Linear Non-Gaussian Component Analysis via Maximum Likelihood

Benjamin B. Risk[1,2,3], David S. Matteson[1], David Ruppert[1]

[1]Department of Statistical Science, Cornell University

[2]SAMSI, Research Triangle Park, North Carolina and the
Department of Biostatistics, University of North Carolina, Chapel Hill

[3]Current Address: Department of Biostatistics and Bioinformatics, Emory University

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 are 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 $\{\boldsymbol{x}_i\}$, $i = 1, \ldots, n$, with corresponding latent $\{\boldsymbol{s}_i\}$. The goal is to estimate $\mathbf{M}$ and $\{\boldsymbol{s}_i\}$. 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, $\{\boldsymbol{s}_i\}$ are correctly recovered if they maximize some measure of non-Gaussianity. Moment or cumulant-based methods (Cardoso and Souloumiac, 1993; Virta et al., 2015), kernel methods (Bach and Jordan, 2003), 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, dimension reduction using PCA is applied to the observations $\{\boldsymbol{x}_i\}$ 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), where here a brain network is a set of locations that exhibit similar temporal behavior. However, 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 certain latent components (LCs; see below) improves estimation.

We propose linear non-Gaussian component analysis (LNGCA). Consider a sample $\{\boldsymbol{x}_i, \boldsymbol{s}_i, \boldsymbol{n}_i\}$, $i = 1, \ldots, n$, of the random variable:

$$\mathbf{X} = \mathbf{M_S}\mathbf{S} + \mathbf{M_N}\mathbf{N} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^T$; $\mathbf{S} \in \mathbb{R}^Q$ is a vector of mutually independent non-Gaussian random variables with $1 \leq Q \leq T$; $\mathbf{M_S} \in \mathbb{R}^{T \times Q}$; $\mathbf{M_N} \in \mathbb{R}^{T \times (T-Q)}$; $\mathbf{M} = [\mathbf{M_S}, \mathbf{M_N}]$ (the concatenation of $\mathbf{M_S}$ and $\mathbf{M_N}$) is full rank; and $\mathbf{N}$ is $(T - Q)$-variate normal. Note that in classic (noise-free) ICA, $\mathbf{S} \in \mathbb{R}^{T-1}$ or $\mathbb{R}^T$ and $\mathbf{N} \in \mathbb{R}^1$ or equals zero. In LNGCA, the dimension of the image of $\mathbf{M_N}$ is $T - Q$, whereas noisy ICA (discussed below) assumes the dimension is T. One observes $\{\boldsymbol{x}_i\}$ while $\{\boldsymbol{s}_i\}$ and $\{\boldsymbol{n}_i\}$ are latent. We assume $\mathrm{E}\,\mathbf{S} = \mathbf{0}$ and $\mathrm{E}\,\mathbf{N} = \mathbf{0}$ in (1), such that it is without loss of generality that we assume $\mathrm{E}\,\mathbf{X} = \mathbf{0}$. In practice, data are centered by their sample mean. Our goal is to estimate $\mathbf{M_S}$ and the realizations $\{\boldsymbol{s}_i\}$ of $\mathbf{S}$, which we call latent components (LCs).

1.1 Motivation for LCA

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. When the motivating scientific problem has a low signal-to-noise ratio, LCA is particularly well-suited to recovering the non-Gaussian signals.

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). Consequently, when the non-Gaussian components in the LNGCA model belong to this class of random variables, the expected values of their marginal likelihoods are larger. Our approach is to constrain the latent distributions to have unit variance, which allows both the marginal likelihoods and $\mathbf{M_S}$ 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 $\mathrm{im}(\mathbf{M_S})$ is T or $T-1$ and $\mathrm{im}(\mathbf{M_N})$ 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 the last component can be determined from the previous components. We ignore this technicality and for clarity, hereafter define classic ICA under the assumption that $\mathbf{M_S}$ is full rank and $\mathbf{M_N} = \mathbf{0}$. The case in which the dimension of $\mathrm{im}(\mathbf{M_N})$ 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 $\mathbf{M_N} = \sigma^2 \mathbf{I}_T$.

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. Allassonniere 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) developed non-parametric density estimators of the component densities in the noisy ICA model but assume $\mathbf{M_S}$ 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 for the case where $Q = T$. One approach to estimating the model in (1) would be to sequentially estimate projection pursuit directions. However, estimates from deflationary fastICA typically have higher asymptotic variance than symmetric fastICA (Miettinen et al., 2017, 2015). 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 uses a likelihood to simultaneously estimate the latent components in the presence of Gaussian noise. See Web Supplement C for additional discussion of these other methods.

In Section 2, we discuss the identifiability of LNGCA and a discrepancy measure to account for unidentifiable signed permutations. In Section 3, we propose parametric LCA.

In Section 4, we propose Spline-LCA where we also estimate the latent densities. In Section 5, we investigate simulations when the observations of the latent variables are iid. In Section 6, we examine model robustness by applying our method to temporally and spatially structured simulated data, and we evaluate the impact of estimating the wrong number of components. In Section 7, we use LCA for data visualization and dimension reduction in multivariate data from leaf characteristics. In Section 8, we estimate brain networks and artifacts from high-resolution fMRI data from the Human Connectome Project. Code implementing our methods and proofs of the theorems appear in the Web Supplement.

2 LNGCA

Throughout this section we assume (for simplicity) all random variables are mean zero. Define the equivalence relation $\mathbf{B} \cong \mathbf{C}$ for matrices $\mathbf{B}$ and $\mathbf{C}$ if $\mathbf{B}$ equals $\mathbf{C}$ up to scaling and permutation of columns. Let "$\overset{d}{=}$" denote equality in distribution. Let $\mathbf{S} = [S_1, \ldots, S_Q]^\top$. We state the assumptions of the LNGCA model below.

Assumption 1. $S_1, \ldots, S_Q$ *are mutually independent, non-Gaussian random variables with* $\mathrm{E}\,\mathbf{S} = \mathbf{0}$ *and* $\mathrm{E}\,\mathbf{S}\mathbf{S}^\top = \mathbf{I}_Q$.

Assumption 2. $\mathrm{rank}([\mathbf{M_S}, \mathbf{M_N}]) = T$

Assumption 3. $\mathbf{N}$ *is* $(T - Q)$*-variate normal with* $\mathrm{E}\,\mathbf{N} = \mathbf{0}$ *and* $\mathrm{E}\,\mathbf{N}\mathbf{N}^\top$ *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* $\mathbf{X} = \mathbf{M_S^*}\mathbf{S^*} + \mathbf{E^*}$ *where* $\mathbf{S^*} \in \mathbb{R}^Q$ *are independent non-Gaussian components and* $\mathbf{E^*}$ *is multivariate normal, we have:* $\mathbf{M_S^*} \cong \mathbf{M_S}$; $\mathbf{S^*} \overset{d}{=} \mathbf{S}$ *up to scaling and permutations;* $\mathbf{M_S}\mathbf{S} \overset{d}{=} \mathbf{M_S^*}\mathbf{S^*}$; *and* $\mathbf{E^*} \overset{d}{=} \mathbf{M_N}\mathbf{N}$.

All proofs appear in Web Supplement A.

From Theorem 1, the signal, $\mathbf{M_S S}$, 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, $\mathbf{M_N N}$, does not have a unique decomposition.

Without loss of generality, we assume that $\mathbf{N}$ is standard multivariate normal. Let $\{f_q\}$ be the true densities of the LCs (the signal components), which are also called the source densities. 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 $\mathbf{\Sigma} = \mathbf{U \Lambda U}^\top$. Let $\mathbf{L} = \mathbf{U \Lambda}^{-1/2} \mathbf{U}^\top$ be a whitening matrix (the covariance matrix of $\mathbf{LX}$ is $\mathbf{I}_T$), and define the unmixing matrix $\mathbf{W} = \mathbf{M}^{-1} \mathbf{L}^{-1}$ 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 $\mathbf{w}_q^\top$ denote the qth 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

$$f_{\mathbf{X}}(\boldsymbol{x}|\mathbf{W}, \mathbf{L}) = \det(\mathbf{L}) \prod_{q=1}^{Q} f_q(\mathbf{w}_q^\top \mathbf{L} \boldsymbol{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{Q+k}^\top \mathbf{L} \boldsymbol{x}). \tag{2}$$

Note that for a density and its corresponding row of the unmixing 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 $f_q^*(\mathbf{w}_q^{*\top} \boldsymbol{x}) = f_q(\mathbf{w}_q^\top \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^T$. 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* $\{f_1, \mathbf{w}_1\}, \ldots, \{f_Q, \mathbf{w}_Q\}$ *are identifiable up to sign and ordering. Note the rows* $\mathbf{w}_{Q+k}$ *for* $k = 1, \ldots, T - Q$ *are not identifiable.*

2.1 Sign- and permutation-invariant discrepancy measure

To accomodate the identifiability limitations, 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). Unlike the Amari or minimum distance (Ilmonen et al., 2010) measures, it applies to non-square unmixing matrices. 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 is also 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 $Q < R$. Let $||\cdot||_F$ denote the Frobenius norm. Define the sign- and permutation-invariant mean-squared error:

$$PMSE(\mathbf{M}_1, \mathbf{M}_2) = \frac{1}{TQ} \operatorname*{argmin}_{\mathbf{P}_\pm \in \mathcal{P}_\pm} ||\mathbf{M}_1 - \mathbf{M}_2 \mathbf{P}_\pm||_F^2, \tag{3}$$

where $\mathbf{P}_\pm$ 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 (3) 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. Also define $PRMSE = \sqrt{PMSE}$, i.e., permutation-invariant root mean squared error.

3 Parametric LCA

Now let $\{\boldsymbol{x}_i\}$ be an iid sample of $\mathbf{X}$, and let $\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$. Assume $n > T$. Let $\widehat{\boldsymbol{\Sigma}}$ be the sample covariance matrix of $\{\boldsymbol{x}_i\}$, with divisor n, not $n-1$. Consider its eigenvalue decomposition, $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^\top$. Then define $\widehat{\mathbf{L}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}^{-1/2}\widehat{\mathbf{U}}^\top$. Let $\mathbf{o}_q^\top$ be the qth row of an orthogonal matrix $\mathbf{O}$. Note that $\sum_{i=1}^n \mathbf{o}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = 0$ and $\sum_{i=1}^n \log \phi \left(\mathbf{o}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right) = -\frac{n}{2}(\log 2\pi + 1)$. Let $\mathcal{O}_{Q \times T}$ be the class of $Q \times T$ semi-orthogonal matrices, which is the Stiefel

Manifold. Let $p_q(x)$ denote a density used in the objective function (possibly mis-specified):

$$\mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) = \frac{1}{n}\sum_{i=1}^{n}\sum_{q=1}^{Q}\log p_q\left(\mathbf{o}_q^\top\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right). \tag{4}$$

Let $h(\boldsymbol{s}) = \sum_{q=1}^{Q}\log p_q(s_q)$, where s_q is the qth element of $\boldsymbol{s}$. Let $\|\boldsymbol{s}\|$ denote the Euclidean (ℓ-2) norm. We make additional assumptions:

Assumption 4. *(i)* $p_q(s_q) < \infty$, $q = 1,\ldots,Q$; *(ii) for all* $\boldsymbol{s}_0$ *and* $\boldsymbol{s}_1 \in \mathbb{R}^Q$, *there exist* $M > 0$ *and* $\alpha \geq 0$ *such that*

$$\|h(\boldsymbol{s}_0) - h(\boldsymbol{s}_1)\| \leq M\|\boldsymbol{s}_0 - \boldsymbol{s}_1\|\left\{1 + \|\boldsymbol{s}_0\|^\alpha + \|\boldsymbol{s}_1\|^\alpha\right\}; \tag{5}$$

and (iii) $\mathrm{E}\,\|\mathbf{S}\|^{1+\alpha} < \infty$.

Note that Assumptions 4 (i) and (ii) define conditions for the densities used in the *objective function* rather than the true densities. A discussion of the densities satisfying these assumptions is in Web Supplement A.2. However, we first consider the case when the Q true component densities are known, i.e., $p_q = f_q$:

$$\widehat{\mathbf{W}}_{\mathbf{S}}^{Or} = \underset{\mathbf{O_S}\in\mathcal{O}_{Q\times T}}{\operatorname{argmax}}\ \sum_{i=1}^{n}\sum_{q=1}^{Q}\log f_q\left(\mathbf{o}_q^\top\widehat{\mathbf{L}}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\right), \tag{6}$$

so that $\widehat{\mathbf{W}}_{\mathbf{S}}^{Or}$ is an oracle (Or) estimator that cannot be used in practice.

Observe that estimating $\mathbf{W_S}$ is equivalent to estimating the LCs because $\hat{\boldsymbol{s}}_i = \widehat{\mathbf{W}}_{\mathbf{S}}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ for all v. Thus we would like a consistent estimator of $\mathbf{W_S}$.

Theorem 2. *Suppose* $\mathbf{X}$ *follows the LNGCA model in* (1) *with Assumptions 1-4. Given an iid sample* $\{\boldsymbol{x}_i\}$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Or} \xrightarrow{a.s.} \mathbf{W_S}$ *on the equivalence class of signed permutations.*

Consider the special case in which p_q in (4) equals the logistic density for all q, hereafter Logis-LCA. The 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.

9

Infomax is popular in fMRI analysis, where it outperforms FastICA and JADE (Correa et al., 2007; Calhoun and Adali, 2006). We define our estimator for some $Q^* \leq T$ such that Q^* may or may not equal Q. After simplifications, the Logis-LCA estimator of $\mathbf{W_S}$ is defined

$$\widehat{\mathbf{W}}_{\mathbf{S}}^{Logis} = \underset{\mathbf{O_S} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} - \sum_{i=1}^{n} \sum_{q=1}^{Q^*} \log \left\{ 1 + \exp\left(-\mathbf{o}_q^\top \widehat{\mathbf{L}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \frac{\pi}{\sqrt{3}} \right) \right\}. \tag{7}$$

We maximize (7) using a modification of the symmetric fixed-point ICA algorithm (Hyvarinen, 1999) discussed in Web Supplement D.

Next, we define sufficient conditions that characterize the extent to which the densities used in the estimator can mismatch the true densities while maintaining consistency. Let $r_q(s)$ and $r_q'(s)$ denote the first and second derivatives of $\log p_q(s)$. Note that $\mathrm{E}\, r_q(S_q) = \int r_q(s) f_q(s) \, ds$ since f_q is the true density.

