Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem
Gatsby Computational Neuroscience Unit, UCL
ilyesk@gatsby.ucl.ac.uk

Diederik P. Kingma
Google Brain
durk@google.com

Aapo Hyvärinen
INRIA-Saclay
Dept of CS, University of Helsinki
aapo.hyvarinen@helsinki.fi

Abstract

The framework of variational autoencoders allows us to efficiently learn deep latent-variable models, such that the model's marginal distribution over observed variables fits the data. Often, we're interested in going a step further, and want to approximate the true joint distribution over observed and latent variables, including the true prior and posterior distributions over latent variables. This is known to be generally impossible due to unidentifiability of the model. We address this issue by showing that for a broad family of deep latent-variable models, identification of the true joint distribution over observed and latent variables is actually possible up to a simple transformation, thus achieving a principled and powerful form of disentanglement. Our result requires a factorized prior distribution over the latent variables that is conditioned on an additionally observed variable, such as a class label or almost any other observation. We build on recent developments in nonlinear ICA, which we extend to the case with noisy, undercomplete or discrete observations, integrated in a maximum likelihood framework. The result also trivially contains identifiable flow-based generative models as a special case.

1 Introduction

The framework of variational autoencoders [18, 24] (VAEs) and its extensions (e.g. [3, 17, 26, 20]) offers a scalable set of techniques for learning deep latent-variable models and corresponding inference models. With VAEs, we can in principle learn flexible models of data such that, after optimization, the model's implicit marginal distribution over the observed variables approximates their true (but unknown) distribution. With VAEs we can also efficiently synthesize pseudo-data from the model.

However, we're often interested in going a step further and want to learn the true joint distribution over both observed and latent variables. This is generally a very difficult task, since by definition we only ever observe the observed variables, never the latent variables, therefore we cannot directly estimate their joint distribution. If we could however somehow achieve this task and learn the true joint distribution, this would imply that we have also learned to approximate the true prior and posterior distributions over latent variables. Learning about these distributions can be very interesting for various purposes, for example in order to learn about latent structure behind the data, or in order to infer the latent variables from which the data originated.

Preprint. Under review.

Learning the true joint distribution is only possible when the model is *identifiable*, as we will explain. The original VAE theory doesn't tell us how or when this is the case; it only tells us how to optimize the model's parameters such that its (marginal) distribution over the observed variables matches the data. The original theory doesn't tell us if or when we learn the correct joint distribution over observed and latent variables.

Almost no literature exists on achieving this goal. A pocket of the VAE literature works towards the related goal of *disentanglement*, but offer no proofs or theoretic guarantees of identifiability of the model or its latent variables. The most prominent of such models are β-VAEs and their extensions [4, 9, 8, 7, 15, 5], in which the authors introduce adjustable hyperparameters in the VAE objective to encourage disentanglement. Other work attempts to find maximally independent components through the GAN framework [2]. However, models in these earlier works are actually non-identifiable due to non-conditional latent priors, as has been seen empirically [19], and we will show formally below.

Recent work in nonlinear Independent Component Analysis (ICA) theory [10, 11, 13] provided the first identifiability results for deep latent-variable models. Nonlinear ICA provides a rigorous framework for recovering independent latents that were transformed by some invertible nonlinear transformation into the data. Some special but not very restrictive conditions are necessary, since it is known that when the function from latent to observed variables is nonlinear, the general problem is ill-posed, and one cannot recover the independent latents [12]. However, existing nonlinear ICA methods do not learn to model the data distribution (pdf), nor do they allow us to synthesize pseudo-data.

In this paper we show that under relatively mild conditions the joint distribution over observed and latent variables in VAEs is identifiable and learnable, thus bridging the gap between VAEs and nonlinear ICA. To this end, we establish a principled connection between VAEs and an identifiable nonlinear ICA model, providing a unified view of two complementary methods in unsupervised representation learning. This integration is achieved by using a latent prior that has a factorized distribution that is conditioned on additionally observed variables, such as a class label, time index, or almost any other further observation. Our theoretical results trivially apply to any consistent parameter estimation method for deep latent-variable models, not just the VAE framework. We found the VAE a logical choice since it allows for efficient latent-variable inference and scales to large datasets and models. Finally, we put our theoretical results to a test in experiments. Perhaps most notably, we find that on a synthetic dataset with known ground-truth model, our method with an identifiable VAE indeed learns to closely approximate the true joint distribution over observed and latent variables, in contrast with a baseline non-identifiable model.

2 Unidentifiability of deep latent variable models

2.1 Deep latent variable models

Consider an observed data variable (random vector) $\mathbf{x} \in \mathbb{R}^d$, and a latent random vector $\mathbf{z} \in \mathbb{R}^n$. A common deep latent variable model has the following structure:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z}) \tag{1}$$

where $\boldsymbol{\theta} \in \Theta$ is a vector of parameters, $p_{\boldsymbol{\theta}}(\mathbf{z})$ is called a prior distribution over the latent variables. The distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, if parameterized with a neural network often called the *decoder*, tells us how the distribution on $\mathbf{x}$ depends on the values of $\mathbf{z}$. The model then gives rise to the observed distribution of the data as:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} \tag{2}$$

Assuming $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is modelled by a deep neural network, this can model a rich class of data distributions $p_{\boldsymbol{\theta}}(\mathbf{x})$.

We assume that we observe data which is generated from an underlying joint distribution $p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}^*}(\mathbf{z})$ where $\boldsymbol{\theta}^*$ are its true but unknown parameters. We then collect a dataset of observations of $\mathbf{x}$:

$$\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\} \text{ where } \begin{aligned} \mathbf{z}^{*(i)} &\sim p_{\boldsymbol{\theta}^*}(\mathbf{z}) \\ \mathbf{x}^{(i)} &\sim p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z}^{*(i)}) \end{aligned}$$

Note that the original values $\mathbf{z}^{*(i)}$ of the latent variables $\mathbf{z}$ are by definition not observed and unknown. The ICA literature, including this work, uses the term *sources* to refer to $\mathbf{z}^{*(i)}$. Also note that we could just as well have written: $\mathbf{x}^{(i)} \sim p_{\boldsymbol{\theta}^*}(\mathbf{x})$.

The VAE framework [18, 24] allows us to efficiently optimize the parameters $\boldsymbol{\theta}$ of such models towards the (approximate) maximum marginal likelihood objective, such that after optimization:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x}) \tag{3}$$

In other words, after optimization we have then estimated the marginal density of $\mathbf{x}$.

2.2 Identifiability

The VAE model actually learns a full generative model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z})$ and an inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ that approximates its posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. The problem is that we generally have no guarantees about what these learned distributions actually are: all we know is that the marginal distribution over $\mathbf{x}$ is meaningful (Eq. 3). The rest of the learned distributions are, generally, quite meaningless.

What we are looking for, is models for which the following implication holds:

$$\forall(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{x}, \mathbf{z}): \quad p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \quad \Longrightarrow \quad p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}'}(\mathbf{x}, \mathbf{z}) \tag{4}$$

That is: if any two different choices of model parameter $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ lead to the same marginal density $p(\mathbf{x})$, then this would imply that they also have matching joint distributions $p(\mathbf{x}, \mathbf{z})$. This means that if we learn a parameter $\boldsymbol{\theta}$ that fits the data perfectly: $p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{x})$ (the ideal case of Eq. 3), then its joint density also matches perfectly: $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z})$. If the joint density matches, this also means that we found the correct prior $p_{\boldsymbol{\theta}}(\mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{z})$ and correct posteriors $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$. In case of VAEs, we can then also use the inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to efficiently perform inference over the sources $\mathbf{z}^*$ from which the data originates.

