EFFICIENT INDEPENDENT COMPONENT ANALYSIS (II)
[1]

By Aiyou Chen and Peter J. Bickel

University of California, Berkeley

Abstract

Independent component analysis (ICA) has been widely used in separating hidden sources from observed linear mixtures 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-dept analysis on the convergence rate and 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 Fisher 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 a $m\times 1$ 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 $m\times m$ matrix θ,

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

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 $S_i=(\theta^{-1}X)_i$. This is also called blind separation in engineering. Let $W=\theta^{-1}$ (called

[1]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.

the demixing matrix). Then the aim is equivalent to find 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.

The model (1) is in fact a special form of a linear structure whose characterization properties have been well studied by Kagan, Linnik and Rao (1973). Comon (1993) showed that θ (W) is identifiable up to scaling and permutation of its columns (rows) iff at most one of S_is is normal. Note that each random variable involved is assumed to be nondegenerate. To illustrate the identifiability invariance to scaling and permutation, suppose $S = WX$ has mutually independent components. Then for a permutation or nonsingular diagonal matrix B, $T = (BW)X$ also has mutually independent components. So to identify W uniquely, it is necessary to define an order and scale for either W's rows or S's components. The model (1) can be seen as a standard semiparametric model with parameters $(W, r_1, \cdots, r_m)$, where r_i parameterizes S_i's density/mass function and $(W, r_1, \cdots, r_m)$ satisfies some identifiability conditions. In this paper, W is the parameter of interest and $\mathbf{r} = (r_1, \cdots, r_m)^T$ is the nuisance parameter. (Note that a math-bold notation stands for a function matrix or vector in this paper.)

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. Most of them are based on estimating equations deduced from some contrast functions, such as MLE [e.g., Pham and Garrat (1997) and Lee, Girolami and Sejnowski (1999)], minimizing mutual information of WX [e.g., Comon (1994)] by parametrizing each S_i's distribution finitely, minimizing higher-order correlation between WX's components [e.g., Cardoso (1999)], and maximizing the non-gaussianity of WX's components [e.g., Hyvarinen (1999)]. 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 and hence that no estimating equation method could be efficient. 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! Recently, some nonparametric methods to estimate W have appeared. For example, Bach and Jordan (2002) proposed: i) To reduce the dimension of the data using a kernel representation; ii) To choose W so as to minimize the empirical *generalized variance* between the components of the vector obtained by applying W to the data. Hastie and Tibshirani (2002) proposed maximizing the penalized likelihood as a function of $(W, r_1, \cdots, r_m)$. Various performance analyses have been made using simulations. But in-depth analysis of the convergence rate and the potential for asymptotic efficiency has to our knowledge not yet been carried out. The construction of Fisher efficient estimates is our concern in this paper. We develop a Fisher efficient estimator by using a sieve profile likelihood technique. This estimator is produced by starting an algorithm at a consistent point. An estimate based on comparison of characteristic functions (CHFICA) first proposed by Eriksson, Kankainen and Koivunen (2001) and studied by us elsewhere is used both theoretically and in our simulations for this purpose. (CHFICA has been shown to be consistent under identifiability conditions and

to be $\sqrt{n}$ consistent under further mild conditions [Chen & Bickel (2003)]. Samarov and Tsybakov (2002) have also proposed a $\sqrt{n}$ consistent estimate.)

In the following, we analyze the ICA model (1) under the framework of semiparametric models [see, e.g., Bickel, Klaassen, Ritov and Wellner (1993)] and propose a new method of estimating W using the efficient score function, as developed in Section 2. Numerical studies are given in Section 3. Asymptotic properties are studied in Section 4. Section 5 and Section 6 contain the technical details.

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

2. Semiparametric inference.

2.1. *Some notation.* Let W_P be one of the nonsingular demixing 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 of wX and define its logarithmic derivative ϕ_w by

$$\phi_w(t) \;=\; -\frac{1}{f_w(t)}\frac{\partial f_w(t)}{\partial t} I(f_w(t) > 0).$$

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

$$\begin{aligned} f_w(t) &= \int_{R^{m-1}} \frac{1}{v_k} r_k\Big(\frac{t - \sum_{j\neq k} v_j s_j}{v_k}\Big) \prod_{j\neq k} r_j(s_j) ds_j \\ &= E\Big[\frac{1}{v_k} r_k\Big(\frac{t - \sum_{j\neq k} v_j S_j}{v_k}\Big)\Big]. \end{aligned} \tag{2}$$

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)]

$$\begin{aligned} \phi_w(t) &= -E\Big[\frac{\partial}{\partial t}\log\Big(r_k\Big(\frac{t - \sum_{j\neq k} v_j S_j}{v_k}\Big)\Big) \Big| vS = t\Big] \\ &= \frac{1}{v_k} E[\phi_k(S_k)|vS = t]. \end{aligned} \tag{3}$$

2.2. *Efficient score function of W.* As it is 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.,

$$\mathrm{med}(|S_i|) = 1 \quad \text{or} \quad 2\int_{|s|\leq 1} r_i(s) ds \;=\; 1. \tag{4}$$

Even after this choice, the correct demixing 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. We use Chen and Bickel (2003)'s estimate of W obtained by using empirical characteristic functions which is consistent under identifiability conditions and $\sqrt{n}$ consistent under mild additional regularity conditions. Define $k(s) = 2I(|s| \leq 1) - 1$, where $I(.)$ is an indicator function. Then (4) is equivalent to

$$\int_{S_i = W_i X} k(S_i) dP_{(W,\mathbf{r})} \quad = \quad 0.$$

By parametrizing the model (1) with $(W, r_1, \cdots, r_m)$, the likelihood function of X can be expressed as $p_X(\mathbf{x}) = |\det(W)| \prod_{i=1}^m r_i(\mathbf{W}_i \mathbf{x})$. In the following we heuristically calculate the efficient score function of W under the framework of semiparametric theory [see Bickel, Klaassen, Ritov and Wellner (1993)]. For simplicity, we assume $E[S_i] = 0$. For the convenience of notation, let $S = WX$ and E be the expectation operator under $P_{(W,\mathbf{r})}$.

