ASYMPTOTIC PERFORMANCE OF THE FASTICA ALGORITHM FOR INDEPENDENT COMPONENT ANALYSIS AND ITS IPROVEMENTS

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

[1]Institute of Information Theory and Automation,
P.O.Box 18, 182 08 Prague 8, Czech Republic

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

[3]Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 5400, 02015 HUT, Finland

ABSTRACT

The fixed point algorithm, known as FastICA, is one of the most successful algorithms for independent component analysis in terms of accuracy and low computational complexity. This paper derives analytic closed form expressions that characterize separating ability of both one-unit and symmetric version of the algorithm in a local sense. Based on the analysis it is possible to combine advantages of the two versions. Although the analysis assumes a "good" initialization of the algorithms and long data records, our computer simulations demonstrate validity of the theoretical expressions in the case of arbitrary initialization and moderate data lengths.

1. INTRODUCTION

The aim of Independent Component Analysis (ICA) is to transform a set of mixed random signals into components that are as mutually independent as possible. ICA serves as a tool for blind source separation, but can be also used as a method for blind deconvolution or blind equalization. It has applications e.g. in speech and image processing and in signal processing for wireless communications and biomedical signals.

FastICA, an algorithm for ICA first proposed by Hyvärinen and Oja, is based on the optimization of some nonlinear contrast function, which indicates the non-gaussianity of the components. Much of the earlier and recent works have studied algorithms based on the kurtosis contrast function: blind deconvolution [13], monotonic convergence [12], global convergence [10], asymptotic performance analysis of the algorithm for one unit [5], and convergence of the symmetric algorithm [9]. The Cramér-Rao bound for ICA was studied in [2, 11, 18, 14, 17] and recently in [7].

[0]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.

2. DATA MODEL AND THE METHOD

The standard linear ICA model for data matrix $\mathbf{X}$ is

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

where $\mathbf{S}$ is a $d \times N$ source matrix composed of d rows, where each row $\mathbf{s}_k^T$, $k = 1, \ldots, d$ contains N independent realizations of a random variable s_k. $\mathbf{A}$ is a $d \times d$ mixing matrix. The goal of independent component analysis is to estimate the matrix $\mathbf{A}$ or, equivalently, the de-mixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ or, equivalently, the original source signals $\mathbf{S}$. It is well known that

1. the separation is unique only up to an unknown scale and order of the components
2. the separation is possible only if at most one original signal has a Gaussian distribution.

Since the scale of the source signals cannot be retrieved, one can assume, without any loss in generality, that the sample variance of the estimated source signals is equal to one. In addition, arithmetic mean of the data is irrelevant for the mutual information and can be removed. Thus, instead of the original source signals $\mathbf{S}$, a normalized source signal matrix denoted $\mathbf{U}$ can be estimated, where

$$\begin{aligned}
\mathbf{U} &= \mathbf{D}^{-1/2}(\mathbf{S} - \overline{\mathbf{S}}) && (2)\\
\mathbf{D} &= \mathrm{diag}[\sigma_1^2, \ldots, \sigma_d^2] && (3)\\
\sigma_k^2 &= (\mathbf{s}_k - \bar{\mathbf{s}}_k)^T(\mathbf{s}_k - \bar{\mathbf{s}}_k)/N\,. && (4)\\
\bar{\mathbf{s}}_k &= \mathbf{s}_k^T \cdot \mathbf{1}_N\mathbf{1}_N/N, \qquad k = 1, \ldots, d && (5)
\end{aligned}$$

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

2.1. Preprocessing

The first step of most variants of the algorithm consists in removing the sample mean and a decorrelation,

$$\mathbf{Z} \;=\; \widehat{\mathbf{C}}^{-1/2}\,(\mathbf{X}-\overline{\mathbf{X}}) \qquad (6)$$

where $\widehat{\mathbf{C}} = (\mathbf{X}-\overline{\mathbf{X}})(\mathbf{X}-\overline{\mathbf{X}})^T/N$ is the sample covariance matrix, and $\overline{\mathbf{X}}$ is the sample mean. The output $\mathbf{Z}$ contains whitened data in the sense that $\mathbf{Z}\mathbf{Z}^T/N = \mathbf{I}$ (identity matrix). Note that $\mathbf{Z}$ can be re-written using (1), (2) as

$$\mathbf{Z} \;=\; \widehat{\mathbf{C}}^{-1/2}\mathbf{A}\mathbf{D}^{1/2}\,\mathbf{U}\,. \qquad (7)$$

The ICA problem can be formulated as finding a de-mixing matrix $\widehat{W}_z$ that separates the signals $\mathbf{U}$ from the whitened mixture $\mathbf{Z}$ as $\widehat{\mathbf{U}} = \widehat{W}_z \cdot \mathbf{Z}$.

