Nonlinear ICA of Temporally Dependent Stationary Sources

Paper by Aapo Hyvärinen and Hiroshi Morioka

20th International Conference on Artificial Intelligence and Statistics - 2017

Presented by

Behrad Moniri

Digital Signal Processing Lab

Sharif University of Technology

Table of Contents

Model Definition

- Mixing Model:

$$\boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{s}(t))$$

f is invertible and *sufficiently* smooth but we do not constrain it in any particular way.

Definition 1: Uniform Dependence

- A two-dimensional random vector (x, y) is called uniformly dependent if the cross-derivative of its log-pdf exists, is continuous, and does not vanish anywhere:

$$q_{x,y}(x,y) := \frac{\partial^2 \log p_{x,y}(x,y)}{\partial x \partial y} \neq 0 \; for \; all \; (x,y).$$

It is stronger than dependence:

Dependence in general only implies that q is **non-zero in some set of non-zero measure**, while we assume here that it is non-zero **everywhere**.

Definition 1: Uniform Dependence

A stationary stochastic process $s(t)$ is called (second order) uniformly dependent if the distribution of $(s(t), s(t\ -\ 1))$ is uniformly dependent.

Definition 2: Quasi-Newton Random Variables

- A two-dimensional random vector (x, y) is called quasi-Gaussian if $q_{x,y}$ exists, is continuous, and it can be factorized as:

$$q_{x,y}(x, y) = c\,\alpha(x)\alpha(y)$$

$$\log p(x, y) = \beta_1(x) + \beta_2(y) + c\bar{\alpha}(x)\bar{\alpha}(y)$$

- A stationary stochastic process $s(t)$ is called (second-order) quasi-Gaussian if the distribution of $(s(t), s(t-1))$ is quasi-Gaussian.

Lemma 1 *If a stochastic process $s(t)$ is quasi-Gaussian, then its instantaneous nonlinear transformation $\tilde{s}(t) = g(s(t))$ is also quasi-Gaussian for any invertible bijective mapping $g : \mathbb{R} \to \mathbb{R}$.*

Proof: For $(\tilde{x}, \tilde{y}) = (g(x), g(y))$, we have

$$\log p(\tilde{x}, \tilde{y}) = \beta_1(g^{-1}(\tilde{x})) + \log|(g^{-1})'(\tilde{x})| + \beta_2(g^{-1}(\tilde{y})) + \log|(g^{-1})'(\tilde{y})| + c\bar{\alpha}(g^{-1}(\tilde{x}))\bar{\alpha}(g^{-1}(\tilde{y})) \quad (6)$$

which is of the same form as Eq. (5), when we regroup the terms and redefine the nonlinearities.

Marginal of Quasi-Gaussian RVs

- An important point is that such factorizability holds for distributions of the type

$$\log p(x, y) = \beta_1(x) + \beta_2(y) - \rho x y$$

- The dependency structure is similar to the Gaussian one, but the marginal distributions can be arbitrarily non-Gaussian.

Separability of Nonlinear Mixtures

- We propose a practical, intuitive learning algorithm for estimating the nonlinear ICA model based on logistic regression with suitably defined input data and labels.

- Although initially only heuristically motivated, we show that in fact the algorithm separates sources which are not quasi Gaussian

Learning Algorithm

- Collect data points in two subsequent time points to construct a sample of a new random vector y:

$$\mathbf{y}(t) = \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{x}(t-1) \end{pmatrix}$$

- For comparison, create a permuted data sample by randomly permuting (shuffling) the time indices:

$$\mathbf{y}^*(t) = \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{x}(t^*) \end{pmatrix}$$

Learning Algorithm

- We propose to learn to discriminate between the sample of $\boldsymbol{y}(t)$ and the sample of $\mathbf{y}^*(t)$. We use logistic regression with a regression function of the form:

$$r(\mathbf{y}) = \sum_{i=1}^{n} B_i(h_i(\mathbf{y}^1), h_i(\mathbf{y}^2))$$

$$r(\boldsymbol{y}) = \log\left(\frac{\Pr(\text{label} = 1 \mid \boldsymbol{Y} = \boldsymbol{y})}{\Pr(\text{label} = 0 \mid \boldsymbol{Y} = \boldsymbol{y})}\right) = \log\left(P_y(\boldsymbol{y})\right) - \log\left(P_{y^*}(\boldsymbol{y})\right)$$

