $\alpha_{g,k}$ for different cases.

$g(\cdot)$	LCE	LEC	CEL	ECL	CLE	ELC
pow3	57	37	73	53	75	35
tanh	75.32	17.33	137.79	79.80	135.55	19.57

Table 2: The limiting values of $n(p-1)\mathrm{E}[MD^2(\hat{\mathbf{W}},\mathbf{A})]$ for the six different extraction orders.

1. Find any equivariant and consistent estimate $\hat{\mathbf{W}}_0$ (e.g. FOBI [2]) such that $\hat{\mathbf{S}}(\hat{\mathbf{W}}_0\mathbf{X}) = \mathbf{I}_p$.
2. Find the estimated sources $\hat{\mathbf{Z}} = \hat{\mathbf{W}}_0(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)$.
3. Find estimates $\hat{\alpha}_{g,k}$, $k = 1,\ldots,p$, based on $\hat{\mathbf{Z}}$ by replacing the expected values by averages in (2).
4. Find the permutation matrix $\hat{\mathbf{P}}$ such that, for the permutated sources, the $\hat{\alpha}_{g,k}$ are in an increasing order.
5. Reload FastICA algorithm 1 with a new initial value: The estimate is $\mathbf{W}(\mathbf{U}(\mathbf{X}),\mathbf{X})$ where $\mathbf{U}(\mathbf{X}) = \hat{\mathbf{P}}\hat{\mathbf{W}}_0\hat{\mathbf{S}}^{1/2}$.

It is easy to see that $\mathbf{W}(\mathbf{U}(\mathbf{X}),\mathbf{X})$ is fully affine equivariant. We conjecture that this new estimator has the same limiting distribution as the simple FastICA estimator which extracts the sources in the (same) optimal order.

5. SIMULATION STUDY

We performed a small simulation study to demonstrate the effect of the extraction order of the sources. We show that reloading FastICA with the data whitened in a new way and with an initial value $\mathbf{U}_{init} = \mathbf{I}_p$ gives the optimal performance among different deflation-based FastICA procedures. The data used in our simulations comes from a three-variate distribution; the independent source distributions are (i) the exponential distribution, (ii) the chi-square distribution with 8 degrees of freedom, and (iii) the Laplace distribution. All three distributions are centered and scaled to have expected value 0 and variance 1. The mixing matrix used in our simulations is $\mathbf{A} = \mathbf{I}_3$. We denote the three sources as E, C, and L, respectively, and the sequence ECL, for example, means the extraction order exponential-chi-square-Laplace. We considered two nonlinearity functions $g = pow3$ and $g = tanh$. The values of corresponding $\alpha_{g,k}$, given in Table 1, were obtained from (2), where the expectations were calculated using numerical integration.

The expected values of $n(p-1)\mathrm{E}[MD^2(\hat{\mathbf{W}},\mathbf{A})]$ for different extraction orders are given in Table 2. The table clearly shows that the extraction order has a large impact on the separation performance. The best extraction order naturally depends on the choice of the nonlinearity function g. Here ELC is the best order for *pow3*, whereas LEC is the best for *tanh* .

To see whether the expected behavior is observed in finite sample sizes we repeated the estimation of the unmixing matrix 5000 times for different sample sizes using all six possible extraction orders for both nonlinearities. The extraction order can be controlled using six different 3×3 permutation matrices $\mathbf{P}$ as initial values $\mathbf{U}_{init}$. For the reloaded deflation-based FastICA we chose FOBI [2] as the initial estimate. The FOBI functional is an affine equivariant IC functional, and the limiting distribution of the unmixing matrix estimate is known to be multivariate normal [7]. FOBI has the advantage that it is easy to compute, and, unlike FastICA, it always gives a solution. In this simulation study we included the FastICA estimators using random initial values as well. Then the extraction order is also random, and hence the performance is expected to be a mixture of the performances of the six possible estimators with different (fixed) extraction orders.

We used the FastICA code [3] for Algorithm 1, and we retained all the default settings except the initial value. One problem worth mentioning is that, unfortunately, the algorithm does not always converge. In applied data analysis the user may be able to change some tuning parameters in order to obtain a solution. However, this is not feasible in a simulation study. In our simulations, we simply ignored the cases when convergence did not occur. (Another option would have been to set the *MD* values to 1 in these cases.)

n	ECL	LCE	CEL	ELC	LEC	CLE	rand	reloaded
1000	20	24	27	0	0	25	5	0
5000	0	0	0	0	0	0	0	0
10000	0	0	0	0	0	0	0	0
$\geq$25000	0	0	0	0	0	0	0	0

Table 3: Number of algorithm failures in 5000 trials for *pow3*.

n	ECL	LCE	CEL	ELC	LEC	CLE	rand	reloaded
1000	340	472	493	0	0	457	145	0
5000	12	79	71	0	0	71	11	0
10000	1	13	10	0	0	10	4	0
$\geq$25000	0	0	0	0	0	0	0	0

Table 4: Number of algorithm failures in 5000 trials for *tanh*.

Table 3 and Table 4 give the number of cases when the algorithm did not converge. These figures clearly illustrate that for small sample sizes the algorithm often fails to converge for the given initial matrix. The problem is more severe in case of *tanh* nonlinearity. However, reloading FastICA seems to help the algorithm to find a solution.

Figure 1 presents the plots of the average values of $n(p-1)MD^2(\hat{\mathbf{W}},\mathbf{A})$ over the sample size n. The black lines in the figure give the results for the deflation-based FastICA with fixed extraction order, and the horizontal lines represent the asymptotic expectations given in Table 2. While for *pow3* convergence is reached quickly, this is not the case for *tanh*. The worse the performance, the slower the convergence seems to be. The performance of the deflation-based FastICA with random initial matrix is somewhere between the optimal and the worst possible case, which supports our conjecture of being a mixture of the six different cases. The strange behavior at the large sample sizes when using *pow3* may be due to the fact that the algorithm often converges to a wrong local maxima. It is clear that the average *MD* of the reloaded FastICA corresponds to the minimum value among the six possible cases. Therefore, the reloaded FastICA behaves as expected and is basically equivalent with the best extraction order for that given nonlinearity.

