Consistent Independent Component Analysis and Prewhitening

Aiyou Chen and Peter J. Bickel

***Abstract*—We study the statistical merits of two techniques used in the literature of independent component analysis (ICA). First, we analyze the characteristic-function based ICA method (CHFICA) and study its statistical properties such as consistency, $\sqrt{n}$-consistency, and robustness against small additive noise. Second, we study the validity of prewhitening: a preprocessing technique used by many ICA algorithms, as applied to the CHFICA method. In particular, we establish the surprising effectiveness of this technique even when some components have heavy tails and others do not. A fast new algorithm implementing the prewhitened CHFICA method is also provided.**

***Index Terms*—Asymptotic normality, characteristic function, consistency, incomplete Cholesky decomposition, independent component analysis, prewhitening.**

I. Introduction

OVER the past decade, independent component analysis (ICA) has received much attention in many different fields, such as signal processing and machine learning [17], [21], [29], [3]. It has been used as a standard statistical tool for blind source separation, e.g., in brain imaging analysis [27]. Formally, the classical ICA model is of the form

$$\mathbf{X} = \mathrm{A}\mathbf{S} \tag{1}$$

where $\mathbf{X} = [X_1, \ldots, X_m]^T$ is a random vector of observations, $\mathbf{S} = [S_1, \ldots, S_m]^T$ is a random vector of hidden sources with mutually independent components, and A is a nonsingular mixing matrix. Define $\mathrm{W} = \mathrm{A}^{-1}$, which is usually called the unmixing matrix. It is well known that A (thus W) is identifiable up to ambiguity of order and scaling if and only if at most one of $\mathbf{S}$'s components is Gaussian [13], [16], [23]. We call these assumptions identifiability conditions. To specify W uniquely, we need to put some scale and permutation constraints either on $\mathbf{S}$ or on W. Having n independently and identically distributed (i.i.d.) samples of $\mathbf{X}$, say $\{\mathbf{X}(j) : 1 \leq j \leq n\}$, ICA aims to estimate the unmixing matrix W and thus to recover each hidden source using $S_k = W_k\mathbf{X}$, where W_k is the kth row of W. This type of problem is also called blind source separation (BSS) in the engineering literature. Many statistical approaches have been proposed for such BSS. Some are based on contrast functions or estimating equations derived from maximal likelihood (ML), mutual information, as well as other criteria under specific parametric models for the sources. These methods use features that do not determine the distribution uniquely, and they are therefore not consistent without further assumptions. An example is the Joint Approximate Diagonalization of Eigenmatrices (JADE) method, which relies on the source distribution's fourth multilinear cumulant not vanishing [10]. Such inconsistency can be readily explained by marginal mismatch of the hidden sources' distributions [9]. Other methods (e.g., [3], [7], [11], and [19]) are based on nonparametric approximation of some feature functions of hidden sources, such as probability density function or density score, which do determine the distribution. Theoretical analysis of these methods is not available (but see Chen and Bickel [11]). See [21] and references therein for an extensive review of ICA methodologies and algorithms.

The characteristic-function based ICA method (CHFICA) [15] has the virture of not requiring an estimation of delicate parameters such as densities and, yet, should be consistent under general conditions. Eriksson and Koivunen [15] showed that the CHFICA method performed very well under simulations and gave a formal argument for consistency.

Prewhitening is a popularly used preprocessing technique in the ICA literature, which speeds algorithms up substantially. Its validity is expected when all hidden sources have finite second moments, and, under second moment constraints, Cardoso [8] obtained a lower bound on estimation errors of the prewhitened ICA algorithms. We found in simulations that even with heavy-tailed data on some sources, prewhitened CHFICA still works well.

In this paper, we address both the question of $\sqrt{n}$-consistency and asymptotic normality for CHFICA and its surprisingly good performance after prewhitening. The paper is organized as follows. In Section II, after reviewing CHFICA, we study its consistency, $\sqrt{n}$-consistency, and asymptotic normality, as defined in Theorem 1, and robustness properties. In Section III, after reviewing prewhitening as a preprocessing technique for ICA, we propose a new algorithm to implement prewhitened CHFICA by using incomplete Cholesky decomposition. In Section IV, we study the consistency of prewhitening in terms of the acting parameter space of the matrix W and show that the prewhitened CHFICA method can be consistent, even when some of the hidden sources do not have finite second moments. In our conclusion in Section V, we review the performances of the procedures we have considered both from the theoretical and numerical point of view, which, as we have noted, are in agreement. Some technical details are given in the Appendix,

Manuscript received March 6, 2004; revised December 23, 2004. This work was supported by the National Science Foundation under Grant DMS-01-04075. Part of the results of this paper appeared in the Fifth International Conference on Independent Component Analysis, August 2004, Granada, Spain. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Behrouz Farhang-Boroujeny.

A. Chen is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: aychen@research.bell-labs.com).

P. J. Bickel is with the Department of Statistics, University of California, Berkeley, CA 94720 USA (e-mail: bickel@stat.berkeley.edu).

Digital Object Identifier 10.1109/TSP.2005.855098

whereas others are in the technical report [12] available on the WEB (http://www.stat.berkeley.edu/tech-reports/index.html).

In this paper, for any matrix, say W, W_j and W^j denote its jth row and jth column separately. $||.||$ denotes the Euclidean norm for a vector and denotes the Frobenius norm for a matrix. We use boldface uppercase to denote random vectors and random matrices.

II. Statistical Properties of CHFICA

A. Review of the Characteristic-Function Based ICA Method

For a random vector ξ, its characteristic function (cf.) is defined by $c_\xi(\mathbf{t}) = E[\exp(i\mathbf{t}^T\xi)]$ and its empirical characteristic function (e.cf.), given n i.i.d. observations $\{\xi(j)\}_{j=1}^n$, is defined by $\hat{c}_\xi(\mathbf{t}) = 1/n\sum_{j=1}^n \exp(i\mathbf{t}^T\xi(j))$, where $i = \sqrt{-1}$. It is well known that $c_\xi(\mathbf{t})$ can be factorized into the product of its marginal characteristic function if and only if ξ has mutually independent components. For $\mathbf{S} = \mathrm{W}\mathbf{X}$, then the difference between $c_\mathbf{S}$ and the product of its marginal cf., for example

$$\Delta_\lambda(\mathrm{W}) = \int_{\mathbf{t}\in\mathcal{R}^m} \left| c_\mathbf{S}(\mathbf{t}) - \prod_{j=1}^m c_{S_j}(t_j) \right|^2 \lambda_m(\mathbf{t})d\mathbf{t} \quad (2)$$

for an m-dim probability density function λ_m, can be considered to be a measurement of the dependence level among $\mathbf{S}$'s components. To make sure that $\Delta_\lambda(\mathrm{W}) = 0$ if and only if $\mathbf{S}$'s components are mutually independent, it is sufficient to choose λ_m such that it is continuous and $\lambda_m(\mathbf{t}) > 0$ for $\forall \mathbf{t} \in \mathcal{R}^m$. We use $\lambda_m(\mathbf{t}) = \prod_{j=1}^m \lambda_1(t_j)$, where λ_1 is a one-dimensional (1-D) positive density function.

