EFFICIENT INDEPENDENT COMPONENT ANALYSIS [*] [‡]

By Aiyou Chen and Peter J. Bickel

Submitted July 2003, Revised Nov. 2004

University of California, Berkeley

Independent component analysis (ICA) has been widely used for blind source separation in many fields such as brain imaging analysis, signal processing, telecommunication. Many statistical techniques based on M-estimates have been proposed in estimating the mixing matrix. Recently a few methods based on nonparametric tools are also available. However, in-depth analysis on asymptotic efficiency has not been available. In this paper we analyze ICA under the framework of semiparametric theories [see Bickel, Klaassen, Ritov and Wellner (1993)] and propose a straightforword estimate based on the efficient score function by using B-spline approximations. This estimate exhibits better performance than stardard ICA methods in a variety of simulations. It is proved that this estimate is asymptotically efficient under moderate conditions.

1. Introduction. Independent component analysis (ICA) aims to separate blind sources from their observed linear mixtures without any prior knowledge, where blind sources are assumed to be mutually independent. This technique has been widely used in the past decade to extract useful features from observed data in many fields such as brain imaging analysis, signal processing, telecommunication. Hyvarinen, Karhunen and Oja (2001) described many effective applications of ICA in different fields. For example the ICA method was shown able to separate artifacts from magnetoencephalography (MEG) data, without modelling the process that generated the artifacts, by Vigario, Jousmaki, Hamalainen, Hari and Oja (1998).

The standard ICA models an $m \times 1$ random vector X (e.g., instantaneous magnetoencephalogical image) by linear mixtures of m mutually independent random variables $(S_1, \cdots, S_m)$ (e.g., artifacts, other brain activities), but each S_i's distribution is totally unknown. That is, for $S = (S_1, \cdots, S_m)^T$ and some $m \times m$

[*]Primarily sponsored by NSF Grant DMS-01-04075.
[†]*AMS 2000 subject classifications.* Primary 62G05; secondary 62H12.
[‡]*Key words and phrases.* Independent component analysis, semiparametric models, efficient score function, asymptotically efficient, generalized M-estimator, B-splines.

matrix θ,

$$X \quad = \quad \theta S. \tag{1}$$

Here θ is called the mixing matrix, assumed nonsingular. Given n independent observations $(X^1, \cdots, X^n)$ from the distribution of X, it is desirable to estimate θ and thus to separate each $S_i = (\theta^{-1}X)_i$. Let $W = \theta^{-1}$ (called the unmixing matrix). Then the aim is equivalent to finding a W such as $S = WX$ has mutually independent components. This can be seen as a projection pursuit problem in seeking for m directions such that the corresponding projections are most mutually independent.

It was shown by Comon (1994) that W is identifiable up to scaling and permutation of its rows if at most one of S's components is normal. The model (1) can be viewed as a semiparametric model with parameters $(W, r_1, \cdots, r_m)$, where r_i parametrizes S_i's density/mass function. In this paper, W is the parameter of interest and $(r_1, \cdots, r_m)$ which themselves can only be identified up to permutation and scale are the nuisance parameters.

Since ICA was motivated by neurophysiological problems in the early 1980s [e.g., Hyvarinen, Karhunen and Oja (2001)], there have been many methods proposed to estimate W. They fall into two classes. One class involves specifying a particular parametric model for each r_i and then optimizing contrast function suggested by the model of the data and $(W, r_1, \cdots, r_m)$ as a function of the latter. The primary examples of this approach are maximum likelihood (ML) [e.g., Pham and Garrat (1997) and Lee, Girolami and Sejnowski (1999)] or equivalently minimizing mutual information [e.g., Comon (1994)], minimizing high-order correlation between WX's components [e.g., Cardoso (1999)], and maximizing the non-gaussianity of WX's components [e.g., Hyvarinen (1999)]. The second approach is to view ICA as a semiparametric model and assume nothing about the distribution of the components of S. Thus two distinct goals can be formulated. (i). To find estimates $\hat{W}$ of W which are consistent or even better $\sqrt{n}$ consistent, that is, $\hat{W} = W + O_p(n^{-1/2})$. (ii). To find procedures which achieve the information bound, that is, estimates of W which are asymptotically normal and have smallest variance-covariance matrix among all estimates which are in a suitable sense uniformly asymptotically normal - see for instance Bickel, Klasseen, Ritov and Wellner (1993) (BKRW). Amari (2002) formally demonstrated that to achieve the information bound in this situation, estimates had to be based on methods which estimated the densities of the sources. In fact it can even be shown [Cardoso (1998)] that for any fixed estimating equation corresponding to maximizing an objective function, there is a possible distribution of sources for which the global maximizer which is a

solution of the estimating equation is inconsistent, despite the consistency of a local solution near the truth.

Recently, some nonparametric methods to estimate W have appeared. For example, Bach and Jordan (2002) proposed: 1) To reduce the dimension of the data using a kernel representation; 2) To choose W so as to minimize the empirical *generalized variance* between the components of WX. Hastie and Tibshirani (2002) proposed maximizing the penalized likelihood as a function of $(W, r_1, \cdots, r_m)$ and Vlassis and Motomura (2001) proposed maximizing the likelihood by using Gaussian kernel density estimation. The Vlassis-Motomura and Hastie-Tibshirani methods are of the same type as ours but in both cases, neither was a method of tuning suggested nor was anything proved about the property of the procedures. Various performance analyses have been made using simulations. However, none of these procedures have been analyzed theoretically. Samarov and Tsybakov (2002) proposed and analyzed an estimate $\sqrt{n}$ consistent under mild conditions. We [Chen and Bickel (2004)] analyzed in detail a characteristic-function based method [Eriksson and Koivunen (2003)] and showed it to be consistent under the minimal identifiability conditions and $\sqrt{n}$ consistent under mild conditions. Our concern in this paper is the construction of efficient estimates. We develop an efficient estimator by using a sieve profile likelihood technique starting the algorithm at a consistent point. The characteristic-function based ICA algorithm PCFICA [Chen and Bickel (2004)] is used both theoretically and in our simulations for this critical starting point.

In the following, we analyze the ICA model (1) in the framework of semiparametric models [see, e.g., BKRW] and propose a new method of estimating W using the efficient score function, as developed in Section 2. The main theorem is given in Section 3. Numerical studies are given in Section 4. Section 5, Section 6 and an appendix contain the technical details.

Notations: In this paper, W denotes an $m \times m$ matrix, W_i and W_{ij} denote the ith row and the (i,j)th element of W separately. The superscript ^{T} denotes the transpose of a matrix or vector. For any matrix A, $||A||_F = \sqrt{tr(A^T A)}$.

2. Semiparametric inference.

2.0. *Efficient estimates for semiparametric models.* In this subsection we review briefly the salient features of estimation in semiparametric models.

Given a semiparametric model, $X^1, \cdots, X^n$ i.i.d $\{P_{(\theta,\eta)} : \theta \in \Omega \subset R^d, \eta \in \mathcal{E}\}$, where $\mathcal{E}$ is a subset of a function space, estimates $\hat{\theta}_n$ of θ are called regular if $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in law uniformly in $P_{(\theta_n,\eta_n)}$ where (θ_n, η_n) converges to (θ_0, η_0) in a smooth way. Then : a) If there is a uniformly best estimate among such

$\hat{\theta}_n$, call it θ_n^*, it must have the form under $P_{(\theta,\eta)}$,

$$\theta_n^* \quad = \quad \theta + \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{l}}(X^i;\theta,\eta) + o_p(n^{-1/2}),$$

where $\tilde{\mathbf{l}}$ may be computed more or less explicitly by suitable projections in $L_2(P_{(\theta,\eta)})$. Closely related is the efficient score function, $\mathbf{l}^* \equiv \mathbf{I}(\theta,\eta)\tilde{\mathbf{l}}$, where $\mathbf{I}(\theta,\eta) = \{E_{(\theta,\eta)}(\tilde{\mathbf{l}}\tilde{\mathbf{l}}^T)\}^{-1}$; b) If $(\tilde{\theta}_n, \tilde{\eta}_n)$ are consistent estimates of (θ,η) and some additional regularity conditions hold, then the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{l}^*(X^i;\theta,\tilde{\eta}_n) \quad = \quad 0 \tag{2}$$

obtained by starting at $(\tilde{\theta}_n, \tilde{\eta}_n)$ is efficient. For a suitable construction of $\hat{\eta}_n(\theta)$ such that $\hat{\eta}_n(\theta)$ is efficient for each η for the model $\mathcal{P}_1 = \{P_{(\theta,\eta)} : \ \eta \in \mathcal{E}\}$, it is possible to replace $\mathbf{l}^*$ by $\frac{\partial l}{\partial \theta}$, where l is the logarithmic likelihood function, if we take $\tilde{\eta}_n$ not fixed but equal to $\hat{\eta}_n(\theta)$. We employ this approach. This and related methods may be found in BKRW Chapter 7 and Murphy and van de Vaart (2000) who call this method profile likelihood.

We note again that this technique is different from that which we have called type (i), known as Quasi ML in the ICA literature which corresponds to a subset of class 1 of the estimates we have considered. The technique is to guess some shape η_0 for η and then do ordinary ML. Of course if η_0 is true, then the resulting estimate is asymptotically Gaussian and has smaller variance than the $\hat{\theta}$ we discuss. But if η_0 is false then the estimate can be inconsistent. The ICA algorithms which we compare ours to in Section 4 such as FastICA [Hyvarinen and Oja (1997)] and Extended Infomax [Lee, Girolami and Senjowski (1998)] are of this type. Pham and Garrat (1997) do consider parametric models such as the logsplines we introduce in Section 2.3. However they do not suggest increasing the dimension of their model with n and hence are subject to the difficulties with consistency that we already discussed.

In the next four subsections, we show how to implement the idea given in (2) for the ICA model. The technical analysis is carried out in Section 3 under the framework of *generalized M-estimates* [see e.g., BKRW].

2.1. *Some notation and further assumptions.* Let W_P be one of the nonsingular unmixing matrices such that $S = W_P X$ has m mutually independent components. Without loss of generality, we may assume that $\det(W_P) > 0$. For any row vector $w \in \mathcal{R}^m$, we use f_w as the probability density function (pdf) of wX and use ϕ_w as the density score function associated with f_w defined by $\phi_w(t) = -\frac{\partial}{\partial t}\log f_w(t) I(f_w(t) > 0)$.

As we mentioned earlier, in the model (1) the order and scaling of either W's rows or S's components need to be defined for the identifiability of W. Here we assume that each S_i has absolute median 1 to control the scaling ambiguity, i.e., $P(|S_i| \leq 1) = \frac{1}{2}$, or equivalently

$$2\int_{-1}^{1} r_i(s)ds \quad = \quad 1. \tag{3}$$

Even after this choice, the correct unmixing matrix requires $2^m m!$ choices due to sign changes and row permutations. This ambiguity can be resolved in many different ways, but we need not strictly specify this since we assume in this paper that we have at hand a raw consistent starting value for W_P, say PCFICA of Chen & Bickel (2004). Define $k(s) = 2I(|s| \leq 1) - 1$, where $I(.)$ is an indicator function. Then (3) is equivalent to

$$\int k(S_i)dP \quad = \quad 0.$$

We proceed to calculate $\mathbf{l}^*$. The terms used in and operations needed for the following calculation may be found in BKRW chapter 3 (page 51, 70).

2.2. *Efficient score function of* W. By parametrizing the model (1) with $(W, r_1, \cdots, r_m)$, the likelihood function of X can be expressed as

$$p_X(\mathbf{x}; W, r_1, \cdots, r_m) \quad = \quad |\det(W)| \prod_{i=1}^{m} r_i(W_i\mathbf{x}).$$

The parameter of interest is W and $(r_1, \cdots, r_m)$ are the nuisance parameters. In the following we heuristically calculate the efficient score function of W as in Section 2.0, but refer to BKRW Chapter 3 for geometric intuition and relevant calculus. For simplicity, we assume $E[S_i] = 0$. For the convenience of notation, let E be the expectation operator under P.

