

Linear Non-Gaussian Component Analysis via Maximum Likelihood

Benjamin B. Risk^{1,2,3}, David S. Matteson¹, David Ruppert¹

¹Department of Statistical Science, Cornell University

²SAMSI, Research Triangle Park, North Carolina

³Department of Biostatistics, University of North Carolina, Chapel Hill

Abstract

Independent component analysis (ICA) is popular in many applications, including cognitive neuroscience and signal processing. Due to computational constraints, principal component analysis is used for dimension reduction prior to ICA (PCA+ICA), which could remove important information. The problem is that interesting independent components (ICs) could be mixed in several principal components that are discarded and then these ICs cannot be recovered. We formulate a linear non-Gaussian component model with Gaussian noise components. To estimate this model, we propose likelihood component analysis (LCA), in which dimension reduction and latent variable estimation is achieved simultaneously. Our method orders components by their marginal likelihood rather than ordering components by variance as in PCA. We present a parametric LCA using the logistic density and a semi-parametric LCA using tilted Gaussians with cubic B-splines. Our algorithm is scalable to datasets common in applications (e.g., hundreds of thousands of observations across hundreds of variables with dozens of latent components). In simulations, latent components are recovered that are discarded by PCA+ICA methods. We apply our method to multivariate data and demonstrate that LCA is a useful data visualization and dimension reduction tool that reveals features not apparent from PCA or PCA+ICA. We also apply our method to an fMRI experiment from the Human Connectome Project and identify artifacts missed by PCA+ICA. We present theoretical results on identifiability of the linear non-Gaussian component model and consistency of LCA.

Keywords: Functional Magnetic Resonance Imaging, Independent Component Analysis, Neuroimaging, Non-Gaussian Component Analysis, Principal Component Analysis, Projection Pursuit

1 Introduction

The classic independent component analysis (ICA) model is $\mathbf{X} = \mathbf{MS}$ where \mathbf{X} is an observed vector, \mathbf{S} is a latent vector of independent random variables, and \mathbf{M} is a square matrix called the mixing matrix. It is assumed that we have a sample $\{\mathbf{x}_v\}$, $v = 1, \dots, V$, with corresponding latent $\{\mathbf{s}_v\}$. The goal is to estimate \mathbf{M} and $\{\mathbf{s}_v\}$. Popular ICA methodology does not directly attempt to find components that are independent but rather components that are as non-Gaussian as possible by maximizing an approximation of negentropy (Hyvärinen and Oja, 2000). The principle here is that any sum of ICs will be closer to Gaussian distributed than the ICs themselves. Thus, $\{\mathbf{s}_v\}$ are correctly recovered if they maximize some measure of non-Gaussianity. Moment or cumulant-based methods (Cardoso and Souloumiac, 1993; Virta et al., 2015), maximum likelihood methods (Chen and Bickel, 2006; Samworth and Yuan, 2012), and methods that directly minimize a measure of dependence (Stögbauer et al., 2004; Matteson and Tsay, 2016) have also been developed.

Transformations that maximize non-Gaussianity play a prominent role in many applications including signal processing (Bell and Sejnowski, 1995), estimating brain networks (Beckmann, 2012), face recognition (Bartlett et al., 2002), and artifact removal (Griffanti et al., 2014). In practice, PCA is applied to the observations $\{\mathbf{x}_v\}$ prior to classic ICA (hereafter, PCA+ICA) to meet the assumption of square mixing and to reduce computational costs (Hyvärinen et al., 2001). PCA+ICA is commonly used to identify brain networks in functional magnetic resonance imaging (fMRI) (Beckmann, 2012), but PCA preprocessing can discard parts of the brain networks (Green et al., 2002). PCA+ICA is also used to identify artifacts in single-subject fMRI to improve sensitivity and specificity in subsequent group-level analyses (Pruim et al., 2015). Even though the results from the two-stage PCA+ICA approach have been useful in the applied sciences, our data applications show that a single analysis that uses non-Gaussianity for both dimension reduction and extracting LCs improves estimation.

We propose a novel model for non-Gaussian signals and Gaussian noise that we call linear

non-Gaussian component analysis (LNGCA). Consider a sample $\{\mathbf{x}_v, \mathbf{s}_v, \mathbf{n}_v\}, v = 1, \dots, V$:

$$\mathbf{X} = \mathbf{M}_\mathbf{S}\mathbf{S} + \mathbf{M}_\mathbf{N}\mathbf{N} \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^T$; $\mathbf{S} \in \mathbb{R}^Q$ is a vector of mutually independent non-Gaussian random variables with $1 \leq Q \leq T$; $\mathbf{M}_\mathbf{S} \in \mathbb{R}^{T \times Q}$; $\mathbf{M}_\mathbf{N} \in \mathbb{R}^{T \times (T-Q)}$; $\mathbf{M} = [\mathbf{M}_\mathbf{S}, \mathbf{M}_\mathbf{N}]$ (the concatenation of $\mathbf{M}_\mathbf{S}$ and $\mathbf{M}_\mathbf{N}$) is full rank; and \mathbf{N} is $(T-Q)$ -variate normal. Note that in LNGCA, the dimension of the image of $\mathbf{M}_\mathbf{N}$ is $T-Q$, whereas other models, such as noisy ICA (discussed below), assume the dimension is T . One observes $\{\mathbf{x}_v\}$ while $\{\mathbf{s}_v\}$ and $\{\mathbf{n}_v\}$ are latent. We assume $\mathbf{E}\mathbf{S} = \mathbf{0}$ and $\mathbf{E}\mathbf{N} = \mathbf{0}$ in (1) (in practice, data are centered by their sample mean). Our goal is to estimate $\mathbf{M}_\mathbf{S}$ and the realizations $\{\mathbf{s}_v\}$ of \mathbf{S} , which we call latent components (LCs).

We estimate the LNGCA model using a maximum-likelihood framework, which we call likelihood component analysis (LCA). We introduce this new term to emphasize that our method uses a likelihood as the pertinent measure of information to achieve dimension reduction. The components are ordered according to a parametric or semi-parametric likelihood rather than by variance as in PCA. By simultaneously performing dimension reduction and latent variable estimation, we will demonstrate through simulations and two real applications that estimation of the proposed model allows the discovery of non-Gaussian signals discarded by other methods. Non-Gaussian signals are often discarded by PCA+ICA when they are associated with small variance. When the motivating scientific problem has a low signal-to-noise ratio, LCA is particularly well-suited to recovering the non-Gaussian signals.

1.1 Model motivation

Consider a simple example with two sensors in which the first sensor measures the latent variable with noise and the second sensor measures the noise only. Formally, $\mathbf{X} = [X_1, X_2]$, $X_1 = S_1 + N_1$, and $X_2 = N_1$. Assuming the signal is non-Gaussian and noise is Gaussian, this process follows the LNGCA model. In applications such as fMRI or EEG, there are multiple

sensors (coils or channels) measuring random variables distributed throughout space, and a given random variable may be more accurately measured by some sensors than others. In such situations, the LNGCA model may provide a useful approximation.

In practice, the dimension of the noise subspace, $\text{im}(\mathbf{M}_{\mathbf{N}})$ (where $\text{im}(\mathbf{A})$ denotes the image or range of a matrix \mathbf{A}), in (1) may be greater than or less than $T - Q$. If the dimension is less than $T - Q$, \mathbf{X} can be transformed to a random variable that follows the LNGCA model without loss of information. If there are more than $T - Q$ noise components, then LNGCA can approximate the noise with a $(T - Q)$ -dimensional subspace.

The idea behind LCA is to use the marginal likelihoods rather than marginal variances as the measure of information when defining latent components, since low-variance signal may be removed by PCA. Among the class of absolutely continuous random variables with mean zero and unit variance, the standard Gaussian density has maximum differential entropy (Cover and Thomas, 2006). By assuming the non-Gaussian components in the LNGCA model belong to this class of random variables, their likelihoods are on average higher. Our approach is to constrain the latent distributions to have unit variance, which allows both the marginal likelihoods and the directions to be estimated. Then the latent component with the highest likelihood, i.e., lowest entropy, contains the most information, and the Gaussian components will have the smallest marginal likelihoods.

1.2 Relation to other methods

The special case in which the dimension of $\text{im}(\mathbf{M}_{\mathbf{S}})$ is T or $T - 1$ and $\text{im}(\mathbf{M}_{\mathbf{N}})$ is zero or one, respectively, is equivalent to the classic ICA model (Hyvärinen and Oja, 2000). Note that one Gaussian component is allowed in classic ICA because if $\text{Cov } \mathbf{X} = \mathbf{I}_T$, where \mathbf{I}_T is the $T \times T$ identity matrix, then \mathbf{M} is orthogonal and so its last column can be determined (up to sign) by the previous $T - 1$ columns. This technicality is generally ignored, and we hereafter define classic ICA under the assumption that $\mathbf{M}_{\mathbf{S}}$ is full rank and $\mathbf{M}_{\mathbf{N}} = \mathbf{0}$. The case in which the dimension of $\text{im}(\mathbf{M}_{\mathbf{N}})$ equals T is the noisy ICA model, which is also called

independent factor analysis (IFA) (Attias, 1999). The noisy ICA model often imposes the additional assumption that $\mathbf{M}_{\mathbf{N}} = \sigma^2 \mathbf{I}_T$.

The noisy ICA model can be approximated using a variant of PCA+ICA (Beckmann and Smith, 2004), where probabilistic PCA is used to estimate the number of components and achieve dimension reduction (Tipping and Bishop, 1999). Alternatively, IFA could be used for simultaneous dimension reduction and latent variable estimation wherein the ICs are modeled as Gaussian mixtures (Attias, 1999). It is difficult to apply IFA because an m^Q -dimensional integral, where m is the number of Gaussian mixtures, must be approximated at each iteration of the EM algorithm, which quickly becomes computationally intractable. Allasonniere and Younes (2012) developed stochastic EM algorithms to estimate the IFA model and proposed parametric methods. Guo and Tang (2013) developed a multi-subject IFA model, and Shi and Guo (2016) extended it to include covariates and an approximate EM algorithm that linearly scales with the number of components, although their application to fMRI uses PCA. Amato et al. (2010) develop non-parametric density estimators of the component densities in the noisy ICA model but assume $\mathbf{M}_{\mathbf{S}}$ is semi-orthogonal, which is not realistic for our application.

Other methods exploring non-Gaussian structure in multivariate data include non-Gaussian component analysis (NGCA) and projection pursuit. NGCA is a more general case of (1) that allows non-linear dependence between the non-Gaussian components. However, this comes at the cost that the latent components are not identifiable. The subspace that contains the non-Gaussian signal is estimated using multiple projection pursuit indices or radial basis functions (Blanchard et al., 2006; Kawanabe et al., 2007). Since it does not estimate latent components, NGCA does not lend itself to identifying brain networks and/or artifacts. Projection pursuit is a method without a generative model that seeks “interesting” directions of information by maximizing projection pursuit indices, such as kurtosis (Huber, 1985). Miettinen et al. (2014) used the deflationary FastICA algorithm to adaptively select the projection pursuit index from a family of indices for each non-Gaussian direction;

however, they only consider the case where $Q = T$. One approach to estimating the model in (1) would be to sequentially estimate projection pursuit directions. However, sequential estimates in ICA are sensitive to the order in which the components are extracted, and errors can accumulate (Ollila, 2010). Overall, the LNGCA model in (1) is unique in that it specifies a latent variable model for the non-Gaussian signal (which we show is identifiable) while also defining a subspace containing Gaussian noise, and the LCA estimation procedure is unique because it simultaneously estimates the latent components in the presence of Gaussian noise. See Web Supplement B for additional discussion of these other methods.

In Section 2, we propose parametric and semi-parametric versions of LCA. In Section 3.1, we investigate simulations when the observations of the latent variables are independently and identically distributed. In Section 3.2, we examine model robustness by applying our method to temporally and spatially structured simulated data. In Section 4, we use LCA for data visualization and dimension reduction in multivariate data. In Section 5, we estimate brain networks and artifacts from high-resolution fMRI data from the Human Connectome Project. In Section 6, we present our conclusions and discuss avenues for future research. Code implementing our method and proofs of the theorems appear in the Web Supplement.

2 Methodology

In Section 2.1, we define the conditions for identifiability of the source densities and mixing matrix of the LNGCA model. In Section 2.2, we define a generic estimator and theory that applies when the densities used in the objective function are fixed. In order to derive an optimization algorithm, we need to assess convergence using a measure of closeness suited to our problem, which we define in Section 2.3. In Section 2.4, we develop an estimator of the mixing matrix under the assumption that the LCs have logistic densities, which is a practical example in which our theory applies. In Section 2.5, we develop a flexible method that estimates the densities and the mixing matrix. (Properties of this estimator will be

examined via simulations in Section 3.)

2.1 Model identifiability

Throughout this section we assume (for simplicity) all random variables are mean zero. Define the equivalence relation $\mathbf{B} \cong \mathbf{C}$ if \mathbf{B} equals \mathbf{C} up to scaling and permutation of columns. Let “ $\stackrel{d}{=}$ ” denote equality in distribution. Let $\mathbf{S} = [\mathbf{S}_1, \dots, \mathbf{S}_Q]'$. We state the assumptions of the LNGCA model below.

Assumption 1. $\mathbf{S}_1, \dots, \mathbf{S}_Q$ mutually independent, non-Gaussian random variables with $\mathbb{E} \mathbf{S} = \mathbf{0}$ and $\mathbb{E} \mathbf{S} \mathbf{S}' = \mathbf{I}_Q$.

Assumption 2. $\text{rank}([\mathbf{M}_\mathbf{S}, \mathbf{M}_\mathbf{N}]) = T$

Assumption 3. \mathbf{N} is $(T - Q)$ -variate normal with $\mathbb{E} \mathbf{N} = \mathbf{0}$ and $\mathbb{E} \mathbf{N} \mathbf{N}'$ non-singular.

The following theorem can be established using Theorem 10.3.9 in Kagan et al. (1973).

Theorem 1. Suppose \mathbf{X} follows the model in (1) with Assumptions 1-3. Then for any other representation $\mathbf{X} = \mathbf{M}_\mathbf{S}^* \mathbf{S}^* + \mathbf{E}^*$ where $\mathbf{S}^* \in \mathbb{R}^Q$ are independent non-Gaussian components and \mathbf{E}^* is multivariate normal, we have: $\mathbf{M}_\mathbf{S}^* \cong \mathbf{M}_\mathbf{S}$; $\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$ up to scaling and permutations; and $\mathbf{E}^* \stackrel{d}{=} \mathbf{M}_\mathbf{N} \mathbf{N}$.

All proofs appear in the Web Supplement Section A.

From Theorem 1, the signal, $\mathbf{M}_\mathbf{S} \mathbf{S}$, has a unique decomposition (on the equivalence class of scalings and permutations) into a fixed matrix and independent components. The assumption that \mathbf{M} is full rank is necessary to ensure the uniqueness of the distributions of the latent components, which in turn is necessary for their identifiability. Note that the noise, $\mathbf{M}_\mathbf{N} \mathbf{N}$, does not have a unique decomposition (e.g., if \mathbf{N} comprises independent normals with equal variance, then $\mathbf{M}_\mathbf{N} \mathbf{O}'$ and $\mathbf{O} \mathbf{N}$ for orthogonal \mathbf{O} is another decomposition with independent components).

Without loss of generality, we will assume that \mathbf{N} is standard multivariate normal. Let $\{f_q\}$ be the true densities of the LCs (the signal components). For the purposes of this paper, we will also assume $\{f_q\}$ are absolutely continuous, although identifiability holds more generally. Denote the eigenvalue decomposition (EVD) of the covariance matrix of \mathbf{X} by $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$. Let $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{U}'$ be a whitening matrix (the covariance matrix of $\mathbf{L}\mathbf{X}$ is \mathbf{I}_T), and define $\mathbf{W} = \mathbf{M}^{-1}\mathbf{L}^{-1}$ where $\mathbf{M} = [\mathbf{M}_S, \mathbf{M}_N]$. Note that $\mathbf{W} \in \mathcal{O}_{T \times T}$, where $\mathcal{O}_{T \times T}$ is the class of $T \times T$ orthogonal matrices. Let \mathbf{w}'_q denote the q th row of \mathbf{W} , and let \mathbf{W}_S denote the first q rows. Let $\phi(x)$ denote the standard normal density. Noting that $|\det \mathbf{W}| = 1$, we have

$$f_{\mathbf{X}}(\mathbf{x}|\mathbf{W}, \mathbf{L}) = \det(\mathbf{L}) \prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{L}\mathbf{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{L}\mathbf{x}). \quad (2)$$

Note that for a density and its corresponding row of the mixing matrix, $\{f_q, \mathbf{w}_q\}$, we can trivially define a density $f_q^*(x) = f_q(-x)$ and vector $\mathbf{w}_q^* = -\mathbf{w}_q$ such that $f_q^*(\mathbf{w}_q^{*'} \mathbf{x}) = f_q(\mathbf{w}'_q \mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^T$. In this sense, we say the density and vector pair, $\{f_q, \mathbf{w}_q\}$, is identifiable up to sign. We can now establish the identifiability of the LNGCA model.

Corollary 1. *Suppose the linear structure model in (1) with density defined in (2) and suppose that Assumptions 1-3 hold. Then $\{f_1, \mathbf{w}_1\}, \dots, \{f_Q, \mathbf{w}_Q\}$ are identifiable up to sign and ordering. Note the rows \mathbf{w}_{Q+k} for $k = 1, \dots, T - Q$ are not identifiable.*

2.2 Consistency of LCA for fixed non-linearity

Now let $\{\mathbf{x}_v\}$ be an iid sample of \mathbf{X} . Since $E\mathbf{X} = \mathbf{0}$, we will demean the data so that $\sum_{v=1}^V \mathbf{x}_v = \mathbf{0}$, and assume such hereafter. Assume $V > T$. Let $\hat{\mathbf{\Sigma}}$ be the sample covariance matrix of $\{\mathbf{x}_v\}$. Consider its eigenvalue decomposition, $\hat{\mathbf{\Sigma}} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}'$. Then define $\hat{\mathbf{L}} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}^{-1/2}\hat{\mathbf{U}}'$. Let \mathbf{o}'_q be the q th row of an orthogonal matrix \mathbf{O} . Note that $\sum_{v=1}^V \mathbf{o}'_q \hat{\mathbf{L}}\mathbf{x}_v = 0$ and $\sum_{v=1}^V \log \phi(\mathbf{o}'_q \hat{\mathbf{L}}\mathbf{x}_v) = -\frac{V}{2}(\log 2\pi + 1)$. Let $\mathcal{O}_{Q \times T}$ be the class of $Q \times T$ semi-orthogonal

matrices, which is the Stiefel Manifold. Define the oracle estimator

$$\widehat{\mathbf{W}}_{\mathbf{S}}^{Or} = \underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} \sum_{v=1}^V \sum_{q=1}^Q \log f_q(\mathbf{o}'_q \widehat{\mathbf{L}} \mathbf{x}_v), \quad (3)$$

where $\widehat{\mathbf{W}}_{\mathbf{S}}^{Or}$ is computed when Q and the true component densities are known, so $\widehat{\mathbf{W}}_{\mathbf{S}}^{Or}$ is an oracle estimator that cannot be used in practice. In our Logis-LCA estimator, $\{f_q\}$ are replaced by the logistic density, while in Spline-LCA we also estimate $\{f_q\}$.

Observe that the problem of estimating $\mathbf{W}_{\mathbf{S}}$ is equivalent to the problem of estimating the LCs because $\hat{\mathbf{s}}_v = \widehat{\mathbf{W}}_{\mathbf{S}} \widehat{\mathbf{L}} \mathbf{x}_v$ for all v . Thus we would like a consistent estimator of $\mathbf{W}_{\mathbf{S}}$. We make an additional assumption (satisfied by the densities considered below):

Assumption 4. $\{\mathbf{S}_q\}$ has a bounded (absolutely continuous) density $\{f_q\}$, $q = 1, \dots, Q$.