To avoid unidentifiability, we consider a true matrix W_P whose rows are scaled and permuted such that I) each of its rows has norm 1; II) the element with maximal modulus in each row is positive; III) the rows are ordered by $\prec$ (for $\forall \mathbf{a}, \mathbf{b} \in \mathcal{R}^m$, $\mathbf{a} \prec \mathbf{b}$ iff there exists $k \in \{1,\ldots,m\}$ such that $a_k < b_k$ and $a_j = b_j$ for $j = 1,\ldots,k-1$). We denote the set of matrices that satisfy conditions I-III by Ω, i.e.,

$$\Omega = \Big\{ \mathrm{W}\ m\times m \text{ matrix} : W_1 \prec \cdots \prec W_m$$
$$||W_k|| = 1,\ \max_{1\le j\le m}(W_{kj}) = \max_{1\le j\le m}(|W_{kj}|),\ \text{for } 1 \le k \le m \Big\}.$$

It is obvious that if W is an $m\times m$ nonsingular matrix, then by rescaling and permuting its rows appropriately the transformed matrix will belong to Ω. We denote such a row-rescaling-permuting transformation as $[\cdot]_\Omega$, i.e, $[\mathrm{W}]_\Omega \in \Omega$. Note that the matrix $\mathrm{W}_P \in \Omega$.

CHFICA makes use of this criterion given by (2) as follows. Using the observations of $\mathbf{X}$, the CHFICA estimator of the matrix W_P is defined by

$$\hat{\mathbf{W}} = \mathrm{argmin}_{\mathrm{W}\in\Omega}\hat{\Delta}_\lambda(\mathrm{W}) \quad (3)$$

where

$$\hat{\Delta}_\lambda(\mathrm{W}) = \int_{\mathbf{t}\in\mathcal{R}^m} \left| \hat{c}_{\mathrm{W}\mathbf{X}}(\mathbf{t}) - \prod_{j=1}^m \hat{c}_{\mathrm{W}_j\mathbf{X}}(t_j) \right|^2 \lambda_m(\mathbf{t})d\mathbf{t}. \quad (4)$$

B. Convergence Rates and Some Simulations

Since $\hat{\Delta}_\lambda(\mathrm{W})$ is a function of W, $\hat{\mathbf{W}}$ can be obtained by directly solving the above optimization problem. Like many other ICA methods, we approximate a feature function of hidden sources: the characteristic function. The advantage here is that the characteristic function can be estimated easily and consistently by the corresponding e.cf. Eriksson and Koivunen [15] argued for the consistency of this estimator using extensive simulations and some heuristics. Here, we give a rigorous proof of this claim and also study the $\sqrt{n}$-consistency and asymptotic normality of the CHFICA estimator.

Let $\rho_1(s_j) = \int_{t\in\mathcal{R}} e^{is_jt}\lambda_1(t)dt$ for $s_j \in \mathcal{R}$, and define $\rho_m(\mathbf{s}) = \prod_{j=1}^m \rho_1(s_j)$ for $\mathbf{s} = (s_1,\ldots,s_m)^T$. Obviously, ρ_k is a cf. corresponding to the density function λ_k for $k \in \{1,m\}$. We will often omit the subscripts of ρ_1, ρ_m, λ_1, and λ_m when the dimension is obvious from their arguments.

For a sequence of random vectors with finite dimension $\{\xi_i\}_{i=1}^\infty$, $\xi_n = o_P(1)$ iff $P(|\xi_n| \ge \delta) \to 0$ as $n \uparrow \infty$ for $\forall\delta > 0$, and $\xi_n = O_P(1)$ iff $\sup_n P(|\xi_n| \ge M) \to 0$ as $M \uparrow \infty$. For two 1-D random sequences, $\xi_n = O_P(\eta_n)$ iff $\xi_n/\eta_n = O_P(1)$.

Theorem 1: Suppose that λ_1 is a symmetric density function with $\lambda_1(t) > 0$ for $\forall t \in \mathcal{R}$. Let $\mathrm{W}_P \in \Omega$ be the true underlying unmixing matrix. Then

i) Under the identifiability conditions, the estimator $\hat{\mathbf{W}}$ of W_P defined by (3) is consistent, that is,

$$||\hat{\mathbf{W}} - \mathrm{W}_P|| = o_P(1). \quad (5)$$

ii) If the first and second derivative functions of ρ_1, which are denoted by $\dot{\rho}_1$ and $\ddot{\rho}_1$ separately, are both bounded, and $E||\mathbf{S}||^2 < \infty$, then $\hat{\mathbf{W}}$ is a $\sqrt{n}$-consistent estimate of W_P, that is,

$$\sqrt{n}||\hat{\mathbf{W}} - \mathrm{W}_P|| = O_p(1). \quad (6)$$

iii) Under the same conditions as for ii), $\hat{\mathbf{W}}$ is asymptotically normal, i.e.,

$$\sqrt{n}(\hat{\mathbf{W}}\mathrm{W}_P^{-1} - \mathrm{I}) \to_d \mathcal{N}(0,\Sigma_P)$$

where $\Sigma_P = \mathrm{cov}(\boldsymbol{\delta})$. Here, $\boldsymbol{\delta}$ is an $m\times m$ matrix of random variables, and its elements are decided by the following equations: For $1 \le k \ne j \le m$

$$\begin{bmatrix} F_{kj} & G_{kj} \\ G_{jk} & F_{jk} \end{bmatrix} \begin{bmatrix} \delta_{kj} \\ \delta_{jk} \end{bmatrix} = \begin{bmatrix} h_{kj}(\mathbf{S}) \\ h_{jk}(\mathbf{S}) \end{bmatrix}$$
$$\text{and} \quad \delta_{kk} = -\sum_{1\le j\le m, j\ne k} \omega_{kj}\delta_{kj}$$

where ω_{kj} is the (k,j)th entry of $\mathrm{W}_P\mathrm{W}_P^T$. The explicit formulae for F_{kj}, G_{kj}, and h_{kj} are given in the supplement [12].

Note that Σ_P can be consistently estimated by estimating ω_{kj}, F_{kj}, and G_{kj} for $1 \le k \ne j \le m$ and the variance-covariance matrix of $\{h_{kj}(\mathbf{S}) : 1 \le k \ne j \le m\}$. Having done this, one can put confidence bands on $\hat{\mathbf{W}}$. Alternatively, the normality result shows that one can use the nonparametric bootstrap distribution of $\hat{\mathbf{W}} - \mathrm{W}_P$ for this purpose.

Theorem 1 is a special case of Theorem 2 below in the case of no additive noise. The complete proof of Theorem 1 is in [12]. Here is the outline of our proof.

Outline of the Proof of i): The main idea is that the contrast function $\hat{\Delta}_\lambda(\cdot)$ can be expressed as a U-process[1] indexed by $\mathrm{W} \in \Omega$. Then, we can apply the Uniform Law of Large Numbers (ULLN) for this U-process [2] such that

$$\sup_{\mathrm{W}\in\Omega} |\hat{\Delta}_\lambda(\mathrm{W}) - \Delta_\lambda(\mathrm{W})| = o_P(1). \tag{7}$$

Then, since W_P can be identified uniquely in Ω by Δ_λ, the consistency of $\hat{\mathbf{W}}_P$ follows from the same arguments of compactness and continuity as for classical likelihood inference.

Outline of the Proof of ii) and iii): The key is to parametrize the 1-D curves from W_P to $\hat{\mathbf{W}}$ indexed by corresponding tangent vectors (a compact set) using the manifold structure of Ω (each row is on a m-D unit ball). Then, by making use of a second-order Taylor expansion, $\|\hat{\mathbf{W}} - \mathrm{W}_P\|$ can be expressed as a ratio of two terms indexed by U-processes separately. Then, the convergence rate can be obtained by applying ULLN and the central limit theorem for the U-processes in the denominator and numerator. Asymptotic normality follows similarly by the delta method. Note that $\hat{\mathbf{W}}$ is not an M estimate so that the analysis using the U processes is needed.

