Academy of Sciences of the Czech Republic
INSTITUTE OF INFORMATION THEORY AND AUTOMATION
Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic

RESEARCH REPORT

Zbyněk Koldovský and Petr Tichavský

Asymptotic Analysis of Bias of FastICA-based Algorithms in Presence of Additive Noise

No. 2181 January 2007

[0]This work was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and through the Grant 102/07/P384 of the Grant Agency of the Czech Republic.

Asymptotic Analysis of Bias of FastICA-based Algorithms in Presence of Additive Noise

Zbyněk Koldovský and Petr Tichavský

Abstract

The idea that common blind techniques based on Independent Component Analysis (ICA) behave in noisy environment like a biased MMSE separator (sometimes called Maximum Ratio Combiner (MRC)) was introduced in our recent work [3]. In this paper, we put this in more precise terms by doing an analysis of the bias of approaches that are based on known ICA algorithm FastICA. We show that the one-unit approach is the best MMSE estimator in terms of the bias.

1 Introduction

The noisy model of Independent Component Analysis (ICA) considered in this paper, is

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

where $\mathbf{S}$ denotes a vector of d independent random variables representing the original signals, $\mathbf{A}$ is an unknown regular $d \times d$ mixing matrix, and $\mathbf{X}$ represents the observed mixed signals. The noise $\mathbf{N}$ denotes a vector of Gaussian variables having the covariance matrix $\boldsymbol{\Sigma}$. Further we will assume that $\boldsymbol{\Sigma}$ equals to the identity matrix $\mathbf{I}$, which is, in some sense, the most difficult case. Consequently, σ^2 is the variance of the added noise to the mixed signals.

Note that we do *not* demand $\mathbf{A}$ to have orthogonal rows, i.e., $\mathbf{A}\mathbf{A}^T$ need not be diagonal. In [1], the case with orthogonal mixing matrix is referred to as the "source noise" model, because, then, (1) can be written as $\mathbf{X} = \mathbf{A}(\mathbf{S} + \sigma\mathbf{A}^{-1}\mathbf{N})$, where $\mathbf{S} + \sigma\mathbf{A}^{-1}\mathbf{N}$ are independent due to the property of $\mathbf{A}$ and gaussianity of $\mathbf{N}$, therefore, (1) becomes the noiseless model with the original independent sources $\mathbf{S} \leftarrow \mathbf{S} + \sigma\mathbf{A}^{-1}\mathbf{N}$.

It is characteristic for most ICA methods that they were derived for the noiseless case, so to solve the task of estimating the mixing matrix $\mathbf{A}$ or its inversion $\mathbf{W} = \mathbf{A}^{-1}$. Then, abilities to separate noised data are studied experimentally, and the non-vanishing estimation error as $N \to +\infty$, N being length of data, is taken for a bias caused by the noise. To compensate such bias, several techniques were proposed [6, 7], but their drawback is that the covariance structure of the noise needs to be known *a priori* [11].

Note that it was shown by Davies in [9] that the mixing matrix $\mathbf{A}$ and the covariance structure of the Gaussian noise $\mathbf{N}$ are not jointly identifiable, therefore, it is not possible (meaningful) to estimate them both from the data.

In our recent work [3], we claim that working separating criteria such as mutual information or entropy aim at signal quality mainly, which generalizes the idea when the noise is present.

Then, the blind source separation problem is defined primarily as the problem of estimating a de-mixing matrix $\mathbf{W}$ such that the difference between the original and estimated sources, $\mathbf{S}$ and $\mathbf{WX}$, after the permutation, scale, and sign or phase ambiguities are resolved, is as small as possible.[1] Note that the purpose of minimizer of $\mathrm{E}[\|\mathbf{S} - \mathbf{WX}\|^2]$, which is referred to as the minimum mean square error (MMSE) estimator, is the same. This gives reasons to regard many classical ICA methods as MMSE estimators.

Hence, we suggest [3] to measure the quality of separation by the interference-plus-noise-to-signal ratio (INSR). In the case $\mathbf{\Sigma} = \mathbf{I}$, the INSR of a k-th estimated signal can be computed as

$$\mathrm{INSR}_k = \frac{\sum_{i\neq k}^{d}(\mathbf{BA})_{ki}^2 + \sigma^2 \sum_{i=1}^{d} \mathbf{B}_{ki}^2}{(\mathbf{BA})_{kk}^2}, \tag{2}$$

where $\mathbf{B}$ is the separating transformation.

First, it is important that (2) takes into account the residual noise included in the estimated signal, thus, it evaluates the signal's quality; it does not simultaneously quantify the accuracy of estimation of system parameters (matrix $\mathbf{A}$ or $\mathbf{A}^{-1}$), like interference-to-signal ratio (ISR) defined for $\sigma = 0$. Second, the solutions that minimize (2) are known to be given by the MMSE separating matrix, denoted by $\mathbf{W}^{\mathrm{MMSE}}$, that takes the form

$$\mathbf{W}^{\mathrm{MMSE}} = \mathbf{A}^T(\mathbf{AA}^T + \sigma^2\mathbf{I})^{-1} . \tag{3}$$

Signals given by $\mathbf{W}^{\mathrm{MMSE}}\mathbf{X}$ will be further called the *MMSE solution*.