The general problem here is a lack of *identifiability* guarantees of the deep latent-variable model. We illustrate this by showing that any model with unconditional latent distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$ is unidentifiable, i.e. that Eq. (4) does not hold. In this case, we can always find transformations of $\mathbf{z}$ that changes its value but does not change its distribution. For a spherical Gaussian distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$, for example, applying a rotation keeps its distribution the same. We can then incorporate this transformation as the first operation in $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. This will not change $p_{\boldsymbol{\theta}}(\mathbf{x})$, but it will change $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, since now the values of $\mathbf{x}$ come from different values of $\mathbf{z}$. This is an example of a broad class of commonly used models that are non-identifiable. We show rigorously in Supplementary Material F that, in fact, models with *any* form of unconditional prior $p_{\boldsymbol{\theta}}(\mathbf{z})$ are unidentifiable.

3 An identifiable model based on conditionally factorial priors

In this section, we define a broad family of deep latent-variable models which is identifiable, and we show how to estimate the model and its posterior through the VAE framework. We call this family of models, together with its estimation method, Identifiable VAE, or iVAE for short.

3.1 Definition of proposed model

The primary assumption leading to identifiability is a conditionally factorized prior distribution over the latent variables $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$, where $\mathbf{u}$ is an additionally observed variable [13]. The variable $\mathbf{u}$ could be, for example, the time index in a time series [10], previous datapoints in a time series, some kind of (possibly noisy) class label [13], or another concurrently observed variable.

Formally, let $\mathbf{x} \in \mathbb{R}^d$, and $\mathbf{u} \in \mathbb{R}^m$ be two observed random variables, and $\mathbf{z} \in \mathbb{R}^n$ (lower-dimensional, $n \leq d$) a latent variable. Let $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ be the parameters of the following conditional generative model:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_{\mathbf{f}}(\mathbf{x}|\mathbf{z})p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) \tag{5}$$

where we first define:

$$p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) = p_{\boldsymbol{\varepsilon}}(\mathbf{x} - \mathbf{f}(\mathbf{z})) \tag{6}$$

which means that the value of $\mathbf{x}$ can be decomposed as $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is an independent noise variable with probability density function $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$, i.e. $\boldsymbol{\varepsilon}$ is independent of $\mathbf{z}$ or $\mathbf{f}$. We assume that the function $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ is bijective; but apart from bijectivity it can be an arbitrarily complicated nonlinear function. For the sake of analysis we treat the function $\mathbf{f}$ itself as a parameter of the model; however in practice we can use flexible function approximators such as neural networks.

We describe the model above with noisy and continuous-valued observations $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}$. However, our identifiability results also apply to non-noisy and discrete observations. Non-noisy observations $\mathbf{x} = \mathbf{f}(\mathbf{z})$ are a special case of Eq. (6) where $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon})$ is Gaussian with infinitesimal variance. Likewise, discrete random variables can viewed as a special case of continuous random variables in the infinitesimal-temperature limit [21, 14]: This case is explained in Supplementary Material E. For these reasons, we can use discrete observations or flow-based generative models [6] for $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, while maintaining identifiability.

The prior on the latent variables $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is assumed to be *conditionally* factorial, where each element of $z_i \in \mathbf{z}$ has a univariate exponential family distribution given conditioning variable $\mathbf{u}$. The conditioning on $\mathbf{u}$ is through an arbitrary function $\boldsymbol{\lambda}(\mathbf{u})$ (such as a look-up table or neural network) that outputs the individual exponential family parameters $\lambda_{i,j}$. The probability density function is thus given by:

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) = \prod_{i=1}^{n} p_i(z_i|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\mathbf{u})} \exp\left[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(\mathbf{u})\right] \tag{7}$$

where Q_i is the base measure, $Z_i(\mathbf{u})$ is the normalizing constant and $T_{i,j}$ are the components of the sufficient statistic and $\lambda_{i,j}(\mathbf{u})$ the corresponding parameters, crucially depending on $\mathbf{u}$. Finally, k, the number of components within each exponential family, is fixed (not estimated). Note that exponential families have universal approximation capabilities, so this assumption is not very restrictive [25].

3.2 Estimation by VAE

Next we consider a practical estimation method for the proposed model. Consider we have a dataset $\mathcal{D} = \left\{\left(\mathbf{x}^{(1)}, \mathbf{u}^{(1)}\right), \ldots, \left(\mathbf{x}^{(N)}, \mathbf{u}^{(N)}\right)\right\}$ of observations generated according to the generative model defined in Eq. (5). We propose to use a VAE as a means of learning the true generating parameters $\boldsymbol{\theta}^* := (\mathbf{f}^*, \mathbf{T}^*, \boldsymbol{\lambda}^*)$, up to some indeterminacies.

VAEs are a framework that simultaneously learns a deep latent generative model and a variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ of its true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$, the latter being often intractable. Denote by $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, |\mathbf{u})\mathrm{d}\mathbf{z}$ the conditional marginal distribution of the observations, and with $q_{\mathcal{D}}(\mathbf{x}, \mathbf{u})$ we denote the empirical data distribution given by dataset $\mathcal{D}$. VAEs learn the vector of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ by maximizing $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$, a lower bound on the data log-likelihood defined by:

$$\mathbb{E}_{q_{\mathcal{D}}(\mathbf{x},\mathbf{u})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})\right] \geq \mathbb{E}_{q_{\mathcal{D}}(\mathbf{x},\mathbf{u})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})\right]\right] := \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) \tag{8}$$

We use the reparameterization trick [18] to sample from $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$. This trick provides a low-variance stochastic estimator for gradients of the lower bound with respect to $\boldsymbol{\phi}$.

3.3 Identifiability and consistency results

As discussed in section 2.2, identifiability as defined by equation (4) is very hard to achieve in deep latent variable models. As a first step towards an identifiable model, we seek to recover the model parameters or the latent variables up to trivial transformations. Here, we state informally our results on this weaker form of identifiability of the model and the consistency of its estimation by VAE—a rigorous treatment is given in Section 4. Consider for simplicity the case of no noise. Then we recover $\mathbf{z}$ which are related to the original $\mathbf{z}^*$ as follows:

$$(T^*_{1,1}(z^*_n), \ldots, T^*_{n,k}(z^*_n))^T = A(T_{1,1}(z_1), \ldots, T_{n,k}(z_n))^T \tag{9}$$

That is, we can recover the original latent variables up to a component-wise (point-wise) transformations $T^*_{i,j}, T_{i,j}$, which are defined as the sufficient statistics of exponential families, and up to a subsequent linear transformation A. Importantly, the linear transformation A can often be resolved by linear ICA methods [10, 22], since the component-wise transformations of the latents are still independent, and typically non-Gaussian. In fact, the linear indeterminacy disappears in some cases, as explained below. Thus, the only real indeterminacy is often the component-wise transformations of the latents, which may be inconsequential in many applications.

3.4 Interpretation as nonlinear ICA

Now we show how the model above is closely related to previous work on nonlinear ICA. In nonlinear ICA, we assume observations $\mathbf{x} \in \mathbb{R}^d$, which are the result of an unknown (but invertible) transformation $\mathbf{f}$ of latent variables $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) \tag{10}$$

where $\mathbf{z}$ are assumed to follow a factorized (but typically unknown) distribution $p(\mathbf{z}) = \prod_{i=1}^{d} p_i(z_i)$. The original $\mathbf{z}$ are also called sources. This model is essentially a deep generative model. The difference to the definition above is mainly in the lack of noise and the equality of the dimensions: The transformation $\mathbf{f}$ is deterministic and invertible. Thus, any posteriors would be degenerate.

The goal is then to recover (identify) $\mathbf{f}^{-1}$, which gives the independent components as $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$, based on a dataset of observations of $\mathbf{x}$ alone. Thus, the goal of nonlinear ICA was always identifiability, which is in general not attained by deep latent variable models, as was discussed in Section 2 above.

To obtain identifiability, we either have to restrict $\mathbf{f}$ (for instance make it linear) and/or we have to introduce some additional constraints on the distribution of the sources $\mathbf{z}$. Recently, three new nonlinear ICA frameworks [10, 11, 13] exploring the latter direction were proposed, in which it is possible to recover identifiable sources, up to some trivial transformations.

