# Independent Component Analysis with Errors by Least Squares Covariance fitting

Alwin Stegeman[1] and Ab Mooijaart[2]

November 6, 2008

**Abstract**

We consider a general Independent Component Analysis (ICA) model with non-Gaussian components with an unknown distribution, and Gaussian measurement errors. For the error covariance matrix, we consider the following structures: isotropic, diagonal, AR(1), and Markov simplex. We propose a two-step estimation procedure. In the first step, the unrotated components and loadings, and the error covariance matrix are estimated by least squares fitting of the model covariance matrix to the data covariance matrix. In the second step, the components are rotated to approximate independence by means of an ICA algorithm. The proposed method is a natural generalization of ICA with isotropic Gaussian errors in which the unrotated components are obtained via Principal Component Analysis. For diagonal error variance, the method is identical to using Factor Analysis to obtain the unrotated components. We assess the performance of the proposed method by means of Monte Carlo simulations.

*Keywords*: Independent Component Analysis, noisy ICA, Factor Analysis, Principal Component Analysis, Factor Rotation, Independent Factor Analysis.

[1] Corresponding author. The author is with the Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, phone: ++31 50 363 6193, fax: ++31 50 363 6304, email: *a.w.stegeman@rug.nl* The author is supported by the Dutch Organisation for Scientific Research (NWO), VENI grant 451-04-102.

[2] The author is with the Institute for Psychological Research (LU-IPR), Department of Psychology, Leiden University, P.O. Box 9555, 2300 RB, Leiden, the Netherlands, email: *mooijaart@fsw.leidenuniv.nl*

# 1 Introduction

The concept of Independent Component Analysis (ICA) has become popular since the 90s of the last century. The goal of ICA is to give a best summary of a set of observed variables in terms of a prespecified number of statistically independent (latent) components and the loadings of the observed variables on these independent components. In the ICA literature, the independent components are also referred to as factors or sources, and the matrix containing the loadings is also referred to as mixing matrix. Conceptually, ICA can be seen as a fine-tuning of the well-known Principal Component Analysis (PCA) method, in which the components are only assumed to be uncorrelated. Applications of ICA are found in various areas. For instance, signal processing (where it is known as Blind Source Separation), neuro-imaging, econometrics and telecommunications.

A crucial assumption in the ICA framework is that the underlying independent components have non-Gaussian distributions. The reason for this is that with normally distributed components uncorrelatedness and independence are equivalent, whereas for non-Gaussian components zero correlation does not imply independence. Hence, ICA with Gaussian components boils down to the PCA method, in which the components can be rotated without affecting the model fit. Contrary to PCA, ICA has, under the assumption of non-Gaussian components, a rotationally unique solution.

We consider the case where the distribution of the components is non-Gaussian but unknown. In this case, a large class of procedures to obtain an ICA solution uses a two-step approach. In the first step, the unrotated components and their loadings are determined by PCA. This is also called the prewhitening step. In the second step, a rotation is determined such that the rotated components are approximately statistically independent. To find such a rotation, the principle of the Central Limit Theorem in Statistics is used. A consequence of the Central Limit Theorem is that a linear combination of two independent non-Gaussian random variables is "more Gaussian" than the variables themselves. Hence, in order to find the underlying independent non-Gaussian components, we must look for linear combinations of the PCA components that are "maximally non-Gaussian". This requires the specification of a measure of non-Gaussianity. Much used measures are cumulants, of which kurtosis is a special case. Comon (1994a) shows that using cumulants as a measure of non-Gaussianity results in the minimization of the mutual information in the joint pdf of the components and their marginal pdfs. Within the class of two-step ICA procedures described above, ICA is a fine-tuning of PCA also algebraically. For a theoretical treatment of ICA, see Comon (1994a). For an introduction to ICA we refer to De Lathauwer, De Moor, and Vandewalle (2000) and Comon and Chevalier (2000). See Hyvärinen, Karhunen, and Oja (2001) for detailed discussions and practical applications of ICA.

When measurement errors are considered in the ICA framework, they are usually assumed to be Gaussian and independent of the components. Throughout the paper, we will work under these assumptions. In the ICA literature, errors are also referred to as noise, and ICA with errors is also known as noisy ICA. In the signal processing literature, early works explicitly taking into account errors include Le Bienvenu and Kopp (1983) and Le Cadre (1989). In both works, the observed signal is equal to the sum of the latent components and an error term. Bienvenu and Kopp (1983) assume a known error covariance matrix, while Le Cadre (1989) considers error modeling via an ARMA process. Signal processing applications featuring correlated errors include radar, underwater passive listening, and general signal detection via an array of sensors.

In the two-step procedure described above, measurement errors are usually only taken into account in the first step. For this, the error covariance matrix needs to be estimated. A commonly used assumption is that the error covariance matrix is proportional to a known matrix (e.g. the identity for isotropic noise). In this case, the prewhitening step can easily be modified; see e.g. Comon and Chevalier (2000). Finding a rotation to approximately independent components is then analogous to the error-free situation.

For an unknown but diagonal error covariance, Ikeda and Toyama (2000) propose to use Factor Analysis (FA) in the prewhitening step instead of PCA. Here, the unrotated components and loadings, and the error covariances are determined such that the observed covariances are best approximated by the covariances of the model. This approximation can be performed in the least squares sense (the Minres approach of Harman and Jones, 1966), but also other criteria have been developed. For instance, the Maximum Likelihood approach of Jöreskog (1977) and Minimum Rank FA of Ten Berge and Kiers (1991).

For a general unknown error covariance, Attias (1999) proposes a Maximum Likelihood framework under the assumption that the components have Gaussian mixtures as distributions. The ICA components, loadings, component density parameters and the error covariance are estimated using an expectation-maximization algorithm. This approach has, however, some serious drawbacks. Ikeda and Toyama (2000) mention the following. It is not clear how to choose the number of Gaussians for modeling the component distributions. Also, the estimation procedure has large computational costs and exhibits slow convergence unless the number of components and the number of Gaussians are small. Finally, it is not clear how to choose the number of Gaussians to model the component distributions.

In this paper, we consider a two-step procedure for estimating an ICA model with measurement errors, where the latter follow a known covariance model. We assume non-Gaussian components with an unknown distribution, Gaussian errors, and independence between components and errors. The first step of our estimation procedure yields the unrotated components and loadings, and the error covariance by least squares fitting of the model covariance to the observed covariance. In the second step, a rotation to approximately independent components is found using standard ICA algorithms. In particular, we consider least squares fitting of the fourth-order cumulants in the second step. The first step of our procedure is analogous to Minres FA, and equivalent to FA for diagonal error covariance. For isotropic errors, our procedure is identical to using PCA to obtain the unrotated components.

The paper is organized as follows. In Section 2 we present and discuss PCA and ICA with measurement errors, and their estimation procedures. In Section 3, we present our estimation procedure for ICA with errors, and discuss models for the error covariance matrix: diagonal, autoregressive error process of order 1, and Markov simplex errors. In Section 4 we conduct a Monte Carlo simulation study to assess the performance of our method using several different error covariance models, and several ICA algorithms. Finally, Section 5 contains a discussion of our findings.

# 2 Model descriptions

## 2.1 PCA with measurement errors

Principal Component Analysis (PCA) is an exploratory data analysis tool for extracting uncorrelated components and their loadings from an $n \times m$ data matrix $\mathbf{X}$ containing $n$ realizations of $m$ variables. The matrix PCA model is

$$\mathbf{X} = \mathbf{A}\,\mathbf{B}^T + \mathbf{E}\,, \tag{2.1}$$

where the columns of $\mathbf{A}$ ($n \times R$) are the latent components, the columns of $\mathbf{B}$ ($m \times R$) contain the loadings of the $m$ variables on the $R$ components and $\mathbf{E}$ is a residual term, see Pearson (1901). It is usually assumed that $m \leq n$, the columns of $\mathbf{X}$ (variables) have mean zero and unit variance, the columns of $\mathbf{A}$ (components) have mean zero and unit variance and are uncorrelated, i.e. $n^{-1}\mathbf{A}^T\mathbf{A} = \mathbf{I}_R$. A PCA solution is found by minimizing the sum-of-squares of $\mathbf{E}$ and can be obtained from the truncated Singular Value Decomposition (SVD) of $\mathbf{X}$, see Eckart and Young (1936).