This work presents deeper study of behavior of ICA techniques in the noisy setup. Specifically, we analyze bias of three FastICA approaches from the MMSE solution. The approaches: a one-unit and a symmetric [10], and EFICA [5], differ in mutual interactions of estimated sources mainly. Similar dependencies are applied in wide family of ICA algorithms, which partly gives general coverage to our conclusions. We show that the bias of the one-unit approach is asymptotically of lower order, in general, than that of the other approaches.

In the performance analysis, also plays an important role the variance of de-mixing matrix estimation, which is however difficult to be derived in a closed-form if the noise is considered. Unless the effect of variance is small, the bias is a sufficient performance criterion only if N is large enough. A resolution in this bias-variance dilemma is given in [11] for separation of noisy mixtures of finite-alphabet signals.

2 Review of FastICA and its variants

In this section, we summarize brief description of the one-unit and the symmetric version of algorithm FastICA [10] and its asymptotically efficient variant, called EFICA [5].

The first step of all variants consists in decorrelating the mixed data $\mathbf{X}$ transforming them into a pre-processed data $\mathbf{Z}$ as

$$\mathbf{Z} \;=\; \mathbf{C}^{-1/2}\mathbf{X}, \tag{4}$$

where $\mathbf{C} = \mathrm{E}[\mathbf{XX}^T]$ is the data covariance matrix.

[1]It can be assumed, without any loss in generality, that the original signals have all unit variances.

In practice, the theoretical matrix $\mathbf{C}$ is replaced by a sample covariance matrix, and, similarly, expectation values are replaced by their sample mean counterparts. However, for the asymptotic analysis of bias presented here, the exact values may be considered, or we can assume that $N = +\infty$.

One-unit version of FastICA estimates one de-mixing vector $\mathbf{w}_k^{1\text{U}}$ iteratively via

$$\mathbf{w}_k^+ \leftarrow \mathrm{E}[\mathbf{Z}g(\mathbf{w}_k^{1\text{U}T}\mathbf{Z})] - \mathbf{w}_k^{1\text{U}}\,\mathrm{E}[g'(\mathbf{w}_k^{1\text{U}T}\mathbf{Z})] \tag{5}$$

$$\mathbf{w}_k^{1\text{U}} \leftarrow \mathbf{w}_k^+/\|\mathbf{w}_k^+\| \tag{6}$$

until convergence is achieved. Here $g(\cdot)$ is a smooth nonlinear function that approximates/surrogates the score function of the distribution of the signal to-be estimated [4]. To estimate the whole matrix $\mathbf{W}$, d one-unit algorithms should be run with suitable initializations. In simulations, we use the initialization taken from the result obtained by the symmetric algorithm. This prevents from estimating any component twice. A more sophisticated procedure is proposed in [11].

The symmetric variant of the algorithm FastICA performs the iteration (5) in parallel for all d separating vectors, but the normalization in (6) is replaced by a symmetric orthogonalization that mediates mutual incidence of the paralel algorithms

$$\mathbf{W}^{\text{sym}} \leftarrow (\mathbf{W}^+\mathbf{W}^{+T})^{-\frac{1}{2}}\mathbf{W}^{+T}, \tag{7}$$

where $\mathbf{W}^+ = [\mathbf{w}_1^+ \dots \mathbf{w}_d^+]^T$. In the original approaches of FastICA we use $g(x) = \tanh(x)$.

The algorithm EFICA [5] is a sophisticated procedure combining the symmetric approach with a choice of the nonlinearity $g_k(\cdot)$, that is chosen for each signal separately, and a refinement step, where the symmetric orthogonalization is performed for each separating vector with suitable optimum weights.

In the noise-less case, EFICA is asymptotically efficient if g_k are score functions of the distributions of $\mathbf{S}$. Even if this is not fulfilled, the algorithm is more accurate than the other variants thanks to the refinement step (assuming that the same g was used in all variants).

Note that at the end of the iteration processes of all approaches, the estimated matrix $\mathbf{W}$ is pre-multiplied by $\mathbf{C}^{-1/2}$ from the right, $\mathbf{W} \leftarrow \mathbf{W} \cdot \mathbf{C}^{-1/2}$.

3 Bias Analysis

In this section, we derive the analysis of bias from the MMSE solution assuming "small" σ. We show that the bias is of order $\mathcal{O}(\sigma^3)$, which is negligible in comparison with bias of other techniques. The analysis is done in a similar fashion like [4], i.e., we assume that the iterative part starts from the MMSE solution, and the algorithm stops after one iteration. This assumption is instrumental in theoretical considerations only; the results are further compared with real algorithms' performances.

