DEFLATION-BASED FASTICA RELOADED

Klaus Nordhausen[†], *Pauliina Ilmonen*[†], *Abhijit Mandal*[†], *Hannu Oja*[†] *and Esa Ollila*[‡,⋆]

[†]School of Health Sciences,
University of Tampere
FIN-33014 Tampere, Finland

[‡]Signal Processing and Acoustics
Aalto University
FIN-00076 Aalto, Finland

[⋆]Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA

ABSTRACT

Deflation-based FastICA, where independent components (IC's) are extracted one-by-one, is among the most popular methods for estimating an unmixing matrix in the independent component analysis (ICA) model. In the literature, it is often seen rather as an algorithm than an estimator related to a certain objective function, and only recently has its statistical properties been derived. One of the recent findings is that the order, in which the independent components are extracted in practice, has a strong effect on the performance of the estimator. In this paper we review these recent findings and propose a new "reloaded" procedure to ensure that the independent components are extracted in an optimal order. The reloaded algorithm improves the separation performance of the deflation-based FastICA estimator as amply illustrated by our simulation studies. Reloading also seems to render the algorithm more stable.

1. INTRODUCTION

The independent component (IC) model is a semiparametric model which has gained increasing interest in various fields of science and engineering during the recent years [6]. The basic IC model assumes that the observed p-variate random vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ is a linear mixture of the p mutually independent sources (IC's) $\mathbf{s} = (s_1, \ldots, s_p)^T$. Then

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{A}$ is assumed to be a full rank $p \times p$ unknown mixing matrix. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote a random sample from the IC model (1). The aim of the independent component analysis (ICA) is to find an estimate $\hat{\mathbf{W}}$ (using the random sample $\mathbf{X}$) of some $p \times p$ unmixing matrix $\mathbf{W}$ verifying $\mathbf{s} = \mathbf{W}\mathbf{x}$ up to permutation, sign and scale changes; see [6]. Naturally $\mathbf{W} = \mathbf{A}^{-1}$ is one possible solution.

In the following, $\mathbf{P}$ denotes a permutation matrix (obtained by permuting the rows or columns of $\mathbf{I}_p$), $\mathbf{J}$ denotes a sign-chance matrix (a $p \times p$ diagonal matrix with entries ± 1), and $\mathbf{D}$ denotes a $p \times p$ diagonal matrix with positive diagonal elements. Let $\mathscr{G}$ denote the set of all full-rank $p \times p$ matrices. Then the set of $p \times p$ matrices, defined as

$$\mathscr{C} = \{\mathbf{C} : \mathbf{C} = \mathbf{P}\mathbf{J}\mathbf{D} \text{ for some } \mathbf{P}, \mathbf{J} \text{ and } \mathbf{D}\},$$

is a subset of $\mathscr{G}$. If a matrix $\mathbf{W} \in \mathscr{G}$ is an unmixing matrix in the IC model (1), then so is $\mathbf{C}\mathbf{W}$ for any $\mathbf{C} \in \mathscr{C}$. We then say that two unmixing matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are (ICA) equivalent if $\mathbf{W}_1 = \mathbf{C}\mathbf{W}_2$ for some $\mathbf{C} \in \mathscr{C}$, and we write $\mathbf{W}_1 \sim \mathbf{W}_2$.

All reasonable estimates $\hat{\mathbf{W}}$ should naturally converge in probability to some population value $\mathbf{W}(F_{\mathbf{x}})$, that is, the value of an independent component (IC) functional $\mathbf{W}$ at $F_{\mathbf{x}}$, where $F_{\mathbf{x}}$ denotes the cumulative distribution function (cdf) of $\mathbf{x}$. A formal (model independent) definition [9] of an IC functional is given below.

Definition 1. *Let $F_{\mathbf{x}}$ denote the cdf of $\mathbf{x}$. The functional $\mathbf{W}(F_{\mathbf{x}}) \in \mathscr{G}$ is an IC functional in the IC model (1) if (i) $\mathbf{W}(F_{\mathbf{x}})\mathbf{A} \sim \mathbf{I}_p$ and (ii) it is affine equivariant in the sense that $\mathbf{W}(F_{\mathbf{B}\mathbf{x}}) = \mathbf{W}(F_{\mathbf{x}})\mathbf{B}^{-1}$ for all $\mathbf{B} \in \mathscr{G}$.*