Assumption 5. *For all q, (i) $\mathrm{E}\, r_q'(S_q) - \mathrm{E}\, S_q\, r_q(S_q) < 0$; (ii) $\mathrm{E}\, r_q'(S_q)$, $\mathrm{E}\, S_q\, r_q(S_q)$, and $\mathrm{E}\, S_q^2 r_q'(S_q)$ are finite; (iii) $\log p_q(s)$ is twice continuously differentiable on the support of S_q; (iv) $\frac{\partial}{\partial o_{qt}} \mathrm{E} \log p_q(\mathbf{o}_q^\top \mathbf{X}) = \mathrm{E}\, \frac{\partial}{\partial o_{qt}} \log p_q(\mathbf{o}_q^\top \mathbf{X})$ and $\frac{\partial}{\partial o_{qt}} \mathrm{E}\, X_t r_q(\mathbf{o}_q^\top \mathbf{X}) = \mathrm{E}\, \frac{\partial}{\partial o_{qt}} X_t r_q(\mathbf{o}_q^\top \mathbf{X})$.*

The interesting assumption here is 5(i), which defines the mis-match criterion. We can check this assumption for a proposed objective function density and a set of hypothetical source densities to gain insight into the robustness of the proposed estimator, which will be done in Section 5. Note the differentiability assumption is for the proposed densities and does not need to hold for the true densities. Now consider compact neighborhoods of $\mathbf{W_S}$ of the form $\mathcal{N}_\epsilon(\mathbf{W_S}) = \{\mathbf{O_S} \in \mathcal{O}_{Q \times T} : ||\mathbf{O_S} - \mathbf{W_S}||_F \leq \epsilon\}$. Note that in place of $\mathbf{W_S}$, we could define this neighborhood for any other $\mathbf{W_S^*} \cong \mathbf{W_S}$ (here the equivalence class is defined for sign changes and permutations of the rows of $\mathbf{W_S}$), which is useful when using a preliminary estimator described below. Let $p(\boldsymbol{x}) = \prod_{q=1}^{Q} p_q(x_q)$.

Proposition 1. *Suppose Assumptions 1-5. There exists $\mathcal{N}_{\epsilon^*}(\mathbf{W_S})$ such that $\mathrm{E} \log p(\mathbf{O_S L X})$ constrained to $\mathbf{O_S} \in \mathcal{N}_{\epsilon^*}(\mathbf{W_S})$ is maximized at $\mathbf{W_S}$.*

Restricting the optimization space is necessary because for many source distributions, the population objective function can contain multiple maxima. In fact, when the wrong density is used in the objective function, the global maximum can correspond to the wrong unmixing matrix (Risk et al., 2014). This notion of localness corresponds to the definition of the theoretical fastICA estimator found in Hyvarinen (1999), Hyvärinen and Oja (1998), and Wei (2015). Formally, define

$$\widehat{\mathbf{W}}_{\mathbf{S}}^{Local} = \underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{N}_{\epsilon^*}(\mathbf{W}_{\mathbf{S}})}{\operatorname{argmax}} \quad \mathcal{J}_n(\mathbf{O}_{\mathbf{S}} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})). \tag{8}$$

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

Theorem 3. *Suppose* $\mathbf{X}$ *follows the LNGCA model in* (1) *with Assumptions 1-5. Given an iid sample* $\{\boldsymbol{x}_i\}$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Local} \xrightarrow{a.s.} \mathbf{W}_{\mathbf{S}}$ *on the equivalence class of signed permutations.*

Under additional assumptions (Assumption 6 in Web Supplement B) and using the methods from Nordhausen et al. (2011), Miettinen et al. (2015), Miettinen et al. (2017), and Virta et al. (2016), we derive $\sqrt{n}$-consistency, asymptotic normality, and the asymptotic variances, which appear as Theorem 4 and Corollary 2 in the Web Supplement B. We also conducted simulations validating the asymptotics on finite samples; see Figure S.1.

We can replace the condition that optimization is over $\mathcal{N}_{\epsilon^*}(\mathbf{W}_{\mathbf{S}})$ with a two-stage estimator in which in the first stage, we use an estimator that is consistent on $\mathcal{O}_{Q \times T}$ and in the second stage, we use an estimator that may improve upon the initial consistent estimate. Virta et al. (2016) propose an estimator for the LNGCA model based on a mixture of squared third and fourth moments in which the global maximum on $\mathcal{O}_{Q \times T}$ is $\sqrt{n}$-consistent under finite eighth moment assumptions. Define the Local+Virta estimator, $\widehat{\mathbf{W}}_{\mathbf{S}}^{LV}$, in which the symmetric estimator from Virta et al. (2016) is updated with a single iteration of the symmetric fixed point algorithm (Algorithm 2 in Web Supplement D, which is an approximate Newton iteration, Hyvarinen 1999) defined for the objective function in (8). Then under the additional moment assumptions, one can obtain an estimator with the wrong likelihood that

11

is consistent on $\mathcal{O}_{Q \times T}$.

For any LCA estimator $\widehat{\mathbf{W}}_{\mathbf{S}}$ and $\hat{\boldsymbol{s}}_i = \widehat{\mathbf{W}}_{\mathbf{S}} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, we also define an estimator of $\mathbf{M}_{\mathbf{S}}$:

$$\widehat{\mathbf{M}}_{\mathbf{S}} = \underset{\mathbf{A} \in \mathbb{R}^{T \times Q}}{\operatorname{argmin}} \sum_{i=1}^{n} \| \boldsymbol{x}_i - \bar{\boldsymbol{x}} - \mathbf{A} \hat{\boldsymbol{s}}_i \|_2^2. \qquad (9)$$

This is the OLS solution, which here is equivalent to $\widehat{\mathbf{M}}_{\mathbf{S}} = \widehat{\mathbf{L}}^{-1} \widehat{\mathbf{W}}_{\mathbf{S}}^{\top}$. Although we assume iid observations in the construction of (6), 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 choose the sign on each row $\widehat{\mathbf{w}}_q$ and corresponding $\{\hat{s}_{iq}\}$ such that $\sum_{i=1}^{n} \hat{s}_{iq}^3 > 0$ for $q = 1, \ldots, Q$. We define the LCA criteria for ordering LCs for a sample $\{\boldsymbol{x}_i\}$:

$$\sum_{i=1}^{n} \log f_1 \left\{ \widehat{\mathbf{w}}_1^{\top} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\} > \sum_{i=1}^{n} \log f_2 \left\{ \widehat{\mathbf{w}}_2^{\top} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\} > \cdots > \sum_{i=1}^{n} \log f_Q \left\{ \widehat{\mathbf{w}}_Q^{\top} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\}. \qquad (10)$$

If we include the Gaussian noise components, then the population analogue of (10), allowing for potentially equal source densities and assuming continuous source densities, is

$$\mathrm{E} \log f_1(\mathbf{w}_1^{\top} \mathbf{L} \mathbf{X}) \geq \cdots \geq \mathrm{E} \log f_Q(\mathbf{w}_Q^{\top} \mathbf{L} \mathbf{X}) > \mathrm{E} \log \phi(\mathbf{w}_{Q+1}^{\top} \mathbf{L} \mathbf{X}) = \cdots = \mathrm{E} \log \phi(\mathbf{w}_T^{\top} \mathbf{L} \mathbf{X}).$$

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

4   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 6 of Web Supplement C.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 $O(n)$, 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 $\mathbf{O} \in \mathcal{O}_{T \times T}$:

$$\ell(\mathbf{O}, g_1, \ldots, g_{Q^*} \mid \widehat{\mathbf{L}}, \bar{\boldsymbol{x}}, Q^*, \{\boldsymbol{x}_i\}) =$$

$$\sum_{i=1}^{n} \left[\sum_{q=1}^{Q^*} \left\{ \log \phi \left(\mathbf{o}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right) + g_q \left(\mathbf{o}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right) \right\} + \sum_{k=1}^{T-Q^*} \log \phi \left(\mathbf{o}_{k+Q^*}^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right) \right].$$

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:

$$\ell_{pen}(\mathbf{O}, g_1, \ldots, g_{Q^*} \mid \widehat{\mathbf{L}}, \bar{\boldsymbol{x}}, Q^*, \{\boldsymbol{x}_i\}) = -\sum_{q=1}^{Q^*} \left\{ \gamma_q \int \{g_q''(u)\}^2 \, du + \int \phi(u)e^{g_q(u)} \, du \right\} \qquad (11)$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \sum_{q=1}^{Q^*} \left\{ \log \phi \left(\mathbf{o}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right) + g_q \left(\mathbf{o}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right) \right\},$$

where we have dropped the noise components since they are constant for all $\mathbf{O}$ but retained the Gaussian contributions to the tilted Gaussian densities, which are not constant when the data are binned as described below. Then we have the following:

Proposition 2. *Let G be the class of all cubic splines $g : \mathbb{R} \to \mathbb{R}$. 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 $\{\hat{f}_q\}$, $q = 1, \ldots, Q^*$, via the fixed point algorithm and estimating $\{f_q\}$ for fixed $\widehat{\mathbf{W}}_{\mathbf{S}}$ using the "Poisson trick". Our account largely follows the description in Hastie et al. (2009) but for semi-orthogonal (rather than orthogonal) matrices.

Suppose $\widehat{\mathbf{W}}_{\mathbf{S}}$ is given and define $s_{vq} = \widehat{\mathbf{w}}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$. Let $u_1^*, \ldots, u_{L+1}^*$ define a discretization, $[u_1^*, u_2^*), [u_2^*, u_3^*), \ldots, [u_L^*, u_{L+1}^*)$, of the support of the tilt function of the non-Gaussian densities such that $\Delta = u_\ell^* - u_{\ell-1}^*$ for all $\ell = 2, \ldots, L+1$. It suffices to take $u_1^* = \min(s_{11}, \ldots, s_{nd}) - 0.1\hat{\sigma}_z$ and $u_{L+1}^* = \max(s_{11}, \ldots, s_{nd}) + 0.1\hat{\sigma}_z$, 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 $q \in \{1, \ldots, Q^*\}$ and $\ell \in \{1, \ldots, L\}$, define $y_{\ell q} = \sum_{i=1}^n \mathbb{1}\{s_{vq} \in [u_\ell^*, u_{\ell+1}^*)\}$.

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

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

for some penalty γ_q. This is proportional to a Poisson generalized additive model (GAM), where $\frac{y_{\ell q}}{n\Delta}$ is the response and the expected response is equal to $\phi(u_\ell) e^{g_q(u_\ell)}$. In practice, we use cubic B-splines in the `gam` package (Hastie, 2013) with `smooth.spline` and the default knot selection, where γ_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 with sample size 1,000. This method also easily scales to tens of thousands of observations. We summarize this procedure in Algorithm 1. 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.

14

Algorithm 1: The Spline-LCA algorithm.

Inputs : The whitened $n \times T$ data matrix $\mathbf{X}_{\mathrm{st}}$; initial $\mathbf{W}_{\mathbf{S}}^0$; tolerance ϵ; and desired effective degrees of freedom.

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

1. Let $(m) = 0$ where (m) denotes the number of update steps. Define $\mathbf{S}^{(m)} = \mathbf{X}_{\mathrm{st}}\mathbf{W}_{\mathbf{S}}^{(m)\top}$.

2. Estimate $\{f_q^{(m+1)}\}$ in which the smoothness penalty is chosen to result in the specified effective df.

3. Update $\mathbf{W}_{\mathbf{S}}^{(m+1)}$ given $f_1^{(m+1)}, \ldots, f_Q^{(m+1)}$ and $\mathbf{S}^{(m)}$ with one step of the symmetric fixed-point algorithm (see Algorithm 2 in Web Supplement D).

4. Let $\mathbf{S}^{(m+1)} = \mathbf{X}_{\mathrm{st}}\mathbf{W}_{\mathbf{S}}^{(m+1)\top}$.

5. If $PMSE(\mathbf{W}_{\mathbf{S}}^{(m+1)\top}, \mathbf{W}_{\mathbf{S}}^{(m)\top}) < \epsilon$, stop, else increment (m) and repeat (2)-(4).

---

5 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 $d_1, \ldots, d_Q$ from the covariance matrix of $\mathbf{M}_{\mathbf{S}}\mathbf{S}$. For LCA, let $d_{\epsilon_1}, \ldots, d_{\epsilon_{T-Q}}$ denote the eigenvalues from the EVD of the covariance matrix of $\mathbf{M}_{\mathbf{N}}\mathbf{N}$. Then, $SNR = \frac{\sum_{q=1}^{Q} d_q}{\sum_{k=1}^{T-Q} d_{\epsilon_k}}$.

For the noisy ICA model, we have T non-zero eigenvalues (equal to σ^2) in the denominator.

We fit D-FastICA using a modification of 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 (9). 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). See Web Supplement C.2.

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 $\mathbf{M_S}\mathbf{S}$ where $\mathbf{M_S}$ 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 differed 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 (volumetric pixels) 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 two $10 \times 10 \times 10$ images: in the first component, activation was represented by a sphere of radius two voxels centered at $(5, 5, 5)$ with voxel-value equal to one and exponential decay rate equal to 0.5; in the second, 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 $n = 1{,}000$ 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 in the principal subspace. Let $\widehat{\mathbf{U}}_{1:Q}^{\top}$ denote the first Q rows from $\widehat{\mathbf{U}}^{\top}$ in the decomposition $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^{\top}$. Then $\mathbf{W_S^0} = \mathbf{O}\widehat{\mathbf{U}}_{1:Q}^{\top}$ for $\mathbf{O} \in \mathcal{O}_{Q \times Q}$ produces a semiorthogonal matrix in the principal subspace, which may help convergence for large SNRs. 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 $\mathbf{W_S}$ (Web Supplement C.2).

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 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 $\widehat{\mathbf{M}_{\mathbf{S}}}$ 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

18

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(i) for Logis-LCA and all values were negative except for the sub-Gaussian mixture of normals; thus the results for $n = 1000$ are in general agreement with Theorem 3. We also evaluated the mis-match criterion between densities (a)-(e) and Spline-LCA densities estimated from a sample from the true densities, and all values were negative.

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 C.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.

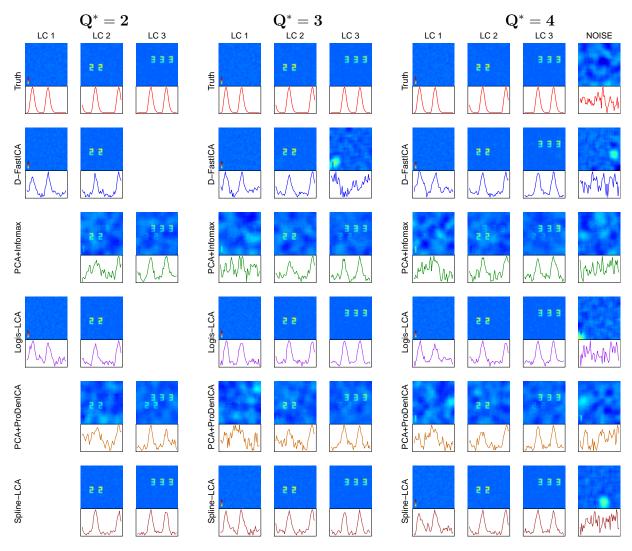# 6 Simulations: Spatio-temporal Signals and $Q^* \neq Q$

Next, we examine the ability of D-FastICA, PCA+Infomax, Logis-LCA, and Spline-LCA to recover simulated spatially structured signals (i.e., sources) whose loadings vary deterministically with time in the presence of spatially and temporally correlated noise. Each spatial source is similar to a brain "network" (or a resting-state network) in fMRI, in which

a set of active locations share the same temporal behavior and each location corresponds to the index i. We also examine the effect of using $Q^* \neq Q$ on source 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. We simulated three sources mixed across fifty time units. The sources were 33×33 images corresponding to $n = 1089$. "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 qth column of $\mathbf{M_S}$. To simulate the temporal activation patterns of brain networks, we used neuRosim (Welvaert et al., 2011) 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.

In the LNGCA scenario, noise components were generated as forty-seven independent 33×33 Gaussian random fields with FWHM=6. 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 8. Additionally, noise components were scaled such that the SNR was 0.4, which approximately equals the SNR estimated in Section 8. 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 $t = 2, \ldots, 50$ 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 5.

