Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA

Hermanni Hälvä[1] * **Sylvain Le Corff**[2] **Luc Lehéricy**[3]

Jonathan So[4] **Yongjie Zhu**[1] **Elisabeth Gassiat**[5] † **Aapo Hyvärinen**[1] †

[1]Department of Computer Science, University of Helsinki, Finland
[2] Samovar, Télécom SudParis, département CITI, Institut Polytechnique de Paris, Palaiseau, France
[3]Laboratoire J. A. Dieudonné, Université Côte d'Azur, CNRS, 06100, Nice, France
[4]Department of Engineering, University of Cambridge, UK
[5]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Abstract

We introduce a new general identifiable framework for principled disentanglement referred to as Structured Nonlinear Independent Component Analysis (SNICA). Our contribution is to extend the identifiability theory of deep generative models for a very broad class of structured models. While previous works have shown identifiability for specific classes of time-series models, our theorems extend this to more general temporal structures as well as to models with more complex structures such as spatial dependencies. In particular, we establish the major result that identifiability for this framework holds even in the presence of noise of unknown distribution. Finally, as an example of our framework's flexibility, we introduce the first nonlinear ICA model for time-series that combines the following very useful properties: it accounts for both nonstationarity and autocorrelation in a fully unsupervised setting; performs dimensionality reduction; models hidden states; and enables principled estimation and inference by variational maximum-likelihood.

1 Introduction

A central tenet of unsupervised deep learning is that noisy and high dimensional real world data is generated by a nonlinear transformation of lower dimensional latent factors. Learning such lower dimensional features is valuable as they may allow us to understand complex scientific observations in terms of much simpler, semantically meaningful, representations (Morioka et al., 2020; Zhou and Wei, 2020). Access to a ground truth generative model and its latent features would also greatly enhance several other downstream tasks such as classification (Klindt et al., 2021; Banville et al., 2021), transfer learning (Khemakhem et al., 2020b), as well as causal inference (Monti et al., 2019; Wu and Fukumizu, 2020).

A recently popular approach to deep representation learning has been to learn *disentangled* features. Whilst not rigorously defined, the general methodology has been to use deep generative models such as VAEs (Kingma and Welling, 2014; Higgins et al., 2017) to estimate semantically distinct factors of variation that generate and encode the data. A substantial problem with the vast majority of work on disentanglement learning is that the models used are not *identifiable* – that is, they do not learn the true generative features, even in the limit of infinite data – in fact, this task has been proven

*hermanni.halva@helsinki.fi
†Equal senior authorship

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

impossible without inductive biases on the generative model (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). Lack of identifiability plagues deep learning models broadly and has been implicated as one of the reasons for unexpectedly poor behaviour when these models are deployed in real world applications (D'Amour et al., 2020). Fortunately, in many applications the data have dependency structures, such as temporal dependencies which introduce inductive biases. Recent advances in both identifiability theory and practical algorithms for nonlinear ICA (Hyvärinen and Morioka, 2016, 2017; Hälvä and Hyvärinen, 2020; Morioka et al., 2021; Klindt et al., 2021; Oberhauser and Schell, 2021) exploit this and offer a principled approach to disentanglement for such data. Learning statistically independent nonlinear features in such models is well-defined, i.e. those models are identifiable.

However, the existing nonlinear ICA models suffer from numerous limitations. First, they only exploit specific types of temporal structures, such as either temporal dependencies or nonstationarity. Second, they often work under the assumption that some 'auxiliary' data about a *latent* process is observed, such as knowledge of the switching points of a nonstationary process as in Hyvärinen and Morioka (2016); Khemakhem et al. (2020a) . Furthermore, all the nonlinear ICA models cited above, with the exception of Khemakhem et al. (2020a), assume that the data are fully observed and noise-free, even though observation noise is very common in practice, and even Khemakhem et al. (2020a) assumes the noise distribution to be exactly known. This approach of modelling observation noise explicitly is in stark contrast to the approach taken in papers, such as Locatello et al. (2020), who instead consider general stochasticity of their model to be captured by latent variables – this approach would be ill-suited to the type of denoising one would often need in practice. Lastly, the identifiability theorems in previous nonlinear ICA works usually restrict the latent components to a specific class of models such as exponential families (but see Hyvärinen and Morioka (2017)).

In this paper we introduce a new framework for identifiable disentanglement, Structured Nonlinear ICA (SNICA), which removes each of the aforementioned limitations in a single unifying framework. Furthermore, the framework guarantees identifiability of a rich class of nonlinear ICA models that is able to exploit dependency structures of any arbitrary order and thus, for instance, extends to spatially structured data. This is the first major theoretical contribution of our paper.

The second important theoretical contribution of our paper proves that models within the SNICA framework are identifiable even in the presence of additive output noise of *arbitrary, unknown* distribution. We achieve this by extending the theorems by Gassiat et al. (2020b,a). The subsequent practical implication is that SNICA models can perform dimensionality reduction to identifiable latent components and de-noise observed data. We note that noisy-observation part of the identifiability theory is not even limited to nonlinear ICA but applies to any system observed under noise.

Third, we give mild sufficient conditions, relating to the strength and the non-Gaussian nature of the temporal or spatial dependencies, enabling identifiability of nonlinear independent components in this general framework. An important implication is that our theorems can be used, for example, to develop models for disentangling identifiable features from spatial or spatio-temporal data.

As an example of the flexibility of the SNICA framework, we present a new nonlinear ICA model called Δ-SNICA . It achieves the following very practical properties which have previously been unattainable in the context of nonlinear ICA: the ability to account for both nonstationarity and autocorrelation in a fully unsupervised setting; ability perform dimensionality reduction; model latent states; and to enable principled estimation and inference by variational maximum-likelihood methods. We demonstrate the practical utility of the model in an application to noisy neuroimaging data that is hypothesized to contain meaningful lower dimensional latent components and complex temporal dynamics.

2 Background

We start by giving some brief background on Nonlinear ICA and identifiability. Consider a model where the distribution of observed data $\mathbf{x}$ is given by $p_X(\mathbf{x};\boldsymbol{\theta})$ for some parameter vector $\boldsymbol{\theta}$. This model is called identifiable if the following condition is fulfilled:

$$\forall(\boldsymbol{\theta},\boldsymbol{\theta}') \qquad p_X(\mathbf{x};\boldsymbol{\theta}) = p_X(\mathbf{x};\boldsymbol{\theta}') \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}' \,. \tag{1}$$

In other words, based on the observed data distribution alone, we can *uniquely* infer the parameters that generated the data. For models parameterized with some nonparametric function estimator $\mathbf{f}$, such as a deep neural network, we can replace $\boldsymbol{\theta}$ with $\mathbf{f}$ in the equation above. In practice, identifiability

might hold for some parameters, not all; and parameters might be identifiable up to some more or less trivial indeterminacies, such as scaling.

In a typical nonlinear ICA setting we observe some $\mathbf{x} \in \mathbb{R}^N$ which has been generated by an invertible nonlinear mixing function $\mathbf{f}$ from latent independent components $\mathbf{s} \in \mathbb{R}^N$, with $p(\mathbf{s}) = \prod_{i=1}^N p(s^{(i)})$, as per:

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

Identifiability of $\mathbf{f}$ would then mean that we can in theory find the true $\mathbf{f}$, and subsequently the true data generating components. Unfortunately, without some additional structure this model is unidentifiable, as shown by Hyvärinen and Pajunen (1999): there is an infinite number of possible solutions and these have no trivial relation with each other. To solve this problem, previous work (Sprekeler et al., 2014; Hyvärinen and Morioka, 2016, 2017) developed models with temporal structure. Such time series models were generalized and expressed in a succinct way by Hyvärinen et al. (2019); Khemakhem et al. (2020a) by assuming the independent components are *conditionally* independent upon some observed auxiliary variable u_t: $p(\mathbf{s}_t|u_t) = \prod_{i=1}^N p(s_t^{(i)}|u_t)$. In a time series context, the auxiliary variable might be history, e.g. $u_t = \mathbf{x}_{t-1}$, or the index of a time segment to model nonstationarity (or piece-wise stationarity). (It could also be data from another modality, such as audio data used to condition video data (Arandjelovic and Zisserman, 2017).)

Notice that the mixing function $\mathbf{f}$ in (2) is assumed bijective and thus *identifiable* dimension reduction is not possible in most of the models discussed above. The only exceptions, we are aware of, are Khemakhem et al. (2020a); Klindt et al. (2021) who choose $\mathbf{f}$ as injective rather than bijective. Further, Khemakhem et al. (2020a) assume additive noise on the observations $\mathbf{x} = \mathbf{f}(\mathbf{s}) + \boldsymbol{\varepsilon}$, which allows to estimate posterior of $\mathbf{s}$ by an identifiable VAE (iVAE). We will take a similar strategy in what follows.

3 Definition of Structured Nonlinear ICA

In this section, we first present the new framework of Structured Nonlinear ICA (SNICA) – a broad class of models for identifiable disentanglement and learning of independent components when data has structural dependencies. Next, we give an example of a particularly useful specific model that fits within our framework, called Δ-SNICA , by using switching linear dynamical latent processes.

3.1 Structured Nonlinear ICA framework

Consider observations $(\mathbf{x}_t)_{t\in\mathbb{T}} = ((x_t^{(1)}, \ldots, x_t^{(M)}))_{t\in\mathbb{T}}$ where $\mathbb{T}$ is a discrete indexing set of arbitrary dimension. For discrete time-series models, like previous works, $\mathbb{T}$ would be a subset of $\mathbb{N}$. Crucially, however, we allow it to be any arbitrary indexing variable that describes a desired structure. For instance, $\mathbb{T}$ could be a subset of $\mathbb{N}^2$ for spatial data.

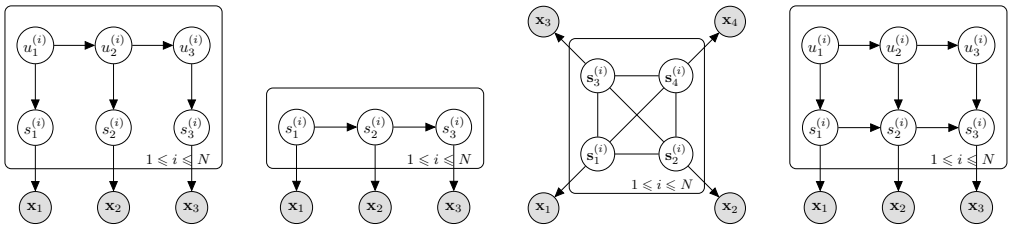
We assume the data is generated according the following nonlinear ICA model. First, there exist latent components $\mathbf{s}^{(i)} = (s_t^{(i)})_{t\in\mathbb{T}}$ for $i \in \{1, \ldots, N\}$ where for any $t, t' \in \mathbb{T}$, the distributions of $(\mathbf{s}_t^{(i)})_{1\leqslant i\leqslant N}$ and $(\mathbf{s}_{t'}^{(i)})_{1\leqslant i\leqslant N}$ are the same, which is a weak form of *stationarity*. Second, we assume that for any $m \in \mathbb{N}^*$ and $(t_1, \ldots, t_m) \in \mathbb{T}^m$, $p(\mathbf{s}_{t_1}, \ldots, \mathbf{s}_{t_m}) = \prod_{i=1}^N p(s_{t_1}^{(i)}, \ldots, s_{t_m}^{(i)})$: that is, the components are unconditionally *independent*. We further assume that the nonlinear mixing function $\mathbf{f} : \mathbb{R}^N \to \mathbb{R}^M$ with $M \geqslant N$ is injective, so there may be more observed variables than components. Finally, denote observational noise by $\boldsymbol{\varepsilon}_t \in \mathbb{R}^M$ and assume that they are i.i.d. for all $t \in \mathbb{T}$ and independent of the signals $\mathbf{s}^{(i)}$. Putting these together, we assume the mixing model where for each $t \in \mathbb{T}$,

$$\mathbf{x}_t = \mathbf{f}(\mathbf{s}_t) + \boldsymbol{\varepsilon}_t \,, \tag{3}$$

where $\mathbf{s}_t = (s_t^{(1)}, \ldots, s_t^{(N)})$. Importantly, $\boldsymbol{\varepsilon}_t$ can have any arbitrary unknown distribution, even with dependent entries; in fact, it may even not have finite moments.

The main appeal of this framework is that, under the conditions given in next section, we can now guarantee identifiability for a very broad and rich class of models.

First, notice that all previous Nonlinear ICA time-series models can be reformulated and often improved upon when viewed through this new unifying framework. In other words, we can create



(a) HMM modulated components c.f. (Hälvä and Hyvärinen, 2020)

(b) Temporal dependencies c.f. (Hyvärinen and Morioka, 2017)

(c) New: Spatial process on a graph (with latent states u_t integrated out)

(d) New: Δ-SNICA , a linear switching dynamics model for components

Figure 1: Graphical models for the SNICA framework

models that are very much like those previous works, and capture their dependency profiles, but with the changes that by assuming unconditional independence and output noise we now allow them to perform dimension reduction (this does also require some additional assumptions needed in our identifiability theorems below). To see this, consider the model in Hälvä and Hyvärinen (2020) which captures nonstationarity in the independent components through a global hidden Markov chain. We can transform this model into the SNICA framework if we instead model *each* independent component as its own HMM (Figure 1a), with the added benefit that we now have marginally independent components and are able to perform dimensionality reduction into low dimensional latent components. Nonlinear ICA with time-dependencies, such as in an autoregressive model, proposed by Hyvärinen and Morioka (2017) is also a special case of our framework (Figure 1b), but again with the extension of dimensionality reduction. Furthermore, this framework allows for a plethora of new Nonlinear ICA models to be developed. As described above, these do not have to be limited to time-series but could for instance be a process on a two-dimensional graph with appropriate (in)dependencies (see Figure 1c). However, we now proceed to introduce a particularly useful time-series model using our framework.

3.2 Δ-SNICA : Nonlinear ICA with switching linear dynamical systems

While the above framework has great generality, any practical application will need a specific model. Next we propose one which combines the following properties of previous nonlinear ICA models into a single model: ability to account for both nonstationarity and autocorrelation in a fully unsupervised setting, to perform dimensionality reduction and model hidden states. Real world processes, such as video/audio data, financial time-series, and brain signals, exhibit these properties – disentangling latent features in such data would hence be very useful.