Let $\phi_i(s_i) = -\frac{\partial}{\partial s_i} \log r_i(s_i) I(r_i(s_i) > 0)$ be the density score function associated with r_i and define Φ by $\Phi(\mathbf{s}) = (\phi_1(s_1), \cdots, \phi_m(s_m))^T$, where $\mathbf{s} = (s_1, \cdots, s_m)^T$. Then the score function of W, $\dot{\mathbf{l}}_W(\mathbf{x}) \equiv \frac{\partial}{\partial W} \log(p_X(\mathbf{x}; W, r_1, \cdots, r_m))$, is equal to

$$\dot{\mathbf{l}}_W(\mathbf{x}) \quad = \quad (I_{m\times m} - \Phi(\mathbf{s})\mathbf{s}^T)W^{-T}, \text{ where } \mathbf{s} = W\mathbf{x}.$$

From this, the minimal regularity conditions for talking about effecient estimation are that each r_i should be absolutely continuous, W nonsingular, and

$$E[\phi_i(S_i)^2] \quad < \quad \infty \text{ and } E[S_i^2] < \infty. \tag{4}$$

To calculate the tangent vectors for each nuisance parameter r_i, we take the representation $r_i(\cdot;t) = r_i(\cdot)e^{th_i(\cdot)}$ for $t \in \mathcal{R}$ close to 0, then the tangent vector w.r.t h_i is

$$\text{limit}_{t\to 0}\frac{\partial}{\partial t}\log p_X(\mathbf{x}; W, r_1, r_i(\cdot;t), r_m) \quad = \quad h_i(W_i\mathbf{x}).$$

Since $r_i(\cdot;t)$ needs to be a probability density function and to satisfy the mean and absolute median assumptions, h_i needs to satisfy $E[h_i(S_i)] = 0, E[h_i(S_i)S_i] = 0, E[h_i(S_i)k(S_i)] = 0$ and is otherwise arbitrary. Thus the tangent space of the nuisance score of r_i can be expressed as

$$TS_i \quad = \quad \{h_i(W_i\mathbf{x}) \in L_2(P_{(W,\mathbf{r})}) | E[h_i(S_i)] = 0, E[h_i(S_i)S_i] = 0, E[h_i(S_i)k(S_i)] = 0\}.$$

Notice that the tangent spaces $\{TS_i : 1 \le i \le m\}$ are perpendicular to each other since S_is are mutually independent. Thus any projection onto the tangent space of $(r_1, \cdots, r_m)$ is equal to the summation of the partial projection onto each TS_i. The efficient score of W can then be expressed as

$$\mathbf{l}^*(.; W, \Phi) \quad = \quad \dot{\mathbf{l}}_W - \sum_{i=1}^{m}\pi(\dot{\mathbf{l}}_W | TS_i),$$

where $\pi(.|L)$ denotes the projection operator in the Hilbert space $L_2(P_{(W,r_1,\cdots,r_m)})$ onto L. After some calculation we find that the efficient score $\mathbf{l}^*(.; W, \Phi)$ is equal to

$$\mathbf{l}^*(\mathbf{x}; W, \Phi) \quad = \quad \mathbf{M}W^{-T}, \tag{5}$$

where $\mathbf{M}$ is a $m \times m$ function matrix and its elements are given by

$$\mathbf{M}_{ij} \quad = \quad -\phi_i(W_i\mathbf{x})W_j\mathbf{x}, \text{ for } 1 \le i \ne j \le m, \tag{6}$$

$$\mathbf{M}_{ii} \quad = \quad \alpha_i W_i\mathbf{x} + \beta_i k(W_i\mathbf{x}), \text{ for } i = 1, \cdots, m, \tag{7}$$

and $\alpha = (\alpha_1, \cdots, \alpha_m)^T, \beta = (\beta_1, \cdots, \beta_m)^T, \sigma^2 = (\sigma_1^2, \cdots, \sigma_m^2)^T$ are defined by

$$\alpha_i = -\frac{(1-u_i)v_i}{\sigma_i^2 - v_i^2},\ \beta_i = \frac{(1-u_i)\sigma_i^2}{\sigma_i^2 - v_i^2},\ \sigma_i^2 = E[S_i^2], \tag{8}$$

$$v_i = E[2S_i I(|S_i| \leq 1)],\ u_i = E[2S_i\phi_i I(|S_i| \leq 1)]. \tag{9}$$

(Note: Most of these formulas have been obtained by Amari and Cardoso (1997).) By the convolution theorem on semiparametric models [BKRW], the information bound for regular estimators of W is $(E[\mathbf{l}^*\mathbf{l}^{*T}(X; W, \Phi)])^{-1}$, where $\mathbf{l}^*(.; W, \Phi)$ is considered as a column vector function, reshaped row by row. It is obvious that the efficient score function only depends on $(r_1, \cdots, r_m)$ through their density score functions $(\phi_1, \cdots, \phi_m)$. In the next subsection, we focus on the estimation of a density score function.

2.3. *Estimating the density score function by B-spline approximations.* Let $\phi = -\frac{r'}{r}$ be the density score associated with a univariate pdf r and $\mathcal{G}$ be a linear space with differentiable basis functions $\mathbf{B} = (B_1, \cdots, B_N)^T$. An estimator of ϕ in $\mathcal{G}$ can then be obtained by minimizing the mean square error $c(\gamma)$ for $\gamma \in R^N$, which is defined as

$$c(\gamma) = \int_R (\phi(s) - \gamma^T \mathbf{B}(s))^2 r(s) ds.$$

By partial integration,

$$c(\gamma) = \gamma^T E_r[\mathbf{B}^T\mathbf{B}]\gamma - 2\gamma^T E_r[\mathbf{B}'] + E_r[\phi^2],$$

where E_r is the expectation operator under the probability measure $r(s)ds$. Thus the optimal γ is $\gamma_\phi = (E_r[\mathbf{B}^T\mathbf{B}])^{-1}E_r[\mathbf{B}']$, where $\mathbf{B}'$ is the derivative of $\mathbf{B}$, and the best approximation of ϕ in $\mathcal{G}$ in the sense of mean square error is $\phi_\mathcal{G} = \gamma_\phi^T\mathbf{B}$. This method was proposed by Jin (1992), as a variant of Cox (1985)'s penalized estimator of ϕ. Given n random samples from the density function r, γ_ϕ can be estimated by combinations of empirical moments. So a natural estimator of ϕ is given by

$$\hat{\phi}_\mathcal{G} = \gamma_n^T\mathbf{B},\ \text{where } \gamma_n = (\hat{E}_r[\mathbf{B}^T\mathbf{B}])^{-1}\hat{E}_r[\mathbf{B}'], \tag{10}$$

and $\hat{E}_r$ denotes the empirical mean operator corresponding to E_r. This method has also been used to

estimate density score functions in the ICA literature, for example by Pham & Garrat (1997).

B-spline basis functions are popular choices for $\mathcal{G}$. In general, the support of r is unknown and we need to choose a working interval $[\underline{b}_n, \overline{b}_n] \subseteq supp(r)$, where some knots are distributed for the construction of the basis functions. To choose $\underline{b}_n$ and $\overline{b}_n$ empirically, we may use for example 1% and 99% empirical quantiles. The basic rule for adaptation is that $[\underline{b}_n, \overline{b}_n] \to supp(r)$ very slowly as $n \to \infty$. The number of basis functions, say N, is an empirical smoothing parameter, which can be dealt with as usual by cross validation. Let

$$PE_n(\gamma) \quad = \quad \gamma^T \hat{E}_r[\mathbf{B}^T\mathbf{B}]\gamma - 2\gamma^T \hat{E}_r[\mathbf{B}']$$

be the empirical prediction error. In this paper, we use two-fold cross validation, that is, by splitting the samples into two half the optimal number of knots is to minimize the average empirical prediction error after alternatively using one half samples to estimate γ and the other half to calculate the empirical prediction error. Jin (1992) used B-spline basis functions for $\mathcal{G}$ and showed that the adaptive choice of N by cross validation under weak conditions on r's smoothness is

$$N \quad = \quad O(n^{\delta}),$$

where $0 < \delta < \frac{1}{6}$ and δ depends on the tail property of r.

2.4. *Estimation of W.* Assume that an available starting estimate $\hat{W}^{(0)}$ is consistent for W_P. We show how to construct an estimate $\hat{\Phi}_W$ of Φ_W and then solve

$$\int \mathbf{l}^*(X; W, \hat{\Phi}_W) dP_n(X) \quad = \quad 0.$$

Here $\hat{\Phi}_W$ is a data dependent function of W. Thus $\mathbf{l}^*(X; W, \hat{\Phi}_W)$ is in fact an approximate efficient score function. Our construction is such that $\hat{\Phi}_W$ is approximately efficient in the sense that smooth functionals of $\hat{\Phi}_W$ such as $\int_{-\infty}^{x_0} \int_{-\infty}^{x} \hat{\Phi}_W(y) dy dx$ are efficient estimates of the corresponding population function.

For each $k \in \{1, \cdots, m\}$, we choose a sieve for $\hat{\phi}_{W_k}$ as follows. Let $[\underline{b}_{nk}, \overline{b}_{nk}] \subset \mathcal{R}$ be a subset of $supp(r_k)$ containing most of the mass of r_k. For an integer n_k, set $n_k + 4$ points $\{\underline{b}_{nk} + (i-1)\delta_{nk} : 1 \leq i \leq n_k + 4\}$ as the knots , where δ_{nk} depends on n_k through

$$\delta_{nk} \quad = \quad (\overline{b}_{nk} - \underline{b}_{nk})/(n_k + 3),$$

and then construct n_k cubic B-spline basis functions as in the appendix. Denote the basis functions as $\mathbf{B}_n^{(k)} \equiv (B_{n1}^{(k)}, \cdots, B_{nn_k}^{(k)})^T$, where the superscript (k) denotes the association with S_k and the first subscript n denotes the dependence with the sample size. Given the random samples $\{W_k X^i : 1 \leq i \leq n\}$ from the density function f_{W_k}, we use (10) to approximate its density score function ϕ_{W_k} by using these basis functions. Here n_k is chosen by cross-validation as described in the previous subsection. To avoid further complications, we assume that both $[\underline{b}_{nk}, \overline{b}_{nk}]$ and n_k are fixed using $\hat{W}^{(0)}$ once for all. That is, the $n_k + 4$ knots are fixed. In the algorithm they are random and depend on the initial estimate $\hat{W}^{(0)}$. Then for any updated W_k we have an available sieve estimator $\hat{\phi}_{W_k}$ for ϕ_{W_k} by (10), that is,

$$\hat{\phi}_{W_k} \quad = \quad [\gamma_n(W_k)]^T \mathbf{B}_n^{(k)},$$

where $\gamma_n(w) = A_n^{-1}(w) D_n(w)$ with $A_n(w) = \int \mathbf{B}_n^{(k)} \mathbf{B}_n^{(k)T}(wX) dP_n$ and $D_n(w) = \int [\mathbf{B}_n^{(k)}]'(wX) dP_n$. Here $[\mathbf{B}_n^{(k)}]'(x) \equiv (\frac{d}{dx} B_{n1}^{(k)}(x), \cdots, \frac{d}{dx} B_{nn_k}(x))^T$ denots its derivative and will be used thereafter.

Now we replace the efficient score function $\mathbf{l}^*(X; W, \Phi)$ defined in (5) by its profile form $\mathbf{l}^*(X; W, \hat{\Phi}_W)$, where $\alpha_i, \beta_i, \sigma_i^2$ defined in (8) and (9) are estimated by moments with plugged-in parameters $(W, \hat{\Phi}_W)$. Denote their estimates by $\hat{\alpha}_i, \hat{\beta}_i, \hat{\sigma}_i^2$ separately, i.e. :

$$\hat{\alpha}_i \quad = \quad -\frac{(1 - \hat{u}_i)\hat{v}_i}{\hat{\sigma}_i^2 - \hat{v}_i^2}, \ \hat{\beta}_i = \frac{(1 - \hat{u}_i)\hat{\sigma}_i^2}{\hat{\sigma}_i^2 - \hat{v}_i^2}, \ \hat{\sigma}_i^2 = \int (W_i X)^2 dP_n, \tag{11}$$

where $\hat{u}_i = \int_{Y = W_i X} 2Y \hat{\phi}_{W_i}(Y) I(|Y| \leq 1) dP_n, \ \hat{v}_i = \int_{Y = W_i X} 2Y I(|Y| \leq 1) dP_n$.

