Independent Component Analysis Based On Mutual Dependence Measures

Ze Jin
Department of Statistical Science
Cornell University
Ithaca, NY, USA
zj58@cornell.edu

David S. Matteson
Department of Statistical Science
Cornell University
Ithaca, NY, USA
matteson@cornell.edu

Tianrong Zhang
Microsoft
Redmond, WA, USA
trzhang@microsoft.com

***Abstract*—We apply both distance-based and kernel-based mutual dependence measures to independent component analysis (ICA), and generalize dCovICA to MDMICA, minimizing empirical dependence measures as an objective function in both deflation and parallel manners. Solving this minimization problem, we introduce Latin hypercube sampling (LHS), and a global optimization method, Bayesian optimization (BO) to improve the initialization of the Newton-type local optimization method. The performance of MDMICA is evaluated in various simulation studies and an image data example. When the ICA model is correct, MDMICA achieves competitive results compared to existing approaches. When the ICA model is misspecified, the estimated independent components are less mutually dependent than the observed components using MDMICA, while the estimated independent components are prone to be even more mutually dependent than the observed components using other approaches.**

***Index Terms*—independent component analysis, mutual dependence measures, multivariate analysis, random sampling, global optimization**

I. Introduction

Since most natural processes have multiple components, multivariate analysis is more compelling than univariate analysis. Nevertheless, multivariate analysis is considerably more complicated than univariate analysis, because it accounts for the mutual dependence between all variables. Due to the curse of dimensionality, it becomes essential to interpret multivariate data through a simplified representation via dimension reduction.

Independent component analysis (ICA) represents multivariate data by mutually independent components (ICs). Thus, linear combinations of ICs capture the structure of multivariate data even when other linear projection methods, such as principal component analysis (PCA), are not sufficient. As a classical unsupervised learning method, ICA has been developed for applications including blind source separation, feature extraction, brain imaging, etc. [1] provide a comprehensive overview of ICA approaches for estimating ICs.

Let $Y = (Y_1, \ldots, Y_d)' \in \mathbb{R}^d$ be a random vector as observations. Assume that Y has a nonsingular, continuous distribution F_Y, with $\mathrm{E}(Y_j) = 0$ and $\mathrm{Var}(Y_j) < \infty$, $j = 1, \ldots, d$. Let $X = (X_1, \ldots, X_d)' \in \mathbb{R}^d$ be a random vector as ICs. According to the fundamental assumption of ICA, the univariate components $X_1, \ldots, X_d$ are mutually independent, and at most one component X_j is Gaussian. Without loss of generality, X is assumed to be standardized such that $\mathrm{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$, $j = 1, \ldots, d$. A linear latent factor model to estimate X from Y is given by

$$Y = MX,$$

where $M \in \mathbb{R}^{d \times d}$ is a nonsingular mixing matrix.

Prewhitened random variables are uncorrelated and thus more convenient to work with from both practical and theoretical perspectives. Let $\Sigma_Y = \mathrm{Cov}(Y)$ be the covariance matrix of Y, and $H = \Sigma_Y^{-1/2}$ be an uncorrelating matrix. Let $Z = HY = (Z_1, \ldots, Z_d)' \in \mathbb{R}^d$ be a random vector as uncorrelated observations, such that $\Sigma_Z = \mathrm{Cov}(Z) = I_d$, the $d \times d$ identity matrix. Then the relation between Z and X is

$$X = M^{-1}Y = M^{-1}H^{-1}Z \triangleq WZ, \tag{1}$$

where $W = M^{-1}H^{-1} \in \mathbb{R}^{d \times d}$ is a nonsingular unmixing matrix. Given that $Z_1, \ldots, Z_d$ are uncorrelated, W is an orthogonal matrix, with $d(d-1)/2$ free elements rather than d^2. We aim to simultaneously estimate W and X, such that the components of X satisfy the assumption of mutual independence.

Many popular ICA approaches minimize the mutual information or maximize the non-Gaussianity of the estimated components under the constraint that they are uncorrelated. Examples include the fourth-moment matrix diagonalization of JADE [2], the information criterion of Infomax [3], the maximum negentropy of FastICA [4], and the maximum likelihood principle of ProDenICA [5] and Spline-LCA [6], [7].

Some other ICA approaches minimize the mutual dependence between the estimated components using a specific dependence measure. While dependence measures have been extensively studied, two classes have attracted a great deal of attention. One is the distance-based energy statistics [8]. [9] proposed distance covariance (dCov) to measure pairwise dependence, and [10] extended it to mutual dependence measures (MDMs). Another is the kernel-based maximum mean discrepancies (MMDs) [11]. [12] proposed Hilbert−Schmidt in-

Research support from an NSF award (DMS-1455172), a Xerox PARC Faculty Research Award, and Cornell University Atkinson Center for a Sustainable Future (AVF-2017).

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00107

dependence criterion (HSIC) to measure pairwise dependence, and [13] generalized it to d-variable Hilbert−Schmidt independence criterion (dHSIC) measuring mutual dependence. [14] showed that these two classes of measures are equivalent in the sense that MMDs can be interpreted as energy statistics with a distance kernel, and energy statistics can be interpreted as MMDs with a negative-type semimetric.

Meanwhile, [15] applied a characteristic function-based dependence measure to ICA, for which [10] provided a closed-form expression as an MDM and studied its asymptotic properties. [16] applied a kernel-based dependence measure to ICA, which was formulated as an HSIC in [12]. Motivated by the properties of HSIC, [17] proposed FastKICA based on a mutual dependence measure extension, which is the sum of all pairwise HSIC while its 0 value does not imply mutual independence. Inspired by the properties of dCov, [18] proposed dCovICA based on another mutual dependence measure extension, which is a sum of squared dCov and equals 0 if and only if mutual independence holds.

However, [18] only demonstrated the results of a single measure from the class of energy-statistics, using multiple values to initialize the local optimization without any comparison. Thus, in this paper, we generalize dCovICA to a new approach, MDMICA, by applying the mutual dependence measures proposed in [10] and [13], and make two contributions as follows. First, we extend its ICA framework to accommodate mutual dependence measures from both classes of energy statistics and MMDs, and compare the performance of these measures in numerical studies. Second, we study the non-convex optimization problem when estimating ICs under this ICA framework, and investigate the improvement of using multiple values over a single value for initialization through Latin hypercube sampling, a random sampling method. In addition, we introduce a global optimization method, Bayesian optimization, to further improve the initialization of local optimization.

