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 $\stackrel{d}{=}$ denote equal in distribution. First consider the theorem on uniqueness of decomposition.

Theorem 10.3.9 from Kagan et al. (1973). *Let $\mathbf{X} = \mathbf{AY}$ be a structural representation of $\mathbf{X}$ and let the columns of $\mathbf{A}$ be linearly independent. Then $\mathbf{X}$ can be expressed as $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent, $\mathbf{X}_1$ has essentially unique structure, and $\mathbf{X}_2$ is multivariate normal with a non-unique structure. Moreover, this decomposition is unique in the sense that if $\mathbf{X} = \mathbf{Z}_1 + \mathbf{Z}_2$ is another decomposition, where $\mathbf{Z}_1$ has essentially unique structure, $\mathbf{Z}_2$ is multivariate normal, and $\mathbf{Z}_1$ is independent of $\mathbf{Z}_2$, then $\mathbf{Z}_1 \stackrel{d}{=} \mathbf{X}_1$ and $\mathbf{Z}_2 \stackrel{d}{=} \mathbf{X}_2$ 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 $\mathbf{Z} \stackrel{d}{=} \mathbf{X}$. Consider their structural representations: $\mathbf{Z} = \mathbf{M_S S}$ and $\mathbf{X} = \mathbf{M}^*_{\mathbf{S}} \mathbf{S}^*$ where $\mathbf{M_S} \in \mathbb{R}^{T \times Q}$ and $\mathbf{M}^*_{\mathbf{S}} \in$*

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

Proof. We have $\mathbf{M_S S} \overset{d}{=} \mathbf{M_S^* S^*}$. 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 $\mathbf{S} \overset{d}{=} \mathbf{BS}^*$. 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 $\mathbf{R} \overset{d}{=} \mathbf{S}^*$, 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 $\mathbf{S}^* \overset{d}{=} \mathbf{S}$ up to scaling and permutations.

Now consider the scaling and permutation such that $\mathbf{S}^* \overset{d}{=} \mathbf{S}$. 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:* $\mathbf{M_S^*} \cong \mathbf{M_S}$*;* $\mathbf{S}^* \overset{d}{=} \mathbf{S}$ *up to scaling and permutations;* $\mathbf{M_S S} \overset{d}{=} \mathbf{M_S^* S^*}$*; and* $\mathbf{E}^* \overset{d}{=} \mathbf{M_N N}$.

Proof. Since $\mathbf{X}$ has a unique decomposition in the sense of Theorem 10.3.9, we have $\mathbf{M_S S} \overset{d}{=} \mathbf{M_S^* S^*}$ and $\mathbf{M_N N} \overset{d}{=} \mathbf{E}^*$. 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

10.3.3 in Kagan et al. (1973), $\mathbf{Z}$ has the decomposition $\mathbf{Z} = \mathbf{K}_1\mathbf{Y}_1 + \mathbf{K}_2\mathbf{Y}_2$ in which $\mathbf{Y}_1$ are independent non-Gaussian and $\mathbf{Y}_2$ are Gaussian. Then from Theorem 1 and the assumption of unit variance, we have that $\mathbf{Y}_1 \stackrel{d}{=} \mathbf{S}$ (up to ordering), and it follows that there exists a subset of $g_1, \ldots, g_T$ equal to $f_1, \ldots, f_Q$. Also from Theorem 1, we have $\mathbf{K}_1 \cong \mathbf{A}_\mathbf{S}$. Note that $\mathbf{K} \in \mathcal{O}_{T\times T}$ since $\mathrm{E}\,\mathbf{Y}\mathbf{Y}^\top = \mathbf{I}$ 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 $\mathbf{K}_1\mathbf{P}_\pm = \mathbf{A}_\mathbf{S}$. Note that $\mathbf{W}_\mathbf{S} = \mathbf{A}_\mathbf{S}^\top$. Define $\mathbf{R}_\mathbf{S} = \mathbf{K}_1^\top$. Then $\mathbf{P}_\pm^\top\mathbf{R}_\mathbf{S} = \mathbf{W}_\mathbf{S}$. □

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,

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

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

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

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

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

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)

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

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

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

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

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

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

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{1+\alpha} &\leq \frac{1}{n}\sum_{i=1}^{n} \left\{ \|\widehat{\mathbf{L}}\| \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\| \right\}^{1+\alpha} \\
&\leq \|\widehat{\mathbf{L}}\|^{1+\alpha} \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^{1+\alpha} \qquad \text{(S.6)}
\end{aligned}$$

Now consider

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^{1+\alpha} &\leq \frac{1}{n}\sum_{i=1}^{n} (\|\boldsymbol{x}_i\| + \|\bar{\boldsymbol{x}}\|)^{1+\alpha} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} (2\|\boldsymbol{x}_i\|)^{1+\alpha} + (2\|\bar{\boldsymbol{x}}\|)^{1+\alpha} \qquad \text{(S.7)}
\end{aligned}$$

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. □

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

$$\begin{aligned}\sup_{\mathbf{O_S}\in B} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{O_S}(\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) - \mathbf{L}\boldsymbol{x}_i)\| &\leq \sup_{\mathbf{O_S}\in B} \|\mathbf{O_S}\| \frac{1}{n}\sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \\ &\leq \sqrt{Q}\frac{1}{n}\sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\|.\end{aligned}$$

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

$$\begin{aligned}&\sup_{\mathbf{O_S}\in B} \frac{1}{n}\sum_{i=1}^{n} \left\| h(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) - h(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) \right\| \leq MQ\Bigg\{ \frac{1}{n}\sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \\ + \quad &\frac{1}{n}\sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \, \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^{\alpha} + \frac{1}{n}\sum_{i=1}^{n} \|\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \mathbf{L}\boldsymbol{x}_i\| \, \|\mathbf{L}\boldsymbol{x}_i\|^{\alpha} \Bigg\}. \quad \text{(S.8)}\end{aligned}$$

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

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

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

□

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).