Let $\Phi_W = [\phi_{W_1}, \cdots, \phi_{W_m}]$. Define

$$\mathbf{e}_n(W) \quad = \quad \int \mathbf{l}^*(X; W, \hat{\Phi}_W) dP_n \text{ and } \mathbf{e}(W) = \int \mathbf{l}^*(X; W, \Phi_W) dP. \tag{12}$$

Let $\hat{W}_n$ be the solution of

$$\mathbf{e}_n(W) \quad = \quad 0. \tag{13}$$

Let $\hat{\mathbf{l}}^*(\mathbf{x}; W) \equiv \mathbf{l}^*(\mathbf{x}; W, \hat{\Phi}_W)$ and $\dot{\mathbf{e}}_n(W) \equiv \frac{\partial}{\partial W} \mathbf{e}_n(W)$. Notice that $-\dot{\mathbf{e}}_n(\hat{W})$ and $\int \hat{\mathbf{l}}^* \hat{\mathbf{l}}^{*T}(X; \hat{W}) dP_n$ have the same limit

$$-\frac{\partial \mathbf{e}(W)}{\partial W}|_{W_P} \quad = \quad E[\mathbf{l}^* \mathbf{l}^{*T}(X; W_P, \Phi_P)]$$

with probability converging to 1, if $\hat{W} \to W_P$, as developed later in Section 5, we employ the following approximate Newton-Rapson scheme :

$$\hat{W}^{(j+1)} \quad = \quad \hat{W}^{(j)} + [\int \hat{\mathbf{l}}^* \hat{\mathbf{l}}^{*T}(X; \hat{W}^{(j)}) dP_n]^{-1} \mathbf{e}_n(\hat{W}^{(j)}), \ j = 0, 1, \cdots \tag{14}$$

Note that this method does not require calculating the Hessian matrix $\dot{\mathbf{e}}_n(W)$ but achieves the same efficiency as using the exact Newton-Rapson algorithm. The convergence and asymptotic properties of (14) is developed in Section 3. Call $\hat{W} \equiv \hat{W}^{(\infty)}$ defined by (14) the EFFICA, which will be used for the simulation in Section 4.

3. **Asymptotic properties.** We are given $\hat{W}^{(0)}$ such that there exists ε_n with $\varepsilon_n \to 0$, $\sqrt{n}\varepsilon_n \to \infty$ such that as $n \to \infty$,

$$P(||\hat{W}^{(0)} - W_P||_F \quad \leq \quad \varepsilon_n) \to 1. \tag{15}$$

Recall that PCFICA does this. Let

$$\Omega_n \quad = \quad \{W \in \mathcal{R}^{m \times m} : ||W - W_P||_F < \varepsilon_n\}. \tag{16}$$

We need the following conditions. Let $\phi_{W_k,n}(x) \equiv \phi_{W_k}(x) I(x \in [\underline{b}_{nk}, \overline{b}_{nk}])$.

C1: W_P is nonsingular.

C2: $E[S_k] = 0$, $E[S_k^2] < \infty$, $med(|S_k|) = 1$ and $E(\phi_k(S_k))^2 < \infty$.

C3: $|r_k|_\infty < \infty, |r_k'|_\infty < \infty$, $sup_{t \in \mathcal{R}} |tr_k'(t)| < \infty$.

C4: The Uniform Law of Large Numbers (ULLN) holds for $\{\phi_{W_k}(W_k X) X_i : W \in \Omega_n\}$, $\{\phi_{W_k}^2(W_k X) X_i^2 : W \in \Omega_n\}$ and for $\{\phi_{W_k}'(W_k X) W_i X X_j : W \in \Omega_n\}$.

C5: For some positive constants c_1, c_2, $r_k(t) \geq c_1 \delta_{nk}$ if $t \in [\underline{b}_{nk}, \overline{b}_{nk}]$, otherwise $r_k(t) \leq c_2 \delta_{nk}$.

C6: $sup_{W \in \Omega_n} |\phi_{W_k,n}|_\infty \delta_{nk} = O(1)$ and $sup_{W \in \Omega_n} |\phi_{W_k,n}'''|_\infty \delta_{nk} = o(1)$.

C7: $\varepsilon_n \delta_{nk}^{-\frac{11}{2}} (\overline{b}_{nk} - \underline{b}_{nk}) = o(1)$, where ε_n, δ_{nk} and $[\underline{b}_{nk}, \overline{b}_{nk}]$ are as in (15) and C5.

(Note: ULLN holds for $\mathcal{G}_n$ iff $\sup_{g \in \mathcal{G}_n} |\int g(X) d(P_n - P)| = o_p(1)$, see for example van de Geer (2000).)

Condition C1-C3 can be considered as the simplified regularity conditions. Condition C1 and the finite second moments on S_ks and its density score functions in Condition C2 are among the minimal regularity

conditions for talking about efficiency, as we mentioned in Section 2.2. The absolute median in Condition C2 is a simple and minimal condition to make the scales of the unmixing matrix identifiable [Comon (1994)]. It should be clear that the zero mean assumption in Condition C2 is in no way crucial to the general argument as the mean can be estimated adaptively, but serves to keep algebraic complication to a minimum. Condition C3 assumes some smoothness on the density score function ϕ_k for each hidden component, which is needed so that it can be well approximated by B-splines.

Condition C4-C7 are technical conditions which we believe are far from necessary but are reasonably easy to check and whose use enables construction of a more compact proof. As an easy example, if $|\phi_k|_\infty < \infty$ and $|\frac{r_k''}{r_k}|_\infty < \infty$ for $k = 1, \cdots, m$, then by (32) $\sup_{W \in \Omega_n} |\phi_{W_k}|_\infty < \infty$ and by (33) $\sup_{\Omega_n} |\phi'_{W_k}|_\infty < \infty$, thus C4 holds. Condition C5 and C6 require that the tail of r_k be not too wiggly. Condition C6 also implies $\delta_{nk} \to 0$. Condition C7 requires that the initial value be reasonably close to the truth and that the domain and the number of knots of the B splines (i.e., $n_k = (\overline{b}_{nk} - \underline{b}_{nk})\delta_{nk}^{-1} - 3$) do not grow so quickly that we lose control of the approximation to Φ_W.

Here is our main theorem.

THEOREM 1. *In the ICA model (1), if (15) and C1-C7 hold for $i, j, k = 1, \cdots, m$, $i \neq k$ and $j \neq k$. Then with probability converging to 1 the algorithm (14) has a limit $\hat{W}^{(\infty)}$* AND

$$\sqrt{n}(\hat{W}^{(\infty)} - W_P) \quad = \quad I_{eff}^{-1} \sqrt{n} \int \mathbf{l}^*(X; W_P, \Phi_P) dP_n + o_P(1), \tag{17}$$

where $I_{eff} = \int \mathbf{l}^* \mathbf{l}^{*T}(X; W_P, \Phi_P) dP$. That is, $\hat{W}^{(\infty)}$ is Fisher efficient. (Note: (17) is considered in a vector form.)

The proof of Theorem 1 is provided in later Sections and the Appendix.

4. Numerical studies and some computational issues. We do two groups of experiments to test the empirical performance of the EFFICA. We generate data from known source distributions listed in Table 1 and then obtain linear mixtures of them by a known mixing matrix $\theta = W_P^{-1}$. In the EFFICA, to choose the boundaries for B-spline approximation of the density score functions, we use the maximal value between the 1% empirical quantile minus Δ_n and 0% quantile as $\underline{b}_{nk}$, and use the minimal value between the 99% empirical quantile plus Δ_n and 100% empirical quantile as $\overline{b}_{nk}$, where $\Delta_n = O(\sqrt{\log \log n})$ is used in our simulation. The choice of the number of knots is a key issue for EFFICA. In practice, we use two-fold cross

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

Table 1: Source distributions used in the simulations

validation (CV) as we mentioned in Section 2.3 to calculate the best empirical prediction error associated with the number of knots n_k, which starts from $n_k = 1, 2, \cdots$, and then we choose the first n_k such that the best empirical prediction error strictly decreases w.r.t the number of knots till n_k. This method was used and shown by Jin (1992) able to find an appropriate n_k, where Jin used *smoothing cross-validation* instead of two-fold CV.

In the first group of experiments, we use 2 hidden components, and $W_P = [2, 1; 2, 3]$. The two components in the first 12 experiments are i.i.d from one of the distributions [1]-[12], and the two components in experiments 13-15 are independent but are from different distributions given in one of cases [13]-[15] in Table 1 separately. Each of these experiments has been replicated 400 times.

In the second group of experiments, we increase the number m of hidden components to 4, 8 and 12 separately (the detailed setup of the sample sizes and replication times is given in Table 3). The m hidden components are chosen in order from the first m source distributions of [0], [1], $\cdots$, [11] in Table 1, and without loss of generality we use the identity matrix for W_P.

Comparisons are made with five existing ICA algorithms: the FastICA algorithm with the options of "symmetric" and "tanh" [Hyvarinen & Oja (1997)], which is equivalent to Quasi ML by specifying the density score function of each hidden source by the optimal one between $-2tanh(\cdot)$ and $tanh(\cdot) - \cdot$, the Jade-ICA algorithm [Cardoso (1999)], the extended Infomax algorithm [Lee, Girolami & Sejnowski (1998)], the KernelICA-Kgv algorithm [Bach & Jordan (2002)], and the PCFICA algorithm [Chen & Bickel 2004)] which has been analyzed thoroughly. We used the estimate obtained by PCFICA as the initial value for EFFICA and KernelICA-Kgv. For computational simplicity, we use FastICA's estimate to initiate PCFICA. Note that restarting is necessary for PCFICA since the algorithm is not convex. The performance of each algorithm is measured by both the Frobenius error, i.e., $d_F(\hat{W}, W_P) = ||\hat{W}W_P^{-1} - I||_F$ after suitable rescaling and permutation on rows of both $\hat{W}$ and W_P, and the so-called *Amari error* $d_A(\hat{W}, W_P)$ [Amari, Cichocki

& Yang (1996)]:

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

where V, W are rescaled into $\bar{V}, \bar{W}$ separately such that each row of $\bar{V}$ and $\bar{W}$ has norm 1, and $a_{ij} = (\bar{V}\bar{W}^{-1})_{ij}$. It is noticed that $d(V,W)$ is invariant to permutation and scaling of the rows of V and W, is always between 0 and $(m-1)$, and is equal to zero if and only if V and W represent the same row components. For each experiment in the first group of simulation with 400 replications, we report in Table 2, the average Amari error, and the square root of mean square error $\sqrt{MSE}$ which is defined by

$$\sqrt{MSE} \quad = \quad \sqrt{\frac{1}{\#(repl)}\sum_{i=1}^{\#(repl)}(d_F^{(i)})^2/m}.$$

(Note: $d_F^{(i)}$ denotes the Frobenius error for the i-th replication and #(repl) is the number of replications). For the second group of simulation, we report the boxplots of the Amari errors (see Figure 1) and $\sqrt{MSE}$ in Table 3.

From the simulation results we can see that in most experiments the parametric methods (FastICA, JADE, ExtImax) behave worse than the nonparametric methods (PCFICA, Kgv, EFFICA), and that the EFFICA has both the smallest Amari errors and smallest Frobenius errors in most experiments, while the KernelICA-Kgv, which we conjecture can be efficient with appropriate regularization, is the best in cases of mixture Gaussians. As a tradeoff of their good statistical performance, the three nonparametric ICA algorithms require heavier computation than the three parametric ICA algorithms.

5. Proof of Theorem 1. The solution of the efficient score equation given by (12) can be viewed as a generalized M-estimator (GM-estimator). The existence/uniqueness, convergence and asymptotic linearity of GM-estimators have been studied in BKRW (the Iteration Theorem in Appendix A.10.2, page 517). Suppose that $M_n(\theta, P_n)$ is a functional of $\theta \in \Omega$ (a subset of a finite Euclidean space) and P_n, but is not necessarily linear with P_n. The subscript n in M_n allows the existence of a possible smoothing or sieve parameter dependent on n. The zero of $M_n(\theta, P_n)$ w.r.t θ is called a generalized M-estimator. Let $M(\theta, P) = M_\infty(\theta, P)$. We review the conditions of the Iteration Theorem.

