Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2007) 392–399
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Finite sample effects of the fast ICA algorithm

Sergio Bermejo

Departament d'Enginyeria Electrònica, Universitat Politècnica de Catalunya, C/Jordi Girona 1-3, 08034 Barcelona, Spain

Received 6 October 2005; received in revised form 24 August 2006; accepted 1 September 2006
Communicated by T. Heskes
Available online 6 March 2007

Abstract

Many algorithms for independent component analysis (ICA) and blind source separation (BSS) can be considered particular instances of a criterion based on the sum of two terms: $C(Y)$, which expresses the decorrelation of the components and $G(Y)$, which measures their non-Gaussianity. Within this framework, the popular FastICA algorithm can be regarded as a technique that keeps $C(Y) = 0$ by first enforcing the whiteness of Y. Because of this constraint, the standard version of FastICA employs the sample-fourth moment as $G(Y)$, instead of the sample-fourth cumulant. Our work analyzes some of the estimation errors introduced by the use of finite date sets in such a higher-order statistics (HOS) contrast and compares FastICA with an alternative version based on the sample-fourth cumulant, which is shown for different probability distributions having a lower variance in the generalization error in the case in which no whitening is performed, e.g. when orthonormal mixing of sources is present.

Keywords: FastICA; Independent component analysis; Blind source separation; Sequential signal extraction; High-order statistics contrast functions

1. Introduction

A recent key discovery in independent component analysis (ICA) [8] computational technique which finds hidden factors in data looking for components that are statistically independent, that the mutual information criterion and related approaches are particular instances of the following weighted criterion:

$$\phi_\lambda(Y) = \lambda C(Y) - G(Y), \tag{1}$$

where $C(Y)$ is a decorrelation term and $G(Y)$ is a measure of non-Gaussianity. Typically, $G(Y)$ can be obtained as a sum of marginal non-Gaussianities based on higher-order statistics (HOS) contrast (or objective) functions, i.e. $G(Y) = \sum_i G(Y_i)$, which leads to successful sequential-independent component (IC) extraction or sequential signal separation from a linear mixture [10]. In this context, and for large values of λ, FastICA [15], one of the most widely employed ICA procedures (see e.g. [2,4,13,21,27] for recent applications of FastICA) mainly due—as its authors claim (p. 179 [17])—to its reliability and very fast convergence, can be understood as a method that enforces $C(Y) = 0$ by means of a whitening process and then extracts ICs with the sample-fourth moment contrast as $G(Y)$. Such orthogonal procedure which is based on a pre-whitening stage has an inherent lower bound in the asymptotic—i.e. when the number of samples (N) tends to infinite—separating performance, due to the errors introduced in the computation of the whitening matrix that cannot be compensated later, as [6,7] pointed out. These phenomena lead to the use of non-orthogonal methods in which $C(Y)$ and $G(Y)$ are simultaneously optimized in order to solving the problem of distorting the optimization criterion introduced by whitening (e.g. [28]). In the case of FastICA, such a distortion is that the learning algorithm is enforced to optimize the sample-fourth moment contrast instead of the sample-fourth cumulant.

Our work studies the consequence of such a distortion when a finite set of observed signals $\mathbf{D}_N = \{x_i, i = 1,\ldots,N\}$ is available in terms of some straightforward distribution-dependent bounds on the generation error in the context of the statistical learning theory (SLT) [24–26], extending some previous results [1] and applying them to FastICA algorithm. Since the formulation of this algorithm given in

E-mail address: sbermejo@eel.upc.edu.

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.09.015

[15,16] and some other works (e.g. [12]) only studied its properties in terms of convergence and the separation of source signals assuming an infinite number of samples, much remains unknown regarding its real convergence and accuracy, topics which are here addressed. In particular, it is shown that the generalization error of an alternative fixed-point algorithm based only on optimizing the sample-fourth cumulant when orthonormal mixing of sources is present has a lower variance for some distributions, when finite data sets are dealt with, in comparison to the standard version of FastICA that is based on whitening and the sample-fourth moment contrast. These distribution-dependent theoretical results are confirmed in an empirical study, which consists of the orthonormal separation of uniform and Laplacian sources with both versions of FastICA, and demonstrate a better estimation of the separating matrix for the algorithm based on the fourth cumulant. A direct consequence of this work is that a fixed-point optimization of contrast functions of the form shown in (1), in which a term $C(Y)$ was employed effectively to ensure decorrelation plus a term $G(Y)$ based on the fourth cumulant, may incur smaller generalization errors than the two-step approach of standard FastICA, in which the whiteness of Y is first imposed and then the constrast function that is to be the sample-fourth moment is restricted later.

2. FastICA REVISITED

2.1. The FastICA algorithm

In a linear BSS model [24], we observe m signals $x_1,\ldots,$ x_m that correspond to a linear mixture of a p source signal $s_1,\ldots,$ s_p, i.e.

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2}$$

where $\mathbf{A}$ is known as the mxp "mixing matrix". According to (2) and given the observable vector $\mathbf{x}$, a linear projection is performed as

$$y = \mathbf{w}^T\mathbf{x}, \tag{3}$$

where, clearly, $\mathbf{w}$ must tend towards one of the column vectors of $\mathbf{A}^{-1}$ in order to obtain one of the p source signals s_i with $i\in\{1,\ldots, p\}$. If there is a minimum of $(p-1)$ non-Gaussian sources and A is forced to be orthonormal by a previously performed whitening process, the so-called fixed-point (or fast) ICA algorithm recovers one of the original signals. The algorithm has the following steps [10,17]:

1. Initialize randomly $\mathbf{w}[0]$ of norm 1. Let $k = 1$.
2. Update $\mathbf{w}[k] = \hat{E}\{\mathbf{x}(\mathbf{w}[k-1]^T\mathbf{x})^3\} - 3\mathbf{w}[k-1]$.
3. Divide $\mathbf{w}[k]$ by its norm.
4. If $|\mathbf{w}[k]^T\mathbf{w}[k-1]|$ is not close to 1, $k = k+1$ and return to step 2. Otherwise, finish with $\mathbf{w}^* = \mathbf{w}[k]$.