(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 $\mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{O_S L X}) \leq \mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{W_S L X}) + \alpha$ 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

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

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 $\mathbf{O_S} \ncong \mathbf{W_S}$.

To show condition (v) is satisfied,

$$\begin{aligned}\mathcal{J}_n(\widehat{\mathbf{W}}_{\mathbf{S}}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) &\geq \mathcal{J}_n(\mathbf{W}_{\mathbf{S}}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) && \text{(by definition)} \\ &= \mathcal{J}_n(\mathbf{W}_{\mathbf{S}}\mathbf{L}\boldsymbol{x}_i) - o_p(1). && \text{(Proposition 3)}\end{aligned}$$

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 $B = \{\mathbf{O}_{\mathbf{S}} \in \mathcal{O}_{Q\times T} : \bigcap_{\mathbf{W}_{\mathbf{S}}^*\in\mathcal{W}_{\mathbf{S}}} ||\mathbf{O}_{\mathbf{S}} - \mathbf{W}_{\mathbf{S}}^*|| \geq \epsilon\}$, 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.

$$\begin{aligned}
\sup_{\mathbf{O_S}\in B} \mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) &\leq \sup_{\mathbf{O_S}\in B} \mathcal{J}_n(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) + o_p(1) && \text{(from Proposition 4)} \quad \text{(S.9)} \\
&\leq \sup_{j=1,\ldots,p} \sup_{\mathbf{O_S}\in U_j} \mathcal{J}_n(\mathbf{O_S}\mathbf{L}\boldsymbol{x}_i) + o_p(1) \\
&\leq \sup_{j=1,\ldots,p} \frac{1}{n}\sum_{i=1}^{n} m_{U_j}(\boldsymbol{x}_i) + o_p(1) \\
&\to \sup_{j=1,\ldots,p} \mathrm{E}\, m_{U_j}(\mathbf{X}) && \text{(law of large numbers)} \\
&< \quad \mathrm{E}\, \log f_{\mathbf{S}}(\mathbf{W_S}\mathbf{L}\mathbf{X}). && \text{(S.10)}
\end{aligned}$$

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

$$\begin{aligned}
\sup_{\mathbf{O_S}\in B} \mathcal{J}_n(\mathbf{O_S}\widehat{\mathbf{L}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})) &\geq \mathcal{J}_n(\mathbf{W_S}\mathbf{L}\boldsymbol{x}_i) - o_p(1) && \text{(from condition (v.))} \\
&= \mathrm{E}\, \log f_{\mathbf{S}}(\mathbf{W_S}\mathbf{L}\mathbf{X}) - o_p(1), && \text{(from LLN)}
\end{aligned}$$

which would imply the following relationship between events:

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

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

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

□

Next we describe conditions for consistency when the density used in the objective function may not be equal to the density of the LCs. We first present a result that is contained in the proof of Theorem 1 in Hyvärinen and Oja (1998), where here the nonlinearity is equal

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

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

$$\begin{aligned}
&\mathrm{E}\,\log p_1[(\boldsymbol{e}_1+\boldsymbol{\epsilon})^\top \mathbf{Z}] = \\
&\mathrm{E}\,\log p_1(S_1) + \boldsymbol{\epsilon}^\top \boldsymbol{e}_1 \mathrm{E}\, S_1 r_1(S_1) + \frac{1}{2}\boldsymbol{\epsilon}^\top \mathrm{diag}\left[\mathrm{E}\, S_1^2 r_1'(S_1), \mathrm{E}\, r_1'(S_1), \ldots, \mathrm{E}\, r_1'(S_1)\right]\boldsymbol{\epsilon} + o(||\boldsymbol{\epsilon}||^2) = \\
&\mathrm{E}\,\log p_1(S_1) + \epsilon_1 \mathrm{E}\, S_1 r_1(S_1) + \frac{1}{2}\epsilon_1^2 \mathrm{E}\, S_1^2 r_1'(S_1) + \frac{1}{2}\mathrm{E}\, r_1'(S_1)\sum_{q>1}\epsilon_q^2 + o(||\boldsymbol{\epsilon}||^2).
\end{aligned}$$

Note that $\epsilon_1 = \sqrt{1-\sum_{q>1}\epsilon_q^2} - 1$. Now we consider the first-order Taylor series expansion of $\sqrt{1-\gamma}$ about 0 which is $1-\gamma/2+o(||\gamma||)$, so $\epsilon_1 = -\frac{1}{2}\sum_{q>1}\epsilon_q^2 + o(\sum_{q>1}\epsilon_q^2)$. By Assumption 5(ii), $|\mathrm{E}\, S_1^2 r_1'(S_1)| < \infty$. Then we can write

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

□

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

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

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

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

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 $\mathcal{G}_{Q\times T} = \left\{\mathbf{G}\in\mathbb{R}^{Q\times T} : \operatorname{diag}\mathbf{G}\mathbf{G}^\top = \mathbf{1}_Q\right\}$. 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 $\mathcal{N}_{\epsilon^*}(\mathbf{W_S})$. □

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. □

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* q*th component. Then (i)* $\int\phi(u)e^{g_q(u)}\,du = 1$ *and (ii)* $\int u\phi(u)e^{g_q(u)}\,du = 0$ *for each* q.