Theorem 2. Suppose \mathbf{X} follows the LNGCA model in (1) with Assumptions 1-4. Additionally assume $\mathbf{E} \mathbf{X} = \mathbf{0}$ and $\mathbf{E} \mathbf{X} \mathbf{X}' = \mathbf{I}$. Let $\mathbf{W}_{\mathbf{S}}$ denote the first Q rows of \mathbf{M}^{-1} . Given an iid sample $\{\mathbf{x}_v\}$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Or} \rightarrow \mathbf{W}_{\mathbf{S}}$ almost surely on the equivalence class of signed permutations.

We can define sufficient conditions that characterize the extent to which the densities used in the estimator can mismatch the true densities while maintaining strong consistency. Let $p_q(x)$ denote a density used in the objective function (possibly mis-specified),

$$J(\mathbf{O}; \mathbf{O} \in \mathcal{O}_{Q \times T}) = \sum_{v=1}^V \sum_{q=1}^Q \log p_q(\mathbf{o}'_q \widehat{\mathbf{L}} \mathbf{x}_v), \quad (4)$$

and let $r_q(s_q)$ and $r'_q(s_q)$ denote the first and second derivatives of $\log p_q(s_q)$:

Assumption 5. For all q , (i) $p_q(s_q)$ is bounded, (ii) $\log p_q(s_q)$ is twice continuously differentiable on the support of \mathbf{S}_q , denoted \mathcal{S}_q , (iii) $\frac{\partial}{\partial o_{qt}} \mathbf{E} \log p_q(\mathbf{o}'_q \mathbf{X}) = \mathbf{E} \frac{\partial}{\partial o_{qt}} \log p_q(\mathbf{o}'_q \mathbf{X})$, (iv) $\frac{\partial}{\partial o_{qt}} \mathbf{E} \mathbf{X}_t r_q(\mathbf{o}'_q \mathbf{X}) = \mathbf{E} \frac{\partial}{\partial o_{qt}} \mathbf{X}_t r_q(\mathbf{o}'_q \mathbf{X})$, and (v) $\mathbf{E} r'_q(s_q) - \mathbf{E} s_q r_q(s_q) < 0$.

Note assumption 5(iii) and (iv) are satisfied if $\log p_q(s_q)$ and $r_q(s_q)$ have bounded derivatives on \mathcal{S}_q . Also note the differentiability assumption is for the proposed densities and does not

need to hold for the true densities. We can check the condition in Assumption 5(v) for a proposed objective function density and hypothetical source density to gain insight into a proposed estimator, which will be done in Section 3.1. Now consider compact neighborhoods of \mathbf{W}_S of the form $\mathcal{N}_\epsilon(\mathbf{W}_S) = \{\mathbf{O}_S \in \mathcal{O}_{Q \times T} : \|\mathbf{O}_S - \mathbf{W}_S\|_2 \leq \epsilon\}$, where here and throughout $\|\cdot\|_2$ denotes the Frobenius norm. Let \mathcal{P}_\pm be the class of $Q \times Q$ signed permutation matrices, \mathbf{P}_\pm . Equivalently, we could consider the set $\cup_{\mathbf{P}_\pm \in \mathcal{P}_\pm} \mathcal{N}_\epsilon(\mathbf{P}_\pm \mathbf{W}_S)$, and then assume that all maxima correspond to a matrix on the equivalence class \mathcal{P}_\pm , since the objective function behaves the same in any neighborhood $\mathcal{N}_\epsilon(\mathbf{P}_\pm \mathbf{W}_S)$. Thus it is without loss of generality that we consider a neighborhood of \mathbf{W}_S . Let $p(\mathbf{x}) = \prod_{q=1}^Q p_q(x_q)$.

Assumption 6. *There exists $\epsilon > 0$ such that $E \log p(\mathbf{O}_S \mathbf{X})$ constrained to $\mathbf{O}_S \in \mathcal{N}_\epsilon(\mathbf{W}_S)$ contains at most one maximum. (There may exist a maximum on $\mathcal{O}_{Q \times T}$ with higher $J(\mathbf{O})$.)* This assumption defines the conditions under which the well-separated criteria for consistency (e.g., van der Vaart 2000) holds. It formalizes the notion of localness that is found in theoretical treatments of the FastICA estimator (Hyvarinen, 1999; Hyvärinen and Oja, 1998; Wei, 2015). Now define the local LCA estimator:

$$\widehat{\mathbf{W}}_S^{Local} = \underset{\mathbf{O}_S \in \mathcal{N}_\epsilon(\mathbf{W}_S)}{\operatorname{argmax}} \sum_{v=1}^V \sum_{q=1}^Q \log p_q(\mathbf{o}'_q \widehat{\mathbf{L}} \mathbf{x}_v). \quad (5)$$

Then we have consistency even when the density is mis-specified.

Theorem 3. *Suppose \mathbf{X} follows the LNGCA model in (1) with Assumptions 1-6. Additionally assume $E\mathbf{X} = \mathbf{0}$ and $E\mathbf{X}\mathbf{X}' = \mathbf{I}$. Given an iid sample $\{\mathbf{x}_v\}$, $\widehat{\mathbf{W}}_S^{Local} \rightarrow \mathbf{W}_S$ almost surely on the equivalence class of signed permutations.*

Though standard in ICA, constraining the optimization space to a neighborhood of the true unmixing matrix is unsatisfactory but necessary to address the situation when the global maxima in the population objective function formulated with the wrong density is not equal to \mathbf{W}_S . For example, this occurs in ICA with certain mixtures of normals and an objective function using the logistic density (distribution “n” in Risk et al. 2014).

One strategy is to develop a preliminary estimator for LNGCA that is consistent on $\mathcal{O}_{Q \times T}$, for example, a moment-based estimator in the spirit of Virta et al. (2015), which would require additional moment assumptions on the source densities. Note in ICA kurtosis-based estimators are in general not recommended for super-Gaussian sources because they are sensitive to outliers (Hyvarinen, 1999). However, one can define an estimator and algorithm that initiates from a preliminary consistent estimator, such that the local maximum of (4) obtained from this initial estimator defines an estimator $\widehat{\mathbf{W}}_{\mathbf{S}}^{*Local}(V)$. Then under additional assumptions the sequence of estimators $\widehat{\mathbf{W}}_{\mathbf{S}}^{*Local}(V)$ will preserve consistency while possibly performing better in finite samples.

Since $\widehat{\mathbf{W}}_{\mathbf{S}}$ is not invertible, we also define an estimator of $\mathbf{M}_{\mathbf{S}}$:

$$\widehat{\mathbf{M}}_{\mathbf{S}} = \underset{\mathbf{M} \in \mathbb{R}^{T \times Q}}{\operatorname{argmin}} \sum_{v=1}^V \|\mathbf{x}_v - \mathbf{M} \hat{\mathbf{s}}_v\|_2^2, \quad (6)$$

where \mathbf{x}_v are the centered data. This is the OLS solution which here is equivalent to $\widehat{\mathbf{M}}_{\mathbf{S}} = \widehat{\mathbf{L}}^{-1} \widehat{\mathbf{W}}_{\mathbf{S}}'$. Although we assume iid observations in the construction of (3), the LNGCA model is capable of recovering many forms of dependent data, as is also the case in ICA. This will be demonstrated in simulations.

There is a natural ordering of the LCs when the component densities are not equal, which can be viewed as ordering components by the information measured by their non-Gaussian likelihood under the constraint of unit variance. Additionally, if the LCs have non-zero finite third moments, we can assume positive skewness and then the LNGCA model is fully identifiable (as in ICA, Eloyan and Ghosh 2013). We define the LCA criteria for ordering LCs for a sample $\{\mathbf{x}_v\}$:

$$\sum_{v=1}^V \log f_1(\mathbf{w}'_1 \mathbf{L} \mathbf{x}_v) > \sum_{v=1}^V \log f_2(\mathbf{w}'_2 \mathbf{L} \mathbf{x}_v) > \cdots > \sum_{v=1}^V \log f_Q(\mathbf{w}'_Q \mathbf{L} \mathbf{x}_v). \quad (7)$$

with $\sum_{v=1}^V (\mathbf{w}'_q \mathbf{L} \mathbf{x}_v)^3 > 0$ for $q = 1, \dots, Q$. For identifiability, we force the sample third

moments to be positive and order components by their likelihoods.

If we include the Gaussian noise components, then the population analogue of (7), allowing for potentially equal source densities and assuming continuous source densities, is

$$\mathbb{E} \log f_1(\mathbf{w}'_1 \mathbf{LX}) \geq \dots \geq \mathbb{E} \log f_Q(\mathbf{w}'_Q \mathbf{LX}) > \mathbb{E} \log \phi(\mathbf{w}'_{Q+1} \mathbf{LX}) = \dots = \mathbb{E} \log \phi(\mathbf{w}'_T \mathbf{LX}).$$

This conveniently characterizes the noise components as containing the least amount of information.

2.3 Sign- and permutation-invariant discrepancy measure

To assess the accuracy of our estimates and/or compare multiple estimates, we need a discrepancy measure that is invariant on the equivalence class of signed permutation matrices, and we would like a measure that can apply to matrices of differing dimensions when the estimated number of components may not equal Q . We cannot use the Amari or the minimum distance (Ilmonen et al., 2010) measures because \mathbf{M}_S is non-square. We propose a novel measure of dissimilarity that uses a modification of the Hungarian algorithm to match rows of the unmixing matrix as in Ilmonen et al. (2010) and Risk et al. (2014) but applies to non-square unmixing. We also generalize the measure to apply to matrices that may have a different number of columns, in which case the measure only compares matching columns. This measure will also be used to assess convergence in our algorithms.

Consider $\mathbf{M}_1 \in \mathbb{R}^{T \times Q}$ and $\mathbf{M}_2 \in \mathbb{R}^{T \times R}$ with $Q \leq R$. With slight abuse of notation, we now let \mathcal{P}_\pm be the class of $R \times Q$ signed permutation matrices, so that post-multiplication of \mathbf{M}_2 by $\mathbf{P}_\pm \in \mathcal{P}_\pm$ results in a subset of Q (permuted) columns of \mathbf{M}_2 for $Q < R$. Define the sign- and permutation-invariant mean-squared error:

$$PMSE(\mathbf{M}_1, \mathbf{M}_2) = \frac{1}{TQ} \underset{\mathbf{P}_\pm \in \mathcal{P}_\pm}{\operatorname{argmin}} \|\mathbf{M}_1 - \mathbf{M}_2 \mathbf{P}_\pm\|_2^2, \quad (8)$$

where the optimal \mathbf{P}_\pm is found using the modified Hungarian algorithm. In practice, we also

standardize the columns of \mathbf{M}_1 and \mathbf{M}_2 to have unit norm, and thus the measure is scale invariant. Then (8) is equivalent to finding \mathbf{P}_\pm such that the sum of the correlations between the columns of \mathbf{M}_1 and $\mathbf{M}_2\mathbf{P}_\pm$ is maximized. Another advantage of this measure is that it can be used to compare independent components directly. If \mathbf{S}_1 is a $V \times Q$ matrix in which each row is a realization of the LC in \mathbb{R}^Q , and if $\mathbf{S}_2 \in \mathbb{R}^{V \times R}$, then we define their discrepancy as $PMSE(\mathbf{S}_1, \mathbf{S}_2)$. We also let $PRMSE = \sqrt{PMSE}$, i.e., permutation-invariant root mean squared error.

2.4 A parametric LCA: Logis-LCA

First we present a parametric method called Logis-LCA in which the densities of the LCs are approximated by logistic densities. The popular Infomax algorithm can be derived as a gradient ascent algorithm for maximum likelihood ICA in which the source densities are assumed to have logistic densities. Infomax is popular in fMRI analysis, where it appears to outperform FastICA and JADE (Correa et al., 2007; Calhoun and Adali, 2006). Under the constraint of zero mean and unit variance, the logistic density has the form

$$f(x) = \frac{\exp(-x/\frac{\sqrt{3}}{\pi})}{\frac{\sqrt{3}}{\pi} \left\{ 1 + \exp(-x/\frac{\sqrt{3}}{\pi}) \right\}^2}. \quad (9)$$

We define our estimator for some $Q^* \leq T$ such that Q^* may or may not equal the true number of LCs, Q . If $Q^* = Q$ and the true source densities are logistic, then it follows from Theorem 2 that the estimator is consistent. We will show robustness to misspecification of Q via simulations in Section 3.2. Applying (9) to (3) and the centered data $\{\mathbf{x}_v\}$, the Logis-LCA estimator of $\mathbf{W}_\mathbf{S}$ can be written as

$$\widehat{\mathbf{W}}_\mathbf{S}^{Logis} = \underset{\mathbf{O}_\mathbf{S} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} - \sum_{v=1}^V \sum_{q=1}^{Q^*} \log \left\{ 1 + \exp \left(-\mathbf{o}'_q \widehat{\mathbf{L}} \mathbf{x}_v \frac{\pi}{\sqrt{3}} \right) \right\}. \quad (10)$$

We maximize (10) using a modification of the symmetric fixed-point ICA algorithm (Hy-

varinen, 1999). An issue is that ICA implementations require the estimator to be a square matrix, and this property is used in orthogonalizing intermediate estimates and in assessing convergence. We orthogonalize intermediate estimates of the rows of $\hat{\mathbf{W}}_{\mathbf{S}}^{Logis}$ by calculating the SVD and setting the singular values equal to one. Additionally, we assess convergence using (8). See Web Supplement Section C for details.

2.5 A semi-parametric LCA: Spline-LCA

In this section, we use the flexible family of tilted Gaussian densities to model the LCs. The proposed model is equivalent to ProDenICA (Hastie and Tibshirani, 2003) when $Q = T$. For $Q < T$, it can be shown that the likelihood extends the semiparametric likelihood in Blanchard et al. (2006) to include an independence model for the LCs (see Proposition 3 of the Web Supplement Section B.1). The independence assumption is necessary for physically and biologically useful interpretations. We chose tilted Gaussian densities with cubic B-splines because ProDenICA generally outperformed parametric and kernel ICA methods (Hastie et al., 2009; Risk et al., 2014) and its algorithmic complexity is $O(V)$, which enables its application to large datasets such as fMRI.

Suppose the LCs have tilted Gaussian distributions of the form $\phi(u)e^{g(u)}$, where $g(u)$ is a twice-differentiable function. Define the log-likelihood for some $\mathbf{O} \in \mathcal{O}_{T \times T}$:

$$\ell(\mathbf{O}, g_1, \dots, g_{Q^*} \mid \hat{\mathbf{L}}, Q^*, \{\mathbf{x}_v\}) = \sum_{v=1}^V \left[\sum_{q=1}^{Q^*} \left\{ \log \phi(\mathbf{o}'_q \hat{\mathbf{L}} \mathbf{x}_v) + g_q(\mathbf{o}'_q \hat{\mathbf{L}} \mathbf{x}_v) \right\} + \sum_{k=1}^{T-Q^*} \log \phi(\mathbf{o}'_{k+Q^*} \hat{\mathbf{L}} \mathbf{x}_v) \right]$$

This log-likelihood does not have an upper bound. We define a penalized log-likelihood that includes a roughness penalty and an additional term to ensure the solution is a density:

$$\begin{aligned} \ell_{pen}(\mathbf{O}, g_1, \dots, g_{Q^*} \mid \hat{\mathbf{L}}, Q^*, \{\mathbf{x}_v\}) = & - \sum_{q=1}^{Q^*} \left\{ \lambda_q \int \{g''_q(u)\}^2 du + \int \phi(u) e^{g_q(u)} du \right\} \quad (11) \\ & + \frac{1}{V} \sum_{v=1}^V \sum_{q=1}^{Q^*} \left\{ \log \phi(\mathbf{o}'_q \hat{\mathbf{L}} \mathbf{x}_v) + g_q(\mathbf{o}'_q \hat{\mathbf{L}} \mathbf{x}_v) \right\}, \end{aligned}$$

where we have dropped the noise components since they are constant for all \mathbf{O} . Then we have the following:

Proposition. *Let G be the class of all cubic splines $g: \mathbb{R} \rightarrow \mathbb{R}$. Consider the argmax of (11) for $g_q \in G$. Then (i) $\int \phi(u) e^{g_q(u)} du = 1$ and (ii) $\int u \phi(u) e^{g_q(u)} du = 0$ for each q .*

We adapt the ProDenICA algorithm of Hastie and Tibshirani (2003) to LCA, in which we alternate between estimating \mathbf{W}_S for fixed $\{\hat{f}_q\}$, $q = 1, \dots, Q^*$, via the fixed point algorithm and estimating $\{f_q\}$ for fixed $\widehat{\mathbf{W}}_S$ using the ‘‘Poisson trick’’. Our account largely follows the description in Hastie et al. (2009) but for semi-orthogonal (rather than orthogonal) matrices.

Suppose \mathbf{W}_S is given and define $s_{vq} = \mathbf{w}'_q \mathbf{z}_v$, where $\mathbf{z}_v = \widehat{\mathbf{L}} \mathbf{x}_v$. Let u_1^*, \dots, u_{L+1}^* define a discretization, $[u_1^*, u_2^*), [u_2^*, u_3^*), \dots, [u_L^*, u_{L+1}^*)$, of the support of the tilt function of the non-Gaussian densities such that $\Delta = u_\ell^* - u_{\ell-1}^*$ for all $\ell = 2, \dots, L+1$. It suffices to take $u_1^* = \min(s_{11}, \dots, s_{nd}) - 0.1\hat{\sigma}_z$ and $u_{L+1}^* = \max(s_{11}, \dots, s_{nd}) + 0.1\hat{\sigma}_z$, where $\hat{\sigma}_z$ denotes the sample standard deviation, which here is equal to one. Next, let $u_\ell = \frac{1}{2}(u_\ell^* + u_{\ell+1}^*)$. For each $q \in \{1, \dots, Q^*\}$ and $\ell \in \{1, \dots, L\}$, define $y_{\ell q} = \sum_{v=1}^V \mathbb{1}\{s_{vq} \in [u_\ell^*, u_{\ell+1}^*)\}$.

We approximate (11) by discretizing the first integral and estimating the sum over V as a weighted sum over L . Restricting our attention to a single q and dividing by Δ ,

$$-\beta_q \int \{g_q''(u)\}^2 du + \sum_{\ell=1}^L \left[\frac{y_{\ell q}}{V\Delta} \{g_q(u_\ell) + \log \phi(u_\ell)\} - \phi(u_\ell) e^{g_q(u_\ell)} \right] \quad (12)$$

for some penalty β_q . This is proportional to a Poisson generalized additive model (GAM), where $\frac{y_{\ell q}}{V\Delta}$ is the response and the expected response is equal to $\phi(u_\ell) e^{g_q(u_\ell)}$. This can be fit using the `gam` package in R (Hastie, 2013) where β_q is chosen to result in a user-specified number of effective degrees of freedom. We find that $df = 8$ and $L = 100$ produce fast and accurate density estimates in simulations for a variety of densities when the sample size V is equal to 1,000. This method also easily scales to tens of thousands of observations.

The algorithm to estimate both \mathbf{W}_S and $\{f_q\}$ is summarized below. Note that step 3 requires the first and second derivatives of the log densities of the LCs, which makes the use

of B-splines convenient.

Algorithm 1: The Spline-LCA algorithm.

Inputs : The whitened $V \times T$ data matrix \mathbf{Z} ; initial $\mathbf{W}_{\mathbf{S}}^0$; tolerance ϵ .

Result: Estimates of the latent components, $\hat{\mathbf{S}}$, and their densities, $\{\hat{f}_q\}$.

