Nonlinear ICA of Temporally Dependent Stationary Sources

Aapo Hyvärinen[1,2]
[1] Gatsby Computational Neuroscience Unit
University College London, UK

Hiroshi Morioka[2]
[2] Department of Computer Science and HIIT
University of Helsinki, Finland

Abstract

We develop a nonlinear generalization of independent component analysis (ICA) or blind source separation, based on temporal dependencies (e.g. autocorrelations). We introduce a nonlinear generative model where the independent sources are assumed to be temporally dependent, non-Gaussian, and stationary, and we observe arbitrarily nonlinear mixtures of them. We develop a method for estimating the model (i.e. separating the sources) based on logistic regression in a neural network which learns to discriminate between a short temporal window of the data vs. a temporal window of temporally permuted data. We prove that the method estimates the sources for general smooth mixing nonlinearities, assuming the sources have sufficiently strong temporal dependencies, and these dependencies are in a certain way different from dependencies found in Gaussian processes. For Gaussian (and similar) sources, the method estimates the nonlinear part of the mixing. We thus provide the first rigorous and general proof of identifiability of nonlinear ICA for temporally dependent stationary sources, together with a practical method for its estimation.

1 INTRODUCTION

Nonlinear independent component analysis (ICA) is one of the biggest unsolved problems in unsupervised learning. The basic idea is to generalize the highly successful linear ICA framework to arbitrary, but usually smooth and invertible, nonlinear mixing functions. Thus, the observed data is assumed to be a nonlinear invertible transformation ("mixing") of statistically independent latent quantities, and the goal is to find the mixing function, or its inverse, solely based on the assumption of the statistical independence of the latent quantities ("independent components", or "sources"). In other words, we want to separate the original sources from the mixed data. Importantly, no prior knowledge on the mixing function should be necessary for the learning.

Nonlinear ICA offers a rigorous framework for unsupervised deep learning, if the nonlinear demixing function (i.e. the inverse of the nonlinear mixing) is modelled by a deep neural network. However, the demixing function could also be modelled by any of the other well-known methods for general nonlinear function approximation, such as kernel methods or Gaussian processes.

The fundamental problem here is that in the basic case the problem is ill-posed: If the latent quantities are random variables, with no temporal structure (i.e. independently and identically distributed, i.i.d., over the set of observations), the original independent components cannot be inferred (Hyvärinen and Pajunen, 1999). In fact, there is an infinite number of possible nonlinear decompositions of a random vector into independent components, and those decompositions are not similar to each other in any trivial way. Assuming the mixing function to be smooth may help (Zhang and Chan, 2008), and in fact estimation of such nonlinear ICA has been attempted by a number of authors (Deco and Brauer, 1995; Tan et al., 2001; Almeida, 2003; Dinh et al., 2015), but it is not clear to what extent such methods are able to separate the sources.

A promising approach to nonlinear ICA is to use the temporal structure of the independent components — this is called by some authors nonlinear blind source separation to emphasize the difference from the temporally i.i.d. case. Using the (linear) autocorrelations of stationary sources enables separation of the sources in the linear mixing case (Tong et al., 1991; Belouchrani et al., 1997). A major advance in the field was to

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

show how this framework can be extended to the nonlinear case (Harmeling et al., 2003). Related proposals have also been made under the heading "slow feature analysis" (Wiskott and Sejnowski, 2002), originating in Földiák (1991), which was extended and thoroughly analysed by Sprekeler et al. (2014). Recent deep learning research (Mobahi et al., 2009; Springenberg and Riedmiller, 2012; Goroshin et al., 2015) as well as blind source separation research (Valpola and Karhunen, 2002; Hosseini and Deville, 2014) use similar ideas.

However, there are two fundamental problems with most research in this direction. First, the objective functions tend to be heuristic, and there is hardly any theoretical justification or proof that they will actually separate the sources. Even the general identifiability of the mixing models considered is not clear: There is no proof in the literature that the sources can be separated under the conditions considered — but see (Sprekeler et al., 2014) for some results in these directions, which are discussed in detail below.

The second problem is that even in the linear case, methods by, or based on Tong et al. (1991) and Belouchrani et al. (1997) can only separate sources which have distinct autocorrelation spectra: If the methods are applied on sources which have identical autocorrelations, they will fail. However, having sources with identical autocorrelations is a very realistic scenario. For example, linear features in video data that only differ by location and/or orientation are likely to have identical autocorrelations; likewise, two electroencephalography (EEG) sources with alpha oscillations may have practically identical autocorrelations. It is clear that the identifiability conditions in the nonlinear case cannot be less strict than in the linear case, and thus sources with identical autocorrelations cannot be separated.

Very recently, a rigorous nonlinear ICA theory was proposed for a different kind of temporal structure, consisting of the nonstationarity of variances or other parameters in an exponential family (Hyvärinen and Morioka, 2017). However, in this paper, we consider stationary sources, which is the "default" class in time series analysis, and widely encountered in real data.

Here, we propose a rigorous framework for nonlinear separation of independent, stationary sources based on temporal dependencies, as well as a practical algorithm. We formulate a nonlinear mixing model with explicit conditions on what kind of temporal dependencies are required in the independent source signals — importantly, without any conditions on the mixing function except for smoothness and invertibility. Essentially, we require that the sources have sufficiently strong temporal dependencies, and in particular these dependencies are different from the dependencies exhibited in Gaussian processes, in a precise sense related to the cross-derivatives of the joint log-pdf in a short time window.

We further propose an algorithm which can be motivated by a simple and intuitive heuristic learning principle: We learn to discriminate between the actual observed data and a corrupted version where the time structure is destroyed by permuting (shuffling) the time points. The discrimination is performed by logistic regression with a multi-layer perceptron. Surprisingly, we show that such learning finds the original sources in a hidden layer, up to trivial or otherwise simple indeterminacies. This also constitutes a constructive identifiability proof of our mixing model.

2 MODEL DEFINITION

In this section, we give a rigorous definition of our generative model, including some illustrative examples of sources.

2.1 General Nonlinear Mixing Model

We assume the n observed signals (i.e. time series or stochastic processes) $x_1(t), \ldots, x_n(t)$ are generated as a nonlinear transformation $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ of n latent signals $s_1(t), \ldots, s_n(t)$:

$$[x_1(t), \ldots, x_n(t)] = \mathbf{f}([s_1(t), \ldots, s_n(t)]) \tag{1}$$

Denoting by $\mathbf{x}(t)$ the vector $[x_1(t), \ldots, x_n(t)]$, and likewise for $\mathbf{s}$, this can be expressed simply as

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)). \tag{2}$$

We assume the function $\mathbf{f}$ is invertible (bijective) and sufficiently smooth but we do not constrain it in any particular way.