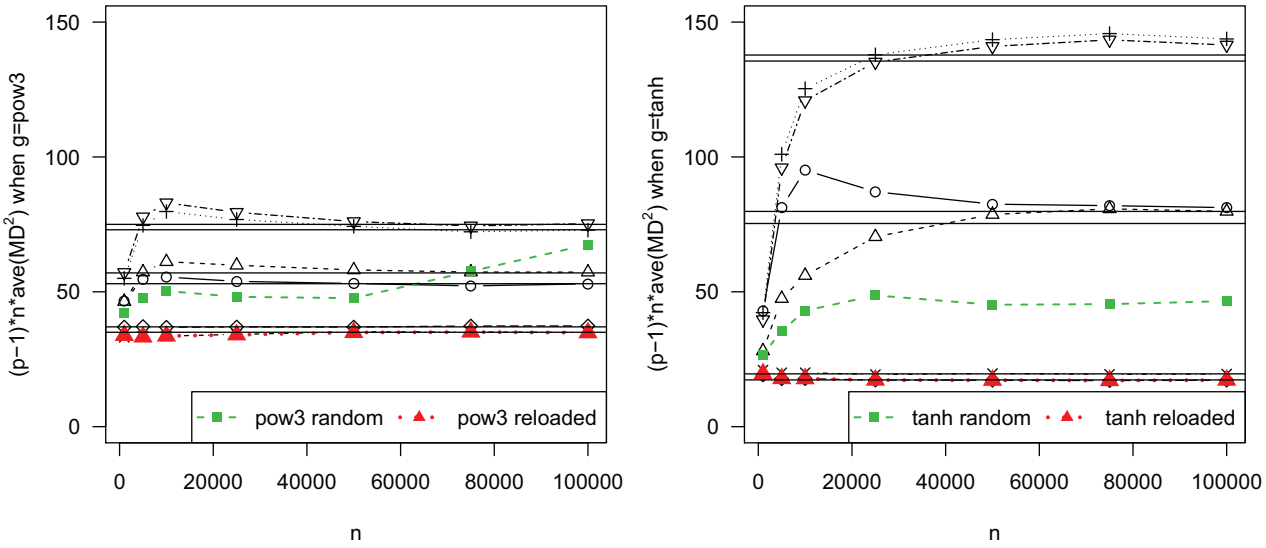

Figure 1: Average performance of the reloaded FastICA and the deflation-based FastICA based on a random initial value. The black curves give the performance of deflation based-FastICA when the extraction order is fixed. Horizontal black lines are asymptotic expectations given in Table 2.

6. CONCLUSIONS

In this paper we reviewed some properties of the deflation-based FastICA. One important curious property of FastICA is that the extraction order has a huge impact on the separation performance. We used this property and suggested the use of the reloaded FastICA to achieve the optimal extraction order. In our approach, we first need to run some ICA procedure that provides a consistent and affine equivariant unmixing matrix estimate. Then the extracted sources are permuted based on the nonlinearity used, and finally the regular deflation-based FastICA is performed using the estimated and permuted sources as whitened data and the identity matrix as an initial value of the rotation matrix. Reloading FastICA this way yields the best extraction order and renders the algorithm more stable at small sample sizes as validated by our simulation studies.

Future research is needed to derive the asymptotic properties of the reloaded FastICA estimator. Above all, more research is needed to derive the optimal choice of the nonlinearity function as well.

REFERENCES

[1] G. Brys, M. Hubert, and P.J. Rousseeuw, "A robustification of independent component analysis". *Chemometrics*, vol. 57 , pp. 364–375, 2006.

[2] J. Cardoso, "Source separation using higher moments," in *Proc. IEEE International Conference on Acustics, Speech and Signal Processing*, Glasgow, 1989, pp. 2109–2112.

[3] *http://www.cis.hut.fi/projects/ica/fastica*.

[4] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, pp.1483-1492, 1997.

[5] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, 1999.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[7] P. Ilmonen, J. Nevalainen, H. Oja, "Characteristics of multivariate distributions and the invariant coordinate system," *Statistics and Probability Letters*, vol. 80, pp. 1844–1853, 2010.

[8] P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila, "A new performance index for ICA: properties, computation and asymptotic analysis," in *Latent Variable Analysis and Signal Processing (Proceedings of 9th International Conference on Latent Variable Analysis and Signal Separation)*. 2010, pp. 229–236.

[9] P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila, "Independent component (IC) functionals and a new performance index", submitted.

[10] K. Nordhausen, E. Ollila and H. Oja, "On the performance indices of ICA and blind source separation," in *Proc. IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2011)*, 2011, pp. 471–475.

[11] E. Ollila, "On the robustness of the deflation-based FastICA estimator," in *Proc. IEEE Workshop on Statistical Signal Processing (SSP'09)*, Cardiff, Wales, Aug. 31–Sep. 3. 2009, pp. 673–676.

[12] E. Ollila, "The deflation-based FastICA estimator: statistical analysis revisited," *IEEE Trans. Signal Processing*, vol. 58, pp. 1527–1541, 2010.

[13] E. Ollila and H.-J. Kim, "'On testing hypotheses of mixing vectors in the ICA model using FastICA," in *Proc. IEEE Int. Symp. on Biomedical Imaging* (ISBI'11), Chicago, USA, Mar. 30 – Apr. 2, 2011, pp. 325-328.

Statistics and Probability Letters 80 (2010) 1844–1853

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Characteristics of multivariate distributions and the invariant coordinate system

Pauliina Ilmonen [a,*], Jaakko Nevalainen [b], Hannu Oja [a]

[a] Tampere School of Public Health, FI-33014 University of Tampere, Finland

[b] Statistics/Department of Social Research, FI-20014 University of Turku, Finland

ARTICLE INFO

Article history:
Received 12 April 2010
Received in revised form 13 August 2010
Accepted 16 August 2010
Available online 21 August 2010

MSC:
62H10
62H12
62G05
62G20
62F12

Keywords:
Asymptotic normality
Independent component analysis
Invariant coordinate selection
Multivariate kurtosis
Multivariate skewness

ABSTRACT

We consider a semiparametric multivariate location–scatter model where the standardized random vector of the model is fixed using simultaneously two location vectors and two scatter matrices. The approach using location and scatter functionals based on the first four moments serves as our main example. The four functionals yield in a natural way the corresponding skewness, kurtosis and unmixing matrix functionals. Affine transformation based on the unmixing matrix transforms the variable to an invariant coordinate system. The limiting properties of the skewness, kurtosis, and unmixing matrix estimates are derived under general conditions. We discuss related statistical inference problems, the role of the sample statistics in testing for normality and ellipticity, and connections to invariant coordinate selection and independent component analysis.

1. Introduction

Consider the $n \times p$ data matrix $X = (x_1, \ldots, x_n)^T$, where $x_1, \ldots, x_n$ is a random sample from a p-variate distribution. Different *location–scatter models* are obtained if one assumes that

$$x_i = \Omega z_i + \mu, \quad i = 1, \ldots, n,$$

where $Z = (z_1, \ldots, z_n)^T$ is an unobservable random sample from a "standardized" distribution, μ is a *location vector* and Ω is a full-rank $p \times p$ matrix, termed the *mixing matrix* in the independent component analysis (ICA) literature. The inverse of Ω, $\Gamma = \Omega^{-1}$, is the *unmixing matrix*, and $\Sigma = \Omega\Omega^T$ is the *scatter matrix*. Posing various assumptions on the distribution of the z_i yields different parametric or semiparametric multivariate models which are parametrized by μ and Σ, or by μ and Ω (Nordhausen et al., 2010).