2.2. The FastICA Algorithm

The FastICA algorithm for one unit estimates one row of the de-mixing matrix $\mathbf{W}_z$ as a vector $\widehat{\mathbf{w}}_z^T$ that is a stationary point (minimum or maximum) of the expression $\mathrm{E}[G(\mathbf{w}^T\mathbf{Z})] \stackrel{\mathrm{def}}{=} G(\mathbf{w}^T\mathbf{Z})\mathbf{1}_N/N$ subject to $\|\mathbf{w}\| = 1$, where $G(\cdot)$ is a suitable nonlinear and non-quadratic function [6].

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

$$\mathbf{w}^+ \;\leftarrow\; \mathbf{Z}g(\mathbf{Z}^T\mathbf{w}) - \mathbf{w}\,g'(\mathbf{w}^T\mathbf{Z})\mathbf{1}_N \qquad (8)$$

$$\mathbf{w} \;\leftarrow\; \mathbf{w}^+/\|\mathbf{w}^+\| \qquad (9)$$

until convergence is achieved. In (8) and also elsewhere in the paper, $g(\cdot)$ and $g'(\cdot)$ denote the first and second derivatives of the function $G(\cdot)$, and they are applied elementwise. Classical widely used functions $g(\cdot)$ include "pow3" (then the algorithm performs kurtosis minimization), "tanh" and "gauss", $g(x) = x\exp(-x^2/2)$.

It is not known in advance which row of $\mathbf{W}_z$ is being estimated: it largely depends on the initialization. If this is the $k-$th row, the resultant $\mathbf{w}^T$ is denoted as $\widehat{W}_{k\cdot}^{1U}$, that is, the $k-$th row of a matrix $\widehat{W}_z^{1U}$.[1] Note that the estimated components are not constrained to be mutually orthogonal in this variant of the algorithm,.

The symmetric FastICA proceeds by applying the recursion (8) for all components in parallel, with the difference that the normalization in (9) is replaced by a symmetrization step: Starting with a random unitary matrix $\mathbf{W}$ iterate

$$\mathbf{W}^+ \;\leftarrow\; g(\mathbf{W}\mathbf{Z})\mathbf{Z}^T - \mathrm{diag}[g'(\mathbf{W}\mathbf{Z})\mathbf{1}_N]\,\mathbf{W} \qquad (10)$$

$$\mathbf{W} \;\leftarrow\; (\mathbf{W}^+\mathbf{W}^{+T})^{-1/2}\mathbf{W}^+ \qquad (11)$$

until convergence is achieved. The result is denoted $\widehat{\mathbf{W}}_z^{SYM}$.

[1] Convergence of the recursion for some components might be problematic. This problem is successfully solved in the algorithm variant Smart FastICA, introduced later in Section 4.

2.3. Measure of the separation quality

The separation ability of ICA algorithms can be characterized by the relative presence of the $k-$th source signal in the estimated $i-$th source signal. It is possible, if the source signals are known. Due to the permutation and sign/phase uncertainty, the estimated sources need to be appropriately sorted to fit the original ones. In this paper, the method proposed in [15] is used. Formally, the estimated source signals can be written using (7) as

$$\widehat{\mathbf{U}} = \widehat{W}_z \cdot \mathbf{Z} = \widehat{W}_z\widehat{\mathbf{C}}^{-1/2}\,\mathbf{A}\mathbf{D}^{1/2}\mathbf{U} = \widehat{\mathbf{G}}\,\mathbf{U} \qquad (12)$$

where $\widehat{\mathbf{G}} = \widehat{W}_z\widehat{\mathbf{C}}^{-1/2}\mathbf{A}\mathbf{D}^{1/2}$ and $\widehat{W}_z$ stands either for $\widehat{\mathbf{W}}_z^{1U}$ or for $\widehat{\mathbf{W}}_z^{SYM}$. It will be called *gain* matrix for easy reference. The de-mixing matrix that separates the independent components from the original data is $\widehat{W}_x = \widehat{W}_z\widehat{\mathbf{C}}^{-1/2}$. With the above notations, $\widehat{\mathbf{G}} = \widehat{W}_x\mathbf{A}\mathbf{D}^{1/2}$.