As the authors pointed out in [10,17] , the estimator of the expectation $\hat{E}$ must employ a large sample set $\mathbf{D}_N = \{\mathbf{x}_r,\ i = 1,\ldots,N\}$ (they postulate 1000 samples). Hence, a batch version of the updated equation typically has the following form:

$$\begin{aligned}\mathbf{w}[k] &= \hat{E}\{\mathbf{x}(\mathbf{w}[k-1]^T\mathbf{x})^3\} - 3\mathbf{w}[k-1]\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(\mathbf{w}[k-1]^T\mathbf{x}_i)^3 - 3\mathbf{w}[k-1].\end{aligned} \tag{4}$$

2.2. The FastICA objective function

Hyvärinen and Oja [16] proposes an optimal objective function suitable for separating one source signal according to the above constraints in $\mathbf{A}$. The source distributions have the following general form:

$$\hat{J}(\mathbf{w}) = \alpha\hat{E}\{(\mathbf{w}^T\mathbf{x})^4\} + \beta F[\hat{E}\{(\mathbf{w}^T\mathbf{x})^2\}], \tag{5}$$

where α, $\beta>0$ are arbitrary scales and F is a suitable penalty function. In particular, (4) is reduced to the fourth cumulant if $\alpha = \beta = 1$. Then, $F[u] = -3u$, whose minimization results in the signal separation of at least one of the available $(p-1)$ platykurtic source signals s_i, as shown in [11,15,16]. If we keep in mind that the normalized update equation of the fixed-point algorithm is given by (p.188; [10]), then

$$\mathbf{w}[k] = \frac{\nabla_{\mathbf{w}}\hat{J}(\mathbf{w}[k-1])}{||\nabla_{\mathbf{w}}\hat{J}(\mathbf{w}[k-1])||}, \tag{6}$$

where $\hat{J}$ is the sample objective function created using $\mathbf{D}_N$. Since in practice a whitening process is performed that enforces $\hat{E}\{(\mathbf{w}^T\mathbf{x})^2\} = 1$, the sample objective function, then what actually minimizes FastICA with the update equation given in (4) is

$$\begin{aligned}\hat{J}_{\mathrm{FastICA}}(\mathbf{w},\mathbf{D}_N) &= \hat{J}_{\mathrm{FastICA}}(y,\mathbf{D}_N)\\ &= \frac{1}{4N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}_i)^4 - 3\frac{||\mathbf{w}||^2}{2}.\end{aligned} \tag{7}$$

Eq. (7) has two terms: the first one is the sample-fourth moment, scaled by a factor of 4; and the second one is a sum of squares of the components of the separating vector, which is the simplest form of a regularizer, a technique for implementing regularization known as weight decay (see §9.2 in [3]). Clearly, the minimization of (7) cannot be similar to that performed with the fourth cumulant, since whitening always ensures that $\hat{E}\{(\mathbf{w}^T\mathbf{x})^2\} = 1$ for all $N<\infty$. The use of fourth moment as a contrast function dates from the late 1980s: it was first introduced for a phase-correction estimation problem [18] (as denoted in [23]), later for blind deconvolution [20] and, subsequently, for blind source separation [5,9]. Differences among all these approaches including FastICA mainly relies in the application field and the optimization procedure that comes from gradient descent [20], relative gradient [9] to the Newton-like optimization procedure of FastICA [17,19].

3. An alternative version of FastICA

3.1. Reformulating the objective function for FastICA

As Section 5.2.4 [15,16] observes, FastICA is derived from an optimal cost function based on the fourth cumulant

$$J_{\mathrm{FastICA}}[\mathbf{w}] = \frac{1}{4}\phi^o_{\kappa 4}[\mathbf{w}] = \frac{1}{4}\phi^o_{\kappa 4}[y] = \frac{1}{4}[m_4(y) - 3m_2(y)^2]$$
$$\text{subject to the constraint } ||\mathbf{w}||^2 = 1, \tag{8}$$

where $\{m_r,\ r = 2{,}4\}$ are the moments about the mean defined by

$$m_r = E\{(y - E[y])^r\}, \tag{9}$$

where the gradient of (8) needed for the update rule is

$$\nabla_{\mathbf{w}} J_{\mathrm{FastICA}}(\mathbf{w}) = E\{y^3\mathbf{x}\} - 3E\{y^2\}E\{y\mathbf{x}\}. \tag{10}$$

If we assume the known constraints of orthonormality in $\mathbf{A}$ and the normalization of the source signal to the unit variance achieved by the whitening process, then the sample covariance matrix $\hat{R}_{XX}$ and $\hat{E}\{y^2\}$ yields

$$\hat{R}_{XX} = \hat{E}\{xx^T\} = I, \tag{11}$$

$$\hat{E}\{y^2\} = 1. \tag{12}$$

Consequently, the estimation of (10) can be simplified to the final form employed in the FastICA algorithm,

$$\nabla_{\mathbf{w}} \hat{J}_{\mathrm{FastICA}}(\mathbf{w}) = \hat{E}\{y^3\mathbf{x}\} - 3\mathbf{w}. \tag{13}$$

However, if no whitening is performed a scaled version of the sample-fourth cumulant to be optimized can be employed, i.e.

$$\hat{J}_{\mathrm{FastICA2}}(\mathbf{w}, \mathbf{D}_N) = \hat{J}_{\mathrm{FastICA2}}(y, \mathbf{D}_N) = -\frac{\beta}{4}[\hat{m}_4(y) - 3\hat{m}_2(y)^2], \tag{14}$$

where $\{\hat{m}_r, r = 2, 4\}$ are the estimations of the moments defined by

$$\hat{m}_r = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}[n])^r \tag{15}$$

and β is the sign of the kurtosis of the source to be extracted ($+1$ for leptokurtic signals and -1 for platykurtic signals).

