Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Independent Component Analysis via Distance Covariance

David S. Matteson & Ruey S. Tsay

To cite this article: David S. Matteson & Ruey S. Tsay (2016): Independent Component Analysis via Distance Covariance, Journal of the American Statistical Association, DOI: 10.1080/01621459.2016.1150851

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1150851

Accepted author version posted online: 22 Mar 2016.
Published online: 22 Mar 2016.

Independent Component Analysis via Distance Covariance

David S. Matteson and Ruey S. Tsay *

March 14, 2016

ABSTRACT

This paper introduces a novel statistical framework for independent component analysis (ICA) of multivariate data. We propose methodology for estimating mutually independent components, and a versatile resampling-based procedure for inference, including misspecification testing. Independent components are estimated by combining a nonparametric probability integral transformation with a generalized nonparametric whitening method based on distance covariance that simultaneously minimizes all forms of dependence among the components. We prove the consistency of our estimator under minimal regularity conditions and detail conditions for consistency under model misspecification, all while placing assumptions on the observations directly, not on the latent components. U-statistics of certain Euclidean distances between sample elements are combined to construct a test statistic for mutually independent components. The proposed measures and tests are based on both necessary and sufficient conditions for mutual independence. We demonstrate the improvements of the proposed method over several competing methods in simulation studies, and we apply the proposed ICA approach to two real examples and contrast it with principal component analysis.

KEY WORDS: Misspecification testing; Multivariate analysis; Mutual independence test; Nonparametric statistics; Principal component analysis; U-Statistics.

Short title: Independent Component Analysis

*Matteson is Assistant Professor, Department of Statistical Science, Cornell University, 1196 Comstock Hall, Ithaca, NY 14853 (Email: matteson@cornell.edu; Web: http://www.stat.cornell.edu/~matteson/). Tsay is H.G.B. Alexander Professor of Econometrics & Statistics, Booth School of Business, University of Chicago, 5807 South Woodlawn Avenue, Chicago, IL 60637 (Email: ruey.tsay@chicagobooth.edu; Web: http://faculty.chicagobooth.edu/ruey.tsay/). Mattesons research is supported in part by a Xerox PARC Faculty Research Award and National Science Foundation grants CMMI-0926814 and DMS-1455172. Tsays research is supported in part by the University of Chicago Booth School of Business.

We would like to acknowledge the Editor, Associate Editor, and two anonymous reviewers for their helpful comments throughout the review process.

1. INTRODUCTION

Most naturally occurring processes are inherently multivariate in their origination. Simultaneous analysis of multiple random variables reveals insights about the relationship between variables. This leads to more compelling analysis than marginal consideration of the components alone. Multivariate analysis is considerably more complicated than univariate analysis, especially when the assumption of multivariate normality does not apply. Methods for reducing the complexity of multivariate observations become essential because of the curse of dimensionality. Independent component analysis (ICA) is a means for finding a suitable representation of multivariate data. We propose methods for measuring mutual independence, introduce a novel statistical framework under minimal prior assumptions for estimation of independent components S from observations Y, establish consistency, and consider misspecification testing of the proposed ICA model.

In statistical analysis, orthogonal components are often used as representations of multivariate data. Principal component analysis (PCA) measures the strength of variabilities of orthogonal linear combinations of components. However, higher-order or nonlinear analyses are often needed to adequately approximate complex joint distributions. Curvilinear component analysis (Demartines and Herault 1997) is a nonlinear extension of PCA that preserves the proximity between observations in the d-dimensional input space as the main features are projected onto a r-dimensional ($r < d$) subspace. Canonical correlation analysis (Hotelling 1936) generalizes PCA to find linear relationships among two sets of variables. Multidimensional scaling (Borg and Groenen 2005) measures the dissimilarities between two sets of variables, but it typically does not consider higher order relationships.

To overcome these weaknesses, we consider modeling multivariate random variables with mutually independent components (ICs). ICA is a method of unsupervised statistical learning that evolved in computer science research on artificial neural networks. Hyvärinen et al. (2001) provide an extensive overview including discussion of non-Gaussianity, some algorithms for estimating ICs, and applications in blind source separation, feature extraction, medical signal processing (fMRI, ECG, EEG), and time series analysis. There are many approaches for ICA estimation, including information theory (see Hyvärinen and Oja 1997), the maximum likelihood principle (see

Hastie and Tibshirani 2003), generalized decorrelation (see Cardoso 1989; Bach and Jordan 2003), characteristic functions (see Eriksson and Koivunen 2003; Chen and Bickel 2005), and ranks (see Nordhausen et al. 2009; Ilmonen and Paindaveine 2011; Hallin and Mehta 2015).

Whereas principal components always exist for variables with finite second moments, a linear decomposition into independent components may not. In an important contrast with the vast majority of ICA developments, we give special consideration to the properties of the proposed methods when the ICA model is misspecified (see Stögbauer et al. 2004). This distinction makes the proposed approach more general, with much greater applicability. In particular, we establish conditions under which our estimators are shown to be consistent under model misspecification. U-statistics of certain Euclidean distances between sample elements are then combined to construct a robust test for mutual independence. Independence testing on estimated components has also been considered by Karvanen (2005); Gretton et al. (2007); Paindaveine et al. (2009); Wu et al. (2009).

We make two more consequential departures from the majority of existing ICA methods. First, our primary assumptions are placed on the observations Y to allow direct validation, not the latent components S. This setting also allows a general assessment to be made when the ICA model is misspecified. Second, our measures and tests of mutual independence are based on both necessary and sufficient conditions for mutual independence. Those based only on necessary conditions for mutual independence, such as Cardoso (1989), are clearly not robust to all forms of dependence. They simply provide no assurances in the estimation of mutually independent components and should be named or categorized as otherwise.

The linear ICA model is applicable in many situations, even when other linear projection methods such as PCA, factor analysis, or projection pursuit are not effective. When the ICA model is correctly specified, one may apply univariate analysis to study or model each component separately. Univariate models may then be combined to obtain a multivariate model for the original observations.

A linear structure model for vector observations Y is given by

$$Y = \boldsymbol{M}S, \tag{1}$$

in which $S = (s_1, \ldots, s_d)'$ is a random vector of the same dimension as Y, and $\boldsymbol{M}$ is a constant, nonsingular *mixing matrix*. The linear ICA model further supposes the univariate components $s_1, \ldots, s_d$ are mutually independent; however, this model is not identified. Traditionally, the components of S are also assumed to be non-Gaussian with finite variance, and without loss of generality, standardized such that $\mathrm{E}(s_k) = 0$ and $\mathrm{Var}(s_k) = 1$, for $k = 1, \ldots, d$. It is assumed that Y has a nonsingular distribution, although Y may represent a given low dimensional projection from some higher dimensional space. In general, the components cannot be identified if the dimension of S exceeds that of Y. When the model is misspecified, the observations Y are projected such that the components of $\boldsymbol{M}^{-1}Y$ are as close to mutually independent as possible, given the dependence measure employed.

In the setting described above, ICs may be estimated jointly or sequentially. Sequential methods, also referred to as deflationary, estimate the components of S one after another. Motivated by potential computational savings, the deflationary approach has been widely promoted in the machine learning literature. However, estimation uncertainty accumulates at each stage in the succession, and joint estimation will always have greater statistical efficiency. For the proposed methodology we also briefly compare the speed and accuracy of joint verses sequential estimation.

In Section 2 we introduce our methodology, propose measures for testing mutual independence, discuss model parameterization and identifiability, propose a versatile inferential framework based on resampling, and state conditions for the consistency of the proposed estimators. In Section 3 we compare the proposed method with several alternatives in simulation studies, detail our practical implementation and discuss empirical performance measures. In Section 4 we apply the proposed approach to two real examples and contrast it with PCA. Concluding remarks are in Section 5 and technical proofs follow in the Appendix.

2. METHODOLOGY

Let $\boldsymbol{Y} = \{Y_i : i = 1, \ldots, n\}$ be an iid sample from the joint distribution of a d-dimensional random vector Y. The theory and methods presented focus on the case in which $d \ll n$. We require Y to obey some regularity conditions regardless of whether the ICA model holds.

Assumption 2.1. *The random vector Y has a nonsingular, continuous distribution function F_Y, with* $\mathrm{E}|Y|^2 < \infty$.

We also make the simplifying assumption that $\mathrm{E}(Y) = 0$ throughout; however, observations are typically centered by their sample mean in practice.

2.1 Measuring Pairwise Multivariate Independence

Distance covariance $\mathcal{I}(Y^{(1)}, Y^{(2)})$ is a multivariate measure of independence between d_1- and d_2-dimensional random vectors $Y^{(1)}$ and $Y^{(2)}$, in which d_1 and d_2 are arbitrary, for all distributions with finite first absolute moments. Letting $|\cdot|$ denote Euclidean distance and let $(\dot{Y}^{(1)}, \dot{Y}^{(2)})$ and $(\ddot{Y}^{(1)}, \ddot{Y}^{(2)})$ denote iid copies of $(Y^{(1)}, Y^{(2)})$, Székely et al. (2007) show that distance covariance may be defined as

$$\begin{aligned}\mathcal{I}(Y^{(1)}, Y^{(2)}) \;=\; & \mathrm{E}|Y^{(1)} - \dot{Y}^{(1)}||Y^{(2)} - \dot{Y}^{(2)}| + \mathrm{E}|Y^{(1)} - \dot{Y}^{(1)}|\mathrm{E}|Y^{(2)} - \dot{Y}^{(2)}| \\ & -\mathrm{E}|Y^{(1)} - \dot{Y}^{(1)}||Y^{(2)} - \ddot{Y}^{(2)}| - \mathrm{E}|Y^{(1)} - \ddot{Y}^{(1)}||Y^{(2)} - \dot{Y}^{(2)}|.\end{aligned}$$

The following properties of $\mathcal{I}$ are the most relevant for ICA: $0 \le \mathcal{I}(Y^{(1)}, Y^{(2)})$; $\mathcal{I}$ is invariant to the group of orthogonal transformations such that $\mathcal{I}(a_1 + b_1\boldsymbol{C}_1Y^{(1)}, a_2 + b_2\boldsymbol{C}_2Y^{(1)}) = \sqrt{|b_1||b_2|}\mathcal{I}(Y^{(1)}, Y^{(2)})$ for all constant vectors a_1, a_2, non-zero scalars b_1, b_2, and orthogonal matrices $\boldsymbol{C}_1, \boldsymbol{C}_2$, of conforming dimensions, respectively; and finally, $\mathcal{I}(Y^{(1)}, Y^{(2)}) = 0$ if and only if $Y^{(1)}$ and $Y^{(2)}$ are independent. Several additional properties of distance covariance, including uniqueness, are discussed in Székely and Rizzo (2012, 2013).

Let ϕ_1 and ϕ_2 denote the characteristic functions of $Y^{(1)}$ and $Y^{(2)}$, respectively, and let $\phi_{1,2}$ denote the joint characteristic function of $Y^{(1)}$ and $Y^{(2)}$. Distance covariance measures the distance between the joint characteristic function and the product of the marginal characteristic functions. It can be applied to test the following hypothesis of independence

$$H_0 : \phi_{1,2}(t) = \phi_1(t_1)\phi_2(t_2) \quad \text{vs.} \quad H_A : \phi_{1,2}(t) \neq \phi_1(t_1)\phi_2(t_2), \quad \forall t_1 \in \mathbb{R}^{d_1}, t_2 \in \mathbb{R}^{d_2},$$

in which $t = (t_1', t_2')'$. The condition stated in H_0 above is both a necessary and sufficient condition for pair-wise multivariate independence.

Let $(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) = \{(Y_i^{(1)}, Y_i^{(2)}) : i = 1, \ldots, n\}$ be an iid sample from the joint distribution of random vectors $Y^{(1)}$ and $Y^{(2)}$, with $\mathrm{E}(|Y^{(1)}| + |Y^{(2)}|) < \infty$. We define an empirical multivariate independence measure as

$$\mathcal{I}_n(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) = T_{1,n}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) + T_{2,n}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) - T_{3,n}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}), \tag{2}$$

which is a function of U-statistics, with

$$\begin{aligned}
T_{1,n}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) &= \binom{n}{2}^{-1} \sum_{i<j} \left|Y_i^{(1)} - Y_j^{(1)}\right| \left|Y_i^{(2)} - Y_j^{(2)}\right|, \\
T_{2,n}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) &= \left[\binom{n}{2}^{-1} \sum_{i<j} \left|Y_i^{(1)} - Y_j^{(1)}\right|\right] \left[\binom{n}{2}^{-1} \sum_{i<j} \left|Y_i^{(2)} - Y_j^{(2)}\right|\right], \text{ and} \\
T_{3,n}(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) &= \binom{n}{3}^{-1} \sum_{i<j<k} \frac{1}{3}\Big(\left|Y_i^{(1)} - Y_j^{(1)}\right| \left|Y_i^{(2)} - Y_k^{(2)}\right| + \left|Y_i^{(1)} - Y_k^{(1)}\right| \left|Y_i^{(2)} - Y_j^{(2)}\right| \\
&\qquad + \left|Y_i^{(1)} - Y_j^{(1)}\right| \left|Y_j^{(2)} - Y_k^{(2)}\right| + \left|Y_j^{(1)} - Y_k^{(1)}\right| \left|Y_i^{(2)} - Y_j^{(2)}\right| \\
&\qquad + \left|Y_i^{(1)} - Y_k^{(1)}\right| \left|Y_j^{(2)} - Y_k^{(2)}\right| + \left|Y_j^{(1)} - Y_k^{(1)}\right| \left|Y_i^{(2)} - Y_k^{(2)}\right|\Big),
\end{aligned}$$

respectively. For more extensive discussion on distance covariance, and an alternative, asymptotically equivalent, empirical measure based on V-statistics, see Székely and Rizzo (2009), from which we note $\lim_{n\to\infty} \mathcal{I}_n(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) \overset{a.s.}{=} \mathcal{I}(Y^{(1)}, Y^{(2)})$, as well as convergence in distribution of $n\mathcal{I}_n(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)})$ to a non-degenerate random variable, under H_0. Additionally, $\mathcal{I}_n$ is invariant to the same group of orthogonal transformations as $\mathcal{I}$.

2.2 Measuring and Testing for Mutual Independence

Here we denote by S an observed d-dimensional random vector. To test whether its univariate components are mutually independent, we propose a statistic based on distance covariance. Let $t = (t_1, \ldots, t_d)' \in \mathbb{R}^d$. A necessary and sufficient condition for $S = (s_1, \ldots, s_d)'$ to consist of mutually independent components is that $\phi_S(t) = \phi_{s_1}(t_1) \cdots \phi_{s_d}(t_d), \ \forall t \in \mathbb{R}^d$, in which ϕ_S is the joint and ϕ_{s_k}, $k = 1, \ldots, d$, are the marginal characteristic functions of S, respectively. Assuming S has a continuous distribution, let F_{s_k}, $k = 1, \ldots, d$, denote the continuous univariate marginal

distribution functions of S. When applied to the corresponding component of s_k, each function $F_{s_k} : \mathbb{R} \to [0,1]$ is a probability integral transformation (PIT). Let $u_k = F_{s_k}(s_k)$, then the marginal distributions for each transformed component u_k is Uniform(0,1). Further, S consists of mutually independent components if and only if $U = (u_1, \ldots, u_d)'$ does.