The rest of this paper is organized as follows. We generalize the ICA framework of dCovICA in Section II. In Section III, we give a brief overview of dCov and MDMs, propose the new ICA approach, MDMICA, based on MDMs, and derive its asymptotic properties. In Section IV, we introduce Latin hypercube sampling and Bayesian optimization to aid the initialization of subsequent local optimization method when estimating ICs. We present the simulation results in Section V, and a real data example in Section VI[1] Finally, Section VII summarizes our work.

II. ICA Framework

For $d \geq 2$, the group of $d \times d$ orthogonal matrices is denoted by $\mathcal{O}(d)$, and its subgroup with determinant 1 is denoted by $\mathcal{SO}(d)$. For $i \neq j$, we start with the identity matrix I_d, and substitute $\cos(\psi)$ for the (i,i) and (j,j) elements, $-\sin(\psi)$ for the (i,j) element, and $\sin(\psi)$ for the (j,i) element, then we obtain a Givens rotation matrix denoted by $G_{i,j}(\psi)$.

[1]An accompanying R package `EDMeasure` [19] is available on CRAN.

Let $\theta = \{\theta_{i,j} : 1 \leq i < j \leq d\}$ denote a vector of rotation angles with length $p = d(d-1)/2$, and let $\theta_i = \{\theta_{i,j} : i < j \leq d\}$ such that $\theta = \{\theta_i : 1 \leq i \leq d-1\}$. Then any rotation matrix $W \in \mathcal{SO}(d)$ can be parameterized via θ as $W(\theta)$, or equivalently a product of p Givens rotation matrices determined by θ as

$$W(\theta) = G^{(d-1)}(\theta_{d-1}) \dots G^{(1)}(\theta_1),$$

where $G^{(k)}(\theta_k) = G_{k,d}(\theta_{k,d}) \dots G_{k,k+1}(\theta_{k,k+1})$ represents the rotations of the kth row with respect to all the ℓth rows, $\ell > k$. We observe that the kth row of $W(\theta)$ is the same as the kth row of the partial product $G^{(k)}(\theta_k) \dots G^{(1)}(\theta_1)$. As a result, let $X(\theta) = W(\theta)Z$, then the subset of angles in $\{\theta_{i,j} : 1 \leq i \leq k, i < j \leq d\} = \{\theta_i : 1 \leq i \leq k\}$ fully determines the kth element of X. We define a support of θ as

$$\Theta = \left\{ \theta_{i,j} : \begin{cases} 0 \leq \theta_{1,j} \leq 2\pi, \\ 0 \leq \theta_{i,j} < \pi, \quad i \neq 1. \end{cases} \right\}, \tag{2}$$

and its subset with respect to θ_i as Θ_i.

Unfortunately, the non-identification issue regarding W and X still exists because the sign and order of the components are not identifiable. Given any signed permutation matrix $P_\pm$, (1) is equivalent to

$$(P_\pm X) = P_\pm X = P_\pm W Z = (P_\pm W)Z,$$

where $P_\pm X$ and $P_\pm W$ become an alternative to X and W, as the new ICs and unmixing matrix. However, the identification up to a signed permutation is adequate in terms of modeling multivariate data by linear combinations of ICs. To make a fair comparison between different estimates, a metric invariant to the three ambiguities, scale, sign, and order of the ICs will be presented in Section V.

Let $\mathbf{Y} \in \mathbb{R}^{n \times d}$ be an i.i.d. sample of observations from F_Y, where $\mathbf{Y}_j \in \mathbb{R}^n$ is an i.i.d. sample of observations from F_{Y_j}, $j = 1, \dots, d$. Let $\widehat{\Sigma}_{\mathbf{Y}}$ be the sample covariance matrix of $\mathbf{Y}$, and $\widehat{H} = \widehat{\Sigma}_{\mathbf{Y}}^{-1/2}$ be the estimated uncorrelating matrix. Although Σ_Y is unknown in practice, the sample covariance is a consistent estimate under the finite second-moment assumption, i.e., $\widehat{\Sigma}_{\mathbf{Y}} \xrightarrow{a.s.} \Sigma_Y$ as $n \to \infty$. Let $\widehat{\mathbf{Z}} = \mathbf{Y}\widehat{H}' \in \mathbb{R}^{n \times d}$ be the estimated uncorrelated observations, such that $\widehat{\Sigma}_{\widehat{\mathbf{Z}}} = I_d$, and $\Sigma_{\widehat{\mathbf{Z}}} \xrightarrow{a.s.} I_d$ as $n \to \infty$.

To simplify notation, we assume that $\mathbf{Z}$, an uncorrelated i.i.d. sample is given, with mean zero and unit variance. Let $\mathbf{X}(\theta) = \mathbf{Z}W(\theta)' \in \mathbb{R}^{n \times d}$ be a sample of X. Then we estimate $W(\theta)$ through θ, and define an ICA estimator as

$$\widehat{\theta} = \underset{\theta \in \Theta}{\arg\min}\ f(\mathbf{X}(\theta)) = \underset{\theta \in \Theta}{\arg\min}\ f(\mathbf{Z}W(\theta)'), \tag{3}$$

where f is an objective function measuring the mutual dependence among $\mathbf{X}(\theta)$. Given the estimate $\widehat{\theta}$, the estimated unmixing matrix is $\widehat{W} = W(\widehat{\theta})$, and the estimated ICs are $\widehat{\mathbf{X}} = \mathbf{X}(\widehat{\theta}) = \mathbf{Z}\widehat{W}' = \mathbf{Z}W(\widehat{\theta})'$.

III. Applying MDM to ICA

We reduce the estimation of ICs to the problem of choosing the function f in (3), which is expected to be a measure of mutual dependence. Following [18], we primarily focus on distance-based energy statistics because of their compact representations as expectations of pairwise Euclidean distances, while all the results can be easily extended to kernel-based MMDs according to the equivalence between these two classes in [14].