3.2. The FastICA2 algorithm

Now, if (14) is minimized instead of (7), the gradient needed for the constrained optimization gives the estimator of (10) computed with D_N, i.e.

$$\begin{aligned}\nabla_{\mathbf{w}} \hat{J}_{\mathrm{FastICA2}}(\mathbf{w}) &= -\beta\{\hat{E}\{y^3\mathbf{x}\} - 3\hat{E}\{y^2\}\hat{E}\{y\mathbf{x}\}\} \\ &= \frac{-\beta}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}_i)^3\mathbf{x}_i + 3\beta\left\{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}_i)^2\right\} \\ &\quad \times \left\{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i\right\}. \end{aligned} \tag{16}$$

Therefore, we can express the FastICA2 algorithm as follows:

1. Initialize randomly $\mathbf{w}[0]$ of norm 1. Let $k = 1$.
2. Update $\mathbf{w}[k] = -\beta\hat{E}\{\mathbf{x}(\mathbf{w}[k-1]^T\mathbf{x})^3\} + 3\beta\hat{E}\{(\mathbf{w}[k-1]^T\mathbf{x})^2\} \times \hat{E}\{\mathbf{x}(\mathbf{w}[k-1]^T\mathbf{x})\}$.
3. Divide $\mathbf{w}[k]$ by its norm
4. If $|\mathbf{w}[k]^T\mathbf{w}[k-1]|$ is not close to 1, $k = k + 1$ and return to Step 2. Otherwise, finish with $\mathbf{w}^* = \mathbf{w}[k]$.

3.3. Estimation errors in the FastICA objective functions

In cases in which the source signals exhibit unit variance and the mixing matrix is orthonormal, the optimization of (7) or (14) is equivalent for $N \to \infty$. However, the estimation errors introduced for employing a finite set $\mathbf{D}_N$ keep this equivalence untrue. The question of how these optimizations differ in the generalization error can be analyzed by studying bounds on the rate of uniform convergence between the sample-based cost functions $\hat{J}_{\mathrm{HOS}}[\mathbf{w}]$ and their mathematical expectations $J_{\mathrm{HOS}}[\mathbf{w}] = E(\hat{J}_{\mathrm{HOS}}[\mathbf{w}])$. According to [24], these bounds have the subsequent form

$$P\{|\hat{J}[\mathbf{w}] - J[\mathbf{w}]| \leqslant \chi\} \geqslant 1 - \eta(N, \chi). \tag{17}$$

The SLT developed in [24], and later in [25,26], is focused on deriving distribution-independent bounds, i.e. bounds based on no constrains about the probability distribution of the cost function. For example, if it is bounded then the Hoeffding inequality can be employed in (17). However, when information about the distribution of $\hat{J}$ is available, distribution-dependent bounds can be obtained. Particularly, the Beinaymé–Chebyshev inequality can be then employed to make statements in probability (Section 10.8 in [22]) substituting the different variables in (17) as follows:

$$J[\mathbf{w}] = E(\hat{J}[\mathbf{w}]), \tag{18}$$

$$\chi = \lambda\sqrt{\mathrm{var}(\hat{J}[\mathbf{w}])}, \tag{19}$$

$$\eta(N, \chi) = 1/\lambda^2 \text{ with } \lambda > 0, \tag{20}$$

where $\mathrm{var}(\hat{J}[\mathbf{w}])$ can be computed according to the distribution of $\hat{J}[\mathbf{w}]$. In source separation, $\hat{J}[\mathbf{w}] = \hat{J}[y]$ (for a given value of $\mathbf{w}$) corresponds to a particular probability

distribution when the distributions of source signals are known. This fact is especially clear for projections that recover one of the source signals. Additionally, in problems with a great number of sources, the distribution of y in points where sources are maximum mixed (e.g. $y = y = (1/\sqrt{2})(s_1 + s_2)$ for sources s_1 and s_2 mixed with an orthonormal matrix), which typically corresponds to those points of $\hat{J}[\mathbf{w}]$ is maximum, will tend towards normality due to the central limit theorem given certain conditions. In all these cases, a simple computation of variances is feasible following Section 10.5 [22], since the variance of a function $g(t1,\ t2,\ldots t_k)$ (abbreviated here as $g(\mathbf{t})$), with statistics t_i computed from N samples that have mean θ_i and variances of the order N^{-r} $(r>0)$, is given by

$$\mathrm{var}\{g(\mathbf{t})\} = \sum_{i=1}^{k}\{g_i'(\boldsymbol{\theta})\}^2\mathrm{var}(t_i) + \sum_{i\neq j}^{k}\sum_{j=1}^{k}\{g_i'(\boldsymbol{\theta})\}\{g_j'(\boldsymbol{\theta})\}^2\mathrm{cov}(t_i,t_j) + O(N^{-r}), \quad (21)$$

where $g_i'(\boldsymbol{\theta})$ denotes $\partial g(\mathbf{t})/\partial t_i$ evaluated at $\theta_1, \theta_2, \ldots, \theta_k$. The analysis of variances in (21) is done with a precision of $O(N^{-r})$ and consequently mean and variances of statistics are also assumed to be of $O(N^{-r})$ with r typically equal to 1. Following [22] , the sample HOS contrasts based on sample moments $\hat{m}_r$ gives a mean $E(\hat{m}_r) = m_r + O(N^{-1})$, which practically converges to the expected value to the order of N^{-1}. Consequently, in the subsequent analysis results are correct up to $O(N^{-1})$. If $t_1 = \hat{m}_4$, $t_2 = \hat{m}_2$ and $g(t_1, t_2) = -4\hat{J}_{\mathrm{FastICA2}}/\beta = t_1 - 3t_2^2$, we achieve

