Hidden Markov Nonlinear ICA: Unsupervised Learning from Nonstationary Time Series

Hermanni Hälvä
University of Helsinki

Aapo Hyvärinen
Université Paris-Saclay, Inria
University of Helsinki

Abstract

Recent advances in nonlinear Independent Component Analysis (ICA) provide a principled framework for unsupervised feature learning and disentanglement. The central idea in such works is that the latent components are assumed to be independent conditional on some observed auxiliary variables, such as the time-segment index. This requires manual segmentation of data into non-stationary segments which is computationally expensive, inaccurate and often impossible. These models are thus not fully unsupervised. We remedy these limitations by combining nonlinear ICA with a Hidden Markov Model, resulting in a model where a latent state acts in place of the observed segment-index. We prove identifiability of the proposed model for a general mixing nonlinearity, such as a neural network. We also show how maximum likelihood estimation of the model can be done using the expectation-maximization algorithm. Thus, we achieve a new nonlinear ICA framework which is unsupervised, more efficient, as well as able to model underlying temporal dynamics.

1 INTRODUCTION

Representation learning – the task of finding useful features from data – is one of the main challenges in unsupervised learning. Recent theoretical and practical advances in Nonlinear ICA provide a principled approach to this problem (Hyvärinen and Morioka, 2016; Hyvärinen and Morioka, 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020; Sorrenson et al., 2020). These works frame Nonlinear ICA as deep generative models, which allows them to harness deep neural networks to recover latent independent components from observed data. Identifiability of the latent components can be guaranteed by explicitly defining probabilistic generative models with appropriate conditional independence structures. A general framework was proposed recently by Hyvärinen et al. (2019), who assumed that the components are independent given some other observed auxiliary variable. For example, in time-series data this can be the time-index or segment-index if the data is non-stationary, as was earlier assumed in Time-Contrastive Learning or TCL (Hyvärinen and Morioka, 2016). Non-stationarity is a fundamental property of many applications, since for example, video, audio, and most neuroscience data are non-stationary.

A crucial limitation of all of the above nonlinear ICA models is that the conditioning auxiliary variable is always assumed observable. In some sense, these models are therefore not fully unsupervised. If, for instance, we wish to exploit the nonstationary temporal structure optimally in estimating independent components, TCL would require segment indices that correspond to the different latent data generative states. In practice we don't observe such states so the default approach is to manually segment the data.

In general, it is unrealistic to assume that we can infer from observed data alone the exact time-points at which the latent data distribution changes. In fact, such change-points may not exist at all. Segmenting data manually is also infeasible for large datasets. The default approach is therefore to segment data at equal intervals, however, this is problematic for various reasons. Consider, for example, a situation where the true latent state switches between five different states. By segmenting the data at equal intervals we will end up with an unnecessarily large number of states where just a few would have done the job. This is computationally expensive, inaccurate and will completely miss out on temporal dynamics in situations where the latent states repeat over time.

In fact, often a reasonable assumption is that non-stationarity can be succinctly summarized using a limited

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

number of segment indices or latent states, and properly modelling such state switching is likely to improve learning. Notice that even if ground-truth nonstationary information was available, the existing methods lack the machinery to perform inference on latent temporal dynamics. In many applications, for example brain imaging, describing the dynamics in terms of latent states could be very useful in its own right.

The points above highlight the need for a nonlinear ICA method that is able to cluster observations and learn latent states and their temporal dynamics in an unsupervised fashion. A well-known approach to modelling hidden latent states in time series is to use a Hidden Markov Model (HMM). HMMs can be viewed as probabilistic mixture models where the discrete latent states, which specify the data generating distribution, are time-dependent with Markov dynamics. HMMs are especially well suited for modelling non-stationary data as they automatically allow for a representation of the time series in terms of a discrete number of states.

In this work, we therefore resolve the above limitations by combining Nonlinear ICA with a HMM. This idea has been proposed earlier for linear ICA (Penny et al., 2000; Zhou and Zhang, 2008) but their identifiability and estimation results do not directly extend to the nonlinear case. In our model, we achieve this by having the latent state act in place of the conditioning auxiliary variable in the framework of Hyvärinen et al. (2019). Importantly, we are able to prove that Hidden Markov Nonlinear ICAs are identifiable. Attaining identifiability has been a major research focus for both Nonlinear ICA (Hyvärinen and Morioka, 2017; Hyvärinen et al., 2019) and HMMs (Allman et al., 2009; Gassiat et al., 2016), and therefore much of our paper is devoted to combining these two research strands. To the best of our knowledge, this is the first fully unsupervised non-linear ICA, in the sense that the model's identifiability comes from an *unobserved* conditioning variable which is inferred from the time series as a part of learning. We further show how the structure of the model allows us to use the Expectation-Maximization (EM) algorithm for parameter estimation. In practice the Hidden Markov Nonlinear ICA is endowed with rich representation learning capabilities that allow it to simultaneously extract independent components and to learn the dynamics of the latent state that drives non-stationarity data, as illustrated by our simulations.

2 BACKGROUND

We start by giving an overview of the problem of unidentifiability in both nonlinear ICA and HMMs, and recently proposed solutions. For both types of models, identifiability arises as a consequence of appropriate temporal structures which suggests a natural synthesis between the two.

2.1 NONLINEAR ICA AND IDENTIFIABILITY

Consider a parametric model of observed data $\mathbf{x}$ with marginal likelihood $p_{\boldsymbol{\theta}}(\mathbf{x})$. This model is *identifiable* if it fulfils below:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}' : \forall(\boldsymbol{\theta}, \boldsymbol{\theta}') \tag{1}$$

In the context of a latent variable model, this is closely connected to the idea of being able to recover the original latent variables, as discussed by Khemakhem et al. (2020).

Assume we observe N-dimensional data at discrete time-steps, $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_N^{(t)})$. Simple nonlinear ICA can be defined as the task of estimating an unobserved N-dimensional independent component vector $\mathbf{s}^{(t)} = (s_1^{(t)}, \ldots, s_N^{(t)})$ such that $p(\mathbf{s}^{(t)}) = \prod_{i=1}^{N} p(s_i)$, as well as the inverse of a mixing function $\mathbf{f}$, that has generated the observed data:

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

Unfortunately, without a temporal structure, that is if $\mathbf{x}^{(t)}$ are i.i.d over the time-index, and if there are no constraints on $\mathbf{f}$, then this model is unidentifiable (Hyvärinen and Pajunen, 1999). In fact, the authors show that there are infinite potential nonlinear transformations and independent components that would satisfy the model, with no criterion for choosing one of them over the others.

In order to make the model identifiable, constraints are thus needed. For time-series data, this comes naturally by placing restrictions on the temporal structure of the model. For linear ICA this approach has been shown to yield identifiable models (Belouchrani et al., 1997; Tong et al., 1991), and extensions to the nonlinear case have been also been proposed in earlier work (Harmeling et al., 2003; Sprekeler et al., 2014). The first fully rigorous proof of an identifiable nonlinear ICA model, along with an estimation algorithm (Time-Contrastive Learning or TCL), was given by Hyvärinen and Morioka (2016). The constraint imposed in that work is that of a non-stationary data generative process such that independent component vectors within different time-segments have different distributional parameters. Specifically, the model assumes that each independent component has an exponential family distribution, where the time segment index τ modulates the natural parameters (denoted as $\boldsymbol{\lambda}$):

$$p_\tau(s_i) = \frac{q_i(s_i)}{Z(\boldsymbol{\lambda}_i)} \exp\{\langle \boldsymbol{\lambda}_i(\tau), \mathbf{T}(s_i) \rangle\} \tag{3}$$

where q_i is the base measure and $\mathbf{T}$ are the sufficient statistics. TCL then assumes that the independent components in all the segments are transformed into observed

variables by some mixing function (2). The authors prove identifiability up to a linear transformation of pointwise functions of the components:

$$\mathbf{T}(\mathbf{s}^{(t)}) = \mathbf{A}\mathbf{h}(\mathbf{x}^{(t)}; \boldsymbol{\theta}) + \mathbf{b} \tag{4}$$

By learning to contrast between the different segments, the TCL algorithm learns the inverse of the mixing function and the independent components.

