Supplement to "Consistent Independent Component Analysis and Prewhitening"*

Aiyou Chen, Peter Bickel

{aychen, bickel}@stat.berkeley.edu

March 2004

Technical Report #656

Department of Statistics, University of California, Berkeley

Abstract

In this paper we study the statistical properties of a characteristic-function based algorithm for independent component analysis (ICA), which was proposed by Eriksson et. al. (2003) and Chen & Bickel (2003) independently. First, statistical consistency of this algorithm with prewhitening is analyzed, especially under heavy-tailed sources. Second, without prewhitening this algorithm is shown to be robust against small additive noise. Finally, $\sqrt{n}$ consistency and asymptotic normality of this method are also established.

1 Introduction to independent component analysis

Suppose that a $m \times 1$ random vector $\mathbf{X}$ can be modeled by

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

where A is a $m \times m$ nonsingular matrix and $\mathbf{S} = (S_1, \cdots, S_m)^T$ is a $m \times 1$ random vector with mutually independent components. Given n independent copies of $\mathbf{X}$, say $\{\mathbf{X}(i) : 1 \leq i \leq n\}$, the objective is to estimate $\mathrm{W} = \mathrm{A}^{-1}$ and to recover each component of $\mathbf{S}$ (or their distributions). This is called *independent component analysis* (ICA), a typical blind source separation problem (Hyvarinen et. al. 2001). The $\mathbf{X}$ is usually called mixed signals obtained by multi-channel sensors, the components of $\mathbf{S}$ are hidden sources of interest and the A is called the mixing matrix. When $\mathbf{S}$ has at most one Gaussian component, W (called the unmixing matrix) is identifiable up to ambiguity of scale and order (Comon, 1994). There has been lots of work in estimating the unmixing matrix W (see for example, Hyvarinen et.

*Supported by NSF Grant DMS-01-04075.

al. 2001 and references therein). Recently a new estimator of the unmixing matrix by using characteristic function (CHFICA) has been proposed and proved to be consistent under the identifiability conditions and to be $\sqrt{n}$-consistent under second moment conditions (Chen & Bickel, 2003, Eriksson et. al. 2003). Unfortunately, the implementation of CHFICA requires $O(n^2)$ operations, which is computationally impossible for a large sample size. Chen & Bickel (2004) proposed a novel algorithm to implement the CHFICA method by using prewhitening and incomplete Cholesky decomposition, which is computationally favorable. In this paper, we study the statistical properties of the prewhitened CHFICA method. We also analyze the robustness of CHFICA with respect to small additive noise.

In the following, we introduce the prewhitened CHFICA method (PCFICA) in Section 2; the consistency property of PCFICA is established in Section 3; and Section 4 contains the analysis of robustness of CHFICA in existence of small additive noise. Some technical details are included in the Appendices.

Notations: for a matrix W, we use W_k, W^k and W_{kj} to denote its kth row, kth column and (k, j)th entry separately. $\mathcal{O}_m$ denotes the set of all $m \times m$ orthogonal matrices and i denotes the square root of -1. $I_{m\times m}$ denotes the $m \times m$ identity matrix and we often omit the subscript.

2 Prewhitened characteristic-function based ICA method

To reduce the ambiguity of identifiability, we need some constraints on the scale and order either on the components of $\mathbf{S}$ or on the rows of W. Without loss of generality, we may focus on W and assume that :

(I). each row of W is normalized;

(II). the element with maximal modulus in each row is positive;

(III). the rows are sorted according to the partial order "$\prec$" (that is, for $\forall a, b \in R^m$, $a \prec b$ if and only if there exists $k \in \{1, \cdots, m\}$ such that $a_k < b_k$ and $a_j = b_j$ for $\forall j < k$).

We call Ω as the set of matrices which satisfy the above conditions (I)-(III). It is clear that on Ω, the unmixing matrix can be identified uniquely. Further, for any nonsingular $m \times m$ matrix, by rescaling and permuting rows appropriately, the transformed matrix will belong to Ω. In the following we will denote such rescaling-permuting transformation by $[\cdot]_\Omega$. Let P be the law of $\mathbf{X}$ under model (1) and $\mathrm{W}_P \in \Omega$ be the true unmixing matrix of interest.

2.1 Introduction to prewhitening

Most ICA algorithms in estimating W_P can be unified under the framework of minimizing some contrast function with respect to (w.r.t) the unmixing matrix with inputs $\{\mathbf{X}(i) : 1 \leq i \leq n\}$. Since the unmixing matrix for model (1) can be essentially arbitrary, naively we have to optimize some contrast function, over all $m \times m$ matrices to obtain an estimate. But prewhitening can project the optimization onto the Stiefel manifold of orthogonal matrices. Optimization on a Stiefel manifold can be solved efficiently (Edelman et. al. 1999). Let $\Sigma_{\mathbf{x}}$ be the covariance matrix of $\mathbf{X}$ and let $\Sigma_{\mathbf{x}}^{\frac{1}{2}}$ be the square root matrix of $\Sigma_{\mathbf{x}}$ obtained by Singular Value Decomposition (SVD), i.e., let $\Sigma_{\mathbf{x}} = \mathbf{UDU^T}$ be the SVD decomposition such that $\mathbf{UU^T} = I_{m\times m}$, an identity matrix, and $\mathbf{D}$ is a diagonal matrix comprised of $\Sigma_{\mathbf{x}}$'s eigenvalues, then $\Sigma_{\mathbf{x}}^{\frac{1}{2}} = \mathbf{UD}^{\frac{1}{2}}\mathbf{U^T}$. Let $\mathbf{Y} = \Sigma_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{X}$. Then $\text{cov}(\mathbf{Y}) = I$ and (1) is equivalent to $\mathbf{S} = \mathrm{W}\Sigma_{\mathbf{x}}^{\frac{1}{2}}\mathbf{Y}$. Without loss of generality, we may assume that each hidden source in $\mathbf{S}$ has unitary variance (so far assume each source has finite variance). By considering the covariance matrix, we have $\mathrm{W}\Sigma_{\mathbf{x}}\mathrm{W}^T = I_{m\times m}$ and thus

$$\mathrm{O} \quad = \quad \mathrm{W}\Sigma_{\mathbf{x}}^{\frac{1}{2}} \tag{2}$$

must be an orthogonal matrix. Notice that $\mathbf{Y} = \mathrm{O}^T\mathbf{S}$, so it is still an ICA model but restricting to orthogonal matrices is computationally very advantageous. Since $\Sigma_{\mathbf{x}}$ can be estimated directly by the sample covariance matrix of $\mathbf{X}$

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

where $\overline{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{X}(k)$, thus a prewhitened ICA algorithm first estimates an orthogonal unmixing matrix $\mathbf{O}$ (say $\hat{\mathbf{O}}$) by fitting the ICA model with inputs $\hat{\mathbf{Y}}(j) = \hat{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{X}(j)$ in some way, and then estimate W by $\hat{\mathrm{W}} = \hat{\mathrm{O}}\hat{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}$ because of (2). This is the so-called prewhitening technique.

2.2 Prewhitened characteristic-function based ICA method

An estimate of W_P, based on empirical characteristic function (e.c.f.), called CHFICA, has been shown (Chen & Bickel, 2003) to be consistent under identifiability conditions. It is defined by

$$\hat{\mathrm{W}} \quad = \quad argmin_{\mathrm{W}\in\Omega}\hat{\Delta}_{\mathbf{x}}(\mathrm{W}), \tag{3}$$

where

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

Here $\hat{c}_{\xi}(\cdot)$ stands for the e.c.f. of a random variable/vector ξ and $\lambda_m(\mathbf{t})$ is an m-dim probability density function. In the following we fix λ as the probability density function of $\mathcal{N}(0,1)$ and define $\lambda_m(\mathbf{t}) = \prod_{j=1}^{m}\lambda(t_j)$ for $\mathbf{t} = (t_1,\cdots,t_m)^T \in R^m$. For simplicity we often write λ_m as λ whenever it does not cause confusion.

For practical use, it is more convenient to work on *prewhitened CHFICA* (PCFICA). The procedure is as follows. First estimate O by

$$\hat{\mathbf{O}} \quad = \quad arg\min_{O\in\mathcal{O}(m)}\hat{\Delta}_{\hat{\mathbf{Y}}}(O), \tag{5}$$

where $\hat{\Delta}_{\hat{\mathbf{Y}}}$ is the same as $\hat{\Delta}_{\lambda}$ defined in (4) except that $\{\mathbf{X}(j)\}$ is now replaced by $\{\hat{\mathbf{Y}}(j)\}$; Second let

$$\hat{\mathrm{W}} \quad = \quad \hat{\mathbf{O}}\hat{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}, \tag{6}$$

and obtain an estimate of W_P by $[\hat{\mathrm{W}}]_\Omega$. By applying the technique of incomplete Cholesky decomposition, $\hat{\Delta}_{\hat{\mathbf{Y}}}(O)$ can be approximated efficiently (Chen & Bickel, 2004).

3 Consistency of Prewhitened CHFICA

It is known that the acting parameter space of any prewhitened ICA algorithm can be used to approximate the unmixing matrix without assuming finite second moments on hidden sources (Chen & Bickel, 2004). And it was conjectured there that PCFICA could provide consistent estimates for the unmixing matrix under certain conditions. We have the following result.

THEOREM 1. *Suppose that the model (1) holds and at most one of* $\mathbf{S}$*'s components is Gaussian. Let* $\mathrm{W}_P \in \Omega$ *be the true unmixing matrix under* P*. Then the estimator defined in (5) and (6) is consistent, i.e.,* $||[\hat{\mathrm{W}}]_\Omega - \mathrm{W}_P|| \to_P 0$*, in the following cases :*

[1]. $m \geq 2$*, all components of* $\mathbf{S}$ *have finite second moments;*

[2]. $m \geq 2$*, all components of* $\mathbf{S}$ *except one have finite second moments;*

[3]. $m = 2$*, both sources have infinite variances and have stable distributions with exponents* $\alpha_j \in (0, 2)$ $(j = 1, 2)$*.*

PROOF. Let $\hat{\Sigma}_{\mathbf{s}}$ be the sample covariance matrix of $\mathbf{S}$. Let $\hat{\sigma}_j = (\hat{\Sigma}_{\mathbf{s}}^{\frac{1}{2}})_{jj}$ be the sample standard deviation of S_j $(j = 1, \cdots, m)$ and let $\hat{I} = \hat{\Sigma}_{\mathbf{s}}^{-\frac{1}{2}} \mathrm{diag}\{\hat{\sigma}_j\}$. By Theorem 5 of Chen & Bickel (2004), we have

$$\hat{I} \quad \to_P \quad I. \tag{7}$$

From (1) we have $\hat{\Sigma}_{\mathbf{x}} = \mathrm{A}\hat{\Sigma}_{\mathbf{s}}\mathrm{A}^T$. Let $\mathbf{O}_n = \hat{\Sigma}_{\mathbf{s}}^{-\frac{1}{2}} \mathrm{A}^{-1} \hat{\Sigma}_{\mathbf{x}}^{\frac{1}{2}}$. Then $\mathbf{O}_n \mathbf{O}_n^T = I$ and

$$[\mathbf{O}_n \hat{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}]_\Omega \quad = \quad [\hat{\Sigma}_{\mathbf{s}}^{-\frac{1}{2}} \mathrm{W}_P]_\Omega = [\hat{I}^T \mathrm{W}_P]_\Omega \to_P \mathrm{W}_P.$$