$$\mathrm{var}(-4\hat{J}_{\mathrm{FastICA2}}/\beta) = \mathrm{var}(\hat{\kappa}_4) = \mathrm{var}(\hat{m}_4) + 36m_2^2\mathrm{var}(\hat{m}_2) - 12m_2\mathrm{cov}(\hat{m}_4, \hat{m}_2), \quad (22)$$

where m_i denote the i th-order moment, $\hat{m}_i$ is the sample i th-order moment and $\hat{\kappa}_4$ is the sample-fourth cumulant. Similarly, we find

$$\mathrm{var}(-4\hat{J}_{\mathrm{FastICA}}/\beta) = \mathrm{var}(\hat{m}_4), \quad (23)$$

where variances of sample moments for populations around zero are given by ([22], p. 347)

$$\mathrm{var}(\hat{m}_r) = \frac{1}{N}(m_{2r} - m_r^2), \quad (24)$$

$$\mathrm{cov}(\hat{m}_\mathrm{q}, \hat{m}_r) = \frac{1}{N}(m_{\mathrm{q}+r} - m_\mathrm{q}m_r). \quad (25)$$

Substituting (25) and (26) in (23) and (24) yields

$$\mathrm{var}(-4\hat{J}_{\mathrm{FastICA2}}/\beta) = \frac{1}{N}\{m_8 - m_4^2 + 36m_2^2(m_4 - m_2^2) - 12m_2(m_6 - m_4m_2)\}, \quad (26)$$

$$\mathrm{var}(-4\hat{J}_{\mathrm{FastICA}}/\beta) = \frac{1}{N}(m_8 - m_4^2). \quad (27)$$

As shown in (27) and (28), the variance of both sample-based cost functions depends differently upon higher moments of even orders of magnitude (2, 4, 6 and 8), so their distribution-dependent probability inequalities obtained substituting (18)–(20) in (17) are not the same. In this way, their optimization will lead to different solutions. The question now is which one will have a better accuracy or, more precisely, which one will lead to a learning algorithm that computes a solution closer to the optimal one, a column vector of $\mathbf{A}^{-1}$. An indirect answer could be provided by first analyzing which sample-based cost function has a lower variance and hence bears a greater resemblance to the optimal counterpart and thus it fulfils with greater probability (17). According to [1] , where the variance of several HOS contrast was computed for several distributions, $\mathrm{var}(-4\hat{J}_{\mathrm{FastICA2}}/\beta)$ is $9.882/N$, $24/N$ and $1656/N$, and $\mathrm{var}(-4\hat{J}_{\mathrm{FastICA}}/\beta)$ is $5.76/N$, $96/N$ and $2484/N$ for uniform, normal and Laplacian distributions, respectively. Here, we extend these previous results studying the zero-mean generalized Gaussian density (p.40, [17]) given by

$$p_y(y) = C\,\exp\left(-\frac{|y|^v}{vE\{|y|^v\}}\right), \quad (28)$$

where C is a scaling constant that ensures $\int p_y(y)\,\mathrm{d}y = 1$ and $v>0$ determines the type of distribution. Impulsive-type distributions are obtained for $0<v<1$, $v=1$ yields a Laplacian distribution, $v=2$ a normal distribution, and $v\to\infty$ the uniform density. Finally, we have sub-Gaussian densities and super-Gaussian densities for $v>2$ and $v<2$, respectively. Fig. 1 shows the evolution of $N\,\mathrm{var}(\hat{\kappa}_4)$ and

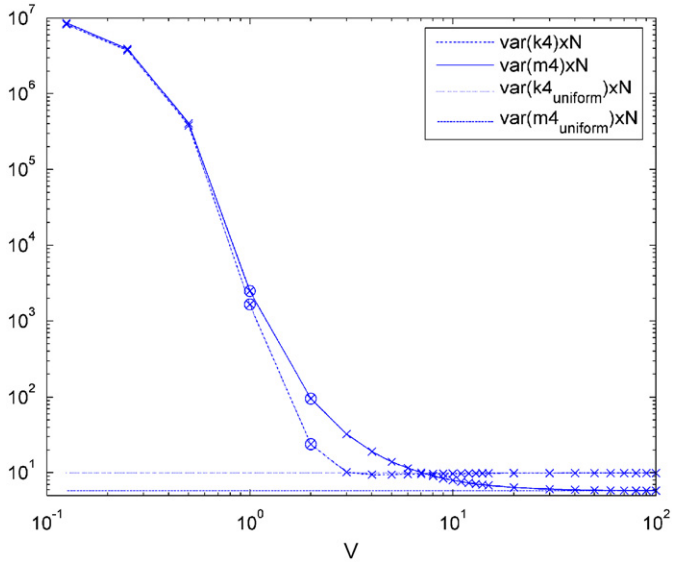

Fig. 1. Variances of the cost functions of FastICA and FastICA2 for those projections which induce a generalized Gaussian distributions in y, with $m_1 = 0$ and $m_2 = 1$ for different values of the shape parameter v. (*Note*: Crosses X denote those points of the curve estimated while circles O are the theoretical values computed for Laplacian and normal distributions. Also the asymptotes $N\,\mathrm{var}(\hat{\kappa}_{4\ \mathrm{uniform}})$ and $N\,\mathrm{var}(\hat{m}_{4\ \mathrm{uniform}})$ are computed analytically.)

