AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 175–189

Scatter Matrices and Independent Component Analysis

Hannu Oja[1], Seija Sirkiä[2], and Jan Eriksson[3]

[1]University of Tampere, Finland
[2]University of Jyväskylä, Finland
[3]Helsinki University of Technology, Finland

Abstract: In the independent component analysis (ICA) it is assumed that the components of the multivariate independent and identically distributed observations are linear transformations of latent independent components. The problem then is to find the (linear) transformation which transforms the observations back to independent components. In the paper the ICA is discussed and it is shown that, under some mild assumptions, two scatter matrices may be used together to find the independent components. The scatter matrices must then have the so called independence property. The theory is illustrated by examples.

Keywords: Affine Equivariance, Elliptical Model, Independence, Independent Component Model, Kurtosis, Location, Principal Component Analysis (PCA), Skewness, Source Separation.

1 Introduction

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ denote a random sample from a p-variate distribution. We also write

$$\mathbf{X} = (\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n)'$$

for the corresponding $n \times p$ data matrix. In statistical modelling of the observed data, one often assumes that the observations $\mathbf{x}_i$ are independent p-vectors "generated" by the model

$$\mathbf{x}_i = \mathbf{A}\mathbf{z}_i + \mathbf{b}\,, \quad i = 1, \ldots, n\,,$$

where the $\mathbf{z}_i$'s are called *standardized vectors*, $\mathbf{b}$ is a *location* p*-vector*, $\mathbf{A}$ is a full-rank $p \times p$ *transformation matrix* and $\mathbf{V} = \mathbf{A}\mathbf{A}'$ is a positive definite $p \times p$ ($PDS(p)$) *scatter matrix*. In most applications (two-samples, several-samples case, linear model, etc.), $\mathbf{b} = \mathbf{b}_i$ may be dependent on the design. In a parametric model approach, one assumes that the distribution of the standardized vectors $\mathbf{z}_i$ are i.i.d. from a distribution known except for a finite number of parameters. A typical assumption then is that $\mathbf{z}_i \sim N_p(\mathbf{0}, \mathbf{I})$ implying that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. from $N(\mathbf{b}, \mathbf{V})$. A semiparametric *elliptical model* is constructed as follows: We assume that $\mathbf{U}\mathbf{z}_i \sim \mathbf{z}_i$ for all orthogonal $\mathbf{U}$. (By $\mathbf{x} \sim \mathbf{y}$ we mean that the probability distributions of $\mathbf{x}$ and $\mathbf{y}$ are the same.) Then the distribution of $\mathbf{z}_i$ is spherical, and the $\mathbf{x}_i$'s are elliptically symmetric. The density of $\mathbf{z}_i$ is then of the form

$$g(\mathbf{z}) = \exp\left\{-\rho(||\mathbf{z}||)\right\}\ .$$

The distribution of the $\mathbf{x}_i$ then depends on unknown location $\mathbf{b}$, scatter $\mathbf{V}$ and function ρ. If $\mathbf{z}$ is spherical then $d\mathbf{U}\mathbf{z}$ is spherical as well, for all $d \neq 0$ and for all orthogonal $\mathbf{U}$.

This implies that $\mathbf{A}$ is not well defined, and extra assumptions, such as $\mathrm{E}(||\mathbf{z}_i||^2) = 1$ or $\mathrm{med}(||\mathbf{z}_i||) = 1$, are needed to uniquely define $\mathbf{V}$ and ρ.

In this paper we consider an alternative semiparametric extension of the multivariate normal model called the *independent component (IC) model*. For this model one assumes that the components $z_{i1}, \ldots, z_{ip}$ of $\mathbf{z}_i$ are independent. The model is used in the so called *independent component analysis (ICA)* (Comon, 1994); recent textbooks provide an interesting tutorial material and partial review on ICA (Hyvärinen et al., 2001; Cichocki and Amari, 2002). In this model, the density of $\mathbf{z}_i$ is

$$g(\mathbf{z}) = \exp\left\{-\sum_{j=1}^{p} \rho_j(z_j)\right\} .$$

The distribution of $\mathbf{x}_i$ now depends on location $\mathbf{b}$, transformation $\mathbf{A}$ and marginal functions $\rho_1, \ldots, \rho_p$. If $\mathbf{z}$ has independent components then $\mathbf{DPz}$ has independent components as well, for all diagonal $p \times p$ matrices $\mathbf{D}$ and for all permutation matrices $\mathbf{P}$. Extra assumptions are then needed to uniquely define $\mathbf{A}$, $\mathbf{b}$ and $\rho_1, \ldots, \rho_p$.

In this paper, under some mild assumptions and using two scatter matrices, we reformulate (and restrict) the model so that $\mathbf{A}$ uniquely defined (up to sign changes of its column vectors) even without specifying $\rho_1, \ldots, \rho_p$. The independent components are then standardized with respect to the first scatter matrix, uncorrelated with respect to the second one, and ordered according to kurtosis. The final aim often is to *separate the sources*, i.e., to estimate the inverse matrix $\mathbf{B} = \mathbf{A}^{-1}$; transformation $\mathbf{B}$ transforms the observed vectors to to vectors with independent components.

Our plan in this paper is as follows. In Section 2 we introduce the concepts of multivariate location and scatter functionals, give several examples and discuss their use in the analysis of multivariate data. We show, for example, how two scatter matrices can be used to describe the multivariate kurtosis. In Section 3, the ICA problem is then discussed and we introduce a new class of estimators of the ICA transformation matrix $\mathbf{B}$. The assumptions and properties of the estimators are shortly discussed. The paper ends with two examples in Section 4. Throughout the paper, notations $\mathbf{U}$ and $\mathbf{V}$ are used for orthogonal $p \times p$ matrices. $\mathbf{D}$ is a $p \times p$ diagonal matrix and $\mathbf{P}$ is a permutation matrix (obtained from the identity matrix $\mathbf{I}$ by successively permuting its rows or columns). Finally, let $\mathbf{J}$ be a sign change matrix, that is, a diagonal matrix with diagonal elements ± 1. For a positive definite symmetric matrix $\mathbf{V}$, the matrices $\mathbf{V}^{1/2}$ and $\mathbf{V}^{-1/2}$ are taken to be symmetric as well.