This seminal work has inspired other frameworks for identifiable nonlinear ICA estimation. Permutation Contrastive Learning (Hyvärinen and Morioka, 2017), for instance, exploits temporal dependencies, rather than non-stationarity, to identify independent components. The unifying tenet of these identifiable nonlinear ICA algorithms is that independent components are *conditionally independent* given some observed auxiliary variable. This general idea was formalized in Hyvärinen et al. (2019), of which both the TCL (segment index as auxiliary variable) and the PCL (past data as auxiliary variable) are special cases.

These identifiable nonlinear ICA models provide a principled approach to finding meaningful data representations. This is in contrast to the majority of recent deep generative models used for representation learning, such as VAEs (Kingma and Welling, 2014) and GANs (Goodfellow et al., 2014), which are all malaised by unidentifiability. In fact, any generative latent variable model with an unconditional prior is unidentifiable. This issue is portrayed in depth by Khemakhem et al. (2020) who resolve it by introducing identifiable VAE (iVAE). Like regular VAE, this model estimates a full generative model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{s})$, but with a factorial conditional prior $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{u})$. As in Hyvärinen et al. (2019), $\mathbf{u}$ is some auxiliary variable, and iVAE provides a novel algorithm to estimate nonlinear independent components in the same identifiable framework. iVAE however suffers from the same problems as TCL, as its auxiliary variable $\mathbf{u}$ has to be observed.

2.2 HIDDEN MARKOV MODELS AND IDENTIFIABILITY

In order to define HMMs, let $\mathbf{x}^{(t)} \in \mathbb{R}^n$ be an observed random variable from a time series with a discrete time index $t \in \{1, \ldots, T\}$. In a standard hidden Markov model, distribution of the observations depends conditionally on a discrete latent state random variable $c^{(t)}$ as per $p(\mathbf{x}^{(t)}|c^{(t)})$; we refer to this as the emission distribution. The latent state $c^{(t)}$ undergoes first-order Markov process governed by a $C \times C$ transition-probability matrix $\mathbf{A}$. $A_{i,j}$ is used to denote the probability of transitioning from state $c^{(t)} = i$ to $c^{(t+1)} = j$, and $\pi(c^{(1)})$ the starting-state probabilities. The likelihood of a typical HMM is hence given by:

$$p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}; \mathbf{A}) = \sum_{c^{(1)}, \ldots, c^{(T)}} \pi(c^{(1)}) p(\mathbf{x}^{(1)}|c^{(1)}) \prod_{t=2}^{T} A_{c^{(t-1)}, c^{(t)}} p(\mathbf{x}^{(t)}|c^{(t)}) \tag{5}$$

HMMs can be viewed as mixture models where the latent state is coupled across time by a Markov process. This observation raises the question of identifiability since mixture models with non-parametric emission distributions are generally unidentifiable, though many commonly used parametric forms are identifiable (Allman et al., 2009). Recently, however, Gassiat et al. (2016) have proven a major result that nonparametric HMMs are in general identifiable under some mild assumptions. We will use this result later and thus reproduce it here (notice that their nonparametric result subsumes parametric HMMs):

Theorem 1. *(Gassiat et al., 2016) Assume the number of latent states, C , is known. Use $\mu_1, \ldots, \mu_C \in \mathbb{R}^N$ to denote nonparametric probability distributions of the C emission distributions. Also assume that the transition-matrix $\mathbf{A}$ is full rank. Then the parameters $\mathbf{A}$ and $M = (\mu_1, \ldots, \mu_C)$ are identifiable given the distribution, $\mathbb{P}^{(3)}_{\mathbf{A},M}$, of at least 3 consecutive observations $\mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}, \mathbf{x}^{(t+1)}$, up to label swapping of the hidden states, that is: if $\widehat{\mathbf{A}}$ is a $C \times C$ transition matrix, if $\widehat{\pi}(c)$ is a stationary distribution of $\widehat{\mathbf{A}}$ with $\widehat{\pi}(c) > 0$ $\forall c \in \{1, \ldots, C\}$, and if $\hat{M} = (\hat{\mu}_1, \ldots, \hat{\mu}_C)$ are C probability distributions on $\mathbb{R}^N$ that verify the equality of the HMM distribution functions $\mathbb{P}^{(3)}_{\widehat{\mathbf{A}},\hat{M}} = \mathbb{P}^{(3)}_{\mathbf{A},M}$, then there exists a permutation σ of the set $\{1, \ldots, C\}$ such that for all $k, l = 1, \ldots, C$ we have $\hat{A}_{k,l} = A_{\sigma(k),\sigma(l)}$ and $\hat{\mu}_k = \mu_{\sigma(k)}$.*

Much like for ICA, identifiability in nonparametric HMMs is a result of temporal structure, namely observations across time are independent conditionally on the latent state—which is in contrast to simple (i.i.d.) mixture models for which similar identifiability results are not available. We show below that this temporal structure of the HMM's, together with nonstationarity similar to TCL, combine to identify the resulting Hidden Markov Nonlinear ICA model.

3 IDENTIFIABLE NONLINEAR ICA FOR NONSTATIONARY DATA

In this section, we propose a combination of a hidden Markov model and nonlinear ICA. Specifically, we propose an HMM which has nonlinear ICA as its emission model, and show how to estimate it by Expectation-Maximization.

3.1 MODEL DEFINITION

To incorporate nonlinear ICA into the standard HMM of (5) we define the emission distribution $p(\mathbf{x}^{(t)}|c^{(t)})$ as a deep latent variable model. First, the latent independent component variables $\mathbf{s}^{(t)} \in \mathbb{R}^N$ are generated from a factorial exponential family prior, given the hidden state $c^{(t)}$, as

$$p(\mathbf{s}^{(t)}|c^{(t)};\boldsymbol{\lambda}_{c^{(t)}}) = \prod_{i=1}^{N} p(s_i^{(t)}|c^{(t)};\boldsymbol{\lambda}_{i,c^{(t)}})$$
$$= \prod_{i=1}^{N} \frac{h(s_i^{(t)})}{Z(\boldsymbol{\lambda}_{i,c^{(t)}})} \exp\{\langle \boldsymbol{\lambda}_{i,c^{(t)}}, \mathbf{T}_i(s_i)\rangle\} \quad (6)$$

where $h(.)$ are the base measures, $Z(\boldsymbol{\lambda}_{i,c^{(t)}})$ the normalizing constants, and $\mathbf{T}_i : \mathbb{R} \to \mathbb{R}^V$ the sufficient statistics. Second, the observed data is generated by a nonlinear mixing function as in Eq. (2).

For remainder of the paper we assume that the exponential family is in minimal representation form so that the sufficient statistics are linearly independent. The corresponding V-dimensional parameter vectors are denoted by $\boldsymbol{\lambda}_{i,c^{(t)}}$. The subscripts indicate that the parameters of the N different components are modulated directly, and independently, by the HMM latent state. Indeed, it is precisely this conditional dependence of the parameters on the discrete latent state that seeps through our model and generates non-stationary observed data. Note that the parameters themselves are time-homogeneous, that is they are constant over time; instead, the latent state evolves over time and determines which set of parameters is active a point in time. In other words, non-stationary arises purely from the dynamics of the latent state $c^{(t)}$. The full set of parameters for the independent components can hence be captured by a $C \times NV$ matrix $\mathbf{L}$ (plus the transition probabilities of the hidden states).