Although different choices of λ_1 can lead to different performance, taking λ_1 Gaussian or Laplace seems to be a good choice. Fig. 1 shows some further simulation results by using MATLAB, where the above characteristic function-based ICA method was implemented by an algorithm, called PCFICA, described in Section III-B. Its performance is compared with that of several other ICA algorithms such as FastICA [22], JADE [10], KGV [3], and EFFICA [11], where EFFICA has been proven to be asymptotically efficient under moderate conditions. Eight hidden sources were used, which were generated from $\mathcal{N}(0,1)$, exponential (1), t-distribution (3), lognormal (0,1), $\mathrm{t}(5)$, logistic (0,1), Weibull (1,1), and exponential$(10) + \mathcal{N}(0,1)$, independently. An 8×8 mixing matrix was generated randomly, the sample size used was 1000, and the experiment was replicated 100 times with the same mixing matrix to obtain the boxplots, where the estimation errors $d(\hat{\mathbf{W}}, \mathrm{W}_P)$ were measured by the so-called "Amari error" [3], i.e., for two $m \times m$ matrices U, V

$$d(\mathrm{U},\mathrm{V}) = \frac{1}{m}\left(\sum_{i=1}^{m}\frac{\sum_{j=1}^{m}|B_{ij}|}{\max_j |B_{ij}|} + \sum_{j=1}^{m}\frac{\sum_{i=1}^{m}|B_{ij}|}{\max_i |B_{ij}|}\right) - 2$$

where $\mathrm{B} = \mathrm{U}\mathrm{V}^{-1}$ (it is necessary to normalize each row of U and V). The simulation also suggests that the CHFICA estimator may not be efficient.

Note that the computational complexity for calculating $\hat{\Delta}_\lambda(\mathrm{W})$ directly from (4) is at least $O(n^2)$ [15]. For PCFICA,

[1]The U-process based on P and indexed by $\mathcal{F}$ is

$$U_m^n(f,P) := U_m^n(f) := \frac{(n-m)!}{n!}\sum_{(j_1,\ldots,j_m)\in I_m^n} f(X(j_1),\ldots,X(j_m))$$

$f \in \mathcal{F}$, where $I_m^n = \{(j_1,\ldots,j_m) : j_k \in \{1,\ldots,n\}$ and $j_k \neq j_l$, for $k \neq l\}$.

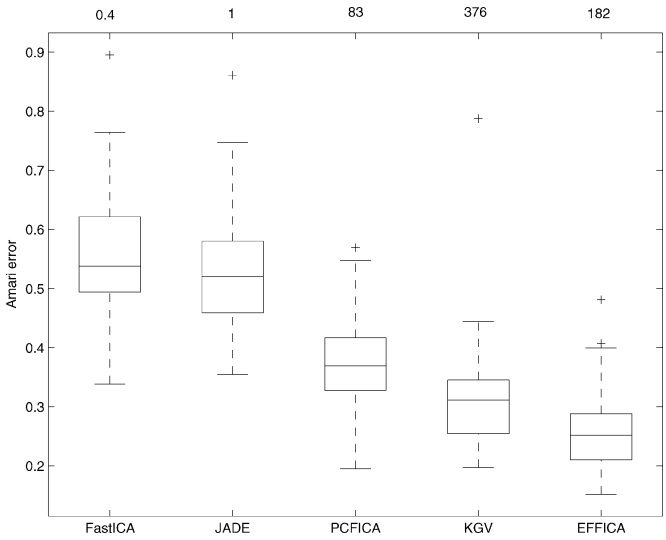

Fig. 1. Performances of different ICA algorithms: $m = 8$ hidden sources and $n = 1000$ sample size were used, and the experiment was replicated 100 times; the boxplots of Amari errors were based on quartiles and the numbers above the boxplots were the average running time (seconds per experiment) of the corresponding algorithms.

we used prewhitening as a preprocessing tool and the incomplete Cholesky decomposition to speed up the computation (see Section III-B).

C. Robustness Against Small Additive Noise

In practice, the model (1) rarely holds exactly. A more realistic situation allows some additive noise. That is, the observation $\mathbf{X}$ can be modeled as

$$\mathbf{X} = \mathrm{A}\mathbf{S} + r\mathbf{n} \tag{8}$$

where A and $\mathbf{S}$ are the same as in the previous sections, $\mathbf{n}$ is an $m\times 1$ random vector, independent of $\mathbf{S}$, standing for an additive noise vector (for example, sensor noise), and r is the magnitude of additive noise. This is usually called the noisy ICA model. In [21], there is a good review of studies of thess types of models. Our objective here is to study how the cf.-based ICA method behaves in the presence of noise. We borrow Bickel and Doksum [4]'s approach in their study of the robustness of Box–Cox transformations. To be more precise, we assume a large sample size n, and further, $r = r(n) \to 0$ as $n \uparrow \infty$. We study how the estimation error behaves in relation to its natural scale $1/\sqrt{n}$ and $r(n)$.

Theorem 2: Suppose that λ satisfies the conditions of Theorem 1. Then, as $r(n) \to 0$, we have the following.

i) $\|\hat{\mathbf{W}} - \mathrm{W}_P\| = o_P(1)$.

ii) If further $E\|\mathbf{S}\|^2 < \infty$ and $E\|\mathbf{n}\|^2 < \infty$, then

$$\|\hat{\mathbf{W}} - \mathrm{W}_P\| = O_P(n^{-(1/2)} + r(n)). \tag{9}$$

The idea of our proof is similar to that outlined for Theorem 1 in Section II-B. We refer to [12] for the complete proof.

Part i) of Theorem 2 is trivial. Part ii) says that as long as $r(n)$ is of smaller order than $n^{-(1/2)}$, the effect of the noise on $\hat{\mathbf{W}}$ is minimal. Put another way, if the ratio of noise scale to the estimation error is small, the ratio of $(\mathrm{Bias})^2/\mathrm{Var}$ for estimation of W_P is also small. Fig. 2 shows a simulation study to illustrate

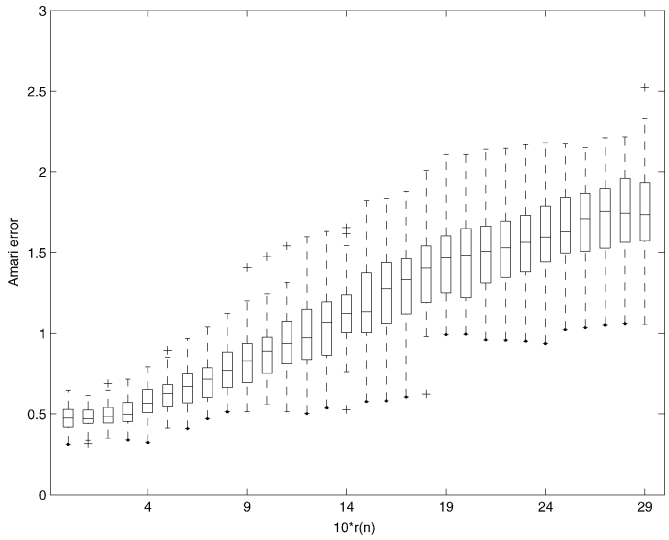

Fig. 2. Performance of PCFICA w.r.t. different noise scales, where i.i.d. $\mathrm{N}(0,1)$ additive noise and the previous $\mathrm{m} = 8$ hidden sources were used to generate noisy mixtures by (8); $\mathrm{n} = 1000$, and each experiment is replicated 100 times to obtain the boxplots; the first boxplot corresponds to no noise, i.e., $\mathrm{r(n)} = 0$, the second boxplot $\mathrm{r(n)} = 0.1$, and so on.

