Independent Component Analysis with Heavy Tails using Distance Covariance

Richard A. Davis* Leon Fernandes †

January 16, 2022

1 Introduction

An *Independent Components Analsis* (ICA) refers to the problem of recovering individual signals from observations that consist of mixtures of the signals. Specifically, we consider observations coming from the model

$$X = AS\,, \tag{1.1}$$

where S is an m-dimensional source vector, X is a random m-vector, and $A \in \mathbb{R}^{m\times m}$ is the unknown mixing matrix. Based on observations $X_1, \ldots, X_n$, the goal is to estimate the unmixing matrix $W = A^{-1}$ which allows on the recover the corresponding signals $S_1, \ldots, S_n$. Independent Component Analysis (ICA) is a popular framework where we answer this question by estimating the a priori unknown mixing matrix. ICA has been used for signal separation in fields such as financial time series, biomedical engineering, neuroscience, speech recognition, and so on, see [12], [6] and [9].

Formally, denote the signal by $S = (S_1, S_2, ..., S_m)^T \in \mathbb{R}^m$ a random vector where the components $\{S_1, S_2, ..., S_m\}$ are mutually independent random variables. For a mixing matrix $A \in \mathbb{R}^{m\times m}$, which we assume to be invertible, we observe n samples $X_1, ..., X_n$ with distribution X where $X = AS$. The question of separating the signals is to estimate $W_0 := A^{-1}$, the unmixing matrix.

The motivation for this body of work comes from studying data like the financial cost and number of deaths due to natural disasters in Davis and Ng, 2021 [**?**]. It is observed that these variables tend to have large values corresponding to disaster type events such as September 11, Hurricane Katrina, Hurricane Sandy, and so on. To fix ideas consider a Vector Autoregressive VAR(p) time series model $Y_t \in \mathbb{R}^m$ represented by

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t \tag{1.2}$$

where $\phi_i \in \mathbb{R}^{m\times m}$ for $1 \le i \le p$. The noise term is taken to be an ICA model $e_t = AS_t$ and independent component of S_t is referred to as a shock. It is observed that the shock component corresponding to disaster-type events has heavy tails. We thus need a method that can reliably return estimates for the unmixing matrix when one of the components of the signal S has infinite variance.

Numerous methods for ICA have been developed and studied since its introduction. These include estimation using maximum likelihood [5], JADE [3], kurtosis maximization/minimization in fastICA [10], product density estimation using splines in ProDenICA [8] and log concave density approximation in LogConICA [14]. Refer [6] and for a more extensive list of methods and techniques developed for ICA. With the exception of LogConICA, all these methods require finite second moments. Anderson et al. 2015 [1] and Anderson et al. 2017 [2] study the heavy tailed ICA problem via a convex geometry approach where they orthogonalize

*Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027. Email: rdavis@stat.columbia.edu The first author acknowledges financial support from the National Science Foundation under grant DMS-2015379.

†Department of Statistics, Columbia University, 1255 Amsterdam Avenue New York, NY 10027. Email: lf2607@columbia.edu

the data using the centroid body. However our method is based on the methods CHFICA Chen and Bickel 2005 [4] and dCovICA from Matteson and Tsay 2017 [11].

CHFICA uses the following method as a criterion for independence. Let $Z \in \mathbb{R}^m$ be a random vector and recall that the components of Z are mutually independent if and only if the joint the characteristic function of Z factorizes: $\mathbb{E}\exp(iz^T Z) = \prod_{k=1}^m \mathbb{E}\exp(iz_k Z_k)$ for each $z \in \mathbb{R}^m$. One can then consider:

$$\int_{\mathbb{R}^m} \left| \mathbb{E}\exp(iz^T Z) - \prod_{k=1}^m \mathbb{E}\exp(iz_k Z_k) \right|^2 d\mu(z) \tag{1.3}$$

where μ is some measure with support $\mathbb{R}^m$. Chen and Bickel [4] considered $\mu(z) = \prod_{j=1}^m \lambda_1(z_j)$ where λ_1 is a 1-D probability measure.

Matteson and Tsay [11] obtained consistency and asymptotic distributions for their estimator by using the distance covariance statistic as the criterion for independence. Distance covariance was developed by Székely et al. [16] and Székely and Rizzo [15]. It is a weighted square norm of the difference between the joint and product of the marginal characteristic functions with respect to a specific weight function. Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors and consider

$$\int_{\mathbb{R}^{p+q}} \left| \mathbb{E}\exp(i\langle s, X\rangle + i\langle t, Y\rangle) - \mathbb{E}\exp(i\langle s, X\rangle) \cdot \mathbb{E}\exp(i\langle t, Y\rangle) \right|^2 d\mu(s,t) \tag{1.4}$$

where μ is a measure supported on $\mathbb{R}^{p+q}$. Suppose we choose μ such that the above integral is finite, for example if μ is a finite measure (as the absolute value of the characteristic functions are bounded by 1). Then (1.4) is zero if and only if X and Y are independent. Distance covariance, however, uses an infinite measure which we will state in Section 2.

This paper is organized as follows. In Section 2 we introduce the distance covariance statistic and prove uniform convergence results for the distance covariance which we shall find useful when tackling ICA. Our main result is in Section 3 where we apply the aforementioned uniform convergence results to the ICA and prove consistency of our estimator. We also show that in the presence of noise, our estimator remains consistent. The ancillary results as well as the more technical results are collected in the Appendix.

2 Distance Covariance

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors and we are interested in studying if X and Y are independent or not. Covariance or correlation between X and Y is one quantity one immediately thinks of when studying independence. However it is well known that covariance or correlation can be zero when X and Y are dependent. An alternative to these quantities is the so called distance covariance (and distance correlation) as introduced by Székely et al. [16] and Székely and Rizzo [15]. This quantity has the property that it is a measure of dependence and in particular this has the very desirable property that it is zero if and only if X and Y are independent.

Notation Before embarking on the main results, we begin with some notation. To avoid heavy notation we suppress the dependence of inner products and norms on p and q. All inner products and norms are compatible with the dimension of the vectors and matrices. For vectors $z, z' \in \mathbb{R}^q$ denote their inner product by $\langle z, z'\rangle$. Denote the Euclidean norm by $|z| := \sqrt{\langle z, z\rangle}$. For a $p \times q$ matrix U, we denote by $||U||$ to be the spectral norm of U. That is,

$$||U|| = \sup_{z \in \mathbb{R}^q} \frac{|Uz|}{|z|}$$

Denote by $\phi_{X,Y}(s,t) = \mathbb{E}\exp(i\langle s, X\rangle + i\langle t, Y\rangle)$, the joint characteristic function. Let $\phi_X(s) = \phi_{X,Y}(s,0)$ and $\phi_Y(t) = \phi_{X,Y}(0,t)$ denote the respective marginal characteristic functions.

We will start by defining the distance covariance metric for studying the factorization of (X, Y) a random vector into its components X and Y. Let (X, Y) have some joint distribution with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ such that $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$. We define the distance covariance to be

$$\mathcal{I}(X, Y) := \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(s, t) - \phi_X(s)\phi_Y(t)|^2 \, d\omega(s, t)$$

where ω is the infinite measure

$$d\omega(s, t) = (c_p c_q |s|^{1+p} |t|^{1+q})^{-1} ds dt,$$

$c_p = \pi^{(1+p)/2}/\Gamma((1 + p)/2)$, $c_q = \pi^{(1+q)/2}/\Gamma((1 + q)/2)$.

The finiteness of $\mathcal{I}(X, Y)$ is guaranteed if $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$ - see (2.5) in Székely, et al. [16]. If in addition we had $\mathbb{E}|XY| < \infty$, then due to Lemma 2.3 in Davis et al. 2018, [7]

$$\mathcal{I}(X, Y) = \mathbb{E}[|X - \dot{X}||Y - \dot{Y}|] + \mathbb{E}[|X - \dot{X}|]\mathbb{E}[|Y - \dot{Y}|] - 2\mathbb{E}[|X - \dot{X}||Y - \ddot{Y}|] \tag{2.1}$$

where $(\dot{X}, \dot{Y})$ and $(\ddot{X}, \ddot{Y})$ are i.i.d copies of (X, Y). If X and Y both have components with infinite variance, then $\mathbb{E}|XY|$ is not necessarily finite. This is a key point in ICA with heavy tails (i.e. source $S \in \mathbb{R}^m$ with one infinite variance component) because for most $m \times m$ matrices B, BS will have multiple components with infinite variance.

