Author(s): Miettinen, Jari; Taskinen, Sara; Nordhausen, Klaus; Oja, Hannu

Title: Fourth Moments and Independent Component Analysis

Year: 2015

Version:

Please cite the original version:

Miettinen, J., Taskinen, S., Nordhausen, K., & Oja, H. (2015). Fourth Moments and Independent Component Analysis. Statistical science, 30(3), 372-390. https://doi.org/10.1214/15-STS520

Statistical Science
2015, Vol. 30, No. 3, 372–390
DOI: 10.1214/15-STS520

Fourth Moments and Independent Component Analysis

Jari Miettinen, Sara Taskinen, Klaus Nordhausen and Hannu Oja

Abstract. In independent component analysis it is assumed that the components of the observed random vector are linear combinations of latent independent random variables, and the aim is then to find an estimate for a transformation matrix back to these independent components. In the engineering literature, there are several traditional estimation procedures based on the use of fourth moments, such as FOBI (fourth order blind identification), JADE (joint approximate diagonalization of eigenmatrices), and FastICA, but the statistical properties of these estimates are not well known. In this paper various independent component functionals based on the fourth moments are discussed in detail, starting with the corresponding optimization problems, deriving the estimating equations and estimation algorithms, and finding asymptotic statistical properties of the estimates. Comparisons of the asymptotic variances of the estimates in wide independent component models show that in most cases JADE and the symmetric version of FastICA perform better than their competitors.

Key words and phrases: Affine equivariance, FastICA, FOBI, JADE, kurtosis.

1. INTRODUCTION

In his system of frequency curves, Pearson (1895) identified different types of distributions, and the classification was based on the use of the standardized third and fourth moments. A measure of degree of kurtosis for the distribution of x was defined as

$$\beta = \frac{E([x - E(x)]^4)}{[E([x - E(x)]^2)]^2} \quad \text{or} \quad \kappa = \beta - 3,$$

and Pearson (1905) called the distribution platykurtic, leptokurtic, or mesokurtic depending on the value of κ. In the case of the normal distribution ($\kappa = 0$, mesokurtic) Pearson also considered the probable error of $\hat{\kappa}$. Later, kurtosis was generally understood simply as a property which is measured by κ, which has raised questions such as “Is kurtosis really peakedness?”; see, for example, Darlington (1970). Van Zwet (1964) proposed kurtosis orderings for symmetrical distributions, and Oja (1981) defined measures of kurtosis as functionals which (i) are invariant under linear transformations and (ii) preserve the van Zwet partial ordering. Most of the measures of kurtosis, including β, can be written as a ratio of two scale measures. Recently, robust measures of kurtosis also have been proposed and considered in the literature; see, for example, Brys, Hubert and Struyf (2006).

It is well known that the variance of the sample mean depends on the population variance only, but the variance of the sample variance depends also on the shape of the distribution through β. The measure β has been used as an indicator of the bimodality, for example, in identifying clusters in the data set (Peña and Prieto, 2001) or as a general indicator for non-Gaussianity, for example, in testing for normality or in independent

Jari Miettinen is a Postdoctoral Researcher, Department of Mathematics and Statistics, 40014 University of Jyväskylä, Jyväskylä, Finland (e-mail: jari.p.miettinen@jyu.fi). Sara Taskinen is an Academy Research Fellow, Department of Mathematics and Statistics, 40014 University of Jyväskylä, Jyväskylä, Finland (e-mail: sara.l.taskinen@jyu.fi). Klaus Nordhausen is a Senior Research Fellow, Department of Mathematics and Statistics, 20014 University of Turku, Turku, Finland (e-mail: klaus.nordhausen@utu.fi). Hannu Oja is a Professor, Department of Mathematics and Statistics, 20014 University of Turku, Turku, Finland (e-mail: hannu.oja@utu.fi).

component analysis (Hyvärinen, 1999). Classical tests for the normality are based on the standardized third and fourth moments. See also DeCarlo (1997) for the meaning and use of kurtosis.

The concept and measures of kurtosis have been extended to the multivariate case as well. The classical skewness and kurtosis measures by Mardia (1970), for example, combine in a natural way the third and fourth moments of a standardized multivariate variable. Mardia's measures are invariant under affine transformations, that is, the p-variate random variables $\mathbf{x}$ and $\mathbf{A}\mathbf{x}+\mathbf{b}$ have the same skewness and kurtosis values for all full-rank $p \times p$ matrices $\mathbf{A}$ and for all p-vectors $\mathbf{b}$. For similar combinations of the standardized third and fourth moments, see also Móri, Rohatgi and Székely (1993). Let next $\mathbf{V}_1$ and $\mathbf{V}_2$ be two $p \times p$ affine equivariant scatter matrices (functionals); see Huber (1981) and Maronna (1976) for early contributions on scatter matrices. Then, in the invariant coordinate selection (ICS) in Tyler et al. (2009), one finds an affine transformation matrix $\mathbf{W}$ such that

$$\mathbf{W}\mathbf{V}_1\mathbf{W}' = \mathbf{I}_p \quad \text{and} \quad \mathbf{W}\mathbf{V}_2\mathbf{W}' = \mathbf{D},$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements in decreasing order. The transformed p variables are then presented in a new invariant coordinate system, and the diagonal elements in $\mathbf{D}$, that is, the eigenvalues of $\mathbf{V}_1^{-1}\mathbf{V}_2$, provide measures of multivariate kurtosis. This procedure is also sometimes called the generalized principal component analysis and has been used to find structures in the data. See Caussinus and Ruiz-Gazen (1993), Critchley, Pires and Amado (2006), Ilmonen, Nevalainen and Oja (2010), Peña, Prieto and Viladomat (2010), and Nordhausen, Oja and Ollila (2011). For the tests for multinormality based on these ideas, see Kankainen, Taskinen and Oja (2007). In independent component analysis, certain fourth moment matrices are used together with the covariance matrix in a similar way to find the transformations to independent components [FOBI by Cardoso (1989) and JADE by Cardoso and Souloumiac (1993)]. See also Oja, Sirkiä and Eriksson (2006).

In this paper, we consider the use of univariate and multivariate fourth moments in independent component analysis (ICA). The basic independent component (IC) model assumes that the observed components of $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ are linear combinations of latent independent components of $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})'$. Hence, the model can be written as

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}_i, \quad i = 1, \dots, n,$$

where the full rank $p \times p$ matrix $\boldsymbol{\Omega}$ is called the mixing matrix and $\mathbf{z}_1, \dots, \mathbf{z}_n$ is a random sample from a distribution with independent components such that $E(\mathbf{z}_i) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{z}_i) = \mathbf{I}_p$. Similarly to the model of elliptically symmetric distributions, the IC model is a semiparametric model, as the marginal distributions of the components of $\mathbf{z}$ are left fully unspecified except for the first two moments. For the identifiability of the parameters, one further assumes that at most one of the components has a normal distribution. Notice also that $\boldsymbol{\Omega}$ and $\mathbf{z}$ are still confounded in the sense that the order and signs of the components of $\mathbf{z}$ are not uniquely defined. The location center, the p-vector $\boldsymbol{\mu}$, is usually considered a nuisance parameter, since the main goal in independent component analysis is, based on a $p \times n$ data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, to find an estimate for an unmixing matrix $\mathbf{W}$ such that $\mathbf{W}\mathbf{x}$ has independent components. Note that all unmixing matrices $\mathbf{W}$ can be written as $\mathbf{C}\boldsymbol{\Omega}^{-1}$, where each row and each column of the $p \times p$ matrix $\mathbf{C}$ has exactly one nonzero element.

The population quantity to be estimated is first defined as an independent component functional $\mathbf{W}(F)$. The estimate $\mathbf{W}(F_n)$, also denoted by $\mathbf{W}(\mathbf{X})$, is then obtained by applying the functional to the empirical distribution F_n of $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$. In the engineering literature, several estimation procedures based on the fourth moments, such as FOBI (fourth order blind identification) (Cardoso, 1989), JADE (joint approximate diagonalization of eigenmatrices) (Cardoso and Souloumiac, 1993), and FastICA (Hyvärinen, 1999), have been proposed and widely used. In these approaches the marginal distributions are separated using various fourth moments. On the other hand, the estimators by Chen and Bickel (2006) and Samworth and Yuan (2012) only need the existence of the first moments and rely on efficient nonparametric estimates of the marginal densities. Efficient estimation methods based on residual signed ranks and residual ranks have been developed recently by Ilmonen and Paindaveine (2011) and Hallin and Mehta (2015). For a parametric model with a marginal Pearson system approach, see Karvanen and Koivunen (2002).

This paper describes in detail the independent component functionals based on fourth moments through corresponding optimization problems and estimating equations, provides fixed-point algorithms and the limiting statistical properties of the estimates, and specifies the needed assumptions. Also, a wide comparison study of the estimates is carried out. As far as we know, most of the results in the paper are new, including the asymptotical properties of the JADE estimate.

The asymptotical properties of the FOBI estimate have been derived earlier in Ilmonen, Nevalainen and Oja (2010). The limiting variances and the limiting multinormality of the deflation-based version of the FastICA estimate have been studied in Ollila (2010) and Nordhausen et al. (2011), respectively.

2. NOTATION AND PRELIMINARY RESULTS

Throughout the paper, we use the following notation. First write, for independent z_{ik}, $k = 1, \dots, p$,

$$E(z_{ik}) = 0, \quad E(z_{ik}^2) = 1, \quad E(z_{ik}^3) = \gamma_k \quad \text{and} \quad E(z_{ik}^4) = \beta_k,$$

and

$$\kappa_k = \beta_k - 3, \quad \pi_k = \operatorname{sign}(\kappa_k) \quad \text{and} \quad \operatorname{Var}(z_{ik}^3) = \sigma_k^2.$$

As seen later, the limiting distributions of the unmixing matrix estimates based on fourth moments depend on the joint limiting distribution of

$$\sqrt{n}\hat{s}_{kl} = n^{-1/2} \sum_{i=1}^n z_{ik} z_{il}, \tag{1}$$
$$\sqrt{n}\hat{r}_{kl} = n^{-1/2} \sum_{i=1}^n (z_{ik}^3 - \gamma_k) z_{il}$$

and

$$\sqrt{n}\hat{r}_{mkl} = n^{-1/2} \sum_{i=1}^n z_{im}^2 z_{ik} z_{il},$$

for distinct $k, l, m = 1, \dots, p$. If the eighth moments of z_i exist, then the joint limiting distribution of $\sqrt{n}\hat{s}_{kl}$, $\sqrt{n}\hat{r}_{kl}$, and $\sqrt{n}\hat{r}_{mkl}$ is a multivariate normal distribution with marginal zero means. The nonzero variances and covariances are

$$\operatorname{Var}(\sqrt{n}\hat{s}_{kl}) = 1, \quad \operatorname{Var}(\sqrt{n}\hat{r}_{kl}) = \sigma_k^2,$$
$$\operatorname{Var}(\sqrt{n}\hat{r}_{mkl}) = \beta_m,$$

and

$$\operatorname{Cov}(\sqrt{n}\hat{s}_{kl}, \sqrt{n}\hat{r}_{kl}) = \beta_k,$$
$$\operatorname{Cov}(\sqrt{n}\hat{r}_{kl}, \sqrt{n}\hat{r}_{lk}) = \beta_k \beta_l,$$

and

$$\operatorname{Cov}(\sqrt{n}\hat{s}_{kl}, \sqrt{n}\hat{r}_{mkl}) = 1,$$
$$\operatorname{Cov}(\sqrt{n}\hat{r}_{kl}, \sqrt{n}\hat{r}_{mkl}) = \beta_k \quad \text{and}$$
$$\operatorname{Cov}(\sqrt{n}\hat{r}_{lk}, \sqrt{n}\hat{r}_{mkl}) = \beta_l.$$

We also often refer to the following sets of $p \times p$ transformation matrices:

1. $\mathcal{D} = \{\operatorname{diag}(d_1, \dots, d_p) : d_1, \dots, d_p > 0\}$ (heterogeneous rescaling),
2. $\mathcal{J} = \{\operatorname{diag}(j_1, \dots, j_p) : j_1, \dots, j_p = \pm 1\}$ (heterogeneous sign changes),
3. $\mathcal{P} = \{\mathbf{P} : \mathbf{P}$ is a permutation matrix$\}$,
4. $\mathcal{U} = \{\mathbf{U} : \mathbf{U}$ is an orthogonal matrix$\}$,
5. $\mathcal{C} = \{\mathbf{C} : \mathbf{C} = \mathbf{PJD}, \mathbf{P} \in \mathcal{P}, \mathbf{J} \in \mathcal{J}, \mathbf{D} \in \mathcal{D}\}$.

Next, let $\mathbf{e}_i$ denote a p-vector with ith element one and other elements zero, and define $\mathbf{E}^{ij} = \mathbf{e}_i \mathbf{e}_j'$, $i, j = 1, \dots, p$, and

$$\mathbf{J}_{p,p} = \sum_{i=1}^p \sum_{j=1}^p \mathbf{E}^{ij} \otimes \mathbf{E}^{ij} = \operatorname{vec}(\mathbf{I}_p) \operatorname{vec}(\mathbf{I}_p)',$$

$$\mathbf{K}_{p,p} = \sum_{i=1}^p \sum_{j=1}^p \mathbf{E}^{ij} \otimes \mathbf{E}^{ji},$$

$$\mathbf{I}_{p,p} = \sum_{i=1}^p \sum_{j=1}^p \mathbf{E}^{ii} \otimes \mathbf{E}^{jj} = \mathbf{I}_{p^2} \quad \text{and}$$

$$\mathbf{D}_{p,p} = \sum_{i=1}^p \mathbf{E}^{ii} \otimes \mathbf{E}^{ii}.$$

Then, for any $p \times p$ matrix $\mathbf{A}$, $\mathbf{J}_{p,p} \operatorname{vec}(\mathbf{A}) = \operatorname{tr}(\mathbf{A}) \cdot \operatorname{vec}(\mathbf{I}_p)$, $\mathbf{K}_{p,p} \operatorname{vec}(\mathbf{A}) = \operatorname{vec}(\mathbf{A}')$, and $\mathbf{D}_{p,p} \operatorname{vec}(\mathbf{A}) = \operatorname{vec}(\operatorname{diag}(\mathbf{A}))$. The matrix $\mathbf{K}_{p,p}$ is sometimes called a commutation matrix. For a symmetric nonnegative definite matrix $\mathbf{S}$, the matrix $\mathbf{S}^{-1/2}$ is taken to be symmetric and nonnegative definite and to satisfy $\mathbf{S}^{-1/2}\mathbf{S}\mathbf{S}^{-1/2} = \mathbf{I}_p$.