The framework in [13] is particularly close to what we proposed above. However, there are several important differences. First, here we define a generative model where posteriors are non-degenerate, which allows us to show an explicit connection to VAE. In fact, we are thus also able to perform maximum likelihood estimation (in terms of evidence lower bound), while previous nonlinear ICA used more heuristic self-supervised schemes. Computing a lower bound on the likelihood is useful, for example, for model selection and validation. We also learn both the forward and backward models, which allows for recovering independent latents from data, but also generating new data. The forward model is also likely to help interrogate the meaning of the latents. At the same time, we are able to provide identifiability results which apply for more general models than earlier theory, and in particular considers the case where the number of latent variables is smaller than the number of observed variables and is corrupted by noise. Given the popularity of VAEs, our current framework should thus be of interest.

4 Identifiability theory

Now we give our main technical results. The proofs are in Supplementary Material.

Notations Let $\mathcal{Z} = \mathcal{Z}_i \times \cdots \times \mathcal{Z}_n$ and $\mathcal{X}$ be the domain and codomain of $\mathbf{f}$ in (6) respectively, and $\mathcal{U}$ the support of the distribution of $\mathbf{u}$. We denote by $\tilde{\mathbf{T}}(\mathbf{z}) = (T_{1,1}(z_1) \ldots, T_{n,k}(z_n)) \in \mathbb{R}^{nk}$ the vector of sufficient statistics of (7), $\mathbf{T}(\mathbf{z}) = (Q_1(z_1), \ldots, Q_n(z_n), T_{1,1}(z_1), \ldots, T_{n,k}(z_n)) \in \mathbb{R}^{n(k+1)}$ is $\tilde{\mathbf{T}}$ to which we append the base measures, $\boldsymbol{\lambda}(\mathbf{u}) = (Z_1(\mathbf{u}), \ldots, Z_n(\mathbf{u}), \lambda_{1,1}(\mathbf{u}), \ldots \lambda_{n,k}(\mathbf{u})) \in \mathbb{R}^{n(k+1)}$ the vector of its parameters and $\Theta = \{\boldsymbol{\theta} := (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})\}$ be the domain of parameters describing (5).

In practice, we are often interested in models that are identifiable up to a class of transformation. Thus, we introduce the following definition:

Definition 1 (Identifiability up to equivalence class) *Let $\sim$ be an equivalence relation on Θ. We say that* (1) *is identifiable up to $\sim$ (or $\sim$-identifiable) if*

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \Rightarrow \boldsymbol{\theta}' \sim \boldsymbol{\theta} \tag{11}$$

The elements of the quotient space $\Theta / \sim$ are called the identifiability classes.

Our proof will end up with the following kind of identifiability relation, so we need to verify it is an equivalence relation:

Proposition 1 *Let $\sim$ be the binary relation on Θ defined as follows:*

$$(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}') \Leftrightarrow \exists A, c \mid \tilde{\mathbf{T}}(\mathbf{f}^{-1}(\mathbf{x})) = A\tilde{\mathbf{T}}'(\mathbf{f}'^{-1}(\mathbf{x})) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{X} \tag{12}$$

where A is an invertible $nk \times nk$ matrix and $\mathbf{c}$ is a vector of size nk. Then $\sim$ is an equivalence relation on Θ.

Our main result is the following Theorem:

Theorem 1 *Assume that we observe data sampled from a generative model defined according to* (5)-(7)*, with parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$. Assume the following holds:*

(i) The set $\{\mathbf{x} \in \mathcal{X} | \varphi_\varepsilon(\mathbf{x}) = 0\}$ has measure zero, where φ_ε is the characteristic function of the density p_ε defined in (6)*.*

(ii) The sufficient statistics $T_{i,j}$ in (7) *are differentiable almost everywhere and $\frac{\mathrm{d}T_{i,j}}{\mathrm{d}z}(z) \neq 0$ almost surely for $z \in \mathcal{Z}_i$ and for all $i \in [1, \ldots, n]$ and $j \in [1, \ldots, k]$.*

(iii) There exist $nk + 1$ distinct points $\mathbf{u}^0, \ldots, \mathbf{u}^{nk}$ such that the matrix

$$L = \begin{pmatrix} \lambda_{1,1}(\mathbf{u}^1) - \lambda_{1,1}(\mathbf{u}^0) & \ldots & \lambda_{1,1}(\mathbf{u}^{nk}) - \lambda_{1,1}(\mathbf{u}^0) \\ \vdots & \ddots & \vdots \\ \lambda_{n,k}(\mathbf{u}^1) - \lambda_{n,k}(\mathbf{u}^0) & \ldots & \lambda_{n,k}(\mathbf{u}^{nk}) - \lambda_{n,k}(\mathbf{u}^0) \end{pmatrix} \tag{13}$$

of size $nk \times nk$ is invertible (where the rows correspond to all possible subscripts for λ).

then the parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ are $\sim$-identifiable.

Moreover, if there exists $(\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}')$ such that $p_{\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}'}(\mathbf{x}|\mathbf{u}) = p_{\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}}(\mathbf{x}|\mathbf{u})$, then $\mathbf{T}'$ and $\boldsymbol{\lambda}'$ verify assumptions (ii) and (iii).

This Theorem guarantees a strong identifiability of the generative model (5). In fact, suppose the data was generated according to the set of parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$. And let $(\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}')$ be the parameters obtained from some learning algorithm (supposed consistent in the limit of infinite data) that perfectly approximates the marginal distribution of the observations. Then the Theorem says that necessarily $(\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}') \sim (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$. If there were no noise, this would mean that the learned transformation $\mathbf{f}'$ transforms the observations into latents $\mathbf{z}' = \mathbf{f}'^{-1}(\mathbf{x})$ that are equal to the true generative latents $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$, up to a linear transformation (the matrix A) and point-wise nonlinearities (in the form of $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{T}}'$). With noise, we obtain the posteriors of the latents up to an analogous indeterminacy.

Next, we show how using a non-negativity constraint, we may be able to get rid of the linear indeterminacy. The basic idea is that if the sufficient statistic is squaring or absolute value, T takes only non-negative values which constrains the indeterminacies, and essentially gets rid of A in Eq. (12). We have the Theorem:

Theorem 2 *Assume the same as in Theorem 1. Furthermore, assume*

(i) $k = 1$, and the function $T_{i,1}$ is the same for all i,

(ii) $T_{i,1}$ has a unique minimum.

Then, the matrix A defining the equivalence class in Theorem 1 is a scaled permutation matrix.

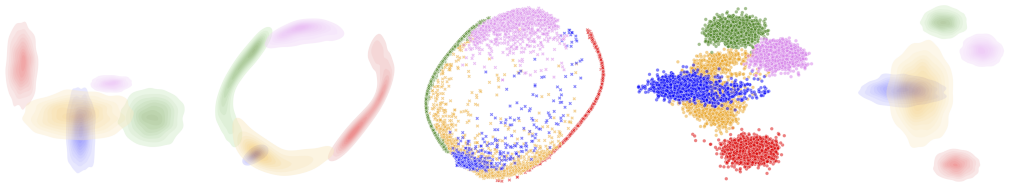
The theory above further implies a consistency result on the VAE. If the variational distribution $q_{\boldsymbol{\phi}}$ is a broad parametric family that includes the true posterior, then we have the following result.

Theorem 3 *Assume the following:*

(i) The family of distributions $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ contains $p_{\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$.