Note that in the PCA solution the first component explains the most variance in the data, the second component explains the most variance in the data after the first component has been subtracted, etcetera. This is due to the ordering of the singular values and the orthogonality of the left- and right singular vectors.

The PCA solution is unique (up to sign changes) in this ordering if the first $R$ singular values are distinct. However, for any $R \times R$ orthogonal matrix $\mathbf{Q}$, the solution $(\mathbf{AQ}, \mathbf{BQ})$ has the same residuals $\mathbf{E}$, since $\mathbf{A}\,\mathbf{B}^T = (\mathbf{AQ})\,(\mathbf{BQ})^T$. Hence, only the space of the PCA components $\mathbf{A}$ is uniquely determined (ignoring the ordering). Within this space, any set of basis vectors can be taken as components. In psychological applications of PCA, the ordering of the components is usually not meaningful and a rotation $\mathbf{Q}$ is calculated which yields interpretable components, see Browne (2001).

The introduction of measurement errors in the PCA model (2.1) is done by interpreting it stochastically. The data matrix $\mathbf{X}$ is interpreted as consisting of $n$ samples or observations of $m$ random variables. We denote the $m \times 1$ vector of these $m$ random variables as $\mathbf{x}$. Analogously, the component matrix $\mathbf{A}$ is interpreted as consisting of $n$ samples of $R$ latent random variables. We denote the $R \times 1$ vector of these $R$ latent random variables as $\mathbf{a}$. The stochastic model is

$$\mathbf{x} = \mathbf{B}\,\mathbf{a} + \mathbf{e}\,, \tag{2.2}$$

where $\mathbf{B}$ contains the loadings and $\mathbf{e}$ denotes the $m \times 1$ vector of measurement errors.

The random vector $\mathbf{a}$ is assumed to have a Gaussian distribution with mean zero and variance $\mathbf{I}_R$, i.e. $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_R)$. Also, it is assumed that $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$, $\mathbf{a}$ and $\mathbf{e}$ are independent and rank($\mathbf{B}$) $= R$. The goal is to estimate $\mathbf{B}$, $\sigma^2$ and $n$ realisations of the random vector $\mathbf{a}$, which are the rows of $\mathbf{A}$ ($n \times R$).

The Maximum Likelihood estimators of $\mathbf{A}$, $\mathbf{B}$ and $\sigma^2$ are given by (e.g. Tipping and Bishop, 1999; Anderson, 2003, Section 11.3)

$$\widehat{\mathbf{A}}_{\mathrm{ML}} = \mathbf{U}_R\,\mathbf{S}_R\,(n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{-1/2}\,\mathbf{Q}\,, \tag{2.3}$$

$$\widehat{\mathbf{B}}_{\mathrm{ML}} = \mathbf{V}_R\,(n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{1/2}\,\mathbf{Q}\,, \tag{2.4}$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{(m-R)\,n}\sum_{j=R+1}^{m} s_{jj}^2\,, \tag{2.5}$$

where $\mathbf{Q}$ is any $R \times R$ orthogonal rotation matrix, $\mathbf{U}_R\,\mathbf{S}_R\,\mathbf{V}_R^T$ is the truncated SVD of $\mathbf{X}$, and $s_{jj}^2$ is the $j$-th diagonal element of $\mathbf{S}^2$ (the SVD of $\mathbf{X}$ is denoted as $\mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$). In the absence of measurement errors, i.e. if $\sigma^2 = 0$, the estimates (2.3) and (2.4) reduce to the ordinary PCA solution. The unrotated PCA solution is obtained by setting $\sigma^2 = 0$ and $\mathbf{Q} = \mathbf{I}_R$. Let the latter be denoted by ($\widehat{\mathbf{A}}_{\mathrm{PCA}}$,$\widehat{\mathbf{B}}_{\mathrm{PCA}}$). Note that $\widehat{\mathbf{A}}_{\mathrm{PCA}}\widehat{\mathbf{B}}_{\mathrm{PCA}}^T = \widehat{\mathbf{A}}_{\mathrm{PCA}}\mathbf{Q}(\widehat{\mathbf{B}}_{\mathrm{PCA}}\mathbf{Q})^T = \widehat{\mathbf{A}}_{\mathrm{ML}}\widehat{\mathbf{B}}_{\mathrm{ML}}^T$. Hence, the unrotated PCA solution, the rotated PCA solution, and the ML estimates in the presence of uncorrelated and homoscedastic measurement errors, all minimize the least squares distance between the data $\mathbf{X}$ and the model part $\mathbf{AB}^T$.

The model (2.2) and the assumptions above imply that

$$\mathrm{Cov}(\mathbf{x}) = E(\mathbf{x}\mathbf{x}^T) = \mathbf{B}\,E(\mathbf{a}\mathbf{a}^T)\,\mathbf{B}^T + E(\mathbf{e}\mathbf{e}^T) = \mathbf{B}\mathbf{B}^T + \sigma^2\,\mathbf{I}_m\,. \tag{2.6}$$

Replacing $E(\mathbf{x}\mathbf{x}^T)$ by its estimator $n^{-1}\mathbf{X}^T\mathbf{X}$ and using the SVD of $\mathbf{X}$ it follows from (2.6) that

$$\mathbf{B}\mathbf{B}^T = \mathbf{V}\,(n^{-1}\mathbf{S}^2 - \sigma^2\,\mathbf{I}_m)\,\mathbf{V}^T\,. \tag{2.7}$$

Since $\mathbf{BB}^T$ is non-negative definite, rank($\mathbf{BB}^T$) = rank($\mathbf{B}$) $= R$ and (2.7) is an eigendecomposition of $\mathbf{BB}^T$, it follows that the last (and smallest) $m - R$ diagonal elements of ($n^{-1}\mathbf{S}^2 - \sigma^2\,\mathbf{I}_m$) are zero. By considering the eigenvalues of $n^{-1}\mathbf{X}^T\mathbf{X}$, this fact can be used to obtain a rough estimate of the number $R$ of components present.

## 2.2 ICA with measurement errors

Here, we discuss the introduction of measurment errors in ICA, where we consider the latter as a refinement of PCA. We have the same model as (2.2) with the same assumptions, except that the components $\mathbf{a}$ have unknown non-Gaussian distributions and are assumed to be statistically independent. For the errors, we still have $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$. A two-step estimation procedure is as follows. In the first step, the unrotated components and loadings, and the error variance are estimated as (2.3)-(2.5). This yields approximately uncorrelated components and fixes $\mathbf{A}$ and $\mathbf{B}$ up to an orthogonal rotation. The second step then finds a rotation $\mathbf{Q}$ such that the components are approximately statistically independent, using some measure of non-Gaussianity as explained in the Introduction.

Note that the estimates (2.3)-(2.5) are the Maximum Likelihood estimates for Gaussian components. However, they can still be used for the situation where the components are non-Gaussian. This is because they are also the least squares estimates, i.e. they minimize $\|\mathbf{X} - \mathbf{A}\mathbf{B}^T\|^2$, as explained in Section 2.1.

