Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound

Zbyněk Koldovský[1,2], Petr Tichavský[2] and Erkki Oja[3]

[1] Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University, Trojanova 13, 120 00 Prague 2

[2] Institute of Information Theory and Automation, Pod vodárenskou věží 4,
P.O. Box 18, 182 08 Praha 8, Czech Republic
E-mail: {koldovsk,tichavsk}@utia.cas.cz

[3] Neural Networks Research Centre, Helsinki University of Technology,
P.O. Box 5400, 02015 HUT, Finland
E-mail: Erkki.Oja@hut.fi

Abstract

FastICA is one of the most popular algorithms for Independent Component Analysis, demixing a set of statistically independent sources that have been mixed linearly. A key question is how accurate the method is for finite data samples. We propose an improved version of the FastICA algorithm which is asymptotically efficient, i.e., its accuracy given by the residual error variance attains the Cramér-Rao lower bound. The error is thus as small as possible. This result is rigorously proven under the assumption that the probability distribution of the independent signal components belongs to the class of generalized Gaussian distributions with parameter α, denoted GG(α) for $\alpha > 2$. We name the algorithm Efficient

[0]Part of this paper is to be presented at the IEEE SSP'05 Workshop in Bordeaux. This work was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M6798555601 and by the Czech Technical University in Prague through the project CTU0508214.

FastICA. Its computational complexity is shown to be only slightly, about three times higher than that of the standard symmetric FastICA. Simulations corroborate these claims and show superior performance of the algorithm compared with JADE and Non-Parametric ICA on separating sources with distribution GG(α) with arbitrary α, as well as on sources with bi-modal distribution, and a good performance in separating linearly mixed speech signals.

Index Terms

Independent component analysis, blind source separation, blind deconvolution, Cramér-Rao lower bound, algorithm FastICA

I. Introduction

Recently, blind techniques such as blind source separation have become popular in the signal processing and machine learning community. One of the central tools for this problem is Independent Component Analysis (ICA) [1], [2]. In this technique, a set of original source signals are retrieved from their mixtures based on the assumption of their mutual statistical independence. The simplest case for ICA is the instantaneous linear noiseless mixing model. In this case, the mixing process can be expressed as

$$\mathbf{X} = \mathbf{AS}, \tag{1}$$

where $\mathbf{X}$ is an $d \times N$ data matrix. The rows of $\mathbf{X}$ are the observed mixed signals, thus d is the number of mixed signals and N is their length or the number of samples in each signal. Similarly, the unknown $d \times N$ matrix $\mathbf{S}$ includes samples of the original source signals. $\mathbf{A}$ is an unknown regular $d \times d$ mixing matrix. It is assumed square because the number of mixtures and sources can always be made equal in this simple model.

A basic assumption in ICA is that the elements of $\mathbf{S}$, denoted s_{ij}, are mutually independent i.i.d. random variables with probability density functions (pdf) $p_i(s_{ij})$ $i = 1, \ldots, d$. The row variables s_{ij} for all $j = 1, \ldots, N$, having the same density, are thus an i.i.d. sample of one of the independent sources denoted by s_i. The key assumptions for the identifiability of the model (1), or solving both $\mathbf{A}$ and $\mathbf{S}$ up to some simple ambiguities, are that all but at most one of the densities $p_i(\cdot)$ are non-Gaussian, and the unknown matrix $\mathbf{A}$ has full rank. In the following, let $\mathbf{W}$ denote the demixing matrix, $\mathbf{W} = \mathbf{A}^{-1}$.

The basic ICA problem and its extensions and applications have been studied widely and many algorithms have been developed. One of the main differences is how the unknown density functions $p_i(\cdot)$ of the original signals are estimated or replaced by suitable nonlinearities in the ICA contrast functions. Non-Gaussianity is the key property. For instance, JADE [4] is based on the estimation of kurtosis via cumulants, NPICA [10] uses a nonparametric model of the density functions, and RADICAL [11] uses an approximation of the entropy of the densities based on order statistics. The FastICA algorithm uses either kurtosis [5] or other measures of non-Gaussianity in entropy approximations in the form of suitable nonlinear functions $G(\cdot)$ [6].

In spite of the success of ICA in solving even large-scale real world problems, some theoretical questions remain partly open. One of the most central questions is the theoretical accuracy of the developed algorithms. Mostly the methods are compared through empirical studies, which may demonstrate the efficacy in various situations. However, the general validity cannot be proven like this. A natural question is, whether it is possible to reach a Cramér Rao lower bound for separation performance, which is widely accepted as measure of efficiency of estimators.

Many of the algorithms can be shown to converge in theory to the correct solution giving the original sources, under the assumption that the sample size N is infinite. This is unrealistic. For finite data sets, like in the model (1), what typically happens is that the sources are not completely unmixed but some traces of the other sources remain in them. This means that the obtained demixing matrix $\widehat{\mathbf{W}}$ is not exactly the inverse of $\mathbf{A}$, and the matrix of estimated sources $\mathbf{Y} = \widehat{\mathbf{W}}\mathbf{X} = \widehat{\mathbf{W}}\mathbf{A}\mathbf{S}$ is only approximately equal to $\mathbf{S}$. A natural measure of error is the deviation of matrix $\widehat{\mathbf{W}}\mathbf{A}$ from the unit matrix, i.e., the variances of its elements.

The present authors published recently an asymptotic performance analysis of the FastICA algorithm in [7], deriving the exact expression for this error variance. Furthermore, it is compared with the Cramér-Rao lower bound (CRB) for ICA (see [3], [8], [14], [16], [17], [18], [21] and references therein) and showed that the accuracy of FastICA is very close, but not equal to, the CRB. The condition for this is that the nonlinearity $G(\cdot)$ in the FastICA contrast function is the integral of the score function $\psi(s)$ of the original signals, or the negative log density, i.e.,

$$G(s) = \int \psi(s)ds = -\int \frac{p_i'(s)}{p_i(s)}ds = -\log p_i(s). \tag{2}$$

The purpose of this paper is to use this analysis to generalize the FastICA algorithm to an improved version so that it becomes asymptotically Fisher efficient, i.e., that the residual error

variance becomes equal to the CRB. When the asymptotic performance achieves the CRB we have reached the absolute accuracy that cannot be improved. We call this new variant Efficient FastICA.

As far as the authors know, there have been few efforts in developing concrete ICA algorithms based on the concept of asymptotic efficiency. A notable approach in this direction is the method of Estimating functions [19], [20], matrix valued functions $\mathbf{F}(\mathbf{X}, \mathbf{W})$ such that their root $\widehat{\mathbf{W}}(\mathbf{X})$ is an estimator for the true demixing matrix $\mathbf{W}$. Amari and Cardoso [20] derived an optimal class of estimating functions whose roots $\widehat{\mathbf{W}}$ are Fisher-efficient. The theory is general and can be applied to stochastic approximation-type learning algorithms.

Another related paper is [18], which studies asymptotic performance of so-called quasi-maximum likelihood estimate (quasi-MLE). In this method, the true probability density function of sources is replaced by an ad-hoc model density. If the model density coincide with the true density, an asymptotically efficient estimate would be obtained.

The contents of this paper are as follows. In the next Section 2, the results of our previous work are briefly summarized. In Section 3, the improved algorithm is derived and its properties are described. In order to demonstrate the efficiency in practice, Section 4 presents computer simulations. The simulations confirm the excellent performance of the Efficient FastICA algorithm and also show that the computational complexity is only about three times that of standard FastICA, which is one of the fastest ICA algorithms. Finally, in Section 5 conclusions are given.

II. Recent results

A. The original FastICA algorithm

The algorithm FastICA was introduced in [5], [6] in two versions: a one-unit approach and a symmetric one. The first preprocessing step, which is common for both versions and for many other ICA algorithms, consists of removing the sample mean and decorrelating the data $\mathbf{X}$, i.e.,

$$\mathbf{Z} \;=\; \widehat{\mathbf{C}}^{-1/2}\,(\mathbf{X} - \overline{\mathbf{X}}) \tag{3}$$

where $\widehat{\mathbf{C}}$ is the sample covariance matrix, $\widehat{\mathbf{C}} = (\mathbf{X}-\overline{\mathbf{X}})(\mathbf{X}-\overline{\mathbf{X}})^T/N$ and $\overline{\mathbf{X}}$ is the sample mean of the mixture data. Now matrix $\mathbf{Z}$ contains the whitened mixtures. The one-unit FastICA algorithm is based on minimization/maximization of the criterion $c(\mathbf{w}) = \mathrm{E}[G(\mathbf{w}^T\mathbf{Z})]$. There $\mathbf{w}$ is the to-be found unitary vector of coefficients that separates one of the independent components from the

mixture $\mathbf{Z}$ (one row of the separating matrix $\widehat{\mathbf{W}}(\mathbf{Z})$). Function $G(\cdot)$ is a suitable nonlinearity, called contrast function, applied elementwise to the row vector $\mathbf{w}^T\mathbf{Z}$; see [1]. The expectation E stands for the sample mean over the elements of the row vector. It is not known in advance which component is being estimated: this mainly depends on the initialization.

To retrieve all the original components, different rows of $\widehat{\mathbf{W}}(\mathbf{Z})$ are estimated under the orthogonality condition, i.e. $\widehat{\mathbf{W}}\widehat{\mathbf{W}}^T = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. In the one-unit deflation method, the independent components are found one by one, and the weight vector $\mathbf{w}$ is always constrained to be orthogonal to the previously found ones. In the symmetric FastICA, the condition is ensured via a symmetric orthogonalization after parallel one-unit iterations:

$$\mathbf{W}^+ \leftarrow g(\mathbf{WZ})\mathbf{Z}^T - \text{diag}[g'(\mathbf{WZ})\mathbf{1}_N]\,\mathbf{W} \tag{4}$$

$$\mathbf{W} \leftarrow (\mathbf{W}^+\mathbf{W}^{+T})^{-1/2}\mathbf{W}^+. \tag{5}$$

There $g(\cdot)$ and $g'(\cdot)$ denote the first and the second derivatives of $G(\cdot)$, respectively, applied elementwise, and $\mathbf{1}_N$ stands for an $N \times 1$ vector of 1's.