Where $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ denote the first and second halves of the vector $\boldsymbol{y}$,

i.e. $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2})$

Learning Algorithm

- Here, the h_i are scalar-valued functions giving a representation of the data, possibly as hidden units in a neural network.
- $B_i : \mathbb{R}^2 \to \mathbb{R}$ are additional nonlinear functions to be learned.

- Here, the h_i are scalar-valued functions giving a representation of the data, possibly as hidden units in a neural network.
- $B_i : \mathbb{R}^2 \to \mathbb{R}$ are additional nonlinear functions to be learned.

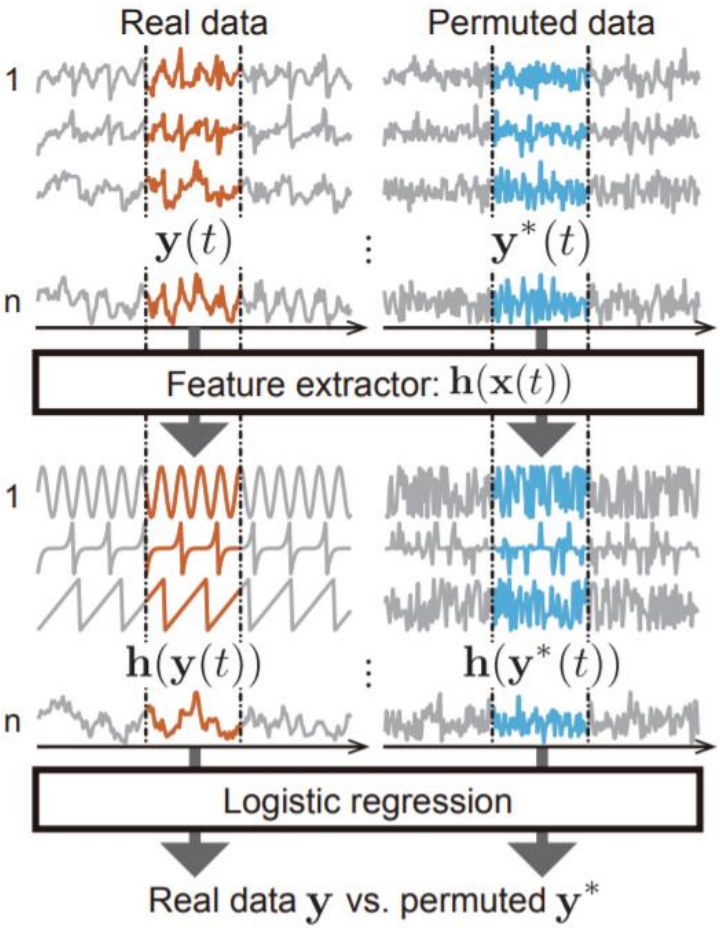


Table of Contents

Separability Theorem

Theorem 1 *Assume that*

1. *The sources $s_i(t), i = 1, \ldots, n$ are mutually independent, stationary ergodic stochastic processes.*

2. *The sources are uniformly dependent (Def. 1).*

3. *None of the sources is quasi-Gaussian (Def. 2).*

4. *We observe a nonlinear mixing $\mathbf{x}(t)$ according to Eq. (2), where the mixing nonlinearity $\mathbf{f}$ is bijective from $\mathbb{R}^n$ onto $\mathbb{R}^n$, twice differentiable, and its inverse is twice differentiable (i.e. $\mathbf{f}$ is a second-order diffeomorphism).*

5. *We learn a logistic regression to discriminate between $\mathbf{y}$ in Eq. (10) and $\mathbf{y}^*$ in Eq. (11) with the regression function in Eq. (12), using function approximators for h_i and B_i both of which are able to approximate any nonlinearities (e.g. a neural network). The functions h_i and B_i have continuous second derivates.*

Then, the hidden representation $h_i(\mathbf{x}(t))$ will asymptotically (i.e. when the length of the observed stochastic process goes infinite) give the original sources $s_i(t)$, up to element-wise transformations, and in arbitrary order with respect to i.

