Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning

Aapo Hyvärinen [1,2]

[1] The Gatsby Unit
UCL, UK

Hiroaki Sasaki [3,1]

[2] Dept. of CS and HIIT
Univ. Helsinki, Finland

[3]Div. of Info. Sci.
NAIST, Japan

Richard E. Turner [4]

[4] Univ. Cambridge &
Microsoft Research, UK

Abstract

Nonlinear ICA is a fundamental problem for unsupervised representation learning, emphasizing the capacity to recover the underlying latent variables generating the data (i.e., identifiability). Recently, the very first identifiability proofs for nonlinear ICA have been proposed, leveraging the temporal structure of the independent components. Here, we propose a general framework for nonlinear ICA, which, as a special case, can make use of temporal structure. It is based on augmenting the data by an auxiliary variable, such as the time index, the history of the time series, or any other available information. We propose to learn nonlinear ICA by discriminating between true augmented data, or data in which the auxiliary variable has been randomized. This enables the framework to be implemented algorithmically through logistic regression, possibly in a neural network. We provide a comprehensive proof of the identifiability of the model as well as the consistency of our estimation method. The approach not only provides a general theoretical framework combining and generalizing previously proposed nonlinear ICA models and algorithms, but also brings practical advantages.

1 INTRODUCTION

Nonlinear ICA is a fundamental problem in unsupervised learning which has attracted a considerable amount of attention recently. It promises a principled

approach to representation learning, for example using deep neural networks. Nonlinear ICA attempts to find nonlinear components, or features, in multidimensional data, so that they correspond to a well-defined generative model (Hyvärinen et al., 2001; Jutten et al., 2010). The essential difference to most methods for unsupervised representation learning is that the approach starts by defining a generative model in which the original latent variables can be recovered, i.e. the model is identifiable by design.

Denote an observed n-dimensional random vector by $\mathbf{x} = (x_1, \ldots, x_n)$. We assume it is generated using n independent latent variables called independent components, s_i. A straightforward definition of the nonlinear ICA problem is to assume that the observed data is an arbitrary (but smooth and invertible) transformation $\mathbf{f}$ of the latent variables $\mathbf{s} = (s_1, \ldots, s_n)$ as

$$\mathbf{x} = \mathbf{f}(\mathbf{s}) \tag{1}$$

The goal is then to recover the inverse function $\mathbf{f}^{-1}$ as well as the independent components s_i based on observations of $\mathbf{x}$ alone.

Research in nonlinear ICA has been hampered by the fact that such simple approaches to nonlinear ICA are not identifiable, in stark contrast to the linear ICA case. In particular, if the observed data $\mathbf{x}$ are obtained as i.i.d. samples, i.e. there is no temporal or similar structure in the data, the model is seriously unidentifiable (Hyvärinen and Pajunen, 1999), although attempts have been made to estimate it nevertheless, often by minimizing the mutual information of outputs of a neural network (Deco and Obradovic, 1995; Tan et al., 2001; Almeida, 2003; Brakel and Bengio, 2017; Hjelm and et al., 2018). This is a major problem since in fact most of the utility of linear ICA rests on the fact that the model is identifiable, or—in alternative terminology—the "sources can be separated". Proving the identifiability of linear ICA (Comon, 1994) was a great advance on the classical theory of factor analysis, where an orthogonal factor rotation could not be

identified.

Fortunately, a solution to non-identifiability in nonlinear ICA can be found by utilizing temporal structure in the data (Harmeling et al., 2003; Sprekeler et al., 2014; Hyvärinen and Morioka, 2017b,a). In recent work, various identifiability conditions have been proposed, assuming that the independent components are actually time series and have autocorrelations (Sprekeler et al., 2014), general non-Gaussian temporal dependencies (Hyvärinen and Morioka, 2017a), or non-stationarities (Hyvärinen and Morioka, 2017b). These generalize earlier identifiability conditions for linear ICA with temporal structure (Belouchrani et al., 1997; Pham and Cardoso, 2001).

Meanwhile, recent work in computer vision has successfully proposed "self-supervised" feature extraction methods from a purely heuristic perspective. The method by Misra et al. (2016) is quite similar to the nonlinear ICA by Hyvärinen and Morioka (2017a), while Oord et al. (2018) proposed a method related to Hyvärinen and Morioka (2017b)—see also further self-supervised methods by Noroozi and Favaro (2016); Larsson et al. (2017). These approaches have allowed unsupervised data to be leveraged for supervised tasks resulting in dramatic performance improvements, but the papers acknowledge that they lack theoretical grounding.