The FastICA algorithm is computationally light, robust, and converges very fast. It is available in public-domain software [12]. Recently, it was proposed to complete the symmetric FastICA by a test of saddle points that eliminates convergence to side minima of the cost function, which may occur for some nonlinearities $G(\cdot)$ [7].

An essential question is the residual error of the algorithm, due to a finite sample of the mixture signals. Let $\widehat{\mathbf{W}}$ be the estimate of the demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ obtained with the FastICA algorithm. The separation quality is analyzed by means of the so called gain matrix, $\mathbf{G} = \widehat{\mathbf{W}}\mathbf{A}$. Theoretically, $\mathbf{G}$ is the unit matrix, but for finite sample sizes there is residual error. The elements of $\mathbf{G}$ characterize the relative remaining presence of the $j-$th original signal component in the estimated $k-$th component, $j, k = 1, \ldots, d$. In the following, their asymptotic variance is compared to the theoretically optimal one.

B. Analysis of FastICA and the Cramér-Rao bound

Let $\mathbf{G}^{1U}$ and $\mathbf{G}^{SYM}$, respectively, be the gain matrix obtained by the one-unit and the symmetric variant of FastICA using a nonlinear function $g(\cdot)$. The main result shown in [7] was the following: Assume that the original signals s_k in the mixture have zero mean and unit

variance, that g is sufficiently smooth, and the following expectations exist:

$$\mu_k \stackrel{\text{def}}{=} \mathrm{E}[s_k g(s_k)] \tag{6}$$

$$\rho_k \stackrel{\text{def}}{=} \mathrm{E}[g'(s_k)] \tag{7}$$

$$\beta_k \stackrel{\text{def}}{=} \mathrm{E}[g^2(s_k)] \tag{8}$$

Then the normalized gain matrix elements $N^{1/2}\mathbf{G}_{k\ell}^{1U}$ and $N^{1/2}\mathbf{G}_{k\ell}^{SYM}$ have asymptotically Gaussian distributions $\mathcal{N}(0, V_{k\ell}^{1U})$ and $\mathcal{N}(0, V_{k\ell}^{SYM})$, with variances

$$V_{k\ell}^{1U} = \frac{\gamma_k}{\tau_k^2} \tag{9}$$

$$V_{k\ell}^{SYM} = \frac{\gamma_k + \gamma_\ell + \tau_\ell^2}{(\tau_k + \tau_\ell)^2}, \tag{10}$$

where

$$\gamma_k = \beta_k - \mu_k^2 \tag{11}$$

$$\tau_k = |\mu_k - \rho_k|. \tag{12}$$

The variances in (9,10) are minimized if the function $g(s)$ equals the score function

$$\psi_k(s) = -\frac{d}{ds}\log p_k(s) = -\frac{p_k'(s)}{p_k(s)} \tag{13}$$

of the corresponding source distribution $p_k(s)$. The minimum variance can be shown to be close, but not to coincide with the Cramér-Rao lower bound (CRB) derived in [4]. It was shown that the CRB is

$$\mathrm{CRB}(\mathbf{G}_{k\ell}) = \frac{1}{N}\frac{\kappa_k}{\kappa_k\kappa_\ell - 1} \tag{14}$$

where $\kappa_k \stackrel{\text{def}}{=} \mathrm{E}\left[\psi_k^2(s_k)\right]$.

III. Efficient FastICA

In this section, first, we define an auxiliary algorithm, called generalized symmetric FastICA, whose asymptotic properties suggest how to optimize the performance of FastICA. Then, the proposed Efficient FastICA method is described.

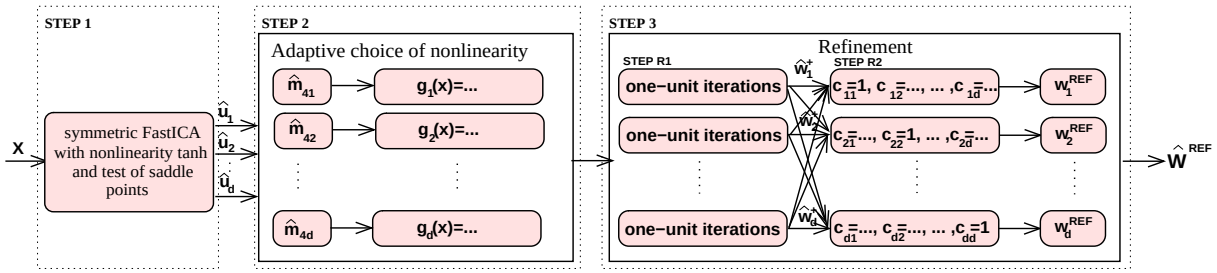


Fig. 1. Flow of the proposed algorithm Efficient FastICA.

A. *Generalizing the symmetric FastICA to attain the Cramér-Rao bound*

Consider now a version of the symmetric version of FastICA where two changes have been made.

First, as it is not possible to attain the CRB if only one nonlinearity $g(\cdot)$ is used, different nonlinear functions $g_k(\cdot)$, $k = 1, 2, \ldots, d$ will be used for estimation of each row of $\mathbf{W}^+$. Denote $g(\mathbf{WZ}) = [g_1(\mathbf{w}_1^T\mathbf{Z}), g_2(\mathbf{w}_2^T\mathbf{Z}), \ldots, g_d(\mathbf{w}_d^T\mathbf{Z})]^T$ where again each function g_j is applied elementwise. The function $g'(\mathbf{WZ})$ is defined likewise. Eventually, the functions $g_k(\cdot)$ should be the score functions of the sources $\mathbf{s}_k$.

Second, the first step (4) of the iteration will be followed by multiplying each row of $\mathbf{W}^+$ with a suitable positive number c_i $i = 1, \ldots, d$ before the symmetric orthogonalization (5). This will change the length (norm) of each row, which will effect the orientations of the rows after orthonormalization.

One iteration of the new Generalized symmetric FastICA algorithm, with the new definition of $g(\cdot)$, can then be written in three steps:

$$\mathbf{W}^+ \leftarrow g(\mathbf{WZ})\mathbf{Z}^T - \mathrm{diag}[g'(\mathbf{WZ})\mathbf{1}_N]\,\mathbf{W} \tag{15}$$

$$\mathbf{W}^+ \leftarrow \mathrm{diag}[c_1, \ldots, c_d] \cdot \mathbf{W}^+ \tag{16}$$

$$\mathbf{W} \leftarrow (\mathbf{W}^+\mathbf{W}^{+T})^{-1/2}\mathbf{W}^+ \tag{17}$$

This algorithm can be analyzed in the same way as the plain symmetric FastICA in [7]. The result is that the non-diagonal normalized gain matrix elements for this method, $N^{1/2}\mathbf{G}_{k\ell}^{GS}$, have asymptotically Gaussian distribution $\mathcal{N}(0, V_{k\ell}^{GS})$, where

$$V_{k\ell}^{GS} = \frac{c_k^2\gamma_k + c_\ell^2(\gamma_\ell + \tau_\ell^2)}{(c_k\tau_k + c_\ell\tau_\ell)^2}. \tag{18}$$

Note that the different choice of nonlinearity for each row of $\mathbf{W}$ also changes the definitions (6)-(8) for μ_k, ρ_k, and β_k in the sense that g in those definitions must be replaced by g_k. The definitions for γ_k and τ_k in (11),(12) remain the same but now they, also, depend on their own nonlinearity $g_k(\cdot)$.

The key thing here is that, since $c_1, \ldots, c_d$ are arbitrary positive numbers, the criterion (18) characterizing the asymptotic error variance of the algorithm can be optimized in terms of these free parameters. The properties of (18) are as follows:

- The variance is invariant with respect to multiplying all parameters c_k by the same factor. Therefore, for a fixed k let c_k be chosen, without any loss of generality, as $c_k = 1$. Let us change the notation so that instead of c_ℓ, we use $c_{k,\ell}$ to denote the other parameters in the case that $c_k = 1$.
- Minimization of (18) with respect to $c_{k,\ell}$, $\ell \neq k$, can be performed analytically in a straightforward way, and it gives

$$c_{k,\ell}^{OPT} = \arg \min_{c_\ell, c_k = 1} V_{k\ell}^{GS} = \frac{\tau_\ell \gamma_k}{\tau_k(\gamma_\ell + \tau_\ell^2)} \ . \tag{19}$$

- Combining (18) and (19), the optimum value of the criterion is

$$V_{k\ell}^{OPT} = \min_{c_\ell, c_k = 1} V_{k\ell}^{GS} = \frac{\gamma_k(\gamma_\ell + \tau_\ell^2)}{\tau_\ell^2 \gamma_k + \tau_k^2(\gamma_\ell + \tau_\ell^2)}. \tag{20}$$

In the special case that $g_k = \psi_k$, i.e., the k-th nonlinearity equals the score function of the $k-$th signal for each $k = 1, \ldots, d$, it is easy to show that $\beta_k = \rho_k = \kappa_k$, $\mu_k = 1$, and $\tau_k = \gamma_k = \kappa_k - 1$. Then (20) simplifies to

$$V_{k\ell}^{OPT} = \frac{\kappa_k}{\kappa_k \kappa_\ell - 1} = N \, \mathrm{CRB}(\mathbf{G}_{k\ell}).$$

This means that in this special case the estimator asymptotically attains the CRB for $N \to \infty$.