Our new model is depicted in Figure 1d. The independent components are generated by a Switching Linear Dynamical System (SLDS) (Ackerson and Fu, 1968; Chang and Athans, 1978; Hamilton, 1990; Ghahramani and Hinton, 2000) with additional latent variables to express rich dynamics. Formally, for each independent component $i \in \{1, \dots, N\}$, consider the following SLDS over some latent vector $\mathbf{y}_t^{(i)}$:

$$\mathbf{y}_t^{(i)} = \mathbf{B}_{u_t}^{(i)} \mathbf{y}_{t-1}^{(i)} + \mathbf{b}_{u_t}^{(i)} + \boldsymbol{\varepsilon}_{u_t}^{(i)}, \tag{4}$$

where $u_t := u_t^{(i)}$ is a state of a first-order hidden Markov chain $(u_t^{(i)})_{t=1:T}$. Crucially, we assume that the independent components at each time-point are the first elements $y_{t,1}^{(i)}$ of $\mathbf{y}_t^{(i)} = (y_{t,1}^{(i)}, \dots, y_{t,d}^{(i)})^T$, i.e. $s_t^{(i)} = y_{t,1}^{(i)}$. The rest of the elements in $\mathbf{y}_t^{(i)}$ are latent variables modelling hidden dynamics. The great utility of using such a higher-dimensional latent variable is that this model allows us, for example, as a special case, to consider higher-order ARMA processes, thus modelling each $s_t^{(i)}$ as switching between ARMA processes of an order determined by the dimensionality of $\mathbf{y}_t$. We call the ensuing model Δ-SNICA ("Delta-SNICA", with delta as in "dynamic").

4 Identifiability

In this section, we present two very general identifiability theorems for SNICA. We basically decouple the problem into two parts. First, we consider identifying the noise-free distribution of $\mathbf{f}(\mathbf{s}_t)$ from noisy data. Theorem 1 states conditions—on tail behaviour, non-degeneracy, and non-Gaussianity—under which it is possible to recover the distribution of a process based on noisy data with unknown noise distribution. Second, we consider demixing of the nonlinearly mixed data. Theorem 2 provides general conditions—on temporal or spatial dependencies, and non-Gaussianity—that allow recovery of the mixing function $\mathbf{f}$ when there is no more noise. We then consider application of these theorems to SNICA.

4.1 Identifiability with unknown noise distribution

Consider the model

$$\mathbf{x}_t = \mathbf{z}_t + \boldsymbol{\varepsilon}_t \,, \tag{5}$$

where $(\mathbf{z}_t)_{t\in\mathbb{T}}$ is a family of random variables in $\mathbb{R}^M$ such that all $\mathbf{z}_t$, $t \in \mathbb{T}$, have the same marginal distribution, and $(\boldsymbol{\varepsilon}_t)_{t\in\mathbb{T}}$ is a family of independent (over t) and identically distributed random variables, independent of $(\mathbf{z}_t)_{t\in\mathbb{T}}$. Let P be the common distribution of each $\boldsymbol{\varepsilon}_t$, for $t \in \mathbb{T}$. Let t_1 and t_2 in $\mathbb{T}$, and consider the following assumptions.

- (A1) [Tail behaviour] For some $\rho < 3$, there exist A and B such that for all $\lambda \in \mathbb{R}^N$,
$$\mathbb{E}[\exp(\langle \lambda, \mathbf{z}_{t_1} \rangle)] \leqslant A \exp(B\|\lambda\|^\rho) \,.$$
- (A2) [Non-degeneracy] For any $\eta \in \mathbb{C}^M$, $\mathbb{E}[\exp\{\langle \eta, \mathbf{z}_{t_2} \rangle\}|\ \mathbf{z}_{t_1}]$ is not the null random variable.
- (A3) [Non-Gaussianity] The following assertion is false: there exist a vector $\eta \in \mathbb{R}^M$ and independent random variables $\tilde{z}$ and u, such that u is a non dirac Gaussian random variable and $\langle \eta, \mathbf{z}_{t_1} \rangle$ has the same distribution as $\tilde{z} + u$.

We defer the detailed discussion on the practical meaning of the assumptions (A1-A3) in the context of SNICA to Section 4.3. We next present Theorem 1 which establishes identifiability under unknown noise (its proof is postponed to Section A.1 in the Supplementary Material):

Theorem 1 *Assume that assumptions (A1), (A2) and (A3) hold for some $(t_1, t_2) \in \mathbb{T}^2$. Then, up to translation, for all $m \geqslant 2$, for all $(t_3, \ldots, t_m) \in \mathbb{T}^{m-2}$, the application that associates the distribution of $(\mathbf{z}_{t_1}, \ldots, \mathbf{z}_{t_m})$ and P to the distribution of $(\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_m})$ is one-to-one.*

Here, up to translation means that adding a constant vector to all ε_t, and substracting this constant to all $\mathbf{z}_t$, $t \in \{t_1, \ldots, t_m\}$, does not change the distribution of $(\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_m})$. The proof of Theorem 1 extends that of Theorem 1 in (Gassiat et al., 2020b), see also (Gassiat et al., 2020a), which assumed sub-Gaussian noise-free data. Our extension allows the noise-free data to have heavier tails, which is important since (noise-free) data in many real-world applications is super-Gaussian, i.e. heavy-tailed, as is well-known in work on linear ICA (Hyvärinen et al., 2001).

Importantly, there is no assumption on the unknown noise distribution in Theorem 1. In fact, it does not even assume a mixing as in ICA, and thus extends greatly outside of the framework of this paper.

4.2 Identifiability of the mixing function

Based on Theorem 1, it is possible to recover the distribution of the noise-free data in SNICA in (3) by setting $\mathbf{z}_t = \mathbf{f}(\mathbf{s}_t)$. Next, we consider under which conditions the mixing function $\mathbf{f}$ is identifiable. Denote by $S = S^{(1)} \times \cdots \times S^{(N)}$ the support of the distribution of all $\mathbf{s}_t$. We consider the situation where each $S^{(i)} \subset \mathbb{R}$, $1 \leqslant i \leqslant N$, is connected, so that each $S^{(i)}$ is an interval. We assume moreover that the injective mixing function $\mathbf{f}$ is a $\mathcal{C}^2$ diffeomorphism between S and a $\mathcal{C}^2$ differentiable manifold $\mathcal{M} \subset \mathbb{R}^M$. Formally, this means that there exists an atlas $\{\varphi_\vartheta : U_\vartheta \to \mathbb{R}^N\}_{\vartheta\in\Theta}$ of $\mathcal{M}$ such that for all $\vartheta, \vartheta' \in \Theta$, the map $\varphi_\vartheta \circ \varphi_{\vartheta'}^{-1}$ is a $\mathcal{C}^2$ map, and $\mathbf{f}$ is a bijection $\mathbb{R}^N \to \mathcal{M}$ such that for all $\vartheta \in \Theta$, $\varphi_\vartheta \circ \mathbf{f}$ and $\mathbf{f}^{-1} \circ \varphi_\vartheta^{-1}$ have continuous second derivatives. The sets U_ϑ, $\vartheta \in \Theta$, cover $\mathcal{M}$ and are open in $\mathcal{M}$. The proof of Theorem 2 is postponed to Section A.2 in the Supplementary Material.

Theorem 2 *Assume that there exist $m \geqslant 2$ and $(t_1, \dots, t_m) \in \mathbb{T}^m$ such that the vector $(s_{t_1}^{(i)}, \dots, s_{t_m}^{(i)})$ has a density $p_m^{(i)}$ which is $\mathcal{C}^2$ on $(S^{(i)})^m$. Assume moreover that there exist $(k, l) \in \{1, \dots, m\}^2$ with $k \neq l$ such that the following assumptions hold with $Q_m^{(i)} = \log p_m^{(i)}$.*

- *(B1) (Uniform (k,l)-dependency). For all $i \in \{1, \dots, N\}$, the set of zeros of $\frac{\partial^2}{\partial s_{t_k}^{(i)} \partial s_{t_l}^{(i)}} Q_m^{(i)}$ is a meagre subset of $(S^{(i)})^m$, i.e. it contains no open subset.*
- *(B2) (Local (k,l)-non quasi Gaussianity). For any open subset $A \subset S^m$, there exists at most one $i \in \{1, \dots, N\}$ such that there exists a function $\alpha : \mathbb{R}^{m-1} \to \mathbb{R}$ and a constant $c \in \mathbb{R}$ such that for all $s \in A$,*

$$\frac{\partial^2}{\partial s_{t_k}^{(i)} \partial s_{t_l}^{(i)}} Q_m^{(i)} = c\, \alpha(s_{t_k}^{(i)}, \mathbf{s}_{(-t_k,-t_l)}^{(i)}) \alpha(s_{t_l}^{(i)}, \mathbf{s}_{(-t_k,-t_l)}^{(i)}), \tag{6}$$

where $\mathbf{s}_{(-t_k,-t_l)}^{(i)}$ is $(s_{t_1}^{(i)}, \dots, s_{t_m}^{(i)})$ without the coordinates t_k and t_l.

Then, $\mathbf{f}^{-1}$ can be recovered up to permutation and coordinate-wise transformations from the distribution of $(\mathbf{f}(\mathbf{s}_{t_1}), \dots, \mathbf{f}(\mathbf{s}_{t_m}))$.

4.3 Applications to SNICA

In this section, we provide additional comments on the assumptions (A1-A3) and (B1-B2) and their verification in the context of SNICA.

Assumption (A1) is a condition on the tails of the noise-free data: it allows tails that are somewhat heavier than Gaussian tails. It is in fact equivalent to assuming that for some $\tilde{\rho} > 3/2$, there exists $A', B' > 0$ such that for all $t > 0$, $\mathbb{P}(\|\mathbf{z}_{t_1}\| \geqslant t) \leqslant A' \exp(-B' t^{\tilde{\rho}})$.

Assumption (A2) is a non-degeneracy condition likely to be fulfilled for any randomly chosen SNICA model parameters. As an example, consider a model such as Fig. 1c, where there exist hidden variables $(u_t)_{t \in \mathbb{T}}$ taking values in a finite set $\{1, \dots, K\}$ such that the pairs of variables $(\mathbf{z}_t, u_t)$ have the same distribution for all $t \in \mathbb{T}$, and such that conditioned on $(u_t)_{t \in \mathbb{T}}$, the variables $(\mathbf{z}_t)_{t \in \mathbb{T}}$ are independent and the distribution of $\mathbf{z}_t$ only depends on u_t. (As a special case, this model includes the temporal HMM setting described in Fig. 1a.) Let $(t_1, t_2) \in \mathbb{T}^2$. For all $u, v \in \{1, \dots, K\}$, let $\pi(u) = p_{u_{t_1}}(u)$ be the mass function of u_{t_1}, $Q(u, v) = p_{u_{t_2}|u_{t_1}}(v|u)$ be the transition matrix from u_{t_1} to u_{t_2}, and $\gamma_u(\mathbf{z}) = p_{\mathbf{z}_{t_1}|u_{t_1}}(\mathbf{z}|u)$ be the density of $\mathbf{z}_{t_1}$ conditionally to $u_{t_1} = u$. By assumption, it is also the density of $\mathbf{z}_{t_2}$ conditionally to $u_{t_2} = u$. Theorem 3 provides sufficient conditions for assumption (A2) to hold:

Theorem 3 *Assume that Q has full rank, $\min_u \pi(u) > 0$ and the $(\gamma_u)_{1 \leqslant u \leqslant K}$ are linearly independent, then (A2) is satisfied as soon as the functions $(\eta \mapsto \int \exp(\langle \eta, \mathbf{z} \rangle) \gamma_v(\mathbf{z}) d\mathbf{z})_{1 \leqslant v \leqslant K}$ do not have simultaneous zeros.*

Besides the non-simultaneous zeros assumption, the assumptions of Theorem 3 are reminiscent of those used for the identifiability of non-parametric hidden Markov models, see for instance Gassiat et al. (2016); Lehéricy (2019). The key element is that $\mathbf{z}_{t_1}$ and $\mathbf{z}_{t_2}$ are not independent. Thus, we see that (A2) holds if the π and the γ are not degenerate (in the precise sense given by Theorem 3), for the latent state models in Figs. 1a,1c.Another situation where (A2) holds is when $\mathbf{z}_{t_2}$ is a complete statistic (Lehmann and Casella, 2006) in the statistical model $\{\mathbb{P}_{\mathbf{z}_{t_2}|\mathbf{z}_{t_1}}(\cdot|\mathbf{z}_{t_1})\}_{\mathbf{z}_{t_1}}$, where $\mathbb{P}_{\mathbf{z}_{t_2}|\mathbf{z}_{t_1}}(\cdot|\mathbf{z}_{t_1})$ is the distribution of $\mathbf{z}_{t_2}$ conditionally to $\mathbf{z}_{t_1}$. Consider the two following examples where this holds: 1) When the model $\{\mathbb{P}_{\mathbf{z}_{t_2}|\mathbf{z}_{t_1}}(\cdot|\mathbf{z}_{t_1})\}_{\mathbf{z}_{t_1}}$ is an exponential family. In this situation, complete statistics are known. 2) Autoregressive models with additive innovation of the form $\mathbf{z}_{t_2} = \mathbf{h}(\mathbf{z}_{t_1}) + \mathbf{v}_{t_2}$ for some bijective function $\mathbf{h}$ when the additive noise $\mathbf{v}_{t_2}$ is a complete statistic in the statistical model $\{\mathbb{P}_{\mathbf{v}_{t_2}|\mathbf{z}_{t_1}}(\cdot|\mathbf{z}_{t_1})\}_{\mathbf{z}_{t_1}}$ (note that $\mathbf{v}_{t_2}$ cannot be independent of $\mathbf{z}_{t_1}$ here). The case in Fig. 1b is typically covered by this example.

Assumption (A3) states that no direction of the noise free data has a non Dirac Gaussian variable component. It holds as soon as $\mathbf{z}_t = \mathbf{f}(\mathbf{s}_t)$ and the range of $\mathbf{f}$ is such that its orthogonal projection on

any line is not the full line. This assumption holds for instance in the following cases: 1) The range of $\mathbf{f}$ is compact, or 2) the range of $\mathbf{f}$ is contained in a half-cylinder, that is, there exists a hyperplane such that the range of $\mathbf{f}$ is only on one side of this hyperplane and the projection of the range of $\mathbf{f}$ on this hyperplane is bounded.

Assumption (B1) and Assumption (B2) are similar to those in (Hyvärinen and Morioka, 2017; Oberhauser and Schell, 2021) in the special case of time-series, i.e. $\mathbb{T} = \mathbb{N}$. (B1) then entails that there must be sufficiently strong statistical dependence between nearby time points. (B2) is a condition which excludes Gaussian processes and processes which can be trivially transformed to be Gaussian. (For treatment of the Gaussian case, see Appendix B in Supplementary Material.) We can further provide a simple and equivalent formulation when the independent components $\mathbf{s}^{(i)}$ follow independent and stationary HMMs with two hidden states, which is a special case of SNICA. Denote by $\gamma_0^{(i)}$ and $\gamma_1^{(i)}$ the densities of $s_t^{(i)}$ conditionally to $\{u_t^{(i)} = 0\}$ and $\{u_t^{(i)} = 1\}$ respectively.

Theorem 4 *Assume that the stationary distribution π of the hidden chain is such that $0 < \pi(0) < 1$ and that its transition matrix is invertible. Then (B1) and (B2) are satisfied with $m = 2$ if and only if on any open interval, $\gamma_0^{(i)}$ and $\gamma_1^{(i)}$ are not proportional.*

Thus, a very simple HMM leads to these conditions being verified. Hyvärinen and Morioka (2017) already showed that the conditions (B1) and (B2) also hold in the case of non-Gaussian autoregressive models. Thus, we see that our identifiability theory applies both in the case HMM's (Fig 1a) and autoregressive models (Fig 1b), the two principal kinds of temporal structure proposed in previous work, while extending them to further cases and combinations such as in Fig 1c,1d.