Let $k^+ = \{\ell : k < \ell \leq d\}$, that is k^+ denotes the indices $(k+1), \ldots, d$, and let $t_{k^+} = (t_{k+1}, \ldots, t_d)'$. We propose simultaneously testing the following joint hypotheses against the stated alternative

$$H_0 : \quad \phi_{u_k,u_{k^+}}(t_k, \ldots, t_d) = \phi_{u_k}(t_k)\phi_{u_{k^+}}(t_{k^+}), \quad \forall t \in \mathbb{R}^d, \quad \text{for all } k = 1, \ldots, d-1,$$
$$H_A : \quad \phi_{u_k,u_{k^+}}(t_k, \ldots, t_d) \neq \phi_{u_k}(t_k)\phi_{u_{k^+}}(t_{k^+}), \quad \forall t \in \mathbb{R}^d, \quad \text{for some } k = 1, \ldots, d-1.$$

Note that H_0 above is both a necessary and sufficient condition for S to consist of mutually independent components by the following Lemma.

Lemma 2.2. *For S, U, t, k^+ and t_{k^+} defined as above,*

$$|\phi_{u_1,\ldots,u_d}(t) - \phi_{u_1}(t_1) \cdots \phi_{u_d}(t_d)| \quad \leq \quad \sum_{k=1}^{d-1} |\phi_{u_k,u_{k^+}}(t_k, t_{k^+}) - \phi_{u_k}(t_k)\phi_{u_{k^+}}(t_{k^+})|, \quad \forall t \in \mathbb{R}^d.$$

The proof follows by the triangle inequality, the multiplicative property of absolute value, and the boundedness of characteristic functions.

Again let S denote an observed d-dimensional random vector, and let $\boldsymbol{S} = \{S_i : i = 1, \ldots, n\}$ denote an iid sample. Let $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_d$ be a partition of the elements of $\boldsymbol{S}$ into samples of the d marginal components, respectively. In practice, the marginal distribution functions of S are unknown, so we replace each PIT with its empirical counterpart. Specifically, for each component $\boldsymbol{S}_k$, we replace each observation $S_{i,k}$ with its normalized marginal rank. That is, each component-wise transformation $\widehat{\boldsymbol{U}}_k$ has elements defined as $\hat{u}_{i,k} = \frac{1}{n}\text{rank}\{S_{i,k} : S_{i,k} \in \boldsymbol{S}_k\}$, for each $k = 1, \ldots, d$. Finally, we define a test statistic for mutual independence as

$$\mathcal{U}_n(\boldsymbol{S}) = n \sum_{k=1}^{d-1} \mathcal{I}_n(\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{U}}_{k^+}). \tag{3}$$

For $d = 2$, $\mathcal{U}_n(\boldsymbol{S})$ is asymptotically distribution-free and its asymptotic distribution can be derived from Theorem 5 of Székely and Rizzo (2009) and the Glivenko-Cantelli theorem. For the more general case, the distribution of $\mathcal{U}_n(\boldsymbol{S})$ depends on the distribution of S, and in practice we

implement a permutation test. The null hypothesis of mutual independence is rejected for a large value of $\mathcal{U}_n(\boldsymbol{S})$. Similar to Székely et al. (2007), we note that if any subsets of S are dependent, then $\mathcal{U}_n(\boldsymbol{S}) \to \infty$ in probability, as $n \to \infty$. Hence, the proposed test of mutual independence is also statistically consistent against all types of dependence. Although this test is widely applicable, others may have more power under specific alternatives, see Székely et al. (2007). In Section 2.6 we apply this statistic for misspecification testing of the ICA model.

2.3 Parameterization and Identifiability

For theoretical and practical considerations it is convenient to work with uncorrelated random variables. Let $\boldsymbol{O}$ denote an *uncorrelating* matrix and let $Z = \boldsymbol{O}Y$ denote uncorrelated observations. When the ICA model (1) holds, the relationship between Z and S is then

$$S = \boldsymbol{M}^{-1}Y = \boldsymbol{M}^{-1}\boldsymbol{O}^{-1}Z = \boldsymbol{W}Z, \tag{4}$$

in which $\boldsymbol{W} = \boldsymbol{M}^{-1}\boldsymbol{O}^{-1}$ is referred to as the *separating* matrix. When the ICA model is misspecified, we seek to estimate components $\boldsymbol{W}Z$ which are as independent as possible.

Let $\boldsymbol{\Sigma}_Y = \mathrm{Cov}(Y)$ denote the covariance matrix of the random variable Y, which is assured to exist by Assumption 2.1. Let $\boldsymbol{\Upsilon}$ be the matrix of eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix of the corresponding eigenvalues of $\boldsymbol{\Sigma}_Y$, then take $\boldsymbol{O} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Upsilon}'$. Without loss of generality, we henceforth assume that $\mathrm{Cov}(Z) = \boldsymbol{I}_d$, the $d \times d$ identity matrix. Given the uncorrelated variable Z, Equation (4) implies that the separating matrix $\boldsymbol{W}$ is necessarily orthogonal, because $\boldsymbol{I} = \mathrm{Cov}(S) = \boldsymbol{W}\mathrm{Cov}(Z)\boldsymbol{W}' = \boldsymbol{W}\boldsymbol{W}'$. Therefore, $\boldsymbol{W}$ has $d(d-1)/2$ free elements, instead of d^2.

For $d \geq 2$, let $O(d)$ denote the group of all $d \times d$ orthogonal matrices and let $\mathcal{SO}(d)$ denote the subgroup (rotation group) with determinant equal to 1. Some relevant properties of $\mathcal{SO}(d)$ are discussed in Matteson and Tsay (2011). Let $\xi_1, \ldots, \xi_d$ denote the canonical basis of $\mathbb{R}^d$. Let $\boldsymbol{Q}_{ij}(\psi)$ denote a rotation of all vectors lying in the (ξ_i, ξ_j)-plane of $\mathbb{R}^d$ by an angle ψ, oriented such that the rotation from ξ_i to ξ_j is assumed to be positive. Specifically, for $i \neq j$, $\boldsymbol{Q}_{ij}(\psi)$ is a Givens (plane) rotation matrix, that is, the identity matrix $\boldsymbol{I}_d$ with the (i,i) and (j,j) elements replaced by $\cos(\psi)$, the (i,j) element replaced by $-\sin(\psi)$, and the (j,i) element replaced by $\sin(\psi)$.

Let θ denote a length $p = d(d-1)/2$ vectorized triangular array of rotation angles, indexed by $\{i, j : 1 \le i < j \le d\}$. Any rotation $\boldsymbol{W} \in \mathcal{SO}(d)$ can be written in the form

$$\boldsymbol{W}_\theta = \boldsymbol{Q}^{(d-1)} \cdots \boldsymbol{Q}^{(1)}, \quad \text{in which} \quad \boldsymbol{Q}^{(k)} = \boldsymbol{Q}_{k,d}(\theta_{k,d}) \cdots \boldsymbol{Q}_{k,k+1}(\theta_{k,k+1}).$$

Although such decompositions are not unique, the one given above has an important invariance property. Specifically, the kth row of $\boldsymbol{W}_\theta$ and the kth row of the partial product $\boldsymbol{Q}^{(k)} \cdots \boldsymbol{Q}^{(1)}$ coincide. Let $\theta^{(\ell:k)} = \{\theta_{i,j} : \ell \le i \le k, i < j \le d\}$, then for $S = \boldsymbol{W}_\theta Z$, we observe that the kth element of S only varies with the subset of angles in $\theta^{(1:k)}$. Let

$$\Theta = \left\{ \theta_{i,j} : \begin{cases} 0 \le \theta_{1,j} < 2\pi, & \\ 0 \le \theta_{i,j} < \pi, & i \ne 1. \end{cases} \right\}. \tag{5}$$

Then, there exists a unique inverse mapping of $\boldsymbol{W} \in \mathcal{SO}(d)$ into $\theta \in \Theta$, such that the mapping is assured to be continuous if either all elements on the main-diagonal of $\boldsymbol{W}$ are positive, or all elements of $\boldsymbol{W}$ are nonzero (see Matteson 2008).

There are two remaining non-identification issues regarding $\boldsymbol{M}$ and S because the sign and the order of the components are not identifiable. Let $\boldsymbol{P}_\pm$ denote a signed permutation matrix and note that model (1) is equivalent to

$$Y = \boldsymbol{M}\boldsymbol{P}_\pm'\boldsymbol{P}_\pm S = (\boldsymbol{M}\boldsymbol{P}_\pm')(\boldsymbol{P}_\pm S),$$

in which $\boldsymbol{P}_\pm S$ are new ICs and $\boldsymbol{M}\boldsymbol{P}_\pm'$ is the new mixing matrix. When identification of ICs up to a signed permutation is sufficient for modeling purposes we may construct an equivalence class and a canonical form for $\boldsymbol{W}$ to conduct inference (see Matteson and Tsay 2011). In general, the non-identification of the scale, sign and order of the ICs must all be taken into account when comparing different estimates; a metric which is invariant to all three is discussed in Section 3.

2.4 Estimation of Independent Components: dCovICA

Let $\boldsymbol{Y}$ be an iid sample from the joint distribution of the continuous random vector Y. In practice, $\boldsymbol{Y}$ is usually replaced by a centered version $\widehat{\boldsymbol{Y}}$, in which the sample mean vector is subtracted from each observation. An uncorrelated variable Z can be defined as $Z = \boldsymbol{O}Y$, in which $\boldsymbol{O}$ denotes an

uncorrelating matrix. In practice, $\mathrm{Cov}(Y)$ is unknown, however, under Assumption 2.1, the sample covariance provides a consistent estimate. That is, $\widehat{\mathrm{Cov}}_n(\boldsymbol{Y}) \xrightarrow{a.s.} \mathrm{Cov}(Y)$, as $n \to \infty$. Using the sample covariance we can approximate the uncorrelating matrix as $\widehat{\boldsymbol{O}}_n = \widehat{\mathrm{Cov}}_n(\boldsymbol{Y})^{-1/2}$, then define approximately uncorrelated observations as $\widehat{\boldsymbol{Z}}_n = \boldsymbol{Y}\widehat{\boldsymbol{O}}_n'$. This is done such that $\widehat{\mathrm{Cov}}_n(\widehat{\boldsymbol{Z}}_n) = \boldsymbol{I}_d, \ \forall n$, and $\mathrm{Cov}(\widehat{\boldsymbol{Z}}_n) \xrightarrow{a.s.} \boldsymbol{I}_d$, as $n \to \infty$.

To simplify notation, we omit the steps described above, and let $\boldsymbol{Z}$, an uncorrelated, mean zero, unit variance, iid sample, be given. We begin by estimating $\boldsymbol{W}_\theta$ via θ. Define $S(\theta) = \boldsymbol{W}_\theta Z$, $\boldsymbol{S}(\theta) = \boldsymbol{Z}\boldsymbol{W}_\theta'$, and let $\boldsymbol{S}_k(\theta)$ denote the kth component of $\boldsymbol{S}(\theta)$. Recall that, by the construction of $\boldsymbol{W}_\theta$, each $\boldsymbol{S}_k(\theta)$ only varies with the subset of angles in $\theta^{(1:k)}$, in which $\theta^{(\ell:k)} = \{\theta_{i,j} : \ell \le i \le k, i < j \le d\}$, and it is invariant to the complementary subset.

Recall $k^+ = \{\ell : k < \ell \le d\}$. To estimate mutually independent components, we define an objective function as

$$\mathcal{J}_n(\theta) = \sum_{k=1}^{d-1} \mathcal{I}_n(\boldsymbol{S}_k(\theta), \boldsymbol{S}_{k^+}(\theta)), \tag{6}$$

and we define the distance covariance ICA estimator (dCovICA) as $\widehat{\theta}_n = \mathrm{argmin}_\theta \mathcal{J}_n(\theta)$. Given an estimate of θ, the separating matrix is estimated as $\widehat{\boldsymbol{W}} = \boldsymbol{W}_{\widehat{\theta}_n}$ and the estimated ICs $\widehat{\boldsymbol{S}}$ are given by $\boldsymbol{S}(\widehat{\theta}_n) = \boldsymbol{Z}\boldsymbol{W}_{\widehat{\theta}_n}'$.

The objective function in Equation (6) has $d(d-1)/2$ parameters which can be estimated jointly. Alternatively, estimation may be performed conditionally in a sequence of $d-1$ minimization problems; the first will have $d-1$ parameters, the second $d-2$, continuing as such until the last, which will have one parameter. This follows by the orthogonal invariance property of $\mathcal{I}_n$ stated in Section 2.1. Specifically, let $\widehat{\theta}^{(1:1)} = \mathrm{argmin}_{\theta^{(1:1)}} \mathcal{I}_n(\boldsymbol{S}_1(\theta), \boldsymbol{S}_{1^+}(\theta))$, in which the elements $\theta^{(2:d)}$ are fixed, but arbitrary. Now, for $k = 2, \ldots, (d-1)$, given $\widehat{\theta}^{(1:(k-1))}$, let

$$\widehat{\theta}_n^{(k:k)} = \{\widehat{\theta}_{k,\ell} : k < \ell \le d\} = \underset{\theta^{(k:k)}}{\mathrm{argmin}}\, \mathcal{I}_n(\boldsymbol{S}_k(\theta), \boldsymbol{S}_{k^+}(\theta)), \tag{7}$$

in which $\theta^{(1:(k-1))}$ are fixed at $\widehat{\theta}_n^{(1:(k-1))}$ and all elements in $\theta^{((k+1):d)}$ are fixed, but arbitrary. Hence, the sequence of estimates from Equation (7), for $k = 1, \ldots, (d-1)$, exactly coincide with the joint estimate $\widehat{\theta}_n = \mathrm{argmin}_\theta \mathcal{J}_n(\theta)$. When the components are estimated in this sequential manner the later

component estimates are restricted to lie within the subspace orthogonal to the span of the earlier estimates, resulting in a tradeoff between computational complexity and statistical efficiency.

An Alternative Estimator: PITdCovICA In general, distance covariance depends on the marginal distributions of the inputs. As described in Section 2.2 (also see Rémillard 2009), for continuous random variables, this dependency can be removed by applying the PIT component-wise. As before, the marginal distribution functions F_{s_k} are unknown in practice, and the PIT must be approximated.

