Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes K [1] Diederik P. Kingma [2] Aapo Hyvärinen [3]

[1]Gatsby [2]Google brain [3]INRIA-Saclay

Gatsby Research Talk

tl;dr

We will learn about:

- Deep latent variable models and identifiability.
- Variational autoencoders (VAEs).
- Nonlinear ICA.
- How to use VAEs to solve nonlinear ICA.
- How to use nonlinear ICA to guarantee identifiability of VAEs.

Outline

Unsupervised learning

- ◇ Unsupervised learning is a fundamental challenge in machine learning.
- ◇ Important because of the the amount of unlabelled data that exists (labels are expensive!).
- ◇ Deep latent variable models are very popular.

Goals of unsupervised learning

(i) Learn an accurate model of the data distribution.
→ Variational autoencoders (VAEs), …

(ii) Generate new samples from the data distribution.
→ Generative adversarial networks (GANs), VAEs, …

(iii) Extract useful features to use for other purposes (e.g.: supervised learning).
→ VAEs, GANs, ICA, and many other methods …

(iv) Identify the true latent quantities.
→ ICA.

Goals of unsupervised learning

(i) Learn an accurate model of the data distribution.
→ Variational autoencoders (VAEs), …

(ii) Generate new samples from the data distribution.
→ Generative adversarial networks (GANs), VAEs, …

(iii) Extract useful features to use for other purposes (e.g.: supervised learning).
→ VAEs, GANs, ICA, and many other methods …

(iv) Identify the true latent quantities.
→ ICA.

$\Longrightarrow$ No method seems to achieve all 4 goals, but we can see that VAEs and ICA achieve complementary goals!

Can they be combined in one unifying framework?

Outline

Deep latent variable models

Consider an observed r.v. $\mathbf{x} \in \mathbb{R}^d$, and a latent r.v. $\mathbf{z} \in \mathbb{R}^n$.
A deep latent variable model (DLVM) commonly has the structure:

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

where $\boldsymbol{\theta} \in \Theta$ is a vector of parameters (often modelled as a neural network, hence the *deep*).

The model then gives rise to the observed distribution of the data as:

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

Deep latent Variable models

In practice, we assume that we have a dataset of observations of $\mathbf{x}$:

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

where $\boldsymbol{\theta}^*$ are true but unknown parameters of the model.
Note that the original values $\mathbf{z}^{*(i)}$ of the latent variables $\mathbf{z}$ are by definition not observed and unknown.

Goal 1: Find a good approximation $\hat{\boldsymbol{\theta}}$ of the true parameter $\boldsymbol{\theta}^*$, based on the observations $\mathcal{D}$ alone.

Goal 2: Recover the latent values $\mathbf{z}^{*(i)}$.