Here, we propose a very general form of nonlinear ICA, based on the idea that the independent components are dependent on some additional auxiliary variable, while being conditionally mutually independent given the auxiliary variable. This unifies and generalizes the methods by Harmeling et al. (2003); Sprekeler et al. (2014); Hyvärinen and Morioka (2017b,a), giving a general framework where it is not necessary to specifically have a temporal (or even spatial) structure in the data. We prove exact identifiability conditions for the new framework, and show how it extends previous conditions, both from the viewpoint of theory and practice. In particular, our theory establishes mathematical principles underlying an important strand of self-supervised approaches by Arandjelovic and Zisserman (2017) and Korbar et al. (2018), showing that under certain conditions, they will extract the underlying latent variables from the data. We further provide a practical algorithm for estimating the model using the idea of contrastive learning (Gutmann and Hyvärinen, 2012; Hyvärinen and Morioka, 2017b,a), and prove its consistency.

2   BACKGROUND

We start by giving some background on nonlinear ICA theory. We explain the central problem of unidentifi-

ability of nonlinear ICA, and discuss some recently proposed solutions using time structure.

A straightforward generalization of ICA to the nonlinear case would assume, as pointed out above, a mixing model (1) with mutually independent latent variables s_i, and a general nonlinear mixing function $\mathbf{f}$, only assumed to be invertible and smooth. Now, if we further assume that the observations of $\mathbf{x}$ are independent and identically distributed (i.i.d.), the model is seriously unidentifiable. A well-known result, see e.g. Hyvärinen and Pajunen (1999), shows how to construct a function g such that for any two random variables x_1 and x_2, the function $g(x_1, x_2)$ is independent of x_1. This leads to the absurd case where based on independence alone, we could consider any of the observed variables an independent component.

Nor can we get any new information based on the non-Gaussianity of the variables, like in linear ICA, because we can trivially create a point-wise transformation $f(x_i)$ to have *any* marginal distribution by well-known theory (compounding the inverse cdf of the target distribution and the cdf of x_i).

One possibility to obtain identifiability is to restrict the nonlinearity $\mathbf{f}$. However, very few results are available in that direction, and usually based on very restrictive conditions, such as adding scalar nonlinearities to a linear mixing (Taleb and Jutten, 1999).

A more promising direction is to relax the assumption of i.i.d. sampling. A fundamental case is to consider time series, and the *temporal structure* of independent components. Thus, we assume

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

where t is the time index. As a first attempt, we can assume that the sources $s_i(t)$ have non-zero *autocorrelations*, which has a long history in the linear case (Tong et al., 1991; Belouchrani et al., 1997). Harmeling et al. (2003) proposed that we could try to find nonlinear transformations which are maximally uncorrelated even over time lags and after nonlinear scalar transformations. Sprekeler et al. (2014) showed that a closely related method enables separation of sources if they all have distinct autocorrelations functions. This constitutes probably the first identifiability proof for nonlinear ICA with general nonlinearities. However, it suffers from the restrictive condition that the sources must have different statistical properties, which is rather unrealistic in many cases.

An alternative framework was proposed by Hyvärinen and Morioka (2017a), where it was first heuristically proposed to transform the problem into a classification problem between two data sets, one constructed by concatenating real data points by taking a time

window of data, and the other by a randomized (permuted) concatenation. In other words, we define two data sets such as:

$$\tilde{\mathbf{x}}(t) = (\mathbf{x}(t), \mathbf{x}(t-1)) \quad \text{vs.} \quad \tilde{\mathbf{x}}^*(t) = (\mathbf{x}(t), \mathbf{x}(t^*)) \quad (3)$$

with a random time index t^*. We then train a neural network to discriminate between these two new data sets. Such Permutation-Contrastive Learning (PCL) (Hyvärinen and Morioka, 2017a) was shown to estimate independent components in the hidden layer, even if they have identical distributions, assuming they have *temporal dependencies* which are, loosely speaking, non-Gaussian enough. In the case of Gaussian sources, PCL estimates the sources under the same conditions as the method by Sprekeler et al. (2014). Thus, the PCL theory proves a stronger version of identifiability based on temporally dependent sources.

Another form of temporal structure that has been previously used in the case of linear ICA is *nonstationarity* (Matsuoka et al., 1995). This principle was extended to the nonlinear case by Hyvärinen and Morioka (2017b). The starting point was a heuristic principle where the time series is divided into a large number of segments. Then, a neural network is trained by multinomial regression so that each data point (i.e. time point) is assigned an artificially defined label given by the index of the time segment to which it belongs. Intuitively speaking, one would expect that the hidden layers of the neural network must learn to represent nonstationarity, since nonstationarity is nothing else than the differences between the distributions of the time segments. The ensuing method, Time-Contrastive Learning (TCL), was actually shown to enable estimation of a nonlinear ICA model where the independent components are assumed to be nonstationary, at the same time constituting another identifiability proof. Note that nonstationarity and temporal dependencies are two completely different properties which do not imply each other in any way.

3 NONLINEAR ICA USING AUXILIARY VARIABLES

Next, we propose our general framework for nonlinear ICA, as well as a practical estimation algorithm.