1. Let $n = 0$ and define $\mathbf{S}^{(n)} = \mathbf{Z}\mathbf{W}_{\mathbf{S}}^{(n) \prime}$.
 2. Estimate $\{f_q^{(n+1)}\}$.
 3. Using (3), update $\mathbf{W}_{\mathbf{S}}^{(n+1)}$ given $f_1^{(n+1)}, \dots, f_Q^{(n+1)}$ and $\mathbf{S}^{(n)}$ with one-step of the fixed-point algorithm (see Algorithm 1 in Web Supplement Section C).
 4. Let $\mathbf{S}^{(n+1)} = \mathbf{Z}\mathbf{W}_{\mathbf{S}}^{(n+1) \prime}$.
 5. If $PMSE(\mathbf{W}_{\mathbf{S}}^{(n+1) \prime}, \mathbf{W}_{\mathbf{S}}^{(n) \prime}) < \epsilon$, stop, else increment n and repeat (2)-(4).
-

3 Simulations

3.1 Simulations: Distributional & Noise-rank Assumptions

In this section, we simulate the LNGCA model [given by (1) with $\mathbf{M}_{\mathbf{S}} \in \mathbb{R}^{T \times Q}$] and the noisy ICA model [again given by (1) with $\mathbf{M}_{\mathbf{S}} \in \mathbb{R}^{T \times Q}$ but now with $\mathbf{M}_{\mathbf{N}}\mathbf{N} \sim N(0, \sigma^2 \mathbf{I}_T)$] under a variety of source distributions in which the components are iid as well as a scenario in which the signals are sparse images. We compare (i) deflationary FastICA with the ‘tanh’ nonlinearity (D-FastICA), where the deflation option estimates components one-by-one such that the algorithm is considered a projection pursuit method (Hyvärinen and Oja, 2000); (ii) two-class IFA with isotropic noise (IFA); (iii) PCA followed by Infomax (PCA+Infomax); (iv) PCA followed by ProDenICA (PCA+ProDenICA); (v) Logis-LCA; and (vi) Spline-LCA. We evaluate the robustness of these methods with respect to assumptions on the rank of the noise components, distribution of the latent components, and the signal-to-noise ratio (SNR). We define the SNR as the ratio of the total variance from the mixed non-Gaussian components to the total variance from the noise components. Formally, consider the non-zero eigenvalues $\lambda_1, \dots, \lambda_Q$ from the covariance matrix of $\mathbf{M}_{\mathbf{S}}\mathbf{S}$. For LCA, let $\lambda_{\epsilon_1}, \dots, \lambda_{\epsilon_{T-Q}}$ denote the

eigenvalues from the EVD of the covariance matrix of $\mathbf{M}_\mathbf{N}\mathbf{N}$. Then, $SNR = \frac{\sum_{q=1}^Q \lambda_q}{\sum_{k=1}^{T-Q} \lambda_{\epsilon_k}}$. For the noisy ICA model, we have T non-zero eigenvalues in the denominator sum.

We fit D-FastICA using the fastICA R package (Marchini et al., 2010). We fit PCA+Infomax using our own implementation of the Infomax algorithm. We fit PCA+ProDenICA using the ProDenICA function from the R package of that name (Hastie and Tibshirani, 2010). Note that these methods can provide an estimate of \mathbf{S} but not the mixing matrix, which we estimated using (6). We fit the IFA model with two-component mixtures of normals using our own implementation, and the ICs were estimated by their conditional means (see equation (81) in Attias 1999). Details are in the Web Supplement Section D.

3.1.1 Simulation Design

Data were generated with $T = 5$ and $Q = 2$ according to a $2^2 \times 6$ full factorial design. The three factors were

- i) **The model:** the levels were (a) the LNGCA model with rank- $(T - Q)$ noise and (b) the noisy ICA model with rank- T noise. In both models the signal was $\mathbf{M}_\mathbf{S}\mathbf{S}$ where $\mathbf{M}_\mathbf{S}$ is $T \times Q$ with $Q < T$.
- ii) **The signal to noise (SNR) ratio:** the levels were (a) high where the ratio of the variance from the signal components to the variance from the noise components was 5:1 and (b) low where that ratio was 1:5.
- iii) **Signal distribution:** the levels were (a) logistic, (b) t, (c) Gumbel, (d) sub-Gaussian mixture of normals, (e) super-Gaussian mixture of normals, (f) with values determined by a sparse image, as described below. The two signal components were each iid and had the same distributions in cases (a)–(e) but different distributions in the sparse signal case.

Since we generated $Q = 2$ signal components for all simulations, there were $T - Q = 3$ and $T = 5$ noise components for the LNGCA model and noisy ICA model, respectively.

Observations in the noise components were iid isotropic normal except for the sparse image scenario, in which we used the R-package neuRosim (Welvaert et al., 2011) to generate three-dimensional Gaussian random fields with full width at half maximum (FWHM) equal to 6 for each noise component.

The signal components had scale parameter equal to $\sqrt{3}/\pi$ for the logistic, three degrees of freedom for the t, and scale parameter equal to $\sqrt{6}/\pi$ for the Gumbel. For the super-Gaussian mixture of normals, we simulated a two-class model with the first centered at 0 with variance 4/9 with probability 0.95 and the second centered at 5 with unit variance (excess kurtosis ≈ 9), which is motivated by a brain network with 5% of voxels activated. For the sub-Gaussian mixture of normals, we used the two-class model with the first centered at -1.7 with unit variance and probability 0.75 and the second centered at 1.7 with unit variance and probability equal to 0.25, which is equivalent to distribution ‘l’ from Hastie and Tibshirani (2003) (excess kurtosis ≈ -0.3). For the sparse image, we used neuRosim to generate a three-dimensional image in which all voxels were iid normal with variance equal to 0.0001 except, in the first component, a sphere of radius two in which the center was located at $(5, 5, 5)$ with voxel-value equal to one and the exponential decay rate set to 0.5. The second sparse image component was similar except the feature was a cube centered at $(7, 7, 7)$ with width equal to two and exponential decay rate equal to one.

We conducted 112 simulations (chosen because we used a cluster with 56 processors) with $V = 1,000$ observations in which \mathbf{M}_S and \mathbf{M}_N were randomly generated to have condition number between one and ten for each combination of factors. Since neither the set of orthogonal matrices (PCA+ICA methods) nor semi-orthogonal matrices (LCA methods) is convex, we approximated the argmax by initializing D-FastICA, PCA+Infomax, Logis-LCA, PCA+ProDenICA, and Spline-LCA from twenty random matrices and selecting the estimate associated with the largest objective function value. For Logis-LCA and Spline-LCA, ten of these twenty initializations were from random matrices constrained to the principal subspace. Let $\hat{\mathbf{U}}_{1:Q}$ denote the first Q rows from $\hat{\mathbf{U}}$ in the decomposition $\hat{\Sigma} = \hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}'$. Then constraining

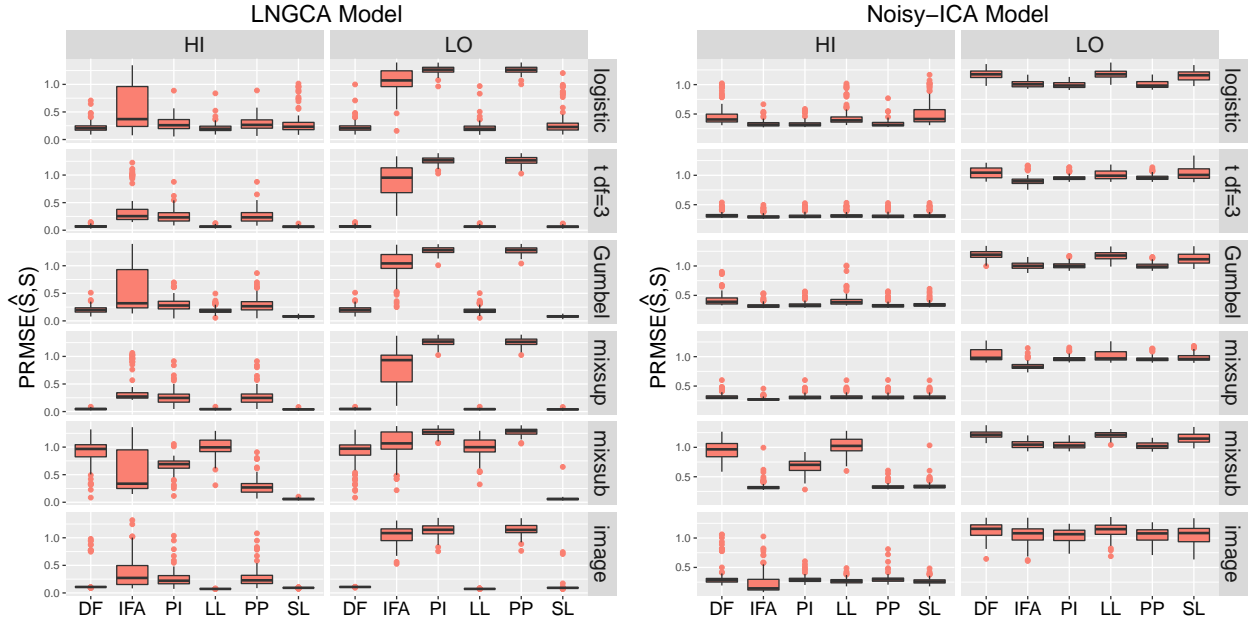
the initial matrix, \mathbf{W}_S^0 , to the principal subspace is equivalent to $\mathbf{W}_S^0 = \widehat{\mathbf{U}}_{1:Q} \mathbf{O}$ with $\mathbf{O} \in \mathcal{O}_{Q \times Q}$. For IFA, one must specify initial values for the unmixing matrix, the variance of the isotropic noise, and the parameters of the Gaussian mixtures, and here we had four strategies to find the argmax including initialization from the true \mathbf{W}_S . See Web Supplement Section D for additional details.

3.1.2 Results

When the LNGCA model was true and there was a high SNR, all methods except IFA generally produced accurate estimates of \mathbf{S} for the logistic, t, Gumbel, super-Gaussian mixture of normals, and sparse images, but only Spline-LCA was accurate for the sub-Gaussian mixture of normals, and the performance of IFA was more variable than other methods for all distributions (Figure 1). PCA+Infomax performed poorly for the sub-Gaussian mixtures because the logistic distribution generally fails for sub-Gaussian distributions (see Lee et al. 1999). Boxplots examining the accuracy of $\widehat{\mathbf{M}}_S$ showed patterns similar to those found in Figure 1 and consequently are not presented.

When the LNGCA model was true and there was a low SNR, Spline-LCA generally outperformed other methods, while IFA, PCA+Infomax, and PCA+ProDenICA failed to recover the LCs for all distributions, and D-FastICA and Logis-LCA recovered all distributions except for the sub-Gaussian mixture of normals. Thus for low SNR, PCA+ICA methods discarded the non-Gaussian signal. This was true even when the correct source density was modeled, as in PCA+Infomax and the logistic density simulation. Spline-LCA was the method most robust to distributional assumptions and was the only method that recovered the sub-Gaussian mixture. We numerically evaluated the condition in Assumption 5(v) for Logis-LCA and Spline LCA and the densities (a)–(e) (with densities centered and scaled to have unit variance and the Spline-LCA densities estimated from a sample from the true densities). For Logis-LCA, all values were negative except for the sub-Gaussian mixture of normals, whereas all values were negative for Spline-LCA. Thus the results for $V = 1000$ are in general agreement with Theorem 3.

Figure 1: Boxplots of permutation-invariant root mean squared error ($PRMSE$) for estimated columns of \mathbf{S} where the rank of the noise was $T - Q$ (LNGCA Model) or T (noisy ICA model) in high SNR (‘HI’) and low SNR (‘LO’) scenarios for various latent distributions. ‘DF’ = D-FastICA; ‘IFA’ = independent factor analysis; ‘PI’ = PCA+Infomax; ‘LL’ = Logis-LCA; ‘PP’ = PCA+ProDenICA; ‘SL’ = Spline-LCA.



When the noisy ICA model was true and there was a high SNR, all methods generally produced reasonably accurate estimates for the logistic, t, Gumbel, super-Gaussian, and sparse image. IFA and Spline-LCA were the only methods that recovered ICs with sub-Gaussian distributions. When the noisy ICA model was true and there was a low SNR, all methods performed poorly, although IFA, PCA+Infomax, and PCA+ProDenICA outperformed LCA algorithms for some distributions. Note that in PCA+ICA methods, PCA decomposes the data into a subspace with the signal and some noise, and a subspace with noise only, see Web Supplement B.2. When the SNR is high, this is an effective strategy because the amount of error that corrupts the ICs is negligible. When there is a low SNR, the components estimated with ICA are highly contaminated with noise.

Overall, LCA methods were robust to the SNR for rank- $(T - Q)$ noise, and performed well in the high SNR scenario for rank- T noise. Additionally, Spline-LCA was most robust to distributional assumptions. In contrast, IFA, PCA+Infomax, and PCA+ProDenICA

performed poorly in the low SNR scenario for both the rank- $(T - Q)$ and rank- T noise.

3.2 Simulations: Spatio-temporal Networks

Next, we examine the ability of D-FastICA, PCA+Infomax, Logis-LCA, and Spline-LCA to recover simulated networks whose loadings vary deterministically with time in the presence of spatially and temporally correlated noise, so that the simulations resemble the structure found in task-based fMRI. We also examine the effect of using $Q^* \neq Q$ on network recovery. We did not include IFA in these simulations because it was difficult to estimate when T was relatively large (e.g., $T = 50$). Additionally, IFA, PCA+Infomax, and PCA+ProDenICA produced similar results for super-Gaussian distributions in the previous simulations.

3.2.1 Simulation Design

We simulated three networks mixed across fifty time units. The networks were 33×33 images where “active” pixels were in the shape of a “1”, “2 2”, or “3 3 3” with values between 0.5 and 1 and “inactive” pixels were mean zero iid normal with variance equal to 0.0001 (see Figure 2). Let \mathbf{m}_q denote the q th column of \mathbf{M}_S . To simulate the temporal activation patterns of brain networks, we used neuRosim to convolve the canonical hemodynamic response function (HRF) with a block-design with a pair of onsets at $\{1, 20.6\}$, $\{10.8, 40.2\}$, and $\{10.8, 30.4\}$ for \mathbf{m}_1 , \mathbf{m}_2 , and \mathbf{m}_3 , respectively, and duration equal to 5 time units (Welvaert et al., 2011).

In the LNGCA scenario, noise components were generated as forty-seven independent 33×33 Gaussian random fields with FWHM=6. Temporal correlation was introduced via the mixing matrix, in which each column of \mathbf{M}_N corresponded to an AR(1) process simulated for fifty time units with AR coefficient equal to 0.47 and unit variance, where the AR coefficient was chosen based on a preliminary analysis of the fMRI data analyzed in Section 5. Additionally, noise components were scaled such that the SNR was 0.4, which approximately equals the SNR estimated in Section 5. In the noisy ICA scenario, a 33×33 Gaussian random field with FWHM=6 was simulated for $t = 1$. Then noise components were defined

recursively for $t = 2, \dots, 50$ to be equal to 0.47 times the noise at time $t - 1$ plus a realization from an independent Gaussian random field with FWHM=6.

We conducted 111 simulations with $Q^* = 2, 3$ or 4 (with fixed $Q = 3$) and initialized all algorithms from twenty random mixing matrices for each simulation and each Q^* . For Logis-LCA and Spline-LCA, ten of the twenty initializations were from random matrices in the principal subspace, as in Section 3.1.1.

3.2.2 Results

By inspecting the images and loadings associated with the median $PRMSE(\hat{\mathbf{S}}, \mathbf{S})$ for each method in the LNGCA scenario, we see that D-FastICA recovers a spurious component when $Q^* = 3$; PCA+Infomax and PCA+ProDenICA generally fail to unmix features; and Logis-LCA and Spline-LCA are highly accurate (Figure 2). Boxplots for D-FastICA indicate higher $PRMSE$ than Logis-LCA or Spline-LCA for $Q^* = 3$ and $Q^* = 4$ (Web Supplement Figure S.1), and the third component was typically not recovered for $Q^* = 3$ (Figure 2). This suggests a deflationary approach to estimating LNGCA may be inaccurate. In contrast, Logis-LCA and Spline-LCA recovered the components in all simulations (Web Supplement Figure S.1). It is notable that estimates from PCA+Infomax, PCA+ProDenICA, and D-FastICA were sensitive to the choice of Q^* , whereas as Logis-LCA and Spline-LCA were robust (Figure 2, Web Supplement Figure S.1).

For the noisy-ICA scenario, the features recovered by Logis-LCA most closely resembled the truth (Figure 3) and Logis-LCA generally outperformed other methods (Web Supplement Figure S.1). Features from component two were again faintly visible in component three for $Q^* = 2$ in both PCA+Infomax and PCA+ProDenICA, again indicating inadequate unmixing of the networks. As seen in the LNGCA scenario, D-FastICA recovered a spurious component for $Q^* = 3$, but accurately estimated component three in the majority of simulations when $Q^* = 4$. Spline-LCA typically failed to recover component one for $Q^* = 3$, although it was quite accurate for components two and three. Spatial correlations in the noise can result in spurious disk-like features, which were estimated in D-FastICA for both scenarios and by

Figure 2: Network recovery from the LNGCA scenario with $Q = 3$ for $Q^* = 2, 3$, or 4. Images depict LCs and time-series depict the loadings ($\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_{Q^*}$) corresponding to the median $PRMSE(\hat{\mathbf{S}}, \mathbf{S})$. In the last column, “Truth” corresponds to an arbitrary noise component whereas the algorithms attempted to estimate a fourth LC.

