Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem
Gatsby Unit
UCL

Diederik P. Kingma
Google Brain

Ricardo Pio Monti
Gatsby Unit
UCL

Aapo Hyvärinen
Université Paris-Saclay,
Inria, Univ. of Helsinki

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 very simple transformations, 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 (Kingma and Welling, 2013; Rezende et al., 2014) (VAEs) and its extensions (e.g. Burda et al. (2015); Kingma et al. (2016); Tucker et al. (2018); Maaløe et al. (2019)) 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.

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 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 offers 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 (Burgess et al., 2018; Higgins et al., 2016, 2018; Esmaeili et al., 2018; Kim and Mnih, 2018; Chen et al., 2018), in which the authors introduce adjustable hyperparameters in the VAE objective to encourage disentanglement. Other work attempts to find maximally independent compo-

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

nents through the GAN framework (Brakel and Bengio, 2017). However, models in these earlier works are actually non-identifiable due to non-conditional latent priors, as has been seen empirically (Locatello et al., 2018), and we will show formally below.

Recent work in nonlinear Independent Component Analysis (ICA) theory (Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019) 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 (Hyvärinen and Pajunen, 1999). 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 $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, often parameterized with a neural network 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 $\mathbf{x}$:

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

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 (Kingma and Welling, 2013; Rezende et al., 2014) 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 Parameter Space vs Function Space

In this work we use slightly non-standard notation and nomenclature: we use $\boldsymbol{\theta} \in \Theta$ to refer to the model parameters in *function space*. In contrast, let $\mathbf{w} \in W$ refer to the space of original neural network parameters (weights, biases, etc.) in which we usually perform gradient ascent.

2.3 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 for all $(\mathbf{x}, \mathbf{z})$:

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

That is: if any two different choices of model parameter $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ lead to the same marginal density $p_{\boldsymbol{\theta}}(\mathbf{x})$, then this would imply that they are equal and thus have matching joint distributions $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$. 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 D 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 (Hyvärinen et al., 2019). The variable $\mathbf{u}$ could be, for example, the time index in a time series (Hyvärinen and Morioka, 2016), previous data points in a time series, some kind of (possibly noisy) class label, 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 $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^d$ is injective; but apart from injectivity 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 be viewed as a special case of continuous random variables in the infinitesimal-temperature limit (Maddison et al., 2016; Jang et al., 2016). This case is explained in Supplementary Material C. For these reasons, we can use discrete observations or flow-based generative models (Dinh et al., 2014) 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:

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

where Q_i is the base measure, $Z_i(\mathbf{u})$ is the normalizing constant and $\mathbf{T}_i = (T_{i,1}, \ldots, T_{i,k})$ are the sufficient statistics and $\boldsymbol{\lambda}_i(\mathbf{u}) = (\lambda_{i,1}(\mathbf{u}), \ldots, \lambda_{i,k}(\mathbf{u}))$ the corresponding parameters, crucially depending on $\mathbf{u}$. Finally, k, the dimension of each sufficient statistic, is fixed (not estimated). Note that exponential families have universal approximation capabilities, so this assumption is not very restrictive (Sriperumbudur et al., 2017).

3.2 Estimation by VAE

Next we propose a practical estimation method for the model introduced above. 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 the indeterminacies discussed below.

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:

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

We use the reparameterization trick (Kingma and Welling, 2013) 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 $\boldsymbol{\phi}$. The training algorithm is the same as in a regular VAE. Estimates of the latent variables can be obtained by sampling from the variational posterior.