The nonlinear mixing function $\mathbf{f}$ in Eq. (2) is assumed to be bijective with inverse given by $\mathbf{s}^{(t)} = \mathbf{g}(\mathbf{x}^{(t)})$. It follows that in our model the conditional emission distribution for observed data is:

$$p(\mathbf{x}^{(t)}|c^{(t)};\mathbf{f},\boldsymbol{\lambda}_{c^{(t)}}) =$$
$$|\mathbf{J}\mathbf{g}(\mathbf{x}^{(t)})| \frac{H(\mathbf{g}(\mathbf{x}^{(t)}))}{Z(\boldsymbol{\lambda}_{c^{(t)}})} \exp\{\langle \boldsymbol{\lambda}_{c^{(t)}}, \mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)}))\rangle\} \quad (7)$$

where $|\mathbf{J}\mathbf{g}(\mathbf{x}^{(t)})|$ is short-hand notation for the absolute value of the determinant of the Jacobian of the inverse (demixing) function, and $H(\mathbf{g}(\mathbf{x}^{(t)})) = \prod_{i=1}^{N} h(g_i(\mathbf{x}^{(t)}))$. We have also simplified notation by stacking the vectors for different components $\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_N)^T$ and $\boldsymbol{\lambda}_{c^{(t)}} = (\boldsymbol{\lambda}_{1,c^{(t)}}, \ldots, \boldsymbol{\lambda}_{N,c^{(t)}})^T$.

We allow $\mathbf{f}$ to be any arbitrary but bijective function. In practice, it can be represented as a neural network. The model can therefore be viewed as a deep generative model for non-stationary data. Finally, using $\boldsymbol{\theta} = \{\mathbf{f}, \mathbf{L}\}$ and $\Theta = \{\boldsymbol{\theta}, \mathbf{A}\}$ our hidden Markov nonlinear ICA model's data-likelihood is given as:

$$p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)};\Theta) =$$
$$\sum_{c^{(1)},\ldots,c^{(T)}} \Big(\pi(c^{(1)}) p(\mathbf{x}^{(1)}|c^{(1)},\boldsymbol{\theta}_{c^{(1)}}) \times$$
$$\prod_{t=2}^{T} A_{c^{(t-1)},c^{(t)}} p(\mathbf{x}^{(t)}|c^{(t)};\boldsymbol{\theta}_{c^{(t)}})\Big) \quad (8)$$

where the emission distributions in Eq. (7) should be plugged in.

3.2 ESTIMATION

Assume we have a sequence of observed data $\mathcal{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(T)}\}$ generated by (8). In order to estimate the model parameters in practice we will choose the factorial prior in (6) from a well-known family such that the normalizing constant is tractable, such as a Gaussian location-scale family. Intractable normalizing constant would make estimation very difficult, even by approximate inference methods such as Variational Bayes or VAEs. However, notice that the choice of distribution for the latent prior does not severely limit the type of data that can be modelled since the non-linear mixing function can be any arbitrary function.

Tractable exponential families also make it easy to estimate the model parameters by maximizing the likelihood in (8) by the EM algorithm. The "free-energy" EM lower bound for our model is given by:

$$\mathcal{L}(q(\mathbf{c}),\Theta) :=$$
$$\mathbb{E}_{q(\mathbf{c})}\left[\log p(\mathbf{c},\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(T)};\Theta)\right] - \mathbb{E}_{q(\mathbf{c})}\left[\log q(\mathbf{c})\right] \quad (9)$$

where $\mathbf{c} = (c^{(1)}, \ldots, c^{(T)})$, such that the first RHS terms is the complete-data likelihood under some distribution $q(\mathbf{c})$. In the E-step one finds $q(\mathbf{c}_\star) := \arg\max_{q(\mathbf{c})} \mathcal{L}(q(\mathbf{c}),\Theta) = p(\mathbf{c}|\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(T)};\Theta)$ which is the standard result for HMMs and can be easily computed using the forward-backward (Baum-Welch) algorithm. In the M-step we aim to find $\Theta_\star = \arg\max_\Theta \mathcal{L}(q(\mathbf{c}_\star),\Theta)$, which reduces to maximizing:

$$\tilde{\mathcal{L}}(q(\mathbf{c}),\Theta) := \mathbb{E}_{q(\mathbf{c}_\star)}\left[\log p(\mathbf{c},\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(T)};\Theta)\right] =$$
$$\sum_{t=1}^{T} \mathbb{E}_{q(c_\star^{(t)})}\left[\log p(\mathbf{x}^{(t)}|c^{(t)};\boldsymbol{\theta}_{c^{(t)}})\right]$$
$$+ \sum_{t=2}^{T} \mathbb{E}_{q(c_\star^{(t-1)},c_\star^{(t)})}\left[\log A_{c^{(t-1)},c^{(t)}}\right] \quad (10)$$

where we have left out the initial-state probability term as we can assume a stationary process and infer them directly from $\mathbf{A}$ as its left eigenvector. The M-step updates for $\mathbf{A}$ are standard:

$$A^{\star}_{i,j} \leftarrow \frac{\sum_{t=2}^{T} q(c^{(t-1)}_{\star} = i, c^{(t)}_{\star} = j)}{\sum_{t=1}^{T} q(c^{(t)}_{\star})} \tag{11}$$

M-step updates for the parameters $\mathbf{L}$ also follow from standard EM results for exponential families:

$$\begin{aligned}
&\nabla_{\boldsymbol{\lambda}_k} \sum_{t=1}^{T} \mathbb{E}_{q(c^{(t)}_{\star})} \left[\log p(\mathbf{x}^{(t)} | c^{(t)}; \boldsymbol{\theta}_{c^{(t)}}) \right] \\
&= \nabla_{\boldsymbol{\lambda}_k} \sum_{t=1}^{T} \mathbb{E}_{q(c^{(t)}_{\star})} \left[\langle \boldsymbol{\lambda}_k, \mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)})) \rangle - \log Z(\boldsymbol{\lambda}_k) \right] \\
&= \sum_{t=1}^{T} q(c^{(t)}_{\star} = k) \left[\mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)})) - \frac{\nabla_{\boldsymbol{\lambda}_k} Z(\boldsymbol{\lambda}_k)}{Z(\boldsymbol{\lambda}_k)} \right] = 0 \\
&\Rightarrow \frac{\nabla_{\boldsymbol{\lambda}_k} Z(\boldsymbol{\lambda}_k)}{Z(\boldsymbol{\lambda}_k)} = \frac{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k) \mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)}))}{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k)}
\end{aligned} \tag{12}$$

where LHS can be rewritten as:

$$\begin{aligned}
&\frac{1}{Z(\boldsymbol{\lambda}_k)} \nabla_{\boldsymbol{\lambda}_k} \int \Big(|\mathbf{J}\mathbf{g}(\mathbf{x}^{(t)})| H(\mathbf{g}(\mathbf{x}^{(t)})) \\
&\times \exp\{ \langle \boldsymbol{\lambda}_{c^{(t)}}, \mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)})) \rangle \} \Big) \\
&= \mathbb{E}_{p(\mathbf{x}^{(t)} | c^{(t)}; \boldsymbol{\theta}_{c^{(t)}})} \left[\mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)})) \right]
\end{aligned} \tag{13}$$

Thus the M-step updates for $\boldsymbol{\lambda}^{\star}_k$ are the ones that solve:

$$\begin{aligned}
&\mathbb{E}_{p(\mathbf{x}^{(t)} | k; \boldsymbol{\lambda}^{\star}_k, \mathbf{f})} \left[\mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)})) \right] \\
&= \frac{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k) \mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)}))}{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k)}
\end{aligned} \tag{14}$$

In practice, (14) has closed-form updates for many usual exponential family members. As an example, if we were to use a Gaussian distribution, the updates for mean and variance vectors would be:

$$\begin{aligned}
\boldsymbol{\mu}^{\star}_k &\leftarrow \frac{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k) \mathbf{g}(\mathbf{x}^{(t)})}{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k)} \\
\boldsymbol{\sigma}^{2\star}_k &\leftarrow \operatorname{diag} \left(\frac{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k) \mathbf{y}^{\star}_k \mathbf{y}^{\star,T}_k}{\sum_{t=1}^{T} q(c^{(t)}_{\star} = k)} \right)
\end{aligned} \tag{15}$$

where $\mathbf{y}^{\star}_k = \mathbf{g}(\mathbf{x}^{(t)}) - \boldsymbol{\mu}^{\star}_k$