Proof. It suffices to consider the case $Q^* = 1$. Let $\mathbf{o}_1$ be given. Let G be the set of cubic splines and note that for any $g\subset G$, 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 $\partial(\int\phi(u)e^{g(u)}du)/\partial\theta_0 =$

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

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:

$$\begin{aligned}
\beta_q &= \mathrm{E}\,S_q^4\\
\eta_q &= \mathrm{E}\,r(S_q)\\
\xi_q &= \mathrm{E}\,r(S_q)^2 - \eta_q^2\\
\lambda_q &= \mathrm{E}\,r(S_q)S_q\\
\delta_q &= \mathrm{E}\,r'(S_q)
\end{aligned}$$

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:

$$\mathcal{L}(\mathbf{C}_{\mathbf{S}}, \boldsymbol{\Theta}) = \sum_{q=1}^{Q} \mathbb{E}_n \left\{ \log p_q(\boldsymbol{c}_q^\top (\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \right\} - \sum_{q=1}^{Q} \frac{\theta_{qq}}{2} (\boldsymbol{c}_q^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{c}_q - 1) - \sum_{q=1}^{Q-1} \sum_{q'=q+1}^{Q} \theta_{qq'} \boldsymbol{c}_q^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{c}_{q'}. \tag{S.12}$$

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

$$\mathcal{L}(\mathbf{O}_{\mathbf{S}}, \boldsymbol{\Theta}) = \sum_{q=1}^{Q} \mathbb{E}_n \left\{ \log p_q(\mathbf{o}_q^\top \widehat{\mathbf{L}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})) \right\} - \sum_{q=1}^{Q} \frac{\theta_{qq}}{2} (\mathbf{o}_q^\top \mathbf{o}_q - 1) - \sum_{q=1}^{Q-1} \sum_{q'=q+1}^{Q} \theta_{qq'} \mathbf{o}_q^\top \mathbf{o}_{q'}. \tag{S.13}$$

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*

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

$$\sqrt{n}(\hat{\boldsymbol{e}}_{qq'}) = \sqrt{n}\frac{\mathbb{E}_n\left\{(r_q(s_{iq})-\eta_q)s_{iq'} - (r_{q'}(s_{iq'})-\eta_{q'})s_{iq} - (\delta_{q'}-\lambda_q)s_{iq}s_{ir}\right\}}{\delta_q - \lambda_q + \delta_{q'} - \lambda_{q'}} + o_p(1),\ q, q' \leq Q \tag{S.14}$$

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

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

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

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

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

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

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

Next, write (S.19) as

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

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

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

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

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

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:

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

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

$$\sqrt{n}\hat{\boldsymbol{e}}_{qr} = \sqrt{n}\frac{1}{\lambda_q-\delta_q}\left[\boldsymbol{e}_r^\top\mathbb{E}_n\,\left\{r_q(\boldsymbol{e}_q^\top\boldsymbol{x}_i)-\eta_q\right\}\boldsymbol{x}_i - \lambda_q\mathbb{E}_n\,(x_{iq}-\bar{x}_q)(x_{ir}-\bar{x}_r)\right] + o_p(1). \tag{S.24}$$

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 $\hat{\boldsymbol{e}}_q$:

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

This gives us the estimating equations:

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

The estimating equation in (S.26) is also found in symmetric FastICA (Miettinen et al., 2015, 2017; Wei, 2015) but here restricted to $q, q' \leq Q$, 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* $\mathrm{E}\,\mathbf{X} = \mathbf{0}$. *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* $q \leq Q$, $\sqrt{n}(\hat{\boldsymbol{e}}_q - \boldsymbol{e}_q) \Rightarrow \mathcal{N}(0, \mathbf{R}_q)$ *with*

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

Proof. Asyptotic 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 $q \leq Q$ 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*

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

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.

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*

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

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

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

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). □

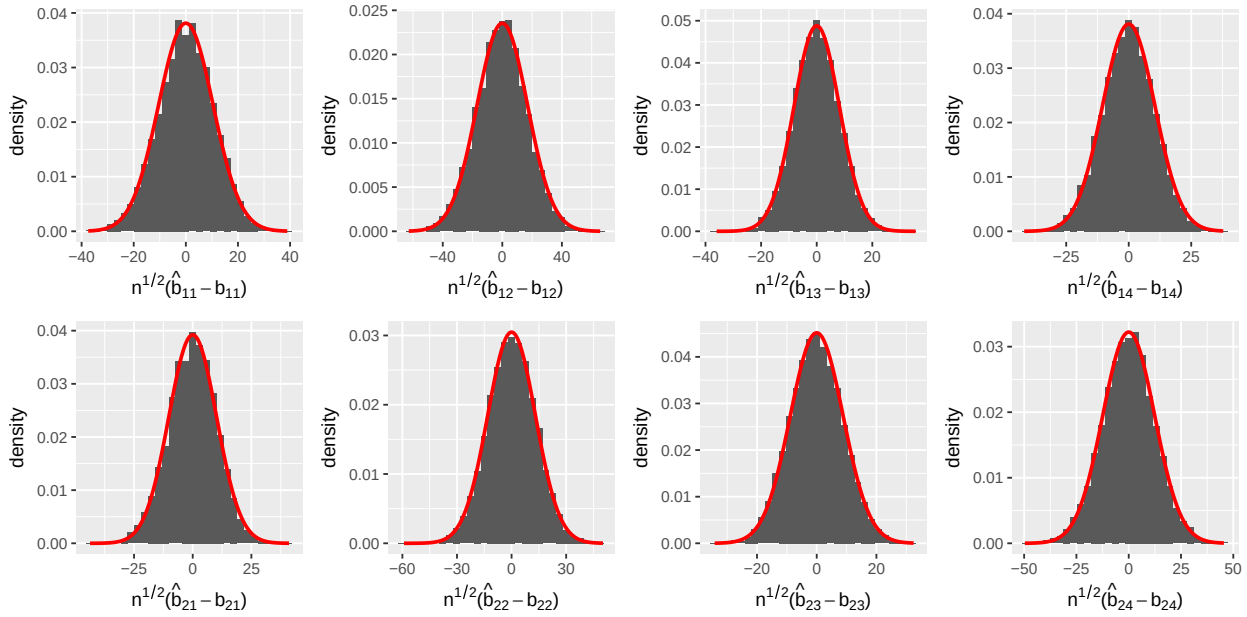


Figure S.1: Theoretical densities versus histograms of $\sqrt{n}(\hat{\boldsymbol{b}}_{qt}^{Logis} - \boldsymbol{b}_{qt})$ where $\widehat{\mathbf{B}}_{\mathbf{S}}^{Logis} = \widehat{\mathbf{W}}_{\mathbf{S}}^{Logis}\widehat{\mathbf{L}}$ from 10,000 simulations with n =10,000, $Q = 2$ with exponential 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 equivariance; 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 semi-parametric 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