2.2 Source Model with Non-Gaussian Temporal Dependencies

In line with mainstream ICA theory, we assume here that the s_i are *mutually independent* stochastic processes (over different i). We further assume the sources are *stationary*, in contrast to time-contrastive learning (Hyvärinen and Morioka, 2017).

Importantly, we assume that the $s_i(t)$ are *temporally dependent*, for example autocorrelated (Tong et al., 1991; Molgedey and Schuster, 1994; Belouchrani et al., 1997; Harmeling et al., 2003; Sprekeler et al., 2014). Next, we formalize rigorously what kind of temporal dependencies are required, defining a class of stochastic processes for which our theory holds.

Definition 1 *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:*

$$q_{x,y}(x, y) := \frac{\partial^2 \log p_{x,y}(x, y)}{\partial x \partial y} \neq 0 \text{ for all } (x, y). \quad (3)$$

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

This definition is stronger than simply assuming that $s(t)$ and $s(t-1)$ are not independent, since 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. Below, we shall always assume the sources s_i are uniformly dependent, which sets our framework apart from ordinary ICA in which the sources are sampled i.i.d.

Conventionally, the analysis of source separation is divided to the Gaussian and the non-Gaussian case. Here, however, the analysis is naturally divided to a class of distributions whose dependencies are similar enough to Gaussian, and the rest. We define:

Definition 2 *A two-dimensional random vector (x, y) is called **quasi-Gaussian** if $q_{x,y}$ in Eq. (3) exists, is continuous, and it can be factorized as*

$$q_{x,y}(x, y) = c\,\alpha(x)\alpha(y) \quad (4)$$

for some real (possibly zero or negative) constant[1] c, and some real-valued function α. A stationary stochastic process $s(t)$ is called (second-order) quasi-Gaussian if the distribution of $(s(t), s(t-1))$ is quasi-Gaussian.

Quasi-Gaussianity is a very interesting further restriction on the temporal dependencies. Below, the strongest results on separability will be obtained for sources which are not quasi-Gaussian; the quasi-Gaussian case has to be considered separately.

Factorizability according to Eq. (4) is equivalent to the joint log-pdf being of the form

$$\log p(x, y) = \beta_1(x) + \beta_2(y) + c\bar{\alpha}(x)\bar{\alpha}(y) \quad (5)$$

where $\bar{\alpha}$ is the integral function of α, and the β_i are some smooth functions. If (x, y) is jointly Gaussian, the log-pdf does have such a form, with $\bar{\alpha}(x)$ being linear and $\beta_i(x)$ quadratic. Furthermore, then $(g(x), g(y))$ has such a factorization for any invertible function g; in general, we have the following result:

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.

However, there are random vectors (and processes) which are quasi-Gaussian but which cannot be obtained as point-wise transformations of Gaussian random vectors. For example, the above logic assumes g is invertible, which restricts the set of nonlinearities considered. An even more important point is that such factorizability holds for distributions of the type

$$\log p(x, y) = \beta_1(x) + \beta_2(y) - \rho xy \quad (7)$$

for *any* non-quadratic functions β_1, β_2, and a constant ρ. For such distributions, loosely speaking, the dependency structure is similar to the Gaussian one, but the marginal distributions can be arbitrarily non-Gaussian.

In fact, it is important to note that assuming non-quasi-Gaussianity only constrains the (temporal) dependencies, while the marginal distribution of s_i (over time) is not restricted in any way; it could be Gaussian. However, taken as a stochastic process, $s_i(t)$ must be non-Gaussian due to "non-Gaussian dependencies".[2]

Examples of Non-Gaussian Sources Next we consider some fundamental models of non-Gaussian processes and their relation to the definitions above. A classic example of a non-Gaussian process is given by a linear autoregressive (AR) model with non-Gaussian innovations:

$$\log p(s(t)|s(t-1)) = G(s(t) - \rho s(t-1)) \quad (8)$$

for some non-quadratic function G corresponding to the log-pdf of innovations, and a regression coefficient $|\rho| < 1$. Another typical model would be a nonlinear AR model with Gaussian innovations:

$$\log p(s(t)|s(t-1)) = -\lambda[s(t) - r(s(t-1))]^2 + \text{const.} \quad (9)$$

with some nonlinear, strictly monotonic regression function r, and a positive precision parameter λ. We

[1] Note that there is some indeterminacy in Eq. (4), since c could be partly absorbed in the functions α. However, it cannot be completely removed from the definition, since a negative c cannot be absorbed into α, and the negative sign has to be taken care of. We shall thus assume $c = \pm 1$.

[2] For clarity, we recall the standard definition of a Gaussian stochastic process which says that the joint probability of any time window, such as $(s(t), s(t-1))$, must be jointly Gaussian, which is much stronger than mere marginal Gaussianity of $s(t)$.

could obviously have a nonlinear AR model with non-Gaussian innovations as well.

Importantly, the examples of non-Gaussian stochastic processes in Eqs (8) and (9) can be proven to be both uniformly dependent and non-quasi-Gaussian under reasonable assumptions. In the case of the non-Gaussian AR model, we assume $G'' < 0$, which is slightly stronger than concavity, and a non-zero value of ρ. For the nonlinear AR model, we assume a strictly monotonic regression function r, and $\lambda > 0$. (These conditions are sufficient but most likely not necessary.) The proofs are in Supplementary Material.

3 SEPARATING SOURCES BY LOGISTIC REGRESSION

In this section, we propose a practical, intuitive learning algorithm for estimating the nonlinear ICA model defined above, 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. For quasi-Gaussian sources, we show that it estimates the model up to a linear mixing.

3.1 Discriminating Real vs. Permuted Data

Collect data points in two subsequent time points to construct a sample of a new random vector $\mathbf{y}$:

$$\mathbf{y}(t) = \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{x}(t-1) \end{pmatrix} \tag{10}$$

which gives a "minimal description" of the temporal dependencies in the data. Here, t is used as the sample index for $\mathbf{y}(t)$. For comparison, create a *permuted* data sample by randomly permuting (shuffling) the time indices:

$$\mathbf{y}^*(t) = \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{x}(t^*) \end{pmatrix} \tag{11}$$

where t^* is a randomly selected time point. In other words, we create data with the same marginal distribution (on the level of the vectors $\mathbf{x}$ instead of single variables), but which does not reflect the temporal structure of the data at all.