Our asymptotic results and our optimization algorithms rely explicitly on our objective function varying continuously in its arguments. This means that approximating F_{s_k} using the empirical cumulative distribution functions (CDF) will not be sufficient because it is a step function. Simply interpolating the empirical CDF between the steps is also insufficient. Instead, we require an estimate of F_{s_k} to depend on the location of all the observations $\{s_{i,k} : i = 1, \ldots, n\}$, not just their relative location. To assure this, we propose applying kernel smoothing to approximate the CDF of each F_{s_k} with a continuous function. Let

$$\widetilde{F}_{s_k,n,\tilde{h}_{k,n}}(s) = \frac{1}{n}\sum_{i=1}^{n} G\left(\frac{s - s_{i,k}}{\tilde{h}_{k,n}}\right) \tag{8}$$

in which G is the integral of a density kernel and $\tilde{h}_{k,n}$ is a data-dependent bandwidth for the kth component. The choice of G and the bandwidth are discussed below.

Given $\boldsymbol{Z}$, for $\boldsymbol{S}(\theta) = \boldsymbol{Z}\boldsymbol{W}'_\theta$, we define $\widetilde{\boldsymbol{U}}_k(\theta)$, as a continuous function of θ, such that $\tilde{u}_{i,k}(\theta) = \widetilde{F}_{s_k(\theta),n,\tilde{h}_{k,n}}[s_{i,k}(\theta)]$, for each $k = 1, \ldots, d$. Now, as an alternative objective function, we consider

$$\widetilde{\mathcal{J}}_n(\theta) = \sum_{k=1}^{d-1} \mathcal{I}_n(\widetilde{\boldsymbol{U}}_k(\theta), \widetilde{\boldsymbol{U}}_{k^+}(\theta)). \tag{9}$$

Finally, we define this PIT and distance covariance based ICA estimator (PITdCovICA) as $\widetilde{\theta}_n = \operatorname{argmin}_\theta \widetilde{\mathcal{J}}_n(\theta)$. Similar to the dCovICA estimator, estimation may also be performed conditionally in a sequence because invariance to orthogonal transformations is preserved when applying the PIT. Many alternative smoothing methods are available for estimating F_{s_k}, but computationally fast methods, such as that proposed below, should be strictly preferred since the approximation

needs to be updated continuously within any optimization algorithm applied to Equation (9). The PITdCovICA estimator is computationally more demanding, but it is even more robust to extreme observations and it remains invariant to component-wise monotone transformations of the observations $\boldsymbol{Y}$. Practical implementation of both estimators is discussed in Section 3.

2.5 Asymptotic Properties of the Proposed Estimators

Asymptotic results for the proposed estimators require some basic assumptions about how the observations are transformed and the parameter space. By Assumption 2.1 and Slutsky's Theorem, without loss of generality, assume throughout this section that $\mathrm{E}(Y) = \mathbf{0}$ and $\mathrm{Cov}(Y) = \boldsymbol{I}_d$, such that $Z = Y$ and $\boldsymbol{Z} = \boldsymbol{Y}$. Let $U(\theta)$ and $\boldsymbol{U}(\theta)$ be defined as a function of θ, such that $U_k(\theta) = F_{s_k}[s_k(\theta)]$, and $u_{i,k}(\theta) = F_{s_k}[s_{i,k}(\theta)]$, for each $k = 1, \ldots, d$. Define the population counterpart of Equation (9) as

$$\widetilde{\mathcal{J}}(\theta) = \sum_{k=1}^{d-1} \mathcal{I}(U_k(\theta), U_{k^+}(\theta)), \tag{10}$$

and let $\overline{\Theta}$ denote a sufficiently large compact subset of the space Θ defined by Equation (5). To establish uniform a.s. convergence of $\widetilde{\mathcal{J}}_n(\theta)$ to $\widetilde{\mathcal{J}}(\theta)$ we require

$$\sup_{y \in \mathbb{R}} |\widetilde{F}_{y_k,n,\tilde{h}_{k,n}}(y) - F_{y_k}(y)| \xrightarrow{a.s.} 0, \quad \text{as} \quad n \to \infty, \tag{11}$$

for each component of Y. The Glivenko-Cantelli theorem does not hold for the standard kernel distribution estimators as defined in Equation (8) with a sample size based h_n replacing $\tilde{h}_{k,n}$. That is, convergence cannot be established uniformly over all $F \in \mathcal{F}$, the class of all continuous distribution functions (Zielinski 2007). To establish uniform in bandwidth consistency for all $F \in \mathcal{F}$, a data-driven bandwidth $\tilde{h}_{k,n}$ is required.

Assumption 2.3. *For each k, the bandwidth $\tilde{h}_{k,n}$ is a measurable function of $\{y_{i,k} : i = 1, \ldots, n\}$, such that $\tilde{h}_{k,n} \xrightarrow{a.s.} 0$ as $n \to \infty$, and the kernel function G is Lipschitz continuous.*

Note that the above assumptions are made on the kernel CDF estimators, not on the observations. Equation (11) holds under Assumptions 2.1 and 2.3 (see Chacón and Rodríguez-Casal 2010, Corollary 1).

Theorem 2.4. *If Assumptions 2.1 and 2.3 hold, if there exists a unique minimizer* $\theta_0 \in \overline{\Theta}$ *of Equation (10), and if* $\boldsymbol{W}_{\theta_0}$ *satisfies the conditions for a unique continuous inverse to exist, then* $\widetilde{\theta}_n \xrightarrow{a.s.} \theta_0$, *as* $n \to \infty$.

Convergence of the PITdCovICA estimator is established on equivalence classes; a proof is given in the Appendix. Note that when the ICA model is misspecified convergence to the pseudo-true value θ_0 is obtained. Under the same conditions, proof that the dCovICA estimator, based on Equation (6), converges a.s. follows from similar arguments. To establish root-*n* consistency of the PITdCovICA estimator we introduce an additional Lemma and Theorem below, while similar arguments establish the same for the dCovICA estimator.

Lemma 2.5. *Let* $x_1, \ldots, x_n$ *be a sample from an (unknown) distribution* $F \in \mathcal{F}$ *in which* $\mathcal{F}$ *denotes the class of all continuous distribution functions. Let*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(x_i)$$

denote the empirical cumulative distribution function (ECDF). Let $x_{1:n} \leq \ldots \leq x_{n:n}$ *denote the order statistics from the sample* $x_1, \ldots, x_n$, *and for some* $b < n$ *let*

$$\tilde{h}_{n,b} = \min\{x_{j:n} - x_{(j-b):n} : j = b+1, ..., n\} \tag{12}$$

denote the minimum bth order spacing *among the order statistics. Define the kernel estimator*

$$\widetilde{F}_{n,b}(x) = \frac{1}{n} \sum_{i=1}^{n} G\left(\frac{x - x_i}{\tilde{h}_{n,b}}\right),$$

in which we assume $G(x) = 0$ *for* $x \leq -1/2$, $G(x) = 1$ *for* $x \geq 1/2$, $G(0) = 1/2$, *and* $G(x)$ *is continuous and nondecreasing in* $(-1/2, 1/2)$. *Then, for* $i = 1, \ldots, n$ *we note that* $|\widetilde{F}_n(x_{i:n}) - F_n(x_{i:n})| = \frac{b}{2n}$, *and*

$$\sup_{x \in \mathbb{R}} |\widetilde{F}_{n,b}(x) - F(x)| = o_P(n^{-1}).$$

This Lemma extends the result of Zielinski (2007), and its proof follows from similar arguments. Examples for $G(x)$ include $\Phi(\text{logit}(x + \frac{1}{2}))$ and $\Phi(\tan(x\pi))$, in which $\Phi(x)$ denotes the

standard Gaussian CDF. We use the former in our simulations and applications. We found that $b = \lfloor \sqrt{n} \rfloor$ worked well in our simulations, with little sensitivity to this choice.

Theorem 2.6. *If the assumptions of Theorem 2.4 and Lemma 2.5 hold, if $F_{y_k}(y)$ is twice continuously differentiable $\forall k$, with derivatives $f_{y_k}(y)$ and $\dot{f}_{y_k}(y)$, respectively, if $\mathrm{E}|f_{y_k}(y_k)|^2 < \infty$ and $\mathrm{E}|\dot{f}_{y_k}(y_k)|^2 < \infty$, $\forall k$, and if the ICA model assumptions hold, then $|\widetilde{\theta}_n - \theta_0| = O_P(n^{-1/2})$.*

Again, convergence of the estimator is established on equivalence classes; a proof of the Theorem above and the misspecification corollary below is given in the Appendix.

Corollary 2.7. *If the ICA model is misspecified, but the remaining conditions stated in Theorem 2.6 hold, and if $\mathrm{E}\left[\frac{\partial}{\partial\theta}\widetilde{\mathcal{J}}_n(\theta)\big|_{\theta=\theta_0}\right] = o_P(n^{-1/2})$, in which θ_0 denotes the pseudo-true value, then $|\widetilde{\theta}_n - \theta_0| = O_P(n^{-1/2})$.*

2.6 Misspecification Testing the ICA Model

Although the minimizers $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ of Equations (6) and (9), respectively, always exist, an important question applicable to all ICA methods is whether or not the ICA model is misspecified. To evaluate this issue statistically, we construct a test of the null hypothesis

$$H_0 : \boldsymbol{Y} = \boldsymbol{S}\boldsymbol{M}',$$

in which $\boldsymbol{M}$ is nonsingular and $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_d$ are mutually independent vectors, each of which is a sequence of iid random variables with mean 0 and variance 1. Under the assumption of nonsingular linear mixing, the null hypothesis above is a sufficient but not necessary condition, as singular or nonlinear transformations may also exist. Each sequence is only required to consist of identically distributed random variables, but independent sequences are required to construct an estimate of the null distribution via resampling. Extension to time series data are also of interest.

Since $\boldsymbol{M}$ is unknown in practice, and the limiting distribution of $\mathcal{U}_n(\widehat{\boldsymbol{S}})$ and $\mathcal{U}_n(\boldsymbol{S})$ differ, we define a resampling based procedure below. This allows us to approximate a reject region for H_0 above. If H_0 fails to be rejected, we may also construct confidence sets for the mixing matrix $\boldsymbol{M}$, and even the ICs $\boldsymbol{S}$, based on the same resampling scheme. The performance of the distance covariance test shown in Székely et al. (2007) is promising, but alternative test statistics may also

be used in the procedure detailed below, and rigorous comparisons of the size and power tradeoff of competing approaches are of interest.

Now, define $\widehat{\boldsymbol{M}}_n = \widehat{\boldsymbol{O}}_n^{-1}\boldsymbol{W}_{\theta_n}^{-1}$ as the estimated mixing matrix, in which $\widehat{\boldsymbol{O}}_n$ is the estimated uncorrelating matrix, and θ_n is either the dCovICA estimator $\widehat{\theta}_n$ or the PITdCovICA estimator $\widetilde{\theta}_n$, as defined in Section 2.4. The proposed resampling scheme consists of the following two steps.

(i) For $k = 1, \ldots, d$, jointly sample the entire sequence $\boldsymbol{S}_k^* = (s_{1,k}^*, \ldots, s_{n,k}^*)'$ by randomly permuting the n elements of $\widehat{\boldsymbol{S}}_k$.

(ii) Let $\boldsymbol{Y}^* = \boldsymbol{S}^*\widehat{\boldsymbol{M}}_n'$, and randomly generate a $d \times d$ signed permutation matrix $\boldsymbol{P}_\pm^*$.

Remark First the observed sample $\boldsymbol{Y}$ is replaced by $\boldsymbol{Y}^*$, then a mixing matrix $\boldsymbol{M}^*$ is estimated via the same procedure used to calculate $\widehat{\boldsymbol{M}}_n$, and the resampled estimates of the ICs are $\widehat{\boldsymbol{S}}^* = \boldsymbol{Y}^*\boldsymbol{M}^{*'^{-1}}$. Let $\mathcal{U}_n^*(\widehat{\boldsymbol{S}}) = \mathcal{U}_n(\widehat{\boldsymbol{S}}^*\boldsymbol{P}_\pm^*)$. Under H_0, the limiting distribution of $\mathcal{U}_n(\boldsymbol{S})$ is invariant with respect to the ordering of the components of $\boldsymbol{S}$, but for small samples, multiplication by $\boldsymbol{P}_\pm^*$ is recommended to eliminate possible order dependence from the statistic's distribution. The resampled observations $\boldsymbol{Y}^*$ are generated under H_0, and conditional on the original observations $\boldsymbol{Y}$, the empirical distribution of $\mathcal{U}_n^*(\widehat{\boldsymbol{S}})$ approximates the distribution of $\mathcal{U}_n(\widehat{\boldsymbol{S}})$. We repeat the above resampling N, a large integer, times, and we reject H_0 if $\mathcal{U}_n(\widehat{\boldsymbol{S}})$ is greater than the $(N\alpha)$th largest value of the $\{\mathcal{U}_n^*(\widehat{\boldsymbol{S}})\}$, in which $\alpha \in (0, 1)$ is the size of the test.

This misspecification test accounts for the uncertainty in estimating ICs given approximately uncorrelated observations $\widehat{\boldsymbol{Z}}_n$, as well as the uncertainty in estimating $\boldsymbol{Z}$. This procedure is independent of the estimation method, hence it may be used with any ICA estimation technique. Although the proposed test statistics are consistent against all types of dependent alternatives, we reiterate that they are not uniformly most powerful for all types of dependent alternatives. If more powerful tests are known to exist for certain applications they may be substituted within the proposed framework, and in finite samples, we may only fail to reject the stated null hypothesis for a stated confidence level.

Finally, let $D(\cdot, \cdot)$ be a suitable metric for comparing two mixing matrices; a specific metric with pertinent invariance properties is described in Section 3 below. Then, a resampling-based

approximation for a $1-\alpha$ confidence set for the mixing matrix $\boldsymbol{M}$ may be constructed,

$$\{\boldsymbol{M} : D(\boldsymbol{M}, \widehat{\boldsymbol{M}}_n) \leq c_\alpha, \boldsymbol{M} \text{ non}-\text{singular}\},$$

in which c_α is the $(N\alpha)$th largest value of $D(\boldsymbol{M}^*, \widehat{\boldsymbol{M}}_n)$ obtained in N replications of the resampling scheme. A confidence set for the separating matrix $\boldsymbol{W}$ may similarly be defined.

3. SIMULATION PERFORMANCE AND PRACTICAL IMPLEMENTATION

In this section we compare the proposed estimation methods with several alternatives in simulation studies. We also detail practical implementation and discuss empirical performance measures for ICA. The proposed methods are available in the `R` (R Development Core Team 2010) package `steadyICA` (Risk et al. 2015).

We evaluate the performance of the proposed dCovICA and PITdCovICA estimators by performing simulations similar to Bach and Jordan (2003) and Hastie and Tibshirani (2003). The left panel in Figure 1 shows the 18 distributions used. These include the Student-t (with 3 and 5 degrees of freedom), uniform, exponential, mixtures of exponentials, as well as symmetric and asymmetric Gaussian mixtures. For each of these distributions, we simulate ICs $\boldsymbol{S}_0$ with length $n = 1{,}000$ and a random mixing matrix $\boldsymbol{M}_0 \in \mathbb{R}^{2\times 2}$ with condition number between 1 and 2 using the `R` package `ProDenICA` (Hastie and Tibshirani 2010). Observations are then defined as $\boldsymbol{Y}_0 = \boldsymbol{S}_0\boldsymbol{M}_0'$. In this experiment, the pairs of simulated independent components have identical distributions. This feature was not incorporated into any of the estimation methods discussed below, but doing so would be expected to improve performance.