Let $\widehat{\mathbf{W}}$ be an estimate of de-mixing matrix separating the data $\mathbf{X}$. Then, the deviation of $\mathrm{E}[\widehat{\mathbf{W}}](\mathbf{W}^{\text{MMSE}})^{-1}$ from a diagonal matrix responds to the bias from the MMSE solution. More precisely, note that $\mathbf{S}^{\text{MMSE}}$ may not be necessarily normalized to have unit variance, unlike outcome of common blind separation methods, that produce normalized components due to the inherent indeterminacy of their scales. Therefore, we study bias of $\widehat{\mathbf{W}}(\mathbf{W}^{\text{MMSE}})^{-1}$ from a diagonal matrix $\mathbf{D}$ such that $\mathbf{D} \cdot \mathbf{S}^{\mathrm{MMSE}}$ are normalized copies of $\mathbf{S}^{\mathrm{MMSE}}$. Thus, the bias of $\widehat{\mathbf{W}}$ is

$$\mathcal{B}[\widehat{\mathbf{W}}] = \mathrm{E}[\widehat{\mathbf{W}}](\mathbf{W}^{\mathrm{MMSE}})^{-1} - \mathbf{D}. \tag{8}$$

The definition of $\mathbf{D}$ gives

$$\begin{aligned} \mathbf{D} &= \mathrm{diag}\left[\mathrm{diag}\left[\mathrm{E}[\mathbf{S}^{\mathrm{MMSE}}\mathbf{S}^{\mathrm{MMSE}T}]\right]\right]^{-\frac{1}{2}} \\ &= \mathrm{diag}\left[\mathrm{diag}\left[(\mathbf{I}+\sigma^2\mathbf{V})^{-1}\right]\right]^{-\frac{1}{2}} \\ &= \mathbf{I} + \frac{1}{2}\sigma^2 \mathrm{diag}[\mathrm{diag}[\mathbf{V}]] + \mathcal{O}(\sigma^3) \end{aligned} \tag{9}$$

From here we use the notation $\mathbf{W} = \mathbf{A}^{-1}$ and $\mathbf{V} = \mathbf{W}\mathbf{W}^T$, and we adopt the Matlab notation for the operator $\mathrm{diag}[\cdot]$, that gives a vector of diagonal elements or a diagonal matrix according to whether the argument is a matrix or a vector, respectively.

Next, thanks to the equivariance property of FastICA, the resulting $\widehat{\mathbf{W}}$ is independent of initialization, and we can put a simplifying assumption that FastICA is started just from the data $\mathbf{S}^{\mathrm{MMSE}} = \mathbf{W}^{\mathrm{MMSE}}\mathbf{X}$. Then, the bias of any equivariant algorithm is

$$\mathcal{B}[\widehat{\mathbf{W}}] = \mathrm{E}[\widehat{\mathbf{W}}] - \mathbf{D}, \tag{10}$$

where $\widehat{\mathbf{W}}$ is the resulting de-mixing matrix that separates the data $\mathbf{S}^{\mathrm{MMSE}}$. For easy notation, we will skip the operator $\mathrm{E}[\cdot]$ further in the paper, because we always assume that $N \to +\infty$ and also the consistency of the estimator.

To begin with the analysis, we consider the algorithm FastICA to be started from the MMSE solution. It holds that

$$\mathbf{S}^{\mathrm{MMSE}} = \mathbf{W}^{\mathrm{MMSE}}\mathbf{X} = \mathbf{S} + \sigma\mathbf{W}\mathbf{N} - \sigma^2\mathbf{V}\mathbf{S} + \mathcal{O}(\sigma^3). \tag{11}$$

The first step of FastICA is the preprocessing (4) of the input data, which can be described in following steps

$$\begin{aligned} \mathbf{C} &= \mathrm{E}[\mathbf{S}^{\mathrm{MMSE}}\mathbf{S}^{\mathrm{MMSE}T}] = \mathbf{I} - \sigma^2\mathbf{V} + \mathcal{O}(\sigma^3) && (12) \\ \mathbf{C}^{-\frac{1}{2}} &= \mathbf{I} + \frac{1}{2}\sigma^2\mathbf{V} + \mathcal{O}(\sigma^3) && (13) \\ \mathbf{Z} &= \mathbf{C}^{-\frac{1}{2}}\mathbf{S}^{\mathrm{MMSE}} = \\ &= \mathbf{S} + \sigma\mathbf{W}\mathbf{N} - \frac{1}{2}\sigma^2\mathbf{V}\mathbf{S} + \mathcal{O}(\sigma^3). && (14) \end{aligned}$$

Now, we will consider one iteration step (5) starting with the initialization $\mathbf{w}_k^{1\mathrm{U}} = \mathbf{e}_k$. Following expansions will be needed, where $\mathbf{e}_\ell^T$, $\mathbf{w}_\ell^T$, and, $\mathbf{v}_\ell^T$ denote, respectively, the ℓ-th row of $\mathbf{I}$, $\mathbf{W}$, and $\mathbf{V}$.

$$\begin{aligned} \mathbf{e}_\ell^T\mathbf{Z} &= s_\ell + \sigma\mathbf{w}_\ell^T\mathbf{N} - \frac{1}{2}\sigma^2\mathbf{v}_\ell^T\mathbf{S} + \mathcal{O}(\sigma^3) && (15) \\ g(\mathbf{e}_k^T\mathbf{Z}) &= g(s_k) + g'(s_k)(\sigma\mathbf{N}^T\mathbf{w}_k - \frac{1}{2}\sigma^2\mathbf{S}^T\mathbf{v}_k) + \\ &\quad + \frac{1}{2}g''(s_k)(\sigma^2\mathbf{w}_k^T\mathbf{N}\mathbf{N}^T\mathbf{w}_k) + \mathcal{O}(\sigma^3), && (16) \end{aligned}$$

Using (15) and (16) in (5), the ℓ-th element of the vector $\mathbf{w}_k^+$, for $\ell \neq k$, is

$$(\mathbf{w}_k^+)_\ell = -\frac{1}{2}\sigma^2 \mathbf{V}_{k\ell}\tau_k + \mathcal{O}(\sigma^3), \tag{17}$$

where $\tau_k = \mathrm{E}[s_k g(s_k) - g'(s_k)]$, and for $\ell = k$ we get

$$(\mathbf{w}_k^+)_k = \tau_k + \sigma^2 \frac{1}{2}\mathbf{V}_{kk}\delta_k + \mathcal{O}(\sigma^3) \tag{18}$$

where δ_k is a scalar that depends on the distribution of s_k and on the nonlinear function g and its derivatives to the third order.