The relative presence of the $k-$th source signal in the estimated $i-$th source signal is represented by the $(k\ell)-$th element of $\widehat{\mathbf{G}}$, denoted $G_{k\ell}$. Then, the total signal-to-interference of the $k-$th source signal is defined as

$$\mathrm{SIR}_k = \frac{\mathrm{E}[G_{kk}^2]}{\mathrm{E}\left[\sum_{\substack{\ell=1\\ \ell\neq k}}^{d} G_{k\ell}^2\right]} \approx \frac{1}{\sum_{\substack{\ell=1\\ \ell\neq k}}^{d}\mathrm{E}[G_{k\ell}^2]} \qquad (13)$$

3. TEST OF SADDLE POINTS

In general, the global convergence of the symmetric FastICA is known to be quite good. Nevertheless, if it is run 10 000 times from random initial demixing matrices, on the average in 1 - 100 cases the algorithm gets stuck in local minima that can be recognized by exceptionally low achieved SIR. A detailed investigation of the false solutions showed that they contain one or more pairs of estimated components, say $(\widehat{\mathbf{u}}_k, \widehat{\mathbf{u}}_\ell)$ such that they are close to $(\mathbf{u}_k + \mathbf{u}_\ell)/\sqrt{2}$ and $(\mathbf{u}_k - \mathbf{u}_\ell)/\sqrt{2}$, respectively, where $(\mathbf{u}_k, \mathbf{u}_\ell)$ is the desired solution. It is because of symmetry, the saddle points of the criterion function mostly lie approximately on half the way between two correct solutions that differ by the order of two of the componets. Thus, an appropriate estimate of $(\mathbf{u}_k, \mathbf{u}_\ell)$ would be $(\widehat{\mathbf{u}}'_k, \widehat{\mathbf{u}}'_\ell)$ where

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

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

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

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

Thus, we suggest to complete the plain symmetric FastICA by the check of all $\binom{d}{2}$ pairs of the estimated independent components for a possible improvement via the saddle points. If the test for saddle point is positive, it is suggested to perform a few more (1-4) additional iterations of the original algorithm, starting from the improved estimate. The improved FastICA imitates a global search for the minimum of the chosen criterion function. It is used in the simulation section.

4. ANALYSIS

To analyze the algorithm it is useful to note that the estimator $\widehat{\mathbf{U}}$ is invariant with respect to orthogonal transformations of the decorrelated data $\mathbf{Z}$, or equivariant [1]. Thus it is possible to show that the output of the algorithm (the gain matrix $\widehat{\mathbf{G}}$) is independent of the mixing matrix $\mathbf{A}$, and is the same as if $\mathbf{Z}$ were equal to $\mathbf{Z} = \mathbf{R}^{-1/2}\mathbf{U}$, where

$$\mathbf{R} = \mathbf{U}\mathbf{U}^T/N \tag{14}$$

This observation simplifies the analysis greatly.

The following Proposition summarizes the main result.

Proposition 1

Assume that (1) all original independent components have zero mean and unit variance, (2) the function g in algorithm FastICA is twice continuously differentiable, (3) the following expectations exist

$$\mathrm{E}[s_k g(s_k)] \stackrel{\mathrm{def}}{=} \mu_k \tag{15}$$

$$\mathrm{E}[g'(s_k)] \stackrel{\mathrm{def}}{=} \rho_k \tag{16}$$

$$\mathrm{E}[g^2(s_k)] \stackrel{\mathrm{def}}{=} \beta_k \tag{17}$$

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

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

$$V_{k\ell}^{1U} = \frac{\beta_k - \mu_k^2}{(\mu_k - \rho_k)^2} \tag{18}$$

$$V_{k\ell}^{SYM} = \frac{\beta_k - \mu_k^2 + \beta_\ell - \mu_\ell^2 + (\mu_\ell - \rho_\ell)^2}{(|\mu_k - \rho_k| + |\mu_\ell - \rho_\ell|)^2} \tag{19}$$

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

Proof: See Appendix. (18) can be found in [1, 5], but (19) is novel.

The assumption 4 may look peculiar at the first glance, but it is not so restrictive as it seems to be. Once the algorithm is started from an initial $\mathbf{W}$ that lies in a right domain of attraction, the resultant stationary point of the recursion, denoted $\widehat{\mathbf{W}}$, is the same, and is approximately equal to $\mathbf{W}^+$ obtained after one step from the ideal solution, thanks to the fact that the convergence is quadratic.

Note some interesting cases:

(1) $\mu_k = \rho_k$ implies infinite variance of $G_{k\ell}^{1U}$. It occurs when the probability distribution is Gaussian, regardless of the choice of $g(\cdot)$.