2 Location Vectors and Scatter Matrices

2.1 Definitions

We first define what we mean by a location vector and a scatter matrix. Let $\mathbf{x}$ be a p-variate random variable with cdf F. A functional $\mathbf{T}(F)$ or $\mathbf{T}(\mathbf{x})$ is a p-variate *location vector* if it is affine equivariant, that is,

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

for all random vectors $\mathbf{x}$, full-rank $p \times p$-matrices $\mathbf{A}$ and p-vectors $\mathbf{b}$. A matrix-valued functional $\mathbf{S}(F)$ or $\mathbf{S}(\mathbf{x})$ is a *scatter matrix* if it is a positive definite symmetric $p \times p$-matrix, write $PDS(p)$, and affine equivariant in the sense that

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

for all random vectors $\mathbf{x}$, full-rank $p \times p$-matrices $\mathbf{A}$ and p-vectors $\mathbf{b}$. "Classical" location and scatter functionals, namely the mean vector $\mathrm{E}(\mathbf{x})$ and the covariance matrix

$$\mathrm{cov}(\mathbf{x}) = E\left((\mathbf{x} - \mathrm{E}(\mathbf{x}))(\mathbf{x} - \mathrm{E}(\mathbf{x}))'\right) ,$$

serve as first examples. If the distribution of $\mathbf{x}$ is elliptically symmetric around $\mathbf{b}$ then $\mathbf{T}(\mathbf{x}) = \mathbf{b}$ for all location vectors $\mathbf{T}$. Moreover, if the distribution of $\mathbf{x}$ is elliptically symmetric and the covariance matrix $\mathrm{cov}(\mathbf{x})$ exists then $\mathbf{S}(\mathbf{x}) \propto \mathrm{cov}(\mathbf{x})$ for all scatter matrices $\mathbf{S}$. There are several alternative competing techniques to construct location and scatter functionals, e.g., M-functionals, S-functionals and τ-functionals just to mention a few. These functionals and related estimates are thoroughly discussed in numerous papers (Maronna, 1976; Davies, 1987; Lopuhaä, 1989; Lopuhaä, 1991; Tyler, 2002); the common feature is that the functionals and related estimates are built for inference in elliptic models only. Next we consider some M-functionals in more details.

2.2 M-Functionals of Location and Scatter

Location and scatter *M-functionals* are sometimes defined as functionals $\mathbf{T}(\mathbf{x})$ and $\mathbf{S}(\mathbf{x})$ which simultaneously satisfy implicit equations

$$\mathbf{T}(\mathbf{x}) = \left[\mathrm{E}[w_1(r)]\right]^{-1} \mathrm{E}\left[w_1(r)\mathbf{x}\right]$$

and

$$\mathbf{S}(\mathbf{x}) = \mathrm{E}\left[w_2(r)(\mathbf{x} - \mathbf{T}(\mathbf{x}))(\mathbf{x} - \mathbf{T}(\mathbf{x}))'\right]$$

for some suitably chosen weight functions $w_1(r)$ and $w_2(r)$. The random variable r is the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{T}(\mathbf{x})$, i.e.

$$r^2 = ||\mathbf{x} - \mathbf{T}(\mathbf{x})||^2_{\mathbf{S}(\mathbf{x})} = (\mathbf{x} - \mathbf{T}(\mathbf{x}))'[\mathbf{S}(\mathbf{x})]^{-1}(\mathbf{x} - \mathbf{T}(\mathbf{x})) .$$

Mean vector and covariance matrix are again simple examples with choices $w_1(r) = w_2(r) = 1$. If $\mathbf{T}_1(\mathbf{x})$ and $\mathbf{S}_1(\mathbf{x})$ are any affine equivariant location and scatter functionals then one-step M-functionals, starting from $\mathbf{T}_1$ and $\mathbf{S}_1$, are given by

$$\mathbf{T}_2(\mathbf{x}) = \left[\mathrm{E}[w_1(r)]\right]^{-1} \mathrm{E}\left[w_1(r)\mathbf{x}\right]$$

and

$$\mathbf{S}_2(\mathbf{x}) = \mathrm{E}\left[w_2(r)(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))(\mathbf{x} - \mathbf{T}_1(\mathbf{x}))'\right] ,$$

where now $r = ||\mathbf{x} - \mathbf{T}_1(\mathbf{x})||_{\mathbf{S}_1(\mathbf{x})}$. It is easy to see that $\mathbf{T}_2$ and $\mathbf{S}_2$ are affine equivariant as well. Repeating this step until it converges often yields the "final" M-estimate with weight functions w_1 and w_2. If $\mathbf{T}_1$ is the mean vector and $\mathbf{S}_1$ is the covariance matrix, then

$$\mathbf{T}_2(\mathbf{x}) = \frac{1}{p}\mathrm{E}[r^2\mathbf{x}] \qquad \text{and} \quad \mathbf{S}_2(\mathbf{x}) = \frac{1}{p+2}E[r^2(\mathbf{x} - \mathrm{E}(\mathbf{x}))(\mathbf{x} - \mathrm{E}(\mathbf{x}))']$$

are interesting one-step location and scatter M-functionals based on third and fourth moments, respectively.

2.3 Multivariate Sign and Rank Covariance Matrices

Consider next multivariate sign and rank covariance matrices. Locantore et al. (1999), Marden (1999), Visuri et al. (2000), and Croux et al. (2002) considered the so called *spatial sign covariance matrix* with a fixed location $\mathbf{T}(\mathbf{x})$

$$\mathrm{E}\left[\frac{(\mathbf{x}-\mathbf{T}(\mathbf{x}))(\mathbf{x}-\mathbf{T}(\mathbf{x}))'}{||\mathbf{x}-\mathbf{T}(\mathbf{x})||^2}\right]$$

and used it as a tool for robust principal component analysis in the elliptic case. The spatial sign covariance matrix is not a genuine scatter matrix, however. It is not affine equivariant but only orthogonally equivariant.

To define a multivariate rank covariance matrix, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two independent copies of $\mathbf{x}$. The *spatial Kendall's tau* matrix (Visuri et al., 2000)

$$\mathrm{E}\left[\frac{(\mathbf{x}_1-\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)'}{||\mathbf{x}_1-\mathbf{x}_2||^2}\right]$$

is not a scatter matrix either. It is again only orthogonally equivariant. Note that no location center is needed to define the spatial Kendall's tau.