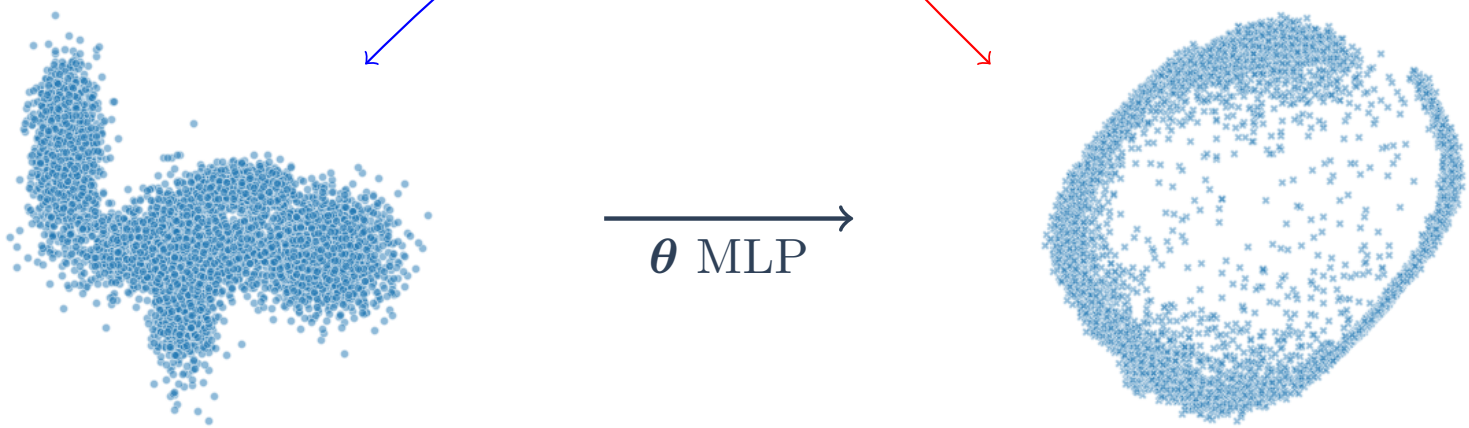
Deep latent Variable models

In practice, we assume that we have a dataset of observations of $\mathbf{x}$:

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

$$\mathbf{x}^{(i)} \sim p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z}^{*(i)})$$

Example:



mixture of Gaussians

observations

Variational autoencoders

Variational autoencoders [Kingma and Welling, 2013] are a framework that combines a DLVM and an estimation method that simultaneously learns:

- ◇ an estimate $\hat{\boldsymbol{\theta}}$ of the true parameter $\boldsymbol{\theta}^*$ s.t. $p_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x})$.
- ◇ a variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ of the posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$.

This is done by maximizing a variational lower bound of the data log-likelihood:

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

Terminology: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is called the *decoder*, and $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ the *encoder*.

⚠ In general, after training, we learn $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ s.t. the only guarantees are $p_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x})$ and $q_{\hat{\boldsymbol{\phi}}}(\mathbf{z}|\mathbf{x}) \approx p_{\hat{\boldsymbol{\theta}}}(\mathbf{z}|\mathbf{x})$.

Identifiability

A desired property for a deep latent variable model is *identifiability*, which can be formulated as follows:

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

This means that if $p_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{x})$, then:

- ◇ the joint densities also match $p_{\hat{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z})$.
- ◇ we learned the correct prior $p_{\hat{\boldsymbol{\theta}}}(\mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{z})$.
- ◇ we learned the correct posterior $p_{\hat{\boldsymbol{\theta}}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x})$
- ◇ in the case of VAEs, we can use $q_{\hat{\boldsymbol{\phi}}}(\mathbf{z}|\mathbf{x})$ to perform inference over the sources $\mathbf{z}^*$ from which the data originates.

The problem is that the identifiability equation doesn't hold for general deep latent variable models!

Identifiability 2

In general: $p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \quad \not\Longrightarrow \quad p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}'}(\mathbf{x}, \mathbf{z}) \;\; \forall (\mathbf{x}, \mathbf{z})$

Let's illustrate this with a simple example:

$$p(\mathbf{z}) = \mathcal{N}(0, I) \quad \text{and} \quad p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\boldsymbol{\theta}(\mathbf{z}), I)$$

Let $\mathbf{z}' = U\mathbf{z}$ for some orthogonal matrix U. Then $\mathbf{z}$ and $\mathbf{z}'$ have the same distribution.
Define $\boldsymbol{\theta}'(\mathbf{z}') = \boldsymbol{\theta}(U^T\mathbf{z}')$ and $p_{\boldsymbol{\theta}'}(\mathbf{x}, \mathbf{z}') = p_{\boldsymbol{\theta}}(\mathbf{x}|U^T\mathbf{z}')p(\mathbf{z}')$. Then

$$p_{\boldsymbol{\theta}'}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\underbrace{U^T\mathbf{z}'}_{=\mathbf{z}})\underbrace{p(\mathbf{z}')\mathrm{d}\mathbf{z}'}_{=p(\mathbf{z})\mathrm{d}\mathbf{z}} = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = p_{\boldsymbol{\theta}}(\mathbf{x})$$

but the posteriors are different, and thus the joint distributions as well.

