Deflation-based FastICA with adaptive choices of nonlinearities

Jari Miettinen, Klaus Nordhausen, Hannu Oja and Sara Taskinen

***Abstract*—Deflation-based FastICA is a popular method for independent component analysis. In the standard deflation-based approach the row vectors of the unmixing matrix are extracted one after another always using the same nonlinearities. In practice the user has to choose the nonlinearities and the efficiency and robustness of the estimation procedure then strongly depends on this choice as well as on the order in which the components are extracted. In this paper we propose a novel adaptive two-stage deflation-based FastICA algorithm that (i) allows one to use different nonlinearities for different components and (ii) optimizes the order in which the components are extracted. Based on a consistent preliminary unmixing matrix estimate and our theoretical results, the algorithm selects in an optimal way the order and the nonlinearities for each component from a finite set of candidates specified by the user. It is also shown that, for each component, the best possible nonlinearity is obtained by using the log-density function. The resulting ICA estimate is affine equivariant with a known asymptotic distribution. The excellent performance of the new procedure is shown with asymptotic efficiency and finite-sample simulation studies.**

***Index Terms*—Independent component analysis, minimum distance index, asymptotic normality, affine equivariance**

EDICS: SSP-SSEP

I. Introduction

An observable p-variate real-valued random vector $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ obeys the basic independent component (IC) model, if it is a linear mixture of p mutually independent latent random variables in $\boldsymbol{z} = (z_1, \ldots, z_p)^T$. The latent variables $z_1, \ldots, z_p$ are also called sources and it is assumed that at most one of the sources is gaussian. We then write

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z},$$

and assume, for simplicity, that $\boldsymbol{A}$ is a full rank $p \times p$ mixing matrix. In this model specification, $\boldsymbol{A}$ and $\boldsymbol{z}$ are confounded, and the mixing matrix $\boldsymbol{A}$ can be identified only up to the order, the signs and heterogenous multiplications of its columns. One can however accept the ambiguity in the model, and define the concepts and analysis tools so that they are independent of the model specification. A $p \times p$ matrix $\boldsymbol{W}$ is called an unmixing matrix in the IC model, if the components of $\boldsymbol{W}\boldsymbol{x}$ are independent. Then all the unmixing matrices $\boldsymbol{W}$ satisfy $\boldsymbol{W} = \boldsymbol{C}\boldsymbol{A}^{-1}$ where $\boldsymbol{C}$ is in the set of $p \times p$ matrices

$$\mathcal{C} = \{\boldsymbol{C} \ : \ \text{each row and column of } \boldsymbol{C} \text{ has exactly one non-zero element.}\}.$$

We also write $\boldsymbol{W}_1 \sim \boldsymbol{W}_2$, if $\boldsymbol{W}_1 = \boldsymbol{C}\boldsymbol{W}_2$ for some $\boldsymbol{C} \in \mathcal{C}$.

We now give a formal definition of a $p \times p$ matrix valued IC functional $\boldsymbol{W}(F)$ such that, for an $\boldsymbol{x}$ coming from the IC model, the components of $\boldsymbol{W}(F_{\boldsymbol{x}})\boldsymbol{x}$ are independent and do not depend on the model specification and the value of $\boldsymbol{A}$ at all.

Definition 1. *Let $F_{\boldsymbol{x}}$ denote the cumulative distribution function of $\boldsymbol{x}$. The functional $\boldsymbol{W}(F_{\boldsymbol{x}})$ is an independent component (IC) functional if (i) $\boldsymbol{W}(F_{\boldsymbol{z}}) \sim \boldsymbol{I}_p$ for any $\boldsymbol{z}$ with independent components and with at most one gaussian component, and (ii) $\boldsymbol{W}(F)$ is affine equivariant in the sense that $\boldsymbol{W}(F_{\boldsymbol{B}\boldsymbol{x}}) \sim \boldsymbol{W}(F_{\boldsymbol{x}})\boldsymbol{B}^{-1}$ for all full rank $p \times p$ matrices $\boldsymbol{B}$.*

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a random sample from a distribution of $\boldsymbol{x}$. An estimate of $\boldsymbol{W} = \boldsymbol{W}(F_{\boldsymbol{x}})$ is obtained if the IC functional is applied to the empirical distribution F_n of $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Then, according to our Definition 1, (ii) is true for F_n as well, and the resulting estimator $\boldsymbol{W}(\boldsymbol{X})$ is affine equivariant in the sense that $\boldsymbol{W}(\boldsymbol{B}\boldsymbol{X}) \sim \boldsymbol{W}(\boldsymbol{X})\boldsymbol{B}^{-1}$.

Write $\boldsymbol{S}(F_{\boldsymbol{x}})$ for the covariance matrix functional of a random vector $\boldsymbol{x}$. The scale ambiguity of the IC functional is usually fixed by requiring that the obtained independent components have unit variances, that is, $\boldsymbol{S}(F_{\boldsymbol{W}(F_{\boldsymbol{x}})\boldsymbol{x}}) = \boldsymbol{I}_p$. This then implies that $\boldsymbol{W}(F_{\boldsymbol{x}}) = \boldsymbol{U}(F_{\boldsymbol{x}})\boldsymbol{S}^{-1/2}(F_{\boldsymbol{x}})$ for some orthogonal matrix $\boldsymbol{U}(F_{\boldsymbol{x}})$, and the estimation problem is reduced to the estimation of an orthogonal matrix $\boldsymbol{U}(F_{\boldsymbol{x}})$. The sample statistic then similarly satisfies $\boldsymbol{W}(\boldsymbol{X}) = \boldsymbol{U}(\boldsymbol{X})\boldsymbol{S}^{-1/2}(\boldsymbol{X})$ for some orthogonal matrix $\boldsymbol{U}(\boldsymbol{X})$.

The outline of this paper is the following. In Section II the classical deflation-based FastICA algorithm and its dependence on the initial value is first discussed. Then the algorithm and estimating equations for a modified deflation-based FastICA procedure that allows us to use different nonlinearity functions for different sources are introduced. The nonlinearities used for the illustration are introduced as well. Section III presents the statistical properties of the new procedure. Based on these theoretical results we further extend in Section IV our procedure by allowing the sources to be found in an optimal order. Finally, the excellent performance of the new estimation procedure is verified by some simulation studies in Section V.

J. Miettinen and S. Taskinen are with the Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, FIN-40014, Finland (e-mail: jari.p.miettinen@jyu.fi).

K. Nordhausen and H. Oja are with the Department of Mathematics and Statistics, University of Turku, Turku, FIN-20014, Finland.

II. Deflation-based FastICA

A. Algorithm

Originally deflation-based FastICA [4] was introduced as an algorithm, which, like many other ICA methods, begins with the whitening of the data. Let $\boldsymbol{x}_{st} = \boldsymbol{S}^{-1/2}(F_{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{\mu}(F_{\boldsymbol{x}}))$ be the data whitened using the mean vector $\boldsymbol{\mu}(F_{\boldsymbol{x}})$ and the covariance matrix $\boldsymbol{S}(F_{\boldsymbol{x}})$. The second step is to find an orthogonal matrix $\boldsymbol{U}$ such that $\boldsymbol{U}\boldsymbol{x}_{st}$ has independent components. In deflation-based FastICA the rows of the matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)^T$ are found one by one, so that $\boldsymbol{u}_k$ maximizes a measure of non-Gaussianity $|\mathrm{E}(G(\boldsymbol{u}_k^T\boldsymbol{x}_{st}))|$ under the constraint that $\boldsymbol{u}_k$ has the length one and that it is orthogonal to rows $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$. The function G is allowed to be any twice continuously differentiable nonquadratic function with $G(0) = 0$ and with first and second derivative functions g and g'. For more details, see Section II-C