We use $(\cdot,\cdot,\ldots,\cdot)$ to concatenate (vector) components into a vector. Let $t=(t_1,\ldots,t_d), X=(X_1,\ldots,X_d)\in\mathbb{R}^p$ where $t_j, X_j\in\mathbb{R}^{p_j}$, p_j is a marginal dimension, $j=1,\ldots,d$, and $p=\sum_{j=1}^d p_j$ is the total dimension. The subset of components to the right of X_c is denoted by $X_{c^+}=(X_{c+1},\ldots,X_d)$, $c=0,1,\ldots,d-1$. The subset of components excluding X_c is denoted by $X_{-c}=(X_1,\ldots,X_{c-1},X_{c^+})$, $c=1,\ldots,d-1$. The "X" under the assumption that $X_1,\ldots,X_d$ are mutually independent is denoted by $\widetilde{X}=(\widetilde{X}_1,\ldots,\widetilde{X}_d)$, where $\widetilde{X}_j\overset{\mathcal{D}}{=}X_j, j=1,\ldots,d$, $\widetilde{X}_1,\ldots,\widetilde{X}_d$ are mutually independent, while $X,\widetilde{X}$ are independent. Let X',X'' be independent copies of X such that X',X'' have the same distribution as X, while they are all independent, i.e., $X,X',X''\overset{i.i.d.}{\sim}F_X$, and $\widetilde{X}'$ be an independent copy of $\widetilde{X}$. The Euclidean norm of X is denoted by $|X|$. The weighted $\mathcal{L}_2$ norm $\|\cdot\|_w$ of any complex-valued function $\eta(t)$ is defined by $\|\eta(t)\|_w^2=\int_{\mathbb{R}^p}|\eta(t)|^2w(t)\,dt$ where $|\eta(t)|^2=\eta(t)\overline{\eta(t)}$, $\overline{\eta(t)}$ is the complex conjugate of $\eta(t)$, and $w(t)$ is any positive weight function for which the integral exists.

Let $\mathbf{X}=\{X^k=(X_1^k,\ldots,X_d^k):k=1,\ldots,n\}$ be an i.i.d. sample from F_X, the joint distribution of X, then $\mathbf{X}_j=\{X_j^k:k=1,\ldots,n\}$ is an i.i.d. sample from F_{X_j}, the marginal distribution of X_j, $j=1,\ldots,d$, such that $\mathbf{X}=\{\mathbf{X}_1,\ldots,\mathbf{X}_d\}$. Denote the joint characteristic function of X as $\phi_X(t)=\mathrm{E}[e^{i\langle t,X\rangle}]$ and its empirical version as $\phi_X^n(t)=\frac{1}{n}\sum_{k=1}^n e^{i\langle t,X^k\rangle}$, and the joint characteristic function of $\widetilde{X}$ as $\phi_{\widetilde{X}}(t)=\prod_{j=1}^d\mathrm{E}[e^{i\langle t_j,X_j\rangle}]$, and its empirical version as $\phi_{\widetilde{X}}^n(t)=\prod_{j=1}^d(\frac{1}{n}\sum_{k=1}^n e^{i\langle t_j,X_j^k\rangle})$. In addition, a simplified empirical version of $\phi_{\widetilde{X}}(t)$ is defined by $\phi_{\widetilde{X}}^{n\star}(t)=\frac{1}{n}\sum_{k=1}^n e^{i\langle t,(X_1^k,\ldots,X_d^{k+d-1})\rangle}$ to substitute $\phi_{\widetilde{X}}^n(t)$ as a simplification, where X_j^{n+k} is interpreted as X_j^k for $k>0$.

We will first review dCov [9] and MDMs [10] in Section III-A, III-B and III-C, and then propose a new ICA approach, MDMICA based on MDMs in Section III-D.

A. Distance Covariance $(d=2)$

[9] proposed distance covariance to capture non-linear and non-monotone pairwise dependence between two random vectors, i.e., $X=(X_1,X_2)$.

The nonnegative distance covariance $\mathcal{V}(X)$ is defined by $\mathcal{V}^2(X)=\|\phi_X(t)-\phi_{\widetilde{X}}(t)\|_{w_1}^2$ where the weight $w_1(t)=(K_{p_1}K_{p_2}|t_1|^{p_1+1}|t_2|^{p_2+1})^{-1}$, $K_q=\frac{2\pi^{q/2}\Gamma(1/2)}{2\Gamma((q+1)/2)}$, and Γ is the gamma function. If $\mathrm{E}|X|<\infty$, then $\mathcal{V}(X)\in[0,\infty)$, and $\mathcal{V}(X)=0$ if and only if X_1,X_2 are pairwise independent. In addition, if $\mathrm{E}|X_1X_2|<\infty$, $\mathcal{V}^2(X)$ can be interpreted as expectations

$$\begin{aligned}\mathcal{V}^2(X) &= \mathrm{E}|X_1-X_1'||X_2-X_2'|\\ &\quad+\mathrm{E}|X_1-X_1'|\mathrm{E}|X_2-X_2'|\\ &\quad-2\mathrm{E}|X_1-X_1'||X_2-X_2''|.\end{aligned}$$

We estimate $\mathcal{V}(X)$ by replacing the characteristic functions with the empirical characteristic functions. The nonnegative empirical distance covariance $\mathcal{V}_n(\mathbf{X})$ is defined by $\mathcal{V}_n^2(\mathbf{X})=\|\phi_X^n(t)-\phi_{\widetilde{X}}^n(t)\|_{w_1}^2$, which can be interpreted as complete V-statistics

$$\begin{aligned}\mathcal{V}_n^2(\mathbf{X}) &= \frac{1}{n^2}\sum_{k,\ell=1}^n|X_1^k-X_1^\ell||X_2^k-X_2^\ell|\\ &\quad+\frac{1}{n^2}\sum_{k,\ell=1}^n|X_1^k-X_1^\ell|\frac{1}{n^2}\sum_{k,\ell=1}^n|X_2^k-X_2^\ell|\\ &\quad-\frac{2}{n^3}\sum_{k,\ell,m=1}^n|X_1^k-X_1^\ell||X_2^k-X_2^m|.\end{aligned}$$

Calculating $\mathcal{V}_n^2(\mathbf{X})$ via the symmetry of Euclidian distances has the time complexity $O(n^2)$. If $\mathrm{E}|X|<\infty$, then we have $\mathcal{V}_n(\mathbf{X})\xrightarrow{a.s.}\mathcal{V}(X)$ as $n\to\infty$.

[10] generalized distance covariance to three mutual dependence measures capturing any form of mutual dependence between multiple random vectors, which include the asymmetric, symmetric, and complete measures below.

B. Asymmetric and Symmetric Measures $(d\geq 2)$

The asymmetric and symmetric measures of mutual dependence $\mathcal{R}(X),\mathcal{S}(X)$ are defined by

$$\mathcal{R}(X)=\sum_{c=1}^{d-1}\mathcal{V}^2((X_c,X_{c^+})),\ \mathcal{S}(X)=\sum_{c=1}^{d}\mathcal{V}^2((X_c,X_{-c})).$$

Analogous to $\mathcal{V}(X)$, if $\mathrm{E}|X|<\infty$, then $\mathcal{R}(X),\mathcal{S}(X)\in[0,\infty)$, and $\mathcal{R}(X),\mathcal{S}(X)=0$ if and only if $X_1,\ldots,X_d$ are mutually independent.

