Improved Variants of the FastICA Algorithm

Zbyněk Koldovský and Petr Tichavský

1 Introduction

Blind Source Separation (BSS) represents a wide class of models and algorithms that have one goal in common: to retrieve unknown original signals from their mixtures [8]. In the instantaneous linear mixture model, the relation between unobserved original signals and observed measured signals is given by

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

where $\mathbf{X}$ and $\mathbf{S}$ are, respectively, matrices containing samples of the measured and the original signals. Their ijth element corresponds to the jth sample of the ith signal. We will consider the regular case where the numbers of rows in $\mathbf{X}$ and $\mathbf{S}$ are the same and are equal to d. $\mathbf{A}$ is a $d \times d$ regular *mixing* matrix representing the mixing system. The instantaneous model says that the jth original signal contributes to the ith measured signal with an attenuation factor of $\mathbf{A}_{ij}$, which is the ijth element of $\mathbf{A}$.

Independent Component Analysis (ICA) solves the BSS task on the basis of an assumption that the original signals $\mathbf{S}$ are statistically *independent*. Since the original signals are mixed through $\mathbf{A}$, the observed signals $\mathbf{X}$ are, in general, dependent. The ICA task thus can be formulated as the one to estimate the mixing matrix $\mathbf{A}$ or, equivalently, $\mathbf{W} \overset{\triangle}{=} \mathbf{A}^{-1}$, called the *de-mixing matrix*, so that signals $\mathbf{Y} = \mathbf{W}\mathbf{X}$ are as independent as possible[1].

[0] This work was supported by the Czech Science Foundation through Project No. 14-13713S.

[1] The beginnings of ICA can be dated to 1986 when Herault and Jutten published their paper [18] on a learning algorithm that was able to separate independent signals. Later, the concept of ICA was most clearly stated by Comon in [10], which is one of the most cited papers on ICA. Presently, there are several books and proceedings devoted to this important topic of signal processing [8, 9, 11, 23, 30].

The solution of the ICA task is not uniquely determined. Any matrix $\mathbf{W}$ of the form

$$\mathbf{W} = \mathbf{\Lambda} \mathbf{P} \mathbf{A}^{-1}, \tag{2}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with nonzero diagonal entries and $\mathbf{P}$ is a permutation matrix, separates the original signals from $\mathbf{X}$ up to their original order, scales, and signs. Therefore we can later assume, without any loss of generality, that the variance of the source signals is equal to one. Furthermore, the mean of the signals is irrelevant for purposes of the signals' independence and can be assumed equal to zero, or may be removed from the data in case it is nonzero.

Statistical (in)dependence can be measured in various ways depending on the assumptions applied to the model of the original signals. There are three basic models used in ICA/BSS[2]. The first one assumes that the signal is a sequence of identically and independently distributed (i.i.d.) random variables. As the condition of separability of such signals requires that no more than one signal is Gaussian, the approach is called *non-Gaussianity*-based [16]. The second approach takes the *nonstationarity* of signals into account by modeling them as independently distributed Gaussian variables whose variances are changing in time. The third basic model considers weakly stationary Gaussian processes. These signals are separable if their spectra are distinct; therefore, it is said to be based on the *spectral diversity* or non-whiteness.

1.1 Non-Gaussianity-based model

In this model, each original signal is modeled as an i.i.d. sequence. Therefore, the nth sample of the ith original signal, which is the nth element of the ith row of $\mathbf{S}$, also denoted as $s_i(n)$, has the probability density function (pdf) f_{s_i}. Since the signals are assumed to be independent, the joint density of $s_1(n), \ldots, s_d(n)$ is equal to the product of the corresponding marginals,

$$f_{s_1,\ldots,s_d} = \prod_{i=1}^{d} f_{s_i}. \tag{3}$$

The corresponding notation of distributions and pdfs will also be used for the measured signals $\mathbf{X}$ and the separated signals $\mathbf{Y}$.

[2] Some authors associate the non-Gaussianity-based model with ICA only. They classify methods using other models as belonging under the general flag of BSS.

A common criterion for measuring independence of separated signals is the *Kullback-Leibler divergence* between their joint density and the product of marginal densities, which is indeed their *mutual information* defined as

$$I(\mathbf{Y}) = \int_{\mathcal{R}^d} f_{y_1,\ldots,y_d}(\xi_1,\ldots,\xi_d) \ln \frac{f_{y_1,\ldots,y_d}(\xi_1,\ldots,\xi_d)}{\prod_{i=1}^{d} f_{y_i}(\xi_i)} d\xi_1,\ldots,d\xi_d. \quad (4)$$

Assume for now that the components of $\mathbf{Y}$ are not correlated and are normalized to have variances equal to one. Then, it holds that

$$I(\mathbf{Y}) = \sum_{i=1}^{d} H(y_i) + \text{const.}, \quad (5)$$

where $H(y_i)$ is the entropy of the ith separated signal defined as

$$H(y_i) = -\int_{\mathcal{R}} f_{y_i}(\xi) \ln f_{y_i}(\xi) d\xi. \quad (6)$$

Hence, the minimization of (4) is equivalent to the minimization of the entropies of all signals, which is the principle also used by FastICA.

1.2 The FastICA Algorithm

FastICA is one of the most widely used ICA algorithms for the linear mixing model, a fixed-point algorithm first proposed by Hyvärinen and Oja [19, 21]. Following (5), it is based on the optimization of a contrast function measuring the non-Gaussianity of the separated source. We will show later that an optimal measure of the non-Gaussianity requires knowledge of the density function. In FastICA, a nonlinear contrast function is chosen so that it can be appropriate for large-scale of densities.

1.2.1 Preprocessing

The first step of many ICA algorithms, including FastICA, consists of removing the sample mean (the mean is irrelevant for the signals' dependence), scaling the signals to have unit variances (the original scale is also irrelevant as it cannot be retrieved due to the indeterminacy of ICA), and de-correlating them. The de-correlation is a necessary condition for independence. Such a transformation is appropriately expressed as

$$\mathbf{Z} \;=\; \widehat{\mathbf{C}}^{-1/2}\,(\mathbf{X} - \overline{\mathbf{X}}) \quad (7)$$

where

$$\widehat{\mathbf{C}} \;=\; (\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^T/N \tag{8}$$

is the sample covariance matrix; $\overline{\mathbf{X}}$ is the sample mean, $\overline{\mathbf{X}} = \mathbf{X} \cdot \mathbf{1}_N \mathbf{1}_N^T/N$ and $\mathbf{1}_N$ denotes the $N \times 1$ vector of ones. Another popular solution is to apply the Principal Component Analysis to rows of $\mathbf{X} - \overline{\mathbf{X}}$.

Now, the output $\mathbf{Z}$ contains de-correlated and unit variance data in the sense that $\mathbf{Z}\mathbf{Z}^T/N = \mathbf{I}$ (the identity matrix) and their mutual information can be written as in (5). This property remains valid if and only if $\mathbf{Z}$ is multiplied by a unitary matrix $\mathbf{U}$. Therefore, the separating transform can be searched through finding an appropriate $\mathbf{U}$ such that $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{U}\mathbf{Z}$ are as independent as possible. The constraint on $\mathbf{U}$ is called the *orthogonal constraint*.

1.2.2 The FastICA algorithm for one unit

The algorithm estimates one row of the de-mixing matrix $\mathbf{U}$ as a vector $\mathbf{u}^T$ that is a stationary point (minimum or maximum) of

$$\hat{\mathrm{E}}[G(\mathbf{u}^T\mathbf{Z})] \overset{\text{def}}{=} G(\mathbf{u}^T\mathbf{Z})\mathbf{1}_N/N$$

subject to $\|\mathbf{u}\| = 1$, where $G(\cdot)$ is a suitable nonlinear and non-quadratic function, is applied element-wise to vector arguments. The latter expression is indeed the sample mean of $G(\cdot)$ over samples of $\mathbf{u}^T\mathbf{Z}$; and $\hat{\mathrm{E}}[\cdot]$ denotes the sample mean operator.