VAEs, like any maximum likelihood estimation method, requires the densities to be normalized. To this end, in practice we choose the prior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ to be a Gaussian location-scale family, which is widely used with VAEs.[1]

3.3 Identifiability and consistency results

As discussed in section 2.3, 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 —a rigorous treatment is given in Section 4. Consider for simplicity the case of no noise and sufficient statistics of size $k = 1$, and define $T_i := T_{i,1}$. Then we can recover $\mathbf{z}$ which are related to the original $\mathbf{z}^*$ as follows:

$$(T_1^*(z_1^*), \ldots, T_n^*(z_n^*)) = A(T_1(z_1), \ldots, T_n(z_n)) \quad (9)$$

for an invertible matrix A. That is, we can recover the original latent variables up to a component-wise (pointwise) transformations T_i^*, T_i, 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 excluding families where, roughly speaking, only the location (mean) is changing. Then A is simply a permutation matrix, and equation (9) becomes

$$T_i^*(z_i^*) = T_{i'}(z_{i'}) \quad (10)$$

for a permuted index i'. 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 $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) \quad (11)$$

where $\mathbf{z}$ are assumed to follow a factorized (but typically unknown) distribution $p(\mathbf{z}) = \prod_{i=1}^d p_i(z_i)$. 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 (Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019) exploring the latter direction were proposed, in which it is possible to recover identifiable sources, up to some trivial transformations.

The framework in Hyvärinen et al. (2019) 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.

[1] As mentioned in section 3.1, our model contains normalizing flows as a special case when $\mathrm{Var}(\varepsilon) = 0$ and the mixing function $\mathbf{f}$ is parameterized as an invertible flow (Rezende and Mohamed, 2015). Thus, as an alternative estimation method, we could then optimize the log-likelihood directly: $\mathbb{E}_{q_{\mathcal{D}}(\mathbf{x},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{u})] = \log p_\theta(\mathbf{f}^{-1}(\mathbf{z})|\mathbf{u}) + \log|J_{\mathbf{f}^{-1}}(\mathbf{x})|$ where $J_{\mathbf{f}^{-1}}$ is easily computable. The conclusion on consistency given in section 4.3 still holds in this case.

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. In addition, we can in fact prove a tight link between maximum likelihood estimation and maximization of independence of latents, as discussed in Supplementary Material F. 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 investigate the meaning of the latents. At the same time, we are able to provide stronger 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. Further discussion can be found in Supplementary Material G.

4 IDENTIFIABILITY THEORY

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

Notations Let $\mathcal{Z} \subset \mathbb{R}^n$ and $\mathcal{X} \subset \mathbb{R}^d$ be the domain and the image of $\mathbf{f}$ in (6), respectively, and $\mathcal{U} \subset \mathbb{R}^m$ the support of the distribution of $\mathbf{u}$. We denote by $\mathbf{f}^{-1}$ the inverse defined from $\mathcal{X} \to \mathcal{Z}$. We suppose that $\mathcal{Z}$, $\mathcal{X}$ and $\mathcal{U}$ are open sets. We denote by $\mathbf{T}(\mathbf{z}) := (\mathbf{T}_1(z_1), \dots, \mathbf{T}_n(z_n)) = (T_{1,1}(z_1) \dots, T_{n,k}(z_n)) \in \mathbb{R}^{nk}$ the vector of sufficient statistics of (7), $\boldsymbol{\lambda}(\mathbf{u}) = (\boldsymbol{\lambda}_1(\mathbf{u}), \dots, \boldsymbol{\lambda}_n(\mathbf{u})) = (\lambda_{1,1}(\mathbf{u}), \dots, \lambda_{n,k}(\mathbf{u})) \in \mathbb{R}^{nk}$ the vector of its parameters. Finally $\Theta = \{\boldsymbol{\theta} := (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})\}$ is the domain of parameters describing (5).

4.1 General results

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

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}) \implies \tilde{\boldsymbol{\theta}} \sim \boldsymbol{\theta} \tag{12}$$

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

We now define two equivalence relations on the set of parameters Θ.

Definition 2 *Let $\sim$ be the equivalence relation on Θ defined as follows:*

$$(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}) \Leftrightarrow$$
$$\exists A, \mathbf{c} \mid \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = A\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{X} \tag{13}$$

where A is an $nk \times nk$ matrix and $\mathbf{c}$ is a vector

If A is invertible*, we denote this relation by $\sim_A$. If A is a block* permutation[2] *matrix, we denote it by $\sim_P$.*

Our main result is the following Theorem[3]:

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 mixing function $\mathbf{f}$ in (6) is injective.