Theorem 2(ii), where the eight hidden sources above were used with $n = 1000$, and without loss of generality, the identity matrix was used for the mixing matrix. Each boxplot, which was obtained by 100 replications, corresponds to a different noise scale $r(n)$. By comparing the medians, the slope before $r(n)$ on the right-hand side of (9) is about 0.4. The variance term (about 0.5 by the first boxplot, which corresponds to $r(n) = 0$), can be roughly interpreted as the signal standard deviation (about $\sqrt{15}$) times $n^{-(1/2)}$. In this example, the error due to additive noise will dominate the estimation error only if $0.4r(n) > 0.5$, i.e., $r(n) > 1.25$.

Both the simulation and the theorem say that CHFICA can provide fairly good estimates of W_P, even in the presence of small additive noise in model (1). It turns out that the bias term due to additive noise is not sensitive to the sample size n nor to the exact distribution of the additive noise for small $r(n)$, but we leave out further simulations due to space limitations. Because of its global consistency and robustness properties, CHFICA can serve as a good starting point for the more efficient techniques of Chen and Bickel [11].

III. New Prewhitened ICA Algorithm

Prewhitening is a popularly used preprocessing technique in the ICA literature. For example, many famous ICA algorithms such as FastICA [22], JADE [10], KCCA, KGV [3] have used this preprocessing technique. In general, prewhitening is expected to be valid when all hidden sources have finite second moments. In this section, we first briefly review this concept and then provide a new algorithm for implementing the characteristic function-based ICA method. Delicate studies of prewhitening and prewhitened CHFICA (PCFICA) are left to Section IV.

A. Introduction to Prewhitening

Since the matrix W for model (1) can be arbitrary, naively, we have to optimize some constrast function, for example, (4), over all $m \times m$ nonsingular matrices to obtain an estimate. However, prewhitening can project the optimization onto the Stiefel manifold of orthogonal matrices [21]. Optimization on a Stiefel manifold can be solved efficiently [14]. Let $\Sigma_{\mathbf{x}} = \mathrm{cov}(\mathbf{X})$, and let $\Sigma_{\mathbf{x}}^{1/2}$ be the square root matrix of $\Sigma_{\mathbf{x}}$ obtained by Singular Value Decomposition (SVD), i.e., let $\Sigma_{\mathbf{x}} = \mathrm{UDU}^T$ be the SVD decomposition such that $\mathrm{UU}^T = \mathrm{I}_{m\times m}$, which is an identity matrix, and D is a diagonal matrix comprised of $\Sigma_{\mathbf{x}}$'s eigenvalues; then, $\Sigma_{\mathbf{x}}^{1/2} = \mathrm{UD}^{1/2}\mathrm{U}^T$. Let $\mathbf{Y} = \Sigma_{\mathbf{x}}^{-(1/2)}\mathbf{X}$. Then, $\mathrm{cov}(\mathbf{Y}) = \mathrm{I}_{m\times m}$, and (1) is equivalent to $\mathbf{S} = \mathrm{W}\Sigma_{\mathbf{x}}^{1/2}\mathbf{Y}$. Without loss of generality, we may assume that each hidden source in $\mathbf{S}$ has unitary variance (we assume currently a finite variance for each source and lose this assumption later). By considering the covariance matrix, we have $\mathrm{W}\Sigma_{\mathbf{x}}\mathrm{W}^T = \mathrm{I}_{m\times m}$, and thus

$$\mathrm{O} = \mathrm{W}\Sigma_{\mathbf{x}}^{1/2} \tag{10}$$

must be an orthogonal matrix. Notice that

$$\mathbf{Y} = \mathrm{O}^T\mathbf{S}$$

is still an ICA model, but restricting to orthogonal matrices is computationally very advantageous. Since $\Sigma_{\mathbf{x}}$ can be estimated directly by the sample covariance matrix of $\mathbf{X}$

$$\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}(j) - \overline{\mathbf{X}})^T(\mathbf{X}(j) - \overline{\mathbf{X}})$$

where $\overline{\mathbf{X}} = 1/n\sum_{k=1}^{n}\mathbf{X}(k)$, a prewhitened ICA algorithm first estimates an orthogonal unmixing matrix O (say $\hat{\mathbf{O}}$) by fitting the ICA model with input $\hat{\mathbf{Y}}(j) = \hat{\Sigma}_{\mathbf{x}}^{-(1/2)}\mathbf{X}(j)$ in some way and then estimates the unmixing matrix by $\hat{\mathbf{W}} = \hat{\mathbf{O}}\hat{\Sigma}_{\mathbf{x}}^{-(1/2)}$ because of (10). This is the so-called prewhitening (or whitening) technique. Note that $\hat{\Sigma}_{\mathbf{x}}$ is the efficient estimate of the variance-covariance matrix of $\mathbf{X}$ if and only if the distribution of $\mathbf{X}$ (thus $\mathbf{S}$) is Gaussian so that the resulting estimate is not efficient. However,if both $\mathbf{O}$ and $\Sigma_{\mathbf{x}}$ are estimated $\sqrt{n}$-consistently, this procedure will lead to a $\sqrt{n}$-consistent estimate of W.

B. New Algorithm for CHFICA Using Prewhitening

Here is our prewhitened CHFICA estimator (PCFICA): First, estimate O defined by (10) using

$$\hat{\mathbf{O}} = \mathrm{argmin}_{\mathrm{O}\in\mathcal{O}(m)}\tilde{\Delta}_{\lambda}(\mathrm{O}) \tag{11}$$

where $\tilde{\Delta}_{\lambda}$ is the same as $\hat{\Delta}_{\lambda}$ defined in (3), except that $\{\mathbf{X}(j)\}$ is now replaced by $\{\hat{\mathbf{Y}}(j)\}$, and $\mathcal{O}(m)$ is the set of $m \times m$ orthogonal matrices; second, let

$$\hat{\mathbf{W}} = \hat{\mathbf{O}}\hat{\Sigma}_{\mathbf{x}}^{-(1/2)} \tag{12}$$

and obtain an estimate of W_P by $[\hat{\mathbf{W}}]_\Omega$. The key problem is the optimization of (11). Algebraic expansion leads to

$$\tilde{\Delta}_\lambda(\mathrm{O}) = \frac{1}{n^2}\sum_{j,l=1}^{n}\rho_m(\mathrm{O}[\hat{\mathbf{Y}}(j)-\hat{\mathbf{Y}}(l)]) - \frac{2}{n^{m+1}}\sum_{l=1}^{n}\prod_{k=1}^{m}\left\{\sum_{j=1}^{n}\rho_1(O_k[\hat{\mathbf{Y}}(j)-\hat{\mathbf{Y}}(l)])\right\} + \frac{1}{n^{2m}}\prod_{k=1}^{m}\left\{\sum_{j,l=1}^{n}\rho_1(O_k[\hat{\mathbf{Y}}(j)-\hat{\mathbf{Y}}(l)])\right\}. \quad (13)$$

This formula was used by Kankainen [24] in the context of testing total independence. Evaluation of this contrast function requires $O(n^2)$ operations, which is computationally impractical with large sample sizes. We provide an algorithm to approximate this function by using the Gaussian kernel and incomplete Cholesky decomposition, which makes the computation feasible for reasonable sizes of n.