(ii) We maximize $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

then in the limit of infinite data, the VAE learns the true parameters $\boldsymbol{\theta}^* := (\mathbf{f}^*, \mathbf{T}^*, \boldsymbol{\lambda}^*)$ up to the equivalence class defined by $\sim$ in (12)*.*



(a) (b) (c) (d) (e)

Figure 1: Visualization of both observation and latent spaces in the case $n = d = 2$ and where the number of segments is $M = 5$ (segments are colour coded). First, data is generated in (a)-(c) as follows: (a) is the true distribution of the sources $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{u})$, (b) is the distribution of the data after transformation with nonlinear MLP, (c) are observations sampled from $p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z})$. Second, after learning the model, we plot in (d) the latent variables sampled from the posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ (where the points $\mathbf{x}$ are from (c), and a single $\mathbf{z}$ is sampled and plotted here); and in (e) the learned prior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$. We see that the estimated prior in (e) is very similar to the generating prior in (a) up to the basic indeterminacies of scaling, global sign, and permutation of the sources, which are even found in linear ICA.

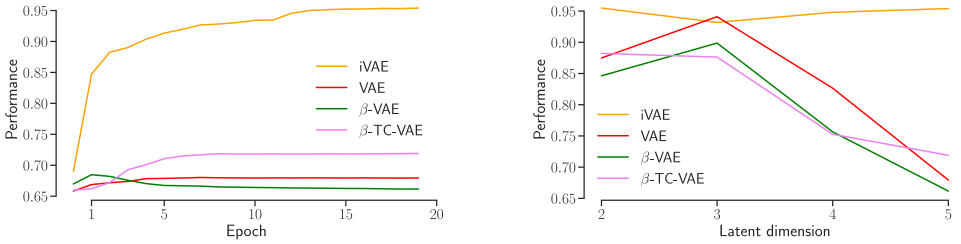
5 Simulation on artificial data

Dataset We generate synthetic datasets where the sources are non-stationary Gaussian time-series: we divide the sources into M segments of L samples each. The conditioning variable $\mathbf{u}$ is the segment label, and its distribution is uniform on the integer set $[\![1, M]\!]$. Within each segment, the conditional prior distribution is chosen from the family (7), where $k = 1$, $T_{i,1}(z_i) = z_i^2$ and $Q_i(z_i) = 1$, and the true λ_i were randomly and independently generated across the segments and the components so that the variances have a uniform distribution on $[.5, 3]$. For visualization purposes (Fig. 1 only), we also create data from an alternative conditional source distribution: $k = 2$, $Q_i(z_i) = 1$, $T_{i,1}(z_i) = z_i^2$, $T_{i,2}(z_i) = z_i$ where we additionally generate the means independently and uniformly from $[-5, 5]$. We finally mix the sources using a 3-layer multi-layer perceptron (MLP).

Choice of distributions and hyperparameters For the decoder (6), we chose $p_\varepsilon = \mathcal{N}\left(0, \sigma^2 I\right)$ a zero mean Gaussian, where the scalar σ^2 controls the noise level. As for the inference model, we let $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) = \mathcal{N}\left(\mathbf{z}|\mathbf{g}(\mathbf{x}, \mathbf{u}; \phi_{\mathbf{g}}), \mathbf{diag}\, \boldsymbol{\sigma}^2(\mathbf{x}, \mathbf{u}; \phi_{\boldsymbol{\sigma}})\right)$ be a multivariate Gaussian with a diagonal covariance. The functional parameters of the decoder and the inference model, as well as the conditional prior are chosen to be MLPs, where the dimension of the hidden layers is chosen in $\{50, 100, 200\}$, the activation function is a leaky ReLU or a leaky hyperbolic tangent, and the number of layers is chosen in $\{3, 4, 5, 6\}$. We fix the noise level $\sigma^2 = 0.01$. We chose a mini-batch of size 64, and an Adam optimizer [16] with learning rate chosen in $\{0.01, 0.001\}$ to update the parameters of the network. The objective we maximize is $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$ (Eq 8).

Performance metric To evaluate the performance of the method, we compute the mean correlation coefficient (MCC) between the original sources the corresponding latents sampled from the learned posterior. To compute this performance metric, we first calculate all pairs of correlation coefficients between source and latent components. We then solve a linear sum assignment problem to assign each latent component to the source component that best correlates with it, thus reversing any permutations in the latent space. Finally, we compute the mean of the best correlation coefficients for each component. A high MCC means that we successfully identified the true parameters and recovered the true sources, up to point-wise transformations. This is a standard measure used in ICA.

Comparisons We compared the performance of iVAE to a vanilla VAE. We used the same network architecture for both models, with the sole exception of the addition of the conditional prior in iVAE . When the data is centered, the VAE prior is Gaussian or Laplace. We also compared the performance to two models from the disentanglement literature, namely a β-VAE [9] and a β-TC-VAE [5]. The parameter β of the β-VAE was chosen in the set $[1, 45]$. The parameters α and γ for β-TC-VAE were set to 1 as suggested in the original paper, and the parameter β was chosen in the set $[1, 35]$. We trained these 4 models on the dataset described above, with $M = 40$, $L = 1000$, $d = 5$ and $n \in [2, 5]$.



(a) (b)

Figure 2: Comparison of the performance of iVAE in recovering the true sources, compared to VAE, β-VAE and β-TC-VAE, for $M = 40$, $L = 1000$ and $d = 5$. a) Evolution of the performance during training, for $n = 5$. b) Performance as a function of the latent dimension.

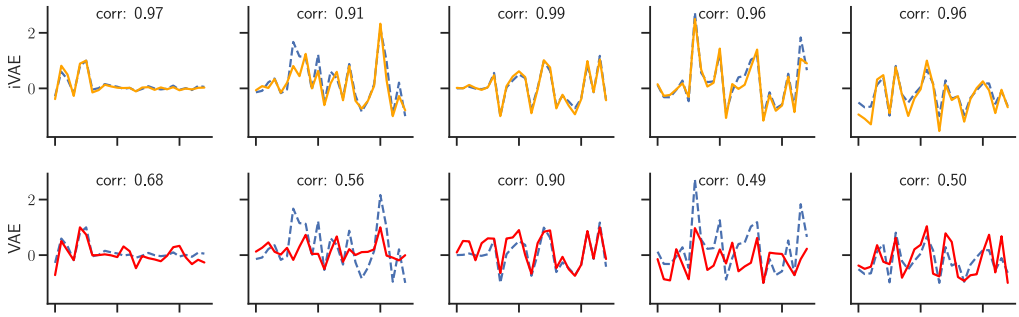


Figure 3: Comparison of the recovered latents of our model to the latents recovered by a vanilla VAE. The dashed blue line is the true source signal, and the recovered latents are in solid coloured lines. We also reported the correlation coefficients for every (source, latent) pair.

Results First, we show a visualization of identifiability of iVAE in a 2D case in Figure 1, where we plot the original sources, observed data, and the prior and posterior distributions learned by our model. The method recovers the original sources and their priors up to trivial indeterminacies. Regarding quantitative results, Figure 2a shows the best performance (obtained from an optimal choice of parameters among those considered) achieved by iVAE and the three models discussed above, when the dimension of the latent space equals the dimension of the data ($n = d = 5$). iVAE reaches a performance above 95%, whereas the other three models fail at finding a good estimation of the true parameters. We further investigated the impact of the latent dimension on the performance in Figure 2b. iVAE has much higher correlations than the three other models, especially as the dimension increases. As a further visualization we show in Figure 3 the recovered latents for VAE and iVAE ; we sampled a random (contiguous) subset of the sources from the dataset, and compared them to the recovered latents (after inverting any permutation in the components). We can see that iVAE has an excellent estimation of the original sources compared to VAE (other models were almost indistinguishable from vanilla VAE). A further comparison in higher dimensions is in Supplementary Material G.

6 Conclusion

Unsupervised learning can have many different goals, such as: (i) approximate the data distribution and (ii) generate new samples, (iii) learn useful features, and above all (iv) learn the original latent code that generated the data (identifiability). Deep latent-variable models typically implemented by VAEs are an excellent framework to achieve (i), and are thus our first building block. The nonlinear ICA model discussed in section 3.4 is the only existing framework to provably achieve (iv). We bring these two pieces together to create our new model termed iVAE . In particular, this is the first rigorous

proof of identifiability in the context of VAEs and is thus a significant contribution as it formalizes the capabilities and limitations of such models, and provides clear guidelines as to the assumptions required in order to reliably recover latent variables. Our model checks all the four boxes above that are desired in unsupervised learning.