Proof of Theorem 1

- We Denote the (true) inverse function of **f** which transforms **x** into **s,** by **g** $\boldsymbol{s}(t) = \boldsymbol{g}(\boldsymbol{x}(t))$

$$\log p(\mathbf{x}(t), \mathbf{x}(t-1)) = \sum_{i=1}^{n} \log p_i^{\tilde{s}}(g_i(\mathbf{x}(t)), g_i(\mathbf{x}(t-1))) + \log |\mathbf{J}\mathbf{g}(\mathbf{x}(t))| + \log |\mathbf{J}\mathbf{g}(\mathbf{x}(t-1))| \quad (22)$$

where $p_i^{\tilde{s}}$ is the pdf of $(s_i(t), s_i(t-1))$, and $\mathbf{J}\mathbf{g}$ denotes the Jacobian of $\mathbf{g}$;

On the other hand, according to well-known theory, when training logistic regression we will asymptotically have

$$r(\mathbf{y}) = \log p_{\mathbf{y}}(\mathbf{y}) - \log p_{\mathbf{y}^*}(\mathbf{y}) \tag{23}$$

This holds in our case in the limit of an infinitely long stochastic process due to the assumption of a stationary ergodic process (Assumption 1).

$$\log p(\mathbf{x}(t), \mathbf{x}(t-1)) = \sum_{i=1}^{n} \log p_i^{\tilde{s}}(g_i(\mathbf{x}(t)), g_i(\mathbf{x}(t-1))) + \log |\mathbf{Jg}(\mathbf{x}(t))| + \log |\mathbf{Jg}(\mathbf{x}(t-1))| \quad (22)$$

Now, based on (22), the probability in the real data class is of the form

$$\log p_{\mathbf{y}}(\mathbf{y}) = \sum_{i=1}^{n} Q_i(g_i(\mathbf{y}^1), g_i(\mathbf{y}^2)) + \log |\mathbf{Jg}(\mathbf{y}^1)| + \log |\mathbf{Jg}(\mathbf{y}^2)| \quad (24)$$

$$Q_i(a, b) = \log p_i^{\tilde{s}}(a, b)$$

$$\log p(\mathbf{x}(t), \mathbf{x}(t-1)) = \sum_{i=1}^{n} \log p_i^{\tilde{s}}(g_i(\mathbf{x}(t)), g_i(\mathbf{x}(t-1))) + \log |\mathbf{Jg}(\mathbf{x}(t))| + \log |\mathbf{Jg}(\mathbf{x}(t-1))| \quad (22)$$

where we denote $Q_i(a, b) = \log p_i^{\tilde{s}}(a, b)$, while in the permuted (time-shuffled) data class the time points are i.i.d., which means that the log-pdf is of the form

$$\log p_{\mathbf{y}^*}(\mathbf{y}) = \sum_{i=1}^{n} \bar{Q}_i(g_i(\mathbf{y}^1)) + \bar{Q}_i(g_i(\mathbf{y}^2)) + \log |\mathbf{Jg}(\mathbf{y}^1)| + \log |\mathbf{Jg}(\mathbf{y}^2)| \quad (25)$$

for some functions $\bar{Q}_i$ which are simply the marginal log-pdf's.

$$r(\mathbf{y}) = \log p_{\mathbf{y}}(\mathbf{y}) - \log p_{\mathbf{y}^*}(\mathbf{y}) \qquad (23)$$

$$\log p_{\mathbf{y}}(\mathbf{y}) = \sum_{i=1}^{n} Q_i(g_i(\mathbf{y}^1), g_i(\mathbf{y}^2)) + \log |\mathbf{Jg}(\mathbf{y}^1)| + \log |\mathbf{Jg}(\mathbf{y}^2)| \qquad (24)$$