Suppose we have samples $\{(X_j, Y_j)\}_{j=1}^n$ where $(X_j, Y_j) \sim (X, Y)$. We can estimate the distance covariance statistic by the sample estimate

$$\mathcal{I}_n(X, Y) = \int_{\mathbb{R}^{p+q}} |\hat{\varphi}_{X,Y}(s, t) - \hat{\varphi}_X(s)\hat{\varphi}_Y(t)|^2 \, d\omega(s, t)$$

where $\hat{\varphi}_{X,Y}(s, t) = \frac{1}{n}\sum_{j=1}^n \exp(i(s^T X_j + t^T Y_j))$, $\hat{\varphi}_X(s) = \frac{1}{n}\sum_{j=1}^n \exp(is^T X_j)$ and $\hat{\varphi}_Y(t) = \frac{1}{n}\sum_{j=1}^n \exp(it^T X_j)$.

Unlike (2.1), the sampling distribution has all the moments. We can thus obtain the expression for the sample distance covariance to be

$$\begin{aligned}\mathcal{I}_n(X, Y) =& \left(\frac{1}{n^2}\sum_{1\le j,k\le n} |X_j - X_k||Y_j - Y_k|\right) + \left(\frac{1}{n^2}\sum_{1\le j,k\le n} |X_j - X_k|\right)\left(\frac{1}{n^2}\sum_{1\le j,k\le n} |Y_j - Y_k|\right) \\ & - 2\left(\frac{1}{n^3}\sum_{1\le j,k,l\le n} |X_j - X_k||Y_j - Y_l|\right)\end{aligned}$$

We obtain a strong consistency result for uniform convergence of the empirical distance covariance over a compact set

$$C_M := \{(U, V) : U \in \mathbb{R}^{p\times p}, V \in \mathbb{R}^{q\times q}, ||U|| \le M, ||V|| \le M\}$$

The consistency $\mathcal{I}_n(X, Y) \xrightarrow{a.s.} \mathcal{I}(X, Y)$ (for i.i.d samples) was shown in Theorem 1 of Székely, et al. [16] and to prove our result we generalize the proof of this consistency result to our case. The details of the proof can be found in Section 4.

Theorem 2.1. *Consider a strictly stationary ergodic time series $\{X_j, Y_j\}_{j\ge1}$ with values in $\mathbb{R}^{p+q}$ with marginal distribution (X, Y) such that $\mathbb{E}|(X, Y)| < \infty$. Then*

(a) For each $M > 0$, we have as $n \to \infty$

$$\sup_{(U,V)\in C_M} |\mathcal{I}_n(UX, VY) - \mathcal{I}(UX, VY)| \xrightarrow{a.s.} 0$$

(b) For $\{(\Gamma_n, \Lambda_n)\}_{n\ge1}$ a sequence of random matrices, where $\Gamma_n \in \mathbb{R}^{p\times p}$ and $\Lambda_n \in \mathbb{R}^{q\times q}$ that converge almost surely to non-random matrices Γ and Λ respectively and $M > 0$, we have as $n \to \infty$

$$\sup_{(U,V)\in C_M} |\mathcal{I}_n(U\Gamma_n X, V\Lambda_n Y) - \mathcal{I}(U\Gamma X, V\Lambda Y)| \xrightarrow{a.s.} 0$$

3 Application of Distance Covariance to ICA

Before stating our results it is important we note that there are identifiability issues that require attention - we employ the same convention from Chen and Bickel [4]. Firstly, to standardize the scaling for each component of S we assume that each row of W_0 has norm 1. Secondly, the signs on each S and corresponding row of S can be flipped to get the same distribution. To solve this we assume that along each row of W_0, the entry with the largest absolute value has positive sign. Finally, we may permute the rows of S - to resolve this we assume that the rows are partially ordered by $\prec$, where for $a, b \in \mathbb{R}^m$ we say $a \prec b$ if whenever there exists $1 \leq k \leq m$ such that $a_k < b_k$ then $a_l = b_l$ for $1 \leq l \leq k-1$. Thus we get the parameter space as in Chen and Bickel [4]

$$\Omega := \{W \in \mathbb{R}^{m \times m} : \forall\, 1 \leq k \leq m, \left|w_k\right| = 1, \max_{1 \leq l \leq m} |W_{k,l}| = \max_{1 \leq l \leq m} W_{k,l}, w_1 \prec w_2 \prec ... \prec w_m\}$$

where he have used the notation w_k to denote the kth row of the matrix W. We shall henceforth assume without loss of generality that $W_0 \in \Omega$. Also note that given any matrix W with non-zero rows, we can rearrange and rescale it's rows to find its Ω projection, which we shall denote by $[W]_\Omega$.

The independence of the components of S implies that the joint characteristic function of S factorizes into its components, that is $\phi_S(z) = \Pi_{k=1}^m \phi_{S_k}(z_k)$. For $W \in \Omega$, if we define $Z := WX$ we can measure the independence of all the components of Z using the difference between the characteristic function $\phi_Z(t)$ and $\prod_{k=1}^m \phi_{Z_k}(t_k)$. Note that Z has independent components if and only if WA is a rescaled, signed permutation matrix which further holds if and only if $W = W_0$, as $W \in \Omega$. Hence it suffices to try and factorize the joint characteristic function of the random variable Z as in (1.3).

Now (1.3) is non-negative and is equal to zero if and only if we have for all $z \in \mathbb{R}^m$ that $\mathbb{E}\exp(iz^T Z) = \Pi_{k=1}^m \mathbb{E}\exp(iz_k Z_k)$ which holds if and only if Z has independent components. As noted above it thus follows that (1.3) is zero if and only if $W = W_0$.

As noted by Matteson and Tsay [11], we can show that the components of Z are mutually independent if and only if for every $1 \leq k \leq m-1$ we have that Z_k and $Z_{k+1:m} = (Z_{k+1}, ..., Z_m)^T$ are independent random vectors. This result is useful because as Matteson and Tsay [11] showed we need only consider,

$$\sum_{k=1}^{m-1} \int_{s \in \mathbb{R}, t \in \mathbb{R}^{m-k}} \left|\mathbb{E}\exp(i(sZ_k + t^T Z_{k+1:m})) - \mathbb{E}\exp(isZ_k)\mathbb{E}\exp(it^T Z_{k+1:m})\right|^2 d\mu_k(s,t) \tag{3.1}$$

where for each $1 \leq k \leq m-1$, $\mu_k(s,t) = \omega(s,t)$ is the distance covariance measure in $\mathbb{R}^{m-k+1}$. By estimating the characteristic functions by the corresponding population quantities and we get the objective function used by Matteson and Tsay [11]

$$\sum_{k=1}^{m-1} \int_{s \in \mathbb{R}, t \in \mathbb{R}^{m-k}} \left|\hat{\varphi}_{Z_{k:m}}(s,t) - \hat{\varphi}_{Z_k}(s)\hat{\varphi}_{Z_{k+1:m}}(t)\right|^2 d\mu_k(s,t) \tag{3.2}$$

where $\hat{\varphi}_{Z_{k:m}}, \hat{\varphi}_{Z_k}, \hat{\varphi}_{Z_{k+1:m}}$ are the estimates of the characteristic functions from the observations $X_1, ..., X_n$. The above is precisely the technique used by Matteson and Tsay [11] for estimating W_0. However their results are only for the case where each component of S has finite variance.

3.1 Prewhitening and the Objective Function

From equation (3.2), optimize over the space of matrices in Ω. To estimate W_0 can be a large dimensional optimization that is difficult to implement. Prewhitening is used to reduce the size of the parameter space which facilitates the optimization. This is often a key first step in an ICA estimation procedure.

Let $X_1, ..., X_n$ be an i.i.d sample from the ICA model $X = AS$. Consider the sample covariance matrix $\hat{\Sigma}_X = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})^T$ where $\bar{X}$ is the mean of the X_j's and let $\hat{\Sigma}_X = UDU^T$ be it's SVD decomposition. Define $\hat{\Sigma}_X^{1/2} := UD^{1/2}U^T$ to be the square root matrix and $\hat{\Sigma}_X^{-1/2} = UD^{-1/2}U^T$. Here and

in further exposition, the square root of a generic symmetric matrix is one that is defined via SVD as shown above for $\hat{\Sigma}_X^{1/2}$.

Then let $Y_j := \hat{\Sigma}_X^{-1/2} X_j$ be the prewhitened data and note that the sample covariance of Y is the identity matrix. With $\mathbf{O}_n := \hat{\Sigma}_S^{-1/2} W_0 \hat{\Sigma}_X^{1/2}$, we have

$$\hat{\Sigma}_S^{-1/2} S_j = \mathbf{O}_n Y_j$$

Again, the sample covariance of $\hat{\Sigma}_S^{-1/2} S$ is the identity so that $I_m = \mathbf{O}_n \mathbf{O}_n^T$ which shows that $\mathbf{O}_n$ is orthogonal. Here is a more explicit verification of this fact:

$$\begin{aligned}\mathbf{O}_n \mathbf{O}_n^T &= \hat{\Sigma}_S^{-1/2} W_0 \hat{\Sigma}_X^{1/2} \hat{\Sigma}_X^{1/2} W_0^T \hat{\Sigma}_S^{-1/2} = \hat{\Sigma}_S^{-1/2} W_0 \hat{\Sigma}_X W_0^T \hat{\Sigma}_S^{-1/2} \\ &= \hat{\Sigma}_S^{-1/2} \hat{\Sigma}_S \hat{\Sigma}_S^{-1/2} = I_m\end{aligned}$$

We expect $\hat{\Sigma}_S^{-1/2}$ to converge to the diagonal matrix $\Sigma_S^{-1/2}$ with each diagonal entry being the inverse standard deviation if the second moment is finite, or zero in case the second moment is infinite. Specifically, the (k,k)th entry of $\Sigma_S^{-1/2}$ is $1/\sqrt{\mathrm{Var}(S^{(k)})}$ if $\mathrm{Var}(S^{(k)}) < \infty$ and 0 otherwise.

