Efficient Independent Component Analysis (I)

Aiyou Chen aychen@stat.berkeley.edu
Peter Bickel bickel@stat.berkeley.edu
Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: TBA

Abstract

In this paper we propose a Fisher Efficient estimator in the model of independent component analysis (ICA). First, we provide a $\sqrt{n}$-consistent estimator using the empirical characteristic function, and then, show that by directly estimating the efficient influence function, we can construct a one-step maximum likelihood estimate (MLE) which reaches asymptotic Fisher efficiency (EFFICA). We compare a variant of EFFICA to standard and state-of-the-art algorithms such as the Kernel ICA method (Bach & Jordan 2002), using benchmark simulations to exhibit its excellent performance.

Keywords: Empirical Characteristic Function, Influence Function, Fisher Information, Semiparametric Model, One-step MLE, Asymptotic Efficient

1. Introduction

Independent component analysis (ICA) has been a powerful tool for engineers to recover blind sources since the 1980s (Hyvarinen, Karhunen & Oja 2001). Bach & Jordan (2002) described it as follows:

> Independent component analysis (ICA) is the problem of recovering a latent random vector $S = (S_1, \cdots, S_m)^T$ from observations of m unknown linear functions of that vector. The components of S are assumed to be mutually independent. Thus, an observation $X = (X_1, \cdots, X_m)^T$ is modeled as:
>
> $$X = AS, \tag{1}$$
>
> where S is a latent random vector with independent components, and where A is an unknown $m \times m$ matrix, called the mixing matrix.

It is well-known that A is identifiable up to scaling and permutation of columns when S has at most one Gaussian component and A is nonsingular (Kagan, Linnik & Rao 1973, Comon 1994). In general, one does not know the probability density (or mass) function of S. Thus this can be viewed as a semiparametric model (Bickel, Klaassen, Ritov & Wellner 1993) with parameter $(A, r = \prod_{k=1}^{m} r_k)$ where r_k is the unknown distribution of S_k. The essential idea is to estimate the mixing matrix A, or equivalently, $W = A^{-1}$, which is usually called the demixing matrix.

In the past decade, there have been many methods used to estimate A. Most of them can be organized in two groups. The first group of methods directly approximates the distributions of hidden sources within a specified class of distributions and minimize contrast functions such as mutual information, likelihood function or equivalents, e.g., Pham & Garrat (1997), Bell & Sejnowski (1995), Amari, Cichocki & Yang (1996), Amari & Cardoso (1997), Cardoso (1999), Comon (1994), Hyvarinen & Oja (1997). The second group of methods optimizes other contrast functions such as nongaussianity (using negentropy or kurtosis), nonlinear correlation among recovered sources without approximating distributions explicitly, e.g., Jutten & Herault (1991), Hyvarinen (1999). The asymptotic properties of the first group of methods are connected with a folk theorem which states that estimation results are robust to the details of distributions, in other words, it will be sufficient for the specified distributions to capture certain properties of true distributions, such as super- or sub-gaussianity. Local consistency of these methods has been shown under such as stability conditions (Amari & Cardoso 1997), where local consistency means that the true value of W is a local optimum of the contrast function if the sample size is large enough. And global consistency in the case of two sources is shown under heavy-tail conditions (MacKay 1996), where global consistency means that the true value of W is the unique global optimum of the contrast function if the sample size is large enough. However, local consistency is often unsatisfactory as it is weaker than consistency (see Definition 1); Further, even local consistency may not hold when we incorrectly specify the properties such as gaussianity. The second group of methods also has difficulties in reaching consistency, for example, the fast ICA method using kurtosis may fail when there are more than one component having zero kurtosis.

Recently, some new methods appear in the ICA literature. For example, Bach & Jordan (2002) minimize a kernel canonical correlation (KCCA) or kernel generalized variance (KGV) among recovered sources. Hastie & Tibshirani (2002) propose a maximum likelihood estimate (MLE) by using spline-based density approximations. These methods are shown to have good performances in simulations. We believe that KGV has nice statistical properties under appropriate conditions, but in depth analysis is not available.

Here, we propose a new estimator and show that it is consistent (see Definition 1) under general conditions; we further show that it is $\sqrt{n}$-consistent (see Definition 2) if hidden sources have finite variances; and then we construct a one-step MLE estimator and show it is Fisher efficient under regularity conditions which will be given in the theorems. Also, we provide applicable algorithms to implement these methods. (In a private communication, we notice that Samarov & Tsybakov (2002) recently have obtained an interesting but different estimator of W by directly approximating the partial derivatives of obvervable X's density function and proved it is $\sqrt{n}$-consistent under stronger conditions.)

The remainder of the paper is organized as follows. In Section 2, we provide an estimator of the demixing matrix W based on empirical characteristic function (CHFICA) and demonstrate its consistency under identifiability conditions and $\sqrt{n}$-consistency if further hidden sources have finite variances. In Section 3, we carry out the so-called one-step MLE estimator (Bickel et al. 1993) by directly estimating the efficient influence function and show that it can reach asymptotic efficiency (EFFICA) under regularity conditions (defined in Theorem 5), that is, $\sqrt{n}$ times the estimation error is asymptotically normal and the asymptotic covariance matrix reaches the lower bound – inverse Fisher information

matrix. In Section 4, we provide applicable implementation of EFFICA, where we compare a variant of EFFICA (we still call it EFFICA) to popular ICA algorithms such as FastICA (Hyvarinen 1999), JADE (Cardoso 1999), KGV (Bach & Jordan 2002) with benchmark simulations. Appendices A, B& C consist of the complete proofs of the theorems and technical lemmas used in the theorems.

In this paper, $|t|$ denotes its absolute value if t is real and denotes its module if t is complex, $||x||$ denotes its l^2 norm if x is a vector of real numbers, and $|W| = \sqrt{tr(W^T W)}$ for a square matrix W. In particular, for a square matrix W, we use W_k to denote its kth row, W_{ij} to denote its (i, j) element, and an upper case ^{T} to denote the transposation of a vector or a matrix.

2. $\sqrt{n}$-consistent estimator by empirical characteristic function

As we mentioned above, A (thus W) is not unique without further conditions. To deal with its lack of identifiability, we put some constraint on the demixing matrix and define its value space Ω by

$$\Omega = \{W : m \times m \text{ real matrix }, ||W_k|| = 1, W_{kk} \geq |W_{ik}|, \text{ for } 1 \leq k \leq i \leq m\}. \qquad (2)$$

However, the parameter is still not uniquely identified on the boundary of Ω. For simplicity, we assume that the true parameter W_0 is nondegenerate and is in the interior of Ω (denoted by Ω^o, i.e., the equalities in $W_{kk} \geq |W_{ik}|$ do not hold for any $k \neq i$). (Note: this condition is not necessary; the essential idea is to define an order of the rows). In this paper, we say the ICA model as above satisfies *the identifiability conditions* if (i) at most one of S's components is Gaussian and none of them has mass 1 on a single point; and (ii) the true demixing matrix $W_0 \in \Omega^o$ is of full rank.

Let $U = (U_1, \cdots, U_m)^T$ be a m-dim random vector. Define $\psi(t; U) = E[e^{itT U}]$, $\psi(t_k; U_k) = E[e^{it_k U_k}]$, $k \in \{1, \cdots, m\}$, where $t = (t_1, \cdots, t_m)^T \in R^m$ and $i = \sqrt{-1}$. Then U's components are mutually independent if and only if

$$\psi(t; U) - \prod_{k=1}^{m} \psi(t_k; U_k) \equiv 0, \text{ for any } t \in R^m.$$

Thus, the difference between U's joint characteristic function and the product of its marginal characteristic functions can be considered as a measurement of mutual independence of its components. In practice, we cannot calculate their characteristic functions with finite samples $U^{(1)}, \cdots, U^{(n)}$. The natural way will be to use empirical characteristic functions (ECF), i.e. for $k = 1, \cdots, m$

$$\psi_n(t; U) = \frac{1}{n} \sum_{j=1}^{n} e^{itT U^{(j)}} \text{ and } \psi_n(t_k; U_k) = \frac{1}{n} \sum_{j=1}^{n} e^{it_k U_k^{(j)}}.$$

The difference is measured by the $L^2(\mu)$ distance, where μ will be chosen later. We assume that $(X^{(1)}, \cdots, X^{(n)})$ are n i.i.d. copies of X and denote the population and empirical distribution of X by P and P_n separately.