3.1 Bias of the one-unit FastICA

In order to receive the de-mixing vector $\mathbf{w}_k^{1\mathrm{U}}$, $\mathbf{w}_k^+$ needs to be normalized and pre-multiplied by $\mathbf{C}^{-1/2}$. Hence, the resulting vector is proportional to

$$\mathbf{C}^{-1/2}\,\mathbf{w}_k^+ = \tau_k \mathbf{e}_k + \frac{1}{2}\sigma^2 \mathbf{V}_{kk}(\tau_k + \delta_k)\mathbf{e}_k + \mathcal{O}(\sigma^3) \tag{19}$$

Since ℓ-th element of (19), for each $\ell \neq k$, is of order $\mathcal{O}(\sigma^3)$, it follows that the bias (10) of the whole separating matrix, that stacks results of d one-unit algorithms $\mathbf{W}^{1\mathrm{U}} = [\mathbf{w}_1^{1\mathrm{U}}, \ldots, \mathbf{w}_d^{1\mathrm{U}}]^T$, is

$$\mathcal{B}[\mathbf{W}^{1\mathrm{U}}] = \mathbf{W}^{1\mathrm{U}} - \mathbf{D} = \mathcal{O}(\sigma^3). \tag{20}$$

3.2 Bias of the inversion solution

It is interesting to compare the previous result with the solution that is given by the exact inversion of the mixing matrix, i.e. $\mathbf{WX} = \mathbf{S} + \sigma\mathbf{WN}$; the signals will be called *the inversion solution*. The relation between $\mathbf{WX}$ and the MMSE solution given by $\mathbf{WX} = \mathbf{W}(\mathbf{W}^{\mathrm{MMSE}})^{-1}\mathbf{S}^{\mathrm{MMSE}}$ reveals the difference between the solutions:

$$\mathbf{W}(\mathbf{W}^{\mathrm{MMSE}})^{-1} = \mathbf{W}(\mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I})\mathbf{W}^T = \mathbf{I} + \sigma^2\mathbf{V}. \tag{21}$$

To get the "bias" of the inversion solution according to (8), normalized signals $\mathbf{WX}$, i.e. $\mathrm{diag}[\mathrm{diag}[\mathrm{E}[\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}]]]^{(-1/2)}\mathbf{WX}$, should be considered. This results in

$$\begin{aligned} \mathcal{B}[\mathbf{W}] &= \Big(\mathrm{diag}[\mathrm{diag}[\mathbf{I} + \sigma^2\mathbf{V}]]\Big)^{-\frac{1}{2}}(\mathbf{I} + \sigma^2\mathbf{V}) - \mathbf{D} = \\ &= \sigma^2\mathbf{V} \odot (\mathbf{1}_{d\times d} - \mathbf{I}) + \mathcal{O}(\sigma^3), \end{aligned} \tag{22}$$

where $\mathbf{1}_{d\times d}$ denotes $d \times d$ matrix of ones, and $\odot$ denotes the element-wise product. The bias is proportional to σ^2, and, in general, it is greater than that of the one-unit FastICA. In other words, **the one-unit algorithm produces components that are closer to the MMSE solution than to the inversion solution**.

3.3 Bias of algorithms using the orthogonal constraint

Large number of ICA algorithms (e.g. JADE, symmetric FastICA, etc.) use an orthogonal constraint, i.e., they enforce the separated components to have sample correlations equal to zero. Since the second-order statistics cannot be estimated perfectly, this constraint compromises the separation quality [8].

Thereunto, the constraint causes estimation bias when the noise is present. For instance, it was shown in [3] that the MMSE components can be significantly correlated, thus, cannot have zero correlations. Here we show that the bias of all ICA algorithms that use the orthogonality constraint, has the asymptotic order $\mathcal{O}(\sigma^2)$.

The orthogonality constraint can be written as

$$\mathrm{E}[\widehat{\mathbf{W}}\mathbf{X}(\widehat{\mathbf{W}}\mathbf{X})^T] = \widehat{\mathbf{W}}(\mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I})\widehat{\mathbf{W}}^T = \mathbf{I}, \tag{23}$$

It follows that the Frobenius norm of the bias of all constrained algorithms is lower bounded by

$$\min_{\widehat{\mathbf{W}}(\mathbf{A}\mathbf{A}^T+\sigma^2\mathbf{I})\widehat{\mathbf{W}}^T=\mathbf{I}} \|\widehat{\mathbf{W}}(\mathbf{W}^{\mathrm{MMSE}})^{-1} - \mathbf{D}\|_F^2, \tag{24}$$

where the minimization proceeds for $\widehat{\mathbf{W}}$.