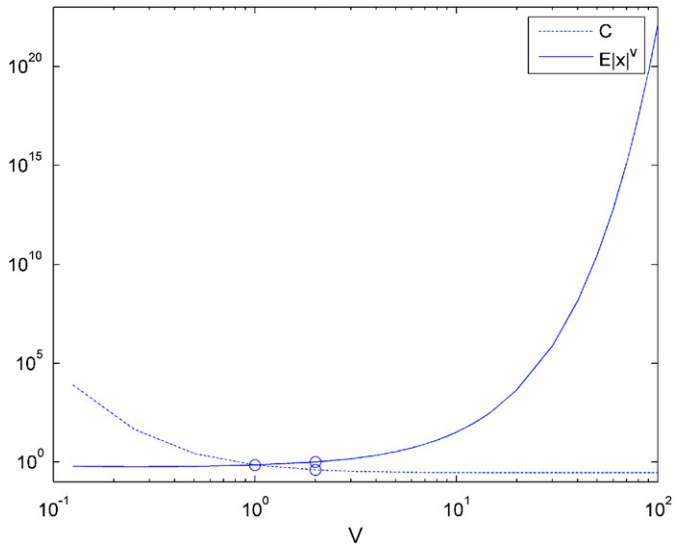

Fig. 2. The estimated parameters C and $E\{|y|^v\}$ of a generalized Gaussian with $m_1 = 0$ and $m_2 = 1$ for different values of the shape parameter v. (*Note*: circles O are the theoretical values computed for Laplacian and normal distributions.)

$N\,\mathrm{var}(\hat{m}_4)$ of a generalized Gaussian random variable with zero mean and $m_2 = 1$ for different values of the shape parameter v. Theses curves have been estimated through numerical integration of (22) and (23) with the parameters C and $E\{|y|^v\}$ computed through a recursive algorithm that involves a simultaneous numerical integration of equations $\int p_y(y)\,\mathrm{d}y = 1$ and $\int y^2 p_y(y)\,\mathrm{d}y = 1$ until a satisfactory solution for these parameters is achieved (see Fig. 2). As it can be observed, both variances are monotonically decreasing functions respect to the shape parameter v and have asymptotes in $v \to \infty$ which are the variance of $N\,\mathrm{var}(\hat{\kappa}_{4\ \mathrm{uniform}})$ and $N\,\mathrm{var}(\hat{m}_{4\ \mathrm{uniform}})$.

However, the question of which cost function has a better accuracy, i.e. a lower estimation error, can be further analyzed having in mind that SLT is really interested in studying a more stringent condition than the more general probability inequality in (17), i.e.

$$P\{\sup_{\mathbf{w}} |\hat{\phi}^o[\mathbf{w}] - \phi^o[\mathbf{w}]| \leqslant \chi\} \geqslant 1 - \eta(N, \chi), \tag{29}$$

where sup denotes supremum. The rationale about analyzing such a worse case is that even large deviation in one single point may incur in a large deviation in achieving the minimum of point of $\hat{\phi}^o$ respect to $\mathbf{w}^* = \arg\min_{\mathbf{w}} \phi^o[\mathbf{w}]$ during optimization (Section 2.6 in [24]). In such a worst-case scenario, the solution $\mathbf{w}_j$ that causes the greatest variance on the random variable y will limit the uniform convergence of the sample cost function to its expected value. According to Fig. 1, if there are many sources with sub-Gaussian distribution (e.g. uniform), this limit will characteristically be controlled by the variance of a normal distribution induced in y for those values of the separating matrix that causes the maximum mixing. On the other hand, if we had sources with a super-Gaussian distribution (e.g. Laplacian) this limit would be established by the variances on points in which a recovery of one of the source signals was accomplished. Consequently, the optimization of $\hat{J}_{FastICA2}$ seems preferable for these two cases since it has a lower variance for normal distributions (the worst case for sub-Gaussian sources) and also for super-Gaussian sources.

Finally, it is worth noting to draw attention to an important consequence of studying the above worst-case setting. If (29) is satisfied then the following inequality is also fulfiled (Section 2.6 in [24]),

$$P\{\phi^o[\mathbf{w}^+] - \phi^o[\mathbf{w}^*] \leqslant 2\chi\} \geqslant 1 - \eta(N, \chi) \text{ with} \tag{30}$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \phi^o[\mathbf{w}], \tag{31}$$

$$\mathbf{w}^+ = \arg\min_{\mathbf{w}} \hat{\phi}^o[\mathbf{w}]. \tag{32}$$

The inequality (30) bounds the error in the expected cost function calculated in those points which minimizes the expected and sample cost functions, and thus it is possible to bound probabilistically the effect of minimizing the sample cost function depending on the number of samples N.

4. Experimental results

In this section, experiments in blind source separation of two and large-scale sources are presented in order to obtain consistent results according to the distribution-dependent theory introduced in Section 3. For this purpose, we only work with artificial data taken from known distribution, i.e. uniform and Laplacian populations, in order to check if the rationale suggested in Subsection 3.3 related to these distributions is supported empirically. The bounds (17), (29) and (30) are probabilistic, i.e. they are fulfiled with a given probability, and does not take into account the particular optimization procedure whose good convergence can only be guaranteed for $N \to \infty$ [12] and, consequently, it can introduce optimization errors in the computed solution. For these reasons, we do not try to estimate directly these bounds but to confirm them quantitatively through the comparison between the separating vector and matrix computed during learning and the mixing matrix.

