Performance Analysis of the FastICA Algorithm and Cramér–Rao Bounds for Linear Independent Component Analysis

Petr Tichavský, *Senior Member, IEEE*, Zbyněk Koldovský, *Member, IEEE*, and Erkki Oja, *Fellow, IEEE*

***Abstract*—The FastICA or fixed-point algorithm is one of the most successful algorithms for linear independent component analysis (ICA) in terms of accuracy and computational complexity. Two versions of the algorithm are available in literature and software: a one-unit (deflation) algorithm and a symmetric algorithm. The main result of this paper are analytic closed-form expressions that characterize the separating ability of both versions of the algorithm in a local sense, assuming a "good" initialization of the algorithms and long data records. Based on the analysis, it is possible to combine the advantages of the symmetric and one-unit version algorithms and predict their performance. To validate the analysis, a simple check of saddle points of the cost function is proposed that allows to find a global minimum of the cost function in almost 100% simulation runs. Second, the Cramér–Rao lower bound for linear ICA is derived as an algorithm independent limit of the achievable separation quality. The FastICA algorithm is shown to approach this limit in certain scenarios. Extensive computer simulations supporting the theoretical findings are included.**

***Index Terms*—Blind-source separation, independent component analysis (ICA), Cramér–Rao lower bound.**

I. Introduction

BLIND-SOURCE separation (BSS), which consists of recovering original signals from their mixtures when the mixing process is unknown, has been a widely studied problem in signal processing for the last two decades (for a review, see [1]). Independent component analysis (ICA), a statistical method for signal separation [2], [3], is also a well-known issue in the community. Its aim is to transform the mixed random signals into source signals or components that are as mutually independent as possible. There are a number of methods intended to solve related problems such as blind deconvolution and blind equalization [4]–[6].

One of the most widely used ICA algorithms for the linear mixing model is FastICA, a fixed-point algorithm first proposed by Hyvärinen and Oja [7], [8]. It is based on the optimization of a nonlinear contrast function measuring the non-Gaussianity of the sources. A widely used contrast function both in FastICA and in many other ICA algorithms is the kurtosis [9]–[11]. This approach can be considered as an extension of the algorithm by Shalvi and Weinstein [6].

There are two varieties of the FastICA algorithm: the deflation, or one-unit algorithm, and the symmetric algorithm. The deflation approach, which is common for many other ICA algorithms [9], estimates the components successively under orthogonality conditions. The symmetric algorithm estimates the components in parallel. This consists of parallel computation of the one-unit updates for each component, followed by subsequent symmetric orthogonalization of the estimated demixing matrix after each iteration. A version of FastICA for complex valued signals was proposed in [12].

An essential question is the convergence of the FastICA algorithm. This can be approached from two directions. First, assuming an ideal infinitely large sample, theoretical expectations for the contrast functions such as the kurtosis can be used in the analysis. Then, the contrast function and the algorithm itself become deterministic, and questions such as asymptotic stability of the extrema and the convergence speed can be discussed. For the kurtosis cost function and the one-unit algorithm, this analysis was done in [7], showing cubic convergence. For a general cost function, the convergence speed is at least quadratic, as shown in [8] (see also [3]). The monotonic convergence and the speed for a general cost function for the related gradient algorithm was considered in [13]. For the kurtosis cost function and the symmetric FastICA algorithm, the cubic convergence was proven in [14] (see also [15]). Different properties of the one-unit version have been illustrated by computer simulations in [16] where the accuracy is also shown to be very good in most cases.

The second question of convergence considers the behavior of the algorithm for a finite sample, which is the practical case. Then, the theoretical expectations in the contrast functions are replaced by sample averages. This results in errors in the estimator for the demixing matrix. A classical measure of the error is the asymptotic variance of the matrix elements. The goal of designing an ICA algorithm is then to make this error as small

Manuscript received November 26, 2004; revised May 19, 2005. The work of P. Tichavský was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M6798555601. Part of this paper (on the CRB) is to be presented at International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, PA, 2005, and another part (analysis of FastICA) was submitted to IEEE SSP'05 Workshop, Bordeaux, France. **<AUTHOR: Please provide full date location of conferences/workshop and full name of SSP Workshop, and update if necessary—ed.>**The associate editor coordinating the review of this paper and approving it for publication was Prof. Jonathon Chambers.

P. Tichavský is with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P 182 08 Prague 8, Czech Republic (e-mail: p.tichavsky@ieee.org' website: http://www.utia.cas.cz/user\rule-data/scientific/SI\ruledept/Tichavsky.html).

Z. Koldovský is with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 182 08 Prague 8, Czech Republic (e-mail: koldovsk@utia.cas.cz; website: http://www.utia.cas.cz/user\rule-data/scientific/SI\ruledept/Tichavsky.html).

E. Oja is with the Neural Networks Research Centre, Helsinki University of Technology, 02015 HUT, Finland (e-mail: Erkki.Oja@hut.fi; website: http://www.cis.hut.fi/~oja/).

Digital Object Identifier 10.1109/TSP.2006.870561

as possible. For the FastICA algorithm, such an asymptotic performance analysis for a general cost function was proposed in [17].

The Cramér–Rao lower bound **<AUTHOR: CRLB?—ed.>**(CRB) provides an algorithm independent bound for parameter estimation. In the context of ICA, a Cramér–Rao-like bound for intersignal interference is derived as asymptotic variance of a maximum-likelihood estimate in [24] and [26]–[29]. A similar result is known for a related problem of blind deconvolution [30].

The purpose of the present paper is to look at the performance of the FastICA algorithm, both the one-unit and symmetric versions, in this latter sense of asymptotic error, and compare it with the exact CRB computed from its definition. The paper is organized as follows. In Section II, the linear ICA model and the FastICA algorithm are described. In addition, a novel check of saddle points of the FastICA cost function is proposed that allows to find the global minimum of the cost function in almost 100% simulation runs. Finally, the following criteria to characterize the performance of the algorithm are introduced: a gain matrix (variance of its elements) and a signal-to-interference ratio (SIR). In Section III, analytic expressions for the variance of the off-diagonal gain matrix elements are derived and discussed. These expressions are asymptotically valid for large data sets when a "good" initialization of the algorithm is assumed. Most of the details of the analysis are deferred to Appendixes. As an example of utilization of the analysis, a novel variant of FastICA is proposed, which combines the one-unit algorithm and the symmetric algorithm adaptively, depending on empirical distribution of the estimated signal components, to improve the performance.

In Section IV, the CRB on the variance of the off-diagonal gain matrix elements is computed via inverse of a Fisher information matrix. Section V compares the CRB with the asymptotic performance of FastICA and explains nonexistence of the CRB for signals with bounded magnitude (e.g., uniform distribution) and for some long-tailed distributions.

Section VI presents a number of computer simulations using artificial data that validate and support the theoretical analysis. The simulations also compare the algorithmic performance with the CRB derived in Section IV. Finally, Section VII summarizes the results and presents the conclusions.

II. Data Model and the Method

Let $\mathbf{S}$ represent a $d \times N$ data matrix, composed of d rows, where each row $\mathbf{s}_k^T, k = 1, \ldots, d$ contains N independent realizations of a random variable s_k. Next assume that s_k has a distribution function $F_k(t) = P(s_k \leq t)$. In a typical case for ICA, the rows $\mathbf{s}_k^T$ are called the source signals, and the d random variables s_k are mutually independent.

The standard linear ICA model of a given $d \times N$ data matrix is

$$\mathbf{X} = \mathbf{A}\mathbf{S} \tag{1}$$

where $\mathbf{A}$ is an unknown, nonsingular $d \times d$ mixing matrix. Thus, each row of $\mathbf{X}$ is a linear mixture of the unknown independent signals $\mathbf{s}_k^T$. The goal of independent component analysis is to estimate the matrix $\mathbf{A}$ or, equivalently, the demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ or, equivalently, the original source signals $\mathbf{S}$. The following are well known:

1) the separation is unique only up to an unknown scaling and ordering of the components $\mathbf{s}_k^T$;
2) the separation is possible only if at most one of the original source variables s_k has a Gaussian distribution.

Since the scale of the source signals cannot be retrieved, one can assume, without any loss in generality, that the sample variance of the estimated source signals is equal to one. Thus, instead of the original source signals $\mathbf{S}$, a normalized source signal matrix denoted $\mathbf{U}$ can be estimated, where

$$\mathbf{U} = \mathbf{D}^{-1/2}(\mathbf{S} - \bar{\mathbf{S}}) \tag{2}$$

$$\mathbf{D} = \mathrm{diag}[\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2] \tag{3}$$

$$\hat{\sigma}_k^2 = (\mathbf{s}_k - \bar{\mathbf{s}}_k)^T(\mathbf{s}_k - \bar{\mathbf{s}}_k)/N. \tag{4}$$

$$\bar{\mathbf{s}}_k = (\mathbf{s}_k^T \cdot \mathbf{1}_N)\mathbf{1}_N/N, \quad k = 1, \ldots, d \tag{5}$$

where $\mathbf{1}_N$ stands for $N \times 1$ vector of 1's.

A. Preprocessing

The first step of many variants of the ICA algorithms consists of removing the sample mean and a whitening (decorrelation and scaling), i.e., the transformation

$$\mathbf{Z} = \hat{\mathbf{C}}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}}) \tag{6}$$

where

$$\hat{\mathbf{C}} = (\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T/N \tag{7}$$

is the sample covariance matrix, and $\bar{\mathbf{X}}$ is the sample mean, $\bar{\mathbf{X}} = \mathbf{X} \cdot \mathbf{1}_N \mathbf{1}_N^T/N$. The output $\mathbf{Z}$ contains decorrelated and unit variance data in the sense that $\mathbf{Z}\mathbf{Z}^T/N = \mathbf{I}$ (identity matrix). Note that $\mathbf{Z}$ can be rewritten using (1) and (2) as

$$\mathbf{Z} = \hat{\mathbf{C}}^{-1/2}\mathbf{A}\mathbf{D}^{1/2}\mathbf{U}. \tag{8}$$

The ICA problem can be formulated as the one to find a demixing matrix $\hat{\mathbf{W}}(\mathbf{Z})$ that separates the original signals from the mixture $\mathbf{Z}$, i.e., $\hat{\mathbf{U}} = \hat{\mathbf{W}}(\mathbf{Z}) \cdot \mathbf{Z}$.

B. FastICA Algorithm for One Unit

The fixed-point algorithm for one=unit estimates one row of the demixing matrix $\mathbf{W}(\mathbf{Z})$ as a vector $\hat{\mathbf{w}}_z^T$ that is a stationary point (minimum or maximum) of the expression $\mathrm{E}[G(\mathbf{w}^T\mathbf{Z})] \overset{\mathrm{def}}{=} G(\mathbf{w}^T\mathbf{Z})\mathbf{1}_N/N$ subject to $\|\mathbf{w}\| = 1$, where $G(\,\cdot\,)$ is a suitable nonlinear and nonquadratic function [3]. In the above expression, $G(\,\cdot\,)$ is applied elementwise.

Finding $\hat{\mathbf{w}}_z$ proceeds iteratively. Starting with a random initial unit norm vector $\mathbf{w}$, iterate

$$\mathbf{w}^+ \leftarrow \mathbf{Z}g(\mathbf{Z}^T\mathbf{w}) - \mathbf{w}g'(\mathbf{w}^T\mathbf{Z})\mathbf{1}_N \tag{9}$$

$$\mathbf{w} \leftarrow \mathbf{w}^+/\|\mathbf{w}^+\| \tag{10}$$

until convergence is achieved. In (9) and also elsewhere in the paper, in accord with the standard notation [3], $g(\,\cdot\,)$ and $g'(\,\cdot\,)$ denote the first and the second derivative of the function $G(\,\cdot\,)$.

The application of $g(\,\cdot\,)$ and $g'(\,\cdot\,)$ to the vector $\mathbf{w}^T\mathbf{Z}$ is elementwise. Classical widely used functions $g(\,\cdot\,)$ include "pow3," i.e., $g(x) = x^3$ (then the algorithm performs kurtosis minimization), "tanh," i.e., $g(x) = \tanh(x)$, and "Gauss," $g(x) = x\exp(-x^2/2)$.

It is not known in advance which column of $\mathbf{W}^T(\mathbf{Z})$ is being estimated: It largely depends on the initialization. Note that the recursion for some components might not converge. In the deflation method [9], which is not studied in this paper, this problem is solved by separating the components from the mixture one by one using orthogonal projections. Here, we shall assume that each signal component can be separated from the *original* signal mixture using suitable initializations. Assume that the separating vectors $\mathbf{w}$ are computed for all components are appropriately sorted [20] and summarized as rows in a matrix denoted $\hat{\mathbf{W}}^{1U}(\mathbf{Z})$. The rows in $\hat{\mathbf{W}}^{1U}(\mathbf{Z})$ may not be mutually orthogonal, in general.

C. Symmetric Fastica Algorithm

The symmetric FastICA proceeds similarly, the estimation of all independent components (or equivalently, of all rows of $\mathbf{W}$) proceeds in parallel, and each step is completed by a symmetric orthonormalization. Starting with a random unitary matrix $\mathbf{W}$, iterate

$$\mathbf{W}^+ \leftarrow g(\mathbf{W}\mathbf{Z})\mathbf{Z}^T - \mathrm{diag}[g'(\mathbf{W}\mathbf{Z})\mathbf{1}_N]\,\mathbf{W} \qquad (11)$$

$$\mathbf{W} \leftarrow (\mathbf{W}^+\mathbf{W}^{+T})^{-1/2}\mathbf{W}^+ \qquad (12)$$

until convergence is achieved. The stopping criterion proposed in [14] is

$$1 - \min(\mathrm{abs}(\mathrm{diag}(\mathbf{W}^T\mathbf{W}_{\mathrm{old}}))) < \epsilon \qquad (13)$$

for a suitable constant ϵ.