Let $\phi_i = -\frac{r_i'}{r_i}$ and define Φ by $\Phi(\mathbf{s}) = (\phi_1(\mathbf{s}_1), \cdots, \phi_m(\mathbf{s}_m))^T$. Then the score function of W, $\dot{l}_W(\mathbf{x}) = \frac{\partial}{\partial W} \log(p_X(\mathbf{x}))$, is equal to

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

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

$$E[\phi_i(S_i)^2] \quad < \quad \infty \text{ and } E[S_i^2] < \infty. \tag{5}$$

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 these tangent spaces are perpendicular to each other since S_is are mutually independent. The efficient score of W can then be expressed as

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

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

$$e(\mathbf{x}; W, \Phi) \quad = \quad \mathbf{M} W^{-T}, \tag{6}$$

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

$$\mathbf{M}_{ij} \quad = \quad -\phi_i(W_i \mathbf{x}) W_j \mathbf{x}, \text{ for } 1 \leq i \neq j \leq m, \tag{7}$$

$$\mathbf{M}_{ii} \quad = \quad \alpha_i W_i \mathbf{x} + \beta_i k(W_i \mathbf{x}), \text{ for } i = 1, \cdots, m, \tag{8}$$

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

$$\alpha_i = -\frac{(1-u_i)v_i}{\sigma_i^2 - v_i^2},\ \beta_i = \frac{(1-u_i)\sigma_i^2}{\sigma_i^2 - v_i^2},\ \sigma_i^2 = E[S_i^2], \tag{9}$$

$$v_i = E[2S_i I(|S_i| \leq 1)],\ u_i = E[2S_i\phi_i I(|S_i| \leq 1)]. \tag{10}$$

(Note: Most of these formulas appear in Amari and Cardoso (1997).) By the convolution theorem on semiparametric models [Bickel, Klaassen, Ritov and Wellner (1993)], the information bound for regular estimators of W is $(E[e(X;W,\Phi)^T e(X;W,\Phi)])^{-1}$, where $e(.;W,\Phi)$ is considered as a column vector function, reshaped row by row. By Theorem 7.8.1 in Bickel, Klaassen, Ritov and Wellner (1993), a one-step MLE will be efficient if there is available a $\sqrt{n}$ consistent initial value of W and $e(.;W,\Phi)$ can be estimated. But the requirement of a $\sqrt{n}$ consistent initial estimate limits a direct implementation of the one-step MLE.

2.3. *Estimating a density score function by B-spline approximations.* Let $\phi = -\frac{r'}{r}$ be the score of a density function r. An approximation of ϕ by a member of an N-dim linear space $\mathcal{G}$ with differentiable basis functions $\mathbf{B} = (B_1, \cdots, B_N)^T$, is to minimize, over $\beta \in R^N$,

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

This can be been as a variant of Cox (1985)'s penalized estimator of ϕ [see, e.g., Jin (1992)]. Notice that by partial integration,

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

where E_r is the expectation operator under the probability measure $r(s)ds$. Thus the optimal β is $\beta_\phi = (E_r[\mathbf{B}^T\mathbf{B}])^{-1}E_r[\mathbf{B}']$ and the best approximation of ϕ in $\mathcal{G}$ in the sense of mean square error is $\phi_\mathcal{G} = \beta_\phi^T\mathbf{B}$. Given n random observations from the density function r, β_ϕ can be estimated by combinations of empirical moments. So a natural estimate of ϕ is given by

$$\hat{\phi}_\mathcal{G} = \hat{\beta}_\phi^T\mathbf{B},\ \text{where } \hat{\beta}_\phi = (\hat{E}_r[\mathbf{B}^T\mathbf{B}])^{-1}\hat{E}_r[\mathbf{B}'], \tag{11}$$

and $\hat{E}_r$ denotes the empirical mean operator corresponding to E_r.

B-spline basis functions because of nice properties such as smoothness are popular choices for $\mathcal{G}$. For example, to construct N quadratic B-splines, we may choose $N+3$ inner points with an equal interim distance $\delta_n = (\overline{b}_n - \underline{b}_n)/(N+2)$ or $N+3$ empirical quantiles in a working interval $[\underline{b}_n, \overline{b}_n] \subseteq supp(r)$. 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 dimension number N is an empirical smoothing parameter, which can be dealt with as usual by cross validation. Jin (1992) used B-spline basis functions for

$\mathcal{G}$ and thoroughly studied the adaptive choice of N; To be precise, under weak conditions on r's smoothness, a good adaptive choice of N by cross-validation is

$$N \;=\; O(n^{\delta}),\ \text{where } 0<\delta<\frac{1}{6},$$

and δ depends on the tail property of r. The advantages of this estimator are its explicit form and its easy implementation with small computational load $O(n^{1+\delta})$.

2.4. *Estimation of W.* Assume that an available starting estimate $W_n^{(0)}$ is consistent for W_P. Following MLE, a plausible efficient estimate of W would be to solve its efficient score equation $\int e(X;W,\Phi)dP_n(X)=0$, where Φ needs to be replaced by a good approximation. Alternatively, we use the sieve profile likelihood technique [e.g., Murphy and Van der Vaart (2000) and Fan and Wong (2000)], that is, to find a path Φ_W (possibly depending on some sieve parameters) indexed by W such that the path Φ_W (approximately) passes Φ at W_P, and then to solve $\int e(X;W,\Phi_W)dP_n(X)=0$. The critical issue is that the approximated efficient score function along this path must be $\sqrt{n}$ *unbiased* at W_P [Bickel & Ritov (2000)]. The following gives a way to construct the sieves and the path Φ_W for W close to W_P.

Choose a sequence of positive constants δ_n with $\delta_n\downarrow 0$. For each $k\in\{1,\cdots,m\}$, let $[\underline{b}_{nk},\overline{b}_{nk}]\subseteq supp(r_k)$ such that for some positive constants c_1 and c_2,

$$r_k(t) \;\geq\; c_1\delta_n \text{ for } t\in[\underline{b}_{nk},\overline{b}_{nk}],\ r_k(t)\leq c_2\delta_n \text{ for } t\in[\underline{b}_{nk},\overline{b}_{nk}]^c. \tag{12}$$

This is always possible for large n when r_k is absolutely continuous and can be obtained by using the initial consistent estimate $W_n^{(0)}$. As $ES_k=0$ by assumption, we have $\underline{b}_{nk}<0<\overline{b}_{nk}$ for large n. Thus we may fix zero as one of the knots. Let us fix a $k\in\{1,\cdots,m\}$. Set the knots for the estimation of ϕ_k as

$$[\underline{b}_{nk}/\delta_n]\delta_n,\cdots,-\delta_n,0,\delta_n,\cdots,[\overline{b}_{nk}/\delta_n]\delta_n$$

and denote them sequentially by

$$\xi_0^{(k)} \;<\; \xi_1^{(k)}<\cdots<\xi_{n_k+3}^{(k)},$$

where $n_k=[\overline{b}_{nk}/\delta_n]-[\underline{b}_{nk}/\delta_n]-3$ and $[t]$ denotes the largest integer no more than t for $t\in\mathcal{R}$. Using these knots sequentially, define the third-order (quadratic) B-splines as follows: for $i=0,\cdots,n_k$ (dropping the superscript k in $\xi^{(k)}$ henceforth)

$$\begin{aligned}B_{ni}^{(k)}(x) \;=\;& \frac{(x-\xi_i)^2}{2\delta_n^2}I(\xi_i<x\leq\xi_{i+1})+\Big(\frac{(x-\xi_{i+1})(\xi_{i+2}-x)}{\delta_n^2}+\frac{1}{2}\Big)I(\xi_{i+1}<x\leq\xi_{i+2})\\ &+\frac{(\xi_{i+3}-x)^2}{2\delta_n^2}I(\xi_{i+2}<x\leq\xi_{i+3}).\end{aligned}$$

Then let

$$\mathbf{B}_n^{(k)} \;=\; (B_{n0}^{(k)},B_{n1}^{(k)},\cdots,B_{nn_k}^{(k)})^T.$$

Note that here the superscript $B^{(k)}$ does not mean kth order derivative, instead we will use B', B'' to denote their first and second order pointwise derivatives separately, similarly for $\mathbf{B}$'s. B-splines have derivatives and nice smoothness properties [see, e.g., De Boor (1978) and Hansen, Huang, Kooperberg, Stone and Truong (2001)].