3. INDEPENDENT COMPONENT MODEL AND FUNCTIONALS

3.1 Independent Component (IC) Model

Throughout the paper, our p-variate observations $\mathbf{x}_1, \dots, \mathbf{x}_n$ follow the independent component (IC) model

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Omega} \mathbf{z}_i, \quad i = 1, \dots, n, \tag{2}$$

where $\boldsymbol{\mu}$ is a mean vector, $\boldsymbol{\Omega}$ is a full-rank $p \times p$ mixing matrix, and $\mathbf{z}_1, \dots, \mathbf{z}_n$ are independent and identically distributed random vectors from a p-variate distribution such that:

ASSUMPTION 1. The components $z_{i1}, \dots, z_{ip}$ of $\mathbf{z}_i$ are independent.

ASSUMPTION 2. Second moments exist, $E(\mathbf{z}_i) = \mathbf{0}$ and $E(\mathbf{z}_i \mathbf{z}_i') = \mathbf{I}_p$.

ASSUMPTION 3. At most one of the components $z_{i1}, \dots, z_{ip}$ of $\mathbf{z}_i$ has a normal distribution.

If the model is defined using Assumption 1 only, then the mixing matrix $\mathbf{\Omega}$ is not well-defined and can at best be identified only up to the order, the signs, and heterogenous multiplications of its columns. Assumption 2 states that the second moments exist, and $E(\mathbf{z}_i) = \mathbf{0}$ and $E(\mathbf{z}_i\mathbf{z}_i') = \mathbf{I}_p$ serve as identification constraints for $\boldsymbol{\mu}$ and the scales of the columns of $\mathbf{\Omega}$. Assumption 3 is needed, as, for example, if $\mathbf{z} \sim N_2(\mathbf{0}, \mathbf{I}_2)$, then also $\mathbf{Uz} \sim N_2(\mathbf{0}, \mathbf{I}_2)$ for all orthogonal $\mathbf{U}$ and the independent components are not well-defined. Still, after these three assumptions, the order and signs of the columns of $\mathbf{\Omega}$ remain unidentified, but one can identify the set of the standardized independent components $\{\pm z_{i1}, \dots, \pm z_{ip}\}$, which is naturally sufficient for practical data analysis.

One of the key results in independent component analysis is the following.

THEOREM 1. *Let $\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Omega z}$ be an observation from an IC model with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma} = \mathbf{\Omega\Omega}'$, and write $\mathbf{x}_{st} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ for the standardized random variable. Then $\mathbf{z} = \mathbf{U}\mathbf{x}_{st}$ for some orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)'$.*

The result says that, starting with standardized observations $\mathbf{x}_{st}$, one only has to search for an unknown $\mathbf{U} \in \mathcal{U}$ such that $\mathbf{U}\mathbf{x}_{st}$ has independent components. Thus, after estimating $\mathbf{\Sigma}$, the estimation problem can be reduced to the estimation problem of an orthogonal matrix $\mathbf{U}$ only.

3.2 Independent Component (IC) Functionals

Write next $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ for a random sample from the IC model (2) with the cumulative distribution function (c.d.f.) $F_\mathbf{x}$. As mentioned in the Introduction, the aim of independent component analysis (ICA) is to find an estimate of some unmixing matrix $\mathbf{W}$ such that $\mathbf{W}\mathbf{x}_i$ has independent components. It is easy to see that all unmixing matrices can be written as $\mathbf{W} = \mathbf{C\Omega}^{-1}$ for some $\mathbf{C} \in \mathcal{C}$. The population quantity, which we wish to estimate, is defined as the value of an independent component functional $\mathbf{W}(F)$ at the distribution of $F_\mathbf{x}$.

DEFINITION 1. The $p \times p$ matrix-valued functional $\mathbf{W}(F)$ is said to be an independent component (IC) functional if (i) $\mathbf{W}(F_\mathbf{x})\mathbf{x}$ has independent components in the IC model (2) and (ii) $\mathbf{W}(F_\mathbf{x})$ is affine equivariant in the sense that

$$\begin{aligned}&\{(\mathbf{W}(F_{\mathbf{Ax+b}})\mathbf{Ax})_1, \dots, (\mathbf{W}(F_{\mathbf{Ax+b}})\mathbf{Ax})_p\} \\ &\quad = \{\pm(\mathbf{W}(F_\mathbf{x})\mathbf{x})_1, \dots, \pm(\mathbf{W}(F_\mathbf{x})\mathbf{x})_p\}\end{aligned}$$

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

Notice that in the independent component model, $\mathbf{W}(F_\mathbf{x})\mathbf{x}$ does not depend on the specific choices of $\mathbf{z}$ and $\mathbf{\Omega}$, up to the signs and the order of the components. Notice also that, in the condition (ii), any c.d.f. F is allowed to be used as an argument of $\mathbf{W}(F)$. The corresponding sample version $\mathbf{W}(F_n)$ is then obtained when the IC functional is applied to the empirical distribution function F_n of $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$. We also sometimes write $\mathbf{W}(\mathbf{X})$ for the sample version. Naturally, the estimator is then also affine equivariant in the sense that, for all nonsingular $p \times p$ matrices $\mathbf{A}$ and for all p-vectors $\mathbf{b}$, $\mathbf{W}(\mathbf{AX} + \mathbf{b1}_n')\mathbf{AX} = \mathbf{PJW}(\mathbf{X})\mathbf{X}$ for some $\mathbf{J} \in \mathcal{J}$ and $\mathbf{P} \in \mathcal{P}$.

REMARK 1. As mentioned before, if $\mathbf{W}$ is an unmixing matrix, then so is $\mathbf{CW}$ for all $\mathbf{C} \in \mathcal{C}$, and we then have a whole set of matrices $\{\mathbf{CW} : \mathbf{C} \in \mathcal{C}\}$ equivalent to $\mathbf{W}$. To find a unique representative in the class, it is often required that $\mathrm{Cov}(\mathbf{CWx}) = \mathbf{I}_p$ but still the order and signs of the rows remain unidentified. Of course, the assumption on the existence of second moments may sometimes be thought to be too restrictive. For alternative ways to identify the unmixing matrix, see then Chen and Bickel (2006), Ilmonen and Paindaveine (2011), and Hallin and Mehta (2015), for example. For a general discussion on this identification problem, see also Eriksson and Koivunen (2004).

4. UNIVARIATE KURTOSIS AND INDEPENDENT COMPONENT ANALYSIS

4.1 Classical Measures of Univariate Skewness and Kurtosis

Let first x be a univariate random variable with mean value μ and variance σ^2. The standardized variable is then $z = (x - \mu)/\sigma$, and classical skewness and kurtosis measures are the standardized third and fourth moments, $\gamma = E(z^3)$ and $\beta = E(z^4)$. For symmetrical distributions, $\gamma = 0$, and for the normal distribution, $\kappa = \beta - 3 = 0$. For a random sample $x_1, \dots, x_n$ from a univariate distribution, write

$$\mu_j = E((x_i - \mu)^j) \quad \text{and}$$

$$m_j = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^j, \quad j = 2, 3, 4.$$

Then the limiting distribution of $\sqrt{n}(m_2 - \mu_2, m_3 - \mu_3, m_4 - \mu_4)'$ is a 3-variate normal distribution with mean vector zero and covariance matrix with the (i, j) element

$$\mu_{i+j+2} - \mu_{i+1}\mu_{j+1} - (i+1)\mu_i\mu_{j+2} - (j+1)\mu_{i+2}\mu_j + (i+1)(j+1)\mu_i\mu_j\mu_2,$$

$i, j = 1, 2, 3$. See Theorem 2.2.3.B in Serfling (1980). Then in the symmetric case with $\mu_2 = 1$, for example,

$$\sqrt{n}\begin{pmatrix} m_2 - 1 \\ m_3 \\ m_4 - \mu_4 \end{pmatrix} \to_d N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mu_4 - 1 & 0 & \mu_6 - \mu_4 \\ 0 & \mu_6 - 6\mu_4 + 9 & 0 \\ \mu_6 - \mu_4 & 0 & \mu_8 - \mu_4^2 \end{pmatrix}\right).$$

If the observations come from $N(0, 1)$, we further obtain

$$\sqrt{n}\begin{pmatrix} m_2 - 1 \\ m_3 \\ m_4 - 3 \end{pmatrix} \to_d N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 12 \\ 0 & 6 & 0 \\ 12 & 0 & 96 \end{pmatrix}\right).$$

The classical skewness and kurtosis statistics, the natural estimates of γ and β, are $\hat{\gamma} = m_3/m_2^{3/2}$ and $\hat{\beta} = m_4/m_2^2$, and then

$$\sqrt{n}\hat{\gamma} = \sqrt{n}m_3 + o_P(1) \quad \text{and}$$
$$\sqrt{n}\hat{\kappa} = \sqrt{n}(\hat{\beta} - 3) = \sqrt{n}(m_4 - 3) - 6\sqrt{n}(m_2 - 1) + o_P(1)$$

and we obtain, in the general $N(\mu, \sigma^2)$ case, that

$$\sqrt{n}\begin{pmatrix} \hat{\gamma} \\ \hat{\kappa} \end{pmatrix} = \sqrt{n}\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} - 3 \end{pmatrix} \to_d N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix}\right).$$

Consider next p-variate observations coming from an IC model. The important role of the fourth moments is stated in the following:

THEOREM 2. *Let the components of* $\mathbf{z} = (z_1, \ldots, z_p)'$ *be independent and standardized so that* $E(\mathbf{z}) = \mathbf{0}$ *and* $\mathrm{Cov}(\mathbf{z}) = \mathbf{I}_p$, *and assume that at most one of the kurtosis values* $\kappa_i = E(z_i^4) - 3$, $i = 1, \ldots, p$, *is zero. Then the following inequalities hold true*:

(i)

$$|E((\mathbf{u}'\mathbf{z})^4) - 3| \le \max\{|E(z_1^4) - 3|, \ldots, |E(z_p^4) - 3|\}$$

for all $\mathbf{u}$ *such that* $\mathbf{u}'\mathbf{u} = 1$. *The equality holds only if* $\mathbf{u} = \mathbf{e}_i$ *for* i *such that* $|E(z_i^4) - 3| = \max\{|E(z_1^4) - 3|, \ldots, |E(z_p^4) - 3|\}$, *and*

(ii)

$$|E[(\mathbf{u}_1'\mathbf{z})^4] - 3| + \cdots + |E[(\mathbf{u}_p'\mathbf{z})^4] - 3| \le |E[z_1^4] - 3| + \cdots + |E[z_p^4] - 3|$$

for all orthogonal matrices $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)'$. *The equality holds only if* $\mathbf{U} = \mathbf{JP}$ *for some* $\mathbf{J} \in \mathcal{J}$ *and* $\mathbf{P} \in \mathcal{P}$.

For the first part of the theorem, see Lemma 2 in Bugrien and Kent (2005). The theorem suggests natural strategies and algorithms in search for independent components. It was seen in Theorem 1 that in the IC model $\mathbf{x}_{st} = \mathbf{Uz}$ with an orthogonal $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$. The first part of Theorem 2 then shows how the components can be found one by one just by repeatedly maximizing

$$|E((\mathbf{u}_k'\mathbf{x}_{st})^4) - 3|, \quad k = 1, \ldots, p$$

(projection pursuit approach), and the second part of Theorem 2 implies that the same components may be found simultaneously by maximizing

$$|E[(\mathbf{u}_1'\mathbf{x}_{st})^4] - 3| + \cdots + |E[(\mathbf{u}_p'\mathbf{x}_{st})^4] - 3|.$$

In the engineering literature, these two approaches are well known and important special cases of the so-called deflation-based FastICA and symmetric FastICA; see, for example, Hyvärinen, Karhunen and Oja (2001). The statistical properties of these two estimation procedures will now be considered in detail.

4.2 Projection Pursuit Approach—Deflation-Based FastICA

Assume that $\mathbf{x}$ is an observation from an IC model (2) and let again $\mathbf{x}_{st} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ be the standardized random variable. Theorem 2(i) then suggests the following projection pursuit approach in searching for the independent components.

DEFINITION 2. The deflation-based projection pursuit (or deflation-based FastICA) functional is $\mathbf{W}(F_{\mathbf{x}}) = \mathbf{U}\mathbf{\Sigma}^{-1/2}$, where $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{x})$ and the rows of an orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)'$ are found one by one by maximizing

$$|E((\mathbf{u}_k'\mathbf{x}_{st})^4) - 3|$$

under the constraint that $\mathbf{u}_k'\mathbf{u}_k = 1$ and $\mathbf{u}_j'\mathbf{u}_k = 0$, $j = 1, \ldots, k-1$.

It is straightforward to see that $\mathbf{W}(F_{\mathbf{x}})$ is affine equivariant. In the independent component model (2), $\mathbf{W}(F_{\mathbf{x}})\mathbf{x}$ has independent components if Assumption 3 is replaced by the following stronger assumption.

ASSUMPTION 4. The fourth moments of $\mathbf{z}$ exist, and at most one of the kurtosis values κ_k, $k = 1, \dots, p$, is zero.

Thus, under this assumption, $\mathbf{W}(F)$ is an independent component (IC) functional. Based on Theorem 2(i), the functional then finds the independent components in such an order that

$$|E((\mathbf{u}_1'\mathbf{x}_{st})^4) - 3| \geq \cdots \geq |E((\mathbf{u}_p'\mathbf{x}_{st})^4) - 3|.$$

The solution order is unique if the kurtosis values are distinct.

The Lagrange multiplier technique can be used to obtain the estimating equations for $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)'$. This is done in Ollila (2010) and Nordhausen et al. (2011) and the procedure is the following. After finding $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$, the solution $\mathbf{u}_k$ thus optimizes the Lagrangian function

$$L(\mathbf{u}_k, \boldsymbol{\theta}_k) = |E((\mathbf{u}_k'\mathbf{x}_{st})^4) - 3| - \sum_{j=1}^{k} \theta_{kj}(\mathbf{u}_j'\mathbf{u}_k - \delta_{jk}),$$

where $\boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{kk})'$ is the vector of Lagrangian multipliers and $\delta_{jk} = 1$ (0) as $j = k$ ($j \neq k$) is the Kronecker delta. Write

$$\mathbf{T}(\mathbf{u}) = E[(\mathbf{u}'\mathbf{x}_{st})^3\mathbf{x}_{st}].$$

The solution for $\mathbf{u}_k$ is then given by the $p + k$ equations

$$4\pi_k\mathbf{T}(\mathbf{u}_k) - \sum_{j=1}^{k-1} \theta_{kj}\mathbf{u}_j - 2\theta_{kk}\mathbf{u}_k = \mathbf{0} \quad \text{and}$$

$$\mathbf{u}_j'\mathbf{u}_k = \delta_{jk}, \quad j = 1, \dots, k,$$

where $\pi_k = \operatorname{sign}(\kappa_k)$. One then first finds the solutions for the Lagrange coefficients in θ_k, and substituting these results into the first p equations, the following result is obtained.