Define

$$\rho(W,P_n)=\int_{R^m}|\psi_n(t;WX)-\prod_{i=1}^m\psi_n(t_i;W_iX)|^2dG(t),$$

$$\rho(W,P_0)=\int_{R^m}|\psi(t;WX)-\prod_{i=1}^m\psi(t_i;W_iX)|^2dG(t),$$

where $G(t)=\prod_{i=1}^m G_i(t_i)$ and each G_i is the standard normal distribution function; by definition, $\psi_n(t;WX)=\frac{1}{n}\sum_{k=1}^n e^{it^TWX_k}$, $\psi_n(t_i;W_iX)=\frac{1}{n}\sum_{k=1}^n e^{it_iW_iX_k}$ for $i=1,\cdots,m$. Thus an estimator of the demixing matrix through empirical characteristic functions is given by

$$\hat{W}=\text{argmin}_\Omega\rho(W,P_n). \tag{3}$$

It is well-known that the one-dimensional empirical characteristic function is bounded and converges uniformly to its population in any compact interval as sample size goes to infinity (Feuerverger & Mureika 1977); this property is also true in multivariate cases. But in our case, we study the asymptotic properties of $\hat{W}$ defined in (3) by empirical process theories since $\rho(W,P_n)$ is a functional of empirical distributions indexed by a compact set Ω. We notice that ECF has been used in the ICA literature, for example, Murata (2001) applies it in testing the independence of recovered sources.

In the remaining of this section, we demonstrate that $\hat{W}$ defined in (3) is consistent and $\sqrt{n}$-consistent to W_0, where consistency and $\sqrt{n}$-consistency are defined formally in the following.

Definition 1 *Given random samples $\{X^i:i=1,\cdots,n\}$, a sequence of estimators $\delta_n(X^1,\cdots,X^n)$ is said to be (weakly) consistent to a parameter θ if $||\delta_n-\theta||\to 0$ in probability, i.e. for any $\epsilon>0$, $Pr(||\delta_n-\theta||>\epsilon)\to 0$ as $n\to\infty$.*

Definition 2 *In the above definition, δ_n is $\sqrt{n}$-consistent to θ if $\sqrt{n}||\delta_n-\theta||=O_p(1)$, i.e., $limsup_n Pr(\sqrt{n}||\delta_n-\theta||>M)\to 0$ as $M\to\infty$.*

Theorem 3 (Consistency) *Suppose that the ICA model is identifiably parametrized and $W_0\in\Omega^o$. Then $\hat{W}$ defined by (3) is a consistent estimate of W_0.*

Proof Let $\hat{W}$ be one solution of $\text{argmin}_\Omega\rho(W,P_n)$. Then

$\rho(\hat{W},P_0)=\int|\psi(t;\hat{W}X)-\prod_{i=1}^m\psi(t_i;\hat{W}_iX)||^2dG(t)$

$\le\int[|\psi(t;\hat{W}X)-\psi_n(t;\hat{W}X)|+|\prod_{i=1}^m\psi(t_i,\hat{W}_iX)-\prod_{i=1}^m\psi_n(t_i;\hat{W}_iX)|+|\psi_n(t;\hat{W}X)-\prod_{i=1}^m\psi_n(t_i;\hat{W}_iX)|]^2dG(t)$

$\le 3\{\sup_\Omega\int|\psi(t;WX)-\psi_n(t;WX)|^2dG(t)+\sup_\Omega\int|\prod_{i=1}^m\psi(t_i;W_iX)-\prod_{i=1}^m\psi_n(t_i;W_iX)|^2dG(t)+\rho(\hat{W},P_n)\},$

(By definition of $\hat{W}$)

$\le 3\{\sup_\Omega\int|\psi(t;WX)-\psi_n(t,WX)|^2dG(t)+\sup_\Omega\int|\prod_{i=1}^m\psi(t_i;W_iX)-\prod_{i=1}^m\psi_n(t_i;W_iX)|^2dG(t)+\rho(W_0,P_n)\}$

$=o_p(1)$, (by Lemma 7(2,4,6) in Appendix A).

Then the consistency follows from Lemma 1 using Wald's argument given the compactness of Ω (Wald 1949). ■

Theorem 4 ($\sqrt{n}$-Consistency) *Suppose that the ICA model (1) is identifiably parametrized, $W_0 \in \Omega^o$, and $E||S||^2 < \infty$, then $\hat{W}$ defined by (3) is a $\sqrt{n}$-consistent estimate of W_0.*

Proof The complete proof is given in appendix B. ■

Remark 1: (i) It is not surprising that under the conditions of Theorem 4 asymptotical normality holds for the free parameters in $W \in \Omega$. (ii) The choice of measurement G for the integral in $\rho(W, P_n)$ is not necessary to be Gaussian; It can be shown similarly that Theorem 3 and 4 still hold when G is some other distribution function, which has support R^m and decays smoothly and moderately fast, for example, the logistic distribution. (iii) But the choice of G does affect the constant term of the asymptotic mean square error (i.e. $E|\hat{W} - W|^2$). Professor Bin Yu suggested the problem how to choose an optimal G which might lead to efficiency. The authors found in simulations that choosing a uniform distribution rather than a Gaussian for G gives almost perfect results if the hidden sources are really uniform, but it is hard to find a generally applicable rule.

3. Construction of efficient estimates

In Section 2, we have proposed an estimate and proved its consistency only requiring the identifiability conditions while $\sqrt{n}$-consistency further requiring finite variance of hidden sources. It is well-known that under regularity conditions, based on a $\sqrt{n}$-consistent initial estimate, a one-step MLE is asymptotically Fisher efficient when the corresponding efficient influence function can be estimated $\sqrt{n}$-unbiasedly (Bickel et al. 1993). However, to calculate the efficient influence function of W in the ICA model, it is more convenient to work on free parameters, i.e., we do not want the normalization constraint in (2). Alternatively, we can re-scale the hidden sources instead of W itself, i.e., for a fixed scaling function K, we may assume that $E[K(S_k)] = 0, k = 1, \cdots, m$, for example, $K(t) = t^2 - 1$ corresponds to assuming second moment to be 1 (Amari & Cardoso 1997). Here instead we use $K(t) = 2I(|t| \leq 1) - 1$, which assumes each source to have absolute median 1. That is, suppose that $W \in \Omega^o$ and S with mutually independent components are such that $S = WX$, then we can define $(\overline{W}, \overline{S})$ by

$$\overline{S}_k = S_k / med(|S_k|) \text{ and } \overline{W}_k = W_k * med(|S_k|).$$

Obviously, $\overline{S} = \overline{W}X$ has mutually independent components, and $med(|\overline{S}_k|) = 1$. Again, for the sake of identifiability, we define

$$\overline{\Omega} = \{W_{m\times m} : ||W_k|| > 0, \frac{W_{kk}}{||W_k||} \geq \frac{|W_{ik}|}{||W_i||}, 1 \leq k \leq i \leq m\}. \quad (4)$$

Then $\overline{W} \in \overline{\Omega}$. This $\overline{W}$ will be of interest in the following construction of efficient estimates.

In the empirical world, we only have an estimate of W (denoted by W_n) and an estimate of S, then the above transformation can be carried out by using empirical medians. It is important to know whether the estimate of $\overline{W}$ by this empirical transformation on W_n has the same asymptotic properties (i.e., consistency and $\sqrt{n}$-consistency) as the estimate W_n of W. This is answered by Lemma 5.

Lemma 5 (Reparametrization) *Let $W \in \Omega$ with $rank(W) = m$ and $\underline{X_n} = \{X^{(1)}, \cdots, X^{(n)}\}$ be n iid copies of X, and $S = WX$ has mutually independent components. Suppose that W_n is an estimate of W. Denote $\Gamma = diag(med(|S|))$ (i.e., a diagonal matrix with diagonals $med((|S|))$, $\Gamma_n = diag(med(|W_n\underline{X_n}|))$ and $\overline{W} = \Gamma W$, $\overline{W_n} = \Gamma_n W_n$, i.e. Γ are Γ_n are both diagonal matrix with diagonals medians of $|S|$ and sample medians of absolute values of $W_n\underline{X_n}$ separately. Then the following results hold:*

(i). If $W_n - W = o_p(1)$, then $\overline{W_n} - \overline{W} = o_p(1)$;

(ii). If $\sqrt{n}(W_n - W) = O_p(1)$ and $E|S_k| < \infty, k = 1, \cdots, m$, then $\sqrt{n}(\overline{W_n} - \overline{W}) = O_p(1)$.

Proof This follows directly from Lemma 8 in Appendix C. ■

To construct an efficient estimate, we will always assume that the hidden sources have finite variance. Then by Theorem 4 and Lemma 5, we can obtain a $\sqrt{n}$-consistent estimate of the demixing matrix $\overline{W} \in \overline{\Omega}$. In the following, we will still use W to denote $\overline{W}$, the parameter of interest, but now $W \in \overline{\Omega}$, for simplicity of notation.

The density function of X in model (1) with respect to (W, r) is

$$p_X(x; W, r) = |det(W)| r(WX),$$

where W is the demixing matrix and $r = \prod_{k=1}^m r_k$ is the joint density function of $S = (S_1, \cdots, S_m)$ with marginal densities $(r_1, \cdots, r_m)$. Let $\phi_k = -r_k'/r_k$, $k \in \{1, \cdots, m\}$, and define $\underline{\phi}$ by $\underline{\phi}(s) = (\phi_1(s_1), \cdots, \phi_m(s_m))^T$, $s \in R^m$. Then the score function of W (i.e partial derivative of log(density) w.r.t W) is equal to

$$S_W = \frac{\partial log(p_X(x; W, r))}{\partial W} = W^{-T} + \underline{\phi}(Wx)x^T = (I + \underline{\phi}(s)s^T)W^{-T},$$

where $s = Wx$ and W^{-T} denotes $[W^{-1}]^T$.