Similarly, the empirical asymmetric and symmetric measures of mutual dependence $\mathcal{R}_n(\mathbf{X}),\mathcal{S}_n(\mathbf{X})$ are defined by $\mathcal{R}_n(\mathbf{X})=\sum_{c=1}^{d-1}\mathcal{V}_n^2((\mathbf{X}_c,\mathbf{X}_{c^+}))$, $\mathcal{S}_n(\mathbf{X})=\sum_{c=1}^{d}\mathcal{V}_n^2((\mathbf{X}_c,\mathbf{X}_{-c}))$, which can be implemented with the time complexity $O(n^2)$. If $\mathrm{E}|X|<\infty$, then we have $\mathcal{R}_n(\mathbf{X})\xrightarrow{a.s.}\mathcal{R}(X)$ and $\mathcal{S}_n(\mathbf{X})\xrightarrow{a.s.}\mathcal{S}(X)$ as $n\to\infty$.

C. Complete Measure $(d\geq 2)$

The complete measure of mutual dependence $\mathcal{Q}(X)$ is defined by $\mathcal{Q}(X)=\|\phi_X(t)-\phi_{\widetilde{X}}(t)\|_{w_2}^2$ where $w_2(t)=(K_p|t|^{p+1})^{-1}$, $K_q=\frac{2\pi^{q/2}\Gamma(1/2)}{2\Gamma((q+1)/2)}$, and Γ is the gamma function. If $\mathrm{E}|X|<\infty$, then $\mathcal{Q}(X)\in[0,\infty)$, and $\mathcal{Q}(X)=0$ if and only if $X_1,\ldots,X_d$ are mutually independent. In addition, $\mathcal{Q}(X)$ can be interpreted as expectations

$$\mathcal{Q}(X)=\mathrm{E}|X-\widetilde{X}'|+\mathrm{E}|X'-\widetilde{X}|-\mathrm{E}|X-X'|-\mathrm{E}|\widetilde{X}-\widetilde{X}'|.$$

We estimate $\mathcal{Q}(X)$ using the simplified empirical complete measure of mutual dependence $\mathcal{Q}_n^\star(\mathbf{X})$, defined by $\mathcal{Q}_n^\star(\mathbf{X})=$

$\|\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)\|_{w_2}^2 = \int_{\mathbb{R}^p} |\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)|^2 w_2(t)\, dt$, which can be interpreted as incomplete V-statistics

$$\begin{aligned}\mathcal{Q}_n^\star(\mathbf{X}) &= \frac{2}{n^2}\sum_{k,\ell=1}^{n} |X^k - (X_1^\ell, \ldots, X_d^{\ell+d-1})| \\ &+ \frac{1}{n^2}\sum_{k,\ell=1}^{n} |X^k - X^\ell| \\ &- \frac{1}{n^2}\sum_{k,\ell=1}^{n} |(X_1^k, \ldots, X_d^{k+d-1}) - (X_1^\ell, \ldots, X_d^{\ell+d-1})|.\end{aligned}$$

The naive implementation of $\mathcal{Q}_n^\star(\mathbf{X})$ has the time complexity $O(n^2)$. If $\mathrm{E}|X| < \infty$, then $\mathcal{Q}_n^\star(\mathbf{X}) \xrightarrow{a.s.} \mathcal{Q}(X)$ as $n \to \infty$.

D. MDMICA Approach and Its Asymptotic Properties

Inspired by the nice equivalence to mutual independence and time complexity $O(n^2)$ of MDMs, we propose an I-CA approach, MDMICA based on MDMs. To be specific, we define three MDMICA estimators, i.e., MDMICA (asy), MDMICA (sym), and MDMICA (com) by applying $f(\mathbf{X}) = \mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X}), \mathcal{Q}_n^\star(\mathbf{X})$ in (3) respectively as

$$\widehat{\theta}_n^{\text{asy}} = \arg\min_{\theta\in\Theta} \mathcal{R}_n(\mathbf{X}(\theta)) = \arg\min_{\theta\in\Theta} \mathcal{R}_n(\mathbf{Z}W(\theta)'), \quad (4)$$

and similar expressions follow for $\widehat{\theta}_n^{\text{sym}}, \widehat{\theta}_n^{\text{com}}$. Further, we define another estimator, MDMICA (hsic), by applying dHSIC in the same way.

Since the ICA model only allows scalar components, we apply a special case of MDMs to ICA where the marginal dimension $p_j = 1$, $j = 1, \ldots, d$, and the total dimension $p = d$. Without loss of generality, we assume that $\mathrm{E}(Y) = 0$ and $\mathrm{Cov}(Y) = I_d$, and therefore $Z = Y$ and $\mathbf{Z} = \mathbf{Y}$ throughout this section. Let $\overline{\Theta}$ denote a large enough compact subset of the space Θ defined by (2). The asymptotic properties of the MDMICA estimators are derived as follows.

Theorem 1: If Y has a nonsingular, continuous distribution F_Y with $\mathrm{E}|Y|^2 < \infty$, if there exists a unique minimizer $\theta_0 \in \overline{\Theta}$ of (4), and if $W(\theta_0)$ satisfies the conditions for a unique continuous inverse to exist, then $\widehat{\theta}_n^{\text{asy}} \xrightarrow{a.s.} \theta_0$ as $n \to \infty$.

When the ICA model is misspecified, convergence to the pseudo-true value θ_0 is obtained. Under similar conditions, $\widehat{\theta}_n^{\text{sym}}, \widehat{\theta}_n^{\text{com}}$ also converges a.s. as $n \to \infty$ due to similar arguments.

We then establish the root-n consistency of the MDMICA estimators under some regularity conditions no matter whether the ICA model holds or is misspecified.

Theorem 2: If the assumptions of Theorem 1 hold, and if the ICA model assumptions hold, then $|\widehat{\theta}_n^{\text{asy}} - \theta_0| = O_P(n^{-1/2})$.

Theorem 3: If the ICA model is misspecified but the remaining assumptions stated in Theorem 2 hold, and if $\mathrm{E}[\frac{\partial}{\partial\theta}\mathcal{R}_n(\mathbf{X}(\theta))|_{\theta=\theta_0}] = o_P(n^{-1/2})$, where θ_0 denotes the pseudo-true value, then $|\widehat{\theta}_n^{\text{asy}} - \theta_0| = O_P(n^{-1/2})$.

Under similar conditions, $\widehat{\theta}_n^{\text{sym}}, \widehat{\theta}_n^{\text{com}}$ are also consistent as $n \to \infty$ due to similar arguments.

The proofs of Theorem 1, 2, and 3 are similar to those of Theorem 2.1, 2.2, and Corollary 2.1 in [18] respectively, considering the same nature of $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X}), \mathcal{Q}_n^\star(\mathbf{X})$ as energy statistics, and replacing the empirical cumulative distribution function (ECDF) with the identity function in derivations.