Note that $\mathbf{W}(F_{\mathbf{B}\mathbf{x}})\mathbf{B}\mathbf{x} = \mathbf{W}(F_{\mathbf{x}})\mathbf{x}$, and therefore $\mathbf{W}(F_{\mathbf{x}})\mathbf{x}$ is invariant under invertible linear transformations of the observation vectors. A finite sample estimator corresponding to an IC functional is obtained if the functional is applied to the empirical distribution based on $\mathbf{X}$. We then write $\hat{\mathbf{W}} = \mathbf{W}(\mathbf{X})$ for the obtained estimator. The estimator is then also affine equivariant in the sense that $\hat{\mathbf{W}}(\mathbf{B}\mathbf{X}) = \hat{\mathbf{W}}(\mathbf{X})\mathbf{B}^{-1}$. Let us denote by $\mathbf{S}(F_{\mathbf{x}}) \equiv \mathrm{COV}(\mathbf{x})$ the covariance matrix (functional) of a random vector $\mathbf{x}$. We note that many IC functionals proposed in the literature are defined either implicitly or explicitly in such a way that the covariance matrix of the obtained source vector is equal to the identity matrix, *i.e.* $\mathrm{COV}(\mathbf{W}(F_{\mathbf{x}})\mathbf{x}) = \mathbf{I}_p$, in which case $\mathbf{W}(F_{\mathbf{x}}) = \mathbf{U}(F_{\mathbf{x}})\mathbf{S}^{-1/2}(F_{\mathbf{x}})$, where $\mathbf{U}(F_{\mathbf{x}})$ is an orthogonal matrix.

The estimator of interest in this paper is the deflation-based FastICA estimator [4, 5]. The paper is organized as follows. Sections 2 recalls the deflation-based FastICA algorithm and estimating equations, while statistical properties of the estimator are discussed in Sections 3. In Section 4, a new novel method is proposed, called the reloaded FastICA, to optimize the extraction order of the sources in succeeding FastICA deflation stages. A Simulation study in Section 5 illustrates the usefulness of our approach, whereas Section 6 presents our conclusions.

2. DEFLATION-BASED FASTICA

Deflation-based FastICA, hereafter FastICA for short, was introduced in [4] and further developed in [5]. Up to date it can be considered among one of the most popular methods to solve the ICA problem.

2.1 FastICA algorithm

Write $\mathbf{z} = \mathbf{S}^{-1/2}(F_{\mathbf{x}})(\mathbf{x} - \mathrm{E}(\mathbf{x}))$ for the whitened random variable, where the square root matrix is chosen to be symmetric. FastICA can be seen as a projection pursuit method, where the directions $\mathbf{u}_k$, maximizing a measure of non-Gaussianity $|\mathrm{E}\left(G(\mathbf{u}_k^T\mathbf{z})\right)|$, are found successively under the constraint that $\mathbf{u}_k$ is orthonormal with the previously found directions $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ (for $k = 1, \ldots, p-1$), where $G(\cdot)$

can be any twice continuously differentiable nonlinear and nonquadratic function with $G(0) = 0$. The unmixing matrix is then $\mathbf{W} = \mathbf{U}\mathbf{S}^{-1/2}$ where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)^T$. Note that the last vector $\mathbf{u}_p$ is set as a unit vector orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_{p-1}$. Let $g(\cdot)$ denote the derivative of $G(\cdot)$, called the nonlinearity. Commonly used nonlinearities are *pow3*: $g(u) = u^3$, *tanh*: $g(u) = \tanh(a_1 u)$, *gaus*: $g(u) = u\exp(-a_2 u^2/2)$ and *skew*: $g(u) = u^2$, where a_1 and a_2 are tuning parameters, usually chosen to be equal to 1.

Due to the whitening, the FastICA method is commonly formulated as an algorithm for finding an estimator $\hat{\mathbf{U}}$. The algorithm (and its slight variations) given below for the directions $\hat{\mathbf{u}}_k$, $k = 1, \ldots, p-1$, is generally accepted in the literature. In the algorithm, $\hat{\mathbf{u}}_j$, $j = 1 \ldots, k-1$, are the previously found directions and the sample mean vector and the sample covariance matrix are denoted by $\bar{\mathbf{x}}$ and $\hat{\mathbf{S}}$, respectively.

Algorithm 1 deflation-based FastICA algorithm for $\hat{\mathbf{u}}_k$

$\mathbf{x}_i \leftarrow \hat{\mathbf{S}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$ {Whiten the data}
$\mathbf{u}_{k,0} \leftarrow \mathbf{u}_{k,init}$ {Choose an initial value}
$\Delta = \infty$
while $\varepsilon < \Delta$ **do**
  $\mathbf{u}_{k,1} \leftarrow \text{ave}(\mathbf{x}_i g(\mathbf{u}_{k,0}^T \mathbf{x}_i)) - \text{ave}(g'(\mathbf{u}_{k,0}^T \mathbf{x}_i))\mathbf{u}_{k,0}$
  $\mathbf{u}_{k,1} \leftarrow \mathbf{u}_{k,1} - \sum_{j=1}^{k-1}(\mathbf{u}_{k,1}^T \hat{\mathbf{u}}_j)\hat{\mathbf{u}}_j$
  $\mathbf{u}_{k,1} \leftarrow \mathbf{u}_{k,1}/\|\mathbf{u}_{k,1}\|$
  $\Delta = \|\mathbf{u}_{k,1} - \mathbf{u}_{k,0}\|$
  $\mathbf{u}_{k,0} \leftarrow \mathbf{u}_{k,1}$
end while
RETURN $\hat{\mathbf{u}}_k = \mathbf{u}_{k,1}$