Recall the definition of $\hat{\mathrm{W}}$ given in (6), thus to show $[\hat{\mathrm{W}}]_\Omega \to_p \mathrm{W}_p$, it is sufficient to show that

$$\hat{\mathbf{O}}\mathbf{O}_n^{-1} \quad \to_p \quad I \text{ (or its permutation).}$$

Define $\tilde{S}_j = S_j / \hat{\sigma}_j$ and $\tilde{\mathbf{S}} = (\tilde{S}_1, \cdots, \tilde{S}_m)^T$. Recall that $\hat{\mathbf{Y}} = \hat{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}} \mathbf{X}$ and $\mathbf{X} = \mathrm{A}\mathbf{S}$, then for $O \in \mathcal{O}_m$ we may write

$$O\hat{\mathbf{Y}} \quad = \quad O\hat{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}} \mathrm{A}\mathbf{S} = O\mathbf{O}_n^{-1} \hat{\Sigma}_{\mathbf{s}}^{-\frac{1}{2}} \mathbf{S} = O\mathbf{O}_n^{-1} \hat{I} \tilde{\mathbf{S}}.$$

Define

$$\tilde{\Delta}_s(O) \quad = \quad \int |\hat{c}_s(t^T O \hat{\Sigma}_s^{-1/2}) - \prod_{j=1}^{m} \hat{c}_s(t_j O_j \hat{\Sigma}_s^{-1/2})|^2 \lambda(t) dt, \tag{8}$$

where $\hat{c}_s$ is the e.c.f. of $\mathbf{S}$. It is obvious that $\tilde{\Delta}_s(O\mathbf{O}_n^{-1}) = \hat{\Delta}_{\hat{\mathbf{Y}}}(O)$ and $O\mathbf{O}_n^{-1}$ is orthogonal for $O \in \mathcal{O}_m$. Let $\tilde{\mathbf{O}} = \hat{\mathbf{O}}\mathbf{O}_n^{-1}$ and then

$$\tilde{\mathbf{O}} \quad = \quad \text{argmin}_{\mathcal{O}_m} \tilde{\Delta}_s(O).$$

It is sufficient to show that $\tilde{\mathbf{O}}$ (or $[\tilde{\mathbf{O}}]_\Omega$) converges in probability to I.

Case [1] : all of $\mathbf{S}$'s components have finite variances.

After algebraic expansion, $\int |\hat{c}_s(t^T O) - \prod_{j=1}^m \hat{c}_s(t_j O_j)|^2 \lambda_m(\mathbf{t}) d\mathbf{t}$ can be expressed as a bounded U-statistic whose kernel function is continuous w.r.t O. Since $\mathcal{O}_m$ is compact and $\hat{\Sigma}_s^{-\frac{1}{2}} \to_P \Sigma_s^{-\frac{1}{2}}$, by Uniform Law of Large Numbers (ULLN) for U-processes (Arcones & Gine, 1993), we have $\sup_{O \in \mathcal{O}_m} |\tilde{\Delta}_s(O) - \Delta_s(O\hat{I})| = o_P(1)$, where $\hat{I} = \hat{\Sigma}_s^{-\frac{1}{2}} \Sigma_s^{\frac{1}{2}}$ and

$$\Delta_s(O) \quad = \quad \int |c_s(t^T O \Sigma_s^{-\frac{1}{2}}) - \prod_{j=1}^m c_s(t_j O_j \Sigma_s^{-\frac{1}{2}})|^2 \lambda_m(\mathbf{t}) d\mathbf{t}. \tag{9}$$

Notice that $\Delta_s(O)$ is uniformly continuous w.r.t. O on $\mathcal{O}_m$ (compact) and $\hat{I} \to_P I$, then $\sup_{O \in \mathcal{O}_m} |\Delta_s(O\hat{I}) - \Delta_s(O)| = o_P(1)$. Thus $\sup_{O \in \mathcal{O}_m} |\tilde{\Delta}_s(O) - \Delta_s(O)| = o_P(1)$. Since $\Delta_s(O) \geq 0$ and the equality holds if and only if $O\mathbf{S}$ has mutually independent components, which by Comon (1994) implies that O must be a rescaled permutation matrix, i.e. $[O]_\Omega = I$. By compactness and continuity we have that as the minimizer of $\hat{\Delta}_s(O)$ over $O \in \mathcal{O}_m$, $[\tilde{\mathbf{O}}]_\Omega \to_P I$.

Case [2] : One of $\mathbf{S}$'s components has infinite variance and all others have finite variances. W.L.O.G, let $E|S_m|^2 = \infty$.

Following the arguments in Case [1] except that $(\Sigma_s^{-\frac{1}{2}})_{m,m} = 0$, we have $\sup_{O \in \mathcal{O}_m} |\tilde{\Delta}_s(O) - \Delta_s(O)| = o_P(1)$. Again $\Delta_s(O) \geq 0$ and it is enough to show that for some O orthogonal, $\Delta_s(O) = 0$ if and only if O is a permutation matrix. Let σ_j be the standard deviation of S_j and let $\xi_j = \frac{S_j}{\sigma_j}$ $(j = 1, \cdots, m)$. Then $\Sigma_s^{-\frac{1}{2}}\mathbf{S} = [\xi_1, \cdots, \xi_{m-1}, 0]^T$.

Write $O = [O1, O2]$, where O_1 contains the first $m-1$ columns of O and $O2$ contains the last column of O. It is clear that $\Delta_s(O) = 0$ if and only if all the components of $O\Sigma_s^{-\frac{1}{2}}\mathbf{S} = O1\xi_{1:m-1}$ are mutually independent. Since $\xi_{1:m-1}$ has at most one Gaussian component, any $m-1$ components of $O1\xi_{m-1}$ are still mutually independent and thus by Comon's (1994) Theorem 1, any matrix comprised of $m-1$ rows of $O1$ must be a rescaled permutation matrix (or degenerated). Hence $O1$ must be a sub-matrix of a $m \times m$ rescaled permutation matrix and O must be a permutation matrix due to its orthogonality.

Case [3] : $m = 2$, both sources have infinite variances and have stable distributions with exponents $\alpha_j \in (0, 1)$ for $j = 1, 2$.

Notice that both $\sup_{\mathcal{O}_m} |\tilde{\Delta}_s(O) - \Delta_s(O)| = o_P(1)$ and $\Delta_s(O) \equiv 0$, thus the arguments in Case [1] and Case [2] do not apply here any more. Instead of using $\Delta_s(O)$, we define a new intermediate function $\Delta_{\tilde{s}}(O)$ as follows :

$$\Delta_{\tilde{s}}(O) \quad = \quad \int |c_s(t^T O \hat{\Sigma}_s^{-\frac{1}{2}}) - \prod_{j=1}^m c_s(t_j O_j \hat{\Sigma}_s^{-\frac{1}{2}})|^2 \lambda(\mathbf{t}) d\mathbf{t}. \tag{10}$$

Let ϕ_j be the c.f. of $\mathbf{S}_j$, for $(x,y) \in R^2$ and $t > 0$,

$$F_j(t;x,y) = \phi_j((x+y)t^{1/\alpha_j}) - \phi_j(xt^{1/\alpha_j})\phi_j(yt^{1/\alpha_j}) \tag{11}$$

and define $F_j'(t;x,y) = \frac{\partial}{\partial t}F_j(t;x,y)$. Then by Proposition 1 and Proposition 2 in Appendix A, we have for $d_n = 1/(\sum_{j=1}^2 \hat{\sigma}_j^{-\alpha_j})^2$ and $h_{jn} = \hat{\sigma}_j^{-\alpha_j}\sqrt{d_n}$ $(j=1,2)$:

(Identifiability) : $\sup_{O\in\mathcal{O}_2} |d_n\Delta_{\tilde{s}}(O) - g(O,h_{1n},h_{2n})| = o_p(1)$, where g is defined by

$$g(O;h_{1n},h_{2n}) = \int |F_1'(0;t_1O_{11},t_2O_{21})h_{2n} + F_2'(0;t_1O_{12},t_2O_{22})h_{1n}|^2\lambda(\mathbf{t})d\mathbf{t}.$$

(Convergence) : $d_n\sup_{O\in\mathcal{O}_m} |\tilde{\Delta}_s(O) - \Delta_{\tilde{s}}(O)| = o_P(1)$.

Let $E_n = d_n\sup_{O\in\mathcal{O}_m} |\tilde{\Delta}_s(O) - \Delta_{\tilde{s}}(O)| = o_P(1)$, $e_n = \sup_{O\in\mathcal{O}_2} |d_n\Delta_{\tilde{s}}(O) - g(O,h_{n1},h_{n2})|$. Then $E_n = o_p(1)$ and $e_n = o_p(1)$.

Let $\mathcal{I}_\Omega = \{O\in\mathcal{O}_2 : [O]_\Omega = I\}$. It is obvious that $\mathcal{I}_\Omega$ has 8 elements. For $\forall\ 1 > \epsilon > 0$, let $\mathcal{O}_\epsilon = \{O\in\mathcal{O}_2 : ||O-\mathcal{I}_\Omega|| \geq \epsilon\}$. Here $||O-\mathcal{I}_\Omega|| = \min\{||O-Q|| : Q\in\mathcal{I}_\Omega\}$. Notice that $\tilde{\Delta}_{\mathbf{s}}(O)$ is invariant to the row permutation of O, then by (Identifiability) and (Convergence) we have

$$\begin{aligned}
P(||\tilde{O}-\mathcal{I}_\Omega|| \geq \epsilon) &\leq P(\tilde{\Delta}_s(I) \geq \sup_{O\in\mathcal{O}_\epsilon}\tilde{\Delta}_s(O)) \\
&\leq P(d_n\Delta_{\tilde{s}}(I) + E_n \geq \sup_{O\in\mathcal{O}_\epsilon} d_n\Delta_{\tilde{s}}(O)) - E_n) \\
&\leq P(2E_n + 2e_n \geq \sup_{O\in\mathcal{O}_\epsilon} g(O,h_{n1},h_{n2})) \\
&\leq P(2E_n + 2e_n \geq \inf_{O\in\mathcal{O}_\epsilon,(h_1,h_2)\in\mathcal{H}} g(O,h_1,h_2)).
\end{aligned}$$

By continuity and compactness there exists $O^{(0)} \in \mathcal{O}_\epsilon$ and $(h_1^{(0)},h_2^{(0)}) \in \mathcal{H} \equiv \{(h_1,h_2) \in [0,1]^2 : h_1+h_2 = 1\}$ such that

$$\inf_{O\in\mathcal{O}_\epsilon,(h_1,h_2)\in\mathcal{H}} g(O,h_1,h_2) = g(O^{(0)},h_1^{(0)},h_2^{(0)}).$$

Since $O^{(0)} \notin \mathcal{I}_\Omega$, thus no entry of $O^{(0)}$ is zero. By Lemma 1 in Appendix B, $g(O^{(0)},h_1^{(0)},h_2^{(0)}) > 0$. Thus $P(2E_n + 2e_n \geq g(O^{(0)},h_1^{(0)},h_2^{(0)})) \to 0$ and $P(||\tilde{O}-\mathcal{I}_\Omega\}|| \geq \epsilon) \to 0$. This concludes Case [3] in Theorem 1.