Next, the demixing function is estimated by parameterizing it as a deep neural network but for notational simplicity we will not write these parameters explicitly and instead subsume them in $\mathbf{g}$. Since an exact M-step is not possible, a gradient ascent step on the lower bound is taken instead, where the gradient is given by:

$$\begin{aligned}
&\nabla_{\mathbf{g}} \tilde{\mathcal{L}}(q(\mathbf{c}), \Theta) = \nabla_{\mathbf{g}} \sum_{t=1}^{T} \mathbb{E}_{q(c^{(t)}_{\star})} \left[\log p(\mathbf{x}^{(t)} | c^{(t)}; \boldsymbol{\theta}_{c^{(t)}}) \right] \\
&= \nabla_{\mathbf{g}} \sum_{t=1}^{T} \log |\mathbf{J}\mathbf{g}| \\
&+ \nabla_{\mathbf{g}} \sum_{t=1}^{T} \mathbb{E}_{q(c^{(t)}_{\star})} \left[\log H(\mathbf{g}) + \langle \boldsymbol{\lambda}_{c^{(t)}}, \mathbf{T}(\mathbf{g}) \rangle \right]
\end{aligned} \tag{16}$$

where we have used $\mathbf{g} = \mathbf{g}(\mathbf{x}^{(t)})$ for brevity. The parameters are then updated as:

$$\mathbf{g}^{\text{new}} \leftarrow \mathbf{g}^{\text{old}} + \eta \nabla_{\mathbf{g}} \tilde{\mathcal{L}}(q(\mathbf{c}), \Theta) \tag{17}$$

See Appendix A for discussion on the convergence of our algorithm.

The gradient term with respect to the determinant of the Jacobian $\log |\mathbf{J}\mathbf{g}|$ deserves special attention. It is widely considered difficult to compute, and therefore, normalizing flows models are often used in literature in order to make the Jacobians more tractable. The problem with this approach is that, to our best knowledge, none of such flow models has universal *function* approximation capabilities (despite some being universal distribution approximators). This would thus restrict the possible set of nonlinear mixing functions that can be estimated. Fortunately modern autograd packages such as JAX make it possible to calculate gradients of the log determinant Jacobian efficiently up to moderate dimensions (see Appendix B) – this is the approach we take. Very recent, promising, alternative for computing the log-determinant is the relative gradient (Gresele et al., 2020) which could easily be implemented in our framework. Finally, notice that the second term (16) is easy to evaluate since the expectation is just a discrete sum over the posteriors that we get from the E-step.

3.3 COMMENT ON ESTIMATION FOR LONG TIME SEQUENCES

The above estimation method may be impractical for very long time sequences since the forward-backward algorithm has computational complexity of $\mathcal{O}(TC^2)$. In such situations we can adapt the subchain sampling approach of Foti et al. (2014). This involves splitting up the full dataset into shorter time sequences and then forming mini-batches over time. The resulting gradient updates would be biased and therefore a scaling term will be applied to them. The forward-backward algorithm applied to the subchains is also only approximate due to loss of information at the ends of the chains but the authors describe a technique to buffer the chains with extra observations to reduce this effect.

3.4 COMMENT ON DIMENSION REDUCTION

An important problem in applying our method on real data is dimension reduction. While in the theory above, we assumed that the number of independent components is equal to the number of observed variables, in many practical cases, we would like to have a smaller number of components than observed variables. We propose here two solutions for this problem.

The first solution, which is widely used in the linear ICA case, is to first reduce the dimension of the data by PCA, and then do ICA in that reduced space with the same dimensions of components and observed variables. In the nonlinear case, a number of nonlinear PCA methods, also called manifold learning methods, has been proposed and could be used for such a two-stage method. In particular, dimension reduction is achieved by even the very simplest autoencoders; recent work has developed the theory further in various directions (Maaten and Hinton, 2008; Vincent et al., 2010). This approach has the advantage of reducing the noise in the data, which is a well-known property of PCA, and allows us to separate the problem of dimension reduction from the problem of developing ICA algorithms. A possible drawback is that such dimension reduction may not be optimal from the viewpoint of estimating independent components.

The second solution is to build an explicit noise model into the nonlinear ICA model, following Khemakhem et al. (2020). Denote by $\mathbf{n}$ a random vector of Gaussian noise which is white both temporally and spatially and of variance σ^2. Instead of the Eq. (2), we would define a mixing model as

$$\mathbf{x}^{(t)} = \mathbf{f}(\mathbf{s}^{(t)}) + \mathbf{n}^{(t)} \tag{18}$$

where the model of the components $\mathbf{s}^{(t)}$ is unchanged. We could then combine the variational estimation method presented by Khemakhem et al. (2020) with the HMM inference procedure presented here. However, we leave the details for future work.

4 IDENTIFIABILITY THEORY

In this section we provide identifiability theory for the model discussed in the previous section. As was discussed above, many deep latent variable models are non-identifiable. In other words, an estimation method such as the EM proposed above might not have a unique solution, or even a small number of solutions which are indistinguishable for any practical purposes.

Fortunately, we are able to combine previous nonlinear ICA theory with the identifiability of Hidden Markov Models to prove the identifiability of our combined model. Albeit our model being different from (Hyvärinen and Morioka, 2017), (Hyvärinen et al., 2019) and (Khemakhem et al., 2020), the identifiability we reach is very similar. We also show that in the case of Gaussian independent components we can get exact identifiability up to linear transformation of the components.

4.1 DEFINITIONS

In order to illustrate the relationship of our model's identifiability to earlier works in the area, we introduce the following definitions from (Khemakhem et al., 2020)

Definition 1. *Let $\sim$ be the equivalence relation on Θ. (8) is said to be identifiable up to $\sim$ if*

$$p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)}; \Theta) = p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)}; \hat{\Theta}) \Rightarrow \Theta \sim \hat{\Theta} \tag{19}$$

Definition 2. *Let $\sim$ be the binary relation on Θ defined by:*

$$(\mathbf{f}, \boldsymbol{\lambda}) \sim (\hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}}) \leftrightarrow$$
$$\exists \mathbf{W}, \mathbf{b} \mid \mathbf{T}(\mathbf{g}(\mathbf{x}^{(t)})) = \mathbf{W}\mathbf{T}(\hat{\mathbf{g}}(\mathbf{x}^{(t)})) + \mathbf{b} \tag{20}$$

where $\mathbf{W}$ is an $NV \times NV$ matrix and $\mathbf{b}$ is an $NV \times 1$ vector.

If $\mathbf{W}$ is invertible, the above relation is denote by $\sim_W$, and if $\mathbf{W}$ is a block permutation matrix, it is denoted by $\sim_\mathcal{P}$. In block permutation, each block linearly transforms $\mathbf{T}_i(g_i(x_i))$ into $\mathbf{T}_j(\hat{g}_j(x_i))$ with each j corresponding to one, and only one, i.

4.2 GENERAL RESULT

Now we present our most general Theorem on identifiability. It will be followed by stronger results in the Gaussian case below.

Theorem 2. *Assume observed data is generated by a Hidden Markov Nonlinear ICA according to (5) - (8). Also, assume:*

(i) The time-homogeneous transition matrix $\mathbf{A}$ has full rank and induces an irreducible [1] Markov chain with a unique stationary state distribution

(ii) The number of latent states, C, is known and $C \geq NV + 1$

(iii) There exists an NV square matrix of the different states' parameters with respect to a pivot state

$$\widetilde{\mathbf{L}} = \begin{pmatrix} (\boldsymbol{\lambda}_{c=1} - \boldsymbol{\lambda}_{c=0})^T \\ \vdots \\ (\boldsymbol{\lambda}_{c=NV} - \boldsymbol{\lambda}_{c=0})^T \end{pmatrix} \tag{21}$$

[1] all states can be reached from every state

which is invertible.