Related scatter matrices may be constructed as follows. The *Tyler (1987) scatter matrix* (with fixed location $\mathbf{T}(\mathbf{x})$) is sometimes referred to as most robust M-functional and is given by implicit equation

$$\mathbf{S}(\mathbf{x}) = p\mathrm{E}\left[\frac{(\mathbf{x}-\mathbf{T}(\mathbf{x}))(\mathbf{x}-\mathbf{T}(\mathbf{x}))'}{||\mathbf{x}-\mathbf{T}(\mathbf{x})||^2_{\mathbf{S}(\mathbf{x})}}\right].$$

Note that Tyler's matrix is characterized by the fact that the spatial sign covariance matrix of the transformed random variable $\mathbf{S}(\mathbf{x})^{-1/2}(\mathbf{x}-\mathbf{T}(\mathbf{x}))$ is $[1/p]\mathbf{I}$.

The *Dümbgen (1998) scatter matrix* is defined in an analogous way but using the spatial Kendall's tau matrix: Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two independent copies of $\mathbf{x}$. Dümbgen's matrix is then implicitly defined by

$$\mathbf{S}(\mathbf{x}) = p\mathrm{E}\left[\frac{(\mathbf{x}_1-\mathbf{x}_2)(\mathbf{x}_1-\mathbf{x}_2)'}{||\mathbf{x}_1-\mathbf{x}_2||^2_{\mathbf{S}(\mathbf{x})}}\right].$$

Tyler's and Dümbgen's matrices are not genuine scatter matrices as they are defined only up to a constant and affine equivariant only in the sense that

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

This is, however, sufficient in most of applications.

2.4 Why do we need different Location and Scatter Functionals?

Write again $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ for a $n \times p$ data matrix with cdf F_n, and write $\mathbf{T}(\mathbf{X})$ and $\mathbf{S}(\mathbf{X})$ for location and scatter statistics at F_n. Different location estimates

$$\mathbf{T}(\mathbf{X}), \mathbf{T}_1(\mathbf{X}), \mathbf{T}_2(\mathbf{X}), \ldots$$

and scatter estimates

$$\mathbf{S}(\mathbf{X}), \mathbf{S}_1(\mathbf{X}), \mathbf{S}_2(\mathbf{X}), \ldots ,$$

possibly with correction factors, often estimate the same population quantities but have different statistical properties (convergence, limiting distributions, efficiency, robustness, computation, etc.) As mentioned before, this is true in the elliptic model, for example. In practice, one can then just pick up an estimate which is most suitable to one's purposes.

Location and scatter statistics may be used to describe the skewness and kurtosis properties of a multivariate distribution as well. Affine invariant multivariate *skewness statistics* may be defined as squared Mahalanobis distances between two location statistics

$$||\mathbf{T}_1 - \mathbf{T}_2||^2_{\mathbf{S}} .$$

If $\mathbf{T}_1$ and $\mathbf{T}_2$ are the multivariate mean vector and an affine equivariant multivariate median, and $\mathbf{S} = Cov$ is the covariance matrix, then an extension of the classical univariate Pearson (1895) measure of asymmetry (mean-median)/σ is obtained. A multivariate generalization of the classical standardized third moment, the most popular measure of asymmetry, is given if one uses $\mathbf{T}_1 = E$ and $\mathbf{S} = \mathbf{S}_1 = Cov$ and $\mathbf{T}_2$ is the one-step location M-estimator with $w_1(r) = r^2$.

As $\mathbf{u}'\mathbf{S}\mathbf{u}$ is a scale measure for linear combination $\mathbf{u}'\mathbf{x}$, the ratio $(\mathbf{u}'\mathbf{S}_2\mathbf{u})/(\mathbf{u}'\mathbf{S}_1\mathbf{u})$ is a descriptive statistic for kurtosis of $\mathbf{u}'\mathbf{x}$, and finally all eigenvalues of $\mathbf{S}_2\mathbf{S}_1^{-1}$, say $d_1 \geq \cdots \geq d_p$ may be used to describe the *multivariate kurtosis*. Again, if $\mathbf{T}_1$ and $\mathbf{S}_1$ are the mean vector and covariance matrix, respectively, and $\mathbf{S}_2$ is the one-step M-estimator with $w_2(r) = r^2$, the eigenvalues of $\mathbf{S}_2\mathbf{S}_1^{-1}$ depend on the fourth moments of the standardized observations only. For a discussion on multivariate skewness and kurtosis statistics with comparisons to Mardia (1970) statistics, see Kankainen et al. (2005).

Scatter matrices are often used to standardize the data. The transformed, standardized data set

$$\mathbf{Z} = \mathbf{X}[\mathbf{S}(\mathbf{X})]^{-1/2}$$

has uncorrelated components with respect to $\mathbf{S}$ (i.e., $\mathbf{S}(\mathbf{Z}) = \mathbf{I}$), and the observations $\mathbf{z}_i$ tend to be spherically distributed in the elliptic case. Unfortunately, the transformed data set $\mathbf{Z}$ is not *coordinate-free*: It is **not** generally true that, for any full rank $\mathbf{A}$,

$$\mathbf{X}\mathbf{A}'[\mathbf{S}(\mathbf{X}\mathbf{A}')]^{-1/2} = \mathbf{X}[\mathbf{S}(\mathbf{X})]^{-1/2} .$$

The *spectral or eigenvalue decomposition* of $\mathbf{S}(\mathbf{X})$ is

$$\mathbf{S}(\mathbf{X}) = \mathbf{U}(\mathbf{X})\ \mathbf{D}(\mathbf{X})\ (\mathbf{U}(\mathbf{X}))' ,$$

where the columns of $p \times p$ orthogonal matrix $\mathbf{U}(\mathbf{X})$ are the eigenvectors of $\mathbf{S}(\mathbf{X})$ and diagonal matrix $\mathbf{D}(\mathbf{X})$ lists the corresponding eigenvalues in a decreasing order. Then the components of the transformed data matrix $\mathbf{Z} = \mathbf{X}\mathbf{U}(\mathbf{X})$ are the so called *principal components*, used in *principal component analysis (PCA)*. The principal components are uncorrelated with respect to $\mathbf{S}$ (as $\mathbf{S}(\mathbf{Z}) = \mathbf{D}(\mathbf{X})$) and therefore *ordered according to their dispersion*. This transformed data set is not coordinate-free either.