4 Robustness of the Characteristic-Function based ICA method

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

$$\mathbf{X} = \mathrm{A}\mathbf{S} + r\mathbf{n}, \tag{12}$$

where A, $\mathbf{S}$ are the same as in the previous sections, $\mathbf{n}$ is a $m \times 1$ random vector, independent of $\mathbf{S}$, standing for an additive noise vector (for example, sensor noise), and r is the magnitude of additive noise. This is usually called the noisy ICA model. Hyvarinen et. al. (2001) have a good review of studies of this type of models. Our objective here is to study how the c.f.-based ICA method behaves in presence of noise. We borrow Bickel and Doksum (1981)'s setup in their study of robustness of Box-Cox transformations. To be more precise, we assume a large sample size n, and further $r = r(n) \to 0$ as $n \uparrow \infty$, an interesting question is how the additive noise affects the performance of the estimator $\hat{\mathrm{W}}$ of W_P defined in (3) in the absence of additive noise. Intuitively, when the noise is too small, the estimate should be close to the true value. Our following theorem confirms and goes beyond this conjecture a little bit.

THEOREM 2. *Let $\hat{\mathrm{W}}$ be given by (3). Suppose that W_P is nonsingular and $\mathbf{S}$ has at most one Gaussian component. If $r(n) = o(1)$. Then*
(i). $||\hat{\mathrm{W}} - \mathrm{W}_P|| = o_P(1)$.
(ii). If further $E||\mathbf{S}||^2 < \infty$ and $E||\mathbf{n}||^2 < \infty$, then

$$||\hat{\mathrm{W}} - \mathrm{W}_P|| \;\; = \;\; O_P(n^{-\frac{1}{2}} + r(n)). \tag{13}$$

The complete proof of Theorem 2 (i) and (ii) are provided in Appendix C and D, separately.

Theorem 2 says that even if there is small additive noise, the c.f.-based ICA methods can provide fairly good estimates for the unmixing system. Thus the c.f.-based ICA method can serve as a good starting point for further separation enhancement.

From Theorem 2, we have the following corollary.

COROLLARY 3. *Suppose that the conditions of Theorem 2 (ii) hold except that $r(n) = 0$. Then $\sqrt{n}(\hat{W}W_P^{-1} - I)$ is asymptotically normal, with variance-covariance matrix Σ_P given in Appendix E.*

The calculation of the variance-covariance matrix is provided in Appendix E. Since the parameter $\mathrm{W}(P)$ defined by minimum contrast is Hadamard differentiable at P, the bootstrap distribution of $\sqrt{n}(\hat{\mathbf{W}} - \mathrm{W}(P))$ has the same limit with probability 1 (See Theorem 4.1 of Bickel and Freedman (1981)).

5 Conclusion

In this paper we have proved two theorems about the characteristic-function based ICA method (CHFICA). We first showed that prewhitened CHFICA can be consistent even when there exists heavy-tailed sources. But this result is in no sense complete and we conjecture that it could be inconsistent in some cases when several very different heavy-tailed sources and one or more not heavy-tailed sources are mixed together. Further studies of consistent estimators are needed under such situations. Second we showed that CHFICA is robust against small additive noise, $\sqrt{n}$-consistent and asymptotically normal.

6 Appendix

6.1 Appendix A

Proposition 1. *Suppose that S_j has a stable distribution with exponent $\alpha_j \in (0,2)$, $j = 1,2$. Let $\hat{\sigma}_j$ is the sample standard deviation of S_j based on n i.i.d. samples and $d_n = 1/(\sum_{j=1}^2 \hat{\sigma}_j^{-\alpha_j})^2$. Let ϕ_j be the c.f. of S_j. For $t > 0$ and $(x,y) \in R^2$, let $F_j(t;x,y)$ and $F_j'(t;x,y)$ be defined by (11). Let $\Delta_{\tilde{s}}$ be the same as in (10). Then for $\forall O \in \mathcal{O}_2$,*

$$\sup_{O \in \mathcal{O}_2} |d_n \Delta_{\tilde{s}}(O) - g(O, h_{1n}, h_{2n})| = o_p(1),$$

where $h_{jn} = \hat{\sigma}_j^{-\alpha_j}\sqrt{d_n}$ and $g(O; h_{1n}, h_{2n})$ is defined by

$$g(O; h_{1n}, h_{2n}) = \int |F_1'(0; t_1 O_{11}, t_2 O_{21}) h_{1n} + F_2'(0; t_1 O_{12}, t_2 O_{22}) h_{2n}|^2 \lambda(\mathbf{t}) d\mathbf{t}.$$

Proof. The characteristic function of S_j can be expressed as

$$\phi_j(t) = \exp\{itc_j - b_j |t|^{\alpha_j}(1 + ik\text{sgn}(t) w_{\alpha_j}(t))\}, \tag{14}$$

where $b_j > 0$, $c_j, -1 \le k_j \le 1$ are constants and $w_\alpha(t) = \tan(\pi\alpha/2)$ if $\alpha \neq 1$ and $w_\alpha(t) = (2/\pi)\log|t|$ if $\alpha = 1$. (c.f. Durrett, 1996). Let $\hat{I} = \hat{\Sigma}_s^{-1/2}\text{diag}\{\hat{\sigma}_j : 1 \le j \le 2\}$, where $\hat{\Sigma}_s$ is the sample covariance matrix of $\mathbf{S}$. Let $\overline{O} = O\hat{I}$. Then

$$\Delta_{\tilde{s}}(O) = \int |\{F_1(\hat{\sigma}_1^{-\alpha_1}; t_1\overline{O}_{11}, t_2\overline{O}_{21})\phi_2(\frac{t^T\overline{O}^2}{\hat{\sigma}_2}) + F_2(\hat{\sigma}_2^{-\alpha_2}; t_1\overline{O}_{12}, t_2\overline{O}_{22})\phi_1(\frac{t_1\overline{O}_{11}}{\hat{\sigma}_1})\phi_1(\frac{t_2\overline{O}_{21}}{\hat{\sigma}_1})\}|^2\lambda(\mathbf{t})d\mathbf{t}.$$

Notice that $g(O; h_{1n}, h_{2n})$ is continuous w.r.t. (O, h_{1n}, h_{2n}) and thus is uniformly continuous on a compact set. Since $\hat{I} \to I_{2\times 2}$ by Theorem 5 of Chen & Bickel (2004), $\mathcal{O}_2$ is compact and $(h_{1n}, h_{2n}) \in \{(h_1, h_2) \in [0,1]^2 : h_1 + h_2 = 1\}$ (compact), we have

$$\sup_{O \in \mathcal{O}_2} |g(\overline{O}; h_{1n}, h_{2n}) - g(O; h_{1n}, h_{2n})| = o_P(1).$$

Thus it is sufficient to show that

$$\sup_{O \in \mathcal{O}_2} |d_n \Delta_{\tilde{s}}(O) - g(\overline{O}, h_{1n}, h_{2n})| = o_p(1). \tag{15}$$

In the following we prove the sufficient conditions for (15) :

$$\sup_{O \in \mathcal{O}_2} |\int |\sqrt{d_n} F_1(\hat{\sigma}_1^{-\alpha_1}; t_1\overline{O}_{11}, t_2\overline{O}_{21})\phi_2(\frac{t^T\overline{O}^2}{\hat{\sigma}_2}) - F_1'(0; t_1\overline{O}_{11}, t_2\overline{O}_{21})h_{2n}|^2\lambda(\mathbf{t})d\mathbf{t} = o_p(1) \tag{16}$$

and

$$\sup_{O\in\mathcal{O}_2} |\int |\sqrt{d_n}F_2(\hat{\sigma}_2^{-\alpha_2}; t_1\overline{O}_{12}, t_2\overline{O}_{22})\phi_1(\frac{t_1\overline{O}_{11}}{\hat{\sigma}_1})\phi_1(\frac{t_2\overline{O}_{21}}{\hat{\sigma}_1}) - F_2'(0; t_1\overline{O}_{12}, t_2\overline{O}_{22})h_{1n}|^2\lambda(\mathbf{t})d\mathbf{t} = o_p(1). \tag{17}$$

Since $|\phi_2| \le 1$ and $0 \le h_{2n} \le 1$, we have

$$\begin{aligned} & |\sqrt{d_n}F_1(\hat{\sigma}_1^{-\alpha_1}; t_1\overline{O}_{11}, t_2\overline{O}_{21})\phi_2(\frac{t^T\overline{O}^2}{\hat{\sigma}_2}) - F_1'(0; t_1\overline{O}_{11}, t_2\overline{O}_{21})h_{2n}| \\ \le\ & |\sqrt{d_n}F_1(\hat{\sigma}_1^{-\alpha_1}; t_1\overline{O}_{11}, t_2\overline{O}_{21}) - F_1'(0; t_1\overline{O}_{11}, t_2\overline{O}_{21})h_{2n}| \\ & +|F_1'(0; t_1\overline{O}_{11}, t_2\overline{O}_{21})| \cdot |1 - \phi_2(\frac{t^T\overline{O}^2}{\hat{\sigma}_2})|. \end{aligned} \tag{18}$$

By Lemma 1 in Appendix B, $\int |F_1'(0; t_1\overline{O}_{11}, t_2\overline{O}_{21})|^4\lambda(t)dt < \infty$ and is continuous w.r.t $(\overline{O}_{11}, \overline{O}_{21})$ and thus

$$\sup_{O\in\mathcal{O}_2} \int |F_1'(0; t_1\overline{O}_{11}, t_2\overline{O}_{21})|^4\lambda(\mathbf{t})d\mathbf{t} \;=\; O_p(1). \tag{19}$$

By using $|\phi_2| \le 1$, $1 - \cos(x) \le \frac{1}{2}x^2$ and $1 - \exp(-|x|) \le |x|$, we have for $O \in \mathcal{O}_2$

$$\begin{aligned} |1 - \phi_2(\frac{t^T\overline{O}^2}{\hat{\sigma}_2})|^4 &\le 8|1 - \mathrm{Re}(\phi_2(\frac{t^T\overline{O}^2}{\hat{\sigma}_2}))|^2 \\ &\le C\{|f_n(t)|^{\alpha_2} + |f_n(t)|^2 + |f_n(t)|^{2\alpha_2}\},\ \text{if } \alpha_2 \ne 1 \\ &\le C\{|f_n(t)|^{\alpha_2} + (|f_n(t)| + |f_n(t)|^{\alpha_2} \cdot |\log||t|| + \log||\hat{I}|| - \log\hat{\sigma}|)^2\},\ \text{if } \alpha_2 = 1, \end{aligned}$$

where $f_n(t) = \hat{\sigma}_2^{-1}||t|| \cdot ||\hat{I}||$ and the right hand side (RHS) of the above does not depend on O. Since $\hat{\sigma}_2^{-1}\log\hat{\sigma}_2 \to_p 0$ and $\int ||t||^2(\log||t||)^2\lambda(t)dt < \infty$, we have

$$\sup_{O\in\mathcal{O}_2} \int |1 - \phi_2(\frac{t^T\overline{O}^2}{\hat{\sigma}_2})|^4\lambda(\mathbf{t})d\mathbf{t} \;=\; o_p(1). \tag{20}$$