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

□

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),

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

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 $\mathbf{M_S S}$ 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

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

Let $\mathbf{X}_{\text{data}}$ be the $n \times T$ data matrix. In PCA+ICA, noise-free ICA is applied to the first Q left singular vectors of $\mathbf{X}_{\text{data}}$ 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 $\Phi\{\boldsymbol{x} \mid \boldsymbol{\mu}^*(\mathbf{k}_j), \boldsymbol{\Sigma}(\boldsymbol{k}_j)\}$ multivariate normal with mean $\boldsymbol{\mu}^*(\mathbf{k}_j)$ and variance $\boldsymbol{\Sigma}(\boldsymbol{k}_j)$ (see (16) and (17) in Attias 1999). Then a likelihood can be constructed from (S.34), and given some $\widehat{\mathbf{M}_\mathbf{S}}$, the ICs can be estimated from their conditional means. Alternatively, maximum a posteriori estimates of the ICs could be obtained, though we pursue the former here.

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 $\mathbf{X}_{\text{st}}$; initial $\mathbf{W}_{\mathbf{S}}^0$; tolerance ϵ.

Result: Estimates of the unmixing matrix, $\widehat{\mathbf{W}}_{\mathbf{S}}$, and latent components, $\widehat{\mathbf{S}} = \mathbf{X}_{\text{st}}\widehat{\mathbf{W}}_{\mathbf{S}}^\top$.

1. Let $\mathbf{S}^0 = \mathbf{X}_{\text{st}}\mathbf{W}_{\mathbf{S}}^{0\top}$ and let $(m) = 0$, where (m) denotes the number of update steps.
2. For each row $\mathbf{w}_q$, $q = 1, \ldots, Q$, of $\mathbf{W}_{\mathbf{S}}$, calculate

$$\mathbf{w}_q^* = \frac{1}{n}\sum_{i=1}^{n}\left\{r_q(\mathbf{w}_q^{(m)\top}\boldsymbol{x}_{\text{st},i})\boldsymbol{x}_{\text{st},i} - r'_q(\mathbf{w}_q^{(m)\top}\boldsymbol{x}_{\text{st},i})\mathbf{w}_q^{(m)}\right\}$$

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).

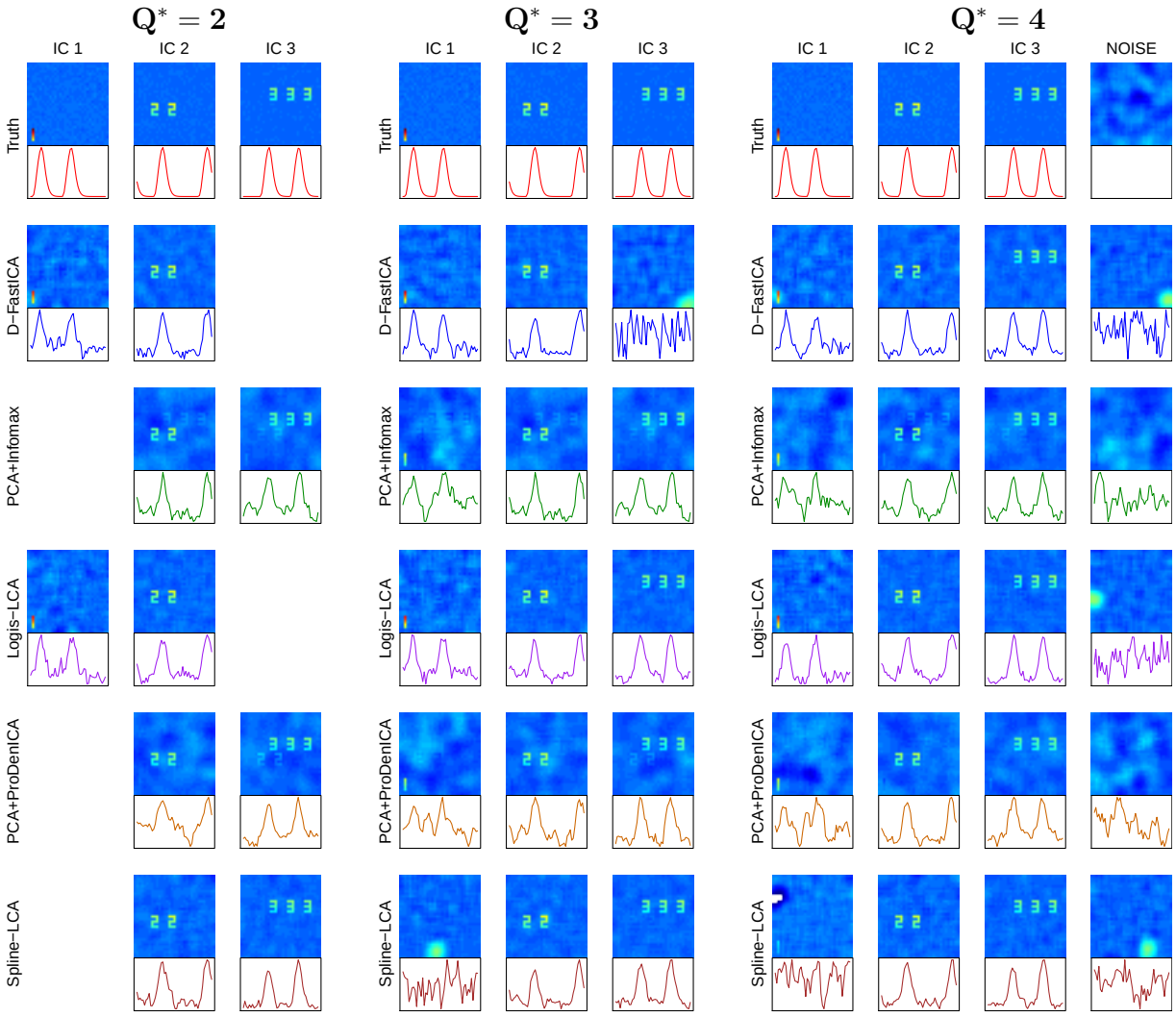
E Supplemental materials for simulations examining distributional and noise-rank assumptions