THEOREM 3. *Write $\mathbf{x}_{st} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ for the standardized random vector, and $\mathbf{T}(\mathbf{u}) = E[(\mathbf{u}'\mathbf{x}_{st})^3 \cdot \mathbf{x}_{st}]$. The orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)'$ solves the estimating equations*

$$(\mathbf{u}_k'\mathbf{T}(\mathbf{u}_k))\mathbf{u}_k = \left(\mathbf{I}_p - \sum_{j=1}^{k-1} \mathbf{u}_j\mathbf{u}_j'\right)\mathbf{T}(\mathbf{u}_k), \quad k = 1, \dots, p.$$

The theorem suggests the following fixed-point algorithm for the deflation-based solution. After finding $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$, the following two steps are repeated until convergence to get $\mathbf{u}_k$:

$$\text{Step 1:} \quad \mathbf{u}_k \leftarrow \left(I_p - \sum_{j=1}^{k-1} \mathbf{u}_j\mathbf{u}_j'\right)\mathbf{T}(\mathbf{u}_k),$$

$$\text{Step 2:} \quad \mathbf{u}_k \leftarrow \|\mathbf{u}_k\|^{-1}\mathbf{u}_k.$$

The deflation-based estimate $\mathbf{W}(\mathbf{X})$ is obtained as above but by replacing the population quantities by the corresponding empirical ones. Without loss of generality, assume next that $|\kappa_1| \geq \cdots \geq |\kappa_p|$. First note that, due to the affine equivariance of the estimate, $\mathbf{W}(\mathbf{X}) = \mathbf{W}(\mathbf{Z})\boldsymbol{\Omega}^{-1}$. In the efficiency studies, it is therefore sufficient to consider $\hat{\mathbf{W}} = \mathbf{W}(\mathbf{Z})$ and the limiting distribution of $\sqrt{n}(\hat{\mathbf{W}} - \mathbf{I}_p)$ for a sequence $\hat{\mathbf{W}}$ converging in probability to $\mathbf{I}_p$. As the empirical and population criterion functions

$$D_n(\mathbf{u}) = \left| n^{-1} \sum_{i=1}^{n} (\mathbf{u}'\mathbf{x}_{st,i})^4 - 3 \right| \quad \text{and}$$

$$D(\mathbf{u}) = |E[(\mathbf{u}'\mathbf{z})^4] - 3|$$

are continuous and $\sup_{\mathbf{u}'\mathbf{u}=1} |D_n(\mathbf{u}) - D(\mathbf{u})| \to_P 0$, one can choose a sequence of solutions such that $\hat{\mathbf{u}}_1 \to_P \mathbf{e}_1$ and similarly for $\hat{\mathbf{u}}_2, \dots, \hat{\mathbf{u}}_{p-1}$. Further, then also $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{S}}^{-1/2} \to_P \mathbf{I}_p$. One can next show that the limiting distribution of $\sqrt{n}(\hat{\mathbf{W}} - \mathbf{I}_p)$ is obtained if we only know the joint limiting distribution of $\sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p)$ and $\sqrt{n}\operatorname{off}(\hat{\mathbf{R}})$, where $\hat{\mathbf{S}} = (\hat{s}_{kl})$ is the sample covariance matrix, $\hat{\mathbf{R}} = (\hat{r}_{kl})$ is given in (1), and $\operatorname{off}(\hat{\mathbf{R}}) = \hat{\mathbf{R}} - \operatorname{diag}(\hat{\mathbf{R}})$. We then have the following results; see also Ollila (2010), Nordhausen et al. (2011).

THEOREM 4. *Let $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ be a random sample from a distribution with finite eighth moments and satisfying the Assumptions 1, 2, and 4 with $|\kappa_1| \geq \cdots \geq |\kappa_p|$. Then there exists a sequence of solutions such that $\hat{\mathbf{W}} \to_P \mathbf{I}_p$ and*

$$\sqrt{n}\hat{w}_{kl} = -\sqrt{n}\hat{w}_{lk} - \sqrt{n}\hat{s}_{kl} + o_P(1), \quad l < k,$$

$$\sqrt{n}(\hat{w}_{kk} - 1) = -1/2\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1) \quad \textit{and}$$

$$\sqrt{n}\hat{w}_{kl} = \frac{\sqrt{n}\hat{r}_{kl} - (\kappa_k + 3)\sqrt{n}\hat{s}_{kl}}{\kappa_k} + o_P(1), \quad l > k.$$

COROLLARY 1. *Under the assumptions of Theorem 4, the limiting distribution of $\sqrt{n}\operatorname{vec}(\hat{\mathbf{W}} - \mathbf{I}_p)$ is* *a multivariate normal with zero mean vector and componentwise variances*

$$\mathrm{ASV}(\hat{w}_{kl}) = \frac{\sigma_l^2 - (\kappa_l + 3)^2}{\kappa_l^2} + 1, \quad \kappa_l \neq 0, l < k,$$
$$\mathrm{ASV}(\hat{w}_{kk}) = (\kappa_k + 2)/4 \quad \textit{and}$$
$$\mathrm{ASV}(\hat{w}_{kl}) = \frac{\sigma_k^2 - (\kappa_k + 3)^2}{\kappa_k^2}, \quad \kappa_k \neq 0, l > k.$$

REMARK 2. Projection pursuit is used to reveal structures in the original data by selecting interesting low-dimensional orthogonal projections of interest. This is done, as above, by maximizing the value of an objective function (projection index). The term "projection pursuit" was first launched by Friedman and Tukey (1974). Huber (1985) considered projection indices with heuristic arguments that a projection is the more interesting, the less normal it is. All his indices were ratios of two scale functionals, that is, kurtosis functionals, with the classical kurtosis measure as a special case. He also discussed the idea of a recursive approach to find subspaces. Peña and Prieto (2001) used the projection pursuit algorithm with the classical kurtosis index for finding directions for cluster identification. For more discussion on the projection pursuit approach, see also Jones and Sibson (1987).

REMARK 3. In the engineering literature, Hyvärinen and Oja (1997) were the first to propose the procedure based on the fourth moments, and later considered an extension with a choice among several alternative projection indices (measures of non-Gaussianity). The approach is called deflation-based or one-unit FastICA and it is perhaps the most popular approach for the ICA problem in engineering applications. Note that the estimating equations in Theorem 3 and the resulting fixed-point algorithm do not fix the order of the components (the order is fixed by the original definition) and, as seen in Theorem 4, the limiting distribution of the estimate depends on the order in which the components are found. Using this property, Nordhausen et al. (2011) proposed a two-stage version of the deflation-based FastICA method with a chosen projection index that finds the components in an optimal efficiency order. Moreover, Miettinen et al. (2014a) introduced an adaptive two-stage algorithm that (i) allows one to use different projection indices for different components and (ii) optimizes the order in which the components are extracted.

4.3 Symmetric Approach—Symmetric FastICA

In the symmetric approach, the rows of the matrix $\mathbf{U}$ are found simultaneously, and we have the following:

DEFINITION 3. The symmetric projection pursuit (or symmetric fastICA) functional is $\mathbf{W}(F_{\mathbf{x}}) = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x})$ and $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)'$ maximizes

$$\left|E\left((\mathbf{u}_1'\mathbf{x}_{st})^4\right) - 3\right| + \dots + \left|E\left((\mathbf{u}_p'\mathbf{x}_{st})^4\right) - 3\right|$$

under the constraint that $\mathbf{U}\mathbf{U}' = \mathbf{I}_p$.

This optimization procedure is called symmetric FastICA in the signal processing community. The functional $\mathbf{W}(F_{\mathbf{x}})$ is again affine equivariant. Based on Theorem 2(ii), in the IC model with Assumption 4 the maximizer is unique up to the order and signs of the rows of $\mathbf{U}$, that is,

$$\{z_1, \dots, z_p\} = \{\pm\mathbf{u}_1'\mathbf{x}_{st}, \dots, \pm\mathbf{u}_p'\mathbf{x}_{st}\}.$$

As in the deflation-based case, we use the Lagrange multiplier technique to obtain the matrix $\mathbf{U}$. The Lagrangian function to be optimized is now

$$L(\mathbf{U}, \boldsymbol{\Theta}) = \sum_{k=1}^{p}\left|E\left((\mathbf{u}_k'\mathbf{x}_{st})^4\right) - 3\right| - \sum_{k=1}^{p}\theta_{kk}(\mathbf{u}_k'\mathbf{u}_k - 1)$$
$$- \sum_{j=1}^{p-1}\sum_{k=j+1}^{p}\theta_{jk}\mathbf{u}_j'\mathbf{u}_k,$$

where the symmetric matrix $\boldsymbol{\Theta} = (\theta_{jk})$ contains all $p(p+1)/2$ Lagrangian multipliers. Write again $\mathbf{T}(\mathbf{u}) = E((\mathbf{u}'\mathbf{x}_{st})^3\mathbf{x}_{st})$. Then the solution $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)'$ satisfies

$$4\pi_k T(\mathbf{u}_k) = 2\theta_{kk}\mathbf{u}_k + \sum_{j<k}\theta_{jk}\mathbf{u}_j + \sum_{j>k}\theta_{kj}\mathbf{u}_j, \quad k = 1, \dots, p,$$

and

$$\mathbf{U}\mathbf{U}' = \mathbf{I}_p.$$

Solving θ_{jk} and using the fact that $\theta_{jk} = \theta_{kj}$ give $\pi_k\mathbf{u}_j'\mathbf{T}(\mathbf{u}_k) = \pi_j\mathbf{u}_k'\mathbf{T}(\mathbf{u}_j)$, $j, k = 1, \dots, p$, and we get the following estimating equations.

THEOREM 5. *Let* $\mathbf{x}_{st} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ *be the standardized random vector from the IC model* (2), $\mathbf{T}(\mathbf{u}) = E((\mathbf{u}'\mathbf{x}_{st})^3\mathbf{x}_{st})$, $\mathbf{T}(\mathbf{U}) = (\mathbf{T}(\mathbf{u}_1), \dots, \mathbf{T}(\mathbf{u}_p))'$ *and* $\boldsymbol{\Pi} = \mathrm{diag}(\pi_1, \dots, \pi_p)$. *The estimating equations for the symmetric solution* $\mathbf{U}$ *are*

$$\mathbf{U}\mathbf{T}(\mathbf{U})'\boldsymbol{\Pi} = \boldsymbol{\Pi}\mathbf{T}(\mathbf{U})\mathbf{U}' \quad \textit{and} \quad \mathbf{U}\mathbf{U}' = \mathbf{I}_p.$$

For the computation of $\mathbf{U}$, the above estimating equations suggest a fixed-point algorithm with the updating step

$$\mathbf{U} \leftarrow \mathbf{\Pi T}(\mathbf{T}'\mathbf{T})^{-1/2}.$$

The symmetric version estimate $\mathbf{W}(\mathbf{X})$ is obtained by replacing the population quantities by their corresponding empirical ones in the estimating equations. Write again $\hat{\mathbf{W}} = \mathbf{W}(\mathbf{Z})$ and let $\hat{\mathbf{S}} = (\hat{s}_{kl})$ and $\hat{\mathbf{R}} = (\hat{r}_{kl})$ be as in (1). Then we have the following:

THEOREM 6. *Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ be a random sample from a distribution of $\mathbf{z}$ satisfying the Assumptions 1, 2, and 4 with bounded eighth moments. Then there is a sequence of solutions such that $\hat{\mathbf{W}} \to_P \mathbf{I}_p$ and*

$$\begin{aligned}
&\sqrt{n}(\hat{w}_{kk} - 1) \\
&\quad = -\frac{1}{2}\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1) \quad \textit{and} \\
&\sqrt{n}\hat{w}_{kl} \\
&\quad = \frac{\sqrt{n}\hat{r}_{kl}\pi_k - \sqrt{n}\hat{r}_{lk}\pi_l - (\kappa_k\pi_k + 3\pi_k - 3\pi_l)\sqrt{n}\hat{s}_{kl}}{|\kappa_k| + |\kappa_l|} \\
&\qquad + o_P(1), \quad k \neq l,
\end{aligned}$$

where $\pi_k = \operatorname{sign}(\kappa_k)$.

COROLLARY 2. *Under the assumptions of Theorem 6, the limiting distribution of $\sqrt{n}\operatorname{vec}(\hat{\mathbf{W}} - \mathbf{I}_p)$ is a multivariate normal with zero mean vector and componentwise variances*

$$\begin{aligned}
&\mathrm{ASV}(\hat{w}_{kk}) = (\kappa_k + 2)/4 \quad \textit{and} \\
&\mathrm{ASV}(\hat{w}_{kl}) = \frac{\sigma_k^2 + \sigma_l^2 - \kappa_k^2 - 6(\kappa_k + \kappa_l) - 18}{(|\kappa_k| + |\kappa_l|)^2}, \\
&\qquad\qquad k \neq l.
\end{aligned}$$

REMARK 4. The symmetric FastICA approach with other choices of projection indices was proposed in the engineering literature by Hyvärinen (1999). The computation of symmetric FastICA estimate was done, as in our approach, by running p parallel one-unit algorithms, which were followed by a matrix orthogonalization step. A generalized symmetric FastICA algorithm that uses different projection indices for different components was proposed by Koldovský, Tichavský and Oja (2006). The asymptotical variances of generalized symmetric FastICA estimates were derived in Tichavsky, Koldovsky and Oja (2006) under the assumption of symmetric independent component distributions.