A simplification of (B1,B2) It is also possible to combine the assumptions (B1) and (B2) in one, while slightly weakening the generality. The key is to notice that (6) in (B2) implies the derivative in (B1) is zero, by setting $c = 0$. But there is still the difference that (B2) considers all but one index while (B1) considers all indices i. If we simply assume (6) does not hold for any i, we can replace (B1) and (B2) by the new condition:

- (B') For any open subset $A \subset S^m$ and for any $i \in \{1, \ldots, N\}$, a function $\alpha : \mathbb{R}^{m-1} \to \mathbb{R}$ and a constant $c \in \mathbb{R}$ do not exist such that (6) would hold for all $s \in A$.

Note that Hyvärinen and Morioka (2017) defined uniform dependency and (non-)quasi-Gaussianity as two separate properties, but in fact their assumption of non-quasi-Gaussianity was weaker than ours: it did not consider all open subsets separately, which is why this simplification was not possible for them. We believe their definition of non-quasi-Gaussianity was in fact not quite sufficient to prove their theorem, and our stronger version may be needed, in line with Oberhauser and Schell (2021).

5 Experiments

Estimation method One challenge is that it is not practically possible to learn Δ-SNICA by exact maximum-likelihood methods. Instead, we perform learning and inference using Structured VAEs (Johnson et al., 2016) – the current state-of-art in variational inference for structured models. Specifically, this consists of assuming that the latent posterior factorizes as per $q(\mathbf{y}_{1:T}^{(1:N)}, u_{1:T}^{(1:N)}) = \prod_{i=1}^{N} q(\mathbf{y}_{1:T}^{(i)}) q(u_{1:T}^{(i)})$, which allows us to optimize the resulting evidence lower bound (ELBO):

$$\log \widehat{\mathcal{L}} = \mathbb{E}_q \left[\sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \mathbf{s}_t^{(1)}, ..., \mathbf{s}_t^{(N)}) \right] + \sum_{i=1}^{N} \Bigg(- \mathrm{KL} \left[q(u_{1:T}^{(i)}) \middle| p(u_{1:T}^{(i)}) \right] + \mathrm{H} \left[q(\mathbf{s}_{1:T}^{(i)}) \right] + \mathbb{E}_q \left[\log p(\mathbf{s}_1^{(i)} \mid u_1^{(i)}) \right] + \sum_{t=2}^{T} \mathbb{E}_q \left[\log p(\mathbf{s}_t^{(i)} \mid \mathbf{s}_{t-1}^{(i)}, u_t^{(i)}) \right] \Bigg). \tag{7}$$

Since all the distributions are in conjugate exponential families (encoder neural network is used to approximate the natural parameters of the nonlinear likelihood term) efficient message passing can be used for inference, and the mixing function is learned as decoder neural network. Even though this method lacks consistency guarantees (but see Wang and Blei (2018)), we find that our model performs very well. A more detailed treatment of estimation and inference of Δ-SNICA is given in Appendix C. Our code will be openly available at https://github.com/HHalva/snica.

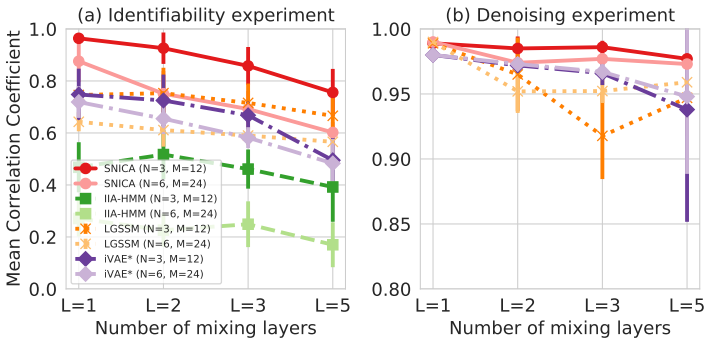


Figure 2: (a) Mean absolute correlation coefficients between ground-truth independent components and their estimates by Δ-SNICA , IIA-HMM, LGSSM and iVAE*, with different orders of complexity (number of layers) and two different dimensions of observed (12, 24) and latent (6, 12) data. (b) Mean absolute correlation coefficient between estimated noise free data and ground-truth noise free data for same set of models except IIA-HMM. Please note the difference in y-axis scales.

5.1 Experiments on simulated data

The identifiability theorems stated above hold in the limit of infinite data. Additionally, a consistent estimator would be required to learn the ground-truth components. In the real world, we are limited by data and estimation methods and hence it is unclear as to what extent we are actually able to estimate identifiable components – and whether identifiability reflects in better performance in real world tasks. To explore this, we first performed experiments on simulated data. We compared the performance of our model to the current state-of-the-art, IIA-HMM (Morioka et al., 2021), as well as identifiable VAE (iVAE) (Khemakhem et al., 2020a) and standard linear Gaussian state-space model (LGSSM). Since iVAE is not able to handle latent auxiliary variables, we allow it to "cheat" by giving it access to the true data generating latent-state, thereby creating a presumably challenging baseline (denoted iVAE* in our figures). LGSSM was included as a naive baseline which is only able to estimate linear mixing function.

Investigating identifiability and consistency We simulated 100K long time-sequences from the Δ-SNICA model and computed the mean absolute correlation coefficient (MCC) between the estimated latent components and ground truth independent components (see Supplementary material for further implementation details). More precisely, to illustrate the dimensionality reduction capabilities we considered two settings where the observed data dimension M, was either 12 or 24 and the number of independent components, N was 3 and 6, respectively. Since IIA-HMM is unable to do dimensionality reduction, we used PCA to get the data dimension to match that of the latent states. We considered four levels of mixing of increasing complexity by randomly initialized MLPs of the following number of layers: 1 (linear ICA), 2, 3, and 5. The results in Figure 2a) illustrate the clearly superior performance of our model. The especially poor performance of IIA-HMM maybe explained by lack of noise model, much simpler latent dynamics, and lost information due to PCA pre-processing. See Appendix D for further discussion and training details.

Application to denoising Δ-SNICA is able to denoise time-series signals by learning the generative model and then performing inference on latent variables. Specifically, SVAE learns the encoder network which is used to perform inference on the posterior of the independent components. We illustrate this using the same settings as above, with the exception that we now use our learned encoder and inference to get the posterior means of the independent components and input these in to the estimated decoder to get predicted noise-free observations, denoted as $\widehat{\mathbf{f}}(\mathbf{s}_t)$ – we measured the correlation between $\widehat{\mathbf{f}}(\mathbf{s}_t)$ and the ground-truth $\mathbf{f}(\mathbf{s}_t)$. Note that IIA-HMM, is not able to perform this task. The results in Figure 2b) show that the other models, designed to handle denoising, perform well at this task, as would be expected – identifiability of the latent state is not necessary for good denoising performance. For LGSSM, denoising is done with the Kalman Smoother algorithm.

5.2 Experiments on real MEG data

To demonstrate real-data applicability, Δ-SNICA was applied to multivariate time series of electrical activity in the human brain, measured by magnetoencephalography (MEG). Recently, many studies have demonstrated the existence of fast transient networks measured by MEG in the resting state and the dynamic switching between different brain networks (Baker et al., 2014; Vidaurre et al., 2017). Additionally, such MEG data is high-dimensional and very noisy. Thus this data provides an excellent target for Δ-SNICA to disentangle the underlying low-dimensional components.

Data and Preprocessing We considered a resting state MEG sessions from the Cam-CAN dataset. During the resting state recording, subjects sat still with their eyes closed. In the task-session data, the subjects carried out a (passive) audio–visual task including visual stimuli and auditory stimuli. We exclusively used the resting-session data for the training of the network, and task-session data was only used in the evaluation. The modality of the sensory stimulation provided a class label that we used in the evaluation, giving in total two classes. We band-pass filtered the data between 4 Hz and 30 Hz (see Supplementary Material for the details of data and settings).

Methods The resting-state data from all subjects were temporally concatenated and used for training. The number of layers of the decoder and encoder were equal and took values 2, 3, 4. We fixed the number of independent components to 5 so that our result can be fairly compared to those in Morioka et al. (2021). To evaluate the obtained features, we performed classification of the sensory stimulation categories by applying feature extractors trained with (unlabeled) resting-state data to (labeled) task-session data. Classification was performed using a linear support vector machine (SVM) classifier trained on the stimulation modality labels and sliding-window-averaged features obtained for each trial. The performance was evaluated by the generalizability of a classifier across subjects. i.e., one-subject-out cross-validation. For comparison, we evaluated the baseline methods: IIA-HMM and IIA-TCL (Morioka et al., 2021). We also visualized the spatial activity patterns obtained by Δ-SNICA , using the weight vectors from encoder neural network across each layer.

Results Figure 3 a) shows the classification accuracies of the stimulus categories, across different methods and the number of layers for each model. The performances by Δ-SNICA were consistently higher than those by the other (baseline) methods, which indicates the importance of the modeling of the MEG signals by Δ-SNICA . Figure 3 b) shows an example of spatial patterns from the encoder network learned by the Δ-SNICA . We used the visualization method presented in (Hyvärinen and Morioka, 2016). We manually picked one out of the hidden nodes from the third layer in encoder network, and plotted its weighted-averaged sensor signals, We also visualized the most strongly contributing second- and first-layer nodes. We see progressive pooling of L1 units to form left lateral frontal, right lateral frontal and parietal patterns in L2 which are then all pooled together in L3 resulting in a lateral frontoparietal pattern. Most of the spatial patterns in the third layer (not shown) are actually similar to those previously reported using MEG (Brookes et al., 2011). Appendix E provides more detail to the interpretation of the Δ-SNICA results.

6 Related work

Previous works on nonlinear ICA have exploited autocorrelations (Hyvärinen and Morioka, 2017; Oberhauser and Schell, 2021) and nonstationarities (Hyvärinen and Morioka, 2016; Hälvä and Hyvärinen, 2020) for identifiability. The SNICA setting provides a unifying framework which allows for both types of temporal dependencies, and further, extends identifiability to other temporal structures as well as any arbitrary higher order data structures which has not previously been considered in the context of nonlinear ICA. Another major theoretical contribution here is to show that identifiability with noise of unknown, arbitrary distribution, while previous work on noisy nonlinear ICA assumed noise of known distribution and known variance (Khemakhem et al., 2020a).

Importantly, the SNICA framework is fully probabilistic and thus accommodates higher order latent variables, leading to "purely unsupervised" learning. This is in large contrast to previous research which have been developed for the case where we are able to observe some additional auxiliary variable, such as audio signals accompanying video (Hyvärinen et al., 2019; Khemakhem et al., 2020a,b), or heuristically define the auxiliary variable based on time structure (Hyvärinen and Morioka, 2016). In practice this means that we are able to estimate our models using (variational)

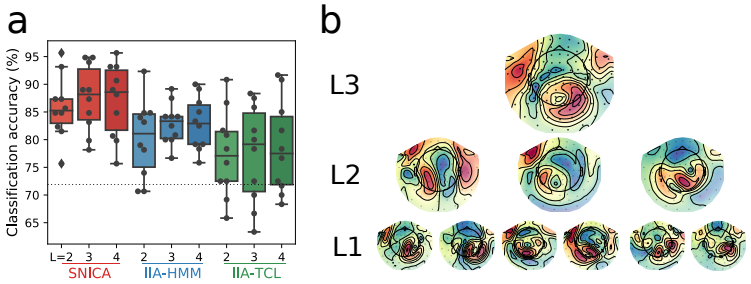


Figure 3: Δ-SNICA on MEG data. (a) Classification accuracies of linear SVMs trained with auditory-visual data to predict stimulus category, with feature extractors trained by Δ-SNICA in advance with resting-state data. Each point represents a testing accuracy on a target subject (chance level: 50%). Horizontal dotted line is PCA-only baseline. (b) Example of spatial patterns of the components learned by Δ-SNICA (L=3). Each topography corresponds to one spatial pattern. L3: approximate total spatial pattern of one third-layer unit. L2: the patterns of the three second-layer units maximally contributing to this L3 unit. L1: for each L2 unit, the two most strongly contributing first-layer units.

MLE, which is more principled than the heuristic self-supervised methods in most earlier papers. The only existing frameworks allowing MLE (Hälvä and Hyvärinen, 2020; Khemakhem et al., 2020a) used model restricted to exponential families, and had either no HMM or a very simple one.

The switching linear dynamical model, Δ-SNICA in Section 3.2, shows the above benefits in the form of a single model. That is, unlike previous nonlinear ICA models, it combines: 1) temporal dependencies and "non-stationarity" (or HMM) in a single model 2) dimensionality reduction within a rigorous maximum likelihood learning and inference framework, and 3) a separate observation equation with general observational noise. This results in a very rich, realistic, and principled model for time series.

Very recently, Morioka et al. (2021) proposed a related model by considering innovations of time series to be nonstationary. However, their model is noise-free, restricted to exponential families of at least order two, and not applicable to the spatial case, thus making our identifiability results significantly stronger. From a more practical viewpoint, their model suffers from the fact that it either does not allow for dimensionality reduction (if an HMM is used) or requires a manual segmentation (if HMM is not used). Nor does it have a clear distinction into a state dynamics equation and a measurement equation which allows for cleaning or denoising of the data.

Limitations Our identifiability theory makes some restrictive assumptions, and it remains to be seen if they could be lifted in future work. In particular, the data is not allowed to have too heavy tails; the noise must be additive, and independent of the signal; and the practical interpretation of some of the assumptions, such as (A3) is difficult. It is also difficult to say whether our assumption of unconditionally independent components is realistic in practice. Regarding practical applications, our specific model only scratches the surface of what is possible in this framework. In particular, we did not develop a model with spatial distributions, nor did we model non-Gaussian observational noise – our main aim was to lay the foundations for the relevant identification theory. Future work should aim to make the estimation more efficient computationally; this is a ubiquitous problem in deep learning, but specific solutions for this concrete problem may be achievable (Gresele et al., 2020).

7 Conclusion

We proposed a new general framework for identifiable disentanglement, based on nonlinear ICA with very general temporal dynamics or spatial structure. Observational noise of arbitrary unknown distribution is further included. We prove identifiability of the models in this framework with high generality and mathematical rigour. For real data analysis, we propose a special case which subsumes the properties of all existing time series models in nonlinear ICA, while generalizing them in many ways (see Section 6 for details). We hope this work will contribute to wide-spread application of identifiable methods for disentanglement in a highly principled, probabilistic framework.

Acknowledgments and Disclosure of Funding

The authors would like to thank Richard Turner for insightful comments and discussion on this work. The authors also wish to thank the Finnish Grid and Cloud Infrastructure (FGCI) for supporting this project with computational and data storage resources. A.H. was supported by a Fellowship from CIFAR, and the Academy of Finland. E.G. would like to acknowledge support for this project from Institut Universitaire de France. J.S. is supported by the University of Cambridge Harding Distinguished Postgraduate Scholars Programme.