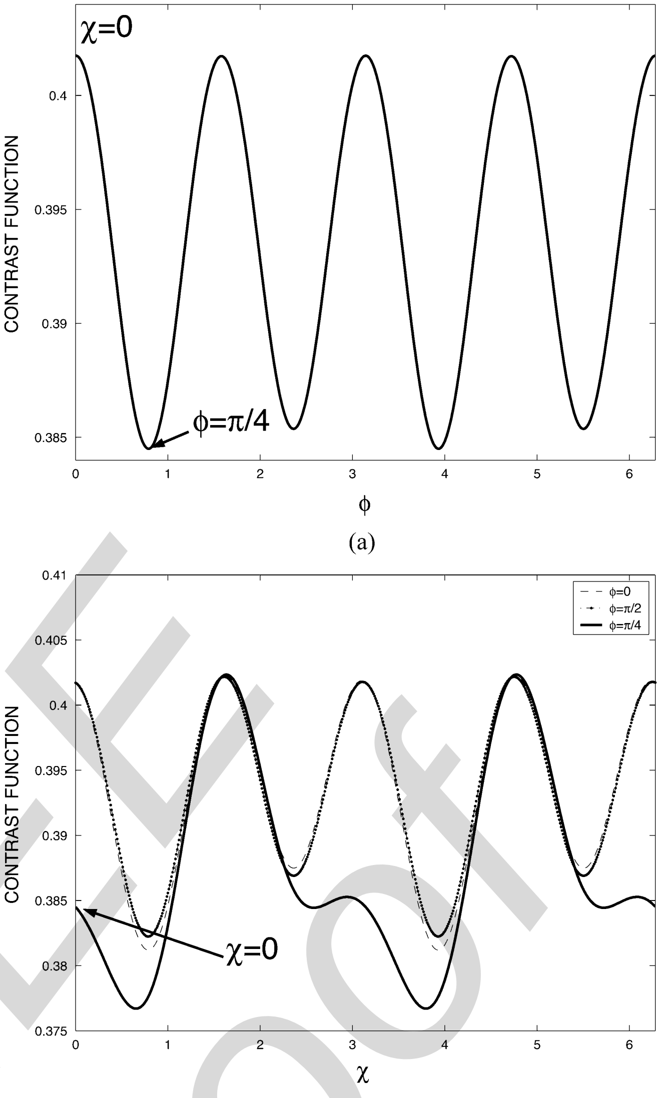
The result of the symmetric FastICA (unlike in the one-unit algorithm without deflation) is a unitary matrix denoted $\hat{\mathbf{W}}^{\mathrm{SYM}}(\mathbf{Z})$. As a consequence, sample correlations between the separated signals are exactly equal to zero.

D. Check of Saddle Points

In general, the global convergence of the symmetric FastICA is known to be quite good. Nevertheless, if it is run 10 000 times from random initial demixing matrices, on the average in 1–100 cases, the algorithm gets stuck at solutions that can be recognized by exceptionally low achieved SIR. The rate of these false solutions depends on the dimension of the model, on the stopping rule, and on the length of the data (see the example at the end of this subsection).

A detailed investigation of the false solutions showed that they contain one or more pairs of estimated components, say $(\hat{\mathbf{u}}_k, \hat{\mathbf{u}}_\ell)$, such that they are close to $(\mathbf{u}_k + \mathbf{u}_\ell)/\sqrt{2}$ and $(\mathbf{u}_k - \mathbf{u}_\ell)/\sqrt{2}$, respectively, where $(\mathbf{u}_k, \mathbf{u}_\ell)$ is the desired solution (see Fig. 1). Due to symmetry, the saddle points of the criterion function lie approximately halfway between two correct solutions that differ in the order of two of their components. Thus, an appropriate estimate of $(\mathbf{u}_k, \mathbf{u}_\ell)$ would be $(\hat{\mathbf{u}}'_k, \hat{\mathbf{u}}'_\ell)$, where

$$\hat{\mathbf{u}}'_k = (\hat{\mathbf{u}}_k + \hat{\mathbf{u}}_\ell)/\sqrt{2} \quad \text{and} \quad \hat{\mathbf{u}}'_\ell = (\hat{\mathbf{u}}_k - \hat{\mathbf{u}}_\ell)/\sqrt{2}.$$



(a)

(b)

Fig. 1. Contrast function $\mathrm{E}\{G(\cos\chi(\cos\varphi\cdot\mathbf{x}_1 + \sin\varphi\cdot\mathbf{x}_2) + \sin\chi\cdot\mathbf{x}_3]\}$ (a) as a function of φ for $\chi = 0$, and (b) as a function of χ for $\varphi = 0, \pi/4$, and $\pi/2$, respectively; $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ were generated as i.i.d. uniformly distributed in $[-\sqrt{3}, \sqrt{3}]$ with the length $N = 10\,000$, and $G(x) = \log\cosh(x)$. The point $[\varphi, \chi] = [\pi/4, 0]$ is a saddle point of the contrast function—it is its local minimum with regard to φ and a local maximum with regard to χ.

A selection between given candidates $(\hat{\mathbf{u}}_k, \hat{\mathbf{u}}_\ell), (\hat{\mathbf{u}}'_k, \hat{\mathbf{u}}'_\ell)$ for a better estimate of $(\mathbf{u}_k, \mathbf{u}_\ell)$ can be done by maximizing the criterion used in the very beginning of derivation of FastICA

$$c(\hat{\mathbf{u}}_k, \hat{\mathbf{u}}_\ell) = \left[G\left(\hat{\mathbf{u}}_k^T\right)\mathbf{1}_N/N - G_0\right]^2 + \left[G\left(\hat{\mathbf{u}}_\ell^T\right)\mathbf{1}_N/N - G_0\right]^2$$

where $G_0 = \mathrm{E}[G(\xi)]$ and ξ is a standard normal random variable. In the case of the nonlinearity "tanh," $G(x) = \log\cosh(x)$ and $G_0 \approx 0.3746$.

Thus, we suggest to complete the plain symmetric FastICA by the check of all $\binom{d}{2}$ pairs of the estimated independent components for a possible improvement via the saddle points. If the test for saddle point is positive, it is suggested to perform one or two additional iterations of the original algorithm, starting from the improved estimate.

The failure rates of the plain symmetric FastICA with three different stopping rules and of the improved FastICA with the check of the saddle points are compared in the following example. The first stopping rule was (13) with $\epsilon = 10^{-4}$, the second stopping rule was the same with $\epsilon = 10^{-5}$, and the third stopping rule required the former condition to be fulfilled in three consecutive steps. The improved algorithm used the first stopping rule and the test of the saddle points.

These four variants of the algorithm were applied to separate $d = 2$, 3, 4, and 5 independent signals with uniform distribution and varying length in 10 000 independent trials with a randomly selected initial demixing matrix. The number of algorithmic failures that are detected by the condition that SIR of some of the separated components is smaller than 3 dB is displayed in Table I. The table shows zero rate of the improved algorithm except for the case of the data with the shortest length, $N = 200$. In the latest case, the rate of failures has significantly dropped compared to the former three variants.

E. Measure of the Separation Quality

The separation ability of ICA algorithms can be characterized by the relative presence of the kth source signal in the estimated ith source signal. It is possible, if the source signals are known. Due to the permutation and sign/phase uncertainty, the estimated sources need to be appropriately sorted to fit the original ones. In this paper, the method proposed in [20] is used. Formally, the estimated source signals can be written using (8) as

$$\hat{\mathbf{U}} = \hat{\mathbf{W}}(\mathbf{Z}) \cdot \mathbf{Z} = \hat{\mathbf{W}}(\mathbf{Z})\hat{\mathbf{C}}^{-1/2}\mathbf{A}\mathbf{D}^{1/2}\mathbf{U} = \mathbf{G}\mathbf{U} \tag{14}$$

where $\mathbf{G} = \hat{\mathbf{W}}(\mathbf{Z})\hat{\mathbf{C}}^{-1/2}\mathbf{A}\mathbf{D}^{1/2}$ and $\hat{\mathbf{W}}(\mathbf{Z})$ stands either for $\hat{\mathbf{W}}^{1U}(\mathbf{Z})$ or for $\hat{\mathbf{W}}^{\mathrm{SYM}}(\mathbf{Z})$. Note that $\mathbf{G}$ has the meaning of the estimated demixing matrix provided that $\mathbf{A} = \mathbf{D} = \mathbf{I}$. It will be called the *gain* matrix for brevity.

The relative presence of the kth source signal in the estimated ith source signal is represented by the (i, k)th element of $\mathbf{G}$, denoted $\mathbf{G}_{ik}$. Then, the total SIR of the kth source signal is defined as follows:

$$\mathrm{SIR}_k = \frac{\mathrm{E}[\mathbf{G}_{kk}^2]}{\mathrm{E}\left[\sum_{\substack{k=1\\k\neq i}}^{d} \mathbf{G}_{ik}^2\right]}. \tag{15}$$

It is important to note that the estimator $\hat{\mathbf{U}}$ is invariant with respect to orthogonal transformations of the decorrelated data $\mathbf{Z}$, or equivariant [10]. It is because the recursions (9) and (10) or (11) and (12) that represent the algorithm are equivalent to the same relations with $\mathbf{Z}$, $\mathbf{W}^{+}$, and $\mathbf{W}$ replaced by $\mathbf{QZ}$, $\mathbf{W}^{+}\mathbf{Q}^{-1}$, and $\mathbf{WQ}^{-1}$, respectively, where $\mathbf{Q}$ is an arbitrary unitary (i.e., obeying $\mathbf{Q}^T = \mathbf{Q}^{-1}$) matrix. Then, the product

$$\hat{\mathbf{U}} = \mathbf{W} \cdot \mathbf{Z} = \mathbf{WQ}^{-1} \cdot \mathbf{QZ}$$

remains independent of $\mathbf{Q}$. From these facts, it follows that the gain matrix $\mathbf{G}$ and consequently the SIR are *independent* of the mixing matrix $\mathbf{A}$.[1]

TABLE I
NUMBER OF FAILURES OF SYMMETRIC FASTICA (tanh) AMONG 10 000 TRIALS

	N=200	N=500	N=1000	N=10000
$d = 2\,\&\,\varepsilon = 10^{-4}$	85	57	59	46
$d = 2\,\&\,\varepsilon = 10^{-5}$	49	16	15	12
$d = 2\,\&$ stop 3×	41	4	1	2
$d = 2\,\&$ s.p.check	**0**	**0**	**0**	**0**
$d = 3\,\&\,\varepsilon = 10^{-4}$	49	5	4	6
$d = 3\,\&\,\varepsilon = 10^{-5}$	43	0	1	0
$d = 3\,\&$ stop 3×	45	0	0	0
$d = 3\,\&$ s.p.check	**0**	**0**	**0**	**0**
$d = 4\,\&\,\varepsilon = 10^{-4}$	95	9	4	11
$d = 4\,\&\,\varepsilon = 10^{-5}$	85	2	0	5
$d = 4\,\&$ stop 3×	90	1	0	1
$d = 4\,\&$ s.p.check	**5**	**0**	**0**	**0**
$d = 5\,\&\,\varepsilon = 10^{-4}$	166	2	4	11
$d = 5\,\&\,\varepsilon = 10^{-5}$	151	1	2	2
$d = 5\,\&$ stop 3×	157	1	2	0
$d = 5\,\&$ s.p.check	**17**	**0**	**0**	**0**

III. ANALYSIS

Due to the above-mentioned equivariant property of FastICA it can be assumed, without any loss in generality, that the recursions (9) and (10) or (11) and (12) begin with the decorrelated data of the form

$$\mathbf{Z} = \mathbf{R}^{-1/2}\mathbf{U} \tag{16}$$

where

$$\mathbf{R} = \frac{1}{N}\mathbf{U}\mathbf{U}^T. \tag{17}$$

The gain matrix of interest is now

$$\mathbf{G} = \hat{\mathbf{W}}(\mathbf{Z}) \cdot \mathbf{R}^{-1/2}. \tag{18}$$

Note that the gain matrix $\mathbf{G}$ (and consequently the SIR as well) is a function of the normalized source signals $\mathbf{U}$ and of the nonlinear function $g(\cdot)$ used in the algorithm only.

The main result of this section can be summarized as follows.

Proposition 1: Assume that 1) all original independent components have zero mean and unit variance and are temporarily

[1] To be exact, a change of the mixing matrix (or a change in the algorithm initialization) may cause a change of the order or sign of the components at the algorithm output. Here, however, we assume that the order and signs of the components are post-processed to fit the original signals [20].

TABLE II
SIR (IN DECIBELS] OF FastICA IN ITS MAIN SIX VARIANTS FOR TWO COMPONENTS WITH THE SAME DISTRIBUTION, AND THE CRAMÉR–RAO BOUND (DERIVED IN SECTION IV) FOR $N = 1000$. THE BEST SIR IS MARKED BY BOLD CHARACTERS

PDF	SYMMETRIC			ONE UNIT			CRB
	TANH	GAUSS	POW3	TANH	GAUSS	POW3	
uniform	32.3	32.2	33.3	31.6	31.5	**33.7**	∞
sinus	34.7	34.7	35.1	37.5	37.6	**39.5**	∞
bpsk	36.0	36.0	36.0	∞	∞	∞	∞
GG(4)	27.6	27.5	**28.0**	25.2	25.1	25.7	28.1
GG(3)	**23.9**	23.7	23.7	21.1	20.1	20.1	24.0
Laplace	29.0	**29.4**	24.9	27.0	27.4	22.2	31.8
GG(0.5)	33.3	33.9	24.9	33.6	**35.0**	22.2	∞
GG(0.1)	35.9	35.9	-3.1	47.8	**50.7**	-6.11	∞

white, 2) the function g in algorithm FastICA is twice continuously differentiable, 3) the following expectations exist:

$$\mathrm{E}[s_k g(s_k)] \overset{\text{def}}{=} \mu_k \tag{19}$$

$$\mathrm{E}[g'(s_k)] \overset{\text{def}}{=} \rho_k \tag{20}$$

$$\mathrm{E}[g^2(s_k)] \overset{\text{def}}{=} \beta_k \tag{21}$$

for $k = 1, \ldots, d$, and 4) the FastICA algorithm (in both variants) is started from the correct demixing matrix and stops after a single iteration.