Finding $\mathbf{u}^T$ proceeds iteratively. Starting with a random initial unit norm vector $\mathbf{u}$, the algorithm iterates

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

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

until convergence is achieved. Here, $g(\cdot)$ and $g'(\cdot)$ denote the first and second derivatives of the function $G(\cdot)$. The application of $g(\cdot)$ and $g'(\cdot)$ to the vector $\mathbf{u}^T\mathbf{Z}$ is also element-wise. 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 $\mathbf{U}$ is being estimated: it largely depends on the initialization. If all independent components were estimated in parallel, the algorithm could be written as

$$\begin{aligned} \mathbf{U}^{+} &\leftarrow g(\mathbf{U}\mathbf{Z})\mathbf{Z}^{T} - \mathtt{diag}[g'(\mathbf{U}\mathbf{Z})\mathbf{1}_{N}]\,\mathbf{U} && (11) \\ \mathbf{u}_{k} &\leftarrow \mathbf{u}_{k}^{+}/\|\mathbf{u}_{k}^{+}\|, \qquad k = 1, \ldots, d, && (12) \end{aligned}$$

where $\mathbf{u}_k^{+}$ stands for the kth row of $\mathbf{U}^{+}$, $\mathbf{u}_k$ stands for the kth row of $\mathbf{U}$, and $\mathtt{diag}[\mathbf{v}]$ stands for a diagonal matrix with diagonal elements taken from the vector $\mathbf{v}$. A result of this iterative process will be denoted as $\mathbf{U}^{1U}$.

1.2.3 The symmetric FastICA algorithm

The symmetric FastICA is designed to estimate all separated signals simultaneously. One step of the parallel estimation proceeds through (11) and each is completed by a symmetric orthonormalization. Specifically, starting with a random unitary matrix $\mathbf{U}$, the method iterates

$$\begin{aligned} \mathbf{U}^{+} &\leftarrow g(\mathbf{U}\mathbf{Z})\mathbf{Z}^{T} - \mathtt{diag}[g'(\mathbf{U}\mathbf{Z})\mathbf{1}_{N}]\,\mathbf{U} && (13) \\ \mathbf{U} &\leftarrow (\mathbf{U}^{+}\mathbf{U}^{+T})^{-1/2}\mathbf{U}^{+} && (14) \end{aligned}$$

until convergence is achieved. Note that $\mathbf{U}$ is orthogonal due to (14). The resulting matrix of the symmetric algorithm will be denoted as $\mathbf{U}^{SYM}$.

The stopping criterion is typically

$$1 - \min(|\mathtt{diag}(\mathbf{U}^{T}\mathbf{U}_{old})|) < \epsilon \qquad (15)$$

for a suitable positive constant ϵ; here $\mathbf{U}_{old}$ denotes the resulting matrix of the previous iteration.

Besides the symmetric algorithm, there also exists Deflation FastICA that estimates all signals. The deflation approach, which is common for many other ICA algorithms [12], estimates the components successively under orthogonality conditions. The accuracy of Deflation FastICA depends on the order of components as they were separated by the algorithm. The order is determined by the initialization.

1.2.4 Summary

The separated signals (independent components) are finally equal to

$$\widehat{\mathbf{S}} = \mathbf{U}\mathbf{Z} = \mathbf{U}\mathbf{D}(\mathbf{X} - \overline{\mathbf{X}}) \qquad (16)$$

where $\mathbf{D}$ stands for the preprocessing transform, e.g., $\mathbf{D} = \mathbf{C}^{-1/2}$ as in (7). The whole separating (de-mixing) matrix is thus equal to

$$\mathbf{W} = \mathbf{U}\mathbf{D}. \tag{17}$$

Note also that the mean $\mathbf{W}\overline{\mathbf{X}}$ could be added back to $\widehat{\mathbf{S}}$.

Since $\mathbf{U}$ is orthogonal, sample correlations of the separated signals $\widehat{\mathbf{S}}$ are exactly equal to zero. This is the consequence of the orthogonality constraint.

1.3 Later Developments of FastICA

Since the first papers on the FastICA algorithm were published, the algorithm has become one of the most successful and most frequently used methods for ICA. It was subject to intensive interest of many researchers, which gave rise to many theoretical and practical analyses of its behavior and modifications. Here we refer to some of them.

Statistical properties of the algorithm, especially its accuracy when finite data are available, were studied in [15, 20, 44] and later in [38, 41, 47]. The results were often compared with the corresponding Cramér-Rao bound derived, e.g., in [5, 11, 24, 44].

The algorithm was also adapted for operation with complex-valued signals [2, 31, 50]; the Cramér-Rao bound for such a case was studied in [32], and identifiability issues were studied in [17].

Speed of the algorithm was improved for FastICA with "pow3" nonlinearity in a novel method called robustICA [49]. Another way of speed enhancement for FastICA with general contrast functions was achieved by replacing the nonlinear contrast functions such as "tanh" or "gauss" by suitable rational functions [45]. With these functions, the statistical properties of FastICA remain nearly the same but the evaluation of rational function is faster on most processors.

Stability issues were studied in [42] where it was shown that in the case that the separated independent components have multimodal distributions (the pdf has two or more peaks), it happens with nonzero probability that deflation FastICA gets stuck in a false solution which does not correspond to the separation of all sources. Only the algorithm with the "pow3" nonlinearity (kurtosis) can guarantee a zero probability of this phenomenon. For the deflation FastICA, the order of the separated components appeared to be crucial for stability of the algorithm. An improved algorithm which optimizes the order of the separated components in FastICA was proposed in [35]. For symmetric FastICA there was a simple test of saddle points proposed to improve the success rate of the algorithm in [44].

An improved FastICA with adaptive choice of nonlinearity is the subject of the EFICA algorithm [25]. The fact that the nonlinearity influences the algorithm's statistical accuracy was already known, and other FastICA variants endowed with an adaptive choice had previously been proposed; see e.g. [7, 13, 33, 35].

FastICA properties were also studied in the presence of additive noise. In [22], an unbiased variant was proposed based on the assumption of known covariances of the noise. Later, it was shown in [27] that One-unit FastICA tends to estimating the minimum mean square solution rather than to identifying the inversion of the mixing matrix.

2 Accuracy of One-unit and Symmetric FastICA

2.1 Performance Evaluation

The accuracy of separation can be evaluated through a comparison of the estimated mixing matrix with the original one or of the separated signals with the original ones. The original quantities must be known, which happens only in simulated experiments: Some independent signals are mixed by a generated mixing matrix, an ICA algorithm is applied to the mixed signals, and the resulting separating matrices or separated signals are evaluated. The evaluation method must take into account the ICA indeterminacy, especially the random order of separated signals.

Let $\mathbf{G}$ be the so-called *gain matrix* defined as

$$\mathbf{G} = \mathbf{W}\mathbf{A}. \tag{18}$$

Ideally, $\mathbf{G}$ is equal to $\mathbf{\Lambda P}$ as follows from (2). Here $\mathbf{\Lambda}$ is the diagonal matrix representing the signals' scale indeterminacy, while $\mathbf{P}$ is the permutation matrix determining their order. In practice, $\mathbf{G} \approx \mathbf{\Lambda P}$ due to estimation errors in $\mathbf{W}$.

The Amari's index evaluates the separation accuracy as a whole with the aid of a non-negative value

$$I = \sum_{i=1}^{d} \left(\frac{\sum_{j=1}^{d} |\mathbf{G}_{ij}|}{\max_k |\mathbf{G}_{ik}|} - 1 \right) + \sum_{j=1}^{d} \left(\frac{\sum_{i=1}^{d} |\mathbf{G}_{ij}|}{\max_k |\mathbf{G}_{kj}|} - 1 \right). \tag{19}$$

The criterion reflects the fact that $\mathbf{G}$ should contain one and only one dominant element per row and column.