5. MULTIVARIATE KURTOSIS AND INDEPENDENT COMPONENT ANALYSIS

5.1 Measures of Multivariate Skewness and Kurtosis

Let $\mathbf{x}$ be a p-variate random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{x}_{st} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. All the standardized third and fourth moments can now be collected into $p \times p^2$ and $p^2 \times p^2$ matrices

$$\begin{aligned}
\boldsymbol{\gamma} &= E(\mathbf{x}_{st}' \otimes (\mathbf{x}_{st}\mathbf{x}_{st}')) \quad \text{and} \\
\boldsymbol{\beta} &= E((\mathbf{x}_{st}\mathbf{x}_{st}') \otimes (\mathbf{x}_{st}\mathbf{x}_{st}')).
\end{aligned}$$

Unfortunately, these multivariate measures of skewness and kurtosis are not invariant under affine transformations: The transformation $\mathbf{x} \to \mathbf{Ax} + \mathbf{b}$ induces, for some unspecified orthogonal matrix $\mathbf{U}$, the transformations

$$\begin{aligned}
&\mathbf{x}_{st} \to \mathbf{U}\mathbf{x}_{st}, \quad \boldsymbol{\gamma} \to \mathbf{U}\boldsymbol{\gamma}(\mathbf{U}' \otimes \mathbf{U}') \quad \text{and} \\
&\boldsymbol{\beta} \to (\mathbf{U} \otimes \mathbf{U})\boldsymbol{\beta}(\mathbf{U}' \otimes \mathbf{U}').
\end{aligned}$$

Notice next that, for any $p \times p$ matrix $\mathbf{A}$,

$$\begin{aligned}
\mathbf{G}(\mathbf{A}) &= E(\mathbf{x}_{st}\mathbf{x}_{st}'\mathbf{A}\mathbf{x}_{st}) \quad \text{and} \\
\mathbf{B}(\mathbf{A}) &= E(\mathbf{x}_{st}\mathbf{x}_{st}'\mathbf{A}\mathbf{x}_{st}\mathbf{x}_{st}')
\end{aligned} \tag{3}$$

provide selected p and p^2 linear combinations of the third and fourth moments as $\operatorname{vec}(\mathbf{G}(\mathbf{A})) = \gamma \operatorname{vec}(\mathbf{A})$ and $\operatorname{vec}(\mathbf{B}(\mathbf{A})) = \beta \operatorname{vec}(\mathbf{A})$. Further, the elements of matrices

$$\mathbf{G}^{ij} = \mathbf{G}(\mathbf{E}^{ij}) \quad \text{and} \quad \mathbf{B}^{ij} = \mathbf{B}(\mathbf{E}^{ij}), \quad i, j = 1, \ldots, p,$$

list all possible third and fourth moments. Also,

$$\mathbf{G} = \mathbf{G}(\mathbf{I}_p) = \sum_{i=1}^{p} \mathbf{G}^{ii} \quad \text{and} \quad \mathbf{B} = \mathbf{B}(\mathbf{I}_p) = \sum_{i=1}^{p} \mathbf{B}^{ii}$$

appear to be natural measures of multivariate skewness and kurtosis. In the independent component model we then have the following straightforward result.

THEOREM 7. *At the distribution of $\mathbf{z}$ with independent components, $E(\mathbf{z}) = \mathbf{0}$, $\operatorname{Cov}(\mathbf{z}) = \mathbf{I}_p$, and $\kappa_i = E(z_i^4) - 3$, $i = 1, \ldots, p$:*

$$\begin{aligned}
&\boldsymbol{\beta} = \sum_{i=1}^{p} \kappa_i(\mathbf{E}^{ii} \otimes \mathbf{E}^{ij}) + \mathbf{I}_{p,p} + \mathbf{J}_{p,p} + \mathbf{K}_{p,p}, \\
&\mathbf{B}^{ij} = \sum_{k=1}^{p} \kappa_k(\mathbf{E}^{kk}\mathbf{E}^{ij}\mathbf{E}^{kk}) + \mathbf{E}^{ij} + \mathbf{E}^{ji} + \operatorname{tr}(\mathbf{E}^{ij})\mathbf{I}_p, \\
&\qquad\qquad i, j = 1, \ldots, p \quad \textit{and} \\
&\mathbf{B} = \sum_{i=1}^{p} (\kappa_i + p + 2)\mathbf{E}^{ii}.
\end{aligned}$$

REMARK 5. The standardized third and fourth moments have been used as building bricks for invariant multivariate measures of skewness and kurtosis. The classical skewness and kurtosis measures by Mardia (1970) are

$$E\big((\mathbf{x}_{st}'\tilde{\mathbf{x}}_{st})^3\big) \quad \text{and} \quad \mathrm{tr}(\mathbf{B}) = E\big((\mathbf{x}_{st}'\mathbf{x}_{st})^2\big),$$

whereas Móri, Rohatgi and Székely (1993) proposed

$$\|\mathbf{G}\|^2 = E\big(\mathbf{x}_{st}'\mathbf{x}_{st}\mathbf{x}_{st}'\tilde{\mathbf{x}}_{st}\tilde{\mathbf{x}}_{st}'\tilde{\mathbf{x}}_{st}\big) \quad \text{and}$$

$$\mathrm{tr}(\mathbf{B}) = E\big((\mathbf{x}_{st}'\mathbf{x}_{st})^2\big),$$

where $\mathbf{x}_{st}$ and $\tilde{\mathbf{x}}_{st}$ are independent copies of $\mathbf{x}_{st}$ (Móri, Rohatgi and Székely, 1993). (The invariance follows as $\mathbf{x} \to \mathbf{A}\mathbf{x} + \mathbf{b}$ induces $\mathbf{x}_{st} \to \mathbf{U}\mathbf{x}_{st}$ for some orthogonal $\mathbf{U}$.) The sample statistics can then be used to test multivariate normality, for example. For their limiting distributions under the normality assumption, see, for example, Kankainen, Taskinen and Oja (2007). For other extensions of multivariate skewness and kurtosis and their connections to skewness and kurtosis measures above, see Kollo (2008) and Kollo and Srivastava (2004). In Sections 5.2 and 5.3, we first use $\mathbf{B}$ alone and then all $\mathbf{B}^{ij}$, $i, j = 1, \ldots, p$, together to find solutions to the independent component problem. In the signal processing literature, these approaches are called FOBI (fourth order blind identification) and JADE (joint approximate diagonalization of eigenmatrices), correspondingly.

5.2 Use of Kurtosis Matrix B—FOBI

The independent component functional based on the covariance matrix $\boldsymbol{\Sigma}$ and the kurtosis matrix $\mathbf{B}$ defined in (3) is known as FOBI (fourth order blind identification) (Cardoso, 1989) in the engineering literature. It is one of the earliest approaches to the independent component problem and is defined as follows.

DEFINITION 4. The FOBI functional is $\mathbf{W}(F_{\mathbf{x}}) = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x})$ and the rows of $\mathbf{U}$ are the eigenvectors of $\mathbf{B} = E(\mathbf{x}_{st}\mathbf{x}_{st}'\mathbf{x}_{st}\mathbf{x}_{st}')$.

First recall that, in the independent component model, $\mathbf{x}_{st} = \mathbf{U}'\mathbf{z}$ for some orthogonal $\mathbf{U}$. This implies that

$$\mathbf{B} = E\big(\mathbf{x}_{st}\mathbf{x}_{st}'\mathbf{x}_{st}\mathbf{x}_{st}'\big) = \mathbf{U}'E\big(\mathbf{z}\mathbf{z}'\mathbf{z}\mathbf{z}'\big)\mathbf{U},$$

where $E(\mathbf{z}\mathbf{z}'\mathbf{z}\mathbf{z}') = \sum_{i=1}^{p}(\kappa_i + p + 2)\mathbf{E}^{ii}$ is diagonal, and therefore the rows of $\mathbf{U}$ are the eigenvectors of $\mathbf{B}$. The order of the eigenvectors is then given by the order of the corresponding eigenvalues, that is, by the kurtosis order. As $\mathbf{W}$ is also affine equivariant, it is an independent component functional if Assumption 3 is replaced by the following stronger assumption.

ASSUMPTION 5. The fourth moments of $\mathbf{z}$ exist and are distinct.

REMARK 6. Notice that Assumption 5 $\Rightarrow$ Assumption 4 $\Rightarrow$ Assumption 3. If Assumption 5 is not true and there are only $m < p$ distinct kurtosis values with multiplicities $p_1, \ldots, p_m$, FOBI still finds these m subspaces, and the FOBI solutions at $\mathbf{z}$ are of the block-diagonal form $\mathrm{diag}(\mathbf{U}_1, \ldots, \mathbf{U}_m)$ with orthogonal $p_i \times p_i$ matrices $\mathbf{U}_i$, $i = 1, \ldots, m$.

It is again sufficient to consider the limiting distribution of the estimator $\hat{\mathbf{W}} = \mathbf{W}(\mathbf{Z})$ only. Then the asymptotical behavior of the FOBI estimator is given as follows.

THEOREM 8. *Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ be a random sample from a distribution of $\mathbf{z}$ with bounded eighth moments and satisfying the Assumptions 1, 2 and 5 with $\kappa_1 > \cdots > \kappa_p$. Then $\hat{\mathbf{W}} \to_P \mathbf{I}_p$ and*

$$\sqrt{n}(\hat{w}_{kk} - 1) = -\frac{1}{2}\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1) \quad \textit{and}$$

$$\begin{aligned}\sqrt{n}\hat{w}_{kl} &= \Big(\sqrt{n}\hat{r}_{kl} + \sqrt{n}\hat{r}_{lk} + \sqrt{n}\sum_{m \neq k,l}\hat{r}_{mlk} \\ &\quad - (\kappa_k + p + 4)\sqrt{n}\hat{s}_{kl}\Big)\Big/(\kappa_k - \kappa_l) + o_P(1), \\ &\qquad k \neq l.\end{aligned}$$

For an alternative asymptotic presentation of the $\sqrt{n}\hat{w}_{kl}$, see Ilmonen, Nevalainen and Oja (2010). The joint limiting multivariate normality of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{S}}, \mathrm{off}(\hat{\mathbf{R}}))$ then implies the following.

COROLLARY 3. *Under the assumptions of Theorem 8, the limiting distribution of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{W}} - \mathbf{I}_p)$ is a multivariate normal with zero mean vector and componentwise variances*

$$\mathrm{ASV}(\hat{w}_{kk}) = (\kappa_k + 2)/4 \quad \textit{and}$$

$$\begin{aligned}\mathrm{ASV}(\hat{w}_{kl}) &= \Big(\sigma_k^2 + \sigma_l^2 - \kappa_k^2 - 6(\kappa_k + \kappa_l) \\ &\quad - 22 + 2p + \sum_{j \neq k,l}\kappa_j\Big)\Big/(\kappa_k - \kappa_l)^2, \\ &\qquad k \neq l.\end{aligned}$$

REMARK 7. Let $\mathbf{x}$ be a p-vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The FOBI procedure may then be seen also as a comparison of two scatter functionals, namely,

$$\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma} \quad \text{and}$$

$$\mathrm{Cov}_4(\mathbf{x}) = E\big((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'\big),$$

and the FOBI functional then satisfies $\mathbf{W}\,\mathrm{Cov}(\mathbf{x})\mathbf{W}' = \mathbf{I}_p$ and $\mathbf{W}\,\mathrm{Cov}_4(\mathbf{x})\mathbf{W}' \in \mathcal{D}$. Other independent component functionals are obtained if Cov and Cov_4 are replaced by any scatter matrices with the independence property; see Oja, Sirkiä and Eriksson (2006) and Tyler et al. (2009).

5.3 Joint Use of Kurtosis Matrices $\mathbf{B}^{ij}$—JADE

The approach in Section 5.2 was based on the fact that the kurtosis matrix $\mathbf{B}$ is diagonal at $\mathbf{z}$. As shown before, the fourth cumulant matrices

$$\mathbf{C}^{ij} = \mathbf{B}^{ij} - \mathbf{E}^{ij} - (\mathbf{E}^{ij})' - \mathrm{tr}(\mathbf{E}^{ij})\mathbf{I}_p, \quad i,j = 1,\dots,p,$$

are also all diagonal at $\mathbf{z}$. Therefore, a natural idea is to try to find an orthogonal matrix $\mathbf{U}$ such that the matrices $\mathbf{U}\mathbf{C}^{ij}\mathbf{U}'$, $i,j = 1,\dots,p$, are all "as diagonal as possible." In the engineering literature this approach is known as joint approximate diagonalization of eigenmatrices (JADE); see Cardoso and Souloumiac (1993). The functional is then defined as follows.

DEFINITION 5. The JADE functional is $\mathbf{W}(F_{\mathbf{x}}) = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x})$ and the orthogonal matrix U maximizes

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\big\|\mathrm{diag}(\mathbf{U}\mathbf{C}^{ij}\mathbf{U}')\big\|^2.$$

First note that

$$\sum_{i=1}^{p}\sum_{j=1}^{p}\big\|\mathrm{diag}(\mathbf{U}\mathbf{C}^{ij}\mathbf{U}')\big\|^2 + \sum_{i=1}^{p}\sum_{j=1}^{p}\big\|\mathrm{off}(\mathbf{U}\mathbf{C}^{ij}\mathbf{U}')\big\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\|\mathbf{C}^{ij}\|^2.$$

The solution thus minimizes the sum of squared off-diagonal elements of $\mathbf{U}\mathbf{C}^{ij}\mathbf{U}'$, $i,j = 1,\dots,p$. Notice that, at $\mathbf{z}$, the only possible nonzero elements of $\mathbf{C}^{ij}$, $i,j = 1,\dots,p$, are $(\mathbf{C}^{ii})_{ii} = \kappa_i$. For the separation of the components, we therefore need Assumption 4 saying that at most one of the kurtosis values κ_i is zero. The JADE functional $\mathbf{W}(F)$ is an IC functional, as we can prove in the following.

THEOREM 9. (i) *Write $\mathbf{x}_{st} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$ for the standardized random vector from the IC model (2) satisfying the Assumptions 1, 2, and 4. If $\mathbf{x}_{st} = \mathbf{U}'\mathbf{z}$, then*

$$D(\mathbf{V}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\big\|\mathrm{diag}(\mathbf{V}\mathbf{C}^{ij}\mathbf{V}')\big\|^2, \quad \mathbf{V} \in \mathcal{U}$$

is maximized by any $\mathbf{PJU}$ where $\mathbf{P} \in \mathcal{P}$ and $\mathbf{J} \in \mathcal{J}$.