To find an estimate, the observed data $\boldsymbol{X}$ are first whitened with the sample mean vector $\boldsymbol{\mu}(\boldsymbol{X})$ and the sample covariance matrix $\boldsymbol{S}(\boldsymbol{X})$. The second step is to find the rotation matrix $\boldsymbol{U}(\boldsymbol{X})$ and the final estimate $\boldsymbol{W}(\boldsymbol{X}) = \boldsymbol{U}(\boldsymbol{X})\boldsymbol{S}^{-1/2}(\boldsymbol{X})$. The so called deflation-based FastICA algorithm for finding the rows of $\boldsymbol{U}(\boldsymbol{X})$ one-by-one is given in [4]. The value $\boldsymbol{U}(\boldsymbol{X})$ provided by the algorithm however depends on the initial value $\boldsymbol{U}_{init} = (\boldsymbol{u}_{init,1}, \ldots, \boldsymbol{u}_{init,p})^T$ used for the computation. Depending on where one starts from, the algorithm may stop at any critical point instead of the global maximum. It is then remarkable that extracting the sources in a different order changes the unmixing matrix estimate more than just the permutation [17], [15]. To be precise in our notation, we therefore write $\boldsymbol{W}(\boldsymbol{U}, \boldsymbol{X}; g)$ for the estimate that is provided by the FastICA algorithm for the data $\boldsymbol{X}$ with the initial value $\boldsymbol{U}_{init} = \boldsymbol{U}$ and the nonlinearity g. Estimates such as $\boldsymbol{W}(\boldsymbol{I}_p, \boldsymbol{X}; g)$ and $\boldsymbol{W}(\boldsymbol{U}_{rand}, \boldsymbol{X}; g)$ with a random orthogonal $\boldsymbol{U}_{rand}$ are often used in practice. Unfortunately, these estimates are not affine equivariant and therefore not IC functionals in the sense of Definition 1. Also, the common practise to use the deflation-based approach to extract only few first components with such fixed or random initial values seems in this light highly questionable.

To obtain an affine equivariant estimator the initial value must be data-based and affine equivariant as well. A preliminary unmixing matrix estimate may be used here as shown by the following lemma.

Lemma 1. *Let $\boldsymbol{W}_0(\boldsymbol{X}) = \boldsymbol{U}_0(\boldsymbol{X})\boldsymbol{S}^{-1/2}(\boldsymbol{X})$ be an IC functional satisfying $\boldsymbol{W}_0(\boldsymbol{B}\boldsymbol{X}) = \boldsymbol{W}_0(\boldsymbol{X})\boldsymbol{B}^{-1}$ for all full-rank $p \times p$ matrices $\boldsymbol{B}$. Then $\boldsymbol{W}(\boldsymbol{U}_0(\boldsymbol{X}), \boldsymbol{X}; g)$ satisfies*

$$\boldsymbol{W}(\boldsymbol{U}_0(\boldsymbol{B}\boldsymbol{X}), \boldsymbol{B}\boldsymbol{X}; g) = \boldsymbol{W}(\boldsymbol{U}_0(\boldsymbol{X}), \boldsymbol{X}; g)\boldsymbol{B}^{-1}$$

for all full-rank $p \times p$ matrices $\boldsymbol{B}$.

The proof follows from the fact that $\boldsymbol{U}_0(\boldsymbol{B}\boldsymbol{X}) = \boldsymbol{U}_0(\boldsymbol{X})\boldsymbol{V}^T$ and $\boldsymbol{S}^{-1/2}(\boldsymbol{B}\boldsymbol{X})\boldsymbol{B}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{S}^{-1/2}(\boldsymbol{X})\boldsymbol{X}$ with an orthogonal $\boldsymbol{V} = \boldsymbol{S}^{-1/2}(\boldsymbol{B}\boldsymbol{X})\boldsymbol{B}\boldsymbol{S}^{1/2}(\boldsymbol{X})$. Thus the transformation $\boldsymbol{X} \to \boldsymbol{B}\boldsymbol{X}$ induces in the algorithm the transformations $\boldsymbol{x}_i \to \boldsymbol{V}\boldsymbol{x}_i$, $i = 1, \ldots, n$, and $\boldsymbol{u}_k \to \boldsymbol{V}\boldsymbol{u}_k$, $k = 1, \ldots, p$, and finally $\boldsymbol{U}(\boldsymbol{X}) \to \boldsymbol{U}(\boldsymbol{X})\boldsymbol{V}^T$. Thus $\boldsymbol{U}(\boldsymbol{B}\boldsymbol{X})\boldsymbol{S}^{-1/2}(\boldsymbol{B}\boldsymbol{X})\boldsymbol{B}\boldsymbol{X} = \boldsymbol{U}(\boldsymbol{X})\boldsymbol{S}^{-1/2}(\boldsymbol{X})\boldsymbol{X}$ and the result follows.

B. Estimating equations

The deflation-based FastICA algorithm discussed in the previous section aims, for $\boldsymbol{u}_k$, $k = 1, \ldots, p-1$, to maximize $|\mathrm{E}(G(\boldsymbol{u}_k^T\boldsymbol{x}_{st}))|$ under the constraint that $\boldsymbol{u}_k^T\boldsymbol{u}_k = 1$ and $\boldsymbol{u}_j^T\boldsymbol{u}_k = 0$, $j = 1, \ldots, k-1$. Consider next the modification of the FastICA procedure that, for $\boldsymbol{u}_k$, $k = 1, \ldots, p-1$, maximizes $|\mathrm{E}(G_k(\boldsymbol{u}_k^T\boldsymbol{x}_{st}))|$ under the constraint that $\boldsymbol{u}_k^T\boldsymbol{u}_k = 1$ and $\boldsymbol{u}_j^T\boldsymbol{u}_k = 0$, $j = 1, \ldots, k-1$. We thus allow that the nonlinearity functions may be different for different components. If $g_k = G_k'$, $k = 1, \ldots, p$, we obtain, using similar Lagrange multiplier technique as in [17], the estimating equations

$$\mathrm{E}[g_k(z_k)z_k]\boldsymbol{u}_k = \left(\boldsymbol{I}_p - \sum_{j=1}^{k-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\right)\mathrm{E}[g_k(z_k)\boldsymbol{x}_{st}],$$

$k = 1, \ldots, p$, and we give the following definition.

Definition 2. *Let $\boldsymbol{U}(\boldsymbol{X})$ be the solution where, after finding $\boldsymbol{u}_1, ..., \boldsymbol{u}_{k-1}$, $\boldsymbol{u}_k$ is found from a fixed point algorithm with the steps*

$$\begin{aligned} z_k &\leftarrow \boldsymbol{u}_k^T\boldsymbol{x}_{st} \quad \textit{and} \\ \boldsymbol{u}_k &\leftarrow \frac{1}{E[g_k(z_k)z_k]}\left(\boldsymbol{I}_p - \sum_{j=1}^{k-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\right)E[g_k(z_k)\boldsymbol{x}_{st}], \end{aligned}$$

and the initial value $\boldsymbol{U}_{init}$ is used in the computation. Then the modified deflation-based FastICA estimator is defined as

$$\boldsymbol{W}_m(\boldsymbol{U}_{init}, \boldsymbol{X}; g_1, \ldots, g_p) = \boldsymbol{U}(\boldsymbol{X})\boldsymbol{S}^{-1/2}(\boldsymbol{X}).$$

It is worth noticing that the estimated equations above do not fix the order of the independent components in the IC model: If $\boldsymbol{U}$ is a solution then so is $\boldsymbol{P}\boldsymbol{U}$ for all permutation matrices $\boldsymbol{P}$. This also means that the algorithm that is solely based on the estimating equations extracts the estimated sources in the order suggested by the initial value $\boldsymbol{U}_{init}$. Also, the rows of $\boldsymbol{U}(\boldsymbol{X})$ are then again changed more than just permuted.

C. Nonlinearities

The derivative function $g = G'$ is called the nonlinearity. Using the classical kurtosis measure as an optimizing criterion gives the nonlinearity $g(z) = z^3$ (*pow3*), see [4]. Functions $g(z) = \tanh(az)$ (*tanh*) and $g(z) = z\exp(-az^2/2)$ (*gaus*) with tuning parameters a were suggested in [5]. The classical skewness measure gives $g(z) = z^2$ (*skew*). The FastICA algorithm thus uses the same nonlinearity for all components with certain general guidelines for its choice. The nonlinearity *pow3*, for example, is considered efficient for sources with light-tailed distributions, whereas *tanh* and *gaus* are preferable for heavy-tailed sources. The nonlinearity *skew* finds skew sources but fails in the case of symmetric sources. In practice, *tanh* and *gaus* seem to be common choices. We will later prove the well-known fact that, for a component with the density function $f(z)$, the function corresponding to $G(z) = \log f(z)$

with the nonlinearity $g(z) = -f'(z)/f(z)$ provides an optimal choice.