$$\log p_{\mathbf{y}^*}(\mathbf{y}) = \sum_{i=1}^{n} \bar{Q}_i(g_i(\mathbf{y}^1)) + \bar{Q}_i(g_i(\mathbf{y}^2)) + \log |\mathbf{Jg}(\mathbf{y}^1)| + \log |\mathbf{Jg}(\mathbf{y}^2)| \qquad (25)$$

$$\sum_{i=1}^{n} B_i(h_i(\mathbf{y}^1), h_i(\mathbf{y}^2)) = \sum_{i=1}^{n} Q_i(g_i(\mathbf{y}^1), g_i(\mathbf{y}^2)) - \bar{Q}_i(g_i(\mathbf{y}^1)) - \bar{Q}_i(g_i(\mathbf{y}^2)) \qquad (26)$$

$$\sum_{i=1}^{n} B_i(h_i(\mathbf{y}^1), h_i(\mathbf{y}^2)) = \sum_{i=1}^{n} Q_i(g_i(\mathbf{y}^1), g_i(\mathbf{y}^2)) - \bar{Q}_i(g_i(\mathbf{y}^1)) - \bar{Q}_i(g_i(\mathbf{y}^2)) \quad (26)$$

We easily notice that one solution to this is given by $h_i(\mathbf{x}) = g_i(\mathbf{x})$, $B_i(x, y) = Q_i(x, y) - \bar{Q}_i(x) - \bar{Q}_i(y)$. In fact, due to the assumption of the universal approximation capability of B and h, such a solution can be reached by the learning process. Next we prove that this is the only solution, up to permutation of the h_i and element-wise transformations.

Make the change of variables

$$\mathbf{z}^1 = \mathbf{g}(\mathbf{y}^1), \quad \mathbf{z}^2 = \mathbf{g}(\mathbf{y}^2) \tag{27}$$

and denote the compound function

$$\mathbf{k} = \mathbf{h} \circ \mathbf{f} = \mathbf{h} \circ \mathbf{g}^{-1} \tag{28}$$

$$\sum_{i=1}^{n} B_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))$$

$$= \sum_{i=1}^{n} Q_i(z_i^1, z_i^2) - \bar{Q}_i(z_i^1) - \bar{Q}_i(z_i^2) \tag{29}$$

Take cross-derivatives of both sides of (29) with respect to z_j^1 and z_k^2. This gives

$$\sum_{i=1}^{n} \frac{\partial^2 B_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))}{\partial z_j^1 \partial z_k^2} = \sum_{i=1}^{n} \frac{\partial^2 Q_i(z_i^1, z_i^2)}{\partial z_j^1 \partial z_k^2}. \tag{30}$$

$$\sum_{i=1}^{n} \frac{\partial^2 B_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))}{\partial z_j^1 \partial z_k^2} = \sum_{i=1}^{n} \frac{\partial^2 Q_i(z_i^1, z_i^2)}{\partial z_j^1 \partial z_k^2}. \quad (30)$$

Denoting cross-derivatives as

$$b_i(a,b) := \frac{\partial^2 B_i(a,b)}{\partial a \partial b}, \quad q_i(a,b) := \frac{\partial^2 Q_i(a,b)}{\partial a \partial b} \quad (31)$$

this gives further

$$\sum_{i=1}^{n} b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2)) \frac{\partial k_i}{\partial z_j^1}(\mathbf{z}^1) \frac{\partial k_i}{\partial z_k^2}(\mathbf{z}^2)$$

$$= \sum_{i=1}^{n} q_i(z_i^1, z_i^2) \delta_{ij} \delta_{ik}$$

which must hold for all j, k.

$$\mathbf{Jk}(\mathbf{z}^1)^T \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))] \mathbf{Jk}(\mathbf{z}^2)$$

$$= \mathrm{diag}_i[q_i(z_i^1, z_i^2)] \quad (32)$$

$$\mathbf{Jk}(\mathbf{z}^1)^T \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))]\mathbf{Jk}(\mathbf{z}^2)$$
$$= \mathrm{diag}_i[q_i(z_i^1, z_i^2)] \quad (32)$$