IV. Improving Initialization of Local Methods

In the literature, there are two primary schemes to estimate ICs with regard to how the optimization is implemented. For one, the components are extracted one at a time, known as the deflation scheme. For another, the components are extracted simultaneously, known as the parallel scheme. The deflation scheme has the advantage of lower computational cost over the parallel scheme. However, the parallel scheme enjoys greater statistical efficiency, as the deflation scheme accumulates estimation uncertainty at each step in its sequential procedure.

For our ICA framework, the objective function f in (3) has $d(d-1)/2$ parameters $\theta_{i,j} \in \theta$, which can be estimated in both deflation (sequential) and parallel (joint) manners. Specifically, the deflation scheme estimates all $\theta_{i,j} \in \theta$ for each i at a time, while the parallel scheme estimates all $\theta_{i,j} \in \theta$ together at once.

In view of the special structures of associated measures, both deflation and parallel schemes are appropriate for MDMICA (asy), denoted by MDMICA (asy, def) and MDMICA (asy, par), while MDMICA (sym), MDMICA (com), and MDMICA (hsic) only fit the parallel scheme. The MDMICA algorithms for both deflation and parallel schemes are described in Alg. 1 below.

Algorithm 1 MDMICA ($\mathbf{Z}, f$)

1. Initialize θ and $W(\theta)$ via θ.
2. (deflation scheme)
   for $i = 1, \cdots, d-1$ **do**
   a. Solve $\widehat{\theta}_i = \arg\min_{\theta_i\in\Theta_i} f(\mathbf{Z}W(\theta)')$ using newton-type local optimization.
   b. Update $\theta_i \leftarrow \widehat{\theta}_i$.
   end for

2'. (parallel scheme)
   Solve $\widehat{\theta} = \arg\min_{\theta\in\Theta} f(\mathbf{Z}W(\theta)')$ using newton-type local optimization.

3. Output $\widehat{\theta} = \{\widehat{\theta}_i : 1 \le i \le d-1\}$, $\widehat{W} = W(\widehat{\theta})$, and $\widehat{\mathbf{X}} = \mathbf{Z}W(\widehat{\theta})'$.

Estimating θ through (3) involves minimization of a non-convex but locally convex objective function f, which requires initialization and iterative algorithms. The default method for MDMICA is a Newton-type local optimization method, for which we explore two ways of finding a good initialization.

The first way is to perform a random sampling method, Latin hypercube sampling (LHS) [20] uniformly over the space Θ to obtain a number of parameter values. Then we evaluate the objective function at each value and record the value minimizing it, which is used to initialize the subsequent local

optimization algorithm. Based on our experience, the number of parameter values sampled should grow with the dimension.

The second way is to take advantage of a global optimization method, Bayesian optimization (BO) [21], where the objective function f is treated as a black box. It is applicable when the function is expensive to evaluate, the derivative is unavailable, or the optimization problem is non-convex. Bayesian optimization is one of the most efficient approaches in terms of the number of function evaluations consumed, as [22] illustrated that it outperforms other state-of-the-art global optimization algorithms on a number of challenging problems.

We will apply LHS and LHS + BO to improve the performance of MDMICA through providing better initialization in Section V. When it comes to the comparison between LHS and LHS + BO, LHS leads to less computation time in some cases, and LHS + BO leads to higher estimation accuracy in some cases.

V. Simulation Studies

In this section, we evaluate the performance of our MDMICA estimators by performing simulations similar to [18], and compare them with the FastICA estimator, the Infomax estimator, and the JADE estimator. MDMICA (asy) is omitted because it is the same as dCovICA. Moreover, we elaborate on the implementation and error metric of ICA.

Furthermore, we try various options for each estimator. For FastICA, we evaluate logarithm of hyperbolic cosine (logcosh) used to approximate negentropy in both deflation and parallel schemes, and omit kurtosis (kur) and exponential (exp) since their performance is similar. For Infomax, we evaluate hyperbolic tangent (tanh) as the nonlinear (squashing) function, and similarly omit logistic (log) and extended Infomax (ext). For MDMICA (hsic), we use the Gaussian (gau) kernel.

We simulate the ICs $\mathbf{X} \in \mathbb{R}^{n\times d}$ from eighteen distributions using `rjordan` in the R package `ProDenICA` [23] with sample size n and dimension d. See Figure 1 for the density functions of the eighteen distributions. Then we generate a mixing matrix $M \in \mathbb{R}^{d\times d}$ with condition number between 1 and 2 using `mixmat` in the R package `ProDenICA` [23], and obtain the observations $\mathbf{Y} = \mathbf{X}M'$, which are centered by their sample mean and then prewhitened by their sample covariance to obtain uncorrelated observations $\mathbf{Z} = \mathbf{Y}\widehat{H}'$. Finally, we obtain the estimate $\widehat{W}$ based on $\mathbf{Z}$ via (3), and evaluate the estimation accuracy by comparing the estimate $\widehat{W}$ to the ground truth $W_0 = (\widehat{H}M)^{-1}$. Moreover, the Newton-type local optimization is implemented by `nlm` in the R package `stats`, and Bayesian optimization is implemented by `mbo` in the R package `mlrMBO` [24] with the Matérn 3/2 kernel.

To take the uncertainty in both prewhitening the observations and estimating the ICs into account when comparing the estimates from different approaches, we use the metric MD proposed by [25] to measure the error between an estimate $\widehat{W}$ and the corresponding truth W_0, which is defined as

$$\mathrm{MD}(\widehat{W}, W_0) = \frac{1}{\sqrt{d-1}} \inf_{P,D} \|PD\widehat{W}W_0^{-1} - I_d\|_F,$$

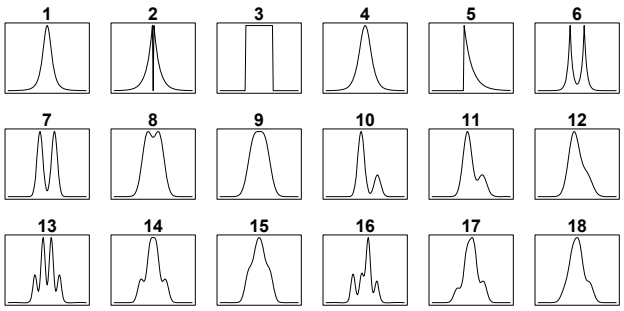

Fig. 1. Density plots of the 18 distributions.

where $\|\cdot\|_F$ denotes the Frobenius norm, P is a $d \times d$ permutation matrix, and D is a $d \times d$ diagonal matrix with nonzero diagonal elements. MD is the minimum distance between components from the estimate and the truth in terms of matching each component from the truth with the closest component from the estimate after permutation and scaling. It is invariant to the three ambiguities associated with ICA as a result of taking the infimum, and is scaled to a value between 0 and 1 where 0 corresponds to a perfect recovery of the truth.