* Corresponding author.
E-mail addresses: pauliina.ilmonen@uta.fi (P. Ilmonen), jaakko.nevalainen@utu.fi (J. Nevalainen), hannu.oja@uta.fi (H. Oja).

0167-7152/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2010.08.010

The following location–scatter models arise from this general structure and are often considered and discussed in the literature.

1. The classical multivariate methods rely on the assumption of multivariate normality, that is, $z_i \sim N_p(0, I_p)$, $i = 1, \ldots, n$. The location parameter μ is the mean vector and the scatter parameter Σ the covariance matrix. As $Oz_i \sim N_p(0, I_p)$ for all orthogonal matrices O, the mixing matrix Ω or the unmixing matrix Γ are defined only up to an orthogonal transformation in the *multivariate normal model*.
2. In the *multivariate elliptical model* it is assumed that $z_i \sim Oz_i$ for all orthogonal matrices O. (Notation $x \sim y$ means that random variables x and y are similarly distributed.) To fix Σ it is often assumed that $E(\|z_i\|^2) = p$ or that $Med(\|z_i\|^2) = \chi^2_{p,1/2}$. (The first configuration naturally requires that finite second moments exist, but the second allows us to avoid any moment assumptions.) As in the multivariate normal model, Ω and Γ are again defined only up to an orthogonal transformation. Elliptical distributions are thus symmetric in the sense that $z_i \sim Oz_i$ for all O, but they may vary in their kurtosis properties. The model permits for heavier (or lighter) tails than the multivariate normal model, and therefore the elliptical models are commonly seen as a more realistic alternative to the multivariate normal model. Robust testing and estimation procedures, for example, often assume ellipticity.
3. Another type of model family is obtained if one presumes that the observations arise from a *parametric independent component (IC) model*. Here the z_i are assumed to have independent and standardized components and the density function $f(z_i) = \prod_{j=1}^p f_j(z_{ij})$ with some known standardized marginal densities $f_1, \ldots, f_p$. Matrix Γ is unique for distinct standardized densities $f_1, \ldots, f_p$.
4. In a generalization of the parametric IC model, the *semiparametric independent component model*, z_i are assumed to consist of independent and standardized components such that $E(z_i) = 0$ and $E(z_i z_i^T) = I_p$, or $Med(z_{ij}) = 0$ and $Med(z_{ij}^2) = \chi^2_{1,1/2}$. But then Ω and Γ are defined only up to permutations and sign changes of the columns and rows, respectively. In ICA the goal is to estimate Γ (up to a permutation, rescaling and sign changes of the rows). In parametric and semiparametric IC models, skewness and kurtosis properties are characteristics of the marginal distributions.

Instead of location–scatter models we will work under a general *semiparametric location–scatter–skewness–kurtosis model* (shortly, semiparametric model), which includes all (continuous as well as discrete) multivariate distributions with finite fourth moments. Thus, many of the more conventional models listed above overlap with the semiparametric model, which was first introduced by Nordhausen et al. (2010).

The paper is organized as follows. The semiparametric model is defined in Section 2 along with a discussion of related unmixing matrix, skewness and kurtosis functionals, G, d and L. Section 3 gives useful asymptotic results for the corresponding estimates $\hat{\Gamma}$, $\hat{\delta}$ and $\hat{\Lambda}$ even outside the semiparametric model. More detailed results for the moment-based estimates are given in Section 4. Statistical properties of the fourth-order blind identification FOBI estimate are obtained as a side-product. Section 5 discusses the uses of sample statistics in testing and estimation problems.

2. Definitions and preliminary results

2.1. Semiparametric model

A multivariate semiparametric model can be defined with natural parameters for multivariate location, scatter, skewness and kurtosis, respectively. The z_i need to be standardized in a special way using two moment-based location functionals, T_1 and T_2, and two moment-based scatter matrix functionals S_1 and S_2. Next we establish the model, the moment-based location and scatter functionals, and their connection to the model parameters.

Semiparametric model. Assume that $X = (x_1, \ldots, x_n)^T$ is random sample from a p-variate distribution such that

$$x_i = \Omega z_i + \mu, \quad i = 1, \ldots, n,$$

where the z_i are standardized so that

$$E(z_i) = 0,$$
$$E(z_i z_i^T) = I_p,$$
$$E(z_i z_i^T z_i) = p \cdot \delta \quad \text{and}$$
$$E(z_i z_i^T z_i z_i^T) = (p+2) \cdot \Lambda$$

where δ is a p-vector with all components $\delta_i \geq 0$, $i = 1, \ldots, p$, and Λ is a diagonal matrix with diagonal elements $\lambda_1 \geq \cdots \geq \lambda_p > 0$.

The semiparametric model was first introduced in Nordhausen et al. (2010). The parameters in the model are the mean vector μ, the covariance matrix $\Sigma = \Omega\Omega^T$, the skewness vector δ based on third moments, and the kurtosis matrix Λ. We will return to these concepts shortly. The model is general in the sense that it includes all p-variate distributions with finite fourth moments, but of course rules out heavy-tailed distributions. The mixing and unmixing matrices, Ω and $\Gamma = \Omega^{-1}$, are uniquely defined if $\delta_i > 0$, $i = 1, \ldots, p$, and $\lambda_1 > \cdots > \lambda_p > 0$. When the model parameters are fixed in this way, the unmixing matrix can be used to transform the random vector to an invariant coordinate system (Tyler et al., 2009, ICS).

If the components of z_i are independent, the unmixing matrix Γ is the fourth-order blind identification (FOBI) functional by Cardoso (1989), a solution in ICA. Alternative models and multivariate skewness and kurtosis measures are obtained if the moment-based location and scatter measures are replaced by some other, *e.g.* robust, multivariate location measures and scatter measures.

The model obviously includes the multivariate normal model with $\delta = 0$ and $\Lambda = I_p$. For elliptical distribution $\delta = 0$ and $\Lambda = \lambda I_p$, where λ is a kurtosis parameter, which may not be finite. In the elliptical model Ω is thus defined only up to an orthogonal transformation. However, $\Sigma = \Omega\Omega^T$ is uniquely defined. IC models are included when the marginal densities possess finite fourth-order moments. Recall that the target parameter of ICA is Γ.