Due to Lemma A.2, $\hat{\Sigma}_S^{1/2}$ is approximately a diagonal matrix for large n. Thus $\hat{\Sigma}_S^{1/2} \mathbf{O}_n$ is approximately a rescaled orthogonal matrix and we can estimate $W_0 = \hat{\Sigma}_S^{1/2} \mathbf{O}_n \hat{\Sigma}_X^{-1/2}$ by $[O\hat{\Sigma}_X^{-1/2}]_\Omega$ where $O \in \mathcal{O}(m)$. Thus, we minimize the objective function (3.2) over the "acting parameter" space

$$\hat{\Omega}_n := \left\{ \left[O\hat{\Sigma}_X^{-1/2} \right]_\Omega : O \in \mathcal{O}(m) \right\}$$

The significance of prewhitening is that now optimizing over $\hat{\Omega}_n$ is easier than the space Ω.

$$\hat{\Delta}_Y(O) := \sum_{k=1}^{m-1} \mathcal{I}_n(Y_k(O), Y_{k^+}(O))$$

where $\hat{\Delta}_Y(O) := \sum_{k=1}^{m-1} \mathcal{I}_n(Y_k(O), Y_{k^+}(O))$ is precisely the characteristic function objective in equation (3.2), with $Y_k(O)$ is the kth component of OY and $Y_{k^+}(O)$ denotes the $k+1, ..., m$ components of OY. We arrive at the optimization problem

$$\hat{O} := \underset{O \in \mathcal{O}(m)}{\arg\min}\, \hat{\Delta}_Y(O) \tag{3.3}$$

The estimate for W_0 is $[\hat{W}]_\Omega$ where $\hat{W} := \hat{O}\hat{\Sigma}_X^{-1/2}$. Now, $O\mathbf{O}_n^{-1}$ is orthogonal for $O \in \mathcal{O}(m)$ and $Z = WX = O\hat{\Sigma}_X^{-1/2} AS = O\mathbf{O}_n^{-1} \hat{\Sigma}_S^{-1/2} S$. Therefore minimizing over orthogonal matrices in (3.2) is equivalent to minimizing

$$\sum_{k=1}^{m-1} \int_{s \in \mathbb{R}, t \in \mathbb{R}^{m-k}} \left| \hat{\varphi}_{\tilde{Z}_{k:m}}(s,t) - \hat{\varphi}_{\tilde{Z}_k}(s) \hat{\varphi}_{\tilde{Z}_{k+1:m}}(t) \right|^2 d\mu_k(s,t) \tag{3.4}$$

over orthogonal matrices $O \in \mathcal{O}(m)$ where now $\tilde{Z} = O\hat{\Sigma}_S^{-1/2} S$. The minimizer of this is given by $\tilde{O} := \hat{O}\mathbf{O}_n^{-1}$. Thus, we study the function in equation (3.4) to get convergence of $\tilde{O}$.

3.2 Consistency of Estimator for ICA

To simplify notation in the objective function (3.4), for $L \in \mathbb{R}^{m \times m}$ a matrix, denote by $S_k(L) \in \mathbb{R}$ to be the kth row of L and $S_{k^+}(L) \in \mathbb{R}^{m-k}$ to be the components $\{k+1, ..., m\}$ of LS for $1 \le k \le m-1$. Define finally

$$\mathcal{J}(L) := \sum_{k=1}^{m-1} \mathcal{I}(S_k(L), S_{k^+}(L))$$

For an i.i.d sample $S_1, ..., S_n$ of S we can similarly define

$$\mathcal{J}_n(L) := \sum_{k=1}^{m-1} \mathcal{I}_n(S_k(L), S_{k^+}(L))$$

Then, the equation (3.4) is precisely $\mathcal{J}_n(O\hat{\Sigma}_S^{-1/2})$. We show that over the compact space of orthogonal matrices $O \in \mathcal{O}(m)$, $\mathcal{J}_n(O\hat{\Sigma}_S^{-1/2})$ converges uniformly to $\mathcal{J}(O\Sigma_S^{-1/2})$.

Theorem 3.1. *Consider a strictly stationary ergodic time series $\{S_j\}_{j\geq 1}$ with values in $\mathbb{R}^m$ with marginal distribution S such that $\mathbb{E}|S| < \infty$. Assume that the components of S are mutually independent, at most one component has infinite variance, at most one component is normal and that none of the components are degenerate. Then as $n \to \infty$*

$$\sup_{O\in\mathcal{O}(m)} \left|\mathcal{J}_n(O\hat{\Sigma}_S^{-1/2}) - \mathcal{J}(O\Sigma_S^{-1/2})\right| \xrightarrow{P} 0$$

Proof. Suffices to show for each $1 \leq k \leq m-1$ that as $n \to \infty$,

$$\sup_{O\in\mathcal{O}(m)} \left|\mathcal{I}_n(S_k(O\hat{\Sigma}_S^{-1/2}), S_{k^+}(O\hat{\Sigma}_S^{-1/2})) - \mathcal{I}(S_k(O\Sigma_S^{-1/2}), S_{k^+}(O\Sigma_S^{-1/2}))\right| \xrightarrow{P} 0$$

Take $p = 1$ and $q = m - k$. Note for any $O \in \mathcal{O}(m)$,

$$\begin{aligned}&\left|\mathcal{I}_n(S_k(O\hat{\Sigma}_S^{-1/2}), S_{k^+}(O\hat{\Sigma}_S^{-1/2})) - \mathcal{I}(S_k(O\Sigma_S^{-1/2}), S_{k^+}(O\Sigma_S^{-1/2}))\right| \\ &= \left|\mathcal{I}_n(U\hat{\Sigma}_S^{-1/2}S, V\hat{\Sigma}_S^{-1/2}S) - \mathcal{I}(U\Sigma_S^{-1/2}S, V\Sigma_S^{-1/2}S)\right|\end{aligned}$$

where $U \in \mathbb{R}^{p\times m}$ is the kth row of O and $V \in \mathbb{R}^{q\times m}$ is the submatrix consisting of rows $\{k+1, ..., m\}$ of O. It is clear that if we take a supremum over $O \in \mathcal{O}(m)$, the RHS above is bounded by the term in Theorem 2.1 with $M = 1$ and taking $X = Y = S$. Taking limit as $n \to \infty$ and using Theorem 2.1, with $\Gamma_n = \Lambda_n = \hat{\Sigma}_S^{-1/2}$ and $\Gamma = \Lambda = \Sigma_S^{-1/2}$, the proof is complete due to Lemma A.2 □

As discussed in Section 2, our estimator is $\hat{W} = \hat{O}\hat{\Sigma}_X^{-1/2}$ and we have $\hat{W} = \tilde{O}\hat{\Sigma}_S^{-1/2}W_0$. We already know the convergence of $\hat{\Sigma}_S^{-1/2}$ due to Lemma A.2. We will show in the proof of our main theorem that $\tilde{O} \xrightarrow{P} I_m$ as a result of the above Theorem 3.1. We now come to our main result:

Theorem 3.2. *Consider an ICA model $\{X_j\}_{j\geq 1}$ taking values in $\mathbb{R}^m$ where $X_j = AS_j$ for some invertible matrix A, $W_0 = A^{-1} \in \Omega$ and $\{S_j\}_{j\geq 1}$ is a strictly stationary ergodic time series with marginal distribution S such that S has mutually independent components and $\mathbb{E}|S| < \infty$. Assume at most one component of S has infinite variance, at most one component is normal and that none of the components are degenerate. Then as $n \to \infty$*

$$\left[\hat{W}\right]_\Omega \xrightarrow{P} W_0$$

Proof. Define,

$$\tilde{O} := \operatorname*{arg\,min}_{O\in\mathcal{O}(m)} \mathcal{J}_n(O\hat{\Sigma}_S^{-1/2})$$

We have previously shown in Section 3.1 that we can express $\hat{O} = \tilde{O}\mathbf{O}_n$ where $\mathbf{O}_n = \hat{\Sigma}_S^{-1/2}W_0\hat{\Sigma}_X^{1/2}$. Hence

$$\hat{W} = \tilde{O}\mathbf{O}_n\hat{\Sigma}_X^{-1/2} = \tilde{O}\hat{\Sigma}_S^{-1/2}W_0$$

We shall show that as $n \to \infty$, $\tilde{O}$ converges to the identity matrix or some permutation of the rows of the identity matrix.