Let

$$\mathcal{G}_n^{(k)} \quad = \quad \{a^T \mathbf{B}_n^{(k)} : a \in \mathcal{R}^{n_k+1}\}.$$

Then for any $m \times m$ matrix W, we use (11) to estimate ϕ_{W_k} in $\mathcal{G}_n^{(k)}$ for $k = 1, \cdots, m$, that is,

$$\hat{\phi}_{W_k} \quad = \quad (A_n^{-1} D_n)^T \mathbf{B}_n^{(k)}(.), \tag{13}$$

where $A_n = \int \mathbf{B}_n^{(k)}(W_k X)\mathbf{B}_n^{(k)}(W_k X)^T dP_n$ and $D_n = \int (\mathbf{B}_n^{(k)})'(W_k X) dP_n$.
$\hat{\Phi}_W = (\hat{\phi}_{W_1}, \cdots, \hat{\phi}_{W_m})^T$.

For the use of later proofs, define

$$\overline{\hat{\phi}}_{W_k} \quad = \quad (A^{-1} D)^T \mathbf{B}_n^{(k)}(.), \tag{14}$$

where $A = \int \mathbf{B}_n^{(k)}(W_k X)\mathbf{B}_n^{(k)}(W_k X)^T dP$ and $D = \int (\mathbf{B}_n^{(k)})'(W_k X) dP$.

Now we replace the efficient score function $e(X; W, \Phi)$ defined in (6) by its profile form $e(X; W, \hat{\Phi}_W)$, where $\alpha_i, \beta_i, \sigma_i^2$ defined in (9)-(10) 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. Thus

$$\hat{\alpha}_i \quad = \quad -\frac{(1-\hat{u}_i)\hat{v}_i}{\hat{\sigma}_i^2 - \hat{v}_i^2},\ \hat{\beta}_i = \frac{(1-\hat{u}_i)\hat{\sigma}_i^2}{\hat{\sigma}_i^2 - \hat{v}_i^2},\ \hat{\sigma}_i^2 = \int (W_i X)^2 dP_n, \tag{15}$$

where

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

Define $\mathbf{e}_n(W) = \int e(X; W, \hat{\Phi}_W) dP_n$ and $\mathbf{e}(W) = \int e(X; W, \Phi_W) dP$.

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

$$\mathbf{e}_n(W) \quad = \quad 0. \tag{16}$$

We use the Newton-Rapson iteration method to solve this equation with the starting value $W_n^{(0)}$. That is, $\hat{W}_n$ is the limit of the iteration sequence

$$W_n^{(j+1)} = W_n^{(j)} - \dot{\mathbf{e}}_n(W_n^{(j)})^{-}\mathbf{e}_n(W_n^{(j)}),\ j = 0, 1, \cdots \tag{17}$$

where $\dot{\mathbf{e}}_n$ is a super matrix with $[\dot{\mathbf{e}}_n]_{ij,kl} = \frac{\partial}{\partial W_{kl}}[\mathbf{e}_n(W)]_{ij}$. Here we use the $m \times m$ matrix $\mathbf{e}_n$ as a column vector and define the its partial derivative matrix $\dot{\mathbf{e}}_n$ correspondingly of dimension $m^2 \times m^2$. But notice that $\hat{\Phi}_W$ is also a function of W. For the convergence of Newton-Rapson algorithm, the initial value needs to be close to the truth. Fortunately, it is not hard to get a consistent estimate for ICA model. For example, CHFICA [Chen & Bickel (2003)] is consistent in general and $\sqrt{n}$ consistent when hidden components have finite variances. The convergence and asymptotic properties are analyzed in Section 4. Call $\hat{W}_n$ the SPMLE.

3. Numerical studies. We do two groups of experiments to test the empirical performance of the SPMLE. We generate data from known source distributions listed in Table 0 and then obtain linear mixtures of them by a known mixing matrix $\theta = W_P^{-1}$.

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 0 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 and the sample sizes are increased sequentially (the detailed setup of the sample sizes and replication times is given in Table 2). The m hidden components are chosen in order from the first m source distributions of [0], [1], $\cdots$, [11] in Table 0, and W_P is the $m \times m$ identity matrix.

Comparisons are made with three existing ICA algorithms: the FastICA algorithm [Hyvarinen & Oja (1997)], the JadeICA algorithm [Cardoso (1999)], and the KernelICA-Kgv algorithm [Bach & Jordan (2002)]. To obtain an initial estimated value of the demixing matrix for the SPMLE, in the first group of experiments, we use the Monte-Carlo version of CHFICA (Mc-ICA) [Chen & Bickel (2003)], while in the second group, we use the KernelICA-Kgv algorithm. The performance of each algorithm is measured by the so-called *Amari error* $d(\hat{W}, W_P)$ [Amari, Cichocki & Yang (1996)]:

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

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.

Table 1 and Table 2 provide the simulation results of the two groups of experiments separately. Table 0 lists all the source distributions used in the simulations.

pdfs	Fast	Jade	Kgv	Mc	SPMLE	Fast	Jade	Kgv	Mc	SPMLE
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: *Reporting the medians of the Amari errors (multiplied by 1000) for two components ICA with 1000 samples(left) and 2000 samples(right) in 400 replications*

```
 Table 0: Source distributions used in the simulations
-------------------------------------------------------------------

 [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)
-------------------------------------------------------------------
```

As we can see from the simulation results, the SPMLE has the smallest Amari errors in most experiments except in cases that all the sources are mixture Gaussians which decay rapidly in the tails. This can be explained by two facts: First, the KernelICA-Kgv itself is very accurate as shown in Bach & Jordan (2002), especially when hidden sources have rapid decaying densities; Second, the efficiency of the SPMLE is in the sense of large sample size.

4. **Asymptotic properties.** Suppose that there exists ε_n with $\varepsilon_n \to 0$, $\sqrt{n}\varepsilon_n \to \infty$ such that as $n \to \infty$,

$$P(|W_n^{(0)} - W_P| \ \leq \ \varepsilon_n) \to 1. \tag{18}$$

m	N	#repl	Fast	Jade	Kgv	SPMLE
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 medians of the Amari errors (multiplied by 1000) for m components with N samples: m components are first m pdfs in the source list*

Let

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

Here is our main theorem. (Let ULLN denote the Uniform Law of Large Numbers [e.g., Van der Geer (2000)].)