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

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 $Q^* = 4$ (Figure S.3), 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
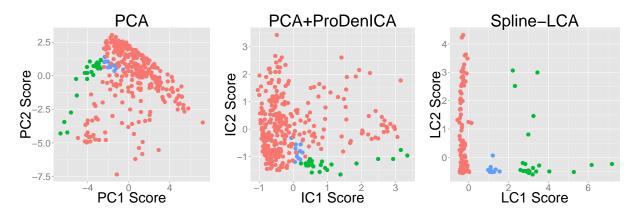
and Spline-LCA recovered the components in all simulations (Figure S.3). It is notable that estimates from PCA+Infomax, PCA+ProDenICA, and D-FastICA were sensitive to the choice of Q^*, whereas Logis-LCA and Spline-LCA were robust (Figures 2, S.3).

For the noisy-ICA scenario, the features recovered by Logis-LCA most closely resembled the truth (Figure S.2) and Logis-LCA generally outperformed other methods (Figure S.3). 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 sources. 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 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 (Figure S.3).

7   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., 2013). Fourteen variables were generated from eight to sixteen images of leaves from each of thirty species (Figure S.4). Many of the covariates are highly correlated (Figure S.5). 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 3). 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 3, Figure S.6). In contrast, Spline-LCA clearly reveals three clusters, where the green dots correspond to two

Figure 3: 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.

plant species that have very thin leaves (species 31 and 34 in Figure S.4), 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 (Figure S.6), although the separation is less than in Spline-LCA. A referee remarked that the NGCA model is not true here, in which the data are a mixture of thirty fourteen-variate distributions corresponding to the thirty species. We agree, but the goal here is to identify useful features. We find the model useful to that end.

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 (Figures S.6 and S.7). 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.
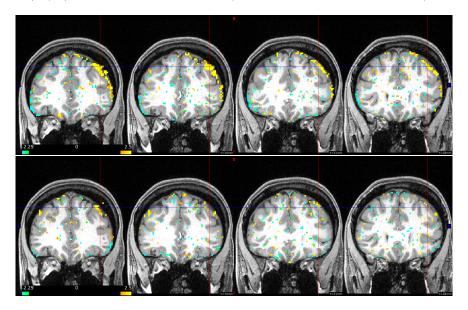
8 Application to fMRI

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 H. 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). We used the minimally preprocessed data from the *fMRIVolume* pipeline (Glasser et al., 2013). The preprocessing pipeline includes rigid-body motion correction of all volumes to a subject's reference image. 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 × 272 data matrix. Each voxel was treated as a replicate with $i = 1, \ldots, n$ for $n = 230,459$, 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).

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 as described in Web Supplement H. Initiating the algorithm from fifty-six matrices resulted in multiple initializations converging to the same estimate of the argmax (Figure S.8). We also completed an analogous PCA+ProDenICA with thirty components using the R package ProDenICA (Hastie and Tibshirani, 2010).

In all subjects, a component highly correlated with the task was found in both Spline-LCA and PCA+ProDenICA, but a number of other components were only detected in Spline-LCA. We discuss the biological interpretation of the task-related network in the Web

Figure 4: 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 $|s_{v,25}| > 2$; yellow indicates $s_{v,25} > 2$ and blue indicates $s_{v,25} < -2$.

Supplement H, and here focus on the application to artifact detection. 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. In one example from subject 103414, LC 25 exhibited activation at the edges of the brain, which is typical of motion artifacts (Salimi-Khorshidi et al., 2014). This artifact was not evident in the matched component from PCA+ProDenICA (Figure 4). Additionally, component two was not correlated with any of the components in PCA+ProDenICA. It exhibited activation in the brainstem and near the edges of the brain, and may correspond to other sources of motion and noise (Figure S.9; Web Supplement H). There were also artifacts that exhibited alternating patterns of positive and negative activation (Figure S.10), which may be due to scanner acquisition and/or air-tissue boundaries (e.g., Figure 6 in Salimi-Khorshidi et al. 2014), and these components were not found in PCA+ProDenICA. Our results suggests that LCA may improve artifact detection.

9 Discussion

We propose a new 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, and artifact detection. 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 performs similarly to LCA for the noisy ICA model. 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 $n =$230,459) and identified artifacts not extracted by PCA+ICA.

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 artifacts 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 4, Figures S.9 and S.10) suggests LCA could be used for more powerful denoising methods over traditional PCA+ICA.

An important advantage of LCA over existing frameworks is its robustness to misspecifi-

cation of the number of estimated components, and future research should examine methods to select Q^*. In contrast to LCA, noisy ICA is sensitive to the choice of Q^* (Section 6, see also Allassonniere 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 brain network into multiple brain networks. However, our simulations suggest that using too few components leads to inappropriately aggregated sources in PCA+ICA methods (Figures 2 and S.2). In contrast, the components recovered for $Q^* \neq Q$ in Logis-LCA across model scenarios and Spline-LCA for the LNGCA scenario accurately represent the spatial features. 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 (Figures S.6 and S.7). To determine Q^* in LNGCA, Virta et al. (2016) suggest the sequential use of the Jarque-Bera test of normality. Nordhausen et al. (2016) develop asymptotic and bootstrap tests of dimensionality using first-order blind identification (FOBI). The use of these criteria in fMRI and other applications is a direction for future research.

10   Acknowledgments

References

Allassonniere, 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.

Attias, H. (1999). Independent factor analysis. *Neural computation*, 11(4):803–851.

Bach, F. R. and Jordan, M. I. (2003). Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48.

Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 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.

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 Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. In *Radar and Signal Processing, IEEE Proceedings F*, volume 140, pages 362–370.

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, 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.

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., and Virta, J. (2017). The squared symmetric fastica estimator. *Signal Processing*, 131:402–411.

Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H., et al. (2015). Fourth moments and independent component analysis. *Statistical science*, 30(3):372–390.

Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., and Ollila, E. (2011). Deflation-based fastica reloaded. In *Signal Processing Conference, 2011 19th European*, pages 1854–1858. IEEE.

Nordhausen, K., Oja, H., and Tyler, D. E. (2016). Asymptotic and bootstrap tests for subspace dimension. *arXiv preprint arXiv:1611.04908*.

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.

Virta, J., Nordhausen, K., and Oja, H. (2015). Joint use of third and fourth cumulants in independent component analysis. *arXiv preprint arXiv:1505.02613*.

Virta, J., Nordhausen, K., and Oja, H. (2016). Projection pursuit for non-gaussian independent components. *arXiv preprint arXiv:1612.05445*.

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.

Supplement to "Linear Non-Gaussian Component Analysis via Maximum Likelihood"

Benjamin B. Risk, David S. Matteson, David Ruppert

A Proofs

A.1 Proofs for Section 2

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 $\mathbf{B} \ncong \mathbf{C}$. Let $\overset{d}{=}$ 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 $\mathbf{Z}_1 \overset{d}{=} \mathbf{X}_1$ and $\mathbf{Z}_2 \overset{d}{=} \mathbf{X}_2$ up to scaling and permutations.*

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 $\mathbf{Z} \overset{d}{=} \mathbf{X}$. Consider their structural representations: $\mathbf{Z} = \mathbf{M_S S}$ and $\mathbf{X} = \mathbf{M_S^* S^*}$ where $\mathbf{M_S} \in \mathbb{R}^{T \times Q}$ and $\mathbf{M_S^*} \in$*

$\mathbb{R}^{T \times Q}$ *for* $Q \leq T$, *and* $\mathrm{rank}(\mathbf{M_S}) = \mathrm{rank}(\mathbf{M_S^*}) = Q$. *Then* $\mathbf{M_S} \cong \mathbf{M_S^*}$ *and* $\mathbf{S} \stackrel{d}{=} \mathbf{S}^*$ *up to scaling and permutations.*

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

$$(\mathbf{M_S}^\top \mathbf{M_S})^{-1} \mathbf{M_S}^\top \mathbf{M_S S} = (\mathbf{M_S}^\top \mathbf{M_S})^{-1} \mathbf{M_S}^\top \mathbf{M_S^* S^*}.$$

Letting $\mathbf{B} = (\mathbf{M_S}^\top \mathbf{M_S})^{-1} \mathbf{M_S}^\top \mathbf{M_S^*}$, we have $\mathbf{S} \stackrel{d}{=} \mathbf{B S^*}$. 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, $\mathbf{S} = \mathbf{IS}$. 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 $\mathbf{BR}$ 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 $(\mathbf{M_S}^\top \mathbf{M_S})^{-1} \mathbf{M_S}^\top \mathbf{M_S^*} = \mathbf{I}$. Now since $(\mathbf{M_S}^\top \mathbf{M_S})^{-1} \mathbf{M_S}^\top$ is full row rank, it has a unique right inverse equal to the Moore-Penrose pseudoinverse, which is equal to $\mathbf{M_S}$, which implies $\mathbf{M_S} = \mathbf{M_S^*}$. For $\mathbf{B} \cong \mathbf{I}$, it follows that $\mathbf{M_S^*} \cong \mathbf{M_S}$. □

We now prove Theorem 1.

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

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

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 $\{f_1, \mathbf{w}_1\}, \ldots, \{f_Q, \mathbf{w}_Q\}$ are identifiable up to sign and ordering. Note the rows $\mathbf{w}_{Q+k}$ for $k = 1, \ldots, T - Q$ are not identifiable.*

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

$$|\det(\mathbf{L})| \prod_{q=1}^{Q} f_q\left(\mathbf{w}_q^\top \mathbf{L} \boldsymbol{x}\right) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{Q+k}^\top \mathbf{L} \boldsymbol{x}) = |\det(\mathbf{C})| \prod_{\ell=1}^{T} g_\ell(\boldsymbol{c}_\ell^\top \boldsymbol{x}) \tag{S.1}$$

then Q of the marginal densities $g_1, \ldots, g_T$ are equivalent up to sign to $f_1, \ldots, f_Q$, where densities $g(x)$ and $f(x)$ are equivalent up to sign if they are equal or if $g(x) = f(-x)$ for all x on $\mathbb{R}$, and that each of the corresponding Q rows of $\mathbf{C}$ equal $\mathbf{w}_1^\top \mathbf{L}, \ldots, \mathbf{w}_Q^\top \mathbf{L}$. Using a change of variable $\mathbf{Z} = \mathbf{L} \mathbf{X}$, we consider the model $\mathbf{Z} = \mathbf{A_S} \mathbf{S} + \mathbf{A_N} \mathbf{N}$, such that $[\mathbf{w}_1^\top; \ldots; \mathbf{w}_Q^\top] = \mathbf{A_S}^\top$ (where $[\mathbf{w}_1^\top; \ldots; \mathbf{w}_Q^\top]$ indicates stacked row vectors) and $[\mathbf{w}_{Q+1}^\top; \ldots; \mathbf{w}_T^\top] = \mathbf{A_N}^\top$. Then (S.1) is equivalent to

$$\prod_{q=1}^{Q} f_q\left(\mathbf{w}_q^\top \boldsymbol{z}\right) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{Q+k}^\top \boldsymbol{z}) = |\det(\mathbf{C})| |\det(\mathbf{L})|^{-1} \prod_{\ell=1}^{T} g_\ell(\boldsymbol{c}_\ell^\top \mathbf{L}^{-1} \boldsymbol{z}).$$

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

$$\prod_{q=1}^{Q} f_q\left(\mathbf{w}_q^\top \boldsymbol{z}\right) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{Q+k}^\top \boldsymbol{z}) = |\det(\mathbf{R})| \prod_{\ell=1}^{T} g_\ell(\boldsymbol{r}_\ell^\top \boldsymbol{z}). \tag{S.2}$$

We have demonstrated identifiability up to signed permutations if we can show that Q of the marginal densities $g_1, \ldots, g_T$ are equivalent to $f_1, \ldots, f_Q$; that each of the corresponding Q rows of $\mathbf{R}$ equal $\pm\mathbf{w}_1, \ldots, \pm\mathbf{w}_Q$; 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 $\mathrm{E}\, \mathbf{Y} \mathbf{Y}^\top = \mathbf{I}$ (there is no loss of generality because we can scale $\mathbf{K}$ such that $\mathrm{E}\, \mathbf{Y} \mathbf{Y}^\top = \mathbf{I}$). From Theorem

3

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 $\mathbf{Y}_1 \overset{d}{=} \mathbf{S}$ (up to ordering), and it follows that there exists a subset of $g_1, \ldots, g_T$ equal to $f_1, \ldots, f_Q$. Also from Theorem 1, we have $\mathbf{K}_1 \cong \mathbf{A_S}$. Note that $\mathbf{K} \in \mathcal{O}_{T \times T}$ since $\mathrm{E}\,\mathbf{Y}\mathbf{Y}^\top = \mathbf{I}$ and $\mathrm{E}\,\mathbf{Z}\mathbf{Z}^\top = \mathbf{I}$, and hence $|\det(\mathbf{R})| = 1$. Then the scaling of $\mathbf{K}_1$ is also identifiable such that there exists a signed permutation matrix, $\mathbf{P}_\pm$, such that $\mathbf{K}_1 \mathbf{P}_\pm = \mathbf{A_S}$. Note that $\mathbf{W_S} = \mathbf{A_S}^\top$. Define $\mathbf{R_S} = \mathbf{K}_1^\top$. Then $\mathbf{P}_\pm^\top \mathbf{R_S} = \mathbf{W_S}$. $\qquad\square$

A.2 Proofs for Section 3

To simplify notation, we assume $\mathrm{E}\,\mathbf{X} = \mathbf{0}$ but include the estimate of the mean $\bar{\boldsymbol{x}}$ in our analysis so this assumption is without loss of generality. Let $f_\mathbf{S}$ denote the joint density of the LCs, and similarly define $p_\mathbf{S}(\boldsymbol{s}) = \prod_{q=1}^{Q} p_q(s_q)$ for the densities used in (4). Let $\|\mathbf{A}\|$ denote the Frobenius norm for $\mathbf{A} \in \mathbb{R}^{Q \times T}$.

Next we discuss Assumption 4 (ii) and inequality (5). The value of α will depend on the tail behavior of $\frac{d}{dx} \log\{p_q(x)\}$, $q = 1, \ldots, Q$. For insight into this assumption, consider $Q = 1$ such that $h(x) = \log p_1(x)$. By the mean value theorem,

$$\|h(x_1) - h(x_0)\| = \|h'(x^*)\| \, \|x_1 - x_0\|$$

with x^* between x_0 and x_1. Then if h' is monotonic,

$$\|h(x_1) - h(x_0)\| \leq \{\|h'(x_1)\| + \|h'(x_0)\|\} \, \|x_1 - x_0\|. \tag{S.3}$$

Therefore, if $\|h'(x)\|$ grows like $\|x\|^\alpha$ as $\|x\| \to \infty$, then (5) will hold.