2.5 Scatter Matrices and Independence

Write $\mathbf{x} = (x_1, \ldots, x_p)'$ and assume that the components $x_1, \ldots, x_p$ are independent. It is then well known that the regular covariance matrix $\mathrm{cov}(\mathbf{x})$ (if it exists) is a diagonal matrix. The M-functionals, S-functionals, τ-functionals, etc., are meant for inference in elliptical models and do not generally have this property:

Definition 1 *If the scatter functional* $\mathbf{S}(\mathbf{x})$ *is a diagonal matrix for all* $\mathbf{x}$ *with independent components, then* $\mathbf{S}$ *is said to have the independence property.*

A natural question then is whether, in addition to the covariance matrix, there are any other scatter matrices with the same independence property. The next theorem shows that, in fact, any scatter matrix yields a symmetrized version which has this property.

Theorem 1 *Let* $\mathbf{S}(\mathbf{x})$ *be any scatter matrix. Then*

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

where $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *are two independent copies of* $\mathbf{x}$*, is a scatter matrix with the independence property.*

Proof $\mathbf{S}_s$ is affine equivariant as $\mathbf{S}$ is affine equivariant. Assume that the components of $\mathbf{x}$ are independent. The components of $\mathbf{x}_1 - \mathbf{x}_2$ are then independent as well and symmetrically distributed around zero implying that $\mathbf{J}(\mathbf{x}_1 - \mathbf{x}_2) \sim (\mathbf{x}_1 - \mathbf{x}_2)$ for all diagonal matrices $\mathbf{J}$ with diagonal elements ± 1. This further implies that $[\mathbf{S}(\mathbf{x}_1 - \mathbf{x}_2)]_{ij} = -[\mathbf{S}(\mathbf{x}_1 - \mathbf{x}_2)]_{ij}$ for all $i \neq j$ and $\mathbf{S}(\mathbf{x}_1 - \mathbf{x}_2)$ must be diagonal. $Q.E.D.$

Another possibility to construct scatter matrix estimates with the independence property might be to use quasi-maximum likelihood estimates ("M-estimates") in the independent component model, that is, the regular maximum likelihood estimates under some specific choices of the marginal distribution (in which one not necessarily believes). See e.g. Pham and Garat (1997) for the use of quasi-ML estimation in the ICA model.

3 Independent Component Analysis (ICA)

3.1 Problem

The ICA problem in its simplest form is as follows. According to the general *independent component model (IC)*, the observed random p-vector $\mathbf{x}$ is generated by

$$\mathbf{x} = \mathbf{A}_0\mathbf{s}\,,$$

where $\mathbf{s} = (s_1, \ldots, s_p)'$ has independent components and $\mathbf{A}_0$ is a full-rank $p \times p$ transformation matrix. For uniqueness of $\mathbf{A}_0$ one usually assumes that at most one component is gaussian (normally distributed). The question then is: Having transformed $\mathbf{x}$, is it possible to retransform to independent components, that is, can one find $\mathbf{B}$ such that $\mathbf{Bx}$ has independent components? See e.g. Hyvärinen et al. (2001).

Clearly the above model is the independent component model (IC model) described in the Introduction. If $\mathbf{D}$ is a $p \times p$ diagonal matrix and $\mathbf{P}$ a $p \times p$ permutation matrix, then one can write

$$\mathbf{x} = (\mathbf{A}_0\mathbf{P}^{-1}\mathbf{D}^{-1})(\mathbf{DPs})\,,$$

where $\mathbf{DPs}$ has independent components as well. Thus $\mathbf{s}$ may be defined only up to multiplying constants and a permutation. In the estimation problem this means that, if $\mathbf{B}$ is a solution, $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ a permutation matrix, then also $\mathbf{DPB}$ is a solution. In fact, it can be shown that $\mathbf{Bx}$ has independent components if and only if $\mathbf{B} = \mathbf{DPA}_0^{-1}$ for some $\mathbf{D}$ and $\mathbf{P}$. The model is then called *separable*; see Comon (1994) and Eriksson and Koivunen (2004).

3.2 Independent Component Models

We now try to fix the parameters in the IC model using a location functional $\mathbf{T}$ and two different scatter functionals $\mathbf{S}_1$ and $\mathbf{S}_2$. Both scatter functionals are assumed to have the independence property.

Definition 2 *The independent component model I (IC-I), formulated using* $\mathbf{T}$, $\mathbf{S}_1$, *and* $\mathbf{S}_2$, *is*

$$\mathbf{x} = \mathbf{Az} + \mathbf{b}$$

where $\mathbf{z}$ *has independent components with* $\mathbf{T}(\mathbf{z}) = \mathbf{0}$,

$$\mathbf{S}_1(\mathbf{z}) = \mathbf{I} \qquad \textit{and} \quad \mathbf{S}_2(\mathbf{z}) = \mathbf{D}(\mathbf{z})\,,$$

and $\mathbf{D}(\mathbf{z})$ *is a diagonal matrix with diagonal elements* $d_1 \geq \cdots \geq d_p$ *in a descending order.*

If $\mathbf{T}$ is the mean vector, $\mathbf{S}_1$ is the covariance matrix, and $\mathbf{S}_2$ is the scatter matrix based on fourth moment, see Section 2.2, then $\mathrm{E}(z_i) = 0$, $\mathrm{var}(z_i) = 1$ and the components are ordered according to the classical univariate kurtosis measure based on standardized fourth moment.

First note that the reformulation of the model in Definition 2 can always be done; it is straightforward to see that $\mathbf{z} = \mathbf{D}^*\mathbf{P}^*\mathbf{s} + \mathbf{b}^*$ for some specific choices $\mathbf{D}^*$, $\mathbf{P}^*$, and $\mathbf{b}^*$ (depending on $\mathbf{T}$, $\mathbf{S}_1$, and $\mathbf{S}_2$). In the model, the diagonal matrix $\mathbf{D}(\mathbf{z})$ lists the eigenvalues of $\mathbf{S}_2(\mathbf{x})[\mathbf{S}_1(\mathbf{x})]^{-1}$ for any $\mathbf{x} = \mathbf{Az} + \mathbf{b}$ in the model. Recall from Section 2.4 also that the ratio

$$\frac{\mathbf{u}'\mathbf{S}_2(\mathbf{z})\mathbf{u}}{\mathbf{u}'\mathbf{S}_1(\mathbf{z})\mathbf{u}} = \sum_{i=1}^{p} u_i^2 d_i$$

gives the kurtosis of $\mathbf{u}'\mathbf{z}$. Therefore, in this formulation of the model, the *independent components are ordered according to their marginal kurtosis*. The order is either from the lowest kurtosis to the highest one or vice versa, depending on the specific choices of $\mathbf{S}_1$ and $\mathbf{S}_2$.