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

$$r(\mathbf{y}) = \sum_{i=1}^{n} B_i(h_i(\mathbf{y}^1), h_i(\mathbf{y}^2)) \tag{12}$$

where $\mathbf{y}^1$ and $\mathbf{y}^2$ denote the first and second halves of the vector $\mathbf{y}$, i.e. $\mathbf{y} = (\mathbf{y}^1, \mathbf{y}^2)$. (That is, $\mathbf{y}^1$ corresponds to $\mathbf{x}(t)$ and $\mathbf{y}^2$ corresponds to either $\mathbf{x}(t-1)$ or $\mathbf{x}(t^*)$ depending on the data set.) Here, the h_i are scalar-valued functions giving a representation of the data, possibly as hidden units in a neural network. The $B_i : \mathbb{R}^2 \to \mathbb{R}$ are additional nonlinear functions to be learned.

Intuitively speaking, it is plausible that h_i somehow recover the temporal structure of the data since recovering such structure is necessary to discriminate real data from permuted data. In particular, since the most parsimonious description of the temporal structure can be found by separating the sources and then modelling the temporal structure of each source separately, it is plausible that the discrimination works best when the h_i separate the sources, and the B_i somehow approximate the distribution of $(s_i(t), s_i(t-1))$. We call the new learning method *"permutation-contrastive learning (PCL)"*.

3.2 Convergence Theory: Non-Quasi-Gaussian Case

While our new method, PCL, was motivated purely heuristically above, it turns out, perhaps surprisingly, that it allows separation of the sources. Rigorously, the correctness of PCL for non-quasi-Gaussian sources is formalized in the following Theorem (proven in Supplementary Material):

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.

This Theorem is the first to provide a clear identifiability condition, and an estimation method, for nonlinear ICA of temporally dependent stationary sources. It should be noted that the assumptions in Theorem 1 are very weak in the context of ICA literature: The nonlinearity is simply assumed to be smooth and invertible, while the obviously necessary assumptions of temporal dependencies and non-Gaussianity are slightly strengthened. The indeterminacy of ordering is ubiquituous in ICA, and the well-known indeterminacy of scaling and signs is naturally generalized to strictly monotonic transformations of the individual sources. Next we discuss the assumptions in detail:

Assumption 1 is standard in ICA, saying that the sources (independent components) are independent. Here, the assumption is simply extended to stochastic processes using standard theory, and the assumptions of stationarity and ergodicity are added by default.

Assumptions 2 and 3 were already discussed above: they contain and extend the assumption of non-Gaussianity (of the stochastic process), and require the temporal dependencies to be strong enough. Gaussian processes in the linear case can be separated only when they have different autocorrelation functions (Tong et al., 1991; Molgedey and Schuster, 1994; Belouchrani et al., 1997); see Section 3.3 for more on Gaussian sources. As a remedy to this situation, non-Gaussian dependencies in the linear case have been considered, e.g. in (Hyvärinen, 2005; Pham, 2002), but here we may be the first to consider them in the nonlinear case.

Assumption 4 is a rather weak assumption of the invertibility and smoothness of the mixing system. Assumption 5 says that we apply PCL on data which comes from the nonlinear ICA model. A smoothness constraint on the nonlinear function approximator is also added, and it is of course necessary to assume that the nonlinear function approximator has some kind of universal approximation capability in order to be able to invert any nonlinear mixing functions; the universal approximation capability of neural networks is well known (Hornik et al., 1989).

What is very interesting, and relevant for practical algorithms, is that Theorem 1 shows that it is enough to consider just one time lag. This is to be contrasted to kTDSEP (Harmeling et al., 2003) and xSFA (Sprekeler et al., 2014) which usually use many time lags in the hope of finding enough differences in the autocorrelations, as may be necessary in the Gaussian case. On the other hand, the theorem here requires independence of the sources, in contrast to kTDSEP and xSFA which use only decorrelation.

Finally, we point out important advantages of our approach with respect to maximum likelihood estimation. The likelihood contains the Jacobian of the transformation $\mathbf{f}$ or its inverse. This Jacobian, and especially its derivatives with respect to any parameters defining the demixing function, are extremely difficult to compute (Deco and Brauer, 1995; Dinh et al., 2015). Our algorithm removes the Jacobian by "contrasting" the likelihoods of two different models which have the same Jacobians, which then cancel out. This leads to a simple method which needs hardly any new algorithmic developments, since it is based on formulating the unsupervised learning problem in terms of logistic regression, in the spirit of noise-contrastive estimation (Gutmann and Hyvärinen, 2012), time-contrastive learning (Hyvärinen and Morioka, 2017), and generative adversarial nets (Goodfellow et al., 2014; Gutmann et al., 2014).

3.3 Convergence Theory: Quasi-Gaussian Case

Next we consider the special case of quasi-Gaussian sources, which includes the case of Gaussian sources. We simplify the situation by using functions B_i which are adapted to the quasi-Gaussian case, i.e. a functional form similar to the log-pdf of quasi-Gaussian sources in Eq. (5). We further simplify by using linear $\bar{\alpha}$ which seems to be sufficient in this special case. Regarding the convergence of PCL we have the following Theorem (proven in Supplementary Material):

Theorem 2 *Assume the same as in Theorem 1, but instead of assuming that none of the sources is quasi-Gaussian (Assumption 3), assume they are all quasi-Gaussian. Furthermore, assume B_i in the regression function has the following form:*

$$B_i(z_1, z_2) = \beta_1^i(z_1) + \beta_2^i(z_2) + a_i z_1 z_2 + \lambda_i \qquad (13)$$

for some functions β_1^i, β_2^i (again, universally approximated) and scalar parameters a_i, λ_i. Then, after learning PCL we asymptotically have

$$h_i(\mathbf{x}(t)) = \sum_{j=1}^{n} b_{ij} \bar{\alpha}_j(s_j(t)) + d_i \qquad (14)$$

for some invertible constant matrix $\mathbf{B}$ with entries b_{ij}, some constants d_i, and the integral functions $\bar{\alpha}_i(y) = \int \alpha_i(y)dy$ of the α_i in the definition of quasi-Gaussianity of each s_i.