(2) $\beta_k = \mu_k^2$, meaning that the variance decays faster with growing N than the usual $O(1/N)$. This condition is fulfilled for signals known under the acronym BPSK in wireless communications, which have $s_k = +1$ or $s_k = -1$ both with probability 0.5 .

(3) If $g(x) = -f'(x)/f(x)$, i.e. if g is chosen as the score function of the pdf of the components, the variances in (18)-(19) are minimized. The minimum variances are compared with the corresponding CRB in [7].

5. EXAMPLE OF UTILIZATION

In this section, the previous analysis is used to derive a novel variant of the FastICA algorithm is proposed, which combines advantages of both previously discussed variants. For easy reference it will be called "Smart FastICA". This algorithm begins with applying symmetric FastICA with nonlinearity "tanh". For each estimated component signal $\widehat{\mathbf{u}}_k^T$, parameters μ_k, ρ_k and β_k are computed according to (15)-(17), namely $\widehat{\mu}_k = \widehat{\mathbf{u}}_k^T g(\widehat{\mathbf{u}}_k)/N$, $\widehat{\rho}_k = \widehat{\mathbf{1}}_N^T g'(\widehat{\mathbf{u}}_k)/N$, $\widehat{\beta}_k = \widehat{\mathbf{1}}_N^T g^2(\widehat{\mathbf{u}}_k)/N$. and then they are plugged in (18)-(19) and (13), namely

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

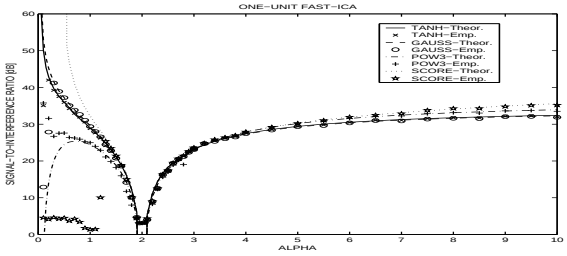
If the former SIR is better than the latter one, the one unit algorithm is performed for the component, taking advantage of a more suitable nonlinearity g for each of particular cases: In the supergaussian case, defined by the condition $\widehat{\mu_k} < \widehat{\rho}_k$, the option "gauss" is selected, and in the subgaussian case with $\widehat{\mu}_k > \widehat{\rho}_k$, "pow3" is applied (see the simulation section for a reason). Then, $\widehat{\mu}_k, \widehat{\rho}_k, \widehat{\beta}_k$ and $\widehat{\mathrm{SIR}}_k$ are computed again. If the new $\widehat{\mathrm{SIR}}_k$ is better than the previous one and if, at the same time, the scalar product between the former separating vector and the new one is higher in absolute value than a constant (we have used 0.95), then the one unit refinement is accepted in favour of the former vector. The condition on the scalar product is intended to eliminate the cases where the one unit algorithm converged to a wrong component. Further improvements of the algorithm are subject of an accompanying paper [8].

6. SIMULATIONS

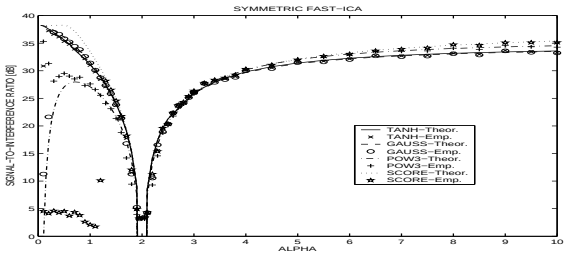
In this section, the validity of the analysis is supported by computer simulations.

Example 1. Four independent random signals with generalized Gaussian distribution [7] with parameter α and length $N = 5000$ were generated in 100 independent trials. The signals were mixed with a matrix that was randomly generated in each trial, and de-mixed again by eight variants of the algorithm: the symmetric FastICA with nonlinearities tanh, gauss, pow3, and with the score function, that is $g(x) = \text{sign}(x) \cdot |x|^{\alpha-1}$, see [7], as well as the one-unit FastICA with the same nonlinearities, implemented like smart FastICA. The resulting theoretical and empirical SIR is plotted in Figure 1 (a) and (b). An erratic behavior of the empirical results is experienced for small α and nonlinearity pow3. Here, the convergence of sample estimates of the expressions in (15)-(17) to their expectations is slow.

We can see that among the α-independent nonlinearities, the "pow3" performs best in the case of $\alpha > 2$ that corresponds to the sub-Gaussian case, and "gauss" is the best one for $\alpha < 2$ where the distribution is super-Gaussian. FastICA with $g(\cdot)$ equal to the score function does not work properly (does not converge at all) for $\alpha \leq 1$, because the score function is not continuous for these α's.