Without loss of generality suppose that only the mth component of S has infinite variance. By the uniqueness of $W_0 \in \Omega$, $[\hat{W}]_\Omega \xrightarrow{P} W_0$.

To prove $\tilde{O} \xrightarrow{P} I_m$ we know from Theorem 3.1, as

$$\tilde{O} = \underset{O \in \mathcal{O}(m)}{\arg\min}\, \mathcal{J}_n(O\hat{\Sigma}_S^{-1/2})$$

converges to $O_* \in \arg\min_{O \in \mathcal{O}(m)} \mathcal{J}(O\Sigma_S^{-1/2})$. But $\mathcal{J}(O_*\Sigma_S^{-1/2}) \geq 0$ and due to Comon [5] and since $O_*\Sigma_S^{-1/2}$ has a zero last column, equality holds if and only if the top $m-1$ columns of $O_*\Sigma_S^{-1/2}$ is a sub-matrix of a rescaled signed permutation matrix, which holds if and only if the top $m-1$ columns of O_* is a sub-matrix of a signed permutation matrix. Furthermore this holds if and only if $O_* = I_m$ or its permutation, as required. □

3.3 ICA with Noise

We now consider a more realistic version of the ICA model where we now include noise. The model will now be $X_j = AS_j + r(n)U_j$ where $U_j \in \mathbb{R}^m$ is the noise with $\mathbb{E}U_j = \mathbf{0}$ and $r(n) \in \mathbb{R}$. We will allow for dependence between U_j and S_j.

Denote by $X^0 = AS$ the noiseless version of the ICA. The objective function with no noise is given by

$$\hat{\Delta}_{X^0}(O) := \sum_{k=1}^{m-1} \mathcal{I}_n(O_k\hat{\Sigma}_{X^0}^{-1/2}X^0, O_{k^+}\hat{\Sigma}_{X^0}^{-1/2}X^0)$$

The objective function for the noisy version we consider is

$$\hat{\Delta}_X(O) := \sum_{k=1}^{m-1} \mathcal{I}_n(O_k\hat{\Sigma}_X^{-1/2}X, O_{k^+}\hat{\Sigma}_X^{-1/2}X)$$

and our estimate for $W_0 = A^{-1}$ is $\hat{W} = \hat{O}\hat{\Sigma}_X^{-1/2}$. By putting conditions on the noise we show that $\hat{\Delta}_X(\cdot)$ is very close to $\hat{\Delta}_{X^0}(\cdot)$ and thus get consistent estimates for W_0.

Theorem 3.3. *Consider a noisy ICA model $\{X_j\}_{j\geq 1}$ taking values in $\mathbb{R}^m$ where $X_j = AS_j + r(n)U_j$ for some invertible matrix A, $W_0 = A^{-1} \in \Omega$, $\{(S_j, U_j)\}_{j\geq 1}$ is a strictly stationary ergodic time series with marginal distribution (S, U) such that S has mutually independent components, $\mathbb{E}|S| < \infty$, $\mathbb{E}U = 0$ and $\mathbb{E}|S||U| < \infty$. Assume at most one component of S has infinite variance, at most one component is normal and that none of the components of S are degenerate. If $r(n) = o_P(1)$ and $||\hat{\Sigma}_X^{-1} - \hat{\Sigma}_{X^0}^{-1}|| \xrightarrow{P} 0$ then,*

1. $\sup_{O \in \mathcal{O}(m)} |\hat{\Delta}_X(O\hat{\Sigma}_X^{-1/2}) - \hat{\Delta}_{X^0}(O\hat{\Sigma}_{X^0}^{-1/2})| \xrightarrow{P} 0$
2. $[\hat{W}]_\Omega \xrightarrow{P} W_0$

We apply this theorem to the VAR(p) time series (1.2) example from Section 1. Consider for simplicity the VAR(1) model:

$$Y_t = \phi Y_{t-1} + e_t$$

Recall that $e_t = AS_t$ is the ICA model where S_t has independent components. From the observations of Y_t, one only has an estimate of ϕ - say the least squares estimator - denoted by $\hat{\phi}$. The ICA model thus becomes $\hat{e}_t = Y_t - \hat{\phi}_1 Y_{t-1} = e_t + (\phi - \hat{\phi})Y_{t-1}$. Or, $\hat{e}_t = AS_t + r(n)U_t$ where $U_t = Y_{t-1}$ and $r(n) = \phi - \hat{\phi}$. Assuming the conditions in the theorem above, we can see that in this case we can show that $\hat{W}$ is a consistent estimator for $W_0 = A^{-1}$. We thus are able to get consistent estimates for the unmixing matrix W_0 in the presence of a heavy-tailed shock even with the noise from estimating ϕ with $\hat{\phi}$.

References

[1] Joseph Anderson, Navin Goyal, Anupama Nandi, and Luis Rademacher. Heavy-tailed independent component analysis. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 290–309. IEEE, 2015.

[2] Joseph Anderson, Navin Goyal, Anupama Nandi, and Luis Rademacher. Heavy-tailed analogues of the covariance matrix for ica. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[3] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET, 1993.

[4] Aiyou Chen and Peter J Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

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

[6] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[7] Richard A Davis, Muneya Matsui, Thomas Mikosch, Phyllis Wan, et al. Applications of distance correlation to time series. *Bernoulli*, 24(4A):3087–3116, 2018.

[8] Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pages 665–672, 2002.

[9] Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. Independent component analysis. *Studies in informatics and control*, 11(2):205–207, 2002.

[10] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

[11] David S Matteson and Ruey S Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112(518):623–637, 2017.

[12] Klaus Nordhausen and Hannu Oja. Independent component analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1440, 2018.

[13] R Ranga Rao. Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, pages 659–680, 1962.

[14] Richard J Samworth, Ming Yuan, et al. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973–3002, 2012.

[15] Gábor J Székely and Maria L Rizzo. Brownian distance covariance. *The annals of applied statistics*, pages 1236–1265, 2009.

[16] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

A Appendix: Proofs

We intend to generalize the results of Székely et al. [16] and use it to prove our main theorem. We will follow the basic outline of the proof of Theorem 2 in Székely et al.[16].

Proof of Theorem 2.1 (a). Let $(U,V) \in C_M$. Define for $t \in \mathbb{R}^p, s \in \mathbb{R}^q$

$$\xi_n(s,t;U,V) := \phi^n_{X,Y}(U^T s, V^T t) - \phi^n_X(U^T s)\phi^n_Y(V^T t)$$
$$\xi(s,t;U,V) := \phi_{X,Y}(U^T s, V^T t) - \phi_X(U^T s)\phi_Y(V^T t)$$

For each $0 < \delta < 1$, define the compact set

$$D(\delta) := \{(s,t) : \delta \le |s| \le 1/\delta, \delta \le |t| \le 1/\delta\} \tag{A.1}$$

We can break the proof of the theorem into three parts since,

$$\begin{aligned}
|\mathcal{I}_n(UX,VY) - \mathcal{I}(UX,VY)| &= \left| \int_{\mathbb{R}^{p+q}} \left|\xi_n(s,t;U,V)\right|^2 d\omega - \int_{\mathbb{R}^{p+q}} \left|\xi(s,t;U,V)\right|^2 d\omega \right| \\
&\le \left| \int_{D(\delta)} \left|\xi_n(s,t;U,V)\right|^2 d\omega - \int_{D(\delta)} \left|\xi(s,t;U,V)\right|^2 d\omega \right| \\
&+ \int_{D^c(\delta)} \left|\xi_n(s,t;U,V)\right|^2 d\omega + \int_{D^c(\delta)} \left|\xi(s,t;U,V)\right|^2 d\omega
\end{aligned}$$

and it suffices to show:

1. For $0 < \delta < 1$,

$$\limsup_{n\to\infty} \sup_{(U,V)\in C_M} \left| \int_{D(\delta)} \left|\xi_n(s,t;U,V)\right|^2 d\omega - \int_{D(\delta)} \left|\xi(s,t;U,V)\right|^2 d\omega \right| = 0 \ a.s. \tag{A.2}$$

2.

$$\limsup_{\delta\downarrow 0} \sup_{(U,V)\in C_M} \int_{D^c(\delta)} \left|\xi(s,t;U,V)\right|^2 d\omega = 0 \tag{A.3}$$

3.