The FastICA estimator of the unmixing matrix is thus $\hat{\mathbf{W}} = \hat{\mathbf{U}}\hat{\mathbf{S}}^{-1/2}$ with $\hat{\mathbf{U}}$ coming from the algorithm. The order in which the sources are found depends heavily on the initial value $\mathbf{U}_{init} = (\mathbf{u}_{1,init}, \ldots, \mathbf{u}_{p,init})^T$. Write next $\mathbf{W}(\mathbf{U}, \mathbf{X})$ for the estimate based on the data $\mathbf{X}$ and the initial value $\mathbf{U}_{init} = \mathbf{U}$. If $\mathbf{U}$ is random, then the estimate $\mathbf{W}(\mathbf{U}, \mathbf{X})$ may get $p!$ different values depending on random $\mathbf{U}$, and the different solutions may not be ICA equivalent.

Let $\mathbf{S}(\mathbf{X})$ be the covariance matrix computed from $\mathbf{X}$. It is well known that $\mathbf{S}(\mathbf{B}\mathbf{X})^{-1/2}(\mathbf{B}\mathbf{X}) = \mathbf{V}_\mathbf{B}\mathbf{S}(\mathbf{X})^{-1/2}\mathbf{X}$ where $\mathbf{V}_B$ is an orthogonal matrix depending on $\mathbf{B}$ (and $\mathbf{X}$). With a fixed choice $\mathbf{U}$, the estimate $\mathbf{W}(\mathbf{U}, \mathbf{X})$ is affine equivariant in the sense that $\mathbf{W}(\mathbf{U}, \mathbf{B}\mathbf{X}) = \mathbf{W}(\mathbf{U}, \mathbf{X})\mathbf{B}^{-1}$ if $\mathbf{W}(\mathbf{U}, \mathbf{X}) = \mathbf{W}(\mathbf{U}\mathbf{V}_\mathbf{B}, \mathbf{X})$, that is, if $\mathbf{W}(\mathbf{U}, \mathbf{X})$ and $\mathbf{W}(\mathbf{U}\mathbf{V}_\mathbf{B}, \mathbf{X})$ find the sources in the same order. (The equalities above are up to sign changes of the rows.) A natural question then is: Is there any choice $\mathbf{U}_{init} = \mathbf{U}(\mathbf{X})$ such that the "reloaded" fastICA estimate $\mathbf{W}(\mathbf{U}(\mathbf{X}), \mathbf{X})$ is fully affine equivariant. We answer this question in Section 4.

2.2 Estimating equations

To facilitate statistical analysis, it is appropriate to formulate the method as an estimator verifying a set of estimating equations. Furthemore, it is useful to formulate the estimator without the pre-whitening stage. Let $\mathbf{T}(F_\mathbf{x}) = \mathrm{E}(\mathbf{x})$ denote the mean vector (functional). The deflation-based FastICA functional $\mathbf{w}_k(F_\mathbf{x})$, $k = 1, \ldots, p-1$, may be seen [11, 12] as an optimizer of

$$|\,\mathrm{E}[G(\mathbf{w}_k^T(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))]\,|$$

under the constraints (i) $\mathbf{w}_k^T \mathbf{S}(F_\mathbf{x})\mathbf{w}_k = 1$ and (ii) $\mathbf{w}_j^T \mathbf{S}(F_\mathbf{x})\mathbf{w}_k = 0$ for $j = 1, \ldots, k-1$. (For $\mathbf{w}_1$, only the first constraint is needed.) Note that, for the definition of the functional $\mathbf{w}_k$, we need functionals $\mathbf{T}$, $\mathbf{S}$, and $\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$.

Using the Lagrange multiplier technique, one can easily show [9, 12] that (under general assumptions) the unmixing matrix functional $\mathbf{W}(F_\mathbf{x}) = (\mathbf{w}_1(F_\mathbf{x}), \ldots, \mathbf{w}_p(F_\mathbf{x}))^T$ satisfies the p estimating equations

$$\mathrm{E}\Big[g\big(\mathbf{w}_k^T(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\big)(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\Big]$$
$$= \mathbf{S}(F_\mathbf{x}) \sum_{j=1}^{k} \mathbf{w}_j \mathbf{w}_j^T \mathrm{E}\Big[g\big(\mathbf{w}_k^T(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\big)(\mathbf{x} - \mathbf{T}(F_\mathbf{x}))\Big],$$

$k = 1, \ldots, p$. Note that, if $\mathbf{s} = \mathbf{W}\mathbf{x}$ has independent components, then $\mathbf{W}$ solves the estimating equations. It is also important to note that, for all permutation matrices $\mathbf{P}$, also $\mathbf{P}\mathbf{W}$ then solves the estimating equations, and therefore the estimating equations do not fix the order of the unmixing vectors $\mathbf{w}_1, \ldots, \mathbf{w}_p$.