The remaining problem is that the densities $p_k(\cdot)$, hence the score functions, are unknown and must be estimated in order to attain the CRB in practice. This is worked out in the next section.

It is worth to mention that the asymptotic performance of the quasi-MLE estimate [18] can be shown to be equal to (18) with $c_k = 1$ and $c_\ell = \mu_k / \rho_\ell$. This choice is not optimum but sub-optimum. It means that if the true density of the sources is not known, but is replaced by a model density, then the generalized symmetric FastICA has the performance which is better or equal to that of quasi-MLE.

B. Proposed algorithm

The proposed algorithm, called for brevity "Efficient FastICA" combines the idea of the generalized symmetric FastICA with an adaptive choice of the function g, in order to attain the CRB. The algorithm consists of three steps:

1) Running the original symmetric FastICA until convergence using a standard nonlinearity such as $g(s) = \tanh(s)$.
2) Adaptive choice of different nonlinearities g_k to estimate the score functions of the found sources, based on the outcome of Step 1.
3) A refinement or fine-tuning for each of the found source components by one-unit FastICA, using the nonlinearities found in Step 2, and another fine-tuning using the optimal c_k parameters as in (16),(17).

For easy reference, the outline of the algorithm is summarized in Figure 1. The three steps are now introduced in detail.

1) Step 1: Running the symmetric FastICA until convergence: The purpose of Step 1 is to quickly and reliably get preliminary estimates of the original signals. In this step, therefore, the optional nonlinearity in the original symmetric FastICA $g(s) = \tanh(s)$ is used due to its universality, but other possibilities seems to give promising results as well, e.g. $g(s) = \frac{s}{1+s^2}$. Also the test for saddle points as introduced in [7] is performed to get reliable source estimates.

2) Step 2: Adaptive choice of the nonlinearities: Assume that $\widehat{\mathbf{u}}_k$ is the $k-$th estimated independent signal obtained in Step 1.

A straightforward solution would be to choose $g_k(\cdot)$ as a score function that belongs to the sample distribution function. Such a nonparametric approach, however, has two problems. First, the score function is difficult to estimate using the given sample distribution of the component, because it needs not only the probability density function but its derivative, too. The second problem is that a suitable g_k needs to be continuous and differentiable. For example, the non-parametric model of the distribution used in [10] seems to be effective, but it is computationally highly demanding.

The distributions of the signals are unimodal and symmetric in many real situations, and thus we suggest instead a parametric choice of g_k that works well for the class of generalized Gaussian distributions with parameter α, denoted GG(α). This class covers a wide variety of

typical distributions including standard Gaussian and Laplacian distributions for $\alpha = 2$ and $\alpha = 1$, respectively, a uniform distribution in the limit as $\alpha \to \infty$, and spiky distributions for $\alpha \to 0+$ (see Appendix C for the definition). The score function of this distribution is proportional to

$$g(x) = \text{sign}(x) \cdot |x|^{\alpha-1}. \tag{21}$$

Note that the scalar multiplicative factor is irrelevant for usage in the FastICA.

A problem with the score function of the GG(α) distribution is that it is not continuous for $\alpha \leq 1$ and thus it is not a valid nonlinearity for FastICA. For these α's the statistical efficiency cannot be achieved by the algorithm using this score function.

Let us look separately at the subgaussian ($\alpha > 2$) and supergaussian ($\alpha < 2$) cases.

Subgaussian case

We propose to use the function $g_k(x) = \text{sign}(x) \cdot |x|^{\alpha-1}$ for subgaussian signals, i.e. for $\alpha > 2$. In this case the parameter α can be well estimated by fitting the theoretical fourth-order moment of the GG(α) distribution

$$m_4(\alpha) = \Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{5}{\alpha}\right)\Gamma^{-2}\left(\frac{3}{\alpha}\right), \tag{22}$$

with the sample fourth-order moment of the k-th signal

$$\widehat{m}_{4k} = \widehat{\mathbf{1}}_N^T \widehat{\mathbf{u}}_k^{\cdot 4}/N. \tag{23}$$

There $\{\cdot\}^{\cdot 4}$ denotes the elementwise fourth order power. In (22), Γ is the Gamma function. The sample fourth moment indicates well the subgaussianity ($\widehat{m}_{4k} < 3$) or supergaussianity ($\widehat{m}_{4k} > 3$) of the signal and can be used to drive the choice of the nonlinearity g_k.

In Appendix A an asymptotic inversion of (22) at point $\alpha = +\infty$ is performed, giving an estimation

$$\widehat{\alpha}_k = [\eta_1\sqrt{\widehat{m}_{4k} - 1.8} - \eta_2(\widehat{m}_{4k} - 1.8)]^{-1} \tag{24}$$

with $\eta_1 \approx 0.2906$ and $\eta_2 \approx 0.1851$. This is valid for $\widehat{m}_{4k} > \lim_{\alpha\to\infty} m_4(\alpha) = 9/5 = 1.8$. If $\widehat{m}_{4k} \leq 1.8$ or $\widehat{\alpha}_k > 15$, maximum power 14 in function g_k is used to improve the stability of the algorithm.

Supergaussian case

In the case where the sources have the distribution GG(α) with $\alpha < 2$, the score function (21)

is not so useful as the nonlinearity in FastICA. For example, it is not continuous for $\alpha \leq 1$. Instead, we suggest to study an ad-hoc choice

$$g_{\eta,\alpha}(x) = x \exp\{-(\eta|x|)^{\alpha}\}. \tag{25}$$

For this function, in order to get analytical results, the asymptotic variance for the Generalized symmetric FastICA has to be recomputed. If α in the definition of $g_{\eta,\alpha}$ coincides with the parameter of the GG distribution, then the expressions in (6)-(8) can be computed analytically.

In order to minimize (18) assume, for simplicity, that all the sources have the same distribution GG(α). Then, (18) is minimized whenever $V_{k\ell}^{1U}$ is minimize since it can be written in a form

$$V_{k\ell}^{GS} = \frac{(c_k^2 + c_\ell^2)V_{k\ell}^{1U} + c_\ell^2}{(c_k + c_\ell)^2}, \tag{26}$$

thus it is just a monotone function of $V_{k\ell}^{1U}$. The same holds for (10) and (20). It follows that the nonlinearity $g(\cdot)$ which minimizes $V_{k\ell}^{1U}$ simultaneously minimizes $V_{k\ell}^{GS}$.

The analytical expression for $V_{k\ell}^{1U}$ (with details given in Appendix B) is

$$V_{k\ell}^{1U}(\eta, \alpha) = \frac{(2\lambda - 1)^{-\frac{3}{\alpha}} - \lambda^{-\frac{6}{\alpha}}}{(\lambda^{-\frac{\alpha+1}{\alpha}} - \lambda^{-\frac{3}{\alpha}})^2} \tag{27}$$

where $\lambda = \lambda(\alpha, \eta) = 1 + (\eta/\beta_\alpha)^\alpha$, with β_α is defined in Appendix C. The variance is shown as a function of λ and α in Figure 2. Note that the variance goes to zero for any $\alpha \in (0, 1/2)$ and $\lambda \to \infty$ (or, equivalently, $\eta \to \infty$). More specifically, for a fixed $\alpha \in (0, 1/2)$ and $\eta \to \infty$ it holds

$$V_{ij}^{1U}(\eta, \alpha) \approx 2^{-3/\alpha}(\beta_\alpha/\eta)^{1-2\alpha} \tag{28}$$

This is in accord with the CRB for separating sources distributed as GG(α) for $\alpha \in (0, 1/2]$. For this case, the CRB does not exist, i.e. the asymptotic estimation variance is not lower bounded.

The above result suggests, that if the estimated α is smaller than 1/2, the parameter η should be as large as possible, or alternatively, that a series of nonlinearity functions of the form (25) with parameters $\widehat{\alpha}$ and $\eta \to \infty$ should be applied. In practice, however, this approach often fails because of numerical problems.

For a fixed α from the interval $(1/2, 2)$, the variance in (27) can be minimized with respect to λ or, equivalently, with respect to η. The optimum η is plotted in Figure 3 as a function of α. We can see that the optimum η grows very fast, when α approaches 1/2.