By Lemma 1 in Appendix B, we have

$$\begin{aligned} & |\sqrt{d_n}F_1(\hat{\sigma}_1^{-\alpha_1}; t_1\overline{O}_{11}, t_2\overline{O}_{21}) - F_1'(0; t_1\overline{O}_{11}, t_2\overline{O}_{21})h_{1n}| \\ \le\ & Ch_{1n}\{B_1(t_1\overline{O}_{11}, t_2\overline{O}_{21})\hat{\sigma}_1^{-\alpha_1} + B_2(t_1\overline{O}_{11}, t_2\overline{O}_{21})\hat{\sigma}_1^{-\alpha}\log\hat{\sigma}_1\}, \end{aligned} \tag{21}$$

where

$$\begin{aligned} B_1(x, y) &= (|x| + |y|) \cdot (|x|^{\alpha_2} + |y|^{\alpha_2}) + (|x|^{\alpha_2} + |y|^{\alpha_2})^2,\ \text{if } \alpha_2 \ne 1 \\ &= |x| + |y| + |l(x)| + |l(y)| + |l(x + y)|,\ \text{if } \alpha_2 = 1 \end{aligned}$$

$$\begin{aligned} B_2(x, y) &= 0,\ \text{if } \alpha_2 \ne 1 \\ &= (|x| + |y|) \cdot (|x| + |y| + |l(x + y) - l(x) - l(y)|),\ \text{if } \alpha_2 = 1 \end{aligned}$$

and $l(x) = x\log|x|$. When $\alpha_2 \neq 1$, it is obvious that the RHS of (21) is a polynomial function of $t_1\overline{O}_{11}$ and $t_2\overline{O}_{21}$ multiplied by $\hat{\sigma}_1^{-\alpha_1} = o_p(1)$ and thus by compactness and continuity

$$\sup_{O\in\mathcal{O}_2}\int|\sqrt{d_n}F_1(\hat{\sigma}_1^{-\alpha_1};t_1\overline{O}_{11},t_2\overline{O}_{21}) - F_1'(0;t_1\overline{O}_{11},t_2\overline{O}_{21})h_{1n}|^2\lambda(\mathbf{t})d\mathbf{t} \quad = \quad o_P(1). \qquad (22)$$

When $\alpha_2 = 1$, the RHS of (21) is two polynomials of $t_1\overline{O}_{11}, t_2\overline{O}_{21}$ and $l(t_1\overline{O}_{11}), l(t_2\overline{O}_{21})$ multiplied by a $\hat{\sigma}_1^{-\alpha_1}$ and $\hat{\sigma}_1^{-\alpha_1}\log\hat{\sigma}_1$ separately, since a polynomial function of $l(x)$ and $\mathbf{X}$ is still integrable w.r.t $\lambda(x)$, by compactness and continuity (22) still holds.

Then (16) is implied by (18)(19)(20)(22). Similarly (17) can be proven.

Hence Proposition 1 holds.

Proposition 2. *Let S_j $(j = 1,2)$ and d_n be the same as in Proposition 1 and $\boldsymbol{S} = (S_1, S_2)^T$. Let $\tilde{\Delta}_s$, $\Delta_{\tilde{s}}$ be defined by (8) and (10), separately. Then*

$$d_n \sup_{O\in\mathcal{O}_2}|\tilde{\Delta}_s(O) - \Delta_{\tilde{s}}(O)| \quad = \quad o_P(1). \qquad (23)$$

Proof. Using the fact that for any complex functions f, g we have

$$|\int|f|^2d\mu - \int|g|^2d\mu| \quad = \quad |\int[|f-g|^2 + 2\mathrm{Re}(g^H(f-g))]d\mu| \leq ||f-g||^2 + 2||f-g||\cdot||g||,$$

where g^H is the conjugate function of g. If $||f-g||^2 = o(||g||^2)$, then $|\int(|f|^2-|g|^2)d\mu| = o(||g||^2)$. Recall the definition of

$$\tilde{\Delta}_s(O) \quad = \quad |\int|\hat{c}_s(t^TO\hat{\Sigma}_s^{-\frac{1}{2}}) - \prod_{j=1}^m \hat{c}_s(t_jO_j\hat{\Sigma}_s^{-\frac{1}{2}})|^2\lambda(\mathbf{t})d\mathbf{t}$$

and

$$\Delta_{\tilde{s}}(O) \quad = \quad |\int|c_s(t^TO\hat{\Sigma}_s^{-\frac{1}{2}}) - \prod_{j=1}^m c_s(t_jO_j\hat{\Sigma}_s^{-\frac{1}{2}})|^2\lambda(\mathbf{t})d\mathbf{t}.$$

By Proposition 1, $d_n\sup_{O\in\mathcal{O}_2}\Delta_{\tilde{s}}(O) = O_p(1)$, thus to show that (23), it is sufficient to show that (after omitting the arguments of the corresponding functions in $\tilde{\Delta}_s$ and $\Delta_{\tilde{s}}$)

$$d_n\sup_{O\in\mathcal{O}_m}\int|c_s - \hat{c}_s + \prod_{j=1}^2 c_s - \prod_{j=1}^2\hat{c}_s|^2\lambda(\mathbf{t})d\mathbf{t} \quad = \quad o_P(1).$$

Hence it is sufficient to show that

$$d_n\sup_{O\in\mathcal{O}_m}\int|c_s(t^TO\hat{\Sigma}_s^{-\frac{1}{2}}) - \hat{c}_s(t^TO\hat{\Sigma}_s^{-\frac{1}{2}})|^2\lambda(\mathbf{t})d\mathbf{t} \quad = \quad o_P(1), \qquad (24)$$

and for $j = 1, 2$,

$$d_n\sup_{O\in\mathcal{O}_m}\int|c_s(t_jO_j\hat{\Sigma}_s^{-\frac{1}{2}}) - \hat{c}_s(t_jO_j\hat{\Sigma}_s^{-\frac{1}{2}})|^2\lambda(\mathbf{t})d\mathbf{t} \quad = \quad o_P(1). \qquad (25)$$

Let $\hat{\sigma}_j = (\hat{\Sigma}_s^{\frac{1}{2}})_{jj}$ be the sample standard deviation of $\mathbf{S}_j$. Then by Theorem 5 of Chen & Bickel (2004), $\hat{\Sigma}_s^{-\frac{1}{2}}\text{diag}\{\hat{\sigma}_j\} \to I$. Let δ_{jn} such that $\delta_{jn}^{\alpha_j} = n^{\alpha_j/2-1+\epsilon}$ for $0 < \epsilon < 1 - \max(\alpha_j)/2$. By Lemma 2 in Appendix B, $P(\hat{\sigma}_j\delta_{jn} \leq 1) = o(1)$. Then we have for $T_n \uparrow \infty$,

$$\begin{aligned}
& d_n \sup_{O\in\mathcal{O}_2} \int |c_s(t^T O\hat{\Sigma}_s^{-\frac{1}{2}}) - \hat{c}_s(t^T O\hat{\Sigma}_s^{-\frac{1}{2}})|^2 \lambda(\mathbf{t})d\mathbf{t} \\
\leq\ & d_n \sup_{O\in\mathcal{O}_2, ||t||\leq T_n} |\hat{c}_s(t^T O\hat{\Sigma}_s^{-\frac{1}{2}}) - c_S(t^T O\hat{\Sigma}_s^{-\frac{1}{2}})|^2 + 4d_n \int_{||t||\geq T_n} \lambda(\mathbf{t})d\mathbf{t} \\
\leq\ & d_n \sup_{||t||\leq 2T_n} |\hat{c}_s(t^T \text{diag}\{\hat{\sigma}_j^{-1}\}) - c_s(t^T \text{diag}\{\hat{\sigma}_j^{-1}\})|^2 + 8d_n \exp(-T_n^2/4) + o_P(1) \\
\leq\ & d_n \sup_{\{|t_j|\leq 2T_n\delta_{jn}:1\leq j\leq 2\}} |\hat{c}_s(t^T) - c_s(t^T)|^2 + 8d_n \exp(-T_n^2/4) + o_P(1) \\
\leq\ & d_n\{ \sup_{\{|t_j|\leq 2T_n\delta_{jn}:1\leq j\leq 2\}} |\int(1-\cos(t^T S))d(P_n - P)|^2 + |\int \sin(t^T S)d(P_n - P)|^2\} \\
& +8d_n \exp(-T_n^2/4) + o_P(1).
\end{aligned}$$

Thus by using Lemma 3 in Appendix B, we have

$$\begin{aligned}
d_n \sup_{O\in\mathcal{O}_2} \int |c_s(t^T O\hat{\Sigma}_s^{-\frac{1}{2}}) - \hat{c}_s(t^T O\hat{\Sigma}_s^{-\frac{1}{2}})|^2 \lambda(t)dt &= O_p(\frac{d_n}{n}\sum_{j=1}^{2} T_n^{\alpha_j}\delta_{jn}^{\alpha_j}))) \\
&\quad +O_p(d_n \exp(-T_n^2/4)).
\end{aligned}$$

By using Lemma 2 in Appendix B, we have

$$\sqrt{d_n} = (\sum_{j=1}^{2} \hat{\sigma}_j^{-\alpha_j})^{-1} = o_P((\sum_{j=1}^{2} n^{\alpha_j/2-1-\epsilon})^{-1}).$$

Thus $\sqrt{d_n}\sum_{j=1}^2 \delta_{jn}^{\alpha_j} = o_P(n^{2\epsilon})$ and $\frac{\sqrt{d_n}}{n} = o_p((\sum_{j=1}^2 n^{\alpha_j/2-\epsilon})^{-1})$. By choosing $T_n^{\alpha_j} = n^\epsilon$ and $\epsilon > 0$ such that $\epsilon < \min(\frac{\alpha_j}{8})$ and $\epsilon < 1 - \max(\frac{\alpha_j}{2})$, the RHS of the above is of order $o_P(1)$. Thus (24) holds. Similarly (25) can be proven. Hence Proposition 2 holds.

6.2 Appendix B

Lemma 1. Let ϕ be the c.f. of a stable distribution with exponent α given in (13) by omitting the subscript j. For $t > 0$ and $x, y \in R$, let

$$F(t; x, y) = \phi(t^{1/\alpha}(x+y)) - \phi(t^{1/\alpha}x)\phi(t^{1/\alpha}y), \tag{26}$$

and $F'(t; x, y) = \frac{\partial}{\partial t}F(t; x, y)$. Then $F(0+; x, y) = 0$ and for $\forall 0 < t < e^{-1}$ we have

(i). if $\alpha \neq 1$, then

$$\begin{aligned}
F'(0+; x, y) &= b(|x|^\alpha + |y|^\alpha - |x+y|^\alpha) \\
&\quad +ibk_\alpha(|x|^\alpha\text{sgn}(x) + |y|^\alpha\text{sgn}(y) - |x+y|^\alpha\text{sgn}(x+y)),
\end{aligned}$$

$$\sup_{s\in(0,t)} |F'(t;x,y)-F'(0+;x,y)| \leq c(\alpha)\{|x+y|(|x|^\alpha+|y|^\alpha)+(|x|^\alpha+|y|^\alpha)^2\}t,$$

where $k_\alpha = k\tan(\pi\alpha/2)$ and $c(\alpha)$ only depends on the constants which decides ϕ.

(ii). if $\alpha = 1$, then

$$F'(0+;x,y) = b(|x|+|y|-|x+y|)+ib\overline{k}(l(x)+l(y)-l(x+y)),$$

$$\begin{aligned}\sup_{s\in(0,t)} |F'(t;x,y)-F'(0+;x,y)| &\leq c(\alpha)\{(|x|+|y|)(|x|+|y|+|l(x+y)-l(x)-l(y)|)\}t\log t\\ &+c(\alpha)\{|x|+|y|+|l(x)|+|l(y)|+|l(x+y)|\}t,\end{aligned}$$

where $\overline{k} = \frac{2k}{\pi}$, and $c(\alpha)$ only depends on the constants which decides ϕ, and $l(x) = x\log|x|$ for $x \neq 0$ and $l(0) = 0$.

(iii). $|\frac{F(t;x,y)}{t} - F'(0+;x,y)| = |\frac{1}{t}\int_0^t[F'(s;x,y)-F'(0+;x,y)]ds| \leq B_1(x,y)t + B_2(x,y)|t\log t|$, where $B_1(x,y)$ and $B_2(x,y)$ are defined obviously by (i) and (ii) and $\int B_j^2(x,y)\lambda(x)\lambda(y)dxdy < \infty$ for $j = 1,2$, where $\lambda(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$.