It is shown in Appendix A that the argument of (24), denoted by $\mathbf{W}^{\mathrm{orth}}$, obeys

$$\mathbf{W}^{\mathrm{orth}}(\mathbf{W}^{\mathrm{MMSE}})^{-1} = \mathbf{I} + \sigma^2\mathbf{\Gamma} + \mathcal{O}(\sigma^3), \tag{25}$$

where $\mathbf{\Gamma}$ is a nonzero matrix such that $\mathbf{\Gamma} + \mathbf{\Gamma}^T = \mathbf{V}$. Hence, the bias of $\mathbf{W}^{\mathrm{orth}}$ from the MMSE solution is

$$\mathcal{B}[\mathbf{W}^{\mathrm{orth}}] = \sigma^2\mathbf{\Gamma} \odot (\mathbf{1}_{d\times d} - \mathbf{I}) + \mathcal{O}(\sigma^3). \tag{26}$$

3.4 Bias of the symmetric FastICA and EFICA

In Appendix B, it is shown that the biases of the symmetric FastICA and of EFICA [5] can be expressed as

$$\mathcal{B}[\widehat{\mathbf{W}}] = \frac{1}{2}\sigma^2\mathbf{V} \odot (\mathbf{1}_{d\times d} - \mathbf{I} + \mathbf{H}) + \mathcal{O}(\sigma^3), \tag{27}$$

where, for the symmetric FastICA,

$$\mathbf{H}_{k\ell} = \mathbf{H}_{k\ell}^{\mathrm{SYM}} = \frac{|\tau_\ell| - |\tau_k|}{|\tau_k| + |\tau_\ell|}, \tag{28}$$

and, for EFICA,

$$\mathbf{H}_{k\ell} = \mathbf{H}_{k\ell}^{\mathrm{EF}} = \frac{c_{k\ell}|\tau_\ell| - |\tau_k|}{|\tau_k| + c_{k\ell}|\tau_\ell|}, \tag{29}$$

where $c_{k\ell} = \frac{|\tau_\ell|\gamma_k}{|\tau_k|(\gamma_\ell+\tau_\ell^2)}$ for $k \neq \ell$ and $c_{kk} = 1$ are the optimum weights used in the symmetric orthogonalization when computing the k-th signal. Next, $\gamma_k = \mathrm{E}[g_k^2(s_k)] - \mathrm{E}^2[s_k g_k(s_k)]$, and g_k is the nonlinear function chosen for the k-th signal.

3.4.1 Discussion of the results given by (27)

In the noiseless case, it is a known fact that the optimal nonlinear functions g_k used in any variant of FastICA are the score functions of distributions of the original signals [1, 4], i.e. $\psi_k(x) = -f_k'(x)/f_k(x)$, where $f_k(\cdot)$ is pdf of the k-th original signal. This is valid under several assumptions: The score functions must be differentiable, and their mean square values must exist, i.e. $\kappa_k = \mathrm{E}[\psi_k^2(x)] < +\infty$. Note that $\kappa_k \geq 1$, where the equality is attained iff the k-th distribution is standard Gaussian [4].

We will adopt this special case in our discussion, i.e., we assume that $g_k = \psi_k$. For simplicity, we use this assumption even for the symmetric FastICA, where, in practice, one nonlinearity g is used for all signals. Then, it holds that the quantities defined in sections 3 and 3.4 are $\gamma_k = |\tau_k| = \kappa_k - 1$, and (28) and (29) take form

$$\mathbf{H}_{k\ell}^{\mathrm{SYM}} = \frac{\kappa_\ell - \kappa_k}{\kappa_k + \kappa_\ell - 2} \tag{30}$$

$$\mathbf{H}_{k\ell}^{\mathrm{EF}} = \frac{2\kappa_\ell - \kappa_k\kappa_\ell - 1}{\kappa_k\kappa_\ell - 1}. \tag{31}$$

Before discussing the latter results, we point out an important property of the symmetric FastICA that follows from skew-symmetric nature of (28). This property is presumably due to the orthogonal constraint used by the method, because $\frac{1}{2}\mathbf{V} \odot (\mathbf{1}_{d\times d} - \mathbf{I} + \mathbf{H}^{\mathrm{SYM}})$ obeys the same property like the matrix $\mathbf{\Gamma}$ in (26). In view of Appendix A, the solution given by the symmetric FastICA is a local minimum of the Frobenius norm in (24), at least.

Our discussion of the results (30) and (31) will proceed in following way. From (27) it follows that the bias of $k\ell$-th element of de-mixing matrix from the MMSE solution is zero iff $\mathbf{H}_{k\ell} = -1$. On the other hand, here, we may consider the bias from the inversion solution, also: by comparing (22) and (27). Easily follows that the bias from the inversion solution is zero iff $\mathbf{H}_{k\ell} = 1$. In conclusion, the bias moves between the MMSE and the inversion solution according to values of elements of $\mathbf{H}$ in $[-1, 1]$. Finally, the overall bias of k-th signal is characterized by all elements of the corresponding row of $\mathbf{H}$.

First, $\mathbf{H}_{k\ell} = -1$ for both (30) and (31) if only $\kappa_\ell = 1$, which means that the ℓ-th signal is Gaussian. It follows that the bias of k-th estimated signal from the MMSE solution is zero if and only if all other signals are Gaussian. This is in accord with the fact that, in such case, both the symmetric FastICA and the EFICA behave like the one-unit algorithm.