Let $\lambda_1(t_1) = (1/\sqrt{2\pi})\exp(-(1/2)t_1^2)$. Then, $\rho_m(\mathbf{t}) = \exp(-(1/2)\|\mathbf{t}\|^2)$, for $\mathbf{t} \in \mathcal{R}^m$. Notice that $\|\mathrm{O}[\hat{\mathbf{Y}}(i) - \hat{\mathbf{Y}}(j)]\|^2 = \|\hat{\mathbf{Y}}(i) - \hat{\mathbf{Y}}(j)\|^2$ so that the first term on the right-hand side of (13) does not depend on O. We only need to consider the second term and last term.

Define $\mathbf{G}^k$ (for $k = 1, \ldots, m$) to be an $n \times n$ matrix such that its (j,l)th entry $(\mathbf{G}^k)_{jl} = \exp(-(1/2)(O_k[\hat{\mathbf{Y}}(j) - \hat{\mathbf{Y}}(l)])^2)$. This is a Gram matrix generated by a 1-D Gaussian kernel, which is known to be non-negative definite. Let $\mathbf{G}^k_{+j}$ be the sum of the jth column of $\mathbf{G}^k$ and $\mathbf{G}^k_{++} = \sum_{j=1}^n \mathbf{G}^k_{+j}$ (sum of all entries of $\mathbf{G}^k$). Then, the contrast function consisting of the second term and the third term on the right-hand side of (13) becomes

$$f(\mathrm{O}) = -2\frac{1}{n^{m+1}}\sum_{j=1}^{n}\prod_{k=1}^{m}\mathbf{G}^k_{+j} + \frac{1}{n^{2m}}\prod_{k=1}^{m}\mathbf{G}^k_{++}.$$

To reduce computational complexity, we propose to approximate $\mathbf{G}^k$ by $\tilde{\mathbf{G}}^k = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$ using incomplete Cholesky (IC) decomposition [18], where $\tilde{\mathbf{L}}$ is a $n \times h_n(k)$ lower triangular matrix. Let ε be the threshold value that controls the summation of the remaining diagonals by IC. Permutation is necessary to make h_n optimal such that $\mathrm{tr}(\mathrm{P}\mathbf{G}^k\mathrm{P}^T - \tilde{\mathbf{G}}^k) \le \varepsilon$, for some permutation matrix P, which is chosen automatically by the IC algorithm. We used the IC algorithm described in [3] and set $\varepsilon = 0.01$ for simulations. Now, the complexity of evaluting the approximate $f(\mathrm{O})$ is only $O(nh_n^2)$, where $h_n = \max\{h_n(k) : 1 \le k \le m\}$. The partial derivative $\partial f/\partial \mathrm{O}$ can also be approximated in $O(nh_n^2)$ operations by using the approximate Gram matrices.

Although h_n is not less than the number of eigenvalues of $\mathbf{G}^k$ such that the remaining eigensum is controlled by ε, Wright [32] showed that a non-negative definite matrix can be approximated well by IC with the same order as that by spectral decomposition if its eigenvalues decrease quickly enough. See Bach and Jordan [3] for empirical studies, and, for instance, to Widom [31] for theoretical results on the rate of decay of the eigenvalues of such Gram matrices. To further understand the magnitude of h_n needed by IC to achieve a given error ε, we generated 1000 random samples $\{z_j : 1 \le j \le 1000\}$ from the mixture Gaussian distribution $0.4\mathcal{N}(-1,1) + 0.6\mathcal{N}(1,1)$ and obtained

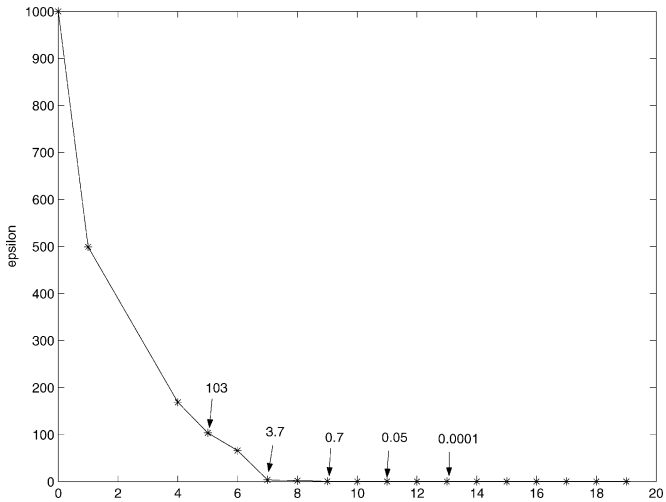

Fig. 3. Report the magnitude of h_n (x-axis) chosen by incomplete Cholesky decomposition with given approximation errors ϵ (y-axis), where the Gram matrix was generated by the 1-D Gaussian kernel and 1000 samples from $0.4\mathcal{N}(-1,1) + 0.6\mathcal{N}(1,1)$.

a 1000×1000 Gram matrix $\mathbf{G}$ with $G_{ij} = \exp(-(1/2)(\tilde{z}_j - \tilde{z}_l)^2)$, where $\tilde{z}_j = z_j/std(z)$, and std(z) is the sample standard deviation of z. Fig. 3 visualizes the IC simulation results.

Since applying the permutation matrix P is equivalent to permuting the order of $\{\mathbf{X}(j) : 1 \le j \le m\}$, which does not change the value of $f(\mathrm{O})$, for simplicity, we write $P\mathbf{G}^kP^T$ as $\mathbf{G}^k$. It is noticeable that $\mathbf{G}^k - \tilde{\mathbf{G}}^k$ is still non-negative definite. Thus, $0 \le \mathbf{G}^k_{+j} - \tilde{\mathbf{G}}^k_{+j} \le \varepsilon$ for $j = 1, \ldots, n$ and $0 \le \mathbf{G}^k_{++} - \tilde{\mathbf{G}}^k_{++} \le n\varepsilon$. Since $\mathbf{G}^k_{+j} = O_P(n)$ and $\mathbf{G}^k_{++} = O_P(n^2)$, the approximation error of $f(\mathrm{O})$ by replacing $\mathbf{G}^k$ with $\tilde{\mathbf{G}}^k$ is bounded by $O_P(\varepsilon/n)$ and, thus, is ignorable with $\varepsilon = O(1)$ due to $f(\mathrm{O}) = O_P(1)$ when at most one source has infinite variance. When all sources have heavy tails, we suggest $\varepsilon = O(1/n)$ by considering the convergence rate of the contrast function (13) (see [12] for such convergence rates).

Once we can evaluate the contrast function $f(\mathrm{O})$ and its partial derivative $\partial f/\partial \mathrm{O}$, the minimization of $f(\mathrm{O})$ in the domain of orthogonal matrices can be done efficiently by using the gradient algorithm described in [14]. Here, the gradient of $f(\mathrm{O})$ is defined by

$$\bigtriangledown f = \frac{\partial f}{\partial \mathrm{O}} - \mathrm{O}\left[\frac{\partial f}{\partial \mathrm{O}}\right]^T \mathrm{O}$$

and the geodesic starting from O in the gradient $\bigtriangledown f$ is determined as

$$G_{\mathrm{O},\bigtriangledown f}(t) = \mathrm{O}\exp(t\mathrm{O}^T \bigtriangledown f)$$

where the matrix exponential can be calculated efficiently by diagonalization. It is noticeable that the contrast function is not convex, mainly due to the identifiability ambiguity. Restarting initial points is necessary to obtain the global minimizer of $f(\mathrm{O})$, which is closest to the truth.