(iii) The sufficient statistics $T_{i,j}$ in (7) are differentiable almost everywhere, and $(T_{i,j})_{1 \le j \le k}$ are linearly independent on any subset of $\mathcal{X}$ of measure greater than zero.

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

$$L = (\boldsymbol{\lambda}(\mathbf{u}_1) - \boldsymbol{\lambda}(\mathbf{u}_0), \dots, \boldsymbol{\lambda}(\mathbf{u}_{nk}) - \boldsymbol{\lambda}(\mathbf{u}_0)) \tag{14}$$

of size $nk \times nk$ is invertible.[4]

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

This Theorem guarantees a basic form of 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 $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$ 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 $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}) \sim_A (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$. If there were no noise, this would mean that the learned transformation $\tilde{\mathbf{f}}$ transforms the observations into latents $\tilde{\mathbf{z}} = \tilde{\mathbf{f}}^{-1}(\mathbf{x})$ that are equal to the true generative latents $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$, up to a linear invertible transformation (the matrix A) and point-wise nonlinearities (in the form of $\mathbf{T}$ and $\tilde{\mathbf{T}}$). With noise, we obtain the posteriors of the latents up to an analogous indeterminacy.

[2] each block linearly transforms $\mathbf{T}_i$ into $\tilde{\mathbf{T}}_{i'}$.

[3] an alternative version is in Supplementary Material E.

[4] the intuition and feasibility of this assumption are discussed in Supplementary Material B.2.3.

4.2 Characterization of the linear indeterminacy

The equivalence relation $\sim_A$ provides a useful form of identifiability, but it is very desirable to remove the linear indeterminacy A, and reduce the equivalence relation to $\sim_P$ by analogy with linear ICA where such matrix is resolved up to a *permutation* and *signed scaling*. We present in this section sufficient conditions for such reduction, and special cases to avoid.

We will start by giving two Theorems that provide sufficient conditions. Theorem 2 deals with the more general case $k \geq 2$, while Theorem 3 deals with the special case $k = 1$.

Theorem 2 ($k \geq 2$) *Assume the hypotheses of Theorem 1 hold, and that $k \geq 2$. Further assume:*

(2.i) The sufficient statistics $T_{i,j}$ in (7) are twice differentiable.

(2.ii) The mixing function $\mathbf{f}$ has all second order cross derivatives.

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

Theorem 3 ($k = 1$) *Assume the hypotheses of Theorem 1 hold, and that $k = 1$. Further assume:*

(3.i) The sufficient statistics $T_{i,1}$ are not monotonic[5].

(3.ii) All partial derivatives of $\mathbf{f}$ are continuous.

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

These two Theorems imply that in most cases $\tilde{\mathbf{f}}^{-1} \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z}$ is a pointwise [6] nonlinearity, which essentially means that the estimated latent variables $\tilde{\mathbf{z}}$ are equal to a permutation and a pointwise nonlinearity of the original latents $\mathbf{z}$.

This kind of identifiability is stronger than any previous results in the literature, and considered sufficient in many applications. On the other hand, there are very special cases where a linear indeterminacy cannot be resolved, as shown by the following:

Proposition 1 *Assume that $k = 1$, and that*

(i) $T_{i,1}(z_i) = z_i$ for all i.

(ii) $Q_i(z_i) = 1$ or $Q_i(z_i) = e^{-z_i^2}$ for all i.

Then A can not be reduced to a permutation matrix.

[5] monotonic means it is strictly increasing or decreasing.
[6] each of its component is a function of only one z_i.

This Proposition stipulates that if the components are Gaussian (or exponential in the case of non-negative components) and *only* the location is changing, we can't hope to reduce the matrix A in $\sim_A$ to a permutation. In fact, to prove this in the Gaussian case, we simply consider orthogonal transformations of the latent variables, which all give rise to the same observational distribution with a simple adjustment of parameters.

4.3 Consistency of Estimation

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 4 *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 $\boldsymbol{\phi}$.