(iv) The emission distributions for the different latent states $p(\mathbf{x}^{(t)}|1;\boldsymbol{\theta}_1),\ldots,p(\mathbf{x}^{(t)}|C;\boldsymbol{\theta}_C)$ *are linearly independent functions of* $\mathbf{x}^{(t)}$

(v) The non-linear mixing function $\mathbf{f}$ *is bijective*

Then the model parameters $(\mathbf{f},\boldsymbol{\lambda})$ *are* $\sim_W$ *identifiable.*

Proof. Suppose we have

$$p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(T)};\Theta)=p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(T)};\hat{\Theta}) \tag{22}$$

Using assumptions (i)-(iv), we can invoke Theorem 1 and apply it to our model to get:

$$\hat{A}_{k,l}=A_{\sigma(k),\sigma(l)} \tag{23}$$

$$p(\mathbf{x}|k;\hat{\boldsymbol{\theta}}_k)=p(\mathbf{x}|\sigma(k);\boldsymbol{\theta}_{\sigma(k)}) \tag{24}$$

where superscript t is dropped for convenience. For notational simplicity, and without loss of generality, we assume the components are ordered such that $k=\sigma(k)$. Substituting in (7) we have:

$$\begin{aligned}&|\mathbf{J}\hat{\mathbf{g}}(\mathbf{x})|\frac{H(\hat{\mathbf{g}}(\mathbf{x}))}{Z(\hat{\boldsymbol{\lambda}}_k)}\exp\{\langle\hat{\boldsymbol{\lambda}}_k,\mathbf{T}(\hat{\mathbf{g}}(\mathbf{x}))\rangle\}\\&=|\mathbf{J}\mathbf{g}(\mathbf{x})|\frac{H(\mathbf{g}(\mathbf{x}))}{Z(\boldsymbol{\lambda}_k)}\exp\{\langle\boldsymbol{\lambda}_k,\mathbf{T}(\mathbf{g}(\mathbf{x}))\rangle\}\end{aligned} \tag{25}$$

for some latent state k. Recall from assumption (iii) that $C\geq NV+1$. We can thus take $C+1$ states and assign one of them, say $c=0$ as a pivot states. Taking logs of (25) for all the other states with respect to the pivot state gives C equations of below form:

$$\begin{aligned}&\langle(\boldsymbol{\lambda}_k-\boldsymbol{\lambda}_1),\mathbf{T}(g_i(\mathbf{x}))\rangle+\log Z(\boldsymbol{\lambda}_1)-\log Z(\boldsymbol{\lambda}_k)\\&=\langle(\hat{\boldsymbol{\lambda}}_k-\hat{\boldsymbol{\lambda}}_1),\mathbf{T}(\hat{g}_i(\mathbf{x}))\rangle+\log Z(\hat{\boldsymbol{\lambda}}_1)-\log Z(\hat{\boldsymbol{\lambda}}_c)\end{aligned} \tag{26}$$

Collecting all the C such equations, we can stack them into a linear system :

$$\widetilde{\mathbf{L}}\mathbf{T}(\mathbf{g}(\mathbf{x}))=\widehat{\widetilde{\mathbf{L}}}\mathbf{T}(\hat{\mathbf{g}}(\mathbf{x}))+\boldsymbol{\beta} \tag{27}$$

where $\widetilde{\mathbf{L}}$ is the invertible square matrix defined in assumption (iii), and the elements of $\widehat{\widetilde{\mathbf{L}}}$ are defined similarly, but no assumption about its invertibility is made. The constants that result from the sums of the log-normalizers are stacked to form $C\times 1$ vector $\boldsymbol{\beta}$. Multiplying both sides by $\widetilde{\mathbf{L}}^{-1}$ results in our desired form:

$$\begin{aligned}\mathbf{T}(\mathbf{g}(\mathbf{x}))&=\widetilde{\mathbf{L}}^{-1}\widehat{\widetilde{\mathbf{L}}}\mathbf{T}(\hat{\mathbf{g}}(\mathbf{x}))+\widetilde{\mathbf{L}}^{-1}\boldsymbol{\beta}\\\mathbf{T}(\mathbf{s})&=\mathbf{W}\mathbf{T}(\hat{\mathbf{g}}(\mathbf{x}))+\mathbf{b}\end{aligned} \tag{28}$$

Recall that we defined the exponential families to be in minimal representation in Section 3.1. It follows that we can find an arbitrary number of points such that the V vectors formed by the sufficient statistic functions of each independent component $(T_{i,1}(s_i),\ldots,T_{i,V}(s_i))$, are linearly independent. This can be done separately for each s_i. Additionally, as s_i and s_j can be changed independently, we can find for $i\neq j$ then $T_l(s_i)$ and $T_m(s_j)$ are linearly independent for all $l,m\in(1,\ldots,V)$. Therefore, all elements of the vector $\mathbf{T}(\mathbf{s})$ are linearly independent which implies that the square matrix $\mathbf{W}$ in (28) is invertible. □

4.2.1 Comments on the assumptions of Theorem 1

The assumptions (i), (ii) are standard HMM assumptions. The assumption of a full rank transition matrix is non-standard but crucial here. Intuitively speaking, it allows the latent states to be distinguished from each other, while the irreducibility assumptions ensures that there is a single unique stationary state distribution.[2] Notice that these assumptions necessarily hold, for example, when the transition matrix is close to identity, as in a case where the states are strongly persistent.

The assumption that the real number of latent components is known, is valid in certain applications, and if not it could be estimated for instance be increasing the number of latent states between each estimation and then detecting the point at which increases in likelihood become marginal (the elbow method). Assumption (iii) is valid in practice as long as the parameters are generated randomly - in that case it almost surely holds as singular solutions will lie in a submanifold of lower dimension. The validity of assumption is less obvious (iv), however, we will below prove that it holds, for instance, in the case of Gaussian independent components.

4.3 IDENTIFIABILITY WITH GAUSSIAN INDEPENDENT COMPONENTS

In this section, we first provide two lemmas which we use to prove the claim, already alluded to above, that assumption (iv) of Theorem 2 is satisfied for Gaussian components. Then, we prove that in this case a stronger form of identifiability can be reached as a special case of above results, namely that we get exact identification of components up to linear transformation. Together these results make a strong case for using Gaussian latent components in practical applications.

We begin by stating two Lemmas (proofs in Appendix C):

[2] technically one aperiodic state is also required. An aperiodic state is one which can be returned to after an irregular number of steps

Lemma 1. *Assumption (iv) of Theorem 1 requires the C emission distributions defined by (7) to be linearly independent. A sufficient, and necessary, condition is that the C conditional source distributions defined by (6) are linearly independent.*

Lemma 2. *Assume K probability density functions of N random variables $p_1(z_1, \ldots, z_N), \ldots, p_K(z_1, \ldots, z_N)$, and that each factorizes across the variables: $p_k(z_1, \ldots, z_N) = \prod_{i=1}^{N} p_k^{(i)}(z_i)\ \forall k \in \{1, \ldots, K\}$. If the K factorial density functions $p_1^{(i)}(z_i), \ldots, p_K^{(i)}(z_i)$ are linearly independent for some $i \in \{1, \ldots, N\}$, then the K joint-density functions $p_1(z_1, \ldots, z_N), \ldots, p_K(z_1, \ldots, z_N)$ are linearly independent.*

Based on these Lemmas, we can prove the following Theorem (proof in Appendix C):

Theorem 3. *Assume that distributions of the independent components conditional on the latent state, as defined by (6), are Gaussian parameterised by mean and variance. Assume also that the means of the C density functions are all different. Then the emission distributions, defined by (7), are linearly independent, thus satisfying assumption (iv) in Theorem 2.*

Finally, we have the following Theorem which proves a stronger form of identifiability—essentially recovering the components with minimum indeterminacy—of our Hidden Markov Nonlinear ICA model in the Gaussian case (proof in AppendixC)

Theorem 4. *Assume that the latent independent components have a conditionally Gaussian distributions, and assume hypotheses (i), (ii),(iii) and (v) of Theorem 2 hold, as well as the assumptions of Theorem 3. Additionally assume that the mixing function $\mathbf{f}$ has all of it second-order cross derivatives, then the components in our Hidden Markov Nonlinear ICA model are exactly identified up to linear transformation.*

Notice that this proof and the identifiability result is similar to that in (Sorrenson et al., 2020), although our models are entirely different. These authors also prove a general version for other distributions with different sufficient statistics.