For example, for the exponential power density centered at 0, which is

$$p_q(x) = \frac{\beta}{2\sigma\Gamma(1/\beta)} \exp\left\{-\left(\frac{|x|}{\sigma}\right)^\beta\right\},$$

we have
$$\frac{d}{dx}\log\{p_q(x)\} = -\beta\,\text{sign}(x)\frac{|x|^{\beta-1}}{\sigma^\beta},\ \ x \neq 0,$$

which is bounded for $\beta = 1$. For $\beta > 1$, we can take $\alpha = \beta - 1$. For $\beta < 1$, the exponential power density has an unbounded score function at zero, but similar densities can be constructed with exponential power law tails such that one can take $\alpha = 0$. The student-t distributions and the logistic distribution are other examples where $\frac{d}{dx}\log\{p_q(x)\}$ is bounded, so $\alpha = 0$. At least in these examples, lighter tails require large values of α, but, fortunately, make it easier for $E(\|\mathbf{S}\|^{1+\alpha}) < \infty$ to hold.

Equation (S.3) shows that (5) cannot be replaced by something like

$$\|h(x) - h(x')\| \leq M\|x - x'\|\left\{1 + \|x - x'\|^\alpha\right\}.$$

The following two propositions are used to prove consistency with pre-whitening. Recall that $\mathcal{J}_n$ is defined in (4) of the main manuscript.

Proposition 3. $\mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \xrightarrow{a.s.} \mathcal{J}_n(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i)$

Proof. First note that

$$\mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) = \mathcal{J}_n(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) + R_n$$

where

$$\begin{aligned}
\|R_n\| &= \left\|\mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) - \mathcal{J}_n(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i)\right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\|h(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) - h(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i)\right\|.
\end{aligned}$$

Using (5),

$$\frac{1}{n}\sum_{i=1}^{n}\left\|h(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}))-h(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i)\right\| \leq M\Bigg\{\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{O_S}(\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})-\mathbf{L}\boldsymbol{x}_i)\| \tag{S.4}$$

$$+\ \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{O_S}(\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})-\mathbf{L}\boldsymbol{x}_i)\|\,\|\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\|^{\alpha}+\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{O_S}(\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})-\mathbf{L}\boldsymbol{x}_i)\|\,\|\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i\|^{\alpha}\Bigg\}$$

Then since $\mathbf{O_S}$ is semi-orthogonal, the right-hand side of (S.4) is at most

$$M\Bigg\{\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})-\mathbf{L}\boldsymbol{x}_i\|+\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})-\mathbf{L}\boldsymbol{x}_i\|\,\|\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\|^{\alpha}$$

$$+\ \frac{1}{n}\sum_{i=1}^{n}\|\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})-\mathbf{L}\boldsymbol{x}_i\|\,\|\mathbf{L}\boldsymbol{x}_i\|^{\alpha}\Bigg\}. \tag{S.5}$$

Note that $\{\mathrm{E}\,\|\boldsymbol{x}_i\|^{1+\alpha}\}^{1/(1+\alpha)}=\{\mathrm{E}\,\|\mathbf{M}\boldsymbol{z}_i\|^{1+\alpha}\}^{1/(1+\alpha)}\leq\|\mathbf{M}\|\,\{\mathrm{E}\,(\|\boldsymbol{s}_i\|+\|\boldsymbol{n}_i\|)^{1+\alpha}\}^{1/(1+\alpha)}\leq$ $\|\mathbf{M}\|(\mathrm{E}\,\|\boldsymbol{s}_i\|^{1+\alpha})^{1/(1+\alpha)}+\|\mathbf{M}\|(\mathrm{E}\,\|\boldsymbol{n}_i\|^{1+\alpha})^{1/(1+\alpha)}<\infty$, where the last inequality uses Assumption 4 (iii) and properties of the normal distribution. For the first term on the right-hand side of (S.5)

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{\mathbf{L}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})-\mathbf{L}\boldsymbol{x}_i\|\leq\Bigg\{\|\widehat{\mathbf{L}}-\mathbf{L}\|\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|+\|\widehat{\mathbf{L}}\|\,\|\bar{\boldsymbol{x}}\|\Bigg\}\overset{a.s.}{\longrightarrow}0,$$

since $\widehat{\mathbf{L}}\overset{a.s.}{\longrightarrow}\mathbf{L}$, $\bar{\boldsymbol{x}}\overset{a.s.}{\longrightarrow}0$, and $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ are iid so we can apply the strong law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|\overset{a.s.}{\longrightarrow}\mathrm{E}\,(\|\boldsymbol{x}_i\|)\leq\{\mathrm{E}\,(\|\boldsymbol{x}_i\|^{1+\alpha})\}^{1/(1+\alpha)}<\infty.$$

For the second term on the right hand side of (S.5),

$$\frac{1}{n} \sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \, \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{\alpha}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left(\|\widehat{\mathbf{L}} - \mathbf{L}\| \, \|\boldsymbol{x}_i\| + \|\widehat{\mathbf{L}}\| \, \|\bar{\boldsymbol{x}}\| \right) \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{\alpha}$$

$$\leq \|\widehat{\mathbf{L}} - \mathbf{L}\| \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i\| \, \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{\alpha} + \|\widehat{\mathbf{L}}\| \, \|\bar{\boldsymbol{x}}\| \frac{1}{n} \sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{\alpha}.$$

To prove this converges to zero, we need to show the means are finite, but we can not directly apply a law of large numbers because the summands are not independent due to prewhitening. First, note that

$$\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i\| \, \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{\alpha} \leq \frac{1}{n} \sum_{i=1}^{n} \left\{ \|\boldsymbol{x}_i\|^{1+\alpha} + \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{1+\alpha} \right\}.$$

Then it remains to be shown that $\lim \frac{1}{n} \sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{1+\alpha} < \infty$. We have

$$\frac{1}{n} \sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{1+\alpha} \leq \frac{1}{n} \sum_{i=1}^{n} \left\{ \|\widehat{\mathbf{L}}\| \, \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\| \right\}^{1+\alpha}$$

$$\leq \|\widehat{\mathbf{L}}\|^{1+\alpha} \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^{1+\alpha} \tag{S.6}$$

Now consider

$$\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^{1+\alpha} \leq \frac{1}{n} \sum_{i=1}^{n} (\|\boldsymbol{x}_i\| + \|\bar{\boldsymbol{x}}\|)^{1+\alpha}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} (2\|\boldsymbol{x}_i\|)^{1+\alpha} + (2\|\bar{\boldsymbol{x}}\|)^{1+\alpha} \tag{S.7}$$

Since $\mathrm{E} \, \|\boldsymbol{x}_i\|^{1+\alpha} < \infty$, we apply the law of large numbers to conclude that (S.7) $< \infty$, and we conclude that (S.6) $< \infty$. Then (S.5) $\xrightarrow{a.s.} 0$ because $\|\widehat{\mathbf{L}} - \mathbf{L}\| \xrightarrow{a.s.} 0$ and $\|\bar{\boldsymbol{x}}\| \xrightarrow{a.s.} 0$.

The third term on the right-hand side of (S.5) can be handled similarly. $\qquad\square$

7

Proposition 4. *Let $B \subseteq \mathcal{O}_{Q \times T}$. Then*

$$\sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \leq \sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) + o(1) \qquad a.s.$$

Proof.

$$\sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \leq \sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) + \sup_{\mathbf{O_S} \in B} \frac{1}{n} \sum_{i=1}^n \left\| h(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \boldsymbol{x})) - h(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) \right\|$$

Note that

$$\sup_{\mathbf{O_S} \in B} \frac{1}{n} \sum_{i=1}^n \|\mathbf{O_S}(\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) - \mathbf{L}\boldsymbol{x}_i)\| \leq \sup_{\mathbf{O_S} \in B} \|\mathbf{O_S}\| \frac{1}{n} \sum_{i=1}^n \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\|$$

$$\leq \sqrt{Q} \frac{1}{n} \sum_{i=1}^n \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\|.$$

Using the inequality in (S.4) and the previous argument, we have

$$\sup_{\mathbf{O_S} \in B} \frac{1}{n} \sum_{i=1}^n \left\| h(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) - h(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) \right\| \leq MQ \left\{ \frac{1}{n} \sum_{i=1}^n \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \right.$$

$$+ \frac{1}{n} \sum_{i=1}^n \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \, \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^\alpha + \frac{1}{n} \sum_{i=1}^n \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \, \|\mathbf{L}\boldsymbol{x}_i\|^\alpha \left. \right\}. (\text{S.8})$$

Using the same arguments as in Proposition 3 to analyze the inequality in (S.5), we have
(S.8) $\xrightarrow{a.s.} 0$. $\qquad \square$

The next proposition is used in the proof of Theorem 2.

Proposition 5. *Consider a random vector $\mathbf{Y} \in \mathbb{R}^T$ with density $f_\mathbf{Y}$ such that $\mathrm{E}\,\mathbf{Y} = \mathbf{0}$ and $\mathrm{E}\,\mathbf{Y}\mathbf{Y}^\top = \mathbf{I}_T$. Then for any $\mathbf{o}$ and $\mathbf{w}$ such that $\mathbf{o}^\top \mathbf{o} = \mathbf{w}^\top \mathbf{w} = 1$, we have*

$$\mathrm{E} \log \phi(\mathbf{o}^\top \mathbf{Y}) = \mathrm{E} \log \phi(\mathbf{w}^\top \mathbf{Y}).$$

Proof. We can ignore the normalizing constants of $\phi(x)$ and consider the quadratic term

8

of the Gaussian kernel. Then we have $\mathrm{E}\left(\mathbf{o}^{\top}\mathbf{Y}\right)^2 = \mathbf{o}^{\top}\mathrm{E}\left(\mathbf{Y}\mathbf{Y}^{\top}\right)\mathbf{o} = \mathbf{o}^{\top}\mathbf{I}\mathbf{o} = \mathbf{o}^{\top}\mathbf{o} = 1$ and similarly for $\mathrm{E}\left(\mathbf{w}^{\top}\mathbf{Y}\right)^2$.

$\square$

Next we prove consistency when the density used in the objective function equals the true density.

Theorem 2. *Suppose* $\mathbf{X}$ *follows the LNGCA model in* (1) *with Assumptions 1-4. Given an iid sample* $\{\boldsymbol{x}_i\}$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Or} \xrightarrow{a.s.} \mathbf{W}_{\mathbf{S}}$ *on the equivalence class of signed permutations.*

Proof. We will include the effects of centering with $\bar{\boldsymbol{x}}$ in the discussion that follows such that it is without loss of generality that we assume $\mathrm{E}\,\mathbf{X} = \mathbf{0}$. Then $\mathbf{X} \sim (0, \boldsymbol{\Sigma})$ and let $\boldsymbol{\Sigma}^{-1/2} = \mathbf{L}$.

We will show the assumptions in Wald's consistency proof as recast in Theorem 5.14 in van der Vaart (2000) hold; a similar proof is in Pollard (2001). Note that this theory applies to a set of maxima of the population objective function, and thus is convenient for the set defined by the equivalence class of signed permutations of $\mathbf{W}_{\mathbf{S}}$. For clarity, we use $o_p(1)$ notation to correspond to van der Vaart, but note that Propositions 3 and 4 hold almost surely and the proof ultimately demonstrates strong consistency as in Wald (1949) and Pollard (2001). Recall $f_{\mathbf{S}}$ denotes the joint density of the LCs. The conditions are not all numbered in van der Vaart (2000), so for ease of reference we now state them.

(i.) The parameter space is compact. This is stated in Pollard (2001), where as van der Vaart proves consistency for all compact subsets, K, of the parameter space.

(ii.) $\log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{L}\boldsymbol{x})$ is upper-semicontinuous for almost all $\boldsymbol{x}$; in van der Vaart, this corresponds to (5.12).

(iii.) For every sufficiently small ball $U \subset \mathcal{O}_{Q \times T}$, the function $\mathbf{O}_{\mathbf{S}} \mapsto \sup_{\mathbf{O}_{\mathbf{s}} \in U} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{L}\boldsymbol{x}_i)$ is measurable and satisfies $\mathrm{E}\sup_{\mathbf{O}_{\mathbf{s}} \in U} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{L}\mathbf{X}) < \infty$; in van der Vaart, this corresponds to (5.13).

9

(iv.) $\mathrm{E}\log f_{\mathbf{S}}(\mathbf{O_S L X}) \leq \mathrm{E}\log f_{\mathbf{S}}(\mathbf{W_S L X})$ for any $\mathbf{O_S} \in \mathcal{O}_{Q \times T}$ with equality if and only if $\mathbf{O_S} \cong \mathbf{W_S}$; this assumption is part of the definition of Θ_0 following assumption (5.13) in van der Vaart and is assumption (i) in Pollard (2001).

(v.) The estimator satisfies:

$$\mathcal{J}_n(\widehat{\mathbf{W}}_{\mathbf{S}}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \geq \mathcal{J}_n(\mathbf{W_S L}\boldsymbol{x}_i) - o_p(1);$$

in van der Vaart's notation, this corresponds to $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$.

In addition to these conditions, we will outline van der Vaart's proof and provide additional justification to apply the law of large numbers, which is required because the observations are not iid due to pre-whitening.

First, $\mathcal{O}_{Q \times T}$ is compact, and (i) is satisfied. Next, we assume continuous densities which implies upper semicontinuity (condition ii). From Assumption 4 (i), the densities are bounded, say by some constant A, and we have $\mathrm{E}\sup_{\mathbf{O_S} \in U}\log f_{\mathbf{S}}(\mathbf{O_S L X}) \leq \mathrm{E}\log A < \infty$ and hence satisfy condition (iii).

We next show condition (iv) is satisfied. Let $\mathbf{W_N}$ denote rows $Q + 1$ to T of $\mathbf{W}$. Note that the fact that $\mathrm{E}\log f_{\mathbf{S}}(\mathbf{O_S L X}) \leq \mathrm{E}\log f_{\mathbf{S}}(\mathbf{W_S L X})$ does not hold trivially can be seen by the following argument:

$$
\begin{aligned}
\mathrm{E}\log\frac{f_{\mathbf{S}}(\mathbf{O_S L X})}{f_{\mathbf{S}}(\mathbf{W_S L X})} &= (\det \mathbf{L})\int \log\left\{\frac{f_{\mathbf{S}}(\mathbf{O_S L}\boldsymbol{x})}{f_{\mathbf{S}}(\mathbf{W_S L}\boldsymbol{x})}\right\}\{f_{\mathbf{S}}(\mathbf{W_S L}\boldsymbol{x})\phi(\mathbf{W_N L}\boldsymbol{x})\}\,d\boldsymbol{x} \\
&\leq (\det \mathbf{L})\log \int\left\{\frac{f_{\mathbf{S}}(\mathbf{O_S L}\boldsymbol{x})}{f_{\mathbf{S}}(\mathbf{W_S L}\boldsymbol{x})}\right\}\{f_{\mathbf{S}}(\mathbf{W_S L}\boldsymbol{x})\phi(\mathbf{W_N L}\boldsymbol{x})\}\,d\boldsymbol{x} \\
&= (\det \mathbf{L})\log \int f_{\mathbf{S}}(\mathbf{O_S L}\boldsymbol{x})\phi(\mathbf{W_N L}\boldsymbol{x})\,d\boldsymbol{x}.
\end{aligned}
$$

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 $\mathbf{O_S}$ and $\mathbf{W_N}$. The term $f_{\mathbf{S}}(\mathbf{O_S}\boldsymbol{x})\phi(\mathbf{W_N}\boldsymbol{x})$ is a density if and only if $|\det(\mathbf{W}^*)| = 1$, which is not true in general because $\mathbf{O_S}$ may not be orthogonal to $\mathbf{W_N}$. Consequently, this quantity could integrate to

greater than one, in which case we would have $\mathrm{E} \log f_{\mathbf{S}}(\mathbf{O_S L X}) \le \mathrm{E} \log f_{\mathbf{S}}(\mathbf{W_S L X}) + \alpha$ for some $\alpha > 0$, and the bound is not tight enough.