4.1. Orthonormal mixing of two uniform and Laplacian sources

We relate, in this subsection, how FastICA2 (Section 3.2) was tested against a modified version of FastICA (Section 2.1) called FastICA1, which includes the term $-\beta$ for detecting the sign of kurtosis of the source signals to be extracted. We employed a linear BSS model in which two source signals $s_1[n]$ and $s_2[n]$ with uniform and Laplacian normalized distributions, i.e. $\mathbf{K}_s = \mathbf{I}$, are mixed

through the orthonormal matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix} = \begin{bmatrix} \cos\alpha & \operatorname{sen}\alpha \\ \operatorname{sen}\alpha & -\cos\alpha \end{bmatrix} = \begin{bmatrix} \cos(\pi/4) & \operatorname{sen}(\pi/4) \\ \operatorname{sen}(\pi/4) & -\cos(\pi/4) \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}. \quad (33)$$

For uniform sources, a triangular distribution is originated when there is maximum mixing instead of normal mixing. This distribution causes a variance in $-4\hat{J}_{\text{FastICA2}}/\beta$ and $-4\hat{J}_{\text{FastICA}}/\beta$ of $9.67/N$ and $23.04/N$, respectively. Experiments were performed on 1000 different data sets D_N of the same size with $N = 5$, 10, 25, 50, 75, 100, 150, 200, 250, 500, 1000, 2500, 5000, 10000 and 25000. The initial separating projection $\mathbf{w} = [\cos\theta \operatorname{sen}\theta]^T$ was chosen randomly within the uniform interval $\theta = [-\pi, \pi]$, and the stopping condition was settled at $|\mathbf{w}[k]^T\mathbf{w}[k-1]| = 1$. Figs. 3–6 display the mean and variance estimated for the 1000 sets for each N of a performance index (PI) and a convergence index (CI) defined as

$$PI = \max\left(|\mathbf{w}^{*T}\mathbf{a}_1|, |\mathbf{w}^{*T}\mathbf{a}_2|\right) = \max\left(|\cos(\alpha-\theta)|, |\operatorname{sen}(\alpha-\theta)|\right), \quad (34)$$

$$CI = (\min_{\mathbf{w}} J_{\text{FastICA}}(\mathbf{w}))/\hat{J}_{\text{FastICA}i}(\mathbf{w}^*), i = 1, 2. \quad (35)$$

Since PI is the absolute value of a scalar product of two one-norm vectors, it is ranged between [0,1] and measures the similarity between the separating vector achieved by the learning algorithm and the nearest-column vector of the mixing matrix $\mathbf{A}$. As observed from (34), it is also a function of the angle error $\varphi = \alpha - \theta$. The convergence index CI measures the divergence between the minimum in the expected cost function and the minimum of the sample cost function obtained during learning. As both minimums come close, CI will tend to unity. Consequently, when the learning algorithms converge to one of column vectors of $\mathbf{A}$ and the solutions achieved by minimizing the expected and sample cost functions are the same, PI and CI tend to unity.

As Figs. 3 and 5 show, FastICA2 visibly outperforms FastICA1 for both separation problems, achieving for almost all the sample sizes N a greater mean (PI) than FastICA1, and thus computing a separating vector closer to one of the column vectors of $\mathbf{A}$. This behavior can be explained in terms of the differences of both algorithms in the evolution of the convergence index CI. For uniform sources, PI of both algorithms are monotonically increasing functions as N augments (Fig. 5) and, accordingly, it suggests that the minimums achieved are increasingly closer to the optimal ones. Accordingly, CIs are monotonically decreasing functions (Fig. 6) but, as it can be observed, CI_{FastICA2} has a greater slope and converges sooner to the unity. In the case of separating Laplacian sources, PI_{FastICA2} (Fig. 3) has a similar evolution than before and its CI (Fig. 4) for $N > 75$ is also again a monotonically decreasing function. Also Fig. 4 shows the unstable evolution of CI_{FastICA1} that justifies the poor performance of FastICA1 for separating two Laplacian sources, even for large data sets (Fig. 3).

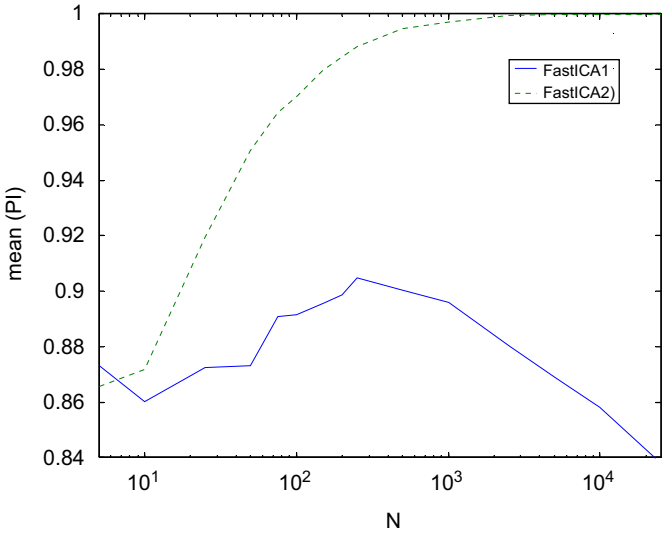

Fig. 3. Mean PI for Laplacian sources.

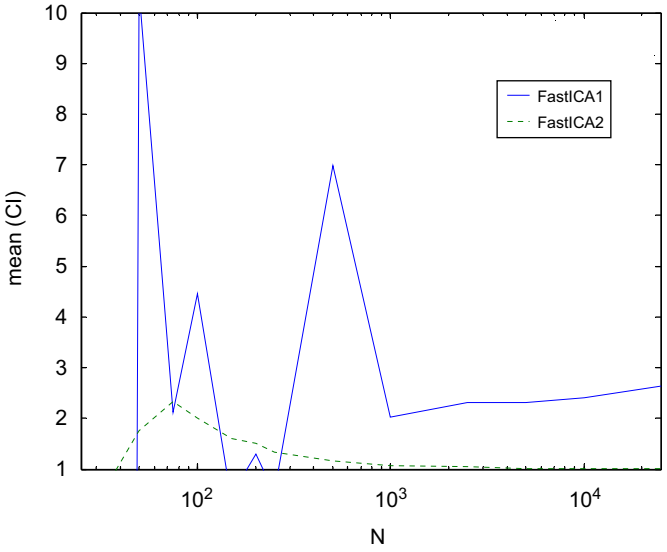

Fig. 4. Mean CI for Laplacian sources.