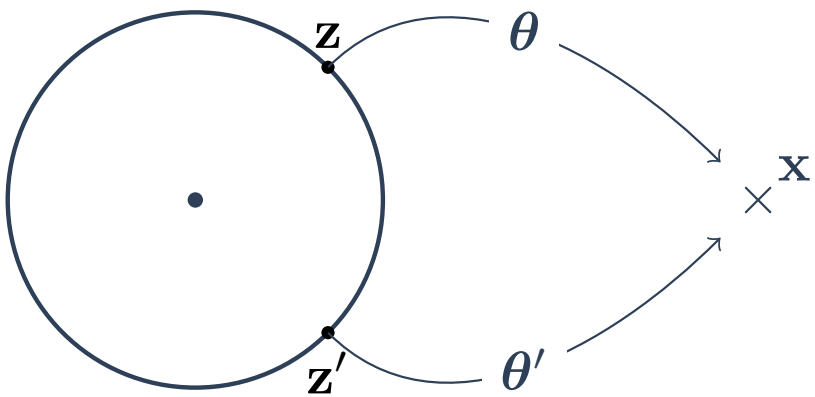
Identifiability 2

In general: $p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \quad \not\Longrightarrow \quad p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}'}(\mathbf{x}, \mathbf{z}) \;\; \forall(\mathbf{x}, \mathbf{z})$

Let's illustrate this with a simple example:

$$p(\mathbf{z}) = \mathcal{N}(0, I) \quad \text{and} \quad p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\boldsymbol{\theta}(\mathbf{z}), I)$$

Let $\mathbf{z}' = U\mathbf{z}$ for some orthogonal matrix U. Then $\mathbf{z}$ and $\mathbf{z}'$ have the same distribution.



Identifiability 3

What about the general case?

Theorem [Hyvärinen and Pajunen, 1999]

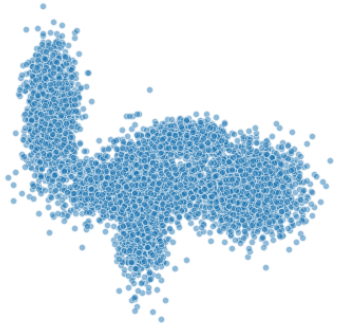
Let $\mathbf{z}$ be an n-dimensional random vector of any distribution. Then there exists an invertible transformation $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ such that the distribution of $\mathbf{z}' := \mathbf{g}(\mathbf{z})$ is a standard Gaussian distribution.

- We transform any latent variable $\mathbf{z}$ into a standard Gaussian $\tilde{\mathbf{z}} = \mathbf{g}(\mathbf{z})$.
- We apply an orthogonal transformation U to $\tilde{\mathbf{z}}$.
- We invert the initial transformation to get $\mathbf{z}' = \mathbf{g}^{-1}(U\mathbf{g}(\mathbf{z}))$ which has the same distribution as $\mathbf{z}$.

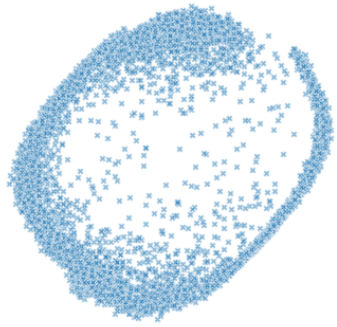
By defining $p_{\boldsymbol{\theta}'}(\mathbf{x}, \mathbf{z}') = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{g}^{-1}(U^T\mathbf{g}(\mathbf{z}')))p(\mathbf{z}')$, then

$$p_{\boldsymbol{\theta}'}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{g}^{-1}(U^T\mathbf{g}(\mathbf{z}')))p(\mathbf{z}')\mathrm{d}\mathbf{z}' = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = p_{\boldsymbol{\theta}}(\mathbf{x})$$