This Theorem tells that PCL will solve the nonlinear part of the mixing, and only leave a linear mixing together with point-wise nonlinearities to be resolved. What is left is essentially a linear ICA model, since the nonlinearities $\bar{\alpha}$ work on the original sources and do not change their independence. Interestingly, the nonlinear transformations $\bar{\alpha}_i$ try to make the sources

Gaussian in the sense that if the sources are pointwise transformations of Gaussian sources as in Lemma 1, they will be transformed to Gaussian, since we will eventually have terms of the form $\bar{\alpha}(\bar{\alpha}^{-1}(\bar{x})) = \bar{x}$ on the RHS of (6).

If some well-known identifiability conditions for linear ICA are met, we can thus perform a further linear ICA on the $h_i(\mathbf{x}(t))$ and identify the nonlinear ICA model completely—but such conditions are rather complicated in the case of general temporal dependencies, so we do not treat them in detail here.

If the sources are Gaussian, the nonlinearities $\bar{\alpha}$ are in fact equal to linear transformations since the α_i are constant. If the autocorrelations are all distinct, Theorem 2 implies that even for Gaussian sources, after applying PCL, we can solve the linear part by linear methods such as SOBI or TDSEP (Belouchrani et al., 1997; Ziehe and Müller, 1998).

In previous literature, there has been no clear proof of whether separation of Gaussian sources is possible from a nonlinear mixture, even in the case of different autocorrelations. Encouraging simulations were presented in the seminal paper by Harmeling et al. (2003), but no identifiability result was provided. An important contribution to the mathematical analysis was made by Sprekeler et al. (2014), who found that some functions of single sources (or products of multiple sources) may be found by SFA. However the rigorous identifiability proof was essentially restricted to the most slowly changing source, and deflating it away (Delfosse and Loubaton, 1995) turned out to be difficult. Thus, it was not clear from that analysis if all the sources can be found and how, although simulations were again promising. Our present result on the Gaussian case extends their results by showing that, indeed, *all* the sources can be estimated by combining PCL and a linear source separation method, if the autocorrelations are distinct.

3.4 Case of Many Time Lags

While the theory above used a single time lag for simplicity, at least Theorem 1 can be extended to multiple, say m, time lags. We define

Definition 3 *An m-dimensional random vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is called **quasi-Gaussian** if for any indices j, k, the cross-derivatives of the log-pdf*

$$q_{j,k}(\mathbf{x}) := \frac{\partial^2 \log p_{\mathbf{x}}(\mathbf{x})}{\partial x_j \partial x_k} \tag{15}$$

exist and are continuous, and can be factorized as

$$q_{j,k}(x, y) = c\, \alpha^{jk}(x_j, \mathbf{x}_{-j-k}) \alpha^{jk}(x_k, \mathbf{x}_{-j-k}) \tag{16}$$

for some real (possibly zero or negative) constant c and some real-valued function α^{jk}, where $\mathbf{x}_{-j-k}$ is the vector $\mathbf{x}$ without the j-th and k-th entries. A stationary stochastic process $s(t)$ is called m-th-order quasi-Gaussian if the distribution of $(s(t), s(t-1), \ldots, s(t-m+1))$ is quasi-Gaussian.

Here, the condition on factorization is more involved than in the case of a single lag. Again, it includes any point-wise nonlinear transformations of jointly Gaussian variables and Gaussian processes, since if $\mathbf{z}$ is Gaussian and g is an invertible nonlinear transformation $\mathbb{R} \to \mathbb{R}$, for $\mathbf{x} = (g(z_1), \ldots, g(z_m))$ we have

$$q_{j,k}(\mathbf{x}) = -\rho_{j,k}(g^{-1})'(x_j)(g^{-1})'(x_k) \tag{17}$$

where $\rho_{j,k}$ is the j,k-th entry in the precision matrix. As another example, the distribution $\log p(\mathbf{x}) = \prod_{i=1}^{m} G(x_i)$ is factorizable for any smooth function G. However, it may be that in the multi-dimensional case the general form of the quasi-Gaussian distributions cannot be given in any simple form, unlike in Eq. (5).

The definition of uniform dependence is generalized in the same way. The estimation method can be extended by defining the two classes as

$$\mathbf{y}(t) = \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{x}(t-1) \\ \vdots \\ \mathbf{x}(t-m+1) \end{pmatrix}, \ \mathbf{y}^*(t) = \begin{pmatrix} \mathbf{x}(t) \\ \mathbf{x}(t^1) \\ \vdots \\ \mathbf{x}(t^{m-1}) \end{pmatrix} \tag{18}$$

where $t^i, i = 1, \ldots, m-1$ are randomly chosen time points (with uniform probability over $1, \ldots, T$). Theorem 1 can be generalized to show the convergence of the method using these generalized definitions, and an obvious generalization of Eq. (12); see Supplementary Material for details.

4 SIMULATIONS

Next we conduct simulations to verify and illustrate the theory above, as well as to compare its performance to other methods.

4.1 Simulation 1: Non-Gaussian AR Model

We first conducted a simulation where the sources in the nonlinear ICA model come from a non-Gaussian AR process, discussed in Section 2. Such sources are not quasi-Gaussian, so Theorem 1 should apply.

Methods First, temporally dependent source signals ($n = 20$) were randomly generated according to Eq. (8) by using $G(u) = -|u|$, and equal autoregressive

coefficients $\rho = 0.7$ for all components.[3] To generate the observed signals from the source signals, we used a multi-layer perceptron (called "mixing-MLP") as a nonlinear mixing function $\mathbf{f}(\mathbf{s})$, following the settings used by Hyvärinen and Morioka (2017); in particular we used leaky ReLU units to make the MLP invertible.

As a feature extractor to be trained by PCL, we adopted an MLP ("feature-MLP") which has the same number of hidden layers as the mixing-MLP, and thus has enough degrees of freedom to represent the inverse-network of the mixing-MLP. The settings for the feature-MLP were basically the same as those used by Hyvärinen and Morioka (2017), except for the activation function of the last hidden layer, which took here the form given in (12), in particular with $r(\mathbf{y})$ defined as the negative of

$$\sum_{i=1}^{n} \left|a_{i,1}h_i(\mathbf{y}^1) + a_{i,2}h_i(\mathbf{y}^2) + b_i\right| - \left(\bar{a}_i h_i(\mathbf{y}^1) + \bar{b}_i\right)^2 + c$$

where $\{a_{i,1}, a_{i,2}, b_i, \bar{a}_i, \bar{b}_i, c\}$ are parameters to be trained simultaneously with the feature-MLP. The squared term is a rough approximation of the marginal log-pdf of the AR process. (The marginal log-pdf of $h_i(\mathbf{y}^2)$ cancels out because the absolute value is a conditional pdf due to the AR model, i.e. $\log p(\mathbf{y}^1|\mathbf{y}^2)$.)