In practical data analysis, however, it does not seem likely that all sources are either light-tailed, heavy-tailed or skew or even that the knowledge about these properties is available. Therefore, the use of only a single nonlinearity g for all different components seems questionable. In this paper we propose a novel algorithm that (i) allows the use of different nonlinearities for different components and (ii) optimizes the order in which the components are extracted. The nonlinearities $g_1, \ldots, g_k$ are selected from a large but finite set of nonlinearities $\mathcal{G}$. In the illustration of our theory, we later choose $\mathcal{G}$ with the four popular nonlinearities mentioned above, namely, *pow3*, *skew*, *tanh* and *gaus*, and the functions

- $g(z) = (z+a)_-^2$ (*left*)
- $g(z) = (z-a)_+^2$ (*right*)
- $g(z) = (z-a)_+^2 - (z+a)_-^2$ (*both*)

with different choices of tuning parameter $a > 0$. The functions *left* and *right* seem useful for extracting skewed sources whereas *both* provides an alternative measure of tail weight (kurtosis). Note that these new functions are simply used to enrich the set $\mathcal{G}$ with different types of nonlinearities for our new estimator in Section IV but may fail, if used alone in traditional deflation-based FastICA.

We end this discussion about the choice of the nonlinearity function with a short note on robustness. As the random vector is first standardized by the regular covariance matrix, the influence function of the functional $\boldsymbol{W}(F)$ is unbounded for any choice of the nonlinearity g and, unfortunately, the FastICA functional is not robust in this sense. See [17].

III. Asymptotics

The statistical properties of the deflation-based FastICA estimator were rigorously discussed only recently in [15], [16] and [17]. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a random sample from a distribution of $\boldsymbol{x}$ obeying the IC model. Thus $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}$ where $\mathrm{E}(\boldsymbol{z}) = \boldsymbol{0}$, $\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{I}_p$ and the components $z_1, \ldots, z_p$ of $\boldsymbol{z}$ are independent. As our unmixing matrix estimate is affine equivariant, we can then assume (wlog) that $\boldsymbol{A} = \boldsymbol{I}_p$ and $\boldsymbol{b} = \boldsymbol{0}$, that is, $\boldsymbol{X} = \boldsymbol{Z}$. In the following, for a sequence of random variables $\boldsymbol{T}_n$, we write in a regular way (i) $\boldsymbol{T}_n = O_P(1)$ if, for all $\epsilon > 0$, there exists an $M_\epsilon > 0$ and N_ϵ such that $\mathbb{P}(||\boldsymbol{T}_n|| > M_\epsilon) \leq \epsilon$ for all $n \geq N_\epsilon$, and (ii) $\boldsymbol{T}_n = o_P(1)$ if $T_n \to_P 0$.

For the asymptotic distribution of the extended FastICA estimator satisfying the estimating equations (2) we need the following assumption. We assume the existence of fourth moments $E(z_k^4)$ and the following expected values

$$\mu_{g_k,k} = \mathrm{E}[g_k(z_k)], \quad \sigma^2_{g_k,k} = \mathrm{Var}[g_k(z_k)],$$
$$\lambda_{g_k,k} = \mathrm{E}[g_k(z_k)z_k] \text{ and}$$
$$\delta_{g_k,k} = \mathrm{E}[g_k'(z_k)] = \mathrm{E}[g_k(z_k)g_{0k}(z_k)],$$

where $g_{0k} = -f_k'/f_k$ is the optimal location score function for the density f_k of z_k, $k = 1, \ldots, p$. We assume that $\delta_{g_k,k} \neq \lambda_{g_k,k}$, $k = 1, \ldots, p-1$. Note that, if the kth component is gaussian, then $\delta_{g_k,k} = \lambda_{g_k,k}$ for all choices of g_k.

The asymptotic distribution $\sqrt{n}\,(\boldsymbol{W}(\boldsymbol{Z}) - \boldsymbol{I}_p)$ is then easily obtained if we only know the joint asymptotic distribution of $\sqrt{n}\,(\boldsymbol{S}(\boldsymbol{Z}) - \boldsymbol{I}_p)$ and $\sqrt{n}\,(\boldsymbol{T}(\boldsymbol{Z}) - \boldsymbol{\Lambda})$, where

$$(\boldsymbol{T}(\boldsymbol{Z}))_{kl} = \frac{1}{n}\sum_{i=1}^{n}(g_k(z_{ik}) - \mu_{g_k,k})\boldsymbol{z}_{il}$$

and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_{k_1,1}, \ldots, \lambda_{k_p,p})$. Write $(\boldsymbol{S}(\boldsymbol{Z}))_{kl} = s_{kl}$, $(\boldsymbol{T}(\boldsymbol{Z}))_{kl} = t_{kl}$ and $(\boldsymbol{W}(\boldsymbol{Z}))_{kl} = w_{kl}$. Under our assumptions,

$$\sqrt{n}\,(\boldsymbol{S}(\boldsymbol{Z}) - \boldsymbol{I}_p) = O_P(1) \text{ and } \sqrt{n}\,(\boldsymbol{T}(\boldsymbol{Z}) - \boldsymbol{\Lambda}) = O_P(1).$$

The following theorem extends Theorem 1 in [15], allowing different nonlinearity functions for different source components. The proof is similar to the proof in [15].

Theorem 1. *Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ be a random sample from a distribution of $\boldsymbol{z}$ satisfying the assumptions stated above. Assume also that $\boldsymbol{W} = \boldsymbol{W}_m(\boldsymbol{U}_{init}, \boldsymbol{Z}; g_1, \ldots, g_p)$ is the solution for the estimating equations (2) with a sequence of initial values $\boldsymbol{U}_{init}$ such that $\boldsymbol{W} \to_P \boldsymbol{I}_p$. Then*

$$\begin{aligned}
\sqrt{n}\,w_{kl} &= -\sqrt{n}\,w_{lk} - \sqrt{n}\,s_{kl} + o_P(1), \qquad l < k,\\
\sqrt{n}\,(w_{kk} - 1) &= -\frac{1}{2}\sqrt{n}\,(s_{kk} - 1) + o_P(1), \qquad l = k,\\
\sqrt{n}\,w_{kl} &= \frac{\sqrt{n}\,t_{kl} - \lambda_{g_k,k}\sqrt{n}\,s_{kl}}{\lambda_{g_k,k} - \delta_{g_k,k}} + o_P(1), \; l > k.
\end{aligned}$$

In Theorem 1 we thus have to assume that the sequence of estimators $\boldsymbol{W}(\boldsymbol{U}_{init}, \boldsymbol{Z}; g_1, \ldots, g_p)$ can be selected in such a way that $\boldsymbol{W} \to_P \boldsymbol{I}_p$. Based on extensive simulations, it seems to us that this can be guaranteed if the initial value $\boldsymbol{U}_{init}$ in the algorithm for $\boldsymbol{W}_m(\boldsymbol{U}_{init}, \boldsymbol{Z}; g_1, \ldots, g_p)$ converges in probability to $\boldsymbol{I}_p$ as well. In the next section IV we propose a new estimator $\boldsymbol{W}_m(\boldsymbol{U}(\boldsymbol{Z}), \boldsymbol{Z}; \mathcal{G})$ that is using a data-based consistent initial value $\boldsymbol{U}(\boldsymbol{Z})$ and then finds the components and nonlinearities in a preassigned, optimal order.

We have the following useful corollaries.