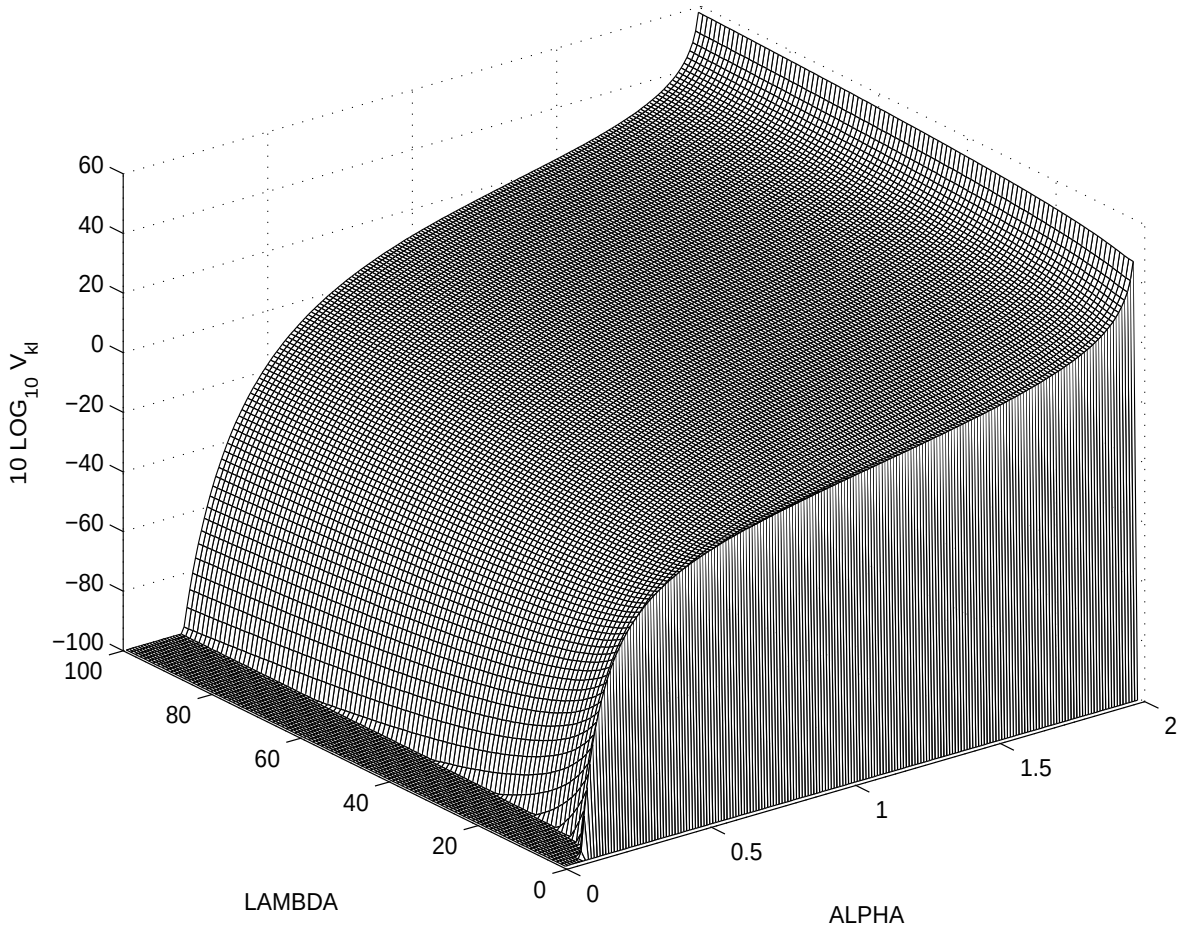


Fig. 2. Asymptotic variance V_{kl}^{1U} in (27) as a function of λ and α, in logarithmic scale.

In the special case of $\alpha = 1$, the optimum η equals

$$\eta^{OPT}(1) \stackrel{\text{def}}{=} \eta_3 \approx 3.348. \tag{29}$$

Numerical experiments show that the nonlinearity (25) with parameters $\alpha = 1$ and $\eta = \eta_3$, i.e.

$$g(x) = x \exp(-\eta_3 |x|) \tag{30}$$

performs quite well for all supergaussian GG(α) distributions with $\alpha < 2$. For easy reference, let us call this nonlinearity "exp3". It is definitely not optimum for $\alpha \leq 1/2$, but appears to be remarkably better than the classical nonlinearity "gauss" with $g(x) = x \exp(-x^2/2)$, which was thought to be best for separation of long-tailed signals. The resultant asymptotic variance is very close to the optimum achievable variance $V_{kl}(\eta^{OPT}(\alpha), \alpha)$ if $\alpha \geq 0.8$, as it is apparent from

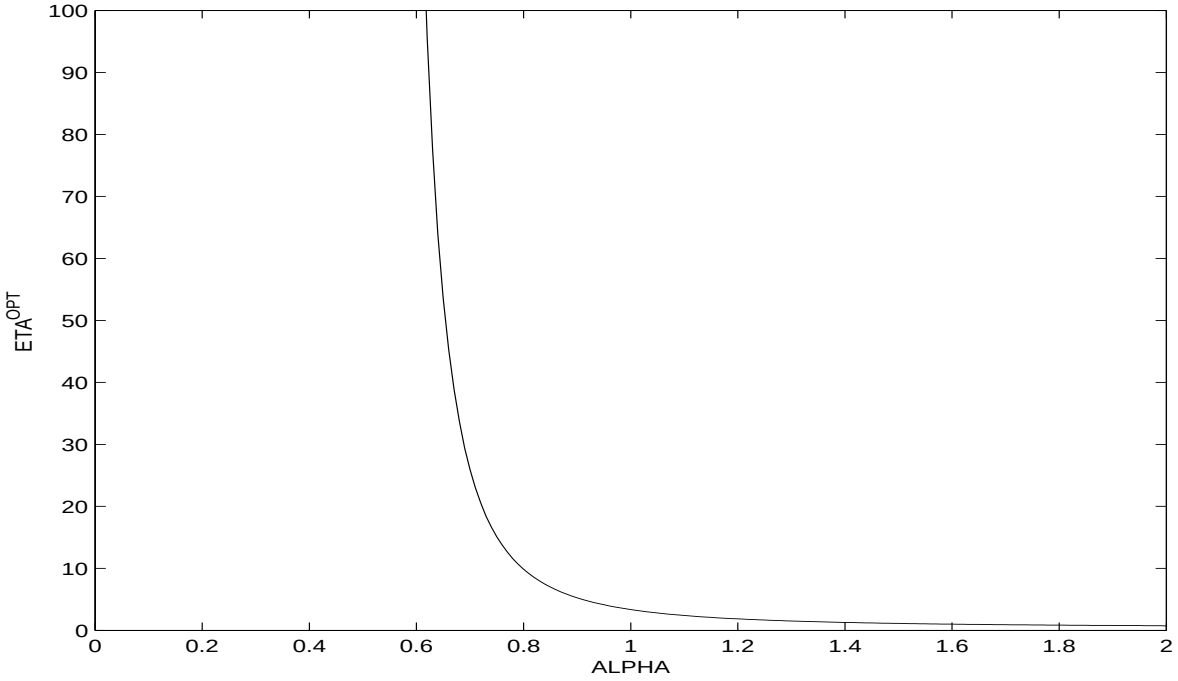


Fig. 3. Parameter η^{OPT} that minimizes V_{kl}^{1U} in (27) as a function of α

Figure 4. Here, the inverse of the variance is shown in logarithmic scale, and compared with similar result with nonlinearity "gauss", with symmetric FastICA, and with the corresponding CRB. In a neighborhood of $\alpha = 2$, the Cramér Rao bound is approached by symmetric FastICA.

The universal choice of the nonlinearity (30) has the advantage that the parameter α need not be estimated in the supergaussian case.

Summary of the Nonlinearity for Efficient FastICA

In summary, the nonlinearity of our choice is

$$g_k(x) = \begin{cases} x \exp(-\eta_3 |x|) & \text{for} \quad \widehat{m}_{4k} > 3 \\ \text{sign}(x) \cdot |x|^{\min\{\widehat{\alpha}_k - 1, 14\}} & \text{for} \quad 1.8 < \widehat{m}_{4k} \leq 3 \\ \text{sign}(x) \cdot |x|^{14} & \text{for} \quad \widehat{m}_{4k} \leq 1.8 \end{cases} \tag{31}$$

where $\widehat{m}_{4k}$ is the estimated fourth-order moment of the k-th source signal, given in eq. (23), the parameter η_3 is given in (29), and $\widehat{\alpha_k}$ is given in (24).

Note that in the vicinity of $\widehat{m}_{4k} = 3$, corresponding to a Gaussian signal, there is a sudden change in nonlinearity, but it does not have any adverse consequences. In separating sources