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 $\sigma^2 = 1$.

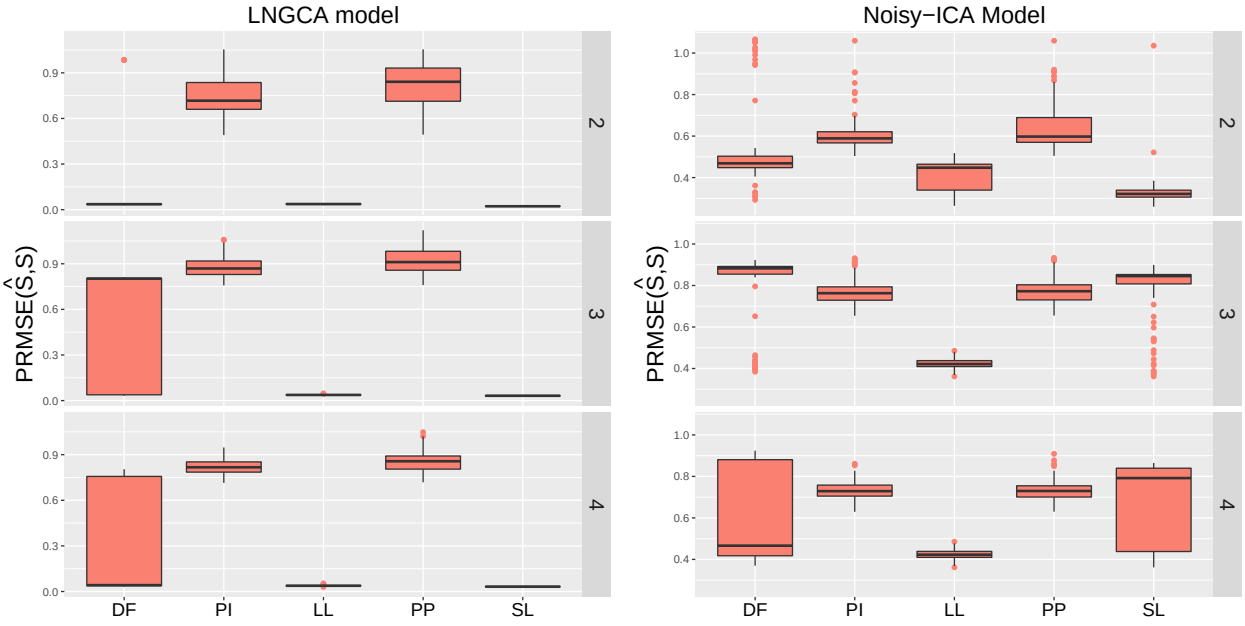
The matrices $\mathbf{M_S}$ and $\mathbf{M_N}$ were generated by first simulating a 5×5 matrix with standard normal entries, taking the singular value decomposition (SVD), then creating a diagonal matrix with five singular values from a uniform(1,10) distribution, followed by multiplying the left singular 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 $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^\top$. 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: Box plots of $PRMSE$ for estimated columns of $\mathbf{S}$ 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.

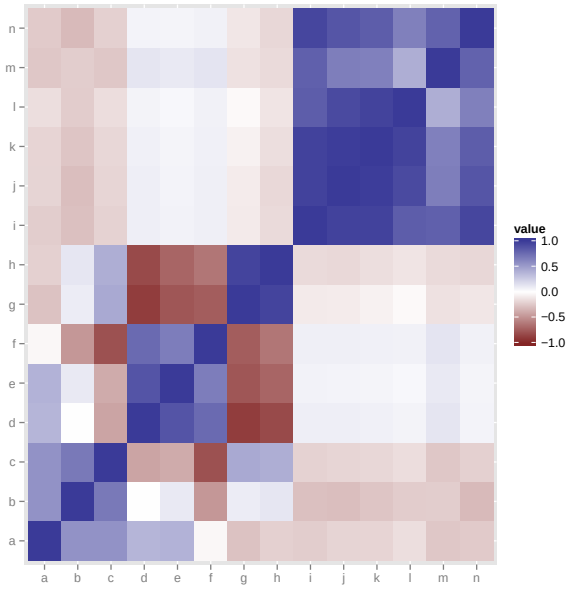
G Supplemental materials for Section: Data Visualization and Dimension Reduction

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.