(a)



(b)

Fig. 1 Performance of One-Unit FastICA (diagram (a)) and of Symmetric FastICA (diagram (b)) in separating signals with distribution GG(α) as a function of α.

Example 2. Here, we have generated signals with four independent components with the distributions: (1) generalized Gaussian with $\alpha = 0.5$, (2) "Sinus", that is the distribution of $\sqrt{2}\sin(u)$, where u is uniformly distributed in $(0, 2\pi)$, (3) Laplace, and (4) Gaussian, of various lengths. Again, the signals were randomly mixed and separated by two methods: symmetric FastICA with nonlinearity tanh, and Smart FastICA. Resultant empirical and theoretical SIR's are plotted in Figure 2 as functions of the data length N. The agreement between the theory and simulation is very good. The smart FastICA is shown to outperform symmetric FastICA in separating the components with "more" non-Gaussian distributions (in the example it is "Sinus" and GG(0.5)), while estimates of the other components remain unchanged.

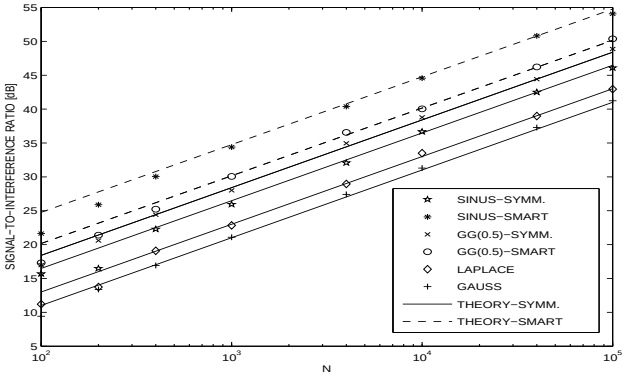


Fig. 2: Performance of symmetric FastICA and smart FastICA versus varying data length.

7. APPENDIX - PROOF OF PROPOSITION 1

The proof utilizes the following lemma. **Lemma 1** Let

$$\mathbf{W} \quad = \quad \mathbf{W}_0 + \Delta\mathbf{W} \tag{20}$$

where $\mathbf{W}_0 = \text{diag}(w_1, \ldots, w_d)$ is a diagonal matrix, with $w_k > 0$ for $k = 1, \ldots, d$. Then, the $(k\ell)-$th element of $(\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W}$ is equal to, for $k, \ell = 1, \ldots, d$, $k \neq \ell$,

$$[(\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W}]_{k\ell} \; = \; \frac{\Delta W_{k\ell} - \Delta W_{\ell k}}{w_k + w_\ell} + O(\|\Delta\mathbf{W}\|^2) \; .$$

Proof of the lemma is skipped for lack of space.

Proof of Proposition 1 proceeds in several steps.

A. Notes on Stochastic Convergence

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

$$N^{-1}\mathbf{s}_k^T g(\mathbf{s}_k) \; = \; \mu_k + o_p(1) \tag{21}$$

$$N^{-1}g'^T(\mathbf{s}_k)\mathbf{1}_N \; = \; \rho_k + o_p(1) \; . \tag{22}$$

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

$$N^{-1}g'^T(\mathbf{s}_k)(\mathbf{s}_\ell \odot \mathbf{s}_\ell) = \rho_k + o_p(1) \tag{23}$$

where $\odot$ denotes the elementwise product. It can be shown, that the same limits are obtained if $\mathbf{s}_k$, $\mathbf{s}_\ell$ in (21)-(23) are replaced by the normalized components $\mathbf{u}_k$, $\mathbf{u}_\ell$, where $\mathbf{u}_k$ is the $k-$th column of $\mathbf{U}$, $k = 1, \ldots, d$. Note from (2) that $\mathbf{u}_k = (\mathbf{s}_k - \bar{\mathbf{s}}_k)/\widehat{\sigma}_k$, $\bar{\mathbf{s}}_k = O_p(N^{-1/2})$, $\widehat{\sigma}_k = 1 + o_p(1)$, consequently $\mathbf{u}_k = \mathbf{s}_k + o_p(1)$, $g(\mathbf{u}_k) = g(\mathbf{s}_k) + o_p(1)$, and

$$\begin{aligned} \mathbf{u}_k^T g(\mathbf{u}_k) &= [\mathbf{s}_k + o_p(1)]^T [g(\mathbf{s}_k) + o_p(1)] \\ &= \mathbf{s}_k^T g(\mathbf{s}_k) + o_p(N) = N\mu_k + o_p(N) \,. \end{aligned} \quad (24)$$