THEOREM 1. *In the ICA model (1), if (18) and the following conditions are satisfied for* $i, j, k = 1, \cdots, m$, $i \neq k$ *and* $j \neq k$ *:*

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

[ii]. $|r_k|_\infty < \infty, |r_k'|_\infty < \infty$, $sup_{t\in\mathcal{R}}|tr_k'(t)| < \infty$*;*

[iii]. W_P *is nonsingular;*

[iv]. $\varepsilon_n \delta_n^{-\frac{9}{2}} n_k^{\frac{1}{2}} = o(1)$, $\varepsilon_n(\overline{b}_{nk} - \underline{b}_{nk}) = o(1)$ *and* $[\underline{b}_{nk}, \overline{b}_{nk}]$ *satisfies (12);*

[v]. $sup_{W\in\Omega_n}|\phi'''_{W_k,n}|_\infty \delta_n = o(1)$*;*

[vi]. ULLN holds for $\{\phi_{W_k}(W_kX)X_i : W \in \Omega_n\}$ *and for* $\{\phi'_{W_k}(W_kX)W_iXX_j : W \in \Omega_n\}$.

Then $\hat{W}_n$ *defined by (25) exists and satisfies*

$$\sqrt{n}(\hat{W}_n - W_P) \;=\; I_{eff}^{-1}\sqrt{n}\int e(X; W_P, \Phi_P)dP_n + o_P(1), \tag{19}$$

where $I_{eff} = \int e(X; W_P, \Phi_P)e(X; W_P, \Phi_P)^T dP$*, that is,* $\hat{W}_n$ *is Fisher efficient. (Note: (19) is considered in a vector form.)*

Remark: Condition [i] assumes that each hidden component has a finite variance and finite Fisher information, together with mean zero and absolute median of 1, where absolute median of 1 is for rescaling so that the demixing matrix may be uniquely identified [Comon (1994)]. It should be clear that the zero mean assumption 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 [ii] assumes that the score function ϕ_k of each hidden component is smooth enough to make its B-spline approximation well behaved; Condition [iv] requires that the initial value be consistent or reasonably close to the truth and that the domain and the number of knots of the B splines do not grow so quickly that we lose control of the approximation to Φ_W. [v]&[vi] looks complicated but are not strong. (For example, if $\frac{\phi_k(t)}{t} = O(1)$ and $\phi_k'(t) = O(1)$ as $t \to \infty$ or $-\infty$, then by (3) ϕ_{W_k} is bounded by a linear form of X and $\phi_{W_k'}$ is bounded from infinity, thus [vi] holds.)

5. Some technical details. The estimator defined by (16)-(17) can be viewed as a generalized M-estimators (GM-estimator). The convergence and asymptotic linearity of GM-estimators have been studied in Bickel, Klaassen, Ritov and Wellner (1993). Define this citation as BKRW and use it thereafter. 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 $M(\theta, P) = M_\infty(\theta, P)$.

THEOREM 2 [BKRW] (ASYMPTOTIC LINEARITY PROPERTIES OF GM-ESTIMATOR) *Let $\hat{\theta}_n$ solve $M_n(\theta, P_n) = o_p(n^{-1/2})$. Let $V_n(\theta) = \sqrt{n}(M_n(\theta, P_n) - M(\theta, P))$. If the following conditions hold:*

[GM0]. $M(\theta_P, P) = 0$ and $\theta_P \in \Omega$ is the unique solution of $M(\theta, P) = 0$ in Ω.

[GM1]. for any $\varepsilon_n \to 0$, $sup_{||\theta - \theta_P|| \le \varepsilon_n} |V_n(\theta) - V_n(\theta_P)|/(1 + \sqrt{n}||\theta - \theta_P||) = o_p(1)$;

[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;

[GM4]. $||\hat{\theta}_n - \theta_P|| = o_P(1)$.

Then, $\sqrt{n}(\hat{\theta}_n - \theta_P) = -\sqrt{n}[\frac{\partial M(\theta_P, P}{\partial \theta}]^{-1} \int \psi_{\theta_P}(X) dP_n + o_p(1)$.

We will prove these are true for our SPMLE under the conditions of Theorem 1 by using the Iteration Theorem in Bickel, Klaassen, Ritov and Wellner (1993). Let $M_n(\theta, P_n) = \mathbf{e}_n(W)$. Note that $\mathbf{e}_n(W)$ is a nonlinear functional of P_n and the sieve parameters include $\delta_n, \underline{b}_{nk}, \overline{b}_{nk}$ for $k \in \{1, \cdots, m\}$. But instead of (GM1), we use the following stronger condition:

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

THEOREM 3 [BKRW]. *Suppose (GM0),(GM2),(GM3) with $M_n(\theta, P_n) = \mathbf{e}_n(W)$ and (U) hold. If the starting point satisfies $P(|W_n^{(0)} - W_P| < \varepsilon_n) \to 1$, then with probability converging to 1, $\mathbf{e}_n(W)$ in (16) has a unique root $W_n^{(\infty)}$ in Ω_n, and $W_n^{(\infty)}$ is asymptotically linear with the influence function $-E[\dot{e}(X; W_P, \Phi_P)]^{-1} e(.; W_P, \Phi_P)$.*

Theorem 3 is called the Iteration Theorem. It is obvious that (GM0) holds under the conditions of Theorem 1 as it is the efficient score function. The following propositions 1, 2 and 3 verify (GM2), (GM3) and (U) separately, which thus proves the convergence and asymptotic linearity of the SPMLE. Futher, for the efficient score function, we have that

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

thus the SPMLE is asymptotic efficient in the setup of BKRW's Fisher efficiency.

6. Proposition 1-3. The following table lists all the notations used in the proofs, for $k \in \{1, \cdots, m\}$, $W \in \Omega_n$:

P, P_n	population, empirical law of X given IID copies $(X^{(1)}, \cdots, X^{(n)})$
W, W_k, W_{ij}	$m \times m$ matrix, its kth row vector, its (i,j) element
W_P, W_{Pk}, W_{Pij}	$m \times m$ matrix (truth for P), its kth row, its (i,j) element
r_k	density function of the kth hidden component $S_k (= W_{Pk}X)$
$\phi_k = -r_k'/r_k$	score function of the kth hidden component $S_k (= W_{Pk}X)$
$\Phi_P = (\phi_1, \cdots, \phi_m)^T$	function vectors
f_{W_k}	density function of W_kX ($f_{W_{Pk}} \equiv r_k$)
$\phi_{W_k} = -f_{W_k}'/f_{W_k}$	score function of W_kX ($\phi_{W_{Pk}} \equiv \phi_k$)
$\Phi_W = (\phi_{W_1}, \cdots, \phi_{W_m})^T$	function vector
$\mathbf{B}_n^{(k)} = (B_{n0}^{(k)}, \cdots, B_{nn_k}^{(k)})^T$	B spline functions defined on $[\underline{b}_{nk}, \overline{b}_{nk}]$ with interim distance δ_n
$A_n(W_k) = \int_{Y=W_kX} \mathbf{B}_n^{(k)}(Y)\mathbf{B}_n^{(k)T}(Y)dP_n$	in coefficients of $\hat{\phi}_{W_k}$
$D_n(W_k) = \int (\mathbf{B}_n^{(k)})'(W_kX)dP_n$	in coefficients of $\hat{\phi}_{W_k}$
$\gamma_n(W_k) = A_n(W_k)^{-1}D_n(W_k)$	coefficients of $\hat{\phi}_{W_k}$
$A(W_k) = \int_{Y=W_kX} \mathbf{B}_n^{(k)}(Y)\mathbf{B}_n^{(k)T}(Y)dP$	in coefficients of $\hat{\phi}_{W_k}$
$D(W_k) = \int (\mathbf{B}_n^{(k)})'(W_kX)dP$	in coefficients of $\hat{\phi}_{W_k}$
$\gamma(W_k) = A(W_k)^{-1}D(W_k)$	coefficients of $\overline{\hat{\phi}}_{W_k}$
$\mathcal{G}_n^{(k)} = \{a^T\mathbf{B}_n^{(k)} : a \in \mathcal{R}^{n_k}\}$	closed linear span of B spline functions

$\hat{\phi}_{W_k} = \gamma_n(W_k)^T \mathbf{B}_n^{(k)}$ estimator of ϕ_{W_k} in $\mathcal{G}_n^{(k)}$, defined in (13)

$\overline{\hat{\phi}}_{W_k} = \gamma(W_k)^T \mathbf{B}_n^{(k)}$ estimator of ϕ_{W_k} in $\mathcal{G}_n^{(k)}$, defined in (14)

$\phi_{k,n}, \phi_{W_k,n}$ truncation of ϕ_k, ϕ_{W_k} on $[\underline{b}_{nk}, \overline{b}_{nk}]$

$e(X; W, \Phi)$ efficient score function of W, defined in (6)

$\mathbf{e}(W) = \int e(X; W, \Phi_W) dP$ expectation

$\mathbf{e}_n(W) = \int e(X; W, \hat{\Phi}_W) dP_n$ empirical expectation

---

It is noted that all the lemmas used in this section are provided and proved in Appendix A.

PROPOSITION 1. *Under the conditions Theorem 1,*
$\mathbf{e}_n(W_P) = \int e(x; W_P, \Phi_P) dP_n + o_P(n^{-1/2})$.

Proof: 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 (9) and (15) separately, and

$$\int \hat{\phi}_{W_{Pi}}(S_i) S_j dP_n = \int \phi_{W_{Pi}}(S_i) S_j dP_n + o_P(n^{-1/2}), \tag{20}$$

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

The first two are not hard to be verified by the Central Limit Theorem and Lemma 10. Here we just show the last argument (20). Observing 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,

$$\begin{aligned} [1] &= |\int (A_n^{-1}(W_{Pi}) D_n(W_{Pi}) - A^{-1}(W_{Pi}) D(W_{Pi}))^T \mathbf{B}_n^{(i)}(S_i) S_j dP_n| \\ &\le ||A_n^{-1} D_n - A^{-1} D||_2 || \int \mathbf{B}_n^{(i)}(S_i) S_j dP_n||_2 \\ &= \varepsilon_n \delta_n^{-4} \sqrt{\frac{n_i \log n_i}{n}} O_P(1) O_P(n^{-1/2}) \\ &= o_P(n^{-1/2}), \end{aligned}$$

where the rate equality is from Lemma 4 and Lemma 6 in Appendix A.

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 in Appendix A, $|\overline{\hat{\phi}}_{W_{Pi}} - \phi_{W_{Pi},n}|_\infty \leq c\delta_n^2 |\phi'''_{W_{Pi},n}|_\infty$, thus

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

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

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

So $[3] = o_P(n^{-1/2})$. ■

PROPOSITION 2. *Under condition [i] & [iii] in Theorem 1, $\frac{\partial \mathbf{e}(W)}{\partial W}|_{W_P}$ is nonsingular and $\mathbf{e}(W)$ is differential w.r.t W in a neighbourhood of W_P.*

Proof. By the classical likelihood theory, we have

$$-\frac{\partial \mathbf{e}(W)}{\partial W}|_{W_P} \;=\; E[e(X; W_P, \Phi_P)e(X; W_P, \Phi_P)^T], \tag{21}$$

and arguments of $e(.; W_P, \Phi_P)$ are linearly independent, so the first claim holds. The second claim holds as all terms in $\mathbf{e}(W)$ are differentiable. ■

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

$$sup_{\Omega_n} |\int \hat{\phi}_{W_i}(W_i X) X_k dP_n(X) - \int \phi_{Pi}(W_{Pi} X) X_k dP| \;=\; o_P(1), \tag{22}$$

and

$$sup_{\Omega_n} |\int \frac{\partial}{\partial W_i}[\hat{\phi}_{W_i}(W_i X)] W_j X dP_n(X) - \int \frac{\partial}{\partial W_i}[\phi_{Pi}(W_i X)]_{W_{Pi}} W_{Pj} X dP| \;=\; o_P(1) \tag{23}$$

That is, condition [U] for theorem 3 holds.

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

$$\begin{aligned} ||\int \mathbf{B}_n^{(i)}(W_i X) X_k dP_n||_2^2 \;&=\; \sum_{l=0}^{n_i} (\int B_{nl}(W_i X) X_k dP_n)^2 \\ &\leq\; ||X_k||_{P_n}^2 \int \sum_{l=0}^{n_i} B_{nl}^2(W_i X) dP_n \\ &\leq\; 3||X_k||_{P_n}^2, \text{ where } ||X_k||_{P_n}^2 = \int |X_k|^2 dP_n. \end{aligned}$$

Then

$$sup_{\Omega_n}||\int \mathbf{B}_n(W_iX)X_kdP_n||_2 \;=\; O_P(1). \tag{24}$$

And by Lemma 4 in Appendix A, (dropping W_k for simplicity),
$sup_{\Omega_n}||A_n^{-1}D_n - A^{-1}D||_2 = o_P(1)$, so

$$\begin{aligned} & sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)X_kdP_n(X) - \int \overline{\hat{\phi}}_{W_i}(W_iX)X_kdP_n| \\ = \;& sup_{\Omega_n}|(A_n^{-1}D_n - A^{-1}D)^T \int \mathbf{B}_n(W_iX)X_kdP_n| \\ \leq \;& sup_{\Omega_n}||A_n^{-1}D_n - A^{-1}D||_2 sup_{\Omega_n}||\int \mathbf{B}_n(W_iX)X_kdP_n||_2 \\ = \;& o_P(1) \end{aligned} \tag{25}$$