To evaluate each separated signal individually, it is popular to use standard measures such as Signal-to-Interference ratio (SIR). However, before the computation of SIR, the separated signals must be correctly assigned to the original ones. A straightforward way is to match the separated and original signals based on dominant elements of $\mathbf{G}$ under the condition that the matched pairs of signals are disjoint. The most common approach, called *greedy*, finds the maximal (in absolute value) element of $\mathbf{G}$, assigns the corresponding signals, and repeats the process until all signals are paired. A more sophisticated non-greedy pairing based on the Kuhn-Munkres algorithm was proposed in [43].

Once the permutation matrix $\mathbf{P}$ is found, and the separated signals are reordered, the kth separated signal, denoted as $\widehat{s}_k(n)$, is equal to

$$\widehat{s}_k(n) = \mathbf{G}_{k1}s_1(n) + \cdots + \mathbf{G}_{kk}s_k(n) + \cdots + \mathbf{G}_{kd}s_d(n).$$

The SIR of the kth separated signal equals

$$\mathrm{SIR}_k = \frac{|\mathbf{G}_{kk}|^2\sigma_k^2}{\sum_{i=1,i\neq k}^{d}|\mathbf{G}_{ki}|^2\sigma_i^2} \tag{20}$$

where σ_i^2 is the variance of the ith original signal. Henceforth, the variances will be assumed equal to one, that is $\sigma_i^2 = 1$, $i = 1, \ldots, d$. This assumption can be used without any loss of generality because of the indeterminacy in signals' scales.

The reciprocal value of SIR is named the Interference-to-Signal Ratio (ISR)

$$\mathrm{ISR}_k = \frac{\sum_{i=1,i\neq k}^{d}|\mathbf{G}_{ki}|^2}{|\mathbf{G}_{kk}|^2}. \tag{21}$$

2.2 Cramér-Rao Lower Bound

Cramér-Rao lower bound (CRLB) is a general bound for the variance of an unbiased estimator [40]. Consider a vector of parameters $\boldsymbol{\theta}$ being estimated from a data vector $\mathbf{x}$, where the latter has probability density $f_{\mathbf{x}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ be an unbiased estimator of $\boldsymbol{\theta}$. If the following *Fisher information matrix* (FIM) exists

$$\mathbf{F}_{\boldsymbol{\theta}} = \mathrm{E}_{\boldsymbol{\theta}}\left[\frac{1}{f_{\mathbf{x}|\boldsymbol{\theta}}^2}\frac{\partial f_{\mathbf{x}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\left(\frac{\partial f_{\mathbf{x}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right], \tag{22}$$

then, under mild regularity conditions, it holds that

$$\operatorname{cov}\hat{\boldsymbol{\theta}} \geq \mathrm{CRLB}_{\boldsymbol{\theta}} = \mathbf{F}_{\boldsymbol{\theta}}^{-1},$$

where $\operatorname{cov}\hat{\boldsymbol{\theta}}$ is the covariance matrix of $\hat{\boldsymbol{\theta}}$.

Now we apply the idea of the Cramér-Rao theory to the elements of an ISR matrix whose ijth element is defined as

$$\mathrm{ISR}_{ij} = \mathrm{E}\left[\frac{|\mathbf{G}_{ij}|^2}{|\mathbf{G}_{ii}|^2}\right], \tag{23}$$

to derive an algorithm-independent lower bound on values of its elements.

Let the separated signals be already re-ordered and scaled so that $\mathbf{G} = \mathbf{I} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a "small" matrix of errors. Then the elements of the ISR matrix can be approximated as

$$\mathrm{ISR}_{ij} \approx \mathrm{E}[|\boldsymbol{\epsilon}_{ij}|^2], \tag{24}$$

and the lower bound can be defined as the CRLB for $\boldsymbol{\epsilon}$; see also [14]. Note that we are only interested in the non-diagonal elements of (24), because the asymptotic behavior of the ISR is independent of the diagonal terms (assuming "small" errors).

2.2.1 Non-Gaussian i.i.d. Signals

Details of the computation of the CRLB is given in [24] with a small correction in [46]. The bound says that

$$\mathrm{ISR}_{ij} \geq \frac{1}{N}\frac{\kappa_j}{\kappa_i\kappa_j - 1}, \qquad i \neq j, \tag{25}$$

where

$$\kappa_i = \mathrm{E}\left[(\psi_i(x))^2\right] \tag{26}$$

and

$$\psi_i(x) = -\frac{f_i'(x)}{f_i(x)} \tag{27}$$

is the so-called score function of f_i. The same result was also observed elsewhere in the literature; see, e.g., [11, 5, 37]; for the complex-domain case see [32].

It can be shown that $\kappa_i \geq 1$ where the equality holds if and only if f_i is Gaussian; see Appendix E in [44]. Hence, the denominator of (25) becomes equal to zero only if both κ_i and κ_j are equal to one, which means that both the ith and jth signals have Gaussian distributions. This is in accordance with the primary requirement that only one original signal can have the Gaussian pdf. It can also be seen that the bound is minimized when $\kappa_i \to +\infty$ and $\kappa_j \to +\infty$, which can be interpreted as the signals being non-Gaussian as much as possible.

2.2.2 Piecewise Stationary Non-Gaussian Signals

The above CRLB, indeed, follows from a more general bound that was derived for a piecewise stationary non-Gaussian model of signals in [6, 29]. In that model, signals are assumed to obey the i.i.d. model separately within M blocks. Let the blocks have, for simplicity, the same length. Then, the bound says that

$$\mathrm{ISR}_{ij} \geq \frac{1}{N} \cdot \frac{A_{ij}}{A_{ij}A_{ji}-1} \cdot \frac{\overline{\sigma}_j^2}{\overline{\sigma}_i^2}, \qquad i \neq j, \tag{28}$$

where

$$A_{ij} = \frac{1}{M}\sum_{\ell=1}^{M} \frac{\sigma_i^{2(\ell)}}{\sigma_j^{2(\ell)}} \kappa_j^{(\ell)} \tag{29}$$

$$\overline{\sigma}_i^2 = \frac{1}{M}\sum_{\ell=1}^{M} \sigma_i^{2(\ell)}. \tag{30}$$

$\sigma_i^{2(\ell)}$ denotes the variance of the ith signal within the ℓth block, and $\kappa_i^{(\ell)}$ is defined as

$$\kappa_i^{(\ell)} = \mathrm{E}\left[\left(\psi_i^{(\ell)}(x)\right)^2\right] \tag{31}$$

where $\psi_i^{(\ell)} = -\left(\bar{f}_i^{(\ell)}\right)' / \bar{f}_i^{(\ell)}$. $\bar{f}_i^{(\ell)}$ denotes the PDF of the ith original signal on the ℓth block, i.e., $f_i^{(\ell)}$, but normalized to the unit variance (the variance of $f_i^{(\ell)}$ is involved in $\sigma_i^{2(\ell)}$). Recall the simplifying assumption that the original signals have unit scales, which means that $\overline{\sigma}_i^2 = 1$, $i = 1, \ldots, d$.

The shapes of the expressions on the right-hand sides of (25) and (28) are analogous to those of other (more general) Cramér-Rao bounds derived for ICA/BSS or related disciplines, such as Independent Vector Analysis (IVA); an interested reader is referred to [1, 48].

2.3 Asymptotic Behavior of FastICA

Let $\mathbf{G}^{1U}$ and $\mathbf{G}^{SYM}$, respectively, be the gain matrices obtained by One-unit and Symmetric FastICA using the nonlinear function $g(\cdot)$. Let the function be even, which means that the corresponding $G(\cdot)$ is symmetric, and also let the pdfs of signals be symmetric. It was shown in [44] that, for $i \neq j$, elements of

$N^{1/2}\mathbf{G}_{ij}^{1U}$ and $N^{1/2}\mathbf{G}_{ij}^{SYM}$ have asymptotically Gaussian distribution $\mathcal{N}(0, V_{ij}^{1U})$ and $\mathcal{N}(0, V_{ij}^{SYM})$, where

$$V_{ij}^{1U} = \frac{\beta_i - \mu_i^2}{(\mu_i - \rho_i)^2} \tag{32}$$

$$V_{ij}^{SYM} = \frac{\beta_i - \mu_i^2 + \beta_j - \mu_j^2 + (\mu_j - \rho_j)^2}{(|\mu_i - \rho_i| + |\mu_j - \rho_j|)^2} \tag{33}$$

with $\mu_i = \mathrm{E}[s_i g(s_i)]$, $\rho_i = \mathrm{E}[g'(s_i)]$, $\beta_i = \mathrm{E}[g^2(s_i)]$, and $g'(\cdot)$ being the first derivative of $g(\cdot)$. It is sufficient to assume that the above derivative and expectations exist. The expressions for non-symmetric distributions were derived in [47].