Back to the example: VAEs

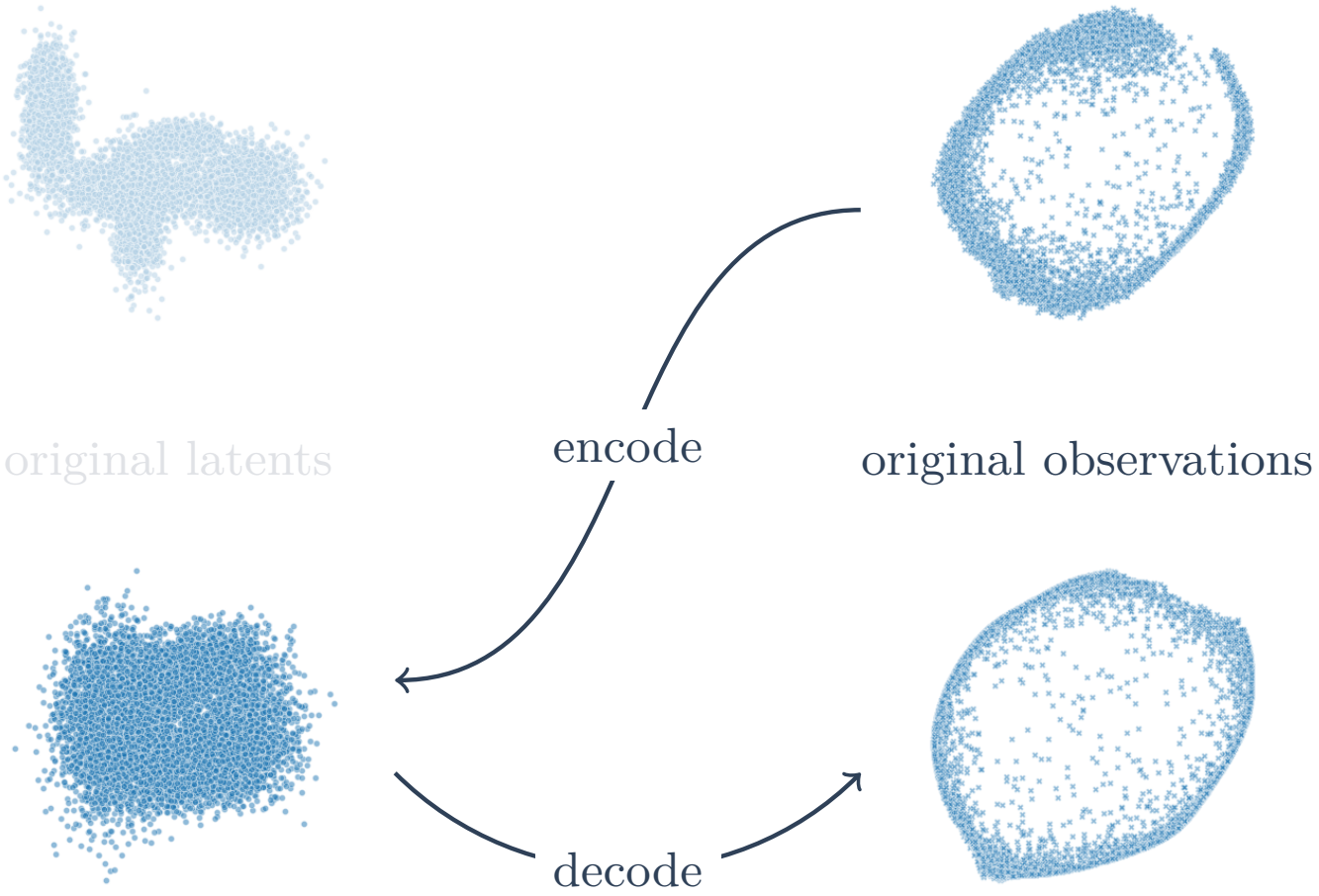


original latents



original observations

Back to the example: VAEs



Outline

Nonlinear ICA

In ICA, the independent components of the latent variables $\mathbf{z} \in \mathbb{R}^d$ are mixed into an observation $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) \quad \text{and} \quad p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i)$$

This is essentially a deterministic deep generative model with degenrate posteriors.

Goal: Recover the original componenets $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$ by inverting $\mathbf{f}$.

The goal of ICA has always been identifiability!

⚠ This model is also unidentifiable when $\mathbf{f}$ is nonlinear!

$\Longrightarrow$ To solve nonlinear ICA, we have to introduce additional constraints on the latent variables or the model in general.

An identifiable nonlinear ICA model

[Hyvarinen et al., 2019] proposes using an auxiliary variable $\mathbf{u} \in \mathbb{R}^m$ that controls the independence of the sources: $p(\mathbf{z}|\mathbf{u}) = \prod_i p(z_i|\mathbf{u})$. They prove that the model is identifiable, and use a self-supervised heuristic scheme to recover the latent variables.

Limitations of this approach:

- $\dim(\mathbf{z}) = \dim(\mathbf{x})$
- $\mathbf{f}$ is deterministic, and thus the posteriors are degenerate.
- hard to cross validate.
- no analysis on its statistical efficiency (compared to MLE for example).
- only recovers the backward model $\mathbf{f}^{-1}$.