Then, the normalized gain matrix elements $N^{1/2}\mathbf{G}_{k\ell}^{1U}$ and $N^{1/2}\mathbf{G}_{k\ell}^{\mathrm{SYM}}$ for the one-unit FastICA and for symmetric FastICA, respectively, have asymptotically Gaussian distribution $\mathcal{N}(0, V_{k\ell}^{1U})$ and $\mathcal{N}(0, V_{k\ell}^{\mathrm{SYM}})$, where

$$V_{k\ell}^{1U} = \frac{\beta_k - \mu_k^2}{(\mu_k - \rho_k)^2} \tag{22}$$

$$V_{k\ell}^{\mathrm{SYM}} = \frac{\beta_k - \mu_k^2 + \beta_\ell - \mu_\ell^2 + (\mu_\ell - \rho_\ell)^2}{(|\mu_k - \rho_k| + |\mu_\ell - \rho_\ell|)^2} \tag{23}$$

for $k, \ell = 1, \ldots, d, k \neq \ell$, provided that the denominators are nonzero.

Proof: See Appendix A. An expression similar to (22) can be found in [10] and [17], but (23) is novel.

The assumption 4 may look peculiar at the first glance, but it is not so restrictive as it seems to be. It reflects the fact that the presented analysis is "local" and assumes a "good" initialization of the algorithm. The algorithm itself may have good global convergence properties (see Section VI), but it is not a subject of this proposition. Once the algorithm is started from an initial $\mathbf{W}$ that lies in a right domain of attraction, the resultant stationary point of the recursion, denoted $\hat{\mathbf{W}}$, is the same and is approximately equal to $\mathbf{W}^+$ obtained after one step from the ideal solution, due to the fact that the convergence is quadratic.[2]

Our numerical simulations presented in Section VII, and also other simulations that were skipped for lack of space, confirm the validity of the asymptotic variances (22) and (23) for the algorithm variant introduced in Section VI *working with arbitrary (random) initialization*. Namely, it is shown that $\mathrm{var}[\mathbf{G}_{k\ell}^{1U}] \approx (1/N)V_{k\ell}^{1U}$ and $\mathrm{var}[\mathbf{G}_{k\ell}^{\mathrm{SYM}}] \approx (1/N)V_{k\ell}^{\mathrm{SYM}}$. The expressions in (22) and (23) are functions of the probability distribution of s_k and of the nonlinear function $g(\cdot)$ via the expectations in (19)–(21). Given the distribution and the nonlinearity, these expressions can be evaluated.

Table II shows the theoretical SIR of the main six variants of FastICA for separation of two components with the same distribution, computed for a few distributions considered frequently in the literature, for sample size $N = 1000$. Here, the distribution "sinus" means the distribution of $\sqrt{2}\sin(u)$, where u is uniformly distributed in $(0, 2\pi)$, "bpsk" is the discrete distribution with values ±1, both with the probability 0.5, and $\mathrm{GG}(\alpha)$ means the generalized Gaussian distribution with parameter α, described in Appendix F. Note that the latter distribution is standard Gaussian for $\alpha = 2$, the Laplace distribution for $\alpha = 1$, sub-Gaussian for $\alpha > 2$, approaching the uniform distribution for $\alpha \to \infty$, and super-Gaussian (spiky) for $\alpha \to 0^+$.

Note that for separation of $d > 2$ components, the SIR would be $(d-1)\times 3$ dB lower than in the table, and if N is increased/decreased ten times, the resultant theoretical SIR is increased/decreased by 10 dB compared with the table.

A. *Example of Utilization*

In this subsection, the previous analysis is used to derive a novel variant of the FastICA algorithm, which combines advantages of both previously discussed variants. For easy reference, it will be called "Smart FastICA." This algorithm begins with applying symmetric FastICA with nonlinearity "tanh." For each estimated component signal $\hat{\mathbf{u}}_k^T$, parameters μ_k, ρ_k, and β_k are computed as sample estimates of the expectations in (19)–(21), namely $\hat{\mu}_k = \hat{\mathbf{u}}_k^T g(\hat{\mathbf{u}}_k)/N, \hat{\rho}_k = \mathbf{1}_N^T g'(\hat{\mathbf{u}}_k)/N, \hat{\beta}_k = \mathbf{1}_N^T g^2(\hat{\mathbf{u}}_k)/N$, and then they are plugged in (22) and (23) and (15), namely

$$\widehat{\mathrm{SIR}}_k^{(1U)} = \frac{N}{\sum_{\substack{\ell=1 \\ \ell\neq k}}^{d} \hat{V}_{k\ell}^{(1U)}}$$

[2] The quadratic convergence means that if the initial difference between the initial $\mathbf{W}$ and $\hat{\mathbf{W}}$ is $\boldsymbol{\Delta}\mathbf{W}$, the distance of $\mathbf{W}^+$ (that is $\mathbf{W}$ after one iteration) is $O(\|\boldsymbol{\Delta}\mathbf{W}\|^2)$.

$$\hat{\mathrm{SIR}}_k^{(\mathrm{SYM})} = \frac{N}{\sum_{\substack{\ell=1\\ \ell\neq k}}^{d} \hat{V}_{k\ell}^{(\mathrm{SYM})}}$$

If the obtained SIR_k for the one-unit algorithm is better than for the former estimate, the algorithm is performed, taking advantage of a more suitable nonlinearity g for each of particular cases: In the super-Gaussian case, defined by the condition $\hat{\mu_k} < \hat{\rho}_k$, the option "Gauss" is selected, and in the sub-Gaussian case with $\hat{\mu}_k > \hat{\rho}_k$, "pow3" is applied (see the simulation section for a reason).

Then, μ_k, ρ_k, β_k, and SIR_k are computed again. If the new SIR_k is better than the previous one and if, at the same time, the scalar product between the former separating vector and the new one is higher in absolute value than a constant (we have used 0.75), then the one=unit refinement is accepted in favor of the former vector. The condition on the scalar product is intended to eliminate the cases where the one=unit algorithm converged to a wrong component. A further optimization of the algorithm exceeds the scope of this paper.

B. Optimum Nonlinearity G

It is interesting to know, which function $g(\cdot)$ would be optimal for given probability density function (pdf) of s_k. If all source signals have the same distribution, the answer is well known. It is the so-called score function of the distribution, defined as $\psi(x) = -f'(x)/f(x)$, where $f(x)$ is the underlying pdf. Introduce the notation

$$\kappa = \mathrm{E}[\psi^2(\xi)] = \int_R \frac{f'^2(x)}{f(x)}\,dx \tag{24}$$

where ξ is a random variable with the pdf $f(\cdot)$. Note that if ξ has zero mean and variance one, it holds $\kappa \geq 1$, where the equality is attained if and only if the underlying distribution is standard Gaussian (see Appendix E). Thus, κ represents a measure of non-Gaussianity.

For the optimum nonlinearity $g_{\mathrm{opt}}(x) = \psi(x)$, a straightforward computation gives $\mu_k = 1$ and $\rho_k = \beta_k = \kappa$, and consequently

$$\min_g V_{k\ell}^{1U} = \frac{1}{\kappa - 1} \tag{25}$$

$$\min_g V_{k\ell}^{\mathrm{SYM}} = \frac{1}{4} + \frac{1}{2}\frac{1}{\kappa - 1}. \tag{26}$$

IV. Cramér–Rao Lower Bound for ICA

Consider a vector of parameters $\boldsymbol{\theta}$ being estimated from a data vector $\mathbf{x}$, having probability density $f_{\mathbf{x}\,|\,\boldsymbol{\theta}}(\mathbf{x}\,|\,\boldsymbol{\theta})$, using some unbiased estimator $\hat{\boldsymbol{\theta}}$. The CRB is the lower bound for the variance of $\hat{\boldsymbol{\theta}}$. Assume that $f_{\mathbf{x}\,|\,\boldsymbol{\theta}}$ is smooth and the following Fisher information matrix exists:

$$\mathbf{F}_{\boldsymbol{\theta}} = \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{1}{f_{\mathbf{x}\,|\,\boldsymbol{\theta}}^2}\frac{\partial f_{\mathbf{x}\,|\,\boldsymbol{\theta}}(\mathbf{x}\,|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\left(\frac{\partial f_{\mathbf{x}\,|\,\boldsymbol{\theta}}(\mathbf{x}\,|\,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right]. \tag{27}$$

Then, under some mild regularity condition,[18],[3] it holds

$$\mathrm{cov}\hat{\boldsymbol{\theta}} \geq \mathrm{CRB}_{\boldsymbol{\theta}} = \mathbf{F}^{-1}.$$

Next, if $\boldsymbol{\varphi} = \boldsymbol{\varphi}(\boldsymbol{\theta})$ is a differentiable function of $\boldsymbol{\theta}$, then the Fisher information matrix for $\boldsymbol{\varphi}$ exists as well and is equal to

$$\mathbf{F}_{\boldsymbol{\varphi}} = \mathbf{J}_{\boldsymbol{\theta}}^{-1}\mathbf{F}_{\boldsymbol{\theta}}\mathbf{J}_{\boldsymbol{\theta}}^{-T} \tag{28}$$

where $\mathbf{J}$ is the Jacobian of the mapping $\boldsymbol{\varphi}(\boldsymbol{\theta})$. If the mapping is linear, or $\boldsymbol{\varphi}(\boldsymbol{\theta}) = \mathbf{M}\boldsymbol{\theta}$ for some regular matrix $\mathbf{M}$, then $\mathbf{J}_{\boldsymbol{\theta}} = \mathbf{M}^T$.

In the context of ICA, we first focus on deriving the CRB for estimation of the demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$, i.e., the parameter vector is $\boldsymbol{\theta} = \mathrm{vec}[\mathbf{W}]$.

The following assumptions will be considered throughout this section:

$$\mathrm{E}\left[s_i^2\right] = \int_R t^2 f_i(t)\,dt = 1 \tag{29}$$

$$\kappa_i \overset{\mathrm{def}}{=} \mathrm{E}\left[\psi_i^2(s_i)\right] = \int_R \psi_i^2(t) f_i(t)\,dt < +\infty \tag{30}$$

$$\eta_i \overset{\mathrm{def}}{=} \mathrm{E}\left[s_i^2\psi_i(s_i)^2\right] = \int_R t^2\psi_i^2(t) f_i(t)\,dt < +\infty \tag{31}$$

where $i = 1,\ldots,d$ and ψ_i denotes the score function of the corresponding pdf, i.e., $\psi_i(x) = -(f_i'(x))/(f_i(x)) \cdot \psi_i$ is assumed to have zero mean for all i, and $f_i(x) > 0$ for all i and x.

A. Fisher Information Matrix

From the independence of the original signals, it follows that their joint pdf is $f_{\mathbf{S}}(\mathbf{S}) = \prod_{i=1}^{d}\prod_{j=1}^{N} f_i(s_{ij})$. Then, using the transformation $\mathbf{X} = \mathbf{W}^{-1}\mathbf{S}$

$$f_{\mathbf{x}}(\mathbf{X}) = |\det\mathbf{W}| f_{\mathbf{S}}(\mathbf{W}\mathbf{X}). \tag{32}$$

Incorporating this density into (27), the mnth element of the $d^2 \times d^2$ Fisher information matrix $\mathbf{F}_{\boldsymbol{\theta}}$, where $m = (i-1)d + j, n = (u-1)d + v$, and w_{ij} denotes the ijth element of the matrix $\mathbf{W}$, is

$$\mathbf{F}_{mn} = \mathrm{E}\left[\frac{|\det\mathbf{W}|^{-2}}{f_{\mathbf{S}}^2(\mathbf{S})}\frac{\partial f_{\mathbf{x}}}{\partial w_{ij}}\frac{\partial f_{\mathbf{x}}}{\partial w_{uv}}\right]. \tag{33}$$

A straightforward computation (see Appendix C) gives

$$\begin{aligned}\mathbf{F}_{mn} = (N-1)^2 a_{ji}a_{vu} + N a_{ju}a_{vi} + \delta_{iu} N a_{ji}a_{vi}(\eta_i - 2)\\ + \delta_{iu} N\kappa_i \sum_{\ell=1,\ell\neq u}^{d} a_{j\ell}a_{v\ell}\end{aligned} \tag{34}$$

with κ_i, η_i defined in (30) and (31), δ_{iu} is the Kronecker's delta, and a_{ij} denotes the ijth element of the mixing matrix $\mathbf{A}$. It can be shown, using (28), that

$$\mathbf{F}_{\boldsymbol{\theta}} = (\mathbf{A}\otimes\mathbf{I})\mathbf{F}_{\mathbf{I}}(\mathbf{A}^T\otimes\mathbf{I}) \tag{35}$$

[3] 1) Support of $f_{\mathbf{x}\,|\,\boldsymbol{\theta}}$ is independent of $\boldsymbol{\theta}$; 2) $\partial f_{\mathbf{x}\,|\,\boldsymbol{\theta}}(\mathbf{x}\,|\,\boldsymbol{\theta})/\partial\theta$ exists for all θ from an open set; and 3) $\mathrm{E}[\partial f_{\mathbf{x}\,|\,\boldsymbol{\theta}}(\mathbf{x}\,|\,\boldsymbol{\theta})/\partial\theta] = 0$

where $\mathbf{F_I}$ stands for the Fisher information matrix derived for a case when $\mathbf{A}=\mathbf{I}$ (identity matrix); $\otimes$ denotes the Kronecker product. Substituting $a_{ij}=\delta_{ij}$ into (34), it easily follows that

$$(\mathbf{F_I})_{mn} = (N-1)^2\delta_{ji}\delta_{vu} + N\delta_{ju}\delta_{vi} + N(\delta_{ji}\delta_{vu}\delta_{vi}(\eta_i-\kappa_i-2)+\delta_{iu}\delta_{vj}\kappa_i). \qquad (36)$$

Some properties of the matrix will be shown in Appendix D.