2.2. Location and scatter functionals

In robust and nonparametric communities the characteristics of a distribution are often described by functionals. Let F_x be the cumulative distribution function (cdf) of a p-variate random variable x. A *location functional* $T(F_x)$ is a vector-valued $(p \times 1)$ functional, which is affine equivariant in the sense that

$$T(F_{Ax+b}) = AT(F_x) + b$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b. A *scatter functional* $S(F_x)$ is a $p \times p$-matrix-valued functional which is positive definite and affine equivariant in the sense that

$$S(F_{Ax+b}) = AS(F_x)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b. A scatter functional S is said to possess the *independence property* if $S(F_x)$ is a diagonal matrix for all x with independent components—a property which not all scatter matrices enjoy, but which is essential in independent component analysis. The first examples of location and scatter functionals are the mean vector and covariance matrix:

$$T_1(F_x) = E(x) \quad \text{and} \quad S_1(F_x) = E((x - E(x))(x - E(x))^T).$$

The covariance matrix S_1 has the independence property. In the semiparametric model $T_1(F_x) = \mu$ and $S_1(F_x) = \Sigma$.

Location and scatter functionals can be based on the third and fourth moments as well. A location functional based on third moments is

$$T_2(F_x) = \frac{1}{p} E((x - E(x))^T S_1(F_x)^{-1}(x - E(x))x).$$

Finally, a scatter matrix based on fourth moments is

$$S_2(F_x) = \frac{1}{p+2} E((x - E(x))(x - E(x))^T S_1(F_x)^{-1}(x - E(x))(x - E(x))^T),$$

which has the independence property (Oja et al., 2006). Note now that these functionals can be used to standardize the random vectors in the semiparametric model as clearly

$$T_1(F_{z_i}) = 0, \qquad T_2(F_{z_i}) = \delta, \qquad S_1(F_{z_i}) = I_p, \quad \text{and} \quad S_2(F_{z_i}) = \Lambda.$$

2.3. Skewness, kurtosis, and unmixing matrix functionals

Without fixing any particular location and scatter functionals, like the moment-based functionals in the above, the *unmixing matrix functional* G $(p \times p)$, *skewness functional* d $(p \times 1)$ and *kurtosis functional* L $(p \times p)$, based on two pairs of some location and scatter functionals, (T_1, S_1) and (T_2, S_2), can be defined as follows.

Definition 2.1. Let a matrix-valued functional G, a vector-valued functional d, and a diagonal matrix-valued functional L be defined so that, if $z = G(F_x)(x - T_1(F_x))$, then

$$T_1(F_z) = 0, \qquad S_1(F_z) = I_p, \qquad T_2(F_z) = d, \quad \text{and} \quad S_2(F_z) = L,$$

where $d \geq 0$ and the diagonal elements of L are in a decreasing order.

Note that G and L are solutions of the eigenvector and eigenvalue problem

$$S_1^{-1} S_2 G^T = G^T L.$$

A solution G is then unique up to a permutation, rescaling and sign changes of the rows. Among these, Definition 2.1 then picks up the solution G for which $S_1(F_z) = I_p$, $T_2(F_z) \geq 0$ and $S_2(F_z)$ is a diagonal matrix with decreasing diagonal elements. The first condition fixes the scales, the second one the signs, and the third one the order of the rows of G. The solution G then also satisfies

$$GS_1G^T = I_p \quad \text{and} \quad GS_2G^T = L,$$

where, as before, L is a diagonal matrix consisting of the eigenvalues of $S_1^{-1}S_2$. The solution is unique if $d > 0$ and L has distinct diagonal elements.

For functionals G, d, and L we then have the following lemma:

Lemma 2.1. *Assume the semiparametric model $x = \Omega z + \mu$, where for some location functionals T_1 and T_2 and for some scatter functionals S_1 and S_2, $T_1(F_z) = 0, S_1(F_z) = I_p$, $T_2(F_z) = \delta > 0$ and $S_2(F_z) = \Lambda$ is a diagonal matrix with diagonal elements $\lambda_1 > \cdots > \lambda_p > 0$. If G is based on (T_1, S_1) and (T_2, S_2), then*

$$G(F_x) = \Gamma, \qquad d(F_x) = \delta, \quad \text{and} \quad L(F_x) = \Lambda.$$

The functionals are affine equivariant and invariant in the sense that

$$G(F_{Ax+b}) = G(F_x)A^{-1}, \qquad d(F_{Ax+b}) = d(F_x), \quad \text{and} \quad L(F_{Ax+b}) = L(F_x)$$

for all nonsingular $p \times p$ matrices A and all p-vectors b.

The values of the functionals at the empirical distribution F_n yield natural Fisher consistent estimates of the corresponding population quantities. For an estimate of $T(F_x)$ we then write $T(F_n)$ or $T(X)$ where X is an $n \times p$ data matrix. To simplify notation, we write

$$T_1 = T_1(F_x), \qquad S_1 = S_1(F_x), \qquad T_2 = T_2(F_x), \quad \text{and} \quad S_2 = S_2(F_x),$$

and

$$\hat{T}_1 = T_1(F_n), \qquad \hat{S}_1 = S_1(F_n), \qquad \hat{T}_2 = T_2(F_n), \quad \text{and} \quad \hat{S}_2 = S_2(F_n).$$

Most of the time, the interest lies in the population parameters

$$\Gamma = G(F_x), \qquad \delta = d(F_x), \quad \text{and} \quad \Lambda = L(F_x)$$

and their Fisher consistent estimates

$$\hat{\Gamma} = G(F_n), \qquad \hat{\delta} = d(F_n), \quad \text{and} \quad \hat{\Lambda} = L(F_n),$$

respectively. Of course, the estimates $\hat{\delta}$, $\hat{\Gamma}$ and $\hat{\Lambda}$ adopt the same equivariance and invariance properties as the corresponding functionals meaning that

$$G(XA^T + 1_n b^T) = G(X)A^{-1},$$
$$d(XA^T + 1_n b^T) = d(X) \quad \text{and}$$
$$L(XA^T + 1_n b^T) = L(X).$$

2.4. Connections to ICA, ICS and classical skewness and kurtosis measures

If T_1, T_2, S_1 and S_2 are the moment-based functionals we can say more, and find similarities in the literature. First, if the observations come from a continuous distribution, the estimates exist and are unique with probability one. Second, the estimate $\hat{\Gamma}$ is the well-known FOBI estimate in the IC model. Third, in the univariate case ($p = 1$), $\|\hat{\delta}\|^2$ and $\hat{\Lambda}$ reduce to the classical univariate skewness and kurtosis measures

$$\frac{[E(x - E(x))^3]^2}{[E(x - E(x))^2]^3} \quad \text{and} \quad \frac{E(x - E(x))^4}{3[E(x - E(x))^2]^2}.$$