Next, it can be shown that (32) achieves its minimum for $g(\cdot)$ being equal to the score function of the distribution f_i, i.e., for

$$g(x) = \psi_i(x) = -\frac{f_i'(x)}{f_i(x)}.$$

In that case, it is easy to compute that $\mu_i = 1$ and $\rho_i = \beta_i = \kappa_i$.

Assume for now that the distributions of all signals are the same, which means that the above quantities are independent of the index i. That is, $g(x) = \psi(x)$ and $\rho_i = \beta_i = \kappa$, and then, according to (32) and (33),

$$\mathrm{var}[\mathbf{G}_{ij}^{1U}] \approx \frac{1}{N} V_{ij}^{1U} = \frac{1}{N}\frac{1}{\kappa - 1} \tag{34}$$

$$\mathrm{var}[\mathbf{G}_{ij}^{SYM}] \approx \frac{1}{N} V_{ij}^{SYM} = \frac{1}{N}\left(\frac{1}{4} + \frac{1}{2}\frac{1}{\kappa - 1}\right). \tag{35}$$

For the same case, the CRLB from (25) takes the form

$$\mathrm{ISR}_{ij} \geq \frac{1}{N}\frac{\kappa}{\kappa^2 - 1}, \qquad i \neq j. \tag{36}$$

Comparisons of (34) and (35) with (36) for $\kappa \geq 1$ are shown in Fig. 1. One-unit FastICA for the optimum case approaches the CRLB when $\kappa \to \infty$, while Symmetric FastICA is nearly efficient for κ lying in a neighborhood of 1. The latter case, however, means that the distributions of signals are close to the Gaussian distribution, so the signals are hard to separate, and the CRLB itself goes to infinity. For $\kappa \to \infty$, the performance of Symmetric FastICA (35) is limited by a constant, which is due to the orthogonal constraint (the sample covariance matrix of the separated signals is exactly equal to the identity matrix).

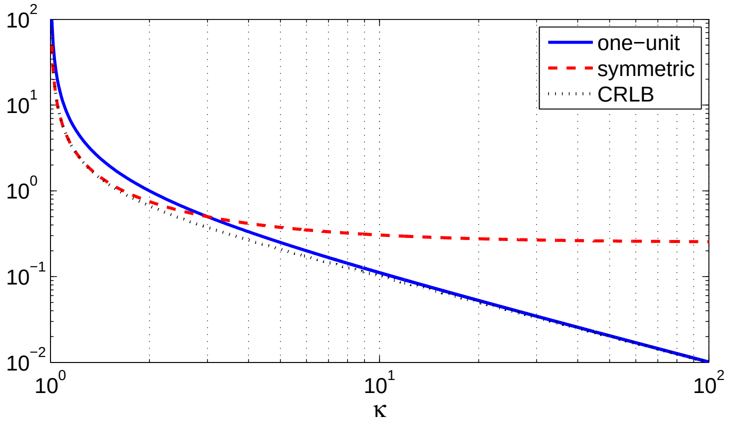

Figure 1: A comparison of One-unit FastICA, Symmetric FastICA and the corresponding CRLB for the case when all signals have the same distributions and $g(x) = \psi(x)$. The expressions are plotted as functions of $\kappa \geq 1$ (here $N = 1$).

2.4 Choice of the Nonlinearity

From the previous analysis it follows that it is not possible to suggest a nonlinearity that would be optimum for all signals, because they need not have the same distribution. The distributions need not be known as we face a blind problem; moreover, score functions of the distributions need not be smooth as required by FastICA. Improved variants of FastICA therefore endow the algorithm by an adaptive choice of the nonlinearity [7, 13, 33, 35].

It is also not possible to choose a nonlinearity that would enable FastICA to separate all non-Gaussian distributions. An example was shown in [45]: Consider signals having the same pdf as $s = \beta b + \sqrt{1-\beta^2}q$ where b and q stand for binary (BPSK) and Laplacean random variables, respectively, and $\beta \in [0, 1]$. For many nonlinearities (e.g. "tanh") it holds that $\mu_i - \rho_i > 0$ for $\beta = 0$ while $\mu_i - \rho_i < 0$ for $\beta = 1$ (if not, other distributions of b and q can be chosen). It then follows that there exists $\beta \in (0, 1)$ such that $\tau_i = \mu_i - \rho_i = 0$. From (32) and (33) it follows that FastICA cannot separate such a distribution (although being non-Gaussian) using the given nonlinearity.

The original variants of FastICA use general-purpose nonlinearities such as "tanh", because it is useful for many signals' distributions that are met in practice. In [45], it was suggested to replace "tanh" by rational functions that are similarly appropriate for separating long-tailed distributions. One such nonlinearity, henceforth referred to as “rati”, is

$$g(x) = \frac{x}{1 + x^2/4}. \tag{37}$$

The advantage of using the rational function is that it requires a significantly lower computational burden than “tanh” due to its evaluation on most CPUs. As a result, FastICA using the rational function is typically twice as fast as the algorithm with “tanh”.

The suitability of any rational function to separate a given distribution with FastICA can be easily inspected and compared with “tanh” using the analytical expressions on the right-hand side of (32) or (33).

3 Global Convergence

The global convergence of FastICA was theoretically proven in special cases only. For example, if the nonlinearity is “pow3”, the global convergence of Symmetric FastICA was proven in [36] but only for the theoretical case in which an infinite amount of samples is available. In practice, the behavior of FastICA is also known to be quite good when “tanh” or other nonlinearities are used.

Nevertheless, if it is run, for instance, 10 000 times from random initial demixing matrices, the algorithm gets stuck in an unwanted solution in 1–100 cases. These cases are recognized by an exceptionally low value of SIR achieved. The rate of false solutions depends on the dimension of the model, on the stopping rule, and on the length of the data. But it never vanishes completely. For example, when separating d signals all having uniform distribution, the failure rates of Symmetric FastICA using the stopping rule (15), respectively, with $\epsilon = 10^{-4}$ and $\epsilon = 10^{-5}$ are shown in Table 1.

3.1 Test of Saddle Points

A detailed investigation of the false solutions showed that they lie approximately halfway (in the angular sense) between a pair of original signals, thus, in saddle points. Although these points are not stable, the algorithm can stop when getting to their close neighborhood as the following iteration step is too small.

Specifically, the false solutions typically contain two components $u_1(n)$ and $u_2(n)$ that are close to $(s_k(n) + s_\ell(n))/\sqrt{2}$ and $(s_k(n) - s_\ell(n))/\sqrt{2}$, for certain

	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
$d = 2\,\&$ 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
$d = 3\,\&$ 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
$d = 4\,\&$ 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
$d = 5\,\&$ s.p.check	**17**	**0**	**0**	**0**

Table 1: Number of failures of Symmetric FastICA with the "tanh" nonlinearity among 10 000 trials; d is the dimension of the signal mixture; ε is the stopping parameter in (15); the acronym "s.p.check" denotes the algorithm endowed by the test of saddle points.

$k, \ell \in \{1, \ldots, d\}$. Thus, they should be transformed into

$$u_1'(n) = (u_1(n) + u_2(n))/\sqrt{2} \quad \text{and} \quad u_2'(n) = (u_1(n) - u_2(n))/\sqrt{2}. \tag{38}$$

It was suggested in [44] to complete the algorithm by checking all $\binom{d}{2}$ pairs of the estimated independent components for a possible improvement via the saddle points. If the test for a saddle point is positive, it is suggested to perform several additional iterations of the original algorithm, starting from the improved estimate (38).