VAEs and nonlinear ICA have complimentary strengths!

Indeterminacies of nonlinear ICA

Nonlinear ICA has indeterminacies that can't be resolved:

- ◇ permutation of the components (similar to linear ICA).
- ◇ component-wise nonlinear transformations (equivalent to scaling in linear ICA).

These indeterminacies are due to a fundamental ambiguity in the setup of nonlinear ICA and do not represent a limitation of nonlinear ICA algorithms.

Outline

Definition of proposed model 1

Let $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathbb{R}^m$ be two observed r.v., and $\mathbf{z} \in \mathbb{R}^n$ ($n \leq d$) a latent variable. Let $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ be the parameters of the following model:

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

where

$$p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) = p_{\varepsilon}(\mathbf{x} - \mathbf{f}(\mathbf{z})), \;\; \varepsilon \sim p_{\varepsilon} \;\; \text{is a noise r.v.,} \;\; \varepsilon \perp \{\mathbf{x}, \mathbf{z}\}$$

$$\mathbf{f} \;\; \text{is bijective}$$

This also includes:

- ◇ non-noisy observations when ε is Gaussian and $\mathbb{V}(\varepsilon) \to 0$.
- ◇ discrete observations as a limit of concrete distributions, *cf.* [Maddison et al., 2016].

Definition of proposed model 2

The prior on the latent variables $p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})$ is a *conditionally factorial exponential family*:

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

where:

- Q_i is the base measure and $T_{i,j}$ are the components of the sufficient statistic.
- $Z_i(\mathbf{u})$ is the normalizing constant and $\lambda_{i,j}(\mathbf{u})$ the corresponding parameters, crucially depending on $\mathbf{u}$.
- k, the number of components within each exponential family, is fixed (not estimated).

N.b. Exponential families have universal approximation capabilities, so this assumption is not very restrictive [Sriperumbudur et al., 2017].

Estimation by VAEs

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 above with parameters $\boldsymbol{\theta}^* = (\mathbf{f}^*, \mathbf{T}^*, \boldsymbol{\lambda}^*)$.

We use a VAE to learn approximations of $\boldsymbol{\theta}^*$ and the intractable posterior $p_{\boldsymbol{\theta}^*}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ by maximizing $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi})$, a lower bound on the data log-likelihood defined by:

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

Is the introduced model identifiable? And can it be estimated using VAEs?

We call this model *iVAE* (identifiable/ica VAE)

Identifiability result

Theorem 1 (simplified: $n = 2, k = 1$, noiseless)

Assume there exist 3 distinct points $\mathbf{u}^0, \mathbf{u}^1, \mathbf{u}^2$ such that the matrix

$$L = \begin{pmatrix} \lambda_1^*(\mathbf{u}^1) - \lambda_1^*(\mathbf{u}^0) & \lambda_1^*(\mathbf{u}^2) - \lambda_1^*(\mathbf{u}^0) \\ \lambda_2^*(\mathbf{u}^1) - \lambda_2^*(\mathbf{u}^0) & \lambda_2^*(\mathbf{u}^2) - \lambda_2^*(\mathbf{u}^0) \end{pmatrix}$$

is invertible. Then:

$$p_{\mathbf{f},\mathbf{T},\boldsymbol{\lambda}}(\mathbf{x}) = p_{\mathbf{f}^*,\mathbf{T}^*,\boldsymbol{\lambda}^*}(\mathbf{x}) \implies \exists A \text{ invertible matrix s.t.}$$
$$(T_1(z_1), T_2(z_2))^T = A(T_1^*(z_1^*), T_2^*(z_2^*))^T$$

The matrix A can be removed by applying linear ICA.

Notation: $\boldsymbol{\lambda}^l = (\lambda_1^*(\mathbf{u}^l), \lambda_2^*(\mathbf{u}^l))^T$. The invertibility of L is equivalent to saying that $\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^0$ and $\boldsymbol{\lambda}^2 - \boldsymbol{\lambda}^0$ are linearly independent.

Intuition behind the Theorem

Suppose $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is an isotropic Gaussian with fixed variance, then $\boldsymbol{\lambda}(\mathbf{u})$ is simply the mean.