5 EXPERIMENTS

In this section we present results from our simulations on artificial non-stationary data. Code, written in JAX, is available at `github.com/HHalva/hmnlica`.

Dataset: We generated a synthetic dataset from model defined in Section 3.1. More specifically, the independent components are created from non-stationary Gaussian distributions, where the non-stationarity comes from an unobserved latent variable that can take C discrete states and follows first order Markov dynamics. The latent state then determines the means and the variances of the independent components at each time point. The transition matrix was defined such that at each time-step there was a 99% probability that the state didn't change a 1% probability the latent state switches to another state.[3] If it switches to another state, it will always go to the next one 'in line', where we define a circular ordering for the states. That is, we essentially defined a circular repeating path for the latent state where transitions could only happen to two states (the same state or another), and where the transition matrix is close to identity (Figure 1 illustrates this). These settings were chosen to ensure the HMM assumptions of Theorem 2 hold, as well as to reflect a situation where a relatively small number of states repeats over time with some interesting, non-random, temporal dynamics, including persistence to stay in the same state. The mean and variance parameters were chosen at random for each latent state before data generation so that assumption (iii) of Theorem 2 holds. Similarly to Hyvärinen et al. (2019), the mixing function (2) was simulated with a randomly initialized, invertible[4] multi-layer perceptron (MLP) – this produced the observed data for our experiments. The sequences that were created are 100,000 time steps long, and were generated for various number of independent components. The number of latent states was set such that $C = 2N + 1$, which ensures that the assumptions (ii) and (iii) were fulfilled.

Model estimation: We estimate the (inverse) mixing function and distribution parameters of our Hidden Markov nonlinear-ICA model using the EM algorithm described in Section 3.2. Mean and variance parameter estimates for the independent components are initialized randomly at the start of the EM algorithm. The inverse mixing function is parameterized with an MLP where the number hidden layers is set to match the number of data generating mixing layers. The gradient M-step are taken with the Adam optimizer (Kingma and Ba, 2017). As typical for the EM algorithm, random restarts were used to avoid inferior local optima and saddle points. Further, we found that a stochastic version of our algorithm (see Section 3.3) converged faster – thus, the experiments here have been run with 100 time-step long sub-sequences in minibatches of 64.

Results – independent component recovery: After estimating the model parameters and independent components, a linear sum assignment problem is solved to op-

[3] for context, the probability of staying in the same state for over 100 time steps with these numbers is around 37%

[4] invertibility achieved by having all layers N units wide and utilizing leaky ReLUs

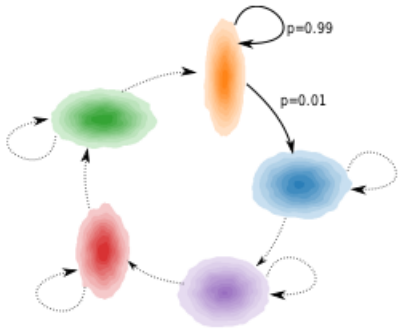

Figure 1: An example of the independent components' distributions from a HMM where the number of components $N = 2$ and the number of latent states $C = 5$. The clusters are ordered to illustrated the dynamics of the hidden Markov model, in particular its circular property. The transition probabilities are the same throughout the data.

timally match each of the estimated components to one of the real ones. This is necessary as the ordering of the components is arbitrary. Mean absolute correlation coefficients over the resulting pairs of true and estimated components are then used to measure how well original components were recovered. This is the methodology taken in previous nonlinear ICA works (Hyvärinen and Morioka, 2016).

Figure 2 shows the mean correlation between the estimated components for our model in comparison to TCL, which is the only other nonlinear ICA model for nonstationary data. For TCL, the data was split into 500 time-step long segments; 500 steps provided best performance relative to other computationally feasible options (100, 250, 750, 1000). We can see that our model outperforms TCL for all levels of nonlinearity. This validates our theoretical arguments that the TCL framework struggles with non-stationary data in which latent states (often a relatively small number) repeat over time since the segments it has access to don't correspond well with the true data generating process.

Results – temporal dynamics Unlike previous models, Hidden Markov Nonlinear ICA is further able to perform unsupervised clustering of latent states and to take into account the learned temporal dynamics in doing so. To estimate this ability, we run the well-know Viterbi algorithm (Viterbi, 1967) which finds the optimal (most likely) path of latent states through time based on our estimated model. The results show that on average for $N = 5$ and $C = 11$ our model's results range in general from 30% to 60% accuracy with a mean of approximately 45%, and therefore its performance is clearly above chance level(9%); for the figure see Appendix.

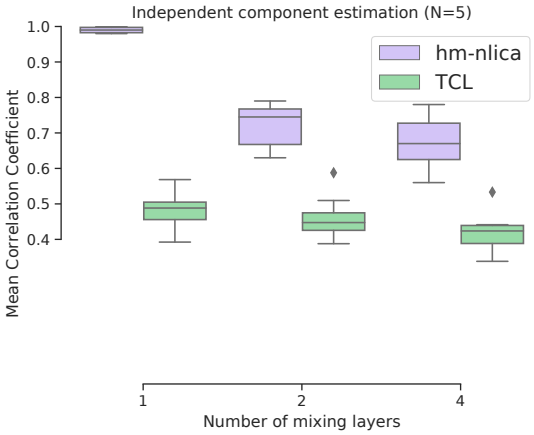

Figure 2: Performance of our Hidden Markov nonlinear ICA vs. TCL in recovering the true sources for N=5 for our synthetic dataset. The amount of nonlinearity is controlled by number of hidden layers in the mixing MLP, so that $L \in [1, 2, 4]$.

6 CONCLUSION

We proposed a framework nonlinear ICA based on a Hidden Markov Model of the temporal dynamics. This improves on existing nonlinear ICA methods in several ways. First, it removes the need for any arbitrary segmentation of the data as in TCL, which is likely to improve the estimation of the demixing function. Second, it leverages the fact that the nonstationary structure is often repeating with a limited number of hidden states, which not only reduces the computation by limiting the number classes, but again is likely to improve estimation of the demixing function. Third, our method estimates the underlying dynamics, which are often interesting in their own right. We believe this in an important advance in order to apply nonlinear ICA methods on real data.

Acknowledgements

The authors would like to thank Elisabeth Gassiat, Ilyes Khemakhem and Ricardo Pio Monti for helpful discussion. I.K.'s help with the experiments is also much appreciated. A.H. was supported by a Fellowship from CIFAR, and from the DATAIA convergence institute as part of the "Programme d'Investissement d'Avenir" (ANR-17-CONV-0003) operated by Inria.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132. arXiv: 0809.5032.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: Non-linear Independent Components Estimation. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Foti, N., Xu, J., Laird, D., and Fox, E. (2014). Stochastic variational inference for hidden Markov models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3599–3607. Curran Associates, Inc.

Gassiat, E., Cleynen, A., and Robin, S. (2016). Inference in finite state space non parametric Hidden Markov Models and applications. *Statistics and Computing*, 26(1-2):61–71.

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.

Gresele, L., Fissore, G., Javaloy, A., Schölkopf, B., and Hyvärinen, A. (2020). Relative gradient optimization of the jacobian term in unsupervised deep learning. *Submitted*.

Harmeling, S., Ziehe, A., Kawanabe, M., and Müller, K.-R. (2003). Kernel-Based Nonlinear Blind Source Separation. *Neural Computation*, 15(5):1089–1124.

Hyvärinen, A. and Morioka, H. (2016). Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *arXiv:1605.06336 [cs, stat]*. arXiv: 1605.06336.

Hyvärinen, A. and Morioka, H. (2017). Nonlinear ICA of Temporally Dependent Stationary Sources. *Proceedings of the 20 th International Con- ference on Artificial Intelligence and Statistics (AISTATS)*, page 14.

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

Hyvärinen, A., Sasaki, H., and Turner, R. E. (2019). Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. *arXiv:1805.08651 [cs, stat]*. arXiv: 1805.08651.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. *arXiv:1907.04809 [cs, stat]*. arXiv: 1907.04809.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*. arXiv: 1312.6114.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Penny, W., R.M., E., and S.J., R. (2000). hidden markov independent component analysis. In *Advances in Independent Component Analysis*.