The selection between given candidates (u_k, u_ℓ) and (u_k', u_ℓ') can be done by maximizing the criterion (a measure of total non-Gaussianity of the components),

$$c(u_k, u_\ell) = (\hat{\mathrm{E}}[G(u_k(n))] - G_0)^2 + (\hat{\mathrm{E}}[G(u_\ell(n))] - G_0)^2$$

where $G_0 = \mathrm{E}[G(\xi)]$ and ξ is a standard Gaussian variable; $\hat{\mathrm{E}}[\cdot]$ stands for the sample mean operator. For example, in the case of the nonlinearity "tanh", $G(x) = \log\cosh(x)$ and $G_0 \approx 0.3746$.

The number of failures after this test of saddle points is compared in Table 1. This Table shows zero rate after the test except for the most difficult case when the data length is $N = 200$. Nevertheless, even in this case the rate of failures has significantly dropped compared to the original FastICA.

4 Approaching Cramér-Rao Bound

The analysis of the FastICA variants and the comparison with the corresponding CRLB showed that there is room for improvements in terms of accuracy. Highly non-Gaussian signals can be accurately separated using an appropriate nonlinearity that is close to the score function of the distribution. Since distributions of signals can be different, the nonlinearity should be chosen different for each signal. However, the accuracy of Symmetric FastICA is limited by the orthogonal constraint, which mainly limits the separation of highly non-Gaussian signals. By contrast, One-unit FastICA is less effective when separating signals having distributions that are close to the Gaussian. Only the symmetric version can guarantee global convergence, that is, the separation of all signals.

These conclusions gave rise to a new, more sophisticated, algorithm named EFICA [25]. EFICA is initialized by the outcome of Symmetric FastICA endowed by the test of saddle points described in the previous section. The partly separated signals are used to select optimal nonlinearities g_i, $i = 1, \ldots, d$, for each separated signal, and used in fine-tuning of rows in $\mathbf{W}$. Finally, the whole $\mathbf{W}$ is refined using weighted symmetric orthogonalizations in a way that the orthogonal constraint is avoided. This is done with optimal weights derived from an analysis of a weighted symmetric algorithm.

4.1 Weighted Symmetric FastICA

Consider a variant of the symmetric algorithm where different nonlinear functions $g_k(\cdot)$, $k = 1, \ldots, d$ are used in (13) to estimate each row of $\mathbf{U}^+$. Then, before the symmetric orthogonalization step (14), the rows of $\mathbf{U}^+$ are re-weighted by positive weights. One iteration of such algorithm is thus

$$\begin{aligned} \mathbf{U}^+ &\leftarrow g(\mathbf{UZ})\mathbf{Z}^T - \text{diag}[g'(\mathbf{UZ})\mathbf{1}_N]\,\mathbf{U} && (39) \\ \mathbf{U}^+ &\leftarrow \text{diag}[c_1, \ldots, c_d] \cdot \mathbf{U}^+ && (40) \\ \mathbf{U} &\leftarrow (\mathbf{U}^+\mathbf{U}^{+T})^{-1/2}\mathbf{U}^+ && (41) \end{aligned}$$

where $g(\cdots)$ is an element-wise function applying $g_k(\cdot)$, $k = 1, \ldots, d$, to the corresponding rows of the argument.

The key step in deriving EFICA is to analyze this algorithm, which was done in [25] in the same way as in [44]. The result is that the non-diagonal normalized gain matrix elements for this method, $N^{1/2}\mathbf{G}_{ij}^{WS}$, have asymptotically Gaussian distribution $\mathcal{N}(0, V_{ij}^{WS})$, where

$$V_{ij}^{WS} \;=\; \frac{c_i^2\gamma_i + c_j^2(\gamma_j + \tau_j^2)}{(c_i\tau_i + c_j\tau_j)^2}, \qquad i \neq j. \tag{42}$$

where $\gamma_i = \beta_i - \mu_i^2$ and $\tau_i = |\mu_i - \rho_i|$.

4.2 EFICA

EFICA utilizes the weighted symmetric orthogonalization within its last refinement stage, that is, after the initialization, choice of nonlinearities and fine-tuning. One such orthogonalization is performed for each separated signal. The weights in (40) are chosen such that V_{ij}^{WS} is minimized, specifically, for the ith separated signal, c_i is put equal to one, and

$$c_j^{OPT} = \arg\min_{c_j, c_i = 1} V_{ij}^{WS} = \frac{\tau_j\gamma_i}{\tau_i(\gamma_j + \tau_j^2)}, \qquad j \neq i. \tag{43}$$

Since c_j^{OPT} also depends on i, the weights must be selected different for each signal. Only the ith row of $\mathbf{U}$ after (41) is then used as the ith row for the final de-mixing transform. Consequently, the rows of the final transform are no more orthogonal in general, which means that the algorithm is not constrained to produce exactly orthogonal components.

By putting (43) into (42), we arrive at the asymptotic variance of the non-diagonal normalized gain matrix elements by EFICA, which is

$$V_{ij}^{EF} \approx \frac{1}{N}\frac{\gamma_i(\gamma_j + \tau_j^2)}{\tau_j^2\gamma_i + \tau_i^2(\gamma_j + \tau_j^2)}, \qquad i \neq j. \tag{44}$$

Comparing (44) with (32) and (33), the former can always be shown to be smaller than the latter two, provided that the same nonlinearity is used for all signals.

If the nonlinearities g_i, $i = 1, \ldots, d$, match the score functions of the signals, then

$$\tau_i = \gamma_i = \kappa_i - 1$$

and (44) becomes equal to the CRLB (25). It means that EFICA is asymptotically efficient in that special case[3].

[3] It should be noted that the analysis of FastICA as well as EFICA is local. Therefore, to be more precise, we should say that the asymptotic efficiency of EFICA is ensured when its global convergence is guaranteed.

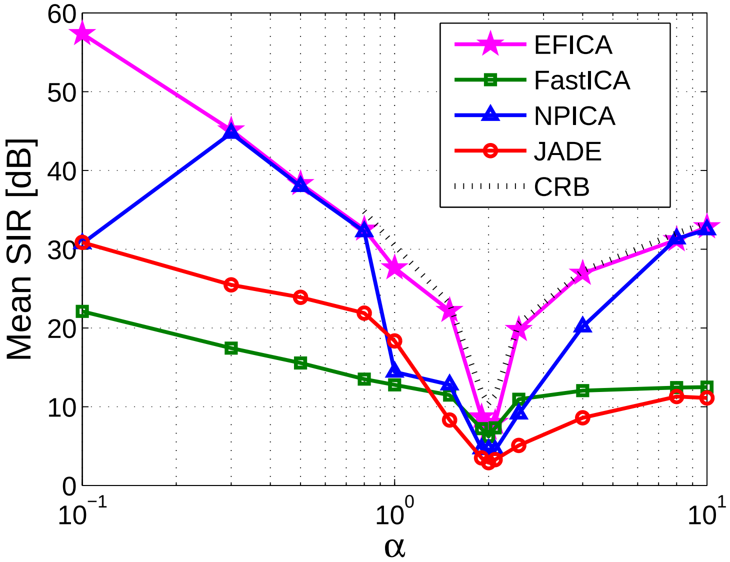

Figure 2: The average SIR of 13 components having the generalized Gaussian distribution with α, respectively, equal to 0.1, 0.3, 0.5, 0.8, 1, 1.5, 1.9, 2, 2.1, 2.5, 4, 8, and 10.

EFICA can be implemented to work efficiently only with a class of distributions for which it is possible to choose appropriate nonlinearities supplying the score functions. The original EFICA implementation from [25] assumes signals having a generalized Gaussian distribution; see Appendix for the definition of this distribution family. A more general implementation using Pham's parametric score function least-square estimator [39] was proposed in [33].

In principle, EFICA does not differ much from FastICA in terms of computational complexity, so it retains its popular property, which is high speed. On the other hand, it outperforms FastICA in terms of accuracy and global convergence (stability), which was demonstrated by various experiments even with real-world signals. Some further improvements of EFICA in terms of speed and accuracy were proposed in [45] and [33].