$$\limsup_{\delta\downarrow 0} \limsup_{n\to\infty} \sup_{(U,V)\in C_M} \int_{D^c(\delta)} \left|\xi_n(s,t;U,V)\right|^2 d\omega = 0 \ a.s. \tag{A.4}$$

Proof of (A.2)**:** Let $\mathcal{A}$ be the collection of functions $f : \mathbb{R}^{p+q} \to \mathbb{C}$ defined by $f(x,y) = \exp(i(s^T U x + t^T V y))$ where $(s,t) \in D(\delta), (U,V) \in C_M$. As

$$\sup_{(U,V)\in C_M,(s,t)\in D(\delta)} |s^T U + t^T V| \le 2M/\delta$$

by continuity of the exponential it follows that $\mathcal{A}$ is equicontinuous. Clearly each f is totally bounded by the constant function 1. Due to Theorem 6.2 in Rao [13] it follows that as $n \to \infty$,

$$\sup_{(U,V)\in C_M,(s,t)\in D(\delta)} \left|\phi^n_{X,Y}(U^T s + V^T t) - \phi_{X,Y}(U^T s + V^T t)\right| \xrightarrow{a.s.} 0$$

Similarly we have as $n \to \infty$,

$$\sup_{(U,V)\in C_M,(s,t)\in D(\delta)} \left|\phi^n_X(U^T s) - \phi_X(U^T s)\right| \xrightarrow{a.s.} 0$$
$$\sup_{(U,V)\in C_M,(s,t)\in D(\delta)} \left|\phi^n_Y(V^T t) - \phi_S(V^T t)\right| \xrightarrow{a.s.} 0$$

Combining these three results above gives us

$$\sup_{(U,V)\in C_m,(s,t)\in D(\delta)} \left| |\xi_n(t,s;U,V)|^2 - |\xi(t,s;U,V)|^2 \right| \xrightarrow{a.s.} 0$$

as $n \to \infty$. As $\int_{(s,t)\in D(\delta)} d\omega < \infty$, $A.2$ follows from above.

Proof of (A.3)**:** Let $(U,V) \in C_M$ and $\delta > 0$. Note that,

$$\begin{aligned}
\xi(s,t;U,V) &= \left|\mathbb{E}\exp(i(s^TUX + t^TVY)) - \mathbb{E}\exp(is^TUX)\mathbb{E}\exp(it^TVY)\right|^2 \\
&\le (1 - \left|\mathbb{E}\exp(is^TUX)\right|^2)(1 - \left|\mathbb{E}\exp(it^TVY)\right|^2) \\
&= (1 - \mathbb{E}\exp(is^TU(X - X')))(1 - \mathbb{E}\exp(it^TV(Y - Y'))) \\
&= \mathbb{E}\left[1 - \cos\langle s, U(X - X')\rangle\right] \cdot \mathbb{E}\left[1 - \cos\langle t, V(Y - Y')\rangle\right]
\end{aligned}$$

For $y \in \mathbb{R}^{\ge 0}$, define

$$G(y) := \int_{z\in\mathbb{R}^p, |z|_p < y} \frac{1 - \cos(z_1)}{c_p|z|^{p+1}} dz$$

where z_1 denotes the first component of $z \in \mathbb{R}^p$. Due to Lemma 1 in [16], $G(y) \le 1$. Clearly G is an increasing function on $y \ge 0$ and $\lim_{y\downarrow 0} G(y) = 0$. We note the following result obtained via change of variables: for $\theta \in \mathbb{R}^p, y \ge 0$,

$$\int_{z\in\mathbb{R}^p, |z|<y} \frac{1 - \cos\langle\theta, z\rangle}{c_p|z|^{p+1}} dz = G(|\theta|y)|\theta|$$

Note that $\left|\mathcal{V}^2_{\cdot,\delta} - \mathcal{I}(UX, VY)\right| = \int_{D^c(\delta)} |\xi(s,t)|^2 \, d\omega$. Thus,

$$\begin{aligned}
\int_{D^c(\delta)} |\xi(s,t;U,V)|^2 \, d\omega \le &\int_{|s|<\delta} |\xi(s,t;U,V)|^2 \, d\omega + \int_{|s|>1/\delta} |\xi(s,t;U,V)|^2 \, d\omega \\
&+ \int_{|t|<\delta} |\xi(s,t;U,V)|^2 \, d\omega + \int_{|t|>1/\delta} |\xi(s,t;U,V)|^2 \, d\omega
\end{aligned}$$

For the first term we have

$$\begin{aligned}
\int_{|s|<\delta} |\xi(s,t;U,V)|^2 \, d\omega &\le \int_{|s|<\delta} \mathbb{E}\left[1 - \cos\langle s, U(X - X')\rangle\right] \frac{dt}{c_p|s|^{p+1}} \\
&\quad \cdot \int_{t\in\mathbb{R}^q} \mathbb{E}\left[1 - \cos\langle t, V(Y - Y')\rangle\right] \frac{ds}{c_q|t|^{q+1}} \\
&= \mathbb{E}\left[G(|U(X - X')|\,\delta)\,|U(X - X')|\right] \cdot \mathbb{E}\,|V(Y - Y')| \\
&\le \mathbb{E}\left[G(M|X - X'|\delta)M\,|X - X'|\right] \cdot M\mathbb{E}|Y - Y'|
\end{aligned}$$

where the equality above follows from Fubini and change of variables as mentioned above. The last inequality follows from the fact that G is an increasing function and that $(U,V) \in C_M$. By continuity of G at zero and DCT, it follows that as $\delta \downarrow 0$, the RHS above goes to zero uniformly over $(U,V) \in C_M$. Next,

$$\begin{aligned}
\int_{|s|>1/\delta} |\xi(s,t;U,V)|^2 \, d\omega &\le \int_{|s|>1/\delta} \mathbb{E}\left[1 - \cos\langle s, U(X - X')\rangle\right] \frac{ds}{c_p|s|^{p+1}} \\
&\quad \cdot \int_{t\in\mathbb{R}^{m-k}} \mathbb{E}\left[1 - \cos\langle t, V(Y - Y')\rangle\right] \frac{dt}{c_q|t|^{q+1}} \\
&\le \int_{|s|>1/\delta} \frac{dt}{c_p|s|^{p+1}} \cdot \mathbb{E}|V(Y - Y')| \\
&\le \int_{|s|>1/\delta} \frac{ds}{c_p|s|^{p+1}} M\mathbb{E}|Y - Y'|
\end{aligned}$$

Finally, $\int_{z\in\mathbb{R}^p,|z|>1/\delta}\frac{dz}{c_p|z|^{p+1}}=o(1)$ as $\delta\downarrow 0$ implies

$$\limsup_{\delta\downarrow 0}\limsup_{n\to\infty}\sup_{(U,V)\in C_M}\int_{|s|>1/\delta}|\xi(s,t;U,V)|^2\,d\omega=0\ \ a.s.$$

Similarly we can show that the integrals over $|t|<\delta$ and $|t|>1/\delta$ also go to zero uniformly over $(U,V)\in C_M$.

Proof of (A.4)**:** Set $u_j=\exp(i\langle s,UX_j\rangle)-\mathbb{E}\exp(i\langle t,UX\rangle)$ and $v_j=\exp(i\langle s,VY_j\rangle)-\mathbb{E}\exp(i\langle t,VY\rangle)$. Note that

$$\xi_n(s,t;U,V)=\frac{1}{n}\sum_{j=1}^n u_jv_j-\frac{1}{n}\sum_{j=1}^n u_j\frac{1}{n}\sum_{j=1}^n v_j$$

Applying $|x+y|^2\le 2|x|^2+2|y|^2$ and Cauchy-Schwarz,

$$|\xi_n(s,t;U,V)|^2\le\frac{4}{n}\sum_{j=1}^n|u_j|^2\frac{1}{n}\sum_{j=1}^n|v_j|^2$$

Now,

$$\begin{aligned}\int_{D^c(\delta)}\left|\xi_n(s,t;U,V)\right|^2d\omega\le&\int_{|s|<\delta}|\xi_n(s,t;U,V)|^2\,d\omega+\int_{|s|>1/\delta}|\xi_n(s,t;U,V)|^2\,d\omega\\&+\int_{|t|<\delta}|\xi_n(s,t;U,V)|^2\,d\omega+\int_{|t|>1/\delta}|\xi_n(s,t;U,V)|^2\,d\omega\end{aligned}\tag{A.5}$$

For the first term above,

$$\int_{|s|<\delta}|\xi_n(s,t;U,V)|^2\,d\omega\le\frac{4}{n}\sum_{j=1}^n\int_{|s|<\delta}\frac{|u_j|^2ds}{c_p|s|^{1+p}}\frac{1}{n}\sum_{j=1}^n\int_{\mathbb{R}^q}\frac{|v_j|^2dt}{c_q|t|^{1+q}}$$