Next note that IC-I model is, in addition to the elliptic model, another possible extension of the multivariate normal model: If a p-variate elliptic distribution is included in the IC-I model, it must be a multivariate normal distribution.

Finally note that the transformation matrix $\mathbf{A}$ is unfortunately not uniquely defined: This happens, e.g., if any two of the independent components of $\mathbf{z}$ have the same marginal distribution. For uniqueness, we need the following restricted model.

Definition 3 *The independent component model II (IC-II) corresponding to* $\mathbf{T}$, $\mathbf{S}_1$, *and* $\mathbf{S}_2$, *is*

$$\mathbf{x} = \mathbf{Az} + \mathbf{b}\,,$$

where $\mathbf{z}$ *has independent components with* $\mathbf{T}(\mathbf{z}) = \mathbf{0}$, $\mathbf{S}_1(\mathbf{z}) = \mathbf{I}$, *and* $\mathbf{S}_2(\mathbf{z}) = \mathbf{D}(\mathbf{z})$ *and* $\mathbf{D}(\mathbf{z})$ *is a diagonal matrix with diagonal elements* $d_1 > \cdots > d_p$ *in a descending order.*

Note that the multivariate normal model is not included in the IC-II model any more; $z_1, \ldots, z_p$ can not be i.i.d. The assumption that $d_1 > \ldots > d_p$ guarantees the uniqueness of the location vector $\mathbf{b}$ and the transformation matrix $\mathbf{A}$ (up to sign changes of its columns). The retransformation matrix $\mathbf{B}$ is then unique up to sign changes of its rows, but could be made unique if one chooses the $\mathbf{JB}$ for which the highest value in each row is positive.

Now we are ready to give the main result of the paper.

Theorem 2 *Assume that the independent component model IC-II in Definition 3 is true. Write*

$$\mathbf{B}(\mathbf{x}) = \left[\mathbf{U}_2\left([\mathbf{S}_1(\mathbf{x})]^{-1/2}\mathbf{x}\right)\right]'[\mathbf{S}_1(\mathbf{x})]^{-1/2},$$

where $\mathbf{U}_2(\mathbf{x})$ *is the matrix of unit eigenvectors of* $\mathbf{S}_2(\mathbf{x})$ *(with corresponding eigenvalues in a decreasing order). Then*

$$\mathbf{B}(\mathbf{x})\left(\mathbf{x} - \mathbf{T}(\mathbf{x})\right) = \mathbf{J}\mathbf{z}$$

for some diagonal matrix $\mathbf{J}$ *with diagonal elements* ± 1.

Proof Assume that the model is true and write (singular value decomposition) $\mathbf{A} = \mathbf{U}\mathbf{L}\mathbf{V}'$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{L}$ is a diagonal matrix (with nonzero diagonal elements). It is not a restriction to assume that $\mathbf{T}(\mathbf{x}) = \mathbf{b} = \mathbf{0}$. Thus

$$\mathbf{x} = \mathbf{U}\mathbf{L}\mathbf{V}'\mathbf{z}\,.$$

Then $\mathbf{S}_1(\mathbf{x}) = \mathbf{U}\mathbf{L}^2\mathbf{U}'$, $[\mathbf{S}_1(\mathbf{x})]^{-1/2} = \mathbf{U}\mathbf{L}^{-1}\mathbf{U}'$,

$$[\mathbf{S}_1(\mathbf{x})]^{-1/2}\mathbf{x} = \mathbf{U}\mathbf{V}'\mathbf{z}\,,$$

and

$$\mathbf{S}_2\left([\mathbf{S}_1(\mathbf{x})]^{-1/2}\mathbf{x}\right) = \mathbf{U}\mathbf{V}'\mathbf{D}\mathbf{V}\mathbf{U}'\,,$$

implying that

$$\mathbf{U}_2\left([\mathbf{S}_1(\mathbf{x})]^{-1/2}\mathbf{x}\right) = \mathbf{U}\mathbf{V}'\,.$$

The result then follows. $Q.E.D.$

Remark 1 *From the proof of Theorem 2 one also sees that it is, in fact, enough to assume that* $\mathbf{S}_2$ *(with the independence property) is only* orthogonal equivariant *in the sense that*

$$\mathbf{S}_2(\mathbf{U}\mathbf{x} + \mathbf{b}) \propto \mathbf{U}\mathbf{S}_2(\mathbf{x})\mathbf{U}'$$

for all random vectors $\mathbf{x}$, *orthogonal* $\mathbf{U}$ *and* p*-vectors* $\mathbf{b}$. *The functional* $\mathbf{S}_2$ *could then be the spatial Kendall's tau, for example.*

3.3 Discussion on ICA Transformation

Theorem 2 proposes a new class of estimators

$$\mathbf{B}(\mathbf{X}) = \left[\mathbf{U}_2\left(\mathbf{X}[\mathbf{S}_1(\mathbf{X})]^{-1/2}\right)\right]'[\mathbf{S}_1(\mathbf{X})]^{-1/2}$$

for the transformation matrix $\mathbf{B}(\mathbf{x})$. In the two examples in Section 4, we use $\mathbf{S}_1 = Cov$ as the first scatter matrix, and the second functional is

$$\mathbf{S}_2(\mathbf{x}) = \text{cov}\left(||\mathbf{x}_1 - \mathbf{x}_2||^{-1}(\mathbf{x}_1 - \mathbf{x}_2)\right) \quad \text{and} \quad \mathbf{S}_2(\mathbf{x}) = \text{cov}\left(||\mathbf{x}_1 - \mathbf{x}_2||(\mathbf{x}_1 - \mathbf{x}_2)\right),$$

respectively. Then $\mathbf{S}_2$ is orthogonally equivariant only, but see Remark 1. The first choice of $\mathbf{S}_2$ is then the Kendall's tau matrix and the second one is a matrix of fourth moments of the differences. Note, however, that the matrices at the second stage might be seen as the (affine equivariant) symmetrized one-step M-estimator as well (with weight functions $w_2(r) = r^{-2}$ and $w_2(r) = r^2$, respectively). Further theoretical work and simulations are needed to consider the efficiency and robustness properties of estimates $\mathbf{B}(\mathbf{X})$ based on different choices of $\mathbf{S}_1$ and $\mathbf{S}_2$.