Experiment 1: [Different distributions of ICs] We sample $\mathbf{X}$ from one distribution in the eighteen distributions, with $d = 3$, $n = 1000$. We obtain $10d$ points using LHS, and select the best initial point. See Figure 2 for the error metrics of all eighteen distributions with 100 trials.

MDMICA achieves competitive results with JADE and dCovICA, and also outperforms FastICA and Infomax in most cases. MDMICA (sym) is equal and often better than dCovICA, while they have similar performance due to their similar structures. Similarly, MDMICA (hsic) is equal and often better than MDMICA (com), while they have similar performance due to their similar structures. Further, MDMICA (com) and MDMICA (hsic) are less sensitive to different distributions than dCovICA and MDMICA (sym) in general. Lastly, there is no remarkable difference between the deflation and parallel schemes.

Experiment 2: [Different dimensions of ICs] We sample $\mathbf{X}$ from one distribution in the eighteen distributions, with $d \in \{2, 3, 4\}$, $n = 1000$. We pick $10d$ points using LHS, and select the best initial point. See Figure 3 for the error metrics of the 1st distribution with 100 trials.

The errors of all estimators increase as the dimension d grows. As in the previous experiment, JADE, dCovICA, and MDMICA have similar performance, and significantly outperform FastICA and Infomax. The computational time of dCovICA and MDMICA grows with the number of parameters, e.g., MDMICA (com) takes 4.75s, 11.50s, 33.28s for $d = 2, 3, 4$ respectively.

Experiment 3: [Different initializations of local optimization] We sample $\mathbf{X}$ from d randomly selected distributions of the eighteen distributions, with $d = 4$, $n = 1000$. We implement three ways to select the initial point for the Newton-

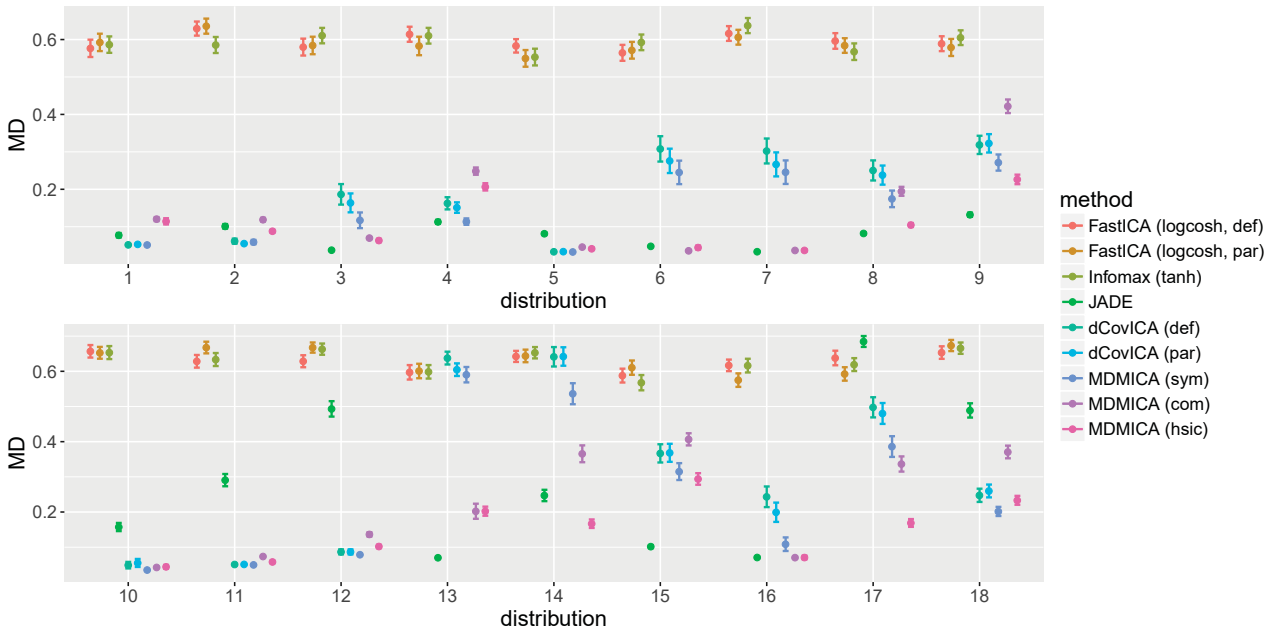

Fig. 2. Error metrics (mean $\pm$ standard error) of all eighteen distributions with 100 trials for Experiment 1.

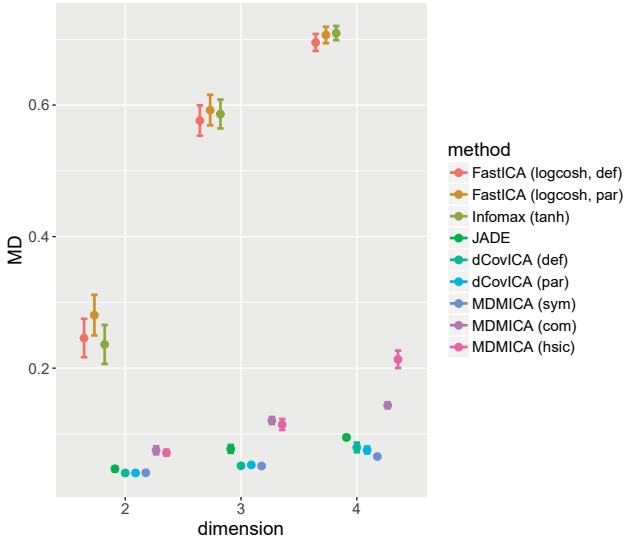

Fig. 3. Error metrics (mean $\pm$ standard error) of the 1st distribution with 100 trials for Experiment 2.

type local optimization method. The first way is to sample one point using LHS. The second way is to sample $20d$ points using LHS, and then select the point out of $20d$ with the lowest objective. The third way is to run $10d$ iterations using BO, with its initial points from $10d$ sampled points using LHS, and then select the point out of $20d$ with the lowest objective. Note that both the second way and third way run $20d$ evaluations on the objective function for a fair comparison. See Table I for the error metrics, objective values, and computational time of the tuple as the (4th, 11th, 12th, 18th) distributions with 100 trials.