Using $|v_j|^2=1+|\phi_S(V^Tt)|^2-e^{i\langle t,VS\rangle}\overline{\phi_S(V^Tt)}-e^{-i\langle t,VS\rangle}\phi_S(V^Tt)$,

$$\int_{\mathbb{R}^q}\frac{|v_j|^2dt}{c_q|t|^{1+q}}=2E_Y\left|V(Y_j-Y)\right|-\mathbb{E}\left|V(Y-\dot{Y})\right|\le 2M\left(|Y_j|+\mathbb{E}|Y|\right)$$

where E_Y is the expectation with respect to Y and $\dot{Y}$ is an independent copy of Y. Next, using change of variables,

$$\begin{aligned}\int_{|s|<\delta}\frac{|u_j|^2ds}{c_p|s|^{1+p}}=&2E_X[\left|U(X_j-X)\right|G(|U(X_j-X)|\delta)]-\mathbb{E}[\left|U(X-\dot{X})\right|G(|U(X-\dot{X})|\delta)]\\\le&2E_X[||U||\cdot|X_j-X|G(||U||\cdot|X_j-X|\delta)]\\\le&2ME_X[|X_j-X|G(M|X_j-X|\delta)]\end{aligned}$$

where E_X is the expectation with respect to X and $\dot{X}$ is an independent copy of X and we have used $||U||\le M$ and $G(\cdot)$ is non-decreasing. Therefore, from (A.5)

$$\int_{|s|<\delta}|\xi_n(s,t;U,V)|^2\,d\omega\le 4\frac{2M}{n}\sum_{j=1}^n(|Y_j|+\mathbb{E}|Y|)\frac{2M}{n}\sum_{j=1}^nE_X[|X_j-X|G(M|X_j-X|\delta)]$$

Now by ergodicity, since $\mathbb{E}[|X-\dot{X}|G(M|X-\dot{X}|\delta)]\le\mathbb{E}[|X-\dot{X}|]\le 2\mathbb{E}|X|<\infty$ and $\mathbb{E}|Y|<\infty$,

$$\limsup_{n\to\infty}\sup_{(U,V)\in C_M}\int_{|s|<\delta}|\xi_n(s,t;U,V)|^2\,d\omega\le 4\cdot 2^2M^2\cdot 2\mathbb{E}|Y|\cdot\mathbb{E}[|X-\dot{X}|G(M|X-\dot{X}|\delta)]\ \ a.s.$$

Therefore by the continuity of G at zero and Lebesgue's DCT,

$$\limsup_{\delta\downarrow 0}\limsup_{n\to\infty}\sup_{(U,V)\in C_M}\int_{|t|<\delta}|\xi_n(s,t;U,V)|^2d\omega = 0\ a.s.$$

Now for the second term in (A.5), $|u_j|^2 \le 4$ which implies $\frac{1}{n}\sum_{j=1}^n |u_j|^2 \le 4$. Hence,

$$\begin{aligned}\int_{|s|>1/\delta}\frac{|u_j|^2ds}{c_p|s|^{1+p}} &\le 16\int_{|s|>1/\delta}\frac{ds}{c_p|s|^{1+p}}\frac{1}{n}\sum_{j=1}^n\int_{\mathbb{R}^q}\frac{|v_j|^2dt}{c_q|t|^{1+q}}\\ &\le 16\int_{|s|>1/\delta}\frac{ds}{c_p|s|^{p+1}}\frac{2}{n}\sum_{j=1}^n M(|Y_j|+\mathbb{E}|Y|)\end{aligned}$$

Taking $n\to\infty$ and then $\delta\downarrow 0$, we have

$$\limsup_{\delta\downarrow 0}\limsup_{n\to\infty}\sup_{(U,V)\in C_M}\int_{|s|>1/\delta}|\xi_n(s,t;U,V)|^2d\omega = 0\ a.s.$$

One can apply a similar argument for the remaining terms in (A.5). □

Proof of Theorem 2.1 (b). First,

$$\begin{aligned}\left|\mathcal{I}_n(U\Gamma_nX,V\Lambda_nY)-\mathcal{I}(U\Gamma X,V\Lambda Y)\right| &\le |\mathcal{I}_n(U\Gamma_nX,V\Lambda_nY)-\mathcal{I}(U\Gamma_nX,V\Lambda_nY)|\\ &\quad+|\mathcal{I}(U\Gamma_nX,V\Lambda_nY)-\mathcal{I}(U\Gamma X,V\Lambda Y)|\end{aligned}$$

Almost surely there exists $L>0$ such that $||\Gamma_n||\le L$ and $||\Lambda_n||\le L$ for all $n\in\mathbb{N}$. With $\tilde{M}:=ML>0$,

$$\begin{aligned}\sup_{(U,V)\in C_M}\left|\mathcal{I}_n(U\Gamma_nX,V\Lambda_nY)-\mathcal{I}(U\Gamma X,V\Lambda Y)\right| &\le \sup_{(\tilde{U},\tilde{V})\in C_{\tilde{M}}}\left|\mathcal{I}_n(\tilde{U}X,\tilde{V}Y)-\mathcal{I}(\tilde{U}X,\tilde{V}Y)\right|\\ &\quad+\sup_{(U,V)\in C_M}|\mathcal{I}(U\Gamma_nX,V\Lambda_nY)-\mathcal{I}(U\Gamma X,V\Lambda Y)|\end{aligned}$$

Using Theorem 2.1 the first term goes to zero almost surely. We need only prove

$$\sup_{(U,V)\in C_M}|\mathcal{I}(U\Gamma_nX,V\Lambda_nY)-\mathcal{I}(U\Gamma X,V\Lambda Y)|\xrightarrow{a.s.}0 \tag{A.6}$$

Let $(U,V)\in C_M$. For $0<\delta<1$, let $D(\delta)$ be as in (A.1). Then, almost surely

$$\begin{aligned}&\left|\mathcal{I}(U\Gamma_nX,V\Lambda_nY)-\mathcal{I}(U\Gamma X,V\Lambda Y)\right|\\ &\quad\le\left|\int_{D(\delta)}\left|\xi(U\Gamma_nX,V\Lambda_nY)\right|^2d\omega-\int_{D(\delta)}\left|\xi(U\Gamma X,V\Lambda Y)\right|^2d\omega\right|\\ &\qquad+\int_{D^c(\delta)}\left|\xi(U\Gamma_nX,V\Lambda_nY)\right|^2d\omega+\int_{D^c(\delta)}\left|\xi(U\Gamma X,V\Lambda Y)\right|^2d\omega\\ &\quad\le\int_{D(\delta)}\left|\left|\xi(U\Gamma_nX,V\Lambda_nY)\right|^2-\left|\xi(U\Gamma X,V\Lambda Y)\right|^2\right|d\omega\\ &\qquad+2\sup_{(\tilde{U},\tilde{V})\in C_{\tilde{M}}}\int_{D^c(\delta)}\left|\xi(\tilde{U}X,\tilde{V}Y)\right|^2d\omega\end{aligned}$$

Taking lim sup as $\delta\downarrow 0$ and using (A.3) we have the last term in the RHS of above going to zero almost surely. To complete the proof, we need only show for $0<\delta<1$,

$$\sup_{(U,V)\in C_M}\int_{D(\delta)}\left|\left|\xi(U\Gamma_nX,V\Lambda_nY)\right|^2-\left|\xi(U\Gamma X,V\Lambda Y)\right|^2\right|d\omega\xrightarrow{a.s.}0$$

Since the characteristic functions are uniformly continuous, given $\epsilon > 0$ there exists $\eta > 0$ such that whenever $|s - \tilde{s}| < \eta$ and $|t - \tilde{t}| < \eta$, we have

$$\left| \left| \phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t) \right|^2 - \left| \phi_{X,Y}(\tilde{s},\tilde{t}) - \phi_X(\tilde{s})\phi_Y(\tilde{t}) \right|^2 \right| < \epsilon'$$

where $\epsilon' > 0$ is chosen later. Almost surely, there exists N large enough such that whenever $n \geq N$, $||\Gamma_n - \Gamma|| \leq \delta\eta/M$ and $||\Lambda_n - \Lambda|| \leq \delta\eta/M$ hold almost surely. Then for $n \geq N$ we have almost surely $|\Gamma_n^T U^T s - \Gamma^T U^T s| < \eta$ and $|\Lambda_n^T V^T t - \Lambda^T V^T t| < \eta$. Thus, almost surely for $n \geq N$

$$\sup_{(U,V)\in C_M} \sup_{(s,t)\in D(\delta)} \left| \left| \xi(s,t;U\Gamma_n, V\Lambda_n) \right|^2 - \left| \xi(s,t;U\Gamma, V\Lambda) \right|^2 \right| < \epsilon'$$