B. Accuracy of the Estimation of $\mathbf{G}_2=\hat{\mathbf{W}}\mathbf{A}$

Let $\hat{\mathbf{W}}$ denote an estimator of the demixing matrix $\mathbf{W}$. Estimated signals $\hat{\mathbf{S}}$ are then $\hat{\mathbf{S}}=\hat{\mathbf{W}}\mathbf{X}=\hat{\mathbf{W}}\mathbf{A}\mathbf{S}$. It is interesting to compute the CRB for the elements of the gain matrix $\mathbf{G}_2=\hat{\mathbf{W}}\mathbf{A}$, which is closely related to the gain matrix $\mathbf{G}$ defined in (14). A comparison of the definition relations gives $\mathbf{G}=\mathbf{G}_2\mathbf{D}^{1/2}$, where $\mathbf{D}$ contains, on its diagonal, sample variances of the original independent signal components. Asymptotically, $\mathbf{D}$ converges to unity matrix, and hence any estimate of $\mathbf{G}$ is at the same time an estimate of $\mathbf{G}_2$, and vice versa. In addition, it follows from the analysis in Appendix A that the asymptotic distribution of nondiagonal elements of $\mathbf{G}$ and those of $\mathbf{G}_2$ is the same.

To compute the CRB for $\mathbf{G}_2$, note that the new parameter vector $\boldsymbol{\theta}_\mathbf{G}=\mathrm{vec}[\mathbf{G}_2]$ is just a linear function of the parameter $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}_\mathbf{G}=\mathrm{vec}[\hat{\mathbf{W}}\mathbf{A}]=(\mathbf{A}^T\otimes\mathbf{I})\mathrm{vec}[\hat{\mathbf{W}}]=(\mathbf{A}^T\otimes\mathbf{I})\boldsymbol{\theta}$. Then, using (28), the Fisher information matrix of $\boldsymbol{\theta}_\mathbf{G}$ is

$$\mathbf{F_G}=(\mathbf{W}\otimes\mathbf{I})\mathbf{F}_{\boldsymbol{\theta}}(\mathbf{W}^T\otimes\mathbf{I})=\mathbf{F_I}. \qquad (37)$$

Note that $\mathbf{F_G}$ is independent of the mixing matrix $\mathbf{A}$. The CRB for the ijth element of $\mathbf{G}$ is

$$\mathrm{var}((\mathbf{G}_2)_{ij}) \geq \mathrm{CRB}((\mathbf{G}_2)_{ij}) = ({\mathbf{F_I}}^{-1})\mathrm{mm}$$

where $m=(i-1)d+j$ and $i\neq j$. In Appendix D, it is proved that for such m

$$({\mathbf{F_I}}^{-1})\mathrm{mm}=\frac{1}{N}\frac{\kappa_j}{\kappa_i\kappa_j-1} \qquad (38)$$

which gives us the desired lower bound

$$\mathrm{CRB}((\mathbf{G}_2)_{ij})=\frac{1}{N}\frac{\kappa_j}{\kappa_i\kappa_j-1}. \qquad (39)$$

The diagonal elements of $\mathbf{G}_2$ are not as important, they just reflect the accuracy of estimating the power of the components, or equivalently, the norm of rows of the demixing matrix.

V. Discussion

A. Comparison of CRB With Performance of FastICA With Optimum G

The Cramér–Rao lower bound in (39) is compared with the asymptotic variance of FastICA in (25) and (26) in Fig. 2. We can see that for κ close to 1, the CRB is close to the variance of the symmetric FastICA with the optimum nonlinearity. In this case, however, the estimation may fail, because the variance of the estimator itself goes to infinity, and convergence of the algorithm may be slow.

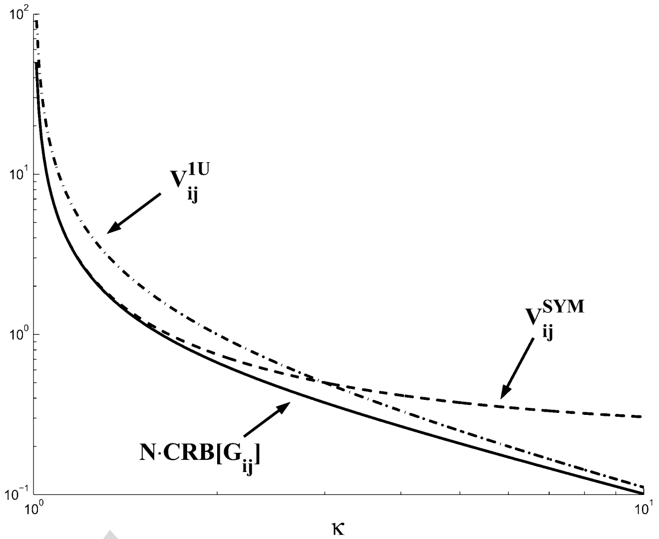


Fig. 2. Asymptotic performance of one=unit and symmetric FastICA and the CRB versus parameter κ.

In the opposite case, for $\kappa \gg 1$, the CRB asymptotically coincides with the variance of the one-unit FastICA with the optimum nonlinearity, because

$$\frac{\mathrm{var}\left[\mathbf{G}_{k\ell}^{1U\text{-opt}}\right]}{\mathrm{CRB}[\mathbf{G}_{k\ell}]} \approx \frac{N^{-1}V_{k\ell}^{1U\text{-opt}}}{\mathrm{CRB}[\mathbf{G}_{k\ell}]} \to 1 \quad \text{for } \kappa\to\infty.$$

We conclude that the FastICA algorithm with the optimum nonlinearity is asymptotically efficient in two cases: 1) one-unit version for $\kappa_i \gg 1$ and 2) symmetric version for $\kappa_i \to 1^+$ provided that all components have the same distribution law.

B. Separation of Sources With the Generalized Gaussian Distribution

Properties of the generalized Gaussian distribution are listed for easy reference in Appendix F. Note that the score function of this distribution is proportional to $|x|^{\alpha-1}\mathrm{sign}(x)$ so that $g(x)=|x|^{\alpha-1}\mathrm{sign}(x)$ is the theoretically optimum nonlinearity for the distribution. However, only for $\alpha>1$ is this function continuous and hence suitable nonlinearity for FastICA. For discontinuous g's, the algorithm appears not to converge.

C. Distributions With Finite Support

The CRB does not exist (the bound is infinite) for the bounded magnitude distributions such as "uniform," "sinus," and "bpsk" in Table II. It happens because these distributions do not have infinite support, as required for existence of the CRB. Since the uniform distribution is a limit of the GGD(α) for α going to infinity, it is natural to study FastICA with nonlinearity $g_k(x)=|x|^k\mathrm{sign}(x)$ with large k. It can be easily shown that the one=unit FastICA with this nonlinearity has asymptotic variance $V_{ij}^{1U}(k)\approx 3/(2k+1)$ that goes to zero for $k\to\infty$. Similar results can be obtained for the distribution "sinus." In other words, the asymptotic variance of FastICA cannot be lower bounded by any bound of the form B/N. Implications of the above observation for an adaptive choice of the nonlinearity exceed the scope of this paper.

D. Distributions With Long Tails

The CRB does not exists for the $\mathrm{GGD}(\alpha)$ distribution with with parameter $\alpha \leq 1/2$ (cf. lines 7 and 8 in Table II). These distributions are sometimes called "long tailed". Instead of the score function, let us consider the nonlinearity $g_{\alpha,k}(x) = x\exp[-(k|x|)^\alpha]$. This choice has the advantage, that the asymptotic variance of FastICA with this nolinearity can be computed analytically. The result is $V_{ij}^{1U}(\alpha,k) \approx 2^{-3/\alpha}(\beta_\alpha/k)^{1-2\alpha}$ for large k and $\alpha \leq 1/2$, with β_α defined in (93). Again, $V_{ij}^{1U}(\alpha,k)$ goes to zero for $k \rightarrow \infty$ and all $0 < \alpha < 1/2$. This explains nonexistence of the CRB in this case. Design of an FastICA-based algorithm taylored **<AUTHOR: Taylored or tailored?—ed.>**for long-tailed distributions exceeds the scope of this paper.

VI. Numerical Results

Example 1: Four independent random signals with generalized Gaussian distribution (see Appendix C) with parameter α and length $N = 5000$ were generated in 100 independent trials. The signals were mixed with a matrix that was randomly generated in each trial, and demixed again by eight variants of the algorithm: the symmetric FastICA with nonlinearities tanh, Gauss, pow3, and with the score function (dependent on α), as well as the one-unit FastICA with the same nonlinearities, implemented like smart FastICA. The resulting theoretical and empirical SIR is plotted in Fig. 3(a) and (b). An erratic behavior of the empirical results is experienced for small α and nonlinearity pow3. Here, the convergence of sample estimates of the expressions in (19)–(21) to their expectations is slow. We can see that among the α-independent nonlinearities, the "pow3" performs best in the case of $\alpha > 2$ that corresponds to the sub-Gaussian case, and "gauss" is the best one for $\alpha < 2$ where the distribution is super-Gaussian. FastICA with $g(\cdot)$ equal to the score function does not work properly (does not converge at all) for $\alpha \leq 1$, because the score function is not continuous for these α's.

Fig. 4 is similar, showing the relative efficiency of the eight methods compared with the corresponding CRB.

Example 2: In the second experiment, we have generated three different components with Gaussian, $\mathrm{GG}(\alpha)$, and Laplace distribution of the fixed length $N = 5000$ in 100 independent trials for each α. Signals were randomly mixed and separated by the symmetric FastICA and Smart FastICA with nonlinearity tanh. Note that this example includes the situation where the mixture includes two Gaussian distributions for $\alpha = 2$. The empirical and theoretical SIR are shown to agree very well. The Smart FastICA outperforms the symmetric version for such α when the one-unit approach has better variance than the symmetric one, and gives the same result otherwise.

Example 3: In the last experiment, we studied performance of two computationally extensive algorithms that are claimed to be more accurate than older algorithms: RADICAL [22] and NPICA [23]. We tested implementations available on the Internet and compared their performance with the CRB. The simulations are obtained from 50 independent separations of a signal of length $N = 1000$ with $d = 3$ components, all having the same distribution function, $\mathrm{GGD}(\alpha)$. In the neighborhood of the point $\alpha = 2$, the symmetric FastICA appears to outperform the other techniques. In general, it appears to give stable results unlike the NPICA.

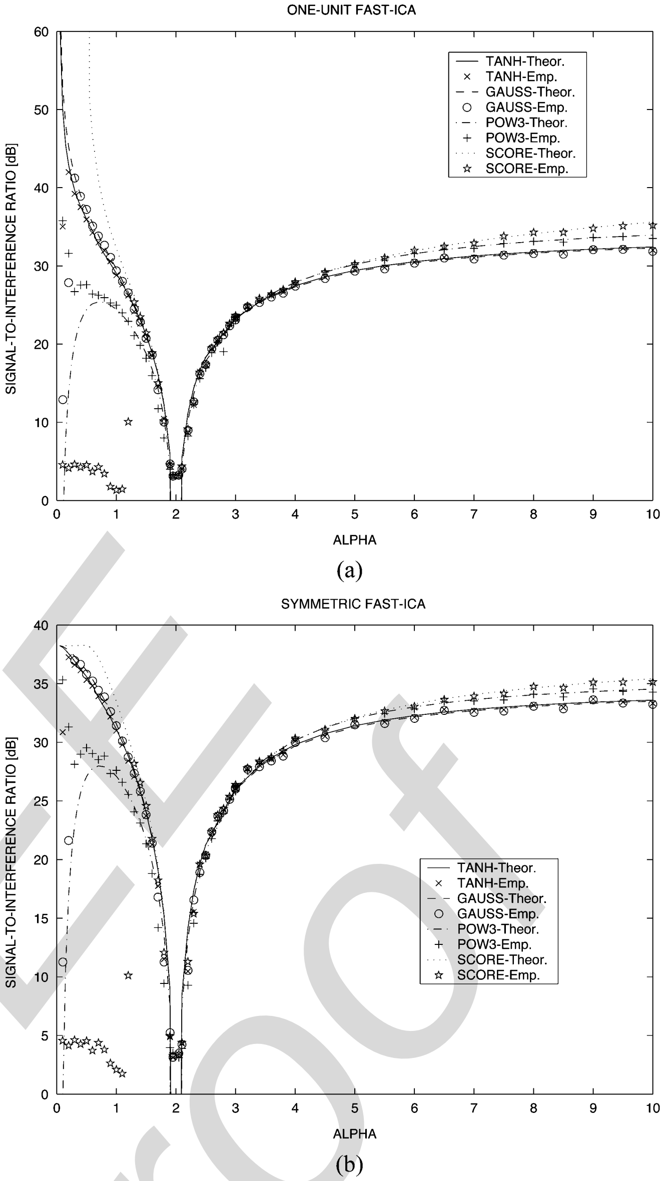
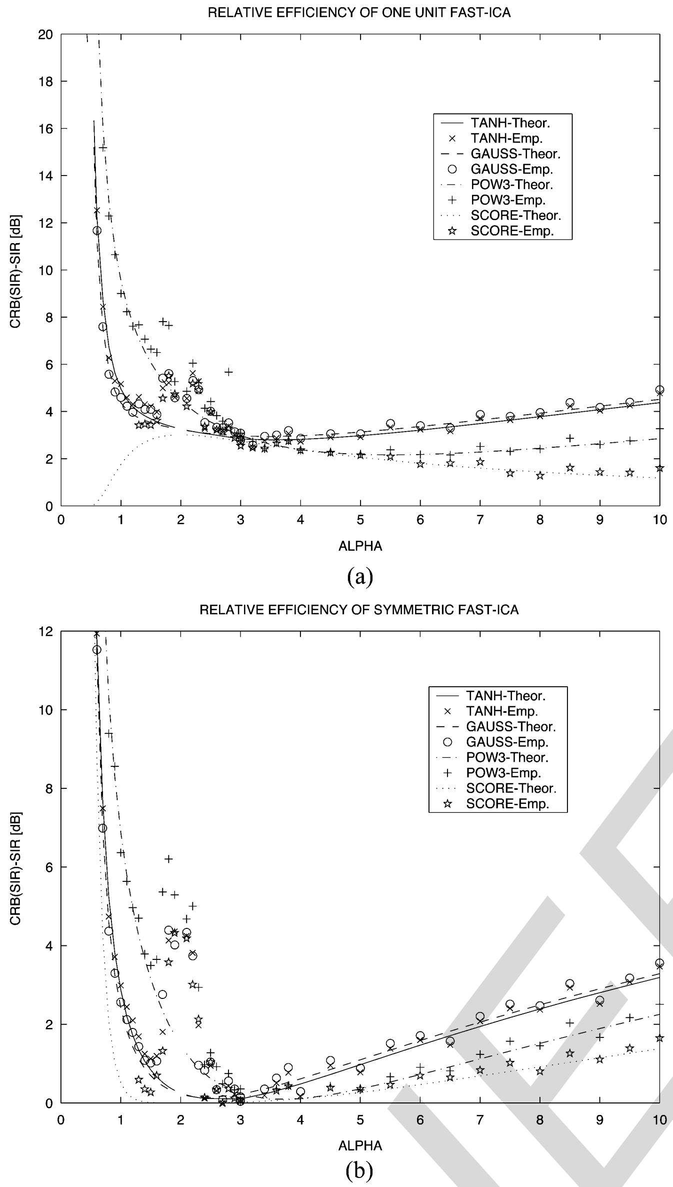