Sorrenson, P., Rother, C., and Köthe, U. (2020). Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN). *arXiv:2001.04872 [cs, stat]*. arXiv: 2001.04872.

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.

Tong, L., Liu, R.-w., Soon, V., and Huang, Y.-F. (1991). Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38(5):499–509.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-a. (2010). stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *journal of machine learning research*, 11(dec):3371–3408.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–267.

Zhou, J. and Zhang, X. (2008). An ica mixture hidden markov model for video content analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1576–1586.

Appendix for

Hidden Markov Nonlinear ICA for Unsupervised Learning from Nonstationary Time Series (published at UAI 2020)

A Note on the convergence of our estimation algorithm

Standard theory (Dempster et al., 1977) shows that each EM iteration increases the likelihood, unless parameters are already at a zero-gradient point. Further, maxima of free-energy and likelihood coincide. This also holds under the gradient M-steps in our algorithm (with classical assumption of sufficiently small step size). Under suitable regularity conditions, theoretical limit of infinite data and universal approximation of the nonlinear transformation, combined with our identifiability proof, MLE guarantees convergence to correct parameters up to the equivalence class identified in our Theorem 2. In practice, however, these assumptions may not be satisfied — for instance, parameters may approach a boundary point and likelihood tend to infinity. Random restarts and regularisation are common strategies to avoid these problems.

B Note on the compute time of the gradients of the logdet Jacobian

We estimate the non-linear mixing function in our model using a multi-layer perceptron without any restrictions on it. As a consequence of the change of variable formula for probability densities, we have to calculate the gradient of the log-determinant of the Jacobian as part of our parameter updates. JAX, a new machine learning package that utilizes autograd, has the ability to calculate the Jacobian in just a single forward pass thus making the computations efficient for typical data dimensions. For our model, we can see that the compute time required for the log-determinant of the Jacobian starts to dominate as we approach 100 dimensions and above.

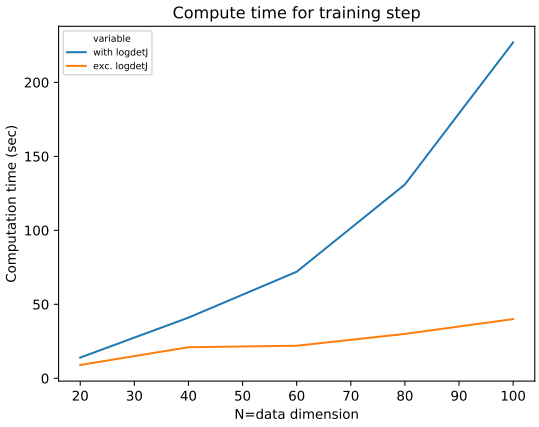

Figure 3: Average computation time over 100 epochs for computing the gradients of the function estimator in our model, including and excluding the log-determinant Jacobian term, in JAX. The function estimator is a four layer deep neural network where the width of the hidden units is always equal to N.

C Proofs

Proof for Lemma 1

Proof. Assume that we have linear independence of the C conditional source distributions as defined by (6). Then we have that:

$$a_1 p_S(\mathbf{s}|1;\boldsymbol{\lambda}_1) + \cdots + a_C p_S(\mathbf{s}|C;\boldsymbol{\lambda}_C) \equiv 0 \Rightarrow \mathbf{a} = \mathbf{0} \tag{29}$$

Above holds if we multiply it through by the Jacobian determinant of the mixing function as we have assumed bijectivity, that is:

$$\begin{aligned} &a_1 |\mathbf{J}\mathbf{g}(\mathbf{x})| p_S(\mathbf{s}|1;\boldsymbol{\lambda}_1) + \dots \\ &+ a_C |\mathbf{J}\mathbf{g}(\mathbf{x})| p_S(\mathbf{s}|C;\boldsymbol{\lambda}_C) \equiv 0 \Rightarrow \mathbf{a} = \mathbf{0} \end{aligned} \tag{30}$$

which is equivalent to:

$$\begin{aligned} &a_1 |\mathbf{J}\mathbf{g}(\mathbf{x})| p_S(\mathbf{g}(\mathbf{x})|1;\boldsymbol{\lambda}_1) + \dots \\ &+ a_C |\mathbf{J}\mathbf{g}(\mathbf{x})| p_S(\mathbf{g}(\mathbf{x})|C;\boldsymbol{\lambda}_C) \equiv 0 \Rightarrow \mathbf{a} = \mathbf{0} \end{aligned} \tag{31}$$

And therefore by (7) we have:

$$\begin{aligned} &a_1 p_X(\mathbf{x}|1;\mathbf{f},\boldsymbol{\lambda}_1) + \dots \\ &+ a_C p_X(\mathbf{x}|C;\mathbf{f},\boldsymbol{\lambda}_C) \equiv 0 \Rightarrow \mathbf{a} = \mathbf{0} \end{aligned} \tag{32}$$

So the emission distributions are linearly independent if the densities for the independence components are linearly independent across the C different latent states. Necessity follows easily by the reverse of above argumentation. □

Proof for Lemma 2

Proof. Assume linear independence of the K joint-density functions for some subset of variables $z_i, \dots, z_{i+n}$, where $i \in \{1, \dots, N\}$ and $0 \leq n \leq N - i - 1$, that is:

$$\begin{aligned} &w_1 p_1(z_i, \dots, z_{i+n}) + \cdots + w_K p_K(z_i, \dots, z_{i+n}) \\ &= w_1 \prod_{j=i}^{i+n} p_1^{(j)}(z_j) + \cdots + w_K \prod_{j=i}^{i+n} p_K^{(j)}(z_j) \equiv 0 \\ &\Rightarrow \mathbf{w} = \mathbf{0} \end{aligned} \tag{33}$$

Now the linear independence for joint of $p_k(z_i, \dots, z_{i+n}, z_{i+n+1})$ requires:

$$\begin{aligned} &w_1 p_1(z_i, \dots, z_{i+n}, z_{i+n+1}) + \\ &\cdots + w_K p_K(z_i, \dots, z_{i+n}, z_{i+n+1}) \equiv 0 \Rightarrow \mathbf{w} = \mathbf{0} \end{aligned}$$

Using the factorial form of the joint, we can rewrite this as:

$$\begin{aligned} &w_1 p_1^{(i+n+1)}(z_{i+n+1}) p_1(z_i, \dots, z_{i+n}) + \dots \\ &+ w_K p_K^{(i+n+1)}(z_{i+n+1}) p_K(z_i, \dots, z_{i+n}) \\ &\equiv 0 \Rightarrow \mathbf{w} = \mathbf{0} \end{aligned} \tag{34}$$

If this didn't hold we could define K constants $v_k := w_k p_k^{(i+n+1)}(z_{i+n+1})$ such that:

$$v_1 p_1(z_i, \dots, z_{i+n}) + \cdots + v_K p_K(z_i, \dots, z_{i+n}) \equiv 0 \tag{35}$$

where the constants are not all zero which would contradict our original assumption. Thus it is sufficient to prove linear independence of $p_1^{(i)}(z_i), \dots, p_K^{(i)}(z_i)$, say for $i = 1$, without loss of generality, and then apply the above induction step to guarantee linear independence of the K joint-density functions $p_1(z_1, \dots, z_N), \dots, p_K(z_1, \dots, z_N)$. □

Proof for Theorem 3

Proof. By Lemma 1 it is sufficient to prove the linear independence of the C different conditional independent component density functions, rather than emission densities. And by Lemma 2 it suffices to prove this only for any one of the N different independent components. In exponential family form, the density is written as (see Appendix D):

$$p(s_i|c) = \frac{1}{\sqrt{2\pi}} \frac{\exp\{\eta_{i,c,1} s_i - \eta_{i,c,2} s_i^2\}}{Z_{i,c}} \tag{36}$$

where $\eta_{i,c,2} > 0 \, \forall c \in \{1, \ldots, C\}$. We drop subscript i for convenience. Consider:

$$\begin{aligned} & w_1 \frac{1}{\sqrt{2\pi}} \frac{\exp\{\eta_{1,1} s - \eta_{1,2} s^2\}}{Z_1} + \ldots \\ & + w_C \frac{1}{\sqrt{2\pi}} \frac{\exp\{\eta_{C,1} s - \eta_{C,2} s^2\}}{Z_C} = 0 \end{aligned} \tag{37}$$