Further, by Lemma 9, $sup_{\Omega_n}|\overline{\hat{\phi}}_{W_i}(W_iX) - \phi_{W_i,n}|_\infty \leq sup_{\Omega_n} c|\phi'''_{W_i,n}|_\infty \delta_n^2 = o(1)$, then

$$\begin{aligned} & sup_{\Omega_n}|\int \overline{\hat{\phi}}_{W_i}(W_iX)X_kdP_n(X) - \int \phi_{W_i,n}(W_iX)X_kdP_n| \\ \leq \;& sup_{\Omega_n}|\phi'''_{W_i,n}|_\infty \delta_n^2 \int |X_k|P_n \\ = \;& o_P(1). \end{aligned} \tag{26}$$

And by Condition [vi], 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

$$\begin{aligned} sup_{\Omega_n}|\int (\phi_{W_i} - \phi_{W_i,n})(W_iX)X_kdP_n| \;&=\; sup_{\Omega_n}|\int \phi_{W_i}(W_iX)X_kI(W_iX \notin [\underline{b}_{ni}, \overline{b}_{ni}])dP_n| \\ &=\; o_P(1). \end{aligned} \tag{27}$$

From (24)-(27), we get

$$sup_{\Omega_n}|\int \hat{\phi}_{W_i}(W_iX)X_kdP_n(X) - \int \phi_{W_i}(W_iX)X_kdP_n(X)| \;=\; o_P(1). \tag{28}$$

Now by the condition [vi],

$$sup_{\Omega_n}|\int \phi_{W_i}(W_iX)X_kd(P_n - P)| \;=\; o_P(1); \tag{29}$$

And by continuity,

$$sup_{\Omega_n}|\int \phi_{W_i}(W_iX)X_kdP - \int \phi_{W_{Pi}}(W_{Pi}X)X_kdP| \;=\; o(1). \tag{30}$$

Then (22) follows from (28)-(30).

In the following, we prove the second one.

Since $\hat{\phi}_{W_i}(W_iX) = \gamma_n^T(W_i)\mathbf{B}_n(W_iX)$, where $\gamma_n(W_i) = A_n^{-1}(W_i)D_n(W_i)$. Thus

$$\frac{\partial}{\partial W_i}\hat{\phi}_{W_i}(W_iX) \;=\; \frac{\partial}{\partial W_i}[\gamma_n^T(W_i)]B_n(W_iX) + \hat{\phi}'_{W_i}(W_iX)X. \tag{31}$$

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

Let's first show [4].

By Lemma 4 in Appendix A and the condition [iv], we have (dropping W_i)

$$\begin{aligned} & sup_{\Omega_n}|\hat{\phi}'_{W_i} - \overline{\hat{\phi}}'_{W_i}|_\infty = sup_{\Omega_n}|(A_n^{-1}D_n - A^{-1}D)^T[\mathbf{B}_n^{(i)}]'|_\infty \\ \leq\; & sup_{\Omega_n}||A_n^{-1}D_n - A^{-1}D||_2\delta_n^{-1} \\ =\; & \delta_n^{-\frac{7}{2}}\sqrt{n_i\log n_i/n}O_P(1) \\ =\; & o_P(1), \end{aligned} \tag{32}$$

so

$$\begin{aligned} & sup_{\Omega_n}|\int[\hat{\phi}'_{W_i} - \overline{\hat{\phi}}'_{W_i}]X_kW_jXdP_n| \\ \leq\; & sup_{\Omega_n}|\hat{\phi}'_{W_i} - \overline{\hat{\phi}}'_{W_i}|_\infty 2\int||X||_2^2dP_n \\ =\; & o_P(1). \end{aligned} \tag{33}$$

By Lemma 9, $|\overline{\hat{\phi}}'_{W_i} - \phi'_{W_i,n}|_\infty \leq c|\phi'''_{W_i,n}|_\infty\delta_n = o_P(1)$, then

$$\begin{aligned} & sup_{\Omega_n}|\int[\overline{\hat{\phi}}'_{W_i}(W_iX) - \phi'_{W_i,n}(W_iX)]X_kW_jXdP_n| \\ \leq\; & sup_{\Omega_n}c|\phi'''_{W_i,n}|_\infty\delta_n\int||X||_2^2P_n \\ =\; & o_P(1). \text{ (by Condition [v])} \end{aligned} \tag{34}$$

Furthermore, by Condition [vi], ULLN holds for $\{\phi'_{W_k}(W_kX)W_iXX_j : W \in \Omega_n\}$, and by Lemma 1, $sup_{\Omega_n}P(W_iX \notin [\underline{b}_{ni}, \overline{b}_{ni}]) = o(1)$, then

$$sup_{\Omega_n}|\int \phi'_{W_i}(W_iX)I(W_iX \notin [\underline{b}_{ni}, \overline{b}_{ni}])X_kW_jXdP_n| \;=\; o_P(1); \tag{35}$$

From Condition [vi],

$$sup_{\Omega_n}|\int \phi'_{W_i}(W_iX)X_kW_jXd(P_n-P)| \;=\; o_P(1). \tag{36}$$

From (33)-(36) we have

$$sup_{\Omega_n}|\int \hat{\phi}'_{W_i}(W_iX)X_kW_jXdP_n(X)-\int \phi'_{W_i}(W_iX)X_kW_jXdP| \;=\; o_P(1). \tag{37}$$

By continuity,

$$sup_{\Omega_n}|\int \phi'_{W_i}(W_iX)X_kW_jXdP-\phi'_{W_{Pi}}(W_{Pi}X)X_kW_{Pj}XdP| \;=\; o(1). \tag{38}$$

Then [4] follows from (37)&(38).

In [5],

$$(LHS)_k \;\leq\; sup_{\Omega_n}||(\frac{\partial}{\partial W_i}\gamma_n(W_i))_k||_2 sup_{\Omega_n}||\int \mathbf{B}_n^{(i)}(W_iX)W_jXdP_n||_2. \tag{39}$$

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

$$sup_{\Omega_n}||(\frac{\partial}{\partial W_i}\gamma_n(W_i))_k||_2\varepsilon_n\delta_n^{-1} \;=\; o_P(1). \tag{40}$$

By taking partial derivatives,

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

and

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

Then

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

Now by Lemma 2-5 in Appendix A, 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 $\varepsilon_n\delta_n^{-\frac{9}{2}}n_i^{\frac{1}{2}} = o(1)$ in Condition [iv], (40) holds. Thus we have in [5] $(LHS)_k = o_P(1)$ for $k = 1, \cdots, m$. ■

ACKNOWLEDGEMENT

The authors would like to thank Prof. Charles Stone for helpful technical discussions and thank Sabrina Soracco for helpful comments on editing.