In the second step, the rotation $\mathbf{Q}$ is usually found by considering the prewhitened data $\widetilde{\mathbf{X}}^T = (n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{-1/2}\,\mathbf{V}_R^T\,\mathbf{X}^T$, which satisfies

$$\widetilde{\mathbf{X}}^T = \mathbf{Q}\,\mathbf{A}^T + \widetilde{\mathbf{E}}^T\,, \tag{2.8}$$

where $\widetilde{\mathbf{E}}^T = (n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{-1/2}\,\mathbf{V}_R^T\,\mathbf{E}^T$. Note that (2.8) follows from $\mathbf{X}^T = \mathbf{B}\,\mathbf{A}^T + \mathbf{E}^T$ and (2.4). In fact, it follows from (2.3) that $\widetilde{\mathbf{X}}$ is equal to the unrotated component estimates. Hence, (2.8) is equivalent to finding a rotation $\mathbf{Q}$ such that the rotated components are approximately independent.

The ICA model with the two-step procedure described above is used by Beckmann and Smith (2004) to extract brain activation patterns from fMRI measurements. As mentioned in the Introduction, when the error covariance is assumed to be proportional to a known matrix (here, the identity), the incorporation of it in the prewhitening step is standard practice in the ICA community.

The two-step ICA estimation procedure described above find an optimal rotation of the principal components according to the criterion of independent components. In psychological applications of PCA, several criteria and algorithms have been developed for finding appropriate rotations of the principal components. A comparison between these and the ICA criterion of independent components is made by Kano, Miyamoto and Shimizu (2003).

## 2.3 ICA algorithms

Next, we describe two different types of ICA algorithms to obtain an optimal rotation $\mathbf{Q}$ of the components (2.3). Let the prewhitened ICA model be given by $\tilde{\mathbf{x}} = \mathbf{Q}\,\mathbf{a} + \tilde{\mathbf{e}}$, see (2.8), with non-Gaussian components $\mathbf{a}$ and Gaussian errors $\tilde{\mathbf{e}}$. Then $\mathbf{a} = \mathbf{Q}^T\tilde{\mathbf{x}} - \mathbf{Q}^T\tilde{\mathbf{e}}$. The goal is to find $\mathbf{Q}$ such that $\mathbf{a}$ are approximately statistically independent.

### Diagonalizing Cumulants

The ICA algorithm of Comon (1994a) considers the $K$-th order cumulants of the components $\mathbf{a}$, which are given by

$$\mathrm{Cum}^{(K)}(a_{j_1},\ldots,a_{j_K}) = \sum(-1)^{k-1}(k-1)!\;E\left(\prod_{j\in S_1} a_j\right)\cdots E\left(\prod_{j\in S_k} a_j\right)\,, \tag{2.9}$$

where $a_{j_1},\ldots,a_{j_K}$ are elements of the random vector $\mathbf{a}$ and the summation involves all possible partitions $\{S_1,\ldots,S_k\}$, $1 \le k \le K$, of the integers $\{j_1,\ldots,j_K\}$. For $K \ge 3$, the cumulants of the components $\mathbf{a}$ are equal to the cumulants of $\mathbf{Q}^T\,\tilde{\mathbf{x}}$. This is due to the multilinearity of the cumulant function and the fact that for $K \geq 3$ the cumulants of Gaussian random variables are zero, see De Lathauwer et al. (2000). To obtain statistically independent components, the following property of cumulants is used. If the mean-zero components $\mathbf{a}$ are statistically independent, then $\mathrm{Cum}^{(K)}(a_{j_1}, \ldots, a_{j_K}) \neq 0$ only if $j_1 = j_2 = \ldots = j_K$, $K \geq 1$.

The algorithm of Comon (1994a) uses the diagonality of the $K$-way array of the $K$-th order cumulants of $\mathbf{a}$ as a measure of non-Gaussianity and independence. The algorithm makes the $K$-way array of (the sample estimates of) the $K$-th order cumulants of $\mathbf{Q}^T\,\tilde{\mathbf{x}}$ as diagonal as possible by varying the rotation $\mathbf{Q}$. A Matlab algorithm for $K = 4$ can be obtained from Comon (1994b). An alternative algorithm for $K = 4$ is called JADE and is due to Cardoso and Souloumiac (1993). It is available from Cardoso (2005). For $K = 4$, each diagonal element of the cumulant array is the kurtosis of one component. Therefore, the method for $K = 4$ is also referred to as "maximizing kurtosis".

For other ICA algorithms using cumulants see De Lathauwer et al. (2000) and the references therein. Comon (1994a) shows that approximate diagonalization of the cumulant $K$-way array of $\mathbf{a}$ is a natural approximation of the minimization of the mutual information of the joint pdf of $\mathbf{a}$ and its marginal distributions. In this paper we will denote the $K$-th order cumulant algorithm of Comon (1994a) as Comon-$K$.

Since $\mathbf{Q}$ is orthonormal, maximizing the sum-of-squares of the diagonal of the cumulant array is equivalent to minimizing the sum-of-squares of the off-diagonal elements of the cumulant array; see Comon (1994a). Moreover, both are equivalent to least squares fitting of the cumulants of $\mathbf{Q}\,\mathbf{a}$ to the cumulants of the prewhitened data $\tilde{\mathbf{x}}$. These results are well-known in the ICA community. For the sake of completeness, we give the proof for $K = 4$ in Proposition 2.1 below. Let $\otimes$ denote the Kronecker product, and let $\mathbf{V}$ be the $R^2 \times R$ matrix such that $(\mathbf{Q} \otimes \mathbf{Q})\,\mathbf{V} = [\mathbf{q}_1 \otimes \mathbf{q}_1 | \ldots | \mathbf{q}_R \otimes \mathbf{q}_R]$.

**Proposition 2.1** *Let $\mathbf{C}_4$ ($R^2 \times R^2$) contain the 4-way array of 4-th order cumulants of the prewhitened data $\tilde{\mathbf{x}}$ in matrix form. Let $\mathbf{Q}$ ($R \times R$) be orthonormal and $\mathbf{K}$ ($R \times R$) be diagonal (containing the kurtosis of the components). Consider the least squares fitting of 4-th order cumulants*

$$\min_{\mathbf{Q},\mathbf{K}} \; \mathrm{tr}\left[(\mathbf{C}_4 - (\mathbf{Q} \otimes \mathbf{Q})\,\mathbf{V}\,\mathbf{K}\,\mathbf{V}^T(\mathbf{Q} \otimes \mathbf{Q})^T)^2\right] . \tag{2.10}$$

*Then, for any $\mathbf{Q}$, the optimal $\mathbf{K}$ is given by*

$$\mathbf{K} = \mathrm{diag}\left[\mathbf{V}^T(\mathbf{Q} \otimes \mathbf{Q})^T\,\mathbf{C}_4\,(\mathbf{Q} \otimes \mathbf{Q})\,\mathbf{V}\right] . \tag{2.11}$$

*Moreover, the optimal $\mathbf{Q}$ can be found by minimizing the sum-of-squares of the off-diagonal of the transformed cumulant 4-way array, i.e.*

$$\min_{\mathbf{Q}} \; \mathrm{tr}\left[((\mathbf{Q} \otimes \mathbf{Q})^T\,\mathbf{C}_4\,(\mathbf{Q} \otimes \mathbf{Q}) - \mathbf{V}\,\mathrm{diag}(\mathbf{V}^T(\mathbf{Q} \otimes \mathbf{Q})^T\,\mathbf{C}_4\,(\mathbf{Q} \otimes \mathbf{Q})\,\mathbf{V})\,\mathbf{V}^T)^2\right] , \tag{2.12}$$

*or, equivalently, by maximizing the sum-of-squares of the diagonal of the transformed cumulant 4-way array, i.e.*

$$\max_{\mathbf{Q}} \; \mathrm{tr}\left[(\mathrm{diag}(\mathbf{V}^T(\mathbf{Q} \otimes \mathbf{Q})^T\,\mathbf{C}_4\,(\mathbf{Q} \otimes \mathbf{Q})\,\mathbf{V}))^2\right] . \tag{2.13}$$