(iv). For $\forall h_1, h_2 \in [0,1]$ such that $h_1+h_2 > 0$, $\alpha_1, \alpha_2 \in (0,2)$, and $a,b,c,d \in R$ such that $bc-ad \neq 0$. If $h_1 F'_{\alpha_1}(0;ax,by)+h_2F'_{\alpha_2}(0;cx,dy) = 0$ for all $(x,y)\in R^2$, then $abcd = 0$.

Proof. Notice that when $\alpha \neq 1$, let $k_\alpha = k\tan(\pi\alpha/2)$, we have (for simplicity we abbreviate $F(t;x,y)$ as $F(t)$)

$$\begin{aligned}F(t) &= e^{itc(x+y)}\{e^{-bt(|x+y|^\alpha+ik_\alpha|x+y|^\alpha \mathrm{sgn}(x+y))}\\ &\quad -e^{-bt(|x|^\alpha+|y|^\alpha+ik_\alpha(|x|^\alpha\mathrm{sgn}(x)+|y|^\alpha\mathrm{sgn}(y)))}\}.\end{aligned}$$

and when $\alpha = 1$, let $\overline{k} = \frac{2k}{\pi}$, we have

$$\begin{aligned}F(t) &= e^{itc(x+y)}\{e^{-bt(|x+y|+i\overline{k}(x+y)\log|x+y|+i\overline{k}(x+y)\log t)}\\ &\quad -e^{-bt(|x|+|y|+i\overline{k}(x\log|x|+y\log|y|)+i\overline{k}(x+y)\log t)}\}\\ &= e^{i(ct+\bar{k}bt\log t)(x+y)}\{e^{-bt(|x+y|+i\overline{k}(x+y)\log|x+y|)}\\ &\quad -e^{-bt(|x|+|y|+i\overline{k}(x\log|x|+y\log|y|))}\}.\end{aligned}$$

Then the results of (i) and (ii) follows by analytic calculation, where $|e^{ix}-e^{iy}| \leq |x-y|$, $|e^{-|x|}-e^{-|y|}| \leq |x-y|$ (for all $x,y \in R$) and the fact that the function $|t\log t|$ is monotone increasing w.r.t $t \in (0, e^{-1})$ are used for the inequalities.

Next for any real function with derivative f' we have $\frac{f(x+y)-f(x)}{y} - f'(x) = \frac{1}{y}\int_0^y[f'(x+t) - f'(x)]dt$, since $B_1(x,y)$ and $B_2(x,y)$ are both L_2-integrable w.r.t $\lambda(x)\lambda(y)$ due to the facts that $\int|x|^c\lambda(x)dx < \infty$ for all $c > 0$ and $\int(x\log|x|)^2\lambda(x)dx < \infty$, then (iii) follows directly from (i) and (ii).

Now we prove (iv) by only using $\mathrm{Re}\{h_1F'_{\alpha_1}(0;ax,by)+h_2F'_{\alpha_2}(0;cx,dy)\} = 0$ for all $(x,y) \in R^2$. When one of $\{h_1,h_2\}$ is zero, say $h_1 = 0$. Then $|cx|^{\alpha_2}+|dy|^{\alpha_2}-|cx+dy|^{\alpha_2} = 0$ for all (x,y). Suppose that $abcd \neq 0$, then by letting $y = -\frac{c}{d}x$ we have $c = d = 0$, contradiction. Thus $abcd = 0$. In the following, we assume that $h_1 > 0$ and $h_2 > 0$. Again suppose that $abcd \neq 0$.

If $\alpha_1 \neq \alpha_2$, say $\alpha_1 < \alpha_2$. By letting $y = -\frac{a}{b}x$ and factoring $|x|^{\alpha_1}$, we have $a = 0$ and have contradiction.

If $\alpha_1 = \alpha_2$, say equal to some $\alpha \in (0, 1]$. When $0 < \alpha \leq 1$, by using the fact that $|x|^\alpha + |y|^\alpha \geq |x + y|^\alpha$ where the equality holds if and only if $xy = 0$ for $\alpha < 1$ and $xy > 0$ for $\alpha = 1$, it is easy to obtain that $abcd = 0$, thus contradiction. When $\alpha \in (1, 2)$, by taking partial derivative $\frac{\partial^2}{\partial x \partial y}$, we have $h_1 ab|ax + by|^{\alpha-2} + h_2 cd|cx + dy|^{\alpha-2} = 0$ for all (x, y) with $(ax + by)(cx + dy) \neq 0$. Since $bc - ad \neq 0$, by taking different values of (x, y), we have $ab = 0$ and $cd = 0$ and thus contradiction.

Thus in all cases we have $abcd = 0$ and (iv) holds.

Lemma 2. *Suppose that ξ has a stable distribution with exponent $\alpha \in (0, 2)$. Let $\hat{\sigma}$ be the sample standard deviation of ξ based on n i.i.d. copies of ξ. Then for $\forall \epsilon > 0$, we have*

$$n^{\frac{1}{\alpha}-\frac{1}{2}-\epsilon}/\hat{\sigma} \quad = \quad o_P(1), \tag{27}$$

and

$$\hat{\sigma}/n^{\frac{1}{\alpha}-\frac{1}{2}+\epsilon} \quad = \quad o_P(1). \tag{28}$$

Proof. Let $\{\xi_k : 1 \leq k \leq n\}$ be i.i.d. copies of ξ and $|\xi|_{(n)} = \max_{1 \leq k \leq n} |\xi_k|$. Since $E\xi^2 = \infty$, by Lemma 1 of Chen & Bickel (2004), there exists a sequence of random variables $\epsilon_n = o_P(1)$ such that

$$\hat{\sigma} \quad = \quad (\frac{1}{n}\sum_{k=1}^{n} |\xi_k|^2)^{\frac{1}{2}}(1 + \epsilon_n).$$

Then $\hat{\sigma} \geq \frac{|\xi|_{(n)}}{\sqrt{n}}(1 + \epsilon_n)$. Let $T_n = n^{\frac{1}{\alpha}-\frac{1}{2}-\epsilon}$ $(0 < \epsilon < \frac{1}{\alpha} - \frac{1}{2})$ and $\delta > 0$. By Theorem 3 in Chapter XVII.5 of Feller (1971), there exists a sequence of numbers $r_n \to 1$ such that its tail probability $P(|\xi| > x_n) = Cr_n x_n^{-\alpha}$ for $x_n \to \infty$. Hence (27) follows from

$$\begin{aligned} P(\frac{T_n}{\hat{\sigma}} \geq \delta) \quad &\leq \quad P(|\xi|_{(n)} \leq \frac{2\sqrt{n}T_n}{\delta}) + P(|\epsilon_n| \geq \frac{1}{2}) \\ &= \quad (1 - Cr_n(\frac{2}{\delta})^{-\alpha} n^{-1+\alpha\epsilon})^n + P(|\epsilon_n| \geq \frac{1}{2}) \\ &\to \quad 0. \end{aligned}$$

By Lemma 2 in Chapter XVII.5 of Feller (1966), for $\forall 0 < a < \alpha$ we have $E|\xi|^a < \infty$. Thus by Gine and Zinn (1992) we have

$$\frac{1}{n}\sum_{k=1}^{n} |\xi_j|^2 \quad = \quad o_P(n^{\frac{2}{\alpha}-1+\epsilon}).$$

Thus (28) follows.

Lemma 3. Suppose that ξ has a stable distribution exponent $\alpha \in (0, 2)$. Let ϕ be the c.f. of ξ. For any sequence of positive numbers $\delta_n = (1)$, we have $1 - \mathrm{Re}(\phi(\delta_n)) = O(\delta_n^\alpha)$.

Proof. Since ϕ can be expressed by (13) by omitting the subscript j, then

$$\begin{aligned} 1 - \mathrm{Re}(\phi(\delta_n)) &= 1 - \exp(-b|\delta_n|^\alpha)\cos(\delta_n c + b\delta_n^\alpha k w_\alpha(\delta_n)) \\ &\leq [1 - \exp(-b\delta_n^\alpha)] + (1 - \cos(\delta_n c + bk\delta_n^\alpha w_\alpha(\delta_n)) \\ &= O(\delta_n^\alpha) + O((\delta_n c + bk\delta_n^\alpha w_\alpha(\delta_n))^2) \\ &= O(\delta_n^\alpha). \end{aligned}$$

Lemma 4. Let $\mathbf{S} = (S_1, S_2)^T$ be the same as in Proposition 1. Denote $\mathbf{t} = (t_1, t_2)^T$. For any sequence of positive numbers $\delta_{jn} = o(1)$, we have

$$\sup_{|t_j| \leq \delta_{jn}} \left| \int (1 - \cos(\mathbf{t}^T\mathbf{S})) d(P_n - P) \right| = O_p(n^{-\frac{1}{2}} (\sum_{j=1}^{2} \delta_{jn}^{\alpha_j})^{\frac{1}{2}}), \quad (29)$$

$$\sup_{|t_j| \leq \delta_{jn}} \left| \int \sin(\mathbf{t}^T\mathbf{S}) d(P_n - P) \right| = O_p(n^{-\frac{1}{2}} (\sum_{j=1}^{2} \delta_{jn}^{\alpha_j})^{\frac{1}{2}}). \quad (30)$$

Proof. We first calculate the generalized bracketing entropy of the above empirical processes and then by using Theorem 5.11 of van de Geer (2001) we can obtain the desired results.

Denote $\mathbf{s} = (s_1, s_2)^T$. Let

$$\mathcal{G}_n = \{g_{(t_1,t_2)}(\mathbf{s}) = 1 - \cos(t_1 s_1 + t_2 s_2) : |t_j| \leq \delta_{jn}\}.$$

and

$$w(\mathbf{s}; \rho_1, \rho_2) = \sup_{\{|t_j - \bar{t}_j| \leq \rho_j : 1 \leq j \leq 2\}} |\cos(t_1 s_1 + t_2 s_2) - \cos(\bar{t}_1 s_1 + \bar{t}_2 s_2)|.$$

Since

$$\begin{aligned} |\cos(t_1 s_1 + t_2 s_2) - \cos(\bar{t}_1 s_1 + \bar{t}_2 s_2)| &\leq |\exp(i(t_1 s_1 + t_2 s_2)) - \exp(i(\bar{t}_1 s_1 + \bar{t}_2 s_2))| \\ &\leq \sum_{j=1}^{2} |1 - \exp(i s_j (t_j - \bar{t}_j))|. \end{aligned}$$

We have $w(\mathbf{s}; \rho_1, \rho_2) = \sum_{j=1}^2 \sup_{|t_j| \leq \rho_j} |1 - \exp(i s_j t_j)|$. Notice that for any $\theta_1, \theta_2 \in (-\pi, \pi)$ with $|\theta_1| \geq |\theta_2|$, $|1 - \exp(i\theta_1)| \leq |1 - \exp(i\theta_2)|$. Thus

$$E(\sup_{|t_j| \leq \rho_j} |1 - \exp(i s_j t_j)|^2) \leq 4P(|S_j| \geq \frac{\pi}{\rho_j}) + E|1 - \exp(i S_j \rho_j)|^2.$$