Fig. 3. Performance of (a) one-unit FastICA and (b) symmetric FastICA in separating signals with distribution $\mathrm{GG}(\alpha)$ as a function of α.

VII. Conclusion

In this paper, 1) a novel technique to improve stability of FastICA is proposed, 2) novel analytical expressions are derived for the variance of gain matrix elements for one-unit and symmetric FastICA, with an arbitrary twice differentiable nonlinear function and arbitrary probability distribution with finite variance of the independent components in the linear mixture, and 3) the Cramér–Rao bound for the above ICA problem is computed. The CRB does not exist for sources with bounded magnitude and for sources with long-tailed distribution. It was shown that asymptotic variance of estimates produced by FastICA with properly selected nonlinearity can approach the CRB, if the CRB exists, or aproach zero, if the CRB does not exist. Good



(a)

(b)

Fig. 4. Relative efficiency of (a) one-unit FastICA and (b) symmetric FastICA.

general performance of this popular algorithm is confirmed and possibilities of its further improvements are indicated.

Computer simulations confirm very well the validity of the theoretical predictions.

Appendix A
Proof of Proposition 1

A. Preliminaries

Invoking assumption (1) of the proposition, and the weak law of large numbers it follows that the sample variance of s_k defined in (4) converges to 1 in probability for N going to infinity, symbolically $\hat{\sigma}_k^2 \xrightarrow{p} 1$, or $\hat{\sigma}_k = 1 + o_p(1)$, where $o_p(\,\cdot\,)$ is the stochastic order symbol (see e.g., Appendix C in [31]). Similarly, thanks to the assumption (3),

$$N^{-1}\mathbf{s}_k^T g(\mathbf{s}_k) \xrightarrow{p} \mu_k \tag{40}$$

$$N^{-1}g'^T(\mathbf{s}_k)\mathbf{1}_N \xrightarrow{p} \rho_k. \tag{41}$$

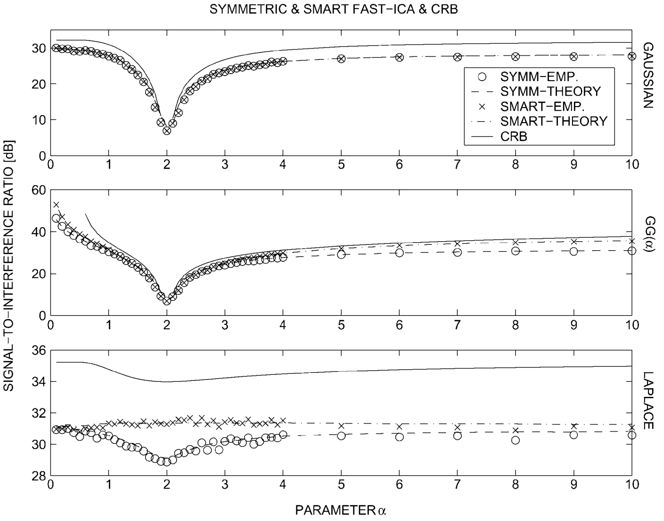


Fig. 5. Performance of symmetric FastICA and smart FastICA separating three different components using "tanh" nonlinearity.

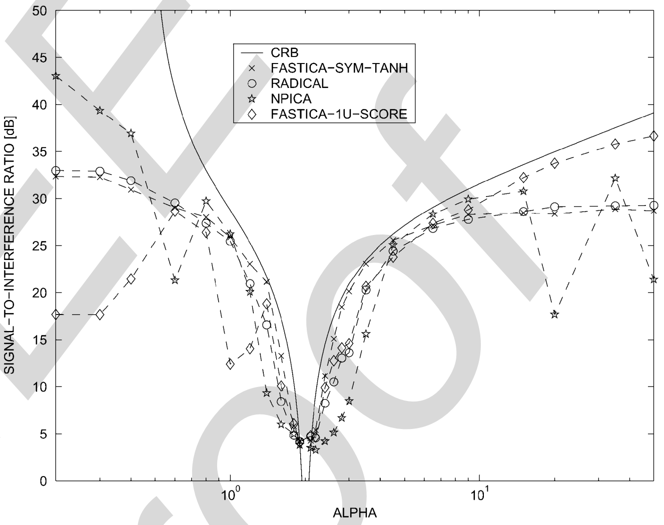


Fig. 6. Comparison of CRB with performance of four ICA techniques.

In addition, due to the mutual independence of components, it holds for $\ell \neq k$

$$N^{-1}g'^T(\mathbf{s}_k)(\mathbf{s}_\ell \odot \mathbf{s}_\ell) \xrightarrow{p} \mathrm{E}[g'(s_k)]\mathrm{E}[s_\ell^2] = \rho_k \tag{42}$$

where $\odot$ denotes the elementwise product. It can be shown, that the same limits are obtained if $\mathbf{s}_k, \mathbf{s}_\ell$ in (40)–(42) are replaced by the normalized components $\mathbf{u}_k, \mathbf{u}_\ell$, where $\mathbf{u}_k$ is the kth column of $\mathbf{U}, k = 1, \ldots, d$. Note from (2) that $\mathbf{u}_k = (\mathbf{s}_k - \bar{\mathbf{s}}_k)/\hat{\sigma}_k, \bar{\mathbf{s}}_k = O_p(N^{-1/2}), \hat{\sigma}_k = 1 + o_p(1)$, consequently $\mathbf{u}_k = \mathbf{s}_k + o_p(1), g(\mathbf{u}_k) = g(\mathbf{s}_k) + o_p(1)$, and

$$\begin{aligned}\mathbf{u}_k^T g(\mathbf{u}_k) &= [\mathbf{s}_k + o_p(1)]^T[g(\mathbf{s}_k) + o_p(1)]\\ &= \mathbf{s}_k^T g(\mathbf{s}_k) + o_p(N) = N\mu_k + o_p(N).\end{aligned} \tag{43}$$

Similarly, it can be shown that

$$g'^T(\mathbf{u}_k)\mathbf{1}_N = N\rho_k + o_p(N) \tag{44}$$

$$g'^T(\mathbf{u}_k)(\mathbf{u}_\ell \odot \mathbf{u}_\ell) = N\rho_k + o_p(N). \tag{45}$$

Moreover, using the asymptotic expression for $\mathbf{R}$, to be derived in the next subsection, it can be shown that the relations (40) and (41) hold true as well, if $\mathbf{s}_k$ is replaced with $\mathbf{z}_k$, that is defined as the kth column of $\mathbf{Z}, k = 1, \ldots, d$

$$\mathbf{z}_k^T g(\mathbf{z}_k) = N\mu_k + o_p(N) \tag{46}$$

$$g'^T(\mathbf{z}_k)\mathbf{1}_N = N\rho_k + o_p(N). \tag{47}$$

B. Asymptotic Behavior of $\mathbf{R}$

As N goes to infinity, the matrix $\mathbf{R}$ defined in (17) approaches identity matrix in the mean square sense. To see this, note that the diagonal elements of $\mathbf{R}$ are equal to one by definition, and that the off-diagonal elements $\mathbf{R}_{k\ell}$ with $k \neq \ell$ have zero mean. Due to assumed independence of $(\tilde{\mathbf{s}}_k, \hat{\sigma}_k)$ and $(\tilde{\mathbf{s}}_\ell, \hat{\sigma}_\ell)$, it holds

$$\begin{aligned}\mathrm{E}[\mathbf{R}_{k\ell}^2] &= \mathrm{E}\left(\frac{\mathbf{u}_k^T\mathbf{u}_\ell}{N}\right)^2 = \frac{1}{N^2}\mathrm{E}\left(\frac{\tilde{\mathbf{s}}_k^T\tilde{\mathbf{s}}_\ell}{\hat{\sigma}_k\hat{\sigma}_\ell}\right)^2\\ &= \frac{1}{N^2}\mathrm{E}\left\{\frac{\tilde{\mathbf{s}}_k^T}{\hat{\sigma}_k}\mathrm{E}\left(\frac{\tilde{\mathbf{s}}_\ell\tilde{\mathbf{s}}_\ell^T}{\hat{\sigma}_\ell^2}\right)\frac{\tilde{\mathbf{s}}_k}{\hat{\sigma}_k}\right\}\end{aligned} \tag{48}$$

where $\tilde{\mathbf{s}}_k = \mathbf{s}_k - \bar{\mathbf{s}}_k$. Let $\mathbf{S}^{(\ell)} = \mathrm{E}[\tilde{\mathbf{s}}_\ell\tilde{\mathbf{s}}_\ell^T/\hat{\sigma}_\ell^2]$. Since all elements of $\tilde{\mathbf{s}}_\ell$ have the same distribution, the diagonal elements of $\mathbf{S}^{(\ell)}$ have all the same value

$$\begin{aligned}\mathbf{S}_{nn}^{(\ell)} &= \mathrm{E}\left[\tilde{\mathbf{s}}_{\ell n}^2/\hat{\sigma}_\ell^2\right] = \frac{1}{N}\sum_{m=1}^{N}\mathrm{E}\left[\tilde{\mathbf{s}}_{\ell m}^2/\hat{\sigma}_\ell^2\right]\\ &= \frac{1}{N}\mathrm{E}\left[\tilde{\mathbf{s}}_\ell^T\tilde{\mathbf{s}}_\ell/\hat{\sigma}_\ell^2\right] = \mathrm{E}[1] = 1\end{aligned} \tag{49}$$

for $n = 1, \ldots, N$. The off-diagonal elements have all the same value as well

$$\begin{aligned}\mathbf{S}_{mn}^{(\ell)} &= \mathrm{E}\left[\tilde{\mathbf{s}}_{\ell m}\tilde{\mathbf{s}}_{\ell n}/\hat{\sigma}_\ell^2\right] = \frac{1}{N-1}\sum_{\substack{q=1\\ q\neq n}}^{N}\mathrm{E}\left[\tilde{\mathbf{s}}_{\ell q}\tilde{\mathbf{s}}_{\ell n}/\hat{\sigma}_\ell^2\right]\\ &= -\frac{1}{N-1}\mathrm{E}\left[\tilde{\mathbf{s}}_{\ell n}^2/\hat{\sigma}_\ell^2\right] = -\frac{1}{N-1}\end{aligned} \tag{50}$$

for $m, n = 1, \ldots, N$. Combining (48), (49), and (50) gives

$$\mathbf{S}^{(\ell)} = \frac{N}{N-1}\mathbf{I} - \frac{1}{N-1}\mathbf{1}_N\mathbf{1}_N^T \tag{51}$$

$$\begin{aligned}\mathrm{E}[\mathbf{R}_{k\ell}^2] &= \frac{1}{N^2}\mathrm{E}\left\{\frac{\tilde{\mathbf{s}}_k^T}{\hat{\sigma}_k}\mathbf{S}^{(\ell)}\frac{\tilde{\mathbf{s}}_k}{\hat{\sigma}_k}\right\}\\ &= \frac{1}{N(N-1)}\mathrm{E}\left\{\frac{\tilde{\mathbf{s}}_k^T}{\hat{\sigma}_k}\frac{\tilde{\mathbf{s}}_k}{\hat{\sigma}_k}\right\} = \frac{1}{N-1}.\end{aligned} \tag{52}$$

It follows from (52) that

$$\Delta\mathbf{R} \overset{\text{def}}{=} \mathbf{R} - \mathbf{I} = O_p(N^{-1/2}) \tag{53}$$

where $O_p(\,\cdot\,)$ denotes a standard stochastic order symbol, or a matrix of stochastic order symbols of appropriate dimension. Using Lemma 1 in Appendix B, it can be derived that

$$\mathbf{R}^{-1/2} = \mathbf{I} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1}). \tag{54}$$