In the multivariate case, Mardia (1970) defined different moment-based measures of skewness and kurtosis for a sample $X = (x_1, \ldots, x_n)'$ as

$$b_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}((x_i - \hat{T}_1)^T\hat{S}_1^{-1}(x_j - \hat{T}_1))^3$$

and

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}((x_i - T_1)^T\hat{S}_1^{-1}(x_i - T_1))^2$$

which use the moments of elements of a maximal invariant statistic, the $n \times n$ matrix

$$(X - 1_n\hat{T}_1^T)\hat{S}_1^{-1}(X - 1_n\hat{T}_1^T)^T.$$

Our skewness and kurtosis statistics are based on the third and fourth moments of another maximal invariant statistic, namely, the $n \times p$ matrix

$$\hat{Z} = (X - 1_n\hat{T}_1^T)\hat{\Gamma}^T.$$

Matrix $\hat{Z}$ gives the observations in an invariant coordinate system (Tyler et al., 2009). Still another invariant coordinate system (maximal invariant statistic) based on $p+1$ observations was proposed by Chakraborty and Chaudhuri (1999). Their approach is known as the transformation–retransformation approach. See also Serfling (2010) for a general discussion on standardization, weak covariance, transformation–retransformation, and strong invariant coordinate system functionals. Bera and John (1983) use the third and fourth moments of the "scaled residuals" in $(X - 1_n\hat{T}_1^T)\hat{S}_1^{-1/2}$ (with a symmetric square root matrix). Unlike Mardia's statistics and our skewness and kurtosis statistics, their statistics are not affine invariant, however.

3. Asymptotical properties

We are interested in the limiting behavior of the estimates $\hat{\delta}$, $\hat{\Gamma}$ and $\hat{\Lambda}$. As the estimates are affine equivariant and invariant, it is not a restriction to assume that

$$T_1 = 0, \qquad S_1 = I_p, \qquad T_2 = \delta \quad \text{and} \quad S_2 = \Lambda, \quad \text{and} \quad \text{therefore } \Gamma = I_p.$$

For uniqueness, we assume that $\delta_i > 0$, $i = 1, \ldots, p$, and the diagonal elements of Λ are strictly ordered so that $\lambda_1 > \cdots > \lambda_p > 0$.

We assume that the location and scatter estimates, not necessarily moment-based yet, are root-n consistent, that is,

$$\sqrt{n}\hat{T}_1 = O_p(1) \quad \text{and} \quad \sqrt{n}(\hat{S}_1 - I_p) = O_p(1)$$

as well as

$$\sqrt{n}(\hat{T}_2 - \delta) = O_p(1) \quad \text{and} \quad \sqrt{n}(\hat{S}_2 - \Lambda) = O_p(1).$$

Then we have the following result.

Theorem 3.1. *If $\hat{T}_1$, $\hat{S}_1$, $\hat{T}_2$ and $\hat{S}_2$ are root-n consistent, then so are $\hat{\delta}$, $\hat{\Gamma}$ and $\hat{\Lambda}$ and*

$$\sqrt{n}(\hat{\delta} - \delta) = \sqrt{n}(\hat{T}_2 - \delta) - \sqrt{n}\hat{T}_1 + \sqrt{n}(\hat{\Gamma} - I_p)\delta + o_p(1)$$

and

$$\sqrt{n}(\hat{\Gamma}_{ii} - 1) = -\frac{1}{2}\sqrt{n}((\hat{S}_1)_{ii} - 1) + o_p(1),$$
$$(\lambda_i - \lambda_j)\sqrt{n}\hat{\Gamma}_{ij} = \sqrt{n}(\hat{S}_2)_{ij} - \lambda_i\sqrt{n}(\hat{S}_1)_{ij} + o_p(1), \quad i \neq j, \quad \textit{and}$$
$$\sqrt{n}(\hat{\Lambda}_{ii} - \lambda_i) = \sqrt{n}((\hat{S}_2)_{ii} - \lambda_i) - \lambda_i\sqrt{n}((\hat{S}_1)_{ii} - 1) + o_p(1).$$

Proof. First note that since the transformation $(T_1, S_1, T_2, S_2) \to (d, G, L)$ is continuous in a neighborhood of $(0, I_p, \delta, \Lambda)$, and $(\hat{T}_1, \hat{S}_1, \hat{T}_2, \hat{S}_2) \to_p (0, I_p, \delta, \Lambda)$, also $(\hat{\delta}, \hat{\Gamma}, \hat{\Lambda}) \to_p (\delta, I_p, \Lambda)$. As

$$\hat{\delta} = T_2((X - 1_n\hat{T}_1^T)\hat{\Gamma}^T),$$

it follows by affine equivariance that

$$\sqrt{n}(T_2((X - 1_n\hat{T}_1^T)\hat{\Gamma}^T) - \delta) = \sqrt{n}\hat{\Gamma}(\hat{T}_2 - \hat{T}_1 - \delta) + \sqrt{n}(\hat{\Gamma} - I_p)\delta.$$

Then by Slutsky's theorem $\sqrt{n}\hat{\Gamma}(\hat{T}_2 - \hat{T}_1 - \delta) - \sqrt{n}(\hat{T}_2 - \delta) + \sqrt{n}\hat{T}_1$ converges to 0 in distribution and therefore in probability as well. Thus

$$\sqrt{n}\hat{\Gamma}(\hat{T}_2 - \hat{T}_1 - \delta) = \sqrt{n}(\hat{T}_2 - \delta) - \sqrt{n}\hat{T}_1 + o_p(1),$$

and the first part of the theorem follows. For $\hat{\Gamma}$ and $\hat{\Lambda}$ we utilize the estimating equations

$$\hat{\Gamma}\hat{S}_1\hat{\Gamma}^T = I_p \quad \text{and} \quad \hat{\Gamma}\hat{S}_2\hat{\Gamma}^T = \hat{\Lambda}.$$

Then

$$(\hat{\Gamma} - I_p)\hat{S}_1\hat{\Gamma}^T + (\hat{S}_1 - I_p)\hat{\Gamma}^T + (\hat{\Gamma} - I_p)^T = 0 \quad \text{and}$$
$$(\hat{\Gamma} - I_p)\hat{S}_2\hat{\Gamma}^T + (\hat{S}_2 - \Lambda)\hat{\Gamma}^T + \Lambda(\hat{\Gamma} - I_p)^T = \hat{\Lambda} - \Lambda$$

and Slutsky's theorem gives

$$\sqrt{n}(\hat{S}_1 - I_p) = -\sqrt{n}(\hat{\Gamma} - I_p) - \sqrt{n}(\hat{\Gamma} - I_p)^T + o_p(1) \quad \text{and}$$
$$\sqrt{n}(\hat{S}_2 - \Lambda) = -\sqrt{n}(\hat{\Gamma} - I_p)\Lambda - \sqrt{n}\Lambda(\hat{\Gamma} - I_p)^T + \sqrt{n}(\hat{\Lambda} - \Lambda) + o_p(1).$$