The MLP was trained by back-propagation with a momentum term. To avoid overfitting, we used ℓ_2 regularization for the parameters. The initial weights of each layer were randomly drawn from a uniform distribution for each layer, scaled as in Glorot and Bengio (2010). The performances were evaluated by averaging over 10 runs for each setting of the number of layers L and the number of data points T.

For comparison, we also applied a linear ICA method based on a temporal decorrelation source separation (TDSEP, Ziehe and Müller (1998)), which is equivalent to SOBI, and a kernel-based nonlinear ICA method (kTDSEP, Harmeling et al. (2003)) to the observed data. In TDSEP, we used 20 time-shifted covariance matrices. The results of kTDSEP were obtained using 20 time-shifted covariance matrices, a polynomial kernel of degree 7, and k-means clustering with a maximum of 10,000 points considered.

Results Figure 1 shows that after training the feature-MLP by PCL, the logistic regression could discriminate *real* data samples from *permuted* ones with high classification accuracies. This implies that the feature-MLP could learn to represent the temporal dependencies of the data at least to some degree. We can see that the larger the number of layers is (which means that the nonlinearity in the mixing-MLP is stronger), the more difficult it is to train the feature-MLP and the logistic regression. The figure also shows that the networks suffered from overfitting (deceptively high classification accuracy on training data) when the number of data points was not sufficient.

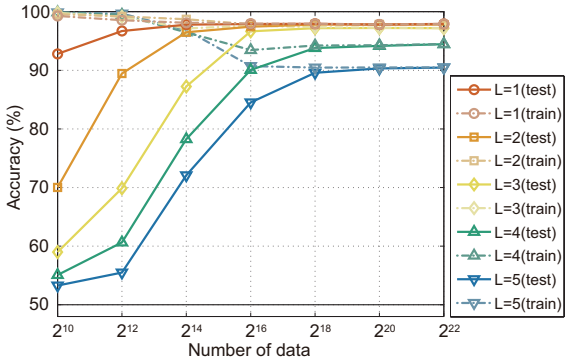

Figure 1: Simulation 1: Mean classification accuracies of the logistic regression trained with the feature-MLP in PCL, as a function of sample size (data length). Solid lines: test data; dash-dotted line: training data. The number of layers L was the same in generation and estimation. The chance level is 50%.

Figure 2 shows that PCL could reconstruct the source signals reasonably well even for the nonlinear mixture case ($L > 1$). Again, we can see that 1) a larger amount of data make it possible to achieve higher performance, and 2) more layers makes learning more difficult. TDSEP performed badly even for the linear-mixture case ($L = 1$). This is because we used the same autoregressive coefficient ρ for all components, so the eigenvalues of time-lagged covariances became equal for all components, which made it impossible for TDSEP to separate the source signals.

4.2 Simulation 2: Gaussian Sources

We further conducted a simulation with Gaussian sources, which is a special case of the quasi-Gaussian sources, treated in Theorem 2.

Methods We generated temporally dependent source signals ($n = 10$) from an Gaussian first-order autoregressive process. We selected the autoregressive coefficients ρ with uniform linear spacing between 0.4 to 0.8, to make sure they are distinct. The standard deviations of the generated signals were normalized to 1 for each component. The other settings are exactly the same as in the previous section, except that we

[3] This G slightly violates the condition of uniform dependence, so we can investigate the robustness of our theory at the same time. A typical smooth approximation of the Laplace density, such as $G(u) = -\sqrt{\epsilon + u^2}$, does give a uniformly dependent source.

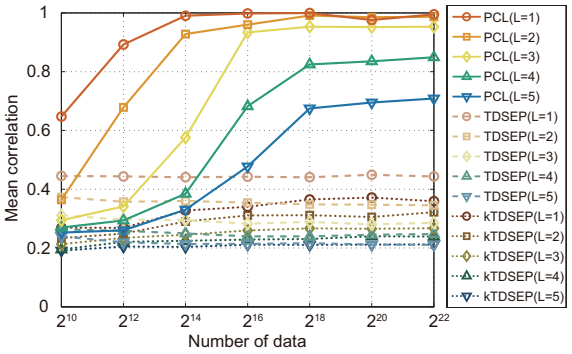

Figure 2: Simulation 1: Mean absolute correlation coefficients between source signals and their estimates given by $h_i(\mathbf{x}(t))$ learned by the proposed PCL method (solid lines), and for comparison, TDSEP (dashed line) and kTDSEP (dotted line), with different settings of the number of layers L and data points T.

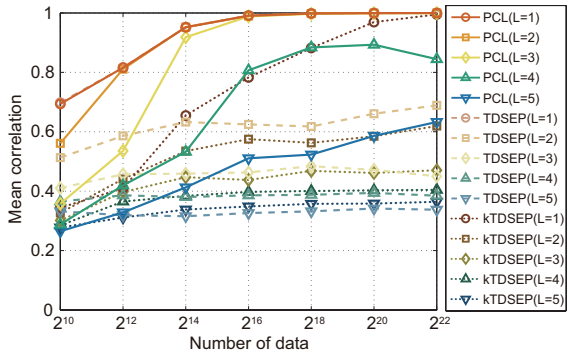

Figure 3: Simulation 2 (Gaussian sources): Mean absolute correlation coefficients between the source signals and their estimates, given by PCL and baseline methods. (See caption of Fig. 2.)

used the regression function of the form given in Eq (13), with all the nonlinearities set as $\beta(u) = -u^2$. Due to the indeterminacy shown in Theorem 2, we further applied (linear) TDSEP (with 20 time shifts) to the feature values $h_i(\mathbf{x}(t))$ obtained by PCL.

Results Figure 3 shows that, in the nonlinear case ($L > 1$), PCL could recover the source signals with much higher accuracies than the other methods. Unlike in Simulation 1, TDSEP (or SOBI) was able to reconstruct the source signals in the linear case ($L = 1$) because we selected all the ρ's of the generative model to be distinct, so the assumptions of TDSEP were fulfilled. In the linear case ($L = 1$), the performance of PCL was, in fact, exactly the same as that of TDSEP, because any "preliminary" processing by PCL did not change the result of the final TDSEP. In contrast to Simulation 1, PCL seems to have had particular problems with highly nonlinear mixing models ($L > 3$).