Similarly, for both methods, $\mathbf{H}_{k\ell} = 1$ if only $\kappa_k = 1$. Consequently, a Gaussian signal is estimated with zero bias from the inversion solution, however, it should be noted than at most one signal may have Gaussian distribution to be identifiable.

It is worth to consider cases such as $\kappa_k \to +\infty$, $\kappa_\ell \to +\infty$, or $\kappa_k, \kappa_\ell \to +\infty$, because large κ means strong non-Gaussianity of the corresponding signal.

1. For $\kappa_k \to +\infty$, $\mathbf{H}_{k\ell}^{\mathrm{SYM}} \to -1$ and $\mathbf{H}_{k\ell}^{\mathrm{EF}} \to -1$ for all $\ell \neq k$. This means that if k-th signal is strongly non-gaussian compared to the others, its estimated counterpart is closed to the MMSE solution. Once again, the performance of the algorithms is similar to that of the one-unit algorithm, in this case.

2. For $\kappa_\ell \to +\infty$, $\mathbf{H}_{k\ell}^{\mathrm{SYM}} \to +1$ and $\mathbf{H}_{k\ell}^{\mathrm{EF}} \to \frac{2-\kappa_k}{\kappa_k}$, $\ell \neq k$. This demonstrates typical behavior of the symmetric FastICA following from the orthogonal constraint used by it:

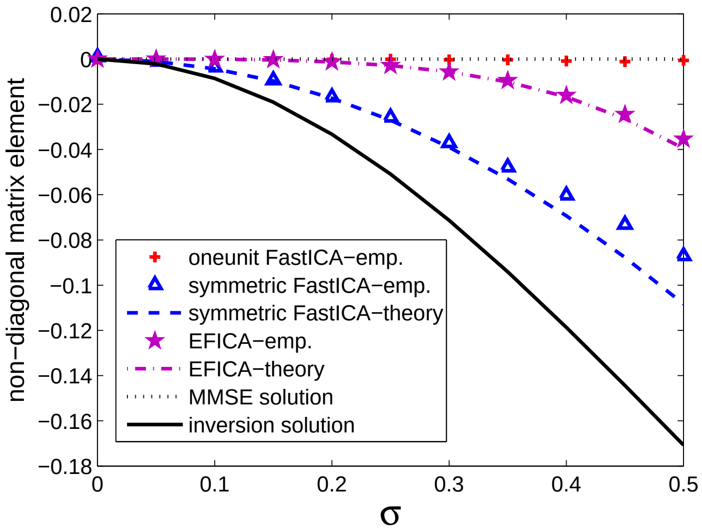

Figure 1: Mean value of the 1,2-th element of estimated separating matrices from the first experiment.

A strongly non-Gaussian signal (ℓ-th signal) brings plus bias from the MMSE solution into the other signals.

The behavior of EFICA is different since it relaxes the orthogonal constraint. The bias of $k\ell$-th element of the de-mixing matrix still depends on (non)gaussianity of the k-th signal through $\frac{2-\kappa_k}{\kappa_k}$. For $\kappa_k \rightarrow +\infty$ also, $\mathbf{H}^{\mathrm{EF}}_{k\ell} \rightarrow -1$. Consequently, if all signals have strongly non-Gaussian distribution, the EFICA's bias of from the MMSE solution is small.

The latter consequence regarding EFICA puts up to belief in its superior ability to separate strongly non-Gaussian components in the noisy environment. However, note that strongly non-gaussian signals such as finite alphabet sources [11] lack existing score function as required at the beginning of this discussion.

In conclusion, the symmetric FastICA and the algorithm EFICA are generally more biased from the MMSE solution then the one-unit approach.

4 Simulations

In this section, we experimentally validate results of our analysis. In the first example, two random BPSK signals were mixed with a fixed matrix

$$\mathbf{A} = \begin{pmatrix} -0.49 & 0.53 \\ 1.93 & 1.92 \end{pmatrix},$$

for which $\mathbf{V} = \left(\begin{smallmatrix} 1.00 & -0.87 \\ -0.87 & 1.00 \end{smallmatrix}\right)$, and a Gaussian noise with covariance $\sigma^2\mathbf{I}$ was added. Then, the MMSE components $\mathbf{S}^{\mathrm{MMSE}}$ were computed, and the above mentioned FastICA approaches were applied to their normalized counterparts. The experiment was repeated in 100 independent trials for each σ taken from $[0, 1/2]$. The large length of generated signals $N = 100000$ was chosen to eliminate the effect of variance of separating matrix estimation. Note that the chosen $\mathbf{A}$ is *not* orthogonal ($\mathbf{V}$ is not diagonal).

Mean values of elements of the resulting matrices, that were normalized, re-signed and row-reshuffled, were computed and compared with the identity matrix, that gives the MMSE components. Since the differences of diagonal elements are of lower order, and the non-diagonal elements are (in this case) the same, we show in Figure 1 mean value of the 1,2-th element only. The proximity of the empirical values to the theoretical ones corroborates validity of our analyses.