We compare empirical performance of the proposed estimators with the FastICA estimator using the negentropy criterion (Hyvärinen and Oja 1997), and the KDICA fast kernel density ICA estimator (Chen 2006), which is initialized from the FastICA estimate as the author describes, although the implementation is our own. KDICA is conjectured (Chen 2006) to have the same semi-parametric efficiency as the EFFICA estimator (Chen and Bickel 2006) while remaining computationally tractable. We also compare with the symmetric (Sym R-est) and asymmetric (Asy R-est) R-estimation ICA methods of Ilmonen and Paindaveine (2011) and Hallin and Mehta

(2015), respectively, each initialized from the consistent dCovICA solution, with R code for both provided by the authors of the latter.

The simulated observations are centered by their sample mean, then pre-whitened using the standardized scores from PCA. In practice, ICA typically requires minimization of a non-linear, locally convex objective function. This is performed using iterative algorithms, any of which requires initialization. To find a suitable initialization for the proposed methods, we perform Latin hypercube sampling uniformly over the space Θ defined in Equation (5) to obtain 1,000 parameter values. We then evaluate the objective function at each value and record which minimizes the objective function for initialization. We recommend that the number of parameter values considered should grow with the dimension.

Each method returns an estimate for the mixing matrix. To jointly measure the uncertainty associated with pre-whitening and estimating ICs, we use the metric proposed by Ilmonen et al. (2010) to measure the error between an estimate $\widehat{\boldsymbol{M}}$ and the known parameter $\boldsymbol{M}_0$. It is defined as

$$D(\boldsymbol{M}_0, \widehat{\boldsymbol{M}}) = \frac{1}{\sqrt{d-1}} \inf_{C \in \mathcal{C}} \|\boldsymbol{C}\widehat{\boldsymbol{M}}^{-1}\boldsymbol{M}_0 - \boldsymbol{I}_d\|_F, \tag{13}$$

in which $\|\cdot\|_F$ denotes the Frobenius norm. Let $\mathcal{M}$ be the set of $d \times d$ nonsingular matrices. Let $\boldsymbol{P}_\pm$ be a signed permutation and let $\boldsymbol{B}$ be a diagonal matrix with positive diagonal elements, both $d \times d$. The infimum above is taken such that the metric D is invariant with respect to the three non-identified quantities associated with ICA by defining

$$\mathcal{C} = \{\boldsymbol{C} \in \mathcal{M} : \boldsymbol{C} = \boldsymbol{P}_\pm \boldsymbol{B} \text{ for some } \boldsymbol{P}_\pm \text{ and } \boldsymbol{B}\}.$$

A function for computing D is available in the R package `JADE` (Nordhausen et al. 2011).

The right panel of Figure 1 shows the mean error for each method and each distribution, based on $N = 1{,}000$ simulations for each distribution, with vertical bars for standard errors. The dCovICA and PITdCovICA results are competitive with the others in all situations. FastICA is dominated for most of the mixture distributions. As expected, Sym R-est performs poorly when the source distributions are asymmetric, and is comparable to Asy R-est for the remaining cases. The Asy R-est could be further improved in the multimodal cases by incorporating our kernel CDF

methods into their estimation procedure. Finally, for $n = 1{,}000$, we see that dCovICA slightly outperforms PITdCovICA in some cases as well.

In our simulations, and applications below, we define the PITdCovICA estimator using the kernel function $G(x) = \Phi(\text{logit}(x + \frac{1}{2}))$, in which $\Phi(x)$ denotes the standard Gaussian CDF, and apply the *spacing* bandwidth from Equation (12) with $b = \lfloor \sqrt{n} \rfloor$. To investigate the finite sample effect the bandwidth choice has on the PITdCovICA estimator, we repeated the previous simulation adjusting the bandwidth by a scale factor of 0.25, 0.5, 1, 1.5, and 2. The difference in mean error for the PITdCovICA method with these bandwidth adjustments was much smaller than the size of the standard errors, so we conclude there is no significant difference between these bandwidths in this simulation.

Finally, with $n = 1{,}000$, we also ran $N = 1{,}000$ simulations in $\mathbb{R}^4$, $\mathbb{R}^8$, and $\mathbb{R}^{16}$ by randomly selecting with replacement 4, 8, or 16 of the 18 distributions, respectively, for each iteration and generating $\boldsymbol{M}_0$ analogously to the previous experiment. The results are shown in Table 1, including mean computation times. FastICA was much faster on average, but its mean error was about twice as large as the others. KDICA was fast for the 4- and 8-dimensional cases, but lost its relative speed advantage in the 16-dimensional case. It performed much better on average than the FastICA estimate it was initialized from. In limited experimentation we found that KDICA was more competitive when it was instead initialized from a dCovICA estimate. The Sym R-est method was not included since asymmetric distributions were frequently included in these simulations. The Asy R-est method was fairly fast and performed reasonably in the 4-dimensional observation vector case, but for this particular simulation study we encountered convergence issues in approximating the suggested reference skewed t-density and were unable to report further results. Alternative parametric families or incorporating our kernel CDF methods might be considered instead.

We included both joint and sequential estimation of the dCovICA and PITdCovICA estimators for further comparison. Joint estimation of the dCovICA and PITdCovICA estimators had the smallest mean error with the smallest increases as the dimension increased, although PITdCovICA was among the slowest overall. The mean error for sequential estimation increased more quickly with the dimension relative to the corresponding joint estimators, and the relative speed improve-

ment diminished as the number of optimizations to be solved also increases with the dimension even though each individually involves many fewer parameters.

4. APPLICATION

In this section we illustrate and discuss application of our methodology to two real examples. Throughout this section the PITdCovICA estimator is calculated using joint estimation, with the kernel function $G(x) = \Phi(\text{logit}(x + \frac{1}{2}))$, in which $\Phi(x)$ denotes the standard Gaussian CDF, and the *spacing* bandwidth from Equation (12) with $b = \lfloor \sqrt{n} \rfloor$, and 1,000 starting values, as outlined in Section 3.

4.1 U.S. Crime Rate

The Freedman data (Freedman 1975), from the U.S. Census Bureau, reports crime rates in U.S. metropolitan areas with 1968 populations of 250,000 or more. The data are available in the `R` package `car` (Fox 2009). We consider four variables: the logarithm of population (1968 total, in thousands); nonwhite (percent nonwhite population, 1960); density (population per square mile, 1968); and crime (crime rate per 100,000, 1969). Below, we consider an exploratory analysis for the joint distribution of these four variables and discuss whether they can be factored into four approximately mutually independent components.

We first remove the 10 observations with missing values and analyze $n = 100$ cities with complete data. Using only the complete data was done for simplicity and our analysis did not differ substantially when imputed values were used. Next, the sample mean was subtracted from each observation. Finally, each of the four marginal variables is divided by its sample standard deviation $(0.79, 10.08, 1441.95, 983.58)'$ to simplify parameter interpretation. Now, we test whether these standardized observations $\widehat{\boldsymbol{Y}}$ are ICs using the statistic from Equation (3). The test statistic is $\mathcal{U}_n(\widehat{\boldsymbol{Y}}) = 2.52$, with p-value ≈ 0, indicating significant dependence.

Next, PCA was applied to obtain approximately uncorrelated components $\widehat{\boldsymbol{Z}}$. The ICs test statistic for these standardized PC scores is $\mathcal{U}_n(\widehat{\boldsymbol{Z}}) = 1.59$, with p-value ≈ 0, hence the PCs are not ICs. Finally, ICs $\widehat{\boldsymbol{S}}$ are estimated using the PITdCovICA method. The ICs test statistic is $\mathcal{U}_n(\widehat{\boldsymbol{S}}) = 0.016$, with p-value ≈ 0.24, hence, we fail to reject the ICA model for this dataset at conventional significance levels. The estimated mixing matrix and its inverse are shown Table 2.

We see that crime, for example, is a weighted average of $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$, with loadings 0.75, −0.43, and −0.50, respectively.

Székely and Rizzo (2009) use this dataset to illustrate a jackknife procedure, based on distance covariance, to identify possible influential observations. Their analysis suggests that Philadelphia is an unusual observation. The PCs are ordered by the proportion of variability they explain in the observations, the lowest of which was above 10%. The first two PCs are shown in Figure 2(a). Estimated contour lines have been drawn for each decile, and Philadelphia is indicated on the plot as a larger solid point. PCA does not identify Philadelphia as an unusual observation, by this plot, or in plots of other pairs of PCs.

The estimated ICs do not have a natural ordering, but $\hat{s}_1$ and $\hat{s}_4$ explain the largest proportion of variability in the observations. They are shown in Figure 2(b), with details similar to 2(a) included. The point corresponding to Philadelphia simultaneously takes large values in magnitude on both $\hat{s}_1$ and $\hat{s}_4$. From Table 2 we see that $\hat{s}_1$ has a negative coefficient for population, but positive for the others, while $\hat{s}_4$ has a negative coefficient for crime, but positive for the others. This corresponds directly with Philadelphia's relatively low crime rate and its relatively high population level, during this time period. Figure 2(c) and Figure 2(d) show the same observations after taking the empirical PIT component-wise. A clear pattern is visible in Figure 2(c) confirming rejection of the ICs test for the PCs, whereas points in 2(d) appear uniformly distributed within the unit square.

4.2 U.S. Unemployment Rate

To further illustrate the proposed approach we consider analysis of statewide, seasonally adjusted monthly unemployment rates from January 1976 through August 2010. We will focus on six states: CA, FL, IL, MI, OH, and WI. The data is available from the U.S. Department of Labor at `http://Data.bls.gov/cgi-bin/surveymost?la`, and also from *FRED* of the Federal Reserve Bank of St. Louis `http://research.stlouisfed.org/fred`.

To begin the analysis we difference each series to remove the observed nonstationarity in mean; no trend is present after differencing. Next, we scale the observations by the reciprocal of their sample monthly standard deviations to remove the observed heteroskedasticity in each series. Let $\widehat{\boldsymbol{Y}}$

denote these standardized observations; they are shown in Figure 3. Assumption 2.1 also requires the observations to be independent, in this case, over time. Let Y_i denote a length d random vector observation occurring at time i and let $Y'_{(i-1):(i-m)} = (Y'_{i-1}, \ldots, Y'_{i-m})$ denote a length dm vector containing m observations occurring at times $i-m, \ldots, i-1$, respectively. We can use distance covariance to simultaneously measure serial dependence by testing whether or not $\mathcal{I}(Y_i, Y_{(i-1):(i-m)}) = 0$. Equivalently, we may preform a PIT and base the test on transformed variables $\boldsymbol{U}$, as in Section 2.4.

For a d dimensional process $\boldsymbol{Y}$, with length n, we define a joint m-lag test statistic as

$$Q_d(\boldsymbol{Y}, m) = (n-m)\mathcal{I}_n\left[\widehat{\boldsymbol{U}}^{(1+m):n}, \left(\widehat{\boldsymbol{U}}^{(m):(n-1)}, \ldots, \widehat{\boldsymbol{U}}^{1:(n-m)}\right)\right], \tag{14}$$

in which $\widehat{\boldsymbol{U}}$ is the component-wise marginal ranks of $\boldsymbol{Y}$ and the superscripts denote the observation indices included in each term. Let ϕ_{u_i} denote the joint characteristic function for the transformed variable U at time i. The hypothesis we are testing is $H_0 : \phi_{u_i,u_{i-1},\ldots,u_{i-m}}(t) = \phi_{u_i}(t_1)\phi_{u_{i-1},\ldots,u_{i-m}}(t_2, \ldots, t_{d+1})$, $\forall t \in \mathbb{R}^{d(m+1)}$ and $\forall i \in \mathbb{N}$. Under the assumption of stationarity, this is equivalent to $H_0 : \phi_{u_i,u_{i-1},\ldots,u_{i-m}}(t) = \phi_{u_i}(t_1)\phi_{u_{i-1}}(t_2)\cdots\phi_{u_{i-m}}(t_{d+1})$, $\forall t \in \mathbb{R}^{d(m+1)}$ and $\forall i \in \mathbb{N}$, that is, mutual independence between any $m+1$ neighboring observations.

Lemma 4.1. *Suppose $\boldsymbol{Y} = \{Y_i : i = 1, \ldots, n\}$ are identically distributed and have a continuous distribution. If they are mutually independent, then for any m,*

$$Q_d(\boldsymbol{Y}, m) \xrightarrow{D} Q, \quad as \quad n \to \infty,$$

in which Q is a non-degenerate random variable.

The definition of Q and its distribution can be derived from Theorem 5 of Székely and Rizzo (2009) and the Glivenko-Cantelli theorem. If $\boldsymbol{Y}$ is a univariate series, then the test statistic will also be asymptotically distribution-free.

Applying this test, we find $Q_6(\widehat{\boldsymbol{Y}}, 12) = 30.92$. By applying a resampling scheme similar to that in Section 2.6, we find this has a p-value ≈ 0. This indicates significant serial dependence in the series. To remove this dependence we fit a vector autoregression (VAR) of order three using ordinary least squares. Let $\widehat{\boldsymbol{E}}$ denote the estimated residuals. We find $Q_6(\widehat{\boldsymbol{E}}, 12) = 0.10$, with p-

value ≈ 0.08. Hence, this simple VAR model is sufficient at the 5% significance level for removing all serial dependence in the series $\widehat{\boldsymbol{Y}}$, and no nonlinear modeling is necessary.

Given the test results above, we proceed under the assumption that the $\widehat{\boldsymbol{E}}$ are iid, and now apply our ICA methodology. First we test whether the components of $\widehat{\boldsymbol{E}}$ are ICs. The ICs test statistic is $\mathcal{U}_n(\widehat{\boldsymbol{E}}) = 5.27$, with p-value ≈ 0, indicating significant dependence. To simplify parameter interpretation, the elements of $\widehat{\boldsymbol{E}}$ are scaled by their standard deviations $(0.45, 0.59, 0.48, 0.41, 0.52, 0.65)'$. Next, PCA was applied to obtain approximately uncorrelated components $\widehat{\boldsymbol{Z}}$. The ICs test statistic for these standardized PC scores is $\mathcal{U}_n(\widehat{\boldsymbol{Z}}) = 0.41$, with p-value ≈ 0, hence the PCs are not ICs. Finally, ICs $\widehat{\boldsymbol{S}}$ are estimated using the PITdCovICA method. The ICs test statistic is $\mathcal{U}_n(\widehat{\boldsymbol{S}}) = -0.42$, with p-value ≈ 0.91, hence we fail to reject the ICA model for the residuals $\widehat{\boldsymbol{E}}$. These results are summarized in Table 3. Note that linear transformation from $\widehat{\boldsymbol{E}}$ to $\widehat{\boldsymbol{Z}}$ did not induce any serial dependence and the transformation from $\widehat{\boldsymbol{Z}}$ to $\widehat{\boldsymbol{S}}$ was an orthogonal rotation, which distance covariance is invariant to, see Table 4.