(ii) *For any $F_{\mathbf{x}}$ with finite fourth moments, $\mathbf{W}(F_{\mathbf{Ax+b}}) = \mathbf{PJW}(F_{\mathbf{x}})\mathbf{A}^{-1}$ for some $\mathbf{P} \in \mathcal{P}$ and $\mathbf{J} \in \mathcal{J}$.*

In this case, the matrix $\mathbf{U} = (\mathbf{u}_1,\dots,\mathbf{u}_p)'$ thus optimizes the Lagrangian function

$$L(\mathbf{U},\boldsymbol{\Theta}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}(\mathbf{u}_k'\mathbf{C}^{ij}\mathbf{u}_k)^2 - \sum_{k=1}^{p}\theta_{kk}(\mathbf{u}_k'\mathbf{u}_k - 1) - \sum_{k=1}^{p-1}\sum_{l=k+1}^{p}\theta_{lk}\mathbf{u}_k'\mathbf{u}_l,$$

where the symmetric matrix $\boldsymbol{\Theta} = (\theta_{ij})$ contains the $p(p+1)/2$ Lagrangian multipliers of the optimization problem. Write

$$\mathbf{T}(\mathbf{u}) = \sum_{i=1}^{p}\sum_{j=1}^{p}(\mathbf{u}'\mathbf{C}^{ij}\mathbf{u})\mathbf{C}^{ij}\mathbf{u} \quad \text{and}$$

$$\mathbf{T}(\mathbf{U}) = \big(\mathbf{T}(\mathbf{u}_1),\dots,\mathbf{T}(\mathbf{u}_p)\big)'.$$

The Lagrangian function then yields the estimating equations

$$\mathbf{u}_i'\mathbf{T}(\mathbf{u}_j) = \mathbf{u}_j'\mathbf{T}(\mathbf{u}_i) \quad \text{and}$$

$$\mathbf{u}_i'\mathbf{u}_j = \delta_{ij}, \quad i,j = 1,\dots,p,$$

and the equations suggest a fixed-point algorithm with the steps $\mathbf{U} \leftarrow \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1/2}$. The estimating equations can also again be used to find the following asymptotical distribution of the JADE estimate $\hat{\mathbf{W}} = \mathbf{W}(\mathbf{Z})$.

THEOREM 10. *Let $\mathbf{Z} = (\mathbf{z}_1,\dots,\mathbf{z}_n)$ be a random sample from a distribution of $\mathbf{z}$ with bounded eighth moments satisfying the Assumptions 1, 2, and 4. Then there is a sequence of solutions $\hat{\mathbf{W}}$ such that $\hat{\mathbf{W}} \to_P \mathbf{I}_p$ and*

$$\sqrt{n}(\hat{w}_{kk} - 1) = -1/2\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1), \quad k = l$$

and

$$\sqrt{n}\hat{w}_{kl} = \frac{\kappa_k\sqrt{n}\hat{r}_{kl} - \kappa_l\sqrt{n}\hat{r}_{lk} + (3\kappa_l - 3\kappa_k - \kappa_k^2)\sqrt{n}\hat{s}_{kl}}{\kappa_k^2 + \kappa_l^2} + o_P(1), \quad k \neq l.$$

COROLLARY 4. *Under the assumptions of Theorem 10, the limiting distribution of* $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{W}} - \mathbf{I}_p)$ *is a multivariate normal with zero mean vector and componentwise variances*

$$\mathrm{ASV}(\hat{w}_{kk}) = (\kappa_k + 2)/4 \quad \textit{and}$$

$$\mathrm{ASV}(\hat{w}_{kl}) = \frac{\kappa_k^2(\sigma_k^2 - \kappa_k^2 - 6\kappa_k - 9) + \kappa_l^2(\sigma_l^2 - 6\kappa_l - 9)}{(\kappa_k^2 + \kappa_l^2)^2}, \quad k \neq l.$$

REMARK 8. In the literature, there are several alternative algorithms available for an approximate diagonalization of several symmetric matrices, but the statistical properties of the corresponding estimates are not known. The most popular algorithm is perhaps the Jacobi rotation algorithm suggested in Clarkson (1988). It appeared in our simulations that the Jacobi rotation algorithm is computationally much faster and always provides the same solution as our fixed-point algorithm. The limiting distribution with variances and covariances of the elements of the JADE estimate (but without the standardization step) was considered also in Bonhomme and Robin (2009).

REMARK 9. The JADE estimate uses p^2 fourth moment matrices in order to be affine equivariant. Therefore, the computational load of JADE grows quickly with the number of components. Miettinen et al. (2013) suggested a quite similar, but faster method, called k-JADE. The k-JADE estimate at $F_{\mathbf{x}}$ is $\mathbf{W} = \mathbf{U}\mathbf{W}_0$, where $\mathbf{W}_0$ is the FOBI estimate and the orthogonal matrix $\mathbf{U}$ maximizes

$$\sum_{|i-j|<k} \|\mathrm{diag}(\mathbf{U}\mathbf{C}^{ij}\mathbf{U}')\|^2,$$

where the $\mathbf{C}^{ij}$'s are calculated for $\mathbf{x}_{st} = \mathbf{W}_0(\mathbf{x} - \boldsymbol{\mu})$. It seems to us that this estimate is asymptotically equivalent to the regular JADE estimate (with much smaller computational load) if the multiplicities of the distinct kurtosis values are at most k. Detailed studies are, however, still missing.

6. COMPARISON OF THE ASYMPTOTIC VARIANCES OF THE ESTIMATES

First notice that, for all estimates, $\sqrt{n}(\mathbf{W}(\mathbf{X}) - \boldsymbol{\Omega}^{-1}) = \sqrt{n}(\mathbf{W}(\mathbf{Z}) - \mathbf{I}_p)\boldsymbol{\Omega}^{-1}$ and the comparisons can be made using $\hat{\mathbf{W}} = \mathbf{W}(\mathbf{Z})$ only. Second, for all estimates, $\sqrt{n}(\hat{w}_{kk} - 1) = -1/2\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1)$ $k = 1, \dots, p$, and therefore the diagonal elements of $\hat{\mathbf{W}}$ should not be used in the comparison. It is then natural to compare the estimates using the sum of asymptotic variances of the off-diagonal elements of $\hat{\mathbf{W}}$, that is,

$$\sum_{k=1}^{p-1} \sum_{l=k+1}^{p} \big(\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})\big). \tag{4}$$

Next note that, for all estimates, except FOBI, the limiting variances of $\sqrt{n}\hat{w}_{kl}$, $k \neq l$, surprisingly depend only on the kth and lth marginal distribution (through κ_k, κ_l, σ_k^2, and σ_l^2) and do not depend either on the number or on the distributions of the other components. Based on the results in the earlier sections, we have the following conclusions:

1. $\sqrt{n}\hat{w}_{kl}$ of the symmetric FastICA estimate and that of the JADE estimate are asymptotically equivalent, that is, their difference converges to zero in probability if the kth and lth marginal distributions are the same.
2. If the independent components are identically distributed, then the symmetric FastICA and JADE estimates are asymptotically equivalent. In this case, their criterium value (4) is one half of that of the deflation-based FastICA estimate. The FOBI estimate fails in this case.
3. $\mathrm{ASV}(\hat{w}_{kl})$ of the FOBI estimate is always larger than or equal to that for symmetric FastICA, $k \neq l$. This follows as $\kappa_k \geq -2$ for all k. The larger the other kurtosis values, the larger is the $\mathrm{ASV}(\hat{w}_{kl})$ of FOBI. The variances are equal when $p = 2$ and $\kappa_k > 0 > \kappa_l$.
4. $\sqrt{n}\hat{w}_{kp}$ of the deflation-based FastICA estimate and of the JADE estimate are asymptotically equivalent if the pth marginal distribution is normal.

The criterium value (4) is thus the sum of the pairwise terms $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$, which do not depend on the number or distributions of other components except for the FOBI estimate. So in most cases the comparison of the estimates can be made only through the values $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$. To make FOBI (roughly) comparable, we use the lower bound of the value $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ with $\kappa_j = -2$, $j \neq k, l$; the lower bound is in fact the exact value in the bivariate case. In Table 1, the values $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ are listed for pairs of independent components from the following five distributions: exponential distribution (EX), logistic distribution (L), uniform distribution (U), exponential power distribution with shape parameter value 4 (EP), and normal or Gaussian (G) distribution. The excess kurtosis values are $\kappa_{\mathrm{EX}} = 6$, $\kappa_{\mathrm{L}} = 1.2$, $\kappa_{\mathrm{U}} = -1.8$, $\kappa_{\mathrm{EP}} \approx -0.81$ and

TABLE 1
The values of $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ *for some selected kth and lth component distributions and for deflation-based FastICA (DFICA), symmetric FastICA (SFICA), FOBI, and JADE estimates. For FOBI, the lower bound of* $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ *is used*

	DFICA	SFICA	FOBI	JADE
EX–EX	11.00	5.50	∞	5.50
EX–L	11.00	8.52	19.18	10.22
EX–U	11.00	7.69	7.69	10.17
EX–EP	11.00	8.63	8.63	10.61
EX–G	11.00	11.33	11.33	11.00
L–L	31.86	15.93	∞	15.93
L–U	31.86	8.43	8.43	8.43
L–EP	31.86	12.38	12.38	15.63
L–G	31.86	40.19	40.19	31.86
U–U	1.86	0.93	∞	0.93
U–EP	1.86	1.80	40.63	1.50
U–G	1.86	10.19	10.19	1.86
EP–EP	6.39	3.20	∞	3.20
EP–G	6.39	34.61	34.61	6.39

$\kappa_G = 0$, respectively. The results in Table 1 are then nicely in accordance with our general notions above and show that none of the estimates outperforms all the other estimates.

Further, in Figure 1, we plot the values $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ when the independent components come (i) from the standardized (symmetric) exponential power distribution or (ii) from the standardized (skew) gamma distribution. The limiting variances then depend only on the shape parameters of the models. In the plot, the darker the point, the higher the value and the worse the estimate. The density function for the exponential power distribution with zero mean and variance one and with shape parameter β is

$$f(x) = \frac{\beta \exp\{-(|x|/\alpha)^{\beta}\}}{2\alpha\Gamma(1/\beta)},$$

where $\beta > 0$, $\alpha = (\Gamma(1/\beta)/\Gamma(3/\beta))^{1/2}$, and Γ is the gamma function. Notice that $\beta = 2$ gives the normal (Gaussian) distribution, $\beta = 1$ gives the heavy-tailed Laplace distribution, and the density converges to an extremely low-tailed uniform density as $\beta \to \infty$. The family of skew distributions for the variables is coming from the gamma distribution with shape parameter α and shifted and rescaled to have mean zero and variance one. For $\alpha = k/2$, the distribution is a chi-square distribution with k degrees of freedom, $k = 1, 2, \ldots$. For $\alpha = 1$, an exponential distribution is obtained, and the distribution is converging to a normal distribution as $\alpha \to \infty$.

For all estimates, Figure 1 shows that $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ gets high values with β close to 2 (normal distribution). Also, the variances are growing with increasing α. The FOBI estimate is poor if the marginal kurtosis values are close to each other. The contours for the deflation-based FastICA estimate illustrate the fact that the criterium function $\mathrm{ASV}(\hat{w}_{12}) + \mathrm{ASV}(\hat{w}_{21})$ is not continuous at the points for which $\kappa_k + \kappa_l = 0$. This is due to the fact that the order in which the components are found changes at that point. The symmetric FastICA and JADE estimates are clearly the best estimates with minor differences.

7. DISCUSSION

Many popular methods to solve the independent component analysis problem are based on the use of univariate and multivariate fourth moments. Examples include FOBI (Cardoso, 1989), JADE (Cardoso and Souloumiac, 1993), and FastICA (Hyvärinen, 1999). In the engineering literature, these ICA methods have originally been formulated and regarded as algorithms only, and therefore the rigorous analysis and comparison of their statistical properties have been missing until very recently. The statistical properties of the deflation-based FastICA method were derived in Ollila (2010) and Nordhausen et al. (2011). The asymptotical behavior of the FOBI estimate was considered in Ilmonen, Nevalainen and Oja (2010), and the asymptotical distribution of the JADE estimate (without the standardization step) was considered in Bonhomme and Robin (2009). This paper describes in detail the independent component functionals based on fourth moments through corresponding optimization problems, estimating equations, fixed-point algorithms and the assumptions they need, and provides for the very first time the limiting statistical properties of the JADE estimate. Careful comparisons of the asymptotic variances revealed that, as was expected, JADE and the symmetric version of FastICA performed best in most cases. It was surprising, however, that the JADE and symmetric FastICA estimates are asymptotically equivalent if the components are identically distributed. The only noteworthy difference between these two estimators appeared when one of the components has a normal distribution. Then JADE outperforms symmetric FastICA. Recall that JADE requires the computation of p^2 matrices of size $p \times p$ and, thus, the use of JADE becomes impractical with a large number of independent components. On the other hand, FastICA

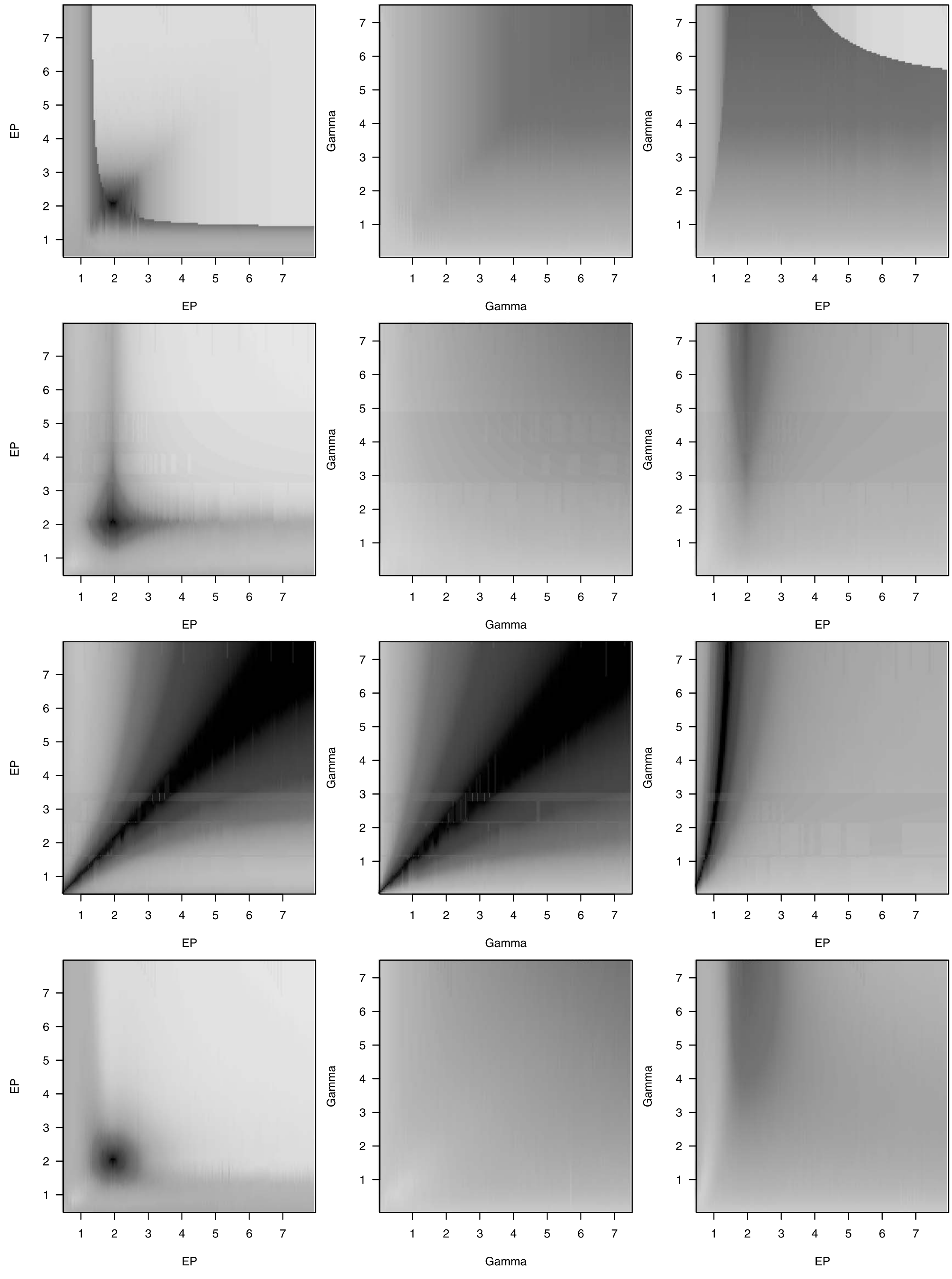

FIG. 1. *Contour maps of* $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ *for different estimates and for different independent component distributions. The distributions are either exponential power distributed* (*EP*) *or gamma distributed* (*Gamma*) *with varying shape parameter values. The estimates, from up to down, are deflation-based FastICA, symmetric FastICA, FOBI, and JADE. For FOBI, the lower bound of* $\mathrm{ASV}(\hat{w}_{kl}) + \mathrm{ASV}(\hat{w}_{lk})$ *is used. The lighter the color is, the lower is the variance.*