3. STATISTICAL PROPERTIES

Despite being such a popular tool, rigorous statistical analysis of the deflation-based FastICA estimator has not been given until quite recently in [9, 11–13]. In this section we discuss the limiting distribution and robustness properties of the deflation-based FastICA estimator. Without loss of generality we assume that $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$, $\mathrm{COV}(\mathbf{x}_i) = \mathbf{I}_p$, and the true mixing matrix is $\mathbf{A} = \mathbf{I}_p = (\mathbf{e}_1, \ldots, \mathbf{e}_p)^T$.

3.1 Limiting distribution

If the first four moments of $\mathbf{s}$ exist, then by the central limit theorem, the joint distribution of $\sqrt{n}\bar{\mathbf{x}}$ and $\sqrt{n}\mathrm{vec}(\hat{\mathbf{S}} - \mathbf{I}_p)$ is asymptotically normal. Furthermore, the existence of the expected values $\mu_{g,k} = \mathrm{E}[g(\mathbf{e}_k^T \mathbf{x}_i)]$,

$$\sigma_{g,k}^2 = Var[g(\mathbf{e}_k^T \mathbf{x}_i)], \quad \lambda_{g,k} = \mathrm{E}[g(\mathbf{e}_k^T \mathbf{x}_i)\mathbf{e}_k^T \mathbf{x}_i]$$

and

$$\delta_{g,k} = \mathrm{E}[g'(\mathbf{e}_k^T \mathbf{x}_i)], \quad \tau_{g,k} = \mathrm{E}[g'(\mathbf{e}_k^T \mathbf{x}_i)\mathbf{e}_k^T \mathbf{x}_i]$$

are required. We also need to assume that $\delta_{g,k} \neq \lambda_{g,k}$, $k = 1, \ldots, p-1$, and we write

$$\alpha_{g,k} = \frac{\sigma_{g,k}^2 - \lambda_{g,k}^2}{(\lambda_{g,k} - \delta_{g,k})^2}, \quad k = 1, \ldots, p. \tag{2}$$

Write $\mathbf{T}_k = \frac{1}{n}\sum_{i=1}^n (g(\mathbf{e}_k^T \mathbf{x}_i) - \mu_{g,k})\mathbf{x}_i$ and $\hat{\mathbf{T}}_k = \frac{1}{n}\sum_{i=1}^n g(\hat{\mathbf{w}}_k^T(\mathbf{x}_i - \bar{\mathbf{x}}))(\mathbf{x}_i - \bar{\mathbf{x}})$. Then, under general assumptions and using Taylor's expansion, we get

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{T}}_k - \lambda_{g,k}\mathbf{e}_k) &= \sqrt{n}\mathbf{T}_k - \tau_{g,k}\mathbf{e}_k\mathbf{e}_k^T\sqrt{n}\bar{\mathbf{x}} \\ &+ \Delta_{g,k}\sqrt{n}(\hat{\mathbf{w}}_k - \mathbf{e}_k) + o_P(1), \end{aligned} \tag{3}$$

where $\Delta_{g,k} = \mathrm{E}[g'(\mathbf{e}_k^T \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T]$.
Now recall that the FastICA unmixing matrix estimator $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_p)^T$ needs to verify the estimating equations

$$\hat{\mathbf{T}}_k = \hat{\mathbf{S}}[\hat{\mathbf{w}}_1\hat{\mathbf{w}}_1^T + ... + \hat{\mathbf{w}}_k\hat{\mathbf{w}}_k^T]\hat{\mathbf{T}}_k, \quad k = 1, \ldots, p. \tag{4}$$

But then

$$(\mathbf{I}_p - \mathbf{U}_k)\sqrt{n}(\hat{\mathbf{T}}_k - \lambda_{g,k}\mathbf{e}_k) = \lambda_{g,k}[\sqrt{n}(\hat{\mathbf{S}} - \mathbf{I}_p)\mathbf{e}_k + \sum_{j=1}^{k} \mathbf{e}_j\mathbf{e}_k^T\sqrt{n}(\hat{\mathbf{w}}_j - \mathbf{e}_j) + \sqrt{n}(\hat{\mathbf{w}}_k - \mathbf{e}_k)] + o_P(1),$$

where $\mathbf{U}_k = \sum_{j=1}^{k} \mathbf{e}_j\mathbf{e}_j^T$, and, using (3), we get the following result.