References

Ackerson, G. and Fu, K. (1968). On state estimation in switching environments. *IEEE Transactions on Automatic Control*, 15:179–188.

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

Baker, A. P., Brookes, M. J., Rezek, I. A., Smith, S. M., Behrens, T., Smith, P. J. P., and Woolrich, M. (2014). Fast transient networks in spontaneous human brain activity. *Elife*, 3:e01867.

Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A., and Gramfort, A. (2021). Uncovering the structure of clinical EEG signals with self-supervised learning. *J. Neural Engineering*, 18(046020).

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.

Brookes, M. J., Woolrich, M., Luckhoo, H., Price, D., Hale, J. R., Stephenson, M. C., Barnes, G. R., Smith, S. M., and Morris, P. G. (2011). Investigating the electrophysiological basis of resting state networks using magnetoencephalography. *Proceedings of the National Academy of Sciences*, 108(40):16783–16788.

Chang, C. B. and Athans, M. (1978). State estimation for discrete systems with switching parameters. *IEEE Transactions on Aerospace and Electronic Systems*, AES-14(3):418–425.

D'Amour, A., Heller, K. A., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C. Y., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395.

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.

Gassiat, E., Le Corff, S., and Lehéricy, L. (2020a). Deconvolution with unknown noise distribution is possible for multivariate signals. *Annals of Statistics*.

Gassiat, E., Le Corff, S., and Lehéricy, L. (2020b). Identifiability and consistent estimation of nonparametric translation hidden markov models with general state space. *Journal of Machine Learning Research*, 21(115):1–40.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural computation*, 12(4):831–864.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267.

Gresele, L., Fissore, G., Javaloy, A., Schölkopf, B., and Hyvärinen, A. (2020). Relative gradient optimization of the jacobian term in unsupervised deep learning. In *Advances in Neural Information Processing Systems (NeurIPS2020)*, Virtual.

Hälvä, H. and Hyvärinen, A. (2020). Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Proc. 36th Conf. on Uncertainty in Artificial Intelligence (UAI2020)*, Toronto, Canada (virtual).

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1-2):39–70.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

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

Hyvärinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NIPS2016)*, Barcelona, Spain.

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

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. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proc. Artificial Intelligence and Statistics (AISTATS2019)*, Okinawa, Japan.

Johnson, M. J., Duvenaud, D., Wiltschko, A. B., Adams, R. P., and Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2946–2954.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. (2020a). Variational autoencoders and nonlinear ICA: A unifying framework. In *Proc. Artificial Intelligence and Statistics (AISTATS2020)*.

Khemakhem, I., Monti, R. P., Kingma, D. P., and Hyvärinen, A. (2020b). ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Advances in Neural Information Processing Systems (NeurIPS2020)*, Virtual.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. M. (2021). Towards nonlinear disentanglement in natural data with temporal sparse coding. In *9th International Conference on Learning Representations, ICLR 2021*.

Lehéricy, L. (2019). Consistent order estimation for nonparametric hidden Markov models. *Bernoulli*, 25(1):464–498.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research.

Locatello, F., Poole, B., Raetsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR.

Monti, R. P., Zhang, K., and Hyvärinen, A. (2019). Causal discovery with general non-linear relationships using non-linear ICA. In *Proc. 35th Conf. on Uncertainty in Artificial Intelligence (UAI2019)*, Tel Aviv, Israel.

Morioka, H., Calhoun, V., and Hyvärinen, A. (2020). Nonlinear ica of fmri reveals primitive temporal structures linked to rest, task, and behavioral traits. *NeuroImage*, 218:116989.

Morioka, H., Hälvä, H., and Hyvärinen, A. (2021). Independent innovation analysis for nonlinear vector autoregressive process. In *Proc. Artificial Intelligence and Statistics (AISTATS2021)*, Virtual.

Oberhauser, H. and Schell, A. (2021). Nonlinear independent component analysis for continuous-time signals. *arXiv preprint arXiv:2102.02876*.

Shabat, B. V. (1992). *Introduction to complex analysis: functions of several variables*, volume 110. American Mathematical Soc.

Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., Calder, A. J., Marslen-Wilson, W. D., Duncan, J., Dalgleish, T., et al. (2014). The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC neurology*, 14(1):1–25.

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.

Stein, E. and Shakarchi, R. (2003). *Complex Analysis*. Princeton University Press, Princeton.

Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., Tyler, L. K., Henson, R. N., et al. (2017). The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262–269.

Vidaurre, D., Smith, S. M., and Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832.

Wang, Y. and Blei, D. M. (2018). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Wu, P. and Fukumizu, K. (2020). Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method. In *International Conference on Artificial Intelligence and Statistics*, pages 1157–1167. PMLR.

Zhou, D. and Wei, X.-X. (2020). Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

A Appendix

A.1 Proof of Theorem 1

Let $m \geqslant 2$ and $(t_1, \dots, t_m) \in \mathbb{T}^m$. Let R_m and $\tilde{R}_m$ be two possible distributions for $(\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_m})$ that satisfy assumptions (A1), (A2) and (A3) and let P and $\tilde{P}$ be two possible distributions for $\boldsymbol{\varepsilon}_{t_1}$. Assume that the distribution of $(\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_m})$ in the model (5) is the same under (R_m, P) and $(\tilde{R}_m, \tilde{P})$.

Write Φ_{R_m} the characteristic function of R_m, and likewise $\Phi_{\tilde{R}_m}$, Φ_P and $\Phi_{\tilde{P}}$. Following the proof of Theorem 1 of Gassiat et al. (2020b) on the distribution of $(\mathbf{z}_{t_1}, \mathbf{z}_{t_2})$, as in Assumption (A1) we have $\rho < 3$, by Hadamard's factorization theorem, there exist a polynomial function Q with total degree at most 2 and a neighborhood V of 0 in $\mathbb{R}^M$ such that for all $\mathbf{u} \in V$,

$$\Phi_P(\mathbf{u}) \exp\{Q(\mathbf{u})\} = \Phi_{\tilde{P}}(\mathbf{u}) \,. \tag{8}$$

For completeness we provide at the end of this section the sketch of proof of (8).

Writing the characteristic function of $(\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_m})$ under the two sets of parameters yields, for all $(\mathbf{u}_1, \dots, \mathbf{u}_m) \in V^m$,

$$\Phi_{R_m}(\mathbf{u}_1, \dots, \mathbf{u}_m) \prod_{k=1}^m \Phi_P(\mathbf{u}_k) = \Phi_{\tilde{R}_m}(\mathbf{u}_1, \dots, \mathbf{u}_m) \left(\prod_{k=1}^m \Phi_P(\mathbf{u}_k) \right) \left(\prod_{k=1}^m \exp(Q(\mathbf{u}_k)) \right) \,. \tag{9}$$

Since Φ_P is continuous and non-zero at 0, we may divide both sides by $\prod_{k=1}^m \Phi_P(\mathbf{u}_k)$ on a neighborhood of zero. Under assumption (A1), Φ_{R_m} and $\Phi_{\tilde{R}_m}$ can be extended into multivariate analytic functions:

$$\begin{aligned} \Phi_{R_m} : \quad (\mathbb{C}^M)^m &\longrightarrow \mathbb{C} \\ (\mathbf{u}_1, \dots, \mathbf{u}_m) &\longmapsto \int \exp\left(i\mathbf{u}_1^\top \mathbf{z}_{t_1} + \cdots + i\mathbf{u}_m^\top \mathbf{z}_{t_m} \right) \mathrm{d}R_m(\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_m}) \,. \end{aligned}$$

We will need the following statement used in Gassiat et al. (2020a) and Gassiat et al. (2020b). We provide a proof at the end of the section for completeness, see also Shabat (1992).

Lemma 1 *If a multivariate function is analytic on the whole multivariate complex space and is the null function on an open set of the multivariate real space or on an open set of the multivariate purely imaginary space, then it is the null function on the whole multivariate complex space.*

Thus, equation (9) can be extended on $(\mathbb{C}^M)^m$, which shows that for all $(\mathbf{u}_1, \dots, \mathbf{u}_m) \in (\mathbb{C}^M)^m$,

$$\Phi_{R_m}(\mathbf{u}_1, \dots, \mathbf{u}_m) = \Phi_{\tilde{R}_m}(\mathbf{u}_1, \dots, \mathbf{u}_m) \prod_{k=1}^m \exp\{Q(\mathbf{u}_k)\} \,.$$

As Φ_{R_m} and $\Phi_{\tilde{R}_m}$ are characteristic functions, Q has no constant term. The degree 1 term corresponds to a translation parameter. Without loss of generality, assume that $\mathbf{z}_{t_1}$ is centered under R_m and $\tilde{R}_m$, then

$$i\,\mathbb{E}_{R_m}[\mathbf{z}_{t_1}] = \nabla_{\mathbf{u}_1} \Phi_{R_m}(0) = \nabla_{\mathbf{u}_1} \Phi_{\tilde{R}_m}(0) + \nabla Q(0) = i\,\mathbb{E}_{\tilde{R}_m}[\mathbf{z}_{t_1}] + \nabla Q(0) \,,$$

which entails $\nabla Q(0) = 0$. Thus, Q only has terms of degree 2, which means it is a quadratic form in $\mathbb{R}^M$. Writing $Q(\mathbf{u}) = \mathbf{u}^\top (Q_+ - Q_-)\mathbf{u}$ where Q_+ and Q_- are the positive semi-definite matrices corresponding to the positive and negative eigenvalues of Q respectively, yields

$$\Phi_{R_m}(\mathbf{u}_1, \dots, \mathbf{u}_m) \prod_{k=1}^m \exp\left\{-\mathbf{u}_k^\top Q_+ \mathbf{u}_k\right\} = \Phi_{\tilde{R}_m}(\mathbf{u}_1, \dots, \mathbf{u}_m) \prod_{k=1}^m \exp\left\{-\mathbf{u}_k^\top Q_- \mathbf{u}_k\right\} \,.$$

From this decomposition, we deduce that if $\mathbf{z} \sim R_m$, $\tilde{\mathbf{z}} \sim \tilde{R}_m$, and $(\mathbf{v}_k)_{1 \leqslant k \leqslant m}$ (resp. $(\tilde{\mathbf{v}}_k)_{1 \leqslant k \leqslant m}$) are i.i.d. multivariate Gaussian random variables with mean 0 and covariance matrices $2Q_+$ (resp. $2Q_-$) that are independent of $\mathbf{z}$ (resp. $\tilde{\mathbf{z}}$), then $(\mathbf{z}_{t_k} + \mathbf{v}_k)_{1 \leqslant k \leqslant m}$ has the same distribution as $(\tilde{\mathbf{z}}_{t_k} + \tilde{\mathbf{v}}_k)_{1 \leqslant k \leqslant m}$. In particular, the supports of the $\mathbf{v}_k$, $1 \leqslant k \leqslant m$ and of the $\tilde{\mathbf{v}}_k$, $1 \leqslant k \leqslant m$, are orthogonal.

Let Π_- be the orthogonal projection on the support of $\tilde{\mathbf{v}}_k$, then $\Pi_- \mathbf{z}_{t_k} = \Pi_- \tilde{\mathbf{z}}_{t_k} + \tilde{\mathbf{v}}_k$, which by assumption (A3) entails $Q_- = 0$ (otherwise, take a non-zero η in the support of $\tilde{\mathbf{v}}_k$). Since $\tilde{\mathbf{z}}$ satisfies the same assumptions as $\mathbf{z}$, $Q_+ = 0$ for the same reason. Thus, $Q = 0$, so that $\Phi_{R_m} = \Phi_{\tilde{R}_m}$, and then $R_m = \tilde{R}_m$, and likewise $P = \tilde{P}$.

Proof of (8). Since the distribution of $(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})$ in the model (5) is the same under (R_2, P) and $(\tilde{R}_2, \tilde{P})$ (likewise for the distribution of $\mathbf{x}_t$ under (R_1, P) and $(\tilde{R}_1, \tilde{P})$ for any t), we get that for all $\mathbf{u} \in \mathbb{R}^M$,

$$\Phi_P(\mathbf{u})\Phi_{R_1}(\mathbf{u}) = \Phi_{\tilde{P}}(\mathbf{u})\Phi_{\tilde{R}_1}(\mathbf{u}) \tag{10}$$

and for all $(\mathbf{u}_1, \mathbf{u}_2) \in (\mathbb{R}^M)^2$,

$$\Phi_P(\mathbf{u}_1)\Phi_P(\mathbf{u}_2)\Phi_{R_2}(\mathbf{u}_1, \mathbf{u}_2) = \Phi_{\tilde{P}}(\mathbf{u}_1)\Phi_{\tilde{P}}(\mathbf{u}_2)\Phi_{\tilde{R}_2}(\mathbf{u}_1, \mathbf{u}_2)\,. \tag{11}$$

There exists a neighborhood W of 0 in $\mathbb{R}^M$ such that Φ_P and $\Phi_{\tilde{P}}$ do not vanish on W, so that equations (10) and (11) give that for all $(\mathbf{u}_1, \mathbf{u}_2) \in W^2$,

$$\Phi_{R_2}(\mathbf{u}_1, \mathbf{u}_2)\Phi_{\tilde{R}_1}(\mathbf{u}_1)\Phi_{\tilde{R}_1}(\mathbf{u}_2) = \Phi_{\tilde{R}_2}(\mathbf{u}_1, \mathbf{u}_2)\Phi_{R_1}(\mathbf{u}_1)\Phi_{R_1}(\mathbf{u}_2)\,. \tag{12}$$

Application of Lemma 1 yields now that (12) holds for all $(\mathbf{u}_1, \mathbf{u}_2) \in (\mathbb{C}^M)^2$. Using Assumption (A2) and Lemma 1 we easily deduce from (12) that the set of zeros of Φ_{R_1} and $\Phi_{\tilde{R}_1}$ are equal. Then, using Assumption (A1) and Hadamard's factorization Theorem, see Stein and Shakarchi (2003) (Chapter 5 Theorem 5.1), and arguing variable by variable, we deduce that there exists a function Q on $\mathbb{C}^M$ such that, for all $i = 1, \ldots, M$, Q is a polynomial function with degree at most 2 (and coefficients depending on $(u^{(1)}, \ldots, u^{(i-1)}, u^{(i+1)}, \ldots, u^{(M)})$) and for all $\mathbf{u} = (u^{(1)}, \ldots, u^{(M)}) \in \mathbb{C}^M$,

$$\Phi_{R_1}(\mathbf{u}) = \Phi_{\tilde{R}_1}(\mathbf{u}) \exp(Q(\mathbf{u})).$$

Using again Assumption (A1) allows to deduce that Q has total degree 2. Coming back to equation (10) yields for all $\mathbf{u} \in \mathbb{R}^M$,

$$\Phi_P(\mathbf{u})\Phi_{\tilde{R}_1}(\mathbf{u}) \exp(Q(\mathbf{u})) = \Phi_{\tilde{P}}(\mathbf{u})\Phi_{\tilde{R}_1}(\mathbf{u}) \tag{13}$$

which, on the neighborhood V of 0 in $\mathbb{R}^M$ where $\Phi_{\tilde{R}_1}$ does not vanish, proves (8).