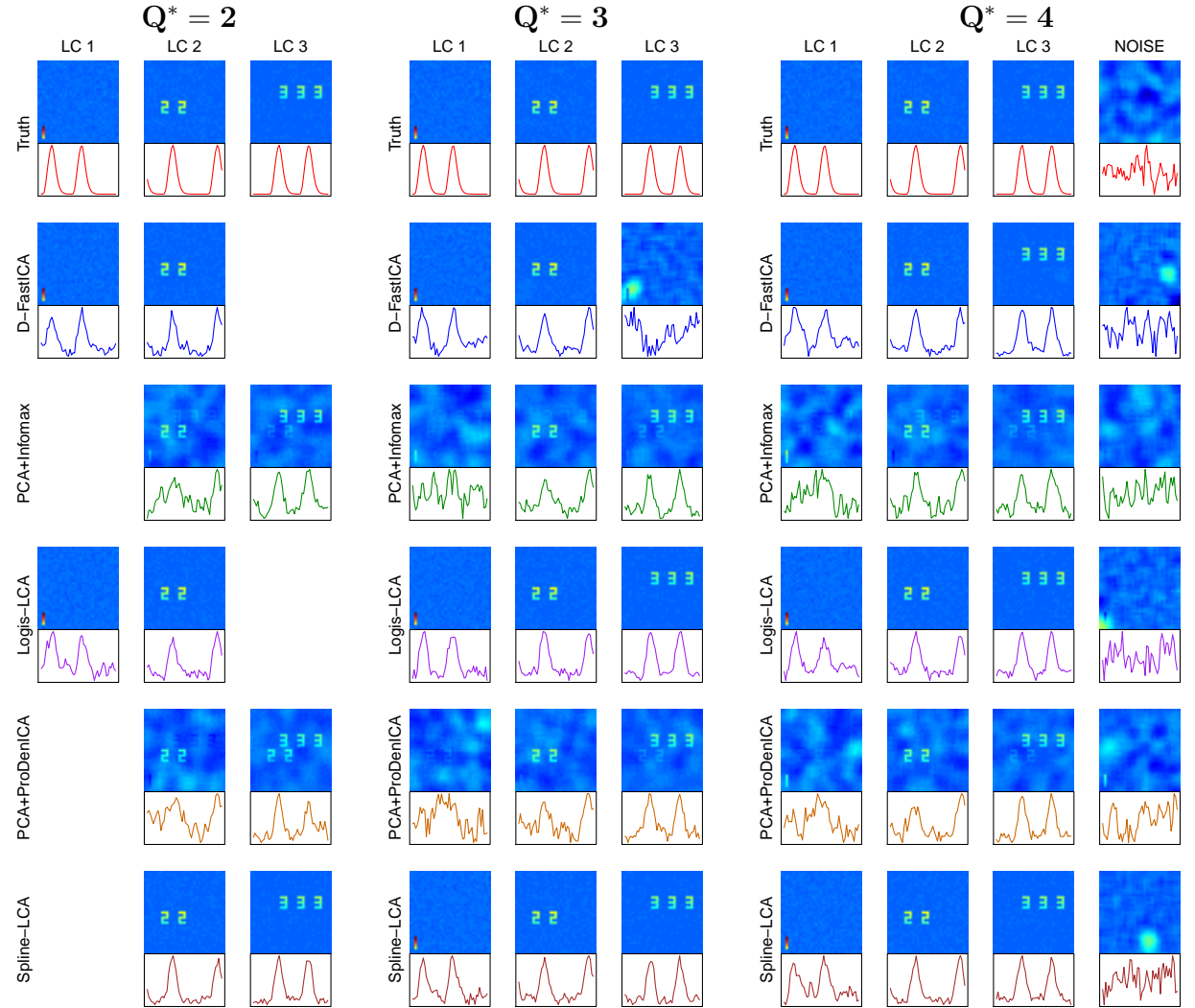
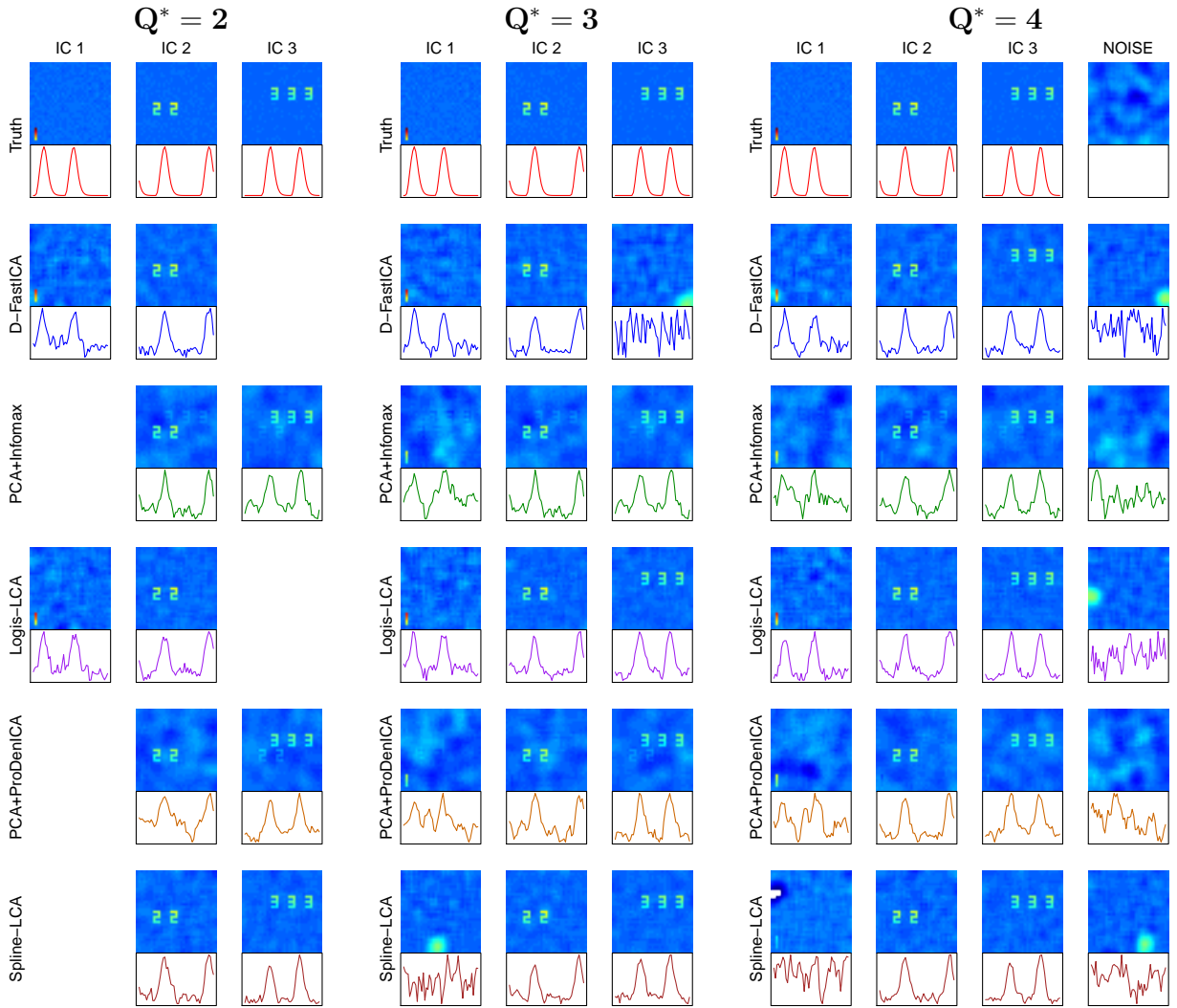


Figure 3: Network recovery from the noisy-ICA scenario with $Q = 3$ for $Q^* = 2, 3$, or 4.

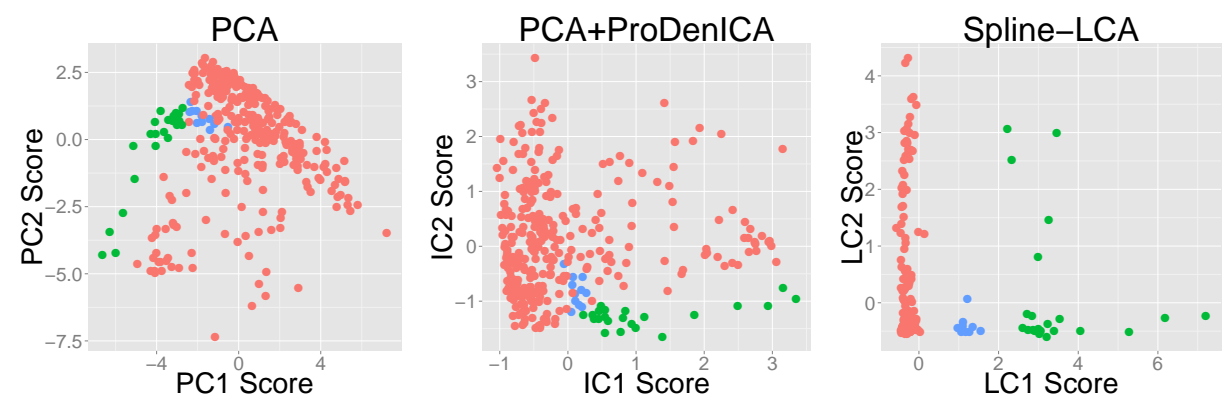


Spline-LCA in the noisy-ICA scenario. For the simulation associated with the median error, an accurate estimate of component one was associated with a local maxima in Spline-LCA, but the spurious component had a higher likelihood. The true component was recovered in some simulations (Web Supplement Figure S.1).

4 Data Visualization and Dimension Reduction

We used Logis-LCA and Spline-LCA for data visualization and dimension reduction in multivariate data comprising measurements from independent leaf samples (Silva et al.,

Figure 4: Data visualization and dimension reduction for the leaf dataset. The original dataset comprises 14 variables, many of which are highly correlated. The green dots correspond to *Podocarpus sp.* and *Pseudosasa japonica*; the blue dots to *Neurium oleander*; the red dots to all other species.



2013). Fourteen variables were generated from eight to sixteen images of leaves from each of thirty species (Web Supplement Figure S.2). Many of the covariates are highly correlated (Web Supplement Figure S.3). We plotted the first two PCs, ICs from PCA+Infomax and PCA+ProDenICA, and LCs from Logis-LCA and Spline-LCA. Two-dimensional PCA does not reveal clear features (Figure 4). Since we are examining two dimensions, the effect of ICA is apparent as a rotation of the X- and Y-axes. Rotating the axes does not reveal any additional insight (Figure 4, Web Supplement Figure S.4). In contrast, Spline-LCA clearly reveals three clusters, where the green dots correspond to two plant species that have very thin leaves (species 31 and 34 in Web Supplement Figure S.2), the blue category corresponds to a species with leaves that are thinner than most species but less than those comprising the green dots (species 8), and the red category corresponds to all other species. Logis-LCA also reveals structure (Web Supplement Figure S.4), although the separation is less than in Spline-LCA.

PCA+ICA methods were sensitive to the number of components estimated whereas the highest ranked components were very similar for different Q^* in the LCA methods. In PCA+Infomax and PCA+ProDenICA, the first two (matched) ICs for $Q^* = 5$ differed from the ICs estimated using two components, demonstrating the sensitivity of PCA+ICA

methods to the number of principal components (Web Supplement Figures S.4 and S.5). In contrast, the two highest-ranked LCs extracted from Logis-LCA and Spline-LCA when five components were estimated were very similar to the LCs estimated using two components.

5 Application to fMRI

We applied Spline-LCA to eleven subjects from the Social Cognition / Theory of Mind experiment of the WU-Minn Human Connectome Project (HCP); additional information is in Web Supplement Section G. Single-subject ICA is an important technique for identifying artifacts in fMRI due to physiology (heart rate, breathing), subject-specific motion, and/or scanner instabilities, and accounting for these artifacts can decrease false positives and increase sensitivity (Pruim et al., 2015). Theory of mind (ToM) refers to the ability of humans to infer the mental states of others. The experiment involved a mentalizing task in which shapes interacted in a goal-directed manner (e.g., a big triangle leading a little triangle out of a box) or according to some complex intentionality (e.g., a shape scaring another shape), and in which the random task involved shapes moving in random directions; for details see Barch et al. (2013). We used the minimally preprocessed data from the *fMRIVolume* pipeline (Glasser et al., 2013) from the first ToM session. The preprocessing pipeline includes rigid-body motion correction of all volumes to a subject’s reference image followed by MNI non-linear registration. Note that even if perfect alignment were possible, motion artifacts may still be present due to spin history effects and/or spatial variation in the coil sensitivities (Friston et al., 1996). The *fMRIVolume* pipeline does not include any spatial smoothing. Three-dimensional volume data were vectorized and non-brain tissue excluded using the mask provided from the HCP. This resulted in a $230,459 \times 272$ data matrix. Each voxel was treated as a replicate with $v = 1, \dots, V$ for $V = 230,459$, which is analogous to ‘spatial’ ICA of fMRI (Calhoun and Adali, 2006). We mean centered and variance normalized each voxel’s time course prior to conducting LCA, as suggested for ICA (Beckmann and

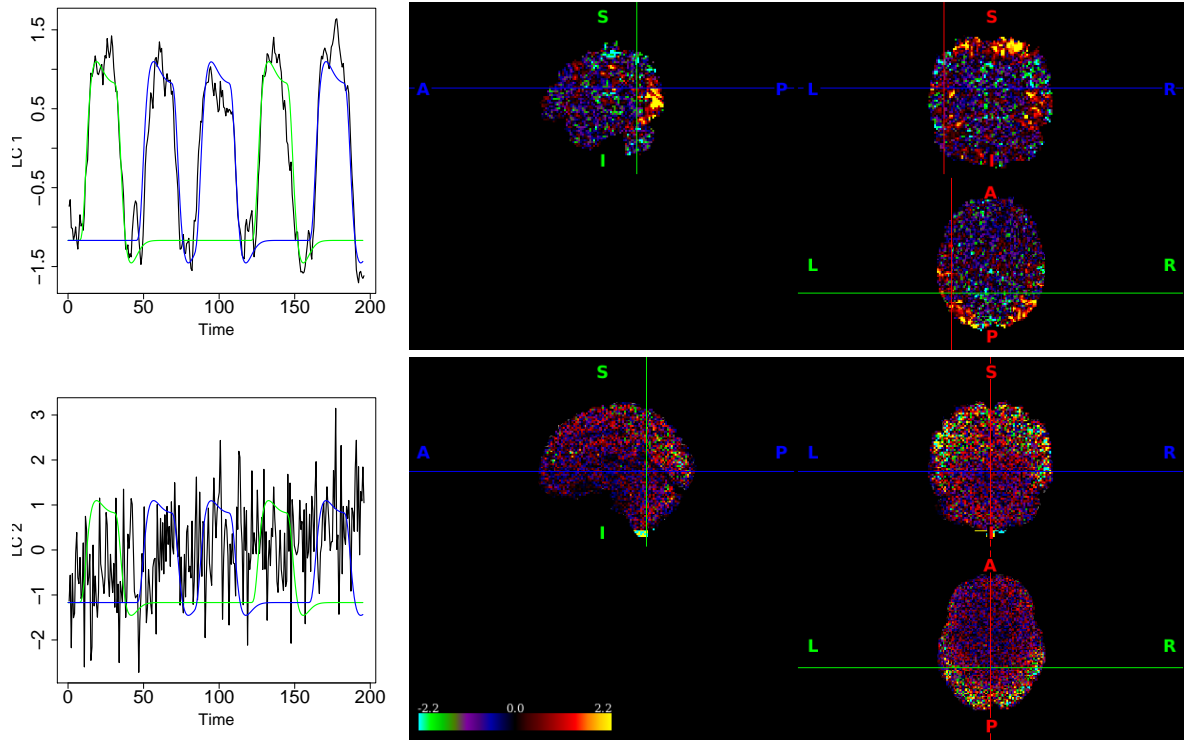
Smith, 2004).

The application of ICA to fMRI usually assumes that voxels are iid (an exception for temporal ICA is Lee et al. 2011). This assumption is often not made explicitly because ICA is usually derived from the perspective of maximizing non-Gaussianity. Since the objective function maximizing non-Gaussianity can also be derived from ML theory where the non-linear function is equivalent to the log likelihood (e.g., Hyvärinen and Oja 2000), summation of the non-linear function over voxels (e.g., Equation 12 in Beckmann and Smith 2004) is mathematically equivalent to assuming the voxels are independent. Despite the violation of model assumptions, ICA recovers simulated brain networks and their loadings (Beckmann and Smith, 2004) and has proven useful in constructing models of functional connectivity that are consistent across subjects and image acquisition centers (Biswal et al., 2010).

We used the ICA software MELODIC (FSL) to determine the number of components that would be used in an analogous ICA of this dataset, which chose thirty components for subject 103414. Thirty components were then estimated for all other subjects. We initiated the algorithm from fifty-six matrices: from the first thirty columns of the FOBI (fourth-order blind identification) estimate of all components (an analytic solution that is fast to compute); twenty-seven semi-orthogonal matrices randomly generated in the principal subspace; and twenty-eight random semi-orthogonal matrices. We selected the estimate corresponding to the largest log likelihood as our estimate of the true argmax. The best estimate corresponded to one of the random matrices from the principal subspace for all subjects. Depending on initialization, the algorithm took between ten minutes and 3.75 hours on a 2666 MHz processor, where 3.75 hours represented initializations that reached the maximum number of iterations, which we conservatively chose to be equal to 300. Initiating the algorithm from fifty-six matrices resulted in multiple initializations converging to the same estimate of the argmax (Web Supplement Figure S.6). We also completed an analogous PCA+ProDenICA with thirty components using the R package ProDenICA (Hastie and Tibshirani, 2010), where one initialization was from the FOBI solution from the PCA-reduced dataset and

fifty-five initializations were from random orthogonal matrices. In PCA+ProDenICA, the best initialization was always from one of the fifty-five random orthogonal matrices. These results suggest that the FOBI solution was not “close enough” to the semiparametric solution to aid detection of the maximum in either Spline-LCA or PCA+ProDenICA.

Figure 5: Selected brain networks estimated from the HCP ToM data using Spline-LCA. The first row depicts a task-activated component that was highly correlated with the mentalizing (green) and random (blue) tasks (MNI coordinates: -50,-56,18); a similar component was found using PCA+ProDenICA (not depicted). The second row appears to be an artifact not found by PCA+ProDenICA (MNI: 0,-50,0). Additional artifacts found in Spline-LCA but not PCA+ProDenICA are in Web Supplement Figures S.7 and S.8.



In all subjects, a component highly correlated with the task was found in both Spline-LCA and PCA+ProDenICA, but a number of components were only detected in Spline-LCA. We examined the correlation between the loadings (columns of $\widehat{\mathbf{M}}_{\mathbf{S}}$) and the mentalizing and random tasks. The mentalizing and random task covariates were generated by convolving each task’s onsets and durations with the canonical HRF in SPM8 (Ashburner et al., 2004). In all subjects, the first component, i.e., the one with the highest likelihood, was highly

correlated with the mentalizing and random tasks (e.g., Figure 5). The most positive values of this component are located in the gray matter, which indicates brain activity. Areas of Brodmann Area 19 in the visual cortex appear activated. This is an area associated with shape recognition and attention, and thus it makes sense that the movies based on moving shapes engaged this area. The same component was found using PCA+ProDenICA. For all subjects, the correlation of the matched PCA+ProDenICA component with the first Spline-LCA component was at least 0.98. Note however that this component does not distinguish between the mentalizing and random tasks. Moreover, the temporal parietal junction (TPJ) is an area often found in ToM studies (Castelli et al., 2000) (the crosshairs in Figure 5 are located near the TPJ) but is not activated in this component, suggesting there exists additional signal in other components.

A number of components containing structure were only identified in Spline-LCA, and these components may correspond to motion and scanner artifacts. Overall, a median of eight components were found in Spline-LCA but not PCA+ProDenICA, as defined by the matched component having a correlation less than 0.5 (minimum number: 4; maximum number: 9). In a detailed examination of subject 103414, component two in Spline-LCA was not correlated with any of the components in PCA+ProDenICA (max correlation among all ICs = 0.01). This component appears to correspond to an artifact due to motion and possibly other sources of noise. Voxels were highly activated in the brainstem and the component's time course was correlated with three of the motion parameters from the rigid-body alignment ($r = 0.32$, 0.32 , and 0.42 for the x-transformation, x-rotation, and z-rotation parameters, respectively). Additionally, there was a positive correlation with time ($r = 0.44$), which could be related to scanner drift. In another example, LC 25 exhibited activation at the edges of the brain, which is typical of motion artifacts (Salimi-Khorshidi et al., 2014), while the matched component from PCA+ProDenICA did not show this pattern. This provides a clear example in which Spline-LCA identified a motion component that was not found in PCA+ProDenICA (Web Supplement Figure S.7). In another subject (100307), two

LC components that were not found in PCA+ProDenICA exhibited alternating patterns of positive and negative activation (Web Supplement Figure S.8), which may be from artifacts due to scanner acquisition/reconstruction (e.g., Figure 6 in Salimi-Khorshidi et al. 2014). Similar artifacts were also found by Spline-LCA but not PCA+ProDenICA in subject 103414 (LC 7, LC 15). Removing artifacts from fMRI detected using PCA+ICA is a popular tool that can increase detection in subsequent mixed-modeling of voxel activation (Pruim et al., 2015). Our results suggests that LCA may improve artifact detection.

6 Discussion

We propose a novel model, LNGCA, and estimation framework, LCA, for non-Gaussian latent components in the presence of Gaussian noise that have many applications including dimension reduction, signal processing, artifact detection, and network estimation. We presented two applications: data visualization and dimension reduction, and identifying brain networks and artifacts from neuroimagery. Our first simulation study indicates that our methods perform well when the LNGCA model is true, even for low SNR, and our methods provide a reasonable approximation to noisy ICA when the SNR is high. Additionally, we found that the popular approach to approximating the noisy ICA model, PCA+ICA, does not approximate the LNGCA model under low SNR, and also only approximates the noisy ICA model under high SNR. In the second simulation study, we examined performance when data contained spatiotemporal dependence and a moderately low SNR. Logis-LCA and Spline-LCA outperformed competing methods for the LNGCA model and Logis-LCA outperformed all other methods for the noisy ICA model. These results suggest that LCA can be used to reveal structure for a large class of non-Gaussian observations. In the leaf example with correlated multivariate data, Spline-LCA revealed biologically meaningful clusters not apparent from PCA+ProDenICA. In our fMRI application, we simultaneously achieved dimension reduction and latent variable extraction for large image data ($T = 272$ and $V = 230,459$) and

identified artifacts not identified by PCA+ICA. Using the LCA criteria in which LCs are ordered by their estimated marginal likelihoods, the component containing the most information coincided with the component most highly correlated with the task for all subjects.