In our independent component model IC-II with $\mathbf{B} = \mathbf{A}^{-1}$,

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

and the estimates $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ solve

$$(\mathbf{S}_2(\mathbf{X}))^{-1}\mathbf{S}_1(\mathbf{X})\hat{\mathbf{B}}' = \hat{\mathbf{B}}'\hat{\mathbf{D}}^{-1}$$

(David Tyler, 2005, personal communication). The rows of the transformation matrix $\mathbf{B}$ and diagonal elements of $\mathbf{D}^{-1}$ thus list the eigenvectors and eigenvalues of $\mathbf{S}_2^{-1}\mathbf{S}_1$. The asymptotical properties (convergence, limiting distributions, limiting variances and covariances) of $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ may then be derived from those of $\mathbf{S}_1(\mathbf{X})$ and $\mathbf{S}_2(\mathbf{X})$. It is also easy to see that, for any two scatter matrices $\mathbf{S}_1$ and $\mathbf{S}_2$, $\mathbf{Z} = \mathbf{X}[\mathbf{B}(\mathbf{X})]'$ allows a coordinate-free presentation of the data cloud up to signs of the components: $\mathbf{X}\mathbf{A}'[\mathbf{B}(\mathbf{X}\mathbf{A}')]' = \mathbf{J}\mathbf{X}[\mathbf{B}(\mathbf{X})]'$ for any full-rank $\mathbf{A}$. Recall also that the components of $\mathbf{Z}$ are then ordered according to their kurtosis. In the literature, most of the algorithms for finding an ICA transformation (i) start with whitening the data (with the regular covariance matrix), and (ii) end with rotating the transformed data to minimize the value of an objective function. The objective function is then typically a measure of dependence between the components, and an iterative algorithm is used for the minimization problem. In our examples in Section 4, we use the regular covariance matrix and one-step M-estimators; the transformation matrix can then be given as an explicit formula, and no iteration is needed. The convergence of the transformation matrix estimate to the true value as well as its distributional behavior may be traced from the corresponding properties of the two scatter matrices.

The idea to use two scatter matrices in estimating the ICA transformation seems to be implicit in many work reported in the literature. One of the first ICA algorithms FOBI, Cardoso (1989), uses the regular covariance matrix $\mathbf{S}_1(\mathbf{x}) = \text{cov}(\mathbf{x})$ to whiten data, and then in the second stage a fourth-order cumulant matrix $\mathbf{S}_2(\mathbf{x}) = \text{cov}(||\mathbf{x} - \text{E}[\mathbf{x}]||(\mathbf{x} -$

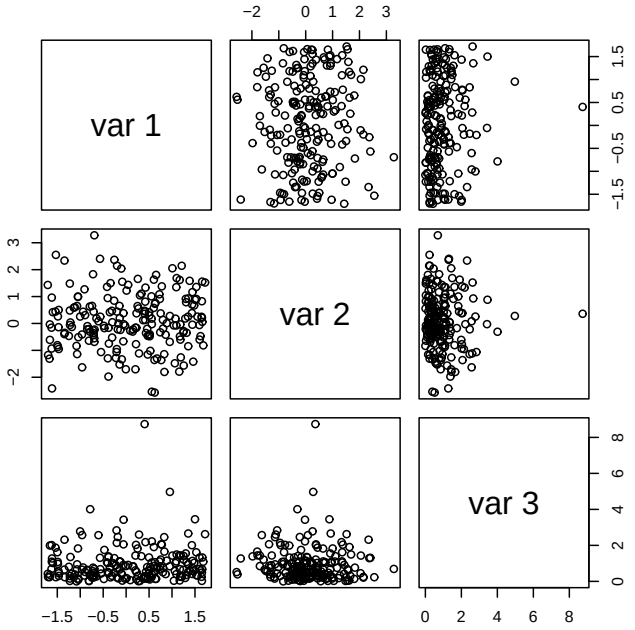

Figure 1: Toy example: Scatter plots for independent components ($\mathbf{Z}$)

$\mathrm{E}[\mathbf{x}])$), which is an orthogonally equivariant scatter matrix with the independence property. The eigenvalues of $\mathbf{S}_2(\mathbf{z})$ are given by $\mathrm{E}[z_k^4] + p - 1$, $k = 1, \ldots, p$, and therefore identically distributed components can not be separated by Theorem 2. This is often seen as a severe restriction from the application point of view. Later FOBI was generalized to JADE (Cardoso and Souloumiac, 1993; see also Cardoso, 1999), where the scatter matrix $\mathbf{S}_2$ is replaced by other cleverly chosen fourth-order cumulant matrices. This generalization allows the separation of identically distributed components, but the independence property is lost, and one needs to resort to computationally ineffective methods instead of straightforward eigenvalue decompositions. We are currently investigating the possibility of generalizing the cumulant-based view of JADE to general scatter matrices. We still wish to mention one alternative approach: Samarov and Tsybakov (2004) simultaneously estimated the transformation matrix and unknown marginal distributions. In our procedure we avoid the estimation of the margins.

4 Two Examples

4.1 A Toy Example

We end this paper with two examples. We first consider a 3-variate random vector $\mathbf{z}$ whose components are independent and have the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$, the standard normal distribution and the exponential distribution with scale parameter 1. A sample of 200 observations from such a distribution is plotted in Figure 1.

To fix the model with the related estimate, choose as $\mathbf{S}_1$ the regular covariance matrix and as $\mathbf{S}_2$ the one step Dümbgen estimator. In this case $\mathbf{S}_1(\mathbf{z})$ is obviously the identity matrix and numerical integration gives that $\mathbf{S}_2(\mathbf{z})$ is

$$\mathbf{D} = \mathrm{diag}(0.37, 0.35, 0.28)\,.$$

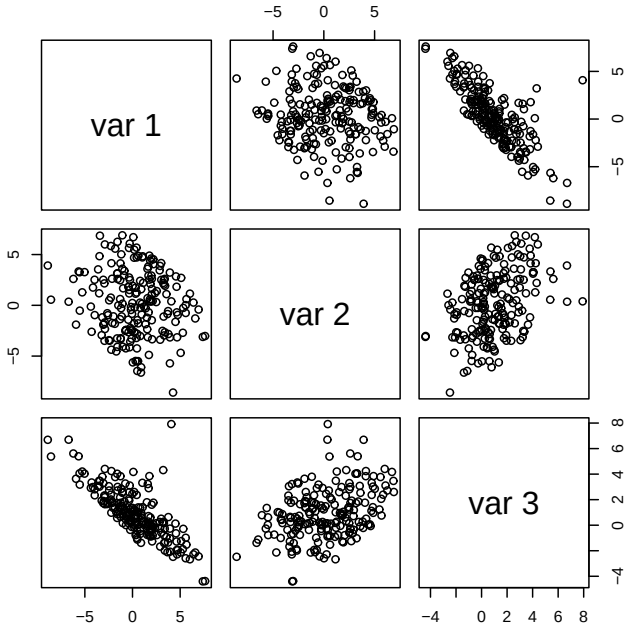

Figure 2: Toy example: Scatter plots for observed data cloud ($\mathbf{X}$)