Let ϕ_j be the c.f. of S_j. Then for $\rho_j > 0$ close to 0, $E|1 - \exp(i S_j \rho_j)|^2 = 2(1 - \mathrm{Re}(\phi_j(\rho_j))) \leq C\rho_j^{\alpha_j}$ by the above Lemma 3, and by Theorem 3 in Chapter XVII.5 of Feller (1966), $P(|S_j| \geq \frac{\pi}{\rho_j}) \leq C\rho_j^{\alpha_j}$. Thus

$$E(\sup_{|t_j| \leq \rho_j} |1 - \exp(i S_j t_j)|^2) \leq C\rho_j^{\alpha_j}.$$

So

$$Ew^2(\mathbf{S};\rho_1,\rho_2) \;\le\; C\sum_{j=1}^{2}\rho_j^{\alpha_j}.$$

Thus the δ-entropy with bracketing of $\mathcal{G}_n$ with $\delta = \sqrt{C\sum_{j=1}^2 \rho_j^{\alpha_j}}$ is bounded by $\log\frac{\delta_{1n}\delta_{2n}}{\rho_1\rho_2}$ for ρ_1,ρ_2 sufficiently small. Since $\sup_{g\in\mathcal{G}_n}|g|\le 1$. By choosing ρ_j such that $\rho_j^{\alpha_j} = \frac{\delta^2}{2C}$, the δ-generalized entropy with bracketing is bounded by $\log(\delta_{1n}\delta_{2n}) - C\log\delta$ for $\delta \le (\delta_{1n}\delta_{2n})^{1/C}$. It is not hard to verify that

$$E|1-\cos(t_1S_1+t_2S_2)|^2 \;\le\; 2\sum_{j=1}^{2}E|1-\exp(it_jS_j)|^2 = 4\sum_{j=1}^{2}(1-\mathrm{Re}(\phi_j(t_j))).$$

Then

$$\sup_{g\in\mathcal{G}_n} Eg^2(\mathbf{S}) \;\le\; C\sum_{j=1}^{2}\delta_{jn}^{\alpha_j}.$$

By Theorem 5.11 of van de Geer (2000), we have

$$\sup_{g\in\mathcal{G}_n}|\int gd(P_n-P)| \;=\; O_p(n^{-\frac{1}{2}}(\sum_{j=1}^{2}\delta_{jn}^{\alpha_j})^{\frac{1}{2}}),$$

i.e., (29) holds. Similarly (30) can be proven. Hence Lemma 4 holds.

6.3 Appendix C

Proof of Theorem 2(i).

We consider the model (11). Let $\mathbf{X}^0(j) = \mathrm{A}\mathbf{S}(j)$. Then $W\mathbf{X}(j) = W[\mathbf{X}^0(j) + r(n)\mathbf{n}(j)]$ for $j = 1,\cdots,n$. We may assume that $r(n)\in[0,,1]$. Define $Z(j) = (\mathbf{X}^0(j),\mathbf{n}(j))^T$ and $z = (\mathbf{x},\mathbf{n})\in\mathbf{R}^m\times\mathbf{R}^m$.

The objective function $\hat{\Delta}_\lambda(\mathrm{W})$ in fact depends on $r(n)$-the magnitude of noise. After expansion and some algebraic organization, we have

$$\begin{aligned}&\hat{\Delta}_\lambda(\mathrm{W})\\ &= \frac{1}{n^{2m}}\sum_{i_1,\cdots,i_{2m}}\{\rho(\mathrm{W}[\mathbf{X}(i_1)-\mathbf{X}(i_2)])\\ &\qquad - 2\prod_{k=1}^{m}\rho(W_k[\mathbf{X}(i_k)-\mathbf{X}(i_{m+1})])\\ &\quad + \prod_{k=1}^{m}\rho(W_k[\mathbf{X}(i_k)-\mathbf{X}(i_{m+k})])\}. \qquad (31)\end{aligned}$$

Define a class of functions $\mathcal{F}:\oplus_{2m}\mathbf{R}^m\to\mathbf{R}$ by : $f\in\mathcal{F}_1$ if and only if there exists $W\in\Omega$ and $r\in[0,1]$ such that

$$f(z(i_1,\cdots,i_{2m})) \;=\; \rho(W\overline{\mathbf{xn}}(i_1,i_2))$$

$$-2\prod_{k=1}^{m}\rho(W_k\overline{\mathbf{xn}}(i_k,i_{m+1}))$$
$$+\prod_{k=1}^{m}\rho(W_k\overline{\mathbf{xn}}(i_k,i_{m+k})), \quad (32)$$

where $\overline{\mathbf{xn}}(i,j)=\mathbf{x}(i)-\mathbf{x}(j)+r(\mathbf{n}(i)-\mathbf{n}(j))$.

Since $|f|\le 4$, $f(z(1:2m))$ is continuous with respect to (W,r) for any fixed $\mathbf{x}(1:2m)$, $\mathbf{n}(1:2m)$, and $\mathcal{F}$'s index set $\Omega\times[0,1]$ is compact, van de Geer (2000)'s Lemma 3.10 tells that the bracketing entropy $H_{1,B}(\delta,\mathcal{F},P)<\infty$ for $\delta>0$. Thus by Uniform Law of Large Numbers (ULLN) for U-processes (see Corollary 3.5 of Arcones & Gine (1993)), we have for $Ef=E[f(Z(1:2m))]$,

$$\sup_{f\in\mathcal{F}_1}[\frac{1}{C(n,2m)}\sum_{\mathcal{I}(2m;n)}f(Z(i_1,\cdots,i_{2m})-Ef]=o_P(1),$$

where $C(n,2m)=n\cdots(n-2m+1)$ and $\mathcal{I}(2m;n)=\{(i_1,\cdots,i_{2m})\in(1:n)^{2m}:\ i_j\neq i_k \text{ for } j\neq k\}$.

Let

$$\Delta_{\lambda,r(n)}(\mathrm{W}) = \int_{\mathbf{t}\in\mathbf{R}^m}|c_{\mathrm{W}\mathbf{X}}(\mathbf{t})-\prod_{j=1}^{m}c_{W_j\mathbf{X}}(t_j)|^2\lambda(\mathbf{t})d\mathbf{t}. \quad (33)$$

By algebraic expansion we have

$$\begin{aligned}\Delta_{\lambda,r(n)}(\mathrm{W}) &= E\rho(\mathrm{W}\mathbf{X}(1)-\mathrm{W}\mathbf{X}(2))\\ &\quad -2E\prod_{k=1}^{m}\rho(W_k\mathbf{X}(k)-W_k\mathbf{X}(m+1)\\ &\quad +\prod_{k=1}^{m}E\rho(W_k\mathbf{X}(1)-W_k\mathbf{X}(2)).\end{aligned}$$

In (31), the sum of terms with $(i_1,\cdots,i_{2m})\notin\mathcal{I}(2m;n)$ is bounded by $\frac{12m^2}{n}\downarrow 0$ as $n\uparrow\infty$. So

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

Since $\Delta_{\lambda,r}(\mathrm{W})\ge 0$ is continuous with respect to (W,r) and $limit_{n\to 0}\Delta_{\lambda,r(n)}(\mathrm{W})=0$ for $\mathrm{W}\in\Omega$ if and only if $\mathrm{W}=\mathrm{W}_P$, thus for any $\varepsilon>0$, there exists $\delta(\varepsilon)>0$ such that for n large enough $min_{\mathrm{W}\in\Omega/d(W_P,\varepsilon)}\Delta_{\lambda,r(n)}(\mathrm{W})\ge\delta(\varepsilon)$, where

$$d(\mathrm{W}_P,\varepsilon) = \{W\in\Omega:\ ||\mathrm{W}-\mathrm{W}_P||\ge\varepsilon\}.$$

Hence for n large enough,

$$
\begin{aligned}
&\Pr(||\hat{\mathrm{W}} - \mathrm{W}_P|| \geq \varepsilon) \\
&\leq \Pr(\min_{\mathrm{W}\in\Omega/d(\mathrm{W}_{P,\varepsilon})} \hat{\Delta}_\lambda(\mathrm{W}) \leq \hat{\Delta}_\lambda(\mathrm{W}_P)) \\
&\leq \Pr(\min_{\Omega/d(\mathrm{W}_{P,\varepsilon})} (\hat{\Delta}_\lambda(\mathrm{W}) - \Delta_{\lambda,r(n)}(W)) \\
&\qquad \leq \hat{\Delta}_\lambda(\mathrm{W}_P) - \delta(\varepsilon)) \\
&\leq \Pr(2\sup_\Omega |\hat{\Delta}_\lambda(\mathrm{W}) - \Delta_{\lambda,r(n)}(W)| \geq \delta(\varepsilon)) \\
&\to 0, \text{ as } n \uparrow \infty.
\end{aligned}
$$

This completes the proof of Theorem 2(i).

6.4 Appendix D

PROOF OF THEOREM 2 (II).

We need to make use of the manifold structure of Ω.

First consider a m-dim unit ball with its center O, on which P and Q are two points and define their angle $\theta > 0$ by the smallest value such that $cos(\theta) =< OP, OQ >$, the Euclidean inner product. If we parameterize the path from OP to OQ by $\gamma(t) = cos(t)OP + sin(t)OR$ for $t \in \mathbf{R}$, where OR is one of the unit tangent vectors at P such that OP, OQ and OR are on the same hyperplane, then $\gamma(0) = P$, $\gamma(\theta) = Q$ and $||\gamma'(t)|| = 1$, i.e., γ is a 1-dim arc curve from P to Q on the m-dim unit ball. Obviously $||r(t_1) - \gamma(t_2)|| \leq |t_1 - t_2|$.,

Since each row of $W \in \Omega$ is on an m-dim unit ball, we may parameterize it row by row as above. Let the angles between kth row of W_0 and $\hat{W}$ be θ_k, $k = 1, \cdots, m$, and $\hat{\eta} = \sqrt{\sum_{k=1}^m \theta_k^2}$. Since $||\hat{W} - \mathrm{W}_P|| = o_P(1)$, $\hat{\eta} = o_P(1)$. W.l.o.g, assume that $\hat{W} \neq W_0$, then $\hat{\eta} > 0$. Let $\gamma : \mathbf{R} \to \Omega$ such that $\gamma(0) = W_0$ and $\gamma(\hat{\eta}) = \hat{W}$, where $\gamma_k(.)$ is parameterize like above on a m-dim unit ball, but rescale it by $\frac{\theta_k}{\hat{\eta}}$. It is easy to see that $||\gamma'(t)|| = \sqrt{\sum_{k=1}^m (\frac{\theta_k}{\hat{\eta}_k})^2} = 1$, and $||\gamma(t_2) - \gamma(t_1)|| \leq |t2 - t1|$. It is obvious that $\gamma(t) \in \Omega$ for small $t \geq 0$ and $||\hat{W} - \mathrm{W}_P|| \leq \hat{\eta}$.