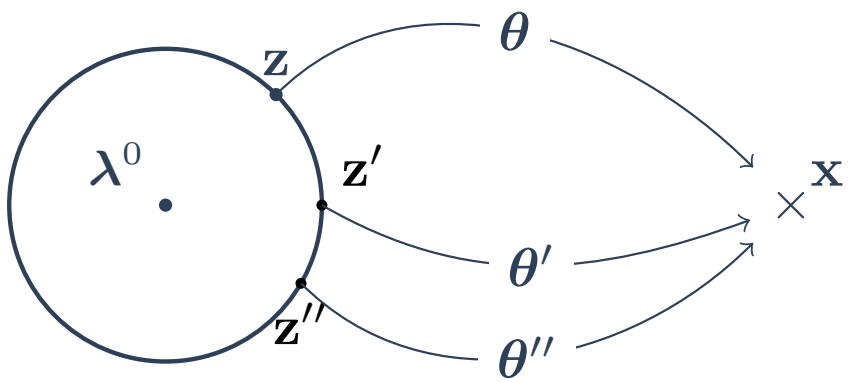
$\rightarrow$ In how many ways can we transform $\mathbf{z}$ without changing any intermediate quantities, thus yielding the same distribution over $\mathbf{x}$?

Intuition behind the Theorem

Suppose $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is an isotropic Gaussian with fixed variance, then $\boldsymbol{\lambda}(\mathbf{u})$ is simply the mean.

$\rightarrow$ In how many ways can we transform $\mathbf{z}$ without changing any intermediate quantities, thus yielding the same distribution over $\mathbf{x}$?

- $\diamond$ prior is unconditional (one Gaussian centered around $\boldsymbol{\lambda}^0$): radial transformations yield the same observations.

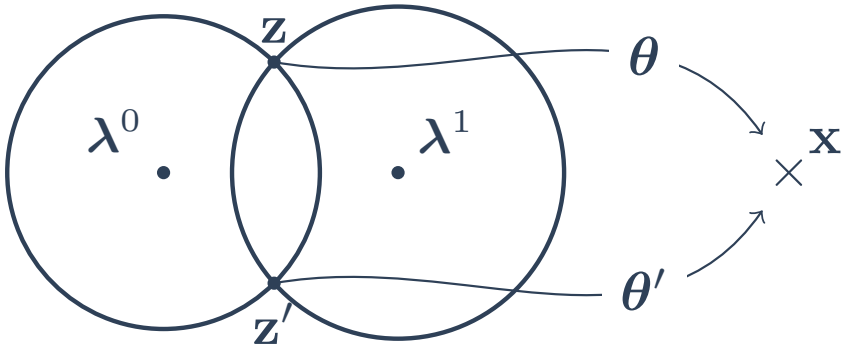


Intuition behind the Theorem

Suppose $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is an isotropic Gaussian with fixed variance, then $\boldsymbol{\lambda}(\mathbf{u})$ is simply the mean.

$\rightarrow$ In how many ways can we transform $\mathbf{z}$ without changing any intermediate quantities, thus yielding the same distribution over $\mathbf{x}$?

- $\diamond$ prior is a mixture of two Gaussians (centered around $\boldsymbol{\lambda}^0$ and $\boldsymbol{\lambda}^1$): a transformation that is mixture agnostic will necessarily be on the intersection.

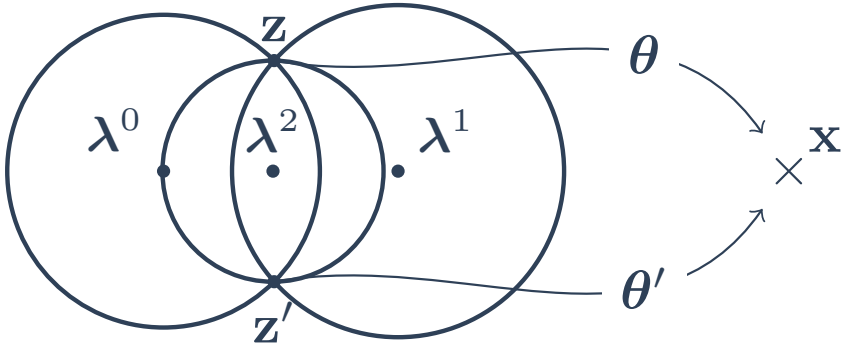


Intuition behind the Theorem

Suppose $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is an isotropic Gaussian with fixed variance, then $\boldsymbol{\lambda}(\mathbf{u})$ is simply the mean.