By definition, the nuisance score function of r along the path $r_t = \prod_{k=1}^m r_k(1 + th_k)$ is

$$S_r = \frac{\partial log(p_X(x; W, r(t)))}{\partial t}|_{t=0} = \sum_{k=1}^m \frac{\partial log(r_k(W_k x)(1 + th_k(W_k x)))}{\partial t}|_{t=0} = \sum_{k=1}^m h_k(s_k)$$

where the variational direction is $(h_1, \cdots, h_m) \in L_2(P)^m$ (Bickel et al. 1993). Let's denote by TS, the closed linear span of the set of nuisance scores in all directions. For simplicity, we assume $E[S_k] = 0$, $k \in \{1, \cdots, m\}$. Then the efficient score function of W, defined by the projection of its score function onto the orthocomplimentary space of tangent space of nuisance parameter r, i.e., $TS^\perp$, is given by

$$S_e(x; W, r) = Proj(S_W | TS^\perp) = [M]W^{-T}, \tag{5}$$

where by making use of orthogonality the matrix M's elements can be expressed as

$$\begin{array}{l} M_{kk} = \alpha_k \widetilde{s_k} + \beta_k K(s_k), \text{ for } k = 1, \cdots, m, \\ M_{ij} = \phi_i(s_i) s_j, \text{ for } 1 \le i \ne j \le m, \end{array} \tag{6}$$

with $s = Wx$, $\widetilde{s_k} = \frac{s_k}{E[S_k^2]}$, $\alpha_k = \frac{E[(\widetilde{S_k} - \lambda_k K(S_k))\phi_k(S_k)S_k]}{1 - E^2(\widetilde{S_k} K(S_k))}, \beta_k = \frac{E[(K(S_k) - \lambda_k \widetilde{S_k})\phi_k(S_k)S_k]}{1 - E^2(\widetilde{S_k} K(S_k)}$, and $\lambda_k = E[\widetilde{S_k} K(S_k)]$ (Amari & Cardoso 1997).

Later in this paper, we consider S_e as a m^2-dim column vector function (reshape the matrix row by row into a vector). As a consequence, its efficient influence function is given by

$$IF = I_e^{-1} S_e, \text{ where } I_e = E[S_e S_e^T].$$

Now we still need to estimate $(\phi_k, \alpha_k, \beta_k, \lambda_k)$ for $k = 1, \cdots, m$ consistently. Notice that $(\alpha_k, \beta_k, \lambda_k)$ are functionals only of (W, ϕ_k) and can be estimated ad hoc by arithmetic combinations of moments if we have estimates of (W, ϕ_k). Given an estimate $\hat{W}$ of W, we can recover hidden sources using $\hat{W}X$. The difficulty is in estimating ϕ_k by using the recovered sources. There are two estimation methods. Since $\phi_k = -\frac{r_k'}{r_k}$, we can use the kernel method to estimate r_k' and r_k separately, and then take the combination. The second method is to use the Cox's method by B-spline approximations (Cox 1985, Jin 1992). Both methods can provide consistent estimates under weak conditions. the Cox's method requires stronger conditions but is easier to implement, which will be defined in (8) and be used for our simulation purpose. For simplicity of proof, we make use of the logistic kernel to estimate the derivative of logarithmic density function. That is, given recovered data of the kth source, say $\underline{\hat{S}_k} = (\hat{S}_k^{(1)}, \cdots, \hat{S}_k^{(n)})$, the estimate of ϕ_k at $t \in \mathcal{R}$ is defined by

$$\hat{\phi}_k(t) = \begin{cases} -\hat{g}'(t)/\hat{g}(t) & \text{if } |t| \le d_n, |\hat{g}'(t)| \le c_n \hat{g}(t), \\ 0 & \text{otherwise} \end{cases}$$

where

$$\hat{g}(t) = \frac{1}{n b_n} \sum_{i=1}^{n} w(\frac{t - \hat{S}_k^{(i)}}{b_n}), \; w(t) = e^{-t}/(1 + e^{-t})^2 \text{ and } \hat{g}'(t) = \frac{\partial}{\partial t} \hat{g}(t)$$

and $b_n = n^{-1/5}, c_n = n^{1/15}, d_n = n^{1/5}$.

This estimator has been well studied in Bickel et al. (1993) and shown to give consistent estimation measured by its integrated square estimation error. Then $(\alpha_k, \beta_k, \lambda_k)$ are estimated by arithmetic combinations of moments with plugged-in estimates of (W, ϕ_k). Thus, we have obtained estimators of $(\alpha, \beta, \lambda, \phi, S_e, I_e)$. Denote them by $(\hat{\alpha}, \hat{\beta}, \hat{\lambda}, \hat{S}_e, \hat{I}_e)$ separately. Finally, we estimate the efficient influence function using

$$\widehat{IF} = \hat{I}_e^{-1} \hat{S}_e. \tag{7}$$

Notice that $\hat{I}_e$ is a $m^2 \times m^2$ matrix and $\hat{S}_e$ is a function vector of dimension m^2. We use a data-splitting scheme (Klaassen 1987) to construct the one-step MLE of the demixing matrix.

Let $n_1 = \frac{1}{3}n, n_2 = \frac{2}{3}n$. Given i.i.d. random samples $\{X^{(i)}, i = 1, \cdots, n\}$, we divide them into three parts and get three estimates of W and IF using the three parts separately. That is, $\hat{W}_1$, an estimate of W, is defined by (3) using the first part $(X^{(1)}, \cdots, X^{(n_1)})$ but is rescaled using recovered absolute medians, $\hat{W}_2, \hat{W}_3$ are defined similarly by using the second part $(X^{(n_1+1)}, \cdots, X^{(n_2)})$ and the third part $(X^{(n_2+1)}, \cdots, X^{(n)})$ separately; And $\widehat{IF}_3(.; \hat{W}_2)$ is defined by (7) using the estimated demixing matrix $\hat{W}_2$ and the third part of data, $\widehat{IF}_1(.; \hat{W}_3)$ using $\hat{W}_3$ and the first part of data, and $\widehat{IF}_2(.; \hat{W}_1)$ using $\hat{W}_1$ and the second part of data. Then the one-step MLE is defined by

$$\begin{aligned}\hat{W}^* &= \tfrac{1}{3}\{\hat{W}_2 + \tfrac{3}{n}\textstyle\sum_{i=1}^{n_1} \widehat{IF}_3(X^{(i)}; \hat{W}_2)\} \\ &\quad + \tfrac{1}{3}\{\hat{W}_3 + \tfrac{3}{n}\textstyle\sum_{i=n_1+1}^{n_2} \widehat{IF}_1(X^{(i)}; \hat{W}_3)\} \\ &\quad + \tfrac{1}{3}\{\hat{W}_1 + \tfrac{3}{n}\textstyle\sum_{i=n_2+1}^{n} \widehat{IF}_2(X^{(i)}; \hat{W}_1)\}.\end{aligned}$$

Theorem 6 (Efficiency) *If the ICA model $X = AS$ satisfies the following conditions (let $W = A^{-1}$):*

(i) W is in the interior of $\overline{\Omega}$;

(ii) S has at most one Gaussian component and has no degenerate component;

(iii) S_k has unknown absolute continous density function r_k with $ES_k^2 < \infty$ and $\int_R \frac{r_k'^2(t)}{r_k(t)}dt < \infty$ for $k = 1, \cdots, m$;

(iv) $ES_k = 0$, $med(|S_k|) = 1$ for $k = 1, \cdots, m$;

Then the estimate $\hat{W}^*$ above is asymptotically efficient.

Proof Under the given conditioins, the ICA model is a regular semiparametric model. From Theorem 7.8.1 of Bickel, et al (1993), it is sufficient to show that for all sequence $\{W_n\}$ with $\sqrt{n}|W_n - W| = O(1)$, the following three claims are true:

(a) $\sqrt{n}\int \hat{S}_e(x; W_n; \underline{X})dP_{(W_n,r)}(x) = o_{P_{(W_n,r)}}(1)$;

(b) $\hat{I}_e(W_n; \underline{X}) - I_e(W_n; r) = o_{P_{(W_n,r)}}(1)$;

(c) $I_e(W; r)$ is positive definite and each entry is finite.

First, condition (iv) implies $\sqrt{n}\int \hat{S}_e(x; W_n; \underline{X})dP_{(W_n,r)}(x) = 0$, that is (a). Second, (b) can be easily verified by using the mutual independence between components of W_nX under the law $P_{(W_n,r)}$ and the facts that kernel estimators and moment estimators are all consistent under our conditions. Third, each entry of $I_e(W, r)$ is finite from condition (iii); suppose that $I_e(W, r)$ is not positive definite, then for some set of real numbers (not all zeros) $\{c_{ij} : i, j = 1, \cdots, m\}$, $var(\sum_{i=1}^m c_{ii}M_{ii} + \sum_{i\neq j} c_{ij}M_{ij}) = 0$, where M_{ij}'s are defined in (6). By taking expansions, $\sum_{i=1}^m c_{ii}^2 E[M_{ii}^2] + \sum_{i<j} E[(c_{ij}M_{ij} + c_{ji}M_{ji})^2] = 0$, and thus there must

be $c_{ij} = 0$ for $i, j = 1, \cdots, m$ since none of S_i's has a degenerate distribution, contradiction! So (c) must hold.