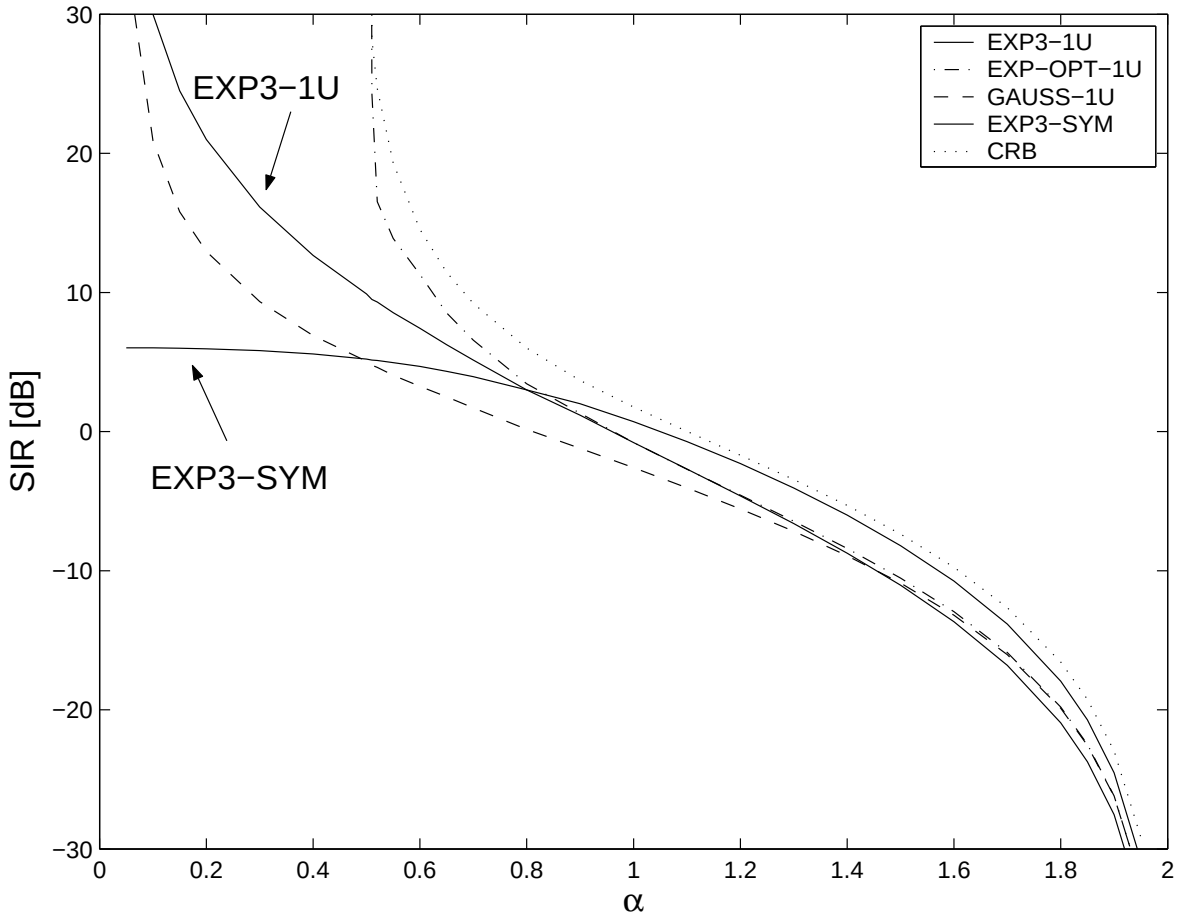
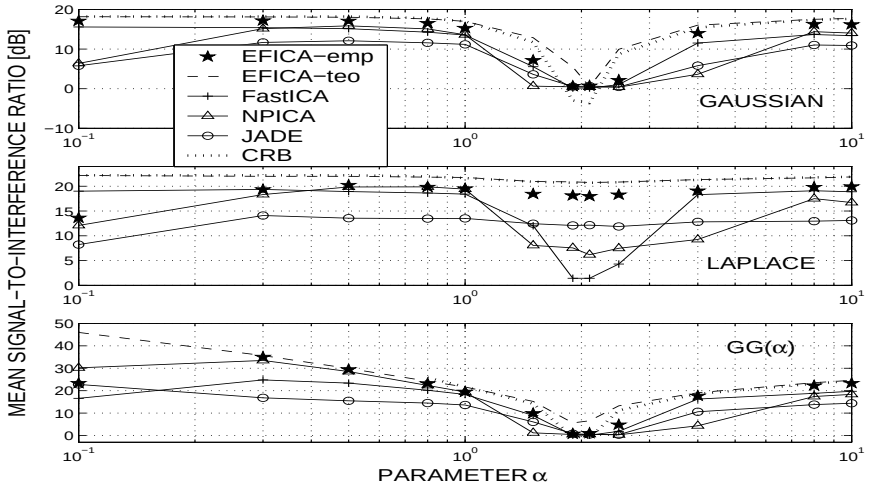


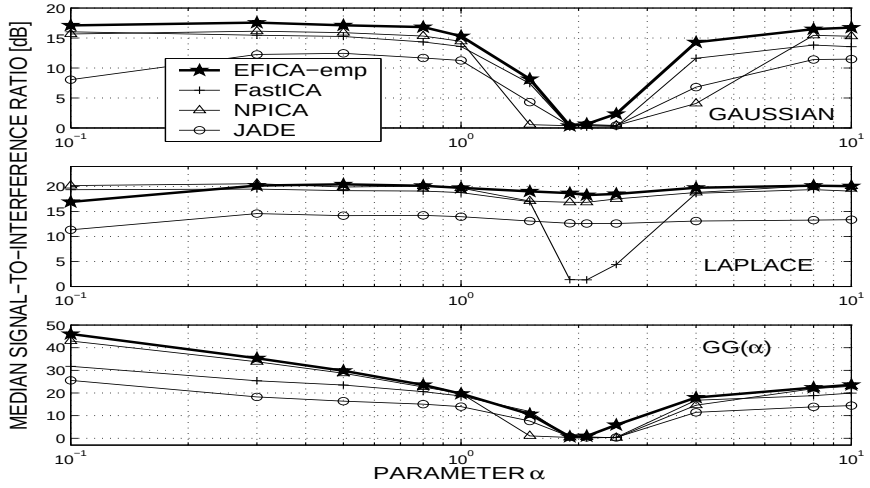
Fig. 4. Theoretically achievable signal-to-interference ratio $-10\log_{10} V_{kl}$ for one unit FastICA with nonlinearity "exp3", with nonlinearity (25) with the optimum η, and with nonlinearity "gauss", and the SIR for symmetric FastICA with nonlinearity "exp3", and the CRB, as a function of α.

with nearly Gaussian distributions, FastICA is known to produce for all common nonlinearities very similar results.

3) Step 3: The refinement: The first step R1 of the refinement or fine-tuning of the algorithm is the analogy of (15), i.e. one-unit iteration, that is run a sufficient number of rounds until convergence is achieved using the optimal nonlinearities $g_k(x)$ of eq. (31). Here, we assume that the preliminary estimates of the original signals $\widehat{\mathbf{u}}_k$ from the first step (symmetric FastICA) of the proposed method lie in the right domain of attraction when using the adaptive nonlinearities. It is the property of the one-unit FastICA algorithm that, suitably initialized, each of the sources is found separately as its own local minimum of the contrast function.



(a)



(b)

Fig. 5. Mean and median SIR of the signal components in Example 1.

The second step R2 of the refinement is based on the optimum choice of numbers $c_1, \ldots, c_d$ in (16) via (19). The fact that the asymptotic variance (18) cannot be minimized, in general, for all $k, \ell = 1, \ldots, d$ jointly, suggests doing the refinement for each independent component separately [15].

Step R1: Let $\widehat{\mathbf{W}}^{SYM+} = [\widehat{\mathbf{w}}_1^{SYM+}, \ldots, \widehat{\mathbf{w}}_d^{SYM+}]^T$ and $\widehat{\mathbf{W}}^{SYM} = [\widehat{\mathbf{w}}_1^{SYM}, \ldots, \widehat{\mathbf{w}}_d^{SYM}]^T$, respectively, be the result of (4) and (5) from the last iteration of symmetric FastICA. Assume that Step 2 has been performed, i.e., for each $k = 1, \ldots, d$, $\widehat{m}_{4k}$ has been computed according to (23) with $\widehat{\mathbf{u}}_k = (\widehat{\mathbf{w}}_k^{SYM})^T \mathbf{Z}$, and the optimal nonlinearity g_k has been chosen via (31).

For each $k = 1, \ldots, d$, initialize

$$\widehat{\mathbf{w}}_k = \widehat{\mathbf{w}}_k^{SYM} \tag{32}$$

and under a condition that $|\widehat{\mathbf{w}}_k^T \widehat{\mathbf{w}}_k^{SYM}| \geq 0.95$ iterate the one-unit FastICA

$$\widehat{\mathbf{w}}_k^+ = \mathbf{Z} g_k(\widehat{\mathbf{w}}_k \mathbf{Z}) - \widehat{\mathbf{w}}_k \, g_k'(\widehat{\mathbf{w}}_k \mathbf{Z}) \mathbf{1}_N \tag{33}$$

$$\widehat{\mathbf{w}}_k = \widehat{\mathbf{w}}_k^+ / \|\widehat{\mathbf{w}}_k^+\| \tag{34}$$

until convergence is achieved. If $|\widehat{\mathbf{w}}_k^T \widehat{\mathbf{w}}_k^{SYM}| < 0.95$, i.e. the new estimate $\widehat{\mathbf{w}}_k$ differs too much from $\widehat{\mathbf{w}}_k^{SYM}$, keep the old result $\widehat{\mathbf{w}}_k^+ = \widehat{\mathbf{w}}_k^{SYM+}$ and put back $g_k(x) = \tanh(x)$. This case occurs when the signal is nearly Gaussian or when the choice of g_k seems to be inappropriate. The condition $\widehat{\mathbf{w}}_k^T \widehat{\mathbf{w}}_k^{SYM} \geq 0.95$ ensures stability of the algorithm and prevents it from doing redundant iterations.

Now, let $\widehat{\mathbf{W}}^+ = [\widehat{\mathbf{w}}_1^+, \ldots, \widehat{\mathbf{w}}_d^+]^T$ be the result after convergence of the one-unit algorithms. The optimum refinement step is completed as follows.

Step R2: For each $k = 1, \ldots, d$, compute

$$\begin{aligned} \widehat{\mu}_k &= \widehat{\mathbf{u}}_k^T g_k(\widehat{\mathbf{u}}_k)/N, & \widehat{\tau}_k &= |\widehat{\mu}_k - \widehat{\rho}_k| \\ \widehat{\rho}_k &= \widehat{\mathbf{1}}_N^T g_k'(\widehat{\mathbf{u}}_k)/N & \widehat{\gamma}_k &= \widehat{\beta}_k - \widehat{\mu}_k^2 \\ \widehat{\beta}_k &= \widehat{\mathbf{1}}_N^T g_k^2(\widehat{\mathbf{u}}_k)/N \end{aligned} \tag{35}$$

For each $k, \ell = 1, \ldots, d$, compute

$$c_{k\ell} = \begin{cases} \frac{\widehat{\tau}_\ell \widehat{\gamma}_k}{\widehat{\tau}_k(\widehat{\gamma}_\ell + \widehat{\tau}_\ell^2)} & \text{for} \quad \ell \neq k \\ 1 & \text{for} \quad \ell = k \end{cases}$$

Next, for each $k = 1, \ldots, d$, compute

$$\widehat{\mathbf{W}}_k^+ = \text{diag}[c_{k1}, \ldots, c_{kd}] \cdot \widehat{\mathbf{W}}^+ \tag{36}$$

$$\widehat{\mathbf{W}}_k^{aux} = (\widehat{\mathbf{W}}_k^+ \widehat{\mathbf{W}}_k^{+T})^{-1/2} \widehat{\mathbf{W}}_k^+ \tag{37}$$

$$\widehat{\mathbf{w}}_k^{REF} = (\widehat{\mathbf{W}}_k^{aux})_{k,\cdot}^T \tag{38}$$

The resulting $k-$th component estimate is $\widehat{\mathbf{u}}_k^{REF} = \mathbf{Z}^T \widehat{\mathbf{w}}_k^{REF}$, and the resulting refined demixing matrix $\widehat{\mathbf{W}}^{REF}$ is

$$\widehat{\mathbf{W}}^{REF} = [\widehat{\mathbf{w}}_1^{REF}, \ldots, \widehat{\mathbf{w}}_d^{REF}]^T \ .$$

This completes the Efficient FastICA algorithm.