3.1 Definition of generative model

Assume the general $\mathbb{R}^n \to \mathbb{R}^n$ mixing model in (1) where the mixing function $\mathbf{f}$ is only assumed invertible and smooth (in the sense of having continuous second derivatives, and the same for its inverse). We emphasize the point that we do not restrict the function $\mathbf{f}$ to

any particular functional form. It can be modelled by a general neural network, since even the assumption of invertibility usually (empirically) seems to hold for the $\mathbf{f}$ estimated by the methods developed here, even without enforcing it.

The key idea here is that we further assume that each s_i is statistically dependent on some fully-observed m-dimensional random variable $\mathbf{u}$, but conditionally independent of the other s_j:

$$\log p(\mathbf{s}|\mathbf{u}) = \sum_{i=1}^n q_i(s_i, \mathbf{u}) \quad (4)$$

for some functions q_i. First, to see how this generalizes previous work on nonlinear ICA using time structure, we note that the auxiliary variable $\mathbf{u}$ could be the past of the component in the time series, giving rise to temporally dependent components as in permutation-contrastive learning or PCL (Hyvärinen and Morioka, 2017a) and the earlier methods by Harmeling et al. (2003); Sprekeler et al. (2014). Alternatively, $\mathbf{u}$ could be the time index t itself in a time series, or the index of a time segment, leading to nonstationary components as in time-contrastive learning or TCL (Hyvärinen and Morioka, 2017b). These connections will be considered in more detail below.

Thus, we obtain a unification of the separation principles of temporal dependencies and non-stationarity. This is remarkable since these principles are well-known in the linear ICA literature, but they have been considered as two distinct principles (Cardoso, 2001; Hyvärinen et al., 2001).

Furthermore, we can define $\mathbf{u}$ in completely new ways. In the case where each observation of $\mathbf{x}$ is an image or an image patch, a rather obvious generalization of TCL would be to assume $\mathbf{u}$ is the pixel index, or any similar spatial index, thus giving rise to nonlinear representation learning by the $\mathbf{s}$. In a visual feature extraction task, $\mathbf{x}$ could be images and $\mathbf{u}$ related audio or text (Arandjelovic and Zisserman, 2017). Moreover, $\mathbf{u}$ could be a class label, giving rise to something more related to conventional representation learning by a supervised neural network (e.g. ImageNet), but now connected to the theory of nonlinear ICA and identifiability (this is also considered in detail below). In a neuroscience context, $\mathbf{x}$ could be brain imaging data, and the $\mathbf{u}$ could be some quantity related to the stimuli in the experiment. Furthermore, $\mathbf{u}$ could be some combination of some of the above, thus providing a very general method. The appropriate definition of $\mathbf{u}$ obviously depends on the application domain, and the list above is by no means exhaustive.

It should be noted that the conditional independence does not imply that the s_i would be marginally inde-

pendent. If $\mathbf{u}$ affects the distributions of the s_i somehow independently (intuitively speaking), the s_i are likely to be marginally independent. This would be case, for example, if each q_i is of the from $q_i(s_i, u_i)$, that is, each source has one auxiliary variable which is not shared with the other sources, and the u_i are independent of each other. Thus, the formulation above is actually generalizing the ordinary independence in ICA to some extent.

3.2 Learning algorithm

To estimate our nonlinear ICA model, we propose a general form of contrastive learning, inspired by the idea of transforming unsupervised learning to supervised learning previously explored by Gutmann and Hyvärinen (2012); Goodfellow et al. (2014); Gutmann et al. (2017). More specifically, we use the idea of discriminating between a real data set and some randomized version of it, as used in PCL. Thus we define two datasets

$$\tilde{\mathbf{x}} = (\mathbf{x}, \mathbf{u}) \quad \text{vs.} \quad \tilde{\mathbf{x}}^* = (\mathbf{x}, \mathbf{u}^*) \tag{5}$$

where $\mathbf{u}^*$ is a random value from the distribution of the $\mathbf{u}$, but independent of $\mathbf{x}$, created in practice by random permutation of the empirical sample of the $\mathbf{u}$. We learn a nonlinear logistic regression system (e.g. a neural network) using a regression function of the form

$$r(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{n} \psi_i(h_i(\mathbf{x}), \mathbf{u}) \tag{6}$$

which then gives the posterior probability of the first class as $1/(1 + \exp(-r(\mathbf{x}, \mathbf{u})))$. Here, the scalar features h_i would typically be computed by hidden units in a neural network. Universal approximation capacity (Hornik et al., 1989) is assumed for the models of h_i and ψ_i. This is a variant of the "contrastive learning" approach to nonlinear ICA (Hyvärinen and Morioka, 2017b,a), and we will see below that it in fact unifies and generalizes those earlier results.

4 THEORETICAL ANALYSIS

In this section, we give exact conditions for the convergence (consistency) of our learning algorithm, which also leads to constructive proofs of identifiability of our nonlinear ICA model with auxiliary variables. It turns out we have two cases that need to be considered separately, based on the property of conditional exponentiality.