The estimated mixing matrix is shown in Table 5(a). Since the components of $\widehat{\boldsymbol{E}}$ have roughly the same variance, and since $\widehat{Cov}_n(\widehat{\boldsymbol{S}}) = \boldsymbol{I}$, we have $\widehat{Cov}_n(\widehat{\boldsymbol{E}}) \approx \widehat{\boldsymbol{M}}\widehat{\boldsymbol{M}}'$. From this, we see that the sum of squares of the kth row of $\widehat{\boldsymbol{M}}$ gives the variance of $\widehat{\boldsymbol{E}}_k$. Thus, the square of each element gives the *proportion* of the variance of $\widehat{\boldsymbol{E}}_k$ explained by the ICs. In this view, we can remove the smaller coefficients to simplify the interpretation and find

CA: $\hat{e}_1 = -0.89\hat{s}_1 - 0.11\hat{s}_2 + 0.36\hat{s}_4 + 0.23\hat{s}_6$, FL: $\hat{e}_2 = -0.24\hat{s}_1 - 0.10\hat{s}_2 - 0.83\hat{s}_5 + 0.48\hat{s}_6$,
IL: $\hat{e}_3 = -0.32\hat{s}_2 - 0.87\hat{s}_3 + 0.27\hat{s}_4 + 0.24\hat{s}_6$, MI: $\hat{e}_4 = -0.33\hat{s}_1 - 0.85\hat{s}_2 - 0.12\hat{s}_3 - 0.16\hat{s}_5 - 0.34\hat{s}_6$,
OH: $\hat{e}_5 = -0.11\hat{s}_1 - 0.65\hat{s}_2 - 0.19\hat{s}_4 + 0.45\hat{s}_5 + 0.56\hat{s}_6$, WI: $\hat{e}_6 = -0.48\hat{s}_2 + 0.32\hat{s}_3 + 0.81\hat{s}_4$.

From these results, we see that $\hat{s}_2$ is related to each state. Time plots of $\hat{s}_2$ and the change series of seasonally adjusted GDP shows a positive association. This supports the hypothesis that $-\hat{s}_2$ is a national component of the unemployment rate. The component $\hat{s}_4$ is largely specific to WI, while MI and OH have the most complicated structure.

The estimated uncorrelating matrix used to estimate $\widehat{\boldsymbol{Z}}$ is shown in Table 5(b). The first component is roughly an equally weighted average of all six series. The second component gives positive loadings to CA and FL, and negative loadings to the midwestern states. The inverse of the

estimated mixing matrix is shown in Table 5(c). Besides the sixth column, the remaining components give much more relative weight to individual states than the PC scores do. We conclude that these six series can adequately be modeled by a vector autoregression, of which the errors can be decomposed into mutually independent components.

5. CONCLUDING REMARKS

In this paper, we extended the pairwise distance covariance dependence measure to a multivariate measure of mutual independence. We developed novel approaches for ICA, dCovICA and PITdCovICA, using a nonparametric probability integral transformation with a generalized nonparametric whitening method that simultaneously minimizes all forms of dependence among the components. We established the limiting properties of the proposed estimators under weak regularity conditions, and under misspecification of the ICA model. We proposed a test statistic and procedure for misspecification testing of the ICA model, and a flexible resampling-based framework for inference. The test procedure is consistent and is found to work well in simulation and real examples. Simulation results showed that the proposed approach to ICA outperforms the competing methods we considered. We then applied the proposed method to two real examples and obtained sensible interpretations for the data. These examples also highlighted the difference between ICA and PCA.

There are several ways to extend the proposed ICA methods. We primarily considered the case of iid observations. However, many ICA applications, especially in economics and neuroscience, have dependent data. Extension of the proposed approach to handle such data can substantially increase its applicability. Second, we only considered lower dimensional applications in this paper. Many applications encounter high dimensional data. Developing an efficient estimation procedure for the proposed ICA methods to handle high dimensional data is challenging, but important. Third, careful consideration of the size and power tradeoff of the proposed tests is needed. Finally, adaptive methods for the proposed estimators may be considered for application of ICA to data which are only locally stationary.

APPENDIX

Proof of Theorem 2.4

Lemma A.1 *Under Assumptions 2.1 and 2.3,* $\widetilde{\mathcal{J}}_n(\theta) \xrightarrow{a.s.} \widetilde{\mathcal{J}}(\theta)$ *as* $n\to\infty$*, for any* $\theta \in \Theta$.

Proof.

$$\begin{aligned}
|\widetilde{\mathcal{J}}_n(\theta) - \widetilde{\mathcal{J}}(\theta)| &= \Big|\sum_{i=1}^{d-1} \mathcal{I}_n(\widetilde{\mathbf{U}}_k(\theta), \widetilde{\mathbf{U}}_{k^+}(\theta)) - \mathcal{I}(U_k(\theta), U_{k^+}(\theta))\Big| \\
&\le \sum_{i=1}^{d-1} \Big|\mathcal{I}_n(\widetilde{\mathbf{U}}_k(\theta), \widetilde{\mathbf{U}}_{k^+}(\theta)) - \mathcal{I}(U_k(\theta), U_{k^+}(\theta))\Big| \\
&\le \sum_{i=1}^{d-1} \Big(\Big|\mathcal{I}_n(\widetilde{\mathbf{U}}_k(\theta), \widetilde{\mathbf{U}}_{k^+}(\theta)) - \mathcal{I}_n(\mathbf{U}_k(\theta), \mathbf{U}_{k^+}(\theta))\Big| \\
&\qquad + \Big|\mathcal{I}_n(\mathbf{U}_k(\theta), \mathbf{U}_{k^+}(\theta)) - \mathcal{I}(U_k(\theta), U_{k^+}(\theta))\Big|\Big)
\end{aligned}$$

for any $\theta \in \Theta$. For each k, $\left|\mathcal{I}_n(\widetilde{\mathbf{U}}_k(\theta), \widetilde{\mathbf{U}}_{k^+}(\theta)) - \mathcal{I}_n(\mathbf{U}_k(\theta), \mathbf{U}_{k^+}(\theta))\right| \xrightarrow{\text{a.s.}} 0$ by Assumption 2.3 and the continuous mapping theorem, and $|\mathcal{I}_n(\mathbf{U}_k(\theta), \mathbf{U}_{k^+}(\theta)) - \mathcal{I}(U_k(\theta), \mathbf{U}_{k^+}(\theta))| \xrightarrow{\text{a.s.}} 0$ by Assumption 2.1, the triangle inequality, Hölder's inequality, the strong law of large numbers for U-statistics (see Hoeffding 1961), and Slutsky's theorem, as $n\to\infty$, thus establishing the assertion. □

Let $\mathcal{D}$ denote any metric on $\mathcal{SO}(d)$, continuous in its first argument, such that for all $\boldsymbol{W}, \boldsymbol{A} \in \mathcal{SO}(d)$, $\mathcal{D}(\boldsymbol{W}, \boldsymbol{A}) = 0$ if and only if there exists a $\boldsymbol{P}_\pm$ such that $\boldsymbol{W} = \boldsymbol{P}_\pm \boldsymbol{A}$, and $\mathcal{D}(\boldsymbol{W}, \boldsymbol{A}) > 0$ otherwise. Partition $\mathcal{SO}(d)$ into equivalence classes via $\mathcal{D}$: the $\mathcal{D}$-distance between any two elements within an equivalence class is 0, and the $\mathcal{D}$-distance between any two elements from different equivalence classes is greater than 0. Let $\mathcal{SO}(d)_\mathcal{D}$ be the quotient space $\mathcal{SO}(d)/\mathcal{D}$ of these equivalence classes. Then $\boldsymbol{W} = \boldsymbol{A}$ on $\mathcal{SO}(d)_\mathcal{D}$ if and only if $\mathcal{D}(\boldsymbol{W}, \boldsymbol{A}) = 0$.

Lemma A.2 *Under Assumptions 2.1 and 2.3,* $\widetilde{\mathcal{J}}_n(\theta)$ *is Lipschitz continuous for* $\theta : \boldsymbol{W}_\theta \in \mathcal{SO}(d)_\mathcal{D}$.

Proof. First, note that the composition of two Lipschitz continuous functions is also Lipschitz continuous. $\boldsymbol{S}_\theta$ is a trigonometric compositions of Lipschitz functions with respect to θ, hence it is

Lipschitz continuous. Lipschitz continuity of $\widetilde{\boldsymbol{U}}(\theta)$ follows from Assumption 2.3.

To establish the Lipschitz continuity of $\widetilde{\mathcal{J}}_n(\theta)$ it is sufficient to show $\mathcal{I}_n(\widetilde{\boldsymbol{U}}_k(\theta), \widetilde{\boldsymbol{U}}_{k^+}(\theta))$ is Lipschitz continuous for $k = 1, \ldots, d-1$. The Euclidean norm is a Lipschitz function, as is a linear combinations of two Lipschitz functions. The product of two bounded Lipschitz functions is a Lipschitz functions as well. It is clear that $\mathcal{I}_n(\widetilde{\boldsymbol{U}}_k(\theta), \widetilde{\boldsymbol{U}}_{k^+}(\theta))$ is uniformly bounded for a fixed dimension d. This establishes the Lipschitz continuity of $\widetilde{\mathcal{J}}_n(\theta)$. □

Lemma A.3 *Under Assumptions 2.1 and 2.3,*

$$\sup_{\theta: W_\theta \in SO(d)_{\mathcal{D}}} |\widetilde{\mathcal{J}}_n(\theta) - \widetilde{\mathcal{J}}(\theta)| \xrightarrow{a.s.} 0 \quad as \quad n \to \infty.$$

Proof. Applying the Arzelá-Ascoli theorem from complex analysis it is sufficient to show:
(i) $\widetilde{\mathcal{J}}_n(\theta) \xrightarrow{\text{a.s.}} \widetilde{\mathcal{J}}(\theta)$ for each $\theta : \boldsymbol{W}_\theta \in \Xi_0$, some countable dense subset of $SO(d)_{\mathcal{D}}$, and
(ii) $\lim_{c\to\infty} \overline{\lim}_n \, m_{\frac{1}{c}}(\widetilde{\mathcal{J}}_n) \overset{\text{a.s.}}{=} 0$, in which

$$m_{\frac{1}{c}}(\widetilde{\mathcal{J}}_n) = \sup\left\{|\widetilde{\mathcal{J}}_n(\theta) - \widetilde{\mathcal{J}}_n(\psi)| : \boldsymbol{W}_\psi, \boldsymbol{W}_\theta \in SO(d)_{\mathcal{D}}, \left\|\boldsymbol{W}_\psi - \boldsymbol{W}_\theta\right\|_F < 1/c\right\}.$$

$SO(d)_{\mathcal{D}}$ is separable since it is compact. Consequently, there exists a countable dense subset, say Ξ_0. Lemma A.1 implies that $\widetilde{\mathcal{J}}_n(\theta) \xrightarrow{\text{a.s.}} \widetilde{\mathcal{J}}(\theta)$ as $n \to \infty$, for each $\boldsymbol{W}_\theta \in SO(d)_{\mathcal{D}}$, and in particular for each $\boldsymbol{W}_\theta \in \Xi_0$.

Let $\tilde{\boldsymbol{u}} = \widetilde{\boldsymbol{U}}(\theta)$, $\tilde{\boldsymbol{v}} = \widetilde{\boldsymbol{U}}(\psi)$, $\boldsymbol{u} = \boldsymbol{U}(\theta)$ and $\boldsymbol{v} = \boldsymbol{U}(\psi)$. Lemma A.2 implies that there exists a constant $0 < L < \infty$ such that for any $\boldsymbol{W}_\psi, \boldsymbol{W}_\theta \in SO(d)_{\mathcal{D}}$, $\left\|\boldsymbol{W}_\psi - \boldsymbol{W}_\theta\right\|_F \leq \delta_1$ implies $|\tilde{\boldsymbol{u}}_i - \tilde{\boldsymbol{v}}_i| < L\delta_1$ for all $i = 1, \ldots, n$. Note that

$$|\widetilde{\mathcal{J}}_n(\theta) - \widetilde{\mathcal{J}}_n(\psi)| = \left|\sum_{\ell=1}^{d-1} \mathcal{I}_n(\widetilde{\boldsymbol{U}}_\ell(\theta), \widetilde{\boldsymbol{U}}_{\ell^+}(\theta)) - \mathcal{I}_n(\widetilde{\boldsymbol{U}}_\ell(\psi), \widetilde{\boldsymbol{U}}_{\ell^+}(\psi))\right|$$

$$
\begin{aligned}
&\leq \sum_{\ell=1}^{d-1} \left| \mathcal{I}_n(\widetilde{\boldsymbol{U}}_\ell(\theta), \widetilde{\boldsymbol{U}}_{\ell^+}(\theta)) - \mathcal{I}_n(\widetilde{\boldsymbol{U}}_\ell(\psi), \widetilde{\boldsymbol{U}}_{\ell^+}(\psi)) \right| \\
&= \sum_{\ell=1}^{d-1} \left| \left(T_{1,n}^{(\ell)}(\theta) + T_{2,n}^{(\ell)}(\theta) - T_{3,n}^{(\ell)}(\theta) \right) - \left(T_{1,n}^{(\ell)}(\psi) + T_{2,n}^{(\ell)}(\psi) - T_{3,n}^{(\ell)}(\psi) \right) \right| \\
&\leq \sum_{\ell=1}^{d-1} \left| T_{1,n}^{(\ell)}(\theta) - T_{1,n}^{(\ell)}(\psi) \right| + \left| T_{2,n}^{(\ell)}(\theta) - T_{2,n}^{(\ell)}(\psi) \right| + \left| T_{3,n}^{(\ell)}(\theta)) - T_{3,n}^{(\ell)}(\psi)) \right|,
\end{aligned}
$$

in which the $T_{j,n}^{(\ell)}(\theta)$ are defined as $T_{j,n}(\widetilde{\boldsymbol{U}}_\ell(\theta), \widetilde{\boldsymbol{U}}_{\ell^+}(\theta))$, analogous to Equation (2). Applying standard Euclidean norm inequalities we note the following inequalities