Corollary 1. *Under the assumptions of Theorem 1, the joint asymptotic distribution of the elements of $\sqrt{n}\,(\boldsymbol{T}(\boldsymbol{Z}) - \boldsymbol{\Lambda})$ and the elements of $\sqrt{n}\,(\boldsymbol{S}(\boldsymbol{Z}) - \boldsymbol{I}_p)$ is a (singular) multivariate normal distribution, and also the asymptotic distribution of $\sqrt{n}\,vec(\boldsymbol{W} - \boldsymbol{I}_p)$ is multivariate normal.*

For an affine equivariant $\boldsymbol{W}(\boldsymbol{X})$, $\boldsymbol{W}(\boldsymbol{X})\boldsymbol{A} = \boldsymbol{W}(\boldsymbol{Z})$ and

$$\mathrm{vec}(\boldsymbol{W}(\boldsymbol{X}) - \boldsymbol{A}^{-1}) = (\boldsymbol{A}^{-T} \otimes \boldsymbol{I}_p)(\boldsymbol{W}(\boldsymbol{Z}) - \boldsymbol{I}_p)$$

which implies that, if $\sqrt{n}\,\mathrm{vec}(\boldsymbol{W}(\boldsymbol{Z}) - \boldsymbol{I}_p) \to_d N_{p^2}(\boldsymbol{0}, \boldsymbol{\Sigma})$ then the asymptotic distribution of $\sqrt{n}\,\mathrm{vec}(\boldsymbol{W}(\boldsymbol{X}) - \boldsymbol{A}^{-1})$ is $N_{p^2}(\boldsymbol{0}, (\boldsymbol{A}^{-T} \otimes \boldsymbol{I}_p)\boldsymbol{\Sigma}(\boldsymbol{A}^{-1} \otimes \boldsymbol{I}_p))$.

The asymptotic variances of the components of $\boldsymbol{W}(\boldsymbol{Z})$ may then be used to compare the efficiencies of the estimates for different choices of nonlinearities.

Corollary 2. *Under the assumptions of Theorem 1, the asymptotic covariance matrix (ASV) of the k-th row $\boldsymbol{w}_k$ of $\boldsymbol{W} = \boldsymbol{W}_m(\boldsymbol{U}_{init}, \boldsymbol{Z}; g_1, \ldots, g_p)$ is*

$$ASV(\boldsymbol{w}_k) = \sum_{j=1}^{k-1}(\alpha_{g_j,j}+1)\boldsymbol{e}_j\boldsymbol{e}_j^T + \kappa_k\boldsymbol{e}_k\boldsymbol{e}_k^T + \alpha_{g_k,k}\sum_{j=k+1}^{p}\boldsymbol{e}_j\boldsymbol{e}_j^T$$

and the sum of the asymptotic variances of the off-diagonal elements is

$$\sum_{i\neq j} ASV(w_{ij}) = 2\sum_{k=1}^{p}(p-k)\alpha_{g_k,k} + \frac{p(p-1)}{2}.$$

Here $\boldsymbol{e}_k$ is a p-vector with the kth element one and other elements zero and

$$\alpha_{g_k,k} = \frac{\sigma^2_{g_k,k} - \lambda^2_{g_k,k}}{(\lambda_{g_k,k} - \delta_{g,k})^2} \quad \textit{and} \quad \kappa_k = \frac{E(z_k^4) - 1}{4},$$

$k = 1, \ldots, p$.

For optimal choices of g_k, we need the following auxiliary result.

Lemma 2. *Let z be a random variable with $E(z) = 0$, $Var(z) = 1$ and assume that its density function f is twice continuously differentiable, the location score function $g_0(z) = -f'(z)/f(z)$ is continuously differentiable and the Fisher information number for the location problem, $I = Var(g_0(z))$, is finite. For a nonlinearity g, write $\sigma^2 = Var[g(z)]$, $\lambda = E[g(z)z]$, $\delta = E[g'(z)] = E[g(z)g_0(z)]$ and*

$$\alpha(g, f) = \frac{\sigma^2 - \lambda^2}{(\lambda - \delta)^2}.$$

Then, for all nonlinearities g,

$$\alpha(g, f) = [(I-1)\rho^2_{g(z)g_0(z)\cdot z}]^{-1},$$

where $\rho^2_{g(z)g_0(z)\cdot z}$ is the squared partial correlation between $g(z)$ and $g_0(z)$ given z. Therefore

$$\alpha(g, f) \geq \alpha(g_0, f) = (I-1)^{-1}.$$

The lemma implies that, for an optimal g, the nonlinearity parts of $g(z)$ and $g_0(z)$ should be the same, that is $g(z) - E[g(z)z]z = g_0(z) - E[g_0(z)z]z$. Note that the function $\tilde{g}(z) = g(z) - E[g(z)z]z$ is orthogonal to linear (function) z in the sense that $E[\tilde{g}(z)z] = 0$. The lemma then implies the following important optimality result for deflation-based FastICA estimates.

Theorem 2. *Under the assumptions of Theorem 1 and Lemma 2, the sum of the asymptotic variances of the off-diagonal elements of $\boldsymbol{W} = \boldsymbol{W}_m(\boldsymbol{U}_{init}, \boldsymbol{Z}; g_1, \ldots, g_p)$,*

$$\sum_{i\neq j} ASV(w_{ij}) = 2\sum_{k=1}^{p}(p-k)\alpha_{g_k,k} + \frac{p(p-1)}{2},$$

is minimized if the components are extracted according to a decreasing order of Fisher information numbers $I_1 \geq \ldots \geq I_p$ and the optimal location scores $g_{01}, \ldots, g_{0p}$ are used as nonlinearities. The minimum value then is

$$2\sum_{k=1}^{p}(p-k)[I_k - 1]^{-1} + \frac{p(p-1)}{2}.$$

One of the implications of Theorem 2 is that the deflation-based FastICA can never be fully efficient: the variances of the components cannot all attain the Cramer-Rao lower point, see [17]. However, Theorem 2 gives us tools to find optimal deflation-based FastICA estimate among all FastICA estimates with different extract orders and different choices of nonlinearities.

IV. The new estimator

The so called reloaded FastICA estimator in [15] optimizes the extraction order for a single nonlinearity function with a data based initial value $\boldsymbol{U}(\boldsymbol{X})$ in the algorithm. In this paper we introduce a new estimator which allows us to use different nonlinearities for different components, optimizes the choice of nonlinearities as well as the order in which the components are found. The nonlinearities are chosen from a finite set of available functions $\mathcal{G}$.

In Corollary 2, the asymptotic covariance matrices of the rows of $\boldsymbol{W} = \boldsymbol{W}_m(\boldsymbol{U}_{init}, \boldsymbol{Z}; g_1, \ldots, g_p)$ were obtained and we found that the asymptotic variances of the diagonal elements do not depend on the choice of the nonlinearities g_k, $k = 1, \ldots, p$. Therefore, the asymptotic efficiencies of the estimates are measured using the sum of asymptotic variances of the off-diagonal elements, that is,

$$\sum_{i\neq j} ASV(w_{ij}) = 2\sum_{k=1}^{p}(p-k)\alpha_{g_k,k} + \frac{p(p-1)}{2}.$$

This is clearly minimized if first (i) $g_1, \ldots, g_p$ satisfy

$$\alpha_{g_k,k} = \min\{\alpha_{g,k} \ : \ g \in \mathcal{G}\}$$

and then (ii) the indices are permuted so that

$$\alpha_{g_1,1} \leq \cdots \leq \alpha_{g_p,p}.$$

These findings suggest the following estimation procedure.