[GM1]. $M(\theta_P, P) = 0$ and $\theta_P \in \Omega$ is the unique solution of $M(\theta, P) = 0$ in Ω.

pdfs	Fast	Jade	ExtImax	Pcf	Kgv	EFFICA
1	37	39	34	18	14	7
	(63)	(47)	(40)	(22)	(17)	(8)
2	36	36	24	35	33	29
	(163)	(48)	(43)	(43)	(39)	(37)
3	33	31	19	16	14	5
	(172)	(49)	(23)	(19)	(17)	(6)
4	39	50	41	60	61	60
	(70)	(61)	(79)	(71)	(72)	(78)
5	71	85	87	109	99	128
	(137)	(108)	(164)	(136)	(120)	(179)
6	42	43	32	18	15	7
	(133)	(59)	(40)	(21)	(17)	(8)
7	43	41	35	18	15	9
	(145)	(51)	(68)	(22)	(18)	(11)
8	36	44	35	21	19	17
	(70)	(68)	(45)	(25)	(22)	(20)
9	35	37	24	16	14	4
	(150)	(59)	(29)	(20)	(17)	(5)
10	46	59	39	44	30	47
	(148)	(73)	(47)	(52)	(35)	(74)
11	28	33	27	29	25	25
	(33)	(38)	(32)	(34)	(29)	(30)
12	50	49	44	44	39	78
	(130)	(58)	(55)	(52)	(47)	(187)
13	65	52	185	24	19	16
	(116)	(63)	(251)	(30)	(25)	(22)
14	35	45	91	20	14	11
	(77)	(63)	(133)	(25)	(17)	(14)
15	69	72	57	32	27	11
	(136)	(130)	(96)	(49)	(41)	(18)

Table 2: *Reporting the mean of the Amari errors without brackt and* $\sqrt{MSE}$ *inside bracket (multiplied by 1000) by using* $m = 2$ *sources and sample size* $n = 1000$*, where the distributions of two sources for the* k*th experiment (*$k = 1, \cdots, 15$ *marked in the first column) are indexed by [k] in Table 0. For* $k = 1, \cdots, 12$*, two sources have the same distribution. The boxed numbers represent the best performance according to each experiment*

case	Fast	Jade	ExtImax	PCF	Kgv	EFFICA
I	7	7	14	3	3	2
II	25	30	42	19	15	14
III	26	32	42	23	23	25

Table 3: *Reporting* $\sqrt{MSE}$ *(multiplied by 100) for ICA algorithms with the same simulations as in Figure 1*

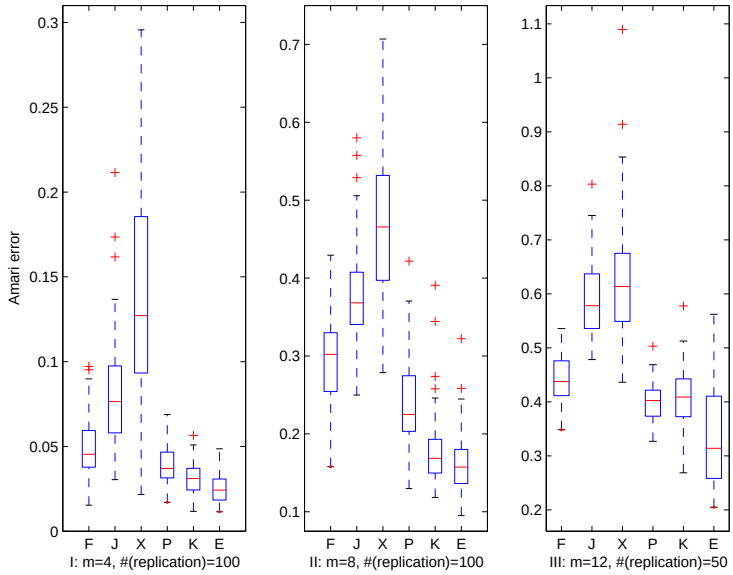

Figure 1: *Reporting the boxplots of Amari errors for ICA algorithms: case I (left) uses pdfs [0]-[3] to generate m=4 hidden sources, case II (middle) uses pdfs [0]-[7] to generate m=8 hidden sources, case III (right) uses [0]-[11] to generate m=12 hidden sources; The X-labels represent ICA algorithms: F-FastICA, J-JadeICA, X-Extended Infomax, P-PCFICA, K-Kgv, E-EFFICA; The sample sizes are 4000 for all the experiments and the replication times are 100, 100, 50 for I, II, III separately*

[GM2]. $M_n(\theta_P, P_n) = \int \psi_{\theta_P}(X)dP_n + o_p(n^{-1/2})$ for some $\psi_{\theta_P} \in L_2(P)$;

[GM3]. $M(\theta, P)$ is differentiable w.r.t θ in a neighbourhood of θ_P and $\frac{\partial M(\theta,P)}{\partial \theta}|_{\theta_P}$ is nonsingular.

For our efficient score equation $M_n(\theta, P_n) = \mathbf{e}_n(W)$ defined by (12), BKRW's condition [U] becomes :

[U]. $sup_{W\in\Omega_n}|\dot{\mathbf{e}}_n(W) - \dot{\mathbf{e}}(W_P)| = o_P(1)$.

THEOREM 2 [BKRW]. *Suppose (GM1),(GM2),(GM3) with $M_n(\theta, P_n) = \mathbf{e}_n(W)$ and (U) hold. If the starting point satisfies $P(|\hat{W}^{(0)} - W_P| < \varepsilon_n) \to 1$, then with probability converging to 1, $\mathbf{e}_n(W)$ in (12) has a unique root $\hat{W}^{(\infty)}$, which is also the limit of the sequence defined by (14) except $\int \hat{\mathbf{l}}^*\hat{\mathbf{l}}^{*T}(X; \hat{W}^{(j)})dP_n$ replaced by $-\dot{\mathbf{e}}_n(\hat{W}^{(j)})$, and $\hat{W}^{(\infty)}$ is asymptotically linear with the influence function $-[\dot{\mathbf{e}}(W_P)]^{-1}\mathbf{l}^*(.; W_P, \Phi_P)$.*

Theorem 2 is called the Iteration Theorem in BKRW. To prove the result of Theorem 1 w.r.t $\hat{W}^{(\infty)}$ defined by (14), we need the following claim:

[V]. $sup_{W\in\Omega_n}|\int \hat{\mathbf{l}}^*\hat{\mathbf{l}}^{*T}(X; W)dP_n - \int \mathbf{l}^*\mathbf{l}^{*T}(X; W_P, \Phi_{W_P})dP| = o_P(1)$.

Proof of Theorem 1. It is obvious that (GM1) holds under the conditions of Theorem 1 as it is the efficient score function. (GM2), (GM3) and (U) are verified by Proposition 1, 2 and 3 below, separately.

P, P_n	population, empirical law of X
W, W_k, W_{ij}	$m \times m$matrix, its kth row, (i,j)th element
W_P, W_{Pk}, W_{Pij}	unmixing matrix, its kth row, (i,j)th element
r_k	density function of S_k
$\phi_k = -r_k'/r_k$	density score function for S_k
$\Phi_P = (\phi_1, \cdots, \phi_m)^T$	function vectors
f_{W_k}	density function of $W_k X$ ($f_{W_{Pk}} \equiv r_k$)
$\phi_{W_k} = -f_{W_k}'/f_{W_k}$	score function of $W_k X$ ($\phi_{W_{Pk}} \equiv \phi_k$)
$\Phi_W = (\phi_{W_1}, \cdots, \phi_{W_m})^T$	function vector
$\mathbf{B}_n^{(k)} = (B_{n0}^{(k)}, \cdots, B_{nn_k}^{(k)})^T$	B-spline functions defined on $[\underline{b}_{nk}, \overline{b}_{nk}]$
$A_n(W_k) = \int_{Y=W_kX} \mathbf{B}_n^{(k)}(Y)\mathbf{B}_n^{(k)T}(Y)dP_n$	served in coefficients of $\hat{\phi}_{W_k}$ in Section 2.4
$D_n(W_k) = \int (\mathbf{B}_n^{(k)})'(W_kX)dP_n$	served in coefficients of $\hat{\phi}_{W_k}$
$\gamma_n(W_k) = A_n(W_k)^{-1}D_n(W_k)$	served as coefficients of $\hat{\phi}_{W_k}$
$A(W_k) = \int_{Y=W_kX} \mathbf{B}_n^{(k)}(Y)\mathbf{B}_n^{(k)T}(Y)dP$	served in coefficients of $\hat{\phi}_{W_k}$
$D(W_k) = \int (\mathbf{B}_n^{(k)})'(W_kX)dP$	served in coefficients of $\hat{\phi}_{W_k}$
$\gamma(W_k) = A(W_k)^{-1}D(W_k)$	served as coefficients of $\overline{\hat{\phi}}_{W_k}$ in (34)
$\mathcal{G}_n^{(k)} = \{a^T\mathbf{B}_n^{(k)} : a \in \mathcal{R}^{n_k}\}$	closed linear span of B spline functions
$\hat{\underline{\phi}}_{W_k} = \gamma_n(W_k)^T\mathbf{B}_n^{(k)}$	estimator of ϕ_{W_k} in $\mathcal{G}_n^{(k)}$ in Section 2.4
$\hat{\overline{\phi}}_{W_k} = \gamma(W_k)^T\mathbf{B}_n^{(k)}$	estimator of ϕ_{W_k} in $\mathcal{G}_n^{(k)}$, defined in (34)
$\phi_{k,n}, \phi_{W_k,n}$	truncation of ϕ_k, ϕ_{W_k} on $[\underline{b}_{nk}, \overline{b}_{nk}]$
$\mathbf{l}^*(X; W, \Phi)$	efficient score function of W, defined in (5)
$\mathbf{e}(W) = \int \mathbf{l}^*(X; W, \Phi_W)dP$	expectation
$\mathbf{e}_n(W) = \int \mathbf{l}^*(X; W, \hat{\Phi}_W)dP_n$	empirical expectation

Table 4: List of all notations used in the proof

Thus the conclusion of the above Iteration Theorem applies here. By Prosition 4, Condition (V) holds. Further by Proposition 2,

$$\dot{\mathbf{e}}(W_P) = -E[\mathbf{l}^*\mathbf{l}^{*T}(X; W_P, \Phi_P)],$$

thus we have

$$\sup_{W\in\Omega_n} |\dot{\mathbf{e}}_n(W) + \int \hat{\mathbf{l}}^*\hat{\mathbf{l}}^{*T}(x; W)dP_n| = o_P(1).$$

Then by following the contraction arguments of BKRW (page 317-319), the iteration sequence given in (14) has the same limit as that replacing $\int \hat{\mathbf{l}}^*\hat{\mathbf{l}}^{*T}(X; \hat{W}^{(j)})dP_n$ by $-\dot{\mathbf{e}}_n(\hat{W}^{(j)})$ with probability converging to 1. ■

6. Proposition 1-4. For convenience we list all the notations used in the following proofs in Table 4, for $k \in \{1, \cdots, m\}$, $W \in \Omega_n$. It is noted that all the lemmas used in this section are provided and proved in the Appendix. For simplicity of notation, we often write δ_{nk} as δ_n below.

PROPOSITION 1. *Under the conditions of Theorem 1,* we have

$$\mathbf{e}_n(W_P) = \int \mathbf{l}^*(X; W_P, \Phi_P)dP_n + o_P(n^{-1/2}).$$

PROOF. Recall the definition of $\mathbf{e}_n(W)$ given by (12) and $\mathbf{l}^*(\mathbf{x}; W, \Phi)$ given by (5)-(9). It is sufficient to show that for $1 \le i \ne j \le m$, $\hat{\alpha_i} - \alpha_i = o_P(1)$, $\hat{\beta}_i - \beta_i = o_P(1)$, where (α_i, β_i) and $(\hat{\alpha_i}, \hat{\beta}_i)$ are defined in (8) and (11) separately, and

$$\int \hat{\phi}_{W_{Pi}}(S_i) S_j dP_n \quad = \quad \int \phi_{W_{Pi}}(S_i) S_j dP_n + o_P(n^{-1/2}), \tag{18}$$

where $S_i = W_{Pi}X,\ S_j = W_{Pj}X$.

The first two are not hard to be verified by the Law of Large Numbers and Lemma 10. Here we just show the last argument (18). Observe that

$$\begin{aligned} |\int \hat{\phi}_{W_{Pi}}(S_i) S_j dP_n - \int \phi_{W_{Pi}}(S_i) S_j dP_n| &= |\int [\hat{\phi}_{W_{Pi}}(S_i) - \overline{\hat{\phi}}_{W_{Pi}}(S_i)] S_j dP_n| \\ &\quad + |\int [\overline{\hat{\phi}}_{W_{Pi}}(S_i) - \phi_{i,n}(S_i)] S_j dP_n| \\ &\quad + |\int (\phi_i(S_i) - \phi_{i,n}(S_i)) S_j dP_n| \\ &= [1] + [2] + [3]. \end{aligned}$$

In the following, we show that all of [1], [2] and [3] are $o_P(n^{-1/2})$.