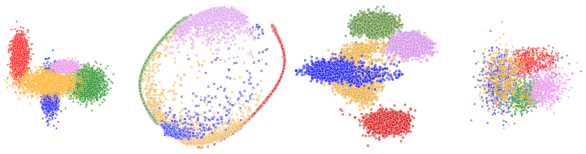
then in the limit of infinite data, the VAE learns the true parameters $\boldsymbol{\theta}^* := (\mathbf{f}^*, \mathbf{T}^*, \boldsymbol{\lambda}^*)$ up to the equivalence class defined by $\sim$ in (13).

5 EXPERIMENTS

5.1 Simulations on artifical data

Dataset We run simulations on data used previously in the nonlinear ICA literature (Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019). 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) for small k. When $k = 2$, we used mean and variance modulated Gaussian distribution. When $k = 1$, we used variance modulated Gaussian or Laplace (to fall within the hypotheses of Theorem 3). The true parameters λ_i were randomly and independently generated across the segments and the components from a non degenerate distributions to satisfy assumption (iv) of Theorem 1. Following Hyvärinen et al. (2019), we mix the sources using a multi-layer perceptron (MLP) and add small Gaussian noise.

Model specification Our estimates of the latent variables are generated from the variational posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{u}, \mathbf{x})$, for which we chose the following form: $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) = \mathcal{N}\left(\mathbf{z}|\mathbf{g}(\mathbf{x}, \mathbf{u}; \phi_{\mathbf{g}}), \mathbf{diag}\, \boldsymbol{\sigma}^2(\mathbf{x}, \mathbf{u}; \phi_{\boldsymbol{\sigma}})\right)$, a multivariate Gaussian with a diagonal covariance. The



(a) $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{u})$ (b) $p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{u})$ (c) $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x},\mathbf{u})$ (d) $p_{\text{VAE}}(\mathbf{z}|\mathbf{x})$

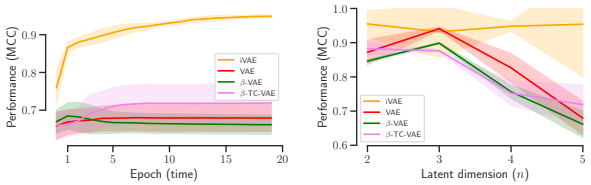
Figure 1: Visualization of both observation and latent spaces in the case $n = d = 2$ and where the number of segments is $M = 5$ (segments are colour coded). First, data is generated in (a)-(b) as follows: (a) samples from the true distribution of the sources $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{u})$: Gaussian with non stationary mean and variance, (b) are observations sampled from $p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z})$. Second, after learning both a vanilla VAE and an iVAE models, we plot in (c) the latent variables sampled from the posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})$ of the iVAE and in (d) the latent variables sampled from the posterior of the vanilla VAE.

noise distribution p_ε is Gaussian with small variance. The functional parameters of the decoder and the inference model, as well as the conditional prior are chosen to be MLPs. We use an Adam optimizer (Kingma and Ba, 2014) to update the parameters of the network by maximizing $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi})$ in equation (8). The data generation process as well as hyperparameter choices are detailed in Supplementary Material H.1.

Performance metric To evaluate the performance of the method, we compute the mean correlation coefficient (MCC) between the original sources and 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. 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.

Results: 2D example 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 posterior distributions learned by our model, compared to a vanilla VAE. Our method recovers the original sources up to trivial indeterminacies (rotation and sign flip), whereas the VAE fails to do a good separation of the latent variables.

Results: Comparison to VAE variants We compared the performance of iVAE to a vanilla VAE. We used the same network architecture for both models,



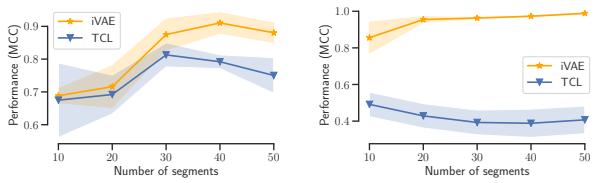
(a) Training dynamics (b) Changing n

Figure 2: Performance of iVAE in recovering the true sources, compared to VAE, β-VAE and β-TC-VAE, for $M = 40$, $L = 1000$ and $d = 5$ (and $n = 5$ for (a)).