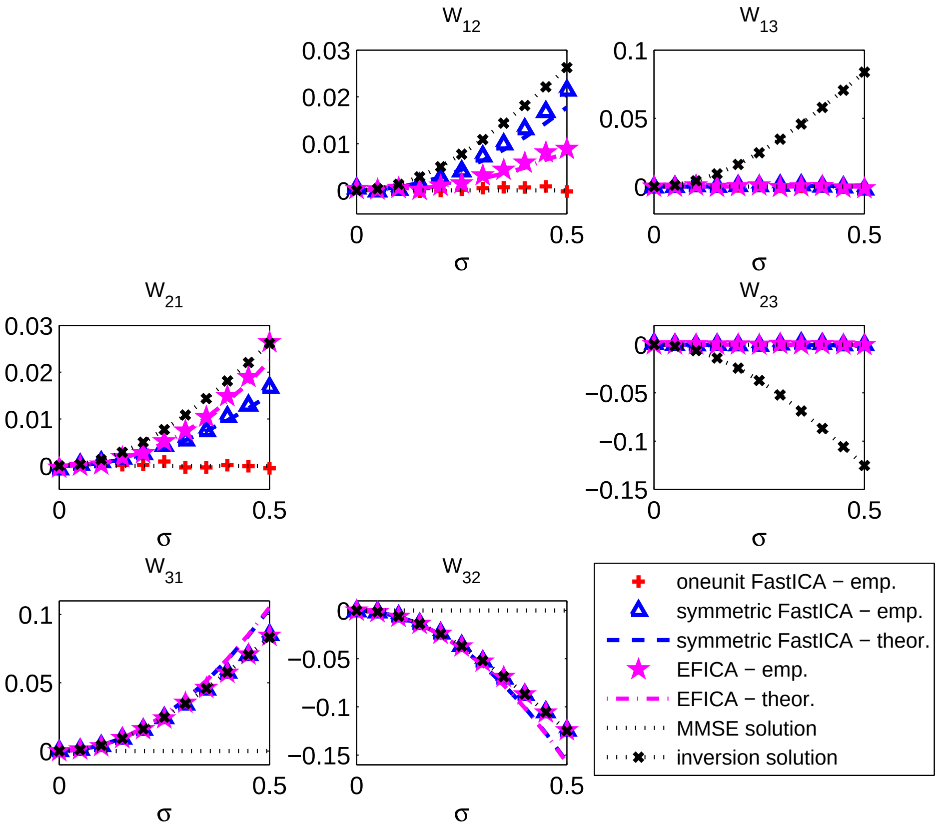

Figure 2: Mean values of elements of estimated separating matrices that separate (row by row), respectively, the uniform, the Laplacean, and the Gaussian component.

The second experiment is similar to the previous one, but three signals are separated. The distribution of the signals are the uniform, the Laplacean, and the Gaussian. The mixing matrix is

$$\mathbf{A} = \begin{pmatrix} 0.23 & -1.42 & -0.89 \\ -0.37 & 0.31 & 1.31 \\ 1.26 & -0.60 & -0.64 \end{pmatrix} \quad \text{for that} \quad \mathbf{V} = \begin{pmatrix} 1.00 & 0.13 & 0.42 \\ 0.13 & 1.00 & -0.63 \\ 0.42 & -0.63 & 1.00 \end{pmatrix}.$$

Again, results shown in Figure 2 support our analyses, and several conclusions given by subsection 3.4.1: The one-unit approach is the best MMSE estimator for all signals up to the

Gaussian one that cannot be estimated thereby, therefore, its results are excluded from graphs showing elements $\mathbf{W}_{31}$ and $\mathbf{W}_{32}$. For the symmetric FastICA and EFICA it holds that the most non-gaussian signal, the uniformly distributed one, is the least one biased from the corresponding MMSE component. Conversely, their estimate of the Gaussian signal is closed to the inversion solution, as expected.

5 Conclusions

This paper presents asymptotic analysis of bias of several variants of well-known algorithm FastICA when separating instantaneous noisy mixtures. The bias is studied subject to the MMSE separator, which is a new insight into noisy ICA. The one-unit FastICA algorithm is shown to estimate the MMSE components with lower bias than the symmetric algorithm and EFICA. Unlike EFICA, the symmetric FastICA numbers among methods using the orthogonal constraint, whose bias is limited as described in section 3.3. In spite of EFICA being more accurate than the symmetric FastICA in the noiseless case, it may not be so in the noisy environment due to the bias, that is asymptotically proportional to σ^2 for both the methods. Computer simulations confirm all theoretical results of our analysis.

References

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

[2] J.-F. Cardoso and A. Souloumiac, “Blind Beamforming from non-Gaussian Signals”, *IEE Proc.-F*, 140(6):362-370, Dec. 1993.

[3] Z. Koldovský, P. Tichavský “Methods of Fair Comparison of Performance of Linear ICA Techniques in Presence of Additive Noise”, *Proc. of ICASSP 2006*, Toulouse, no. V., pp. 873-876, May 2006.

[4] 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.

[5] Z. Koldovský, P. Tichavský and E. Oja, “Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramr-Rao Lower Bound”, *IEEE Trans. on Neural Networks*, Vol. 17, No. 5, Sept 2006.

[6] A. Hyvärinen, ”Gaussian Moments for Noisy Independent Component Analysis”. *IEEE Signal Processing Letters*, vol. 6, pp. 145-147, June 1999.

