Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA

Anonymous Author(s)
Affiliation
Address
email

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. The SNICA setting therefore subsumes all the existing nonlinear ICA models for time-series and also allows for new much richer identifiable models. 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., 2020a; 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., 2020; 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 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., 2020; Oberhauser and Schell, 2021)

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

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 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. Lastly, the identifiability theorems in those 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 for a very rich class of models, in a much more general sense than done previously. Importantly, our identifiability results are able to exploit dependency structures of any arbitrary order, and therefore can extend identifiability for instance 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, previously unattainable, very practical properties: the ability to account for both nonstationarity and autocorrelation in a fully unsupervised setting; 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 dimension reduction is not possible in most of the above models. The only exception is Khemakhem et al. (2020a) who achieve this by assuming that we know the distribution of some additive noise on the observations $\mathbf{x} = \mathbf{f}(\mathbf{s}) + \boldsymbol{\varepsilon}$, and by choosing $\mathbf{f}$ as injective rather than bijective. This 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

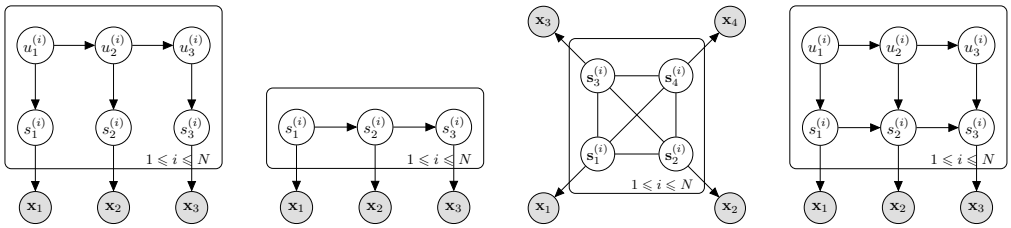
Consider observations $(\mathbf{x}_t)_{t\in\mathbb{T}} = ((x_t^{(1)}, \dots, x_t^{(M)}))_{t\in\mathbb{T}}$ where $\mathbb{T}$ is a discrete indexing set of arbitrary dimension. For discrete time-series models, like previous works, $\mathbb{T}$ would be a subset of $\mathbb{N}$. Crucially, however, we 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, which no previous work has allowed for.

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 $i \in \{1, \dots, N\}$ 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 $(t_1, \dots, t_m) \in \mathbb{T}^m$, $p(\mathbf{s}_{t_1}, \dots, \mathbf{s}_{t_m}) = \prod_{i=1}^{N} p(s_{t_1}^{(i)}, \dots, s_{t_m}^{(i)})$: 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 $\mathbf{s}_t = (s_t^{(1)}, \dots, s_t^{(N)})$. 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 recast and often improved upon when viewed through this new unifying framework. 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.



(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

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, again with the goal of subsuming previous models used in nonlinear ICA. In particular, we combine the two statistical properties of "non-stationarity"(e.g HMMs) and stationary temporal dependencies (e.g. autoregressive models). No model has combined these two aspects in the context of nonlinear ICA. Yet, real world processes, such as video/audio data, financial time-series, and brain signals, exhibit these properties – disentangling latent features in such models 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. In particular, they enable identifiability based on either the "nonstationarities" or the temporal dependencies, thus generalizing results of previous work.

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.2 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 $(t_3, \dots, t_m) \in \mathbb{T}^{m-2}$, the application that associates the distribution of $(\mathbf{z}_{t_1}, \dots, \mathbf{z}_{t_m})$ and P to the distribution of $(\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_m})$ is one-to-one.*

Here, up to translation means that adding a constant vector to all ε_t, and substracting this constant to all $\mathbf{z}_t$, $t \in \{t_1, \dots, t_m\}$, does not change the distribution of $(\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_m})$. 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.3 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 $(s^{(i)}_{t_1}, \dots, s^{(i)}_{t_m})$ has a density $p^{(i)}_m$ 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 $Q^{(i)}_m = \log p^{(i)}_m$.*

- *(B1) (Uniform (k,l)-dependency). For all $i \in \{1, \dots, N\}$, the set of zeros of $\frac{\partial^2}{\partial s^{(i)}_{t_k} \partial s^{(i)}_{t_l}} Q^{(i)}_m$ 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 $s \in A$,*
$$\frac{\partial^2}{\partial s^{(i)}_{t_k} \partial s^{(i)}_{t_l}} Q^{(i)}_m = c\,\alpha(s^{(i)}_{t_k}, \mathbf{s}^{(i)}_{(-t_k,-t_l)})\alpha(s^{(i)}_{t_l}, \mathbf{s}^{(i)}_{(-t_k,-t_l)})\,, \tag{6}$$
where $\mathbf{s}^{(i)}_{(-t_k,-t_l)}$ is $(s^{(i)}_{t_1}, \dots, s^{(i)}_{t_m})$ 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 statistics 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.*