These equations yield the desired results for $\hat{\Gamma}$ and $\hat{\Lambda}$. □

In the matrix form, we can write

$$\sqrt{n}\, diag(\hat{\Gamma} - I_p) = -\frac{1}{2}\sqrt{n}\, diag(S_1 - I_p) + o_p(1),$$
$$\sqrt{n}\,(\hat{\Gamma} - diag(\hat{\Gamma})) = \sqrt{n}\, H \odot ((\hat{S}_2 - \Lambda) - (\hat{S}_1 - I_p)\Lambda) + o_p(1), \quad \text{and}$$
$$\sqrt{n}\,(\hat{\Lambda} - \Lambda) = \sqrt{n}\, diag((\hat{S}_2 - \Lambda) - (\hat{S}_1 - I_p)\Lambda) + o_p(1),$$

where H is a $p \times p$ matrix with elements

$$H_{ij} = 0, \quad \text{if } i = j, \quad \text{and} \quad H_{ij} = (\lambda_i - \lambda_j)^{-1}, \quad \text{if } i \neq j,$$

$diag(\Gamma)$ for example is a diagonal matrix with the same diagonal elements as Γ, and $\odot$ means the Hadamard (entrywise) product.

Note that the principal component analysis is a special case here: if one takes $S_1 = I_p$ (constant) and $S_2 = S$, the theorem gives the limiting behavior of the eigenvectors and eigenvalues of S.

4. Limiting distributions of the moment-based estimates

We next establish the limiting distributions of the estimates $\hat{\delta}$, $\hat{\Gamma}$ and $\hat{\Lambda}$ obtained by using the moment-based location and scatter statistics

$$\hat{T}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \text{and} \quad \hat{S}_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{T}_1)(x_i - \hat{T}_1)^T$$

and

$$\hat{T}_2 = \frac{1}{np}\sum_{i=1}^{n}(x_i - \hat{T}_1)^T \hat{S}_1^{-1}(x_i - \hat{T}_1)x_i$$

and

$$\hat{S}_2 = \frac{1}{n(p+2)}\sum_{i=1}^{n}(x_i - \hat{T}_1)(x_i - \hat{T}_1)^T \hat{S}_1^{-1}(x_i - \hat{T}_1)(x_i - \hat{T}_1)^T.$$

Again, as the estimates are affine equivariant and invariant, we may assume that the population values are

$$T_1 = 0, \qquad S_1 = I_p, \qquad T_2 = \delta \quad \text{and} \quad S_2 = \Lambda, \quad \text{and} \quad \text{therefore } \Gamma = I_p.$$

Thus $x_i = z_i$, $i = 1, \ldots, n$, and we write

$$\tilde{T}_1 = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad \text{and} \quad \tilde{S}_1 = \frac{1}{n}\sum_{i=1}^{n} z_i z_i^T$$

and

$$\tilde{T}_2 = \frac{1}{np}\sum_{i=1}^{n} z_i z_i^T z_i \quad \text{and} \quad \tilde{S}_2 = \frac{1}{n(p+2)}\sum_{i=1}^{n} z_i z_i^T z_i z_i^T.$$

If the first eight moments of z_i exist, the joint distribution of

$$\sqrt{n}\begin{pmatrix} \tilde{T}_1 \\ vec(\tilde{S}_1 - I_p) \\ \tilde{T}_2 - \delta \\ vec(\tilde{S}_2 - \Lambda) \end{pmatrix}$$

is, by the central limit theorem, a $2(p + p^2)$ variate (singular) normal distribution with mean zero and covariance matrix given by

$$D = E\left(\begin{pmatrix} z_i \\ vec(z_i z_i^T - I_p) \\ \frac{1}{p} z_i z_i^T z_i - \delta \\ vec\left(\frac{1}{p+2} z_i z_i^T z_i z_i^T - \Lambda\right) \end{pmatrix}\begin{pmatrix} z_i \\ vec(z_i z_i^T - I_p) \\ \frac{1}{p} z_i z_i^T z_i - \delta \\ vec\left(\frac{1}{p+2} z_i z_i^T z_i z_i^T - \Lambda\right) \end{pmatrix}^T\right).$$

One can show that

$$\sqrt{n}\begin{pmatrix} \hat{T}_1 \\ vec(\hat{S}_1 - I_p) \\ \hat{T}_2 - \delta \\ vec(\hat{S}_2 - \Lambda) \end{pmatrix} = C\sqrt{n}\begin{pmatrix} \tilde{T}_1 \\ vec(\tilde{S}_1 - I_p) \\ \tilde{T}_2 - \delta \\ vec(\tilde{S}_2 - \Lambda) \end{pmatrix} + o_p(1),$$

where

$$C = \begin{pmatrix} I_p & 0 & 0 & 0 \\ 0 & I_{p^2} & 0 & 0 \\ C_{31} & C_{32} & I_p & 0 \\ C_{41} & C_{42} & 0 & I_{p^2} \end{pmatrix}$$

with

$$C_{31} = -\frac{2}{p}I_p \quad \text{and} \quad C_{32} = -\frac{1}{p}E(z_i^T \otimes (z_i z_i^T))$$

and

$$C_{41} = -\frac{1}{p+2}[E((z_i^T z_i)(I_p \otimes z_i)) + E((z_i^T z_i)(z_i \otimes I_p)) + 2 \cdot E(z_i \otimes (z_i z_i^T))]$$

and

$$C_{42} = -\frac{1}{p+2}E((z_i z_i^T) \otimes (z_i z_i^T)).$$

The asymptotic normality of

$$\sqrt{n}\begin{pmatrix} \hat{T}_1 \\ vec(\hat{S}_1 - I_p) \\ \hat{T}_2 - \delta \\ vec(\hat{S}_2 - \Lambda) \end{pmatrix}$$

then follows.

Finally, if $\delta_i > 0$, $i = 1, \ldots, p$, and the diagonal elements of Λ are strictly ordered, we get

$$\sqrt{n}\begin{pmatrix} \hat{\delta} - \delta \\ vec(\hat{\Lambda} - \Lambda) \\ vec(\hat{\Gamma} - \Gamma) \end{pmatrix} = B\sqrt{n}\begin{pmatrix} \hat{T}_1 \\ vec(\hat{S}_1 - I_p) \\ \hat{T}_2 - \delta \\ vec(\hat{S}_2 - \Lambda) \end{pmatrix},$$

where

$$B = \begin{pmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ 0 & B_{22} & 0 & B_{24} \\ 0 & B_{32} & 0 & B_{34} \end{pmatrix}$$

with

$$B_{11} = -I_p, \quad \text{and} \quad B_{12} = [\delta^T \otimes I_p]\left[-\frac{1}{2}diag(vec(I_p)) - diag(vec(H))(\Lambda \otimes I_p)\right]$$

and

$$B_{13} = I_p \quad \text{and} \quad B_{14} = [\delta^T \otimes I_p]diag(vec(H))$$

and

$$B_{22} = -diag(vec(I_p))(\Lambda \otimes I_p) \quad \text{and} \quad B_{24} = diag(vec(I_p))$$

and

$$B_{32} = -\frac{1}{2}diag(vec(I_p)) - diag(vec(H))(\Lambda \otimes I_p) \quad \text{and} \quad B_{34} = diag(vec(H)).$$

(The matrix H is given in Section 3.)