■

Remark 2: (i) In condition (iv), mean zero is not necessary since the mean can be estimated by its sample mean and be removed adaptively. (ii) To control the scaling, we use the absolute median 1 instead of unit variance because the efficient score function in using unit variances will require stronger conditions – finite fourth moments. (iii) Iterating the one-step MLE more than once still provides efficient estimates.

In application, to estimate the derivative of a logarithmic density function, the kernel method has large computational complexity, so we use the Cox's method. That is, given samples $(t_1, \cdots, t_n)$ from some distribution with density f, $h = -f'/f$ can be approximated by

$$\hat{h} = (A_n^{-1} D_n)^T \mathbf{B_n}, \tag{8}$$

where $\mathbf{B_n}$ is the B-spline basis (a column vector of B-splines) with knots chosen in the range of $(t_1, \cdots, t_n)$, $A_n = \frac{1}{n}\sum_{i=1}^n \mathbf{B_n}(\mathbf{t_i})\mathbf{B_n^T}(\mathbf{t_i})$ and $D_n = \frac{1}{n}\sum_{i=1}^n \mathbf{B'_n}(\mathbf{t_i})$ (note: $\mathbf{B'_n}$ is the pointwise derivatives of $\mathbf{B_n}$). We may use empirical quantiles or equal distances to choose the knots. The number of knots is a usual smooth parameter and can be decided by cross-validation. Jin (1992) has studied this problem thoroughly and shows that under weak conditions the optimal choice of the number of knots by smoothing cross-validation is of order n^δ with $0 < \delta < \frac{1}{6}$ both theoretically and with a variety of simulations.

As a variant of the estimator defined in Theorem 6, we do not use the data splitting scheme. Instead we use the whole data $(X^{(1)}, \cdots, X^{(n)})$ to estimate the efficient influence function IF (say $\widehat{IF}$) with an initial estimate $W^{(0)}$ of W, then one-step MLE becomes

$$W^{(1)} = W^{(0)} + \frac{1}{n}\sum_{i=1}^n \widehat{IF}(X^{(i)}; W^{(0)}). \tag{9}$$

We further update the above estimation several times and denote the final estimate in the left hand by $\hat{W}$, in abusing the name we call this estimator as EFFICA. The further study of this variant estimator is beyond the scope of this paper.

4. Realization of CHFICA and EFFICA in simulations

To carry out CHFICA defined in (3), we can simply choose G to be multivariate Gaussian distribution with mean zeros and covariance$m \times m$ identity matrix. By expanding the contrast function $\rho(W, P_n)$, we have

$$\begin{aligned}\rho(W, P_n) = \quad & \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^n \exp\{-\sum_{k=1}^m \frac{|W_k(X_i - X_j)|^2}{2}\} + \frac{1}{n^{2m}}\prod_{k=1}^m (\sum_{i,j=1}^n \exp\{-\frac{|W_k(X_{i_k} - X_{j_k})|^2}{2}\}) \\ & -\tfrac{2}{n^{m+1}}\textstyle\sum_{i=1}^n \prod_{k=1}^m (\sum_{j=1}^n \exp\{-\frac{|W_k(X_i - X_{j_k})|^2}{2}\})\end{aligned}$$

Thus the computational complexity of directly minimizing $\rho(W, P_n)$ is $O(n^2 m^2)$, which is slower than some well known ICA algorithms. We propose a Montel Carlo (MC) version

of CHFICA which approximates the Gaussian integration in the contrast function $\rho(W, P_n)$ of CHFICA by MC method (MCCHFICA). It is not surprising that a larger Monte-Carlo sample size will provide a better performance. In general, $O(log(N))$ will be enough for MC sample size to give a good initial estimate. To approximate ϕ_k with efficient computation, we use the Cox's method with the B-spline approximation defined in (7), where the number of knots are chosen by smoothing cross-validation (Jin 1992). Furthermore, instead of one-step MLE, we update the one-step MLE iteration defined in (8) till convergence or a fixed number of steps (EFFICA). The computational complexity for MCCHFICA is $O(m^2 N log(N))$ and for EFFICA is $O(m^2 N^{1+\delta})$, where $0 < \delta < \frac{1}{6}$ depends on distributions of hidden sources (Jin 1992). We notice that essentially, EFFICA is equivalent to solving the efficient score equations in sieves using Newton-Rapson method (Murphy & van der Vaart 2000). The algorithms for MCCHFICA and EFFICA are given in figure 1 & 2 separately. As usual, restarting initials are needed to reach global optimum for MCCHFICA.

Algorithm **MCCHFICA**

Input: m-dim Data vectors $x^1, x^2, \cdots, x^n$ (data matrix X)

1. Generate random vector $t^1, \cdots, t^K$ from $MN(0, I_m)$
2. Minimize (w.r.t W) the contrast function $f(W)$ defined as:

$$f(W) = E_K[|E_n(exp(it^T WX)) - \prod_{i=1}^m E_n(exp(it_i W_i X))|^2],$$

where $K = \max(50, 10\log(nm))$, E_n calculates the empirical mean over X for each fixed t and E_K calculates the empirical mean over $(t^1, \cdots, t^K)$.

Output: W.

Figure 1: A high-level description of MCCHFICA algorithm for estimating W

Note: In practice, prewhitening data X can make the optimization much easier since we can concentrate W on the Stief manifold of orthogonal matrix. Efficient algorithms of optimization on Stief manifold have been well studied in Edelman, Arias & Smith (1999), which have also been successfully implemented in Bach & Jordan (2002). Further, the estimation accuracy depends on the choice of K, the larger the better. But the choice of K eventually does not hurt the consistency as long as it goes up to infinity with the sample size n. For computational purposes, we find that the choice in the algorithm usually give satisfactory estimates as initials for EFFICA when the number of source is not too large.

Algorithm **EFFICA**

Input: m-dim Data vectors $x^1, x^2, \cdots, x^n$ (data matrix X) and initial W, set $\overline{K} = 5 \times m$

0. Calculate $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x^i$, set $x^i = x^i - \overline{x}$
1. $S = WX$, rescale W and S with absolute medians of S's rows
2. For $k = 1, \cdots, m$, estimate $\phi_k = r_k'/r_k$ using B-splines approximation as in Jin (1992)
3. Calculate Plug-in estimates of α, β, λ in (6);
4. Calculate Plug-in estimates of S_e in (5) and $I_e = \frac{1}{n}\sum_{i=1}^{n} S_e(X_i)S_e^T(X_i)$
5. Update $\hat{W} = W + \frac{1}{n}\sum_{i=1}^{n} I_e^{-1} S_e(X_i)$, k=k+1
6. if $||\hat{W} - W|| < \epsilon$ or $k = \overline{K}$, stop; otherwise, $W = \hat{W}$, repeat 2-6

Output: W

Figure 2: A high-level description of EFFICA algorithm for estimating W

We have done an extensive set of simulation experiments using data obtained from a variety of source distributions which are very common in statistics. Comparisons were made with three existing ICA algorithms: the FastICA algorithm (Hyvarinen & Oja 1997), the JADE algorithm (Cardoso 1999) and KernelICA-KGV (Bach & Jordan 2002). These three algorithms were used with their default settings, and EFFICA used the estimates by MCCHFICA as initialization for case $m = 2$ (results show in Table 1 and figure 3) and used estimates by KernelICA-KGV as initialization in other simulations (results show in Table 2). The performance of each algorithm in the simulations is defined by the estimation error $d(\hat{W}, W_0)$ measured by the so-called *Amari error* :

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

where $a_{ij} = (VW^{-1})_{ij}$, which is invariant to permutation and scaling of the columns of V and W, is always between 0 and $(m-1)$, and is equal to zero if and ony if V and W represent the same components (Bach & Jordan 2002, Amari et al. 1996) .