C. Approximation for $\mathbf{Z}, g(\mathbf{Z})$

Obviously, $\mathbf{U} = O_p(1)$ and

$$\begin{aligned}\mathbf{Z} = \mathbf{R}^{-1/2}\mathbf{U} &= \left(\mathbf{I} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1})\right)\mathbf{U}\\ &= \mathbf{U} - \frac{1}{2}\Delta\mathbf{R}\mathbf{U} + O_p(N^{-1}).\end{aligned} \tag{55}$$

A Taylor series expansion of function $g(\,\cdot\,)$ in a neighborhood of $\mathbf{Z} = \mathbf{U}$ gives

$$g(\mathbf{Z}) = g(\mathbf{U}) + g'(\mathbf{U}) \odot \Delta\mathbf{Z} + O_p(N^{-1}) \tag{56}$$

where $\odot$ denotes the elementwise product and

$$\Delta\mathbf{Z} \overset{\text{def}}{=} \mathbf{Z} - \mathbf{U} = -\frac{1}{2}\Delta\mathbf{R}\mathbf{U} + O_p(N^{-1}). \tag{57}$$

Using (17), the kth column of $\Delta\mathbf{Z}^T$ is

$$\Delta\mathbf{z}_k = -\frac{1}{2N}\sum_{\substack{m=1\\ m\neq k}}^{d}\mathbf{u}_k^T\mathbf{u}_m\mathbf{u}_m + O_p(N^{-1}). \tag{58}$$

D. Approximation for $\mathbf{W}^+, \hat{\mathbf{W}}$

Inserting $\mathbf{W} = \mathbf{I}$ in (11), the $k\ell$th element of $\mathbf{W}^+$ reads

$$\mathbf{W}_{k\ell}^+ = \begin{cases} g(\mathbf{z}_k^T)\mathbf{z}_\ell & \text{for } k \neq \ell\\ g(\mathbf{z}_k^T)\mathbf{z}_k - g'(\mathbf{z}_k^T)\mathbf{1}_N & \text{for } k = \ell\end{cases} \tag{59}$$

For $k = \ell$, we get using (46) and (47)

$$\mathbf{W}_{kk}^+ = N(\mu_k - \rho_k) + o_p(N). \tag{60}$$

For $k \neq \ell$, we get using (56)

$$\begin{aligned}\mathbf{W}_{k\ell}^+ &= g\left(\mathbf{z}_k^T\right)\mathbf{z}_\ell\\ &= \left[g\left(\mathbf{u}_k^T\right) + g'\left(\mathbf{u}_k^T\right) \odot \Delta\mathbf{z}_k^T + O_p(N^{-1})\right]\left[\mathbf{u}_\ell + \Delta\mathbf{z}_\ell\right]\\ &= \left[g\left(\mathbf{u}_k^T\right) - \frac{1}{2N}\sum_{\substack{m=1\\ m\neq k}}^{d} g'\left(\mathbf{u}_k^T\right) \odot \left(\mathbf{u}_k^T\mathbf{u}_m\mathbf{u}_m^T\right)\right]\\ &\quad\cdot\left[\mathbf{u}_\ell - \frac{1}{2N}\sum_{\substack{m=1\\ m\neq \ell}}^{d}\mathbf{u}_\ell^T\mathbf{u}_m\mathbf{u}_m\right] + O_p(1).\end{aligned} \tag{61}$$

The reminder term in (61) has the stochastic order $O_p(1)$ for the following reason. It holds that $\mathbf{u}_k = O_p(1)$, and the remainder in the expansion of $g(\mathbf{z}_k^T)$, that is $O_p(N^{-1})$, are N-element vectors. The stochastic order notation is valid uniformly over elements of these vectors. Hence, scalar product of these two vectors is $O_p(1)$. Similarly, $\Delta\mathbf{z}_\ell = O_p(N^{-1/2}), g'(\mathbf{u}_k^T) \odot \Delta\mathbf{z}_\ell = O_p(N^{-1/2})$, and $(g'(\mathbf{u}_k^T) \odot \Delta\mathbf{z}_k^T)\Delta\mathbf{z}_\ell = O_p(1)$.

In the following, let $\mathbf{g}_k$ and $\mathbf{g}_k'$ stand for $g(\mathbf{u}_k)$ and $g'(\mathbf{u}_k)$, respectively, $k = 1, \ldots, d$. Note that, due to (21) and due to independence of $\mathbf{u}_k, \mathbf{u}_\ell$ for $k \neq \ell$, it holds

$$\begin{aligned}\mathrm{E}\left[(\mathbf{g}_k^T\mathbf{u}_\ell)^2\right] &= \mathrm{E}\left[\mathbf{g}_k^T\mathbf{u}_\ell\mathbf{u}_\ell^T\mathbf{g}_k\right] = \mathrm{E}\left[\mathbf{g}_k^T\mathrm{E}\left(\mathbf{u}_\ell\mathbf{u}_\ell^T\right)\mathbf{g}_k\right]\\ &= \mathrm{E}\left[\mathbf{g}_k^T\mathbf{g}_k\right] = N\beta_k.\end{aligned} \tag{62}$$

It follows from (19) and (62) that

$$\mathbf{g}_k^T\mathbf{u}_\ell = N\mu_k\delta_{k\ell} + o_p(N). \quad (63)$$

Similarly

$$\mathbf{u}_k^T\mathbf{u}_\ell = N\delta_{k\ell} + O_p(N^{1/2}). \quad (64)$$

Applying (63) and (64) and (43)–(45) in (61) gives

$$\begin{aligned}\mathbf{W}_{k\ell}^{+} &= \mathbf{g}_k^T\mathbf{u}_\ell - \frac{1}{2N}\mathbf{g}_k^T\mathbf{u}_k\mathbf{u}_k^T\mathbf{u}_\ell \\ &\quad - \frac{1}{2N}\left[\mathbf{g}_k'^T \odot \left(\mathbf{u}_k^T\mathbf{u}_\ell\mathbf{u}_\ell^T\right)\right]\mathbf{u}_\ell + O_p(1) \\ &= \mathbf{g}_k^T\mathbf{u}_\ell - \frac{1}{2N}\left(\mathbf{g}_k^T\mathbf{u}_k\right)\mathbf{u}_k^T\mathbf{u}_\ell \\ &\quad - \frac{1}{2N}\left(\mathbf{u}_k^T\mathbf{u}_\ell\right)\mathbf{g}_k'^T(\mathbf{u}_\ell\odot\mathbf{u}_\ell) + O_p(1) \\ &= \mathbf{g}_k^T\mathbf{u}_\ell - \frac{\mu_k+\rho_k}{2}\mathbf{u}_k^T\mathbf{u}_\ell + o_p(N^{1/2}). \quad (65)\end{aligned}$$

E. Approximation for $\hat{\mathbf{W}}, \mathbf{G}$

Note that if $\hat{\mathbf{W}}_{kk}^{+} < 0$ for some k, the kth diagonal element of the demixing matrices $\hat{\mathbf{W}}_{kk}^{1U}$ and $\hat{\mathbf{W}}_{kk}^{\mathrm{SYM}}$ may have the wrong sign, i.e., it might be close to -1 instead of 1. It corresponds to reversed sign of the kth estimated independent component. In the one-unit version of the algorithm, the sign can be corrected by replacing the normalization in (10) by an equivalent formula

$$\hat{\mathbf{W}}_{k\ell}^{1U} = \frac{\mathbf{W}_{k\ell}^{+}}{\mathbf{W}_{kk}^{+}} = \frac{\mathbf{W}_{k\ell}^{+}}{N(\mu_k-\rho_k)} + o_p(N^{-1/2}). \quad (66)$$

Similarly, using Lemma 2 in Appendix B, the asymptotically equivalent sign corrected expression for the estimated demixing matrix is

$$\begin{aligned}\hat{\mathbf{W}}_{k\ell}^{\mathrm{SYM}} &= \delta_{k\ell} + \frac{\mathbf{W}_{k\ell}^{+}\mathrm{sign}(\mathbf{W}_{kk}^{+}) - \mathbf{W}_{\ell k}^{+}\mathrm{sign}(\mathbf{W}_{\ell\ell}^{+})}{|\mathbf{W}_{kk}^{+}| + |\mathbf{W}_{\ell\ell}^{+}|} \\ &\qquad + o_p(N^{-1/2}). \quad (67)\end{aligned}$$

For both estimator variants,$\hat{\mathbf{W}}^{1U}$ and $\hat{\mathbf{W}}^{\mathrm{SYM}}$ we can write

$$\Delta\mathbf{W} = \hat{\mathbf{W}} - \mathbf{I} = O_p(N^{-1/2}). \quad (68)$$

Since

$$\begin{aligned}\mathbf{G} = \hat{\mathbf{W}}\mathbf{R}^{-1/2} &= (\mathbf{I}+\Delta\mathbf{W})\left(\mathbf{I} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1})\right) \\ &= \mathbf{I} + \Delta\mathbf{W} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1}) \quad (69)\end{aligned}$$

the gain matrix off-diagonal elements read

$$\mathbf{G}_{k\ell} = \hat{\mathbf{W}}_{k\ell} - \frac{1}{2N}\mathbf{u}_k^T\mathbf{u}_\ell + O_p(N^{-1}). \quad (70)$$

For the one-unit variant, we get

$$\begin{aligned}N^{1/2}\mathbf{G}_{k\ell}^{1U} &= N^{1/2}\frac{\mathbf{W}_{k\ell}^{+}}{N(\mu_k-\rho_k)} - \frac{1}{2N^{1/2}}\mathbf{u}_k^T\mathbf{u}_\ell + o_p(1) \\ &= \frac{N^{-1/2}}{\mu_k-\rho_k}(\mathbf{g}_k^T\mathbf{u}_\ell - \mu_k\mathbf{u}_k^T\mathbf{u}_\ell) + o_p(1). \quad (71)\end{aligned}$$

Finally, we show that (71) can be rewritten in terms of $\mathbf{s}_k, \mathbf{s}_\ell$ in an asymptotically equivalent formula

$$N^{1/2}\mathbf{G}_{k\ell}^{1U} = \frac{N^{-1/2}}{\mu_k-\rho_k}(g(\mathbf{s}_k^T)\mathbf{s}_\ell - \mu_k\mathbf{s}_k^T\mathbf{s}_\ell) + o_p(1). \quad (72)$$

To see that, note that

$$\begin{aligned}\mathbf{u}_k^T\mathbf{u}_\ell &= \left(\frac{\mathbf{s}_k - \bar{\mathbf{s}}_k}{\hat{\sigma}_k}\right)^T\frac{\mathbf{s}_\ell - \bar{\mathbf{s}}_\ell}{\hat{\sigma}_\ell} = \frac{\mathbf{s}_k^T\mathbf{s}_\ell - \bar{\mathbf{s}}_k^T\bar{\mathbf{s}}_\ell}{\hat{\sigma}_k\hat{\sigma}_\ell} \\ &= \frac{\mathbf{s}_k^T\mathbf{s}_\ell - O_p(1)}{1+o_p(1)} = \mathbf{s}_k^T\mathbf{s}_\ell + o_p(N^{1/2}). \quad (73)\end{aligned}$$

Similarly, it can be shown that

$$g(\mathbf{u}_k^T)\mathbf{u}_\ell = g(\mathbf{s}_k^T)\mathbf{s}_\ell + o_p(N^{1/2}). \quad (74)$$

Equation (74) concludes the proof of (72). Now, applying the central limit theorem to (72) implies that the distribution of $N^{1/2}\mathbf{G}_{k\ell}^{1U}$ is asymptotic normal with zero mean and variance equal to the variance of the leading term in (72). Using (62)–(64) gives

$$\begin{aligned}V_{k\ell}^{1U} &= \mathrm{var}\left[\frac{N^{-1/2}}{\mu_k-\rho_k}(g(\mathbf{s}_k^T)\mathbf{s}_\ell - \mu_k\mathbf{s}_k^T\mathbf{s}_\ell)\right] \\ &= \frac{N^{-1}}{(\mu_k-\rho_k)^2}\mathrm{var}\left[(g(\mathbf{s}_k^T)\mathbf{s}_\ell - \mu_k\mathbf{s}_k^T\mathbf{s}_\ell)\right] \\ &= \frac{\beta_k - \mu_k^2}{(\mu_k-\rho_k)^2}. \quad (75)\end{aligned}$$

Similarly, for symmetric FastICA, it holds using (67) that

$$\begin{aligned}N^{1/2}\mathbf{G}_{k\ell}^{\mathrm{SYM}} &= N^{1/2}\hat{\mathbf{W}}_{k\ell}^{\mathrm{SYM}} - \frac{1}{2N^{1/2}}\mathbf{u}_k^T\mathbf{u}_\ell + o_p(1) \\ &= \frac{\mathbf{W}_{k\ell}^{+}\mathrm{sign}(\mu_k-\rho_k) - \mathbf{W}_{\ell k}^{+}\mathrm{sign}(\mu_\ell-\rho_\ell)}{N^{1/2}(|\mu_k-\rho_k| + |\mu_\ell-\rho_\ell|)} \\ &\quad - \frac{1}{2N^{1/2}}\mathbf{u}_k^T\mathbf{u}_\ell + o_p(1). \quad (76)\end{aligned}$$

The variance of the leading term in (76) results, after some algebra using (63)–(65), in

$$V_{k\ell}^{\mathrm{SYM}} = \frac{\beta_k - \mu_k^2 + \beta_\ell - \mu_\ell^2 + (\mu_\ell-\rho_\ell)^2}{(|\mu_k-\rho_k| + |\mu_\ell-\rho_\ell|)^2} \quad (77)$$

as desired.