The performance of dCovICA and MDMICA is greatly improved by selecting the best point from multiple initial points, as LHS and LHS + BO produce smaller objective values and more accurate estimates than a single initial point. The reason is two-fold. First, LHS and BO offer the subsequent local optimization method better initial points in terms of lower objective, which leads to a better estimate in terms of lower objective as well. Second, a better estimate with lower objective is likely to be a better solution with lower MD, since the objective is a truly mutual dependence measure. Moreover, LHS + BO has noticeable advantage over LHS alone for MDMICA (com) and MDMICA (hsic), while it is similar to LHS alone for dCovICA (def), dCovICA (par), and MDMICA (sym).

dCovICA and MDMICA take remarkably longer computational time than the others, which makes sense because the optimization problem of dCovICA and MDMICA has $d(d-1)/2$ parameters and is much more difficult to solve. This obstacle in turn motivates us to improve the local optimization by choosing a better initialization point. As LHS and BO provide better initial points for the subsequent local optimization method, the local optimization time is reduced and the total time is not necessarily longer compared to using a single initial point.

Experiment 4: [Misspecified ICA model] We sample $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$ from one distribution in the eighteen distributions, with $n = 1000$. Let $\mathbf{Y_1} = \mathbf{X_1}$, $\mathbf{Y_2} = (\mathbf{X_2})^2$. We pick $10d$ points using LHS, and select the best initial point. See Table II for the results of the 1st distribution with 1 trial.

We use $\mathcal{R}_n, \mathcal{S}_n, \mathcal{Q}_n^\star$ to measure the mutual dependence

TABLE I

ERROR METRICS (MEAN ± STANDARD ERROR), OBJECTIVE VALUES (MEAN ± STANDARD ERROR), COMPUTATIONAL TIME (MEAN) IN INITIALIZATION (LHS, BO) AND LOCAL OPTIMIZATION (NEWTON-TYPE), AND TOTAL COMPUTATIONAL TIME (MEAN) OF THE TUPLE AS THE (4TH, 11TH, 12TH, 18TH) DISTRIBUTIONS WITH 100 TRIALS FOR EXPERIMENT 3.

Estimator	Initialization	MD (10^{-1})	Obj (10^{-3})	Total Time (Init + Local Opt) (s)
FastICA (logcosh, def)	LHS ($20d$)	6.780 ± 0.124	-	0.16 (0.13 + 0.03)
FastICA (logcosh, par)	LHS ($20d$)	6.978 ± 0.106	-	0.18 (0.11 + 0.07)
Infomax (tanh)	LHS (1)	6.861 ± 0.113	-	0.05 (0.00 + 0.05)
JADE	LHS (1)	3.992 ± 0.156	-	0.01 (0.00 + 0.01)
dCovICA (def)	LHS (1)	1.334 ± 0.105	4.090 ± 0.124	210.76 (0.00 + 210.76)
	LHS ($20d$)	1.133 ± 0.047	3.959 ± 0.047	395.96 (221.19 + 174.77)
	LHS ($10d$) + BO ($10d$)	1.128 ± 0.050	3.941 ± 0.048	381.18 (207.03 + 174.15)
dCovICA (par)	LHS (1)	1.458 ± 0.111	4.029 ± 0.060	910.03 (0.00 + 910.03)
	LHS ($20d$)	1.356 ± 0.086	3.944 ± 0.047	980.55 (147.49 + 833.06)
	LHS ($10d$) + BO ($10d$)	1.375 ± 0.099	3.963 ± 0.064	1014.75 (205.96 + 808.79)
MDMICA (sym)	LHS (1)	1.134 ± 0.066	6.961 ± 0.065	1179.84 (0.00 + 1179.84)
	LHS ($20d$)	1.057 ± 0.035	6.947 ± 0.063	1072.22 (196.65 + 875.57)
	LHS ($10d$) + BO ($10d$)	1.094 ± 0.052	6.952 ± 0.065	1115.78 (258.58 + 857.20)
MDMICA (com)	LHS (1)	2.725 ± 0.186	2.037 ± 0.093	92.58 (0.00 + 92.58)
	LHS ($20d$)	2.064 ± 0.097	1.671 ± 0.015	54.88 (16.19 + 38.69)
	LHS ($10d$) + BO ($10d$)	1.964 ± 0.096	1.673 ± 0.014	112.10 (78.71 + 33.39)
MDMICA (hsic)	LHS (1)	3.981 ± 0.243	1.385 ± 0.091	267.13 (0.00 + 267.13)
	LHS ($20d$)	2.521 ± 0.152	0.834 ± 0.019	306.03 (40.29 + 265.74)
	LHS ($10d$) + BO ($10d$)	2.208 ± 0.135	0.797 ± 0.017	402.13 (106.10 + 296.03)

TABLE II

MUTUAL DEPENDENCE MEASURES OF OBSERVED COMPONENTS (BEFORE OPTIMIZATION, $\mathbf{Z}$) AND ESTIMATED INDEPENDENT COMPONENTS (AFTER OPTIMIZATION, $\widehat{\mathbf{X}}$) WITH 1 TRIAL FOR EXPERIMENT 4 (MISSPECIFIED ICA MODEL).

Estimator	$\mathcal{R}_n(\mathbf{Z})$ (10^{-3})	$\mathcal{R}_n(\widehat{\mathbf{X}})$	$\mathcal{S}_n(\mathbf{Z})$ (10^{-3})	$\mathcal{S}_n(\widehat{\mathbf{X}})$	$\mathcal{Q}_n^\star(\mathbf{Z})$ (10^{-4})	$\mathcal{Q}_n^\star(\widehat{\mathbf{X}})$
FastICA (logcosh, def)		0.531		1.062		3.088
FastICA (logcosh, par)		0.588		1.176		2.786
Infomax (tanh)		0.606		1.212		3.081
JADE		1.031		2.062		3.330
dCovICA (def)	0.548	0.441	1.097	0.882	2.797	2.677
dCovICA (par)		0.441		0.882		2.677
MDMICA (sym)		0.441		0.882		2.677
MDMICA (com)		0.446		0.892		2.672
MDMICA (hsic)		0.443		0.887		2.687

between the components before (w.r.t. $\mathbf{Z}$) and after (w.r.t. $\widehat{\mathbf{X}}$) the optimization. dCovICA and MDMICA successfully decreases the mutual dependence between the components through optimization, while FastICA, Infomax, and JADE are unable to and even increase it. Thus, ICA methods based on mutual dependence measures outperform others in reducing the mutual dependence when the ICA model is misspecified.