Definition 3. *The deflation-based FastICA estimate $\boldsymbol{W}_m(\boldsymbol{U}_0(\boldsymbol{X}), \boldsymbol{X}; \mathcal{G})$ with adaptive choices of nonlinearities is obtained using the following steps.*

1) *Transform the data using a preliminary affine equivariant estimate $\boldsymbol{W}_0(\boldsymbol{X}) = \boldsymbol{U}_0(\boldsymbol{X})\boldsymbol{S}^{-1/2}(\boldsymbol{X})$ and write $\hat{\boldsymbol{Z}} = \boldsymbol{W}_0(\boldsymbol{X})(\boldsymbol{X} - \boldsymbol{\mu}(\boldsymbol{X})1_n^T)$.*
2) *Use $\hat{\boldsymbol{Z}}$ to find $\hat{\alpha}_{g,k}$ for all $g \in \mathcal{G}$ and for all $k = 1, \ldots, p$.*
3) *Find $\hat{g}_1, \ldots, \hat{g}_p \in \mathcal{G}$ that minimize $\hat{\alpha}_{g,1}, \ldots, \hat{\alpha}_{g,p}$, resp.*
4) *Permute the rows of $\boldsymbol{U}_0(\boldsymbol{X})$, $\boldsymbol{U}_0(\boldsymbol{X}) \to \hat{\boldsymbol{P}}\boldsymbol{U}_0(\boldsymbol{X})$, so that, after the permutation, $\hat{\alpha}_{\hat{g}_1,1} \leq \cdots \leq \hat{\alpha}_{\hat{g}_p,p}$.*
5) *The estimate is the modified deflation-based estimate $\boldsymbol{W}_m(\hat{\boldsymbol{P}}\boldsymbol{U}_0(\boldsymbol{X}), \boldsymbol{X}; \hat{g}_1, \ldots, \hat{g}_p)$.*

The estimate $\boldsymbol{W}_m(\boldsymbol{U}_0(\boldsymbol{X}), \boldsymbol{X}; \mathcal{G})$ is affine equivariant, see Lemma 1. The following theorem gives the asymptotic distribution of the new estimator $\boldsymbol{W}_m(\boldsymbol{U}_0(\boldsymbol{X}), \boldsymbol{X}; \mathcal{G})$. The theorem is proved in the Appendix.

Theorem 3. *Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ be a random sample from a distribution of $\boldsymbol{z}$ satisfying the assumptions of Theorem 1. Assume that the components of $\boldsymbol{z}$ are ordered so that $\alpha_{g_1,1} < \cdots < \alpha_{g_p,p}$. If the initial estimate $\boldsymbol{W}_0(\boldsymbol{Z}) \to_P \boldsymbol{P}^T$ for some permutation matrix $\boldsymbol{P}^T$, then under general assumptions (see the Appendix)*

$$P\left(\boldsymbol{W}_m(\boldsymbol{U}_0(\boldsymbol{Z}), \boldsymbol{Z}; \mathcal{G}) = \boldsymbol{W}_m(\boldsymbol{P}\boldsymbol{U}_0(\boldsymbol{Z}), \boldsymbol{Z}; g_1, \ldots, g_p)\right) \to 1.$$

Remark 1. *The theorem thus shows that the asymptotic behavior of the deflation-based FastICA with adaptive choices of nonlinearities is similar to that of the modified deflation-based FastICA with known optimal choices of the nonlinearities and known optimal extraction order of the components,*

see Theorem 1. Note also that the reloaded deflation-based FastICA estimator [15] with a single nonlinearity is a special case here.

Remark 2. *The estimate $\boldsymbol{W}_m(\boldsymbol{U}_0(\boldsymbol{X}), \boldsymbol{X}; \mathcal{G})$ is the best possible fastICA estimate as soon as the optimal marginal nonlinearities $\log f_k$, $k = 1, \ldots, p$, are all in the set $\mathcal{G}$. In practice $f_1, \ldots, f_k$ are of course unknown, but instead of trying to estimate optimal nonlinearities $g_1, \ldots, g_k$, we hope to get a high efficiency with a careful flexible choice of possible candidates. In this way, our estimate is made fast in computation with a small loss of efficiency (as compared to the estimate with maximum efficiency).*

V. Simulations

A. The minimum distance index

To measure the separation accuracies in our simulation studies we use the minimum distance index [11] defined by

$$\hat{D} = \frac{1}{\sqrt{p-1}} \inf_{\boldsymbol{C} \in \mathcal{C}} \|\boldsymbol{C}\boldsymbol{W}(\boldsymbol{X})\boldsymbol{A} - I_p\| \qquad (1)$$

with the matrix (Frobenius) norm $\|\cdot\|$. The index is affine invariant as $\boldsymbol{W}(\boldsymbol{X})\boldsymbol{A}$ does not depend on the mixing matrix $\boldsymbol{A}$. The minimum distance index is a natural choice for our simulation studies, as it is the only performance criterium with known asymptotical behavior. If $\sqrt{n}\,\mathrm{vec}(\boldsymbol{W}(\boldsymbol{X})\boldsymbol{A} - \boldsymbol{I}_p)$ has asymptotic normal distribution with zero mean, then

$$n\hat{D}^2 = \frac{n}{p-1}\|\mathrm{off}(\boldsymbol{W}(\boldsymbol{X})\boldsymbol{A})\|^2 + o_p(1),$$

where $\mathrm{off}(\boldsymbol{A}) = \boldsymbol{A} - \mathrm{diag}(\boldsymbol{A})$. The expected value of its asymptotic distribution is the sum of the asymptotic variances of $\mathrm{off}(\boldsymbol{W}(\boldsymbol{Z}))$. This relates the finite sample efficiencies to the asymptotic efficiencies considered in Section III.

B. Models and asymptotic behavior

For the rest of the paper we assume for demonstration purposes that the set $\mathcal{G}$ will consist of the functions: *pow3*, *tanh* with $a = 1$, *gaus* with $a = 1$, *left* with $a = 0.6$ (*lt0.6*), *right* with $a = 0.6$ (*rt0.6*), and *both* with several values of a (*bt0*, *bt0.2*, *bt0.4*, *bt0.6*, *bt0.8*, *bt1.0*, *bt1.2*, *bt1.4* and *bt1.6*).

We will consider the performance of our adaptive deflation-based FastICA method in four different settings of source distributions:

- Setting 1: The log-normal distribution with variance parameter value 0.25 (LN), exponential power distribution with shape parameter value 4 (EP4), uniform distribution (U) and t_5-distribution (T).
- Setting 2: The exponential distribution (E), the chi-square distribution with 8 degrees of freedom (C), the Laplace distribution (L) and the gaussian distribution (G).
- Setting 3: The gaussian distribution (G), exponential power distribution with shape parameter value 3 (EP3), exponential power distribution with shape parameter value 6 (EP6) and the nonsymmetric mixture of two gaussian distributions as defined as distribution (l) in [1] (MG).
- Setting 4: The distributions (T), (EP3) and (EP6).

For all settings the distributions are standardized to have mean value zero and variance one.

We used numerical integration to calculate $\alpha_{g,k}$ for each nonlinearity function $g \in \mathcal{G}$ and for each distribution in our four settings. See Section II-C for our set of nonlinearities $\mathcal{G}$. Also $\alpha_{g,k}$ for the optimal nonlinearity function (*optim*) is given if only the density function is twice continuously differentiable. The values are given in Table I.

TABLE I
THE VALUES $\alpha_{g,k}$ FOR ALL NONLINEARITY FUNCTIONS IN $\mathcal{G}$ AND ALL SOURCES USED IN SETTING 1 -SETTING 4.