We have thus proved the following theorem.

Theorem 4.1. *Assume that X is a random sample from the semiparametric model with $\Omega = I_p$, $\mu = 0$, $\delta_i > 0$, $i = 1, \ldots, p$, and $\lambda_1 > \cdots > \lambda_p > 0$. Assume also that the first eight moments of z_i are finite. Then*

$$\sqrt{n}\begin{pmatrix} \hat{\delta} - \delta \\ vec(\hat{\Lambda} - \Lambda) \\ vec(\hat{\Gamma} - I_p) \end{pmatrix}$$

has the limiting (singular) $p + 2p^2$-variate normal distribution with mean value zero and covariance matrix $BCDC^TB^T$.

By affine equivariance and invariance properties of the estimates this generalizes to

Corollary 4.1. *Assume that X is a random sample from the semiparametric model with $\delta_i > 0$, $i = 1, \ldots, p$, and $\lambda_1 > \cdots > \lambda_p > 0$. Assume also that the first eight moments of z_i are finite. Then*

$$\sqrt{n}\begin{pmatrix} \hat{\delta} - \delta \\ vec(\hat{\Lambda} - \Lambda) \\ vec(\hat{\Gamma} - \Gamma) \end{pmatrix}$$

has the limiting (singular) $p + 2p^2$-variate normal distribution with mean value zero and covariance matrix $ABCDC^TB^TA^T$ where

$$A = \begin{pmatrix} I_p & 0 & 0 \\ 0 & I_{p^2} & 0 \\ 0 & 0 & \Gamma^T \otimes I_p \end{pmatrix}.$$

Remark 4.1. Under the assumption of central symmetry (i.e. $z_i \sim -z_i$) the value of $\delta = 0$, but even then the unmixing matrix functional satisfying $\hat{\Gamma}\hat{S}_1\hat{\Gamma}^T = I_p$ is affine equivariant up to the signs of its row vectors. In order to fix the signs of the unmixing matrix functional and the estimate of it, we can require for example that $\Gamma 1_p > 0$ (and set $\hat{\Gamma} 1_p > 0$). Now, even under the assumption of central symmetry, the limiting joint distribution of $\hat{\Gamma}$ and $\hat{\Lambda}$ is the one given in Theorem 4.1 and Corollary 4.1. In the elliptic case $\delta = 0$ and $\Lambda = \lambda I_p$ and the limiting behavior of $\hat{\delta}$, $\hat{\Lambda}$, and $\hat{\Gamma}$ is unknown. However, the limiting properties of $\|\hat{\delta}\|^2$ and the mean and variance of the elements of $\hat{\Lambda}$ are known, see Section 5.2.

Remark 4.2. In this section we derived the asymptotic joint distribution of the skewness, kurtosis and unmixing matrix estimators for moment-based functionals. Since Theorem 3.1 is not restricted to moment-based functionals, the same methodology can be used when other location and scatter functionals are considered, as long as the joint limiting distribution of the corresponding location and scatter estimates is known.

5. Applications and concluding remarks

5.1. Statistical inference

The results in Section 4 can be used to find estimates of the limiting covariance matrices of the estimates $\hat{\Gamma}$, $\hat{\delta}$ and $\hat{\Lambda}$. These could then in turn be used in the construction of confidence ellipsoids for the parameters. The results can also be employed in the development and the conduct of interesting testing procedures, which we are currently working on.

To estimate the limiting distribution in practice, one can proceed as follows.

1. Calculate $\hat{T}_1$, $\hat{T}_2$, $\hat{S}_1$ and $\hat{S}_2$.
2. Find estimates $\hat{\Gamma}$, $\hat{\delta}$ and $\hat{\Lambda}$.
3. Transform observations to the invariant coordinate system:
$$\hat{Z} = (X - 1_n\hat{T}_1^T)\hat{\Gamma}^T.$$
4. Find estimates $\hat{D}$, $\hat{C}$, $\hat{B}$ and $\hat{A}$: in the formulas for D and C replace the expectations by averages and the (unknown) z_i by $\hat{z}_i$, $i = 1, \ldots, n$.
5. Then, approximately,
$$\begin{pmatrix} \hat{\delta} - \delta \\ vec(\hat{\Lambda} - \Lambda) \\ vec(\hat{\Gamma} - \Gamma) \end{pmatrix} \sim N_{p+2p^2}\left(0, \frac{1}{n}\hat{A}\hat{B}\hat{C}\hat{D}\hat{C}^T\hat{B}^T\hat{A}^T\right).$$

A bootstrap technique can also be used to estimate the distributions of the sample statistics: Let U be a random $n \times n$ matrix such that the rows are independent and the row vectors have $Multin(1; (1/n, \ldots, 1/n))$ distribution. Then U is

called a bootstrap matrix and UX is a bootstrap sample. The bootstrap estimates of the covariance matrices of $\hat{\Gamma}$, $\hat{\delta}$ and $\hat{\Lambda}$, for example, can be found as follows.

1. Calculate $\hat{T}_1, \hat{T}_2, \hat{S}_1$ and $\hat{S}_2$.
2. Find estimates $\hat{\Gamma}, \hat{\delta}$ and $\hat{\Lambda}$.
3. Choose M independent bootstrap matrices $U_1, \ldots, U_M$.
4. Calculate M bootstrap samples $X_i^* = U_i X, i = 1, \ldots, M$.
5. Calculate M bootstrap estimates
$$\hat{\delta}_i^* = d(X_i^*), \qquad \hat{\Gamma}_i^* = G(X_i^*) \quad \text{and} \quad \hat{\Lambda}_i^* = L(X_i^*), \quad i = 1, \ldots, M.$$
6. Calculate the sample covariance matrix of
$$\begin{pmatrix} \hat{\delta}_i^* - \hat{\delta} \\ vec(\hat{\Lambda}_i^* - \hat{\Lambda}) \\ vec(\hat{\Gamma}_i^* - \hat{\Gamma}) \end{pmatrix}, \quad i = 1, \ldots, M.$$

5.2. Tests for normality and ellipticity

Skewness and kurtosis statistics can be used to test for normality and/or ellipticity. In the elliptic case $\delta = 0$ and $\Lambda = \lambda I_p$. In the multivariate normal case $\lambda = 1$. Our assumptions for δ and Λ stated in Theorem 3.1 are thus not true, and the limiting behavior of $\hat{\delta}$, $\hat{\Lambda}$, and $\hat{\Gamma}$ is unknown. However, the limiting properties of $\|\hat{\delta}\|^2$ and the mean and variance of the elements of $\hat{\Lambda}$ are known, and may be used in the following way.