$$
\begin{aligned}
\left| T_{1,n}^{(\ell)}(\theta) - T_{1,n}^{(\ell)}(\psi) \right| &= \left| \binom{n}{2}^{-1} \sum_{i<j} |\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}| |\tilde{\boldsymbol{u}}_{i,\ell^+} - \tilde{\boldsymbol{u}}_{j,\ell^+}| - \binom{n}{2}^{-1} \sum_{i<j} |\tilde{v}_{i,\ell} - \tilde{v}_{j,\ell}| |\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+}| \right| \\
&\leq \binom{n}{2}^{-1} \sum_{i<j} \left| |\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}| |\tilde{\boldsymbol{u}}_{i,\ell^+} - \tilde{\boldsymbol{u}}_{j,\ell^+}| - |\tilde{v}_{i,\ell} - \tilde{v}_{j,\ell}| |\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+}| \right| \\
&\leq \binom{n}{2}^{-1} \sum_{i<j} |\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}| |(\tilde{\boldsymbol{u}}_{i,\ell^+} - \tilde{\boldsymbol{u}}_{j,\ell^+}) - (\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+})| + \\
&\quad \binom{n}{2}^{-1} \sum_{i<j} |(\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}) - (\tilde{v}_{i,\ell} - \tilde{v}_{j,\ell})| |\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+}| \\
&\leq \left(\binom{n}{2}^{-1} \sum_{i<j} |\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}| \right) \left(\binom{n}{2}^{-1} \sum_{i<j} |(\tilde{\boldsymbol{u}}_{i,\ell^+} - \tilde{\boldsymbol{u}}_{j,\ell^+}) - (\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+})| \right) + \\
&\quad \left(\binom{n}{2}^{-1} \sum_{i<j} |(\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}) - (\tilde{v}_{i,\ell} - \tilde{v}_{j,\ell})| \right) \left(\binom{n}{2}^{-1} \sum_{i<j} |\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+}| \right) \\
&\leq \left(\binom{n}{2}^{-1} \sum_{i<j} |\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}| \right) \left(\binom{n}{2}^{-1} \sum_{i<j} (|\tilde{\boldsymbol{u}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{i,\ell^+}| + |\tilde{\boldsymbol{u}}_{j,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+}|) \right) + \\
&\quad \left(\binom{n}{2}^{-1} \sum_{i<j} (|\tilde{u}_{i,\ell} - \tilde{v}_{i,\ell}| + |\tilde{u}_{j,\ell} - \tilde{v}_{j,\ell}|) \right) \left(\binom{n}{2}^{-1} \sum_{i<j} |\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+}| \right) \\
&\leq \left(\binom{n}{2}^{-1} \sum_{i<j} |\tilde{\boldsymbol{u}}_i - \tilde{\boldsymbol{u}}_j| \right) \left(\binom{n}{2}^{-1} \sum_{i<j} (|\tilde{\boldsymbol{u}}_i - \tilde{\boldsymbol{v}}_i| + |\tilde{\boldsymbol{u}}_j - \tilde{\boldsymbol{v}}_j|) \right) + \\
&\quad \left(\binom{n}{2}^{-1} \sum_{i<j} (|(\tilde{\boldsymbol{u}}_i - \tilde{\boldsymbol{v}}_i| + |\tilde{\boldsymbol{u}}_j - \tilde{\boldsymbol{v}}_j)|) \right) \left(\binom{n}{2}^{-1} \sum_{i<j} |\tilde{\boldsymbol{v}}_i - \tilde{\boldsymbol{v}}_j| \right),
\end{aligned}
$$

$$
\begin{aligned}
\left| T_{2,n}^{(\ell)}(\theta) - T_{2,n}^{(\ell)}(\psi) \right| &= \left| \binom{n}{2}^{-1} \sum_{i<j} |\tilde{u}_{i,\ell} - \tilde{u}_{j,\ell}| \binom{n}{2}^{-1} \sum_{i<j} |\tilde{\boldsymbol{u}}_{i,\ell^+} - \tilde{\boldsymbol{u}}_{j,\ell^+}| \right. \\
&\quad \left. - \binom{n}{2}^{-1} \sum_{i<j} |\tilde{v}_{i,\ell} - \tilde{v}_{j,\ell}| \binom{n}{2}^{-1} \sum_{i<j} |\tilde{\boldsymbol{v}}_{i,\ell^+} - \tilde{\boldsymbol{v}}_{j,\ell^+}| \right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left|\binom{n}{2}^{-1}\sum_{i<j}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}|\right|\left|\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{u}}_{j,\ell^+}|-\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{j,\ell^+}|\right| \\
&\quad+\left|\binom{n}{2}^{-1}\sum_{i<j}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}|-\binom{n}{2}^{-1}\sum_{i<j}|\tilde{v}_{i,\ell}-\tilde{v}_{j,\ell}|\right|\left|\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{j,\ell^+}|\right| \\
&\leq \left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|(\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{u}}_{j,\ell^+})-(\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{j,\ell^+})|\right) \\
&\quad+\left(\binom{n}{2}^{-1}\sum_{i<j}|(\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell})-(\tilde{v}_{i,\ell}-\tilde{v}_{j,\ell})|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{j,\ell^+}|\right) \\
&\leq \left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{i,\ell^+}|+|\tilde{\boldsymbol{u}}_{j,\ell^+}-\tilde{\boldsymbol{v}}_{j,\ell^+}|)\right) \\
&\quad+\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{u}_{i,\ell}-\tilde{v}_{i,\ell}|+|\tilde{u}_{j,\ell}-\tilde{v}_{j,\ell}|)\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{j,\ell^+}|\right) \\
&\leq \left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{u}}_{i}-\tilde{\boldsymbol{u}}_{j}|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_{i}-\tilde{\boldsymbol{v}}_{i}|+|\tilde{\boldsymbol{u}}_{j}-\tilde{\boldsymbol{v}}_{j}|)\right) \\
&\quad+\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_{i}-\tilde{\boldsymbol{v}}_{i}|+|\tilde{\boldsymbol{u}}_{j}-\tilde{\boldsymbol{v}}_{j}|)\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_{i}-\tilde{\boldsymbol{v}}_{j}|\right),
\end{aligned}
$$

$$
\begin{aligned}
&\left|\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}||\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{u}}_{k,\ell^+}|-\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{v}_{i,\ell}-\tilde{v}_{j,\ell}||\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+}|\right| \\
&\quad\leq \binom{n}{3}^{-1}\sum_{i<j<k}\left||\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}||\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{u}}_{k,\ell^+}|-|\tilde{v}_{i,\ell}-\tilde{v}_{j,\ell}||\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+}|\right| \\
&\quad\leq \binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}||(\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{u}}_{k,\ell^+})-(\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+})|+ \\
&\qquad\binom{n}{3}^{-1}\sum_{i<j<k}|(\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell})-(\tilde{v}_{i,\ell}-\tilde{v}_{j,\ell})||\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+}| \\
&\quad\leq \left(\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}|\right)\left(\binom{n}{3}^{-1}\sum_{i<j<k}|(\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{u}}_{k,\ell^+})-(\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+})|\right)+ \\
&\qquad\left(\binom{n}{3}^{-1}\sum_{i<j<k}|(\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell})-(\tilde{v}_{i,\ell}-\tilde{v}_{j,\ell})|\right)\left(\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+}|\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left(\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{u}_{i,\ell}-\tilde{u}_{j,\ell}|\right)\left(\binom{n}{3}^{-1}\sum_{i<j<k}(|\tilde{\boldsymbol{u}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{i,\ell^+}|+|\tilde{\boldsymbol{u}}_{k,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+}|)\right)+\\
&\quad\left(\binom{n}{3}^{-1}\sum_{i<j<k}(|\tilde{u}_{i,\ell}-\tilde{v}_{i,\ell}|+|\tilde{u}_{j,\ell}-\tilde{v}_{j,\ell}|)\right)\left(\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{\boldsymbol{v}}_{i,\ell^+}-\tilde{\boldsymbol{v}}_{k,\ell^+}|\right)\\
&\leq \left(\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{u}}_j|\right)\left(\binom{n}{3}^{-1}\sum_{i<j<k}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_k-\tilde{\boldsymbol{v}}_k|)\right)+\\
&\quad\left(\binom{n}{3}^{-1}\sum_{i<j<k}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_j-\tilde{\boldsymbol{v}}_j|)\right)\left(\binom{n}{3}^{-1}\sum_{i<j<k}|\tilde{\boldsymbol{v}}_i-\tilde{\boldsymbol{v}}_k|\right),\\
&\leq \left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{u}}_j|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_j-\tilde{\boldsymbol{v}}_j|)\right)\\
&\quad+\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_j-\tilde{\boldsymbol{v}}_j|)\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_i-\tilde{\boldsymbol{v}}_j|\right),
\end{aligned}
$$

and similarly for the remaining terms in $T_{3,n}$. Hence,

$$
\begin{aligned}
\left|T^{(\ell)}_{3,n}(\theta)-T^{(\ell)}_{3,n}(\psi)\right| &\leq 2\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{u}}_j|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_j-\tilde{\boldsymbol{v}}_j|)\right)\\
&\quad+2\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_j-\tilde{\boldsymbol{v}}_j|)\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_i-\tilde{\boldsymbol{v}}_j|\right).
\end{aligned}
$$

Therefore, $\left\|\boldsymbol{W}_\psi-\boldsymbol{W}_\theta\right\|_F\leq\delta_1$ implies

$$
\begin{aligned}
|\widetilde{\mathcal{J}}_n(\theta)-\widetilde{\mathcal{J}}_n(\psi)| &\leq 4\sum_{\ell=1}^{d-1}\left(\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{u}}_j|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_j-\tilde{\boldsymbol{v}}_j|)\right)+\right.\\
&\qquad\left.\left(\binom{n}{2}^{-1}\sum_{i<j}(|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|+|\tilde{\boldsymbol{u}}_j-\tilde{\boldsymbol{v}}_j|)\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_i-\tilde{\boldsymbol{v}}_j|\right)\right)\\
&\leq 4\sum_{\ell=1}^{d-1}\left(\frac{2}{n}\sum_{i=1}^{n}|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{v}}_i|\right)\left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{u}}_j|+\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_i-\tilde{\boldsymbol{v}}_j|\right)\\
&\leq \left(\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{u}}_i-\tilde{\boldsymbol{u}}_j|+\binom{n}{2}^{-1}\sum_{i<j}|\tilde{\boldsymbol{v}}_i-\tilde{\boldsymbol{v}}_j|\right)8(d-1)L\delta_1\\
&= B_n 8(d-1)L\delta_1
\end{aligned}
$$

For each $c\in\mathbb{N}$, let $\delta=\min\{1/c,\delta_1\}$. Now observe that for $n\in\mathbb{N}$

$$
m_\delta(\widetilde{\mathcal{J}}_n)=\sup_{\|W_\psi-W_\theta\|_F<\delta}|\widetilde{\mathcal{J}}_n(\theta)-\widetilde{\mathcal{J}}_n(\psi)|\leq B_n 8(d-1)L\delta.
$$

Let $B=\mathrm{E}|\boldsymbol{u}-\boldsymbol{u}'|+\mathrm{E}|\boldsymbol{v}-\boldsymbol{v}'|$ in which $\boldsymbol{u}'$ and $\boldsymbol{v}'$ are iid copies of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. By Assumption 2.1 we have $B<\infty$, and by the SLLN for U-statistics $B_n\xrightarrow{\text{a.s.}}B$, as $n\to\infty$. Therefore, $\overline{\lim}_n\, m_\delta(\widetilde{\mathcal{J}}_n)\leq\overline{\lim}_n\, B_n 8(d-1)L\delta\overset{\text{a.s.}}{=}B8(d-1)L\delta$. As $c\to\infty$, $\delta=\min\{1/c,\delta_1\}=1/c$. Therefore,

the claim is established by noting

$$\lim_{c\to\infty} \overline{\lim_n}\, m_{\frac{1}{c}}(\widetilde{\mathcal{J}}_n) \leq_{a.s.} \lim_{c\to\infty} (B8(d-1)L)/c = 0. \qquad \square$$

Proof of Theorem 2.4 Under Assumptions 2.1 and 2.3, note that for any $n \in \mathbb{N}$, $\widetilde{\mathcal{J}}_n(\theta_0) \geq \widetilde{\mathcal{J}}_n(\widetilde{\theta}_n)$ and $\widetilde{\mathcal{J}}(\widetilde{\theta}_n) \geq \widetilde{\mathcal{J}}(\theta_0)$. Hence,

$$\widetilde{\mathcal{J}}_n(\theta_0) - \widetilde{\mathcal{J}}(\theta_0) \geq \widetilde{\mathcal{J}}_n(\widetilde{\theta}_n) - \widetilde{\mathcal{J}}(\theta_0) \geq \widetilde{\mathcal{J}}_n(\widetilde{\theta}_n) - \widetilde{\mathcal{J}}(\widetilde{\theta}_n),$$

and

$$\begin{aligned} |\widetilde{\mathcal{J}}_n(\widetilde{\theta}_n) - \widetilde{\mathcal{J}}(\theta_0)| &\leq \max\left(|\widetilde{\mathcal{J}}_n(\theta_0) - \widetilde{\mathcal{J}}(\theta_0)|, |\widetilde{\mathcal{J}}_n(\widetilde{\theta}_n) - \widetilde{\mathcal{J}}(\widetilde{\theta}_n)|\right) \\ &\leq \sup_{\theta: W_\theta \in \mathcal{SO}(d)_{\mathcal{D}}} |\widetilde{\mathcal{J}}_n(\theta) - \widetilde{\mathcal{J}}(\theta)|. \end{aligned}$$

Therefore, Lemma A.3 implies that $\widetilde{\mathcal{J}}_n(\widetilde{\theta}_n) \xrightarrow{\text{a.s.}} \widetilde{\mathcal{J}}(\theta_0)$ as $n\to\infty$ for $\theta : \boldsymbol{W}_{\theta_0} \in \mathcal{SO}(d)_{\mathcal{D}}$. Note that the argmin mapping is continuous on $\mathcal{SO}(d)_{\mathcal{D}}$. Since $\mathcal{SO}(d)_{\mathcal{D}}$ is compact, the argmin of $\widetilde{\mathcal{J}}_n$ and $\widetilde{\mathcal{J}}$ exists in $\mathcal{SO}(d)_{\mathcal{D}}$; therefore, $\boldsymbol{W}_{\theta_n} \xrightarrow{\text{a.s.}} \boldsymbol{W}_{\theta_0}$, as $n\to\infty$, for $\boldsymbol{W}_{\theta_0} \in \mathcal{SO}(d)_{\mathcal{D}}$. If $\theta_0 \in \overline{\Theta}$, in which $\overline{\Theta}$ is a sufficiently large compact subset of the space Θ, then Lemma A.3 and the continuous mapping theorem imply $\widetilde{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$ as $n\to\infty$. $\square$

Proof of Theorem 2.6 and Corollary 2.7

We first note some properties specified by the manifold structure of Ω (cf. Chen and Bickel, 2005). Consider a unit ball in $\mathbb{R}^d$ centered at $\boldsymbol{A}$, which contains the points $\boldsymbol{B}$ and $\boldsymbol{C}$; let the angle $\xi > 0$ denote the smallest value such that $\cos(\xi) = \langle \boldsymbol{AB}, \boldsymbol{AC} \rangle$. For $\tau \in \mathbb{R}$, let $\gamma(\tau) = \cos(\tau)\boldsymbol{AB} + \sin(\tau)\boldsymbol{AD}$ denote a path from $\boldsymbol{AB}$ to $\boldsymbol{AC}$, in which $\boldsymbol{AD}$ is a unit tangent vector at $\boldsymbol{B}$ such that $\boldsymbol{AB}$, $\boldsymbol{AC}$, and $\boldsymbol{AD}$ are on the same hyperplane. Then, $\gamma(0) = \boldsymbol{B}$, $\gamma(\xi) = \boldsymbol{C}$, and $\|\frac{\partial}{\partial\tau}\gamma(\tau)\| = 1$; further, $\|\gamma(\tau_2) - \gamma(\tau_1)\| \leq |\tau_2 - \tau_1|$.