$g(\cdot)$	LN	U	EP4	T	E	C	L	EP3	EP6	MG
pow3	8.39	0.43	2.70	∞	5	15	6	7.97	1.16	14.53
tanh	4.33	0.69	2.98	4.00	3.14	32.13	2.01	7.73	1.47	11.50
gaus	6.08	0.71	3.09	4.32	3.94	86.95	1.82	7.95	1.53	11.47
skew	1.58	∞	∞	∞	1	2.5	∞	∞	∞	16.26
lt0.6	0.55	1.67	8.89	20.63	0.08	1.23	11.49	24.92	4.05	1475
rt0.6	3.28	1.67	8.89	20.63	2.33	5.92	11.49	24.92	4.05	8.38
bt0	4.55	0.60	2.82	7.61	3.31	16.85	3.00	7.57	1.35	12.10
bt0.2	4.52	0.58	2.80	7.72	3.26	16.4	3.13	7.58	1.33	12.21
bt0.4	4.44	0.53	2.77	8.06	3.11	15.25	3.43	7.63	1.27	12.62
bt0.6	4.36	0.45	2.72	8.66	2.95	13.79	3.87	7.79	1.20	13.46
bt0.8	4.37	0.37	2.73	9.55	2.87	12.45	4.43	8.15	1.13	14.88
bt1.0	4.51	0.29	2.82	10.74	3.04	11.51	5.13	8.79	1.09	17.17
bt1.2	4.85	0.20	3.04	12.38	3.47	11.15	5.97	9.87	1.09	20.68
bt1.4	5.41	0.12	3.49	14.18	3.96	11.52	6.99	11.63	1.18	26.13
bt1.6	6.12	0.05	4.38	16.51	4.52	12.72	8.23	14.54	1.45	34.74
optim	0.50	-	2.70	4	-	1	-	7.57	1.07	3.69

Hence, in Setting 1, the adaptive procedure aims to extract first the uniformly distributed component using *bt1.6*, then the log-normally distributed one with *lt0.6*, before using *pow3* for EP4 distributed component. This yields, as Table II shows, a performance value of 16.18 which is about one half of the value obtained using *pow3*, *tanh* or *gaus* alone in the reloaded deflation-based FastICA procedure. In Setting 2 the optimal performance is obtained by first finding the exponentially distributed and chi-squared component (in this order) with *lt0.6* and then Laplace distributed component with *gaus*. The expected value of the asymptotic distribution of $n(p-1)\hat{D}^2$ is then 15.04 which is a highly significant improvement as compared to the reloaded procedures with traditional nonlinearities. In Setting 3 the separation order is EP6, EP3 and MG using *bt1.0*, *bt* and *rt0.6*, respectively. Finally, in Setting 4, one finds first EP6 and EP3. In the last two settings, the adaptive procedure again outperforms reloaded procedures but, of course, are not as good as the optimal procedure as the optimal nonlinearities are not included in $\mathcal{G}$.

C. Simulation study

We next consider the performance of the adaptive FastICA procedure for finite data sets from the same Settings 1-4. This then allows us to make comparisons between finite sample and asymptotic behaviors as well. For the adaptive procedure we thus need an equivariant and consistent initial ICA estimate

TABLE II
THE ASYMPTOTIC VALUES OF $n(p-1)E[D(\hat{W}, A)^2]$ FOR THE FOUR SETTINGS OF RELOADED DEFLATION-BASED FASTICA WITH NONLINEARITIES *pow3*, *tanh* AND *gaus* AND ADAPTIVE DEFLATION-BASED FASTICA WITH $\mathcal{G}$ (*adaptive*) AND $\mathcal{G}_0$ (*optimal*).

	pow3	*tanh*	*gaus*	*adaptive*	*optimal*
Setting 1	36.16	30.06	31.26	16.18	-
Setting 2	90.00	94.88	206.58	15.04	-
Setting 3	73.91	68.75	69.91	59.57	42.32
Setting 4	23.59	16.90	17.75	15.37	15.27

$\boldsymbol{W}_0(\boldsymbol{X})$. Of course any estimate meeting the requirements will do - but in the following we will consider the following three estimates

1) FOBI [2] is the initial estimate that is easiest to compute. For its convergence, the fourth moments of $\boldsymbol{z}$ must exist and the kurtosis values $E(z_k^4)-3$, $k=1,\ldots,p$ must be distinct.
2) For the convergence of the JADE [3] estimator, fourth moments must exist with at most one zero kurtosis value. JADE is, however, computationally expensive for large dimensions p.
3) k-JADE [12], $k=1,\ldots,p$, may be seen as a compromise between FOBI and JADE where. The smaller k the faster is its computation. For the consistency, the procedure allows at most k components with equal kurtosis values and at most one zero kurtosis value.

For Setting 1, we first consider the performance of adaptive FastICA when different initial estimators are used. Figure 1 shows, for different sample sizes n and with 10000 repetitions, the average criterium values for the three initial estimates FOBI, JADE and 2-JADE as well as for the adaptive FastICA estimates FI-FOBI, FI-JADE and FI-2-JADE with different initial estimates. The difference between JADE and 2-JADE is negligible in this low-dimensional case and they both clearly outperform FOBI. Despite the differences in initial estimates, the average behavior of the adaptive FastICA seems similar in all cases and it is in accordance with the asymptotic theory. (The horizontal line yields the expected value of the asymptotic distribution of $n(p-1)\hat{D}^2$.) Based on these results, we recommend to use JADE for low-dimensional problems and k-JADE for higher dimensions. In the simulations that follow we always use JADE as an initial estimate.

In all four settings, we compare the performance of our new adaptive FastICA method to the reloaded procedure that uses only one popular linearity function, namely, *pow3*, *tanh* and *gaus*. For the comparison we also include, in Setting 1 and Setting 2, the original deflation-based FastICA with *tanh* and a random initial value (denoted here *rand*), which may be the most common version used in practice. In Setting 3 and Setting 4 all the marginal densities are continuously differentiable, and, in adaptive fastICA, we can compare two sets of nonlinearity functions, $\mathcal{G}$ and $\mathcal{G}_0$ where the latter includes the optimal marginal nonlinearities as well as the established nonlinearities *pow3*, *tanh*, *gaus* and *skew*. The adaptive FastICA using $\mathcal{G}_0$ is called *optimal*.

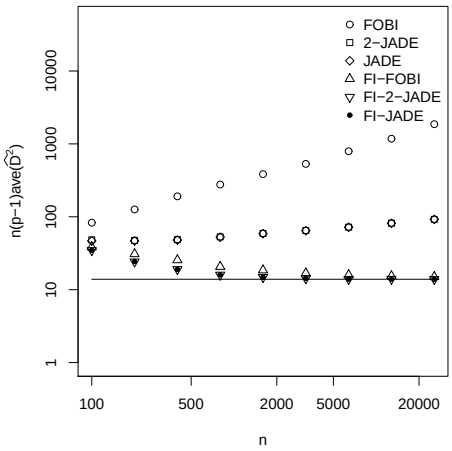

Fig. 1. The averages of $n(p-1)\hat{D}^2$ from 10 000 repetitions in Setting 1. The horizontal line gives the expected value of the asymptotic distribution of $n(p-1)\hat{D}^2$ for the adaptive fastICA method.

The averages of $n(p-1)\hat{D}^2$ for all four settings and for different estimates are depicted in Figure 2. Naturally, the performance of the new adaptive method is better than *pow3*, *tanh* and *gaus* alone and seems to reach the asymptotic level quite fast. No asymptotic level can be given for *rand* since its asymptotic distribution is unknown. It is remarkable that adaptive FastICA also makes the computations more reliable, see Tables III-VI for the number of runs that failed to converge in 10000 repetitions. While reloaded *pow3*, *gaus* and *tanh* with a data-based initial value are already more stable than *rand*, see also [15], the adaptive algorithm has the clearly lowest failure rates.

All computations were done using our freely available R-package fICA [13], which implements the new deflation-based FastICA method allowing the user to choose the initial estimator and to provide the set of nonlinearities $\mathcal{G}$. The computation times for the adaptive estimates (using the default set $\mathcal{G}$) were in our simulation studies only 5-10 times longer than the computation times for the traditional FastICA estimates. This extra computational load is a very small price to pay considering the efficiency gain of the adaptive method. Finally notice that, the asymptotic as well as estimated covariances can be computed using our R-package BSSasymp [14].