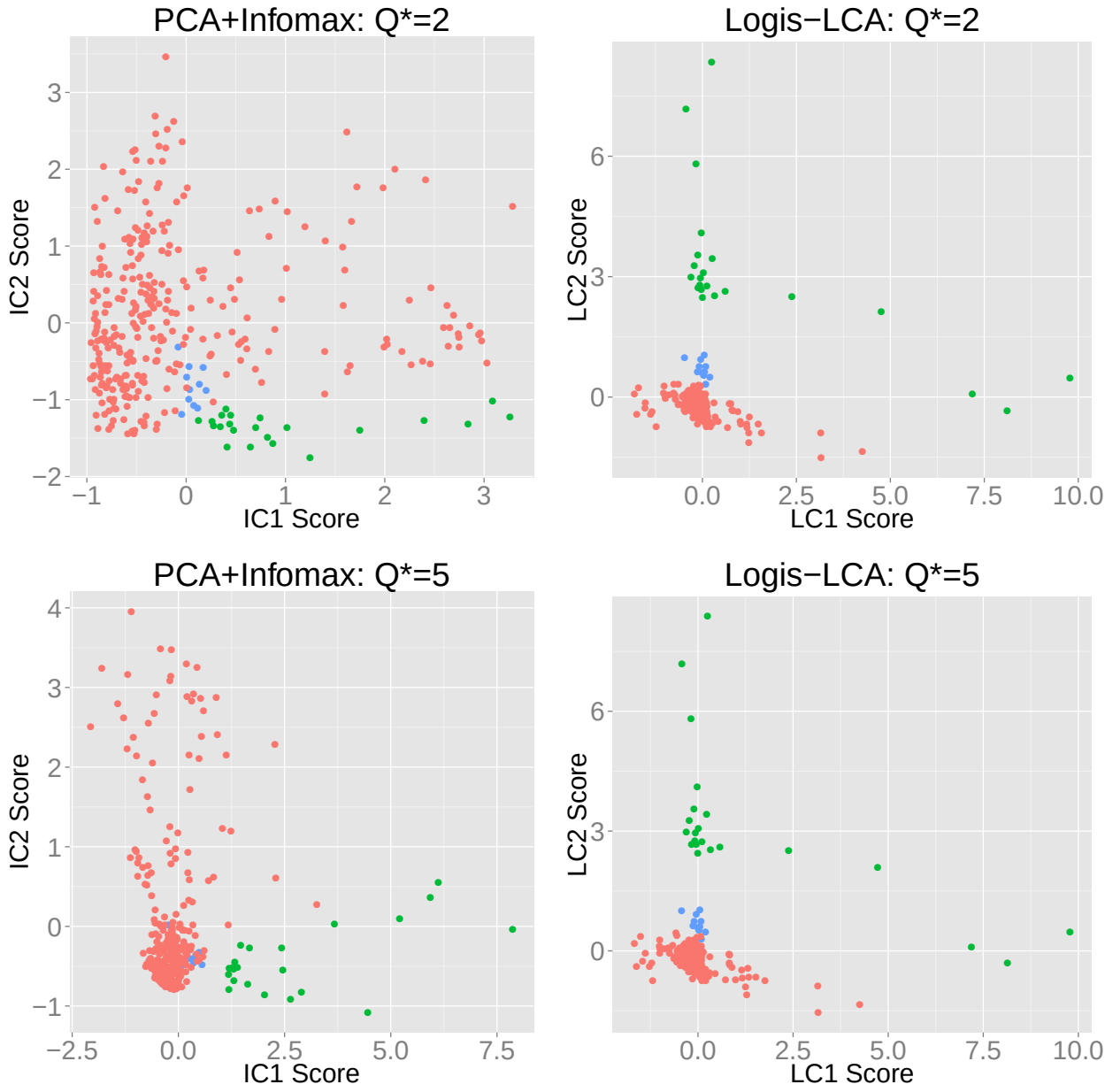
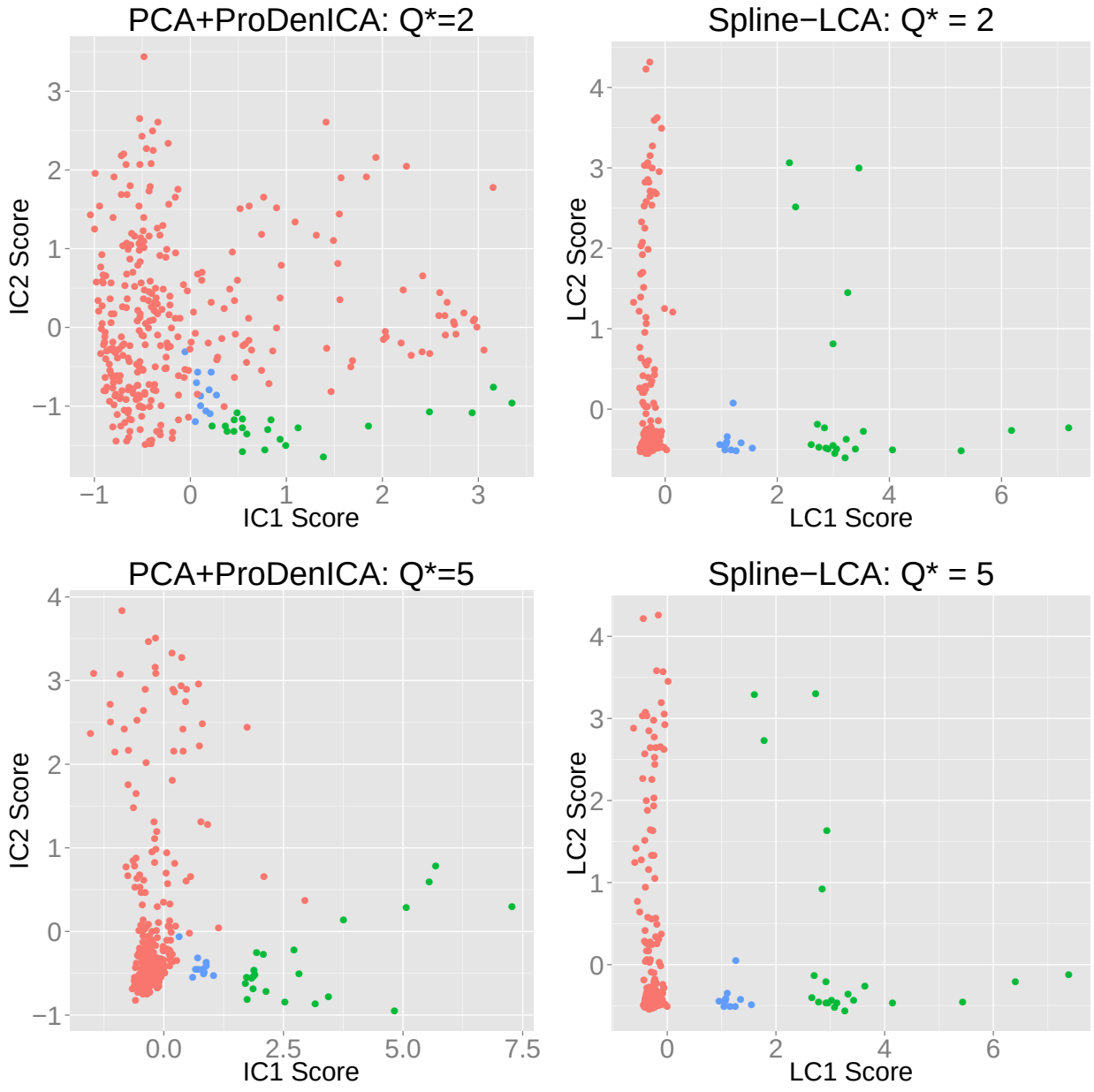