4.2. Orthonormal mixing of large-scale uniform sources

A second set of experiments was carried out to compare the performance of optimizing the sample-fourth cumulant contrast with the sample-fourth moment in a large-scale separation problem with uniform sources. For this purpose: (1) the symmetrical implementation of FastICA2 was employed following the symmetrical version included in the FastICA package [14], since, as shown in Oja, this

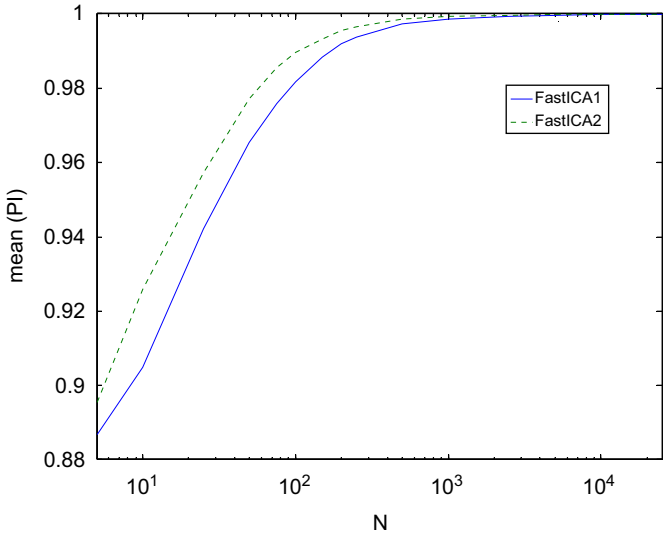

Fig. 5. Mean PI for uniform sources.

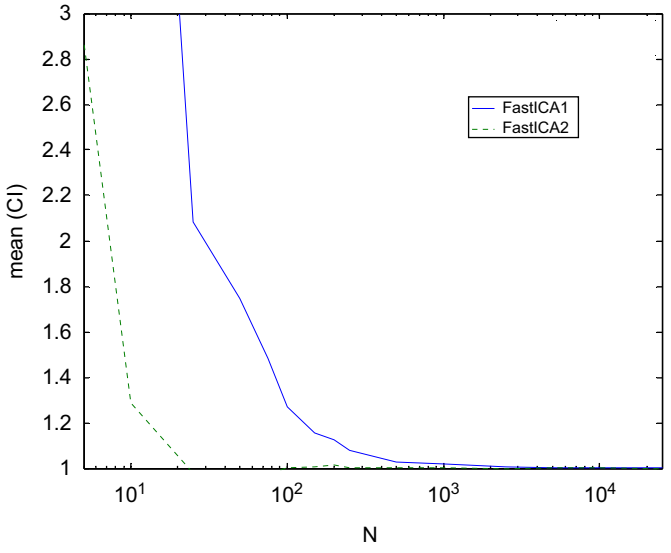

Fig. 6. Mean CI for uniform sources.

version shares the good convergence properties of the one-unit case and, as we observed in a previous set of experiments, all these algorithms achieve better generation results than those based on the deflation approach, (2) the optimization of the sample-fourth moment was ensured enforcing a pre-whitening in the standard FastICA algorithm supplied in [14].

The symmetrical FastICA2 was tested against the following symmetrical FastICA algorithms based on two non-linearities: (1) third power and (2) tanh. (The skew and gauss non-linearities were discarded since they behave very similar than tanh.) The experiments were performed on 1000 different data sets $\mathbf{D}_N$ of the same size with $N = 400$, 1000, 2000, 2500, 6000 and 8000 and a number of sources $p = 4$, 8, 10, 20, 40 and 50. Uniform sources were mixed through an orthonormal matrix A that was randomly chosen. The initial values of the separating matrix and the stopping criterion were settled as those employed by default in the FastICA package. Table 1 shows the mean performance index PI, computed as in ([10] p. 219), and its variance. As can be seen, the PI of FastICA2 is smaller than the rest, which implies a closer estimate of the separating matrix.

5. Conclusions and future work

Our work has analyzed some of the estimation errors introduced by the use of finite date sets in the FastICA algorithm introducing distribution-dependent bounds on the generalization error. Such a two-step approach for extracting independent components, in which the whiteness of output variable is first imposed and then the sample-fourth moment constrast is optimized later, may suffer greater generalization errors than one-step methods in which a term was employed effectively to ensure decorrelation plus a second term based on the sample-fourth cumulant. In particular, it is shown that the sample-fourth cumulant has a lower variance in the generalization error for some distributions which has been confirmed by experiments in separating orthornormal uniform and Laplacian sources provided with several finite sample sets

Table 1
Mean and variance PI of different versions of symmetrical FastICA for large-scale orthornormal mixing of uniform sources with different data set sizes (N) and number of sources (p)

N	p	No whitening	With whitening	
		FastICA2	FastICA with third power	FastICA with tanh
400	4	$2.3 \times 10^{-3}/1.1 \times 10^{-2}$*	$3.7 \times 10^{-3}/1.6 \times 10^{-3}$	$4.7 \times 10^{-3}/2.1 \times 10^{-3}$
1000	8	$1.7 \times 10^{-3}/5 \times 10^{-4}$	$3.3\ \text{x}10^{-3}/7 \times 10^{-4}$	$4.3 \times 10^{-3}/9 \times 10^{-4}$
2000	10	$1.1 \times 10^{-3}/2 \times 10^{-4}$	$2.1 \times 10^{-3}/3 \times 10^{-4}$	$2.7 \times 10^{-3}/4 \times 10^{-4}$
2500	20	$1.9 \times 10^{-3}/2 \times 10^{-4}$	$3.7\ \text{x}10^{-3}/3 \times 10^{-4}$	$4.7 \times 10^{-3}/3 \times 10^{-4}$
6000	40	$1.6 \times 10^{-3}/1 \times 10^{-4}$	$3.1 \times 10^{-3}/8 \times 10^{-4}$	$4.0 \times 10^{-3}/1 \times 10^{-4}$
8000	50	$1.5 \times 10^{-3}/1 \times 10^{-4}$	$3.0 \times 10^{-3}/8 \times 10^{-4}$	$3.8\ \text{x}10^{-3}/7 \times 10^{-4}$

[*mean/variance].

confirms greater accuracy and robustness than the standard version.