The following source distributions were used for a bunch of simulations given in Table 1, Table 2 and figure 3:

Table 0: Distribution used in first group of simulations (output: Table 1)

[0]. N(0,1)	[8]. IID exp.(1)+ U(0,1)
[1]. IID exp.(1)	[9]. IID mixture exp.
[2]. IID t(3)	[10]. IID mixture of exp. and normal
[3]. IID lognormal(1,1)	[11]. IID mixture Gaussians: multimodal
[4]. IID t(5)	[12]. IID mixture Gaussians: unimodal
[5]. IID logistic(0,1)	[13]. exp. vs normal
[6]. IID Weibull(3,1)	[14]. lognormal vs normal
[7]. IID exp.(10)+normal(0,1)	[15]. Weibull(3,1) vs exp(1)

In the first group of simulations, we do experiments with two independent sources listed in Table 0 from [1] to [15] with sample size 1000 and 2000 separately with true de-mixing

pdfs	Fast	Jade	Kgv	MCCHF	EFF	Fast	Jade	Kgv	MCCHF	EFF
1	57	44	16	20	11	48	29	14	16	8
2	53	44	36	45	32	23	23	28	41	21
3	38	34	16	18	8	34	23	13	14	6
4	80	63	68	75	48	54	44	39	39	30
5	117	86	109	121	85	54	44	69	73	42
6	44	35	15	19	10	35	25	11	14	6
7	59	48	17	20	13	37	33	13	15	9
8	65	50	17	20	14	45	36	13	16	11
9	53	38	15	18	7	30	25	10	13	5
10	85	68	33	55	49	60	52	22	35	34
11	48	36	28	35	44	34	26	16	22	30
12	85	58	49	61	71	48	36	31	32	33
13	61	60	17	24	17	41	41	11	18	10
14	62	62	17	26	17	43	43	12	19	8
15	49	38	17	20	9	36	26	11	14	6

Table 1: Report of the medians of the Amari errors (multiplied by 1000) for two components ICA with 1000 samples(left) and 2000 samples(right) in 400 replications.

m	N	#repl	Fast	Jade	Kgv	EFF
4	1000	100	146	135	62	58
	4000	100	85	77	31	27
8	2000	50	455	430	205	162
	4000	50	322	305	138	114
12	4000	25	515	492	385	250

Table 2: Reporting the median of the Amari errors (multiplied by 1000) for m components with N samples:m components are first m pdfs in the source list

matrix $W = [2, 1; 2, 3]$ and the output of estimator errors of different estimators is given in Table 1 (EFFICA uses MCCHFICA estimates as initial values). As we can see that EFFICA has small amari errors in most cases except in cases that the sources are mixture Gaussians which decay rapidly in the tails. This can be explained by two facts: First, KGV itself is very accurate, especially when hidden sources have rapid decaying densities; Second, the efficiency of EFFICA is in the sense of large sample size. In the second group of simulations, we use $m \times m$ identity matrix as the mixing matrix and increase sources number m to 4, 8 and 12 with different sample sizes: for $m = 4$, we use 4 source distributions from [0]-[3] separately given in Table 0; for $m = 8$, we use [0]-[7]; and for $m = 12$, we use [0]-[11]. The output is given in Table 2 (to speed up the computation, we have used JADE as initial estimates of KGV and used KGV as initial estimates of EFFICA), where EFFICA

has uniformly smaller Amari errors. This phenomena can be explained by the asymptotic properties of EFFICA since we have used relatively large sample sizes.

5. Conclusion

In this paper, we have provided a $\sqrt{n}$-consistent estimate (CHFICA) of the ICA model and showed that the classical one-step MLE estimate reaches Fisher efficiency. For practical uses, we have proposed an initial estimate using Montel-Carlo version of CHFICA, which is consistent and costs less computation, and then, proposed iterating one-step MLE several times by directly estimating efficient score function using B-spline approximations (EFFICA). Benchmark simulations have exhibited the excellent performance of EFFICA in comparison with standard and state-of-the-art ICA algorithms such as the Kernel ICA method.

Techniques of using characteristic function have been widely used in deconvolution problems (Fan 1989). It is not hard to extend CHFICA to noisy ICA models when the covariance of sensor noises are known or can be estimated in other ways, and again the rate of convergence can reach $n^{-\frac{1}{2}}$. We conjecture that deriving efficient estimators through likelihood for ICA models requires more or less constraints on moments as its score function shows. Deeper analysis of efficient estimators that requires less conditions, by using other methods such as kernel methods or characteristic function, is still needed in this literature.

Acknowledgement

The authors are grateful to Prof. Michael Jordan, Prof. Bin Yu and Prof. Anant Sahai for helpful discussions and appreciate Francis Bach for his generosity in sharing all his simulation codes on Kernel-ICA. The authors also acknowledge the support for this project from the National Science Foundation.

Appendix A

This appendix provides Lemma 7 used in the previous theorems.

Lemma 7 *The following results hold:*

1. *$argmin_\Omega \rho(W, P_0)$ has unique solution W_0 if $W_0 \in \Omega^o$;*
2. *$\rho(W_0, P_n) = O_p(\frac{1}{n})$;*
3. *The distance ρ satisfies Lipschitz condition, i.e if W_0 is in the interior of Ω, then there exists $c(W_0) < \infty$ and $\epsilon(W_0) > 0$, such that ($B(W_0; \epsilon(W_0))$ is a ball in R^{m^2} with center W_0 and radius $\epsilon(W_0)$)*

$$||W - W_0||^2 \leq c(W_0)\rho(W, P_0), \text{ for any } W \in B(W_0; \epsilon(W_0)) \cap \Omega;$$

4. *$sup_\Omega \int |\psi_n(t; WX) - \psi(t; WX)|^2 dG(t) = o_p(1)$;*
5. *$sup_\Omega \int |\psi_n(t_i; W_i X) - \psi(t_i; W_i X)|^2 dG_i(t_i) = o_p(1)$ for $i = 1, \cdots, m$;*

6. $sup_\Omega \int |\prod_{1=1}^m \psi_n(t_i; W_i X) - \prod_{i=1}^m \psi(t_i; W_i X)|^2 dG(t) = o_p(1)$.

Proof Claim 1 is obvious by identifiability condition. Claim 4 is implied by $|\psi_n(t; WX) - \psi(t; WX)| \leq 2$ and $\sup_\Omega |\psi_n(t; WX) - \psi(t; WX)| = o_p(1)$ which is assured by boundedness of envelope, compactness of Ω and continuity of ψ w.r.t W by Lemma 3.10 of Van der Geer (2000). Similarly, Claim 5 holds. And hence Claim 6 follows from

$$\begin{aligned} & \sup_\Omega \int |\textstyle\prod_{1=1}^m \psi_n(t_i; W_i X) - \prod_{i=1}^m \psi(t_i; W_i X)|^2 dG(t) \\ \leq \quad & \sup_\Omega \int (\textstyle\sum_{i=1}^m |\psi_n(t_i; W_i X) - \psi(t_i; W_i X)|)^2 dG(t) \\ \leq \quad & m \textstyle\sum_{i=1}^m \sup_\Omega \int |\psi_n(t_i; W_i X) - \psi(t_i; W_i X)|^2 dG_i(t_i) \\ = \quad & o_p(1). \end{aligned}$$

Claim 2 and 3 are proved in the following. ■

Claim 2: $\rho(W_0, P_n) = O_p(\frac{1}{n})$.
Proof It is enough to show that for any $W \in \Omega$,

$$\int |\psi_n(t; WX) - \psi(t; WX)|^2 dG(t) = O_p(\frac{1}{n}) \tag{10}$$

and

$$\int |\prod_{i=1}^m \psi_n(t_i, W_i X) - \prod_{0=1}^m \psi(t_i, W_i X)|^2 dG(t) = O_p(\frac{1}{n}). \tag{11}$$

Notice that

$$\begin{aligned} E[\textstyle\int |\psi_n(t; WX) - \psi(t; WX)|^2 dG(t)] \quad & = \textstyle\int E[|\psi_n(t; WX) - \psi(t; WX)|^2] dG(t) \\ & = \tfrac{1}{n} \textstyle\int E[|e^{it^T WX} - E(e^{it^T WX})|^2] dG(t) \\ & \leq \tfrac{4}{n} \end{aligned}$$

which proves (11) by using Chebyshev inequality.
Using the fact $|\psi_n| \leq 1, |\psi| \leq 1$,

$$\begin{aligned} & |\textstyle\prod_{i=1}^m \psi_n(t_i, W_i X) - \prod_{0=1}^m \psi(t_i, W_i X)| \\ = \quad & |(\psi_n^1 - \psi^1) \textstyle\prod_{i=2}^m \psi_n^i + (\psi_n^2 - \psi^2)\psi^1 \prod_{i=3}^m \psi_n^i + \cdots + (\psi_n^m - \psi^m) \prod_{i=1}^{m-1} \psi^i| \\ & \text{(here } \psi_n^i = \psi_n(t_i, W_i X) \text{ and } \psi^i = \psi(t_i, W_i X) \text{ for } i = 1, \cdots, m) \\ \leq \quad & |\psi_n^1 - \psi^1| + |\psi_n^2 - \psi^2| + \cdots + |\psi_n^m - \psi^m|, \end{aligned}$$

thus

$$E[\int |\prod_{i=1}^m \psi_n(t_i, W_i X) - \prod_{i=1}^m \psi(t_i, W_i X)|^2 dG(t)] \leq m \int \sum_1^m E|\psi_n(t_i, W_i X) - \psi(t_i, W_i X)|^2 dG(t)$$
$$\leq \frac{4m^2}{n}$$

which implies (12) by using Chebyshev inequality. ■

Claim 3: *The distance ρ satisfies a Lipschitz condition, i.e if W_0 is in the interior of Ω, then there exists $c(W_0) < \infty$ and $\epsilon(W_0) > 0$, such that ($B(W_0; \epsilon(W_0))$) is a ball in R^{m^2} with center W_0 and radius $\epsilon(W_0)$)*

$$||W - W_0||^2 \leq c(W_0)\rho(W, P_0), \text{ for any } W \in B(W_0; \epsilon(W_0)) \cap \Omega.$$

Proof For each $W \in \Omega$ such that $||W - W_0|| \leq 1$, notice that for each k, W_k is on a m-dimensional unit ball, let Δ_k be the unit tangent vector at W_{0k} in the space spanned by W_k and W_{0k}, then let $\gamma_k^o(t) = cos(t)W_{0k} + sin(t)\Delta_k$ (choosing the sign of Δ_k such that the angle between Δ_k and W_k is acute), it is easy to check that $\gamma_k(0) = W_{0k}$ and $\gamma_k^o(\theta_k) = W_k$ where θ_k is the angle between W_{0k} and W_k. It is obvious that γ_k^o defines the "shortest" path on the unit ball from W_{0k} and W_k.