IV. Statistical Properties of Prewhitened CHFICA

Since we are, in this paper, interested in $\mathrm{W}_P \in \Omega$, we need to estimate W_P by $[\hat{\mathbf{W}}]_\Omega$, where $\hat{\mathbf{W}}$ is an estimator obtained by any prewhitened ICA algorithm described in Section III-A. It is

clear that $\Omega_n = \{[\mathrm{O}\hat{\Sigma}_{\mathbf{x}}^{-(1/2)}]_\Omega : \mathrm{O} \in \mathcal{O}(m)\}$ is the set of all possible values of $[\hat{\mathbf{W}}]_\Omega$. We call Ω_n the acting parameter space for the estimation of W_P using prewhitening.

When $E\|\mathbf{S}\|^2 < \infty$, by classical theories, $\hat{\Sigma}_{\mathbf{x}} \to \Sigma_{\mathbf{x}}$ almost surely. Then, we can approximate W_P in Ω_n. It can be shown that for the prewhitened CHFICA algorithm, if all sources have finite second moments, the properties claimed in Theorem 1 continue to hold. However, when $E\|\mathbf{S}\|^2 = \infty$, i.e., some sources have heavy tails, $\hat{\Sigma}_{\mathbf{x}}$ diverges, and how W_P is approximated in Ω_n is unclear. One would think that if one or more sources do not have finite second moments, then prewhitening would cause a breakdown of these prewhitened ICA algorithms. To our surprise, simulations of Kernel ICA (KGV) [3] and prewhitened CHFICA gave excellent results, even when there exist heavy-tailed sources (for example, one is uniformly distributed on [0,1], and the other is Cauchy distributed). This is different from the super-efficiency phenomena for i.i.d. heavy-tailed sources studied in [30]. The following subsections develop some statistical theory for this phenomenon.

A. Consistent Acting Space

In this section, we prove that with prewhitening, the acting parameter space is consistent in the sense of Theorem 3, regardless of whether all hidden sources have finite second moments or not.

Theorem 3: Under the identifiability conditions

$$d(\mathrm{W}_P, \Omega_n) \to_P 0$$

where $d(\mathrm{W}_P, \Omega_n) = inf_{\mathrm{W}\in\Omega_n}\|\mathrm{W}_P - \mathrm{W}\|$.

Proof: Suppose that $\mathbf{S}$ has sample covariance matrix $\hat{\Sigma}_{\mathbf{s}}$. Then, $\hat{\Sigma}_{\mathbf{s}} = \mathrm{W}_P\hat{\Sigma}_{\mathbf{x}}(\mathrm{W}_P)^T$ since $\mathbf{S} = \mathrm{W}_P\mathbf{X}$. Thus, $\mathbf{O}_n \equiv \hat{\Sigma}_{\mathbf{s}}^{-(1/2)}\mathrm{W}_P\hat{\Sigma}_{\mathbf{x}}^{1/2}$ is orthogonal, and $[\mathbf{O}_n\hat{\Sigma}_{\mathbf{x}}^{-(1/2)}]_\Omega \in \Omega_n$. Hence, it is enough to prove that

$$\|\mathrm{W}_P - [\hat{\Sigma}_{\mathbf{s}}^{-(1/2)}\mathrm{W}_P]_\Omega\| \quad \to_P \quad 0.$$

This is completed by Theorem 5, which we have developed and proven in the Appendix. ■

This theorem says that for all kinds of hidden sources, there exists at least a sequence of points in the acting parameter space with prewhitening, which converges to W_P in probability. This result is independent of the particular ICA algorithm.

B. Consistent Prewhitened CHFICA

The previous subsection provides the possibility that some prewhitened ICA algorithm may be able to obtain consistent estimates of the unmixing matrix, even with the existence of heavy-tailed sources. This begs the question of whether an implemented algorithm is consistent. Our goal in this subsection is to study the prewhitened CHFICA method. Fig. 4 shows some simulation results in the case of two sources: One has a uniform distribution on [0,1], and the other is Cauchy distributed. To detect whether ICA algorithms can obtain consistent estimates in such a situation, the sample size was increased from I ($n = 1000$) to II ($n = 8000$). We compare PCFICA with three other ICA algorithms: FastICA, JADE, and KGV. From Fig. 4, we can see that as the sample size increases, the estimation error measured by the Amari error for PCFICA and KGV decreases toward zero more significantly than that for FastICA and JADE. However, the simulation also suggests that the convergence rate of the PCFICA estimator is slower than $n^{-1/2}$.

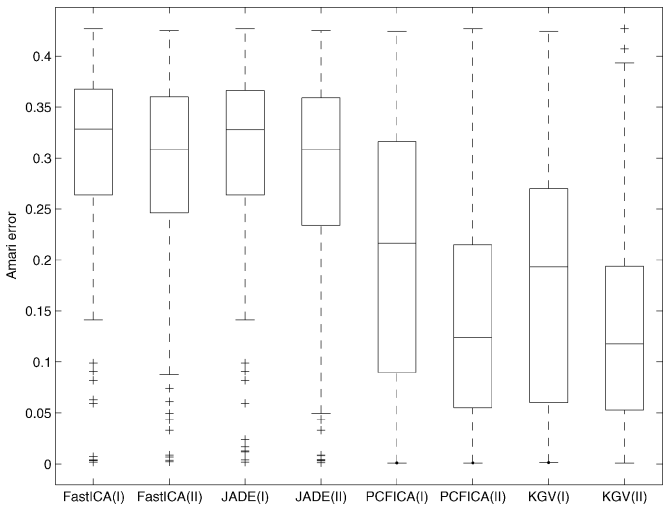

Fig. 4. Consistency of different ICA algorithms with prewhitening when $m = 2$: One is uniform on [0, 1], and the other is Cauchy: One hundred replications were used to obtain the boxplots based on quartiles, where the sample sizes were 1000 for case I and 8000 for case II.

The main result of this study is given in the following Theorem.

Theorem 4: Suppose that the identifiability conditions hold in model (1). Let $\mathrm{W}_P \in \Omega$ be as usual. The estimator of $\mathrm{W}_P \in \Omega$ defined by [(11) and (12)] is consistent, i.e., $\|[\hat{\mathbf{W}}]_\Omega - \mathrm{W}_P\| \to_P 0$, in either of the three cases.

i) All components of $\mathbf{S}$ have finite variances.
ii) All but one component of $\mathbf{S}$ have finite variances.
iii) $m = 2$, and both components of $\mathbf{S}$ have heavy tails and stable distributions.