[7] S.C. Douglas, A. Cichocki, and S. Amari, “A Bias Removal Technique for Blind Source Separation with Noisy Measurements,” *Electronics Letters*, vol. 34, no. 14, pp. 1379-1380, July 1998.

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

[9] M. Davies, "Identifiability Issues in Noisy ICA", *IEEE Signal Processing Letters*, vol. 11, no. 5, pp. 470-473, May 2005.

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

[11] Z. Koldovský and P. Tichavský, "Blind Instantaneous Noisy Mixture Separation with Best Interference-plus-noise Rejection", submitted to *SPAWC2007*, Jan 2007.

Appendix A

Using the substitution $\mathbf{F} = \widehat{\mathbf{W}}(\mathbf{W}^{\mathrm{MMSE}})^{-1}$, (24) can be written as

$$\min_{\mathbf{FGF}^T=\mathbf{I}} \|\mathbf{F} - \mathbf{D}\|_F^2, \tag{32}$$

where $\mathbf{G} = (\mathbf{I} + \sigma^2\mathbf{V})^{-1}$. Then the Lagrangian of the optimization problem is

$$L = \|\mathbf{F} - \mathbf{D}\|_F^2 - (\mathrm{vec}[\mathbf{\Lambda}])^T(\mathrm{vec}[\mathbf{FGF}^T]), \tag{33}$$

where $\mathbf{\Lambda}$ is a matrix of Lagrange multipliers. A simple calculus gives

$$\frac{dL}{d\mathbf{F}} = 2\mathbf{F} - 2\mathbf{D} - (\mathbf{\Lambda} + \mathbf{\Lambda}^T)\mathbf{FG}. \tag{34}$$

Putting (34) equal to zero and using the constraint in (32) we get

$$\mathbf{\Lambda} + \mathbf{\Lambda}^T = 2\mathbf{FF}^T - 2\mathbf{DF}^T.$$

It follows that the Lagrangian (33) is zero if and only if

$$(\mathbf{F} - \mathbf{D})(\mathbf{I} - \mathbf{F}^T\mathbf{FG}) = \mathbf{0} \tag{35}$$

Since $\mathbf{F} = \mathbf{D}$ does not fulfill the constraint $\mathbf{FGF}^T = \mathbf{I}$, it follows that each local minimum of $\|\mathbf{F} - \mathbf{D}\|_F^2$ satisfying the constraint obeys

$$\mathbf{F}^T\mathbf{F} = \mathbf{G}^{-1} = \mathbf{I} + \sigma^2\mathbf{V}. \tag{36}$$

Now, (25) follows.

Appendix B

Using (17) and (18), the sign-corrected vector $\mathbf{w}_k^+$ can be written as

$$\mathbf{w}_k^+ = |\tau_k|\mathbf{e}_k - \frac{1}{2}\sigma^2|\tau_k|\mathbf{v}_k + \frac{1}{2}\sigma^2\mathbf{V}_{kk}(|\tau_k| + \mathrm{sign}(\tau_k)\delta_k)\mathbf{e}_k + \mathcal{O}(\sigma^3) \tag{37}$$

Then, the symmetric orthogonalization should be applied to the matrix $\mathbf{W}^+$ defined in (7). Similar expansion to that in [4] (Appendix A, expression (67)) can be used, which gives

$$\begin{aligned}\mathbf{W}^{\text{SYM+}}_{k\ell} &= \mathbf{I}_{k\ell} + \frac{\mathbf{W}^+_{k\ell} - \mathbf{W}^+_{\ell k}}{\mathbf{W}^+_{kk} + \mathbf{W}^+_{\ell\ell}} + \mathcal{O}(\sigma^3) = &(38)\\ &= \mathbf{I}_{k\ell} + \frac{1}{2}\sigma^2 \mathbf{V}_{k\ell}\frac{|\tau_\ell| - |\tau_k|}{|\tau_k| + |\tau_\ell|} + \mathcal{O}(\sigma^3). &(39)\end{aligned}$$

Finally, the resulting matrix is pre-multiplied by $\mathbf{C}^{-1/2}$ from the right

$$\begin{aligned}\mathbf{W}^{\text{SYM}} &= \mathbf{W}^{\text{SYM+}}\mathbf{C}^{-1/2} = \\ &= \mathbf{I} + \frac{1}{2}\sigma^2 \mathbf{V} \odot (\mathbf{1}_{d\times d} + \mathbf{H}^{\text{SYM}}) + \mathcal{O}(\sigma^3) &(40)\end{aligned}$$

The definition (28) and the result (27) follow.

In EFICA [5], the computation of each row of the resulting de-mixing matrix is done separately via the symmetric orthogonalization of weighted matrix $\mathbf{W}^+$. Specifically, the r-th row equals the r-th row of a matrix that is computed via the symmetric orthogonalization of $\text{diag}[c_{r1}, \ldots, c_{rd}] \cdot \mathbf{W}^+$. The weights $c_{k\ell}$ are defined in section 3.4. Hence, to compute the bias of the r-th row, a substitution $\mathbf{W}^+_{k\ell} \leftarrow c_{rk}\mathbf{W}^+_{k\ell}$ should be used in (38). Then, (29) easily follows.