First by Lemma 4 and Lemma 6,

$$\begin{aligned} [1] &= |\int (\gamma_n(W_{Pi}) - \gamma(W_{Pi}))^T \mathbf{B}_n^{(i)}(S_i) S_j dP_n| \\ &\le ||\gamma_n(W_{Pi}) - \gamma(W_{Pi})||_2 || \int \mathbf{B}_n^{(i)}(S_i) S_j dP_n||_2 \\ &= o_P(1) O_P(n^{-1/2}). \end{aligned}$$

Further, $E([2])^2 = \frac{1}{n} E(\overline{\hat{\phi}}_{W_{Pi}}(S_i) - \phi_{W_{Pi},n}(S_i))^2 E(S_j^2)$. By Lemma 9, $|\overline{\hat{\phi}}_{W_{Pi}} - \phi_{W_{Pi},n}|_\infty \le c\delta_n^2 |\phi'''_{W_{Pi},n}|_\infty$, thus by Condition C6

$$[2] \quad = \quad n^{-1/2} \delta_n^2 |\phi'''_{W_{Pi},n}|_\infty O_P(1) = o_P(n^{-1/2}).$$

For [3], since $P(S_i \notin [\underline{b}_{ni}, \overline{b}_{ni}]) \to 0$, we have

$$E([3])^2 \quad = \quad \frac{1}{n} E(\phi_i(S_i)^2 I(S_i \notin [\underline{b}_{ni}, \overline{b}_{ni}])) E(S_j^2) = o(\frac{1}{n}).$$

So $[3] = o_P(n^{-1/2})$. ■

PROPOSITION 2. *Under Condition C1, C2, C4 in Theorem 1,* $\mathbf{e}(W)$ *is differential w.r.t* W *in a neighbourhood of* W_P *and* $\dot{\mathbf{e}}(W_P) = -E[\mathbf{l}^*\mathbf{l}^{*T}(X; W_P, \Phi_P)]$*, nonsingular.*

PROOF. Let $T_w(\cdot) = \frac{\partial}{\partial w}\phi_w(\cdot)$, for any nonzero $w \in R^m$. By (32) after exchanging the order of derivative and integration we have $E[T_w(wX)] = 0$. Then by (6) we have

$$E[\frac{\partial}{\partial W}\mathbf{l}^*(X; W, \Phi_W)]|_{W_P} \quad = \quad E[\frac{\partial}{\partial W}\mathbf{l}^*(X; W, \Phi_P)]|_{W_P}.$$

Since the left hand side (LHS) of the above is $\dot{\mathbf{e}}(W_P)$, hence by Lemma 11 the right hand side (RHS) is equal to

$$\dot{\mathbf{e}}(W_P) \quad = \quad -E[\mathbf{l}^*\mathbf{l}^{*T}(X; W_P, \Phi_P)]. \tag{19}$$

Notice that the elements of $\mathbf{l}^*(.; W_P, \Phi_P)$ are linearly independent, $\dot{\mathbf{e}}(W_P)$ must be nonsingular. ■

PROPOSITION 3. *Under the conditions of Theorem 1, for* $k = 1, \cdots, m$*, we have*

$$sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)X_k dP_n(X) - \int \phi_{Pi}(W_{Pi}X)X_k dP| \quad = \quad o_P(1), \tag{20}$$

and

$$sup_{\Omega_n}|\int \frac{\partial}{\partial W_i}[\hat{\phi}_{W_i}(W_iX)]W_jX dP_n(X) - \int \frac{\partial}{\partial W_i}[\phi_{Pi}(W_iX)]_{W_{Pi}}W_{Pj}X dP| \quad = \quad o_P(1). \tag{21}$$

Then, Condition [U] holds.

PROOF. Notice that (dropping superscript (i) in $B_n^{(i)}$ henceforth)

$$\begin{aligned} ||\int \mathbf{B}_n^{(i)}(W_iX)X_k dP_n||_2^2 \quad &= \quad \sum_{l=1}^{n_i}(\int B_{nl}(W_iX)X_k dP_n)^2 \\ &\leq \quad \int \sum_{l=1}^{n_i} B_{nl}^2(W_iX)dP_n \int |X_k|^2 dP_n \\ &\leq \quad \int |X_k|^2 dP_n, \text{ (by Property III, page 23).} \end{aligned}$$

Then

$$sup_{\Omega_n}||\int \mathbf{B}_n(W_iX)X_k dP_n||_2 \quad = \quad O_P(1). \tag{22}$$

And by Lemma 4, $sup_{\Omega_n}||\gamma_n(W_k) - \gamma(W_k)||_2 = o_P(1)$, so

$$\begin{aligned} & sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)X_k dP_n(X) - \int \overline{\hat{\phi}}_{W_i}(W_iX)X_k dP_n| \\ = \quad & sup_{\Omega_n}|(\gamma_n(W_k) - \gamma(W_k))^T \int \mathbf{B}_n(W_iX)X_k dP_n| \\ \leq \quad & sup_{\Omega_n}||\gamma_n(W_k) - \gamma(W_k)||_2 sup_{\Omega_n}||\int \mathbf{B}_n(W_iX)X_k dP_n||_2 \\ = \quad & o_P(1)O_P(1). \end{aligned} \tag{23}$$

Further, by Lemma 9, $sup_{\Omega_n}|\overline{\hat{\phi}}_{W_i}(W_iX) - \phi_{W_i,n}|_\infty \leq sup_{\Omega_n} c|\phi'''_{W_i,n}|_\infty \delta_n^2$, then

$$\begin{aligned} & sup_{\Omega_n}|\int \overline{\hat{\phi}}_{W_i}(W_iX)X_k dP_n(X) - \int \phi_{W_i,n}(W_iX)X_k dP_n| \\ \leq \quad & sup_{\Omega_n}|\phi'''_{W_i,n}|_\infty \delta_n^2 \int |X_k| P_n \\ = \quad & o_P(1), \text{ (by Condition C6).} \end{aligned} \tag{24}$$

And by Condition C4, ULLN holds for $\{\phi_{W_i}(W_iX)X_k : W \in \Omega_n\}$, and by Lemma 1 $sup_{\Omega_n} P(W_iX \notin [\underline{b}_{ni}, \overline{b}_{ni}]) = o(1)$, then

$$\begin{aligned} sup_{\Omega_n}|\int (\phi_{W_i} - \phi_{W_i,n})(W_iX)X_k dP_n| \quad &= \quad sup_{\Omega_n}|\int \phi_{W_i}(W_iX)X_k I(W_iX \notin [\underline{b}_{ni}, \overline{b}_{ni}])dP_n| \\ &= \quad o_P(1). \end{aligned} \tag{25}$$

From (23)-(25), we get

$$sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)X_k dP_n(X) - \int \phi_{W_i}(W_iX)X_k dP_n(X)| \quad = \quad o_P(1). \tag{26}$$

Now by Condition C4,

$$sup_{\Omega_n}|\int \phi_{W_i}(W_iX)X_k d(P_n - P)| \quad = \quad o_P(1); \tag{27}$$

And by continuity,

$$sup_{\Omega_n}|\int \phi_{W_i}(W_iX)X_kdP - \int \phi_{W_{Pi}}(W_{Pi}X)X_kdP| \quad = \quad o(1). \tag{28}$$

Then (20) follows from (26)-(28).

In the following, we prove (21).

Notice that

$$\frac{\partial}{\partial W_i}\hat{\phi}_{W_i}(W_iX) \quad = \quad \frac{\partial}{\partial W_i}[\gamma_n^T(W_i)]\mathbf{B}_n^{(i)}(W_iX) + \hat{\phi}'_{W_i}(W_iX)X. \tag{29}$$

It is enough to show that the following [4]&[5] hold:

[4]. $sup_{\Omega_n}|\int \hat{\phi}'_{W_i}(W_iX)X_kW_jXdP_n(X) - \int \phi'_{Pi}(W_{Pi}X)X_kW_{Pj}XdP| = o_P(1)$;

[5]. $sup_{\Omega_n}|\int \frac{\partial}{\partial W_i}[\gamma_n^T(W_i)]\mathbf{B}_n^{(i)}(W_iX)W_jXdP_n(X)| = o_P(1)$.

Similar to (20), the uniform convergence of [4] can be verified by using Condition C4, C6, C7 and Lemma 1, 4, 9. Thus we only prove [5] in the following.

Notice that in [5],

$$(LHS)_k \quad \leq \quad sup_{\Omega_n}||(\frac{\partial}{\partial W_i}\gamma_n(W_i))_k||_2 sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXdP_n||_2. \tag{30}$$

By Lemma 7, $sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXdP_n||_2 = O_p(\varepsilon_n\delta_n^{-1}n_i^{1/2})$. Thus it is enough to show that

$$sup_{\Omega_n}||(\frac{\partial}{\partial W_i}\gamma_n(W_i))_k||_2\varepsilon_n\delta_n^{-1}n_i^{1/2} \quad = \quad o_P(1). \tag{31}$$

By taking partial derivatives,

$$\frac{\partial}{\partial W_{ik}}\gamma_n(W_i) \quad = \quad \frac{\partial}{\partial W_{ik}}A_n^{-1}(W_i)D_n(W_i) + A_n^{-1}(W_i)\frac{\partial}{\partial W_{ik}}D_n(W_i),$$

and

$$\frac{\partial}{\partial W_{ik}}A_n^{-1}(W_i) \quad = \quad -A_n^{-1}\frac{\partial}{\partial W_{ik}}A_n(W_i)A_n^{-1}.$$

Then

$$\frac{\partial}{\partial W_{ik}}\gamma_n(W_i) \quad = \quad -A_n^{-1}\frac{\partial}{\partial W_{ik}}A_n(W_i)\gamma_n(W_i) + A_n^{-1}(W_i)\frac{\partial}{\partial W_{ik}}D_n(W_i).$$

Now by Lemma 2-5, we get

$$\begin{aligned} sup_{\Omega_n}||\frac{\partial}{\partial W_{ik}}\gamma_n(W_i)||_2 &\leq sup_{\Omega_n}||A_n^{-1}||_2(||\frac{\partial}{\partial W_{ik}}A_n(W_i)||_2||\gamma_n(W_i)||_2 + ||\frac{\partial}{\partial W_{ik}}D_n(W_i)||_2) \\ &= O_p(\delta_n^{-2})\{O_p(\delta_n^{-\frac{1}{2}})O_p(\delta_n^{-1}\sqrt{n_i}) + \delta_n^{-2}O_p(1)\} \\ &= \delta_n^{-\frac{7}{2}}\sqrt{n_i}O_P(1). \end{aligned}$$

Provided that $\varepsilon_n\delta_n^{-\frac{9}{2}}n_i = o(1)$ implied by Condition C7, (31) holds. Thus we have in [5] $(LHS)_k = o_P(1)$ for $k = 1, \cdots, m$. ■

PROPOSITION 4. *Under the conditions of Theorem 1, Condition [V] holds, i.e.,*

$$sup_{W\in\Omega_n}|\int \hat{\mathbf{l}}^*\hat{\mathbf{l}}^{*T}(X;W)dP_n - \int \mathbf{l}^*\mathbf{l}^{*T}(X;W_P,\Phi_{W_P})dP| \quad = \quad o_P(1).$$

PROOF. By checking the elements of $\hat{\mathbf{l}}^*\hat{\mathbf{l}}^{*T}(\mathbf{x};W)$, it is enough to show that for $1 \leq i,j,k,l \leq m$,

$$sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)\hat{\phi}_{W_j}(W_jX)X_kX_ldP_n(X) - \int \phi_{Pi}(W_{Pi}X)\phi_{Pj}(W_{Pj}X)X_kX_ldP| = o_P(1),$$

$$sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)X_kX_ldP_n(X) - \int \phi_{Pi}(W_{Pi}X)X_kX_ldP| = o_P(1),$$

and

$$sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)X_kk(W_jX)dP_n(X) - \int \phi_{Pi}(W_{Pi}X)X_kk(W_{Pj}X)dP| = o_P(1).$$

Each of these can be verified by using Lemma 1, 4, 9 and Condition C4, C7 with the similar arguments in proving (20). ■

7. Conclusion. In this paper, we put the classical ICA model under the framework of semiparametric models and obtained an asymptotically efficient estimator for the unmixing matrix, by solving an approximate efficient score equation. The main difference between this method and popular parametric ICA methods is that we estimate the density score functions of hidden sources adaptively. A variety of simulations have illustrated statistical efficiency of this estimator in comparison with state-of-the-art ICA algorithms.