4.1 Definition of conditional exponentiality

We start by a basic definition describing distributions which are in some sense pathological in our theory.

Definition 1 *A random variable (independent component) s_i is conditionally exponential of order k given random vector $\mathbf{u}$ if its conditional pdf can be given in the form*

$$p(s_i | \mathbf{u}) = \frac{Q_i(s_i)}{Z_i(\mathbf{u})} \exp[\sum_{j=1}^{k} \tilde{q}_{ij}(s_i) \lambda_{ij}(\mathbf{u})] \tag{7}$$

almost everywhere in the support of $\mathbf{u}$, with $\tilde{q}_{ij}$, λ_{ij}, Q_i, and Z_i scalar-valued functions. The sufficient statistics $\tilde{q}_{ij}$ are assumed linearly independent (over j, for each fixed i).

This definition is a simple variant of the conventional theory of exponential families, adding conditioning by $\mathbf{u}$ which comes through the parameters only.

As a simple illustration, consider a (stationary) Gaussian time series as s_i, and define $\mathbf{u}$ as the past of the time series. The past of the time series can be compressed in a single statistic $\lambda(\mathbf{u})$ which essentially gives the conditional expectation of s_i. Thus, models of independent components using Gaussian autocorrelations lead to the conditionally exponential case, of order $k = 1$. As is well-known, the basic theory of linear ICA relies heavily on non-Gaussianity, the intuitive idea being that the Gaussian distribution is too "simple" to support identifiability. Here, we see a reflection of the same idea. Note also that if s_i and $\mathbf{u}$ are independent, s_i is conditionally exponential, since then we simply set $k = 1, \tilde{q}_{i1} \equiv 0$.

In the following, we analyse our algorithm separately for the general case, and for conditionally exponential independent components of low order k. The fundamental result is that for sufficiently complex source distributions, the independent components are estimated up to component-wise nonlinear transformations; if the data comes from an exponential family of low order, there is an additional linear transformation that remains to be determined (by linear ICA, for example).

4.2 Theory for general case

First, we consider the much more general case of distributions which are not "pathological". Our main theorem, proven in Supplementary Material A, is as follows:

Theorem 1 *Assume*

1. *The observed data follows the nonlinear ICA model with auxiliary variables in Eqs. (1,4).*

2. *The conditional log-pdf q_i in (4) is sufficiently smooth as a function of s_i, for any fixed $\mathbf{u}$.*

3. **[Assumption of Variability]** *For any* $\mathbf{y} \in \mathbb{R}^n$, *there exist* $2n+1$ *values for* $\mathbf{u}$, *denoted by* $\mathbf{u}_j$, $j = 0...2n$ *such that the* $2n$ *vectors in* $\mathbb{R}^{2n}$ *given by*

$$(\mathbf{w}(\mathbf{y}, \mathbf{u}_1) - \mathbf{w}(\mathbf{y}, \mathbf{u}_0)), (\mathbf{w}(\mathbf{y}, \mathbf{u}_2) - \mathbf{w}(\mathbf{y}, \mathbf{u}_0)),$$
$$..., (\mathbf{w}(\mathbf{y}, \mathbf{u}_{2n}) - \mathbf{w}(\mathbf{y}, \mathbf{u}_0)) \quad (8)$$

with

$$\mathbf{w}(\mathbf{y}, \mathbf{u}) = \left(\frac{\partial q_1(y_1, \mathbf{u})}{\partial y_1}, \ldots, \frac{\partial q_n(y_n, \mathbf{u})}{\partial y_n}, \right.$$
$$\left. \frac{\partial^2 q_1(y_1, \mathbf{u})}{\partial y_1^2}, \ldots, \frac{\partial^2 q_n(y_n, \mathbf{u})}{\partial y_n^2} \right) \quad (9)$$

are linearly independent.

4. *We train some nonlinear logistic regression system with universal approximation capability to discriminate between* $\tilde{\mathbf{x}}$ *and* $\tilde{\mathbf{x}}^*$ *in (5) with regression function in (6).*

5. *In the regression function in Eq. (6), we constrain* $\mathbf{h} = (h_1, ..., h_n)$ *to be invertible, as well as smooth, and constrain the inverse to be smooth as well.*

Then, in the limit of infinite data, $\mathbf{h}$ *in the regression function provides a consistent estimator of demixing in the nonlinear ICA model: The functions (hidden units)* $h_i(\mathbf{x})$ *give the independent components, up to scalar (component-wise) invertible transformations.*

Essentially, the Theorem shows that under mostly weak assumptions, including invertibility of $\mathbf{h}$ and smoothness of the pdfs, and of course independence of the components, our learning system will recover the independent components given an infinite amount of data. Thus, we also obtain a constructive identifiability proof of our new, general nonlinear ICA model.