**Proof.** Since $(\mathbf{Q}\otimes\mathbf{Q})^T(\mathbf{Q}\otimes\mathbf{Q})=\mathbf{I}_R$ and $\mathbf{V}^T\mathbf{V}=\mathbf{I}_R$, we have

$$\mathrm{tr}\left[(\mathbf{C}_4-(\mathbf{Q}\otimes\mathbf{Q})\,\mathbf{V}\,\mathbf{K}\,\mathbf{V}^T(\mathbf{Q}\otimes\mathbf{Q})^T)^2\right]=\mathrm{tr}\left[\mathbf{C}_4^2\right]-2\;\mathrm{tr}\left[\mathbf{V}^T(\mathbf{Q}\otimes\mathbf{Q})^T\,\mathbf{C}_4\,(\mathbf{Q}\otimes\mathbf{Q})\,\mathbf{V}\,\mathbf{K}\right]+\mathrm{tr}\left[\mathbf{K}^2\right]\;. \tag{2.14}$$

Setting the derivative of (2.14) with respect to $\mathbf{K}$ equal to zero yields that, for any $\mathbf{Q}$, the optimal $\mathbf{K}$ in (2.10) is given by (2.11). Rewriting (2.10) as

$$\mathrm{tr}\left[(\mathbf{Q}\otimes\mathbf{Q})^T\,\mathbf{C}_4\,(\mathbf{Q}\otimes\mathbf{Q})-\mathbf{V}\,\mathbf{K}\,\mathbf{V}^T\right]\;, \tag{2.15}$$

and substituting (2.11) for $\mathbf{K}$ yields (2.12). Note that the term $\mathbf{V}\,\mathbf{K}\,\mathbf{V}^T$ with $\mathbf{K}$ given by (2.11) is the matrix form of the transformed 4-way cumulant array in which only the diagonal elements, i.e. the kurtosis, are nonzero.

Next, we show the equivalence of (2.10) and (2.13). Let $\mathbf{F}=\mathbf{V}^T(\mathbf{Q}\otimes\mathbf{Q})^T\,\mathbf{C}_4\,(\mathbf{Q}\otimes\mathbf{Q})\,\mathbf{V}$. Substituting $\mathbf{K}=\mathrm{diag}(\mathbf{F})$ in the right-hand side of (2.14) yields

$$\mathrm{tr}\left[\mathbf{C}_4^2\right]-2\;\mathrm{tr}\left[\mathbf{F}\,\mathrm{diag}(\mathbf{F})\right]+\mathrm{tr}\left[(\mathrm{diag}(\mathbf{F}))^2\right]=\mathrm{tr}\left[\mathbf{C}_4^2\right]-\mathrm{tr}\left[(\mathrm{diag}(\mathbf{F}))^2\right]\;. \tag{2.16}$$

Minimizing the right-hand side of (2.16) over $\mathbf{Q}$ yields (2.13). The diagonal of the transformed 4-way cumulant array is given by $\mathbf{V}\,\mathrm{diag}(\mathbf{F})\,\mathbf{V}^T$, and we have

$$\mathrm{tr}\left[(\mathbf{V}\,\mathrm{diag}(\mathbf{F})\,\mathbf{V}^T)^2\right]=\mathrm{tr}\left[(\mathrm{diag}(\mathbf{F}))^2\right]\;. \tag{2.17}$$

This completes the proof. □

### Approximate maximization of negentropy

Negentropy is used as a measure of "non-Gaussianity" in information theory and statistics. It is negatively related to mutual information, see Hyvärinen et al. (2001). Hence, the maximization of negentropy implies the minimization of mutual information. The following approximation of negentropy has been proposed by Hyvärinen (1999) and reads as

$$[E(G(a))-E(G(z))]^2\;, \tag{2.18}$$

where $a$ is one component of $\mathbf{a}$, $z$ is a standard Gaussian variable and $G$ is some non-quadratic function. Clearly, (2.18) is a measure of non-Gaussianity. The algorithm maximizing the sample estimate of the objective function (2.18) over $\mathbf{Q}$ is known as FastICA (Hyvärinen & Oja, 2000; Hyvärinen, 2005). FastICA uses approximative Newton iterations and can be used for each component separately or for all components simultaneously. For remarks on the influence of the choice of $G$ on the performance of the FastICA algorithm, see Hyvärinen (1999). In Hyvärinen and Oja (2000) it is shown that the objective function (2.18) is indeed an approximation of negentropy. See also Hyvärinen et al. (2001).

# 3 ICA with measurement errors by least squares covariance fitting

Here, we discuss a general ICA model with measurement errors. We have the same assumptions as in Section 2.2 except that the errors may be correlated and their variances are allowed to be different, i.e. $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$, where $\mathbf{\Psi}$ is the error covariance matrix. The estimation procedure consists of two steps. In the first step, the unrotated components and loadings, and the error covariance $\mathbf{\Psi}$ are determined by least squares fitting of the model covariance to the observed covariance. In the second step, an ICA algorithm determines a rotation such that the rotated components are approximately independent. When $\mathbf{\Psi} = \sigma^2 \, \mathbf{I}_m$ this yields the same estimation procedure as described in Section 2.2. This will be shown below. When $\mathbf{\Psi}$ is diagonal this is the approach of Ikeda and Toyama (2000) with Minres FA in the first step of the estimation procedure. In case the error covariance $\mathbf{\Psi}$ is not diagonal, we assume a parametric covariance model to ensure identifiability. Error covariance models are discussed in Section 3.2 below.

## 3.1 Estimation procedure

Here, we explain the estimation procedure in more detail. We have the model (2.2) with non-Gaussian components $\mathbf{a}$ with an unknown distribution, and errors $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$. We assume $\mathbf{a}$ and $\mathbf{e}$ to be independent. Analogous to (2.6), we have

$$\mathrm{Cov}(\mathbf{x}) = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi} \, . \tag{3.1}$$

Let $\mathbf{C} = n^{-1}\mathbf{X}^T\mathbf{X}$ be the estimate of the data covariance matrix. In the first step, we minimize the function

$$f(\mathbf{B}, \mathbf{\Psi}) = \mathrm{tr}\left[(\mathbf{C} - \mathbf{B}\mathbf{B}^T - \mathbf{\Psi})^2\right] \, , \tag{3.2}$$

over $\mathbf{B}$ and $\mathbf{\Psi}$. For given $\mathbf{\Psi}$ and eigendecomposition $\mathbf{C} - \mathbf{\Psi} = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^T$, where the eigenvalues in $\mathbf{D}$ are in decreasing order, we obtain

$$\widehat{\mathbf{B}}_{LS} = \mathbf{V}_R\,\mathbf{D}_R^{1/2} \, , \tag{3.3}$$

where $\mathbf{V}_R$ contains the first $R$ columns of $\mathbf{V}$, and $\mathbf{D}_R$ contains the first $R$ eigenvalues of $\mathbf{D}$. The estimate of $\mathbf{\Psi}$ is then obtained by minimizing

$$\begin{aligned} g(\mathbf{\Psi}) &= \mathrm{tr}\left[(\mathbf{C} - \mathbf{V}_R\,\mathbf{D}_R\,\mathbf{V}_R^T - \mathbf{\Psi})^2\right] \\ &= \mathrm{tr}\left[\mathbf{V}\,\mathbf{D}\,\mathbf{V}^T - \mathbf{V}_R\,\mathbf{D}_R\,\mathbf{V}_R^T)^2\right] \\ &= \mathrm{tr}\left[(\mathbf{V}_{m-R}\,\mathbf{D}_{m-R}\,\mathbf{V}_{m-R}^T)^2\right] \\ &= \mathrm{tr}\left[\mathbf{D}_{m-R}^2\right] \\ &= d_{R+1}^2 + \cdots + d_m^2 \, , \end{aligned} \tag{3.4}$$

where $\mathbf{V}_{m-R}$ contains the last $m-R$ columns of $\mathbf{V}$, $\mathbf{D}_{m-R}$ contains the last $m-R$ eigenvalues of $\mathbf{D}$, and $d_j$ is the $j$-th diagonal element of $\mathbf{D}$. Hence, the function value of $g(\mathbf{\Psi})$ can be computed as the sum-of-squares of the $m-R$ smallest eigenvalues of $\mathbf{C} - \mathbf{\Psi}$.