estimates are sometimes difficult to find due to convergence problems of the algorithms, when the sample size is small.

In this paper we considered only the most basic IC model, where the number of independent components equals the observed dimension and where no additive noise is present. In further research we will consider also these cases. Note that some properties of JADE for noisy ICA were considered in Bonhomme and Robin (2009).

APPENDIX: PROOFS OF THE THEOREMS

PROOF OF THEOREM 1. Let $\boldsymbol{\Omega} = \mathbf{ODV}'$ be the singular value decomposition of full-rank $\boldsymbol{\Omega}$. Then $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}' = \mathbf{OD}^2\mathbf{O}'$, and $\boldsymbol{\Sigma}^{-1/2} = \mathbf{OJD}^{-1}\mathbf{O}'$ for some $\mathbf{J} \in \mathcal{J}$. ($\mathbf{J}$ is needed to make $\boldsymbol{\Sigma}^{-1/2}$ positive definite.) Then

$$\begin{aligned}\mathbf{x}_{st} &= \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{OJD}^{-1}\mathbf{O}'\mathbf{ODV}'\mathbf{z}\\ &= \mathbf{OJV}'\mathbf{z} = \mathbf{Uz}\end{aligned}$$

with an orthogonal $\mathbf{U} = \mathbf{OJV}'$. □

PROOF OF THEOREM 2. If $\mathbf{u}'\mathbf{u} = 1$, then it is straightforward to see that

$$E[(\mathbf{u}'\mathbf{z})^4 - 3] = \sum_{i=1}^{p} u_i^4[E(z_i^4) - 3].$$

It then easily follows that

$$\begin{aligned}|E[(\mathbf{u}'\mathbf{z})^4] - 3| &\le \sum_{i=1}^{p} u_i^4 |E(z_i^4) - 3|\\ &\le \max_{i=1,\dots,p} |E(z_i^4) - 3|\end{aligned}$$

and that, for any orthogonal $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)'$,

$$\begin{aligned}\sum_{j=1}^{p} |E[(\mathbf{u}_j'\mathbf{z})^4] - 3| &\le \sum_{i=1}^{p}\Bigl(\sum_j u_{ji}^4\Bigr)|E(z_i^4) - 3|\\ &\le \sum_{i=1}^{p} |E(z_i^4) - 3|.\end{aligned}$$

For the first result, see also Lemma 2 in Bugrien and Kent (2005). □

PROOF OF THEOREM 6. As the functions

$$D_n(\mathbf{U}) = \sum_{j=1}^{p}\left| n^{-1}\sum_{i=1}^{n}(\mathbf{u}_j'\mathbf{x}_{st,i})^4 - 3\right| \quad \text{and}$$

$$D(\mathbf{U}) = \sum_{j=1}^{p} |E[(\mathbf{u}_j'\mathbf{z})^4] - 3|$$

are continuous and $D_n(\mathbf{U}) \to_P D(\mathbf{U})$ for all $\mathbf{U}$, then, due to the compactness of $\mathcal{U}$, also

$$\sup_{\mathbf{U}\in\mathcal{U}} |D_n(\mathbf{U}) - D(\mathbf{U})| \to_P 0.$$

$D(\mathbf{U})$ attains its maximum at any $\mathbf{JP}$, where $\mathbf{J} \in \mathcal{J}$ and $\mathbf{P} \in \mathcal{P}$. This further implies that there is a sequence of maximizers that satisfy $\hat{\mathbf{U}} \to_P \mathbf{I}_p$, and therefore also $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{S}}^{-1/2} \to_P \mathbf{I}_p$.

For the estimate $\hat{\mathbf{W}}$, the estimating equations are

$$\begin{aligned}\hat{\mathbf{w}}_k'\hat{\mathbf{T}}(\hat{\mathbf{w}}_l)\hat{\pi}_l &= \hat{\mathbf{w}}_l'\hat{\mathbf{T}}(\hat{\mathbf{w}}_k)\hat{\pi}_k \quad \text{and}\\ \hat{\mathbf{w}}_k'\hat{\mathbf{S}}\hat{\mathbf{w}}_l &= \delta_{ij}, \quad k, l = 1, \dots, p,\end{aligned}$$

where $\hat{\mathbf{T}}(\hat{\mathbf{w}}_k) = n^{-1}\sum_i(\hat{\mathbf{w}}_k'(\mathbf{z}_i - \bar{\mathbf{z}}))^3(\mathbf{z}_i - \bar{\mathbf{z}})$. It is straightforward to see that the second set of estimating equations gives

$$\begin{aligned}&\sqrt{n}(\hat{w}_{kk} - 1) = -2^{-1}\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1) \quad \text{and}\\ &\sqrt{n}(\hat{w}_{kl} + \hat{w}_{lk}) = -\sqrt{n}\hat{s}_{kl} + o_P(1).\end{aligned} \tag{5}$$

Consider then the first set of estimating equations for $k \neq l$. To shorten the notation, write $\hat{\mathbf{T}}(\hat{\mathbf{w}}_k) = \hat{\mathbf{T}}_k$. Now

$$\sqrt{n}\hat{\mathbf{w}}_k'\hat{\mathbf{T}}_l = \sqrt{n}(\hat{\mathbf{w}}_k - \mathbf{e}_k)'\hat{\mathbf{T}}_l + \sqrt{n}\mathbf{e}_k'(\hat{\mathbf{T}}_l - \beta_l\mathbf{e}_l).$$

Using equation (2) in Nordhausen et al. (2011) and Slutsky's theorem, the above equation reduces to

$$\begin{aligned}&\sqrt{n}\hat{\mathbf{w}}_k'\hat{\mathbf{T}}_l\hat{\pi}_l\\ &\quad = \bigl(\sqrt{n}(\hat{\mathbf{w}}_k - \mathbf{e}_k)'\beta_l\mathbf{e}_l + \mathbf{e}_k'\bigl(\sqrt{n}\hat{\mathbf{T}}_l^* - \gamma_l\mathbf{e}_l\mathbf{e}_l'\sqrt{n}\bar{\mathbf{x}}\\ &\qquad\qquad + \boldsymbol{\Delta}_l\sqrt{n}(\hat{\mathbf{w}}_l - \mathbf{e}_l)\bigr)\bigr)\pi_l\\ &\qquad + o_P(1),\end{aligned}$$

where $\hat{\mathbf{T}}_l^* = n^{-1}\sum_i((\mathbf{e}_l'\mathbf{z}_i)^3 - \gamma_l)\mathbf{z}_i$ and $\boldsymbol{\Delta}_l = 3E[(\mathbf{e}_l'\mathbf{z}_i)^2\mathbf{z}_i\mathbf{z}_i']$. According to our estimating equation, the above expression should be equivalent to

$$\begin{aligned}\sqrt{n}\hat{\mathbf{w}}_l'\hat{\mathbf{T}}_k\hat{\pi}_k = \bigl(&\sqrt{n}(\hat{\mathbf{w}}_l - \mathbf{e}_l)'\beta_k\mathbf{e}_k\\ &+ \mathbf{e}_l'\bigl(\sqrt{n}\hat{\mathbf{T}}_k^* - \gamma_k\mathbf{e}_k\mathbf{e}_k'\sqrt{n}\bar{\mathbf{z}}\\ &+ \boldsymbol{\Delta}_k\sqrt{n}(\hat{\mathbf{w}}_i - \mathbf{e}_i)\bigr)\bigr)\pi_k + o_P(1).\end{aligned}$$

This further implies that

$$\begin{aligned}&(\beta_l\pi_l - 3\pi_k)\sqrt{n}\hat{w}_{kl} - (\beta_k\pi_k - 3\pi_l)\sqrt{n}\hat{w}_{lk}\\ &\quad = \sqrt{n}(\hat{r}_{kl}\pi_k + \hat{r}_{lk}\pi_l) + o_P(1),\end{aligned}$$

where $\hat{r}_{kl} = \sum_i(z_{ik}^3 - \gamma_k)z_{il}$. Now using (5), we have that

$$\begin{aligned}&(\beta_l\pi_l - 3\pi_k)\sqrt{n}\hat{w}_{kl}\\ &\qquad + (\beta_k\pi_k - 3\pi_l)(\sqrt{n}\hat{s}_{kl} + \sqrt{n}\hat{w}_{kl})\\ &\quad = \sqrt{n}(\hat{r}_{kl}\pi_k + \hat{r}_{lk}\pi_l) + o_P(1).\end{aligned}$$

Then

$$\begin{aligned}(|\beta_k - 3| + |\beta_l - 3|)\sqrt{n}\hat{w}_{kl} \\ &= \sqrt{n}(\hat{r}_{kl}\pi_k + \hat{r}_{lk}\pi_l) + (3\pi_l - \beta_k\pi_k)\sqrt{n}\hat{s}_{kl} \\ &\quad + o_P(1),\end{aligned}$$

which gives the desired result. □

PROOF OF THEOREM 8. As mentioned in Remark 7, the FOBI functional diagonalizes the scatter matrices $\mathrm{Cov}(\mathbf{x})$ and $\mathrm{Cov}_4(\mathbf{x}) = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})']$ simultaneously. Then $\mathrm{Cov}(\mathbf{z}) = \mathbf{I}_p$ and $\mathrm{Cov}_4(\mathbf{z}) = \mathbf{D}$ with strictly decreasing diagonal elements. Next write $\hat{\mathbf{S}}$ and $\hat{\mathbf{S}}_4$ for the empirical scatter matrices. Then $\hat{\mathbf{S}} \to_P \mathbf{I}_p$ and $\hat{\mathbf{S}}_4 \to_P \mathbf{D}$, and, as $\hat{\mathbf{W}}$ is a continuous function of $(\hat{\mathbf{S}}, \hat{\mathbf{S}}_4)$ in a neighborhood of $(\mathbf{I}_p, \mathbf{D})$, also $\hat{\mathbf{W}} \to_P \mathbf{I}_p$.

Let $\tilde{\mathbf{Z}} = (\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_n) = (\mathbf{z}_1 - \bar{\mathbf{z}}, \dots, \mathbf{z}_n - \bar{\mathbf{z}})$ denote the centered sample,

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{S}}_4 - \mathbf{D}) &= n^{-1/2}\sum_{i=1}^n (\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i' - \mathbf{D}) \\ &= -n^{-1}\sum_{i=1}^n \tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p)\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i' \\ &\quad + n^{-1/2}\sum_{i=1}^n (\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i' - \mathbf{D}),\end{aligned}$$

where the (k,l) element, $k \neq l$, of the first matrix is

$$\begin{aligned}-n^{-1}\sum_{i=1}^n \tilde{z}_{ki}\tilde{z}_i'\sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p)\tilde{z}_i\tilde{z}_{li} \\ &= -2n^{-1}\sum_{i=1}^n \tilde{z}_{ki}^2\tilde{z}_{li}^2\sqrt{n}\hat{s}_{kl} + o_P(1) \\ &= -2\sqrt{n}\hat{s}_{kl} + o_P(1),\end{aligned}$$

and the (k,l) element of the second matrix is

$$\begin{aligned}n^{-1/2}\sum_{i=1}^n \tilde{z}_{ki}^3\tilde{z}_{li} + n^{-1/2}\sum_{i=1}^n \tilde{z}_{ki}\tilde{z}_{li}^3 \\ + n^{-1/2}\sum_{i=1}^n\sum_{m\neq k,l} \tilde{z}_{mi}^2\tilde{z}_{ki}\tilde{z}_{li}.\end{aligned}$$

Thus,

$$\sqrt{n}(\hat{\mathbf{S}}_4)_{kl} = \sqrt{n}\hat{r}_{kl} + \sqrt{n}\hat{r}_{lk} + \sum_{m\neq k,l}\hat{r}_{mkl} + o_P(1).$$

Then Theorem 3.1 of Ilmonen, Nevalainen and Oja (2010) gives

$$\begin{aligned}\sqrt{n}\hat{w}_{kl} \\ &= \frac{\sqrt{n}(\hat{\mathbf{S}}_4)_{kl} - (\kappa_k + p + 2)\sqrt{n}\hat{s}_{kl}}{\kappa_k + p + 2 - (\kappa_l + p + 2)} + o_P(1) \\ &= \Bigl(\sqrt{n}\hat{r}_{kl} + \sqrt{n}\hat{r}_{lk} + \sum_{m\neq k,l}\sqrt{n}\hat{r}_{mkl} \\ &\quad - (\kappa_k + p + 4)\sqrt{n}\hat{s}_{kl}\Bigr)\Big/(\kappa_k - \kappa_l) + o_P(1).\end{aligned}$$

□

To prove Theorem 9, we need the following lemma.