By definition, each row of $\boldsymbol{W} \in \Omega$ is on the unit ball in $\mathbb{R}^d$. Let $\xi_1, \ldots, \xi_d$ denote the angles between the corresponding rows of $\boldsymbol{W}_0 = \boldsymbol{W}_{\theta_0}$ and $\widehat{\boldsymbol{W}} = \boldsymbol{W}_{\tilde{\theta}_n}$. Let $\hat{\eta} = \sqrt{\sum_{k=1}^d \xi_k^2}$, and note that $\|\widehat{\boldsymbol{W}} - \boldsymbol{P}_\pm \boldsymbol{W}_0\| = o_P(1)$ implies $\hat{\eta} = o_P(1)$. Assume, w.o.l.o.g., that $\|\widehat{\boldsymbol{W}} - \boldsymbol{P}_\pm \boldsymbol{W}_0\| > 0$, then $\hat{\eta} > 0$.

Now let $\gamma : \mathbb{R} \to \Omega$, such that $\gamma(0) = \boldsymbol{W}_0$ and $\gamma(\hat{\eta}) = \widehat{\boldsymbol{W}}$, by considering $\gamma_k(\cdot)$ for the kth rows, as described above, but rescaled by $\xi_k/\hat{\eta}$. Then, we similarly note that $\|\frac{\partial}{\partial\tau}\gamma(\tau)\| = \sqrt{\sum_{k=1}^{d}(\xi_k/\hat{\eta})^2} = 1$, and $\|\gamma(\tau_2) - \gamma(\tau_1)\| \leq |\tau_2 - \tau_1|$. As such, for $\|\widehat{\boldsymbol{W}} - \boldsymbol{P}_{\pm}\boldsymbol{W}_0\| \leq \hat{\eta}$ and sufficiently small $\tau \geq 0$, we note $\gamma(\tau) \in \Omega$.

Now let $\widetilde{\mathcal{J}}_n(\gamma(\hat{\eta})) = \widetilde{\mathcal{J}}_n(\gamma(\tau))|_{\tau=\hat{\eta}} = \widetilde{\mathcal{J}}_n(\widehat{\boldsymbol{W}})$, and similarly let $\mathcal{J}(\gamma(\tau)) = \mathcal{J}(\boldsymbol{W})$, for $\boldsymbol{W} = \gamma(\tau)$. We apply a first-order Taylor expansion, and by the mean value theorem there exists a $\overline{\tau} \in [0, \hat{\eta}]$ such that

$$\widetilde{\mathcal{J}}_n(\gamma(\hat{\eta})) = \widetilde{\mathcal{J}}_n(\gamma(0)) + \hat{\eta}\frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau})).$$

Next, we note that

$$0 = \frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\gamma(\hat{\eta})) = \frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\gamma(0)) + \hat{\eta}\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau})),$$

which implies

$$\hat{\eta} = -\frac{\frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\gamma(0))}{\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau}))}. \tag{A.1}$$

First we consider the numerator in Equation (A.1); by definition, $\widetilde{\mathcal{J}}_n(\boldsymbol{W})$ is a simple function of U-statistics, and we note that $\frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\boldsymbol{W}_0) = \frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\gamma(\tau))|_{\tau=0}$ is as well. In general, the mean of $\frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\boldsymbol{W}_0)$ depends on the distribution of the observations; however, its mean is zero under the ICA model and in the misspecification corollary it is assumed to be $o_P(n^{-1/2})$. Under the assumptions, we may apply Lemma 12, the Central Limit Theorem, and Cramér's Theorem to note that $\frac{\partial}{\partial\tau}\widetilde{\mathcal{J}}_n(\gamma(0)) = O_P(n^{-1/2})$.

Next we consider the denominator in Equation (A.1). Since $\frac{\partial^2}{\partial\tau^2}\gamma(\tau) = -\gamma(\tau)$, we note that the ranges of $\gamma(\tau)$, $\frac{\partial}{\partial\tau}\gamma(\tau)$, and $\frac{\partial^2}{\partial\tau^2}\gamma(\tau)$ are compact. We may express $\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau}))$ as a simple linear expression of U-statistics. The kernel function for each term is indexed by a compact set given the tuple $\left(\gamma(\tau), \frac{\partial}{\partial\tau}\gamma(\tau), \frac{\partial^2}{\partial\tau^2}\gamma(\tau)\right)$. For each term in the sum, there exists a dominating function with finite expectation, for all $\overline{\tau} \in [0, \hat{\eta}]$, since the assumptions and the CauchySchwarz inequality imply $\mathrm{E}|s_k f_{s_k}(s_k)\dot{f}_{s_k}(s_k)|^2 < \infty,\ \forall k$. We may then apply the triangle inequality and the ULLN for

U-statistics to note that

$$\sup_{\overline{\tau}\in[0,\hat{\eta}]}\left|\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau}))-\mathrm{E}\left(\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau}))\right)\right|=o_P(1).$$

Since $\overline{\tau}\xrightarrow{P}0$, the limit of $\lim_{n\to\infty}\mathrm{E}\left(\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau}))\right)=\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}(\gamma(0))$, by continuity. Hence,

$$\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}_n(\gamma(\overline{\tau}))\geq\min_{\frac{\partial}{\partial\tau}\gamma(0),\frac{\partial^2}{\partial\tau^2}\gamma(0)}\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}(\gamma(0))+o_P(1).$$

Given the identifiability assumptions, $\gamma(0)=\boldsymbol{P}_{+}\boldsymbol{W}_0$ is the unique local minimizer of $\widetilde{\mathcal{J}}(\gamma(\tau))$, and by differentiability we have $\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}(\gamma(0))>0$, for any $\frac{\partial}{\partial\tau}\gamma(0),\frac{\partial^2}{\partial\tau^2}\gamma(0)$. By compactness, $\min\frac{\partial^2}{\partial\tau^2}\widetilde{\mathcal{J}}(\gamma(0))>0$. Finally, combining the results for the numerator and denominator we have $\hat{\eta}=O_P(n^{-1/2})$, which establishes the claim. □

REFERENCES

Bach, F., and Jordan, M. (2003), "Kernel Independent Component Analysis," *The Journal of Machine Learning Research*, 3, 1–48.

Borg, I., and Groenen, P. (2005), *Modern Multidimensional Scaling: Theory and Applications*, New York: Springer Verlag.

Cardoso, J. F. (1989), Source Separation Using Higher Order Moments,, in *ICASSP Proceedings*, IEEE, pp. 2109–2112.

Chacón, J., and Rodríguez-Casal, A. (2010), "A Note on the Universal Consistency of the Kernel Distribution Function Estimator," *Statistics & Probability Letters*, In Press.

Chen, A. (2006), Fast kernel density independent component analysis,, in *Proceedings of the 6th international conference on Independent Component Analysis and Blind Signal Separation*, Springer-Verlag, pp. 24–31.

Chen, A., and Bickel, P. (2005), "Consistent Independent Component Analysis and Prewhitening," *IEEE Trans. Signal Processing*, 53(10), 3625–3632.

Chen, A., and Bickel, P. J. (2006), "Efficient independent component analysis," *The Annals of Statistics*, 34(6), 2825–2855.

Demartines, P., and Herault, J. (1997), "Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets," *IEEE Transactions on Neural Networks*, 8(1), 148.

Eriksson, J., and Koivunen, V. (2003), "Characteristic-Function-Based Independent Component Analysis," *Signal Process*, 83, 2195–2208.

Fox, J. (2009), *car: Companion to Applied Regression*. R Package Version 1.2-16.

Freedman, J. (1975), *Crowding and Behavior*, New York: Viking Press.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007), A Kernel Statistical Test of Independence.,, in *NIPS*, Vol. 20, pp. 585–592.

Hallin, M., and Mehta, C. (2015), "R-estimation for asymmetric independent component analysis," *Journal of the American Statistical Association*, 110(509), 218–232.

Hastie, T., and Tibshirani, R. (2003), "Independent Components Analysis Through Product Density Estimation," *Advances in Neural Information Processing Systems*, 15, 665–672.

Hastie, T., and Tibshirani, R. (2010), *ProDenICA: Product Density Estimation for ICA using Tilted Gaussian Density Estimates*. R Package Version 1.0.

Hoeffding, W. (1961), "The Strong Law of Large Numbers for U-Statistics,", Technical Report, North Carolina State University, Department of Statistics.

Hotelling, H. (1936), "Relations Between Two Sets of Variates," *Biometrika*, 28(3-4), 321.

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

Hyvärinen, A., and Oja, E. (1997), "A Fast Fixed-Point Algorithm for Independent Component Analysis," *Neural Computation*, 9(7), 1483–1492.

Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2010), "A New Performance Index for ICA: Properties, Computation and Asymptotic Analysis," *Latent Variable Analysis and Signal Separation*, pp. 229–236.

Ilmonen, P., and Paindaveine, D. (2011), "Semiparametrically efficient inference based on signed ranks in symmetric independent component models," *Annals of Statistics*, 39(5), 2448–2476.

Karvanen, J. (2005), "A resampling test for the total independence of stationary time series: Application to the performance evaluation of ica algorithms," *Neural Processing Letters*, 22(3), 311–324.

Matteson, D. S. (2008), Statistical Inference for Multivariate Nonlinear Time Series, PhD thesis, The University of Chicago.

Matteson, D. S., and Tsay, R. S. (2011), "Dynamic Orthogonal Components for Multivariate Time Series," *Journal of the American Statistical Association*, 106(496), 1450–1463.

Nordhausen, K., Cardoso, J.-F., Oja, H., and Ollila, E. (2011), *JADE: JADE and ICA Performance Criteria*. R Package Version 1.0-4.

Nordhausen, K., Oja, H., and Paindaveine, D. (2009), "Signed-rank tests for location in the symmetric independent component model," *Journal of Multivariate Analysis*, 100(5), 821–834.

Paindaveine, D., Oja, H., and Taskinen, S. (2009), "Rank Tests for Multivariate Independence in Independent Component Models," *ECARES working paper 018*, .

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rémillard, B. (2009), "Discussion of: Brownian Distance Covariance," *Annals of Applied Statistics*, 3(4), 1295–1298.

Risk, B. B., James, N. A., and Matteson, D. S. (2015), *steadyICA: ICA and Tests of Independence via Multivariate Distance Covariance*. R Package Version 1.0.
URL: https://cran.r-project.org/web/packages/steadyICA/

Stögbauer, H., Kraskov, A., Astakhov, S. A., and Grassberger, P. (2004), "Least-dependent-component analysis based on mutual information," *Physical Review E*, 70(6), 066123.

Székely, G. J., and Rizzo, M. L. (2009), "Brownian Distance Covariance," *Annals of Applied Statistics*, 3(4), 1236–1265.

Székely, G. J., and Rizzo, M. L. (2012), "On the uniqueness of distance covariance," *Statistics & Probability Letters*, 82(12), 2278–2282.

Székely, G. J., and Rizzo, M. L. (2013), "Energy statistics: A class of statistics based on distances," *Journal of Statistical Planning and Inference*, 143(8), 1249–1272.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *Annals of Statistics*, 35(6), 2769–2794.

Wu, E. H., Yu, P. L., and Li, W. (2009), "A smoothed bootstrap test for independence based on mutual information," *Computational Statistics & Data Analysis*, 53(7), 2524–2536.

Zielinski, R. (2007), "Kernel Estimators and the Dvoretzky-Kiefer-Wolfowitz Inequality," *Applicationes Mathematicae*, 34(4), 401.

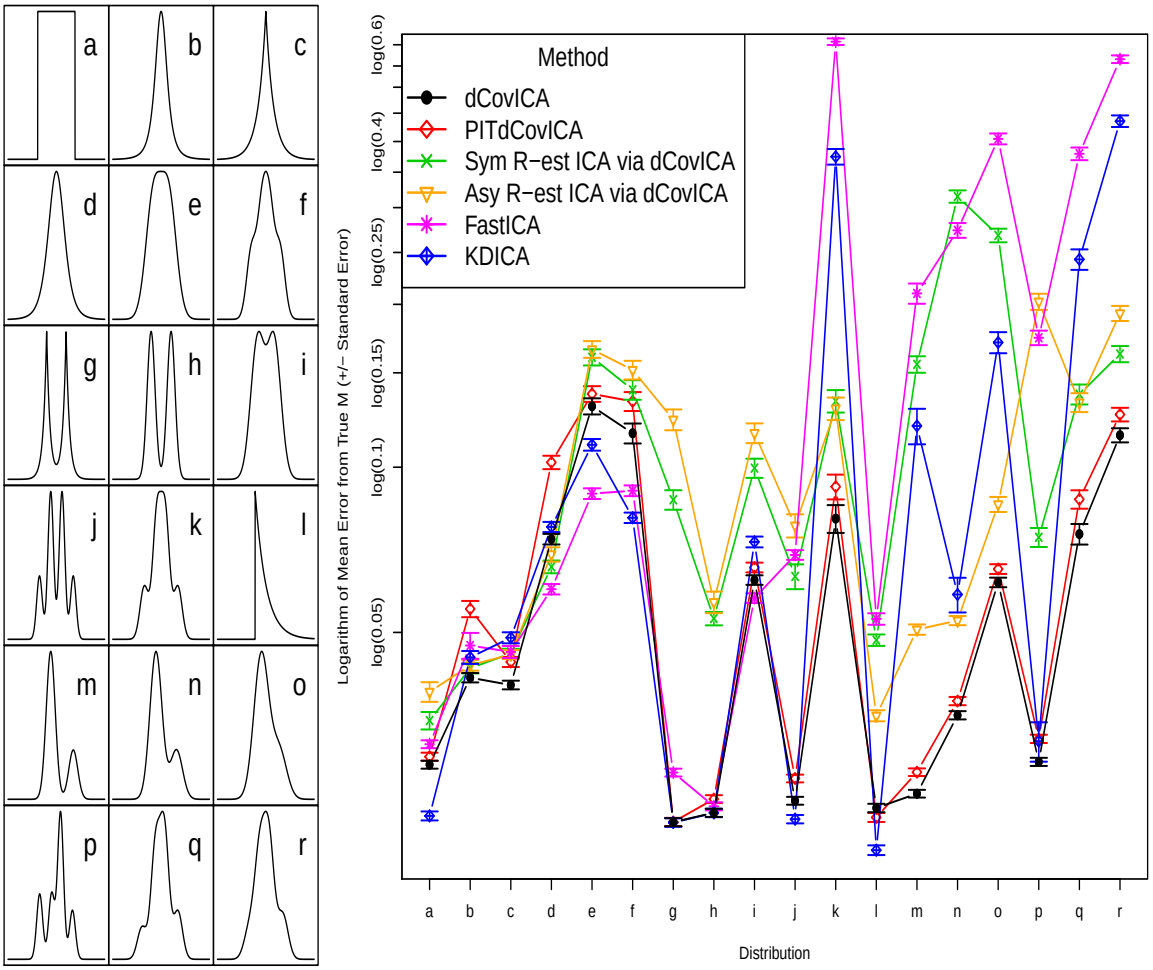

Figure 1: The left panel shows 18 distributions used for comparisons. These include the uniform, Student-t, exponential, mixtures of exponentials, and Gaussian mixtures; distributions $a - k$ are symmetric and $l - r$ are asymmetric The right panel show the logarithm of mean error distance, Equation (13), for each method and each distribution, based on $N = 1{,}000$ simulations in $\mathbb{R}^2$ with sample size $n = 1{,}000$ for each distribution. Vertical bars denote approximate standard errors.