IV. Simulations

In this section, we provide an experimental comparison of the proposed algorithm with other well-known methods for ICA. For this purpose, we choose the original version of the symmetric FastICA [5] with nonlinearity "tanh", JADE [4], and the nonparametric algorithm NPICA [10], generally believed to give very good results in a variety of separation tasks.

In order to measure the asymptotic errors of the algorithms, we use the achieved signal-to-interference ratio of the k-th signal. It can be estimated as

$$\overline{\mathrm{SIR}}_k = \frac{N}{\sum_{\ell=1 \ell \neq k}^{d} V_{k\ell}^{OPT}} \tag{39}$$

using (35) and (20).

There occur some random convergence failures in the original FastICA and NPICA, which in our method are fixed by the test of saddle points. To eliminate their effect in the comparisons, both mean and median SIRs from each experiment are shown.

A. Experiments with GG(α) distributions

Three simulation experiments with artificial data were carried out to demonstrate the efficiency of the proposed algorithm and its superiority above other methods when all the original signals have generalized Gaussian distributions.

Example 1. One Gaussian component, 10 Laplacian, and 10 components with Generalized Gaussian distribution with parameter $\alpha \in [0.1, 10]$ were generated in 100 independent trials. We have chosen the length of data $N = 2000$, which is quite small for separation of 21 signals. In each trial, the signals were mixed with a randomly generated matrix, and separated by the proposed method and the algorithms listed above. Averaged results over the 100 trials were computed.

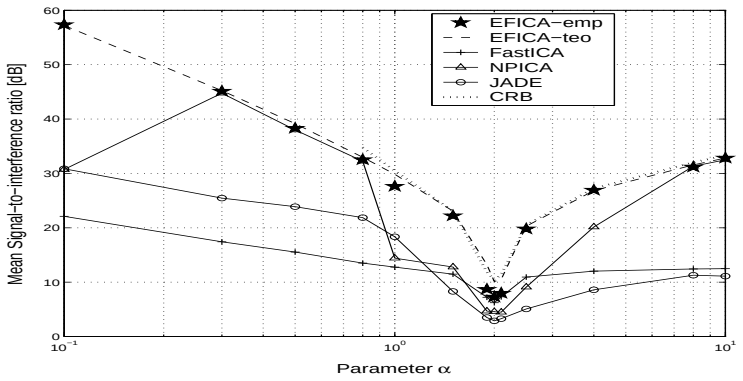
In Figure 5(a), the three diagrams show the average SIRs of the Gaussian signal, of the first Laplacian, and of the first signal with Generalized Gaussian distribution. For the proposed

[0]The matlab code for the algorithm is made available on the Internet at the first author's web page www.utia.cas.cz/user_data/scientific/SI_dept/Tichavsky.html.

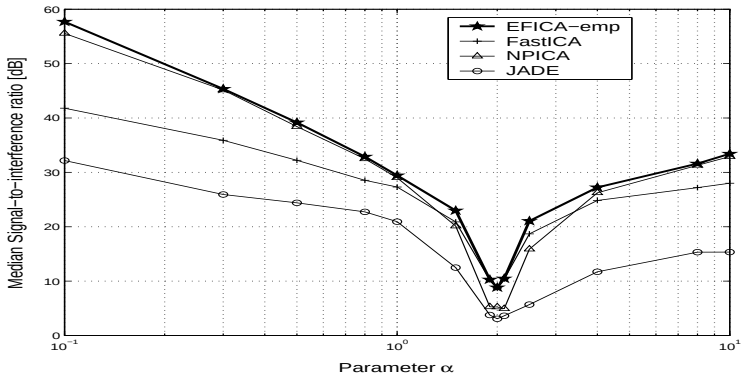
Efficient FastICA method (designated as EFICA), both the theoretical SIRs computed via (39) and the empirical ones are presented. The results can be compared with those in Figure 5(b) where median SIRs from the same experiment are presented. The theoretical SIRs and the CRB are not included there, since they are derived only for mean SIR.

As can be seen from the Figure, the empirical and theoretical SIR (corresponding to the CRB bound) of the Efficient FastICA are in good agreement. The empirical SIR is considerably higher than for the comparison methods.

Example 2. 13 signals of Generalized Gaussian distribution, each with a different value of the parameter α taken from interval $[0.1, 10]$, were mixed with a random mixing matrix and demixed. The experiment was repeated 100 times with fixed length of data $N = 5000$. The results are plotted in Figure 6. The same conclusions hold as in Example 1.



(a) Mean SIR



(b) Median SIR

Fig. 6. Quality of separation of 13 GG(α) components with α, respectively, equal to 0.1, 0.3, 0.5, 0.8, 1, 1.5, 1.9, 2, 2.1, 2.5, 4, 8, and 10.

Example 3. To demonstrate the performance of the method for different lengths of data three signals with Gaussian, Laplacian, and uniform distribution were mixed with a random mixing matrix and demixed. The average and median SIRs from 100 independent trials for each length of data are plotted in Figure 7. Again, EFICA outperforms the other methods.

B. Separation of noisy BPSK signals

In this example, we consider 10 BPSK signals distorted by Gaussian noise, i.e., i.i.d. data distributed as $\sqrt{1-\varepsilon^2}b+\varepsilon n$, where b is a Bernoulli random variable equal to 1 or -1 with equal probabilities, and n is a standard Gaussian variable. Now the probability density of each signal is

$$f_\varepsilon(s) = \frac{1}{2\varepsilon\sqrt{2\pi}}\left(e^{-\frac{\left(s-\sqrt{1-\varepsilon^2}\right)^2}{2\varepsilon^2}} + e^{-\frac{\left(s+\sqrt{1-\varepsilon^2}\right)^2}{2\varepsilon^2}}\right). \tag{40}$$

Data sequences of length $N = 1000$ were generated in 100 independent trials, mixed with a random mixing matrix, and separated. In order to compute the CRB from (14), κ was numerically estimated via a Monte Carlo method. SIR of the first estimated signal as a function of ε is shown in Figure 9.

The results of the experiment need further comments. First, for small ε's the theoretical SIR underestimates the empirical SIR. The reason is that the former SIR is computed from not ideally separated signal components. We have used the estimated signal components to mimic the real situation, when the original signals are not available to predict the estimation accuracy. Performance of the symmetric FastICA and of JADE is limited by the orthogonality constraint, which requires that the separated signals must have mutual correlation exactly zero [14]. The NPICA would perform well in separating two or three signals, as it is tailored to separating signals with multimodal distributions. However, in our example with 10 components it fails, even in median, probably because of some implementation or numerical problems.

C. Speech signals separation

To show the performance of the algorithm on real data, although with artificial mixtures, 10 speech signals of length $N = 5000$ were randomly selected from a database of isolated words[1]

[1]http://noel.feld.cvut.cz/vyu/dzr/cislovky/OBRACENE_BYTY/

containing about 200 samples. After centering and normalization, the data were mixed with a random matrix, and consequently separated. Mean and median SIR computed from estimated gain matrices $\mathbf{G} = \widehat{\mathbf{W}}\mathbf{A}$ obtained in 1000 independent trials are summarized in Table I. The proposed EFICA method gives better results than FastICA and JADE. It is outperformed by the algorithm NPICA, though, but note that this is at the expense of much higher computational cost, as shown in the following section.

TABLE I

MEAN AND MEDIAN SIR OF SEPARATED SPEECH SIGNALS

Algorithm	MEAN [dB]	MEDIAN [dB]
EFICA	21.43	28.03
FastICA	19.87	25.16
NPICA	25.32	30.58
JADE	17.57	21.65

D. Complexity of the algorithm

In order to demonstrate the computational complexity, a similar experiment to that in [10] was done. The average CPU time[2] required by the compared methods is shown in Fig. 10(a) for a varying length of data when 6 generalized Gaussian signals with a random parameter $\alpha \in [0.1, 10]$ were separated. The results for a fixed length of data $N = 1000$ and a variable number of signals are in Fig. 10(b).

The complexity of the algorithm is only slightly higher than that of the original symmetric FastICA. The test of saddle points has complexity $\mathcal{O}(d^2N)$; the adaptive choice of nonlinearity (step R1) has complexity $\mathcal{O}(dN)$. Note that the latter two steps have a fixed number of operations, provided that the test of saddle points is negative. Otherwise, only a few additional iterations are needed since the algorithm is initialized almost in the correct solution.

Similarly, only a few additional one-unit iterations in the refinement step R1 are needed.

[2]The experiment was performed in Matlab^{TM} on a Pentium IV 2.4GHz PC with 512 MB of RAM.

V. Conclusion

An improved version of the FastICA algorithm was proposed, based on the concept of statistical efficiency. This means that the asymptotic variance of the gain matrix, defined as the product of the estimated unmixing matrix and the original mixing matrix, attains the Cramér-Rao lower bound (CRB) which is the theoretical minimum for the variance. The algorithm was named Efficient FastICA (EFICA).