4.2.1 Example

A simulated example was conducted where 13 signals of the generalized Gaussian distribution, each with a different value of the parameter 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, were mixed by a random mixing matrix and separated. The experiment was repeated 100 times with a fixed length of data $N = 5000$. The achieved average SIR of the signals separated by EFICA and by other ICA methods (Symmetric FastICA with the "tanh" nonlinearity, JADE by J. F. Cardoso [4], and NPICA by Boscolo et al. [3]) was computed and is shown in Fig. 2 as a function of α (one value per separated signal). Likewise, the CRLB computed using (25) and (62) is shown in Fig. 2.

The CRLB exists only for $\alpha > 1$. EFICA approaches the bound, which confirms its efficiency for the generalized Gaussian family. FastICA and JADE do not approach the CRLB, which is mainly caused by the orthogonal constraint. NPICA is close to the CRLB up to some failures that deteriorate the average SIR. However, NPICA requires a much higher computational load than EFICA as it utilizes a nonparametric modeling of the signals' distributions.

4.3 Block EFICA

Block EFICA is a generalization of the EFICA algorithm for piecewise stationary non-Gaussian signals proposed in [29]. The model, first mentioned in Section 2.2.2, assumes that the original signal can be partitioned into a set of M blocks, so that the signals are i.i.d. within each block. The distributions may have different variances and even different distributions on distinct blocks.

Block EFICA searches for appropriate nonlinearities similarly to EFICA, but separately for each block of the pre-separated signals. Assuming that the selected nonlinearities match true score functions and that variance of the signals is constant over the blocks, the asymptotic variance of the non-diagonal normalized gain matrix elements by Block EFICA was shown to be

$$V_{ij}^{\mathrm{BEF}} = \frac{\overline{\kappa}_j}{\overline{\kappa}_i\,\overline{\kappa}_j - 1}, \qquad i \neq j \tag{45}$$

where $\overline{\kappa}_i = \frac{1}{M}\sum_{\ell=1}^{M}\kappa_i^{(\ell)}$. This result corresponds with the CRLB in (28) when taking $(\sigma^2)_i^{(\ell)} = 1$ for all i and ℓ.

5 FastICA in Presence of Additive Noise

In this section, we will assume that the mixed signals also contain additive noise, so the mixing model is

$$\mathbf{X} = \mathbf{AS} + \mathbf{N} \tag{46}$$

where $\mathbf{N}$ has the same size as $\mathbf{X}$ and its rows contain samples of noise. The noise signals are assumed to be Gaussian i.i.d. and uncorrelated[4] with the covariance matrix equal to $\sigma^2\mathbf{I}$. It is worth noting that when $\sigma^2 > 0$, the tasks to identify $\mathbf{A}$ and to separate $\mathbf{S}$ are no longer equivalent. We will henceforth focus on the separation of $\mathbf{S}$; an unbiased estimation of $\mathbf{A}$ through FastICA assuming known σ^2 was studied in [22].

5.1 Signal-to-Interference-plus-Noise Ratio

An appropriate criterion for the evaluation of separated signals by $\mathbf{W}$ is now the Signal-to-Interference-plus-Noise ratio (SINR)[5]. For the kth separated signal, the SINR is equal to [26]

$$\mathrm{SINR}_k = \frac{|\mathbf{G}_{kk}|^2}{\sum_{i=1,i\neq k}^{d} |\mathbf{G}_{ki}|^2 + \sigma^2 \sum_{i=1}^{d} |\mathbf{W}_{ki}|^2}. \tag{47}$$

The values of SINR are bounded unless $\sigma^2 = 0$. The maximum SINR is achieved for

$$\mathbf{W}^{\mathrm{MMSE}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I})^{-1}, \tag{48}$$

which simultaneously minimizes the mean square distance between the original and separated signals, i.e.,

$$\mathbf{W}^{\mathrm{MMSE}} = \arg\min_{\mathbf{W}} \mathrm{E}[\|\mathbf{S} - \mathbf{W}\mathbf{X}\|_F^2]. \tag{49}$$

By putting $\mathbf{W}^{\mathrm{MMSE}}$ into (47), the ultimate bound for the SINR of the kth signal is

$$\frac{\mathbf{V}_{kk}^2}{\sum_{i\neq k}^{d} \mathbf{V}_{ki}^2 + \sigma^2 \sum_{i=1}^{d} (\mathbf{V}\mathbf{A}^{-1})_{ki}^2} \tag{50}$$

where $\mathbf{V} = (\mathbf{I} + \sigma^2(\mathbf{A}^T\mathbf{A})^{-1})^{-1}$. It is worth noting that the latter bound depends on $\mathbf{A}$ unlike the Cramér-Rao bounds for the noise-free cases (Section 2.2).

The asymptotic expansion of (50) for "small" σ^2 was derived in [26] and gives

$$\min \mathrm{SINR}_k = \frac{1}{\sigma^2\|\mathbf{b}_k\|^2} - B_k + \mathcal{O}(\sigma^2), \tag{51}$$

[4] The case when noise signals have general covariance matrix $\mathbf{C}_\mathbf{N}$ can be transformed into the mixing model with uncorrelated noise where the unknown mixing matrix is $\sigma\mathbf{C}_\mathbf{N}^{-1/2}\mathbf{A}$.

[5] Note that SIR does not take into account the presence of the residual noise in separated signals.

where

$$B_k = 2 + \frac{1}{\|\mathbf{b}_k\|^4}\left(\sum_{i\neq k}^{d}(\mathbf{B}\mathbf{B}^T)_{ki}^2 - 2\sum_{i=1}^{d}\mathbf{B}_{ki}(\mathbf{B}\mathbf{B}^T\mathbf{B})_{ki}\right),$$

$\mathbf{B} = \mathbf{A}^{-1}$, and $\mathbf{b}_k^T$ denotes the kth row of $\mathbf{B}$.

The first term in (51) reveals that if the rows of $\mathbf{A}^{-1}$ have the same norm, the ultimate bound (50) is approximately the same for each signal (provided that $\mathbf{A}$ is well conditioned).

5.2 Bias from the Minimum Mean-Squared Error Solution

Without analysis, it is not clear whether FastICA aims to approach the de-mixing transform $\mathbf{W}^{\mathrm{MMSE}}$ or $\mathbf{A}^{-1}$ when noise is present. A more practical method seems to be the former transform as it yields the optimum signals in terms of SINR, that is, the minimum mean-squared error solution

$$\mathbf{S}^{\mathrm{MMSE}} = \mathbf{W}^{\mathrm{MMSE}}\mathbf{X}. \tag{52}$$

We therefore define the bias of an estimated separating matrix $\mathbf{W}$ as

$$\mathrm{E}[\mathbf{W}](\mathbf{W}^{\mathrm{MMSE}})^{-1} - \mathbf{D} \tag{53}$$

where $\mathbf{D}$ is the diagonal matrix that normalizes $\mathbf{S}^{\mathrm{MMSE}}$ to unit scales. The definition of $\mathbf{D}$ comes from the fact that an optimum blind algorithm is expected to yield normalized $\mathbf{S}^{\mathrm{MMSE}}$ since their original scales are unknown to it.

It was shown in [27] that, for “small” σ^2, $\mathbf{D}$ satisfies

$$\mathbf{D} = \mathbf{I} + \frac{1}{2}\sigma^2\mathtt{diag}[\mathbf{H}_{11}, \ldots, \mathbf{H}_{dd}] + \mathcal{O}(\sigma^3), \tag{54}$$

where $\mathbf{H} = (\mathbf{A}^T\mathbf{A})^{-1}$.