Theorem 1. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from the IC model (1) with $\mathbf{A} = \mathbf{I}_p$, $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$, and $\mathrm{COV}(\mathbf{x}_i) = \mathbf{I}_p$. Let $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_p)^T$ be the solution for the estimating equations in (4) such that $\hat{\mathbf{W}} \to_P \mathbf{I}_p$. Then, under the general assumptions,*

$$\begin{aligned}\sqrt{n}\hat{\mathbf{w}}_{kl} &= \frac{1}{\lambda_{g,k} - \delta_{g,k}}\left[\mathbf{e}_l^T\sqrt{n}\mathbf{T}_k - \lambda_{g,k}\sqrt{n}\hat{\mathbf{S}}_{kl}\right] \\ &+ o_P(1) \quad \textit{for } l > k, \\ \sqrt{n}\hat{\mathbf{w}}_{kl} &= -\sqrt{n}\hat{\mathbf{w}}_{lk} - \sqrt{n}\hat{\mathbf{S}}_{kl} + o_P(1) \quad \textit{for } l < k,\end{aligned}$$

and

$$\sqrt{n}(\hat{\mathbf{w}}_{kk} - 1) = -\frac{1}{2}\sqrt{n}(\hat{\mathbf{S}}_{kk} - 1) + o_P(1).$$

Remark 1. *It follows from Theorem 1 that, for $\mathbf{A} = \mathbf{I}_p$, the asymptotic covariance matrix (ASV) of the k-th source $\hat{\mathbf{w}}_k$ is*

$$ASV(\hat{\mathbf{w}}_k) = \sum_{j=1}^{k-1}(\alpha_{g,j} + 1)\mathbf{e}_j\mathbf{e}_j^T + \kappa_k\mathbf{e}_k\mathbf{e}_k^T + \alpha_{g,k}\sum_{l=k+1}^{p}\mathbf{e}_l\mathbf{e}_l^T.$$

where $\kappa_k = (\mathrm{E}(x_{ik}^4) - 1)/4$ and $\alpha_{g,j}$ is defined in (2)*. We note that this result is in accordance with [12, Corollary 1]. Note that the asymptotic variances of the diagonal elements of $\hat{\mathbf{W}}$ do not depend on the choice of the function $g(\cdot)$, but only on the kurtosis of the corresponding source.*

Remark 2. *Theorem 1 implies that, if $\sqrt{n}\mathbf{T}_k$, $k = 1, ..., p$, and $\sqrt{n}vec(\hat{\mathbf{S}} - \mathbf{I}_p)$ have a joint limiting multivariate distribution, the limiting distribution of $\sqrt{n}vec(\hat{\mathbf{W}} - \mathbf{I}_p)$ is also multivariate normal. Interestingly, the limiting distributions of the estimated directions $\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_p$ depend on the order in which they are found; see [12] for details and illustrations. The initial value $\mathbf{U}_{init}$ in the FastICA algorithm mainly determines the order of the extracted sources in practice and hence plays a crucial role in the performance of the estimator.*

3.2 Robustness

Due to the different options for the nonlinearity function $g(\cdot)$, FastICA is often called robust when used with 'robust' nonlinearity functions, for example, *tanh* or *gaus* function. The influence function (IF) of the FastICA functional $\mathbf{w}_k$, $k = 1, \ldots, p$, in the IC model (1) is given in [12] as

$$\mathrm{IF}(\mathbf{z}; \mathbf{w}_k, F) = -p_k\sum_{j=1}^{k-1}(q_j + p_j)\mathbf{w}_j - \frac{p_k^2 - 1}{2}\mathbf{w}_k + q_k\sum_{l=k+1}^{p} p_l\mathbf{w}_l,$$

where $p_k = \mathbf{w}_k^T(\mathbf{z} - \mathrm{E}(\mathbf{x}))$ and

$$q_k = \frac{g(p_k) - \mu_{g,k} - \lambda_{g,k}p_k}{\lambda_{g,k} - \delta_{g,k}}.$$

Since the IF is a weighted sum of the sources $\mathbf{w}_1, \ldots, \mathbf{w}_p$, where the weights are unbounded functions of p_k, any large value of p_j, $j = 1, \ldots, p$ can have unbounded impact on $\mathbf{w}_k$ - irrelevant of the choice of the nonlinearity $g(\cdot)$. Thus, according to its IF, the deflation-based FastICA will never be robust - independently of the choice of $g(\cdot)$ (see [12] for details).

Note also that it is not straightforward to robustify deflation-based FastICA by replacing mean vector and covariance matrix with their more robust counterparts as reported in [1].