Similarly, it can be shown that

$$g'^T(\mathbf{u}_k)\mathbf{1}_N = N\rho_k + o_p(N) \quad (25)$$

$$g'^T(\mathbf{u}_k)(\mathbf{u}_\ell \odot \mathbf{u}_\ell) = N\rho_k + o_p(N) \quad (26)$$

Moreover, using the asymptotic expression for $\mathbf{R}$, to be derived in the next subsection, it can be shown that the relations (21) and (22) hold true as well, if $\mathbf{s}_k$ is replaced with $\mathbf{z}_k$, that is defined as the $k-$th column of $\mathbf{Z}$, $k = 1, \ldots, d$,

$$\mathbf{z}_k^T g(\mathbf{z}_k) = N\mu_k + o_p(N) \quad (27)$$

$$g'^T(\mathbf{z}_k)\mathbf{1}_N = N\rho_k + o_p(N) \,. \quad (28)$$

B. Asymptotic behaviour of R

As N goes to infinity, the matrix $\mathbf{R}$ defined in (14) approaches identity matrix in the mean square sense. To see this, note that the diagonal elements R_{kk} of $\mathbf{R}$ are equal to one by definition, and that the off-diagonal elements $R_{k\ell}$ with $k \neq \ell$ have zero mean. Variance of these $R_{k\ell} = \mathbf{u}_k^T \mathbf{u}_\ell / N$ can be shown to be equal to $1/(N-1)$ [16]. Hence

$$\Delta\mathbf{R} \stackrel{\text{def}}{=} \mathbf{R} - \mathbf{I} = O_p(N^{-1/2}) \,. \quad (29)$$

Using a Taylor series expansion it can be derived that

$$\mathbf{R}^{-1/2} = \mathbf{I} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1}) \,. \quad (30)$$

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

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

$$\begin{aligned} \mathbf{Z} = \mathbf{R}^{-1/2}\mathbf{U} &= \left(\mathbf{I} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1})\right)\mathbf{U} \\ &= \mathbf{U} - \frac{1}{2}\Delta\mathbf{R}\mathbf{U} + O_p(N^{-1}) \,. \end{aligned} \quad (31)$$

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

$$g(\mathbf{Z}) = g(\mathbf{U}) + g'(\mathbf{U}) \odot \Delta\mathbf{Z} + O_p(N^{-1}) \quad (32)$$

where $\odot$ denotes the elementwise product and

$$\Delta\mathbf{Z} \stackrel{\text{def}}{=} \mathbf{Z} - \mathbf{U} = -\frac{1}{2}\,\Delta\mathbf{R}\,\mathbf{U} + O_p(N^{-1}) \,. \quad (33)$$

Using (14), the $k-$th column of $\Delta\mathbf{Z}^T$ is

$$\Delta\mathbf{z}_k = -\frac{1}{2N}\sum_{\substack{m=1 \\ m\neq k}}^{d} \mathbf{u}_k^T\mathbf{u}_m\mathbf{u}_m + O_p(N^{-1}) \,. \quad (34)$$

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

Inserting $\mathbf{W} = \mathbf{I}$ in (10), the $k\ell-$th element of $\mathbf{W}^+$ reads

$$W_{k\ell}^+ = \begin{cases} g(\mathbf{z}_k^T)\mathbf{z}_\ell & \text{for} \quad k \neq \ell \\ g(\mathbf{z}_k^T)\mathbf{z}_k - g'(\mathbf{z}_k^T)\mathbf{1}_N & \text{for} \quad k = \ell \end{cases} \quad (35)$$

For $k = \ell$ we get using (27)-(28)

$$W_{kk}^+ = N(\mu_k - \rho_k) + o_p(N) \,. \quad (36)$$

For $k \neq \ell$ we get using (32), (34)

$$\begin{aligned} W_{k\ell}^+ &= g(\mathbf{z}_k^T)\,\mathbf{z}_\ell \\ &= [g(\mathbf{u}_k^T) + g'(\mathbf{u}_k^T) \odot \Delta\mathbf{z}_k^T + O_p(N^{-1})][\mathbf{u}_\ell + \Delta\mathbf{z}_\ell] \\ &= \left[g(\mathbf{u}_k^T) - \frac{1}{2N}\sum_{\substack{m=1 \\ m\neq k}}^{d} g'(\mathbf{u}_k^T) \odot (\mathbf{u}_k^T\mathbf{u}_m\mathbf{u}_m^T)\right] \\ &\quad \cdot \left[\mathbf{u}_\ell - \frac{1}{2N}\sum_{\substack{m=1 \\ m\neq \ell}}^{d} \mathbf{u}_\ell^T\mathbf{u}_m\mathbf{u}_m\right] + O_p(1) \,. \end{aligned} \quad (37)$$