5 DISCUSSION

We considered the popular principle for unsupervised feature learning based on "temporal coherence", "temporal stability", or "slowness" of the features. We proposed the first rigorous treatment of the identifiability of such models in a nonlinear generative model setting. Lack of rigorous theory has been a major impediment for development of nonlinear unsupervised learning methods.

Our proof was constructive in the sense that we actually proposed a practical algorithm, PCL, for estimating the generative model, and showed its convergence (i.e. statistical consistency). Essentially, we introduced "uniform dependence" as a sufficient condition for identifiability and convergence. We further introduced the concept of quasi-Gaussianity and treated separately the cases of quasi-Gaussian and non-quasi-Gaussian sources. We showed convergence of our algorithm either to the final solution or a linear ICA model, respectively.

It should be noted that the conditions we proposed for identifiability were sufficient but not necessary. In particular, the condition of uniform dependence can presumably be relaxed to some extent. Furthermore, the case where some sources are quasi-Gaussian and others not, remains to be investigated.

The work most related to ours is the framework of time-contrastive learning (TCL) by Hyvärinen and Morioka (2017), which is based on non-stationarity. The two methods share the idea of using logistic regression, but they are used on datasets defined in very different ways. In fact, our contribution here is strictly complementary to TCL, since we assume stationary sources with autocorrelations, while TCL assumes non-stationarity and sources which are i.i.d. given the time point or segment. In the theory of linear ICA, it is widely accepted that autocorrelations and non-stationarity are two very different kinds of temporal structure: Together with non-Gaussianity of i.i.d. signals, these constitute what Cardoso (2001) called the "three easy routes to [linear] ICA".

For each practical application, it is thus worthwhile to consider which of the two models (TCL or PCL) gives a better match to the statistical properties of the data. This is a strictly empirical question: One can apply both methods and analyse the statistical properties of the obtained sources to see which assumptions made in the models are a better fit. An important question for future work is to combine the two principles in a single theory and a single algorithm.[4]

[4]This research was supported by JSPS KAKENHI 16J08502 (H.M.) and the Academy of Finland (A.H.,H.M.)

References

Almeida, L. B. (2003). MISEP—linear and nonlinear ICA based on mutual information. *J. of Machine Learning Research*, 4:1297–1318.

Belouchrani, A., Meraim, K. A., Cardoso, J.-F., and Moulines, E. (1997). A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444.

Cardoso, J.-F. (2001). The three easy routes to independent component analysis: contrasts and geometry. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California.

Deco, G. and Brauer, W. (1995). Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures. *Neural Networks*, 8:525–535.

Delfosse, N. and Loubaton, P. (1995). Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83.

Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: Non-linear independent components estimation. *arXiv:1410.8516 [cs.LG]*.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *AISTATS'10*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Goroshin, R., Bruna, J., Tompson, J., Eigen, D., and LeCun, Y. (2015). Unsupervised learning of spatiotemporally coherent metrics. In *IEEE Int. Conf. on Computer Vision*.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2014). Likelihood-free inference via classification. *arXiv:1407.4981 [stat.CO]*.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. of Machine Learning Research*, 13:307–361.

Harmeling, S., Ziehe, A., Kawanabe, M., and Müller, K.-R. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Hosseini, S. and Deville, Y. (2014). Blind separation of parametric nonlinear mixtures of possibly autocorrelated and non-stationary sources. *IEEE Transactions on Signal Processing*, 62(24):6521–6533.

Hyvärinen, A. (2005). A unifying model for blind separation of independent sources. *Signal Processing*, 85(7):1419–1427.

Hyvärinen, A. and Morioka, H. (2017). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NIPS2016)*.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Mobahi, H., Collobert, R., and Weston, J. (2009). Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 737–744.

Molgedey, L. and Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3636.

Pham, D.-T. (2002). Exploiting source non-stationary and coloration in blind source separation. In *Proc. Int. Conf. on Digital Signal Processing (DSP2002)*, pages 151 – 154.

Sprekeler, H., Zito, T., and Wiskott, L. (2014). An extension of slow feature analysis for nonlinear blind source separation. *J. of Machine Learning Research*, 15(1):921–947.

Springenberg, J. T. and Riedmiller, M. (2012). Learning temporal coherent features through life-time sparsity. In *Neural Information Processing*, pages 347–356. Springer.

Tan, Y., Wang, J., and Zurada, J. (2001). Nonlinear blind source separation using a radial basis function network. *IEEE Transactions on Neural Networks*, 12(1):124–134.

Tong, L., Liu, R.-W., Soon, V. C., and Huang, Y.-F. (1991). Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38:499–509.

Valpola, H. and Karhunen, J. (2002). An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.

Zhang, K. and Chan, L. (2008). Minimal nonlinear distortion principle for nonlinear independent component analysis. *J. of Machine Learning Research*, 9:2455–2487.

Ziehe, A. and Müller, K.-R. (1998). TDSEP—an efficient algorithm for blind separation using time structure. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675–680, Skövde, Sweden.

Supplementary Material for
Nonlinear ICA of Temporally Dependent Stationary Sources
by Aapo Hyvärinen and Hiroshi Morioka, AISTATS2017

Analysis of Processes in Eq. (8) and Eq. (9)

For the nonlinear AR model in Eq. (9), quasi-Gaussianity and uniform dependence are easy to see (under the assumptions given in the main text) since we have $\frac{\partial^2 \log p_{x,y}(x,y)}{\partial x \partial y} = 2\lambda r'(y)$. This implies uniform dependence by the strict monotonicity of r, and non-quasi-Gaussianity by its functional form.

For the non-Gaussian AR model in Eq. (8) we proceed as follows: First, we have

$$\frac{\partial^2 \log p_{x,y}(x,y)}{\partial x \partial y} = -\rho\, G''(x - \rho y). \tag{19}$$

This is always non-zero by the assumption on G'', and non-zero ρ, so uniform dependence holds. Assume a factorization as in (4) holds:

$$-G''(x - \rho y) = c\alpha(x)\alpha(y). \tag{20}$$

By assumption, $-G''$ is always positive, so we can take logarithms on both sides of (20), and again cross-derivatives. We necessarily have $\frac{\partial \log -G''(x-\rho y)}{\partial x \partial y} = 0$, since the RHS is separable. This can be evaluated as $(\log -G'')''(x - \rho y) = 0$ which implies $\log -G''(u) = du + b$ and

$$G''(u) = -\exp(du + b) \tag{21}$$

for some real parameters d, b. Now, if we have $d = 0$ and thus $G''(u)$ constant, we have a Gaussian process. On the other hand, if we have $d \neq 0$, we can plug this back in (20) and see that it cannot hold because the exponents for x and y would be different unless $\rho = -1$, which was excluded by assumption (as is conventional to ensure stability of the process). Thus, only a Gaussian linear AR process can be quasi-Gaussian under the given assumptions.