For an optimization method using gradients, we have

$$\frac{\partial g}{\partial \psi_{ii}} \;=\; 2\,\left(\psi_{ii} + \sum_{l=1}^{R} d_l\, v_{il}^2 \;-\; c_{ii}\right)\,, \tag{3.5}$$

$$\frac{\partial g}{\partial \psi_{ij}} \;=\; 4\,\left(\psi_{ij} + \sum_{l=1}^{R} d_l\, v_{il}\, v_{jl} \;-\; c_{ij}\right)\,. \tag{3.6}$$

From $\mathbf{X} = \mathbf{A}\,\mathbf{B}^T + \mathbf{E}$ it follows that, for given $\mathbf{B}$, the least squares estimate of the components is given by

$$\widehat{\mathbf{A}}_{LS} = \mathbf{X}\,\mathbf{B}\,(\mathbf{B}^T\mathbf{B})^{-1}\,. \tag{3.7}$$

This equality also holds for the unrotated PCA, rotated PCA, and ML estimates (2.3)-(2.4). As before, the least squares estimates of $\mathbf{A}$ and $\mathbf{B}$ can be rotated without affecting the model fit. The rotation $\mathbf{Q}$ is determined in the second step of the estimation procedure. As in Section 2.2, an ICA algorithm is used to find a $\mathbf{Q}$ such that the rotated components are approximately independent.

Below, we summarize the steps of our estimation method.

ICA WITH ERRORS BY LEAST SQUARES COVARIANCE FITTING

Input: data matrix $\mathbf{X}$ $(n \times m)$, number of components $R$.
Output: estimates of (rotated) components $\widehat{\mathbf{A}}_{\mathrm{LS}}\mathbf{Q}$, (rotated) loadings $\widehat{\mathbf{B}}_{\mathrm{LS}}\mathbf{Q}$, and error covariance $\widehat{\boldsymbol{\Psi}}_{\mathrm{LS}}$.

1. Estimate the data covariance as $\mathbf{C} = n^{-1}\mathbf{X}^T\mathbf{X}$.
2. Minimize tr $\left[(\mathbf{C} - \mathbf{B}\mathbf{B}^T - \boldsymbol{\Psi})^2\right]$ over $\mathbf{B}$ and $\boldsymbol{\Psi}$. This yields the least squares estimates $\widehat{\mathbf{B}}_{\mathrm{LS}}$ and $\widehat{\boldsymbol{\Psi}}_{\mathrm{LS}}$.
3. Compute the unrotated components as $\widehat{\mathbf{A}}_{LS} = \mathbf{X}\,\mathbf{B}\,(\mathbf{B}^T\mathbf{B})^{-1}$, where $\mathbf{B} = \widehat{\mathbf{B}}_{\mathrm{LS}}$.
4. Using an ICA algorithm, determine a rotation $\mathbf{Q}$ such that the rotated components $\widehat{\mathbf{A}}_{\mathrm{LS}}\mathbf{Q}$ are approximately independent.

The least squares estimates $\widehat{\mathbf{A}}_{LS}$ and $\widehat{\mathbf{B}}_{LS}$ do not have a closed form representation unless $\boldsymbol{\Psi} = \sigma^2\,\mathbf{I}_m$. This also holds for the prewhitened data $\widetilde{\mathbf{X}}$, which we define as $\widetilde{\mathbf{X}}^T = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{X}^T$ with $\mathbf{B} = \widehat{\mathbf{B}}_{LS}$. Note that the prewhitened data are defined in the same way as in Section 2.2. Again, the prewhitened data $\widetilde{\mathbf{X}}$ are equal to the estimate of the unrotated components.

Note that by assuming $\boldsymbol{\Psi} \neq \sigma^2\,\mathbf{I}_m$, the number of components in the ICA model cannot be estimated by considering the eigenvalues of the sample covariance matrix $n^{-1}\mathbf{X}^T\mathbf{X}$, while this is possible for $\boldsymbol{\Psi} = \sigma^2\,\mathbf{I}_m$. For the general ICA model, information criteria such as the Akaike information criterion or minimum description length may be used to estimate the number of components, see Ikeda and Toyama (2000) for diagonal $\boldsymbol{\Psi}$.

As a final result in this section, we show that if $\mathbf{\Psi} = \sigma^2 \mathbf{I}_m$, then the first step of the estimation procedure yields the ML estimates (2.3)-(2.5) with $\mathbf{Q} = \mathbf{I}_R$. Hence, in this case the minimization of $\|\mathbf{X} - \mathbf{A}\mathbf{B}^T\|^2$, i.e. the distance between the data and the model, produces the same results as the minimization of $\|n^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{B}\mathbf{B}^T - \mathbf{\Psi}\|^2$, i.e. the distance between the data covariance and the model covariance. Note that $\mathbf{A} = \mathbf{X}\,\mathbf{B}\,(\mathbf{B}^T\mathbf{B})^{-1}$ also holds for (2.3)-(2.4).

**Proposition 3.1** *The expressions* (2.4)-(2.5) *for* $\mathbf{B}$ *and* $\sigma^2$ *minimize*

$$f(\mathbf{B}, \sigma^2) = \mathrm{tr}\left[(n^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{B}\mathbf{B}^T - \sigma^2\,\mathbf{I}_m)^2\right] .$$

**Proof.** See Anderson (2003), Section 14.3. □

## 3.2 Error covariance models

Here, we discuss the error covariance models we consider. For diagonal error covariance matrix $\mathbf{\Psi}$, we discuss two known identifiability results (in the least squares sense) of $\mathbf{\Psi}$ when fitting the model covariance to the observed covariance using least squares. Next, we discuss the cases where the errors follow an AR(1) process or a Markov simplex process.

### Diagonal $\mathbf{\Psi}$

For diagonal $\mathbf{\Psi}$, the first step of the estimation procedure is exactly the Minres FA procedure. Regarding the (global) identifiability of a diagonal $\mathbf{\Psi}$ in (3.1), we state the following two known results. It is assumed that the loadings matrix $\mathbf{B}$ has full column rank $R$.

**Theorem 3.2** *(Anderson and Rubin, 1956) Consider the factor analysis model* (3.1) *with diagonal error covariance matrix. Suppose there are two solutions, i.e.*

$$\mathbf{B}_1\,\mathbf{B}_1^T + \mathbf{\Psi}_1 = \mathbf{B}_2\,\mathbf{B}_2^T + \mathbf{\Psi}_2\,. \tag{3.8}$$

*Suppose that if any row of* $\mathbf{B}_1$ *is deleted there remain two disjoint submatrices of rank* $R$*. Then* $\mathbf{\Psi}_1 = \mathbf{\Psi}_2$ *and* $\mathbf{B}_1 = \mathbf{B}_2\,\mathbf{Q}$ *for some orthogonal rotation* $\mathbf{Q}$*.* □

**Theorem 3.3** *(Bekker and Ten Berge, 1997) Consider the factor analysis model* (3.1) *with diagonal error covariance matrix. Suppose there are two solutions, i.e.*

$$\mathbf{B}_1\,\mathbf{B}_1^T + \mathbf{\Psi}_1 = \mathbf{B}_2\,\mathbf{B}_2^T + \mathbf{\Psi}_2\,. \tag{3.9}$$

*If*

$$R \;<\; \frac{2m + 1 - (8m+1)^{1/2}}{2}\,, \tag{3.10}$$

*then the set of* $(\mathbf{B}_1, \mathbf{\Psi}_1)$ *for which* (3.9) *holds with* $\mathbf{\Psi}_1 \neq \mathbf{\Psi}_2$ *has Lebesgue measure zero. That is, we have* $\mathbf{\Psi}_1 = \mathbf{\Psi}_2$ *and* $\mathbf{B}_1 = \mathbf{B}_2\,\mathbf{Q}$ *for some orthogonal rotation* $\mathbf{Q}$ *almost everywhere.* □

The condition of Theorem 3.2 implies that $2R + 1 \leq m$. Theorem 3.3 shows that a diagonal $\mathbf{\Psi}$ is identified almost surely if $R$ is below the so-called Ledermann bound (3.10) (Ledermann, 1937). For $m \geq 4$ the latter is less restrictive than $2R+1 \leq m$. Under $2R+1 \leq m$, however, Theorem 3.3 follows from Theorem 3.2 (without the almost everywhere restriction).