Let $\eta_w = \sqrt{\sum_{k=1}^m \theta_k^2}$, define $\gamma_k(t) = \gamma_k^o(\frac{\theta_k}{\eta_w}t)$ and $\gamma(t) = (\gamma_1(t), \cdots, \gamma_m(t))'$, then $\gamma(0) = W_0$, $\gamma(\eta_w) = W$, $|W - W_0| = \sqrt{\sum_{k=1}^m (sin^2(\theta_k) + (1 - cos(\theta_k))^2)} = \sqrt{\sum_{k=1}^m 4sin^2(\frac{\theta_k}{2})} = \sqrt{(\sum_{k=1}^m \theta_k^2)(1 + o(1))} = \eta_w(1 + o(1))$, thus $\frac{1}{2}\eta_w \leq |W - W_0| \leq \eta_w$ if $|W - W_0| \leq 0.05$. Furthermore $|\gamma'(t)| = \sqrt{\sum_{k=1}^m (\frac{\theta_k}{\eta_w})^2} = 1$ and $|\gamma''(t)| = \sqrt{\sum_{k=1}^m (\frac{\theta_k}{\eta_w})^4}$ is in $[\frac{1}{m}, 1]$.

We will use the following fact (Page 98, Murray & Rice): given the variation γ defined as above on the manifold $\Omega^* = \{W : m \times m \text{ matrix, each row has norm 1}\}$ which passes through W_0, then the Taylor expression of $\rho(\gamma(t), P_0)$ about $t = 0$ is given by

$$\rho(W_0, P_0) + t\rho'(\gamma(0), P_0) + \frac{t^2}{2}(\rho''(\gamma(0), P_0) + o(1))$$

as t converges to 0 , where Δ is the tangent vector determined by W.

By taking $t = \eta_w$ in the above expansion and noticing that $\gamma(0)$ only depending on the unit tangent vector Δ, it will be enough to show the following two facts:

(a) $\rho'(\gamma(0), P_0) = 0$ on the tangent space of W_0;

(b) $\rho''(\gamma(0), P_0) \geq \epsilon(W_0)$ uniformly on the tangent space of W_0 and $\epsilon(W_0) > 0$ only depends on W_0. Notice that $\Omega \subset \Omega^*$, the result will follow by using Taylor expansion.

For simplicity, we first show the result holds when W_0 is the identity matrix I. In this case, the tangent space at W_0 on the manifold Ω^* is given by $TS(I) = \{\Delta : m \times m\text{matrixwithtotalnorm1,}\Delta_{kk} = 0, k = 1, \cdots, m\}$. In the following, we let $D(t, \eta, \Delta) \equiv E[e^{it^T(I+\eta\Delta)S}] - \prod_{k=1}^m E[e^{it_k S_k + it_k \eta \Delta_k S}]$, then $D(0, \Delta) \equiv 0$. Condition (a) is verified by the following calculation (Let Δ_k, Δ^k be the kth row and kth column of Δ separately):

$$\begin{aligned}\rho'(\gamma(0),P_0) &= \tfrac{d}{d\eta}\rho(I+\eta\Delta,P_0)|_{\eta=0}\\ &= \tfrac{d}{d\eta}\int |E[e^{it^T(I+\eta\Delta)S}]-\textstyle\prod_{k=1}^m E[e^{it_kS_k+it_k\eta\Delta_kS}]|^2dG(t)|_{\eta=0}\\ &= 2\int \mathrm{Re}(\tfrac{d}{d\eta}D(t,\eta,\Delta)\times\overline{D(t,\eta,\Delta)})|_{\eta=0}dG(t)\\ &= 0.\end{aligned}$$

Furthermore,

$$\begin{aligned}\rho''(\gamma(0),P_0) &= \tfrac{d^2}{d\eta^2}\rho(I+\eta\Delta,P_0)|_{\eta=0}\\ &= \tfrac{d^2}{d^2\eta}\int |E[e^{it^T(I+\eta\Delta)S}]-\textstyle\prod_{k=1}^m E[e^{it_kS_k+it_k\eta\Delta_kS}]|^2dG(t)|_{\eta=0}\\ &= 2\int |E[e^{it^TS}]|^2|\textstyle\sum_{k=1}^m \frac{E[\sum_{i\neq k}(t_i\Delta_{ik}S_k-t_k\Delta_{ki}S_i)e^{it_kS_k}]}{E[e^{it_kS_k}]}|^2dG(t);\end{aligned}$$

So, $\rho''(\gamma(0),P_0)\geq 0$, and equality holds if and only if

$$\sum_{k=1}^m \frac{E[\sum_{i\neq k}(t_i\Delta_{ik}S_k-t_k\Delta_{ki}S_i)e^{it_kS_k}]}{E[e^{it_kS_k}]}\equiv 0,\ \text{for any } t\in R^m,$$

or (Let $E_k(t_k)\equiv\frac{E[S_ke^{it_kS_k}]}{E[e^{it_kS_k}]}$)

$$\sum_{k=1}^m\sum_{i\neq k} t_i\Delta_{ik}(E_k(t_k)-E[S_k])\equiv 0.$$

By taking second-order partial derivative w.r.t t_i,t_k

$$\Delta_{ik}\frac{dE_k(t_k)}{dt_k}+\Delta_{ki}\frac{dE_i(t_i)}{dt_i}\equiv 0,\ \text{for any } i\neq k$$

which implies that $\Delta_{ik}\frac{dE_k(t_k)}{dt_k}$ is constant for any $i\neq k$; since there exists Δ_{ik} nonzero (W.O.L.G, say $\Delta_{21}\neq 0$), then $\frac{dE_1(t_1)}{dt_1}\equiv c_0$ for some constant c_0, i.e

$$\frac{d}{dt_1}(\frac{E[S_1e^{it_1S_1}]}{E[e^{it_1S_1}]})\equiv c_0 \text{ or } \frac{E[S_1e^{it_1S_1}]}{E[e^{it_1S_1}]}\equiv c_0t_1+d_0,$$

which implies that $\log E[e^{it_1S_1}]$ is a polynomial of degree at most 2, thus by Lemma 1.2.1 in Kagan et al. (1973), S_1 is normal (or degenerate with mass 1 at its mean). Now if $\Delta_{12}\neq 0$, then similarly we get that S_2 must also be normal; otherwise, c_0 must be 0 which implies that S_1 is degenerate with mass 1 at its mean. However, neither of these is possible by our previous assumptions.

Hence, $\rho''(\gamma(0),P_0)>0$. Notice that $TS(I)$ is compact, thus $\rho''(\gamma(0),P_0)\geq\epsilon_0$ for some $\epsilon_0>0$, i.e., condition (b) holds.

When W_0 is not the identity matrix, we can base our analysis on the parameter space of $V=WW_0^{-1}$. It can be similarly shown that the results hold. ■

Appendix B

Proof of theorem 4.

Proof First consider a m-dim unit ball with its center O, on which P and Q are close to each other, let θ be the acute angle between OP and OQ, i.e $cos(\theta) =< OP, OQ >$ the Euclidean inner product. If we parameterize the path from OP to OQ by $\gamma(t) = cos(t)OP + sin(t)OR$, where OR is the unit tangent vector at P such that OP,OQ and OR are on the same hyperplane (OR is not unique when $\theta = 0$), then $\gamma(0) = P$, $\gamma(\theta) = Q$ and $|\gamma'(t)| = 1$ for $0 \leq t \leq \theta$, i.e γ is the arc from P to Q.

For $\hat{W}$, let the angles between kth row of W_0 and $\hat{W}$ be θ_k, $k = 1, \cdots, m$. Consider the path γ from W_0 to $\hat{W}$ whose component corresponding to the kth row of W_0 is an arc centering at origin on m-dim unit ball starting from W_0 and ending at $\hat{W}$. Parameterize kth component of γ like above but rescale it by $\frac{\theta_k}{\hat{\eta}}$ where $\hat{\eta} = \sqrt{\sum_{k=1}^m \theta_k^2}$ such that $\gamma_k(0) = W_{0k}$ and $\gamma_k(\hat{\eta}) = \hat{W}_k$ and $\gamma_k(.)$ has derivative of norm $\frac{\theta_k}{\hat{\eta}}$. Then it is easy to see that $|\gamma'(t)| = \sqrt{\sum_{k=1}^m (\frac{\theta_k}{\hat{\eta}_k})^2} = 1$, and $|\gamma(t_2) - \gamma(t_1)| \leq |t2 - t1|$ for $|t_2 - t_1|$ small enough by the Mean Value Theorem. It is obvious that $\gamma(t) \in \Omega$ for any $t \in R$.