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 (Higgins et al., 2016) and a β-TC-VAE (Chen et al., 2018). The parameter β of the β-VAE and the parameters α, β and γ for β-TC-VAE were chosen by following the instructions of their respective authors. We trained these 4 models on the dataset described above, with $M = 40$, $L = 1000$, $d = 5$ and $n \in [2, 5]$. Figure 2a compares performances obtained from an optimal choice of parameters 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 achieved an MCC score of 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. Further visualization are in Supplementary Material I.4.

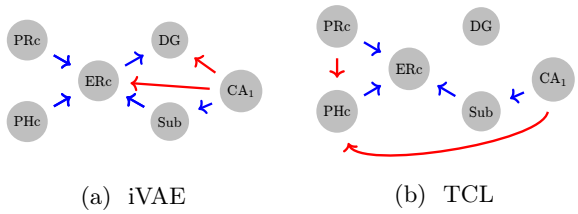
Results: Comparison to TCL Next, we compared our method to previous nonlinear ICA methods, namely TCL by Hyvärinen and Morioka (2016), which is based on a self supervised classification task (see Supplementary Material G.1). We run simulations on the same dataset as Figure 2a, where we varied the number of segments from 10 to 50. Our method slightly outperformed TCL in our experiments. The results are reported in Figure 3a. Note that according to Hyvärinen et al. (2019), TCL performs best among previously proposed methods for this kind of data.

Finally, we wanted to show that our method is robust to some failure modes which occur in the context of self-supervised methods. The theory of TCL is premised on the notion that in order to accurately classify observations into their relative segments, the model must learn the true log-densities of sources within each segment. While such theory will hold in the limit of infinite data, we considered here a special case where accurate classi-



(a) Normal (b) Easy to classify

Figure 3: (*a*) Performance of iVAE in comparison to TCL in recovering the true sources on normal data (*b*) Performance of iVAE in comparison to TCL in recovering the true sources on easy to classify data.



(a) iVAE (b) TCL

Figure 4: Estimated causal graph on hippocampal fMRI data unmixing of sources is achieved via iVAE (left) or TCL (right). Blue edges are feasible given anatomical connectivity, red edges are not.

fication did not require learning the log-densities very precisely. This was achieved by generating synthetic data where x_2 alone contained sufficient information to perform classification, by making the mean of x_2 significantly modulated across segments; further details in Supplementary Material H.2. In such a setting, TCL is able to obtain high classification accuracy without unmixing observations, resulting in its failure to recover latent variables as reflected in Figure 3b. In contrast, the proposed iVAE, by virtue of optimizing a maximum likelihood objective, does not suffer from such degenerate behaviour.

Further simulations on hyperparameter selection and discrete data are in Supplementary Material I.

5.2 Nonlinear causal discovery in fMRI

An important application of ICA methods is within the domain of causal discovery (Peters et al., 2017). The use of ICA methods in this domain is premised on the equivalence between a (nonlinear) ICA model and the corresponding structural equation model (SEM). Such a connection was initially exploited in the linear case (Shimizu et al., 2006) and extended to the nonlinear case by Monti et al. (2019) who employed TCL.

Briefly, consider data $\mathbf{x} = (x_1, x_2)$. The goal is to establish if the causal direction is $x_1 \to x_2$, or $x_2 \to x_1$, or conclude that no (acyclic) causal relationship exists. Assuming $x_1 \to x_2$, then the problem can be described by the following SEM: $x_1 = f_1(n_1)$, $x_2 = f_2(x_1, n_2)$ where $\mathbf{f} = (f_1, f_2)$ is a (possibly nonlinear) mapping and $\mathbf{n} = (n_1, n_2)$ are latent disturbances that are assumed to be independent. The above SEM can be seen as a nonlinear ICA model where latent disturbances, $\mathbf{n}$, are the sources. As such, we may perform causal discovery by first recovering latent disturbances (using TCL or iVAE) and then running a series of independence tests. Formally, if $x_1 \to x_2$ then, denoting statistical independence by $\perp\!\!\!\perp$, it suffices to verify that $x_1 \perp\!\!\!\perp n_2$ whereas $x_1 \not\perp\!\!\!\perp n_1$, $x_2 \not\perp\!\!\!\perp n_1$ and $x_2 \not\perp\!\!\!\perp n_2$. Such an approach can be extended beyond two-dimensional observations as described in Monti et al. (2019).