The assumptions of Theorem 2 are then met. With these choices, the order of the components is then from the lowest kurtosis to the highest one. As a mixing matrix $\mathbf{A}$ consider a combination of permutation of the components to the order $3, 1, 2$, a rescaling of the (new) second and third components by 3 and 5, respectively, and finally rotations of $\pi/4$ around the axis z, y, and x, in this order. The approximate value of the true unmixing matrix, transformation matrix $\mathbf{B}$ is then

$$\mathbf{B} = \begin{pmatrix} 0.17 & 0.28 & -0.05 \\ -0.20 & 0.14 & 0.14 \\ 0.50 & -0.15 & 0.85 \end{pmatrix} .$$

See Figure 2 for a plot of the observed data set. Following our estimation procedure gives

$$\hat{\mathbf{B}} = \begin{pmatrix} 0.14 & 0.30 & -0.07 \\ 0.23 & -0.11 & -0.11 \\ 0.49 & -0.15 & 0.89 \end{pmatrix}$$

as the estimated ICA transformation matrix (unmixing matrix) with the estimated kurtosis matrix

$$\hat{\mathbf{D}} = \text{diag}(0.39, 0.36, 0.25) .$$

Both are close to the true values but the second row of the unmixing matrix has changed sign. The successful estimation is clearly visible in Figure 3 which shows a plot of the estimated independent components. Apart from the flipped axis the plot is almost identical to the original plot.

4.2 Example: Swiss Bank Notes

Assume that the distribution of $\mathbf{x}$ is a mixture of two multivariate normal distribution differing only in location: $\mathbf{x}$ has a $N_p(\mu_1, \mathbf{\Sigma})$-distribution with probability $1 - \epsilon$ and a

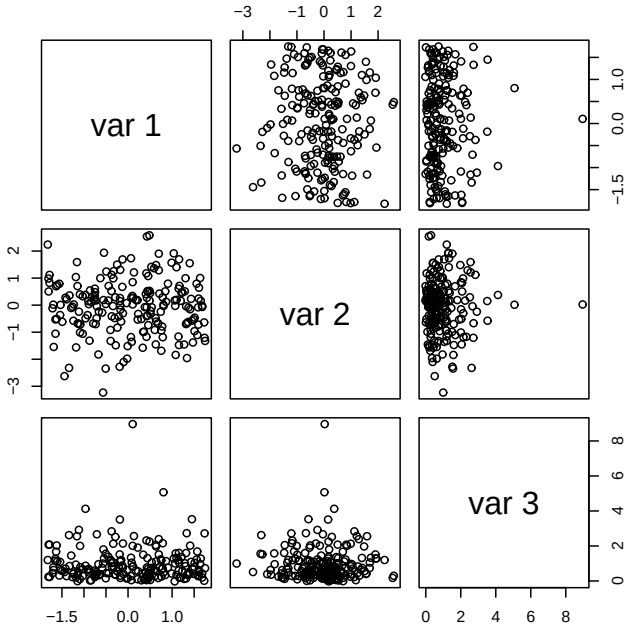

Figure 3: Toy example: Scatter plots for estimated independent components $(\mathbf{X}(\mathbf{B}(\mathbf{X}))')$

$N_p(\mu_2, \mathbf{\Sigma})$-distribution with probability ϵ. The distribution lies in the IC-I model, and one possible vector of independent components $\mathbf{s}$ is the mixture of multivariate normal distributions $N_p(\mathbf{0}, \mathbf{I})$ and $N_p(\mu, \mathbf{I})$, where $\mu' = (0, \ldots, 0, c)$ with $c^2 = (\mu_2 - \mu_1)'\mathbf{\Sigma}^{-1}(\mu_2 - \mu_1)$. Note that the last component with lowest kurtosis can be identified with two scatter matrices; it is found in the direction of $\mu_2 - \mu_1$. In this example $\mathbf{S}_1$ is again chosen to be the regular covariance matrix and $\mathbf{S}_2(\mathbf{x}) = \text{cov}(||\mathbf{x}_1 - \mathbf{x}_2||(\mathbf{x}_1 - \mathbf{x}_2))$ (a matrix of fourth moments of the differences). With these choices, the last row of the transformation matrix $\mathbf{B}$ yields the direction of lowest kurtosis.

In this example we analyze the data set appeared in Flury and Riedwyl (1988). The data contain measurements on 100 genuine and 100 forged thousand franc bills. Each row in the data matrix contains the six measurements for a bill ($n = 200, p = 6$), namely length of bill (x_1), width of bill measured on the left (x_2), width of bill measured on the right (x_3), width of the margin at the bottom (x_4), width of the margin at the top (x_5), and length of the image diagonal (x_6). See Figure 4 for the data set of 200 observations. *We analyze the data without knowing which of the bills are genuine and which are forged*. So the distribution of the 6-variate random variable $\mathbf{x}$ may be thought to be a mixture of two normal distribution.

The estimates of $\mathbf{B}$ and $\mathbf{D}$ are now

$$\hat{\mathbf{B}} = \begin{pmatrix} -1.18 & 1.74 & -0.07 & -0.78 & -0.71 & -0.98 \\ 2.40 & -2.33 & 1.03 & -0.19 & -0.29 & -0.91 \\ 0.80 & 1.36 & -3.61 & 0.17 & 0.71 & -0.15 \\ 0.98 & 2.64 & -1.24 & 0.15 & -0.79 & 0.06 \\ -0.39 & -1.37 & -0.50 & 0.09 & -1.06 & -0.68 \\ -0.27 & 0.43 & -0.20 & 0.57 & 0.35 & -0.03 \end{pmatrix}$$