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

$$
\begin{aligned}
\mathrm{E} \log \frac{f_{\mathbf{S}}(\mathbf{O_S L X})}{f_{\mathbf{S}}(\mathbf{W_S L X})} &= \mathrm{E} \log \frac{f_{\mathbf{S}}(\mathbf{O_S L X})\phi(\mathbf{O_N L X})}{f_{\mathbf{S}}(\mathbf{W_S L X})\phi(\mathbf{O_N L X})} \\
&= \mathrm{E} \log \frac{f_{\mathbf{S}}(\mathbf{O_S L X})\phi(\mathbf{O_N L X})}{f_{\mathbf{S}}(\mathbf{W_S L X})\phi(\mathbf{W_N L X})},
\end{aligned}
$$

where the second line follows from Proposition 5. Then applying Jensen's inequality, we have

$$
\begin{aligned}
\mathrm{E} \log \frac{f_{\mathbf{S}}(\mathbf{O_S L X})\phi(\mathbf{O_N L X})}{f_{\mathbf{S}}(\mathbf{W_S L X})\phi(\mathbf{W_N L X})} &\le (\det \mathbf{L}) \log \int \left(\frac{f_{\mathbf{S}}(\mathbf{O_S L X})\phi(\mathbf{O_N L X})}{f_{\mathbf{S}}(\mathbf{W_S L X})\phi(\mathbf{W_N L X})} \right) f_{\mathbf{S}}(\mathbf{W_S L X})\phi(\mathbf{W_N L X}) d\boldsymbol{x} \\
&= (\det \mathbf{L}) \log \int f_{\mathbf{S}}(\mathbf{O_S L}\boldsymbol{x})\phi(\mathbf{O_N L}\boldsymbol{x}) d\boldsymbol{x} \\
&= 0,
\end{aligned}
$$

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

$$
\mathrm{E} \log f_{\mathbf{S}}(\mathbf{O_S L X}) < \mathrm{E} \log f_{\mathbf{S}}(\mathbf{W_S L X})
$$

for all $\mathbf{O_S} \not\cong \mathbf{W_S}$.

To show condition (v) is satisfied,

$$\mathcal{J}_n(\widehat{\mathbf{W}}_{\mathbf{S}}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \geq \mathcal{J}_n(\mathbf{W}_{\mathbf{S}}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \qquad \text{(by definition)}$$

$$= \mathcal{J}_n(\mathbf{W}_{\mathbf{S}}\mathbf{L}\boldsymbol{x}_i) - o_p(1). \qquad \text{(Proposition 3)}$$

In other words, our estimator $\widehat{\mathbf{W}}_{\mathbf{S}}$ with $\mathcal{J}_n$ defined using the sequence $\{\widehat{\mathbf{L}}, \bar{\boldsymbol{x}}\}$ is an approximate maximum of the exact maximum of the function $\mathcal{J}_n(\mathbf{O}_{\mathbf{S}}\mathbf{L}\boldsymbol{x}_i)$.

In this paragraph, we recount the first half of the proof of van der Vaart (2000) 5.14. Let $\mathcal{W}_{\mathbf{S}}$ be the set of signed permutations of $\mathbf{W}_{\mathbf{S}}$. Fix some $\mathbf{O}_{\mathbf{S}}^{\dagger} \notin \mathcal{W}_{\mathbf{S}}$ with $\mathbf{O}_{\mathbf{S}}^{\dagger} \in \mathcal{O}_{Q \times T}$, and let U_ℓ be a decreasing sequence of open balls around $\mathbf{O}_{\mathbf{S}}^{\dagger}$ with diameter converging to zero. Define the function: $m_{U_\ell}(\boldsymbol{x}_i) = \sup_{\mathbf{O}_{\mathbf{S}} \in U_\ell} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}\mathbf{L}\boldsymbol{x}_i)$. Then using (ii) we have $m_{U_\ell}(\boldsymbol{x}_i) \downarrow \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}^{\dagger}\mathbf{L}\boldsymbol{x}_i)$ and from (iii) we can apply the monotone convergence theorem to obtain $\mathrm{E}\, m_{U_\ell}(\boldsymbol{x}_i) \downarrow \mathrm{E} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}^{\dagger}\mathbf{L}\boldsymbol{x}_i)$. From (iv), we have $\mathrm{E} \log f_{\mathbf{S}}(\mathbf{O}_{\mathbf{S}}^{\dagger}\mathbf{L}\mathbf{X}) < \mathrm{E} \log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{L}\mathbf{X})$. Then with the previous argument, for any $\mathbf{O}_k \in \mathcal{O}_{Q \times T} \setminus \mathcal{W}_{\mathbf{S}}$, we can define a set $U_{\mathbf{O}_k}$ such that $\mathrm{E}\, m_{U_{\mathbf{O}_k}}(\boldsymbol{x}_i) < \mathrm{E} \log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{L}\mathbf{X})$. Now let ϵ be given and consider the set $B = \{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T} : \cap_{\mathbf{W}_{\mathbf{S}}^* \in \mathcal{W}_{\mathbf{S}}} \|\mathbf{O}_{\mathbf{S}} - \mathbf{W}_{\mathbf{S}}^*\| \geq \epsilon\}$, which is compact. This set is covered by the balls $U_{\mathbf{O}_k}$. Then there exists a finite subcover $U_1, \ldots, U_p$.

Next, we detail the second half of the proof of van der Vaart (2000) 5.14, where we incorporate Proposition 4 to account for pre-whitening. In the argument that follows, note that if $\mathrm{E}\, m_{U_k}(X) = -\infty$ for some k, then we can discard the set U_k, and since we have $\mathrm{E}\, m_{U_j}(X) < \infty$ from (iii), we have $\mathrm{E}\, |m_{U_j}(X)| < \infty$ for all remaining sets, and

$\frac{1}{n} \sum_{i=1}^{n} m_{U_j}(\boldsymbol{x}_i) \xrightarrow{a.s.} \mathrm{E}\, m_{U_j}$ from the law of large numbers.

$$\sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \leq \sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S} \mathbf{L} \boldsymbol{x}_i) + o_p(1) \qquad \text{(from Proposition 4)} \quad \text{(S.9)}$$

$$\leq \sup_{j=1,\ldots,p} \sup_{\mathbf{O_S} \in U_j} \mathcal{J}_n(\mathbf{O_S} \mathbf{L} \boldsymbol{x}_i) + o_p(1)$$

$$\leq \sup_{j=1,\ldots,p} \frac{1}{n} \sum_{i=1}^{n} m_{U_j}(\boldsymbol{x}_i) + o_p(1)$$

$$\rightarrow \sup_{j=1,\ldots,p} \mathrm{E}\, m_{U_j}(\mathbf{X}) \qquad \text{(law of large numbers)}$$

$$< \mathrm{E} \log f_\mathbf{S}(\mathbf{W_S L X}). \qquad (\text{S}.10)$$

Now if $\widehat{\mathbf{W}}_\mathbf{S} \in B$, then we have

$$\sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \geq \mathcal{J}_n(\mathbf{W_S L} \boldsymbol{x}_i) - o_p(1) \qquad \text{(from condition (v.))}$$

$$= \mathrm{E} \log f_\mathbf{S}(\mathbf{W_S L X}) - o_p(1), \qquad \text{(from LLN)}$$

which would imply the following relationship between events:

$$\left\{ \widehat{\mathbf{W}}_\mathbf{S} \in B \right\} \subset \left\{ \sup_{\mathbf{O_S} \in B} \mathcal{J}_n(\mathbf{O_S} \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \geq \mathrm{E} \log f_\mathbf{S}(\mathbf{W_S L X}) - o_p(1) \right\}. \qquad (\text{S}.11)$$

In view of (S.9) and (S.10), the probability of the event on the right-hand side of (S.11) converges to zero as $n \to \infty$. Note the $o_p(1)$ inequalities hold almost surely from Propositions 3 and 4. Then

$$P \left(\lim_{n \to \infty} \bigcap_{\mathbf{W_S^*} \in \mathcal{W_S}} \left\{ \|\widehat{\mathbf{W}}_\mathbf{S} - \mathbf{W_S^*}\| \geq \epsilon \right\} \right) \to 0.$$

$\square$

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

13

to the log of the density used in the objective function.

Recall that $r_q(\cdot)$ denotes the score function of $\log f_q(\cdot)$ and $r_q'(\cdot)$ denotes the derivative of the score function. Additionally, define $\mathbf{Z} = [\mathbf{S}^\top, \mathbf{N}^\top]^\top$.

Lemma 2. *Let $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^\top$ and let $\boldsymbol{\epsilon}$ be given such that $||\boldsymbol{e}_1 + \boldsymbol{\epsilon}|| = 1$. Then*

$$\mathrm{E} \log p_1 \left[(\boldsymbol{e}_1 + \boldsymbol{\epsilon})^\top \mathbf{Z} \right] = \mathrm{E} \log p_1(S_1) + \frac{1}{2} \left[\mathrm{E}\, r_1'(S_1) - \mathrm{E}\, S_1 r_1(S_1) \right] \sum_{q=2}^{T} \epsilon_q^2 + o(||\boldsymbol{\epsilon}||^2).$$

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

$$\nabla \mathrm{E} \log p_1(\mathbf{o}^\top \mathbf{Z}) = \mathrm{E}\, \mathbf{Z} r_1(\mathbf{o}^\top \mathbf{Z}),$$

where we have applied Assumption 5(iv) to interchange differentiation and integration. Evaluating this at $\boldsymbol{e}_1$, and using the fact that $\mathrm{E}\, S_q = \mathrm{E}\, N_k = 0$, $q = 1, \ldots, Q$, $k = 1, \ldots, T - Q$, and the fact that $\mathbf{S}_1$ is independent of $\mathbf{S}_q$, $q > 1$, and $\mathbf{N}_k$,

$$\nabla \mathrm{E} \log p_1(\boldsymbol{e}_1^\top \mathbf{Z}) = \boldsymbol{e}_1 \mathrm{E}\, S_1 r_1(S_1).$$

We also have

$$\nabla^2 \mathrm{E} \log p_1(\boldsymbol{e}_1^\top \mathbf{Z}) = \mathrm{diag} \left[\mathrm{E}\, S_1^2 r_1'(S_1), \mathrm{E}\, r_1'(S_1), \ldots, \mathrm{E}\, r_1'(S_1) \right]$$

where as before we have interchanged integration and differentiation using Assumption 5(iv) and applied independence and the fact that $\mathrm{E}\, S_q^2 = \mathrm{E}\, N_k^2 = 1$.

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

$$\mathrm{E} \log p_1[(\boldsymbol{e}_1 + \boldsymbol{\epsilon})^\top \mathbf{Z}] =$$

$$\mathrm{E} \log p_1(S_1) + \boldsymbol{\epsilon}^\top \boldsymbol{e}_1 \mathrm{E}\, S_1 r_1(S_1) + \frac{1}{2} \boldsymbol{\epsilon}^\top \mathrm{diag} \left[\mathrm{E}\, S_1^2 r_1'(S_1), \mathrm{E}\, r_1'(S_1), \ldots, \mathrm{E}\, r_1'(S_1) \right] \boldsymbol{\epsilon} + o(||\boldsymbol{\epsilon}||^2) =$$

$$\mathrm{E} \log p_1(S_1) + \epsilon_1 \mathrm{E}\, S_1 r_1(S_1) + \frac{1}{2} \epsilon_1^2 \mathrm{E}\, S_1^2 r_1'(S_1) + \frac{1}{2} \mathrm{E}\, r_1'(S_1) \sum_{q>1} \epsilon_q^2 + o(||\boldsymbol{\epsilon}||^2).$$

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 $\epsilon_1 = -\frac{1}{2} \sum_{q>1} \epsilon_q^2 + o(\sum_{q>1} \epsilon_q^2)$. By Assumption 5(ii), $|\mathrm{E}\, S_1^2 r_1'(S_1)| < \infty$. Then we can write

$$\mathrm{E} \log p_1 \left[(\boldsymbol{e}_1 + \boldsymbol{\epsilon})' \mathbf{Z} \right] = \mathrm{E} \log p_1(S_1) + \frac{1}{2} \left[\mathrm{E}\, r_1'(S_1) - \mathrm{E}\, S_1 r_1(S_1) \right] \sum_{q>1} \epsilon_q^2 + o(||\boldsymbol{\epsilon}||^2).$$

$\square$

Proposition. *(Proposition 1 in the main manuscript.) Suppose Assumptions 1-3 and 5. There exists $\mathcal{N}_{\epsilon^*}(\mathbf{W_S})$ such that $\mathrm{E} \log p(\mathbf{O_S L X})$ constrained to $\mathbf{O_S} \in \mathcal{N}_{\epsilon^*}(\mathbf{W_S})$ is maximized at $\mathbf{W_S}$.*

Proof. We consider a perturbation of $\mathbf{W_S}$. Using the change of variables $\mathbf{Z} = \mathbf{WLX} = [\mathbf{S}^\top, \mathbf{N}^\top]^\top$, it suffices to consider the case where $\mathbf{w}_q = \boldsymbol{e}_q$, where $\boldsymbol{e}_{qt} = 1$ for $q = t$ and 0 otherwise. For $q = 1$, consider a perturbation $\boldsymbol{\epsilon}_1 \in \mathbb{R}^T$ with $||\boldsymbol{e}_1 + \boldsymbol{\epsilon}_1|| = 1$. From Lemma 2, we have

$$\mathrm{E} \log p_1[(\boldsymbol{e}_1 + \boldsymbol{\epsilon}_1)^\top \mathbf{Z}] = \mathrm{E} \log p_1(S_1) + \frac{1}{2} \mathrm{E}\, [r_1'(S_1) - S_1 r_1(S_1)] \sum_{q>1} \epsilon_{1q}^2 + o(||\boldsymbol{\epsilon}_1||^2).$$

By Assumption 5(i), which states $\mathrm{E}\, r_q'(S_q) - \mathrm{E}\, S_q\, r_q(S_q) < 0$, and for sufficiently small $\boldsymbol{\epsilon}_1$, we have

$$\frac{1}{2} \mathrm{E}\, [r_1'(S_1) - S_1 r_1(S_1)] \sum_{q>1} \epsilon_{1q}^2 + o(||\boldsymbol{\epsilon}_1||^2) < 0,$$

which makes $\boldsymbol{e}_1$ a local maximum for $\mathrm{E}\log p_1(\mathbf{o}^\top\mathbf{Z})$. Since this also true for $\mathrm{E}\log p_q(\mathbf{o}^\top\mathbf{Z})$, $q = 2,\ldots,Q$, we have that $\mathbf{I}_{Q\times T}$ (the $Q\times Q$ identity matrix padded with zeros) is a local maximum on the set $\mathcal{G}_{Q\times T} = \{\mathbf{G}\in\mathbb{R}^{Q\times T} : \mathrm{diag}\,\mathbf{G}\mathbf{G}^\top = \mathbf{1}_Q\}$. 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). Then for the perturbations $\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_Q$, it suffices to let $\epsilon^* = \min_{q=1}^Q\min_{t=1}^T\epsilon_{qt}$, and define $\mathcal{N}_{\epsilon^*}(\mathbf{W_S})$. $\qquad\square$

Theorem 3. *Suppose* $\mathbf{X}$ *follows the LNGCA model in* (1) *with Assumptions 1-5. Given an iid sample* $\{\boldsymbol{x}_i\}$, $\widehat{\mathbf{W}}_{\mathbf{S}}^{Local}\xrightarrow{a.s.}\mathbf{W_S}$ *on the equivalence class of signed permutations.*

Proof. We restrict the parameter space to $\mathcal{N}_{\epsilon^*}(\mathbf{W_S})$. 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 (condition (iv) in Theorem 2), which was proven in Proposition 1 for the restricted parameter space $\mathcal{N}_{\epsilon^*}(\mathbf{W_S})$. The other conditions are satisfied using the previous arguments in the proof of Theorem 2. $\qquad\square$

A.3 Proofs for Section 4

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

Proposition. *(Proposition 2 in the main manuscript.) Let* G *be the class of all cubic splines* $g : \mathbb{R}\to\mathbb{R}$. *Consider the argmax of* (11) *of the main manuscript for* $g_q\in G$ *with* g_q *denoting the tilt function for the* qth *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 $g\subset G$, 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 = $

16

$\partial(e^{\theta_0} \int \phi(u)e^{\theta_1 u+j(u)}du)/\partial\theta_0 = \int \phi(u)e^{g(u)}du$, 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,

$$\frac{\partial \ell_{pen}}{\partial \theta_1} = \frac{1}{n}\sum_{i=1}^n \mathbf{o}_1^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \int u\phi(u)e^{g(u)}\, du,$$

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

B   Additional Asymptotics for Section 3

In this section, we examine $\sqrt{n}$-consistency, asymptotic normality, and the asymptotic variances of the parametric LCA estimators.