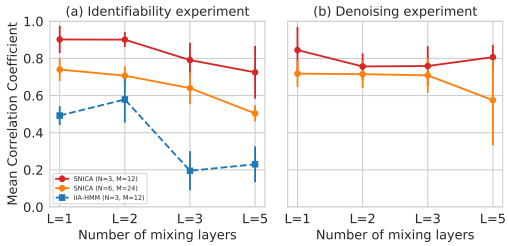


Figure 2: (a) Mean absolute correlation coefficients between ground true independent components and their estimates by Δ-SNICA (solid lines), with different orders of complexity (number of layers) and two different dimensions of observed (12, 24) and latent (6, 12) data. Results for IIA-HMM (dashed line) shown for comparison. (b) Mean absolute correlation coefficient between estimated noise free data and ground true noise free data for Δ-SNICA .

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.

5 Experiments

Estimation method One challenge is that it is not practically possible to learn Δ-SNICA by exact maximum-likelihood methods. However, by framing the model within conjugate exponential families we are able to perform learning and inference using Structured VAEs (Johnson et al., 2017) – the current state-of-art in variational inference for structured data. Despite lacking consistency guarantees (but see Wang and Blei (2018)), we find that our model performs very well. A detailed treatment of estimation and inference of Δ-SNICA is given in Supplementary Material. Our code is openly available at [address redacted for anonymity].

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., 2020b).

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. This is expected as IIA-HMM has a much simpler model of dynamics, and no noise model, and likely lost information due to PCA pre-processing. Details and more evaluations are provided in the Supplementary Material.

Application to denoising Δ-SNICA is able to denoise time-series signals by learning the generative model and then performing inference on latent variables. 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 then use these in the ground-truth 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, or any other latent variable nonlinear ICA model, is not able to perform this task. The results in Figure 2b) show that our model performs well in this task.

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. 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).

6 Related work

The SNICA setting is much broader than any previous work, in fact it subsumes most existing time-series nonlinear ICA models (Hyvärinen and Morioka, 2017; Oberhauser and Schell, 2021; Hälvä and Hyvärinen, 2020). Furthermore, we extend identifiability to models exploiting any higher ordered structures in data rather than just time-dependencies used in previous work. 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 accomodates for 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

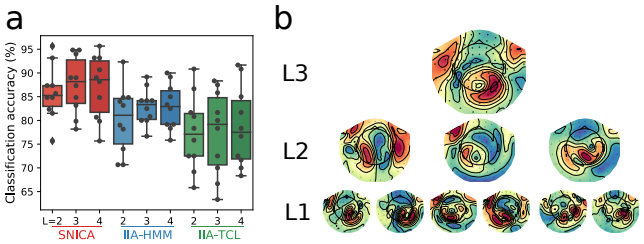


Figure 3: Δ-SNICA on MEG data. (a) Classification accuracies of linear SVMs newly trained with auditory-visual task data to predict stimulus category, with feature extractors trained by Δ-SNICA in advance with resting-state data. Each point represents a testing accuracy on a target subject (chance level: 50%). (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 selected 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.

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) 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 any existing model, 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. 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 still subsumes 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.

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. *arXiv:2006.14226*.

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.

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., Datta, S. R., and Adams, R. P. (2017). Composing graphical models with neural networks for structured representations and fast inference.

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.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. (2020). Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*.

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.

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. (2020a). Nonlinear ICA of fMRI reveals primitive temporal structures linked to rest, task, and behavioral traits. *NeuroImage*, 218:116989.

Morioka, H., Calhoun, V., and Hyvärinen, A. (2020b). 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*.

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.

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. *arXiv preprint arXiv:2011.04798*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work is mainly theoretical, with the aim of providing theoretical guarantees for the identifiability of a new framework for deep models. Identifiability is curcial for reproducible science and interpretable results and generally a desirable property. Our theoretical guarantees abstract away the nature of the data and the practical implementation. We are strictly against any malicious uses of such deep models.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] All authors have read this in depth and adhered fully to it in the paper
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See section 4 and supplementary material
   (b) Did you include complete proofs of all theoretical results? [Yes] These are all in the supplementary material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have provided link though its redacted for the initial review period to protect anonymity. Experimental details are also given in supplementary material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] These are given in the supplementary material and more concretely in the code
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] see Section 5
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Experimental details in Supplementary material and the Acknowledgement section for provider
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Cam-CAN was cited in the real data experiments section

(b) Did you mention the license of the assets? [Yes] See supplementary for experiments; we mention that Cam-CAN is released under creative commons license. License for our model's code is on the github page.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provided URL to the code for our model but it has been anonymized for review

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] As explained in Supplementary material, we applied and were approved for data access at Cam-CAN

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] As explained in the Supplementary material, the Cam-CAN data is anonymized

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] no experiments with human subjects were conducted

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] no experiments with human subjects were conducted

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] no experiments with human subjects were conducted