The presence of local maxima in LCA can increase computational expenses, and more initializations are required for larger values of T . Since the set of orthogonal matrices is non-convex, local optima are also a problem in PCA+ICA (e.g., Risk et al. 2014). For fMRI data, fifty initializations appeared to be adequate when estimating thirty components with nearly three hundred time points (Web Supplement Figure S.6). In general, we found that Logis-LCA was less sensitive to initialization than Spline-LCA (results not shown). However, we favor Spline-LCA because it can more accurately model source densities.

An important advantage of LCA over existing frameworks is its robustness to misspecification of the number of estimated components. This robustness suggests LCA could be used to improve estimates of brain networks in fMRI studies. In contrast, estimating the correct number of components in noisy ICA is a pre-requisite to recovering valid components (Section 3.2, see also Allasonniere and Younes 2012). Beckmann and Smith (2004) explored the use of probabilistic PCA to estimate the number of brain networks prior to ICA in order to avoid model over-fitting, which addresses the concern that over fitting may separate a single network into multiple networks. However, our simulations suggest that using too few components leads to inappropriately aggregated networks in PCA+ICA methods (Figures 2 and 3). In contrast, the components recovered for $Q^* \neq Q$ in Logis-LCA across model scenarios (Figures 2 and 3) and Spline-LCA for the LNGCA scenario (Figure 2) accurately represent functional connectivity. Moreover, in the leaf data example, the first two components were nearly identical for $Q^* = 2$ and $Q^* = 5$ for LCA but differed for PCA+ICA (Web Supplement Figures S.4 and S.5).

LCA offers a computationally tractable alternative to one of the most common applications of ICA to fMRI: artifact detection. Currently, PCA+ICA is used as a pre-processing step to reveal biologically implausible loadings and/or loadings resembling physiological arti-

facts that can be used to de-noise data for subsequent analyses (Beckmann, 2012). In LCA, these artifacts appear as LCs since they have non-Gaussian distributions. Our improved detection of artifacts (Figure 5, Web Supplement Figures S.7 and S.8) suggests LCA could be used for more powerful denoising methods over traditional PCA+ICA.

Code implementing LCA is available in the Web Supplement.

7 Acknowledgments

We thank Dr. Nathan Spreng, Department of Human Development, Cornell University, for scientific guidance and assistance with the HCP data. DSM was supported by a Xerox PARC Faculty Research Award and NSF grant DMS-1455172. BBR was partially supported by the NSF grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657).

A Proofs

We assume all random variables are mean zero. In Kagan et al. (1973), a random variable $\mathbf{X} \in \mathbb{R}^T$ is said to have a *linear structure* if it can be represented as $\mathbf{X} = \mathbf{BY}$ where the elements of \mathbf{Y} are mutually independent random variables and no two columns of \mathbf{B} are proportional. We say a linear-structure random vector \mathbf{X} has *essentially unique structure* if for any two representations $\mathbf{X} = \mathbf{BY}$ and $\mathbf{X} = \mathbf{CZ}$, we have \mathbf{B} equals \mathbf{C} up to scaling and permutation of the columns, which we denote as $\mathbf{B} \cong \mathbf{C}$. A random variable \mathbf{X} is non-unique if there exist representations $\mathbf{X} = \mathbf{BY} = \mathbf{CZ}$ but $\mathbf{B} \not\cong \mathbf{C}$. Let $\stackrel{d}{=}$ denote equal in distribution. First consider the theorem on uniqueness of decomposition.

Theorem 10.3.9 from Kagan et al. (1973). *Let $\mathbf{X} = \mathbf{AY}$ be a structural representation of \mathbf{X} and let the columns of \mathbf{A} be linearly independent. Then \mathbf{X} can be expressed as $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$, where \mathbf{X}_1 and \mathbf{X}_2 are independent, \mathbf{X}_1 has essentially unique structure and \mathbf{X}_2*

is multivariate normal with a non-unique structure. Moreover, this decomposition is unique in the sense that if $\mathbf{X} = \mathbf{Z}_1 + \mathbf{Z}_2$ is another decomposition, where \mathbf{Z}_1 has essentially unique structure, \mathbf{Z}_2 is multivariate normal, and \mathbf{Z}_1 is independent of \mathbf{Z}_2 , then $\mathbf{Z}_1 \stackrel{d}{=} \mathbf{X}_1$ and $\mathbf{Z}_2 \stackrel{d}{=} \mathbf{X}_2$.

For a proof see Kagan et al. (1973).

Before proving Theorem 1, we consider the following lemma.

Lemma 1. *Suppose \mathbf{Z} and \mathbf{X} each have essentially unique structure and $\mathbf{Z} \stackrel{d}{=} \mathbf{X}$. Consider their structural representations: $\mathbf{Z} = \mathbf{M}_\mathbf{S} \mathbf{S}$ and $\mathbf{X} = \mathbf{M}_\mathbf{S}^* \mathbf{S}^*$ where $\mathbf{M}_\mathbf{S} \in \mathbb{R}^{T \times Q}$ and $\mathbf{M}_\mathbf{S}^* \in \mathbb{R}^{T \times Q}$ for $Q \leq T$, and $\text{rank}(\mathbf{M}_\mathbf{S}) = \text{rank}(\mathbf{M}_\mathbf{S}^*) = Q$. Then $\mathbf{M}_\mathbf{S} \cong \mathbf{M}_\mathbf{S}^*$ and $\mathbf{S} \stackrel{d}{=} \mathbf{S}^*$ up to scaling and permutations.*

Proof. We have $\mathbf{M}_\mathbf{S} \mathbf{S} \stackrel{d}{=} \mathbf{M}_\mathbf{S}^* \mathbf{S}^*$. Then,

$$(\mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S})^{-1} \mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S} \mathbf{S} = (\mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S})^{-1} \mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S}^* \mathbf{S}^*.$$

Letting $\mathbf{B} = (\mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S})^{-1} \mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S}^*$, we have $\mathbf{S} \stackrel{d}{=} \mathbf{B} \mathbf{S}^*$. Note by assumption $\mathbf{S} \in \mathbb{R}^Q$ and $\mathbf{S}^* \in \mathbb{R}^Q$. Now \mathbf{S} has non-Gaussian independent components and thus has essentially unique structure for the given number of components Q (Theorem 10.3.5 in Kagan et al. 1973); in particular, $\mathbf{S} = \mathbf{I} \mathbf{S}$. We can define a random variable $\mathbf{R} = \mathbf{B}^{-1} \mathbf{S}$, and note that $\mathbf{R} \stackrel{d}{=} \mathbf{S}^*$, and \mathbf{S}^* has independent components, which implies \mathbf{R} has independent components, which implies $\mathbf{B} \mathbf{R}$ is a structural representation of \mathbf{S} . Since \mathbf{S} has essentially unique structure, $\mathbf{B} \cong \mathbf{I}$. It follows that $\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$ up to scaling and permutations.

Now consider the scaling and permutation such that $\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$. Then we have $\mathbf{B} = \mathbf{I}$, so $(\mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S})^{-1} \mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S}^* = \mathbf{I}$. Now since $(\mathbf{M}_\mathbf{S}' \mathbf{M}_\mathbf{S})^{-1} \mathbf{M}_\mathbf{S}'$ is full row rank, it has a unique right inverse equal to the Moore-Penrose pseudoinverse, which is equal to $\mathbf{M}_\mathbf{S}$, which implies $\mathbf{M}_\mathbf{S} = \mathbf{M}_\mathbf{S}^*$. For $\mathbf{B} \cong \mathbf{I}$, it follows that $\mathbf{M}_\mathbf{S}^* \cong \mathbf{M}_\mathbf{S}$. \square

We now prove Theorem 1.

Theorem 1. *Suppose \mathbf{X} follows the model in (1) of the main manuscript with Assumptions*

1-3. Then for any other representation $\mathbf{X} = \mathbf{M}_{\mathbf{S}}^* \mathbf{S}^* + \mathbf{E}^*$ where $\mathbf{S}^* \in \mathbb{R}^Q$ are independent non-Gaussian components and \mathbf{E}^* is multivariate normal, we have: $\mathbf{M}_{\mathbf{S}}^* \cong \mathbf{M}_{\mathbf{S}}$; $\mathbf{S}^* \stackrel{d}{=} \mathbf{S}$ up to scaling and permutations; and $\mathbf{E}^* \stackrel{d}{=} \mathbf{M}_{\mathbf{N}} \mathbf{N}$.

Proof. Since \mathbf{X} has a unique decomposition in the sense of Theorem 10.3.9, we have $\mathbf{M}_{\mathbf{S}} \mathbf{S} \stackrel{d}{=} \mathbf{M}_{\mathbf{S}}^* \mathbf{S}^*$ and $\mathbf{M}_{\mathbf{N}} \mathbf{N} \stackrel{d}{=} \mathbf{E}^*$. Moreover, $\mathbf{M}_{\mathbf{S}} \mathbf{S}$ and $\mathbf{M}_{\mathbf{S}}^* \mathbf{S}^*$ have essentially unique structure (Theorem 10.3.5 in Kagan et al. 1973). Applying Lemma 1, we obtain the desired result. \square

Corollary 1. *Suppose the linear structure model in (1) of the main manuscript with density defined in (2) and suppose that Assumptions 1-3 hold. Then $\{f_1, \mathbf{w}_1\}, \dots, \{f_Q, \mathbf{w}_Q\}$ are identifiable up to sign and ordering. Note the rows \mathbf{w}_{Q+k} for $k = 1, \dots, T - Q$ are not identifiable.*

Proof. For identifiability, we need to show that if there exist densities g_1, \dots, g_T and a matrix \mathbf{C} such that

$$|\det(\mathbf{L})| \prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{L} \mathbf{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{L} \mathbf{x}) = |\det(\mathbf{C})| \prod_{\ell=1}^T g_{\ell}(\mathbf{c}'_{\ell} \mathbf{x}) \quad (\text{S.1})$$

then Q of the marginal densities g_1, \dots, g_T are equivalent to f_1, \dots, f_Q and the rest of the distributions are Gaussian, and that each of the corresponding Q rows of \mathbf{C} equal $\mathbf{w}'_1 \mathbf{L}, \dots, \mathbf{w}'_Q \mathbf{L}$. Using a change of variable $\mathbf{Z} = \mathbf{L} \mathbf{X}$, we consider the model $\mathbf{Z} = \mathbf{A}_{\mathbf{S}} \mathbf{S} + \mathbf{A}_{\mathbf{N}} \mathbf{N}$, such that $[\mathbf{w}'_1; \dots; \mathbf{w}'_Q] = \mathbf{A}'_{\mathbf{S}}$ (where $[\mathbf{w}'_1; \dots; \mathbf{w}'_Q]$ indicates stacked row vectors) and $[\mathbf{w}'_{Q+1}; \dots; \mathbf{w}'_T] = \mathbf{A}'_{\mathbf{N}}$. Then (S.1) is equivalent to

$$\prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{z}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{z}) = |\det(\mathbf{C})| |\det(\mathbf{L})|^{-1} \prod_{\ell=1}^T g_{\ell}(\mathbf{c}'_{\ell} \mathbf{L}^{-1} \mathbf{z}).$$

We define $\mathbf{R} = \mathbf{C} \mathbf{L}^{-1}$ such that we have

$$\prod_{q=1}^Q f_q(\mathbf{w}'_q \mathbf{z}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{z}) = |\det(\mathbf{R})| \prod_{\ell=1}^T g_{\ell}(\mathbf{r}'_{\ell} \mathbf{z}). \quad (\text{S.2})$$

We have demonstrated identifiability up to signed permutations if we can show that Q of the marginal densities g_1, \dots, g_T are equivalent to f_1, \dots, f_Q ; that each of the corresponding Q rows of \mathbf{R} equal $\pm \mathbf{w}_1, \dots, \pm \mathbf{w}_Q$; and that $|\det(\mathbf{R})| = 1$.

Define $\mathbf{K} = \mathbf{R}^{-1}$. Given the relationship in (S.2), then there exists another *linear structure* representation of \mathbf{Z} such that $\mathbf{Z} = \mathbf{K}\mathbf{Y}$. Without loss of generality, we have $\mathbf{E}\mathbf{Y}\mathbf{Y}' = \mathbf{I}$ (there is no loss of generality because we can scale \mathbf{K} such that $\mathbf{E}\mathbf{Y}\mathbf{Y}' = \mathbf{I}$). From Theorem 10.3.3 in Kagan et al. (1973), \mathbf{Z} has the decomposition $\mathbf{Z} = \mathbf{K}_1\mathbf{Y}_1 + \mathbf{K}_2\mathbf{Y}_2$ in which \mathbf{Y}_1 are independent non-Gaussian and \mathbf{Y}_2 are Gaussian. Then from Theorem 1 and the assumption of unit variance, we have that $\mathbf{Y}_1 \stackrel{d}{=} \mathbf{S}$ (up to ordering), and it follows that there exists a subset of g_1, \dots, g_T equal to f_1, \dots, f_Q . Also from Theorem 1, we have $\mathbf{K}_1 \cong \mathbf{A}_\mathbf{S}$. Note that $\mathbf{K} \in \mathcal{O}_{T \times T}$ since $\mathbf{E}\mathbf{Y}\mathbf{Y}' = \mathbf{I}$ and $\mathbf{E}\mathbf{Z}\mathbf{Z}' = \mathbf{I}$. Then the scaling of \mathbf{K}_1 is also identifiable such that there exists a signed permutation matrix, \mathbf{P}_\pm , such that $\mathbf{K}_1\mathbf{P}_\pm = \mathbf{A}_\mathbf{S}$. Note that $\mathbf{W}_\mathbf{S} = \mathbf{A}'_\mathbf{S}$. Define $\mathbf{R}_\mathbf{S} = \mathbf{K}'_1$. Then $\mathbf{P}'_\pm \mathbf{R}_\mathbf{S} = \mathbf{W}_\mathbf{S}$. \square

Proposition 1. *Consider a random vector $\mathbf{Y} \in \mathbb{R}^T$ with density $f_\mathbf{Y}$ such that $\mathbf{E}\mathbf{Y} = \mathbf{0}$ and $\mathbf{E}\mathbf{Y}\mathbf{Y}' = \mathbf{I}_T$. Then for any \mathbf{o} and \mathbf{w} such that $\mathbf{o}'\mathbf{o} = \mathbf{w}'\mathbf{w} = 1$, we have*

$$\mathbf{E} \log \phi(\mathbf{o}'\mathbf{Y}) = \mathbf{E} \log \phi(\mathbf{w}'\mathbf{Y}).$$

Proof. We can ignore the normalizing constants of $\phi(x)$ and consider the quadratic term of the Gaussian kernel. Then we have $\mathbf{E}(\mathbf{o}'\mathbf{Y})^2 = \mathbf{o}'\mathbf{E}(\mathbf{Y}\mathbf{Y}')\mathbf{o} = \mathbf{o}'\mathbf{I}\mathbf{o} = \mathbf{o}'\mathbf{o} = 1$ and similarly for $\mathbf{E}(\mathbf{w}'\mathbf{Y})^2$. \square

Next we describe conditions for consistency when the density used in the objective function equals the true density.

Theorem 2. *Suppose \mathbf{X} follows the LNGCA model in (1) of the main manuscript with Assumptions 1-4. Additionally assume $\mathbf{E}\mathbf{X} = \mathbf{0}$ and $\mathbf{E}\mathbf{X}\mathbf{X}' = \mathbf{I}$. Let $\mathbf{W}_\mathbf{S}$ denote the first Q*

rows of \mathbf{M}^{-1} . Given an iid sample $\{\mathbf{x}_v\}$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Or} \rightarrow \mathbf{W}_{\mathbf{S}}$ almost surely on the equivalence class of signed permutations.

Proof. Note that $\mathcal{O}_{Q \times T}$ is compact. We will show the four assumptions in Wald's consistency proof as recast in Pollard (2001) are satisfied. (A similar proof is in Theorem 5.14 in van der Vaart (2000) but we refer to the assumptions as conveniently enumerated in Pollard (2001)). Let $f_{\mathbf{S}}$ denote the joint density of the LCs. First, we show $\mathbb{E} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X}) \leq \mathbb{E} \log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})$ for any $\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}$ with equality if and only if $\mathbf{O}_{\mathbf{S}} \cong \mathbf{W}_{\mathbf{S}}$. Let $\mathbf{W}_{\mathbf{N}}$ denote rows $T - Q$ to T of \mathbf{W} .

Note that the fact that $\mathbb{E} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X}) \leq \mathbb{E} \log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})$ does not hold trivially can be seen by the following argument:

$$\begin{aligned} \mathbb{E} \log \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})} &\leq \log \mathbb{E} \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})} \\ &= \log \int \left\{ \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{x})} \right\} \{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{x})\} d\mathbf{x} \\ &= \log \int f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{x}) d\mathbf{x}. \end{aligned}$$

We would like the last quantity to be equal to zero, in which case we would obtain the desired bound. Let \mathbf{W}^* be the $T \times T$ matrix formed by stacking $\mathbf{O}_{\mathbf{S}}$ and $\mathbf{W}_{\mathbf{N}}$. The term $f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{x})$ is a density if and only if $|\det(\mathbf{W}^*)| = 1$, which is not true in general because $\mathbf{O}_{\mathbf{S}}$ may not be orthogonal to $\mathbf{W}_{\mathbf{N}}$. Consequently, this quantity could integrate to greater than one, in which case we would have $\mathbb{E} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X}) \leq \mathbb{E} \log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X}) + \alpha$ for some $\alpha > 0$, and thus our bound is not tight enough.

Thus define an orthogonal matrix in $\mathcal{O}_{T \times T}$ such that rows 1 to Q are equal to $\mathbf{O}_{\mathbf{S}}$ and the other rows are arbitrary. Then

$$\begin{aligned} \mathbb{E} \log \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})} &= \mathbb{E} \log \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{X})} \\ &= \mathbb{E} \log \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{X})}, \end{aligned}$$

where the second line follows from the previous proposition. Then applying Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E} \log \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{X})} &\leq \log \mathbb{E} \frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{X})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{X})} \\
&= \log \int \left(\frac{f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{x})}{f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{x})} \right) f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{x}) d\mathbf{x} \\
&= \log \int f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{x}) d\mathbf{x} \\
&= 0,
\end{aligned}$$

which holds with equality if and only if $f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{x}) = f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{x})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{x})$, where the only if direction is a consequence of absolute continuity. Now suppose equality holds for the matrix $\mathbf{O}_{\mathbf{S}}^*$ and let \mathbf{Y} be a random variable with density $f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}^*\mathbf{y})\phi(\mathbf{O}_{\mathbf{N}}\mathbf{y}) = f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{y})\phi(\mathbf{W}_{\mathbf{N}}\mathbf{y})$. Let $\mathbf{O}_+ = [\mathbf{O}_{\mathbf{S}}^{*'}, \mathbf{O}_{\mathbf{N}}']'$. Then there exist random variables \mathbf{R}_+ and \mathbf{R} such that $\mathbf{Y} = \mathbf{O}_+\mathbf{R}_+$ and $\mathbf{Y} = \mathbf{W}\mathbf{R}$. Applying Theorem 1, we have $\mathbf{O}_{\mathbf{S}}^* \cong \mathbf{W}_{\mathbf{S}}$. It follows that

$$\mathbb{E} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{X}) < \mathbb{E} \log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})$$

for all $\mathbf{O}_{\mathbf{S}} \not\cong \mathbf{W}_{\mathbf{S}}$. The other three conditions are satisfied since we assume continuous densities which implies upper semicontinuity (condition ii); we assume the source densities are bounded, say by some constant A , so $\mathbb{E} \sup_{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{x}) \leq \mathbb{E} \log A < \infty$ (condition iii); and our estimator is an exact M-estimator (condition iv). \square

Next we describe conditions for consistency when the density used in the objective function may not be equal to the density of the LCs. We first present a result that is contained in the proof of Theorem 1 in Hyvärinen and Oja (1998), where here the nonlinearity is equal to the log of the density used in the objective function. Additionally, define $\mathbf{Z} = [\mathbf{S}', \mathbf{N}']'$.