A straightforward outcome of this work is that a two-step approach in which the whiteness of Y is first imposed and then a HOS constrast function is restricted later may incur greater generalization errors than one-step methods in which a term $C(Y)$ was employed effectively to ensure decorrelation plus a term $G(Y)$ based on HOS contrasts. Future work will be concentrated in studying which $C(Y)$ and λ in (1) should be employed in order to obtain a suitable learning algorithm based on such contrasts.

Acknowledgments

The author wishes to thank the anonymous reviewers for their comments, which have helped to improve the final version of this paper. This work has been partly supported by a MEC project (TEC2004-05127-C02-01) and also by the EU through FEDER funding.

References

[1] S. Bermejo, Finite sample effects in higher order statistics contrast functions for sequential blind source separation, IEEE Signal Proc. Lett. 12 (6) (2005) 481–484.
[2] S. Bermejo, C. Jutten, J. Cabestany, ISFET source separation: foundations and techniques, Sensors Actuators B: Chem. 113 (2006) 222–233.
[3] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[4] E. Briselli, et al., An independent component analysis-based approach on ballistocardiogram artifact removing, Magn. Resonance Imag. 24 (4) (2006) 393–400.
[5] J.-F. Cardoso, Iterative techniques for blind source separation using only fourth order cumulants, in: Proceeding EUSIPCO, 1992, pp. 739–742.
[6] J.-F. Cardoso, On the performance of orthogonal source separation algorithms, in: M. Holt, et al. (Eds.), Procceedings. EUSIPCO, VII European Signal Proceeding Conference, Edinburgh, Scotland, UK, September 1994, pp. 776–779.
[7] J.-F. Cardoso, Blind signal separation: statistical principles, Proc. IEEE 86 (1998) 2009–2025.
[8] J.-F. Cardoso, Dependence, correlation and Gaussianity in independent component analysis, J. Machine Learn. Res. 4 (2003) 1177–1203.
[9] J.-F. Cardoso, B. Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Process. 44 (12) (1996) 3017–3030.
[10] A. Cichocki, S. Amari, Adaptive Blind Signal and Image Processing, Wiley, New York, 2002.
[11] N. Delfosse, P. Loubaton, Adaptive blind separation of independent sources: a deflation approach, Signal Process. 45 (1995) 59–83.
[12] S. Douglas, On the convergence behavior of the FastICA algorithm. in: Proceedings of the fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Japan.
[13] K. Fan, M. Wang, W. Mo, X. Zhao, Novel copyright protection scheme for digital content, J. Syst. Eng. Electron. 17 (2) (2006) 423–429.
[14] H. Gävert, J. Hurri, J. Särelä, A. Hyvärinen, FastICA for Matlab 5.x, version 2.1, January 15, 2001.
[15] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (1997) 1483–1492.
[16] A. Hyvärinen, E. Oja, One-unit learning rules for independent component analysis, in: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems (NIPS'1996), vol. 9, MIT Press, Cambridge, MA, 1997.
[17] A. Hyvärinen, K. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[18] J. Longbottom, A.T. Walden, R.E. White, Principles and applications of maximum kurtosis phase estimation, Geophys. Prospect. 36 (1988) 115–138.
[19] E. Oja, Convergence of the symmetrical FastICA algorithm. in: Proceedings of the Ninth International Conference on Neural Information Processing (ICONIP'02), vol. 3, pp. 1368–1372.
[20] O. Shalvi, E. Weinstein, New criteria for blind deconvolution of nonminimum phase systems (channels), IEEE Trans. Inform. Theor. 36 (2) (1990) 312–321.
[21] L. Shang, D.S. Huang, J.-X. Du, C.-H. Zheng, Palmprint recognition using FastICA algorithm and radial basis probabilistic neural network, Neurocomputing 69 (13-15) (2006) 1782–1786.
[22] A. Stuart, K. Ord, Kendall's advanced theory of statistics, Distribution Theory, sixth ed., vol. I. Arnold, London, 1994.
[23] J.K. Tugnait, Comments on new criteria for blind deconvolution of nonminimum phase systems (channels), IEEE Trans. Inform. Theor. 38 (1) (1992) 210–212.
[24] V.N. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, Berlin, 1982.
[25] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[26] V.N. Vapnik, The Nature of Statistical Learning Theory, second ed, Springer, Berlin, 2000.
[27] R.D.S. Yadava, R. Chaudhary, Solvation, transduction and independent component analysis for pattern recognition in SAW electronic nose, Sensors Actuators B: Chem. 113 (1) (2006) 1–21.
[28] A. Yeredor, Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation, IEEE Trans. Signal Process. 50 (7) (2002) 1545–1553.

Sergio Bermejo received the M.Sc. and Ph.D. degrees in telecommunication engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1996 and 2000, respectively. He joined the Department of Electronics Engineering (DEE) of UPC as a Researcher in 1996. Currently, he is an Associate Professor in the DEE and teaches at the School of Telecommunications Engineering of Barcelona (ETSETB). His research interests are statistical learning, with a special focus on large-margin classification, unsupervised learning, and their application to signal processing, software agents, smart sensors, and autonomous robotics.