$\rightarrow$ In how many ways can we transform $\mathbf{z}$ without changing any intermediate quantities, thus yielding the same distribution over $\mathbf{x}$?

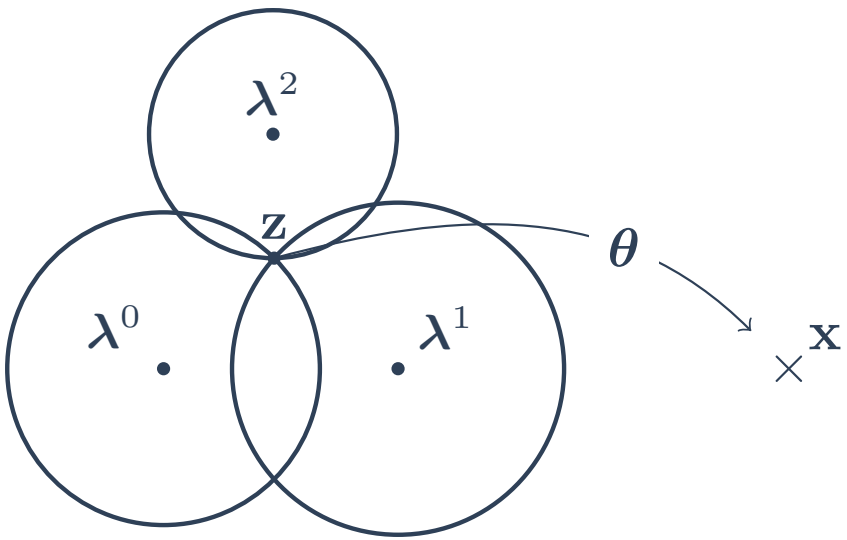
- $\diamond$ prior is a mixture of three aligned Gaussians: same thing!



Intuition behind the Theorem

Suppose $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is an isotropic Gaussian with fixed variance, then $\boldsymbol{\lambda}(\mathbf{u})$ is simply the mean.
$\rightarrow$ In how many ways can we transform $\mathbf{z}$ without changing any intermediate quantities, thus yielding the same distribution over $\mathbf{x}$?

- $\diamond$ prior is a mixture of three unaligned Gaussians: only one intersection!



Estimation by VAEs 2

Theorem 2

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 indeterminacies of Theorem 1.

Proof: The loss can be written as:

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

By first optimizing the loss over $\boldsymbol{\phi}$, the KL term reaches 0 and the loss will be equal to the log-likelihood.

The VAE in this case inherits all the properties of MLE. In particular, it is a consistent estimator of the parameters $\boldsymbol{\theta}^*$ *i.e.* in the limit of infinite data. □