After some simplifications, using (24)-(26), it can be shown that

$$W_{k\ell}^+ = g(\mathbf{u}_k^T)\,\mathbf{u}_\ell - \frac{\mu_k + \rho_k}{2}\,\mathbf{u}_k^T\mathbf{u}_\ell + o_p(N^{1/2}) \,. \quad (38)$$

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

Note that if $\widehat{W}_{kk}^+ < 0$ for some k, the $k-$th diagonal element of the de-mixing matrices $\widehat{W}_{kk}^{1U}$ and $\widehat{W}_{kk}^{SYM}$ may have wrong sign, i.e. it might be close to -1 instead of 1. It corresponds to reversed sign of the $k-$th estimated independent component. In the one-unit version of the algorithm, the sign can be corrected by replacing the normalization in (9) by an equivalent formula

$$\widehat{W}_{k\ell}^{1U} = \frac{W_{k\ell}^+}{W_{kk}^+} = \frac{W_{k\ell}^+}{N(\mu_k - \rho_k)} + o_p(N^{-1/2}) \,. \quad (39)$$

Similarly, using Lemma 1, the asymptotically equivalent sign corrected expression for the estimated de-mixing matrix element $\widehat{W}_{k\ell}^{SYM}$ with $k \neq \ell$ is

$$\widehat{W}_{k\ell}^{SYM} = \frac{W_{k\ell}^+ \text{sign}(W_{kk}^+) - W_{\ell k}^+ \text{sign}(W_{\ell\ell}^+)}{|W_{kk}^+| + |W_{\ell\ell}^+|} + o_p(N^{-\frac{1}{2}}) \quad (40)$$

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

$$\Delta\mathbf{W} = \widehat{\mathbf{W}} - \mathbf{I} = O_p(N^{-1/2}) \,. \tag{41}$$

Since

$$\begin{aligned}\mathbf{G} &= \widehat{\mathbf{W}}\mathbf{R}^{-1/2} = (\mathbf{I}+\Delta\mathbf{W})\left(\mathbf{I} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1})\right) \\ &= \mathbf{I} + \Delta\mathbf{W} - \frac{1}{2}\Delta\mathbf{R} + O_p(N^{-1}) \end{aligned} \tag{42}$$

the gain matrix off-diagonal elements read

$$G_{k\ell} = \widehat{W}_{k\ell} - \frac{1}{2N}\,\mathbf{u}_k^T\mathbf{u}_\ell + O_p(N^{-1}) \,. \tag{43}$$

For the one unit variant we get using (39)

$$\begin{aligned} N^{1/2}G_{k\ell}^{1U} &= N^{1/2}\,\frac{W_{k\ell}^+}{N(\mu_k-\rho_k)} - \frac{1}{2N^{1/2}}\,\mathbf{u}_k^T\mathbf{u}_\ell + o_p(1) \\ &= \frac{N^{-1/2}}{\mu_k-\rho_k}\,(g(\mathbf{u}_k^T)\mathbf{u}_\ell - \mu_k\,\mathbf{u}_k^T\mathbf{u}_\ell) + o_p(1) \end{aligned} \tag{44}$$

Finally we will show that (44) can be re-written in terms of $\mathbf{s}_k$, $\mathbf{s}_\ell$ in an asymptotically equivalent formula

$$N^{1/2}G_{k\ell}^{1U} = \frac{N^{-1/2}}{\mu_k-\rho_k}\,(g(\mathbf{s}_k^T)\mathbf{s}_\ell - \mu_k\,\mathbf{s}_k^T\mathbf{s}_\ell) + o_p(1) \,. \tag{45}$$

To prove (45), note that

$$\begin{aligned}\mathbf{u}_k^T\mathbf{u}_\ell &= \left(\frac{\mathbf{s}_k-\overline{\mathbf{s}}_k}{\widehat{\sigma}_k}\right)^T\frac{\mathbf{s}_\ell-\overline{\mathbf{s}}_\ell}{\widehat{\sigma}_\ell} = \frac{\mathbf{s}_k^T\mathbf{s}_\ell - \overline{\mathbf{s}}_k^T\overline{\mathbf{s}}_\ell}{\widehat{\sigma}_k\widehat{\sigma}_\ell} \\ &= \frac{\mathbf{s}_k^T\mathbf{s}_\ell - O_p(1)}{1+o_p(1)} = \mathbf{s}_k^T\mathbf{s}_\ell + o_p(N^{1/2}) \,. \end{aligned} \tag{46}$$