Recall that $r_q(\cdot)$ is the score function of $\log f_q(\cdot)$ and $r'_q(\cdot)$ is the derivative of the score function. Define the following quantities:

$$\beta_q = \mathrm{E}\, S_q^4$$

$$\eta_q = \mathrm{E}\, r(S_q)$$

$$\xi_q = \mathrm{E}\, r(S_q)^2 - \eta_q^2$$

$$\lambda_q = \mathrm{E}\, r(S_q)S_q$$

$$\delta_q = \mathrm{E}\, r'(S_q)$$

Also define the empirical expectation: $\mathbb{E}_n f(x_i) = \frac{1}{n}\sum_{i=1}^n f(x_i)$. Recall that $\boldsymbol{e}_q \in \mathbb{R}^T$ such that $\boldsymbol{e}_{qq'} = 0$ for $q' \neq q$ and 1 for $q' = q$.

We apply the approach used in Virta et al. (2016) to derive asymptotic variances based on rewriting the objective function using Lagrange multipliers. Virta et al. (2016) find non-Gaussian components using a modified version of symmetric fastICA but with the measure of non-Gaussianity equal to a convex combination of squared skewness and kurtosis. We adapt their approach to log likelihoods. For an arbitrary consistent estimator of the LNGCA model, $\widehat{\mathbf{W}}_{\mathbf{S}}$, define $\widehat{\mathbf{B}}_{\mathbf{S}} = \widehat{\mathbf{W}}_{\mathbf{S}}\widehat{\mathbf{L}}$. Let $\mathbf{B}_{\mathbf{S}}$ be the first Q rows of $\mathbf{M}^{-1}$. Consistency of $\widehat{\mathbf{B}}_{\mathbf{S}}$ follows from Slutsky's theorem. Throughout the remainder of this section, we focus on $\widehat{\mathbf{B}}_{\mathbf{S}}$ rather than $\widehat{\mathbf{W}}_{\mathbf{S}}$.

First, consider:

$$\mathcal{L}(\mathbf{C_S}, \boldsymbol{\Theta}) = \sum_{q=1}^{Q} \mathbb{E}_n \left\{ \log p_q(\boldsymbol{c}_q^\top(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \right\} - \sum_{q=1}^{Q} \frac{\theta_{qq}}{2}(\boldsymbol{c}_q^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{c}_q - 1) - \sum_{q=1}^{Q-1} \sum_{q'=q+1}^{Q} \theta_{qq'} \boldsymbol{c}_q^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{c}_{q'}.$$

(S.12)

Consider the substitution $\mathbf{o}_q^\top \widehat{\mathbf{L}} = \boldsymbol{c}_q$. Then we rewrite (S.12):

$$\mathcal{L}(\mathbf{O_S}, \boldsymbol{\Theta}) = \sum_{q=1}^{Q} \mathbb{E}_n \left\{ \log p_q(\mathbf{o}_q^\top \widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \right\} - \sum_{q=1}^{Q} \frac{\theta_{qq}}{2}(\mathbf{o}_q^\top \mathbf{o}_q - 1) - \sum_{q=1}^{Q-1} \sum_{q'=q+1}^{Q} \theta_{qq'} \mathbf{o}_q^\top \mathbf{o}_{q'}.$$

(S.13)

Then the partial derivatives of (S.12) at $\widehat{\mathbf{B}}_{\mathbf{S}}$ equal zero.

In the special case where $\mathbf{M} = \mathbf{I}$, let $\hat{\boldsymbol{e}}_q$ be the estimate of the qth row of the true unmixing matrix $\mathbf{I}_{Q \times T}$.

Next we define the conditions for $\sqrt{n}$-consistency and asymptotic normality.

Assumption 6. *For all q, the following expectations are finite: (i) $\mathrm{E}\, S_q^4$; (ii) $\mathrm{E}\, r_q^2(S_q)$; (iii) $\mathrm{E}\, r_q'(S_q)$; (iv) $\mathrm{E}\, r_q(S_q)S_q$; and (v) $\mathrm{E}\, r_q'(S_q)S_q$.*

Lemma 3. *Suppose $\mathrm{E}\, \mathbf{X} = 0$, $\mathbf{M} = \mathbf{I}$, and Assumptions 1-6. Consider a consistent estimator, $\widehat{\mathbf{E}}_{\mathbf{S}}$, of the first Q rows of $\mathbf{M}^{-1}$ with the rows permuted and signs specified such that*

$\hat{e}_q \to e_q$. Let $\hat{e}_{qq'}$ be the q'th element of $\hat{e}_q$. Then

$$\sqrt{n}(\hat{e}_{qq'}) = \sqrt{n}\frac{\mathbb{E}_n\left\{(r_q(s_{iq}) - \eta_q)s_{iq'} - (r_{q'}(s_{iq'}) - \eta_{q'})s_{iq} - (\delta_{q'} - \lambda_q)s_{iq}s_{ir}\right\}}{\delta_q - \lambda_q + \delta_{q'} - \lambda_{q'}} + o_p(1), \ q,q' \leq Q$$

(S.14)

$$\sqrt{n}(\hat{e}_{qq} - 1) = -\sqrt{n}\frac{1}{2}\mathbb{E}_n\left(s_{iq}^2 - 1\right) + o_p(1), \ q \leq Q \tag{S.15}$$

$$\sqrt{n}(\hat{e}_{qr}) = \sqrt{n}\,\frac{\mathbb{E}_n\left[\{r_q(s_{iq}) - \eta_q\}\,n_{i,r-Q} - \lambda_q s_{iq}n_{i,r-Q}\right]}{\lambda_q - \delta_q} + o_p(1), \ q \leq Q, Q < r < T. \tag{S.16}$$

Proof. At the estimates $\hat{e}_q$, the Lagrangian in (S.12) enforces the constraints

$$\hat{e}_q^\top\widehat{\Sigma}\hat{e}_{q'} = 0, q \neq q' \tag{S.17}$$

$$\hat{e}_q^\top\widehat{\Sigma}\hat{e}_q = 1. \tag{S.18}$$

Now we differentiate the Lagrangian with respect to c_q and set the result equal to zero, and replace c_q with the estimates $\hat{e}_q$, $q = 1, \ldots, Q$:

$$\mathbb{E}_n\, r_q(\hat{e}_q^\top(x_i - \bar{x}))(x_i - \bar{x}) = \theta_{qq}\widehat{\Sigma}\hat{e}_q + \sum_{q' \neq q}\theta_{qq'}\widehat{\Sigma}\hat{e}_{q'}. \tag{S.19}$$

Next, write (S.19) as

$$\mathbb{E}_n\, r_q(\hat{e}_q^\top(x_i - \bar{x}))(x_i - \bar{x}) = \widehat{\Sigma}\sum_{q'=1}^{Q}\hat{e}_{q'}\theta_{qq'}. \tag{S.20}$$

Multiplying (S.19) by $\hat{e}_{q'}$ and applying (S.17) and (S.18), we get

$$\hat{e}_{q'}^\top\mathbb{E}_n\, r_q(\hat{e}_q^\top(x_i - \bar{x}))(x_i - \bar{x}) = \theta_{qq'}.$$

Then substituting this expression into (S.20), we write

$$\mathbb{E}_n \, r_q(\hat{\boldsymbol{e}}_q^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}))(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \widehat{\boldsymbol{\Sigma}} \left(\sum_{q'=1}^Q \hat{\boldsymbol{e}}_{q'} \hat{\boldsymbol{e}}_{q'}^\top \right) \left[\mathbb{E}_n \left\{ r_q(\hat{\boldsymbol{e}}_q^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}))(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\} \right]. \quad \text{(S.21)}$$

This is the same estimating equation that appears in deflationary fastICA when $q = Q$, see equation (4) in Nordhausen et al. (2011), but here it applies to all $q \leq Q$, and we replace the non-linearities with the log likelihoods. Then we apply Theorem 1 from Nordhausen et al. (2011); see similar theorems in Miettinen et al. (2017, 2015) and Virta et al. (2016), which requires Assumption 6:

$$\sqrt{n}\hat{\boldsymbol{e}}_{qq'} = -\sqrt{n}\hat{\boldsymbol{e}}_{q'q} - \sqrt{n}\mathbb{E}_n \, (x_{iq} - \bar{x}_q)(x_{iq'} - \bar{x}_{q'}) + o_p(1), \; q \neq q', q, q' \leq Q \quad \text{(S.22)}$$

$$\sqrt{n}(\hat{\boldsymbol{e}}_{qq} - 1) = -\frac{1}{2}\sqrt{n} \left\{ \mathbb{E}_n \, (x_{iq} - \bar{x}_q)^2 - 1 \right\} + o_p(1), q \leq Q \quad \text{(S.23)}$$

$$\sqrt{n}\hat{\boldsymbol{e}}_{qr} = \sqrt{n}\frac{1}{\lambda_q - \delta_q} \left[\boldsymbol{e}_r^\top \mathbb{E}_n \left\{ r_q(\boldsymbol{e}_q^\top \boldsymbol{x}_i) - \eta_q \right\} \boldsymbol{x}_i - \lambda_q \mathbb{E}_n \, (x_{iq} - \bar{x}_q)(x_{ir} - \bar{x}_r) \right] + o_p(1).$$

$$\text{(S.24)}$$

Next note that

$$\sqrt{n} \left[\mathbb{E}_n \, (x_{iq} - \bar{x}_q)^2 \right] = \sqrt{n}\mathbb{E}_n \, x_{iq}^2 + o_p(1),$$

since $\sqrt{n}\bar{x}_q^2 = o_p(1)$. Similarly, $\sqrt{n}\bar{x}_q\bar{x}_r = o_p(1)$. Then applying $x_{iq} = s_{iq}$ and $x_{ir} = n_{i,r-Q}$, we obtain (S.15) and (S.16).

To obtain (S.14), we derive a second expression for $\theta_{qq'}$ by performing the differentiation with respect to $\boldsymbol{c}_{q'}$ and multiplying by $\hat{\boldsymbol{e}}_q$:

$$\mathbb{E}_n \, r_q(\hat{\boldsymbol{e}}_{q'}^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}))\hat{\boldsymbol{e}}_q^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \theta_{qq'}. \quad \text{(S.25)}$$

This gives us the estimating equations:

$$\mathbb{E}_n \left[r_q \left\{ \hat{\boldsymbol{e}}_q^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\} \hat{\boldsymbol{e}}_{q'}^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right] = \mathbb{E}_n \left[r_{q'} \left\{ \hat{\boldsymbol{e}}_{q'}^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right\} \hat{\boldsymbol{e}}_q^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \right], \ q, q' \le Q. \quad \text{(S.26)}$$

The estimating equation in (S.26) is also found in symmetric fastICA (Miettinen et al., 2015, 2017; Wei, 2015) but here restricted to $q, q' \le Q$, and we replace the nonlinearities by the log likelihoods. Then (S.14) is a special case of the symmetric case in Theorem 1 in Miettinen et al. (2017) with additional details in the proof of Theorem 6 in Miettinen et al. (2015), where here we revise the sign modification, π_j, in their theorem, to be equal to -1 when we use the log likelihood in lieu of their objective function. Again, we use the fact that the terms arising from centering converge at a faster rate and thus vanish from the asymptotic variances. Then we restate the symmetric case from Theorem 1 in Miettinen et al. (2017) in terms of the iid non-Gaussian components. □

Theorem 4. *Suppose Assumptions 1-6 and additionally let $\mathbf{M} = \mathbf{I}$ and $\mathbb{E}\,\mathbf{X} = \mathbf{0}$. Consider a consistent estimator, $\widehat{\mathbf{E}}_\mathbf{S}$, of the first Q rows of $\mathbf{M}^{-1}$ with the rows permuted and signs specified such that $\hat{\boldsymbol{e}}_q \to \boldsymbol{e}_q$. Then for $q \le Q$, $\sqrt{n}(\hat{\boldsymbol{e}}_q - \boldsymbol{e}_q) \Rightarrow \mathcal{N}(0, \mathbf{R}_q)$ with*

$$\begin{aligned}
\mathbf{R}_q &= \frac{\beta_q - 1}{4} \boldsymbol{e}_q \boldsymbol{e}_q^\top \\
&+ \sum_{q' \ne q}^Q \frac{\xi_q + \xi_{q'} + \delta_{q'}^2 - \lambda_q^2 - 2\delta_{q'}\lambda_{q'}}{(\delta_q - \lambda_q + \delta_{q'} - \lambda_{q'})^2} \boldsymbol{e}_{q'} \boldsymbol{e}_{q'}^\top \\
&+ \frac{\xi_q - \lambda_q^2}{(\lambda_q - \delta_q)^2} \left(\mathbf{I} - \sum_{q'=1}^Q \boldsymbol{e}_{q'} \boldsymbol{e}_{q'}^\top \right).
\end{aligned} \quad \text{(S.27)}$$

Proof. Asymptotic normality follows from the central limit theorem for the iid observations on the right-hand side of equations (S.14)-(S.16) together with Slutsky's theorem. The variances can be calculated directly from the previous lemma and correspond to the variances of symmetric fastICA for $q \le Q$ and the variances of deflationary fastICA for $r > Q$. □

Note that the asymptotic variances for symmetric fastICA are also derived in Theorem 8

in Wei (2015) using a modified M-estimator approach. They are equivalent to Miettinen et al. (2017) except the sign modification is replaced by the sign of the term $\mathrm{E}\, r'_q(S_q) - \mathrm{E}\, S_q r_q(S_q)$. In LCA, this is always equal to negative one due to Assumption 5(i), and then Wei (2015) Theorem 8 is equivalent to the result presented here for $q, q' \leq Q$ and $\mathbf{M} = \mathbf{I}$.