VI. IMAGE DATA

Fulfilling the task of unmixing vectorized images, we consider the three gray-scale images in the R package `ICS` [26], depicting a cat, a forest road, and a sheep respectively. Each image is represented by a 130×130 matrix, where each element indicates the intensity value of a pixel. We standardize the three images such that the intensity values across all the pixels in each image have mean zero and unit variance. Then we vectorize each image into a vector of length 130^2, and combine the vectors from all three images as a matrix $\mathbf{X}$, with $d = 3$, $n = 130^2$.

We use `mixmat` in the R package `ProDenICA` [23] again to generate a mixing matrix $A \in \mathbb{R}^{p\times p}$, and mix the three images to obtain the observations $\mathbf{Y} = \mathbf{X}A^T$, which are centered by their sample mean, then prewhitened by their sample covariance to obtain uncorrelated observations $\mathbf{Z} = \mathbf{Y}\widehat{H}^T$.

We estimate the intensity values $\widehat{\mathbf{S}}$ initialized from $10d$ points using LHS. See Figures 4 for the recovered images, where the Euclidean norm of vectorized errors is computed to evaluate the estimation accuracy. Indicated by the estimated images and errors, dCovICA and MDMICA outperforms JADE. Moreover, MDMICA (com) achieves the best overall performance.

VII. CONCLUSION.

Resorting to recently proposed mutual dependence measures including MDMs in [10] and dHSIC in [13], we generalize dCovICA in [18] to a new ICA approach, MDMICA, taking empirical dependence measures as an objective function for the estimation of ICs. In addition, we study the asymptotic properties of MDMICA.

When solving the non-convex minimization problem to estimate ICs, we apply LHS and BO to select a better initial point for the Newton-type local optimization method, and improve the performance of MDMICA.

MDMICA achieves competitive results with JADE and dCovICA, and outperforms FastICA and Infomax in numerical

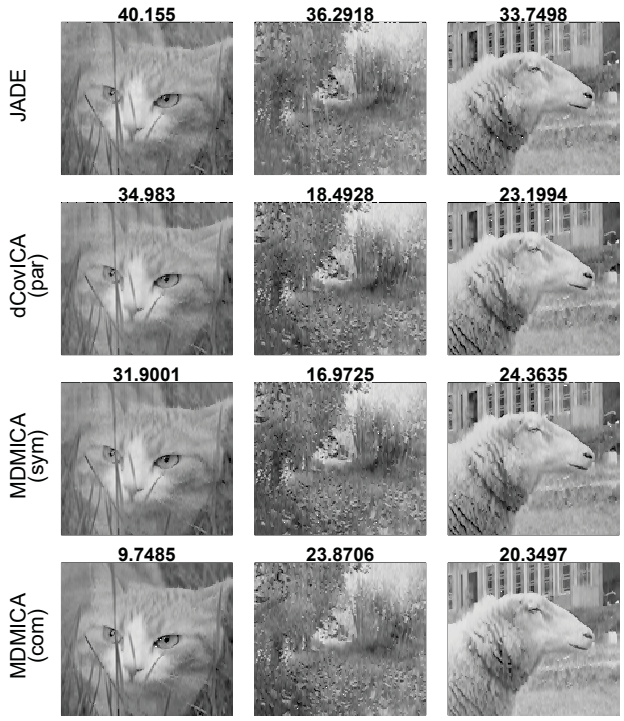

Fig. 4. Recovered images with $d = 3$, $n = 130^2$ for the image data. Each value on title is the Euclidean norm of the vectorized errors of the recovered image. A signed permutation is applied to the images for illustration.

studies, under different distributions and dimensions of ICs. When the ICA model is misspecified, MDMICA decreases the mutual dependence between components via optimization, while other approaches cannot and even increase it. We illustrate the advantage of using multiple initial points from LHS and BO over a single initial point.

During the image recovery task from mixed image data, MDMICA not only nicely recovers the true images, but also achieves lower overall errors than other approaches, which demonstrates the value of MDMICA in real data applications.

REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[2] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," in *IEE proceedings F (radar and signal processing)*, vol. 140, no. 6. IET, 1993, pp. 362–370.

[3] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[4] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[5] T. Hastie and R. Tibshirani, "Independent components analysis through product density estimation," in *Advances in neural information processing systems*, 2003, pp. 665–672.

[6] B. B. Risk, D. S. Matteson, and D. Ruppert, "Linear non-gaussian component analysis via maximum likelihood," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 332–343, 2019.

[7] Z. Jin, B. B. Risk, and D. S. Matteson, "Optimization and testing in linear non-gaussian component analysis," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 12, no. 3, pp. 141–156, 2019.

[8] G. J. Székely and M. L. Rizzo, "Energy statistics: A class of statistics based on distances," *Journal of statistical planning and inference*, vol. 143, no. 8, pp. 1249–1272, 2013.

[9] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The annals of statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[10] Z. Jin and D. S. Matteson, "Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete v-statistics," *Journal of Multivariate Analysis*, vol. 168, pp. 304–322, 2018.

[11] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Advances in neural information processing systems*, 2007, pp. 513–520.

[12] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2075–2129, 2005.

[13] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters, "Kernel-based tests for joint independence," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 1, pp. 5–31, 2018.

[14] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and rkhs-based statistics in hypothesis testing," *The Annals of Statistics*, pp. 2263–2291, 2013.

[15] A. Chen and P. J. Bickel, "Consistent independent component analysis and prewhitening," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3625–3632, 2005.

[16] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002.

[17] H. Shen, S. Jegelka, and A. Gretton, "Fast kernel ica using an approximate newton method," in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 476–483.

[18] D. S. Matteson and R. S. Tsay, "Independent component analysis via distance covariance," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 623–637, 2017.

[19] Z. Jin, S. Yao, D. S. Matteson, and X. Shao, *EDMeasure: Energy-Based Dependence Measures*, 2018, R package version 1.2.

[20] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 42, no. 1, pp. 55–61, 2000.

[21] J. Mockus, "Application of bayesian approach to numerical methods of global and stochastic optimization," *Journal of Global Optimization*, vol. 4, no. 4, pp. 347–365, 1994.

[22] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[23] T. Hastie and R. Tibshirani, *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*, 2010, R package version 1.0.

[24] B. Bischl, J. Richter, J. Bossek, D. Horn, M. Lang, and J. Thomas, *mlrMBO: Bayesian Optimization and Model-Based Optimization of Expensive Black-Box Functions*, 2018, R package version 1.1.

[25] P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila, "A new performance index for ica: properties, computation and asymptotic analysis," *Latent Variable Analysis and Signal Separation*, pp. 229–236, 2010.

[26] K. Nordhausen, H. Oja, and D. E. Tyler, "Tools for exploring multivariate data: The package ics," *Journal of Statistical Software*, vol. 28, no. 6, pp. 1–31, 2008.