Proof of Lemma 1 We prove the statement by induction on the number d of variables. If h is analytic on $\mathbb{C}$ and is not the null function, then h has isolated zeros, so that Lemma 1 holds for $d = 1$. Assume that the lemma holds for analytic functions on $\mathbb{C}^d$ and let h be an analytic function on $\mathbb{C}^{d+1}$ which is the null function on an open set A of $\mathbb{R}^{d+1}$. Then, there exists open sets $B_1, \ldots, B_{d+1}$ of $\mathbb{R}$ such that $B_1 \times \cdots \times B_{d+1} \subset A$. For any $t \in B_{d+1}$, let $h_t : \mathbb{C}^d \to \mathbb{C}$ such that $h_t(\cdot) = h(\cdot, t)$, then h_t is analytic on $\mathbb{C}^d$ and is the null function on $B_1 \times \cdots \times B_d$ so that by the induction hypothesis, for all $z \in \mathbb{C}^d$, $h_t(z) = 0$, that is $h(z, t) = 0$ for all $z \in \mathbb{C}$ and for all $t \in B_{d+1}$. Therefore, for any $z \in \mathbb{C}^d$, the function $h(z, \cdot)$ is analytic on $\mathbb{C}$ and is the null function on B_{d+1} so that for any $z_0 \in \mathbb{C}$, $h(z, z_0) = 0$ and h is the null function. The proof when h is the null function on an open set of the multivariate purely imaginary space is similar.

A.2 Proof of Theorem 2

In the following, the index m may be dropped in the notations $p_m^{(i)}$ and $Q_m^{(i)}$ when there is no confusion. Let $p^{(i)}$, $\tilde{p}^{(i)}$, $\mathbf{f}$ and $\tilde{\mathbf{f}}$ be such that if $\mathbf{s} \sim p^{(i)}$ and $\tilde{\mathbf{s}} \sim \tilde{p}^{(i)}$, then $\mathbf{f}(\mathbf{s})$ and $\tilde{\mathbf{f}}(\tilde{\mathbf{s}})$ have the same distribution. Write $\mathbf{g} = \mathbf{f}^{-1}$ and $\tilde{\mathbf{g}} = \tilde{\mathbf{f}}^{-1}$.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{M}$. For each $k \in \{1, \ldots, m\}$, let $\vartheta_k \in \Theta$ such that $\mathbf{x}_k \in U_{\vartheta_k}$ and let $\mathbf{w}_k = \varphi_{\vartheta_k}(\mathbf{x}_k)$. Writing the density of the random vector $(\varphi_{\vartheta_1}(\mathbf{f}(\mathbf{s}_{t_1})), \ldots, \varphi_{\vartheta_m}(\mathbf{f}(\mathbf{s}_{t_m})))$ at $(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ with respect to the Lebesgue measure for the two parameterizations, yields

$$\begin{aligned}\prod_{k=1}^{m} |J_{\mathbf{g}\circ\varphi_{\vartheta_j}^{-1}}(\mathbf{w}_k)| \prod_{i=1}^{N} p^{(i)}((\mathbf{g}^{(i)} \circ \varphi_{\vartheta_1}^{-1})(\mathbf{w}_1), \ldots, (\mathbf{g}^{(i)} \circ \varphi_{\vartheta_m}^{-1})(\mathbf{w}_m)) \\ = \prod_{k=1}^{m} |J_{\tilde{\mathbf{g}}\circ\varphi_{\vartheta_k}^{-1}}(\mathbf{w}_k)| \prod_{i=1}^{N} \tilde{p}^{(i)}((\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_1}^{-1})(\mathbf{w}_1), \ldots, (\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_m}^{-1})(\mathbf{w}_m))\,. \end{aligned} \tag{14}$$

Let $k, \ell \in \{1, \ldots, m\}$ and $u, v \in \{1, \ldots, N\}$ be such that $k \neq \ell$, then by (14),

$$\begin{aligned}\sum_{i=1}^{N} \frac{\partial^2}{\partial w_k^{(u)} \partial w_\ell^{(v)}} \log p^{(i)}((\mathbf{g}^{(i)} \circ \varphi_{\vartheta_1}^{-1})(\mathbf{w}_1), \ldots, (\mathbf{g}^{(i)} \circ \varphi_{\vartheta_m}^{-1})(\mathbf{w}_m)) \\ = \sum_{i=1}^{N} \frac{\partial^2}{\partial w_k^{(u)} \partial w_\ell^{(v)}} \log \tilde{p}^{(i)}((\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_1}^{-1})(\mathbf{w}_1), \ldots, (\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_m}^{-1})(\mathbf{w}_m))\,,\end{aligned}$$

that is

$$\sum_{i=1}^{N} \frac{\partial^2 \log p^{(i)}}{\partial s_k^{(i)} \partial s_\ell^{(i)}} \left((\mathbf{g}^{(i)} \circ \varphi_{\vartheta_1}^{-1})(\mathbf{w}_1), \dots, (\mathbf{g}^{(i)} \circ \varphi_{\vartheta_m}^{-1})(\mathbf{w}_m) \right) \frac{\partial (\mathbf{g}^{(i)} \circ \varphi_{\vartheta_k}^{-1})}{\partial w^{(u)}}(\mathbf{w}_k) \frac{\partial (\mathbf{g}^{(i)} \circ \varphi_{\vartheta_\ell}^{-1})}{\partial w^{(v)}}(\mathbf{w}_\ell)$$
$$= \sum_{i=1}^{N} \frac{\partial^2 \log \tilde{p}^{(i)}}{\partial s_k^{(i)} \partial s_\ell^{(i)}} \left((\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_1}^{-1})(\mathbf{w}_1), \dots, (\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_m}^{-1})(\mathbf{w}_m) \right) \frac{\partial (\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_k}^{-1})}{\partial w^{(u)}}(\mathbf{w}_k) \frac{\partial (\tilde{\mathbf{g}}^{(i)} \circ \varphi_{\vartheta_\ell}^{-1})}{\partial w^{(v)}}(\mathbf{w}_\ell) .$$

For all $(\mathbf{s}_1, \dots, \mathbf{s}_m) \in S^m$, let

$$q_{i,(k,\ell)} = \frac{\partial^2 \log p^{(i)}}{\partial s_k^{(i)} \partial s_\ell^{(i)}} , \quad \tilde{q}_{i,(k,\ell)} = \frac{\partial^2 \log \tilde{p}^{(i)}}{\partial s_k^{(i)} \partial s_\ell^{(i)}} ,$$
$$D_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_m) = \operatorname{diag} \left(q_{i,(k,\ell)} \left(s_1^{(i)}, \dots, s_m^{(i)} \right) \right)_{1 \leqslant i \leqslant N} ,$$
$$\tilde{D}_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_m) = \operatorname{diag} \left(\tilde{q}_{i,(k,\ell)} \left((\tilde{\mathbf{g}}^{(i)} \circ \mathbf{g}^{-1})(\mathbf{s}_1), \dots, (\tilde{\mathbf{g}}^{(i)} \circ \mathbf{g}^{-1})(\mathbf{s}_m) \right) \right)_{1 \leqslant i \leqslant N} ,$$

so that, writing $(J_a)_{ij} = \partial a_i / \partial x_j$ the Jacobian matrix of the map a and $\mathbf{s}_j = \mathbf{g}(\mathbf{x}_j)$ for each $j \in \{1, \dots, m\}$,

$$J_{\mathbf{g} \circ \varphi_{\vartheta_k}^{-1}}(\mathbf{w}_k)^\top D_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_m) J_{\mathbf{g} \circ \varphi_{\vartheta_\ell}^{-1}}(\mathbf{w}_\ell) = J_{\tilde{\mathbf{g}} \circ \varphi_{\vartheta_k}^{-1}}(\mathbf{w}_k)^\top \tilde{D}_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_m) J_{\tilde{\mathbf{g}} \circ \varphi_{\vartheta_\ell}^{-1}}(\mathbf{w}_\ell) .$$

Note that for all $\mathbf{w} \in \varphi_{\vartheta_k}(U_{\vartheta_k})$,

$$J_{\tilde{\mathbf{g}} \circ \varphi_{\vartheta_k}^{-1}}(\mathbf{w}) (J_{\mathbf{g} \circ \varphi_{\vartheta_k}^{-1}}(\mathbf{w}))^{-1} = J_{\tilde{\mathbf{g}} \circ \mathbf{g}^{-1}}((\mathbf{g} \circ \varphi_{\vartheta_k}^{-1})(\mathbf{w})) ,$$

so that for all $(\mathbf{s}_1, \dots, \mathbf{s}_m) \in S^m$,

$$D_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_m) = J_{\tilde{\mathbf{g}} \circ \mathbf{g}^{-1}}(\mathbf{s}_k)^\top \tilde{D}_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_m) J_{\tilde{\mathbf{g}} \circ \mathbf{g}^{-1}}(\mathbf{s}_\ell) . \tag{15}$$

Consider the following assertion.

- (P) For all $\mathbf{s}$ in a dense subset of S, there exist integers $k, \ell \in \{1, \dots, m\}$ with $k \neq \ell$ and $\mathbf{s}_1, \dots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \dots, \mathbf{s}_m \in S$ such that all entries of the vector

$$\left(\frac{q_{i,(k,\ell)}(\dots, s^{(i)}, \dots, s^{(i)}, \dots) q_{i,(k,\ell)}(\dots, s_\ell^{(i)}, \dots, s_\ell^{(i)}, \dots)}{q_{i,(k,\ell)}(\dots, s^{(i)}, \dots, s_\ell^{(i)}, \dots)^2} \right)_{1 \leqslant i \leqslant N}$$

  are distinct ($s^{(i)}$ and $s_\ell^{(i)}$ are in the positions k and ℓ in the equation above).

Assume that (P) holds. [We shall prove below that (P) holds under the assumptions of Theorem 2]. Let $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_m) \in S$ such that $D_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_m)$ is invertible (any $\mathbf{s}$ in a dense subset of S works thanks to assumption B1). For ease of notation in the following sequence of equations, we drop all unused subscripts and parameters, thus writing $J(\mathbf{s}_k)$ instead of $J_{\tilde{\mathbf{g}} \circ \mathbf{g}^{-1}}(\mathbf{s}_k)$ and $D(\mathbf{s}_k, \mathbf{s}_\ell)$ instead of $D_{k,\ell}(\mathbf{s}_1, \dots, \mathbf{s}_k, \dots, \mathbf{s}_\ell, \dots, \mathbf{s}_m)$ (and likewise for $\tilde{J}$ and $\tilde{D}$). We follow the arguments of the proof of Lemma 2 in Hyvärinen and Morioka (2017) to deduce from (15) an eigenvalue decomposition. Write (15) for several parameters:

$$\begin{aligned} D(\mathbf{s}_k, \mathbf{s}_k) &= J(\mathbf{s}_k)^\top \tilde{D}(\mathbf{s}_k, \mathbf{s}_k) J(\mathbf{s}_k), \\ D(\mathbf{s}_k, \mathbf{s}_\ell) &= J(\mathbf{s}_k)^\top \tilde{D}(\mathbf{s}_k, \mathbf{s}_\ell) J(\mathbf{s}_\ell) \\ &= J(\mathbf{s}_\ell)^\top \tilde{D}(\mathbf{s}_k, \mathbf{s}_\ell) J(\mathbf{s}_k) \quad \text{by symmetry}, \\ D(\mathbf{s}_\ell, \mathbf{s}_\ell) &= J(\mathbf{s}_\ell)^\top \tilde{D}(\mathbf{s}_\ell, \mathbf{s}_\ell) J(\mathbf{s}_\ell), \end{aligned}$$

which altogether entails

$$D(\mathbf{s}_k, \mathbf{s}_\ell)^{-1} D(\mathbf{s}_\ell, \mathbf{s}_\ell) D(\mathbf{s}_k, \mathbf{s}_\ell)^{-1} D(\mathbf{s}_k, \mathbf{s}_k)$$
$$= J(\mathbf{s}_k)^{-1} \left[\tilde{D}(\mathbf{s}_k, \mathbf{s}_\ell)^{-1} \tilde{D}(\mathbf{s}_\ell, \mathbf{s}_\ell) \tilde{D}(\mathbf{s}_k, \mathbf{s}_\ell)^{-1} \tilde{D}(\mathbf{s}_k, \mathbf{s}_k) \right] J(\mathbf{s}_k).$$

The vector in assertion (P) contains the diagonal entries of this diagonal matrix. If they are all distinct, the eigenvalue decomposition is unique, which means that $J(\mathbf{s}_k)$ is the product of a permutation matrix and a diagonal matrix.

Thus, $J_{\tilde{\mathbf{g}}\circ\mathbf{g}^{-1}}$ is the product of a permutation matrix with a diagonal matrix on a dense subset of S, and hence on S by regularity of $\mathbf{g}$ and $\tilde{\mathbf{g}}$.

For any permutation matrix P, the set of all $\mathbf{s} \in S$ where $J_{\tilde{\mathbf{g}}\circ\mathbf{g}^{-1}}(\mathbf{s})$ is the product of P with an invertible diagonal matrix $D(\mathbf{s})$ is both open (by continuity of $J_{\tilde{\mathbf{g}}\circ\mathbf{g}^{-1}}$) and closed (if $\mathbf{s}_n \to \mathbf{s}$ are such that $J_{\tilde{\mathbf{g}}\circ\mathbf{g}^{-1}}(\mathbf{s}_n) = PD_n$ for all n, then by continuity the permutation matrix at $\mathbf{s}$ is also P and since the jacobian is always invertible by the diffeomorphism assumption, $\lim_n D_n$ exists and is invertible). Thus, by connexity of S, the permutation is the same for all $\mathbf{s} \in S$. For the next paragraph, we assume without loss of generality that it is the identity permutation.

Therefore, since for all j and $s^{(j)} \in S^{(j)}$, the set $S^{(1)} \times \cdots \times S^{(j-1)} \times \{s_j\} \times S^{(j+1)} \times \cdots \times S^{(N)}$ is connected, $(\tilde{\mathbf{g}} \circ \mathbf{g}^{-1})^{(j)}$ is constant on this set, and thus it depends on $s^{(j)}$ only. It is bijective on $S^{(j)}$ because both $\mathbf{g}$ and $\tilde{\mathbf{g}}$ are. Thus, $\mathbf{g} = \tilde{\mathbf{g}}$ up to a permutation of the coordinates and a bijective transformation of each coordinate.