Proof of Theorem 1

Denote by $\mathbf{g}$ the (true) inverse function of $\mathbf{f}$ which transforms $\mathbf{x}$ into $\mathbf{s}$, i.e. $\mathbf{s}(t) = \mathbf{g}(\mathbf{x}(t))$. We can easily derive the log-pdf of an observed $(\mathbf{x}(t), \mathbf{x}(t-1))$ as

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

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 $\mathbf{g}$; its log-determinant appears twice because the transformation is done twice, separately for $\mathbf{x}(t)$ and $\mathbf{x}(t-1)$.

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

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

i.e. the regression function will asymptotically give the difference of the log-probabilities in the two classes. 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).

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

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

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

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

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

The equality in (23) means the regression function (12) is asymptotically equal to the difference of (24) and (25), i.e.

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

where we see that the Jacobian terms vanish because we "contrast" two data sets with the same Jacobian terms.

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

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

and denote the compound function

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

This is the compound transformation of the attempted demixing by $\mathbf{h}$ and the original mixing by $\mathbf{f}$. Such a compound function is of main interest in the theory of ICA, since it tells how well the original sources were separated. Our goal here is really to show that this function is a permutation with component-wise nonlinearities. So, we consider the transformed version of (26) given by

$$\sum_{i=1}^{n} B_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2)) = \sum_{i=1}^{n} Q_i(z_i^1, z_i^2) - \bar{Q}_i(z_i^1) - \bar{Q}_i(z_i^2) \quad (29)$$

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

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

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. We can collect these equations in a matrix form as

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

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

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

Next, we use the assumption of non-quasi-Gaussianity, in the form of the following Lemma (proven below):

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 $\mathbb{R}^n \to \mathbb{R}^{n \times n}$, 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 $\mathbf{u}$.

We apply this Lemma on Eq. (33) with $\mathbf{M}(\mathbf{z}) = \mathbf{Jk}(\mathbf{z})^{-1}$. The assumptions of the Lemma are included in the assumptions of the Theorem, except for the non-singularity of $\mathbf{M}$ which was just proven above, and the continuity of $\mathbf{M}$. If $\mathbf{Jk}(\mathbf{z})^{-1}$ were not continuous, the fact that the diagonal matrix on the LHS of (33) is continuous would imply that the diagonal matrix on the RHS is discontinuous, and this contradicts the assumptions on smoothness of $\mathbf{h}$, $\mathbf{g}$ and B_i.

Thus the Lemma shows that $\mathbf{Jk}(\mathbf{z})^{-1}$ must be a rescaled permutation matrix for all $\mathbf{z}$, with the same locations of the non-zero elements; the same applies to $\mathbf{Jk}(\mathbf{z})$. Thus, by (28), $\mathbf{g}$ and $\mathbf{h}$ must be equal up to a permutation and element-wise functions, plus a constant offset which can be absorbed in the element-wise functions. The fact that the signs of the elements in $\mathbf{M}$ stay the same implies the transformations are strictly monotonic, which proves the Theorem.

Proof of Lemma 2

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

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

with $p, q \in \{1, 2\}$. Evaluating (34) with all the possible combinations of setting $\mathbf{u}^1$ and $\mathbf{u}^2$ to $\bar{\mathbf{u}}^1$ and $\bar{\mathbf{u}}^2$, that is the four combinations $\mathbf{u}^1 := \bar{\mathbf{u}}^1, \mathbf{u}^2 := \bar{\mathbf{u}}^2$; $\mathbf{u}^1 := \bar{\mathbf{u}}^2, \mathbf{u}^2 := \bar{\mathbf{u}}^1$; $\mathbf{u}^1 := \bar{\mathbf{u}}^1, \mathbf{u}^2 := \bar{\mathbf{u}}^1$; and $\mathbf{u}^1 := \bar{\mathbf{u}}^2, \mathbf{u}^2 := \bar{\mathbf{u}}^2$, we have three different equations (the first one being obtained twice):

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

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

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

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

We will show that for any given $\bar{\mathbf{u}}^1$, it is always possible to find a $\bar{\mathbf{u}}^2$ such that the conditions (36–38) lead to an eigenvalue problem which has only a trivial solution consisting of a scaled permutation matrix.

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

[5] In this lemma, the q_i need not have anything to do with pdf's, so we do not directly use the assumption of quasi-Gaussianity, but the conditions on q are identical.

linear algebra, we can thus solve from the first equation (36)

$$\mathbf{M}_2 = \mathbf{D}_{12}^{-1}\mathbf{M}_1^{-T}\mathbf{D} \tag{39}$$

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

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

Next we multiply both sides of (38) by the respective sides of (40) from the left, and denoting $\mathbf{D}''' = \mathbf{D}^{-1}\mathbf{D}'\mathbf{D}^{-1}\mathbf{D}''$ we have

$$\mathbf{M}_1^{-1}[\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}]\mathbf{M}_1 = \mathbf{D}''' \tag{41}$$

Here, we see a kind of eigenvalue decomposition.

The rest of the proof of this lemma is based on the uniqueness of the eigenvalue decomposition, which requires that the eigenvalues are distinct (i.e. no two of them are equal). 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

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

For simplicity of notation, drop the index i and denote $a := \bar{u}_i^1, b = \bar{u}_i^2$. The diagonal entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ can be chosen distinct if ψ is not a function of a alone (which was fixed above since $\bar{\mathbf{u}}^1$ was fixed). Suppose ψ is a function of a alone: Then we would have

$$\frac{q(a,a)q(b,b)}{q^2(a,b)} = f(a) \tag{43}$$

for some function f. Since this holds for any b, we can set $b = a$, we see that f must be identically equal to one. So, we would have

$$q^2(a,b) = q(a,a)q(b,b) \tag{44}$$

or

$$q(a,b) = c\sqrt{|q(a,a)|}\sqrt{|q(b,b)|} \tag{45}$$

with the constant $c = \pm 1$. But a factorizable form in (45) with $\alpha(y) = \sqrt{|q(y,y)|}$ is exactly the same as in (4) in the definition of quasi-Gaussianity, or, equivalently, in the assumptions of the Lemma, and thus excluded by assumption.