4. RELOADING FASTICA BY OPTIMIZING THE EXTRACTION ORDER

In this section, we first discuss the properties of the performance index *MD* for the ICA estimates, and show how it is connected to the asymptotic distribution of the estimate. We then suggest a two-step modified FastICA procedure which optimizes the extraction order.

4.1 Minimum distance performance criterion

Many different performance measures for the IC estimates have been suggested in the literature, see, for example, [10]. In this paper we use the so called minimum distance (*MD*) measure which was recently suggested in [8,9]. The measure is defined as

$$MD(\hat{\mathbf{W}}, \mathbf{A}) = \frac{1}{\sqrt{p-1}}\inf_{\mathbf{C}\in\mathscr{C}}\|\mathbf{C}\hat{\mathbf{W}}\mathbf{A} - \mathbf{I}_p\|.$$

This index is independent of the model specification and surprisingly easy to compute in practice (for details see [8, 9]). The asymptotic behavior of the index *MD* is as follows. If an equivariant estimator $\hat{\mathbf{W}}$ satisfies $\sqrt{n}\mathrm{vec}(\hat{\mathbf{W}} - \mathbf{I}_p) \to_d N_{p^2}(0, \Sigma)$, then

$$nMD^2(\hat{\mathbf{W}}, \mathbf{A}) = \frac{n}{p-1}\|\mathrm{off}(\hat{\mathbf{W}})\|^2 + o_P(1),$$

and the limiting distribution of $nMD^2(\hat{\mathbf{W}}, \mathbf{A})$ is that of a weighted sum of independent chi-square variables [9]. Also, the expected value $n(p-1)\mathrm{E}[MD^2(\hat{\mathbf{W}}, \mathbf{A})]$ converges to the sum of the limiting variances of the off-diagonal elements of $\hat{\mathbf{W}}$ as $n \to \infty$.

4.2 Reloaded FastICA

In order to achieve optimal performance in terms of the *MD* measure, we thus should minimize the sum of the variances of the off-diagonal elements of the FastICA estimator. Using Remark 1 it is easy to see that, for $\mathbf{A} = \mathbf{I}_p$,

$$\sum_{i\neq j} ASV(\hat{\mathbf{w}}_{ij}) = 2\sum_{i=1}^{p}(p-i)\alpha_{g,i} + \frac{p(p-1)}{2},$$

which is minimized if the $\alpha_{g,i}$'s are in the increasing order of magnitude.

To optimize the performance of the deflation-based FastICA, we therefore suggest the following simple procedure.

$g(\cdot)$	$\alpha_{g,E}$	$\alpha_{g,C}$	$\alpha_{g,L}$
pow3	5	15	6
tanh	3.14	32.13	2.01

Table 1: The theoretical values of $\alpha_{g,k}$ for different cases.

$g(\cdot)$	LCE	LEC	CEL	ECL	CLE	ELC
pow3	57	37	73	53	75	35
tanh	75.32	17.33	137.79	79.80	135.55	19.57

Table 2: The limiting values of $n(p-1)\mathrm{E}[MD^2(\hat{\mathbf{W}},\mathbf{A})]$ for the six different extraction orders.

1. Find any equivariant and consistent estimate $\hat{\mathbf{W}}_0$ (e.g. FOBI [2]) such that $\hat{\mathbf{S}}(\hat{\mathbf{W}}_0\mathbf{X}) = \mathbf{I}_p$.
2. Find the estimated sources $\hat{\mathbf{Z}} = \hat{\mathbf{W}}_0(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)$.
3. Find estimates $\hat{\alpha}_{g,k}$, $k = 1,\ldots,p$, based on $\hat{\mathbf{Z}}$ by replacing the expected values by averages in (2).
4. Find the permutation matrix $\hat{\mathbf{P}}$ such that, for the permutated sources, the $\hat{\alpha}_{g,k}$ are in an increasing order.
5. Reload FastICA algorithm 1 with a new initial value: The estimate is $\mathbf{W}(\mathbf{U}(\mathbf{X}),\mathbf{X})$ where $\mathbf{U}(\mathbf{X}) = \hat{\mathbf{P}}\hat{\mathbf{W}}_0\hat{\mathbf{S}}^{1/2}$.

It is easy to see that $\mathbf{W}(\mathbf{U}(\mathbf{X}),\mathbf{X})$ is fully affine equivariant. We conjecture that this new estimator has the same limiting distribution as the simple FastICA estimator which extracts the sources in the (same) optimal order.