For fixed $\{x_i : i = 1, \cdots, n\}$, (in the latter of this paper, we use $\rho'(\gamma(\eta), .)$ to denote $\frac{\partial}{\partial t}\rho(\gamma(t), .)|_{t=\eta}$ and $\rho''(\gamma(\eta), .)$ to denote $\frac{\partial^2}{\partial t^2}\rho(\gamma(t), .)|_{t=\eta}$,) expansion of $\rho'(\gamma(\hat{\eta}), P_n)$ about 0 gives:

$$\rho'(\gamma(\hat{\eta}), P_n) = \rho'(\gamma(0), P_n) + \hat{\eta}\rho''(\gamma(\eta^*), P_n)$$

where η^* lies between 0 and $\hat{\eta}$. By assumption, the left side is zero, so that

$$\sqrt{n}\hat{\eta} = -\frac{\sqrt{n}\rho'(\gamma(0), P_n)}{\rho''(\gamma(\eta*), P_n)}.$$

It will be sufficient to show that

$$\sqrt{n}\rho'(\gamma(0), P_n) = O_p(1) \tag{12}$$

and that

$$\rho''(\gamma(\eta^*), P_n) \text{ is bounded away from 0 in probability.} \tag{13}$$

In the following, for any functin f we use $E_n(f(S))$ as empirical mean of $f(S), S = W_0X$. Let Δ denote $\gamma'(\hat{\eta})$. Also we denote $\varsigma_n(t, S) = \psi_n(-t, S) - \prod_{k=1}^m \psi_n(-t_k, S_k)$ in the following proofs.

Of the above statements, (10) follows from the fact that by taking expansions

$$
\begin{aligned}
& \rho'(\gamma(0), P_n) \\
= \; & 2\int \mathrm{Re}\{(E_n[it^T\Delta S e^{it^T S}] - \textstyle\sum_{k=1}^m E_n[it_k\Delta_k S e^{it_k S_k}]\prod_{j\neq k}\psi_n(t_j S_j))\varsigma_n(t,S)\}dG(t) \\
= \; & \tfrac{1}{n^{2m}}\textstyle\sum_{i_1,\cdots,i_{2m}=1}^n\{(S^{i_1}-S^{i_2})^T\Delta(S^{i_1}-S^{i_2})e^{-\frac{||S^{i_1}-S^{i_2}||^2}{2}} \\
& -2(S^{i_{m+1}}-[S_1^{i_1};\cdots;S_m^{i_m}])^T(\Delta S^{i_{m+1}} - diag_{col}(\Delta[S^{i_1},\cdots,S^{i_m}]))e^{-\frac{||S^{i_{m+1}}-(S_1^{i_1},\cdots,S_m^{i_m})^T||^2}{2}} \\
& +[S_1^{i_1}-S_1^{i_{m+1}},\cdots,S_m^{i_m}-S_m^{i_{2m}}]\Delta[S^{i_1}-S^{i_{m+1}},\cdots,S^{i_m}-S^{i_{m+m}}]e^{-\frac{||(S_1^{i_1}-S_1^{i_{m+1}},\cdots,S_m^{i_m}-S_m^{i_{2m}})||^2}{2}}\} \\
= \; & O_p(n^{-\frac{1}{2}}), \text{ from U-statistics' asymptotic normality (Koroljuk \& Borovskich 1994).}
\end{aligned} \tag{14}
$$

Next, for simplicity let $W^* = \gamma(\eta^*), \Delta^* = \gamma'(\eta^*)$, then

$$
\begin{aligned}
& \rho''(\gamma(\eta^*), P_n) \\
= \; & 2\textstyle\int_{R^m}\{|E_n[(it^T(\Delta^*)S)e^{it^T W^* X}] - \sum_{k=1}^m E_n[(it_k(\Delta_k^*)S)e^{it_k W_k^* X}]\prod_{j\neq k}E_n[e^{it_j W_j^* X}]|^2 \\
& -\mathrm{Im}\{(E_n[(t^T(\Delta^*)S)^2 e^{it^T W^* X}] - \textstyle\sum_{k=1}^m E_n[(t_k(\Delta_k^*)S)^2 e^{it_k W_k^* X}]\prod_{j\neq k}E_n[e^{it_j W_j^* X}] \\
& - \textstyle\sum_{k=1}^m E_n[t_k\Delta_k^* S e^{it_k W_k^* X}]\sum_{j\neq k}E_n[it_j\Delta_j^* S e^{it_j W_j^* X}]\prod_{l\neq k,j}E_n[e^{it_l W_l^* X}]) * \varsigma_n(t, W^* X)\}dG(t),
\end{aligned}
$$

where by checking weak convergence

$$
\int |E_n[(it^T(\Delta^*)S)e^{it^T W^* X}] - \sum_{k=1}^m E_n[(it_k(\Delta_k^*)S)e^{it_k W_k^* X}]\prod_{j\neq k}E_n[e^{it_j W_j^* X}]|^2 dG(t)
$$

is bounded below in probability by $\frac{1}{2}min_{||\Delta||=1,\Delta\in TS(W_0)W_0^{-T}}\frac{d^2}{d\eta^2}\rho(I+\eta\Delta, P_0)|_{\eta=0}$ (see Claim 3 in Appendix A) which is greater than 0, and that

$$
\begin{aligned}
& |\textstyle\int_{R^m}\mathrm{Im}\{(E_n[(t^T(\Delta^*)S)^2 e^{it^T W^* X}] - \sum_{k=1}^m E_n[(t_k(\Delta_k^*)S)^2 e^{it_k W_k^* X}]\prod_{j\neq k}E_n[e^{it_j W_j^* X}] \\
& \quad - \textstyle\sum_{k=1}^m E_n[t_k\Delta_k^* S e^{it_k W_k^* X}]\sum_{j\neq k}E_n[it_j\Delta_j^* S e^{it_j W_j^* X}]\prod_{l\neq k,j}E_n[e^{it_l W_l^* X}]) * \varsigma_n(t, W^* X)\}\}dG(t)| \\
& \leq \textstyle\int_{R^m}\{(|t|^2 + m|t|)E_n(|S|^2) + m\prod_{k=1}^m t_k(E_n(|S|))^2\} * |\varsigma_n(t, W^* X)|dG(t)| \\
& = E_n|S|^2\textstyle\int_{R^m}|t|^2|\varsigma_n(t, W^* X)|dG(t) + m(E_n|S|)^2\int_{R^m}\prod_{k=1}^m t_k|\varsigma_n(t, W^* X)|dG(t) \\
& = o_p(1),
\end{aligned}
$$

by recalling that $\varsigma_n(t, W^* X) \to_P 0$ and is bounded by 2, and hence (13) holds.

The desired result follows. ■

Appendix C

Lemma 8 *Let $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$ be IID sequence from two independent population with cdf F and G separately, $m = F^{-1}(\frac{1}{2})$ and $\hat{m} = med(\alpha_n X_i + \beta_n Y_i : i = 1, \cdots, n)$. Assume that F has a first derivative, $F'(x)$ is continuous at m with $F'(m) > 0$. The following results hold:*

(i). If $\alpha_n - 1 = o_p(1)$ and $\beta_n - 1 = o_p(1)$, then $\hat{m} - m = o_p(1)$;

(ii). If $\sqrt{n}(\alpha_n - 1) = O_p(1)$, $\sqrt{n}\beta_n = O_p(1)$, $sup_x F'(x) < \infty$ and

$$\int [F(m - ty) - \frac{1}{2}]dG(y) = O(t) \text{ as } t \to 0, \tag{15}$$

then $\sqrt{n}(\hat{m} - m) = O_p(1)$;

(iii). Especially, under conditions of (ii) with (15) replaced by $E|Y| < \infty$, $\sqrt{n}(\hat{m} - m) = O_p(1)$; this can be extended straightforwardly to the sum of more than two random variables.

Proof Define $H(t, a, b) = P(aX + bY \le t) = \int_R F(\frac{t-by}{a})dG(y)$,$a \neq 0$, then it follows that $H(m, 1, 0) = \frac{1}{2}$. Let $\Psi = \{f : f(x) = I(\alpha x_1 + \beta x_2 \le t), t \in R, x \in R^2\}$, then it is easy to see that Ψ is a VC-subgraph with VC-dimension 4.