Let us now prove that assertion (P) is true. The negation of (P) is that there exists an open set $A \subset S$ such that for all $\mathbf{s} \in A$, for all $k, \ell \in \{1, \dots, m\}$ with $k \neq \ell$ and for all $(\mathbf{s}_1, \dots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \dots, \mathbf{s}_m) \in S^{m-1}$, there exists $i, j \in \{1, \dots, N\}$ with $i \neq j$ such that

$$\frac{q_{i,(k,\ell)}(\dots, s^{(i)}, \dots, s^{(i)}, \dots) q_{i,(k,\ell)}(\dots, s_\ell^{(i)}, \dots, s_\ell^{(i)}, \dots)}{q_{i,(k,\ell)}(\dots, s^{(i)}, \dots, s_\ell^{(i)}, \dots)^2} = \frac{q_{j,(k,\ell)}(\dots, s^{(j)}, \dots, s^{(j)}, \dots) q_{j,(k,\ell)}(\dots, s_\ell^{(j)}, \dots, s_\ell^{(j)}, \dots)}{q_{j,(k,\ell)}(\dots, s^{(j)}, \dots, s_\ell^{(j)}, \dots)^2}. \quad (16)$$

Let $\mathbf{s} \in A$, $k, \ell \in \{1, \dots, m\}$ with $k \neq \ell$. For all $(i, j) \in \{1, \dots, N\}^2$ with $i \neq j$, define $\tilde{S}_{i,j}$ the subset of S^{m-1} such that for all $(\mathbf{s}_1, \dots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \dots, \mathbf{s}_m) \in \tilde{S}_{i,j}$, equation (16) holds. Since the sets $\tilde{S}_{i,j}$, $(i, j) \in \{1, \dots, N\}^2$, $i \neq j$, form a partition of S^{m-1}, which has non-empty interior, there exists at least one pair (i, j) such that the closure of $\tilde{S}_{i,j}$ contains a non-empty open subset $O_{i,j}$. Since $q_{i,(k,\ell)}$ and $q_{j,(k,\ell)}$ are non zero almost everywhere by the uniform (k, ℓ)-dependency assumption, we may assume without loss of generality that the denominators of equation (16) are non zero for all $(\mathbf{s}_1, \dots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \dots, \mathbf{s}_m) \in O_{i,j}$. Thus, by continuity of $q_{i,(k,\ell)}$ and $q_{j,(k,\ell)}$, the terms of equation (16) do not depend on the choice of element in $O_{i,j}$: write $f_{i,(k,\ell)}(s^{(i)}, O_{i,j})$ the left hand term and $f_{j,(k,\ell)}(s^{(j)}, O_{i,j})$ the right hand term.

Let $k, \ell \in \{1, \dots, m\}$ with $k \neq \ell$. Let $(V_n)_{n \geqslant 1}$ be a basis of open sets of $(\mathbb{R}^N)^{m-1}$. For all $(i, j) \in \{1, \dots, N\}$ with $i \neq j$ and $n \in \mathbb{N}^*$, let $A_{(i,j),n}$ be the subset of A such that for all $\mathbf{s} \in A_{(i,j),n}$ and all $(\mathbf{s}_1, \dots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \dots, \mathbf{s}_m) \in V_n$, equation (16) holds. Then, $A = \bigcup_{n \geqslant 1} \bigcup_{i \neq j} A_{(i,j),n}$ (since $O_{i,j}$ contains at least one of the sets of the basis $(V_n)_{n \geqslant 1}$) and thus there exists $i \neq j$ and n such that the interior of the closure of $A_{(i,j),n}$ is non-empty (otherwise A would be a meagre set and thus have empty interior by Baire's category theorem, which is absurd since A is a non-empty open set). Let i, j, n be such that the closure of $A_{(i,j),n}$ has non-empty interior, and B be a non-empty subset of the closure of $A_{(i,j),n}$. Since $q_{i,(k,\ell)}$ and $q_{j,(k,\ell)}$ are non zero almost everywhere by the uniform (k, ℓ)-dependency assumption, we may take an open set $V \subset V_n$ and assume without loss of generality that the denominators of equation (16) are non zero for all $(\mathbf{s}_1, \dots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \dots, \mathbf{s}_m) \in V$ and all $\mathbf{s} \in B$. Thus, by continuity of $q_{i,(k,\ell)}$ and $q_{j,(k,\ell)}$, the terms of equation (16) do not depend on the choice of element in B or V.

To summarize, this means that for all $k, \ell \in \{1, \dots, m\}$ with $k \neq \ell$, there exists $(i, j) \in \{1, \dots, N\}$ with $i \neq j$, a constant c and an open set $A' \subset S^m$ such that for all $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_m) \in A'$,

$$q_{i,(k,\ell)}(\dots, s_k^{(i)}, \dots, s_\ell^{(i)}, \dots)^2 = c q_{i,(k,\ell)}(\dots, s_k^{(i)}, \dots, s_k^{(i)}, \dots) q_{i,(k,\ell)}(\dots, s_\ell^{(i)}, \dots, s_\ell^{(i)}, \dots),$$
$$q_{j,(k,\ell)}(\dots, s_k^{(j)}, \dots, s_\ell^{(j)}, \dots)^2 = c q_{j,(k,\ell)}(\dots, s_k^{(j)}, \dots, s_k^{(j)}, \dots) q_{j,(k,\ell)}(\dots, s_\ell^{(j)}, \dots, s_\ell^{(j)}, \dots).$$

This situation is excluded by the local (k, ℓ)-non quasi Gaussianity assumption, therefore the negation of (P) is false, therefore $\mathbf{g} = \tilde{\mathbf{g}}$ up to permutation and bijective transformation of each coordinate.

A.3 Proof of Theorem 3

For all $\eta \in \mathbb{C}^m$,

$$\mathbb{E}\left[\exp\left\{\langle \eta, \mathbf{z}_{t_2}\rangle\right\} | \ \mathbf{z}_{t_1}\right] = \frac{\sum_{u,v} \pi(u) Q(u,v) \gamma_u(\mathbf{z}_{t_1}) \int \exp(\langle \eta, \mathbf{z}\rangle) \gamma_v(\mathbf{z}) d\mathbf{z}}{\sum_u \pi(u) \gamma_u(\mathbf{z}_{t_1})}$$
$$= \frac{\sum_u \alpha_u(\eta) \pi(u) \gamma_u(\mathbf{z}_{t_1})}{\sum_u \pi(u) \gamma_u(\mathbf{z}_{t_1})},$$

with $\alpha_u(\eta) = \sum_v Q(u,v) \int \exp(\langle \eta, \mathbf{z}\rangle) \gamma_v(\mathbf{z}) d\mathbf{z}$.

Assume that the emission densities $(\gamma_u)_{1 \leqslant u \leqslant K}$ are linearly independent and $\pi(u) > 0$ for all $u \in \{1, \ldots, K\}$, then the only situation where $\mathbb{E}[\exp\{\langle \eta, \mathbf{z}_{t_2}\rangle\} | \mathbf{z}_{t_1}]$ is the null random variable is when $\alpha_u(\eta) = 0$ for all $u \in \{1, \ldots, K\}$. If the functions $(\eta \mapsto \int \exp(\langle \eta, \mathbf{z}\rangle) \gamma_v(\mathbf{z}) d\mathbf{z})_{1 \leqslant v \leqslant K}$ do not have simultaneous zeros and Q has full rank, this is not possible.

A.4 Proof of Theorem 4

We prove that the result holds for all $i = 1, \ldots, N$ and drop the index i in this proof for ease of notation. Denote by

$$\Lambda := \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

the transition matrix of the hidden chain. Then, the stationary distribution is given by $\pi(0) = q/(p+q)$, $\pi(1) = p/(p+q)$, and the distribution of 2 consecutive observations is given by, for all (a,b) in the support:

$$p_2(a,b) = \frac{q(1-p)}{p+q} \gamma_0(a)\gamma_0(b) + \frac{qp}{p+q} \gamma_0(a)\gamma_1(b) + \frac{pq}{p+q} \gamma_1(a)\gamma_0(b) + \frac{p(1-q)}{p+q} \gamma_1(a)\gamma_1(b)\,.$$

If $Q_2 = \log p_2$ then simple computations lead to

$$(p+q)^2 p_2(a,b)^2 \frac{\partial^2 Q_2}{\partial a \partial b} = pq(1-p-q)(\gamma_0(a)\gamma_1'(a) - \gamma_0'(a)\gamma_1(a))(\gamma_0(b)\gamma_1'(b) - \gamma_0'(b)\gamma_1(b))\,.$$

Since $\gamma_0(a)\gamma_1'(a) - \gamma_0'(a)\gamma_1(a) = 0$ for a in an open subset of the support if and only if on this interval γ_0 and γ_1 are proportional, assumption (B1) is satisfied if and only if on any open interval $\gamma_0^{(i)}$ and $\gamma_1^{(i)}$ are not proportional. Moreover, on the set of couples (a,b) such that $\frac{\partial^2 Q_2}{\partial a \partial b} \neq 0$,

$$\log\left(\frac{\partial^2 Q_2}{\partial a \partial b}\right) = \log[|pq(1-p-q)|] - 2\log(p+q) - 2\log p_2(a,b) + h(a) + h(b)\,,$$

where $h(a) = |\gamma_0(a)\gamma_1'(a) - \gamma_0'(a)\gamma_1(a)|$. We deduce easily that (B2) is satisfied if and only if on any open interval $\gamma_0^{(i)}$ and $\gamma_1^{(i)}$ are not proportional.

B Identifiability in Gaussian case

Theorem 2 has a condition on "non-quasi-Gaussianity" which is a generalization of the property of non-Gaussianity typical in ICA. Here, we consider the case of Gaussian noise-free data. Separation is actually possible by the temporal dependencies, but under a stricter condition. We put together results by Hyvärinen and Morioka (2017) and Belouchrani et al. (1997), and arrive at the following result:

Theorem 5 *Assume the data follows the noise-free mixing model $\mathbf{x}_t = \mathbf{f}(\mathbf{s}_t)$ where $\mathbf{s}_t$ is a Gaussian process with independent components, and $\mathbf{f}$ is a $\mathcal{C}^2$ diffeomorphism with $M = N$. Assume further that*

- *The autocovariance functions $c_i(\tau) = cov(s_t^{(i)}, s_{t-\tau}^{(i)})$ are all distinct (i.e. any two of them for i, i' are not equal). (Here, τ takes values in the set allowed by the definition of the index set.)*

Then, $\mathbf{f}^{-1}$ and $\mathbf{f}$ can be recovered up to permutation and coordinate-wise linear transformations (applied on the components $\mathbf{s}_t^{(i)}$) from the distribution of $\mathbf{x}_t$.

The proof is a straightforward implication of two theorems proven earlier: The nonlinear part is identifiable according to Theorem 2 by Hyvärinen and Morioka (2017) but a linear indeterminacy remains; here we need to note that $\bar{\alpha}$ in (Hyvärinen and Morioka, 2017) is a linear function for a Gaussian process. Subsequently the linear part can be identified, thanks to the autocovariance assumption above, as in Theorem 2 of Belouchrani et al. (1997).

Note that in the Gaussian case, it is not possible to apply Theorem 1 since (A3) cannot hold. Thus, Theorem 5 only applies for noise-free data.

C Learning and inference for Δ-SNICA

The Δ-SNICA generative model, as introduced in Section 3.2 can be written as:

$$p(u_1^{(i)}) = \prod_{k=1}^{K} (\pi_k^{(i)})^{\delta(u_1^{(i)}=k)} \tag{17}$$

$$p(u_t^{(i)} \mid u_{t-1}^{(i)}) = \prod_{k=1}^{K} \prod_{\ell=1}^{K} (A_{k\ell}^{(i)})^{\delta(u_t^{(i)}=k)\delta(u_{t-1}^{(i)}=\ell)} \tag{18}$$

$$p(\mathbf{y}_1^{(i)} \mid u_1^{(i)}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}_1^{(i)}; \bar{\mathbf{b}}_k^{(i)}, \bar{\mathbf{Q}}_k^{(i)})^{\delta(u_1^{(i)}=k)} \tag{19}$$

$$p(\mathbf{y}_t^{(i)} \mid \mathbf{y}_{t-1}^{(i)}, u_t^{(i)}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}_t^{(i)}; \mathbf{B}_k^{(i)}\mathbf{y}_{t-1}^{(i)} + \mathbf{b}_k^{(i)}, \mathbf{Q}_k^{(i)})^{\delta(u_t^{(i)}=k)} \tag{20}$$

$$p(\mathbf{x}_t \mid \mathbf{s}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{f}(\mathbf{s}_t), \mathbf{R}) \tag{21}$$

where the superscript (i) again denotes that each independent component $i \in \{1, \ldots, N\}$ follows its own switching linear dynamical system. Also, as explained in Section 3.2, each independent component is part of a higher dimensional latent component $\mathbf{y}_t^{(i)} = (s_t^{(i)}, y_{t,2}^{(i)}, \ldots, y_{t,d}^{(i)})$. The mixing function $\mathbf{f}$ and other variables are defined as in the main text. The log-joint $\log \mathcal{L} = \log p(\mathbf{x}_{1:T}^{(1:N)}, \mathbf{y}_{1:T}^{(1:N)}, u_{1:T}^{(1:N)})$ can be written as:

$$\begin{aligned} \log \mathcal{L} = \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \mathbf{s}_t) + \sum_{i=1}^{N} \Big(& \log p(u_1^{(i)}) + \log p(\mathbf{y}_1^{(i)} \mid u_1^{(i)}) \\ & \sum_{t=2}^{T} \log p(u_t^{(i)} \mid u_{t-1}^{(i)}) + \log p(\mathbf{y}_t^{(i)} \mid \mathbf{y}_{t-1}^{(i)}, u_t^{(i)}) \Big) . \end{aligned} \tag{22}$$

The marginal likelihood is intractable and hence we instead optimize the variational evidence lower bound (ELBO), denoted here $\log \widehat{\mathcal{L}}$, under the assumption that the posterior factorizes as per

$$q(\mathbf{y}_{1:T}^{(1:N)}, u_{1:T}^{(1:N)}) = \prod_{i=1}^{N} q(\mathbf{y}_{1:T}^{(i)}) q(u_{1:T}^{(i)}). \tag{23}$$

The ELBO can thus be written as:

$$
\begin{aligned}
\log \widehat{\mathcal{L}} &= \mathbb{E}_q \left[\log \frac{p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}^{(1:N)}, u_{1:T}^{(1:N)})}{q(\mathbf{y}_{1:T}^{(1:N)}, u_{1:T}^{(1:N)})} \right] \\
&= \mathbb{E}_q \left[\sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \mathbf{s}_t^{(1)}, ..., \mathbf{s}_t^{(N)}) + \sum_{i=1}^{N} \log \frac{p(\mathbf{y}_{1:T}^{(i)} \mid u_{1:T}^{(i)}) p(u_{1:T}^{(i)})}{q(\mathbf{y}_{1:T}^{(i)}) q(u_{1:T}^{(i)})} \right] \\
&= \mathbb{E}_q \left[\sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \mathbf{s}_t^{(1)}, ..., \mathbf{s}_t^{(N)}) \right] + \sum_{i=1}^{N} \left(- \mathrm{KL} \left[q(u_{1:T}^{(i)}) \middle| p(u_{1:T}^{(i)}) \right] + \mathrm{H} \left[q(\mathbf{y}_{1:T}^{(i)}) \right] \right. \\
&\quad \left. + \mathbb{E}_q \left[\log p(\mathbf{y}_{1:T}^{(i)} \mid u_{1:T}^{(i)}) \right] \right) \\
&= \mathbb{E}_q \left[\sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \mathbf{s}_t^{(1)}, ..., \mathbf{s}_t^{(N)}) \right] + \sum_{i=1}^{N} \left(- \mathrm{KL} \left[q(u_{1:T}^{(i)}) \middle| p(u_{1:T}^{(i)}) \right] + \mathrm{H} \left[q(\mathbf{s}_{1:T}^{(i)}) \right] \right. \\
&\quad \left. + \mathbb{E}_q \left[\log p(\mathbf{s}_1^{(i)} \mid u_1^{(i)}) \right] + \sum_{t=2}^{T} \mathbb{E}_q \left[\log p(\mathbf{s}_t^{(i)} \mid \mathbf{s}_{t-1}^{(i)}, u_t^{(i)}) \right] \right) \qquad (24)
\end{aligned}
$$

where H denotes Gaussian differential entropy, and q is always with respect to the relevant variables. As long as all the distributions are conjugate-exponential families, we can use the Structured VAE Johnson et al. (2016) framework for inference and learning. We provide further detail on these two steps below.