5.2.1 Bias of algorithms using the orthogonal constraint

The orthogonal constraint requires that

$$\mathrm{E}[\mathbf{W}\mathbf{X}(\mathbf{W}\mathbf{X})^T] = \mathbf{W}(\mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I})\mathbf{W}^T = \mathbf{I}, \tag{55}$$

so the bias of all constrained algorithms is lower-bounded by

$$\min_{\mathbf{W}} \|\mathbf{W}(\mathbf{W}^{\mathrm{MMSE}})^{-1} - \mathbf{D}\|_F \quad \text{w.r.t.} \quad \mathbf{W}(\mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I})\mathbf{W}^T = \mathbf{I}. \tag{56}$$

It was shown in [28] that $\mathbf{W}$ solving the minimization problem (56) has the property that

$$\mathbf{W}(\mathbf{W}^{\mathrm{MMSE}})^{-1} = \mathbf{I} + \sigma^2\mathbf{\Gamma} + \mathcal{O}(\sigma^3)$$

where $\mathbf{\Gamma}$ is a nonzero matrix obeying $\mathbf{\Gamma} + \mathbf{\Gamma}^T = \mathbf{H}$.

It follows that the bias (53) of ICA algorithms which use the orthogonal constraint has the asymptotic order $\mathcal{O}(\sigma^2)$.

5.2.2 Bias of One-unit FastICA

Consider the situation that FastICA is applied to $\mathbf{S}^{\mathrm{MMSE}}$. An optimum unbiased solution in the sense of (53) is the diagonal matrix $\mathbf{D}$. It was shown in [28] that

$$\mathrm{E}[\mathbf{w}_k^{1\mathrm{U}}] \propto \mathbf{e}_k + O(\sigma^3) \tag{57}$$

where $\mathbf{w}_k^{1\mathrm{U}}$ denotes the kth row of the de-mixing transform by One-unit FastICA (when initialized by $\mathbf{D}$ and then applied to $\mathbf{S}^{\mathrm{MMSE}}$); $\mathbf{e}_k$ denotes the kth row of the identity matrix.

It follows that the asymptotic bias of the one-unit approach has the order $O(\sigma^3)$, that is, lower than $O(\sigma^2)$.

5.2.3 Bias of Symmetric FastICA and EFICA

The biases of FastICA and EFICA derived in the same way satisfy [28]

$$\mathrm{E}[\mathbf{W}^{\mathrm{alg}}](\mathbf{W}^{\mathrm{MMSE}})^{-1} - \mathbf{D} = \frac{1}{2}\sigma^2\mathbf{H} \odot (\mathbf{1}_{d\times d} - \mathbf{I} + \mathbf{M}^{\mathrm{alg}}) + O(\sigma^3), \tag{58}$$

where the superscript $^{\mathrm{alg}}$ signifies the algorithm (either Symmetric FastICA or EFICA). In both cases, $\mathbf{M}$ is not diagonal; $\mathbf{1}_{d\times d}$ is the $d \times d$ matrix of ones. It follows that the bias of both algorithms has the order $O(\sigma^2)$; hence the bias is asymptotically higher than that of One-unit FastICA.

5.3 1FICA

EFICA is an optimal estimator of the separating matrix in terms of the estimation variance when the mixed signals do not contain any noise. However, if the noise is present, the estimate by EFICA is biased and need not be optimal in terms of SINR. By contrast, the above results show that the bias of One-unit FastICA has at least the order $\mathcal{O}(\sigma^3)$.

The only problem is to modify One-unit FastICA to guarantee the estimation of all components. The 1FICA algorithm derived in [27] (also for complex-valued signals) was designed to meet this requirement. It proceeds in three steps.

1. Because of a good global convergence behavior, the initialization is taken from Symmetric FastICA using nonlinearity "tanh" or "rati" followed by the test of saddle points.
2. Each row of the de-mixing transform is fine-tuned through performing few one-unit iterations using an adaptively chosen nonlinearity.
3. To restrain the global solution, the resulting row is accepted if not being too distant from the initialization; otherwise the solution will be the outcome of the first step.

Under mild assumptions, it follows that 1FICA has the same asymptotic bias as One-unit FastICA.

Acknowledgment

We thank Prof. Erkki Oja for his hospitality and fruitful discussions that we had during our visit at HUT in 2006 and later while writing joint papers on statistical analysis of FastICA and on EFICA.

Appendix - Generalized Gaussian Distributions

The normalized random variable distributed according to the generalized Gaussian law has the density function with a shape parameter $\alpha > 0$ defined as

$$f_\alpha(x) = \frac{\alpha\beta_\alpha}{2\Gamma(1/\alpha)} \exp\{-(\beta_\alpha|x|)^\alpha\} \tag{59}$$

where $\Gamma(\cdot)$ is the Gamma function, and

$$\beta_\alpha = \sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}. \tag{60}$$

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 $\alpha \to \infty$.

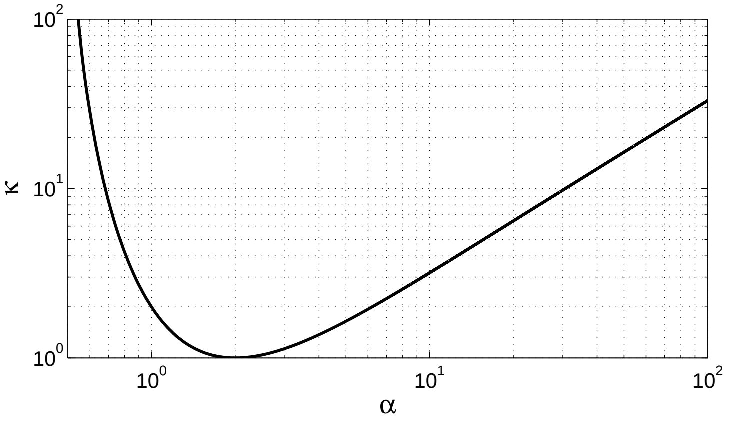

Figure 3: The moment κ of the generalized Gaussian pdf as a function of the shape parameter α according to (62).

The score function of the distribution is

$$\psi_\alpha(x) = -\frac{\frac{\partial f_\alpha(x)}{\partial x}}{f_\alpha(x)} = \frac{|x|^{\alpha-1}\mathrm{sign}(x)}{\mathrm{E}_\alpha[|x|^\alpha]}, \tag{61}$$

which is continuous only for $\alpha > 1$. It can be shown that κ defined similar to (26) depends on α as

$$\kappa_\alpha \;=\; \mathrm{E}_\alpha[\psi_\alpha^2(x)] = \{\mathrm{E}_\alpha[|x|^\alpha]\}^2 = \begin{cases} \frac{\Gamma\left(2-\frac{1}{\alpha}\right)\Gamma\left(\frac{3}{\alpha}\right)}{\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^2} & \text{for} \quad \alpha > 1/2 \\ +\infty & \text{otherwise.} \end{cases} \tag{62}$$

The dependence of κ_α on $\alpha \in [0.5, 100]$ is displayed in Fig. 3. For $\alpha < 0.5$, κ_α goes to infinity and the CRLB does not exist. It may follow that, for $\alpha < 0.5$, there might be estimators whose variances decrease faster than N^{-1} as $N \to +\infty$.

References

[1] T. Adali, M. Anderson, G.-S. Fu, "Diversity in Independent Component and Vector Analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18–33, May 2014.

[2] E. Bingham and A. Hyvärinen, "A Fast Fixed-Point Algorithm for Independent Component Analysis of Complex Valued Signals," *International Journal of Neural Systems*, vol 10, No. I, pp. 1–8, Feb. 2000.

[3] 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.

[4] J.-F. Cardoso and A. Souloumiac, "Blind Beamforming from non-Gaussian Signals", *Radar and Signal Processing, IEE Proceedings F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[5] J.-F. Cardoso, "Blind Signal Separation: Statistical Principles", *Proceedings of the IEEE*, vol. 90, n. 8, pp. 2009-2026, October 1998.

[6] J.-F. Cardoso and D. T. Pham, "Separation of non Stationary Sources. Algorithms and Performance.," in *Independent Components Analysis: Principles and Practice*, pp. 158–180. S. J. Roberts and R. M. Everson (editors), Cambridge University Press, 2001.

[7] J.-C. Chao, S. C. Douglas, "Using Piecewise Linear Nonlinearities in the Natural Gradient and FastICA Algorithms for Blind Source Separation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1813–1816, 2008.

[8] A. Cichocki and S.-I. Amari, *Adaptive Signal and Image Processing: Learning Algorithms and Applications*, Wiley, New York, 2002.