Among the assumptions above, the only one which cannot considered weak or natural is clearly Assumption of Variability (#3), which is central in the our developments. It is basically saying that the auxiliary variable must have a sufficiently strong and diverse effect on the distributions of the independent components. To further understand this condition, we give the following Theorem, proven in Supplementary Material B:

Theorem 2 *Assume the independent components are conditionally exponential given* $\mathbf{u}$, *with the same order* k *for all components. Then,*

1. *If* $k = 1$, *the Assumption of Variability cannot hold.*

2. *Assume* $k > 1$ *and for each component* s_i, *the vectors* $\left(\frac{\partial \tilde{q}_{ij}(s_i, \mathbf{u})}{\partial s_i}, \frac{\partial^2 \tilde{q}_{ij}(s_i, \mathbf{u})}{\partial s_i^2} \right)$ *are not all proportional*

to each other for different $j = 1, \ldots, k$, *for* $\mathbf{s}$ *almost everywhere. Then, the Assumption of Variability holds almost surely if the* λ's *are statistically independent and follow a distribution whose support has non-zero measure.*

Loosely speaking, the Assumption of Variability holds if the sources, or rather their modulation by $\mathbf{u}$, is not "too simple", which is here quantified as the order of the exponential family from which the s_i are generated. Furthermore, for the second condition of Theorem 2 to hold, the sufficient statistics cannot be linear (which would lead to zero second derivatives), thus excluding the Gaussian scale-location family as too simple as well. (See Supplementary Material D for an alternative formulation of the assumption.)

Another non-trivial assumption in Theorem 1 is the invertibility of $\mathbf{h}$. It is hoped that the constraint of invertibility is only necessary to have a rigorous theory, and not necessary in any practical implementation. Our simulations below, as well as our next Theorem, seem to back up this conjecture to some extent.

4.3 Theory for conditionally exponential case

The theory above excluded the conditionally exponential case of order one (Theorem 2). This is a bit curious since it is actually the main model considered in TCL (Hyvärinen and Morioka, 2017b). In fact, the exponential family model of nonstationarities in that work is nothing other than a special case of our "conditionally exponential" family of distributions; we will consider the connection in detail in the next section.

There is actually a fundamental difference between Theorem 1 above and the TCL theory in (Hyvärinen and Morioka, 2017b). In TCL, and in contrast to our current results, a linear indeterminacy remains—but the TCL theory never showed that such an indeterminacy is a property of the model and not only of the particular TCL algorithm employed by Hyvärinen and Morioka (2017b).

Next, we construct a theory for conditionally exponential families adapting our current framework, and indeed, we see the same kind of linear indeterminacy as in TCL appear. We give the result for general k, although the case $k = 1$ is mainly of interest:

Theorem 3 *Assume*

1. *The data follows the nonlinear ICA model with auxiliary variables in Eqs. (1,4).*

2. *Each* s_i *is conditionally exponential given* $\mathbf{u}$ *(Def. 1).*

3. *There exist $nk+1$ points $\mathbf{u}_0, \ldots, \mathbf{u}_{nk}$ such that the matrix of size $nk \times nk$*

$$\bar{\mathbf{L}} = \begin{pmatrix} \lambda_{11}(\mathbf{u}_1) - \lambda_{11}(\mathbf{u}_0), ..., \lambda_{11}(\mathbf{u}_{nk}) - \lambda_{11}(\mathbf{u}_0) \\ \vdots \\ \lambda_{nk}(\mathbf{u}_1) - \lambda_{nk}(\mathbf{u}_0), ..., \lambda_{nk}(\mathbf{u}_{nk}) - \lambda_{nk}(\mathbf{u}_0) \end{pmatrix} \quad (10)$$

is invertible (here, the rows corresponds to all the nk possible subscript pairs for λ).

4. *We train a nonlinear logistic regression system with universal approximation capability to discriminate between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^*$ in (5) with regression function in (6).*

Then,

1. *The optimal regression function can be expressed in the form*

$$r(\mathbf{x}, \mathbf{u}) = \tilde{\mathbf{h}}(\mathbf{x})^T \mathbf{v}(\mathbf{u}) + a(\mathbf{x}) + b(\mathbf{u}) \quad (11)$$

for some functions $\mathbf{v} : \mathbb{R}^m \to \mathbb{R}^{nk}$, $\tilde{\mathbf{h}} : \mathbb{R}^n \to \mathbb{R}^{nk}$ and two scalar-valued functions a, b.

2. *In the limit of infinite data, $\tilde{\mathbf{h}}(\mathbf{x})$ provides a consistent estimator of the nonlinear ICA model, up to a linear transformation of point-wise scalar (not necessarily invertible) functions of the independent components. The point-wise nonlinearities are given by the sufficient statistics $\tilde{q}_i$. In other words,*

$$\big(\tilde{q}_{11}(s_1), \tilde{q}_{12}(s_1), \ldots, \tilde{q}_{21}(s_2), \ldots, \tilde{q}_{nk}(s_n)\big)^T \\ = \mathbf{A}\tilde{\mathbf{h}}(\mathbf{x}) - \mathbf{c} \quad (12)$$

for some unknown matrix $\mathbf{A}$ and an unknown vector $\mathbf{c}$.