APPENDIX B
LEMMAS

Lemma 1: Let $\mathbf{R}_0$ and $\mathbf{R}$ be positive definite matrices of the same dimension and $\Delta\mathbf{R} = \mathbf{R} - \mathbf{R}_0$. Then, for $\|\Delta\mathbf{R}\| \to 0$ (in any matrix norm), it holds

$$\mathbf{R}^{-1/2} = \mathbf{R}_0^{-1/2} + \Delta\mathbf{M} + O(\|\Delta\mathbf{R}\|^2) \tag{78}$$

where

$$\Delta\mathbf{M} = -\text{unvec}\left\{\left[\mathbf{I} \otimes \mathbf{R}_0^{-1/2} + \mathbf{R}_0^{-1/2} \otimes \mathbf{I}\right]^{-1} \times \text{vec}\left(\mathbf{R}_0^{-1}\Delta\mathbf{R}\mathbf{R}_0^{-1}\right)\right\}. \tag{79}$$

Here, "vec" denotes the operation that reshapes columns of a matrix in one long column vector, and "unvec" is the corresponding inverse operation.

In the case that $\mathbf{R}_0$ is diagonal, $\mathbf{R}_0 = \text{diag}(r_1, \ldots, r_d)$ is a diagonal matrix with $r_k > 0$ for $k = 1, \ldots, d$, then $\Delta\mathbf{M}$ has elements

$$\Delta\mathbf{M}_{k\ell} = -\frac{\Delta\mathbf{R}_{k\ell}}{\sqrt{r_k}\sqrt{r_\ell}(\sqrt{r_k} + \sqrt{r_\ell})}. \tag{80}$$

In the case that $\mathbf{R}_0 = \mathbf{I}$, (80) gives $\Delta\mathbf{M} = -(1/2)\Delta\mathbf{R}$.

Proof: The identity

$$\mathbf{I} = \mathbf{R}(\mathbf{R}^{-1/2})^2 = (\mathbf{R}_0 + \Delta\mathbf{R})(\mathbf{R}_0^{-1/2} + \Delta\mathbf{M})^2 \tag{81}$$

leads, after neglecting higher than first-order terms in $\Delta\mathbf{R}$ and $\Delta\mathbf{M}$, to the relation

$$\Delta\mathbf{M}\mathbf{R}_0^{-1/2} + \mathbf{R}_0^{-1/2}\Delta\mathbf{M} = -\mathbf{R}_0^{-1}\Delta\mathbf{R}\mathbf{R}_0^{-1} \tag{82}$$

or, equivalently

$$\left[\mathbf{I} \otimes \mathbf{R}_0^{-1/2} + \mathbf{R}_0^{-1/2} \otimes \mathbf{I}\right] \text{vec}\Delta\mathbf{M} = -\text{vec}\left[\mathbf{R}_0^{-1}\Delta\mathbf{R}\mathbf{R}_0^{-1}\right].$$

The desired solution (79) follows. ∎

Lemma 2: Let

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} \tag{83}$$

where $\mathbf{W}_0 = \text{diag}(w_1, \ldots, w_d)$ is a diagonal matrix, and let $w_k > 0$ for $k = 1, \ldots, d$. Then, for $\|\Delta\mathbf{W}\| \to 0$ it holds

$$\mathbf{S} \overset{\text{def}}{=} (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W} = \mathbf{I} + \Delta\mathbf{S} + O(\|\Delta\mathbf{W}\|^2) \tag{84}$$

where $\Delta\mathbf{S}$ has elements

$$\Delta\mathbf{S}_{k\ell} = \frac{\Delta\mathbf{W}_{k\ell} - \Delta\mathbf{W}_{\ell k}}{w_k + w_\ell}. \tag{85}$$

Proof: Using Lemma 1 gives

$$(\mathbf{W}\mathbf{W}^T)^{-1/2} = \mathbf{V}_0 + \Delta\mathbf{V} + O(\|\Delta\mathbf{W}\|^2) \tag{86}$$

where

$$\mathbf{V}_0 = (\mathbf{W}_0\mathbf{W}_0^T)^{-1/2} = \text{diag}\left(\frac{1}{w_1}, \ldots, \frac{1}{w_d}\right) \tag{87}$$

and $\Delta\mathbf{V}$ has as elements

$$\begin{aligned}\Delta\mathbf{V}_{k\ell} &= -\frac{(\mathbf{W}\mathbf{W}^T - \mathbf{W}_0\mathbf{W}_0^T)_{k\ell}}{w_k w_\ell(w_k + w_\ell)} \\ &= -\frac{(\mathbf{W}_0\Delta\mathbf{W}^T + \Delta\mathbf{W}\mathbf{W}_0^T)_{k\ell}}{w_k w_\ell(w_k + w_\ell)} + O(\|\Delta\mathbf{W}\|^2) \\ &= -\frac{w_k\Delta\mathbf{W}_{\ell k} + \Delta\mathbf{W}_{k\ell}w_\ell}{w_k w_\ell(w_k + w_\ell)} + O(\|\Delta\mathbf{W}\|^2).\end{aligned} \tag{88}$$

Then

$$\begin{aligned}\mathbf{S} &= (\mathbf{V}_0 + \Delta\mathbf{V} + O(\|\Delta\mathbf{W}\|^2))(\mathbf{W}_0 + \Delta\mathbf{W}) \\ &= \mathbf{I} + \mathbf{V}_0\Delta\mathbf{W} + \Delta\mathbf{V}\mathbf{W}_0 + O(\|\Delta\mathbf{W}\|^2)\end{aligned} \tag{89}$$

and hence the leading term $\Delta\mathbf{S}$ has elements

$$\Delta\mathbf{S}_{k\ell} = \frac{1}{w_k}\Delta\mathbf{W}_{k\ell} + \Delta\mathbf{V}_{k\ell}w_\ell = \frac{\Delta\mathbf{W}_{k\ell} - \Delta\mathbf{W}_{\ell k}}{w_k + w_\ell}.$$

∎

APPENDIX C
COMPUTING FISHER INFORMATION MATRIX

Applying the fact that $(\partial \det \mathbf{W})/(\partial w_{ij}) = a_{ji} \det \mathbf{W}$, we get from (32)

$$\begin{aligned}\frac{\partial f_x}{\partial w_{uv}} &= \frac{\partial f_{\mathbf{s}}(\mathbf{W}\mathbf{x})|\det \mathbf{W}|}{\partial w_{uv}} \\ &= \frac{\partial|\det \mathbf{W}|}{\partial w_{uv}} f_{\mathbf{s}}(\mathbf{W}\mathbf{x}) + |\det \mathbf{W}|\frac{\partial f_{\mathbf{s}}(\mathbf{W}\mathbf{x})}{\partial w_{uv}} \\ &= |\det \mathbf{W}|a_{vu}f_{\mathbf{s}}(\mathbf{W}\mathbf{x}) + |\det \mathbf{W}| \\ &\quad \times \sum_{k=1}^{d}\sum_{l=1}^{N}\left(\prod_{\substack{i=1,\ldots,d \\ j=1,\ldots,N \\ \neg(i=k \wedge j=l)}} f_i((\mathbf{W}\mathbf{x})_{ij})\right)\frac{\partial f_k((\mathbf{W}\mathbf{x})_{kl})}{\partial w_{uv}} \\ &= |\det \mathbf{W}|a_{vu}f_{\mathbf{s}}(\mathbf{W}\mathbf{x}) + |\det \mathbf{W}| \\ &\quad \times \sum_{k=1}^{d}\sum_{l=1}^{N} f_{\mathbf{s}}(\mathbf{W}\mathbf{x})\frac{\frac{\partial}{\partial w_{uv}}f_k((\mathbf{W}\mathbf{x})_{kl})}{f_k((\mathbf{W}\mathbf{x})_{kl})}.\end{aligned}$$

Next

$$\begin{aligned}\frac{\partial f_k((\mathbf{W}\mathbf{x})_{kl})}{\partial w_{uv}} &= f_k'((\mathbf{W}\mathbf{x})_{kl})\frac{\partial(\mathbf{W}\mathbf{x})_{kl}}{\partial w_{uv}} \\ &= f_k'((\mathbf{W}\mathbf{x})_{kl})\sum_{\kappa=1}^{d}\frac{\partial w_{k\kappa}}{\partial w_{uv}}x_{\kappa l} \\ &= f_k'((\mathbf{W}\mathbf{x})_{kl})\delta_{ku}x_{vl} \\ &= f_k'((\mathbf{W}\mathbf{x})_{kl})\delta_{ku}\sum_{k=1}^{d}a_{vk}s_{kl}.\end{aligned}$$

Returning to the above formula, we get

$$\frac{\partial f_{\mathbf{x}}}{\partial w_{uv}} = |\det \mathbf{W}| f_{\mathbf{s}}(\mathbf{W}\mathbf{x}) \times \left[a_{vu} + \sum_{l=1}^{N} \sum_{k=1}^{d} \frac{f_u'((\mathbf{W}\mathbf{x})_{ul})}{f_u((\mathbf{W}\mathbf{x})_{ul})} a_{vk} s_{kl} \right].$$

From (1), it follows that $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} = \mathbf{W}\mathbf{x}$, and consequently

$$\frac{\partial f_{\mathbf{x}}}{\partial w_{uv}} = |\det \mathbf{W}| f_{\mathbf{s}}(\mathbf{s}) \left[a_{vu} + \sum_{l=1}^{N} \sum_{k=1}^{d} \frac{f_u'(s_{ul})}{f_u(s_{ul})} a_{vk} s_{kl} \right].$$

Using this, we can directly compute the mnth entry of the Fisher information matrix.

$$\begin{aligned}
\mathbf{F}_{mn} &= \mathrm{E}\left[\left(\frac{|\det \mathbf{W}|^{-2}}{f^2} \right) \frac{\partial f_{\mathbf{x}}}{\partial w_{uv}} \frac{\partial f_{\mathbf{x}}}{\partial w_{pr}} \right] \\
&= a_{vu} a_{rp} + a_{rp} \mathrm{E}\left[\sum_{l=1}^{N} \sum_{k=1}^{d} \frac{f_u'(s_{ul})}{f_u(s_{ul})} a_{vk} s_{kl} \right] \\
&\quad + a_{vu} \mathrm{E}\left[\sum_{i=1}^{N} \sum_{j=1}^{d} \frac{f_p'(s_{pi})}{f_p(s_{pi})} a_{rj} s_{ji} \right] \\
&\quad + \mathrm{E}\left[\sum_{l=1}^{N} \sum_{i=1}^{N} \sum_{k=1}^{d} \sum_{j=1}^{d} \frac{f_u'(s_{ul})}{f_u(s_{ul})} \frac{f_p'(s_{pi})}{f_p(s_{pi})} s_{kl} s_{ji} a_{vk} a_{rj} \right].
\end{aligned}$$

The second and the third term are equal to $-N a_{vu} a_{rp}$, because $\mathrm{E}[(f_u'(s_{ul}))/(f_u(s_{ul}))s_{kl}] = -\delta_{ku}$. To simplify the last term, we shall consider two cases:

1) $u \neq p$, then

$$\sum_{l=1}^{N} \sum_{i=1}^{N} \sum_{k=1}^{d} \sum_{j=1}^{d} \underbrace{\mathrm{E}\left[\frac{f_u'(s_{ul})}{f_u(s_{ul})} \frac{f_p'(s_{pi})}{f_p(s_{pi})} s_{kl} s_{ji} \right]}_{\delta_{ku}\delta_{jp} + \delta_{kp}\delta_{ju}\delta_{il}} a_{vk} a_{rj} = N^2 a_{vu} a_{rp} + N a_{vp} a_{ru}$$

2) $u = p$, then

$$\begin{aligned}
&\mathrm{E}\left[\sum_{l=1}^{N} \sum_{i=1}^{N} \sum_{k=1}^{d} \sum_{j=1}^{d} \frac{f_u'(s_{ul})}{f_u(s_{ul})} \frac{f_u'(s_{ui})}{f_u(s_{ui})} s_{kl} s_{ji} a_{vk} a_{rj} \right] \\
&= \sum_{i=1}^{N} \sum_{k=1}^{d} \sum_{j=1}^{d} \underbrace{\mathrm{E}\left[\left[-\frac{f_u'(s_{ui})}{f_u(s_{ui})} \right]^2 s_{ki} s_{ji} \right]}_{\delta_{kj} \times \ldots} a_{vk} a_{rj} \\
&\quad + \sum_{\substack{i,l=1 \\ i \neq l}}^{N} \sum_{k=1}^{d} \sum_{j=1}^{d} \underbrace{\mathrm{E}\left[\frac{f_u'(s_{ul})}{f_u(s_{ul})} \frac{f_u'(s_{ui})}{f_u(s_{ui})} s_{kl} s_{ji} \right]}_{\delta_{ku}\delta_{ju} \times \ldots} a_{vk} a_{rj} \\
&= \sum_{i=1}^{N} \underbrace{\mathrm{E}\left[-\frac{f_u'(s_{ui})}{f_u(s_{ui})} \right]^2}_{\mathrm{E}[\psi_u^2(\xi_u)]} \sum_{j=1 j \neq u}^{d} \underbrace{\mathrm{E}[s_{ji}^2]}_{1} a_{vj} a_{rj} \\
&\quad + a_{vu} a_{ru} \sum_{i=1}^{N} \underbrace{\mathrm{E}\left[-\frac{f_u'(s_{ui})}{f_u(s_{ui})} s_{ui} \right]^2}_{\mathrm{E}[\psi_u^2(\xi_u)\xi_u^2]} + \sum_{\substack{i,l=1 \\ i \neq l}}^{N} a_{vu} a_{ru} \\
&= N \left(\mathrm{E}[\psi_u^2(\xi_u)] \sum_{\substack{j=1 \\ j \neq u}}^{d} a_{vj} a_{rj} + \left(\mathrm{E}\left[\psi_u^2(\xi_u) \xi_u^2 \right] + (N-1) \right) a_{vu} a_{ru} \right).
\end{aligned}$$