To demonstrate the benefits of iVAE as compared to TCL, both algorithms were employed to learn causal structure from fMRI data (details in Supplementary Material I.3). The recovered causal graphs are shown in Figure 4. Blue edges are anatomically feasible whilst red edges are not. There is significant overlap between the estimated causal networks, but in the case of iVAE both anatomically incorrect edges correspond to indirect causal effects. This is in contrast with TCL where incorrect edges are incompatible with anatomical structure and cannot be explained as indirect effects.

6 CONCLUSION

Unsupervised learning can have many different goals, such as: (*i*) approximate the data distribution, (*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. Our model in fact 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 Hyvärinen et al. (2019) are several; briefly, we significantly strengthen the identifiability results, 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

Ben-Israel, A. (1999). The Change-of-Variables Formula Using Matrix Volume. *SIAM J. Matrix Anal. Appl.*, 21(1):300–312.

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

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance Weighted Autoencoders. *arXiv:1509.00519 [cs, stat]*.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. (2018). Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv:1802.04942 [cs, stat]*.

Dinh, L., Krueger, D., and Bengio, Y. (2014). NICE: Non-linear Independent Components Estimation. *arXiv:1410.8516 [cs]*.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and van de Meent, J.-W. (2018). Structured Disentangled Representations. *arXiv:1804.02086 [cs, stat]*.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. (2018). Towards a Definition of Disentangled Representations. *arXiv:1812.02230 [cs, stat]*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.

Hyvärinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pages 3765–3773.

Hyvärinen, A. and Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. In *The 20th International Conference on Artificial Intelligence and Statistics*.

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 *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kim, H. and Mnih, A. (2018). Disentangling by Factorising. *arXiv:1802.05983 [cs, stat]*.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improving Variational Inference with Inverse Autoregressive Flow. *arXiv:1606.04934 [cs, stat]*.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. In *arXiv:1312.6114 [Cs, Stat]*.

Lee, J. M. (2003). *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer-Verlag, New York.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2018). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv:1811.12359 [cs, stat]*.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. (2019). BIVA: A very deep hierarchy of latent variables for generative modeling. In *Advances in Neural Information Processing Systems*, pages 6548–6558.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *arXiv:1611.00712 [cs, stat]*.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2018). Disentangling Disentanglement in Variational Autoencoders. *arXiv:1812.02833 [cs, stat]*.

Monti, R. P., Zhang, K., and Hyvarinen, A. (2019). Causal discovery with general non-linear relationships using non-linear ICA. In *35th Conference on Uncertainty in Artificial Intelligence, UAI 2019*, volume 35.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT press.

Poldrack, R. A., Laumann, T. O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y., Gorgolewski, K. J., Luci, J., Joo, S. J., and Boyd, R. L. (2015). Long-term neural and physiological phenotyping of a single human. *Nature communications*, 6(1):1–15.

Rezende, D. J. and Mohamed, S. (2015). Variational Inference with Normalizing Flows. *arXiv:1505.05770 [cs, stat]*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]*.

Rolinek, M., Zietlow, D., and Martius, G. (2018). Variational Autoencoders Pursue PCA Directions (by Accident). *arXiv:1812.06775 [cs, stat]*.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density Estimation in Infinite Dimensional Exponential Families. *Journal of Machine Learning Research*, 18(57):1–59.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. (2018). Doubly reparameterized gradient estimators for monte carlo objectives. *arXiv preprint arXiv:1810.04152*.