Inference Notice that we can write the latent variable part of our generative model in the following useful exponential family forms:

$$
\begin{aligned}
p(u_1^{(i)}) &= \prod_{k=1}^{K} \pi_k^{(i)^{\delta(u_1^{(i)}=k)}} = \exp \left\{ \sum_{i=1}^{K} \delta(u_1^{(i)} = k) \log \pi_k^{(i)} \right\} = \exp \left\{ \langle \boldsymbol{\eta}_{\boldsymbol{\pi}}^{(i)}, \boldsymbol{\delta}_{u_1}^{(i)} \rangle \right\} \\
p(u_t^{(i)} \mid u_{t-1}^{(i)}) &= \prod_{k=1}^{K} \prod_{\ell=1}^{K} A_{k\ell}^{(i)^{\delta(u_{t-1}^{(i)}=k)\delta(u_t^{(i)}=\ell)}} = \exp \left\{ \langle \boldsymbol{\eta}_{\mathbf{A}}^{(i)}, \boldsymbol{\delta}_{u_{t-1}, u_t}^{(i)} \rangle \right\} \qquad (25) \\
p(\mathbf{y}_1^{(i)} \mid u_1^{(i)}) &= \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}_1^{(i)}; \bar{\mathbf{b}}_k^{(i)}, \bar{\mathbf{Q}}_k^{-1^{(i)}})^{\delta(u_1^{(i)}=k)} \\
&= \exp \left\{ \sum_{k=1}^{K} \delta(u_1^{(i)} = k) \left(\langle \mathbf{h}_{1,k}^{(i)}, \mathbf{y}_1^{(i)} \rangle + \mathbf{y}_1^{(i)^T} \mathbf{J}_{1,k}^{(i)} \mathbf{y}_1^{(i)} - \log Z_{1,k}^{(i)} \right) \right\} \\
\mathbf{h}_{1,k}^{(i)} &= \bar{\mathbf{Q}}_k^{(i)} \bar{\mathbf{b}}_k^{(i)} \\
\mathbf{J}_{1,k}^{(i)} &= -\frac{1}{2} \bar{\mathbf{Q}}_k^{(i)} ,
\end{aligned}
$$

where $\log Z_{1,k}^{(i)}$ is the log-normalizer, and similarly

$$
\begin{aligned}
p(\mathbf{y}_t^{(i)} \mid \mathbf{y}_{t-1}^{(i)}, u_t^{(i)}) &= \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}_t^{(i)}; \mathbf{B}_k^{(i)} \mathbf{y}_{t-1}^{(i)} + \mathbf{b}_k^{(i)}, \mathbf{Q}_k^{-1^{(i)}})^{\delta(u_t^{(i)}=k)} \\
&= \exp\left\{ \sum_{k=1}^{K} \delta(u_t^{(i)} = k) \left(\left\langle \mathbf{h}_k^{(i)}, \mathbf{y}_{t-1,t}^{(i)} \right\rangle + \mathbf{y}_{t-1,t}^{(i)^T} \mathbf{J}_k^{(i)} \mathbf{y}_{t-1,t}^{(i)} - \log Z_k^{(i)} \right) \right\} \\
\mathbf{y}_{t-1,t}^{(i)} &= (\mathbf{y}_{t-1}^{(i)}, \mathbf{y}_t^{(i)})^T \\
\mathbf{h}_k^{(i)} &= \begin{pmatrix} \mathbf{B}_k^{(i)^T} \mathbf{Q}_k^{(i)} \mathbf{B}_k^{(i)} & -\mathbf{B}_k^{(i)^T} \mathbf{Q}_k^{(i)} \\ -\mathbf{Q}_k^{(i)} \mathbf{B}_k^{(i)} & \mathbf{Q}_k^{(i)} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{b}_k^{(i)} \end{pmatrix} \\
\mathbf{J}_k^{(i)} &= -\frac{1}{2} \begin{pmatrix} \mathbf{B}_k^{(i)^T} \mathbf{Q}_k^{(i)} \mathbf{B}_k^{(i)} & -\mathbf{B}_k^{(i)^T} \mathbf{Q}_k^{(i)} \\ -\mathbf{Q}_k^{(i)} \mathbf{B}_k^{(i)} & \mathbf{Q}_k^{(i)} \end{pmatrix}.
\end{aligned}
$$

Applying standard results from structured mean-field inference, the updates for the approximate posterior of the HMM latent variables is as follows:

$$
\begin{aligned}
q(u_{1:T}^{(i)}) \propto \exp\Bigg\{ & \log p(u_1^{(i)}) + \sum_{t=2}^{T} \log p(u_t^{(i)} \mid u_{t-1}^{(i)}) \\
& + \mathbb{E}_{q(\mathbf{y}_1^{(i)})} \left[\log p(\mathbf{y}_1^{(i)} \mid u_1^{(i)}) \right] + \mathbb{E}_{q(\mathbf{y}_{t-1,t}^{(i)})} \left[\log p(\mathbf{y}_t^{(i)} \mid \mathbf{y}_{t-1}^{(i)}, u_t^{(i)}) \right] \Bigg\}.
\end{aligned}
$$

And by plugging in the distributions explicitly gives

$$
q(u_{1:T}^{(i)}) \propto \exp\left\{ \langle \boldsymbol{\eta}_{\boldsymbol{\pi}^{(i)}}, \boldsymbol{\delta}_{u_1}^{(i)} \rangle + \langle \boldsymbol{\delta}_{u_1}^{(i)}, \boldsymbol{\rho}_1^{(i)} \rangle + \sum_{t=2}^{T} \langle \boldsymbol{\eta}_{\mathbf{A}^{(i)}}, \mathrm{vec}\left(\boldsymbol{\delta}_{u_{t-1}}^{(i)} \boldsymbol{\delta}_{u_t}^{(i)^T} \right) \rangle + \langle \boldsymbol{\delta}_{u_t}^{(i)}, \boldsymbol{\rho}_t^{(i)} \rangle \right\}, \tag{26}
$$

where we have defined

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{y}_{t-1,t}^{(i)})} \left[\log p(\mathbf{y}_t^{(i)} \mid \mathbf{y}_{t-1}^{(i)}, u_t^{(i)}) \right] &= \sum_{k=1}^{K} \delta(u_t^{(i)} = k) \, \mathbb{E}_{q(\mathbf{y}_{t-1,t}^{(i)})} \Big[\left\langle \mathbf{h}_{t,k}^{(i)}, \mathbf{y}_{t-1,t}^{(i)} \right\rangle + \\
&\qquad \mathbf{y}_{t-1,t}^{(i)^T} \mathbf{J}_{t,k}^{(i)} \mathbf{y}_{t-1,t}^{(i)} - \log Z_{t,k}^{(i)} \Big] \\
&= \langle \boldsymbol{\delta}_{u_t}^{(i)}, \boldsymbol{\rho}_t^{(i)} \rangle .
\end{aligned}
$$

Equation (26) can be viewed as a factor graph of unnormalized potentials – we can therefore use standard message passing algorithms for efficient inference. For instance, the forward-pass is:

$$
\alpha(u_t^{(i)}) = \sum_{u_{t-1}} \exp\left\{ \sum_{t=2}^{T} \langle \boldsymbol{\eta}_{\mathbf{A}^{(i)}}, \mathrm{vec}\left(\boldsymbol{\delta}_{u_{t-1}}^{(i)} \boldsymbol{\delta}_{u_t}^{(i)^T} \right) \rangle + \langle \boldsymbol{\delta}_{u_t}^{(i)}, \boldsymbol{\rho}_t^{(i)} \rangle \right\} \alpha(u_{t-1}^{(i)}). \tag{27}
$$

Similarly, the standard mean-field updates for the dynamical system latent variables gives:

$$
\begin{aligned}
q(\mathbf{y}_{1:T}^{(i)}) \propto \exp\Bigg\{ & \sum_{t=1}^{T} \mathbb{E}_{\prod_{j=1}^{N \setminus i} q(\mathbf{y}_t^{(j)})} \left[\log p(\mathbf{x}_t \mid \mathbf{s}_t) \right] + \mathbb{E}_{q(u_1^{(i)})} \left[\log p(\mathbf{y}_1^{(i)} \mid u_1^{(i)}) \right] \\
& + \sum_{t=2}^{T} \mathbb{E}_{q(u_t^{(i)})} \left[\log p(\mathbf{y}_t^{(i)} \mid \mathbf{y}_{t-1}^{(i)}, u_t^{(i)}) \right] \Bigg\}.
\end{aligned} \tag{28}
$$

The problem here is that we would like to write all the factors in terms of $\mathbf{s}_t$ and $\mathbf{y}_t$ conditonal on $\mathbf{x}_t$. However, due to the nonlinear mixing function, we can't write this directly in conjugate exponential family form. To resolve this, we follow Johnson et al. (2016) and use a decoder neural network to predict approximate natural parameters such that they are in conjugate form, namely:

$$
\mathbb{E}_{\prod_{N \setminus i} q(\mathbf{y}_t^{(j)})} \left[\log p(\mathbf{x}_t \mid \mathbf{s}_t) \right] \propto \langle \mathbf{v}_t(\mathbf{x}_t; \boldsymbol{\phi}), \mathbf{s}_t \rangle + \mathbf{s}_t^T \mathbf{W}_t(\mathbf{x}_t; \boldsymbol{\phi}) \mathbf{s}_t ,
$$

where $\mathbf{v}_t, \mathbf{W}_t$ are thus the outputs of the decoder network, with the latter term assumed to have diagonal structure with negative entries to ensure it's an appropriate Gaussian natural parameter. Further, due to the factored approximation assumption over $\mathbf{y}_t^{(1)}, \ldots, \mathbf{y}_t^{(N)}$ and thus $\mathbf{s}_t^{(1)}, \ldots, \mathbf{s}_t^{(N)}$, above can be written as:

$$\begin{aligned}\mathbb{E}_{\prod_{N\setminus i} q(\mathbf{y}_t^{(j)})}\left[\log p(\mathbf{x}_t \mid \mathbf{s}_t)\right] &\propto \left(v_{t,i} + 2\sum_{j\setminus i}^{N} w_{t,j,i}\, \mathbb{E}_{q(\mathbf{y}_t^{(j)})}\left[y_{t,1}^{(j)}\right]\right) y_{t,1}^{(i)} + w_{t,i,i} y_{t,1}^{(i)^2} \\ &= \langle \tilde{\mathbf{v}}_t{}^{(i)}, \mathbf{y}_t^{(i)} \rangle + \mathbf{y}_t^{(i)^T} \widetilde{\mathbf{W}}^{(i)} \mathbf{y}_t^{(i)} \end{aligned} \tag{29}$$

where $\tilde{\mathbf{v}}_t{}^{(i)}, \widetilde{\mathbf{W}}^{(i)}$ are zero everywhere except in their first indices. The other expectations in Equation (28) are just responsibility weighted natural parameters. For instance:

$$\begin{aligned}\mathbb{E}_{q(u_t^{(i)})}\left[\log p(\mathbf{y}_t^{(i)} \mid \mathbf{y}_{t-1}^{(i)}, u_t^{(i)})\right] &\propto \sum_{k=1}^{K} \mathbb{E}_{q(u_t^{(i)})}\left[\delta(u_t^{(i)} = k)\right] \left(\left\langle \mathbf{h}_{t,k}^{(i)}, \mathbf{y}_{t-1,t}^{(i)} \right\rangle + \mathbf{y}_{t-1,t}^{(i)^T} \mathbf{J}_{t,k}^{(i)} \mathbf{y}_{t-1,t}^{(i)}\right) \\ &\propto \left\langle \tilde{\mathbf{h}}_t^{(i)}, \mathbf{y}_{t-1,t}^{(i)} \right\rangle + \mathbf{y}_{t-1,t}^{(i)^T} \tilde{\mathbf{J}}_t^{(i)} \mathbf{y}_{t-1,t}^{(i)} \\ \tilde{\mathbf{h}}_t^{(i)} &= \sum_{k=1}^{K} \mathbb{E}_{q(u_t^{(i)})}\left[\delta(u_t^{(i)} = k)\right] \mathbf{h}_{t,k}^{(i)} \\ \tilde{\mathbf{J}}_t^{(i)} &= \sum_{k=1}^{K} \mathbb{E}_{q(u_t^{(i)})}\left[\delta(u_t^{(i)} = k)\right] \mathbf{J}_{t,k}^{(i)}\end{aligned}$$

The approximate posterior in (28) can therefore be written as:

$$\begin{aligned}q(\mathbf{y}_{1:T}^{(i)}) \propto \exp\Big\{ &\langle \tilde{\mathbf{v}}_1^{(i)}, \mathbf{y}_1^{(i)} \rangle + \mathbf{y}_1^{(i)^T} \widetilde{\mathbf{W}}^{(i)} \mathbf{y}_1^{(i)} + \langle \tilde{\mathbf{h}}_1^{(i)}, \mathbf{y}_1^{(i)} \rangle + \mathbf{y}_1^{(i)^T} \tilde{\mathbf{J}}_1^{(i)} \mathbf{y}_1^{(i)} \\ &+ \sum_{t=2}^{T} \langle \tilde{\mathbf{v}}_t^{(i)}, \mathbf{y}_t^{(i)} \rangle + \mathbf{y}_t^{(i)^T} \widetilde{\mathbf{W}}^{(i)} \mathbf{y}_t^{(i)} + \left\langle \tilde{\mathbf{h}}_t^{(i)}, \mathbf{y}_{t-1,t}^{(i)} \right\rangle + \mathbf{y}_{t-1,t}^{(i)^T} \tilde{\mathbf{J}}_t^{(i)} \mathbf{y}_{t-1,t}^{(i)} \Big\}. \end{aligned} \tag{30}$$