Lemma 2. Let $\mathbf{e}_1 = [1, 0, \dots, 0]'$ and define $\boldsymbol{\epsilon}$ such that $\|\mathbf{e}_1 + \boldsymbol{\epsilon}\|_2 = 1$. Then

$$\mathbb{E} \log p_1[(\mathbf{e}_1 + \boldsymbol{\epsilon})' \mathbf{Z}] = \mathbb{E} \log p_1(s_1) + \frac{1}{2} [\mathbb{E} r_1'(s_1) - \mathbb{E} s_1 r_1(s_1)] \sum_{q=2}^T \epsilon_q^2 + o(\|\boldsymbol{\epsilon}\|_2^2).$$

Proof. Calculating the gradient with respect to \mathbf{o} ,

$$\nabla \mathbb{E} \log p_1(\mathbf{o}' \mathbf{Z}) = \mathbb{E} \mathbf{Z} r_1(\mathbf{o}' \mathbf{Z}),$$

where we have applied Assumption 5 to interchange differentiation and integration. Evaluating this at \mathbf{e}_1 , and using the fact that $\mathbb{E} \mathbf{S}_q = \mathbb{E} \mathbf{N}_k = 0$, $q = 1, \dots, Q$, $k = 1, \dots, T - Q$, and the fact that \mathbf{S}_1 is independent of \mathbf{S}_q and \mathbf{N}_k ,

$$\left. \nabla \mathbb{E} \log p_1(\mathbf{o}' \mathbf{Z}) \right|_{\mathbf{e}_1} = \mathbf{e}_1 \mathbb{E} s_1 r_1(s_1).$$

We also have

$$\left. \nabla^2 \mathbb{E} \log p_1(\mathbf{o}' \mathbf{Z}) \right|_{\mathbf{e}_1} = \text{diag} [\mathbb{E} s_1^2 r_1'(s_1), \mathbb{E} r_1'(s_1), \dots, \mathbb{E} r_1'(s_1)]$$

where we have interchanged integration and differentiation using Assumption 5 and applied independence and the fact that $\mathbb{E} \mathbf{S}_q^2 = \mathbb{E} \mathbf{N}_k^2 = 1$.

Now for some small $\boldsymbol{\epsilon}$ with $\|\mathbf{e}_1 + \boldsymbol{\epsilon}\|_2 = 1$, we have

$$\begin{aligned} \mathbb{E} \log p_1[(\mathbf{e}_1 + \boldsymbol{\epsilon})' \mathbf{Z}] &= \\ \mathbb{E} \log p_1(s_1) + \boldsymbol{\epsilon}' \mathbf{e}_1 \mathbb{E} s_1 r_1(s_1) + \frac{1}{2} \boldsymbol{\epsilon}' \text{diag} [\mathbb{E} s_1^2 r_1'(s_1), \mathbb{E} r_1'(s_1), \dots, \mathbb{E} r_1'(s_1)] \boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2^2) &= \\ \mathbb{E} \log p_1(s_1) + \epsilon_1 \mathbb{E} s_1 r_1(s_1) + \frac{1}{2} \epsilon_1^2 \mathbb{E} s_1^2 r_1'(s_1) + \frac{1}{2} \mathbb{E} r_1'(s_1) \sum_{q>1} \epsilon_q^2 + o(\|\boldsymbol{\epsilon}\|_2^2). \end{aligned}$$

Note that $\epsilon_1 = \sqrt{1 - \sum_{q>1} \epsilon_q^2} - 1$. Now we consider the first-order Taylor series expansion of $\sqrt{1 - \gamma}$ about 0 which is $1 - \gamma/2 + o(\gamma)$, so $\epsilon_1 = -\frac{1}{2} \sum_{q>1} \epsilon_q^2 + o(\sum_{q>1} \epsilon_q^2)$. Then we can

write

$$\mathbb{E} \log p_1[(\mathbf{e}_1 + \boldsymbol{\epsilon})'\mathbf{Z}] = \mathbb{E} \log p_1(s_1) + \frac{1}{2} [\mathbb{E} r'_1(s_1) - \mathbb{E} s_1 r_1(s_1)] \sum_{q>1} \epsilon_q^2 + o(\|\boldsymbol{\epsilon}\|_2^2).$$

□

Theorem 3. *Suppose \mathbf{X} follows the LNGCA model in (1) of the main manuscript with Assumptions 1-6. Additionally assume $\mathbb{E} \mathbf{X} = \mathbf{0}$ and $\mathbb{E} \mathbf{X} \mathbf{X}' = \mathbf{I}$. Given an iid sample $\{\mathbf{x}_v\}$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Local} \rightarrow \mathbf{W}_{\mathbf{S}}$ almost surely on the equivalence class of signed permutations.*

Proof. Wald's method for consistency of the MLE can be applied to the more general setting in which the wrong likelihood is used if the supremum of the population objective function corresponds to the set of true parameters. Thus we need to show $\sup_{\mathbf{O}_{\mathbf{S}} \in \mathcal{N}_{\epsilon}(\mathbf{W}_{\mathbf{S}})} \mathbb{E} \log p(\mathbf{O}_{\mathbf{S}} \mathbf{X})$ occurs at $\mathbf{W}_{\mathbf{S}}$. It suffices to consider the case where $\mathbf{W} = \mathbf{I}_{T \times T}$, since we can use the change of variables $\mathbf{z} = \mathbf{W} \mathbf{x} = [\mathbf{s}', \mathbf{n}']'$. Let $\mathbf{e}_1 = [1, 0, \dots, 0]' \in \mathbb{R}^T$ and consider a perturbation $\boldsymbol{\epsilon}^* \in \mathbb{R}^T$ with $\|\mathbf{e}_1 + \boldsymbol{\epsilon}^*\|_2 = 1$. From the previous lemma, we have

$$\mathbb{E} \log p_1[(\mathbf{e}_1 + \boldsymbol{\epsilon}^*)'\mathbf{Z}] = \mathbb{E} \log p_1(s_1) + \frac{1}{2} \mathbb{E} [r'_1(s_1) - s_1 r_1(s_1)] \sum_{q>1} \epsilon_q^{*2} + o(\|\boldsymbol{\epsilon}^*\|_2^2).$$

By Assumption 5 and for sufficiently small $\boldsymbol{\epsilon}^*$, we have

$$\frac{1}{2} \mathbb{E} [r'_1(s_1) - s_1 r_1(s_1)] \sum_{q>1} \epsilon_q^2 + o(\|\boldsymbol{\epsilon}^*\|_2^2) < 0,$$

which makes \mathbf{e}_1 a local maximum for $\mathbb{E} \log p_1(\mathbf{o}'\mathbf{Z})$. Since this also true for $\mathbb{E} \log p_q(\mathbf{o}'\mathbf{Z})$, $q = 2, \dots, Q$, we have that $\mathbf{I}_{Q \times T}$ (the $Q \times Q$ identity matrix padded with zeros) is a local maximum on the set $\mathcal{G}_{Q \times T} = \{\mathbf{B} \in \mathbb{R}^{Q \times T} : \text{diag } \mathbf{B} \mathbf{B}' = \mathbf{1}_Q\}$. Since $\mathcal{O}_{Q \times T} \subset \mathcal{G}_{Q \times T}$ and $\mathbf{I}_{Q \times T} \in \mathcal{O}_{Q \times T}$, $\mathbf{I}_{Q \times T}$ is also a local maximum on $\mathcal{O}_{Q \times T}$. (For a similar argument in ICA, see Wei 2015). Now by Assumption 6, $\mathbf{I}_{Q \times T}$ is the unique maximum on $\mathcal{N}_{\epsilon}(\mathbf{W}_{\mathbf{S}})$. This establishes condition (i) in Wald's theorem presented in Pollard (2001). Conditions (ii) - (iv) follow from the argument used in our Theorem 2, and thus we obtain $\widehat{\mathbf{W}}_{\mathbf{S}}^{Local} \rightarrow \mathbf{W}_{\mathbf{S}}$

Next we show that the solution to the Spline-LCA objective function corresponds to a mean-zero density.

Proposition 2. *Let G be the class of all cubic splines $g : \mathbb{R} \rightarrow \mathbb{R}$. Consider the argmax of (11) of the main manuscript for $g_q \in G$ with g_q denoting the tilt function for the q th component. Then (i) $\int \phi(u)e^{g_q(u)} du = 1$ and (ii) $\int u\phi(u)e^{g_q(u)} du = 0$ for each q .*

Proof. It suffices to consider the case $Q^* = 1$. Let \mathbf{o}_1 be given. Let G be the set of cubic splines and note that for any $g \in G$, we can write $g(u) = \theta_0 + \theta_1 u + j(u)$ with $\theta_0 \in \mathbb{R}$, $\theta_1 \in \mathbb{R}$, and $j(u)$ does not depend on θ_0 or θ_1 . Noting that $\partial(\int \phi(u)e^{g(u)} du)/\partial\theta_0 = \partial(e^{\theta_0} \int \phi(u)e^{\theta_1 u + j(u)} du)/\partial\theta_0 = \int \phi(u)e^{g(u)} du$, we have

$$\frac{\partial \ell_{pen}}{\partial \theta_0} = 1 - \int \phi(u)e^{g(u)} du,$$

from which it follows that at the optimum g^* , $\phi(u)e^{g^*(u)}$ is a density. Next, note that $\partial(\phi(u)e^{\theta_0 + \theta_1 u + j(u)})/\partial\theta_1 = u\phi(u)e^{g(u)}$. Then,

$$\frac{\partial \ell_{pen}}{\partial \theta_1} = \frac{1}{V} \sum_{v=1}^V \mathbf{o}'_1 \widehat{\mathbf{L}} \mathbf{x}_v - \int u\phi(u)e^{g(u)} du,$$

where we have assumed $\int |u|\phi(u)e^{g(u)} du < \infty$ to interchange integration and differentiation. Then it follows that $\mathbb{E} U = 0$ for U with density $\phi(u)e^{g^*(u)}$. □

B Additional Background

B.1 Projection Pursuit, D-FastICA, and Non-Gaussian Component Analysis

Projection pursuit is an exploratory method for finding low-dimensional representations of multivariate data that reveal interesting patterns and structure (Huber, 1985). Let $\{\mathbf{x}_v\}$, $v = 1, \dots, V$ be a data sample with $\mathbf{x}_v \in \mathbb{R}^T$, and assume $\sum_{v=1}^V \mathbf{x}_v = \mathbf{0}$, where $\mathbf{0}$ is the vector of T zeros, and $\frac{1}{V} \sum_{v=1}^V \mathbf{x}_v^2 = \mathbf{1}$, where $\mathbf{1}$ is a length T vector of ones. Let Q be the number of projection pursuit directions that are estimated. In FastICA in deflation mode (D-FastICA), the projection pursuit index is equivalent to an approximation of negentropy (Hyvarinen, 1999):

$$\mathbf{w}_q = \underset{\mathbf{w} \in \mathbb{R}^T}{\operatorname{argmax}} \left\{ \frac{1}{V} \sum_{v=1}^V G(\mathbf{w}' \mathbf{x}_v) - \mathbb{E} G(n) \right\}^2, \quad (\text{S.3})$$

where \mathbf{w} is orthogonal to $\widehat{\mathbf{w}}_1, \dots, \widehat{\mathbf{w}}_{q-1}$ and $\|\mathbf{w}\| = 1$ with $\|\cdot\|$ denoting the L2-norm, G is a non-linear function, and n is a standard normal random variable. A common choice for G is $\log \cosh(x)$, which is used to estimate projection pursuit directions in our simulations.

NGCA uses multiple projection pursuit indices (Blanchard et al., 2006) or radial basis functions (Kawanabe et al., 2007) to find a non-Gaussian subspace that is assumed to contain the interesting features of data. NGCA can be formulated using a semiparametric likelihood,

$$f_{\mathbf{X}}(\mathbf{x}) = h(\mathbf{W}_{\mathbf{S}} \mathbf{x}) \phi_{\mathbf{0}, \mathbf{\Sigma}}(\mathbf{x}) \quad (\text{S.4})$$

where $\phi_{\mathbf{0}, \mathbf{\Sigma}}$ is multivariate normal with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$; $\mathbf{W}_{\mathbf{S}}$ is a $Q \times T$ matrix; and $h(\cdot)$ is a function that captures departures from Gaussianity under the constraint that $f_{\mathbf{X}}(\mathbf{x})$ is a density. The NGCA model does not assume independent factors, and thus we do not consider it in our simulations.

The density in the Spline-LCA model can be considered an extension of (S.4) with the additional assumption of independence.

Proposition 3. *Let \mathbf{X} be a random variable from the LCA model where the LCs have tilted Gaussian densities. Then the density of \mathbf{X} is*

$$f_{\mathbf{X}}(\mathbf{x}) = \phi_{\mathbf{0}, \Sigma}(\mathbf{x}) \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})}$$

where $\phi_{\mathbf{0}, \Sigma}$ is the mean zero multivariate distribution with covariance Σ .

Proof. Using the tilted Gaussian density, we have

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \det \mathbf{L} \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})} \phi(\mathbf{w}'_q \mathbf{L} \mathbf{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}'_{Q+k} \mathbf{L}' \mathbf{x}) \\ &= \left\{ \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})} \right\} (2\pi)^{-T/2} (\det \mathbf{L}) \exp \left\{ -\frac{1}{2} \sum_{k=1}^T \mathbf{x}' \mathbf{L} \mathbf{w}_k \mathbf{w}'_k \mathbf{L} \mathbf{x} \right\} \\ &= (\det \Sigma)^{-1/2} (2\pi)^{-T/2} \exp \left\{ -\frac{1}{2} \mathbf{x}' \Sigma^{-1} \mathbf{x} \right\} \prod_{q=1}^Q e^{g_q(\mathbf{w}'_q \mathbf{L} \mathbf{x})}. \end{aligned}$$

□

Writing the likelihood in this way, one notes that we are using the Gaussian density to model the covariance between components and we are using the tilt functions to model deviations from the Gaussian model.

B.2 Noisy ICA and IFA

In the noisy ICA model, Q ICs are corrupted by rank- T Gaussian noise, where $Q \leq T$ (Hyvärinen et al., 2001),

$$\mathbf{X} = \mathbf{M}_S \mathbf{S} + \mathbf{E} \tag{S.5}$$

with $\mathbf{X} \in \mathbb{R}^T$, $\mathbf{M}_\mathbf{S}$ is $T \times Q$ with $Q \leq T$, \mathbf{E} is mean-zero multivariate normal with covariance matrix Ψ , and \mathbf{E} is independent of \mathbf{S} .

Assume that $\Psi = \sigma^2 \mathbf{I}$. Let $\lambda_1, \dots, \lambda_Q$ denote the eigenvalues from the covariance matrix of $\mathbf{M}_\mathbf{S} \mathbf{S}$ and let $\lambda_{\epsilon_1}, \dots, \lambda_{\epsilon_T}$ denote the eigenvalues from the decomposition of \mathbf{E} . Under the assumption of isotropic noise, we have $\lambda_{\epsilon_i} = \lambda_{\epsilon_j} = \sigma^2$ for all $i, j = 1, \dots, T$. Then the eigenvalue decomposition can be written as

$$\text{Cov } \mathbf{X} = \mathbf{U} \text{diag}(\lambda_1 + \sigma^2, \dots, \lambda_Q + \sigma^2, \sigma^2, \dots, \sigma^2) \mathbf{U}'. \quad (\text{S.6})$$

Let \mathbf{X}_{data} be the $V \times T$ data matrix. In PCA+ICA, noise-free ICA is applied to the first Q left singular vectors of \mathbf{X}_{data} multiplied by \sqrt{V} , which is equivalent to the first Q standardized principal components.

In IFA, (S.5) is estimated under the assumption that the densities of the ICs are Gaussian mixtures (Attias, 1999). In its original formulation, Ψ was an arbitrary positive definite matrix, the IC densities had K_q classes, and the variance of each IC was standardized to unity after each iteration. In our presentation and estimation, we assume that the covariance of the noise is $\sigma^2 \mathbf{I}$ and IC densities are mixtures of two Gaussians, which has been assumed elsewhere (e.g., Guo and Tang 2013; Beckmann and Smith 2004), and enforce the constraint that the IC densities are mean zero with unit variance. Let π_{q1} be the probability that an observation of the q th IC comes from the first class, where the first class has a normal distribution with mean μ_{q1} and variance ν_{q1} . Then the probability, mean, and variance for the second class are $\pi_{q2} = 1 - \pi_{q1}$, $\mu_{q2} = -\frac{\pi_{q1}\mu_{q1}}{\pi_{q2}}$, and $\nu_{q2} = \frac{1 - \pi_{q1}\nu_{q1} - \pi_{q1}\mu_{q1}^2}{\pi_{q2}} - \mu_{q2}^2$, respectively. Then the joint density of \mathbf{x}_v can be written

$$f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{M}_\mathbf{S}) = \prod_{t=1}^T \int \phi_{0, \sigma^2}(\mathbf{x}_t - \mathbf{m}'_t \mathbf{s}) f_{\mathbf{S}}(\mathbf{s}) d\mathbf{s}, \quad (\text{S.7})$$

where ϕ_{0,σ^2} is a normal density with mean zero and variance σ^2 and

$$f_{\mathbf{s}}(\mathbf{s}) = \prod_{q=1}^Q \left\{ \pi_{q1} \phi_{\mu_{q1}, \nu_{q1}}(s_q) + \pi_{q2} \phi_{\mu_{q2}, \nu_{q2}}(s_q) \right\}.$$

Analytic integration across \mathbf{s} is possible. Let k_q equal one if s_q is in the first class and zero otherwise. Let \mathcal{K} be the set of all possible states for the Q components composed from the Cartesian product Q -times of the singletons $\{\{0\}, \{1\}\}$. Let $\mathbf{k}_j = \{k_1, \dots, k_Q\}$ denote an element of \mathcal{K} , where $j \in \{1, \dots, 2^Q\}$. Let $\boldsymbol{\mu}(\mathbf{k}_j)$ and $\boldsymbol{\nu}(\mathbf{k}_j)$ denote the conditional means of \mathbf{s} given the states \mathbf{k}_j . Now define

$$\boldsymbol{\Sigma}(\mathbf{k}_j) = \mathbf{M}_{\mathbf{S}} \text{diag}\{\boldsymbol{\nu}(\mathbf{k}_j)\} \mathbf{M}_{\mathbf{S}}' + \sigma^2 \mathbf{I}$$

and

$$\boldsymbol{\mu}^*(\mathbf{k}_j) = \mathbf{M}_{\mathbf{S}} \boldsymbol{\mu}(\mathbf{k}_j).$$

Then the density is

$$f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{M}_{\mathbf{S}}) = \sum_{\mathbf{k}_j \in \mathcal{K}} \Phi\{\mathbf{x} \mid \boldsymbol{\mu}^*(\mathbf{k}_j), \boldsymbol{\Sigma}(\mathbf{k}_j)\} \prod_{q=1}^Q \pi_{q1}^{k_q} \pi_{q2}^{1-k_q} \quad (\text{S.8})$$

with $\Phi\{\mathbf{x} \mid \boldsymbol{\mu}^*(\mathbf{k}_j), \boldsymbol{\Sigma}(\mathbf{k}_j)\}$ multivariate normal with mean $\boldsymbol{\mu}^*(\mathbf{k}_j)$ and variance $\boldsymbol{\Sigma}(\mathbf{k}_j)$ (see (16) and (17) in Attias 1999). Then a likelihood can be constructed from (S.8), and given some $\widehat{\mathbf{M}}_{\mathbf{S}}$, the ICs can be estimated from their conditional means. Alternatively, maximum a posteriori estimates of the ICs could be obtained, though we pursue the former here.