First we show (i). By ULLN (van der Geer 2000), $\sup_{f\in\Psi} |E_n f(X) - Ef(X)| \to 0$ in probability, thus $E_n I(\alpha_n X + \beta_n Y \le \hat{m}) - H(\alpha_n, \beta_n, \hat{m}) = o_p(1)$; and by definition $E_n I(\alpha_n X + \beta_n Y \le \hat{m}) - \frac{1}{2} = O_p(n^{-1})$, then $H(\hat{m}, \alpha_n, \beta_n) - \frac{1}{2} = o_p(1)$. Since F is uniform continuous, for any $y \in R$, $F(\frac{\hat{m}}{\alpha_n} - \frac{\beta_n}{\alpha_n} y) - F(\frac{\hat{m}}{\alpha_n}) \to_p 0$ as $n \to +\infty$, then by Dominated Convergence Theorem we have

$$\int_R [F(\frac{\hat{m} - \beta_n y}{\alpha_n}) - F(\frac{\hat{m}}{\alpha_n})]dG(y) = o_p(1), \text{ i.e., } H(\hat{m}, \alpha_n, \beta_n) - F(\frac{\hat{m}}{\alpha_n}) = o_p(1),$$

thus $F(\frac{\hat{m}}{\alpha_n}) - F(m) = o_p(1)$. Since $F'(x)$ is continuous at m and $F'(m) > 0$, it must follow that $\hat{m} = m + o_p(1)$.

Next prove (ii). By P-Donsker's property (van der Geer 2000, van der Vaart & Wellner 1996)

$$sup_{|f_1 - f_2| < \delta} \sqrt{n}|[E_n f_1(X) - Ef_1(X)] - [E_n f_2(X) - Ef_2(X)]| \to_p 0, \text{ as } \delta \to 0, n \to \infty.$$

From (i) we have $\hat{m} = m + o_p(1)$, thus

$$\sqrt{n}|[E_n I(\alpha_n X + \beta_n Y \le \hat{m}) - H(\hat{m}, \alpha_n, \beta_n] - [E_n I(X \le m) - \frac{1}{2}]| = o_p(1).$$

By definition of $\hat{m}$, $\sqrt{n}(E_n I(\alpha_n X + \beta_n Y \le \hat{m}) - \frac{1}{2}) = o_p(1)$, so

$$\sqrt{n}(H(\hat{m}, \alpha_n, \beta_n) - \frac{1}{2}) = \sqrt{n}(E_n I(X \le m) - \frac{1}{2}) + o_p(1),$$

On the other side,

$$\begin{aligned} & \sqrt{n}(H(\hat{m}, \alpha_n, \beta_n) - \tfrac{1}{2}) \\ = \; & \sqrt{n}(\hat{m} - m)\tfrac{H(\hat{m}, \alpha_n, \beta_n) - H(m, \alpha_n, \beta_n)}{\hat{m} - m} \\ & + \sqrt{n}(\alpha_n - 1)\tfrac{H(m, \alpha_n, \beta_n) - H(m, 1, \beta_n/\alpha_n)}{\hat{\alpha}_n - 1} \\ & + \sqrt{n}(H(m, 1, \tfrac{\beta_n}{\alpha_n}) - \tfrac{1}{2}). \end{aligned}$$

By dominated convergence theorem,

$$\frac{H(\hat{m},\alpha_n,\beta_n)-H(m,\alpha_n,\beta_n)}{\hat{m}-m}=\int\frac{F(\frac{\hat{m}-\beta_n y}{\alpha_n})-F(\frac{m-\beta_n y}{\alpha_n})}{\hat{m}-m}dG(y)=F'(m)+o_p(1),$$

and

$$\frac{H(m,\alpha_n,\beta_n)-H(m,1,\beta_n/\alpha_n)}{\hat{\alpha}_n-1}=-mF'(m)+o_p(1).$$

And by assumption there is a sequence of r.v.s $Z_n=O_p(1)$ such that $H(m,1,\beta_n/\alpha_n)-\frac{1}{2}=\frac{\beta_n}{\alpha_n}Z_n$.
Hence,

$$\sqrt{n}(\hat{m}-m)=\frac{\sqrt{n}(E_nI(X\le m)-\frac{1}{2})-\sqrt{n}(\alpha_n-1)mF'(m)+\sqrt{n}\frac{\beta_n}{\alpha_n}Z_n+o_p(1)}{F'(m)+o_p(1)}=O_p(1). \tag{16}$$

For (iii), since $E|Y|<\infty$ implies that in the proof of (ii) $Z_n=F'(m)E[Y]+o_p(1)$, (15) still holds. In case of sum of more than two variables with one coefficient converging to 1 and others to 0, under conditions similar to (iii), $\sqrt{n}$-consistency of the median still holds and the proof is very similar.

■

Note: Even if $E|Y|=\infty$, the lemma's condition of (ii) may still hold, e.g: let X be $U[0,1]$ and Y be Cauchy, $m=0.5$, W.L.O.G let $t>0$, then

$$\int\frac{F(m-ty)-\frac{1}{2}}{t}dG(y)=\int\int_{|y|\le\frac{1}{2t}}ydG(y)-\int_{y>\frac{1}{2t}}dG(y)*\frac{1}{2t}+\int_{y<-\frac{1}{2t}}dG(y)*\frac{1}{2t}=0.$$

In general cases, we need to check whether or not $\int_{|ty|\le1}\frac{F(m-ty)-0.5}{t}dG(y)$ and $\int_{|ty|>1}\frac{F(m-ty)-0.5}{t}dG(y)$ are both bounded from ∞.

References

Amari, S. (2002). Independent component analysis (ica) and method of estimating functions, *IEICE Trans. Fundamentals* **E85-A**(3): 540–547.

Amari, S. & Cardoso, J. (1997). Blind source separation–semiparametric statistical approach, *IEEE Trans. Signal Processing* **45**(11): 2692–2700.

Amari, S., Cichocki, A. & Yang, H. (1996). A new learning algorithm for blind signal separation, *Advances in Neural Information Processing Systems* **8**: 752–763. Cambridge, MA. MIT Press.

Bach, F. & Jordan, M. (2002). Kernel independent component analysis, *Journal of Machine Learning Research* **3**: 1–48.

Bell, A. & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution, *Neural Computation* **7**(6): 1129–1159.

Bickel, P. (1982). On adaptive estimation, *Ann. Statist.* **10**: 647–671.

Bickel, P., Klaassen, C., Ritov, Y. & Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Springer Verlag, New York, NY.

Breiman, L. & Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation, *J. Amer. Statist. Assoc.* **80**: 580–598.

Cardoso, J.-F. (1999). High-order contrasts for independent component analysis, *Neural Computation* **11**(1): 157–192.

Comon, P. (1994). Independent component analysis, a new concept?, *Signal processing* **36**(3): 287–314.

Cox, D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function, *Ann. Inst. Statist. Math.* **37**: 271–288.

Edelman, A., Arias, T. & Smith, S. (1999). The geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications* **20**(2): 303–353.

Fan, J. (1989). *Contributions to the estimation of nonregular functionals*, PhD thesis, University of California, Berkeley.

Feuerverger, A. & Mureika, R. (1977). The empirical characteristic function and its applications, *Ann. Statist.* **5**(1): 88–97.

Hastie, T. & Tibshirani, R. (2002). Independent component analysis through product density estimation, *Technical report*, Stanford University.

Huber, P. (1967). *The behavior of maximum likelihood estimates under nonstandard conditions*, Vol. 1, Univ. California Press, Berkeley.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis, *IEEE Transactions on Neural Networks* **10**(3): 626–634.

Hyvarinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, John Wiley & Sons, New York, NY.

Hyvarinen, A. & Oja, E. (1997). A fast fixed point algorithm for independent component analysis, *Neural Computation* **9**(7): 1483–1492.

Jin, K. (1992). Empirical smoothing parameter selection in adaptive estimation, *Ann. Statist.* **20**(4): 1844–1874.

Jutten, C. & Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture, *Signal Processing* **24**: 1–10.

Kagan, A., Linnik, Y. & Rao, C. (1973). *Characterization Problems in Mathematical Statistics*, John Wiley & Sons, USA.

Klaassen, C. (1987). Consistent estimation of the influence function of locally asymptotically linear estimates, *Ann. Statist.* **15**: 1548–1562.

Koroljuk, V. & Borovskich, Y. (1994). *Theory of U-Statistics*, Kluwer Academic Publishers, the Netherlands.

MacKay, D. (1996). Maximum likelihood and covariant algorithms for independent component analysis, *Unpublished* .

Murata, N. (2001). Properties of the empirical characteristic function and its application to testing for independence, *Proceedings of 3rd International Conference on Independent Component Analysis and Signal Separation*, San Diego, CA.

Murphy, S. & van der Vaart, A. (2000). On profile likelihood, *Journal of the American Statistical Association* **95**: 449–485.

Pham, D.-T. & Garrat, P. (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood appoach, *IEEE Trans. on Signal Processing* **45**(7): 1712–1725.

Samarov, A. & Tsybakov, A. (2002). Nonparametric independent component analysis, *Technical report*, Université Paris.

van der Geer, S. (2000). *Applications of Empirical Process Theory*, Cambridge University Press, UK.

van der Vaart, A. & Wellner, J. (1996). *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

Wald, A. (1949). Note on the consistency of maximum likelihood estimate, *Ann. Math. Statist.* **20**: 595–601.