Now, the q_i are non-zero for all $\mathbf{z}^1, \mathbf{z}^2$ by assumption of uniform dependence. Since the RHS of (32) is invertible at any point, also $\mathbf{Jk}$ must be invertible at any point. We can thus obtain

$$[\mathbf{Jk}(\mathbf{z}^1)^{-1}]^T \mathrm{diag}_i[q_i(z_i^1, z_i^2)]\mathbf{Jk}(\mathbf{z}^2)^{-1}$$
$$= \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))] \quad (33)$$

Lemma 2 *Assume the continuous functions* $q_i(a, b)$ *are non-zero everywhere, and not factorizable as in Eq. (4) in the definition of quasi-Gaussianity.*[5] *Assume* $\mathbf{M}$ *is any continuous matrix-valued function* $\mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$, *such that the matrix* $\mathbf{M}(\mathbf{u})$ *is non-singular for any* $\mathbf{u}$. *Assume we have*

$$\mathbf{M}(\mathbf{u}^1)^T \, diag_i[q_i(u_i^1, u_i^2)] \, \mathbf{M}(\mathbf{u}^2) = \mathbf{D}(\mathbf{u}^1, \mathbf{u}^2) \quad (34)$$

for any $\mathbf{u}^1, \mathbf{u}^2$ *in* $\mathbb{R}^n$, *and for some unknown matrix-valued function* $\mathbf{D}$ *which takes only diagonal values. Then, the function* $\mathbf{M}(\mathbf{u})$ *is such that every row and column has exactly one non-zero entry, and the locations and signs of the non-zero entries are the same for all* $\mathbf{u}$.

$$[\mathbf{Jk}(\mathbf{z}^1)^{-1}]^T \mathrm{diag}_i[q_i(z_i^1, z_i^2)] \mathbf{Jk}(\mathbf{z}^2)^{-1} = \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))] \quad (33)$$

- Apply the lemma with $\mathbf{M}(\mathbf{z}) = \mathbf{Jk}(\mathbf{z})^{-1}$

The assumptions of the Lemma are included in the assumptions of the Theorem and the non-singularity was proven in the previous slide.

Thus the Lemma shows that $\mathbf{Jk}(\mathbf{z})^{-1}$ must be a **rescaled permutation matrix** for all **z**, with the same locations of the non-zero elements; the same applies to $\mathbf{Jk}(\mathbf{z})$

$$\mathbf{k} = \mathbf{h} \circ \mathbf{f} = \mathbf{h} \circ \mathbf{g}^{-1}$$

Thus, **g** and **h** must be equal up to a permutation and element-wise functions.

The fact that the signs of the elements in **M** stay the same implies the transformations are strictly monotonic, which proves the Theorem. ■

Proof of Lemma 2

Lemma 2 *Assume the continuous functions* $q_i(a, b)$ *are non-zero everywhere, and not factorizable as in Eq. (4) in the definition of quasi-Gaussianity.*[5] *Assume* $\mathbf{M}$ *is any continuous matrix-valued function* $\mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$, *such that the matrix* $\mathbf{M}(\mathbf{u})$ *is non-singular for any* $\mathbf{u}$. *Assume we have*

$$\mathbf{M}(\mathbf{u}^1)^T \, diag_i[q_i(u_i^1, u_i^2)] \, \mathbf{M}(\mathbf{u}^2) = \mathbf{D}(\mathbf{u}^1, \mathbf{u}^2) \quad (34)$$

for any $\mathbf{u}^1, \mathbf{u}^2$ *in* $\mathbb{R}^n$, *and for some unknown matrix-valued function* $\mathbf{D}$ *which takes only diagonal values. Then, the function* $\mathbf{M}(\mathbf{u})$ *is such that every row and column has exactly one non-zero entry, and the locations and signs of the non-zero entries are the same for all* $\mathbf{u}$.