VI. CONCLUSION

In this paper we extend recent theoretical results for deflation-based FastICA and suggest a novel adaptive deflation-based FastICA method that, for each component, picks up the best nonlinearity function among the nonlinearity functions specified by the user, and finds the sources in an optimal order. The approach is based on new theoretical results for the asymptotic distribution of the FastICA estimate; the asymptotic efficiency is shown to depend (i) on the marginal distributions of the sources, (ii) on the used nonlinearity functions, and (iii) on the order in which the sources are found. It is shown that the best possible set of nonlinearities then includes the optimal location scores of the twice differentiable marginal densities. For the optimization step of the algorithm

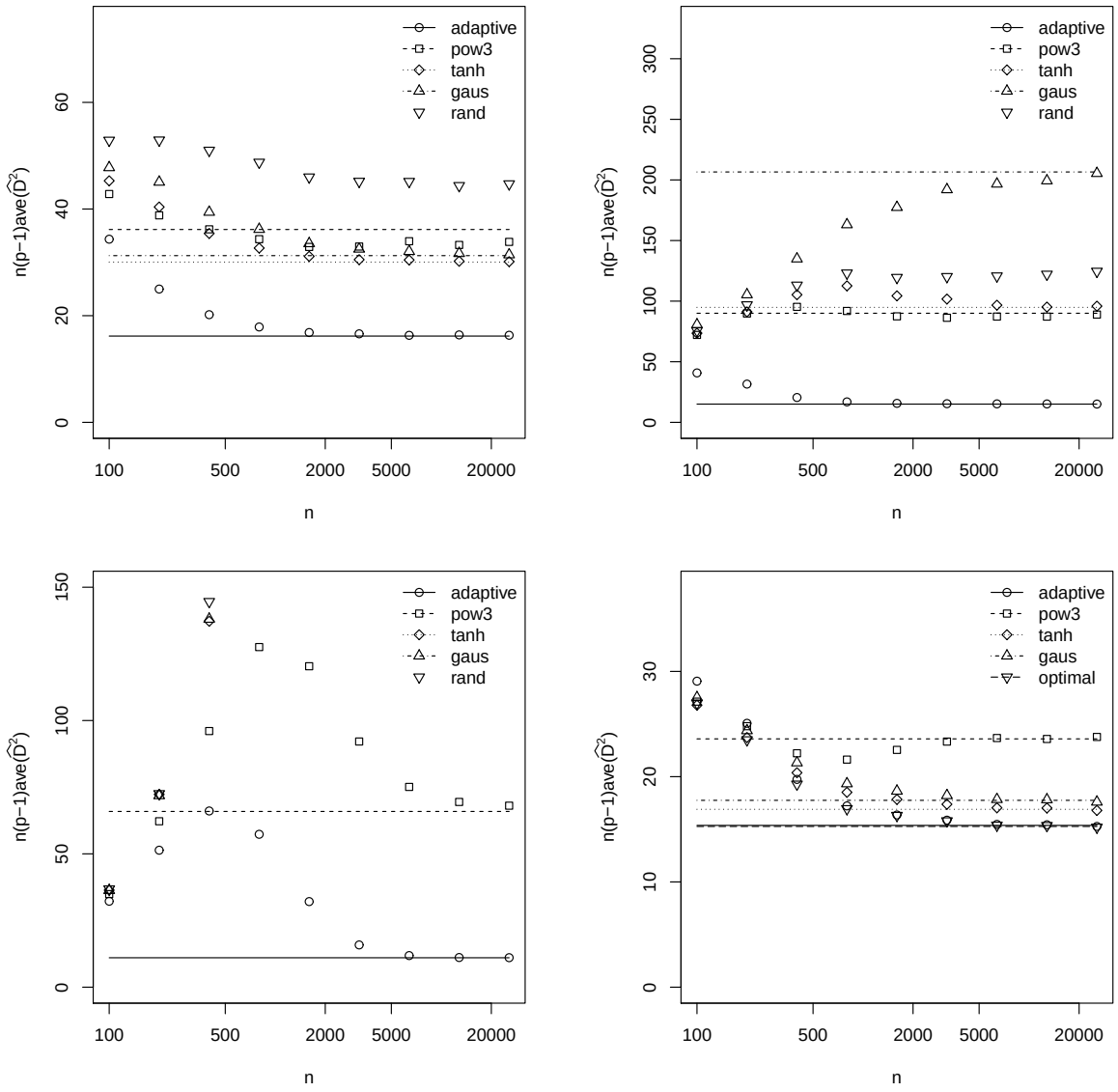

Fig. 2. The averages of $n(p-1)\hat{D^2}$ from 10 000 repetitions in all four settings (Setting 1 in the top left panel, Setting 2 in the top right panel, Setting 3 in the bottom left panel and Setting 4 in the bottom right panel. The horizontal lines give the expected values of the asymptotic distributions of $n(p-1)\hat{D^2}$.

and for the affine equivariance of the procedure, an affine equivariant preliminary ICA estimate such as FOBI, JADE or k-JADE is needed. With the sparse set of nonlinearities used in our simulations, the new estimate clearly outperforms estimates that are based on the use of a single nonlinearity only and is more stable in simulations. We thus think that the adapted version of FastICA developed in the paper is the best possible approach available if one wishes to find the sources one after another using FastICA method.

APPENDIX A
PROOF OF LEMMA 2

Since $\alpha(g,f) = \alpha(ag+b,f)$ for any nonzero real number a and any real number b, we may assume that $\mathrm{E}[g(z)] = 0$ and $Var[g(z)] = \sigma^2 = 1$. The assumptions and integration by parts then gives

$$\begin{aligned}
&\mathrm{E}[g_0(z)] = 0,\\
&\rho_{g(z)g_0(z)} = I^{-1/2}\mathrm{E}[g(z)g_0(z)] = I^{-1/2}\delta,\\
&\rho_{g(z)z} = \mathrm{E}[g(z)z] = \lambda \text{ and}\\
&\rho_{g_0(z)z} = I^{-1/2}\mathrm{E}[g_0(z)z] = I^{-1/2}.
\end{aligned}$$

Hence

$$\begin{aligned}
\alpha(g,f) &= \frac{1-\lambda^2}{(\lambda-\delta)^2} = \frac{I(1-I^{-1})(1-\lambda^2)}{(I-1)(\lambda-\delta)^2}\\
&= \frac{(1-I^{-1})(1-\lambda^2)}{(I-1)(I^{-1/2}\lambda - I^{-1/2}\delta)^2}\\
&= \frac{(1-\rho^2_{g_0(z)z})(1-\rho^2_{g(z)z})}{(I-1)(\rho_{g_0(z)z}\rho_{g(z)z} - \rho_{g(z)g_0(z)})^2}\\
&= [(I-1)\rho^2_{g(z)g_0(z)\cdot z}]^{-1}.
\end{aligned}$$

TABLE III
NUMBER OF NON-CONVERGENT RUNS IN 10000 RUNS FOR SETTING 1.

	n=100	200	400	800	1600	$\geq$ 3200
adaptive	127	12	1	0	0	0
pow3	503	43	0	0	0	0
gaus	981	245	18	0	0	0
tanh	658	98	3	0	0	0
rand	1541	626	164	18	2	0

TABLE IV
NUMBER OF NON-CONVERGENT RUNS IN 10000 RUNS IN SETTING 2.

	n=100	200	400	800	1600	3200	6400	12800	25600
adaptive	341	73	6	1	0	0	0	0	0
pow3	1698	700	236	37	0	0	0	0	0
gaus	2391	1704	1379	1268	938	433	101	4	0
tanh	1932	1164	848	507	177	14	0	0	0
rand	2739	1957	1477	1002	477	187	79	36	28

TABLE V
NUMBER OF NON-CONVERGENT RUNS IN 10000 RUNS FOR SETTING 3.

	n=100	200	400	800	1600	$\geq$ 3200
adaptive	1113	722	277	33	1	0
pow3	2140	1371	658	165	15	0
tanh	1579	1016	429	91	6	0
gaus	1463	975	433	100	5	0
optimal	980	560	167	14	0	0

TABLE VI
NUMBER OF NON-CONVERGENT RUNS IN 10000 RUNS FOR SETTING 4.

	n=100	200	400	$\geq$ 800
adaptive	281	47	1	0
pow3	440	49	3	0
tanh	329	51	0	0
gaus	373	72	0	0
optimal	253	37	0	0

APPENDIX B
PROOF OF THEOREM 3

It is not a restriction to assume that $\boldsymbol{W}_0(\boldsymbol{Z}) = (\boldsymbol{w}_{01}, \ldots, \boldsymbol{w}_{0p})^T \rightarrow_P \boldsymbol{I}_p$. The algorithm then uses for $\beta_{h,k} = E[h(z_k)]$ the estimates of the type

$$\hat{\beta}_{h,k} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{w}_{0k}^T \boldsymbol{z}_i).$$