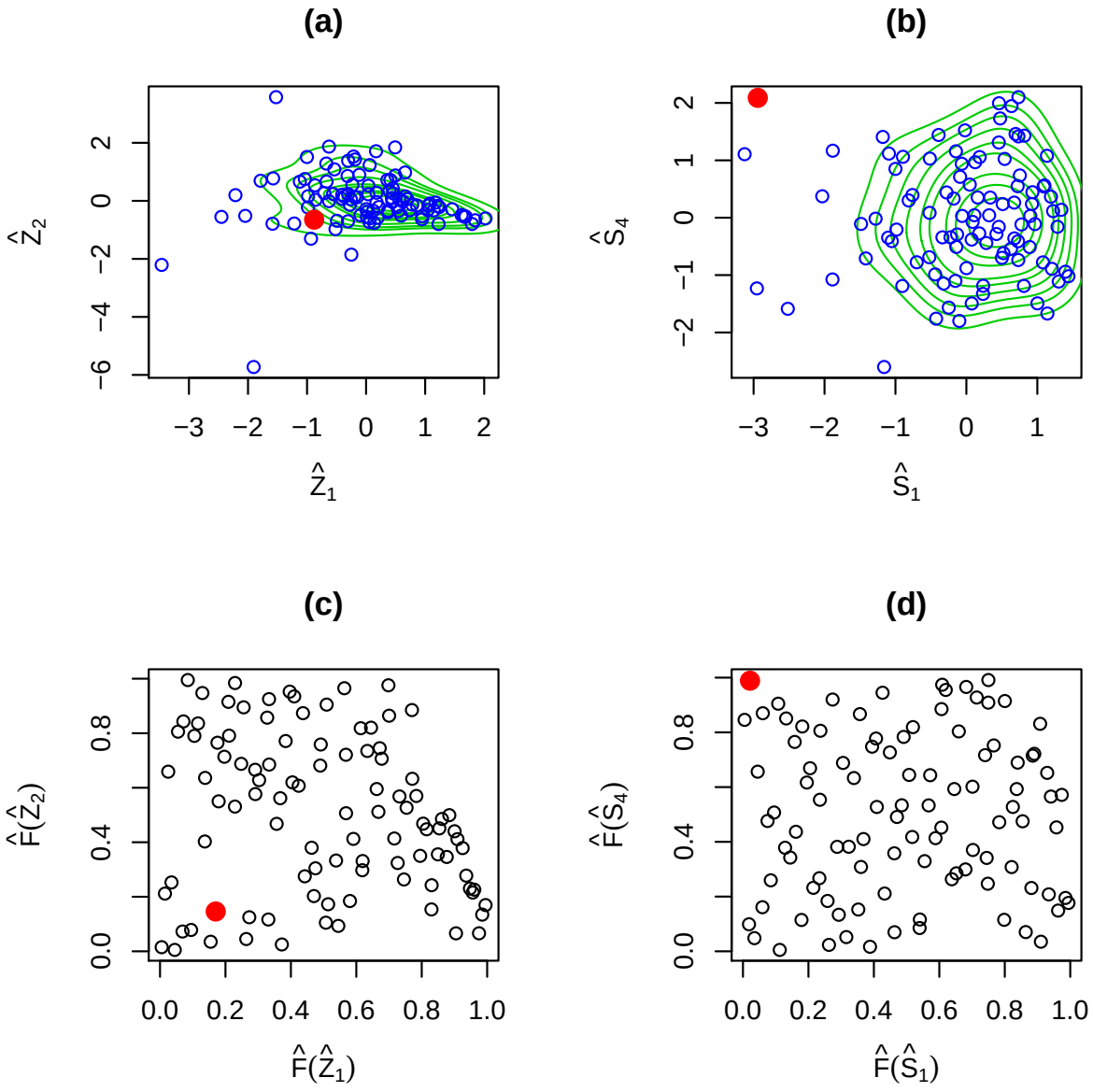

Figure 2: The Freedman data based on crime rates in US metropolitan areas with 1968 populations of 250,000 or more. We consider four variables: the logarithm of population (total 1968, in thousands), nonwhite (percent nonwhite population, 1960), density (population per square mile, 1968), crime (crime rate per 100,000, 1969). (a) first two principal component scores; (b) two estimated independent components; (c) first two principal component scores and (d) two estimated independent components, each after taking the probability integral transformation defined by Equation (8). Estimated contour lines have been drawn for each decile and Philadelphia is indicated on the plot as a larger solid point.

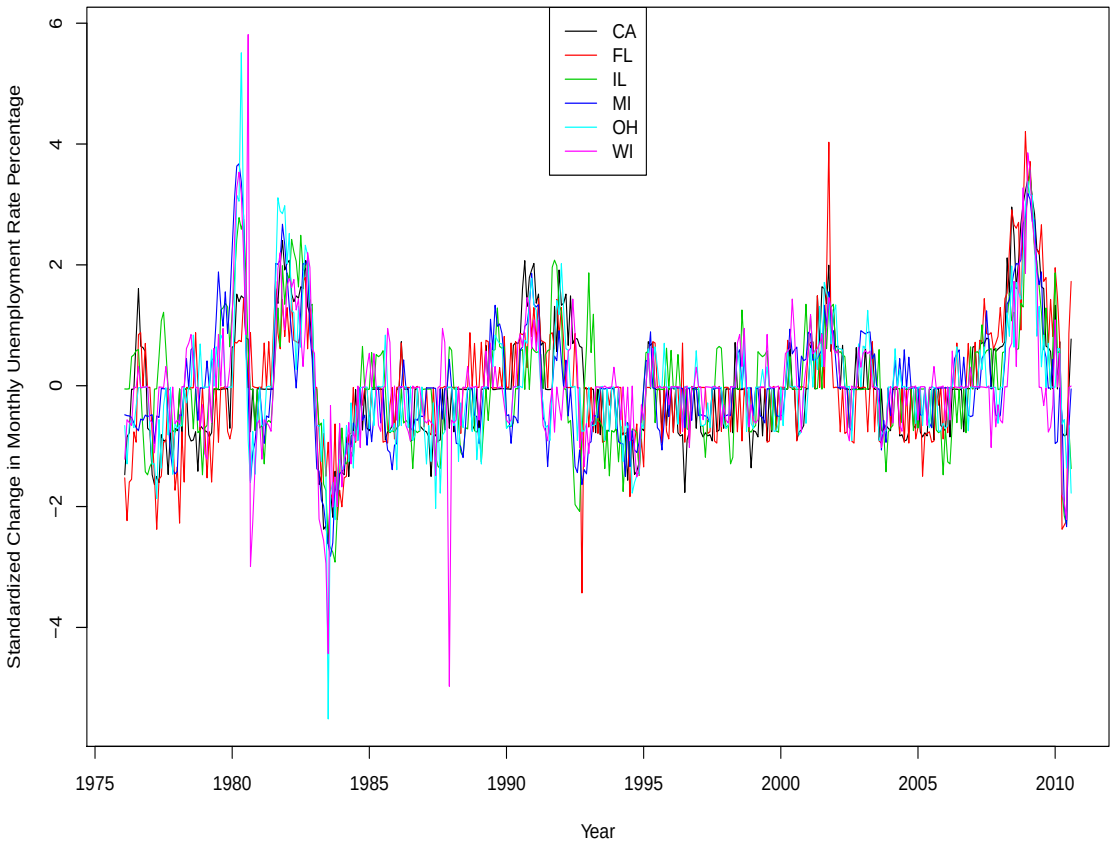

Figure 3: Standardized Change in Monthly Unemployment Rate Percentage for California, Florida, Illinois, Michigan, Ohio, and Wisconsin. This vector series appears stationary, but exhibits serial dependence.

Table 1: Mean error distance (×100), Equation (13), approximate standard error, and mean computation time in seconds (s) for $N = 1{,}000$ simulations in $\mathbb{R}^4, \mathbb{R}^8$, and $\mathbb{R}^{16}$ with sample size $n = 1{,}000$: randomly selecting with replacement from the 18 distributions shown in Figure 1.

		Joint Estimation		Sequential Estimation				Asymmetric
		dCovICA	PITdCovICA	dCovICA	PITdCovICA	FastICA	KDICA	R-est ICA
$\mathbb{R}^4$	Mean Error	8.075	8.036	8.335	9.020	19.381	9.951	17.664
	Standard Error	0.150	0.132	0.175	0.196	0.512	0.362	0.415
	Mean Time (s)	8.96	23.77	1.64	9.92	0.02	0.18	1.67
$\mathbb{R}^8$	Mean Error	8.600	8.628	14.844	14.070	32.070	20.439	*N/A*
	Standard Error	0.040	0.039	0.196	0.167	0.476	0.479	*N/A*
	Mean Time (s)	26.77	64.80	18.40	146.06	0.07	5.54	*N/A*
$\mathbb{R}^{16}$	Mean Error	8.884	8.878	23.505	19.760	48.396	39.930	*N/A*
	Standard Error	0.019	0.019	0.180	0.127	0.364	0.473	*N/A*
	Mean Time (s)	66.56	124.86	73.14	534.57	0.12	156.61	*N/A*

Table 2: ICA of the Freedman crime data, using the PITdCovICA estimator. The standardized observations $\widehat{\boldsymbol{Y}}$ consist of: the logarithm of population (total 1968, in thousands), nonwhite (percent nonwhite population, 1960), density (population per square mile, 1968), crime (crime rate per 100,000, 1969). The fitted mixing matrix and its inverse are shown below. They define the relationship between the observations and the estimated ICs $\widehat{\boldsymbol{S}}$.

$\widehat{\boldsymbol{M}}_n' : \widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{S}}\widehat{\boldsymbol{M}}_n'$				$\widehat{\boldsymbol{M}}_n'^{-1} : \widehat{\boldsymbol{S}} = \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{M}}_n'^{-1}$			
0.10	0.55	0.40	0.75	-0.56	-0.33	-0.53	0.91
-0.55	0.66	-0.33	-0.43	0.36	0.84	-0.18	0.48
-0.40	-0.31	0.68	-0.50	0.49	-0.17	0.88	0.29
0.73	0.40	0.52	0.01	0.88	-0.49	-0.27	-0.64

Table 3: Test statistic $\mathcal{U}_n(\cdot)$ (see Equation (3)), and approximate p-value (based on 1999 permutations) for joint test of mutually independent components for the seasonally adjusted monthly unemployment rates from January 1976 through August 2010 for 6 states: $\widehat{\boldsymbol{Y}}$ standardized observations; $\widehat{\boldsymbol{E}}$ VAR(3) residuals; $\widehat{\boldsymbol{Z}}$ estimated PCs from $\widehat{\boldsymbol{E}}$; and $\widehat{\boldsymbol{S}}$ estimated ICs from $\widehat{\boldsymbol{E}}$.

$\mathcal{U}_n(\cdot)$	$\widehat{\boldsymbol{Y}}$	$\widehat{\boldsymbol{E}}$	$\widehat{\boldsymbol{Z}}$	$\widehat{\boldsymbol{S}}$
Test Statistic	39.7	5.27	0.41	-0.42
Approx. p-value	0	0	0	0.91

Table 4: Test statistic $Q_d(\cdot, m)$, see Equation (14), and approximate p-value (based on 1999 permutations) for $m = 12$ lag joint test of multivariate serial dependence of the seasonally adjusted monthly unemployment rates from January 1976 through August 2010 for $d = 6$ states: $\widehat{\boldsymbol{Y}}$ standardized observations; $\widehat{\boldsymbol{E}}$ VAR(3) residuals; $\widehat{\boldsymbol{Z}}$ estimated PCs from $\widehat{\boldsymbol{E}}$; and $\widehat{\boldsymbol{S}}$ estimated ICs from $\widehat{\boldsymbol{E}}$.

$Q_6(\cdot, m = 12)$	$\widehat{\boldsymbol{Y}}$	$\widehat{\boldsymbol{E}}$	$\widehat{\boldsymbol{Z}}$	$\widehat{\boldsymbol{S}}$
Test Statistic	30.92	0.10	-0.02	-0.02
Approx. p-value	0	0.08	0.54	0.54

Table 5: ICA of the standardized change in monthly unemployment rate percentage, using the PITdCovICA estimator. The standardized observations $\widehat{\boldsymbol{E}}$ consist of state level unemployment for: CA, FL, IL, MI, OH, and WI. These series were rescaled by $\widehat{\boldsymbol{D}}$ to have unit variance. The fitted mixing matrix and its inverse are shown below, along with the estimated uncorrelating matrix. They define the relationship between the observations and the estimated PCs $\widehat{\boldsymbol{Z}}$ and ICs $\widehat{\boldsymbol{S}}$.

(a) $\widehat{\boldsymbol{M}}_n' : \widehat{\boldsymbol{E}} = \widehat{\boldsymbol{S}}\widehat{\boldsymbol{M}}_n'\widehat{\boldsymbol{D}}$

-0.89	-0.24	0.05	-0.33	-0.11	0.07
-0.11	-0.10	-0.32	-0.85	-0.65	-0.48
-0.09	0.09	-0.87	-0.12	0.06	0.32
0.36	-0.11	0.27	-0.11	-0.19	0.81
-0.01	-0.83	-0.07	-0.16	0.45	-0.07
0.23	0.48	0.24	-0.34	0.56	0.09

(b) $\widehat{\boldsymbol{O}}_n' : \widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{E}}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{O}}_n'$

-0.34	0.32	0.11	0.03	0.58	-0.84
-0.19	0.76	-0.17	-0.23	-0.15	0.62
-0.26	-0.17	-0.67	0.72	0.06	0.19
-0.34	-0.19	-0.09	-0.34	-0.91	-0.38
-0.29	-0.43	-0.03	-0.59	0.56	0.49
-0.27	-0.07	0.79	0.50	-0.15	0.39

(c) $\widehat{\boldsymbol{M}}_n'^{-1} : \widehat{\boldsymbol{S}} = \widehat{\boldsymbol{E}}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{M}}_n'^{-1}$

-1.01	0.30	-0.08	0.31	0.24	0.06
0.04	-0.06	0.17	-0.23	-0.86	0.53
0.27	-0.08	-0.95	0.28	-0.08	0.22
-0.22	-0.69	-0.05	-0.31	-0.25	-0.75
0.07	-0.41	0.20	-0.37	0.53	0.75
0.35	-0.31	0.42	0.85	-0.11	0.04