Acknowledgements. This paper was presented at the Joint Statistical Meeting, in San Francisco, USA, August 2003. We would like to thank Professor Charles Stone for helpful technical discussions and thank Sabrina Soracco for helpful comments on editing. We also wish to thank the Editors and four referees for their remarks that helped to improve the paper significantly.

APPENDIX

Some useful formulae. Let $v = wW_P^{-1}$. Then $wX = vS$. If $v_k \neq 0$ for some $k \in \{1, \cdots, m\}$, then

$$\begin{aligned} f_w(t) &= \int_{R^{m-1}} \frac{1}{v_k} r_k\left(\frac{t - \sum_{j\neq k} v_j s_j}{v_k}\right) \prod_{j\neq k} r_j(s_j) ds_j \\ &= E\left[\frac{1}{v_k} r_k\left(\frac{t - \sum_{j\neq k} v_j S_j}{v_k}\right)\right]. \end{aligned}$$

Since $f_w(t)$ is a marginal density function of $(vS, S_j : 1 \leq j \neq k \leq m)$, by a standard formula [see, e.g., Bickel and Doksum (2001)]

$$\begin{aligned} \phi_w(t) &= -\frac{1}{v_k} E\left[\frac{r_k'}{r_k}\left(\frac{t - \sum_{j\neq k} v_j S_j}{v_k}\right) | vS = t\right] \\ &= \frac{1}{v_k} E[\phi_k(S_k) | vS = t], \end{aligned} \tag{32}$$

and further calculation gives

$$\frac{\partial}{\partial t}\phi_w(t) = \phi_w^2(t) - \frac{1}{v_k^2} E\left[\frac{r_k''}{r_k}\left(\frac{t - \sum_{j\neq k} v_j S_j}{v_k}\right) | vS = t\right]. \tag{33}$$

Some properties of cubic B-splines. Let $\xi_1 < \xi_2 < \cdots < \xi_N$ be fixed points. The first order B-spline basis functions based on these knots can be expressed as $B_i^1(x) = I(x \in [\xi_i, \xi_{i+1}))$, $i = 1, \cdots, N-1$ and the kth order B-spline basis functions can be obtained recursively ($k \geq 2$) by

$$B_i^k(x) = \frac{x - \xi_i}{\xi_{i+k-1} - \xi_i} B_i^{k-1}(x) + \frac{\xi_{i+k} - x}{\xi_{i+k} - \xi_{i+1}} B_{i+1}^{k-1}(x),$$

for $i = 1, \cdots, N-k$. It is well known that $B_i^k(x)$ is differentiable w.r.t. x up to order $k-2$. The first order

derivative can be expressed as

$$\frac{d}{dx}B_i^k(x) = \frac{k-1}{\xi_{i+k-1}-\xi_i}B_i^{k-1}(x) - \frac{k-1}{\xi_{i+k}-\xi_{i+1}}B_{i+1}^{k-1}(x).$$

We use the 4th order, so-called cubic B-splines $\{B_i^4 : 1 \le i \le N-4\}$ with equally spaced knots, i.e., $\xi_{i+1}-\xi_i = \delta$ $(i = 1, \cdots, N-1)$ for some algorithm-determined δ. For simplicity, the superscript in B_i^4 is omitted below. The following properties of cubic B-splines will be frequently used in proving the lemmas below (see de Boor (1978) for the details).

I). $0 \le B_i(x) < 1$, $B_i(x)B_j(x) = 0$ if $|i-j| > 3$.

II). $|\frac{d}{dx}B_i(x)| < \delta^{-1}$, $|\frac{d^2}{dx^2}B_i(x)| < 2\delta^{-2}$.

III). $\sum_{i=1}^N [B_i(x)]^2 < 1$, $\sum_{i=1}^N [\frac{d^2}{dx^2}B_i(x)]^2 < 6\delta^{-4}$.

Supporting lemmas for Proposition 1-4. In this subsection, we prove all the lemmas used in the proof of Propositon 1-4. Recall that for each ϕ_k $(k = 1, \cdots, m)$, we have an interval $[\underline{b}_{nk}, \overline{b}_{nk}]$ and n_k cubic B-spline basis functions defined on it using equally spaced knots on it, say $\mathbf{B}_n^{(k)} = (B_{n1}^{(k)}, \cdots, B_{nn_k}^{(k)})^T$ as in Section 2.4. Thus we have constructed a sequence of sieves $\mathcal{G}_n^{(k)}$ using $\mathbf{B}_n^{(k)}$ as basis functions. For any $W \in \Omega_n$, we have a class of estimates $\hat{\phi}_{W_k} \in \mathcal{G}_n^{(k)}$ for ϕ_{W_k} as defined in Section 2.4.

Let $\Omega_n^{(k)} = \{W_k : W \in \Omega_n\}$ for $k = 1, \cdots, m$. We also need an intermediate approximation function $\overline{\hat{\phi}}_{W_k} \in \mathcal{G}_n^{(k)}$ as follows. As a little confusion, for $w \in \Omega_n^{(k)}$,

$$\overline{\hat{\phi}}_w = \gamma(w)^T \mathbf{B}_n^{(k)}, \tag{34}$$

where $\gamma(w) = A(w)^{-1}D(w)$ with $A(w) = \int \mathbf{B}_n^{(k)}(wX)[\mathbf{B}_n^{(k)}(wX)]^T dP$ and $D(w) = \int [\mathbf{B}_n^{(k)}]'(wX)dP$. Note that the subscript w of $\overline{\hat{\phi}}_w$ should always associate with some $\Omega_n^{(k)}$ for $k \in \{1, \cdots, m\}$, similarly for $\hat{\phi}_w$.

In the following c denotes a constant (only dependent on the population law P), but its exact value may vary in different places even in a line without clarification. For a column vector $\mathbf{x} \in \mathcal{R}^m$, $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$. For an $m \times m$ real matrix A, $||A||_1 = max_{1\le i\le m}||A_i||_2$, $||A||_2 = max_{x\in R^m, |x|=1}|Ax|$, $||A||_F = \sqrt{tr(A^TA)}$. Then $||A||_2 \le ||A||_1$.

The following Lemma 1-10 hold under the conditions of Theorem 1. Jin (1992) had similar results as Lemma 2-4 and Lemma 8-10 about the B-spline approximation but under generally different settings.

LEMMA 1. $sup_{w\in\Omega_n^{(k)}}|f_w|_\infty < \infty$, $sup_{w\in\Omega_n^{(k)}}|f_w'|_\infty < \infty$, $sup_{w\in\Omega_n^{(k)}} min_{t\in[\underline{b}_{nk},\overline{b}_{nk}]} f_w(t) \ge c\delta_n$, and

$sup_{w\in\Omega_n^{(k)}} P(wX \notin [\underline{b}_{ni}, \overline{b}_{ni}]) = o(1)$.

PROOF. Remember that $min_{t\in[\underline{b}_{nk},\overline{b}_{nk}]} r_k(t) \geq c\delta_n$. For any $w \in \Omega_n^{(k)}$, $||w - W_{Pk}||_2 \leq \varepsilon_n$. Let $v = wW_P^{-1}$, then $|v_j| \to 0$ for $1 \leq j \neq k \leq m$ and $|v_k - 1| \to 0$ as $n \to \infty$. Fix a $t \in [\underline{b}_{nk}, \overline{b}_{nk}]$.
Since $f_w(t) = E[\frac{1}{v_k} r_k(\frac{t - \sum_{j\neq k} v_j S_j}{v_k})]$, consider the right hand side as a function (say h) of v. By the first order Taylor expansion,

$$|f_w(t) - r_k(t)| \quad \leq \quad \varepsilon_n ||W_P^{-1}||_2 \{\sum_{j=1}^{m} max_{w\in\Omega_n^{(k)}} |\frac{\partial}{\partial v_j} h(v)|\} \leq c\varepsilon_n = o(\delta_n),$$

where by direct calculation and using C3-C4, $|\frac{\partial}{\partial v_j} h(v)|$ is uniformly bounded with $w \in \Omega_n^{(k)}$. Thus

$$sup_{w\in\Omega_n^{(k)}} min_{t\in[\underline{b}_{nk},\overline{b}_{nk}]} f_w(t) \geq c\delta_n \text{ and } sup_{w\in\Omega_n^{(k)}} |f_w|_\infty < \infty.$$

Further, $sup_{w\in\Omega_n^{(k)}} |f'_w|_\infty < \infty$ follows from $|r'_k|_\infty < \infty$.

Finally,

$$\begin{aligned} P(wX \in [\underline{b}_{ni}, \overline{b}_{ni}]) &= \int_{[\underline{b}_{nk},\overline{b}_{nk}]} f_w(t)dt \\ &\geq \int_{[\underline{b}_{nk},\overline{b}_{nk}]} (r_k(t) - c\varepsilon_n)dt \\ &= P(S_k \in [\underline{b}_{nk}, \overline{b}_{nk}]) - c\varepsilon_n(\overline{b}_{nk} - \underline{b}_{nk}). \end{aligned}$$

Since $\varepsilon_n(\overline{b}_{nk} - \underline{b}_{nk}) = o(1)$ and $P(S_k \in [\underline{b}_{nk}, \overline{b}_{nk}]) \uparrow 1$, thus

$$inf_{w\in\Omega_n^{(k)}} P(wX \quad \in \quad [\underline{b}_{ni}, \overline{b}_{ni}]) \geq P(S_k \in [\underline{b}_{nk}, \overline{b}_{nk}]) - c\varepsilon_n(\overline{b}_{nk} - \underline{b}_{nk}) \to 1.$$

■

Recall the definition of $\hat{\phi}_{W_k}$, $\gamma_n(W_k)$, $A_n(W_k)$ and $D_n(W_k)$ in Section 2.4 and $\gamma(w) = [A(w)]^{-1}D(w)$ in (34).

LEMMA 2. $sup_{w\in\Omega_n^{(k)}} ||D(w)||_2 \leq c\sqrt{n_k}\delta_n$; $c\delta_n^2 \leq eig(A(w)) \leq c\delta_n$ for $w \in \Omega_n^{(k)}$.

PROOF. By taking the derivative of the cubic B-splines, $(B_{ni}^{(k)})'(t) = \delta_n^{-1}(B_{ni}^3(t) - B_{n,i+1}^3(t))$, where B_{ni}^3 is the third-order B-splines defined on the same knots, $(i = 1, \cdots, n_k)$, then

$$|D_i(w)| \quad = \quad \delta_n^{-1} |\int (B_{n,i}^3(t) - B_{n,i+1}^3(t)) f_w(t)dt|$$

$$
\begin{aligned}
&= \delta_n^{-1} \left| \int B_{n,i}^3(t)(f_w(t) - f_w(t+\delta_n))dt \right| \\
&\leq \int B_{n,i}^3(t)dt |f'_w|_\infty < 3|f'_w|_\infty \delta_n.
\end{aligned}
$$

So the first result holds by using Lemma 1. By Lemma 5.1 in Jin (1992), $c\delta_n min_{t\in[\underline{b}_{nk},\overline{b}_{nk}]} f_w(t) \leq eig(A(w)) \leq c\delta_n max_{t\in[\underline{b}_{nk},\overline{b}_{nk}]} f_w(t)$, thus $c\delta_n^2 \leq eig(A(w)) \leq c\delta_n$. ■

LEMMA 3. $sup_{w\in\Omega_n^{(k)}} ||D_n(w) - D(w)||_2 = \sqrt{\frac{n_k \log n_k}{n\delta_n}} O_P(1)$; $sup_{w\in\Omega_n^{(k)}} ||A_n(w) - A(w)||_2 = \sqrt{\frac{\delta_n \log n_k}{n}} O_P(1)$.