As skewness and kurtosis are affine invariant, it is not a restriction to assume that the distribution is spherical. Then
$$T_1 = T_2 = 0, \qquad S_1 = I_p \quad \text{and} \quad S_2 = \lambda I_p,$$
for some $\lambda > 0$. It is well known that in the spherical case the location statistics often satisfy
$$\sqrt{n}\hat{T}_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \gamma_i(r_j) u_j + o_p(1), \quad i = 1, 2,$$
and the scatter statistics satisfy
$$\sqrt{n}(\hat{S}_i - S_i) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\alpha_i(r_j) u_j u_j^T - \beta_i(r_j) S_i) + o_p(1), \quad i = 1, 2,$$
where $r_i = \|x_i\|$ and $u_i = \|x_i\|^{-1} x_i$, $i = 1, \ldots, n$. Functions γ_i, α_i and β_i then give the influence functions for T_i and S_i, respectively.

Kankainen et al. (2007) proposed the use of the skewness and kurtosis statistics
$$U = (\hat{T}_2 - \hat{T}_1)^T \hat{S}_1^{-1} (\hat{T}_2 - \hat{T}_1) \quad \text{and} \quad W = \|\hat{S}_1^{-1} \hat{S}_2 - I_p\|^2$$
for testing multivariate normality. It is then straightforward to see that
$$U = \|\hat{\delta}\|^2 \quad \text{and} \quad W = \|\hat{\Lambda} - I_p\|^2 = tr((\hat{\Lambda} - I_p)^2).$$
Kankainen et al. (2007) proved that

Theorem 5.1. *In the multivariate normal model,*

(i) *the limiting distribution of nU is that of $\eta_1 U_1$, where $U_1 \sim \chi_p^2$ and $\eta_1 = (1/p)E[(\gamma_1(r) - \gamma_2(r))^2]$ with $r^2 \sim \chi_k^2$;*
(ii) *the limiting distribution of nW is that of*
$$\eta_2 W_1 + \eta_3 W_2,$$
where $W_1 \sim \chi^2_{p(p+1)/2-1}$ and $W_2 \sim \chi_1^2$ are independent,
$$\eta_2 = \frac{2}{p(p+2)} E[(\alpha_2(r) - \alpha_1(r))^2]$$
and
$$\eta_3 = \frac{1}{p} E[(\alpha_2(r) - \alpha_1(r))^2] - 2E[(\alpha_2(r) - \alpha_1(r))(\beta_2(r) - \beta_1(r))] + pE[(\beta_2(r) - \beta_1(r))^2].$$
The expected values are calculated for $r^2 \sim \chi_p^2$.

The statistics U and W_1 can be used to test for ellipticity as well (but with different limiting distributions). Mardia (1970) advocated using his skewness and kurtosis statistics to test for multivariate normality. Under the null hypothesis of multivariate normality

$$\frac{nb_1}{6} \quad \text{and} \quad \frac{n(b_2 - p(p+1))^2}{8p(p+2)}$$

have limiting chi-square distributions with $p(p+1)(p+2)/6$ and 1 degrees of freedom, correspondingly. Kankainen et al. (2007) obtained the limiting Pitman efficiencies of U and W with respect to Mardia's statistics for contiguous sequences of contaminated normal distributions.

Nordhausen et al. (2010) discuss the general idea of using of $\hat{\delta}$ and $\hat{\Lambda}$ in the selection of an appropriate model for the data. Our results now provide the basic elements to convert their ideas into formal inference tools.

5.3. Invariant coordinate selection

Tyler et al. (2009) introduced a general method for exploring multivariate data called the *invariant coordinate selection*. In their approach, they used two shape (not scatter) matrices to find invariant coordinate system; the resulting coordinate system is invariant up to coordinatewise location, sign, and scale. Here, by associating two scatter statistics together with two location statistics, we also fix the location, sign, and the scale, and obtain a fully invariant coordinate system. The invariant coordinate system is useful in many ways. Plotting the observations in the new coordinate system

$$\hat{Z} = (X - 1_n \hat{T}_1^T)\hat{\Gamma}^T,$$

helps in finding outlying observations and clusters in the data. In the case of mixtures of elliptical distributions, a subset of invariant coordinates corresponds to Fisher's linear discriminant subspace (Tyler et al., 2009). Invariant coordinate selection may thus be seen as a tool for dimension reduction as well. Note that $\hat{Z}$ is a maximal invariant statistic under the group of affine transformations.

5.4. Independent component model

In the semiparametric independent component model, matrix $\hat{\Gamma}$ is a solution to the ICA problem. If T_1, S_1, T_2 and S_2 are the moment-based estimates, then $\hat{\Gamma}$ is the well-known FOBI estimate. Our approach thus gives a whole family of unmixing matrix estimates for the ICA problem. Furthermore, the limiting properties of the estimates can be considered and compared in our approach. Nordhausen et al. (2009) and Oja et al. (2010) found optimal signed-rank tests for location and independence in the independent component model.

Acknowledgements

The authors wish to thank the editor and the anonymous referee for their insightful comments and suggestions that helped to improve the paper.

This research was supported by the Academy of Finland.

References

Bera, A., John, S., 1983. Tests for multivariate normality with Pearson alternatives. Communications in Statistics — Theory and Methods 12, 103–117.

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

Chakraborty, B., Chaudhuri, P., 1999. On affine invariant sign and rank tests in one sample and two sample multivariate problems. In: Ghosh, S. (Ed.), Multivariate, Design and Sample Survey. Marcel-Dekker, New York, pp. 499–522.

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

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530.

Nordhausen, K., Oja, H., Ollila, E., 2010. Multivariate models and the first four moments. In: Hunter, D., Rosenberger, J., Richards, D. (Eds.), Festschrift for Thomas P. Hettmansperger (in press).

Nordhausen, K., Oja, H., Paindaveine, D., 2009. Signed-rank tests for location in the symmetric independent component model. Journal of Multivariate Analysis 100, 821–834.

Oja, H., Paindaveine, D., Taskinen, S., 2010. Parametric and nonparametric tests for multivariate independence in the independent component model (submitted for publication).

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

Serfling, R.J., 2010. Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. Journal of Nonparametric Statistics (in press).

Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H., 2009. Invariant co-ordinate selection. Journal of the Royal Statistical Society, Series B 71, 549–592.