Choose $\epsilon' = \epsilon / \int_{D(\delta)} d\omega > 0$. Thus whenever $n \geq N$, almost surely

$$\sup_{(U,V)\in C_M} \int_{D(\delta)} \left| \left| \xi(U\Gamma_n X, V\Lambda_n Y) \right|^2 - \left| \xi(U\Gamma X, V\Lambda Y) \right|^2 \right| d\omega \leq \epsilon' \int_{D(\delta)} d\omega = \epsilon$$

□

Lemma A.1. *Let $\Gamma_n \in \mathbb{R}^{m\times m}$ be a sequence of symmetric random matrices converging in probability to a non-random symmetric matrix Γ. Then $\Gamma_n^{1/2} \xrightarrow{P} \Gamma^{1/2}$.*

Proof of Lemma A.1. We have the SVD $\Gamma_n = U_n^T D_n U_n$ where D_n is a diagonal matrix such that the entries of the diagonal are non-increasing. Let $\Gamma = V^T D V$ be the eigenvalue decomposition with entries of D non-increasing. As the eigenvalue is a continuous functions of the entries of Γ_n, we have due to hypothesis that $D_n \xrightarrow{P} D$. Suppose we take an arbitrary subsequence of $\mathbb{N}$. As the sequence of U_n's belong to $\mathcal{O}(m)$ which is compact, consider a further subsequence $\{n_k\}_{k\geq 1}$ along which U_{n_k}'s converge to (a possibly random) matrix U_* almost surely and as $l \to \infty$, $\Gamma_{n_k} \xrightarrow{a.s.} \Gamma$ and $D_{n_k} \xrightarrow{a.s.} D$. But, $\Gamma_{n_k} = U_{n_k}^T D_{n_k} U_{n_k} \xrightarrow{a.s.} U_*^T D U_*$ so that $U_*^T D U_* = V^T D V$ or $D(U_* V^T) = (U_* V^T) D$. We claim that the last equality implies that $D^{1/2}(U_* V^T) = (U_* V^T) D^{1/2}$, so that $U_*^T D^{1/2} U_* = V^T D^{1/2} V = \Gamma^{1/2}$ and hence

$$\Gamma_{n_k}^{1/2} = U_{n_k}^T D_{n_k}^{1/2} U_{n_k} \xrightarrow{a.s.} U_*^T D^{1/2} U_* = \Gamma^{1/2}$$

To finish the proof we need to show $D^{1/2}(U_* V^T) = (U_* V^T) D^{1/2}$ when $D(U_* V^T) = (U_* V^T) D$ holds. To see this, for notational ease set $\tilde{U} := U_* V^T$ and denote its columns by $\tilde{u}_1, ..., \tilde{u}_m$. We then have for $1 \leq i \leq m$ that $D\tilde{u}_i = D_{i,i}\tilde{u}_i$. Thus for $1 \leq j \leq m$, $D_{j,j}\tilde{u}_{i,j} = D_{i,i}\tilde{u}_{i,j}$ which then implies that $\sqrt{D_{j,j}}\tilde{u}_{i,j} = \sqrt{D_{i,i}}\tilde{u}_{i,j}$ for every $1 \leq i, j \leq m$. Thus $D^{1/2}\tilde{U} = \tilde{U} D^{1/2}$. □

Lemma A.2. *Let $S \in \mathbb{R}^m$ is a random vector. Assume that the components of S are mutually independent, none of the components are degenerate and that at most one component has infinite variance. If $\{S_j\}_{j\geq 1}$ is a strictly stationary ergodic time series with stationary distribution S then as $n \to \infty$,*

$$\hat{\Sigma}_S^{-1/2} \xrightarrow{P} \Sigma_S^{-1/2}$$

Proof of Lemma A.2. If all the components of S have finite variance then we know from the ergodic theorem that $\hat{\Sigma}_S \xrightarrow{P} \Sigma_S$ so that $\hat{\Sigma}_S^{-1} \xrightarrow{P} \hat{\Sigma}_S^{-1}$. Due to Lemma A.1 we are done.
Now suppose that only the mth component has infinite variance. Without loss of generality assume that the first $m-1$ components are non-zero and that the entries of Σ_S^{-1} are non-increasing along the diagonal. We know by ergodicity that as $n \to \infty$, $\hat{\Sigma}_{S_{1:m-1}}^{-1} \xrightarrow{P} \Sigma_{S_{1:m-1}}^{-1}$ where $\hat{\Sigma}_{S_{1:m-1}}$ is the sample standard variance of

the first $m-1$ components of S and $\Sigma_{S_{1:m-1}}$ is the diagonal matrix of variance of the first $m-1$ components of S. Let $A = \hat{\Sigma}_{S_{1:m-1}}$ and b, c be such that $\hat{\Sigma}_S = \begin{pmatrix} A & b \\ b^T & c \end{pmatrix}$. Note that,

$$\hat{\Sigma}_S^{-1} = \begin{pmatrix} A^{-1} + A^{-1}bb^TA^{-1}/r & -A^{-1}b/r \\ -b^TA^{-1}/r & 1/r \end{pmatrix}$$

where $r = c - b^TA^{-1}b$. Note that for $1 \le t \le m-1$, $b_t = \frac{1}{n}\sum_{j=1}^n S_j^{(t)}S_j^{(m)}$. By the ergodic theorem, $b_j \xrightarrow{a.s.} \mathbb{E}S^{(t)}\mathbb{E}S^{(m)}$ so that $b^T \xrightarrow{a.s.} \mathbb{E}S^{(m)}\left(\mathbb{E}S^{(1)}, \mathbb{E}S^{(2)}, ..., \mathbb{E}S^{(m-1)}\right)^T$. Also, $\lim_{n\to\infty} c = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^n \left(S_j^{(m)}\right)^2 = \infty$ almost surely. We know $A^{-1} = \hat{\Sigma}_{S_{1:m-1}}^{-1} \xrightarrow{P} \Sigma_{S_{1:m-1}}^{-1}$. Thus, $r \xrightarrow{a.s.} \infty$, and hence $\hat{\Sigma}_S^{-1} \xrightarrow{P} \Sigma_S^{-1}$. Applying Lemma A.1 finishes the proof. □

Proof of Theorem 3.3. We first show a stronger result, that under $r(n) \xrightarrow{a.s.} 0$,

$$\sup_{(U,V)\in C_M} |\mathcal{I}_n(UX, VX) - \mathcal{I}_n(UX^0, VX^0)| \xrightarrow{a.s.} 0 \tag{A.7}$$

where $C_M = \{(U,V) : U \in \mathbb{R}^{p\times m}, V \in \mathbb{R}^{q\times m}, ||U|| \le M, ||V|| \le M\}$. Using the subsequence argument for the case $r(n) = o_P(1)$, this proves part 1 of the theorem. We make the claim:

$$\limsup_{\delta\downarrow 0} \sup_{(U,V)\in C_M} \int_{D^c(\delta)} |\xi_n(s,t;U,V)|^2 d\omega(s,t) = 0 \ \ a.s.$$
$$\limsup_{\delta\downarrow 0} \sup_{(U,V)\in C_M} \int_{D^c(\delta)} |\xi_n^0(s,t;U,V)|^2 d\omega(s,t) = 0 \ \ a.s.$$

where

$$\xi_n(s,t;U,V) = \frac{1}{n}\sum_{j=1}^n e^{is^TUX_j + it^TVX_j} - \frac{1}{n}\sum_{j=1}^n e^{is^TUX_j} \cdot \frac{1}{n}\sum_{j=1}^n e^{it^TVX_j},$$
$$\xi_n^0(s,t;U,V) = \frac{1}{n}\sum_{j=1}^n e^{is^TUX_j^0 + it^TVX_j^0} - \frac{1}{n}\sum_{j=1}^n e^{is^TUX_j^0} \cdot \frac{1}{n}\sum_{j=1}^n e^{it^TVX_j^0}.$$

Note that from the proof of Theorem 2.1, specifically (A.4) the second equation above holds. For the first, appealing to the proof of (A.4) we need to show

$$\limsup_{\delta\downarrow 0}\limsup_{n\to\infty} \frac{1}{n}\sum_{j=1}^n E_{X^0}|X_j - X^0|G(M|X_j - X^0|\delta) = 0 \ \ a.s.$$

Since $G(\cdot)$ is increasing and bounded by 1, for every j by the triangle inequality

$$\begin{aligned} E_{X^0}[|X_j - X^0|G(M|X_j - X^0|\delta)] &\le E_{X^0}[|X_j^0 - X^0|G(M|X_j - X^0|)] + r|U_j| \\ &\le E_{X^0}[|X_j^0 - X^0|G(M|X_j^0 - X^0|\delta + rM|u_j|\delta)] + r|U_j| \end{aligned}$$

By a change of variables to polar coordinates, one can express

$$G(y) = \int_0^\pi \int_0^y \frac{1-\cos(r\cos\theta)}{\pi r^2}\sin^{m-2}(\theta)drd\theta$$