[9] A. Cichocki, R. Zdunek, A. H. Phan and S. I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 2009.

[10] P. Comon, "Independent Component Analysis: A New Concept?", *Signal Processing*, 36(3):287-314, Apr. 1994.

[11] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, Elsevier Ltd., 859 pp., 2010.

[12] N. Delfosse and P. Loubaton, "Adaptive Blind Separation of Independent Sources: A Deflation Approach", *Signal Processing*, Vol. 45, pp. 59-83, 1995.

[13] A. Dermoune, T. Wei, “FastICA Algorithm: Five Criteria for the Optimal Choice of the Nonlinearity Function,” *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2078–2087, 2013.

[14] E. Doron, A. Yeredor and P. Tichavský, “Cramér-Rao Lower Bound for Blind Separation of Stationary Parametric Gaussian Sources”, *IEEE Signal Processing Letters*, vol. 14, no. 6, pp. 417–420, June 2007.

[15] S. C. Douglas, “A Statistical Convergence Analysis of the FastICA Algorithm for Two-Source Mixtures,” *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, pp. 335–339, Nov. 2005.

[16] J. Eriksson, V. Koivunen, “Identifiability, Separability, and Uniqueness of Linear ICA Models,” *IEEE Signal Processing Letters*, vol. 11, no. 7, pp. 601–604, July 2004.

[17] J. Eriksson, V. Koivunen, “Complex Random Vectors and ICA Models: Identifiability, Uniqueness, and Separability,” *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1017–1029, March 2006.

[18] J. Herault and C. Jutten, “Space or Time Adaptive Signal Processing by Neural Network Models, ” *AIP Conference Proceedings*, Vol. 151, pp. 206–211, USA, April 13-16, 1986.

[19] A. Hyvärinen and E. Oja, “A Fast Fixed-Point Algorithm for Independent Component Analysis”, *Neural Computation*, vol. 9, pp. 1483- 1492, 1997.

[20] A. Hyvärinen, “One-Unit Contrast Functions for Independent Component Analysis: A Statistical Analysis,” in *Neural Networks for Signal Processing VII (Proc. IEEE NNSP Workshop 1997)*, Amelia Island, Florida, pp. 388–397, 1997.

[21] A. Hyvärinen, “Fast and Robust Fixed-Point Algorithms for Independent Component Analysis”. *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, 1999.

[22] A. Hyvärinen, “Gaussian Moments for Noisy Independent Component Analysis,” *IEEE Signal Processing Letters*, vol. 6, no. 6, 145–147, 1999.

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

[24] Z. Koldovský, P. Tichavský and E. Oja, “Cramér-Rao lower Bound for Linear Independent Component Analysis”, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, Philadelphia, vol. III, pp. 581–584, March 2005.

[25] Z. Koldovský, P. Tichavský and E. Oja, “Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound”, *IEEE Trans. on Neural Networks*, vol. 17, no. 5, pp. 1265–1277, Sept 2006.

[26] Z. Koldovský, P. Tichavský, “Methods of Fair Comparison of Performance of Linear ICA Techniques in Presence of Additive Noise,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, Toulouse, no. V., pp. 873–876, May 2006.

[27] Z. Koldovský and P. Tichavský, “Blind Instantaneous Noisy Mixture Separation with Best Interference-plus-noise Rejection,” *Proceedings of the 7th International Conference on Independent Component Analysis (ICA2007)*, pp. 730–737, Sept. 2007.

[28] Z. Koldovský and P. Tichavský, “Asymptotic Analysis of Bias of FastICA-based Algorithms in Presence of Additive Noise,” *Technical report no. 2181*, ÚTIA, AV ČR, 2007.

[29] Z. Koldovský, J. Málek, P. Tichavský, Y. Deville, and S. Hosseini, “Blind Separation of Piecewise Stationary NonGaussian Sources,” *Signal Processing*, vol. 89, no. 12, Pages 2570–2584, December 2009.

[30] Te-Won Lee, *Independent Component Analysis: Theory and Applications*, MA: Kluwer, Boston, 237 pp., 1998.

[31] H. L. Li and T. Adali , ”Algorithms For Complex ML ICA And Their Stability Analysis Using Wirtinger Calculus”, *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6156–6167, Dec. 2010.

[32] B. Loesch and B. Yang, “Cramér-Rao Bound for Circular and Noncircular Complex Independent Component Analysis,” *IEEE Transactions on Signal Processing*, vol. 61, no. 2, pp. 365–379, 2013.

[33] J. Málek, Z. Koldovský, S. Hosseini, and Y. Deville, “A Variant of EFICA Algorithm with Adaptive Parametric Density Estimator”, *8th International Workshop on Electronics, Control, Modelling, Measurement, and Signals (ECMS 2007)*, pp. 79–84, Liberec, Czech Republic, May 2007.

[34] [online] Matlab codes at `http://itakura.ite.tul.cz/zbynek/downloads.htm`.

[35] J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, “Deflation-Based FastICA With Adaptive Choices of Nonlinearities,” *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5716–5724, 2014.

[36] E. Oja, Z. Yuan, “The FastICA Algorithm Revisited: Convergence Analysis,” *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1370–1381, 2006.

[37] E. Ollila, K. Hyon-Jung, V. Koivunen, “Compact Cramér-Rao Bound Expression for Independent Component Analysis,” *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1421–1428, April 2008.

[38] E. Ollila, “The Deflation-based FastICA Estimator: Statistical Analysis Revisited,” *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1527–1541, March 2010.

[39] D. T. Pham, P. Garat, “Blind separation of mixture of independent sources through a quasi-maximum likelihood approach”, *IEEE Trans. on Signal Processing*, vol. 45, no. 7, pp. 1712–1725, July 1997.

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

[41] H. Shen, M. Kleinsteuber, K. Huper, “Local Convergence Analysis of FastICA and Related Algorithms,” *IEEE Transactions on Neural Networks*, vol. 19, no. 6, pp. 1022–1032, June 2008.

[42] Tianwen Wei, “On the Spurious Solutions of the FastICA Algorithm”, *IEEE Statistical Signal Processing Workshop 2014*, pp. 161–164, 2014.

[43] P. Tichavský and Z. Koldovský, “Optimal Pairing of Signal Components Separated by Blind Techniques”, *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 119–122, 2004.

[44] P. Tichavský, Z. Koldovský, and E. Oja, “Performance Analysis of the FastICA Algorithm and Cramér-Rao Bounds for Linear Independent Component Analysis”, *IEEE Trans. on Signal Processing*, vol. 54, no. 4, April 2006.

[45] P. Tichavský, Z. Koldovský, and E. Oja, “Speed and Accuracy Enhancement of Linear ICA Techniques Using Rational Nonlinear Functions,” *Proceedings of the 7th International Conference on Independent Component Analysis (ICA2007)*, pp. 285–292, Sept. 2007.

[46] P. Tichavský, Z. Koldovský, and E. Oja, “Corrections of the ’Performance Analysis of the FastICA Algorithm and Cramér-Rao Bounds for Linear Independent Component Analysis’ ”, *IEEE Trans. on Signal Processing*, vol. 56, no. 4, pp. 1715–1716, April 2008.

[47] T. Wei, “Asymptotic Analysis of the Generalized Symmetric FastICA Algorithm,” *IEEE Workshop on Statistical Signal Processing (SSP 2014)*, pp. 460–463, 2014.

[48] A. Yeredor, “Blind Separation of Gaussian Sources With General Covariance Structures: Bounds and Optimal Estimation,” *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5057–5068, Oct. 2010.

[49] V. Zarzoso, P. Comon, and M. Kallel, “How Fast is FastICA?”, *Proceedings of the 14th European Signal Processing Conference (EUSIPCO-2006)*, Florence, Italy, Sept. 4-8, 2006.

[50] Y. Zhang and S. A. Kassam, “Optimum Nonlinearity and Approximation in Complex FastICA,” *Proc. of the 46th Conf. on Information Sciences and Systems (CISS)*, pp. 1–6, 2012.