Here, ξ_u denotes a random variable with pdf f_u, and ψ_u denotes its score function, i.e., $\psi_u(x) = -(f_u'(x)/f_u(x))$. After a few simplifications, (34) follows. ■

Appendix D
Computing Matrix Inversion of $\mathbf{F_I}$

Definition (36) can be rewritten as $\mathbf{F_I} = (N-1)^2 \mathbf{F}_1 + N(\mathbf{P} + \mathbf{\Sigma})$, where mnth element of $\mathbf{F}_1, \mathbf{P}$ and $\mathbf{\Sigma}$ are $\delta_{ji}\delta_{vu}, \delta_{ju}\delta_{vi}$, and $\delta_{ji}\delta_{vu}\delta_{vi}(\eta_i - \kappa_i - 2) + \delta_{iu}\delta_{vj}\kappa_i$, respectively, for $m = (i-1)d + j$ and $n = (u-1)d + v$. Note that $\mathbf{F}_1$ is a rank-one matrix, $\mathbf{F}_1 = \mathbf{e}\mathbf{e}^T$, where $\mathbf{e} = \mathrm{vec}(\mathbf{I})$. Applying the matrix inversion lemma gives

$$\mathbf{F_I}^{-1} = \frac{1}{N} \left[(\mathbf{P} + \mathbf{\Sigma})^{-1} - \frac{(\mathbf{P} + \mathbf{\Sigma})^{-1} \mathbf{e}\mathbf{e}^T (\mathbf{P} + \mathbf{\Sigma})^{-1}}{N(N-1)^{-2} + \mathbf{e}^T (\mathbf{P} + \mathbf{\Sigma})^{-1} \mathbf{e}} \right].$$

To compute the inversion $(\mathbf{P} + \mathbf{\Sigma})^{-1}$, note that $\mathbf{\Sigma}$ is diagonal

$$\mathbf{\Sigma} = \mathrm{diag}(\underbrace{\eta_1 - 2, \kappa_1, \ldots, \kappa_1}_{d}, \underbrace{\kappa_2, \eta_2 - 2, \kappa_2, \ldots, \kappa_2}_{d}, \ldots) \tag{90}$$

and $\mathbf{P}$ is a special permutation matrix such that $\mathbf{P}\mathrm{vec}(\mathbf{M}) = \mathrm{vec}(\mathbf{M}^T)$ for any $d \times d$ matrix $\mathbf{M}$. Moreover, $\mathbf{P}$ obeys $\mathbf{P}\mathbf{P} = \mathbf{I}$, and for any diagonal matrix $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ it holds that

$$\mathbf{P}\mathbf{D} = \mathbf{D}'\mathbf{P}$$

where $\mathbf{D}' = \mathrm{diag}(\mathbf{P}\mathbf{d}) = \mathbf{P}\mathbf{D}\mathbf{P}$. These facts can be used to show that the inversion of $\mathbf{P} + \mathbf{\Sigma}$ can be written in the form $\mathbf{D}_1 + \mathbf{D}_2\mathbf{P}$ for suitable diagonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$. The equality

$$(\mathbf{P} + \mathbf{\Sigma})(\mathbf{D}_1 + \mathbf{D}_2\mathbf{P}) = \mathbf{I}$$

is fulfilled for $\mathbf{\Sigma}\mathbf{D}_1 + \mathbf{D}_2' = \mathbf{I}$ and $\mathbf{D}_1' + \mathbf{\Sigma}\mathbf{D}_2 = \mathbf{0}$. Hence

$$\mathbf{D}_1 = (\mathbf{\Sigma}'\mathbf{\Sigma} - \mathbf{I})^{-1}\mathbf{\Sigma}' \quad \text{and} \quad \mathbf{D}_2 = -\mathbf{\Sigma}^{-1}\mathbf{D}_1'$$

where $\mathbf{\Sigma}' = \mathbf{P}\mathbf{\Sigma}\mathbf{P}$ and $\mathbf{D}_1' = \mathbf{P}\mathbf{D}_1\mathbf{P}$. Finally, it can be shown that $(\mathbf{F_I}^{-1})\mathrm{mm} = N^{-1}(\mathbf{D}_1)\mathrm{mm}$ for $m = (i-1)d + j, i \neq j$. (38) easily follows.

Appendix E
Proof That $\kappa \geq 1$

Assume that f is a positive probability density function of a random variable with zero mean and variance 1, such that κ in (24) exists. Then, integration *per partes* and the Cauchy–Schwartz inequality gives

$$\begin{aligned}1 &= \int_R f(x)\,dx = -\int_R x f'(x)\,dx \\ &\leq \sqrt{\int_R x^2 f(x)\,dx}\sqrt{\int_R \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\,dx} \\ &= 1\cdot\sqrt{\kappa}.\end{aligned} \tag{91}$$

The equality in (91) is attained if f'/f is proportional to x, which necessarily means that that the distribution is Gaussian.

Appendix F
Generalized Gaussian Distribution Family

Consider the generalized Gaussian density function with parameter α, zero mean and variance one, as [19]

$$f_\alpha(x) = \frac{\alpha\beta_\alpha}{2\Gamma(1/\alpha)}\exp\{-(\beta_\alpha|x|)^\alpha\} \tag{92}$$

where $\alpha > 0$ is a positive parameter that controls the distribution's exponential rate of decay, $\Gamma(\cdot)$ is the Gamma function, and

$$\beta_\alpha = \sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}. \tag{93}$$

This generalized Gaussian family encompasses the ordinary standard normal distribution for $\alpha = 2$, the Laplacean distribution for $\alpha = 1$, and the uniform distribution in the limit $\alpha \to \infty$.

The kth absolute moment for the distribution is

$$\mathrm{E}_\alpha\{|x|^k\} = \int_\infty^\infty |x|^k f_\alpha(x)dx = \frac{1}{\beta_\alpha^k}\frac{\Gamma\left(\frac{k+1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)}. \tag{94}$$

The score function of the distribution is

$$\psi_\alpha(x) = -\frac{\frac{\partial f_\alpha(x)}{\partial x}}{f_\alpha(x)} = \frac{|x|^{\alpha-1}\mathrm{sign}(x)}{\mathrm{E}_\alpha[|x|^\alpha]}. \tag{95}$$

Then, simple computations give

$$\begin{aligned}\kappa_\alpha &= \mathrm{E}_\alpha[\psi_\alpha^2(x)] = \frac{\mathrm{E}_\alpha[|x|^{2\alpha-2}]}{\{\mathrm{E}_\alpha[|x|^\alpha]\}^2} \\ &= \begin{cases}\frac{\Gamma\left(2-\frac{1}{\alpha}\right)\Gamma\left(\frac{3}{\alpha}\right)}{\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^2} & \text{for } \alpha > 1/2 \\ +\infty & \text{otherwise.}\end{cases}\end{aligned} \tag{96}$$

References

[1] S.-I. Amari and A. Cichocki, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[2] P. Comon, "Independent components analysis: A new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

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

[4] D. Donoho, "On minimum entropy deconvolution," in *Applied Time-Series Analysis II*. New York: Academic, 1981, pp. 565–609.

[5] P. Comon, "Contrasts for multichannel blind deconvolution," *IEEE Signal Process. Lett.*, vol. 3, no. 7, pp. 209–211, Jul. 1996.

[6] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 312–321, Mar. 1990.

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

[8] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.

[9] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Process.*, vol. 45, pp. 59–83, 1995.

[10] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[11] C. B. Papadias, "Globally convergent blind source separation based on a multiuser kurtosis maximization criterion," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3508–3519, Dec. 2000.

[12] A. Hyvärinen and E. Bingham, "A fast fixed-point algorithm for independent component analysis of complex-valued signals," *Int. J. Neural Syst.*, vol. 10, no. 1, pp. 1–8, 2000.

[13] P. A. Regalia and E. Kofidis, "Monotonic convergence of fixed-point algorithms for ICA," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 943–949, Jul. 2003.

[14] E. Oja, "Convergence of the symmetrical FastICA algorithm," in 9th Int. Conf. Neural Information Processing (ICONIP)**<AUTHOR: Location?—ed.>**, Nov. 18–22, 2002.

[15] S. Douglas, "On the convergence behavior of the FastICA algorithm," in *Proc. 4th Symp. Independent Component Analysis Blind Source Separation*, Nara, Japan, Apr. 2003, pp. 409–414.

[16] X. Giannakopoulos, J. Karhunen, and E. Oja, "Experimental comparison of neural algorithms for independent component analysis and blind separation," *Int. J. Neural Syst.*, vol. 9, pp. 651–656, 1999.

[17] A. Hyvärinen, "One-unit contrast functions for independent component analysis: A statistical analysis," in *Proc. IEEE Neural Networks for Signal Processing (NNSP) Workshop*, Amelia Island, FL, 1997, pp. 388–397.

[18] R. C. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley, 1973.

[19] K. Waheed and F. M. Salam, "Blind source recovery using an adaptive generalized Gaussian score function," in *Proc. 45th Midwest Symp. Circuits Systems (MWSCAS)*, vol. 3, Aug. 4–7, 2002, pp. 656–659.

[20] P. Tichavský and Z. Koldovský, "Optimal pairing of signal components separated by blind techniques," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 119–122, Feb. 2004.

[21] Z. Koldovský, P. Tichavský, and E. Oja, "Cramér–Rao lower bound for linear independent component analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. III, Philadelphia, PA, Mar. 2005, pp. 581–584.

[22] E. G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *J. Mach. Learning Res.*, vol. 4, pp. 1271–1295, 2003.

[23] R. Boscolo, H. Pan, and V. P. Roychowdhury, "Independent component analysis based on nonparametric density estimation," *IEEE Tran. Neural Netw.*, vol. 15, no. 1, pp. 55–65, Jan. 2004.

[24] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 90, no. 8, pp. 2009–2026, Oct. 98.

[25] ——, "On the performance of orthogonal source separation algorithms," in *Proc. EUSIPCO*, Edinburgh, U.K., Sep. 1994, pp. 776–779.

[26] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1712–1725, Jul. 1997.

[27] D. Yellin and B. Friedlander, "Multichannel system identification and deconvolution: performance bounds," *IEEE Trans. Signal Process.*, vol. 47, no. 5, pp. 1410–1414, May 1999.

[28] O. Shalvi and E. Weinstein, "Maximum likelihood and lower bounds in system identification with non-Gaussian inputs," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 328–339, Mar. 1994.

[29] V. Vigneron and Ch. Jutten, "Fisher information in source separation problems," in *Proc. ICA 2004****<AUTHOR: Please define ICA—ed.>***, Granada, Spain, pp. 168–176.

[30] A. M. Bronstein, M. M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi, "QML, blind deconvolution: Asymptotic analysis," in *Proc. ICA 2004*, Granada, Spain, pp. 677–684.

[31] B. Porat, *Digital Processing of Random Signals: Theory and Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

Petr Tichavský (M'98–SM'04) graduated from the **<AUTHOR: Please specify degree—ed.>**Czech Technical University, Prague, Czechoslovakia, in 1987 and received the Ph.D. degree in theoretical cybernetics from the Czechoslovak Academy of Sciences, Prague, in 1992.

Since that time, he has been with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague. He is author and coauthor of research papers in the area of sinusoidal frequency/frequency-rate estimation, adaptive filtering and tracking of time-varying signal parameters, algorithm-independent bounds on achievable performance, sensor-array processing, independent component analysis, and blind signal separation.

Dr. Tichavský received the Fulbright grant for a ten-month fellowship at the Department of Electrical Engineering, Yale University, New Haven, CT, In 1994 and the Otto Wichterle Award from the Academy of Sciences of the Czech Republic in 2002. He served as Associate Editor of the IEEE SIGNAL PROCESSING LETTERS from 2002 to 2004, and since 2005, has served as Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Zbyněk Koldovský (S'03–M'04) was born in Jablonec nad Nisou, Czech Republic, in 1979. He received the M.S. degree in mathematical modeling from the Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague, in 2002. He is currently working toward the Ph.D. degree with the Department of Mathematics, Czech Technical University, Prague.

He has also been with the Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic, Prague, since 2002. His main research interests are in audio processing, independent component analysis, and blind deconvolution.

Erkki Oja (S'75–M'78–SM'90–F'00) received the Dr.Sc. degree **<AUTHOR: From which institution?—ed.>** in 1977.

He is Director of the Neural Networks Research Centre and Professor of Computer Science at the Laboratory of Computer and Information Science, Helsinki University of Technology, Helsinki, Finland. He has been Research Associate at Brown University, Providence, RI, and visiting Professor at the Tokyo Institute of Technology, Tokyo, Japan. He is the author or coauthor of more than 280 articles and book chapters on pattern recognition, computer vision, and neural computing, and three books: *Subspace Methods of Pattern Recognition* (New York: RSP and Wiley, 1983)**<AUTHOR: Please provide full name for RSP and location—ed.>**, which has been translated into Chinese and Japanese; *Kohonen Maps* (Amsterdam, The Netherlands: Elsevier, 1999); and *Independent Component Analysis* (New York: Wiley, 2001). His research interests are in the study of principal component and independent component analysis, self-organization, statistical pattern recognition, and applying artificial neural networks to computer vision and signal processing.

Prof. Oja is a member of the Finnish Academy of Sciences, Founding Fellow of the International Association of Pattern Recognition (IAPR), and President of the European Neural Network Society (ENNS). He is also a member of the editorial boards of several journals and has been in the program committees of several recent conferences, including ICANN, IJCNN, and International Conference Neural Information Processing (ICONIP).**<AUTHOR: Please define conference names—ed.>**