It is straightforward to extend this result to arbitrary mixing matrices when the estimators are affine equivariant, and this property is used in the estimators considered in Virta et al. (2016) and related works by Nordhausen et al. (2011) and Miettinen et al. (2015). Let $F_{\mathbf{X}}$ be the cumulative distribution of $\mathbf{X}$, and let $\mathcal{B}(F_{\mathbf{X}}) \in \mathbb{R}^{Q \times T}$ be a functional. As defined in Nordhausen et al. (2011),

Definition 1. *A functional $\mathcal{B}(F_{\mathbf{X}})$ is affine equivariant if*

$$\mathcal{B}(F_{\mathbf{A}\mathbf{X}}) = \mathcal{B}(F_{\mathbf{X}})\mathbf{A}^{-1}.$$

Wei (2015) proves that an estimator is affine equivariant if and only if it does not depend on initialization, and thus our estimators are not in general affine equivariant. In practice, we satisfy this requirement by initializing from a sufficiently large number of random orthogonal matrices, such that if we were to estimate the unmixing matrix with another set of random initial values, we would obtain the same estimate with high probability. Alternatively, one can use the two-stage estimator, $\widehat{\mathbf{W}}_{\mathbf{S}}^{LV}$, since the estimator from Virta et al. (2016) is affine equivariant.

For the following theorem, we additionally assume the estimator is globally consistent, for example, under finite eighth moment assumptions with $\widehat{\mathbf{W}}_{\mathbf{S}}^{LV}$, which simplifies the exposition by avoiding the dependency between the optimization space and the choice of mixing matrix.

Corollary 2. *Suppose Assumptions 1-6. Let $\widehat{\mathbf{B}}_{\mathbf{S}}$ be a globally consistent and affine equivariant estimator of the LCA model for any full rank $\mathbf{M} \in \mathbb{R}^{T \times T}$ with $\mathbf{M}^{-1} = \mathbf{B}$, and let $\widehat{\mathbf{B}}_{\mathbf{S}}$ have rows permuted and signs chosen such that $\widehat{\mathbf{B}}_{\mathbf{S}} \to \mathbf{B}_{\mathbf{S}}$. Then for $q \leq Q$,*

$\sqrt{n}(\hat{\boldsymbol{b}}_q - \boldsymbol{b}_q) \Rightarrow \mathcal{N}(0, \mathbf{R}_q)$ *with*

$$\mathbf{R}_q = \frac{\beta_q - 1}{4} \boldsymbol{b}_q \boldsymbol{b}_q^\top + \sum_{q' \neq q}^{Q} \frac{\xi_q + \xi_{q'} + \delta_{q'}^2 - \lambda_q^2 - 2\delta_{q'}\lambda_{q'}}{(\delta_q - \lambda_q + \delta_{q'} - \lambda_{q'})^2} \boldsymbol{b}_{q'} \boldsymbol{b}_{q'}^\top \qquad \text{(S.28)}$$
$$+ \frac{\xi_q - \lambda_q^2}{(\lambda_q - \delta_q)^2} \left(\boldsymbol{\Sigma}^{-1} - \sum_{q'=1}^{Q} \boldsymbol{b}_{q'} \boldsymbol{b}_{q'}^\top \right).$$

Proof. Consider the trivial model: $\boldsymbol{z}_i = \mathbf{I}\boldsymbol{z}_i$ and let $\hat{\mathbf{I}}_{\mathbf{S}} = \underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} \ \mathcal{J}_n(\mathbf{O}_{\mathbf{S}}; \{\boldsymbol{z}_i\})$. Define $\widehat{\mathbf{W}}_{\mathbf{S}} = \underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} \ \mathcal{J}_n(\mathbf{O}_{\mathbf{S}}; \{\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\})$. Then

$$\widehat{\mathbf{B}}_{\mathbf{S}} = \widehat{\mathbf{W}}_{\mathbf{S}} \widehat{\mathbf{L}}$$
$$= \left[\underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} \ \mathcal{J}_n(\mathbf{O}_{\mathbf{S}}; \widehat{\mathbf{L}}\{\boldsymbol{x}_i - \bar{\boldsymbol{x}}\}) \right] \widehat{\mathbf{L}}$$
$$= \left[\underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} \ \mathcal{J}_n(\mathbf{O}_{\mathbf{S}}; \widehat{\mathbf{L}}\mathbf{M}\boldsymbol{z}_i) \right] \widehat{\mathbf{L}}$$
$$= \left[\underset{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q \times T}}{\operatorname{argmax}} \ \mathcal{J}_n(\mathbf{O}_{\mathbf{S}}; \boldsymbol{z}_i) \right] \mathbf{B}\widehat{\mathbf{L}}^{-1}\widehat{\mathbf{L}}.$$

Then $\widehat{\mathbf{B}}_{\mathbf{S}}$ is a linear transformation of the estimator in Theorem 4 and $\sqrt{n}$-consistency and asymptotic normality follow.

The asymptotic variance is a linear transformation of the asymptotic variance of the previous theorem. Define the $QT \times QT$ covariance matrix: $\operatorname{Var}\{\operatorname{vec}(\widehat{\mathbf{W}}_{\mathbf{S}})\} = \mathbf{R}$. Using the fact $\operatorname{vec}(\mathbf{ACB}) = (\mathbf{B}^\top \otimes \mathbf{A})\operatorname{vec}(\mathbf{C})$, we have

$$\operatorname{Var}\{\operatorname{vec}(\mathbf{I}_Q \widehat{\mathbf{W}}_{\mathbf{S}}\mathbf{B})\} = (\mathbf{B}^\top \otimes \mathbf{I}_Q)\mathbf{R}(\mathbf{B} \otimes \mathbf{I}_Q)$$

Now let $\mathbf{B}_{\mathbf{N}}$ be the rows of the full unmixing matrix corresponding to the Gaussian components. Restricting our attention to the block of this matrix corresponding to the covariance matrix for $\boldsymbol{b}_q$, then applying simplifications and the property that $\mathbf{B}_{\mathbf{N}}^\top\mathbf{B}_{\mathbf{N}} = \boldsymbol{\Sigma}^{-1} - \mathbf{B}_{\mathbf{S}}^\top\mathbf{B}_{\mathbf{S}}$, we obtain (S.28). $\qquad \square$
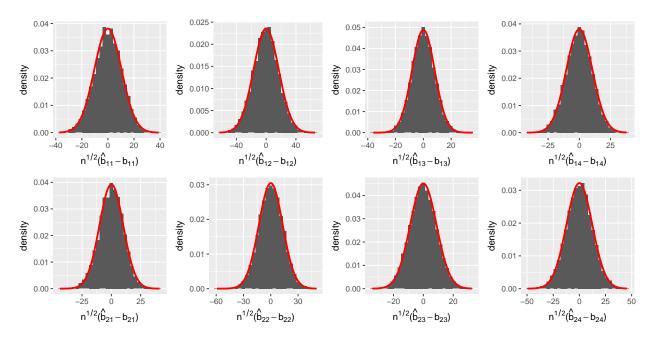
Figure S.1: Theoretical densities versus histograms of $\sqrt{n}(\hat{\boldsymbol{b}}_{qt}^{Logis} - \boldsymbol{b}_{qt})$ where $\widehat{\mathbf{B}}_{\mathbf{S}}^{Logis} = \widehat{\mathbf{W}}_{\mathbf{S}}^{Logis}\widehat{\mathbf{L}}$ from 10,000 simulations with n =10,000, $Q = 2$ with exponential and logistic densities, $T = 4$, and the true $\mathbf{B}$ is fixed at a randomly generated matrix.

Wei (2015) develop similar asymptotics for estimators using the theory of M-estimation without requiring affine equivalence; however, his approach does not readily extend to LNGCA and LCA. In particular, the identifiability issues created by the Gaussian components precludes the direct application to LNGCA. For $T = Q$, $\sum_{q'=1}^{Q} \boldsymbol{b}_{q'}\boldsymbol{b}_{q'}^{\top} = \boldsymbol{\Sigma}^{-1}$, and Corollary 2 is equivalent to Theorem 8 in Wei (2015) for the special case specified by our Assumption 5(i).

We validated the asymptotic approximation of the distribution of the Logis-LCA estimator on a finite sample through simulations. Here we present the results from a single random choice of $\mathbf{M}$ with 10,000 simulations, $n = 10,000$, $Q = 2$, and $T = 4$ in which the true densities were exponential and logistic. In Figure S.1, we can see that the histograms are in general agreement with the theoretical results.

C Additional Background

C.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 $\{\boldsymbol{x}_{\text{st},\,i}\}$, $i = 1, \ldots, n$ be the standardized data sample with $\boldsymbol{x}_i \in \mathbb{R}^T$, $\sum_{i=1}^n \boldsymbol{x}_{\text{st},\,i} = \mathbf{0}$, where $\mathbf{0}$ is the vector of T zeros, and $\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_{\text{st},\,i}^2 = \mathbf{1}$, 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):

$$\mathbf{w}_q = \underset{\mathbf{w} \in \mathbb{R}^T}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^n R(\mathbf{w}^\top \boldsymbol{x}_{\text{st},\,i}) - \int R(n)\phi(n)\,dn \right\}^2, \tag{S.29}$$

where $\mathbf{w}$ is orthogonal to $\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_{q-1}$ and $||\mathbf{w}|| = 1$ with $||\cdot||$ denoting the L2-norm, R is a non-linear function (in likelihood-based ICA, $R = \log f(x)$), and $\phi(n)$ is the standard normal density. A common choice for R 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 the data. NGCA can be formulated using a semiparametric likelihood,

$$f_{\mathbf{X}}(\boldsymbol{x}) = h^*(\mathbf{B}_{\mathbf{S}}\boldsymbol{x})\phi_{\mathbf{0},\boldsymbol{\Sigma}}(\boldsymbol{x}) \tag{S.30}$$

where $\phi_{\mathbf{0},\boldsymbol{\Sigma}}$ is multivariate normal with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$; $\mathbf{B}_{\mathbf{S}}$ is a $Q \times T$ matrix; and $h^*(\cdot)$ is a function that captures departures from Gaussianity under the constraint that $f_{\mathbf{X}}(\boldsymbol{x})$

is a density. NGCA does not assume linear mixing of independent factors, and consequently the factors are not identifiable. Thus we do not consider it in our simulations.

The density in the Spline-LCA model can be considered an extension of (S.30) with the additional assumption of independence.

Proposition 6. *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*

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

where $\phi_{\mathbf{0},\boldsymbol{\Sigma}}$ *is the mean zero multivariate distribution with covariance* $\boldsymbol{\Sigma} = \mathbf{L}^{-2}$.

Proof. Using the tilted Gaussian density, we have

$$
\begin{aligned}
f_{\mathbf{X}}(\boldsymbol{x}) &= \det \mathbf{L} \prod_{q=1}^{Q} e^{g_q(\mathbf{w}_q^\top \mathbf{L}\boldsymbol{x})} \phi(\mathbf{w}_q^\top \mathbf{L}\boldsymbol{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{Q+k}^\top \mathbf{L}\boldsymbol{x}) \\
&= \left\{ \prod_{q=1}^{Q} e^{g_q(\mathbf{w}_q^\top \mathbf{L}\boldsymbol{x})} \right\} (2\pi)^{-T/2} (\det \mathbf{L}) \exp \left\{ -\frac{1}{2} \sum_{k=1}^{T} \boldsymbol{x}^\top \mathbf{L}\mathbf{w}_k \mathbf{w}_k^\top \mathbf{L}\boldsymbol{x} \right\} \\
&= (\det \boldsymbol{\Sigma})^{-1/2} (2\pi)^{-T/2} \exp \left\{ -\frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \right\} \prod_{q=1}^{Q} e^{g_q(\mathbf{w}_q^\top \mathbf{L}\boldsymbol{x})}.
\end{aligned}
$$

$\square$

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.

C.2 Noisy ICA and IFA

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

$$\mathbf{X} = \mathbf{M_S}\mathbf{S} + \mathbf{E} \tag{S.31}$$

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

Assume that $\mathbf{\Psi} = \sigma^2\mathbf{I}$. Let $d_1, \ldots, d_Q$ denote the eigenvalues from the covariance matrix of $\mathbf{M_S}\mathbf{S}$ and let $d_{\epsilon_1}, \ldots, d_{\epsilon_T}$ denote the eigenvalues from the decomposition of $\mathbf{E}$. Under the assumption of isotropic noise, we have $d_{\epsilon_i} = \sigma^2$ for all $i, j = 1, \ldots, T$. Then the eigenvalue decomposition can be written as

$$\mathrm{Cov}\,\mathbf{X} = \mathbf{U}\,\mathrm{diag}(d_1 + \sigma^2, \ldots, d_Q + \sigma^2, \sigma^2, \ldots, \sigma^2)\,\mathbf{U}^\top. \tag{S.32}$$

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

In IFA, (S.31) is estimated under the assumption that the densities of the ICs are Gaussian mixtures (Attias, 1999). In its original formulation, $\mathbf{\Psi}$ 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 qth IC comes from the first class, where the first class has a normal distribution with mean μ_{q1} and variance ρ_{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 $\rho_{q2} = \frac{1 - \pi_{q1}\rho_{q1} - \pi_{q1}\mu_{q1}^2}{\pi_{q2}} - \mu_{q2}^2$, respectively.

Then the joint density of $\mathbf{X}$ can be written

$$f_{\mathbf{X}}(\boldsymbol{x} \mid \mathbf{M_S}) = \prod_{t=1}^{T} \int \phi_{0,\sigma^2}\left(x_t - \mathbf{m}_t^\top \boldsymbol{s}\right) f_{\mathbf{S}}(\boldsymbol{s}) \, d\boldsymbol{s}, \tag{S.33}$$

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

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

Analytic integration across $\boldsymbol{s}$ 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 $\mathbf{k}_j = \{k_1, \ldots, k_Q\}$ denote an element of $\mathcal{K}$, where $j \in \{1, \ldots, 2^Q\}$. Let $\boldsymbol{\mu}(\mathbf{k}_j)$ and $\boldsymbol{\rho}(\mathbf{k}_j)$ denote the conditional means of $\boldsymbol{s}$ given the states $\mathbf{k}_j$. Now define

$$\boldsymbol{\Sigma}(\mathbf{k}_j) = \mathbf{M_S}\,\mathrm{diag}\{\boldsymbol{\rho}(\mathbf{k}_j)\}\,\mathbf{M_S}^\top + \sigma^2\mathbf{I}$$

and

$$\boldsymbol{\mu}^*(\mathbf{k}_j) = \mathbf{M_S}\boldsymbol{\mu}(\mathbf{k}_j).$$

Then the density is

$$f_{\mathbf{X}}(\boldsymbol{x} \mid \mathbf{M_S}) = \sum_{\mathbf{k}_j \in \mathcal{K}} \Phi\{\boldsymbol{x} \mid \boldsymbol{\mu}^*(\mathbf{k}_j), \boldsymbol{\Sigma}(\mathbf{k}_j)\} \prod_{q=1}^{Q} \pi_{q1}^{k_q}\pi_{q2}^{1-k_q} \tag{S.34}$$

with $\Phi\{\boldsymbol{x} \mid \boldsymbol{\mu}^*(\mathbf{k}_j), \boldsymbol{\Sigma}(\mathbf{k}_j)\}$ 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.34), and given some $\widehat{\mathbf{M_S}}$, 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.