5. SIMULATION STUDY

We performed a small simulation study to demonstrate the effect of the extraction order of the sources. We show that reloading FastICA with the data whitened in a new way and with an initial value $\mathbf{U}_{init} = \mathbf{I}_p$ gives the optimal performance among different deflation-based FastICA procedures. The data used in our simulations comes from a three-variate distribution; the independent source distributions are (i) the exponential distribution, (ii) the chi-square distribution with 8 degrees of freedom, and (iii) the Laplace distribution. All three distributions are centered and scaled to have expected value 0 and variance 1. The mixing matrix used in our simulations is $\mathbf{A} = \mathbf{I}_3$. We denote the three sources as E, C, and L, respectively, and the sequence ECL, for example, means the extraction order exponential-chi-square-Laplace. We considered two nonlinearity functions $g = pow3$ and $g = tanh$. The values of corresponding $\alpha_{g,k}$, given in Table 1, were obtained from (2), where the expectations were calculated using numerical integration.

The expected values of $n(p-1)\mathrm{E}[MD^2(\hat{\mathbf{W}},\mathbf{A})]$ for different extraction orders are given in Table 2. The table clearly shows that the extraction order has a large impact on the separation performance. The best extraction order naturally depends on the choice of the nonlinearity function g. Here ELC is the best order for *pow3*, whereas LEC is the best for *tanh* .

To see whether the expected behavior is observed in finite sample sizes we repeated the estimation of the unmixing matrix 5000 times for different sample sizes using all six possible extraction orders for both nonlinearities. The extraction order can be controlled using six different 3×3 permutation matrices $\mathbf{P}$ as initial values $\mathbf{U}_{init}$. For the reloaded deflation-based FastICA we chose FOBI [2] as the initial estimate. The FOBI functional is an affine equivariant IC functional, and the limiting distribution of the unmixing matrix estimate is known to be multivariate normal [7]. FOBI has the advantage that it is easy to compute, and, unlike FastICA, it always gives a solution. In this simulation study we included the FastICA estimators using random initial values as well. Then the extraction order is also random, and hence the performance is expected to be a mixture of the performances of the six possible estimators with different (fixed) extraction orders.

We used the FastICA code [3] for Algorithm 1, and we retained all the default settings except the initial value. One problem worth mentioning is that, unfortunately, the algorithm does not always converge. In applied data analysis the user may be able to change some tuning parameters in order to obtain a solution. However, this is not feasible in a simulation study. In our simulations, we simply ignored the cases when convergence did not occur. (Another option would have been to set the *MD* values to 1 in these cases.)

n	ECL	LCE	CEL	ELC	LEC	CLE	rand	reloaded
1000	20	24	27	0	0	25	5	0
5000	0	0	0	0	0	0	0	0
10000	0	0	0	0	0	0	0	0
$\geq$25000	0	0	0	0	0	0	0	0

Table 3: Number of algorithm failures in 5000 trials for *pow3*.

n	ECL	LCE	CEL	ELC	LEC	CLE	rand	reloaded
1000	340	472	493	0	0	457	145	0
5000	12	79	71	0	0	71	11	0
10000	1	13	10	0	0	10	4	0
$\geq$25000	0	0	0	0	0	0	0	0

Table 4: Number of algorithm failures in 5000 trials for *tanh*.

Table 3 and Table 4 give the number of cases when the algorithm did not converge. These figures clearly illustrate that for small sample sizes the algorithm often fails to converge for the given initial matrix. The problem is more severe in case of *tanh* nonlinearity. However, reloading FastICA seems to help the algorithm to find a solution.

Figure 1 presents the plots of the average values of $n(p-1)MD^2(\hat{\mathbf{W}},\mathbf{A})$ over the sample size n. The black lines in the figure give the results for the deflation-based FastICA with fixed extraction order, and the horizontal lines represent the asymptotic expectations given in Table 2. While for *pow3* convergence is reached quickly, this is not the case for *tanh*. The worse the performance, the slower the convergence seems to be. The performance of the deflation-based FastICA with random initial matrix is somewhere between the optimal and the worst possible case, which supports our conjecture of being a mixture of the six different cases. The strange behavior at the large sample sizes when using *pow3* may be due to the fact that the algorithm often converges to a wrong local maxima. It is clear that the average *MD* of the reloaded FastICA corresponds to the minimum value among the six possible cases. Therefore, the reloaded FastICA behaves as expected and is basically equivalent with the best extraction order for that given nonlinearity.