LEMMA 1. *Denote*

$$\mathbf{C}(\mathbf{x}, \mathbf{A}) = E[(\mathbf{x}'\mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}'] - \mathbf{A} - \mathbf{A}' - \mathrm{tr}(\mathbf{A})\mathbf{I}_p,$$
$$\mathbf{C}^{ij}(\mathbf{x}) = E[(\mathbf{x}'\mathbf{E}^{ij}\mathbf{x})\mathbf{x}\mathbf{x}'] - \mathbf{E}^{ij} - \mathbf{E}^{ji} - \mathrm{tr}(\mathbf{E}^{ij})\mathbf{I}_p,$$

where $\mathbf{E}^{ij} = \mathbf{e}_i\mathbf{e}_j'$, $i, j = 1, \dots, p$. *Then* $\mathbf{C}(\mathbf{x}, \mathbf{A})$ *is additive in* $\mathbf{A} = (a_{ij})$, *that is*,

$$\mathbf{C}(\mathbf{x}, \mathbf{A}) = \sum_{i=1}^p\sum_{j=1}^p \mathbf{a}_{ij}\mathbf{C}^{ij}(\mathbf{x}).$$

Also, for an orthogonal $\mathbf{U}$, *it holds that*

$$\mathbf{C}(\mathbf{U}\mathbf{x}, \mathbf{A}) = \mathbf{U}\mathbf{C}(\mathbf{x}, \mathbf{U}'\mathbf{A}\mathbf{U})\mathbf{U}'.$$

PROOF. For additivity, it is straightforward to see that, for all $\mathbf{A}$, $\mathbf{A}_1$, $\mathbf{A}_2$, and b,

$$\mathbf{C}(\mathbf{x}, b\mathbf{A}) = b\mathbf{C}(\mathbf{x}, \mathbf{A}) \quad \text{and}$$
$$\mathbf{C}(\mathbf{x}, \mathbf{A}_1 + \mathbf{A}_2) = \mathbf{C}(\mathbf{x}, \mathbf{A}_1) + \mathbf{C}(\mathbf{x}, \mathbf{A}_2).$$

For orthogonal $\mathbf{U}$, we obtain

$$\begin{aligned}\mathbf{C}(\mathbf{U}\mathbf{x}, \mathbf{A}) \\ &= E[(\mathbf{x}'\mathbf{U}'\mathbf{A}\mathbf{U}\mathbf{x})(\mathbf{U}\mathbf{x}\mathbf{x}'\mathbf{U}')] - \mathbf{A} - \mathbf{A}' - \mathrm{tr}(\mathbf{A})\mathbf{I}_p \\ &= \mathbf{U}(E[(\mathbf{x}'(\mathbf{U}'\mathbf{A}\mathbf{U})\mathbf{x})\mathbf{x}\mathbf{x}'] - (\mathbf{U}'\mathbf{A}\mathbf{U}) - (\mathbf{U}'\mathbf{A}\mathbf{U}) \\ &\quad - \mathrm{tr}((\mathbf{U}'\mathbf{A}\mathbf{U}))\mathbf{I}_p)\mathbf{U}' \\ &= \mathbf{U}\mathbf{C}(\mathbf{x}, \mathbf{U}'\mathbf{A}\mathbf{U})\mathbf{U}'.\end{aligned}$$

□

PROOF OF THEOREM 9. (i) First notice that

$$\mathbf{C}^{ij}(\mathbf{z}) = \mathbf{0}, \quad \text{for } i, j = 1, \dots, p \text{ and } i \neq j$$
$$\mathbf{C}^{ii}(\mathbf{z}) = \kappa_i\mathbf{E}^{ii}, \quad \text{for } i = 1, \dots, p.$$

It then follows that, for an orthogonal $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)$,

$$\begin{aligned}C^{ij}(\mathbf{U}'\mathbf{z}) &= \mathbf{U}'\mathbf{C}(\mathbf{z}, \mathbf{U}\mathbf{E}^{ij}\mathbf{U}')\mathbf{U} \\ &= \mathbf{U}'\mathbf{C}\Bigl(\mathbf{z}, \sum_{k=1}^p\sum_{l=1}^p u_{ki}u_{lj}\mathbf{E}_{kl}\Bigr)\mathbf{U}\end{aligned}$$

$$= \mathbf{U}'\left(\sum_{k=1}^{p}\sum_{l=1}^{p} u_{ki}u_{lj}\mathbf{C}(\mathbf{z},\mathbf{E}_{kl})\right)\mathbf{U}$$

$$= \mathbf{U}'\left(\sum_{k=1}^{p} \kappa_k u_{ki}u_{kj}\mathbf{E}_{kk}\right)\mathbf{U}.$$

Now

$$D(\mathbf{V}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\|\mathrm{diag}(\mathbf{V}\mathbf{C}^{ij}(\mathbf{U}'\mathbf{z})\mathbf{V}')\|^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p}\left\|\mathbf{V}\mathbf{U}'\left(\sum_{k=1}^{p}\kappa_k u_{ki}u_{kj}E_{kk}\right)(\mathbf{V}\mathbf{U}')'\right\|^2.$$

If we write $\mathbf{G} = \mathbf{V}\mathbf{U}' = (\mathbf{g}_1,\dots,\mathbf{g}_p)$, then $D(\mathbf{V})$ simplifies to

$$D(\mathbf{V}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\sum_{l=1}^{p} g_{kl}^2\kappa_l u_{li}u_{lj}\right)^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{l^*=1}^{p} g_{kl}^2 g_{kl^*}^2\kappa_l\kappa_{l^*}u_{li}u_{lj}u_{l^*i}u_{l^*j}$$

$$= \sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{l^*=1}^{p} g_{kl}^2 g_{kl^*}^2\kappa_l\kappa_{l^*}$$

$$\cdot\sum_{i=1}^{p}(u_{li}u_{l^*i})\sum_{j=1}^{p}(u_{lj}u_{l^*j})$$

$$= \sum_{k=1}^{p}\sum_{l=1}^{p} g_{kl}^4\kappa_l^2,$$

which is maximized by $\mathbf{V} = \mathbf{PJU}$ for any $\mathbf{P}\in\mathcal{P}$ and $\mathbf{J}\in\mathcal{J}$.

(ii) Write $\mathbf{y} = \mathbf{Ax}+\mathbf{b}$, and let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean vector and covariance matrix of $\mathbf{x}$, respectively. As

$$(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')^{-1/2}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')^{-1/2} = \mathbf{I}_p,$$

we have that $(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')^{-1/2}\mathbf{A} = \mathbf{Q}\boldsymbol{\Sigma}^{-1/2}$ for some $\mathbf{Q}\in\mathcal{U}$, and therefore $\mathbf{y}_{st} = \mathbf{Q}\mathbf{x}_{st}$ with the same $\mathbf{Q}\in\mathcal{U}$.

We thus define $\mathbf{W}(F_{\mathbf{x}}) = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$, where $\mathbf{U}$ maximizes the function

$$D_{\mathbf{x}_{st}}(\mathbf{V}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\|\mathrm{diag}(\mathbf{V}\mathbf{C}^{ij}(\mathbf{x}_{st})\mathbf{V}')\|^2.$$

The maximizer $\mathbf{U}$ is not unique, as the maximum is then attained for any $\mathbf{PJU}$ where $\mathbf{P}\in\mathcal{P}$ and $\mathbf{J}\in\mathcal{J}$.

Consider next the criterium function for the standardized transformed random variable $\mathbf{y}_{st}$. Then

$$D_{y_{st}}(\mathbf{V}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\|\mathrm{diag}(\mathbf{V}\mathbf{C}^{ij}(\mathbf{Q}\mathbf{x}_{st})\mathbf{V}')\|^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p}\|\mathbf{VQC}(\mathbf{x}_{st},\mathbf{Q}'\mathbf{E}^{ij}\mathbf{Q})\mathbf{Q}'\mathbf{V}'\|^2.$$

If we write $\mathbf{G} = \mathbf{VQ} = (\mathbf{g}_1,\dots,\mathbf{g}_p)'$, then

$$D_{y_{st}}(\mathbf{V})$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}(\mathbf{g}_k\mathbf{C}(\mathbf{x}_{st},\mathbf{Q}'\mathbf{E}^{ij}\mathbf{Q})\mathbf{g}_k')^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\sum_{l=1}^{p}\sum_{m=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p} g_{ks}g_{kt}q_{il}q_{jm}\right.$$

$$\left.\cdot\,\mathbf{C}(\mathbf{x}_{st},\mathbf{E}^{lm})_{st}\right)^2$$

$$= \sum_{i,j,k,l,l^*,m,m^*,s,s^*,t,t^*=1}^{p} g_{ks}g_{kt}g_{ks^*}g_{kt^*}q_{il}$$

$$\cdot\, q_{jm}q_{il^*}q_{jm^*}\mathbf{C}(\mathbf{x}_{st},\mathbf{E}^{lm})_{st}$$

$$\cdot\,\mathbf{C}(\mathbf{x}_{st},\mathbf{E}^{l^*m^*})_{s^*t^*}$$

$$= \sum_{k,l,l^*,m,m^*,s,s^*,t,t^*=1}^{p} g_{ks}g_{kt}g_{ks^*}g_{kt^*}\mathbf{C}(\mathbf{x}_{st},\mathbf{E}^{lm})_{st}$$

$$\cdot\,\mathbf{C}(\mathbf{x}_{st},\mathbf{E}^{l^*m^*})_{s^*t^*}$$

$$\cdot\sum_i(u_{il}u_{il^*})\sum_j(u_{jm}u_{jm^*})$$

$$= \sum_{k,l,m,s,s^*,t,t^*=1}^{p} g_{ks}g_{kt}g_{ks^*}g_{kt^*}\mathbf{C}(\mathbf{x}_{st},\mathbf{E}^{lm})_{st}$$

$$\cdot\,\mathbf{C}(\mathbf{x}_{st},\mathbf{E}^{lm})_{s^*t^*}$$

$$= D_{\mathbf{x}_{st}}(\mathbf{G}).$$

Hence, $D_{y_{st}}(\mathbf{V}) = D_{x_{st}}(\mathbf{VQ}) \le D_{\mathbf{x}_{st}}(\mathbf{U})$, with equality, if $\mathbf{V} = \mathbf{PJUQ}'$ for any $\mathbf{P}\in\mathcal{P}$ and $\mathbf{J}\in\mathcal{J}$. Thus, $W(F_{\mathbf{y}}) = \mathbf{PJUQ}'\mathbf{Q}\boldsymbol{\Sigma}^{-1/2}\mathbf{A}^{-1} = \mathbf{PJU}\boldsymbol{\Sigma}^{-1/2}\mathbf{A}^{-1}$ for any $\mathbf{P}\in\mathcal{P}$ and $\mathbf{J}\in\mathcal{J}$. □

For Theorem 10 we need the following lemma.

LEMMA 2. *Assume that* $\hat{\mathbf{S}}_k$, $k = 1,\dots,K$ *are* $p\times p$ *matrices such that* $\sqrt{n}(\hat{\mathbf{S}}_k - \boldsymbol{\Lambda}_k)$ *are asymptotically normal with mean zero and* $\boldsymbol{\Lambda}_k = \mathrm{diag}(\lambda_{k1},\dots,\lambda_{kp})$.

Let $\hat{\mathbf{U}} = (\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_p)$ *be the orthogonal matrix that maximizes*

$$\sum_{k=1}^{K} \|\mathrm{diag}(\hat{\mathbf{U}}'\hat{\mathbf{S}}_k\hat{\mathbf{U}})\|^2.$$

Then

$$\sqrt{n}\hat{u}_{ij} = \frac{\sum_{k=1}^{K}(\lambda_{ki} - \lambda_{kj})\sqrt{n}(\hat{\mathbf{S}}_k)_{ij}}{\sum_{k=1}^{K}(\lambda_{ki} - \lambda_{kj})^2} + o_P(1).$$

PROOF. The proof is similar to the proof of Theorem 4.1 of Miettinen et al. (2014b). □

PROOF OF THEOREM 10. As the criterium functions

$$D_n(\mathbf{U}) = \sum_{j=1}^{p}\sum_{j=1}^{p} \|\mathrm{diag}(\mathbf{U}\hat{\mathbf{C}}^{ij}\mathbf{U}')\|^2 \quad \text{and}$$

$$D(\mathbf{U}) = \sum_{j=1}^{p}\sum_{j=1}^{p} \|\mathrm{diag}(\mathbf{U}\mathbf{C}^{ij}\mathbf{U}')\|^2$$

are continuous and $D_n(\mathbf{U}) \to_P D(\mathbf{U})$ for all $\mathbf{U}$, then, due to the compactness of $\mathcal{U}$,

$$\sup_{\mathbf{U}\in\mathcal{U}} |D_n(\mathbf{U}) - D(\mathbf{U})| \to_P 0.$$

$D(\mathbf{U})$ attains its maximum at any $\mathbf{JP}$ where $\mathbf{J} \in \mathcal{J}$ and $\mathbf{P} \in \mathcal{P}$. This further implies that there is a sequence of maximizers that satisfy $\hat{\mathbf{U}} \to_P \mathbf{I}_p$, and therefore also $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{S}}^{-1/2} \to_P \mathbf{I}_p$.

Let $\tilde{\mathbf{Z}} = (\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_n) = (\mathbf{z}_1 - \bar{\mathbf{z}}, \dots, \mathbf{z}_n - \bar{\mathbf{z}})$ denote the centered sample, and write

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\tilde{\mathbf{Z}}) = n^{-1}\sum_{i=1}^{n}(\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i') \otimes (\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i').$$

As the eighth moments of $\mathbf{z}$ exist, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normal with the expected value zero, and $\boldsymbol{\beta}$ as in Theorem 7.