The proof, found in Supplementary Material C, is quite similar to the proof of Theorem 1 by Hyvärinen and Morioka (2017b). Although the statistical assumptions made here are different, the very goal of modelling exponential sources by logistic regression means the same linear indeterminacy appears, based on the linearity of the log-pdf in exponential families.

5 DIFFERENT DEFINITIONS OF AUXILIARY VARIABLES

Next, we consider different possible definitions of the auxiliary variable, and show some exact connections and generalization of previous work.

First, we want to emphasize that any arbitrary definition of $\mathbf{u}$ is not possible since many definitions are likely to violate the central assumption of conditional independence of the components. For example, one might be tempted to choose a $\mathbf{u}$ which is some deterministic function of $\mathbf{x}$; in Supplementary Material E we give a simple example showing how this violates the conditional independence.

5.1 Using time as auxiliary variable

A real practical utility of the new framework can be seen in the case of nonstationary data. Assume we observe a time series $\mathbf{x}(t)$ as in Eq. (2). Assume the n independent components are nonstationary, with densities $p(s_i|t)$. For analysing such nonstationary data in our framework, define $\mathbf{x} = \mathbf{x}(t)$ and $\mathbf{u} = t$. We can easily consider the time index as a random variable, observed for each data point, and coming from a uniform distribution. Thus, we create two new datasets by augmenting the data by adding the time index:

$$\tilde{\mathbf{x}} = (\mathbf{x}(t), t) \quad \text{vs.} \quad \tilde{\mathbf{x}}^* = (\mathbf{x}(t), t^*) \quad (13)$$

We analyse the nonstationary structure of the data by learning to discriminate between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^*$ by logistic regression. Directly applying the general theory above, we define the regression function to have the following form:

$$r(\mathbf{x}, t) = \sum_i \psi_i(h_i(\mathbf{x}), t) \quad (14)$$

where each ψ_i is $\mathbb{R}^2 \to \mathbb{R}$. Intuitively, this means that the nonstationarity is separately modelled for each component, with no interactions.

Theorems 1 and 3 above give exact conditions for the consistency of such a method. This provides an alternative way of estimating the nonstationary nonlinear ICA model proposed in (Hyvärinen and Morioka, 2017b) as a target for the TCL method.

A practical advantage is that if the assumptions of Theorem 1 hold, the method actually captures the independent components directly: There is no indeterminacy of a linear transformation unlike in TCL. Nor is there any nonlinear non-invertible transformation (e.g. squaring) as in TCL, although this may come at the price of constraining $\mathbf{h}$ to be invertible. The Assumption of Variability in Theorem 1 is quite comparable to the corresponding full rank condition in the convergence theory of TCL. Another advantage of our new method is that there is no need to segment the data, although in our simulations below we found that segmentation is computationally very useful. From a theoretical perspective, the current theory in Theorem 1 is also much more general than the TCL theory since no assumption of an exponential family is needed — "too simple" exponential families are in fact considered separately in Theorem 3.

5.2 Using history as auxiliary variables

Next, we consider the theory in the case where $\mathbf{u}$ is the history of each variable. For the purposes of our present theory, we define $\mathbf{x} = \mathbf{x}(t)$ and $\mathbf{u} = \mathbf{x}(t-1)$ based on a time-series model in (2). So, the nonlinear ICA model in Eqs. (1, 4) holds. Note that here, it does not make any difference if we use the past of $\mathbf{x}$ or of $\mathbf{h}(\mathbf{x})$ as $\mathbf{u}$ since they are invertible functions of each other. Each component follows a distribution

$$q_i(s_i, \mathbf{u}) = q_i(s_i(t), s_i(t-1)) \qquad (15)$$

This model is the same as in PCL (Hyvärinen and Morioka, 2017a), and in fact PCL is thus a special case of the discrimination problem we formulated in this paper. Likewise, the restriction of the regression function in (6) is very similar to the form imposed in Eq. (12) of (Hyvärinen and Morioka, 2017a). Thus, essentially, Theorem 1 above provides an alternative identifiability proof of the model in (Hyvärinen and Morioka, 2017a), with quite similar constraints. See Supplementary Material F for a detailed discussion on the connection. Our goal here is thus not to sharpen the analysis of (Hyvärinen and Morioka, 2017a), but merely to show that that model falls into the present framework with minimal modification.