Figure S.7: Components in the leaf data from PCA-ProDenICA and Spline-LCA when two components were estimated and when five components were estimated (when five components were estimated, the two components with the highest marginal likelihood are plotted). The green dots correspond to *Podocarpus sp.*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

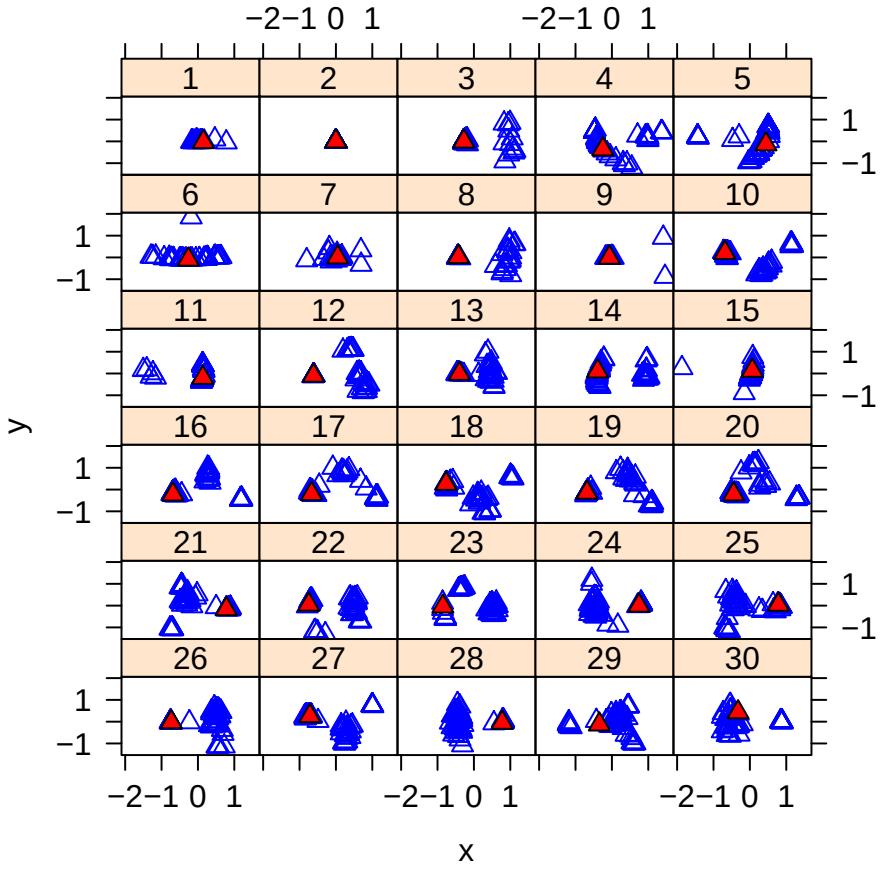
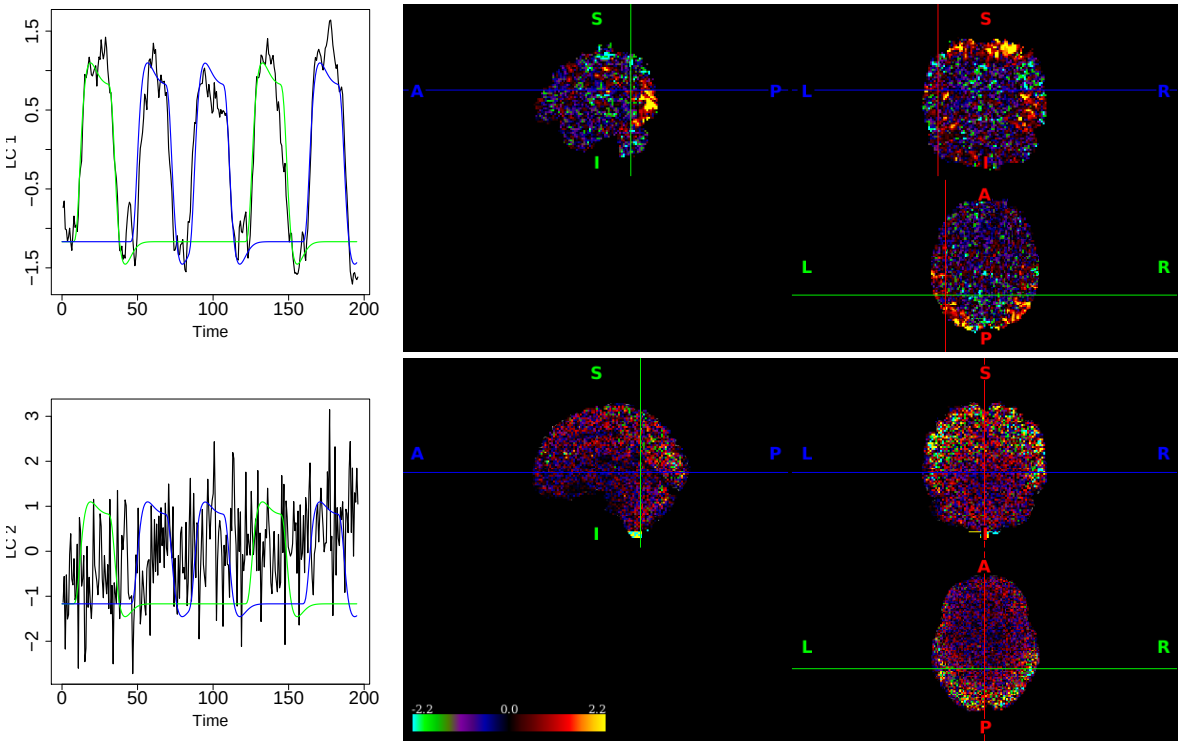