First assume, all the $\boldsymbol{\eta}_c$ are distinct. Also we can assume, without loss of generality, that the C latent states are ordered such that $\eta_{1,2} < \eta_{2,2} < \cdots < \eta_{C,2}$. We can divide all the terms with the first density to give:

$$\begin{aligned} & w_1 + w_2 \frac{Z_1}{Z_2} \exp\{(\eta_{1,1} - \eta_{2,1}) s + (\eta_{1,2} - \eta_{2,2}) s^2\} + \ldots \\ & + w_C \frac{Z_1}{Z_C} \exp\{(\eta_{1,1} - \eta_{C,1}) s + (\eta_{1,2} - \eta_{C,2}) s^2\} = 0 \end{aligned} \tag{38}$$

taking $\lim_{s \to +\infty}$ of above gives $w_1 = 0$. Repeatedly performing this process for remaining terms eventually gives $\mathbf{w} = \mathbf{0}$. Consider now the opposite case in which all the $\boldsymbol{\eta}_c$ are equal. Then we have:

$$w_1 \frac{1}{\sqrt{2\pi}} \frac{\exp\{\eta_{1,1}\}}{Z_1} + \cdots + w_C \frac{1}{\sqrt{2\pi}} \frac{\exp\{\eta_{C,1}\}}{Z_C} = 0 \tag{39}$$

If we re-order the terms such that $\eta_{1,1}$ is the largest (recall we assumed that the means are different). We can again divide everything by this term, take $\lim_{s \to +\infty}$, and establish $w_1 = 0$ and repeat the process to get $\mathbf{w} = \mathbf{0}$. In the the final case where more than one component has the highest variance, but rest are unequal, (only the equality of the largest variances of the remaining terms matters), we can first perform the variance division followed by division by the largest mean, repeatedly until $\mathbf{w} = \mathbf{0}$. □

Proof for Theorem 4

Proof. By Theorem 3, the above assumptions suffice for Theorem 2 to hold. Next, note that the sufficient statistics of a Gaussian distribution are twice differentiable. This, combined with the assumption about the existence of $\mathbf{f}$'s cross-derivatives fulfils the conditions of Theorem 2 of Khemakhem et al. (2020) and thus our model's parameters are $\sim_{\mathcal{P}}$ identifiable (as per Definition 2). We therefore have:

$$\begin{pmatrix} s_i \\ s_i^2 \end{pmatrix} = \mathbf{W}_j \begin{pmatrix} g_j(\mathbf{x}) \\ g_j(\mathbf{x})^2 \end{pmatrix} + \mathbf{b_i} \tag{40}$$

for some i, j. Hence, we have

$$\begin{aligned} &(w_{11} g_j(\mathbf{x}) + w_{12} g_j(\mathbf{x})^2 + b_{11})^2 \\ &= w_{21} g_j(\mathbf{x}) + w_{22} g_j(\mathbf{x})^2 + b_{21} \\ &w_{12}^2 z^4 + 2 w_{11} w_{12} z^3 + (w_{11}^2 - w_{22}) z^2 - w_{21} z + b = 0 \end{aligned} \tag{41}$$

Above has to hold for all values of $z = g_j(\mathbf{x})$. The trivial solution of all $w_{ij} = 0$ is impossible as $\mathbf{W}$ would not be invertible. Therefore, it must be that $w_{12} = w_{21} = 0$ and $w_{11}^2 = w_{22}$. Thus we have exact identification (up to linear transformation) $s_i = w_{ij} g_j(\mathbf{x}) + b_i$ for some constants w_{ij}, b_i. □

D Model with Gaussian independent components

$$p(s_i|c) = \frac{1}{\sqrt{2\pi\sigma_{i,c}^2}} \exp\{-\frac{1}{2\sigma_{i,c}^2}(s_i - \mu_{i,c})^2\} \tag{42}$$

$$= \frac{1}{\sqrt{2\pi\sigma_{i,c}^2}} \exp\{-\frac{1}{2\sigma_{i,c}^2}(s_i^2 - 2 s_i \mu_{i,c} + \mu_{i,c}^2)\} \tag{43}$$

$$= \frac{1}{\sqrt{2\pi\sigma_{i,c}^2}} \exp\{s_i \frac{\mu_{i,c}}{\sigma_{i,c}^2} - s_i^2 \frac{1}{2\sigma_{i,c}^2} - \frac{\mu_{i,c}^2}{2\sigma_{i,c}^2}\} \tag{44}$$

$$= Z_{i,c}^{-1} \exp\{s_i \frac{\mu_{i,c}}{\sigma_{i,c}^2} - s_i^2 \frac{1}{2\sigma_{i,c}^2}\} \tag{45}$$

Therefore by independence of components:

$$p(\mathbf{s}|c) = \exp\{\sum_{i=1}^{N} (s_i \frac{\mu_{i,c}}{\sigma_{i,c}^2} - s_i^2 \frac{1}{2\sigma_{i,c}^2})\} \prod_{i=1}^{N} Z_{i,c}^{-1} \tag{46}$$

$$= \exp\{\sum_{i=1}^{N} (s_i \frac{\mu_{i,c}}{\sigma_{i,c}^2} - s_i^2 \frac{1}{2\sigma_{i,c}^2})\} Z_c^{-1} \tag{47}$$

And change of variable gives:

$$p(\mathbf{x}|c) = |\mathbf{J}\mathbf{g}(\mathbf{x})| \exp\{\sum_{i=1}^{N} (g_i(\mathbf{x}) \frac{\mu_{i,c}}{\sigma_{i,c}^2} - g_i(\mathbf{x})^2 \frac{1}{2\sigma_{i,c}^2})\} Z_c^{-1} \tag{48}$$

$$= |\mathbf{J}\mathbf{g}(\mathbf{x})| \exp\{\langle \boldsymbol{\lambda}_c, \mathbf{T}(\mathbf{g}(x)) \rangle\} Z_c^{-1} \tag{49}$$

where $\boldsymbol{\lambda}_c = \begin{bmatrix} \frac{\mu_{1,c}}{\sigma_{1,c}^2} \\ -\frac{1}{2\sigma_{1,c}^2} \\ \vdots \\ \frac{\mu_{N,c}}{\sigma_{N,c}^2} \\ -\frac{1}{2\sigma_{N,c}^2} \end{bmatrix}$ and $\mathbf{T}(\mathbf{g}(\mathbf{x})) = \begin{bmatrix} g_1(\mathbf{x}) \\ g_1^2(\mathbf{x}) \\ \vdots \\ g_N(\mathbf{x}) \\ g_N^2(\mathbf{x}) \end{bmatrix}$.

E Latent state prediction

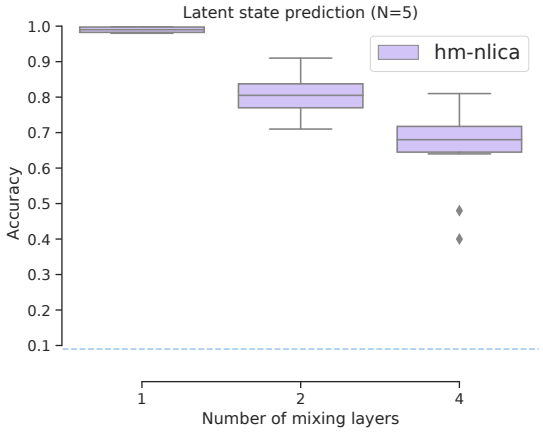

Figure 4: Performance of our Hidden Markov nonlinear ICA vs. chance level (dotted line = 0.09) for different levels of nonlinearity in latent state prediction. The number of latent states is $11 = 2N + 1$