Now we can express $\hat{\Delta}_\lambda(\hat{W})$ by $\hat{\Delta}_\lambda(\gamma(\hat{\eta}))$, which can be seen as the value of $\hat{\Delta}_\lambda(\gamma(t))$ (a function of t) at $t = \hat{\eta}$. Also we rewrite $\Delta_\lambda(\gamma(t)) \equiv \Delta_{\lambda,0}(W)$ for $W = \gamma(t)$. Using the first-order Taylor expansion, we have for a $0 \leq \bar{t} \leq \hat{\eta}$ such that

$$0 = \frac{\partial}{\partial t}\hat{\Delta}_\lambda(\gamma(\hat{\eta})) = \frac{\partial}{\partial t}\hat{\Delta}_\lambda(\gamma(0)) + \hat{\eta}\frac{\partial^2}{\partial t^2}\hat{\Delta}_\lambda(\gamma(\bar{t})).$$

Thus

$$\hat{\eta} = -\frac{\frac{\partial}{\partial t}\hat{\Delta}_\lambda(\gamma(0))}{\frac{\partial^2}{\partial t^2}\hat{\Delta}_\lambda(\gamma(\bar{t}))}. \tag{34}$$

In the proof of Theorem 2(i), we have used the fact that $\hat{\Delta}_\lambda(W)$ can be expressed in terms of a U-statistic in (31) and thus ULLN for U-processes can be used. Again, it is not hard to verify that $\frac{\partial}{\partial t}\hat{\Delta}_\lambda(\gamma(0))$ can also be expressed in terms of a U-statistic. By algebraic calculation, we have precisely

$$\frac{\partial}{\partial t}\hat{\Delta}_\lambda(\gamma(0)) = \frac{1}{n^{2m}} \sum_{i_1,\cdots,i_{2m}} g(Z(i_1,\cdots,i_{2m}); \gamma'(0), r(n)),$$

where g is given by

$$\begin{aligned}
& g(z(i_1,\cdots,i_{2m}); \gamma'(0), r) \\
& = (\gamma'(0)\overline{\mathbf{xn}}(i_1,i_2))^T \rho'_{i1,i2} \\
& \quad -2\sum_{k=1}^{m} \gamma'_k(0)\overline{\mathbf{xn}}(i_k, i_{m+1})\rho'_{i_k,i_{m+1}} \prod_{j\neq k} \rho_{i_j,i_{m+1}} \\
& \quad +\sum_{k=1}^{m} \gamma'_k(0)\overline{\mathbf{xn}}(i_k, i_{m+k})\rho'_{i_k,i_{m+k}} \prod_{j\neq k} \rho_{i_j,i_{m+j}}. \qquad (35)
\end{aligned}$$

with $\rho'_{i_1,i_2} = \rho'(\gamma(0)\overline{\mathbf{xn}}(i_1,i_2))$, $\rho_{i_1,i_2} = \rho(\gamma(0)\overline{\mathbf{xn}}(i_1,i_2))$, and

$$\overline{\mathbf{xn}}(i_1,i_2) = [\mathbf{x}(i_1) - \mathbf{x}(i_2) + r(\mathbf{n}(i_1) - \mathbf{n}(i_2))].$$

Let $\overline{\mathbf{x}}(i_1,i_2) = \mathbf{x}(i_1) - \mathbf{x}(i_2)$ and $\overline{\mathbf{n}}(i_1,i_2) = \mathbf{n}(i_1) - \mathbf{n}(i_2)$.

Since $r(n) \to 0$, by taking a first-order Taylor expansion of g at $r = 0$, we have

$$\begin{aligned}
& g(z(i_1,\cdots,i_{2m}); \gamma'(0), r(n)) \\
& = g(z(i_1,\cdots,i_{2m}; \gamma'(0), 0) \\
& \quad +r(n)\{\overline{\mathbf{n}}(i_1,i_2)^T d_1 + \overline{\mathbf{x}}(i_1,i_2)^T d_2 \overline{\mathbf{n}}(i_1,i_2) \\
& \quad -2\sum_{k=1}^{m} [\overline{\mathbf{n}}(i_k,i_{m+1})^T d_{3k} + \overline{\mathbf{x}}(i_k,i_{m+1})^T d_{4k}\overline{\mathbf{n}}(i_k,i_{m+1})] \\
& \quad +\sum_{k=1}^{m} [\overline{\mathbf{n}}(i_k,i_{m+k})^T d_{5k} + \overline{\mathbf{x}}(i_k,i_{m+k})^T d_{6k}\overline{\mathbf{n}}(i_k,i_{m+k})]\},
\end{aligned}$$

where d_1, d_{3k}, d_{5k} are totally bounded random vectors, and d_2, d_{4k}, d_{6k} are totally bounded random matrices (since ρ, ρ' and ρ'' are all bounded).

Furthermore, it is not hard to see that

$$g(z(i_1,\cdots,i_{2m}; \gamma'(0), 0)$$

can be written as $\sum_{k=1}^{m} \gamma'_k(0) h_k(\mathbf{x}(1:2m))$, where h_k is a function of $(\mathbf{x}(1),\cdots,\mathbf{x}(2m))$ and does not depend on $\gamma'(0)$, and $Eh_k(\mathbf{X}^0(1:2m)) = 0$. By the central limit theorem, we have $\frac{1}{n^{2m}}\sum_{i_1,\cdots,i_{2m}} h_k(\mathbf{X}^0(1:2m)) = O_P(\frac{1}{\sqrt{n}})$. By applying ULLN on the bounded U-statistics, the remaining term in $\frac{\partial}{\partial t}\hat{\Delta}_\lambda(\gamma(0))$ is $O_P(r(n))$. Thus

$$\frac{\partial}{\partial t}\hat{\Delta}_\lambda(\gamma(0)) = O_P(\frac{1}{\sqrt{n}} + r(n)). \qquad (36)$$

Finally let us consider the magnitude of $\frac{\partial^2}{\partial t^2}\hat{\Delta}_\lambda(\gamma(\bar{t}))$. Using the fact that $\gamma''(t) = -\gamma(t)$. Thus the sets for all $\gamma(t)$, $\gamma'(t)$ and $\gamma''(t)$ are compact. Let $\overline{x}(i_1, i_2) = x(i_1) - x(i_2)$.

$$\frac{\partial^2}{\partial t^2}\hat{\Delta}_\lambda(\gamma(\bar{t})) \quad = \quad \frac{1}{n^{2m}} \sum_{i_1,\cdots,i_{2m}} h(X(i_1, \cdots, i_{2m}; \gamma(\bar{t}), \gamma'(\bar{t}), \gamma''(\bar{t})). \tag{37}$$

where, denoting $(\gamma(\bar{t}), \gamma'(\bar{t}), \gamma''(\bar{t}))$ by $(w, \overline{w}, \overline{\overline{w}})$,

$$\begin{aligned}
& h(x(i_1, \cdots, i_{2m}); w, \overline{w}, \overline{\overline{w}}) \\
&= \{(\overline{w}\ \overline{x}(i_1, i_2))^T \rho''(w\overline{x}(i_1, i_2))(\overline{w}\ \overline{x}(i_1, i_2)) \\
&\quad +(\overline{\overline{w}}\ \overline{x}(i_1, i_2))^T \rho'(w\overline{x}(i_1, i_2))\} \\
&\quad -2\{\sum_{k=1}^{m} \overline{\overline{w}}_k \overline{x}(i_k, i_{m+1})\rho'_{k,1} \prod_{j\neq k} \rho_{j,1} \\
&\qquad + \sum_{k=1}^{m} (\overline{w}_k \overline{x}(i_k, i_{m+1}))^2 \rho''_{k,1} \prod_{j\neq k} \rho_{j,1} \\
&\qquad + \sum_{j\neq k} \prod_{l\in\{j,k\}} \overline{w}_l \overline{x}(i_l, i_{m+1})\rho'_{l,1} \prod_{l\notin\{j,k\}} \rho_{l,1}\} \\
&\quad +\{\sum_{k=1}^{m} \overline{\overline{w}}_k \overline{x}(i_k, i_{m+k})\rho'_{k,m} \prod_{j\neq k} \rho_{j,m} \\
&\qquad + \sum_{k=1}^{m} (\overline{w}_k \overline{x}(i_k, i_{m+k}))^2 \rho''_{k,m} \prod_{j\neq k} \rho_{j,m} \\
&\qquad + \sum_{j\neq k} \prod_{l\in\{j,k\}} \overline{w}_l \overline{x}(i_l, i_{m+l})\rho'_{l,m} \prod_{l\notin\{j,k\}} \rho_{l,m}\},
\end{aligned}$$

with $\rho_{j,m} = \rho(w_j\overline{x}(i_j, i_{m+j})$, $\rho'_{j,m} = \rho'(w_j\overline{x}(i_j, i_{m+j})$, $\rho''_{j,m} = \rho''(w_j\overline{x}(i_j, i_{m+j})$, $\rho_{j,1} = \rho(w_j\overline{x}(i_j, i_{m+1})$, $\rho'_{j,1} = \rho'(w_j\overline{x}(i_j, i_{m+1})$ and $\rho''_{j,1} = \rho''(w_j\overline{x}(i_j, i_{m+1})$.

Notice that since $w, \overline{w}, \overline{\overline{w}}, \rho, \rho', \rho''$ are all bounded, then $h(x(i_1, \cdots, i_{2m}); w, \overline{w}, \overline{\overline{w}})$ is bounded by $C_1||\overline{x}(i_1, i_2)|| + C_2 \sum_{k=1}^{m} ||\overline{x}(i_k, i_{m+1})|| + C_3 \sum_{k=1}^{m} ||\overline{x}(i_k, i_{m+1})||^2 + C_4 \sum_{k=1}^{m} ||\overline{x}(i_k, i_{m+k})||^2$. Further the U-process defined by the kernel function h is indexed by a compact set defined by $(\gamma(t), \gamma'(t), \gamma''(t), r(n))$, thus ULLN can be applied and we have

$$\sup_{\bar{t}, r(n)} |\frac{\partial^2}{\partial t^2}\hat{\Delta}_\lambda(\gamma(\bar{t})) - Eh(X(1:2m); \gamma(\bar{t}), \gamma'(\bar{t}), \gamma''(\bar{t}))| \quad = \quad o_P(1).$$

Since $\bar{t} \to_P 0$ and $r(n) \to 0$, by continuity we can replace $(\bar{t}, r(n))$ with $(0, 0)$ for the limit of $Eh(X(1:2m); \gamma(\bar{t}), \gamma'(\bar{t}), \gamma''(\bar{t}))$. After some algebraic calculation, the limit is equal to $\frac{\partial^2}{\partial t^2}\Delta_\lambda(\gamma(0))$. Thus

$$\frac{\partial^2}{\partial t^2}\hat{\Delta}_\lambda(\gamma(\bar{t})) \quad \geq \quad \min_{\gamma'(0), \gamma''(0)} \frac{\partial^2}{\partial t^2}\Delta_\lambda(\gamma(0)) + o_P(1).$$

Since $\mathbf{W_P} = \gamma(0)$ is the unique local minimizer due to identifiability, by differentiability we have $\frac{\partial^2}{\partial t^2}\Delta_\lambda(\gamma(0)) > 0$ in any direction $(\gamma'(0), \gamma''(0))$. Again by compactness,

$$\min_{\gamma'(0), \gamma''(0)} \frac{\partial^2}{\partial t^2}\Delta_\lambda(\gamma(0)) \quad > \quad 0.$$

Hence from (34)(36) we have

$$\hat{\eta} \;=\; O_P(\frac{1}{\sqrt{n}} + r(n)).$$

This completes the proof of Theorem 2(ii).