Similarly it can be shown that

$$g(\mathbf{u}_k^T)\mathbf{u}_\ell = g(\mathbf{s}_k^T)\mathbf{s}_\ell + o_p(N^{1/2}) \,. \tag{47}$$

Combining (44), (46) and (47) gives (45). Next, applying the central limit theorem to (45) implies that the distribution of $N^{1/2}G_{k\ell}^{1U}$ is asymptotic normal with zero mean and variance equal to the variance of the leading term in (45). A simple computation gives $\mathrm{E}[(g(\mathbf{s}_k^T)\mathbf{s}_\ell)^2] = N\beta_k$, $\mathrm{E}[(\mathbf{s}_k^T\mathbf{s}_\ell)^2] = N$, $\mathrm{E}[(g(\mathbf{s}_k^T)\mathbf{s}_\ell)(\mathbf{s}_k^T\mathbf{s}_\ell)] = N\mu_k$, and hence

$$\begin{aligned} V_{k\ell}^{1U} &= \mathrm{var}\left[\frac{N^{-1/2}}{\mu_k-\rho_k}\,(g(\mathbf{s}_k^T)\mathbf{s}_\ell - \mu_k\,\mathbf{s}_k^T\mathbf{s}_\ell)\right] \\ &= \frac{N^{-1}}{(\mu_k-\rho_k)^2}\,\mathrm{var}\left[(g(\mathbf{s}_k^T)\mathbf{s}_\ell - \mu_k\,\mathbf{s}_k^T\mathbf{s}_\ell)\right] \\ &= \frac{\beta_k-\mu_k^2}{(\mu_k-\rho_k)^2} \end{aligned} \tag{48}$$

Similarly for symmetric FastICA it holds using (40) that

$$\begin{aligned} N^{1/2}G_{k\ell}^{SYM} &= N^{1/2}\widehat{W}_{k\ell}^{SYM} - \frac{1}{2N^{1/2}}\,\mathbf{u}_k^T\mathbf{u}_\ell + o_p(1) \\ &= \frac{W_{k\ell}^+\mathrm{sign}(\mu_k-\rho_k) - W_{\ell k}^+\mathrm{sign}(\mu_\ell-\rho_\ell)}{N^{1/2}(|\mu_k-\rho_k| + |\mu_\ell-\rho_\ell|)} \\ &\quad -\frac{1}{2N^{1/2}}\,\mathbf{u}_k^T\mathbf{u}_\ell + o_p(1) \end{aligned} \tag{49}$$

The variance of the leading term in (49) results, using (38), after some algebra in (19).

8. REFERENCES

[1] J.-F. Cardoso and B. H. Laheld, "Equivariant Adaptive Source Separation", *IEEE Trans. on Signal Processing*, vol. 44, pp. 3017-3030, 1996.

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

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

[4] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach", *Signal Processing*, Vol. 45, pp. 59-83, 1995.

[5] A. Hyvärinen, "One-Unit Contrast Functions for Independent Component Analysis: A Statistical Analysis". *Proc. IEEE NNSP Workshop '97, Amelia Island, Florida*, pp. 388–397, 1997.

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

[7] 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.

[8] Z. Koldovský, P. Tichavský, "Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound", *Proc. SSP'2005*, Bordeaux.

[9] E. Oja, "Convergence of the Symmetrical FastICA Algorithm", *Proc. 9th Int. Conf. on Neural Information Processing (ICONIP '02)*, Nov. 18-22, 2002.

[10] C. B. Papadias, "Globally Convergent Blind Source Separation Based on a Multiuser Kurtosis Maximization Criterion", *IEEE Tr. Signal Processing*, vol. 48, pp. 3508-3519, 2000.

[11] 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 .

[12] P. A. Regalia and E. Kofidis, "Monotonic Convergence of Fixed-Point Algorithms for ICA", *IEEE Trans. on Neural Networks*, vol. 14, pp. 943-949, 2003.

[13] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)", *IEEE Trans. Inform. Theory*, vol. 36, pp. 312-321, 1990.

[14] 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.

[15] 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, Feb. 2004.

[16] 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.

[17] V. Vigneron, Ch. Jutten, "Fisher information in source separation problems", *Proc. ICA 2004*, Granada, pp. 168-176.

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