Two changes have to be made in the standard symmetrical FastICA: first, the nonlinearities must be approximations of the score functions of the true sources. This is achieved by running FastICA in two passes, using the separation results from the first pass to estimate a parametric model of the source densities. Second, a set of extra parameters are added to the algorithm in the form of multiplying the lengths of the weight vectors by some numbers before orthogonalization. By optimizing these numbers, this allows to adjust the asymptotic variance so that it becomes equal to the CRB.

The asymptotic efficiency of EFICA was rigorously proven under the assumption that the source signals follow a Generalized Gaussian density with parameter $\alpha > 0$. In simulations with source signals drawn from this density, the agreement between the theoretical and experimental results was proven. Also, in comparisons to some other ICA algorithms, the EFICA algorithm is superior in this case as predicted by the theory.

A key problem in EFICA, as in ICA algorithms in general, is how to estimate the score functions of the sources. While the algorithms perform well with a wide range of approximations or contrasts such as kurtosis in the ideal large-sample case, the effect of the nonlinearity becomes pronounced in the case of small training samples. This is a separate problem, however. The EFICA algorithm hopefully points the way how to extend the methods towards minimum variance estimators.

Appendix A

Consider $\alpha \in (2, +\infty)$. The following second order asymptotic expansion of of (22) was derived for $\alpha \to +\infty$ in Maple^{TM},

$$z(\alpha) \stackrel{\text{def}}{=} \sqrt{m_4(\alpha) - \frac{9}{5}} = \frac{A}{\alpha} - \frac{B}{\alpha^2} + o(\frac{1}{\alpha^2}), \tag{41}$$

where $A = \frac{\sqrt{30}\pi}{5}$, $B = \frac{18\sqrt{30}\zeta(3)}{5\pi}$, and $\zeta(\cdot)$ is a Riemann function obeying $\zeta(3) \doteq 1.202$. To invert the above relation, first, substitute $y = \frac{1}{\alpha}$. Then, we can write

$$y = \frac{1}{A}z + \frac{B}{A}y^2 + o(y^2) = \\ = \frac{1}{A}z + \frac{B}{A}\left(\frac{1}{A}z + \frac{B}{A}y^2 + o(y^2)\right)^2 + o(y^2) = \\ = \frac{1}{A}z + \frac{B}{A^3}z^2 + o(z^2) \quad (42)$$

Using the definition (41) of z gives

$$\alpha \approx \left(\frac{1}{A}\sqrt{m_4(\alpha) - \frac{9}{5}} - \frac{B}{A^3}\left(m_4(\alpha) - \frac{9}{5}\right)\right)^{-1} \quad (43)$$

APPENDIX B

The definition of the nonlinearity (25) allows us to compute (6)-(8) analytically. Using a substitution $\eta = (\lambda - 1)^{1/\alpha}\beta_\alpha$, where β_α is defined in Appendix C and $\lambda > 1$, it can be shown that

$$\mu_{\eta,\alpha} = \mathrm{E}_\alpha[x g_{\eta,\alpha}(x)] = \frac{1}{\lambda^{\frac{3}{\alpha}}} \quad (44)$$

$$\rho_{\eta,\alpha} = \mathrm{E}_\alpha[g'_{\eta,\alpha}(x)] = \frac{1}{\lambda^{\frac{\alpha+1}{\alpha}}} \quad (45)$$

$$\beta_{\eta,\alpha} = \mathrm{E}_\alpha[g^2_{\eta,\alpha}(x)] = \frac{1}{(2\lambda - 1)^{\frac{3}{\alpha}}} \quad (46)$$

Here, E_α denotes expectation with respect to the distribution GG(α), (47). Inserting (44)-(46) in (10) gives (27).

APPENDIX C

Generalized Gaussian distribution family

The generalized Gaussian density function with parameter α, zero mean and variance one, is defined as

$$f_\alpha(x) = \frac{\alpha\beta_\alpha}{2\Gamma(1/\alpha)} \exp\left\{-(\beta_\alpha|x|)^\alpha\right\} \quad (47)$$

where $\alpha > 0$ is a positive parameter that controls the distribution's exponential rate of decay, and $\Gamma(\cdot)$ is the Gamma function and $\beta_\alpha = \sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}$.

The $k-$th absolute moment for the distribution is

$$\mathrm{E}_\alpha\{|x|^k\} = \int_\infty^\infty |x|^k \, f_\alpha(x) dx = \frac{1}{\beta_\alpha^k} \frac{\Gamma\left(\frac{k+1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)} \tag{48}$$

The score function of the distribution is

$$\psi_\alpha(x) = -\frac{\frac{\partial f_\alpha(x)}{\partial x}}{f_\alpha(x)} = \frac{|x|^{\alpha-1}\mathrm{sign}(x)}{\mathrm{E}_\alpha[|x|^\alpha]} \tag{49}$$

Then, simple computations give

$$\kappa_\alpha = \mathrm{E}_\alpha[\psi_\alpha^2(x)] = \frac{\mathrm{E}_\alpha[|x|^{2\alpha-2}]}{\{\mathrm{E}_\alpha[|x|^\alpha]\}^2} = \frac{\Gamma\left(2-\frac{1}{\alpha}\right)\Gamma\left(\frac{3}{\alpha}\right)}{\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^2} \tag{50}$$

Note that $\kappa_\alpha = +\infty$ for $\alpha \leq \frac{1}{2}$, $\kappa_\alpha = 1$ for $\alpha = 2$ (the distribution is standard Gaussian), and $\kappa_\alpha \to +\infty$ for $\alpha \to +\infty$.

REFERENCES

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

[2] A. Cichocki and S.-I. Amari, *Adaptive signal and image processing: learning algorithms and applications*, Wiley, New York, 2002.

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

[4] J.-F. Cardoso, "High-order Contrasts for Independent Component Analysis", *Neural Computat.*, vol. 11, no. 1,pp. 157-192, 1999.

[5] A. Hyvärinen and E. Oja, "A Fast Fixed-Point Algorithm for Independent Component Analysis". *Neural Computation*, 9(7):1483-1492, 1997.

[6] A. Hyvärinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis". *IEEE Transactions on Neural Networks*, 10(3):626-634, 1999.

[7] P. Tichavský, Z. Koldovský, and E. Oja, "Asymptotic Performance Analysis of the Fixed-Point Algorithm (Fast-ICA) for Independent Component Analysis", *IEEE Statistical Signal Processing Workshop*, Bordeaux, July 2005.

[8] Z. Koldovský, P. Tichavský and E. Oja, "Cramér-Rao Lower Bound for Linear Independent Component Analysis", *Proc. ICASSP-2005*, Philadelphia, Vol. III, pp. 581 - 584, March 2005.

[9] P. Tichavský, Z. Koldovský, and E. Oja, "Performance Analysis of the FastICA Algorithm and Cramér-Rao Bounds for Linear Independent Component Analysis", accepted for publication in *IEEE Trans. on Signal Processing*, May 2005.

[10] R. Boscolo, H. Pan, and V. P. Roychowdhury, "Independent Component Analysis Based on Nonparametric Density Estimation", *IEEE Trans. on Neural Networks*, vol. 15, no. 1, pp. 55-65, 2004.

[11] E. G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy", *Journal of Machine Learning Research*, vol. 4, no. 7-8, pp. 1271-1295, 2004.

[12] `http://www.cis.hut.fi/projects/ica/fastica/`

[13] Waheed, K.; Salam, F.M., "Blind source recovery using an adaptive generalized Gaussian score function", Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on, 4-7 Aug. 2002 Vol.3, pp. 656-659.

[14] J.-F. Cardoso, "Blind signal separation: statistical principles", *Proceedings of the IEEE*, vol. 90, n. 8, pp. 2009-2026, October 1998.

[15] J.-F. Cardoso, "On the performance of orthogonal source separation algorithms", *Proc. EUSIPCO*, pp. 776-779, Edinburgh, September 1994.

[16] D. Yellin, B. Friedlander, "Multichannel system identification and deconvolution: performance bounds", *IEEE Tr. Signal Processing*, Vol. 47, No. 5, May 1999, pp. 1410 - 1414.

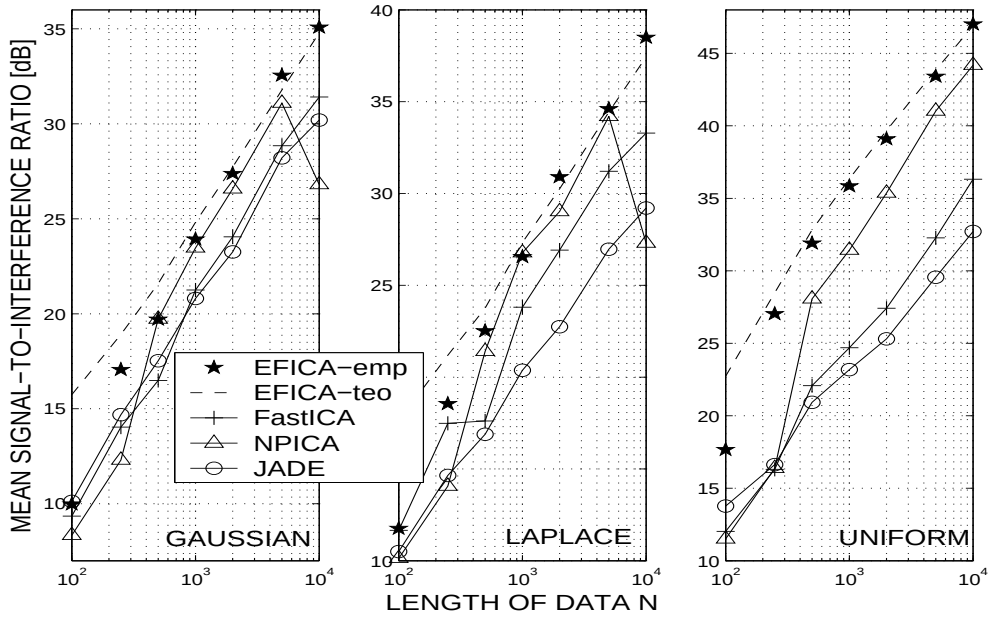
[17] O. Shalvi, E. Weinstein, "Maximum likelihood and lower bounds in system identification with non-Gaussian inputs", *IEEE Tr. Information Theory*, Vol. 40, No. 2, March 1994, pp. 328 - 339.