Recall that, for all $g \in \mathcal{G}$ and k, we need the estimates for $\sigma^2 = E[(g(z_k))^2] - (E[g(z_k)])^2$, $\lambda = \mathrm{E}[g(z_k)z_k]$, $\delta = \mathrm{E}[g'(z_k)]$ and, finally,

$$\alpha_{g,k} = \frac{\sigma^2 - \lambda^2}{(\lambda - \delta)^2}.$$

In the following we therefore assume that h is in

$$\mathcal{H} = \left\{ h \ : \ h(z) = (g(z))^2, \ g(z), \ g(z)z \text{ or } g'(z), \ g \in \mathcal{G} \right\}.$$

First note that, assuming that the $\beta_{h,k}$ exist,

$$\tilde{\beta}_{h,k} = \frac{1}{n}\sum_{i=1}^{n} h(z_{ik}) \rightarrow_P \beta_{h,k}, \quad \text{for all } k \text{ and } h \in \mathcal{H}.$$

Then if, for all $h \in \mathcal{H}$ and all components z_k of $\boldsymbol{z}$, there exists an integer $s > 0$ such that

$$\sup_z |h^{(s)}(z)| \leq M \text{ and } E(||h^{(r)}(z_k) z_k^r||) < \infty, \ r = 0, \ldots, s-1,$$

and the sth moments exist, then, using Taylor expansions, it easily follows that $\hat{\beta}_{h,k} - \tilde{\beta}_{h.k} \rightarrow_P 0$ and, consequently, $\hat{\beta}_{h,k} \rightarrow_P \beta_{h,k}$ for all h and k. If, for example, $h(z) = z^2$, then $h''(z) \equiv 2$,

$$\begin{aligned} h(\boldsymbol{w}_{0k}^T \boldsymbol{z}_i) &= h(z_{ik}) + (\boldsymbol{w}_{0k} - \boldsymbol{e}_k)^T 2 z_{ik} \boldsymbol{z}_i \\ &+ (\boldsymbol{w}_{0k} - \boldsymbol{e}_k)^T (\boldsymbol{z}_i \boldsymbol{z}_i^T)(\boldsymbol{w}_{0k} - \boldsymbol{e}_k) \end{aligned}$$

and therefore

$$\begin{aligned} \hat{\beta}_{h,k} - \tilde{\beta}_{h,k} &= (\boldsymbol{w}_{0k} - \boldsymbol{e}_k)^T \frac{2}{n}\sum_{i=1}^{n} z_{ik} \boldsymbol{z}_i \\ &+ (\boldsymbol{w}_{0k} - \boldsymbol{e}_k)^T (\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T)(\boldsymbol{w}_{0k} - \boldsymbol{e}_k) \\ &\rightarrow_P 0. \end{aligned}$$

Note that if the assumption above holds true for all $h \in \mathcal{H}$ and for all k, then we obtain also the convergence

$$\hat{\alpha}_{g,k} \rightarrow_p \alpha_{g,k}, \quad \text{for all } g \in \mathcal{G} \text{ and for all } k.$$

This implies that the probability for

$$\hat{\alpha}_{g_k,k} = \min\{\hat{\alpha}_{g,k} \ : \ g \in \mathcal{G}\}$$

and

$$\hat{\alpha}_{g_1,1} < \cdots < \hat{\alpha}_{g_p,p}$$

goes to one. This further means that, in the algorithm, the rows of $\boldsymbol{U}_0(\boldsymbol{z})$ are permuted with a probability going to zero and therefore 'permuted' $\boldsymbol{U}_0(\boldsymbol{z})$ converges in probability to $\boldsymbol{I}_p$ as well. The asymptotic distribution is then given in Theorem 1.

ACKNOWLEDGMENT

This work was supported by the Academy of Finland (grants 256291 and 268703).

REFERENCES

[1] F. Bach and M. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.
[2] J.F. Cardoso, "Source Separation Using Higher Order Moments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 1989), Glasgow,
[3] J.C. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *IEE Proceedings-F*, vol. 140, pp. 362–370, 1993.
[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, "Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound", *IEEE Trans. Neural Networks*, vol. 17, no. 5, pp. 1265-1277, 1999.

[7] A. Hyvärinen, "One-Unit Contrast Functions for Independent Component Analysis: A Statistical Analysis," in *Neural Networks for Signal Processing VII* (Proc. IEEE NNSP Workshop 1997), Amelia Island, Florida, pp. 388-397.
[8] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications.* Wiley , Cichester.
[9] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis.* New York: Wiley, 2001.
[10] P. Ilmonen, J., Nevalainen, and H. Oja, "Characteristics of multivariate distributions and the invariant coordinate system," *Statistics and Probability Letters* vol. 80, pp. 1844-1853, 2010.
[11] 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)*, 229-236, 2010.
[12] J. Miettinen, K. Nordhausen, H. Oja and S. Taskinen, "Fast Equivariant JADE," in *Proceedings of "38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)"*, Vancouver, 2013, pp. 6153–6157.
[13] J. Miettinen, K. Nordhausen, H. Oja and S. Taskinen, "fICA: Classic and adaptive FastICA algorithms," R package version 1.0-0, *http://cran.r-project.org/web/packages/fICA*, 2013.
[14] J. Miettinen, K. Nordhausen, H. Oja and S. Taskinen, "BSSasymp: Covariance matrices of some BSS mixing and unmixing matrix estimates," R package version 1.0-0, *http://cran.r-project.org/web/packages/BSSasymp*, 2013.
[15] K. Nordhausen, P. Ilmonen, A. Mandal, H. Oja and E. Ollila, "Deflation-based FastICA reloaded," in *Proc. "19th European Signal Processing Conference 2011 (EUSIPCO 2011)"*, Barcelona, 2011, pp. 1854-1858.
[16] E. Ollila, "On the robustness of the deflation-based FastICA estimator," in *Proc. IEEE Workshop on Statistical Signal Processing* (SSP'09), pp. 673-676, 2009.
[17] E. Ollila, "The deflation-based FastICA estimator: statistical analysis revisited," *IEEE Transactions on Signal Processing*, vol. 58, pp. 1527-1541, 2010.

Jari Miettinen received the M.Sc. degree in mathematics from the University of Jyväskylä, in 2011, and the Ph.D. degree in statistics from the University of Jyväskylä, in 2014. Currently he is a postdoctoral researcher at the University of Jyväskylä. His research interest is statistical signal processing.

Klaus Nordhausen received the M.Sc. degree in statistics from the University of Dortmund in 2003, and the Ph.D. degree in biometry from the University of Tampere in 2008. From 2010 to 2012 he was a Postdoctoral Researcher of the Academy of Finland. He has also been a Lecturer at the University of Tampere. Currently he is a Senior Research Fellow at the University of Turku. His research interest are nonparametric and robust multivariate methods, statistical signal processing, computational statistics and dimension reduction.

Hannu Oja received the M.Sc. degree in statistics from the University of Tampere, Finland, in 1973, and the Ph.D. degree in statistics from the University of Oulu in 1981. He is currently a Professor in the Department of Mathematics and Statistics at the University of Turku, Finland He has been a Professor in statistics at the Universities of Oulu and Jyväskylä and a Professor in biometry and an Academy Professor at the University of Tampere. His research interests include nonparametric and robust multivariate methods, dimensional reduction and blind source separation. Dr. Oja is a fellow of IMS, the Institute of Mathematical Statistics. He has served as an Associate Editor for various statistical journals.

Sara Taskinen received the M.Sc. degree in mathematics from the University of Jyväskylä, in 1999, and the Ph.D. degree in statistics from the University of Jyväskylä, in 2003. From 2004 to 2007, she was a Postdoctoral Researcher of the Academy of Finland. She has also been a Senior Assistant at the University of Jyväskylä. Currently, she is appointed as an Academy Research Fellow of the Academy of Finland at the University of Jyväskylä. She is also a University Lecturer of the same university. Her research interests are nonparametric and robust multivariate methods, statistical signal processing and robust methods in biology and ecology.