Note that $\mathbf{B}_1 \mathbf{B}_1^T = \mathbf{B}_2 \mathbf{B}_2^T$ implies $\mathbf{B}_1 = \mathbf{B}_2 \mathbf{Q}$ for some orthogonal rotation $\mathbf{Q}$.

In the proofs of Theorems 3.2 and 3.3 it is not assumed that $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ are positive semi-definite. Therefore, for diagonal $\mathbf{\Psi}$, Theorems 3.2 and 3.3 can be used as uniqueness conditions for Minres solutions to (3.1).

### Two covariance models for $\mathbf{\Psi}$

We consider two parametric models for a non-diagonal $\mathbf{\Psi}$, namely the autoregressive model of order 1 and the Markov simplex model. For AR(1), we have

$$\psi_{ij} = \frac{\sigma^2}{1-\rho^2} \, \rho^{|i-j|} \,, \tag{3.11}$$

with $|\rho| < 1$. For the Markov simplex model, we have

$$\psi_{ij} = \sigma^2 \, \rho^{|i-j|} \,, \tag{3.12}$$

with $|\rho| < 1$. The gradients of the parameters in (3.11) and (3.12) can be obtained as

$$\frac{\partial g}{\partial \mathbf{q}^T} = \frac{\partial g}{\partial \mathrm{vech}(\mathbf{\Psi})^T} \, \frac{\partial \mathrm{vech}(\mathbf{\Psi})}{\partial \mathbf{q}^T} \,, \tag{3.13}$$

where vech denotes the vector of non-duplicate elements of a symmetric matrix and $\mathbf{q}$ is the vector of parameters to be estimated. The elements of $\partial g / \partial \mathrm{vech}(\mathbf{\Psi})^T$ are given by (3.5)-(3.6).

## 4 Monte Carlo experiments

To assess the performance of our ICA estimation method, we conduct a series of Monte Carlo experiments. For various error covariance models for $\mathbf{\Psi}$, we compare the estimation accuracy of the loadings matrix $\mathbf{B}$ for the correct $\mathbf{\Psi}$ model versus assuming a simpler $\mathbf{\Psi}$ model. Throughout, we have $R = 3$ independent components which follow a Laplace distribution, i.e.

$$f(x) = \frac{\sqrt{2}}{2} \exp\left(-\sqrt{2}\, |x|\right) \,. \tag{4.1}$$

From (4.1) it follows that the components have mean zero and variance 1. Their kurtosis equals 3.

We have $m = 9$ observed variables and loadings matrix

$$\mathbf{B} = \begin{bmatrix} 0.9 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 0.9 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0.1 & 0.1 & 0.1 \\ 0.5 & 0.5 & 0.5 \\ 0.9 & 0.9 & 0.9 \end{bmatrix} . \tag{4.2}$$

In each of the cases below, we generate $n = 500$ realizations of the independent components, and use 200 repetitions.

### Diagonal $\boldsymbol{\Psi}$

The true error covariance equals $\boldsymbol{\Psi} = \mathrm{diag}(0.8 \;\; 0.4 \;\; 0.8 \;\; 0.4 \;\; 0.8 \;\; 0.4 \;\; 0.8 \;\; 0.4 \;\; 0.8)$. In Table 1 in the Appendix we summarize the performance of correctly assuming a diagonal $\boldsymbol{\Psi}$ versus incorrectly assuming $\boldsymbol{\Psi} = \sigma^2 \, \mathbf{I}_m$. In the second step, we use four different ICA algorithms, namely Comon-4, JADE, FastICA and the least squares fitting of 4-th order cumulants (LSCUM-4). Recall that Comon-4 and JADE are designed to maximize the sum-of-squares of the kurtosis of the rotated components, and this objective is equivalent to LSCUM-4; see Proposition 2.1. For the 27 parameters of the mixing matrix $\mathbf{B}$ we give the maximum average bias for the 200 repetitions and the corresponding average relative bias. Also, we report the maximum average standard deviation and the maximum average root mean squared error for the parameters of $\mathbf{B}$.

For the FastICA algorithm we used $g(x) = x^3$ as derivative of $G$ in (2.18). For $g(x) = \tanh(x)$, which is recommended for components with positive kurtosis (Hyvärinen, 1999), the FastICA algorithm suffered from convergence problems.

As can be seen from Table 1, the correct error model has considerably lower bias than when incorrectly assuming $\boldsymbol{\Psi} = \sigma^2 \, \mathbf{I}_m$. The standard deviations and root mean squared errors are of the same order of magnitude for both models. JADE has the highest maximum bias, which is still 9.4 percent for the correct error model. Also, JADE has larger standard deviations than the other ICA algorithms.

Note that, with for $\mathbf{B}$ in (4.2) the identifiability condition for diagonal $\boldsymbol{\Psi}$ of Theorem 3.2 is satisfied.

### AR(1) model for $\boldsymbol{\Psi}$

The true values of the AR(1) parameters are $\sigma^2 = 0.8$ and $\rho = 0.2$. In Table 2 in the Appendix we summarize the performance of incorrectly assuming $\boldsymbol{\Psi} = \sigma^2 \, \mathbf{I}_m$ versus correctly assuming an AR(1) model for $\boldsymbol{\Psi}$. The results are comparable to those in Table 1. The maximum bias is substantially decreased by choosing the correct model for $\boldsymbol{\Psi}$, while the standard deviations and root mean squared errors are of the same order of magnitude for both models. If we take $\rho = 0.4$ instead of $\rho = 0.2$ (results not presented), then incorrectly choosing $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}_m$ results in a root mean squared error which is approximately twice as large as when we assume $\boldsymbol{\Psi}$ follows an AR(1) model.

Out of the four ICA methods, JADE gives in the highest bias and standard deviation.

In Table 4 in the Appendix we summarize the performance of incorrectly assuming a diagonal $\boldsymbol{\Psi}$ versus correctly assuming an AR(1) model for $\boldsymbol{\Psi}$. Clearly, incorrectly assuming uncorrelated errors results in a much higher bias and standard deviation. This holds for all four ICA algorithms.

For the estimation of the AR(1) parameters of the error model random starting values were used.

### Markov simplex model for $\boldsymbol{\Psi}$

The true values of the Markov simplex parameters are $\sigma^2 = 0.8$ and $\rho = 0.2$. In Table 3 in the Appendix we summarize the performance of incorrectly assuming $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}_m$ versus correctly assuming a Markov simplex model for $\boldsymbol{\Psi}$. As before, the maximum bias is substantially decreased by choosing the correct model for $\boldsymbol{\Psi}$, while the standard deviations and root mean squared errors are of the same order of magnitude for both models. If we take $\rho = 0.4$ instead of $\rho = 0.2$ (results not presented), then incorrectly choosing $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}_m$ results in a root mean squared error which is more than 1.5 times as large as when we assume $\boldsymbol{\Psi}$ follows a Markov simplex model.

The best performing ICA algorithms are Comon-4 and LSCUM-4, while JADE does worst.

In Table 5 in the Appendix we summarize the performance of incorrectly assuming a diagonal $\boldsymbol{\Psi}$ versus correctly assuming a Markov simplex model for $\boldsymbol{\Psi}$. Clearly, incorrectly assuming uncorrelated errors results in a much higher bias and standard deviation. This holds for all four ICA algorithms, although JADE has higher bias and standard deviation than the other three algorithms.