Thus, we have proven by contradiction that ψ cannot be a function of a alone. The functions involved are continuous by assumption, so since ψ takes more than one value for any given a, it takes an infinity of values for any given a. 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 $\bar{\mathbf{u}}^1$.

Since the entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ can be assumed to be distinct, the eigenvectors of the (product) matrix on the LHS of (41) are equal to the columns of $\mathbf{M}_1^{-1}$, and uniquely defined up to a multiplication by a scalar constant which is always indetermined for eigenvectors. The diagonal entries on both sides are equal to the eigenvalues of the corresponding matrices, because eigenvalues are invariant to change of basis by $\mathbf{M}_1$, so we have $d_i''' = d_i^{11}d_i^{22}/(d_i^{12})^2$, up to permutation. On the other hand, the eigenvectors on the RHS of (41) are equal to the canonical basis vectors, and they are also uniquely defined (up to scalar multiplication) since the d_i''' are also distinct. 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.

Proof of Theorem 2

First, since we have a restricted form of regression function, we have to prove that it can actually converge to the optimal theoretical regression function in (23). This is true because the regression function in (13) can still approximate all quasi-Gaussian densities which have uniform dependence, after suitable transformation. Namely, uniform dependence together with quasi-Gaussianity implies that $\bar{\alpha}$ must be monotonic. Thus, by a pointwise transformation inverting such monotonic $\bar{\alpha}$, we can transform the data so that $\bar{\alpha}$ is linear, and the regression function in the Theorem can be learned to be optimal.

The proof of Theorem 1 is then valid all the way until (33), since we didn't use non-quasi-Gaussianity up to that point. We have from (33), (13), and the definition of quasi-Gaussianity

$$[\mathbf{Jk}(\mathbf{z}^1)^{-1}]^T\mathrm{diag}_i[\alpha_i(z_i^1)]\mathrm{diag}_i[c_i]\mathrm{diag}_i[\alpha_i(z_i^2)]\mathbf{Jk}(\mathbf{z}^2)^{-1} = \mathrm{diag}_i[a_i] \tag{46}$$

which must hold for any $\mathbf{z}^1, \mathbf{z}^2$. The matrices in this equation are invertible by the proof of Theorem 1. Now, define

$$\mathbf{V}(\mathbf{z}) = \mathrm{diag}_i[\alpha_i(z_i)]\mathbf{Jk}(\mathbf{z})^{-1} \tag{47}$$

so the condition above takes the form

$$\mathbf{V}(\mathbf{z}^1)^T\mathrm{diag}_i[c_i]\mathbf{V}(\mathbf{z}^2) = \mathrm{diag}_i[a_i]. \tag{48}$$

Setting $\mathbf{z}^2 = \mathbf{z}^1$, we can solve

$$\mathbf{V}(\mathbf{z}^1)^T = \mathrm{diag}_i[a_i]\mathbf{V}(\mathbf{z}^1)^{-1}\mathrm{diag}_i[1/c_i]. \quad (49)$$

Plugging this back into (48), we have

$$\mathrm{diag}_i[a_i]\mathbf{V}(\mathbf{z}^1)^{-1}\mathrm{diag}_i[1/c_i]\mathrm{diag}_i[c_i]\mathbf{V}(\mathbf{z}^2) = \mathrm{diag}_i[a_i] \quad (50)$$

which gives equivalently

$$\mathbf{V}(\mathbf{z}^1) = \mathbf{V}(\mathbf{z}^2). \quad (51)$$

That is, $\mathbf{V}(\mathbf{z})$ does not depend on $\mathbf{z}$. Denote its constant value by $\mathbf{V}$.

Solving for $\mathbf{Jk}(\mathbf{z})$ in (47) with such a constant $\mathbf{V}$, we have

$$\mathbf{Jk}(\mathbf{z}) = \mathbf{V}^{-1}\mathrm{diag}_i[\alpha_i(z_i)]. \quad (52)$$

Now, substitute, by (28), $\mathbf{J}(\mathbf{h}\circ\mathbf{f})(\mathbf{z})$ for the LHS, and change the dummy variable $\mathbf{z}$ to $\mathbf{s}$. Then we can integrate both sides to obtain

$$(\mathbf{h}\circ\mathbf{f})(\mathbf{s}) = \mathbf{h}(\mathbf{x}) = \mathbf{V}^{-1}\begin{pmatrix}\bar{\alpha}_1(s_1)\\ \bar{\alpha}_2(s_2)\\ \vdots\\ \bar{\alpha}_n(s_n)\end{pmatrix} + \mathbf{d} \quad (53)$$

for some integration constant vector $\mathbf{d}$. Thus we get the form given in the Theorem, with $\mathbf{B} = \mathbf{V}^{-1}$.

Theory and Proof for Multiple Time Lags

In the case of multiple lags, the assumptions in a theorem corresponding to Theorem 1 are apparently identical to those in Theorem 1, but we use the general definition of quasi-Gaussianity in Definition 3, and the general definition of uniform dependence, which is that the cross-derivative $q_{j,k}(\mathbf{x})$ is non-zero for any j, k and any $\mathbf{x}$. We further define the discrimination problem using (18) and use the obvious generalization of the regression function given by

$$r(\mathbf{y}) = \sum_{i=1}^{m} B_i(h_i(\mathbf{y}^1), h_i(\mathbf{y}^2), \ldots, h_i(\mathbf{y}^m)). \quad (54)$$

We can then use the proof of Theorem 1 with minimal changes. Non-quasi-Gaussianity implies that for some j, k, factorizability is impossible. Fix j, k to those values. Fix $\mathbf{y}^p$ for $p \neq j, k$ to any arbitrary values. The proof proceeds in the same way, largely ignoring any $\mathbf{y}^p$ with p not equal to j or k. In particular, the derivative in (30) is taken with respect to those j, k. Furthermore, (33) has the form

$$[\mathbf{Jk}(\mathbf{z}^j)^{-1}]^T\mathrm{diag}_i[q_i(z_i^1, \ldots, z_i^m))]\mathbf{Jk}(\mathbf{z}^k)^{-1} = \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), \ldots, k_i(\mathbf{z}^m))] \quad (55)$$

where both q_i and b_i are functions of $\mathbf{z}^j$ and $\mathbf{z}^k$ (or, equivalently, of $\mathbf{y}^j$ and $\mathbf{y}^k$) only, since all the other $\mathbf{z}^p$ (or $\mathbf{y}^p$) are fixed.

A version of Theorem 2 for multiple time lags is left as a question for future research.