[18] D.T. Pham, Garat, P., "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach", *IEEE Tr. Signal Processing*, Vol. 45, No. 7, July 1997, pp. 1712 - 1725.

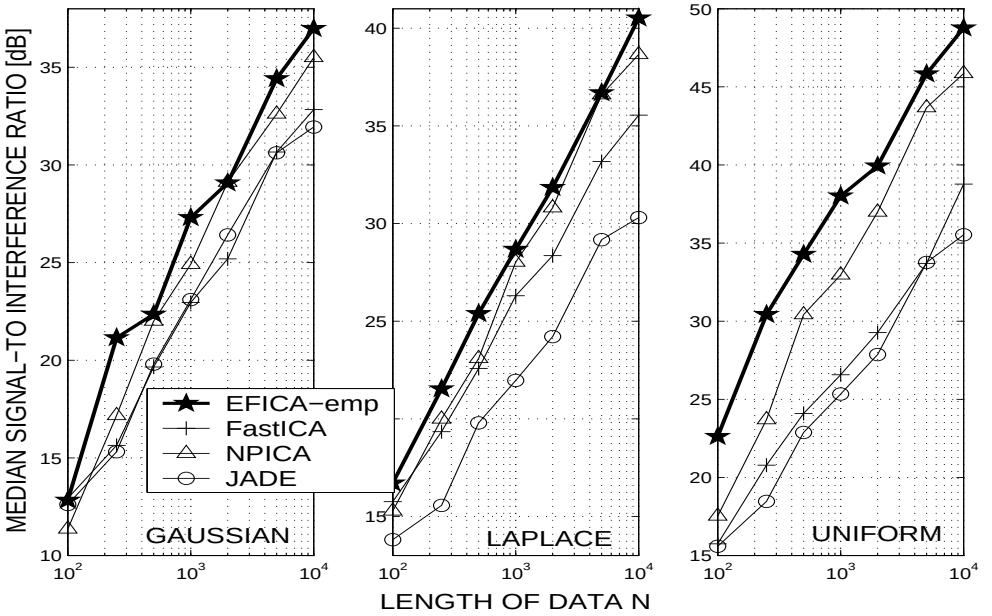
[19] S.-I. Amari, "Superefficiency in blind source separation", *IEEE Trans. on Signal Processing*, vol. 47, no. 4, pp. 936 -944, 1999.

[20] S.-I. Amari and J.- F. Cardoso, "Blind source separation - semiparametric statistical approach", *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2692-2700, 1997.

[21] V. Vigneron and C. Jutten, "Fisher information in source separation problems", *Proc. Fifth Intern. Conf. on ICA and BSS*, Granada, Spain, September 2004, pp. 168-176.



(a)



(b)

Fig. 7. Mean and median SIRs obtained in separating three signal components with varying length.

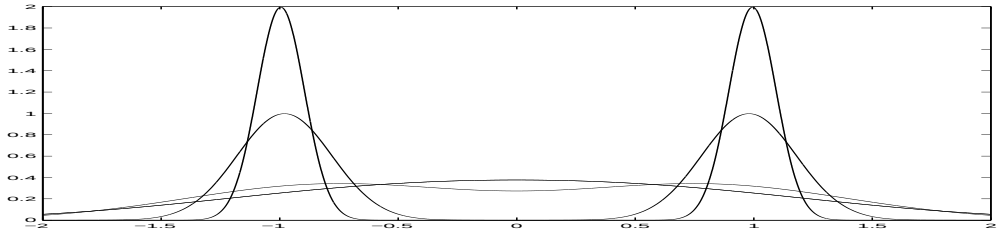


Fig. 8. Probability density (40) for $\varepsilon = 0.1, 0.2, 0.6$ and 0.8, respectively.

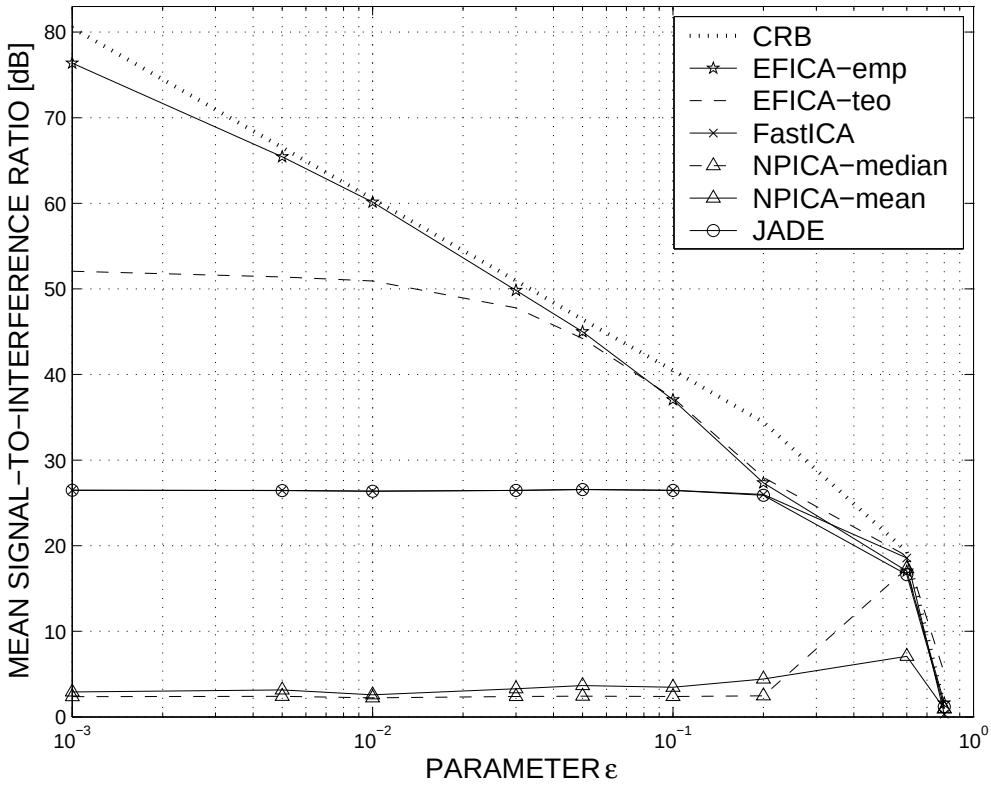


Fig. 9. Separation of noisy BPSK signals.

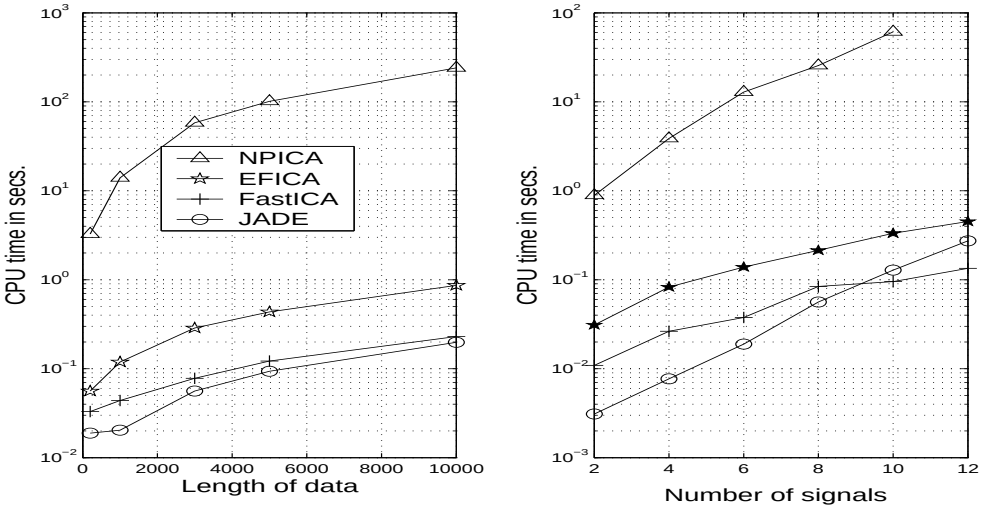


Fig. 10. CPU time needed to separate (a) 6 signals of a various length (b) various number of signals of a fixed length $N = 1000$.

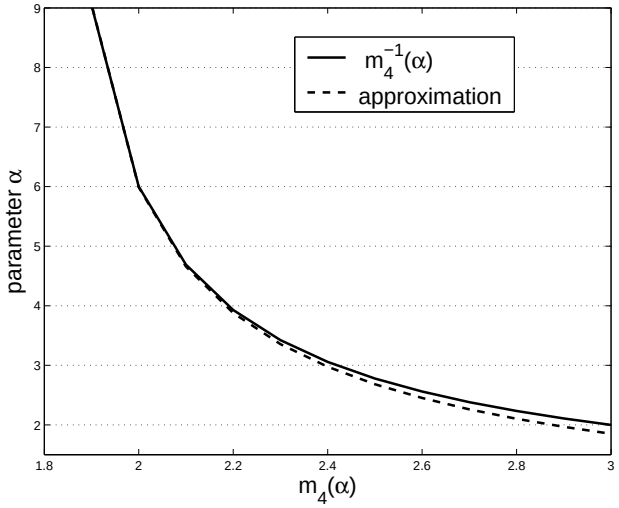


Fig. 11. Comparison of the inversion of (22) with the approximation (43).