6.5 Appendix E

In the following, we formally calculate the variance-covariance matrix for $\sqrt{n}(\hat{\mathbf{W}}\mathrm{W}_P^{-1} - \mathrm{I})$, where I is the $m \times m$ identity matrix. Denote $\hat{\delta} = \hat{\mathbf{W}}\mathrm{W}_P^{-1} - \mathrm{I}$, i.e. $\hat{\mathrm{W}}_k = \hat{\delta}_k \mathrm{W}_P + W_{Pk}$. Since $\hat{\mathrm{W}}, \mathrm{W}_P \in \Omega$, we have $||\hat{\delta}_k \mathrm{W}_P + W_{Pk}||^2 = 1$. By Theorem 2 (i), $\hat{\delta} = o_P(1)$. By taking partial derivative w.r.t δ_{kj} for $1 \le k \ne j \le m$, we have $\frac{\partial \hat{W}_k}{\partial \delta_{kj}} = W_{Pj} + W_{Pk}\frac{\partial \hat{\delta}_{kk}}{\partial \delta_{kj}}$, $\frac{\partial \hat{\delta}_{kk}}{\partial \delta_{kj}} = -\frac{\hat{\delta}_k \mathrm{W}_P W_{Pj}^T + \omega_{kj}}{\hat{\delta}_k \mathrm{W}_P W_{Pk}^T + 1}$ and $\frac{\partial^2 \hat{\delta}_{kk}}{\partial \delta_{kj}^2}|_{\hat{\delta}_k = 0} = (\omega_{kj})^2 - 1$, where $\omega_{kj} = W_{Pk}W_{Pj}^T$. Then

$$\frac{\partial \hat{\delta}_{kk}}{\partial \delta_{kj}}|_{\hat{\delta}_k=0} = -\omega_{kj} \quad \text{and} \quad \frac{\partial^2 \hat{\delta}_{kk}}{\partial \delta_{kj}^2}|_{\hat{\delta}_k=0} = \omega_{kj}^2 - 1.$$

Thus $\hat{\delta}_{kk} = -\sum_{1 \le j \le m, j \ne k} \omega_{kj}\hat{\delta}_{kj} + o_P(\sum_{j \ne k} \hat{\delta}_{kj})$.

We choose ρ such that it is an even function and define $\dot{\rho}(t) = \frac{\partial \rho(t)}{\partial t}$, $\ddot{\rho}(t) = \frac{\partial^2 \rho(t)}{\partial t^2}$. Let $\mathcal{P}(2m)$ denote the set of all possible permutation of $(1, \cdots, 2m)$. $\hat{\Delta}(\hat{\mathrm{W}})$ is a function of δ through $\hat{\mathrm{W}} = \delta \mathrm{W}_P + \mathrm{W}_P$. Denote $\hat{\Delta}(\delta) \equiv \hat{\Delta}((\delta + \mathrm{I})\mathrm{W}_P)$.

Let

$$\begin{aligned} K(\mathrm{W}; \mathbf{X}(i_1, \cdots, i_{2m})) \;=\; & \prod_{k=1}^{m} \rho(W_k[\mathbf{X}(i_1) - \mathbf{X}(i_2)]) \\ & -2\prod_{k=1}^{m} \rho(W_k[\mathbf{X}(i_k) - \mathbf{X}(i_{m+1})]) \\ & +\prod_{k=1}^{m} \rho(W_k[\mathbf{X}(i_k) - \mathbf{X}(i_{m+k})]), \end{aligned}$$

which should be considered as a function of δ through $\mathrm{W} = \delta \mathrm{W}_P + \mathrm{W}_P$. Then

$$\hat{\Delta}_\lambda(\delta) \;=\; \frac{1}{n^{2m}} \sum_{i_1, \cdots, i_{2m}} K((\delta + \mathrm{I})\mathrm{W}_P; \mathbf{X}(i_1, \cdots, i_{2m})),$$

where $\mathbf{X}(i_1, \cdots, i_{2m}) = (\mathbf{X}(i_1), \cdots, \mathbf{X}(i_{2m}))$.

Denote $\overline{\mathbf{X}}(i_1, i_2) = \mathbf{X}(i_1) - \mathbf{X}(i_2)$, $\overline{X}_k(i_1, i_2) = X_k(i_1) - X_k(i_2)$, similarly for $\overline{\mathbf{S}}(i_1, i_2)$ and $\overline{S}_k(i_1, i_2)$. Since we have, for $1 \le k \ne j \le m$,

$$\sqrt{n}\frac{\partial \hat{\Delta}(\hat{\delta})}{\partial \delta_{kj}} \;=\; 0,$$

by using Taylor expansion and ignore the $o_P(\sqrt{n}\hat{\delta}_{kj})$ terms, the left-hand side approximately equals

$$\sqrt{n}\frac{\partial\hat{\Delta}(0)}{\partial\delta_{kj}} + \sum_{1\le p\neq q\le m}\sqrt{n}\hat{\delta}_{pq}\frac{\partial^2\hat{\Delta}(0)}{\partial\delta_{kj}\partial\delta_{pq}},$$

where

$$\frac{\partial\hat{\Delta}(0)}{\partial\delta_{kj}} = \frac{1}{n^{2m}}\sum_{i_1,\cdots,i_{2m}}\dot{K}_{kj}(\mathbf{S}(i_1,\cdots,i_{2m})),$$

$$\frac{\partial^2\hat{\Delta}(0)}{\partial\delta_{kj}\partial\delta_{pq}} = \frac{1}{n^{2m}}\sum_{i_1,\cdots,i_{2m}}\frac{\partial^2 K((\delta+\mathrm{I})\mathrm{W}_P;\mathbf{X}(i_1,\cdots,i_{2m}))}{\partial\delta_{kj}\partial\delta_{pq}}|_{\delta=0},$$

and

$$\begin{aligned}\dot{K}_{kj}(\mathbf{S}(i_1,\cdots,i_{2m})) &\equiv \frac{\partial K((\delta+\mathrm{I})\mathrm{W}_P;\mathbf{X}(i_1,\cdots,i_{2m}))}{\partial\delta_{kj}}|_{\delta=0}\\ &= (\overline{S}_j(i_1,i_2)-\omega_{kj}\overline{S}_k(i_1,i_2))\dot{\rho}(\overline{S}_k(i_1,i_2))\prod_{l\neq k}\rho(\overline{S}_l(i_1,i_2))\\ &\quad -2(\overline{S}_j(i_k,i_{m+1})-\omega_{kj}\overline{S}_k(i_k,i_{m+1}))\dot{\rho}(\overline{S}_k(i_k,i_{m+1}))\prod_{l\neq k}\rho(\overline{S}_l(i_l,i_{m+1}))\\ &\quad +(\overline{S}_j(i_k,i_{m+k})-\omega_{kj}\overline{S}_k(i_k,i_{m+k}))\dot{\rho}(\overline{S}_k(i_k,i_{m+k}))\prod_{l\neq k}\rho(\overline{S}_l(i_l,i_{m+l})).\end{aligned}$$

By using the Hajek decomposition,

$$\sqrt{n}\frac{\partial\hat{\Delta}(0)}{\partial\delta_{kj}} = \frac{1}{\sqrt{n}}\int h_{kj}(\mathbf{S})dP_n + o_P(1),$$

where after simplification

$$\begin{aligned}h_{kj}(\mathbf{s}) &\equiv \frac{1}{(2m)!}\sum_{(i_1,\cdots,i_{2m})\in\mathcal{P}(2m)}E[\dot{K}_{kj}(\mathbf{S}(i_1,\cdots,i_{2m}))|\mathbf{S}(1)=\mathbf{s}]\\ &= \frac{1}{m}\mathrm{cov}(S_j-s_j,\rho(S_j-s_j))E[\dot{\rho}(S_k-s_k)\prod_{l\neq(k,j)}\rho(S_l-s_l)]\end{aligned}$$

does not depend on W_P.

Since $\dot{\rho}$ and $t\rho(t)$ are odd functions and $\{S_k : 1\le k\le m\}$ are mutually independent, by using the Law of Large Numbers for U-statistics, we have for $1\le p\neq q\le m$ with $(p,q)\notin\{(k,j),(j,k)\}$,

$$\frac{\partial^2\hat{\Delta}(0)}{\partial\delta_{kj}\partial\delta_{pq}} \to_P 0,$$

for $p=k, q=j$,

$$\frac{\partial^2\hat{\Delta}(0)}{\partial\delta_{kj}\partial\delta_{kj}} \quad \to_P \quad F_{kj},$$

where

$$\begin{aligned} F_{kj} &= \{E[(\overline{S}_j(1,2))^2\rho(\overline{S}_j(1,2))] - 2E[\overline{S}_j(1,2))^2\rho(\overline{S}_j(3,2))] + E[(\overline{S}_j(1,2))^2]E[\rho(\overline{S}_j(1,2))]\} \\ &\quad \times E[\ddot{\rho}(\overline{S}_k(1,2))] \prod_{l\neq k,j} E\rho(\overline{S}_l(1,2)) \end{aligned}$$

and for $(p,q)=(j,k)$,

$$\frac{\partial^2\hat{\Delta}(\mathrm{I})}{\partial\delta_{kj}\partial\delta_{jk}} \quad \to_P \quad G_{kj},$$

where

$$\begin{aligned} G_{kj} &= \{\prod_{l\in\{k,j\}} E[\overline{S}_l(1,2)\dot{\rho}(\overline{S}_l(1,2))] - 2\prod_{l\in\{k,j\}} E[\overline{S}_l(2,3)\dot{\rho}(\overline{S}_l(1,3))]\} \\ &\quad \times \prod_{l\neq k,j} E[\rho(\overline{S}_l(1,2))]. \end{aligned}$$

Hence we have for $1\leq k\neq j\leq m$,

$$\sqrt{n}\mathrm{H}(\hat{\delta}_{kj},\hat{\delta}_{jk})^T = \sqrt{n}\int(h_{kj}(\mathbf{S}),h_{jk}(\mathbf{S}))^T dP_n + o_P(1),$$

where $\mathrm{H}=[F_{kj},G_{kj};F_{jk},G_{jk}]$. Thus by the central limit theorem, we have

$$\sqrt{n}\hat{\delta} \to_d \mathcal{N}(0,\Sigma_P),$$

where $\Sigma_P = cov(\varepsilon)$. Here ε is an $m\times m$ matrix of random variables and its elements are decided by the following equations: For $1\leq k\neq j\leq m$,

$$\begin{bmatrix} F_{kj} & G_{kj} \\ G_{jk} & F_{jk} \end{bmatrix} \begin{bmatrix} \varepsilon_{kj} \\ \varepsilon_{jk} \end{bmatrix} = \begin{bmatrix} h_{kj}(\mathbf{S}) \\ h_{jk}(\mathbf{S}) \end{bmatrix},$$

and

$$\varepsilon_{kk} = -\sum_{1\leq j\leq m, j\neq k} \omega_{kj}\varepsilon_{kj}.$$

Note that $\{\varepsilon_{kj}: 1\leq k\neq j\leq m\}$ do not depend on W_P.

References

[1] Arcones, M. and Gine, E. (1993). Limit theorems for U-processes, *Annals of Probability,* Vol. 21, No. 3, pp.1494-1542.

[2] Bickel, P. J. and Doksum, K. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association,* **76** 296-311.

[3] Bickel, P. J. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, Vol. 9, No. 6, pp.1196-1217.

[4] Chen, A., Bickel, P. J. (2004). Consistent Independent Component Analysis and Prewhitening. *Submitted.*

[5] Chen, A., Bickel, P. J. (2003). Efficient Independent Component Analysis. *Technical Report #634, Department of Statistics, UC Berkeley.*

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

[7] Durrett, R. (1996). *Probability: Theory and Examples.* Second edition. Duxbury Press.

[8] Edelman, A., Arias, T. and Smith, S. (1999). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, **20**(2): 303-353.

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

[10] Feller, W. (1966). *An Introduction t o Probability Theory and Its Applications, Volume II.* John Wiley & Sons, Inc.

[11] Gine, E., Zinn, J. (1992). Marcinkiewicz type laws of large numbers and convergence of moments for U-statistics. *Probability in Banach Spaces, 8: proceedings of the eighth international conference*, edited by Richard M. Dudley, Marjorie G. Hahn, James Kuelbs.

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

[13] van de Geer, S. (2000). *Applications of Empirical Process Theory.* Cambridge University Press.