Proof of Lemma 2

$$\mathbf{M}(\mathbf{u}^1)^T \; diag_i[q_i(u_i^1, u_i^2)] \; \mathbf{M}(\mathbf{u}^2) = \mathbf{D}(\mathbf{u}^1, \mathbf{u}^2) \quad (34)$$

Consider (34) for two different points $\bar{\mathbf{u}}^1$ and $\bar{\mathbf{u}}^2$ in $\mathbb{R}^n$. Denote for simplicity

$$\mathbf{M}_p = \mathbf{M}(\bar{\mathbf{u}}^p), \qquad \mathbf{D}_{pq} = \mathrm{diag}_i[q_i(\bar{u}_i^p, \bar{u}_i^q)] \qquad (35)$$

$$p, q \in \{1, 2\}$$

$$\mathbf{M}(\mathbf{u}^1)^T \; diag_i[q_i(u_i^1, u_i^2)] \; \mathbf{M}(\mathbf{u}^2) = \mathbf{D}(\mathbf{u}^1, \mathbf{u}^2) \quad (34)$$

$$\mathbf{M}_1^T \mathbf{D}_{12} \mathbf{M}_2 = \mathbf{D} \quad (36)$$

$$\mathbf{M}_2^T \mathbf{D}_{22} \mathbf{M}_2 = \mathbf{D}' \quad (37)$$

$$\mathbf{M}_1^T \mathbf{D}_{11} \mathbf{M}_1 = \mathbf{D}'' \quad (38)$$

for some diagonal matrices $\mathbf{D}, \mathbf{D}', \mathbf{D}''$.

By the assumption that q_i is non-zero, $\mathbf{D}_{12}$ is invertible, which also implies $\mathbf{D}$ is invertible.

$$\mathbf{M}_2 = \mathbf{D}_{12}^{-1} \mathbf{M}_1^{-T} \mathbf{D}$$

and plugging this into the second equation (37) we have

$$\mathbf{M}_1^{-1} \mathbf{D}_{22} \mathbf{D}_{12}^{-2} \mathbf{M}_1^{-T} = \mathbf{D}^{-1} \mathbf{D}' \mathbf{D}^{-1} \quad (40)$$

$$\mathbf{M}_1^{-1}[\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}]\mathbf{M}_1 = \mathbf{D}'''$$

- The rest of the proof of this lemma is based on the uniqueness of the eigenvalue decomposition which requires that the eigenvalues are distinct.

So, next we show that the assumption of non-factorizability of q_i implies that for any given $\bar{\mathbf{u}}^1$ we can find a $\bar{\mathbf{u}}^2$ such that the diagonal entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ are distinct. The diagonal entries are given by the function ψ defined as

$$\psi(\bar{u}_i^1, \bar{u}_i^2) = \frac{q_i(\bar{u}_i^1, \bar{u}_i^1)q_i(\bar{u}_i^2, \bar{u}_i^2)}{q_i^2(\bar{u}_i^1, \bar{u}_i^2)}. \qquad (42)$$

Claim: This function is not constant with respect to any of its arguments

Thus, it is possible to choose $\bar{\mathbf{u}}^2$ (corresponding to n choices of b for given n values of a) so that the diagonal entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ are all distinct, for any given $\bar{\mathbf{u}}^1$.