For the estimation of the Markov simplex parameters of the error model random starting values were used.

## 5 Discussion

In this paper, we discussed ICA with non-Gaussian components that have an unknown distribution, and Gaussian measurement errors, and we proposed a two-step estimation procedure. In the first step, least squares fitting of the model covariance to the observed covariance yields the unrotated components and loadings, and the error covariance matrix. In the second step, an ICA algorithm is used to find a rotation such that the rotated components are approximately independent. We considered isotropic errors, uncorrelated errors, AR(1) errors, and Markov simplex errors. For uncorrelated errors, two global identifiability conditions for error covariance $\boldsymbol{\Psi}$ are known in the literature (see Theorems 3.2 and 3.3). Further research is needed to obtain global identifiability results for non-diagonal $\boldsymbol{\Psi}$ and parametrized covariance models such as AR(1) and Markov simplex.

An interesting reference is Davies (2004) who discusses general identifiability issues in ICA with errors, especially related to the error covariance.

In case of isotropic errors, our approach is equivalent to using PCA to obtain the unrotated components (see Proposition 3.1). In case of uncorrelated and heteroscedastic errors, our approach is identical to Ikeda and Toyama (2000) who use Minres FA as the first step of the estimation procedure. In this case, also other FA techniques may be used. For example, Minimum Rank FA of Ten Berge and Kiers (1991) guarantees that both $\mathbf{\Psi}$ and $\mathbf{C} - \mathbf{\Psi}$ are positive semi-definite. This is not the case for Minres FA.

By Monte Carlo experiments, we compared four ICA algorithms for obtaining an optimal rotation: Comon-4, JADE, FastICA and LSCUM-4 (least squares fitting of the fourth order cumulants). Also, we assessed the consequences of misspecification of the error covariance model. As a measure of performance, we used the accuracy in estimating the loadings matrix. Three independent components were present, all following a Laplace distribution. The number of observed variables was nine. From the results of our Monte Carlo experiments, we may conclude that Comon-4 and LSCUM-4 gave the best overall performance in terms of bias and mean squared error of estimates of the loadings. JADE performed worst overall.

The results from the Monte Carlo experiments regarding the misspecification of the error covariance model were as follows. Considerably lower bias of the loadings estimates is obtained when the correct error covariance model is chosen, especially when choosing $\mathbf{\Psi}$ diagonal if the true error covariance model is AR(1) or Markov simplex. In that case, also the mean squared errors of the loadings estimates are considerably lower. Choosing $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ while the correct model for $\mathbf{\Psi}$ is diagonal, AR(1) or Markov simplex results in a mean squared error of the same order as when the correct model for $\mathbf{\Psi}$ is chosen. This is because the mean squared error is dominated by the variance of the estimator since the bias is relatively small. Considering only mean squared error one may therefore draw the conclusion that choosing $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ is not bad even in the presence of correlated errors. However, this only holds if the correlation parameter $\rho$ in (3.11) and (3.12) is small. If we take $\rho = 0.4$ instead of $\rho = 0.2$, then incorrectly choosing $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ results in a root mean squared error which is more than 1.5 times as large as for the correct error covariance model.

Our method does not have the drawbacks of Attias (1999), who assumes Gaussian mixtures for the component distributions, since our estimation procedure is efficient and information criteria can be used to infer on the number of independent components present in the data. Once more general identifiability results of the error covariance are obtained, least squares covariance fitting can be used for a large variety of ICA applications featuring correlated errors. The latter include radar, underwater passive listening, and general signal detection via an array of sensors.

# References

Anderson, T.W., & Rubin, H. (1956). Statistical inference in factor analysis. In Neyman, J. (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 5* (pp. 111–150) University of California Press.

Anderson, T.W. (2003) *An Introduction to Multivariate Statistical Analysis, 3rd edition.* New Jersey: Wiley.

Attias, H. (1999). Independent factor analysis. *Neural Computation*, *11*, 803–851.

Beckmann, C.F., & Smith, S.M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, *24*, 137–152.

Bekker, P.A., & Ten Berge, J.M.F. (1997). Generic global identification in factor analysis. *Linear Algebra and its Applications*, *264*, 255–263.

Bienvenu, G., & Kopp, L. (1983). Optimality of high resolution array processing using the eigensystem approach. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *31*, 1235–1248.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.

Cardoso, J.-F., & Souloumiac, A. (1993). Blind beamforming for non Gaussian signals. In *IEE Proceedings-F*, *140*, 362–370.

Cardoso, J.-F. (2005). JADE algorithm. Available online at
http://www.tsi.enst.fr/∼cardoso/Algo/Jade/jadeR.m

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

Comon, P. (1994b). Fourth order cumulants ICA algorithm. Available online at
http://www.i3s.unice.fr/∼comon/codesICA.txt

Comon, P., & Chevalier, P. (2000). Blind Source Separation: Models, Concepts, Algorithms and Performance. In Haykin, S. (Ed.), *Unsupervised Adaptive Filtering, Vol. I, Blind Source Separation, Series on Adaptive and Learning Systems for Communications, Signal Processing and Control* (pp. 191–236) Wiley.

Davies, M. (2004). Identifiability issues in noisy ICA. *IEEE Signal Processing Letters*, *11*, 470–473.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). An introduction to independent component analysis. *Journal of Chemometrics*, *14*, 123–149.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218.

Harman, H.H., & Jones, W.H. (1966). Factor analysis by minimizing residuals (Minres). *Psychometrika*, *31*, 351–369.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*, 626–634.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, *13*, 411–430.

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

Hyvärinen, A. (2005). FastICA routine. Available online at
http://www.cis.hut.fi/projects/ica/fastica

Ikeda, S., & Toyama, K. (2000) Independent component analysis for noisy data - MEG data analysis. *Neural Networks*, *13*, 1063–1074.

Jöreskog, K.G. (1977). Factor analysis by least-suqares and maximum-likelihood methods. In Enslein, K., Ralston, A., & Wilf, H.S. (Eds.), *Statistical Methods for Digital Computers* (pp. 125–153) New York: Wiley.

Kano, Y., Miyamoto, Y., & Shimizu, S. (2003). Factor rotation and ICA. *Proceedings of the 4th International Symposium on Independent Component analysis and Blind Signal Separation (ICA2003)*, Nara, Japan

Le Cadre, J.P. (1989). Parametric methods of spatial signal processing in the presence of unknown colored noise fields. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*, 965–983.

Ledermann, W. (1937). On the rank of the reduced correlation matrix in multiple-factor analysis. *Psychometrika*, *2*, 85–93.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572.

Ten Berge, J.M.F., & Kiers, H.A.L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, *56*, 309–315.

Tipping, M.E., & Bishop, C.M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Computation*, *11*, 443–482.