Figure S.8: Multidimensional scaling of $||\widehat{\mathbf{S}}_j^{(k)} - \widehat{\mathbf{S}}_j^{(\ell)}||_F$ for components $j = 1, \ldots, 30$ and initializations $k \neq \ell \in \{1, \ldots, 56\}$. The coordinates corresponding to the initialization with the highest likelihood are depicted by solid 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 $\widehat{\mathbf{M}}_{\mathbf{S}}$) and the mentalizing and random tasks. The mentalizing and random task covariates were generated by convolving each task's onsets and durations with the canonical HRF in SPM8 (Ashburner et al., 2004). In all subjects, the first component, i.e., the one with the highest likelihood, was highly correlated with the mentalizing and random tasks (e.g., Figure 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 cross hairs 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).

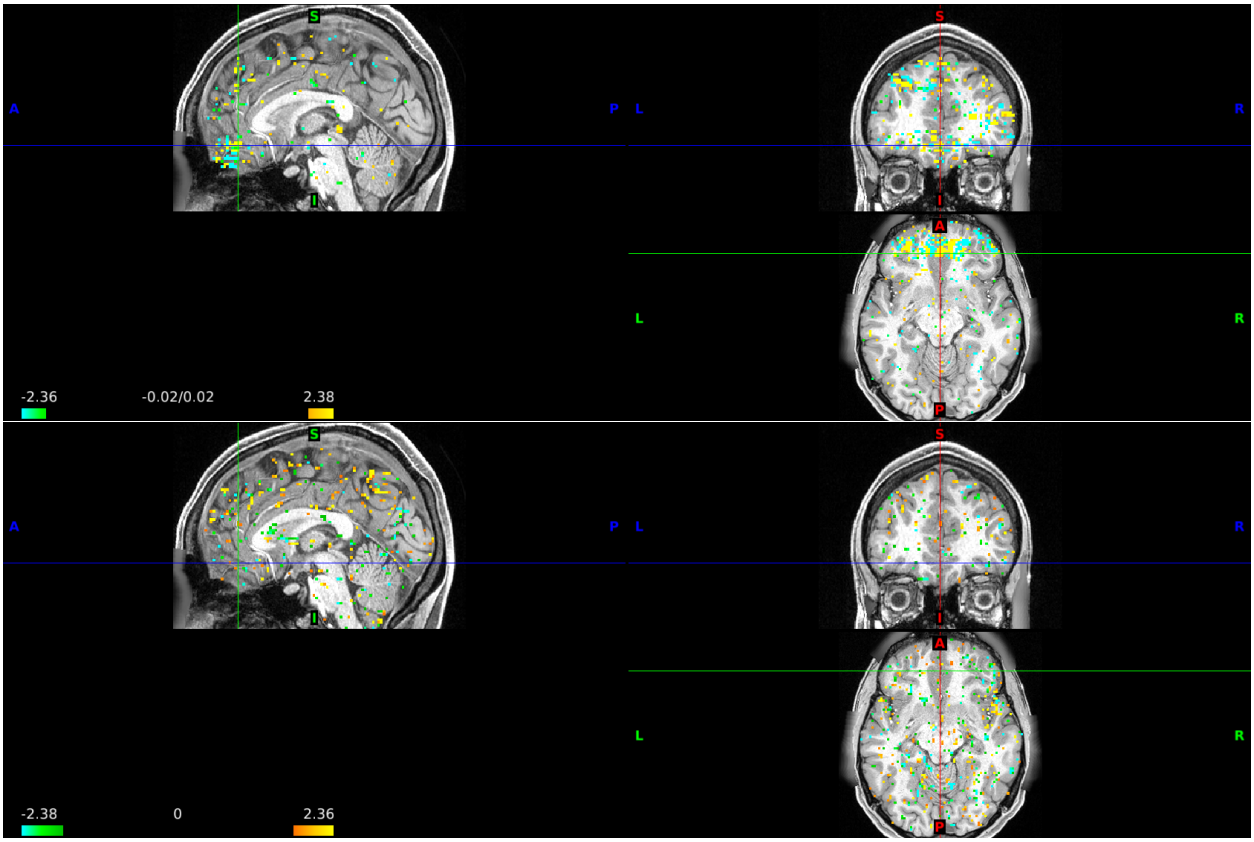


Voxels were highly activated in the brainstem and the component's time course was correlated with three of the motion parameters from the rigid-body alignment ($r = 0.32$, 0.32, and 0.42 for the x-transformation, x-rotation, and z-rotation parameters, respectively). This may be related to a gradual relaxation 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.