PROOF.

$$
\begin{aligned}
P(sup_{w\in\Omega_n^{(k)}} ||D_n(w) - D(w)||_2 \geq t) &= Pr(sup_{W\in\Omega_n^{(k)}} || \int \mathbf{B}'_n(wX)d(P_n - P)||_2 \geq t) \\
&\leq \sum_{i=1}^{n_k} Pr(sup_{w\in\Omega_n^{(k)}} | \int B'_{n,i}(wX)d(P_n - P)| \geq \frac{t}{\sqrt{n_k}}).
\end{aligned}
$$

For a fixed pair (i,k), let $\mathcal{F}_n = \{g_w(\mathbf{x}) = B'_{n,i}(w\mathbf{x}) : w \in \Omega_n^{(k)}\}$. Then $\sup_{g_w\in\mathcal{F}_n} ||\frac{\partial}{\partial w} g_w(\mathbf{x})||_2 \leq 2\delta_n^{-2}||\mathbf{x}||_2$. By definition and the Euclidean ball theory (see for example Definition 2.2 and Lemma 2.5 of van de Geer (2000)), the *bracketing entropy* of $\mathcal{F}_n$ is bounded by for $0 < u < c\varepsilon_n\delta_n^{-2}$,

$$
\mathcal{H}_B(u, \mathcal{F}_n, P) \leq m \log(c\varepsilon_n\delta_n^{-2}/u).
$$

Further by the property I of cubic B-splines, $\sup_{g_w\in\mathcal{F}_n} |g_w|_\infty \leq \delta_{nk}^{-1}$, and by Lemma 1, $\sup_{g_w\in\mathcal{F}_n} \int |g_w(X)|^2 dP \leq 4\delta_n^{-1} \sup_{w\in\Omega_n^{(k)}} |f_w|_\infty$. Then by Theorem 5.11 of van de Geer (2000, page 75), we have for $c\max(\delta_n^{-1/2}, \varepsilon_n\delta_n^{-2}) \leq a \leq c\sqrt{n}$,

$$
P(sup_{w\in\Omega}\sqrt{n} | \int B'_{n,i}(wX)d(P_n - P)| \geq a) \leq \exp(-ca^2\delta_n).
$$

Notice that $\varepsilon_n \ll \delta_{nk}^{3/2}$ by Condition C7, so

$$
P(sup_{w\in\Omega_n^{(k)}} ||D_n(w) - D(w)||_2 \geq t) \leq n_k \exp(-ct^2 n\delta_n/n_k).
$$

Thus

$$
sup_{\Omega_n^{(k)}} ||D_n - D||_2 = O_p(\sqrt{\frac{n_k \log n_k}{n\delta_n}}).
$$

Similarly by using the properties I-II of cubic B-splines and the bracketing entropy tool we get

$$sup_{\Omega_n^{(k)}}||A_n - A||_2 \quad \leq \quad sup_{\Omega_n^{(k)}}||A_n - A||_1 = O_p(\sqrt{\delta_n \log n_k / n}).$$

■

LEMMA 4. $sup_{w\in\Omega_n^{(k)}}||D_n(w)||_2 = O_p(\delta_n\sqrt{n_k})$, $sup_{w\in\Omega_n^{(k)}}||\gamma_n(w)||_2 = O_p(\sqrt{n_k}/\delta_n)$, and $sup_{\Omega_n^{(k)}}||\gamma_n(w) - \gamma(w)||_2 = o_P(1)$.

PROOF. The first result directly follows from Lemma 2 and 3. The following proves the second and third results.

Since $A_n^{-1} = (A + A_n - A)^{-1} = A^{-1}(I - (A_n - A)A^{-1})^{-1}$, and by Lemma 2 and 3

$$sup_{w\in\Omega_n^{(k)}}||A_n - A||_2||A^{-1}||_2 \quad = \quad o_p(1),$$

then

$$sup_{w\in\Omega_n^{(k)}}||A_n^{-1}||_2 \quad \leq sup_{w\in\Omega_n^{(k)}} \quad ||A^{-1}||_2(1 - ||A_n - A||_2||A^{-1}||_2)^{-1} = \delta_n^{-2}O_P(1).$$

(Here we use the inequality of matrix norm $||(I + A)||_2 \leq (1 - ||A||_2)^{-1}$for any square matrix A with $||A||_2 < 1$, where I is the identity matrix.) Thus $sup_{w\in\Omega_n^{(k)}}|A_n^{-1}(w)D_n(w)| = O_p(\sqrt{n_k}/\delta_n)$.

For the last one, by Lemma 2 and 3, we have

$$\begin{aligned}
& sup_{w\in\Omega_n^{(k)}}||\gamma_n(w) - \gamma(w)||_2 \\
= \quad & sup_{w\in\Omega_n^{(k)}}||A^{-1}(D_n - D) - A_n^{-1}(A_n - A)A^{-1}D_n||_2 \\
\leq \quad & sup_{\Omega_n^{(k)}}\{||A^{-1}||_2||D_n - D||_2\} + sup_{\Omega_n^{(k)}}\{||A_n - A||_2||D_n||_2||A^{-1}||_2^2\}(1 + o_p(1)) \\
= \quad & O_p(\delta_n^{-2})O_p(\sqrt{\frac{n_k \log n_k}{n\delta_n}}) + O_p(\sqrt{\frac{\delta_n \log n_k}{n}})O_p(\delta_n\sqrt{n_k})O_p(\delta_n^{-4}) \\
= \quad & O_p(\delta_n^{-\frac{5}{2}}\sqrt{\frac{n_k \log n_k}{n}}) = o_P(1). \text{ (by Condition C7)}
\end{aligned}$$

■

LEMMA 5. $sup_{\Omega_n^{(i)}}||\frac{\partial}{\partial w_k}A_n(w)||_2 = O_P(\delta_n^{-\frac{1}{2}})$, $sup_{\Omega_n^{(i)}}||\frac{\partial}{\partial w_k}D_n(w)||_2 = O_P(\delta_n^{-2})$ for $i, k = 1, \cdots, m$.

PROOF. First notice that (dropping $^{(i)}$ in $\mathbf{B}_n^{(i)}$)

$$\frac{\partial}{\partial w_k} A_n(w) \quad = \quad \int (\mathbf{B}_n \mathbf{B}_n'^T + \mathbf{B}_n' \mathbf{B}_n^T)(wX) X_k dP_n.$$

By the Cauchy-Schwartz inequality,

$$|\int [\mathbf{B}_n \mathbf{B}_n'^T + \mathbf{B}_n' \mathbf{B}_n^T]_{jl}(wX) X_k dP_n| \quad \leq \quad \sqrt{\int (B_{nj} B'_{nk} + B'_{nj} B_{nk})^2 (wX) dP_n} \sqrt{\int X_k^2 dP_n}.$$

Following the proof in Lemma 3 by using the bracketing entropy, we have

$$sup_{0 \leq j,l \leq n_i, |j-l| \leq 3} sup_{w \in \Omega_n^{(i)}} \int (B_{nj} B'_{nl} + B'_{nj} B_{nl})^2 (wX) d(P_n - P) \quad = \quad o_p(1).$$

Further from Lemma 1 $\sup_{w \in \Omega_n^{(k)}} |f_w|$ is bounded, after algebraic expanion we have

$$sup_{w \in \Omega_n^{(i)}} \int (B_{nj} B'_{nl} + B'_{nj} B_{nl})^2 (wX) dP \quad \leq \quad c \delta_n^{-1}.$$

Thus $|[\frac{\partial}{\partial w_k} A_n(w)]_{jl}| \leq c/\sqrt{\delta_{nk}}$. By the property I of cubic B-splines, $[\mathbf{B}_n \mathbf{B}_n'^T]_{jl} \equiv 0$ for $|j-l| > 3$, thus each row of $\frac{\partial}{\partial w_k} A_n(w)$ has at most 7 nonzero elements. So

$$sup_{\Omega_n^{(i)}} || \frac{\partial}{\partial w_k} A_n(w) ||_2 \quad \leq \quad sup_{\Omega_n^{(i)}} || \frac{\partial}{\partial w_k} A_n(w) ||_1 = O_p(\delta_n^{-\frac{1}{2}}).$$

For the second result, since $\frac{\partial}{\partial w_k} D_n(w) = \int \mathbf{B}_n''(wX) X_k dP_n$, we have

$$\begin{aligned} || \frac{\partial}{\partial w_k} D_n(w) ||_2 \quad &\leq \quad \sqrt{\int |X_k|^2 dP_n} \sqrt{\int \sum_{l=1}^{n_k} (B''_{nl})^2 (wX) dP_n} \\ &\leq \quad \sqrt{\int |X_k|^2 dP_n} \sqrt{\int 6 \delta_n^{-4} dP_n} \\ &= \quad O_P(\delta_n^{-2}). \end{aligned}$$

■

LEMMA 6. $|| \int \mathbf{B}_n^{(i)}(S_i) S_j dP_n ||_2 = O_P(n^{-1/2})$, where $S_i = W_{Pi} X$, $1 \leq i \neq j \leq m$.

PROOF. (dropping $^{(i)}$ in $\mathbf{B}_n^{(i)}, B_{nk}^{(i)}$)

$$\begin{aligned} E(||\int \mathbf{B}_n(S_i)S_j dP_n||_2^2) &= E(\sum_{k=1}^{n_i}(\int B_{nk}(S_i)S_j dP_n)^2) \\ &= \frac{1}{n}E(\sum_{k=1}^{n_i} B_{nk}(S_i)^2 S_j^2) \\ &\leq \frac{4}{n}E(S_j^2). \end{aligned}$$

■

LEMMA 7. $sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXdP_n||_2 = O_p(\varepsilon_n\delta_n^{-1}\sqrt{n_i})$, for $1 \leq i \neq j \leq m$.

PROOF. First

$$P(\sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXd(P_n-P)||_2 > t) \leq \sum_{k=1}^{n_i} P(\sup_{\Omega_n}|\int B_{nk}^{(i)}(W_iX)W_jXd(P_n-P)| > \frac{t}{\sqrt{n_i}}).$$

By using the similar arguments of bracketing entropy as that of Lemma 3, we have for $c\max(\delta_n, \varepsilon_n\delta_n^{-1}) < a < c\sqrt{n}\delta_n^2$,

$$P(\sup_{\Omega_n}\sqrt{n}|\int B_{nk}^{(i)}(W_iX)W_jXd(P_n-P)| > a) \leq \exp(-ca^2\delta_n^{-2}).$$

Thus

$$P(\sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXd(P_n-P)||_2 > t) \leq n_i\exp(-ct^2n\delta_n^{-2}n_i^{-1}).$$

Then $\sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXd(P_n-P)||_2 = O_P(\sqrt{\frac{\delta_n^2 n_i \log n_i}{n}})$.

Second, notice that $|B_{nk}^{(i)}(x) - B_{nk}^{(i)}(y)| \leq \delta_n^{-1}|x-y|$, then

$$\begin{aligned} sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXdP||_2 &= \sup_{\Omega_n}(\sum_{k=1}^{n_i}|\int (B_{nk}^{(i)}(W_iX)W_jX - B_{nk}^{(i)}(W_{Pi}X)W_{Pj}X)dP|^2)^{1/2} \\ &\leq \sup_{\Omega_n}(\sum_{k=1}^{n_i}(\delta_n^{-1}E||X||_2^2||W_i-W_{Pi}||_2||W_i||_2 + E||X||_2||W_i-W_{Pi}||_2)^2)^{1/2} \\ &= O(\varepsilon_n\delta_n^{-1}\sqrt{n_i}). \end{aligned}$$

Thus $sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXdP_n||_2 = O_P(\sqrt{\frac{\delta_n^2 n_i \log n_i}{n}} + \varepsilon_n\delta_n^{-1}\sqrt{n_i})$.

■

LEMMA 8. $E(\overline{\hat{\phi}}_{W_i}(W_iX)-\phi_{W_i,n}(W_iX))^2 \leq \delta_n^6|\phi'''_{W_i,n}|^2_\infty$.

PROOF. Since for any $h \in \mathcal{G}_n^{(i)}$,

$$E(\overline{\hat{\phi}}_{W_i}(W_iX) - \phi_{W_i,n}(W_iX))^2 \quad \leq \quad E(h(W_iX) - \phi_{W_i,n}(W_iX))^2,$$

then

$$E(\overline{\hat{\phi}}_{W_i}(W_iX) - \phi_{W_i,n}(W_iX))^2 \quad \leq \quad d(\phi_{W_i,n}, \mathcal{G}_n)^2,$$

where $d(\phi_{W_i,n}, \mathcal{G}_n) = inf_{h\in\mathcal{G}_n}|\phi_{W_i,n} - h|_\infty$. Now the result follows by the Jackson type theorem [de Boor (1978)],

$$d(\phi_{W_i,n}, \mathcal{G}_n) \quad \leq \quad c\delta_n^3|\phi'''_{W_i,n}|_\infty.$$

■

LEMMA 9. $|\overline{\hat{\phi}}_{W_i} - \phi_{W_i,n}|_\infty \leq c\delta_n^2|\phi'''_{W_i,n}|_\infty$; $|\overline{\hat{\phi}}'_{W_i} - \phi'_{W_i,n}|_\infty \leq c|\phi'''_{W_i,n}|_\infty\delta_n$.