C Using the fixed-point algorithm to fit the LCA model

Here we describe the fixed-point algorithm from Hyvarinen (1999). Our account is equivalent to Hyvarinen (1999) except we use our novel discrepancy measure ($PMSE$) and a different orthogonalization method. Under the constraint that the noise components follow a standard

normal distribution, we can ignore rows $Q^* + 1 : T$ in $\widehat{\mathbf{W}}$. For now, we assume the densities of the latent components f_1, \dots, f_{Q^*} are known. Define the scalar $h_q(x) = \log f_q(x)$, and let $h'(x)$ denote its derivative. Algorithm 1 provides details on estimating $\widehat{\mathbf{W}}_{\mathbf{S}}$.

Algorithm 2: The fastICA algorithm for LCA.

Inputs : The whitened $V \times T$ data matrix \mathbf{Z} ; initial $\mathbf{W}_{\mathbf{S}}^0$; tolerance ϵ .

Result: Estimates of the unmixing matrix, $\widehat{\mathbf{W}}_{\mathbf{S}}$, and latent components, $\widehat{\mathbf{S}} = \mathbf{Z}\widehat{\mathbf{W}}_{\mathbf{S}}'$.

1. Let $\mathbf{S}^0 = \mathbf{Z}\mathbf{W}_{\mathbf{S}}^{0'}$ and let $n = 0$.
2. For each $q = 1, \dots, Q$, calculate

$$\mathbf{w}_q^* = \frac{1}{V} \sum_{v=1}^V \{h'_q(\mathbf{w}_q^{(n)'} \mathbf{z}_v) \mathbf{z}_v - h''_q(\mathbf{w}_q^{(n)'} \mathbf{z}_v) \mathbf{w}_q^{(n)}\}$$

3. Calculate the thin SVD of $\mathbf{W}_{\mathbf{S}}^* = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*'}$.
 4. Let $\mathbf{W}^{(n+1)} = \mathbf{U}^* \mathbf{V}^{*'}$.
 5. If $PMSE(\mathbf{W}_{\mathbf{S}}^{(n+1)'}, \mathbf{W}_{\mathbf{S}}^{(n)'}) < \epsilon$, stop, else increment n and repeat (2)-(4).
-

D Supplemental materials for simulations examining distributional and noise-rank assumptions

We fit D-FastICA using the ‘deflation’ option in the fastICA R package (Marchini et al., 2010). However, this popular function does not include an option to use projection pursuit for dimension reduction. If one specifies some $Q < T$ number of components, PCA is performed prior to the ICA. Consequently, one must estimate all T directions and then subset to the first two.

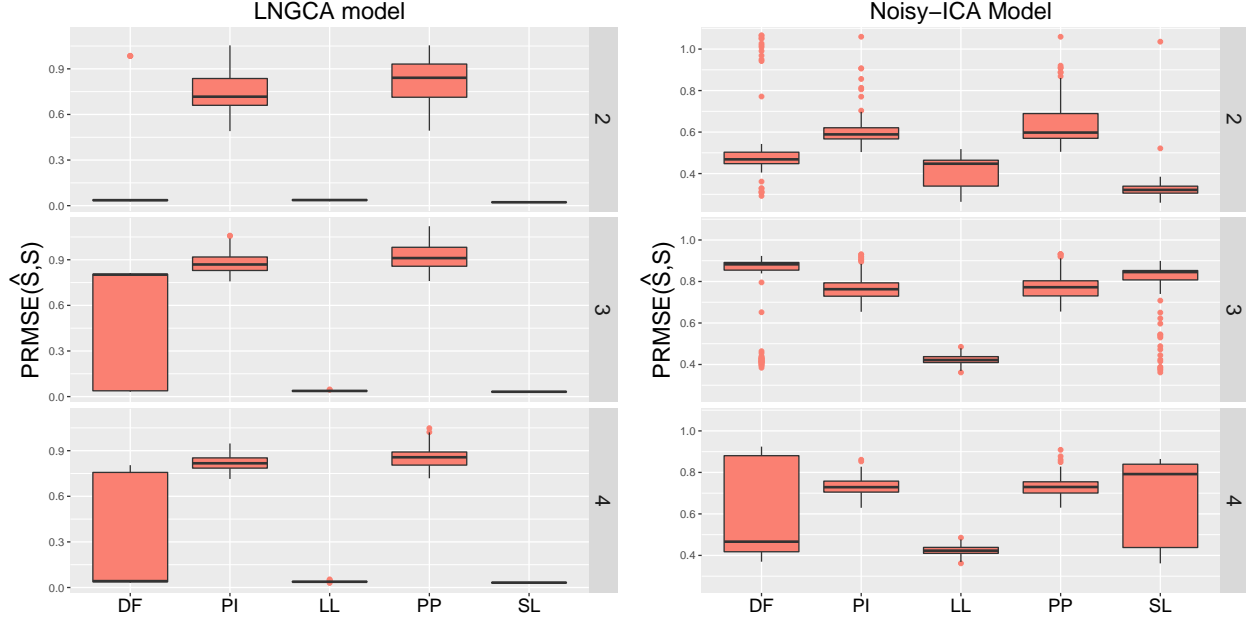
We fit the IFA model with two-class mixtures of normals by maximizing the log likelihood using a numerical optimizer. This contrasts with methods using approximating EM algorithms, as described in the introduction. Our implementation is not scalable to large Q or T (nor is the exact EM algorithm) but suffices for the simulation experiments. For IFA,

one must specify initial values for the unmixing matrix, the variance of the isotropic noise, and the parameters of the Gaussian mixtures. We had four strategies to find the argmax as detailed here. In our function, we constrain the latent component distributions to have zero expectation and unit norm, and as a result, the number of parameters to estimate for each latent component distribution is three. First, we estimated the parameters of the model proposed in Beckmann and Smith (2004) (BS-PICA) and used this solution to initialize the IFA. We then estimated the model from six additional random matrices but with density parameters initialized from the BS-PICA solution. Secondly, when the IFA model was true, we initialized it from the true mixing matrix and true density parameters and also from six additional random matrices with density parameters initialized from their true values. When the IFA model was not true, we initialized it from the true mixing matrix but with the density parameters initialized from their BS-PICA estimates and an additional six random matrices. Thirdly, we initialized the algorithm from seven random matrices but with initial Gaussian mixture densities defined by the parameters $(0.7, 0.7, -0.5, -0.5, 0.5, 0.5)$ (super-Gaussian distribution) for $\pi_{11}, \pi_{21}, \mu_{11}, \mu_{21}, \nu_{11}, \nu_{21}$ and $\sigma^2 = 1$. Finally, we initialized the algorithm from seven random matrices but with initial Gaussian mixture densities defined by the parameters $(0.3, 0.3, -1, -1, 0.5, 0.5)$ (sub-Gaussian distribution) with $\sigma^2 = 1$.

The matrices \mathbf{M}_S and \mathbf{M}_N were generated by first simulating a 5×5 matrix with standard normal entries, taking the singular value decomposition (SVD), then creating a diagonal matrix with five singular values from a $\text{uniform}(1,10)$ distribution, followed by multiplying the left singular vectors from the SVD, the diagonal matrix, and the right singular vectors, which created $[\mathbf{M}_S, \mathbf{M}_N]$. For the noisy ICA model, we generated a random mixing matrix in the same manner, then retained the first two columns.

To generate semi-orthogonal random matrices to initiate the fixed point algorithm, matrices were generated by taking the left eigenvectors from the SVD of a 2×5 matrix with entries simulated from a standard normal. We generated random matrices constrained to the principal subspace in the following manner. Let $\hat{\mathbf{U}}_{1:Q}$ denote the first Q rows from $\hat{\mathbf{U}}$ in

Figure S.1: Boxplots of $PRMSE$ for estimated columns of \mathbf{S} from simulations of spatial networks with temporal dependence and $Q = 3$ with $Q^* = 2, 3$, or 4. ‘DF’ = D-FastICA; ‘PI’ = PCA+Infomax; ‘LL’ = Logis-LCA; ‘PP’ = PCA+ProDenICA; ‘SL’ = Spline-LCA.



the decomposition $\hat{\Sigma} = \hat{\mathbf{U}}\hat{\Lambda}\hat{\mathbf{U}}'$. Then constraining the initial matrix, $\mathbf{W}_{\mathbf{S}}^0$, to the principal subspace is equivalent to $\mathbf{W}_{\mathbf{S}}^0 = \hat{\mathbf{U}}_{1:Q}\mathbf{O}$ where \mathbf{O} is a random $Q \times Q$ orthogonal matrix.

E Supplemental figures for the spatio-temporal network simulations

The permutation-invariant root mean squared errors for the components estimated from the spatio-temporal network simulations are much lower for Logis-LCA and Spline-LCA when the noise rank is $T - Q$ (Figure S.1). When the noise is rank- T , Logis-LCA performs best. Spline-LCA is excellent at finding two of the three components, but appears to sometimes find spurious components that were produced from the correlated noise when three or four components are estimated.

Figure S.2: Species 1-15 and 22-36 are included in the leaf dataset. Species 8 corresponds to *Neurium oleander* (blue dots in Figure 4 and Supplemental Figures 4 and 5); species 31 and 34 correspond to *Podocarpus sp.* and *Pseudosasa japonica* (green dots in Figure 4 and Supplemental Figures S.4 and S.5).

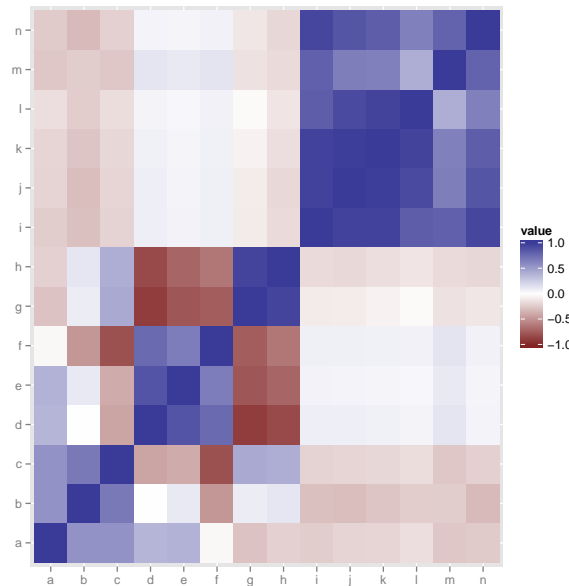


F Supplemental materials for correlated multivariate data

Covariates were generated from photographs of leaf samples from thirty species (Figure S.2). Many of these covariates are highly correlated (Figure S.3).

Logis-LCA and Spline-LCA reveal features in the data (Figures S.4, S.5), while PCA+Infom and PCA+ProDenICA simply rotate the principal components. Additionally, when five components are estimated using the LCA methods, the first two components are nearly equivalent to the components obtained from $Q^* = 2$. This is not the case with the PCA+ICA methods. Thus, the components in LCA appear less sensitive to the number of estimated components

Figure S.3: Correlation matrix of the variables in the leaf dataset: a) eccentricity, b) aspect ratio, c) elongation, d) solidity, e) stochastic convexity, f) isoperimetric factor, g) maximal indentation depth, h) lobedness, i) average intensity, j) average contrast, k) smoothness, l) third moment, m) uniformity, and n) entropy.



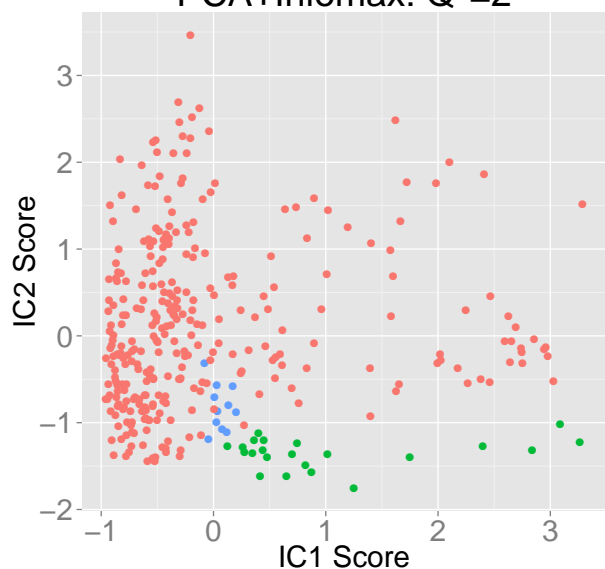
than the components from PCA+ICA methods.

G Supplemental materials for the fMRI analysis

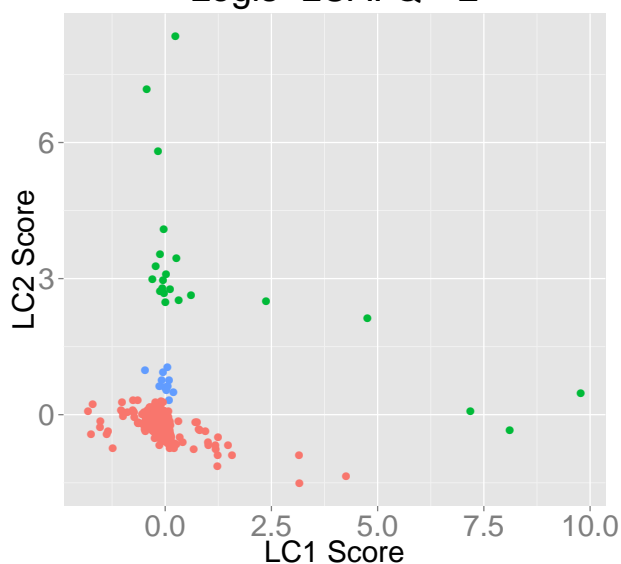
We analyzed the following subjects from the HCP 900-subject release dataset: 100206, 100307, 100408, 100610, 101006, 101107, 101309, 101410, 101915, 102008, and 103414. Whole-brain data were acquired from two sessions with 274 volumes each using gradient-echo EPI with multiband acceleration factor equal to eight and 2 x 2 x 2 mm voxels (repetition time (TR) = 720 ms; echo time (TE) = 33.1 ms; flip angle=52°; field of view = 208 x 180 mm (readout x phase-encoding); acquisition matrix = 104 x 90; slice thickness = 2.0 mm). Only the first session was used in our analyses (the session with right-left phase encoding). Inspection revealed that the first two TRs contained BOLD signals that were higher than other time points. Consequently, we removed the first two TRs. After vectorization, the voxels were standardized across time to have mean zero and unit variance.

Figure S.4: Components in the leaf data from PCA+Infomax and Logis-LCA when two components were estimated and when five components were estimated (when five components were estimated, the two components with the highest marginal likelihood are plotted). The green dots correspond to *Podocarpus sp.* and *Pseudosasa japonica*; the blue dots correspond to *Neurium oleander*; the red dots correspond to all other species.

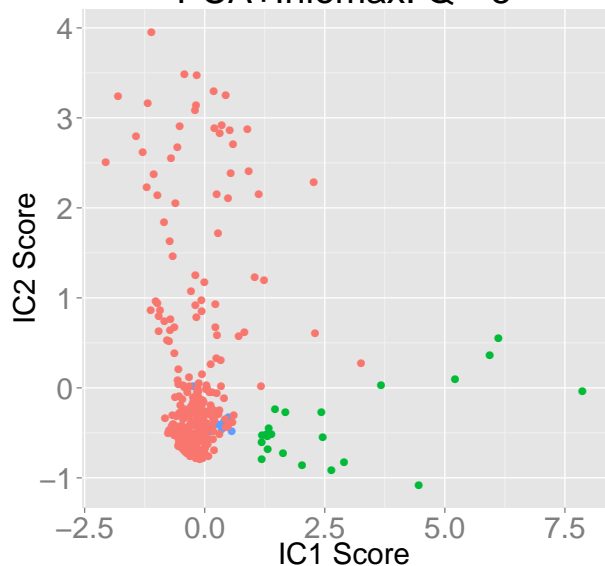
PCA+Infomax: $Q^*=2$



Logis-LCA: $Q^*=2$



PCA+Infomax: $Q^*=5$



Logis-LCA: $Q^*=5$

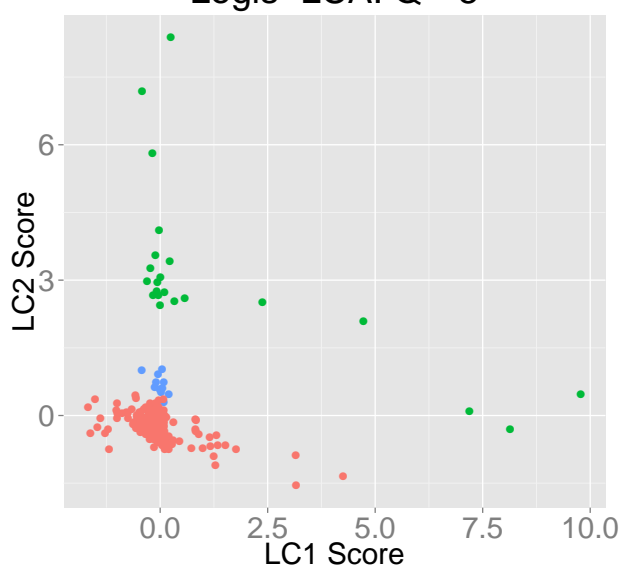
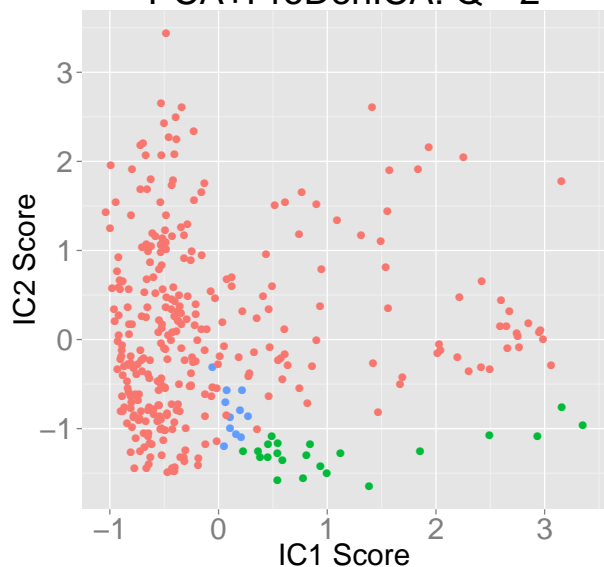
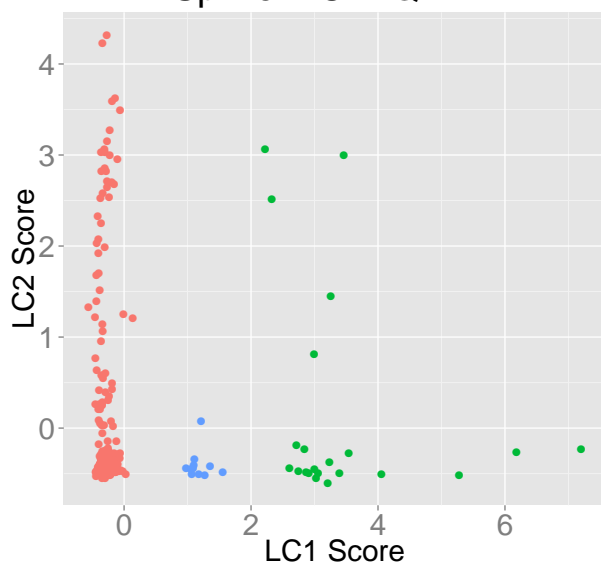


Figure S.5: Components in the leaf data from PCA-ProDenICA and Spline-LCA when two components were estimated and when five components were estimated (when five components were estimated, the two components with the highest marginal likelihood are plotted). The green dots correspond to *Podocarpus sp.* and *Pseudosasa japonica*; the blue dots correspond to *Neurium oleander*; the red dots correspond to all other species. The plots in the first row also appear in Figure 4 of the main manuscript.