The advantage of the new framework over typical deep latent-variable models used with VAEs is that we actually recover the original latents, thus providing principled "disentanglement". On the other hand, the advantages of this algorithm for solving nonlinear ICA over [13] are several; briefly, we obtain the likelihood and can use MLE, we learn a forward model as well and can generate new data, and we consider the more general cases of noisy data with fewer components, and even discrete data.

References

[1] Adi Ben-Israel. The Change-of-Variables Formula Using Matrix Volume. 21(1):300–312, 1999.

[2] Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ICA. *arXiv preprint arXiv:1710.05050*, 2017.

[3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. *arXiv:1509.00519 [cs, stat]*, September 2015.

[4] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

[5] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv:1802.04942 [cs, stat]*, February 2018.

[6] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear Independent Components Estimation. *arXiv:1410.8516 [cs]*, October 2014.

[7] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N. Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured Disentangled Representations. *arXiv:1804.02086 [cs, stat]*, April 2018.

[8] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv:1812.02230 [cs, stat]*, December 2018.

[9] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016.

[10] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

[11] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. Proceedings of Machine Learning Research, 2017.

[12] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999.

[13] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. *arXiv preprint arXiv:1805.08651*, 2018.

[14] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[15] Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. *arXiv:1802.05983 [cs, stat]*, February 2018.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014.

[17] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving Variational Inference with Inverse Autoregressive Flow. *arXiv:1606.04934 [cs, stat]*, June 2016.

[18] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. December 2013.

[19] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[20] Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. BIVA: A very deep hierarchy of latent variables for generative modeling. *arXiv preprint arXiv:1902.02102*, 2019.

[21] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *arXiv:1611.00712 [cs, stat]*, November 2016.

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

[23] M. D. Plumbley. Algorithms for non-negative independent component analysis. *IEEE Transactions on Neural Networks*, 14(3):534–543, 2003.

[24] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]*, January 2014.

[25] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *J. of Machine Learning Research*, 18:1–59, 2017.

[26] George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for Monte Carlo objectives. *arXiv preprint arXiv:1810.04152*, 2018.

Supplementary Material

A Proof of Proposition 1

It is clear that $\sim$ is reflexive and symmetric. Let $((\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}), (\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}'), (\mathbf{f}'', \mathbf{T}'', \boldsymbol{\lambda}'')) \in \Theta^3$, s.t. $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}')$ and $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\mathbf{f}'', \mathbf{T}'', \boldsymbol{\lambda}'')$. Then $\exists A_1, A_2$ and $\mathbf{c}_1, \mathbf{c}_2$ s.t.

$$\begin{aligned} \tilde{\mathbf{T}}(\mathbf{f}^{-1}(\mathbf{x})) &= A_1 \tilde{\mathbf{T}}'(\mathbf{f}'^{-1}(\mathbf{x})) + \mathbf{c}_1 \text{ and} \\ \tilde{\mathbf{T}}''(\mathbf{f}''^{-1}(\mathbf{x})) &= A_2 \tilde{\mathbf{T}}(\mathbf{f}^{-1}(\mathbf{x})) + \mathbf{c}_2 \\ &= A_2 A_1 \tilde{\mathbf{T}}'(\mathbf{f}'^{-1}(\mathbf{x})) + A_2 \mathbf{c}_1 + \mathbf{c}_2 \\ &= A_3 \tilde{\mathbf{T}}'(\mathbf{f}'^{-1}(\mathbf{x})) + \mathbf{c}_3 \end{aligned} \tag{14}$$

and thus $(\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}') \sim (\mathbf{f}'', \mathbf{T}'', \boldsymbol{\lambda}'')$. □

B Proof of Theorem 1

We first start by proving a useful Lemma.

Lemma 1 *Consider a univariate exponential family distribution with sufficient statistic $T(x) = (T_1(x), \dots, T_k(x))$ where $x \in \mathbb{R}$, and such that $\frac{\mathrm{d}T_i}{\mathrm{d}x}(x) \neq 0$ almost surely on $\mathbb{R}$, for all values of i. Then there exist k distinct values x_1 to x_k such that $(\frac{\mathrm{d}T}{\mathrm{d}x}(x_1), \dots, \frac{\mathrm{d}T}{\mathrm{d}x}(x_k))$ are linearly independent in $\mathbb{R}^k$.*

Proof of Lemma We proceed by contradiction. Consider arbitrary $k-1$ distinct points $x_1, \dots, x_{k-1}$, and let $x \in \mathbb{R}$. We suppose that for every choice of such points, and every x, the vectors $(\frac{\mathrm{d}T}{\mathrm{d}x}(x), \frac{\mathrm{d}T}{\mathrm{d}x}(x_1), \dots, \frac{\mathrm{d}T}{\mathrm{d}x}(x_{k-1}))$ are not linearly independent. Then there exist $\alpha_1, \dots, \alpha_k$ such that:

$$\frac{\mathrm{d}T}{\mathrm{d}x}(x) = \sum_{i=1}^{k-1} \alpha_i \frac{\mathrm{d}T}{\mathrm{d}x}(x_i) \tag{15}$$

In particular, this equality holds for every component of $\frac{\mathrm{d}T}{\mathrm{d}x}(x)$, and so we have k such equations. From the first equation, we write α_1 as a function of $\frac{\mathrm{d}T_1}{\mathrm{d}x}(x)$ (possible because the derivatives of the components are non-zero almost everywhere), and we inject this expression in the second equation. Then from the second equation, we write α_2 as a function of $\frac{\mathrm{d}T_1}{\mathrm{d}x}(x)$ and $\frac{\mathrm{d}T_2}{\mathrm{d}x}(x)$ that we then inject in the third equation. We continue in such fashion until all α_i are functions of $\frac{\mathrm{d}T_i}{\mathrm{d}x}(x)$ for $i \in [1, \dots, k-1]$. We finally inject all this into the final equation for the last component to get:

$$\frac{\mathrm{d}T_k}{\mathrm{d}x}(x) = \sum_{i=1}^{k-1} \beta_i \frac{\mathrm{d}T_i}{\mathrm{d}x}(x) \tag{16}$$

where β_i is a combination of α_j and $\frac{\mathrm{d}T_l}{\mathrm{d}x_t}(x)$ for some combination of indices (j, l, t). By integrating (16), we get:

$$T_k(x) = \sum_{i=1}^{k-1} T_i(x) + b \tag{17}$$

The k-th component of the sufficient statistic is a linear combination of the rest of the components. This can not happen in exponential families by definition, as such a term will factor with the others and the effective size of the sufficient statistic is $k-1$. We thus conclude that there exists a set of points $(x_1, \dots, x_k)$ such that $(\frac{\mathrm{d}T}{\mathrm{d}x}(x_1), \dots, \frac{\mathrm{d}T}{\mathrm{d}x}(x_k))$ are linearly independent. □