Taking the derivative under the integral sign,

$$G'(y) = \int_0^\pi \frac{1-\cos(y\cos\theta)}{\pi y^2}\sin^{m-2}(\theta)d\theta$$

which can be bounded by a positive constant c using the inequality $1-\cos(x) \leq x^2/2$. The mean value theorem gives $G(M|X_j^0 - X^0|\delta + rM|u_j|\delta) \leq G(M|X_j^0 - X^0|\delta) + crM|U_j|\delta$ so that

$$\begin{aligned} E_{X^0}[|X_j - X^0|G(M|X_j - X^0|\delta)] \leq E_{X^0}[|X_j^0 - X^0|G(M|X_j^0 - X^0|\delta)] \\ + rM|U_j| \cdot ||A|| \cdot E_S[|S_j^0 - S|] + r|U_j| \end{aligned}$$

Now $r(n) \xrightarrow{a.s.} 0$ and ergodic theorem shows, $\limsup_{n\to\infty} r(n)\frac{1}{n}\sum_{j=1}^n |U_j| = 0$ almost surely. Similarly, the ergodicity of $\{S_j, U_j\}$ shows $\limsup_{n\to\infty} r(n)\frac{1}{n}\sum_{j=1}^n |U_j| E_S[|S_j - S|] = 0$ almost surely. For the first term, it suffices to show

$$\limsup_{\delta\downarrow 0} \limsup_{n\to\infty} \frac{1}{n}\sum_{j=1}^n E_S[|S_j - S|G(M||A||\delta|S_j - S|)] = 0 \ \ a.s.$$

By the ergodic theorem,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{j=1}^n E_S[|S_j - S|G(M||A||\delta|S_j - S|)] = \mathbb{E}[|S - \dot{S}|G(M||A||\delta|S - \dot{S}|)]$$

where $\dot{S}$ is an independent copy of S. By the continuity of $G(\cdot)$ at zero and DCT we are done.

With the claim above, we are left to show for $0 < \delta < 1$ that

$$\sup_{(U,V)\in C_M} \int_{D(\delta)} |\xi_n(s,t;U,V) - \xi_n^0(s,t;U,V)|^2 d\omega(s,t) \xrightarrow{a.s.} 0$$

Note that for $(s,t) \in D(\delta)$ by the triangle inequality and $|e^{ix} - 1| \leq |x|$,

$$\begin{aligned} \left|\frac{1}{n}\sum_{j=1}^n \left(e^{is^TUX_j + it^TVX_j} - e^{is^TUX_j^0 + it^TVX_j^0}\right)\right| &\leq \frac{1}{n}\sum_{j=1}^n \left|e^{irs^TUU_j + irt^TVU_j} - 1\right| \\ &\leq \frac{r}{n}\sum_{j=1}^n |s^TU + t^TV||U_j| \\ &\leq \frac{M}{\delta}\cdot\frac{r}{n}\sum_{j=1}^n |U_j| \end{aligned}$$

Similarly $\left|\frac{1}{n}\sum_{j=1}^n \left(e^{is^TUX_j} - e^{is^TUX_j^0}\right)\right| \leq \frac{M}{\delta}\frac{r}{n}\sum_{j=1}^n |U_j|$ and $\left|\frac{1}{n}\sum_{j=1}^n \left(e^{it^TVX_j} - e^{it^TVX_j^0}\right)\right| \leq \frac{M}{\delta}\frac{r}{n}\sum_{j=1}^n |U_j|$. But $r(n) \xrightarrow{a.s.} 0$ and the ergodic theorem implies $\frac{r}{n}\sum_{j=1}^n |U_j| \xrightarrow{a.s.} 0$. Thus,

$$\sup_{(U,V)\in C_M} \sup_{(s,t)\in D(\delta)} \left||\xi_n(s,t;U,V)|^2 - |\xi_n^0(s,t;U,V)|^2\right| \xrightarrow{a.s.} 0$$

so that

$$\begin{aligned} &\sup_{(U,V)\in C_M} \int_{D(\delta)} \left||\xi_n(s,t;U,V)|^2 - |\xi_n^0(s,t;U,V)|^2\right| d\omega(s,t) \\ &\leq \sup_{(U,V)\in C_M} \sup_{(s,t)\in D(\delta)} \left||\xi_n(s,t;U,V)|^2 - |\xi_n^0(s,t;U,V)|^2\right| \int_{D(\delta)} d\omega \end{aligned}$$

goes to zero almost surely. Thus (A.7) holds.

If $\hat{\Sigma}_{X^0}^{-1/2} \xrightarrow{P} \Sigma_{X^0}^{-1/2}$ and $\hat{\Sigma}_{X}^{-1/2} \xrightarrow{P} \Sigma_{X^0}^{-1/2}$, then via Theorem 2.1 we get

$$\sup_{(U,V)\in C_M} |\mathcal{I}_n(U\hat{\Sigma}_{X^0}^{-1/2}X^0, V\hat{\Sigma}_{X^0}^{-1/2}X^0) - \mathcal{I}(U\Sigma_{X^0}^{-1/2}X^0, V\Sigma_{X^0}^{-1/2}X^0)| \xrightarrow{P} 0$$

$$\sup_{(U,V)\in C_M} |\mathcal{I}_n(U\hat{\Sigma}_{X}^{-1/2}X^0, V\hat{\Sigma}_{X}^{-1/2}X^0) - \mathcal{I}(U\Sigma_{X^0}^{-1/2}X^0, V\Sigma_{X^0}^{-1/2}X^0)| \xrightarrow{P} 0$$

Combining both these results show us that

$$\sup_{(U,V)\in C_M} |\mathcal{I}_n(U\hat{\Sigma}_{X}^{-1/2}X^0, V\hat{\Sigma}_{X}^{-1/2}X^0) - \mathcal{I}_n(U\hat{\Sigma}_{X^0}^{-1/2}X^0, V\hat{\Sigma}_{X^0}^{-1/2}X^0)| \xrightarrow{P} 0 \tag{A.8}$$

We thus need to prove $\hat{\Sigma}_{X^0}^{-1/2} \xrightarrow{P} \Sigma_{X^0}^{-1/2}$ and $\hat{\Sigma}_{X}^{-1/2} \xrightarrow{P} \Sigma_{X^0}^{-1/2}$. Due to Lemma A.1 it suffices to show $\hat{\Sigma}_{X^0}^{-1} \xrightarrow{P} \Sigma_{X^0}^{-1}$ and $\hat{\Sigma}_{X}^{-1} \xrightarrow{P} \Sigma_{X^0}^{-1}$. The first one is an easy result due to

$$\hat{\Sigma}_{X^0}^{-1} = W_0^T \hat{\Sigma}_S^{-1} W_0 \xrightarrow{P} W_0^T \Sigma_S^{-1} W_0 = \Sigma_{X^0}^{-1}$$

Finally, $||\hat{\Sigma}_{X}^{-1} - \hat{\Sigma}_{X^0}^{-1}|| \xrightarrow{P} 0$ implies $\hat{\Sigma}_{X}^{-1} \xrightarrow{P} \Sigma_{X^0}^{-1}$.

Now by triangle inequality

$$|\hat{\Delta}_X(O) - \hat{\Delta}_{X^0}(O)| \leq \sum_{k=1}^{m-1} |\mathcal{I}_n(O_k\hat{\Sigma}_{X}^{-1/2}X, O_{k+1:m}\hat{\Sigma}_{X}^{-1/2}X) - \mathcal{I}_n(O_k\hat{\Sigma}_{X^0}^{-1/2}X^0, O_{k+1:m}\hat{\Sigma}_{X^0}^{-1/2}X^0)|$$

Taking for each fixed k, U to be kth row of O and V to be the $k+1, .., m$th rows of O and apply (A.8) we get

$$\sup_{(U,V)\in C_M} |\hat{\Delta}_X(O) - \hat{\Delta}_{X^0}(O)| = o_P(1)$$

This finishes the proof of Part 1. For Part 2, we define $\hat{O}^0$ and $\hat{W}^0$ to be the noiseless estimates from the previous sections. Then,

$$\hat{W} = (\hat{O} - \hat{O}^0)\hat{\Sigma}_X^{-1/2} + \hat{W}^0(\hat{\Sigma}_{X^0}^{1/2}\hat{\Sigma}_X^{-1/2})$$

Due to Part 1 of this theorem $\hat{O} - \hat{O}^0 \xrightarrow{P} \mathbf{0}$ and we have already seen $\hat{\Sigma}_{X^0}^{1/2}\hat{\Sigma}_X^{-1/2} \xrightarrow{P} I$. From Theorem 3.2 we conclude that $[\hat{W}^0]_\Omega \xrightarrow{P} W_0$ and we thus thus get consistent estimation of W_0 since it now follows that $[\hat{W}]_\Omega \xrightarrow{P} W_0$. □