# Appendix: Monte Carlo results

| model $\boldsymbol{\Psi}$ | ICA method | max bias | max st. dev. | max rmse |
|---|---|---|---|---|
| $\boldsymbol{\Psi} = \sigma^2\,\mathbf{I}_m$ | Comon-4 | -0.1234 (13.7%) | 0.2190 | 0.2219 |
| $\boldsymbol{\Psi} = \sigma^2\,\mathbf{I}_m$ | JADE | -0.1299 (14.4%) | 0.2476 | 0.2487 |
| $\boldsymbol{\Psi} = \sigma^2\,\mathbf{I}_m$ | FastICA | -0.1198 (13.3%) | 0.2147 | 0.2163 |
| $\boldsymbol{\Psi} = \sigma^2\,\mathbf{I}_m$ | LSCUM-4 | -0.1235 (13.7%) | 0.2157 | 0.2208 |
| $\boldsymbol{\Psi}$ diagonal | Comon-4 | -0.0320 (3.6%) | 0.2152 | 0.2162 |
| $\boldsymbol{\Psi}$ diagonal | JADE | -0.0841 (9.4%) | 0.2378 | 0.2516 |
| $\boldsymbol{\Psi}$ diagonal | FastICA | -0.0380 (4.2%) | 0.1895 | 0.1914 |
| $\boldsymbol{\Psi}$ diagonal | LSCUM-4 | -0.0408 (4.5%) | 0.2011 | 0.2052 |

Table 1: Performance of ICA with errors by least squares covariance fitting: incorrectly assuming $\boldsymbol{\Psi} = \sigma^2\,\mathbf{I}_m$ versus correctly assuming diagonal $\boldsymbol{\Psi}$. The true error covariance equals $\boldsymbol{\Psi} = \text{diag}(0.8 \quad 0.4 \quad 0.8 \quad 0.4 \quad 0.8 \quad 0.4 \quad 0.8 \quad 0.4 \quad 0.8)$. Four different ICA algorithms have been used. The last three columns present the maximal bias, maximal standard deviation and maximal root mean squared error among the 27 parameters of the mixing matrix $\mathbf{B}$. For the maximal bias, also the relative bias is given.

| model $\mathbf{\Psi}$ | ICA method | max bias | max st. dev. | max rmse |
|---|---|---|---|---|
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | Comon-4 | -0.1234 (13.7%) | 0.2190 | 0.2219 |
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | JADE | -0.1299 (14.4%) | 0.2476 | 0.2487 |
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | FastICA | -0.1198 (13.3%) | 0.2147 | 0.2163 |
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | LSCUM-4 | -0.1235 (13.7%) | 0.2157 | 0.2208 |
| $\mathbf{\Psi}$ AR(1) | Comon-4 | -0.0438 (4.9%) | 0.2156 | 0.2168 |
| $\mathbf{\Psi}$ AR(1) | JADE | -0.0575 (6.4%) | 0.2635 | 0.2676 |
| $\mathbf{\Psi}$ AR(1) | FastICA | -0.0546 (6.1%) | 0.2260 | 0.2325 |
| $\mathbf{\Psi}$ AR(1) | LSCUM-4 | -0.0365 (4.1%) | 0.2045 | 0.2052 |

Table 2: Performance of ICA with errors by least squares covariance fitting: incorrectly assuming $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ versus correctly assuming an AR(1) model for $\mathbf{\Psi}$. The true model for $\mathbf{\Psi}$ is AR(1) with $\sigma^2 = 0.8$ and $\rho = 0.2$. Four different ICA algorithms have been used. The last three columns present the maximal bias, maximal standard deviation and maximal root mean squared error among the 27 parameters of the mixing matrix $\mathbf{B}$. For the maximal bias, also the relative bias is given.

| model $\mathbf{\Psi}$ | ICA method | max bias | max st. dev. | max rmse |
|---|---|---|---|---|
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | Comon-4 | 0.1157 (Inf) | 0.2223 | 0.2250 |
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | JADE | -0.1110 (12.3%) | 0.3603 | 0.3761 |
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | FastICA | 0.1086 (Inf) | 0.2381 | 0.2444 |
| $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ | LSCUM-4 | 0.1145 (Inf) | 0.2066 | 0.2099 |
| $\mathbf{\Psi}$ Markov | Comon-4 | -0.0396 (4.4%) | 0.1969 | 0.1973 |
| $\mathbf{\Psi}$ Markov | JADE | -0.0510 (5.7%) | 0.2330 | 0.2372 |
| $\mathbf{\Psi}$ Markov | FastICA | -0.0502 (5.6%) | 0.2233 | 0.2289 |
| $\mathbf{\Psi}$ Markov | LSCUM-4 | -0.0336 (3.7%) | 0.1968 | 0.1975 |

Table 3: Performance of ICA with errors by least squares covariance fitting: incorrectly assuming $\mathbf{\Psi} = \sigma^2\,\mathbf{I}_m$ versus correctly assuming a Markov simplex model for $\mathbf{\Psi}$. The true model for $\mathbf{\Psi}$ is Markov simplex with $\sigma^2 = 0.8$ and $\rho = 0.2$. Four different ICA algorithms have been used. The last three columns present the maximal bias, maximal standard deviation and maximal root mean squared error among the 27 parameters of the mixing matrix $\mathbf{B}$. For the maximal bias, also the relative bias is given. The relative bias is reported as Inf if the true parameter value is zero.

| model $\mathbf{\Psi}$ | ICA method | max bias | max st. dev. | max rmse |
|---|---|---|---|---|
| $\mathbf{\Psi}$ diagonal | Comon-4 | 0.4432 (49%) | 1.2708 | 1.3459 |
| $\mathbf{\Psi}$ diagonal | JADE | 0.4527 (50%) | 1.2546 | 1.3338 |
| $\mathbf{\Psi}$ diagonal | FastICA | 0.4480 (50%) | 1.1833 | 1.2653 |
| $\mathbf{\Psi}$ diagonal | LSCUM-4 | 0.4411 (49%) | 1.2563 | 1.3315 |
| $\mathbf{\Psi}$ AR(1) | Comon-4 | -0.0426 (4.7%) | 0.2134 | 0.2145 |
| $\mathbf{\Psi}$ AR(1) | JADE | -0.0399 (4.4%) | 0.2565 | 0.2583 |
| $\mathbf{\Psi}$ AR(1) | FastICA | -0.0413 (4.6%) | 0.2113 | 0.2153 |
| $\mathbf{\Psi}$ AR(1) | LSCUM-4 | -0.0431 (4.8%) | 0.2119 | 0.2156 |

Table 4: Performance of ICA with errors by least squares covariance fitting: incorrectly assuming diagonal $\mathbf{\Psi}$ versus correctly assuming an AR(1) model for $\mathbf{\Psi}$. The true model for $\mathbf{\Psi}$ is AR(1) with $\sigma^2 = 0.8$ and $\rho = 0.2$. Four different ICA algorithms have been used. The last three columns present the maximal bias, maximal standard deviation and maximal root mean squared error among the 27 parameters of the mixing matrix $\mathbf{B}$. For the maximal bias, also the relative bias is given.

| model $\mathbf{\Psi}$ | ICA method | max bias | max st. dev. | max rmse |
|---|---|---|---|---|
| $\mathbf{\Psi}$ diagonal | Comon-4 | 0.3367 (37%) | 1.2992 | 1.3421 |
| $\mathbf{\Psi}$ diagonal | JADE | 0.3448 (38%) | 1.3312 | 1.3751 |
| $\mathbf{\Psi}$ diagonal | FastICA | 0.3670 (41%) | 1.3333 | 1.3829 |
| $\mathbf{\Psi}$ diagonal | LSCUM-4 | 0.3346 (37%) | 1.2938 | 1.3363 |
| $\mathbf{\Psi}$ Markov | Comon-4 | -0.0367 (4.1%) | 0.2389 | 0.2405 |
| $\mathbf{\Psi}$ Markov | JADE | -0.0929 (10%) | 0.2877 | 0.3023 |
| $\mathbf{\Psi}$ Markov | FastICA | -0.0313 (3.5%) | 0.2319 | 0.2334 |
| $\mathbf{\Psi}$ Markov | LSCUM-4 | -0.0409 (4.5%) | 0.2428 | 0.2462 |

Table 5: Performance of ICA with errors by least squares covariance fitting: incorrectly assuming diagonal $\mathbf{\Psi}$ versus correctly assuming a Markov simplex model for $\mathbf{\Psi}$. The true model for $\mathbf{\Psi}$ is Markov simplex with $\sigma^2 = 0.8$ and $\rho = 0.2$. Four different ICA algorithms have been used. The last three columns present the maximal bias, maximal standard deviation and maximal root mean squared error among the 27 parameters of the mixing matrix $\mathbf{B}$. For the maximal bias, also the relative bias is given.