Proof of Theorem We introduce here the volume of a matrix denoted $\mathrm{vol}\, A$ as the product of the singular values of A. When A is full rank, $\mathrm{vol}\, A = \sqrt{\det A^T A}$, and when A is invertible, $\mathrm{vol}\, A = |\det A|$. The matrix volume can be used in the change of variable formula as a replacement for the absolute determinant of the Jacobian ([1]). This is most useful when the Jacobian is a

rectangular matrix ($n < d$). Suppose we have two sets of parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ and $(\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}')$ such that $p_{\mathbf{f},\mathbf{T},\boldsymbol{\lambda}}(\mathbf{x}|\mathbf{u}) = p_{\mathbf{f}',\mathbf{T}',\boldsymbol{\lambda}'}(\mathbf{x}|\mathbf{u})$ for all pairs $(\mathbf{x}, \mathbf{u})$. Then:

$$
\begin{aligned}
& \int_{\mathcal{Z}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{z} = \int_{\mathcal{Z}} p_{\mathbf{T}',\boldsymbol{\lambda}'}(\mathbf{z}|\mathbf{u}) p_{\mathbf{f}'}(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{z} \\
\Rightarrow \quad & \int_{\mathcal{Z}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) p_{\varepsilon}(\mathbf{x} - \mathbf{f}(\mathbf{z})) \mathrm{d}\mathbf{z} = \int_{\mathcal{Z}} p_{\mathbf{T}',\boldsymbol{\lambda}'}(\mathbf{z}|\mathbf{u}) p_{\varepsilon}(\mathbf{x} - \mathbf{f}'(\mathbf{z})) \mathrm{d}\mathbf{z} \\
\Rightarrow \quad & \int_{\mathcal{X}} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{f}^{-1}(\bar{\mathbf{x}})|\mathbf{u}) \operatorname{vol} J_{\mathbf{f}^{-1}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} = \\
& \qquad \int_{\mathcal{X}} p_{\mathbf{T}',\boldsymbol{\lambda}'}(\mathbf{f}'^{-1}(\bar{\mathbf{x}})|\mathbf{u}) \operatorname{vol} J_{\mathbf{f}'^{-1}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} \\
\Rightarrow \quad & \int_{\mathbb{R}^d} \tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} = \int_{\mathbb{R}^d} \tilde{p}_{\mathbf{T}',\boldsymbol{\lambda}',\mathbf{f}',\mathbf{u}}(\bar{\mathbf{x}}) p_{\varepsilon}(\mathbf{x} - \bar{\mathbf{x}}) \mathrm{d}\bar{\mathbf{x}} \\
\Rightarrow \quad & (\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}} * p_{\varepsilon})(\mathbf{x}) = (\tilde{p}_{\mathbf{T}',\boldsymbol{\lambda}',\mathbf{f}',\mathbf{u}} * p_{\varepsilon})(\mathbf{x}) \\
\Rightarrow \quad & F[\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}](\omega) \varphi_{\varepsilon}(\omega) = F[\tilde{p}_{\mathbf{T}',\boldsymbol{\lambda}',\mathbf{f}',\mathbf{u}'}](\omega) \varphi_{\varepsilon}(\omega) \\
\Rightarrow \quad & F[\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}](\omega) = F[\tilde{p}_{\mathbf{T}',\boldsymbol{\lambda}',\mathbf{f}',\mathbf{u}'}](\omega) \\
\Rightarrow \quad & \tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}(\mathbf{x}) = \tilde{p}_{\mathbf{T}',\boldsymbol{\lambda}',\mathbf{f}',\mathbf{u}}(\mathbf{x})
\end{aligned}
\tag{18}
$$

where:

- in line 3, J denotes the Jacobian, and we made the change of variable $\bar{\mathbf{x}} = \mathbf{f}(\mathbf{z})$ on the left hand side, and $\bar{\mathbf{x}} = \mathbf{f}'(\mathbf{z})$ on the right hand side.
- in line 4, we introduced
$$
\tilde{p}_{\mathbf{T},\boldsymbol{\lambda},\mathbf{f},\mathbf{u}}(\mathbf{x}) = \begin{cases} p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{f}^{-1}(\mathbf{x})|\mathbf{u}) \operatorname{vol} J_{\mathbf{f}^{-1}}(\mathbf{x}) \text{ if } \mathbf{x} \in \mathcal{X} \\ 0 \text{ instead} \end{cases} \tag{19}
$$
on the left hand side, and similarly on the right hand side.
- in line 5, we used $*$ for the convolution operator.
- in line 6, we used $F[.]$ to designate the Fourier transform, and where $\varphi_{\varepsilon} = F[p_{\varepsilon}]$ (by definition of the characteristic function).
- in line 7, we dropped $\varphi_{\varepsilon}(\omega)$ from both sides as it is non-zero almost everywhere (by assumption (i)).

This equation is valid for all $(\mathbf{x}, \mathbf{u}) \in \mathcal{X} \times \mathcal{U}$. What is basically says is that for the distributions to be the same after adding the noise, the noise-free distributions have to be the same. Note that $\mathbf{x}$ here is a general variable and we are actually dealing with the noise-free probability densities.

By taking the logarithm on both sides and replacing $p_{\mathbf{T},\boldsymbol{\lambda}}$ by its expression from (7), we get:

$$
\begin{aligned}
& \log \operatorname{vol} J_{\mathbf{f}^{-1}}(\mathbf{x}) + \sum_{i=1}^{n} \left(\log Q_i(f_i^{-1}(\mathbf{x})) - \log Z_i(\mathbf{u}) + \sum_{j=1}^{k} T_{i,j}(f_i^{-1}(\mathbf{x})) \lambda_{i,j}(\mathbf{u}) \right) = \\
& \log \operatorname{vol} J_{\mathbf{f}'^{-1}}(\mathbf{x}) + \sum_{i=1}^{n} \left(\log Q'_i(f_i'^{-1}(\mathbf{x})) - \log Z'_i(\mathbf{u}) + \sum_{j=1}^{k} T'_{i,j}(f_i'^{-1}(\mathbf{x})) \lambda'_{i,j}(\mathbf{u}) \right)
\end{aligned}
\tag{20}
$$

Let $\mathbf{u}_0, \ldots, \mathbf{u}_{nk}$ be the points provided by assumption (iii) of the Theorem. We plug each of those $\mathbf{u}_l$ in (20) to obtain $nk + 1$ such equations. We subtract the first equation for $\mathbf{u}_0$ from the remaining nk equations to get for $l = 1, \ldots, nk$:

$$
\begin{aligned}
& \sum_i \log \frac{Z_i(\mathbf{u}_0)}{Z_i(\mathbf{u}_l)} + \sum_j T_{i,j}(f_i^{-1}(\mathbf{x}))(\lambda_{i,j}(\mathbf{u}_l) - \lambda_{i,j}(\mathbf{u}_0)) = \\
& \qquad \sum_i \log \frac{Z'_i(\mathbf{u}_0)}{Z'_i(\mathbf{u}_l)} + \sum_j T'_{i,j}(f_i'^{-1}(\mathbf{x}))(\lambda'_{i,j}(\mathbf{u}_l) - \lambda'_{i,j}(\mathbf{u}_0))
\end{aligned}
\tag{21}
$$

Let L bet the matrix defined in assumption (iii), and L' similarly defined for $\boldsymbol{\lambda}'$ (L' is not necessarily invertible). Define $b_l = \sum_i \log \frac{Z_i'(\mathbf{u}_0)Z_i(\mathbf{u}_l)}{Z_i(\mathbf{u}_0)Z_i'(\mathbf{u}_l)}$ and $\mathbf{b}$ the vector of all b_l for $l = 1, \ldots, nk$. Expressing (21) for all points $\mathbf{u}_l$ in matrix form, we get:

$$L^T(\tilde{\mathbf{T}}(\mathbf{f}^{-1}(\mathbf{x}))) = L'^T(\tilde{\mathbf{T}}'(\mathbf{f}'^{-1}(\mathbf{x}))) + \mathbf{b} \tag{22}$$

We multiply both sides of (22) by the transpose of the inverse of L^T from the left to find:

$$\tilde{\mathbf{T}}(\mathbf{f}^{-1}(\mathbf{x})) = A(\tilde{\mathbf{T}}'(\mathbf{f}'^{-1}(\mathbf{x})) + \mathbf{c} \tag{23}$$

where $A = L^{-T}L'$ and $\mathbf{c} = L^{-T}\mathbf{b}$.