We provide some heuristics for the above results. The crucial step of a prewhitened ICA algorithm is to separate, by Theorem 3, the whitened mixture $\hat{\mathbf{Y}} = \mathbf{O}_n^T\hat{\Sigma}_{\mathbf{s}}^{-(1/2)}\mathbf{S}$, where $\mathbf{O}_n$ is a data-dependent orthogonal matrix defined in the previous subsection and, thus, is equivalent to separating $\tilde{\mathbf{S}} = \hat{\Sigma}_{\mathbf{s}}^{-(1/2)}\mathbf{S}$ for which the identity matrix is a true unmixing matrix. By Theorem 5 below, $\hat{\Sigma}_{\mathbf{s}}^{-(1/2)}$ is almost diagonal, and thus, by prewhitening, we essentially separate mixtures of individually rescaled hidden sources, which are still mutually independent, despite their weak time dependence. The result for Case i) becomes obvious. In Case ii), the heavy-tailed rescaled source is asymptotically zero; fortunately, the orthogonal unmixing structure makes sure that the one (almost zero) can be separated well if the other $m - 1$ sources can be separated consistently. Thus, in Cases i) and ii), the consistency results can also apply to other prewhitened ICA algorithms, which can estimate W_P consistently without using prewhitening. The simulation shown in Fig. 4 is for Case ii). Zibulevsky and Pearlmutter [33] had some different heuristics for Case ii) by considering the sparseness property of heavy-tailed hidden sources. Case iii) is more sophisticated. See [12] for the complete proof. As a special case of iii), when both components have the same symmetric and stable distribution, Shereshevski *et al.* [30] showed that without prewhitening, the unmixing matrix can be estimated super efficiently, for example, by a quasi maximum likelihood estimation (MLE). Under such a situation, we conjecture that the corresponding estimates using prewhitening would also be super-efficient, but we leave this for further analysis.

V. Conclusion

In this paper, we have analyzed the CHFICA method both theoretically and numerically under the setup of classical ICA models. First, the CHFICA estimate is consistent under minimal identifiability conditions, whereas many well-known ICA algorithms such as FastICA and JADE, which are based on parametric methods and thus can be unified under the framework of quasi MLE or equivalently by parametrizing hidden sources with particular density families (see Lee *et al.* [25]), are shown by Cardoso [9] to be inconsistent when the hidden sources do not belong to such families. Second, CHFICA is $\sqrt{n}$-consistent, asymptotically normal, and robust against small additive noise under mild conditions. Third, the acting parameter space for prewhitened ICA algorithms is shown to always be able to capture the true value of the unmixing matrix asymptotically, and in particular, prewhitened CHFICA is consistent even in the presence of heavy-tailed sources. Numerically, although the computational complexity of CHFICA is $O(n^2)$, we have proposed a fast algorithm (PCFICA) to implement it by using prewhitening and incomplete Cholesky decomposition. Simulation results of PCFICA are in agreement with the above theoretical analysis.

Appendix

Let $\hat{\Sigma}_{\mathbf{s}}$ be the sample covariance matrix of an m-D random vector $\mathbf{S}$, which has mutually independent components. Denote its diagonal elements by $\{\sigma_{i,n}^2\}_{i=1}^m$ with $\sigma_{i,n} \geq 0$, and define $\Gamma_n = diag(\{\sigma_{i,n}\}_{i=1}^m)$.

Theorem 5: If none of $\mathbf{S}$'s component is degenerate, then

$$\Gamma_n^{-1}\hat{\Sigma}_{\mathbf{s}}^{1/2} \rightarrow_P \mathrm{I}_{m\times m}.$$

Proof: Suppose the sample correlation matrix of $\mathbf{S}$ is $\mathbf{R}_n = [r_{ij,n}]_{i,j=1}^m$. Then, $\hat{\Sigma}_{\mathbf{s}} = \Gamma_n \mathbf{R}_n \Gamma_n$. Let $\mathbf{R}_n^0$ be the matrix such that $(\mathbf{R}_n^0)_{ij} = (\mathbf{R}_n)_{ij}$ for $i \neq j$ and $(\mathbf{R}_n^0)_{ii} = 0$ for $1 \leq i \leq m$. Then, $\hat{\Sigma}_{\mathbf{s}} = \Gamma_n^2 + \eta_n \Delta_n$, where $\Delta_n = \Gamma_n \mathbf{R}_n^0 \Gamma_n/\eta_n$ and $\eta_n = ||\mathbf{R}_n^0||$. It is clear that $||\Gamma_n^{-1}\Delta_n\Gamma_n^{-1}|| = \mathbf{R}_n^0/\eta_n$. From the following Proposition, we have $\mathbf{R}_n \rightarrow_P \mathrm{I}_{m\times m}$, and thus, $\eta_n = o_P(1)$. By Mathias's theorem [25], we have $||\Gamma_n^{-1}\hat{\Sigma}_{\mathbf{s}}^{1/2} - \mathrm{I}_{m\times m}|| \leq \eta_n + O_P(\eta_n^2) = o_P(1)$. ∎

Let (X, Y) be independent. $\{X_i, Y_i\}_{i=1}^n$ are i.i.d. copies of them. Define the sample correlation coefficient

$$r_n = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \overline{X}\,\overline{Y}}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2\right)\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \overline{Y}^2\right)}}.$$

When $EX^2 < \infty$ and $EY^2 < \infty$, it is well known that $r_n \rightarrow_P 0$.

Proposition: Let (X, Y) be independent. Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d. realizations of them. Then, $r_n = o_P(1)$.

Proof: When both X and Y are finite second moment, the result is well-known (see, for example, Bickel and Doksum [5]). If either is infinite, the result follows from the following Lemmas 1. ∎

Lemma 1 (Klass): If $EX^2 = \infty$, then

$$r_{n1} = \frac{\overline{X}}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}} \quad \rightarrow_P \quad 0.$$

Thus, $\overline{X}/\sqrt{(1/n)\sum_{i=1}^n X_i^2 - \overline{X}^2} = o_P(1)$ by monotone transformation.

Proof: It is enough to consider the case $X \geq 0$.

Notice that $(1/n)\sum_{i=1}^n X_i^2 \rightarrow \infty$ a.s.; then, there exists $a_n \uparrow \infty$ s.t. $\sum_{i=1}^n X_i^2/na_n^2 \rightarrow \infty$ a.s. By truncation

$$\begin{aligned} r_{n1} &= \frac{\frac{1}{n}\sum_{i=1}^n X_i I(X_i \leq a_n)}{\sqrt{\frac{1}{n}\sum X_i^2}} + \frac{\sum_{i=1}^n X_i I(X_i > a_n)}{\sqrt{n\sum_{i=1}^n X_i^2}} \\ &\leq \frac{a_n \sum_{i=1}^n \frac{I(X_i \leq a_n)}{n}}{\sqrt{\sum_{i=1}^n \frac{X_i^2}{n}}} + \sqrt{\frac{\sum_{i=1}^n I(X_i > a_n)}{n}} \\ &\leq \frac{1}{\sqrt{\frac{\sum_{i=1}^n X_i^2}{na_n^2}}} + o_P(1), [P(X > a_n) = o(1)] = o_P(1). \end{aligned}$$

∎

Lemma 2: If X and Y are independent, and $EX^2 = \infty$, then

$$r_{n2} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)}} = o_P(1).$$

This, together with Lemma 1, implies that if further Y is not degenerated, then

$$\frac{1}{n}\frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2\right)\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \overline{Y}^2\right)}} = o_P(1).$$

Proof: It is enough to prove the result for non-negative X and Y. As in the proof of Lemma 1, we have $a_n \uparrow \infty$ and $\sum_{i=1}^n X_i^2/na_n^2 \rightarrow \infty$ a.s.