APPENDIX

Some lemmas and their proofs used in Proposition 1-3. In this section, we provide and prove all the lemmas used in the proof of Propositon 1-3. Note that for each ϕ_k, we has constructed a sequence of sieves $\mathcal{G}_n^{(k)}$ using $\mathbf{B}_n^{(k)} = (B_{n0}^{(k)}, \cdots, B_{nn_k}^{(k)})^T$ and a class of estimates $\hat{\phi}_{W_k} \in \mathcal{G}_n^{(k)}$ and $\overline{\hat{\phi}}_{W_k} \in \mathcal{G}_n^{(k)}$ for ϕ_{W_k} with $W \in \Omega_n$according to (13) and (14), where ϕ_{W_k} is given in (3). By assumption, $S = W_P X$, where $S = (S_1, \cdots, S_m)^T$ has independent components and $X = (X_1, \cdots, X_m)^T$. In the following c will denote a constant (only dependent on the population law P), but its exact value may vary in different places (even in a line) without clarifying.

If $x \in \mathcal{R}^m$ is a column vector, $|x| \equiv ||x||_2 = \sqrt{x^T x}$.

If A is a $m \times m$ real matrix, $||A||_1 = max_{1 \le i \le m} ||A_i||_2$, $||A||_2 = max_{x \in R^m, |x|=1} |Ax|$, $|A| = \sqrt{tr(A^T A)}$.

Let $\Omega_n^{(k)} = \{W_k : W \in \Omega_n\}$ for $k = 1, \cdots, m$.

For $w \in \Omega_n^{(k)}$, recall the definition of $\hat{\phi}_w = [A_n^{(k)}(w)^{-1} D_n^{(k)}(w)]' \mathbf{B}_n^{(k)}$ and $\overline{\hat{\phi}}_w = [A^{(k)}(w)^{-1} D^{(k)}(w)]' \mathbf{B}_n^{(k)}$, where $A_n^{(k)}(w) = \int [\mathbf{B}_n^{(k)}(wX)][\mathbf{B}_n^{(k)}(wX)]^T dP_n$, $A^{(k)}(w) = \int [\mathbf{B}_n^{(k)}(wX)][\mathbf{B}_n^{(k)}(wX)]^T dP$, $D_n^{(k)}(w) = \int (\mathbf{B}_n^{(k)})'(wX) dP_n$ and $D^{(k)}(w) = \int (\mathbf{B}_n^{(k)})'(wX) dP$. In the following, we often drop the superscript k and the argument w in $\mathbf{B}_n^{(k)}, A_n^{(k)}(w)$, $D_n^{(k)}(w), A^{(k)}(w)$,$D^{(k)}(w)$ whenever possible.

The following Lemma 1-10 hold under the conditions of Theorem 1.

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

Proof. Remember that $min_{t \in [\underline{b}_{nk}, \bar{b}_{nk}]} r_k(t) \ge c\delta_n$. For any $w \in \Omega_n^{(k)}$, $|w - W_{Pk}| \le \varepsilon_n$. Let $v = w W_P^{-1}$, then $|v_j| \to 0$ for $1 \le j \ne k \le m$ and $|v_k - 1| \to 0$ as $n \to \infty$. Fix a $t \in [\underline{b}_{nk}, \overline{b}_{nk}]$.

Since $f_w(t) = E[\frac{1}{v_k} r_k(\frac{t - \sum_{j \ne k} v_j S_j}{v_k})]$, consider the right hand side as a function (say h) of v. By the first order Taylor expansion,

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

where by direct calculation, $|\frac{\partial}{\partial v_j} h(v)|$ is uniformly bounded with $w \in \Omega_n^{(k)}$. Thus $sup_{w \in \Omega_n^{(k)}} min_{t \in [\underline{b}_{nk}, \bar{b}_{nk}]} f_w(t) \ge c\delta_n$ 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

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

■

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. Notice that $D(w) = (D_0(w), \cdots, D_{n_k}(w))^T$and by direct calculation we have

$$\begin{aligned} |D_i(w)| &= |\int_0^1 t(f_w(\xi_i + \delta_n t) - f_w(\xi_{i+1} + \delta_n t) + f_w(\xi_{i+2} - \delta_n t) - f_w(\xi_{i+3} - \delta_n t))dt| \\ &\leq |f'_w|_\infty \delta_n \end{aligned}$$

where the inequality is by the mean value theorem. So the first result holds. 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.

$$\begin{aligned} Pr(sup_{w\in\Omega_n^{(k)}}||D_n(w) - D(w)||_2 \geq t) &= Pr(sup_{W\in\Omega_n^{(k)}}||\int \mathbf{B}'_n(wX)d(P_n - P)||_2 \geq t) \\ &\leq \sum_{i=0}^{n_k} Pr(sup_{w\in\Omega_n^{(k)}}|\int B'_{n,i}(wX)d(P_n - P)| \geq \frac{t}{\sqrt{n_k + 1}}). \end{aligned}$$

By calculating *the generalized bracketing entropy* [see, e.g., Van de Geer (2000)] with the facts that $|B'_{n,i}| \leq \delta_n^{-1}, |B''_{n,i}| \leq \delta_n^{-2}$,

$$\mathcal{H}_{B,\delta_n^{-1}}(u, \{B'_{n,i}(w\mathbf{x}) : w \in \Omega_n^{(k)}\}, P) \leq cm\log(\varepsilon_n\delta_n^{-2}/u), 0 < u < \varepsilon_n\delta_n^{-2}.$$

Then using Theorem 5.11 in Van Der Geer (2000) , we have for $\sqrt{\frac{n_k}{n\delta_n}} \le t \le \sqrt{n_k}$,

$$Pr(sup_{w\in\Omega}|\int B'_{n,i}(wX)d(P_n - P)| \;\; \geq \;\; t/\sqrt{n_k+1}) \le \exp(-cnt^2\delta_n/n_k).$$

So $Pr(sup_{w\in\Omega_n^{(k)}}||D_n(w) - D(w)||_2 \ge t) \le (n_k+1)\exp(-cnt^2\delta_n/n_k)$.

Thus

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

Similarly we get

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

(Notice the fact that at most 7 elements in each row of $A_n - A$ are nonzero).■

LEMMA 4. $sup_{w\in\Omega_n^{(k)}}|D_n(w)| = O_p(\delta_n\sqrt{n_k})$, $sup_{w\in\Omega_n^{(k)}}|A_n^{-1}(w)D_n(w)| = O_p(\sqrt{n_k}/\delta_n)$, and $sup_{\Omega_n^{(k)}}||A_n^{-1}(w)D_n(w) - A^{-1}(w)D(w)||_2 = \delta_n^{-\frac{5}{2}}\sqrt{\frac{n_k\log n_k}{n}}O_P(1)$.