Now by definition of $\tilde{\mathbf{T}}$ and according to assumption (ii), its Jacobian exists and is an $nk \times n$ matrix of rank n. This implies that the Jacobian of $\tilde{\mathbf{T}}' \circ \mathbf{f}'^{-1}$ exists and is of rank n and so is A. We distinguish two cases:

- If $k = 1$, then this means that A is invertible (because A is $n \times n$).
- If $k > 1$, define $\bar{\mathbf{x}} = \mathbf{f}^{-1}(\mathbf{x})$ and $\tilde{\mathbf{T}}_i(\bar{x}_i) = (T_{i,1}(\bar{x}_i), \ldots T_{i,k}(\bar{x}_i))$. According to Lemma 1, for each $i \in [1, \ldots, n]$ there exist k points $\bar{x}_i^1, \ldots, \bar{x}_i^k$ such that $(\frac{\mathrm{d}\tilde{\mathbf{T}}_i}{\mathrm{d}x}(\bar{x}_i^1), \ldots, \frac{\mathrm{d}\tilde{\mathbf{T}}}{\mathrm{d}x\ i}(\bar{x}_i^k))$ are linearly independent. Collect those points into k vectors $(\bar{\mathbf{x}}^1, \ldots, \bar{\mathbf{x}}^k)$, and concatenate the k Jacobians $J_{\tilde{\mathbf{T}}}(\bar{\mathbf{x}}^l)$ evaluated at each of those vectors horizontally into the matrix $Q = [J_{\tilde{\mathbf{T}}}(\bar{\mathbf{x}}^1), \ldots, J_{\tilde{\mathbf{T}}}(\bar{\mathbf{x}}^k)]$ (and similarly define Q' as the concatenation of the Jacobians of $\mathbf{T}'(\mathbf{f}'^{-1} \circ \mathbf{f}(\bar{\mathbf{x}}))$ evaluated at those points). Then the matrix Q is invertible (through a combination of Lemma 1 and the fact that each component of $\tilde{T}$ is univariate). By differentiating (23) for each $\mathbf{x}^l$, we get (in matrix form):

  $$Q = AQ' \tag{24}$$

  The invertibility of Q implies the invertibility of A and Q'.

Hence, (23) and the invertibility of A mean that $(\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}') \sim (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$.

Moreover, we have the following observations:

- the invertibility of A and L imply that L' is invertible,
- because the Jacobian of $\tilde{\mathbf{T}}' \circ \mathbf{f}'^{-1}$ is full rank and $\mathbf{f}'$ is bijective (hence its Jacobian is full rank too), $J_{\tilde{\mathbf{T}}'}$ has to be full rank too, and $\frac{\mathrm{d}T_{i,j}'}{\mathrm{d}z}(z) \neq 0$ almost everywhere.
- the real equivalence class of identifiability may actually be narrower that what is defined by $\sim$, as the matrix A and the vector $\mathbf{c}$ here have very specific form, and are functions of $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$.

□

Remark: Understanding assumption (iii) in Theorem 1 Let $\mathbf{u}^0$ be an arbitrary point in its support $\mathcal{U}$, and $h(\mathbf{u}) = \big(\lambda_{1,1}(\mathbf{u}) - \lambda_{1,1}(\mathbf{u}^0), \ldots, \lambda_{n,k}(\mathbf{u}) - \lambda_{n,k}(\mathbf{u}^0)\big)^T \in \mathbb{R}^{nk}$. Saying that there exists nk distinct points $\mathbf{u}^1$ to $\mathbf{u}^{nk}$ (all different from $\mathbf{u}^0$) such that L is invertible is equivalent to saying that the vectors $\mathbf{h} := (h(\mathbf{u}^1), \ldots, h(\mathbf{u}^{nk}))$ are linearly independent in $\mathbb{R}^{nk}$. Let's suppose for a second that for any such choice of points, these vectors are not linearly independent. This means that $h(\mathcal{U})$ is necessarily included in a subspace of $\mathbb{R}^{nk}$ of dimension at most $nk - 1$. Such a subspace has measure zero in $\mathbb{R}^{nk}$. Thus, if $h(\mathcal{U})$ isn't included in a subset of measure zero in $\mathbb{R}^{nk}$, this can't be true, and there exists a set of points $\mathbf{u}^1$ to $\mathbf{u}^{nk}$ (all different from $\mathbf{u}^0$) such that L is invertible. This implies that as long as the $\lambda_{i,j}(\mathbf{u})$ are generated randomly and independently, then almost surely, $h(\mathcal{U})$ won't be included in any such subset with measure zero, and the assumption holds.

C Proof of Theorem 2

Denote for simplicity $T := T_{i,1}$. If T is squaring or absolute value, $(T(z_1, \ldots, z_n))$ would typically has non-negative quadrant as its support, and this implies, adapting the theory of non-negative ICA [23], that A has to be a scaled permutation matrix.

Formally, the proof is as follows. Subtract the minimum from T, to get a new function $\tilde{T}$, which is non-negative and has a minimum value of zero. The distribution of each $\tilde{T}(z_i)$ then has the non-negative half-line as its support. Thus the pdf of $(\tilde{T}(z_1), \tilde{T}(z_2), \ldots, \tilde{T}(z_n))$ has the non-negative quadrant as its support. Using Theorem 1 in [23] (noting that the $T(z_i)$ are "well-grounded" in the sense assumed in that theorem), we can prove that A must be a scaled permutation matrix, since only a scaled permutation matrix can map the non-negative quadrant onto itself. □

D Proof of Theorem 3

The loss (8) can be written as follows:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) \| p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})\right) \tag{25}$$

If the family $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ is large enough to include $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})$, then by optimizing the loss over its parameter $\boldsymbol{\phi}$, we will minimize the KL term, eventually reaching zero, and the loss will be equal to the log-likelihood. The VAE in this case inherits all the properties of maximum likelihood estimation. In particular, it is a consistent estimator of the equivalence class (Th 1) of true parameter $\boldsymbol{\theta}^*$ *i.e.* in the limit of infinite data, the estimated parameter $\boldsymbol{\theta} \sim \boldsymbol{\theta}^*$. □

E Discrete observations

As explained in [21, 14], categorical distributions can be viewed as a infinitesimal-temperature limit of continuous distributions. We can use this fact to extend our theory to discrete latent variables.

For example, let:

$$\mathbf{m} = \mathbf{f}(\mathbf{z}) \tag{26}$$

$$\mathbf{x} = \text{sigmoid}((\mathbf{m} + \boldsymbol{\varepsilon})/T) \tag{27}$$

$$\forall \varepsilon_i \in \boldsymbol{\varepsilon} : \varepsilon_i \sim \text{Logistic}(0, 1) \tag{28}$$

where sigmoid() is the element-wise sigmoid nonlinearity, and $T \in (0, \infty)$ is a temperature variable.

If we let T approach 0 from above, then:

$$\mathbf{x} \sim \text{Bernoulli}(\mathbf{p}) \;\text{ with }\; \mathbf{p} = \text{sigmoid}(\mathbf{m}) \tag{29}$$

For proof that this holds, we refer to [21], appendix B.

The sigmoid$(\cdot/T)$ function is invertible, and the Logistic distribution has a probability density function that allows for deconvolution since its Fourier transform is non zero almost everywhere. As a result, for a given value of T, the distribution $p(\mathbf{x})$ has a one-to-one mapping to a distribution $p(\mathbf{m})$. This means that we can apply a small change to (18) and arrive at the same identifiability result. This example with a Bernoulli distribution can be extended to a categorical distribution with any number of components [21, 14].

F Unidentifiability of generative models with factorial prior

In this section, we present two well-known proofs of unidentifiability of generative models. The first proof is simpler and considers factorial priors, which are widely-used in deep generative models and the VAE literature. The second proof is extremely general, and shows how any random vector can be transformed into independent components, in particular components which are standardized Gaussian. Thus, we see how in the general nonlinear case, there is little hope of finding the original latent variables based on the (unconditional, marginal) statistics of $\mathbf{x}$ alone.