This can again be viewed as a factor graph on which to perform message passing. The initial forward message is

$$\begin{aligned}\alpha(\mathbf{y}_1) &= \exp\left\{ \langle \tilde{\mathbf{v}}_1 + \tilde{\mathbf{h}}_1, \mathbf{y}_1 \rangle + \mathbf{y}_1^T \left(\widetilde{\mathbf{W}} + \tilde{\mathbf{J}}_1\right) \mathbf{y}_1 \right\}, \\ &= \exp\left\{ \langle \boldsymbol{\eta}_1, \mathbf{y}_1 \rangle + \mathbf{y}_1^T \mathbf{P}_1 \mathbf{y}_1 \right\},\end{aligned}$$

which is an unnormalized Gaussian distribution, and we have dropped superscripts for convenience. The forward equations can be derived as follows, shown here for $t-1=1, t=2$:

$$\alpha(\mathbf{y}_2) = \exp\{\langle \tilde{\mathbf{v}}_2, \mathbf{y}_2 \rangle + \mathbf{y}_2^T \widetilde{\mathbf{W}} \mathbf{y}_2\} \int_{\mathbf{y}_1} \exp\left\{ \left\langle \tilde{\mathbf{h}}_2, \mathbf{y}_{1,2} \right\rangle + \mathbf{y}_{1,2}^T \tilde{\mathbf{J}}_2 \mathbf{y}_{1,2} + \langle \boldsymbol{\eta}_1, \mathbf{y}_1 \rangle + \mathbf{y}_1^T \mathbf{P}_1 \mathbf{y}_1 \right\}.$$

Define $\boldsymbol{\eta}_2^* = (\tilde{\mathbf{h}}_2^1 + \boldsymbol{\eta}_1, \tilde{\mathbf{h}}_2^2)^T$ and $\mathbf{P}_2^* = \begin{pmatrix} \tilde{\mathbf{J}}_2^{11} + \mathbf{P}_1 & \tilde{\mathbf{J}}_2^{12} \\ \tilde{\mathbf{J}}_2^{21} & \tilde{\mathbf{J}}_2^{22} \end{pmatrix}$ with the superscripts denoting block partitions corresponding to $\mathbf{y}_1$ and $\mathbf{y}_2$, so that

$$\alpha(\mathbf{y}_2) = \exp\{\langle \tilde{\mathbf{v}}_2, \mathbf{y}_2 \rangle + \mathbf{y}_2^T \widetilde{\mathbf{W}} \mathbf{y}_2\} \int_{\mathbf{y}_1} \exp\left\{ \langle \boldsymbol{\eta}_2^*, \mathbf{y}_{1,2} \rangle + \mathbf{y}_{1,2}^T \mathbf{P}_2^* \mathbf{y}_{1,2} \right\},$$

where the integral is (unnormalized) joint Gaussian on $(\mathbf{y}_1, \mathbf{y}_2)^T$ with $\boldsymbol{\mu} = -\frac{1}{2} \mathbf{P}_2^{*-1} \boldsymbol{\eta}_2^*$ and $\boldsymbol{\Lambda} = -2\mathbf{P}_2^*$. The block marginalization properties of Gaussian distributions gives:

$$\alpha(\mathbf{y}_2) = \exp\{\langle \tilde{\mathbf{v}}_2, \mathbf{y}_2 \rangle + \mathbf{y}_2^T \widetilde{\mathbf{W}} \mathbf{y}_2\} \exp\left\{ \langle \boldsymbol{\eta}_2, \mathbf{y}_2 \rangle + \mathbf{y}_2^T \mathbf{P}_2 \mathbf{y}_2 \right\},$$

with

$$\begin{aligned}\boldsymbol{\eta}_2 &= \tilde{\mathbf{h}}_2^2 - \tilde{\mathbf{J}}_2^{21} (\tilde{\mathbf{J}}_2^{11} + \mathbf{P}_1)^{-1} (\tilde{\mathbf{h}}_2^1 + \boldsymbol{\eta}_1) \\ \mathbf{P}_2 &= \tilde{\mathbf{J}}_2^{22} - \tilde{\mathbf{J}}_2^{21} (\tilde{\mathbf{J}}_2^{11} + \mathbf{P}_1)^{-1} \tilde{\mathbf{J}}_2^{12}\end{aligned}$$

Thus, the message passing on the linear dynamical system ends up as updates on the natural parameters:

$$\alpha(\mathbf{y}_2) = \exp\left\{\langle \tilde{\mathbf{v}}_2 + \boldsymbol{\eta}_2, \mathbf{y}_2\rangle + \mathbf{y}_2^T\left(\widetilde{\mathbf{W}} + \mathbf{P}_2\right)\mathbf{y}_2\right\},$$

which is analogous to the Kalman filter updates. Similar update equations can be derived for the backward pass and the marginal posteriors are given by the normalized product of the forward and backward passes. Since the resulting distributions are Gaussian, it is easy to compute the expected sufficient statistics required in the inference step described above for $q(u_{1:T})$. In practice, we will cycle between these two inference steps until convergence, after which the M-step is carried out.

Learning After repeating the inference step until convergence, we perform stochastic gradient updates by maximizing the ELBO (Equation (24)) with respect to all the model parameters. In particular, to optimize the first term:

$$\mathbb{E}_q\left[\sum_{t=1}^{T}\log p(\mathbf{x}_t \mid \mathbf{s}_t^{(1)}, ..., \mathbf{s}_t^{(N)})\right]$$

we sample $\mathbf{s}_t^{(1:N)} \sim q(\mathbf{s}_t^{(1:N)}), \forall t \in (1, \ldots, T)$, and parameterize the mixing function with a decoder neural network $\mathbf{f}(\cdot; \boldsymbol{\theta})$:

$$p(\mathbf{x}_t \mid \mathbf{s}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{f}(\mathbf{s}_t; \boldsymbol{\theta}), \mathbf{R})\,. \tag{31}$$

D Details on experiments on simulated data

Simulated data We simulated 100K long time-sequences from the Δ-SNICA and computed the mean absolute correlation coefficient (MCC) between the estimated latent components and ground true independent components. The switching linear dynamical system was simulated to have two latent hmm states, one that induced strong mean reverting behaviour upon the linear dynamical system, and another with oscillatory dynamics. The dimension of the linear dynamical system state-space was also set to 2 (1 + independent component). The HMM transition matrix was close to diagonal with 0.99 probability of staying in current state and 0.01 probability of transitioning to the other state, at each time step of the 100k long sequence. The code at [redacted for anonymity] provides the exact simulation details. To illustrate the dimensionality reduction capabilities we considered two settings where the observed data dimension M, was either 12 or 24 and the number of independent components, N was 3 and 6, respectively. Therefore the model consist of N independent processes of Equation (4). Observations were created by the mixing function (Eq. (3)) and additive Gaussian diagonal noise. We considered four levels of mixing of increasing complexity by randomly initialized MLPs of the following number of layers: 1 (linear ICA), 2, 3, and 5.

Training details All the experiments were run on ten different randomly simulated data sets to compute error bars. The model parameters, including the mixing function, were estimated using the inference and learning algorithm described above. All parameters were trained in ordered to increase the ELBO of the model; Adam with learning rate 1e-2 was used. The number of layers in the decoder networks was set equal to the number of mixing layers for both Δ-SNICA and IIA-HMM benchmark. The number of layers in the encoder Δ-SNICA was always one more than that for the decoder. We suspect this extra nonlinearity in the encoder helped training since VAEs have tendency to over-emphasize learning the likelihood term, which this may have alleviated. The number of hidden units was set at 128 and 64 for the decoder and encoder respectively. In order to avoid local minima, we started training from 20 different inital seeds and chose the model that reached the highest ELBO, or likelihood. The models were trained on University of Helsinki SLURM cluster until convergence, which in practice was approximately 12 hours on most settings. All training was done on CPUs only. Memory used for a single model to be trained was 15G RAM.

Further discussion of results One possible reason for the relatively poor performance of IIA-HMM on the simulated data experiment (Figure 2) was suspected to be loss of information that resulted from the PCA preprocessing step. We explored this in additional experiments where there was no dimension reduction: Δ-SNICA still outperforms IIA-HHM also in this setting, although the latter's performance is now improved for small dimensionality (in 3-dimensions: MCC avg.

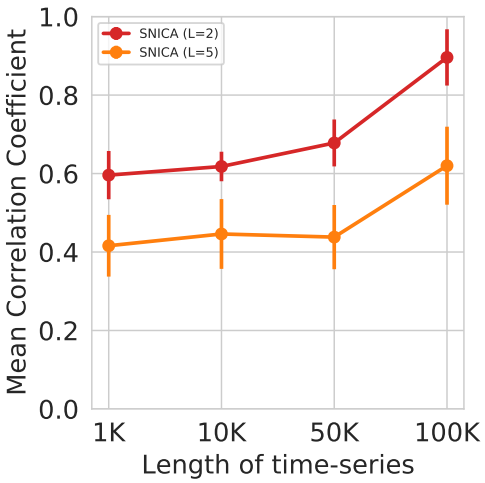


Figure 4: Mean absolute correlation coefficient between estimated and ground true independent components for varying lengths of training data for Δ-SNICA (N=3, M=12), for equal training time. Result shown for two different numbers of mixing layers L=2 and L=5

0.4 for 3 mixing layers), though remains clearly below Δ-SNICA (3-dimensions: MCC avg. 0.7). IIA-HMM performance for dimensions above 6, even without dimension reduction, was very poor (MCC < 0.3) suggesting its poor performance is not solely due to PCA, bur rather likely due to it lacking observation noise model (unlike all the other models we considered) and simpler model of latent dynamics (original iVAE also has no latent dynamics model but was here supplied with the ground-truth HMM latent state thus giving it a substantial advantage over IIA-HMM).

Size of training data The theoretical identifiability results presented in this paper hold in the limit of infinite data. Hence, we hypothesized that the amount of training data may have large impact in any practical situations – in addition to the usual benefits of increased dataset size. To explore this, we trained our model for varying lengths of datasets, with the results shown in Figure 4. We observed much better results for the largest dataset. Due to limited compute available to us, we leave it for future works to investigate even larger data sizes.

E Details on MEG experiment

Data and Preprocessing The MEG data used were from the open Cam-CAN data repository[3] (available at http://www.mrc-cbu.cam.ac.uk/datasets/camcan/), and released under Creative Commons license. (Taylor et al., 2017; Shafto et al., 2014). The MEG dataset was collected using a 306-channel VectorView MEG system (Elekta Neuromag, Helsinki), consisting of 102 magnetometers and 204 orthogonal planar gradiometers with sampling 1000Hz. MEG data was preprocessed by temporal signal space separation (tsss; MaxFilter 2.2, Elekta Neuromag Oy, Helsinki, Finland) to remove noise from external sources and from HPI coils and head-motion was corrected (see (Taylor et al., 2017) for more details of the preprocessing). During the resting state recording, subjects sat still with their eyes closed for at least 8 min and 40 s. In the task-session data, the subjects carried out a (passive) audio–visual task including 120 trials of unimodal stimuli (60 visual stimuli: bilateral/full-field circular checkerboards; 60 auditory stimuli: binaural tones), presented at a rate of approximately 1 per second. In this study, We applied the method to 10 subjects' data and downsampled it to 128 Hz for saving computational resources, and only data from the planar gradiometers (204 channels) were used. We further band-pass filtered the data between 4 Hz and 30 Hz and normalized them to have

[3]Acknowledgment for Cam-CAN data: Data collection and sharing for this project was provided by the Cambridge Centre for Ageing and Neuroscience (CamCAN). CamCAN funding was provided by the UK Biotechnology and Biological Sciences Research Council (grant number BB/H008217/1), together with support from the UK Medical Research Council and University of Cambridge, UK.

zero-mean and unit variance. For the task-session data, we cropped each trial from -300ms to 600ms after the onset. The MNE package (Gramfort et al., 2013) was used for preprocessing.

SNICA setting We only used resting-state data for training. For saving memory, we selected 5-min long resting-state data from each subject. We temporally concatenated segments of each subject to form a dataset (5*60*128*10 = 384k time points) for training. We fixed the number of independent components to 5, and set the number of hidden markov states and the dimension of the linear dynamical system to 2. The number of layesr in the encoder and decoder networks was set equal, and the number of hidden units was set to 32. Otherwise, all the settings were as in Simulation.

Evaluation Methods For evaluation, we used the model trained with (unlabeled) resting-state data as feature extractors to perform a downstream task for classification of (labeled) task-session data. We carried out classification of the stimulus modality (auditory or visual) by using the estimated features. Classification was performed using a linear support vector machine (SVM) classifier trained on the stimulation modality labels and sliding-window-averaged features (width=10 and stride=3 samples) for each trial. The performance was evaluated by the generalizability of a classifier across subjects, i.e., one-subject-out cross-validation (OSO-CV). The hyperparameters of the SVM were determined by nested OSO-CV without using the test data. For comparison, IIA-HMM and IIA-TCL for the nonlinear vector autoregressive model (NVAM) were applied as baseline methods. Since IIA-HMM is not able to reduce the dimensionality, PCA was performed on the concatenated resting-state data to reduce the dimension to 5 for fair comparison. For IIA-TCL, we used segments of equal size, of length 10 s or 1280 data points, and also set the number of independent innovation to 5 for fair comparison.

We visualized the spatial patterns of the estimated features by plotting the weight vectors of units from encoder MLP in the topography map space. For the first layer, we have weight vectors (columns of the weight matrix $\mathbf{W}_1$) across sensors for each unit, and directly mapped them into brain topography space. And the weight matrix $\mathbf{W}_2$ multiplied by $\mathbf{W}_1$ to obtain weight vectors (columns of $\mathbf{W}_1\mathbf{W}_2$) of sensors for each unit in the second layer, and so on for subsequent layers.

Interpreting the latent dynamics in the MEG experiments The learned parameters for the Δ-SNICA 's latent dynamics, namely HMM-style switching, provide interesting interpretations in the MEG data experiment. Since we fix the number of HMM states to be two for each component, our assumption is that they can be interpreted as on/off or activity/inactivity. Such long-term on/off switching of the sources thus characterizes the nonstationary of the brain signal, as is quite often assumed in brain imaging. In particular, the components can be interpreted to represent different dynamic brain processes that are well-known to exist in the resting brain: visual, auditory, and other sensory networks; executive networks, attentional networks, and default mode network. The specific transition matrices for the hidden Markov discrete states can be interpreted to represent the movement between the transient brain states (process) in the real data. In particular, we found the HMM transition matrix to be close to diagonal, which suggests that we are capturing relatively slowly evolving states. The precise figures from the transition matrix suggest that on average a given state (active/inactive) lasts between 0.8 and 7 seconds. The marginal probabilities of the different states are fairly similar to each other, ranging between 0.3 to 0.6, thus all the states are relatively common in this sense. The hidden continuous states, on the other hand, are used here mainly as an algorithmic trick to easily model higher-order AR processes, and are thus harder to interpet.