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

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

then

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

(Here we use the inequality of matrix norm $||(I+N)^{-1}||_2 \le (1 - ||N||_2)^{-1}$for a square matrix N with $||N||_2 < 1$, 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

$$\begin{aligned}
& sup_{w\in\Omega_n^{(k)}}||A_n^{-1}D_n - A^{-1}D||_2 \\
\le \;\; & sup_{w\in\Omega_n^{(k)}}||A^{-1}(D_n - D) - A_n^{-1}(A_n - A)A^{-1}D_n||_2 \\
\le \;\; & sup_{\Omega_n^{(k)}}\{||A^{-1}||_2||D_n - D||_2\} + sup_{\Omega_n^{(k)}}\{||A_n - A||_2||D_n||_2||A^{-1}||_2^2\}(1 + o_p(1)) \\
= \;\; & O_p(\delta_n^{-2})O_p(\sqrt{\frac{n_k\log n_k}{n\delta_n}}) + O_p(\sqrt{\frac{\delta_n\log n_k}{n}})O_p(\delta_n\sqrt{n_k})O_p(\delta_n^{-4}) \\
= \;\; & O_p(\delta_n^{-\frac{5}{2}}\sqrt{\frac{n_k\log n_k}{n}}).
\end{aligned}$$

■

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 $\mathbf{B}_n^{(i)}$)

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

has less than $7n_i$ nonzero elements since $[\mathbf{B}_n\mathbf{B}_n'^T]_{jl} = 0$ for $0 \le j,l \le n_i$ with $|j-l| > 3$. By the Cauchy-Schwartz inequality,

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

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

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

Further from Lemma 1, we have $sup_{w\in\Omega_n^{(i)}}\int(B_{nj}B'_{nl} + B'_{nj}B_{nl})^2(wX)dP \le c\delta_n^{-1}$.

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

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

$$\begin{aligned}
||\frac{\partial}{\partial w_k}D_n(w)||_2 &\le \sqrt{\int|X_k|^2 dP_n}\sqrt{\int\sum_{l=0}^{n_k}(B''_{nl})^2(wX)dP_n} \\
&\le \sqrt{\int|X_k|^2 dP_n}\sqrt{\int 4\delta_n^{-4}dP_n} \\
&= 2\delta_n^{-2}\sqrt{\int|X_k|^2 dP_n}.
\end{aligned}$$

■

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

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

$$\begin{aligned} E(||\int \mathbf{B}_n(S_i)S_j dP_n||_2^2) &= E(\sum_{k=0}^{n_i}(\int B_{nk}(S_i)S_j dP_n)^2) \\ &= \sum_{k=0}^{n_i}\frac{1}{n^2}E[\sum_{l=1}^{n}(B_{nk}(S_i^{(l)})S_j^{(l)})^2 \\ &\quad + \sum_{1\le l_1\neq l_2\le n}(B_{nk}(S_i^{(l_1)})S_j^{(l_1)})(B_{nk}(S_i^{(l_2)})S_j^{(l_2)})] \\ &= \sum_{k=0}^{n_i}\frac{1}{n}E(B_{nk}(S_i)S_j)^2 \\ &= \frac{1}{n}E(\sum_{k=0}^{n_i}B_{nk}(S_i)^2S_j^2) \\ &\le \frac{3}{n}E(S_j^2). \end{aligned}$$

■

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

Proof.

$$\begin{aligned} \int \mathbf{B}_n(W_iX)W_jXdP_n &= \int \mathbf{B}_n(W_{Pi}X)W_{Pj}XdP_n + \int \mathbf{B}_n(W_iX)(W_j - W_{Pj})XdP_n \\ &\quad + \int [\mathbf{B}_n(W_iX) - \mathbf{B}_n(W_{Pi}X)]W_{Pj}XdP_n. \end{aligned}$$

By Lemma 6, $||\int \mathbf{B}_n(W_{Pi}X)W_{Pj}XdP_n||_2 = O_P(n^{-1/2})$. And

$$\begin{aligned} ||\int \mathbf{B}_n(W_iX)(W_j - W_{Pj})XdP_n||_2 &\le ||\int \mathbf{B}_n(W_iX)|X|dP_n||_2|W_j - W_{Pj}| \\ &\le \varepsilon_n\sqrt{\sum_{k=0}^{n_i}(\int B_{nk}(W_iX)|X|dP_n)^2} \\ &\le \varepsilon_n\sqrt{\int |X|^2dP_n\int\sum_{k=0}^{n_i}B_{nk}^2(W_iX)dP_n} \\ &\le \varepsilon_n\sqrt{3\int |X|^2dP_n}. \end{aligned}$$

Further,

$$||\int [\mathbf{B}_n(W_iX) - \mathbf{B}_n(W_{Pi}X)]W_{Pj}XdP_n||_2$$

$$
\begin{aligned}
&\leq \; || \int [\mathbf{B}_n(W_iX) - \mathbf{B}_n(W_{Pi}X)]W_{Pj}XdP_n||_2 \\
&= \; \sqrt{\sum_{k=0}^{n_i}(\int [B_{nk}(W_iX) - B_{nk}(W_{Pi}X)]W_{Pj}XdP_n)^2} \\
&\leq \; \sqrt{\int |W_{Pj}X|^2dP_n \int \sum_{k=0}^{n_i}(B_{nk}(W_iX) - B_{nk}(W_{Pi}X))^2dP_n} \\
&\leq \; \delta_n^{-1}\varepsilon_n||W_{Pj}X||_{2,P_n}\sqrt{3\int |X|^2dP_n}.
\end{aligned}
$$

Thus

$$
sup_{\Omega_n}|| \int \mathbf{B}_n(W_iX)W_jXdP_n||_2 \;\; = \;\; O_p(n^{-1/2}) + O_P(\varepsilon_n) + O_P(\delta_n^{-1}\varepsilon_n).
$$

■

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

Proof. Since for any $h \in \mathcal{G}_n$,

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

then

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

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)],

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

■

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 $a \in \mathcal{R}^{n_i+1}$,

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

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

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

and

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

So

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

Together with Lemma 8, we have

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

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

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

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

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

and

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

So

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

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

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

Thus

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

■

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

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

First, (dropping W_{Pk} in $A_n(W_{Pk}), D_n(W_{Pk}), A(W_{Pk}), D(W_{Pk})$), 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

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

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

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. (2003). Efficient independent component analysis. *Technical report* #634, Department of Statistics, University of California, Berkeley.

[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, Carl (1978). A practical guide to splines. Springer-Verlag.

[14] Eriksson, J., Kankainen, A. and Koivunen, V. (2001). Novel characteristic function based criteria for ICA. *Proceedings of 3rd International Conference on Independent Component Analysis and Signal Separation*. San Diego, California.

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

[28] 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.

[29] Samarov, A. and Tsybakov, A. (2002). Nonparametric independent component analysis. *Submitted to Bernoulli*.

[30] Van der Geer, S. (2000). *Applications of Empirical Process Theory*. Cambridge University Proess, UK.

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

[32] 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.