$$\begin{aligned} r_{n2} &= \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i I(X_i \leq a_n)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)}} + \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i I(X_i > a_n)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)}} \\ &\leq \frac{a_n \sum_i Y_i I(X_i \leq a_n)}{\sqrt{\left(\sum_i X_i^2\right)\left(\sum_i Y_i^2\right)}} + \frac{\sqrt{\sum_i X_i^2}\sqrt{\sum_i Y_i^2 I(X_i > a_n)}}{\sqrt{\left(\sum_i X_i^2\right)\left(\sum_i Y_i^2\right)}} \\ &\leq \frac{a_n\sqrt{\sum_i I(X_i \leq a_n)}}{\sqrt{\sum_i X_i^2}} + \sqrt{\frac{\sum_i Y_i^2 I(X_i > a_n)}{\sum_i Y_i^2}} \\ &\leq o_P(1) + o_P(1) \end{aligned}$$

where the second $o_P(1)$ in the last inequality is due to

$$E\left[\frac{\sum_i Y_i^2 I(X_i > a_n)}{\sum_i Y_i^2}\right] = E\left[E\left(\frac{\sum_i Y_i^2 I(X_i > a_n)}{\sum_i Y_i^2}\middle| \{Y_i\}\right)\right] = P(X > a_n) = o(1).$$

■

ACKNOWLEDGMENT

The authors are very grateful to M. Klass for technical discussions and the proof of Lemma 1 and would like to thank four referees and the associate editor for very helpful comments.

REFERENCES

[1] S. Amari, "Independent component analysis and method of estimating functions," *IEICE Trans. Funda.*, vol. E85-A, no. 3, pp. 540–547, 2002.
[2] M. Arcones and E. Gine, "Limit theorems for U-processes," *Ann. Prob.*, vol. 21, no. 3, pp. 1494–1542, 1993.
[3] F. Bach and M. Jordan, "Kernel independent component analysis," *J. Machine Learning Res.*, vol. 3, pp. 1–48, 2002.
[4] P. Bickel and K. Doksum, "An analysis of transformations revisited," *J. Amer. Stat. Assoc.*, vol. 76, pp. 296–311, 1981.
[5] ——, *Mathematical Statistics*, Second ed. Upper Saddle River, NJ: Prentice-Hall, 2001, vol. I.
[6] P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag, 1993.
[7] R. Boscolo, H. Pan, and V. P. Roychowdhury, "Independent component analysis based on nonparametric density estimation," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 55–65, Jan. 2004.
[8] J. F. Cardoso, "On the performance of orthogonal source separation algorithms," in *Proc. EUSIPCO*, Edinburgh, U.K., 1994, pp. 776–779.
[9] ——, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
[10] ——, "High-order contrasts for independent component analysis," *Neural Comput.*, vol. 11, no. 1, pp. 157–192, 1999.
[11] A. Chen and P. J. Bickel, "Efficient Independent Component Analysis (II)," Univ, Calif,, Dept. Statist., Berkeley, CA, Tech. Rep. 645, 2003, submitted for publication.
[12] ——, "Supplement to "Consistent Independent Component Analysis and Prewhitening"," Univ. Calif., Dept. Statist., Berkeley, CA, Tech. Rep. 656, 2004.
[13] P. Comon, "Independent component analysis, A new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.
[14] A. Edelman, T. Arias, and S. Smith, "The geometry of algorithms with orghogonality constriants," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1999.
[15] J. Eriksson and V. Koivunen, "Characteristic-function based independent component analysis," *Signal Process.*, vol. 83, pp. 2195–2208, 2003.
[16] ——, "Identifiability, separability and uniqueness of linear ICA models," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 601–604, Jul. 2004.
[17] M. Girolami, *Advances in Independent Component Analysis*. New York: Springer-Verlag, 2000.
[18] G. Golub, *Matrix Computation*. Baltimore, MD: Johns Hopkins Univ. Press, 1996.
[19] T. Hastie and R. Tibshirani. (2002) Independent Component Analysis Through Product Density Estimation. Tech. Rep.. Stanford Univ., Dept. Stat.. [Online]. Available: www-stat.stanford.edu/~hastie/pub.htm
[20] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.
[21] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.
[22] A. Hyvarinen and E. Oja, "A fast fixed point algorithm for independent component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, 1997.
[23] A. Kagan, Y. Linnik, and C. Rao, *Characterization Problems in Mathematical Statistics*. NY: Wiley, 1973.
[24] A. Kankainen, "Consistent Testing of Total Independence based on the Empirical Characteristic Function," Ph.D. dissertation, Univ. Jyvaskyla, Jyvaskyla, Finland, 1995.
[25] T. W. Lee, M. Girolami, M. Bell, and R. Sejnowski, "A unifying information theoretic framework for independent component analysis," *Comput. Math. Appl.*, vol. 39, pp. 1–21, 2000.
[26] T. W. Lee, M. Girolami, and T. Sejnowski, "Independent component analysis using an extended informax algorithm for mixed subGaussian and superGaussian sources," *Neural Comput.*, vol. 11, no. 2, pp. 417–441, 1999.
[27] S. Makeig, M. Westerfield, T.-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Dynamic brain sources of visual evoked responses," *Science*, vol. 295, pp. 690–694, 2002.
[28] R. Mathias, "A bound for the matrix square root with application to eigenvector perturbation," *SIAM J. Matrix Anal. Appl.*, vol. 18, pp. 861–867, 1997.
[29] S. Roberts and R. Everson, *Independent Component Analysis – Principles and Practice*. Cambridge, UK: Cambridge Univ. Press, 2001.
[30] Y. Shereshevsk, A. Yeredor, and H. Messer, "Super-efficiency in blind signal separation of symmetric heavy-tailed sources," in *Proc. IEEE Workshop Stat. Signal Process.*, Singapore, Aug. 2001, pp. 78–81.
[31] H. Widom, "Asymptotic behavior of the eigenvalues of certain integral equations," *Trans. Amer. Math. Soc.*, vol. 109, no. 2, pp. 278–295, 1963.
[32] S. Wright, "Modified Cholesky factorizations in interior-point algorithms for linear programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 1159–1191, 1999.
[33] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Comput.*, vol. 13, no. 4, pp. 863–882, 1999.

Aiyou Chen received the B.S. degree in mathematics from Wuhan University, Wuhan, China, in 1997, the M.S. degree in probability and mathematical statistics from Peking University, Beijing, China, in 2000, and the Ph.D. degree in statistics from University of California, Berkeley, in 2004.

He is currently a Member of Technical Staff with Bell Laboratories, Lucent Technologies, Murray Hill, NJ. His research interests include nonparametric inference, independent component analysis, statistical learning, and applications in network tomography and sensor networks.

Peter J. Bickel is a Professor in the statistics department at University of California, Berkeley. His research spans a number of areas. In his work on semiparametric models [he is a co-author of the recent book *Efficient and Adaptive Estimation for Semiparametric Models* (New York: Springer, 1998)], he uses asymptotic theory to guide development and assessment of such models. His studies of hidden Markov models, which are important in such diverse fields as speech recognition and molecular biology, are directed toward understanding how well the method of maximum likelihood performs. He is also interested in the bootstrap, in particular, in constructing diagnostic measures to detect the malfunction of this technique. Recently, he has become involved in developing empirical statistical models for genomic sequences. He is a co-author of the well known book *Mathematical Statistics: Basic Ideas and Selected Topics (Upper Saddle River, NJ: Prentice-Hall, 2000).*

Prof. Bickel is past President of the Bernoulli Society and of the Institute of Mathematical Statistics, a MacArthur Fellow, and a member of the American Academy of Arts and Sciences and of the National Academy of Sciences.