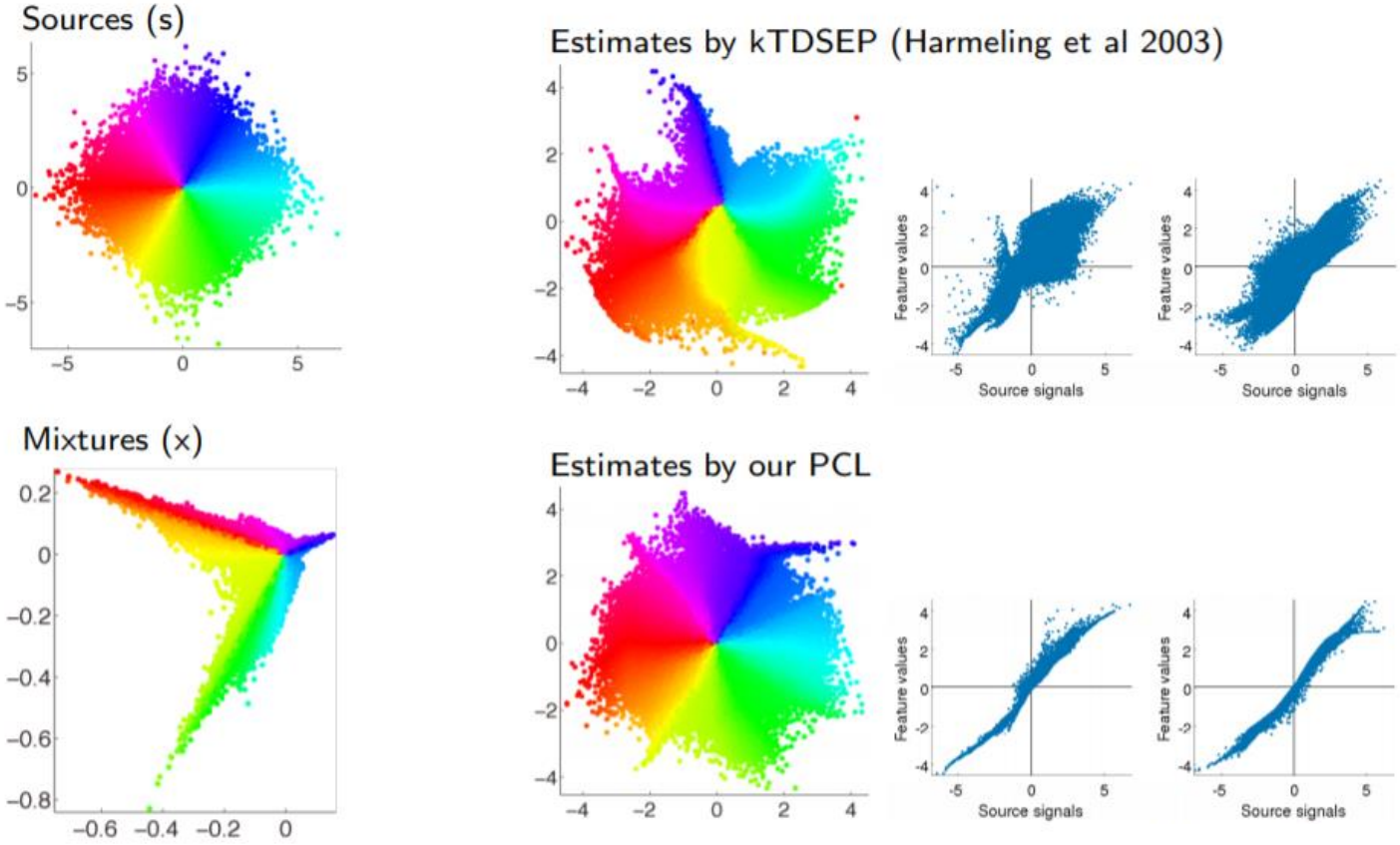
$$\mathbf{M}_p = \mathbf{M}(\bar{\mathbf{u}}^p) \qquad \mathbf{M}_1^{-1}[\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}]\mathbf{M}_1 = \mathbf{D}'''$$

The eigenvectors on both sides must be equal, and thus, $\mathbf{M}(\bar{\mathbf{u}}^1)$ must be equal to a permutation matrix, up to multiplication of each row by a scalar which depends on $\bar{\mathbf{u}}^1$.

Since $\bar{\mathbf{u}}^1$ could be freely chosen, $\mathbf{M}(\mathbf{u})$ is equal to such a rescaled permutation matrix everywhere. By continuity the non-zero entries in $\mathbf{M}(\mathbf{u})$ must be in the same locations everywhere; if they switched locations, $\mathbf{M}(\mathbf{u})$ would have to be singular at one point at least, which is excluded by assumption. With the same logic, we see the signs of the entries cannot change. Thus the Lemma is proven. ■

Table of Contents

- AR Model with Laplacian innovations, $n = 2$
  $\log p(s(t)|s(t-1)) = -|s(t) - \rho s(t-1)|$
- Nonlinearity is MLP. Mixing: leaky ReLU's; Demixing: maxout



- AR Model with Laplacian innovations, $n = 20$