5.3 Combining time and history

Another generalization of previously published theory which could be of great interest in practice is to combine the nonstationarity-based model in TCL (Hyvärinen and Morioka, 2017b) with the temporal dependencies model in PCL (Hyvärinen and Morioka, 2017a). Clearly, we can combine these two by defining $\mathbf{u} = (\mathbf{x}(t-1), t)$, and thus discriminating between

$$\tilde{\mathbf{x}}(t) = (\mathbf{x}(t), \mathbf{x}(t-1), t) \text{ vs. } \tilde{\mathbf{x}}^*(t) = (\mathbf{x}(t), \mathbf{x}(t^*-1), t^*)$$

with a random time index t^*, and accordingly defining the regression function as

$$r_{\mathrm{comb}}(\mathbf{x}(t), \mathbf{x}(t-1), t) = \sum_{i=1}^{n} \psi_i(h_i(\mathbf{x}(t)), h_i(\mathbf{x}(t-1)), t)$$

Such a method now has the potential of using both nonstationarity and temporal dependencies for nonlinear ICA. Thus, there is no need to choose which method to use, since this combined method uses both properties. (See Supplementary Material G for an alternative formulation.)

5.4 Using class label as auxiliary variable

Finally, we consider the very interesting case where the data includes class labels as in a classical supervised setting, and we use them as the auxiliary variable. Let us note that the existence of labels does not mean a nonlinear ICA model is not interesting, because our interest might not be in classifying the data using these labels, but rather in understanding the structure of the data, or possibly, finding useful features for classification using some other labels as in transfer learning. In particular, with scientific data, the main goal is usually to understand its structure; if the labels correspond to different treatments, or experimental conditions, the classification problem in itself may not be of great interest. It could also be that the classes are somehow artificially created, as in TCL, and thus the whole classification problem is of secondary interest.

Formally, denote by $c \in \{1, .., k\}$ the class label with k different classes. As a straight-forward application of the theory above, we learn to discriminate between

$$\tilde{\mathbf{x}} = (\mathbf{x}, c) \text{ vs. } \tilde{\mathbf{x}}^* = (\mathbf{x}, c^*) \qquad (16)$$

where c is the class label of $\mathbf{x}$, and c^* is a randomized class label; "one-hot" coding of c could also be used. Note that we could also apply the TCL method and theory on such data, simply using the c as class labels instead of the time segment indices as in (Hyvärinen and Morioka, 2017b). Applying Theorem 1, we see that, interestingly, we have no linear indeterminacy, unlike in TCL (unless the data follows a conditionally exponential source model of low rank, in which case we fall back to Theorem 3.) Thus, the current theorem seems to be in some sense stronger than the TCL theory, although it is not a strict generalization. In either case, we use the class labels to estimate independent components, thus combining supervised and unsupervised learning in an interesting, new way.

6 SIMULATIONS

To test the performance of the method, we applied it on non-stationary sources similar to those used in TCL. This is the case of main interest here since for temporally correlated sources, the framework gives PCL. It is not our goal to claim that the new method performs better than TCL, but rather to confirm that our new very general framework includes something similar to TCL as well.

First, we consider the non-conditionally-exponential case in Theorem 1, where the data does not follow a conditionally exponential family, and the regression function has the general form in (6). We artificially generated nonstationary sources s_i on a 2D grid indexed by ξ, η by a scale mixture model: $s_i(\xi, \eta) = \sigma_i(\xi, \eta) \cdot z_i(\xi, \eta)$, $i = 1, \ldots, n$, where z_i is a standardized Laplacian variable, and the scale components $\sigma_i(\xi, \eta)$ were generated by creating Gaussian

blobs in random locations to represent areas of higher variance. The number of dimensions was 5 and the number of data points 2^{16}. The mixing function $\mathbf{f}$ was a random three-layer feedforward neural network as in (Hyvärinen and Morioka, 2017b). We used the spatial index pair as $\mathbf{u} := (\xi, \eta)$. We modelled $\mathbf{h}(\mathbf{x})$ by a feedforward neural network with three layers: The number of units in the hidden layers was $2n$, except in the final layer where it was n; the nonlinearity was max-out except for the last layer where absolute values were taken; L_2 regularization was used to prevent overlearning. The function ψ_i was also modelled by a neural network. In contrast to the assumptions of Theorem 1, no constraint related to the invertibility of h was imposed. After learning the neural network, we further applied FastICA to the estimated features (heuristically inspired by Theorem 3). Performance was evaluated by the Pearson correlation between the estimated sources and the original sources (after optimal matching and sign flipping). The results are shown in Fig. 1 a). Our method has performance similar to TCL.

Second, we considered the conditionally exponential family case as in Theorem 3. We generated nonstationary sources s_i as above, but we generated them as time-series, and divided the time series into equispaced segments. We used a simple random neural network to generate separate variances σ_i inside each segment. The mixing function was as above. Here, we used the index of the segment as $\mathbf{u}$. This means we are also testing the applicability of using a class label as the auxiliary variable as in Section 5.4. We modelled $\mathbf{h}(\mathbf{x})$ as above. The $\mathbf{v}$ and b in (11) were modelled by constant parameter vectors inside each segment, and a by another neural network. Performance was evaluated by the Pearson correlation of the absolute values of the components, since the sign remains unresolved in this case. The results are shown in Fig. 1 b). Again our method has performance similar to TCL, confirming that source separation by nonstationarity, as well as using class labels as in Section 5.4, can be modelled in our new framework.