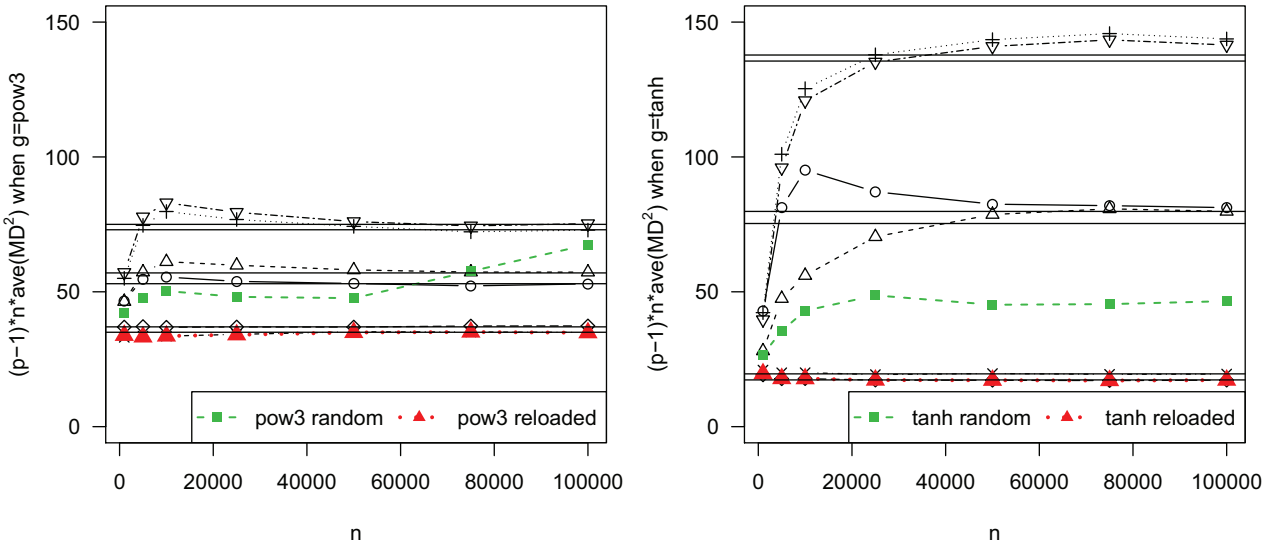

Figure 1: Average performance of the reloaded FastICA and the deflation-based FastICA based on a random initial value. The black curves give the performance of deflation based-FastICA when the extraction order is fixed. Horizontal black lines are asymptotic expectations given in Table 2.

6. CONCLUSIONS

In this paper we reviewed some properties of the deflation-based FastICA. One important curious property of FastICA is that the extraction order has a huge impact on the separation performance. We used this property and suggested the use of the reloaded FastICA to achieve the optimal extraction order. In our approach, we first need to run some ICA procedure that provides a consistent and affine equivariant unmixing matrix estimate. Then the extracted sources are permuted based on the nonlinearity used, and finally the regular deflation-based FastICA is performed using the estimated and permuted sources as whitened data and the identity matrix as an initial value of the rotation matrix. Reloading FastICA this way yields the best extraction order and renders the algorithm more stable at small sample sizes as validated by our simulation studies.

Future research is needed to derive the asymptotic properties of the reloaded FastICA estimator. Above all, more research is needed to derive the optimal choice of the nonlinearity function as well.

REFERENCES

[1] G. Brys, M. Hubert, and P.J. Rousseeuw, "A robustification of independent component analysis". *Chemometrics*, vol. 57 , pp. 364–375, 2006.

[2] J. Cardoso, "Source separation using higher moments," in *Proc. IEEE International Conference on Acustics, Speech and Signal Processing*, Glasgow, 1989, pp. 2109–2112.

[3] *http://www.cis.hut.fi/projects/ica/fastica*.

[4] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, pp.1483-1492, 1997.

[5] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, 1999.

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

[7] P. Ilmonen, J. Nevalainen, H. Oja, "Characteristics of multivariate distributions and the invariant coordinate system," *Statistics and Probability Letters*, vol. 80, pp. 1844–1853, 2010.

[8] P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila, "A new performance index for ICA: properties, computation and asymptotic analysis," in *Latent Variable Analysis and Signal Processing (Proceedings of 9th International Conference on Latent Variable Analysis and Signal Separation)*. 2010, pp. 229–236.

[9] P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila, "Independent component (IC) functionals and a new performance index", submitted.

[10] K. Nordhausen, E. Ollila and H. Oja, "On the performance indices of ICA and blind source separation," in *Proc. IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2011)*, 2011, pp. 471–475.

[11] E. Ollila, "On the robustness of the deflation-based FastICA estimator," in *Proc. IEEE Workshop on Statistical Signal Processing (SSP'09)*, Cardiff, Wales, Aug. 31–Sep. 3. 2009, pp. 673–676.

[12] E. Ollila, "The deflation-based FastICA estimator: statistical analysis revisited," *IEEE Trans. Signal Processing*, vol. 58, pp. 1527–1541, 2010.

[13] E. Ollila and H.-J. Kim, "'On testing hypotheses of mixing vectors in the ICA model using FastICA," in *Proc. IEEE Int. Symp. on Biomedical Imaging* (ISBI'11), Chicago, USA, Mar. 30 – Apr. 2, 2011, pp. 325-328.