D   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 $\widehat{\mathbf{W}}$. Recall $r_q(x)$ and $r_q'(x)$ are the first and second derivatives of $\log f_q(x)$. Algorithm 1 provides details on estimating $\widehat{\mathbf{W}}_{\mathbf{S}}$.

Algorithm 2: The fastICA algorithm (symmetric fixed point) for LCA.

Inputs : The whitened $n \times T$ data matrix $\mathbf{X}_{\mathrm{st}}$; initial $\mathbf{W}_{\mathbf{S}}^0$; tolerance ϵ.
Result: Estimates of the unmixing matrix, $\widehat{\mathbf{W}}_{\mathbf{S}}$, and latent components, $\widehat{\mathbf{S}} = \mathbf{X}_{\mathrm{st}} \widehat{\mathbf{W}}_{\mathbf{S}}^\top$.

1. Let $\mathbf{S}^0 = \mathbf{X}_{\mathrm{st}} \mathbf{W}_{\mathbf{S}}^{0\top}$ and let $(m) = 0$, where (m) denotes the number of update steps.

2. For each row $\mathbf{w}_q$, $q = 1, \dots, Q$, of $\mathbf{W}_{\mathbf{S}}$, calculate

$$\mathbf{w}_q^* = \frac{1}{n} \sum_{i=1}^{n} \left\{ r_q(\mathbf{w}_q^{(m)\top} \boldsymbol{x}_{\mathrm{st},i}) \boldsymbol{x}_{\mathrm{st},i} - r_q'(\mathbf{w}_q^{(m)\top} \boldsymbol{x}_{\mathrm{st},i}) \mathbf{w}_q^{(m)} \right\}$$
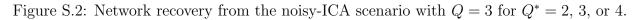
3. Calculate the thin SVD of $\mathbf{W}_{\mathbf{S}}^* = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top}$.

4. Let $\mathbf{W}^{(m+1)} = \mathbf{U}^* \mathbf{V}^{*\top}$.

5. If $PMSE(\mathbf{W}_{\mathbf{S}}^{(m+1)\top}, \mathbf{W}_{\mathbf{S}}^{(m)\top}) < \epsilon$, stop, else increment (m) and repeat (2)-(4).

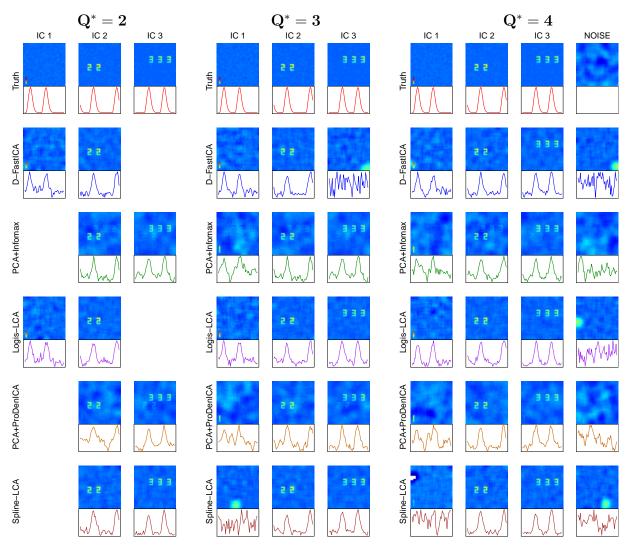# E   Supplemental materials for simulations examining distributional and noise-rank assumptions

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 $\pi_{11}, \pi_{21}, \mu_{11}, \mu_{21}, \rho_{11}, \rho_{21}$ 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 $\sigma^2 = 1$.
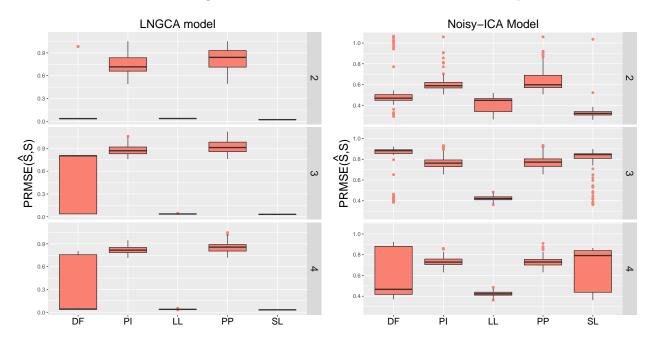
The matrices $\mathbf{M_S}$ and $\mathbf{M_N}$ were generated by first simulating a 5×5 matrix with standard normal entries, taking the singular value decomposition (SVD), then creating a diagonal matrix with five singular values from a uniform(1,10) distribution, followed by multiplying the left singular 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.

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

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 $\widehat{\mathbf{U}}_{1:Q}^{\top}$ denote the first Q rows from $\widehat{\mathbf{U}}^{\top}$ in the decomposition $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{U}}^{\top}$. Then constraining the initial matrix, $\mathbf{W}_{\mathbf{S}}^{0}$, to the principal subspace is equivalent to $\mathbf{W}_{\mathbf{S}}^{0} = \mathbf{O}\widehat{\mathbf{U}}_{1:Q}^{\top}$ where $\mathbf{O}$ is a random $Q \times Q$ orthogonal matrix.

Figure S.3: Boxplots of $PRMSE$ for estimated columns of **S** from simulations of spatial sources 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.

F Supplemental figures for the spatio-temporal sources

The permutation-invariant root mean squared errors for the components estimated from the spatio-temporal source simulations are much lower for Logis-LCA and Spline-LCA when the noise rank is $T - Q$ (Figure S.3). 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.

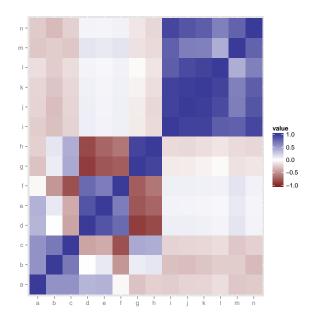# G Supplemental materials for Section: Data Visualization and Dimension Reduction

Silva et al. (2013) generated covariates from photographs of leaf samples from thirty species (Figure S.4). Many of these covariates are highly correlated (Figure S.5).

Logis-LCA and Spline-LCA reveal features in the data (Figures S.6, S.7), while PCA+Infomax

Figure S.4: 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 3 and Supplemental Figures S.6 and S.7). Figure from Silva et al. (2013).

Figure S.5: 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.

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 than the components from PCA+ICA methods.

H Supplemental materials for Section: Application to fMRI

We analyzed task data from the theory of mind experiment in the HCP dataset. 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

Figure S.6: 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.
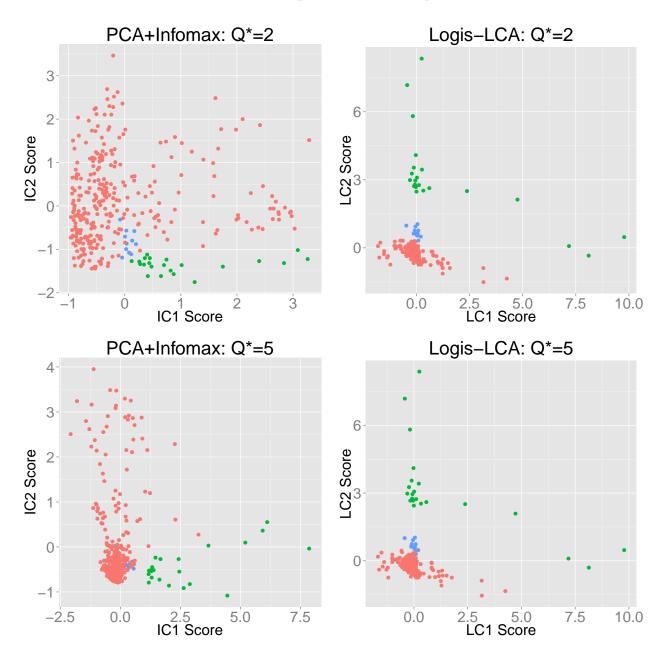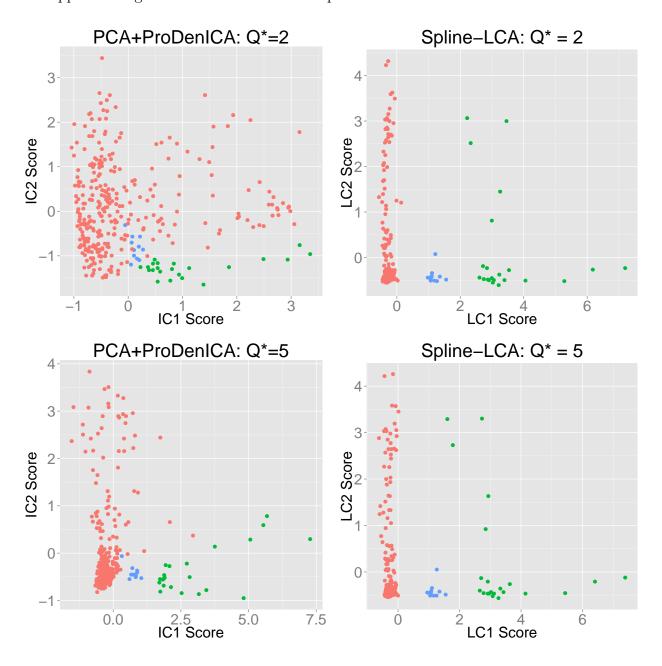
Figure S.7: Components in the leaf data from PCA-ProDenICA and Spline-LCA when two components were estimated and when five components were estimated (when five components were estimated, the two components with the highest marginal likelihood are plotted). The green dots correspond to *Podocarpus sp.* 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 3 of the main manuscript.

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).

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 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 (i.e., brain images) 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) in which the sessions differed in phase-encoding direction (right-left versus left-right). 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 resulting in 272 time points for each voxel. After vectorization, the voxels were standardized across time to have mean zero and unit variance.

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. 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.

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 (Figure S.8). 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.

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.8).
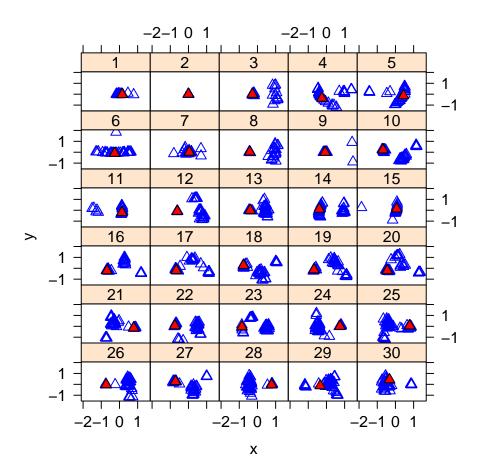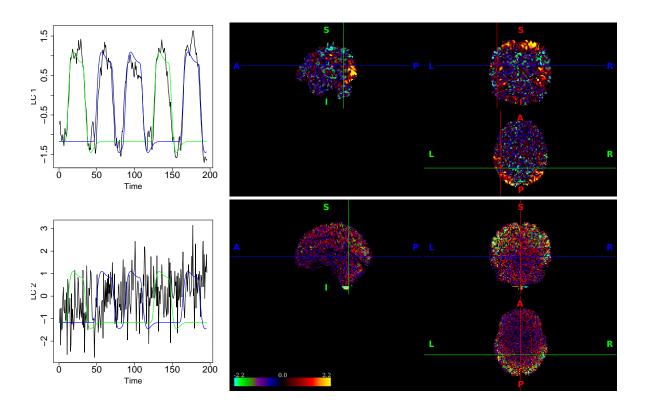
Figure S.8: Multidimensional scaling of $||\widehat{\mathbf{S}}_j^{(k)} - \widehat{\mathbf{S}}_j^{(\ell)}||_F$ for components $j = 1, \ldots, 30$ and initializations $k \neq \ell \in \{1, \ldots, 56\}$. The coordinates corresponding to the initialization with the highest likelihood are depicted by solid red triangles. In all instances, the red triangle appears in a cluster of other triangles, indicating agreement between a subset of initializations.

Components one, two, and nine were relatively robust to initialization and contained only one (main) basin of attraction.

We examined the correlation between the loadings (columns of $\widehat{\mathbf{M_S}}$) 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 S.9). 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 S.9 are located near the TPJ) but is not activated in this component, suggesting there exists additional signal in other components.

Figure S.9: Selected components 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).
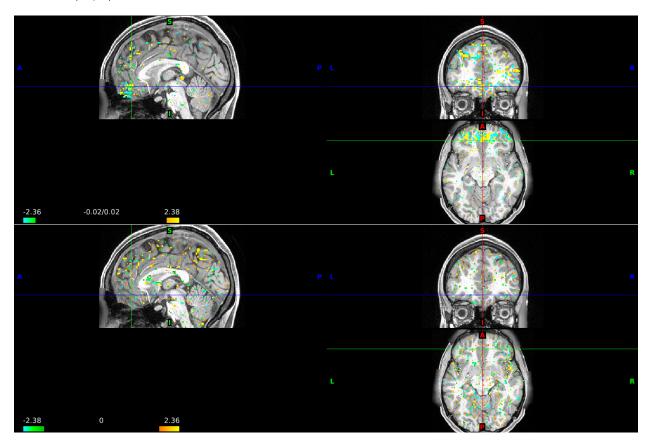
Voxels were highly activated in the brainstem and the component's time course was correlated with three of the motion parameters from the rigid-body alignment ($r = 0.32$, 0.32, and 0.42 for the x-transformation, x-rotation, and z-rotation parameters, respectively). This may be related to a gradual relaxation of the neck or spine over the course of the subject's session. Additionally, there was a positive correlation with time ($r = 0.44$), which could also be related to scanner drift.

LCA also identified a type of artifact that did not seem to be found in PCA+ProDenICA. Some components had alternating bands of positive and negative values, in particular in axial slices through orbitofrontal regions (Figure S.10). 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.10: Artifact (component 14) identified using Spline-LCA (top) and the matched component from PCA+ProDenICA (bottom; correlation = 0.08) in subject 100307. Thresholded at $|s_{v,14}| > 1.75$.

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.

References

Ashburner, J., Friston, K., and Penny, W. (2004). Part II – Imaging Neuroscience – Theory and Analysis. In Frackowiak, R., editor, *Human Brain Function*. Academic Press, 2nd edition.

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.

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.

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.

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.

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. and Tibshirani, R. (2010). *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*. R package version 1.0.

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.

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.

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.

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., and Virta, J. (2017). The squared symmetric fastica estimator. *Signal Processing*, 131:402–411.

Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H., et al. (2015). Fourth moments and independent component analysis. *Statistical science*, 30(3):372–390.

Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., and Ollila, E. (2011). Deflation-based fastica reloaded. In *Signal Processing Conference, 2011 19th European*, pages 1854–1858. IEEE.

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.

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).

van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Virta, J., Nordhausen, K., and Oja, H. (2016). Projection pursuit for non-gaussian independent components. *arXiv preprint arXiv:1612.05445*.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.

Wei, T. (2015). A convergence and asymptotic analysis of the generalized symmetric FastICA algorithm. *IEEE Transactions on Signal Processing*, 63(24):6445–6458.