F.1 Factorial priors

Let us start with factorial, gaussian priors. In other words, let $\mathbf{z} \sim p_{\boldsymbol{\theta}}(\mathbf{z}) = N(\mathbf{0}, \mathbf{I})$. Now, a well-known result says that any orthogonal transformation of $\mathbf{z}$ has exactly the same distribution. Thus, we could transform the latent variable by any orthogonal transformation $\mathbf{z}' = M\mathbf{z}$, and cancel that transformation in $p(\mathbf{x}|\mathbf{z})$ (e.g. in the first layer of the neural network), and we would get exactly the same observed data (and thus obviously the same distribution of observed data) with $\mathbf{z}'$.

Formally we have

$$p_{\mathbf{z}'}(\boldsymbol{\xi}) = p_{\mathbf{z}}(M^T\boldsymbol{\xi})|\det M| = \frac{1}{(2\pi)^{d/2}}\exp(-\frac{1}{2}\|M^T\boldsymbol{\xi}\|^2) = \frac{1}{(2\pi)^{d/2}}\exp(-\frac{1}{2}\|\boldsymbol{\xi}\|^2) = p_{\mathbf{z}}(\boldsymbol{\xi}) \tag{30}$$

where we have used the fact that the determinant of an orthogonal matrix is equal to unity.

This result applies easily to any factorial prior. For z_i of any distribution, we can transform it to a uniform distribution by $F_i(z_i)$ where F_i is the cumulative distribution function of z_i. Next, we can transform it into standardized Gaussian by $\Phi^{-1}(F_i(z_i))$ where Φ is the standardized Gaussian cdf. After this transformation, we can again take any orthogonal transformation without changing the distribution. And we can even transform back to the same marginal distributions by $F_i^{-1}(\Phi(.))$. Thus, the original latents are not identifiable.

F.2 General priors

The second proof comes from the theory of nonlinear ICA [12], from which the following Theorem is adapted.

Theorem 4 ([12]) *Let $\mathbf{z}$ be a d-dimensional random vector of any distribution. Then there exists a transformation $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ such that the components of $\mathbf{z}' := \mathbf{g}(\mathbf{z})$ are independent, and each component has a standardized gaussian distribution. In particular, z'_1 equals a monotonic transformation of z_1.*

The proof is based on an iterative procedure reminiscent of Gram-Schmidt, where a new variable can always be transformed to be independent of any previously considered variables, which is why z_1 is essentially unchanged.

This theorem means that there are infinitely many ways of defining independent components $\mathbf{z}$ that nonlinearly generated an observation $\mathbf{x}$. This is because we can first transform $\mathbf{z}$ any way we like and then apply the Theorem. The arbitrariness of the components is seen in the fact that we will always find that one arbitrary chosen variable in the transformation is one of the independent components. This is in some sense an alternative kind of indeterminacy to the one in the previous subsection.

In particular, we can even apply this theorem on the observed data, taking $\mathbf{x}$ instead of $\mathbf{z}$. Then, in the case of factorial priors, just permuting the data variables, we would arrive at the conclusion that any of the x_i can be taken to be one of the independent components, which is absurd.

Now, to apply this theory in the case of a general prior on $\mathbf{z}$, it is enough to point out that we can transform any variable into independent Gaussian variables, apply any orthogonal transformation, then invert the transformation in the theorem, and we get a nonlinear transformation $\mathbf{z}' = \mathbf{g}^{-1}(M\mathbf{g}(\mathbf{z}))$ which has exactly the same distribution as $\mathbf{z}$ but is a complex nonlinear transformation. Thus, no matter what the prior may be, by looking at the data alone, it is not possible to recover the true latents based an unconditional prior distribution, in the general nonlinear case.

G Visualisations

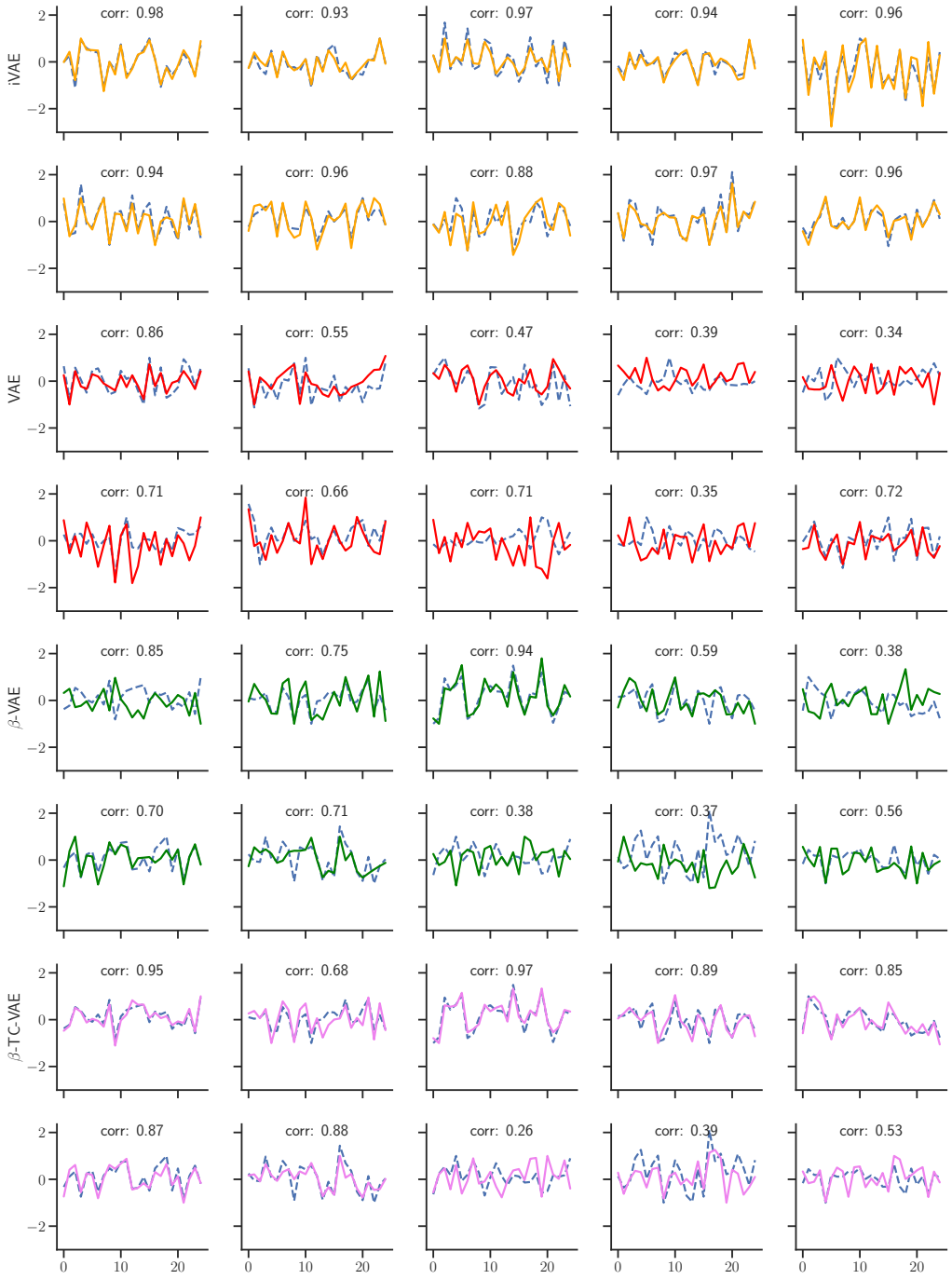


Figure 4: Comparison of the recovered latents of our model to the latents recovered by a vanilla VAE, a β-VAE and a β-TC-VAE, where the dimension of the data is $d = 40$, and the dimension of the latents is $n = 10$, the number of segments is $M = 40$ and the number of samples per segment is $L = 4000$. The dashed blue line is the true source signal, and the recovered latents are in solid coloured lines. We reported the correlation coefficients for every (source, latent) pair. We can see that iVAE have an excellent estimation of the original sources compared to the other models.