Consider first a general sample whitening matrix $\hat{\mathbf{V}}$ satisfying $\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p) = O_P(1)$. For the whitened data we obtain

$$\begin{aligned}\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\mathbf{V}}\tilde{\mathbf{Z}}) &= n^{-1}\sum_{i=1}^{n}(\hat{\mathbf{V}}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\hat{\mathbf{V}}') \otimes (\hat{\mathbf{V}}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\hat{\mathbf{V}}') \\ &= (\hat{\mathbf{V}} \otimes \hat{\mathbf{V}})\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}}' \otimes \hat{\mathbf{V}}'),\end{aligned}$$

and, further,

$$\begin{aligned}&\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\qquad + [(\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p) \otimes \mathbf{I}_p) + (\mathbf{I}_p \otimes \sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p))]\boldsymbol{\beta} \\ &\qquad + \boldsymbol{\beta}[(\sqrt{n}(\hat{\mathbf{V}}' - \mathbf{I}_p) \otimes \mathbf{I}_p) \\ &\qquad\quad + (\mathbf{I}_p \otimes \sqrt{n}(\hat{\mathbf{V}}' - \mathbf{I}_p))].\end{aligned}$$

Write next

$$\hat{\mathbf{B}}^{kl} = \mathbf{B}(\mathbf{E}^{kl}, \tilde{\mathbf{Z}}) = n^{-1}\sum_{i=1}^{n}(\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\mathbf{E}^{kl}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i') \quad \text{and}$$

$$\hat{\mathbf{T}}^{kl} = \mathrm{vec}(\hat{\mathbf{B}}^{kl}) = \hat{\boldsymbol{\beta}}\,\mathrm{vec}(\mathbf{E}^{kl}).$$

Then $\sqrt{n}(\hat{\mathbf{T}}^{kl} - \mathrm{vec}(\mathbf{B}^{kl}))$ is asymptotically normal with expected value zero and $\mathbf{B}^{kl}$ as given in Theorem 7. Also,

$$\sqrt{n}\hat{b}_{kl}^{kk} = \sqrt{n}(\hat{\mathbf{B}}^{kk})_{kl} = \sqrt{n}\hat{r}_{kl} + o_P(1).$$

Next, let

$$\tilde{\mathbf{B}}^{kl} = \mathbf{B}(\mathbf{E}^{kl}, \hat{\mathbf{V}}\tilde{\mathbf{Z}}) = n^{-1}\sum_{i=1}^{n}(\hat{\mathbf{V}}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\hat{\mathbf{V}}'\mathbf{E}^{kl}\hat{\mathbf{V}}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\hat{\mathbf{V}}')$$

and

$$\tilde{\mathbf{T}}^{kl} = \mathrm{vec}(\tilde{\mathbf{B}}^{kl}) = \tilde{\boldsymbol{\beta}}\,\mathrm{vec}(\mathbf{E}^{kl})$$

denote the standardized counterparts of $\hat{\mathbf{B}}^{kl}$ and $\hat{\mathbf{T}}^{kl}$, respectively. Then

$$\begin{aligned}&\sqrt{n}(\tilde{\mathbf{T}}^{kl} - \mathrm{vec}(\mathbf{B}^{kl})) \\ &\quad = \sqrt{n}(\hat{\mathbf{T}}^{kl} - \mathrm{vec}(\mathbf{B}^{kl})) \\ &\qquad + [(\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p) \otimes \mathbf{I}_p) \\ &\qquad\quad + (\mathbf{I}_p \otimes \sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p))]\,\mathrm{vec}(\mathbf{B}^{kl}) \\ &\qquad + \beta[(\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p) \otimes \mathbf{I}_p) \\ &\qquad\quad + (\mathbf{I}_p \otimes \sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p))]\,\mathrm{vec}(\mathbf{E}^{kl}).\end{aligned}$$

It turns out that for the asymptotics of $\hat{\mathbf{W}}$, we only need

$$\begin{aligned}&\sqrt{n}(\tilde{\mathbf{B}}^{kk} - \mathbf{B}^{kk})_{kl} \\ &\quad = \sqrt{n}(\hat{\mathbf{B}}^{kk} - \mathbf{B}^{kk})_{kl} + 3\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p)_{kl} \\ &\qquad + (\kappa_k + 3)\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p)_{lk}\end{aligned} \tag{6}$$

and

$$\begin{aligned}&\sqrt{n}(\tilde{\mathbf{B}}^{ll} - \mathbf{B}^{ll})_{kl} \\ &\quad = \sqrt{n}(\hat{\mathbf{B}}^{ll} - \mathbf{B}^{ll})_{kl} + 3\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p)_{lk} \\ &\qquad + (\kappa_l + 3)\sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p)_{kl}.\end{aligned} \tag{7}$$

Next, note that in the JADE procedure the matrices to be diagonalized are

$$\tilde{\mathbf{C}}^{kl} = \tilde{\mathbf{B}}^{kl} - \mathbf{E}^{kl} - \mathbf{E}^{lk} - \mathrm{tr}(\mathbf{E}^{kl})I_p, \quad k, l = 1, \dots, p.$$

As $\sqrt{n}(\mathrm{vec}(\hat{\mathbf{C}}^{kl}) - \mathrm{vec}(\mathbf{C}^{kl}))$ are asymptotically normal with mean zero and $\mathbf{C}^{kl} = 0$, for $k \neq l$, and $\mathbf{C}^{kk} = \kappa_k \mathbf{E}^{kk}$, then by Lemma 2, $\sqrt{n}u_{kl}$ reduces to

$$\begin{aligned}\sqrt{n}\hat{u}_{kl} &= \frac{\kappa_k\sqrt{n}\tilde{c}_{kl}^{kk} - \kappa_l\sqrt{n}\tilde{c}_{kl}^{ll}}{\kappa_k^2+\kappa_l^2} + o_P(1)\\ &= \frac{\kappa_k\sqrt{n}\tilde{b}_{kl}^{kk} - \kappa_l\sqrt{n}\tilde{b}_{kl}^{ll}}{\kappa_k^2+\kappa_l^2} + o_P(1),\end{aligned} \tag{8}$$

where $\tilde{c}_{kl}^{kk} = (\tilde{\mathbf{C}}^{kk})_{kl}$ and $\tilde{b}_{kl}^{kk} = (\tilde{\mathbf{B}}^{kk})_{kl}$. So, asymptotically, all the information is in the matrices $\tilde{\mathbf{B}}^{kk}$, $k = 1, \dots, p$. As $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{V}}$, where $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are the rotation matrix and the whitening matrix, respectively, we have that

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{W}} - \mathbf{I}_p) &= \sqrt{n}(\hat{\mathbf{U}}\hat{\mathbf{V}} - \mathbf{I}_p)\\ &= \sqrt{n}(\hat{\mathbf{U}} - \mathbf{I}_p) + \sqrt{n}(\hat{\mathbf{V}} - \mathbf{I}_p) + o_P(1).\end{aligned}$$

The asymptotics of the regular JADE unmixing matrix is then obtained with $\hat{\mathbf{V}} = \hat{\mathbf{S}}^{-1/2}$, where $\hat{\mathbf{S}}$ is the sample covariance matrix.

Notice first that

$$\sqrt{n}(\hat{\mathbf{S}}^{-1/2} - \mathbf{I}_p) = -1/2\sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p) + o_P(1).$$

Then substituting (6) and (7) into (8), we have that, for $k \neq l$,

$$\begin{aligned}&\sqrt{n}\hat{w}_{kl}\\ &\quad= \frac{\kappa_k\sqrt{n}\hat{r}_{kl} - \kappa_l\sqrt{n}\hat{r}_{lk} + (3\kappa_l - 3\kappa_k - \kappa_k^2)\sqrt{n}\hat{s}_{kl}}{\kappa_k^2+\kappa_l^2}\\ &\qquad + o_P(1).\end{aligned}$$

For the diagonal elements we have simply

$$\sqrt{n}\hat{w}_{kk} = -1/2(\sqrt{n}\hat{s}_{kk} - 1) + o_P(1). \qquad \square$$

ACKNOWLEDGMENTS

The authors wish to thank the editors and the referees for valuable comments and suggestions.

Research supported in part by the Academy of Finland (Grants 251965, 256291 and 268703).

REFERENCES

BONHOMME, S. and ROBIN, J.-M. (2009). Consistent noisy independent component analysis. *J. Econometrics* **149** 12–25. MR2515042

BRYS, G., HUBERT, M. and STRUYF, A. (2006). Robust measures of tail weight. *Comput. Statist. Data Anal.* **50** 733–759. MR2207005

BUGRIEN, J. B. and KENT, J. T. (2005). Independent component analysis: An approach to clustering. In *Proceedings in Quantitative Biology, Shape Analysis and Wavelets* (S. Barber, P. D. Baxter, K. V. Mardia and R. E. Walls, eds.) 111–114. Leeds Univ. Press, Leeds, UK.

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

CARDOSO, J. F. and SOULOUMIAC, A. (1993). Blind beamforming for non Gaussian signals. *IEE Proc. F* **140** 362–370.

CAUSSINUS, H. and RUIZ-GAZEN, A. (1993). Projection pursuit and generalized principal component analyses. In *New Directions in Statistical Data Analysis and Robustness* (*Ascona*, 1992). *Monte Verità* 35–46. Birkhäuser, Basel. MR1280272

CHEN, A. and BICKEL, P. J. (2006). Efficient independent component analysis. *Ann. Statist.* **34** 2825–2855. MR2329469

CLARKSON, D. B. (1988). A least squares version of algorithm AS 211: The F-G diagonalization algorithm. *Appl. Stat.* **37** 317–321.

CRITCHLEY, F., PIRES, A. and AMADO, C. (2006). Principal axis analysis. Technical Report 06/14, The Open Univ., Milton Keynes, UK.

DARLINGTON, R. B. (1970). Is kurtosis really "peakedness?" *Amer. Statist.* **24** 19–22.

DECARLO, L. T. (1997). On the meaning and use of kurtosis. *Psychol. Methods* **2** 292–307.

ERIKSSON, J. and KOIVUNEN, V. (2004). Identifiability, separability and uniqueness of linear ICA models. *IEEE Signal Process. Lett.* **11** 601–604.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput. C* **23** 881–890.

HALLIN, M. and MEHTA, C. (2015). R-estimation for asymmetric independent component analysis. *J. Amer. Statist. Assoc.* **110** 218–232. MR3338498

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–525. MR0790553

HYVÄRINEN, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10** 626–634.

HYVÄRINEN, A., KARHUNEN, J. and OJA, E. (2001). *Independent Component Analysis*. Wiley, New York.

HYVÄRINEN, A. and OJA, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9** 1483–1492.

ILMONEN, P., NEVALAINEN, J. and OJA, H. (2010). Characteristics of multivariate distributions and the invariant coordinate system. *Statist. Probab. Lett.* **80** 1844–1853. MR2734250

ILMONEN, P. and PAINDAVEINE, D. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *Ann. Statist.* **39** 2448–2476. MR2906874

JONES, M. C. and SIBSON, R. (1987). What is projection pursuit? *J. Roy. Statist. Soc. Ser. A* **150** 1–36. MR0887823

KANKAINEN, A., TASKINEN, S. and OJA, H. (2007). Tests of multinormality based on location vectors and scatter matrices. *Stat. Methods Appl.* **16** 357–379. MR2399867

KARVANEN, J. and KOIVUNEN, V. (2002). Blind separation methods based on pearson system and its extensions. *Signal Process*. **82** 663–673.

KOLDOVSKÝ, Z., TICHAVSKÝ, P. and OJA, E. (2006). Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér–Rao lower bound. *IEEE Trans*. *Neural Netw*. **17** 1265–1277.

KOLLO, T. (2008). Multivariate skewness and kurtosis measures with an application in ICA. *J. Multivariate Anal*. **99** 2328–2338. MR2463392

KOLLO, T. and SRIVASTAVA, M. S. (2004). Estimation and testing of parameters in multivariate Laplace distribution. *Comm*. *Statist*. *Theory Methods* **33** 2363–2387. MR2104118

MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57** 519–530. MR0397994

MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann*. *Statist*. **4** 51–67. MR0388656

MIETTINEN, J., NORDHAUSEN, K., OJA, H. and TASKINEN, S. (2013). Fast equivariant JADE. In *Proc*. 38*th IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing* (*ICASSP* 2013) 6153–6157. Vancouver, BC.

MIETTINEN, J., NORDHAUSEN, K., OJA, H. and TASKINEN, S. (2014a). Deflation-based FastICA with adaptive choices of nonlinearities. *IEEE Trans*. *Signal Process*. **62** 5716–5724. MR3273526

MIETTINEN, J., ILLNER, K., NORDHAUSEN, K., OJA, H., TASKINEN, S. and THEIS, F. J. (2014b). Separation of uncorrelated stationary time series using autocovariance matrices. Available at arXiv:1405.3388.

MÓRI, T. F., ROHATGI, V. K. and SZÉKELY, G. J. (1993). On multivariate skewness and kurtosis. *Theory Probab. Appl*. **38** 547–551.

NORDHAUSEN, K., OJA, H. and OLLILA, E. (2011). Multivariate models and the first four moments. In *Nonparametric Statistics and Mixture Models* 267–287. World Scientific, Singapore. MR2838731

NORDHAUSEN, K., ILMONEN, P., MANDAL, A., OJA, H. and OLLILA, E. (2011). Deflation-based FastICA reloaded. In *Proc*. 19*th European Signal Processing Conference* 2011 (*EUSIPCO* 2011) 1854–1858. World Scientific, Singapore.

OJA, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scand*. *J*. *Stat*. **8** 154–168. MR0633040

OJA, H., SIRKIÄ, S. and ERIKSSON, J. (2006). Scatter matrices and independent component analysis. *Aust*. *J*. *Stat*. **35** 175–189.

OLLILA, E. (2010). The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE Trans*. *Signal Process*. **58** 1527–1541. MR2758026

PEARSON, K. (1895). Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material. *Philos*. *Trans*. *R*. *Soc*. **186** 343–414.

PEARSON, K. (1905). Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder. *Biometrika* **4** 169–212.

PEÑA, D. and PRIETO, F. J. (2001). Cluster identification using projections. *J*. *Amer. Statist*. *Assoc*. **96** 1433–1445. MR1946588

PEÑA, D., PRIETO, F. J. and VILADOMAT, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *J*. *Multivariate Anal*. **101** 1995–2007. MR2671197

SAMWORTH, R. J. and YUAN, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *Ann*. *Statist*. **40** 2973–3002. MR3097966

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

TICHAVSKY, P., KOLDOVSKY, Z. and OJA, E. (2006). Performance analysis of the FastICA algorithm and Cramer–Rao bounds for linear independent component analysis. *IEEE Trans*. *Signal Process*. **54** 1189–1203.

TYLER, D. E., CRITCHLEY, F., DÜMBGEN, L. and OJA, H. (2009). Invariant co-ordinate selection. *J*. *R*. *Stat*. *Soc*. *Ser. B Stat*. *Methodol*. **71** 549–592. MR2749907

VAN ZWET, W. R. (1964). *Convex Transformations of Random Variables*. *Mathematical Centre Tracts* **7**. Mathematical Centre, Amsterdam.