PROOF. By Theorem XII.4 of de Boor (1978), there exists a quasi-interpolant with some $a \in \mathcal{R}^{n_i}$,

$$\tilde{\hat{\phi}}_{W_i}(t) \quad = \quad a^T\mathbf{B}_n^{(i)}(t),$$

such that $\tilde{\hat{\phi}}_{W_i}$ simultaneously approximates $\phi_{W_i,n}$ and its first derivative to optimal order, that is

$$|\tilde{\hat{\phi}}_{W_i} - \phi_{W_i,n}|_\infty \quad = \quad c|\phi'''_{W_i,n}|_\infty\delta_n^3$$

and

$$|\tilde{\hat{\phi}}'_{W_i} - \phi'_{W_i,n}|_\infty \quad = \quad c|\phi'''_{W_i,n}|_\infty\delta_n^2.$$

So

$$E(\tilde{\hat{\phi}}_{W_i}(W_iX) - \phi_{W_i,n}(W_iX))^2 \quad \leq \quad c|\phi'''_{W_i,n}|^2_\infty\delta_n^6.$$

Together with Lemma 8, we have

$$E(\overline{\hat{\phi}}_{W_i} - \tilde{\hat{\phi}}_{W_i})^2 \quad \leq \quad E(\tilde{\hat{\phi}}_{W_i} - \phi_{W_i,n})^2 + E(\overline{\hat{\phi}}_{W_i} - \phi_{W_i,n})^2 \leq c|\phi'''_{W_i,n}|^2_\infty \delta_n^6.$$

Let $coef(\tilde{\hat{\phi}}_{W_i})$, $coef(\overline{\hat{\phi}}_{W_i})$ be coefficients of $\mathbf{B}_n^{(i)}$ in $\tilde{\hat{\phi}}_{W_i}$ and $\overline{\hat{\phi}}_{W_i}$separately, then

$$E(\overline{\hat{\phi}}_{W_i} - \tilde{\hat{\phi}}_{W_i})^2 \quad = \quad E((coef(\tilde{\hat{\phi}}_{W_i}) - coef(\overline{\hat{\phi}}_{W_i}))^T \mathbf{B}_n^{(i)})^2 \geq \lambda_n ||coef(\tilde{\hat{\phi}}_{W_i}) - coef(\overline{\hat{\phi}}_{W_i})||_2^2,$$

where λ_n is the minimum eigenvalue of $A(W_i) = E[\mathbf{B}_n^{(i)}(W_iX)\mathbf{B}_n^{(i)}(W_iX)^T]$. By Lemma 2, $\lambda_n \geq c\delta_n^2$. Thus

$$||coef(\tilde{\hat{\phi}}_{W_i}) - coef(\overline{\hat{\phi}}_{W_i})||_2 \quad \leq \quad c|\phi'''_{W_i,n}|_\infty \delta_n^2.$$

and

$$|\overline{\hat{\phi}}_{W_i} - \tilde{\hat{\phi}}_{W_i}|_\infty \quad \leq \quad ||coef(\tilde{\hat{\phi}}_{W_i}) - coef(\overline{\hat{\phi}}_{W_i})||_2 \leq c|\phi'''_{W_i,n}|_\infty \delta_n^2.$$

Hence

$$sup_{\Omega_n}|\overline{\hat{\phi}}_{W_i} - \phi_{W_i,n}| \quad \leq \quad sup_{\Omega_n} c|\phi'''_{W_i,n}|_\infty \delta_n^2.$$

Further by observing $|(B_{nk}^{(i)})'|_\infty \leq \delta_n^{-1}$, we have

$$|\overline{\hat{\phi}}'_{W_i} - \tilde{\hat{\phi}}'_{W_i}|_\infty \quad \leq \quad ||coef(\tilde{\hat{\phi}}_{W_i}) - coef(\overline{\hat{\phi}}_{W_i})||_2 \delta_n^{-1} \leq c|\phi'''_{W_i,n}|_\infty \delta_n.$$

Thus

$$|\overline{\hat{\phi}}'_{W_i} - \phi'_{W_i,n}|_\infty \quad \leq \quad |\tilde{\hat{\phi}}'_{W_i} - \phi'_{W_i,n}|_\infty + |\overline{\hat{\phi}}'_{W_i} - \tilde{\hat{\phi}}'_{W_i}|_\infty \leq c|\phi'''_{W_i,n}|_\infty \delta_n.$$

■

LEMMA 10. $\int (\hat{\phi}_{W_{Pk}}(S_k) - \phi_k(S_k))^2 dP_n = o_p(1)$.

PROOF. Observe that

$$\begin{aligned}\int(\hat{\phi}_{W_{Pk}}(S_k)-\phi_k(S_k))^2dP_n &\leq 3\{\int(\hat{\phi}_{W_{Pk}}(S_k)-\overline{\hat{\phi}}_{W_{Pk}}(S_k))^2dP_n+\int(\overline{\hat{\phi}}_{W_{Pk}}(S_k)-\phi_{k,n}(S_k))^2dP_n\\ &\quad+\int\phi_k(S_k)^2I(S_k\notin[\underline{b}_{nk},\overline{b}_{nk}])dP_n\}.\end{aligned}$$

First, (dropping W_{Pk} in $A_n(W_{Pk}), D_n(W_{Pk}), A(W_{Pk})$ and $D(W_{Pk})$), by Lemma 4, $||A_n^{-1}D_n-A^{-1}D||_2=o_p(1)$, and by Lemma 2 and Lemma 3, $||A_n||_2\leq||A_n-A||_2+||A||_2=o_p(1)$, then

$$\begin{aligned}\int(\hat{\phi}_{W_{Pk}}(S_k)-\overline{\hat{\phi}}_{W_{Pk}}(S_k))^2dP_n &= \int[(\gamma_n-\gamma)^T\mathbf{B}_n^{(k)}(S_k)]^2dP_n\\ &\leq ||\gamma_n-\gamma||_2^2||A_n||_2\\ &= o_p(1).\end{aligned}$$

By Lemma 9, $|\overline{\hat{\phi}}_{W_{Pk}}-\phi_{k,n}|_\infty=o(1)$, then $\int(\overline{\hat{\phi}}_{W_{Pk}}(S_k)-\phi_{k,n}(S_k))^2dP_n=o_p(1)$. Further since $P(S_k\notin[\underline{b}_{nk},\overline{b}_{nk}])\downarrow 0$, $\int\phi_k(S_k)^2I(S_k\notin[\underline{b}_{nk},\overline{b}_{nk}])dP_n=o_p(1)$. Hence the result follows. ■

LEMMA 11. Let $\{p(\cdot;\theta,\eta):\theta\in\Omega\subset R^d,\eta\in\mathcal{E}\}$ be a parameteric or semiparametric model, where θ is the parameter of interest. Suppose that moderate regularity conditions are satisfied and $\mathbf{l}^*(\cdot;\theta,\eta)$ is the efficient score function of θ as defined in BKRW. Then

$$\int\frac{\partial}{\partial\theta}\mathbf{l}^*(X;\theta,\eta)dP_{(\theta,\eta)} = -\int[\mathbf{l}^*\mathbf{l}^{*T}](X;\theta,\eta)dP_{(\theta,\eta)}.$$

PROOF. We only prove it for the parametric case $\mathcal{E}\subset R^m$. Let $I(\theta,\eta)$ be the information matrix of (θ,η). Then by classic likelihood theory (for example, Proposition 2.4.1 of BKRW), $\mathbf{l}^*(\cdot;\theta,\eta)=\dot{l}_1-(I_{12}I_{22}^{-1})(\theta,\eta)\dot{l}_2$, where $\dot{l}_1$ and $\dot{l}_2$ are the partial derivatives of $l(\cdot;\theta,\eta)\equiv\log p(\cdot;\theta,\eta)$ w.r.t θ and η separately. Thus $\frac{\partial}{\partial\theta}\mathbf{l}^*(X;\theta,\eta)=\ddot{l}_{11}-(I_{12}I_{22}^{-1})(\theta,\eta)\ddot{l}_{21}-\frac{\partial}{\partial\theta}\{(I_{12}I_{22}^{-1})(\theta,\eta)\}\dot{l}_2$. Since $\int\dot{l}_2(X;\theta,\eta)dP_{(\theta,\eta)}=0$, we have

$$\int\frac{\partial}{\partial\theta}\mathbf{l}^*(X;\theta,\eta)dP_{(\theta,\eta)} = \int\ddot{l}_{11}dP_{(\theta,\eta)}-(I_{12}I_{22}^{-1})(\theta,\eta)\int\ddot{l}_{21}dP_{(\theta,\eta)}.$$

For the information matrix we have $I_{ij}=-\int\ddot{l}_{ij}(X;\theta,\eta)dP_{(\theta,\eta)}$ $(i,j=1,2)$, hence the result follows by $\int\mathbf{l}^*\mathbf{l}^{*T}dP=I_{11}-I_{12}I_{22}^{-1}I_{21}$ (see Proposition 2.4.1 of BKRW, page 32). For the semiparametric case, the reader is referred to BKRW for a generalization of this proof. ■

References

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

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

[3] Amari, S., Cichocki, A. and Yang, H. (1996). A new learning algorithm for blind signal separation. In Touretzky, D.S., Mozer, M.C. and Hasselmo, M.E., editors, *Advances in Neural Information Processing Systems, **8***. Cambridge, MA: MIT Press.

[4] Bach, F. and Jordan, M. (2002). Kernel independent component analysis. *Journal of Machine Learning Research* **3** 1-48.

[5] Bickel, P. and Doksum, K. (2001). *Mathematical Statistics, Volume I,* second edition. Prentice Hall.

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

[7] Bickel, P. and Ritov, Y. (2000). Comment (on Profile Likelihood). *Journal of the American Statistical Association* **95**(450) 466-468.

[8] Cardoso, J. F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE* **86**(10) 2009-2025.

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

[10] Chen, A. and Bickel, P.J. (2004). Consistent independent component analysis and prewhitening. Accepted by *IEEE Trans. Signal Processing.*

[11] Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing* 36(3):287-314.

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

[13] de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag.

[14] Eriksson, J. and Koivunen, V. (2003). Characteristic-function based independent component analysis. *Signal Processing*, **83**: 2195-2208.

[15] Fan, J. and Wong, W. (2000). Comment (on Profile Likelihood). *Journal of the American Statistical Association* **95**(450) 468-471.

[16] Faraway, J.J. (1992). Smoothing in adaptive estimation. *Ann. Statist.* **20**(1) 414-427.

[17] Hansen, M. H., Huang, J., Kooperberg, C., Stone, C. J., and Truong, Y.K. (2001). *Statistical Modeling with Spline Functions Methodology and Theory*. Springer-Verlag, New York.

[18] Hastie, T. and Tibshirani, R. (2002). Independent component analysis through product density estimation, *Technical report*, Department of Statistics, Stanford University.

[19] Huber, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435-525.

[20] Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks* **10**(3) 626-634.

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

[22] Hyvarinen, A. and Oja, E. (1997). A fast fixed point algorithm for independent component analysis. *Neural Computation,* **9**(7) 1483-1492.

[23] Jin, K. (1992). Empirical smoothing parameter selection in adaptive estimation. *Ann. Statist.* **20**(4) 1844-1874.

[24] Jung, T-P, Makeig, S., Westerfield, M., Townsend, J., Courchesne, E. and Sejnowski. (2001). Independent component analysis of single-trial event-related potentials. *Human Brain Mapping* **14**(3) 168-185.

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

[26] Lee, T. W., Girolami, M. and Sejnowski, T. (1999). Independent component analysis using an extended informax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation* **11**(2) 417-441.

[27] Lee, T. W., Girolami, M., Bell, A. and Sejnowski, T. (2000). A unifying information-theoretic framework for independent component analysis. *Computers and Mathematics with Application* **39** 1-21.

[28] Murphy, S. and Van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association* **95** 449-485.

[29] Pham, D. T. and Garrat, P. (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing* **45**(7) 1712-1725.

[30] Samarov, A. and Tsybakov, A. (2002). Nonparametric independent component analysis. *Bernoulli* **10**(4) 565-582.

[31] van de Geer, S. (2000). *Applications of Empirical Process Theory*. Cambridge University Proess, UK.

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

[33] Vigario, R., Jousmaki, V., Hamalainen, M., Hari, R. and Oja, E. (1998). Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in Neural Information Processing Systems* **10** 229-235. MIT Press.

[34] Vlassis, N. and Motomura, Y. (2001). Efficient source adaptivity in independent component analysis. *IEEE Trans. Neural Networks* **12**(3) 559-565.