and

$$\hat{\mathbf{D}} = \text{diag}(41.0, 39.7, 35.1, 32.9, 31.5, 26.4)\,,$$

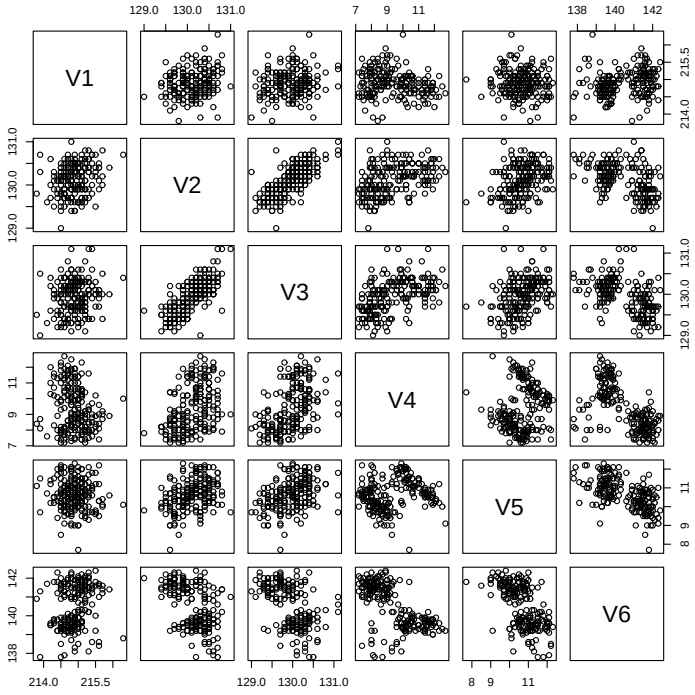

Figure 4: Swiss bank notes: Original data set ($\mathbf{X}$)

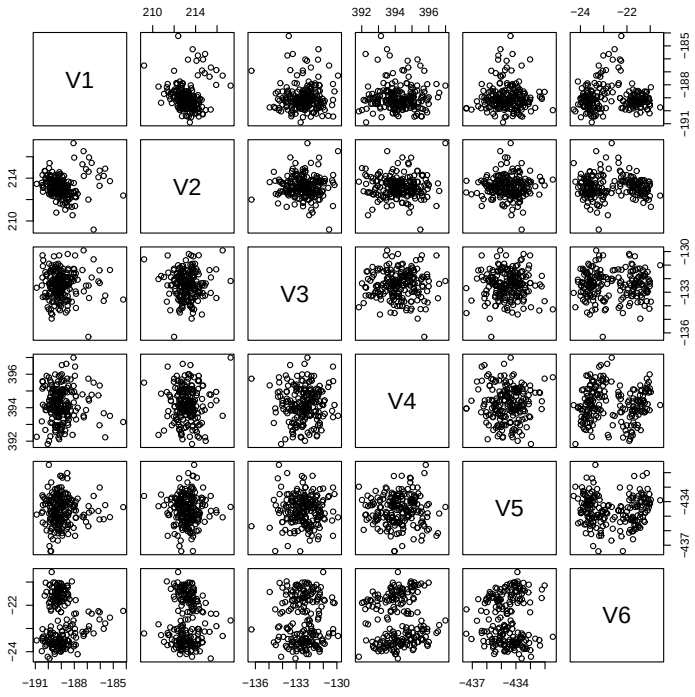

Figure 5: Swiss bank notes: ICA transformed data set ($\mathbf{X}(\mathbf{B}(\mathbf{X}))'$)

respectively, ordered according to kurtosis. The bimodality of the last estimated component with lowest kurtosis (caused by the clusters of genuine and forged notes) is clearly seen in Figure 5. Few outliers (forged bills) seem to cause the high kurtosis of the first component.

Acknowledgements

The authors wish to thank the referee for careful reading and valuable comments which helped to write the final version. The work was partly supported by grants from Academy of Finland. The first author wish thank David Tyler and Frank Critchley for many helpful discussions.

References

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

Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, *11*, 157-192.

Cardoso, J. F., and Souloumiac, A. (1993). Blind beamforming for non gaussian signals. *IEE Proceedings-F*, *140*(6), 362-370.

Cichocki, A., and Amari, S. (2002). *Adaptive blind signal and image processing: Learning algorithms and applications*. J. Wiley.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*(3), 287-314.

Croux, C., Ollila, E., and Oja, H. (2002). Sign and rank covariance matrices: Statistical properties and application to principal component analysis. In Y. Dodge (Ed.), *Statistical data analysis based on l_1-norm and related methods* (p. 257-269). Basel: Birkhäuser.

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

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

Eriksson, J., and Koivunen, V. (2004). Identifiability, separability and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, *11*, 601–604.

Flury, B., and Riedwyl, H. (1988). *Multivariate statistics. a practical approach*. London: Chapman and Hall.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. J. Wiley.

Kankainen, A., Taskinen, S., and Oja, H. (2005). Tests of multinormality based on location vectors and scatter matrices. *submitted*.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Kohen, K. L. (1999). Robust principal components for functional data. *Test*, *8*, 1-73.

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

Lopuhaä, H. P. (1991). Multivariate τ-estimators of location and scatter. *Canadian Journal of Statistics*, *19*, 310-321.

Marden, J. (1999). Some robust estimates of principal components. *Statistics and Probability Letters*, *43*, 349-359.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519-530.

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

Pearson, K. (1895). Contributions to the mathematical theory of evolution ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, *186*, 343-414.

Pham, D. T., and Garat, P. (1997). Blind separation of mixture of independent sources through quasi-maximum likelihood approach. *IEEE Transactions of Signal Processing*, *45*(7), 1712-1725.

Samarov, A., and Tsybakov, A. (2004). Nonparametric independent component analysis. *Bernoulli*, *10*, 565-582.

Tyler, D. E. (1987). A distribution-free m-estimator of multivariate scatter. *Annalc of Statistics*, *15*, 234-251.

Tyler, D. E. (2002). High breakdown point multivariate estimation. *Estadística*, *54*, 213-247.

Visuri, S., Koivunen, V., and Oja, H. (2000). Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, *91*, 557-575.

Authors' addresses:

Hannu Oja
Tampere School of Public Health
FIN-33014 University of Tampere
Finland

E-mail: Hannu.Oja@uta.fi

Seija Sirkiä
Department of Mathematics and Statistics
P.O Box 35 (MaD)
FIN-40014 University of Jyväskylä
Finland
E-mail: ssirkia@maths.jyu.fi

Jan Eriksson
Sgnal Processing Laboratory
P.O. Box 3000
FIN-02015 Helsinki University of Technology
Finland
E-mail: jan.erikssson@hut.fi