7 CONCLUSION

We introduced a new framework for nonlinear ICA. To solve the problem of non-identifiability central to nonlinear ICA theory, we assume there is an external, auxiliary variable, such that conditioning by the auxiliary variables changes the distributions of the independent components. In a time series, the auxiliary variable can correspond to the history, or the time index, thus unifying the previous frameworks (Sprekeler et al., 2014; Hyvärinen and Morioka, 2017b,a) both in theory and practice.
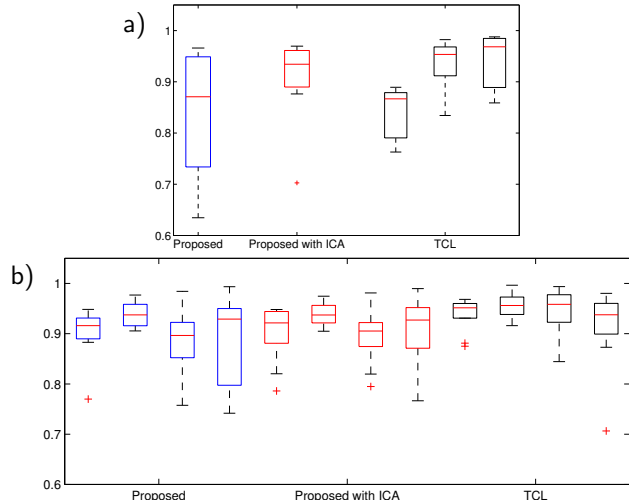

Figure 1: Performance measured by correlations between estimates and original quantities (see text). The non-conditionally-exponential case is given in (a) and the exponential family case in (b). "Proposed" is taking raw outputs from neural network learned by our new method, "Proposed with ICA" is adding final linear ICA, "TCL" is time-contrastive learning (with final linear ICA) given for comparison. In a), TCL was performed with $16, 64$, and 256 time segments. In b), for each method, we report four cases, with $10, 50, 100$, and 300 time segments.

We gave exact conditions for identifiability, showing how the definition of conditional exponentiality divides the problem into two domains. Conditional exponentiality interestingly corresponds to the simplest case of TCL theory in (Hyvärinen and Morioka, 2017b). In the special case of nonstationary components like in TCL, we actually relaxed the assumption of an exponential family model for the independent components, and removed the need to segment the data, which may be difficult in practice; nor was there any remaining linear mixing, unlike in TCL. This result carried over to the case where we actually have class labels available; we argued that the identifiability theory of nonlinear ICA is interesting even in such an apparently supervised learning case. We also provided a learning algorithm based on the idea of contrastive learning by logistic regression, and proved its consistency.

Recent work has successfully used a very similar idea for "self-supervised" audio-visual feature extraction from a purely heuristic perspective (Arandjelovic and Zisserman, 2017; Korbar et al., 2018), and our theory hopefully elucidates the mathematical principles underlying such methods. Yet, our framework is quite versatile, and the auxiliary variables can be defined in many different ways depending on the application.

References

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

Arandjelovic, R. and Zisserman, A. (2017). Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617. IEEE.

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.

Brakel, P. and Bengio, Y. (2017). Learning independent features with adversarial nets for non-linear ICA. *arXiv preprint arXiv:1710.05050*.

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.

Comon, P. (1994). Independent component analysis—a new concept? *Signal Processing*, 36:287–314.

Deco, G. and Obradovic, D. (1995). Linear redundancy reduction learning. *Neural Networks*, 8(5):751–755.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer: New York.

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.

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Likelihood-free inference via classification. *Statistics and Computing*. doi:10.1007/s11222-017-9738-6.

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.

Hjelm, R. D. and et al. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

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

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley Interscience.

Hyvärinen, A. and Morioka, H. (2017a). Nonlinear ICA of temporally dependent stationary sources. In *Proc. Artificial Intelligence and Statistics (AISTATS2017)*, Fort Lauderdale,Florida.

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

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

Jutten, C., Babaie-Zadeh, M., and Karhunen, J. (2010). Nonlinear mixtures. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pages 549–592.

Korbar, B., Tran, D., and Torresani, L. (2018). Cotraining of audio and video representations from self-supervised temporal synchronization. *arXiv preprint arXiv:1807.00230*.

Larsson, G., Maire, M., and Shakhnarovich, G. (2017). Colorization as a proxy task for visual understanding. In *CVPR*, volume 2, page 8.

Matsuoka, K., Ohya, M., and Kawamoto, M. (1995). A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419.

Misra, I., Zitnick, C. L., and Hebert, M. (2016). Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer.

Noroozi, M. and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proc. European Conference on Computer Vision*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pham, D.-T. and Cardoso, J.-F. (2001). Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Trans. Signal Processing*, 49(9):1837–1848.

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.

Taleb, A. and Jutten, C. (1999). Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820.

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.