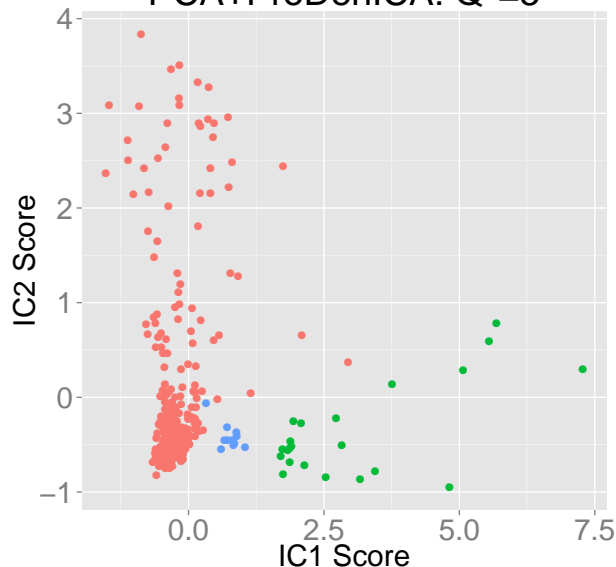
PCA+ProDenICA: $Q^*=2$



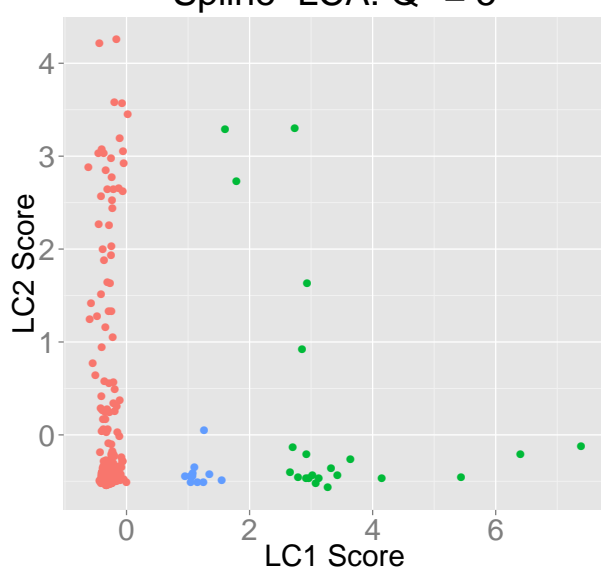
Spline-LCA: $Q^* = 2$



PCA+ProDenICA: $Q^*=5$



Spline-LCA: $Q^* = 5$



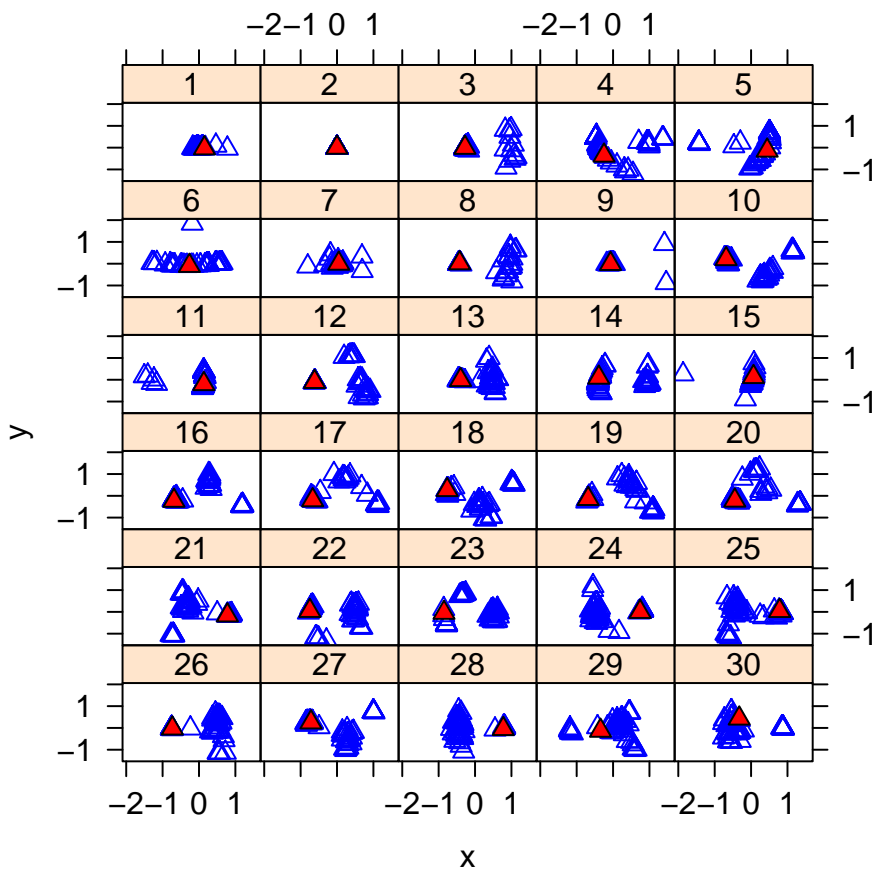


Figure S.6: Multidimensional scaling of $\|\widehat{\mathbf{S}}_j^{(k)} - \widehat{\mathbf{S}}_j^{(\ell)}\|_2$ for components $j = 1, \dots, 30$ and initializations $k \neq \ell \in \{1, \dots, 56\}$. The coordinates corresponding to the initialization with the highest likelihood are depicted by solid red triangles. In all instances, the red triangle appears in a cluster of other triangles, indicating agreement between initializations.

For subject 103414, we examined the effect of initialization in detail. Following Risk et al. (2014), we assessed the reliability of individual components by matching components from all other initializations to the components corresponding to the argmax using the modified Hungarian algorithm. We then created dissimilarity matrices for each component based on the MSE and visualized basins of attraction using multidimensional scaling. Generally, there were at least two basins of attraction corresponding to initializations from the principal subspace and initializations from the entire column space (Supplemental Figure S.6). Components one, two, and nine were relatively robust to initialization and contained only one (main) basin of attraction. Note that Figure 5 in the main manuscript depicts components one and two.

Component 25 is a clear example from subject 103414 of a motion artifact identified by Spline-LCA but not PCA+ProDenICA (Figure S.7). Here, the correlation between Spline-LCA and the matched component in PCA+ProDenICA was 0.38. Voxels near the edge of the cortex had high positive values on one side of the brain and negative values on the opposite side, which is typical of motion artifacts.

LCA also identified a type of artifact that did not seem to be found in PCA+ProDenICA. Some components had alternating lines of positive and negative lines, in particular in axial slices through orbitofrontal regions (Figure S.8). The patterns of activation ignored gray and white matter tissue boundaries, which is evidence of an artifact. This type of pattern is described as an “MRI acquisition/reconstruction related artifact” in Salimi-Khorshidi et al. (2014).

Figure S.7: Motion artifact (component 25) identified using Spline-LCA (top) and the matched component from PCA+ProDenICA (bottom; correlation = 0.38) in subject 103414. Note the component exhibited activation near the edge of the brain in the LC but not the IC. Thresholded at $|s_{v,25}| > 2$.

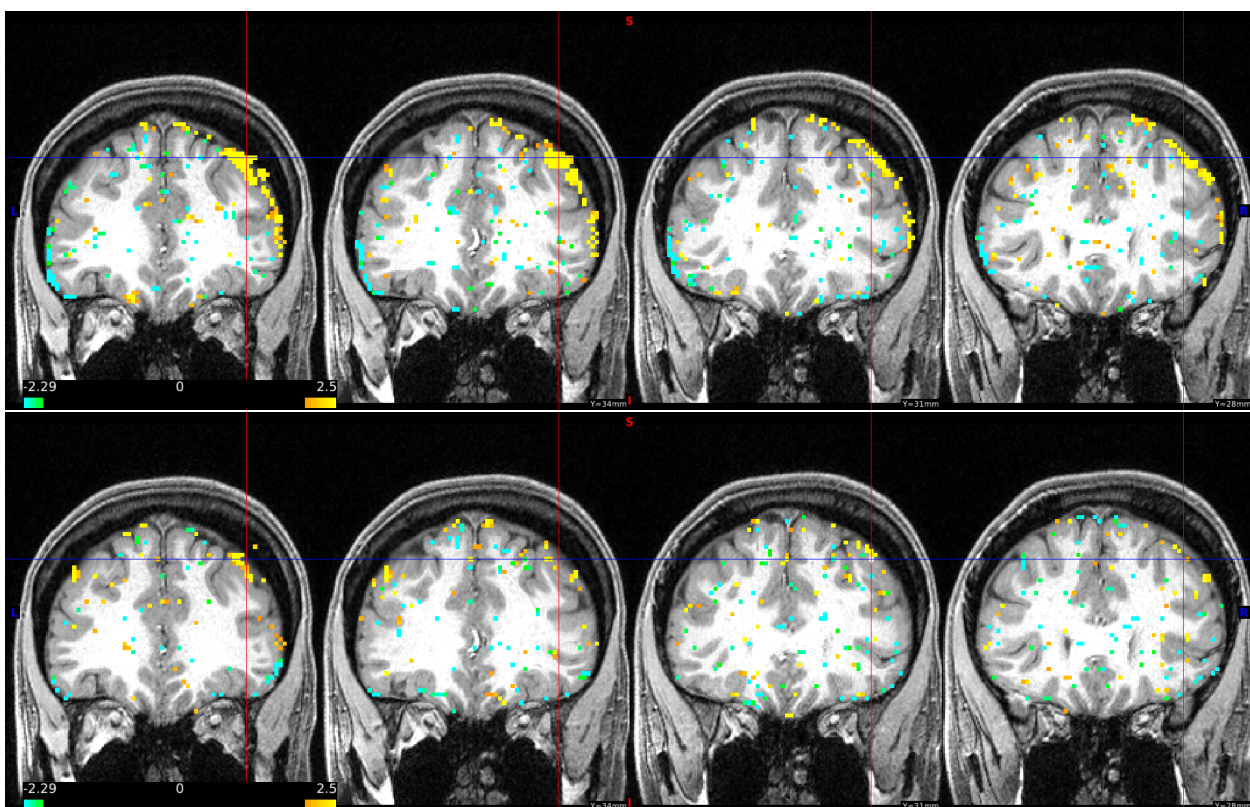
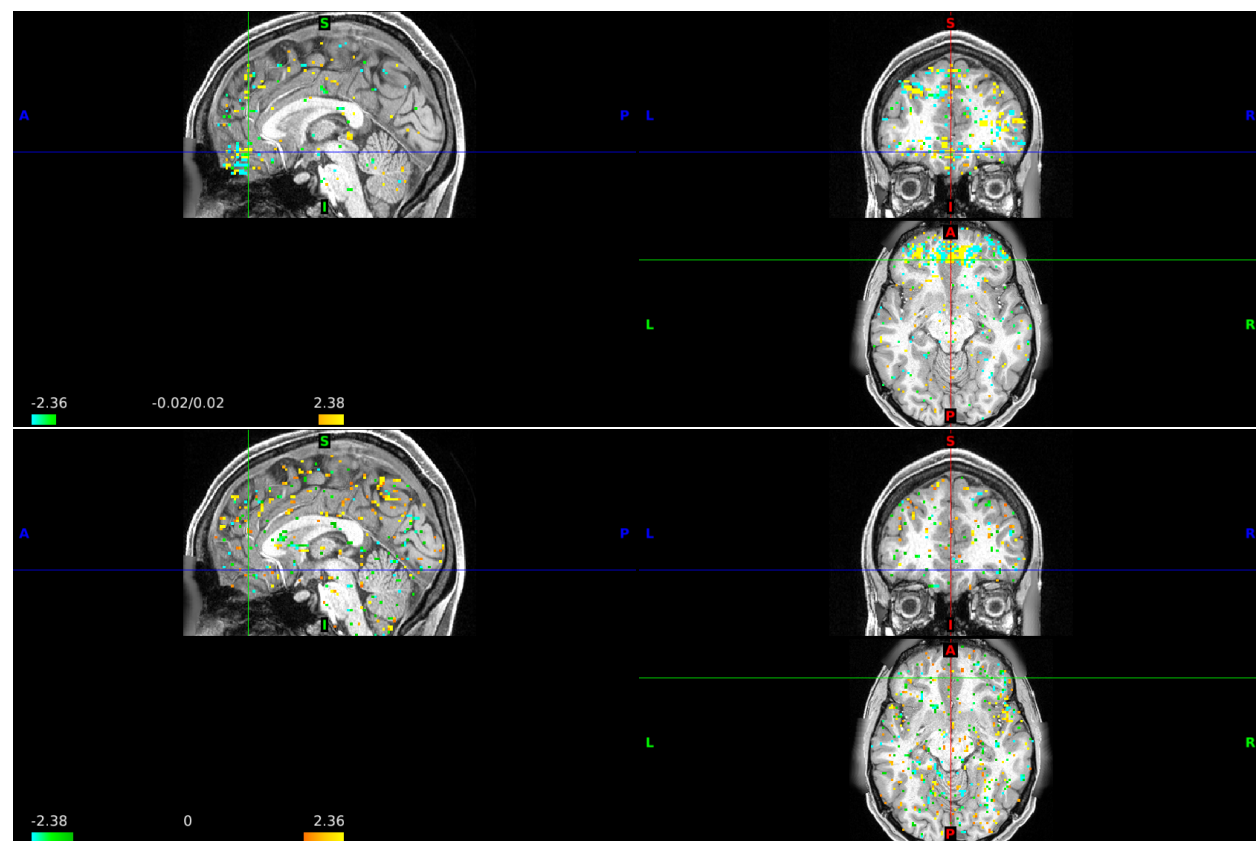


Figure S.8: Artifact (component 14) identified using Spline-LCA (top) and the matched component from PCA+ProDenICA (bottom; correlation = 0.08) in subject 100307. Thresholded at $|s_{v,14}| > 1.75$.



References

- Allasonniere, S. and Younes, L. (2012). A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, 6(1):125–160.
- Amato, U., Antoniadis, A., Samarov, A., and Tsybakov, A. (2010). Noisy independent factor analysis model for density estimation and classification. *Electronic Journal of Statistics*, 4:707–736.
- Ashburner, J., Friston, K., and Penny, W. (2004). Human brain function. *Academic press*, 1:2.
- Attias, H. (1999). Independent factor analysis. *Neural computation*, 11(4):803–851.
- Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 80:169–189.

- Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face recognition by independent component analysis. *Neural Networks, IEEE Transactions on*, 13(6):1450–1464.
- Beckmann, C. F. (2012). Modelling with independent components. *NeuroImage*, 62(2):891–901.
- Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2):137–152.
- Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.
- Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739.
- Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., and Müller, K.-R. (2006). In search of non-Gaussian components of a high-dimensional distribution. *The Journal of Machine Learning Research*, 7:247–282.
- Calhoun, V. D. and Adali, T. (2006). Unmixing fMRI with independent component analysis. *Engineering in Medicine and Biology Magazine, IEEE*, 25(2):79–90.
- Cardoso, J. F. and Soughoumiac, A. (1993). Blind beamforming for non-Gaussian signals. In *Radar and Signal Processing, IEEE Proceedings F*, volume 140, pages 362–370.
- Castelli, F., Happé, F., Frith, U., and Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage*, 12(3):314–325.
- Chen, A. and Bickel, P. J. (2006). Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825–2855.
- Correa, N., Adali, T., and Calhoun, V. D. (2007). Performance of blind source separation algorithms for fMRI analysis using a group ICA method. *Magnetic Resonance Imaging*, 25(5):684–694.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, New Jersey.
- Eloyan, A. and Ghosh, S. K. (2013). A semiparametric approach to source separation using independent component analysis. *Computational Statistics and Data Analysis*, 58:383 – 396.
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magnetic Resonance in Medicine*, 35(3):346–355.

- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.
- Green, C. G., Nandy, R. R., and Cordes, D. (2002). PCA-preprocessing of fMRI data adversely affects the results of ICA. In *Proceedings of International Society of Magnetic Resonance in Medicine*, page 10.
- Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., et al. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, 95:232–247.
- Guo, Y. and Tang, L. (2013). A hierarchical model for probabilistic independent component analysis of multi-subject fMRI studies. *Biometrics*, 69(4):970–981.
- Hastie, T. (2013). *GAM: Generalized Additive Models*. R package version 1.08.
- Hastie, T. and Tibshirani, R. (2003). Independent components analysis through product density estimation. *Advances in Neural Information Processing Systems*, 15:649–656.
- Hastie, T. and Tibshirani, R. (2010). *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*. R package version 1.0.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, pages 435–475.
- Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. Wiley-Interscience.
- Hyvärinen, A. and Oja, E. (1998). Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313.
- Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430.
- 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*, pages 229–236.
- Kagan, A. M., Rao, C. R., and Linnik, Y. V. (1973). *Characterization Problems in Mathematical Statistics*. Wiley.

- Kawanabe, M., Sugiyama, M., Blanchard, G., and Müller, K. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1):57–75.
- Lee, S., Shen, H., Truong, Y., Lewis, M., and Huang, X. (2011). Independent component analysis involving autocorrelated sources with an application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 106(495):1009–1024.
- Lee, T. W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441.
- Marchini, J. L., Heaton, C., and Ripley, B. D. (2010). *FastICA: FastICA Algorithms to perform ICA and Projection Pursuit*. R package version 1.1-13.
- Matteson, D. S. and Tsay, R. S. (2016). Independent component analysis via distance covariance. *Journal of the American Statistical Association*, in press.
- Miettinen, J., Nordhausen, K., Oja, H., and Taskinen, S. (2014). Deflation-based FastICA with adaptive choices of nonlinearities. *IEEE Transactions on Signal Processing*, 62(21):5716–5724.
- Ollila, E. (2010). The deflation-based FastICA estimator: statistical analysis revisited. *Signal Processing, IEEE Transactions on*, 58(3):1527–1541.
- Pollard, D. (2001). Chapter 13 from Asymptopia work-in-progress.
- Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., and Beckmann, C. F. (2015). ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, 112:267–277.
- Risk, B. B., Matteson, D. S., Ruppert, D., Eloyan, A., and Caffo, B. S. (2014). An evaluation of independent component analyses with an application to resting-state fMRI. *Biometrics*, 70(1):224–236.
- Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468.
- Samworth, R. J. and Yuan, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973–3002.
- Shi, R. and Guo, Y. (2016). Investigating differences in brain functional networks using hierarchical covariate-adjusted independent component analysis. *Annals of Applied Statistics*, in press.
- Silva, P. F., Marcal, A. R., and da Silva, R. M. A. (2013). Evaluation of features for leaf discrimination. *Springer Lecture Notes in Computer Science*, Vol. 7950(197-204).

- 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.
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.
- van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.
- Virta, J., Nordhausen, K., and Oja, H. (2015). Joint use of third and fourth cumulants in independent component analysis. *arXiv preprint arXiv:1505.02613*.
- Wei, T. (2015). A convergence and asymptotic analysis of the generalized symmetric FastICA algorithm. *IEEE Transactions on Signal Processing*, 63(24):6445–6458.
- Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., and Rosseel, Y. (2011). neuRosim: An R package for generating fMRI data. *Journal of Statistical Software*, 44(10):1–18.