Back to the example: iVAE

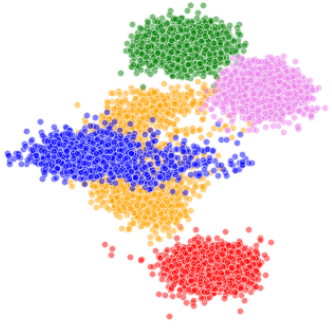
original latents

original observations

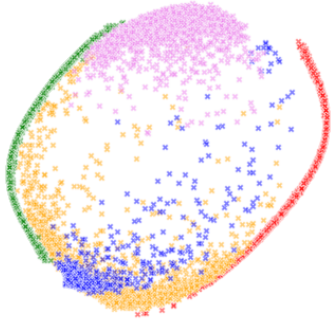
Back to the example: iVAE

original latents

original observations



recovered latents

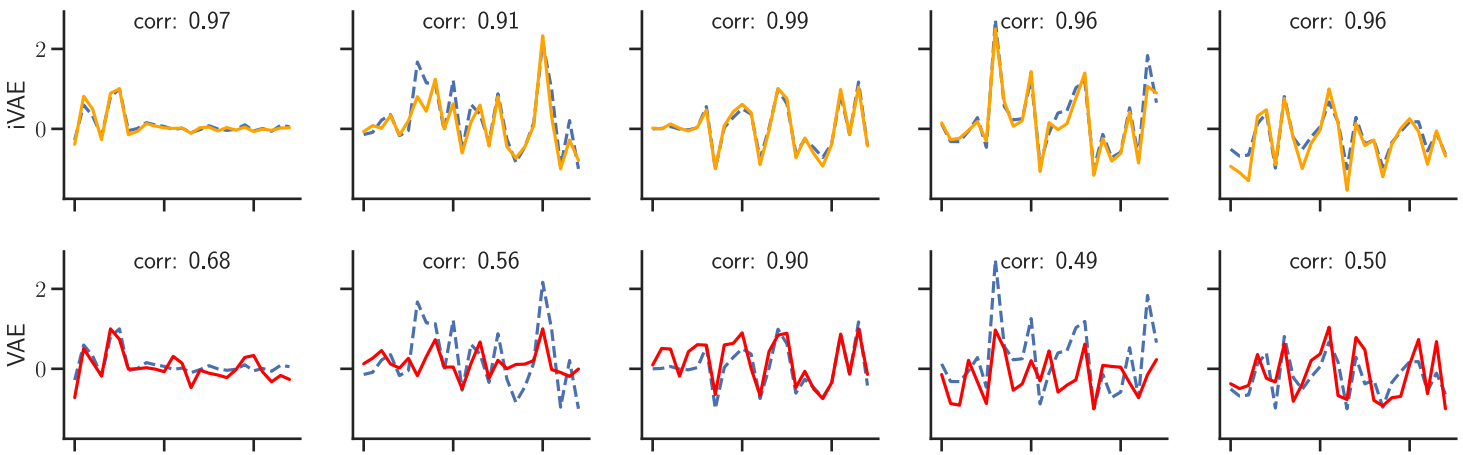


reconstructed observations

Second Example

We sample 5-d latents from a conditional Gaussian distribution where $p(\mathbf{u}) = \text{Unif}(\{1, \ldots, 40\})$, and we mix them into 25-d observations.

The dashed blue line is the true source signal, and the recovered latents are in solid coloured lines. We also reported the mean correlation coefficients for every (source, latent) pair.



Outline

Conclusion: contributions

- ◇ Draw attention to the (un)-identifiability of popular deep latent variable models.
- ◇ First identification proof withing the VAE framework.
- ◇ First proof of solvability of nonlinear ICA by MLE.
- ◇ Extend the nonlinear ICA framework to noisy and discrete observations, and lower dimensional latent variables.

Conclusion: future work

- ◇ Have a more flexible decoding distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.
- ◇ Have a more general form of the prior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$.
- ◇ Extend to the case $\dim(\mathbf{z}) > \dim(\mathbf{x})$.

References i

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

Hyvarinen, A., Sasaki, H., and Turner, R. (2019).
Nonlinear ica using auxiliary variables and generalized contrastive learning.
In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR.

Kingma, D. P. and Welling, M. (2013).
Auto-encoding variational bayes.
arXiv preprint arXiv:1312.6114.

References ii

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].

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

From continuous to discrete

Categorical distributions can be viewed as a infinitesimal-temperature limit of continuous distributions.

For example, let:

$$\begin{aligned} \mathbf{m} &= \mathbf{f}(\mathbf{z}) \\ \mathbf{x} &= \text{sigmoid}((\mathbf{m} + \boldsymbol{\varepsilon})/T) \\ \forall \varepsilon_i \in \boldsymbol{\varepsilon} : \varepsilon_i &\sim \text{Logistic}(0, 1) \end{aligned}$$

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

If $T \to 0^+$, then:

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

For proof that this holds, *cf.* [Maddison et al., 2016], appendix B.

Identifiability result 1

Definition 1

Let $\eqsim$ be an equivalence relation on Θ. We say that a probabilistic model is identifiable up to $\eqsim$ (or $\eqsim$-identifiable) if

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \implies \boldsymbol{\theta}' \eqsim \boldsymbol{\theta}$$

Proposition 1

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

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

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.

Identifiability result 2

Theorem 1

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 p_ε.

(ii) The sufficient statistics $T_{i,j}$ are differentiable almost everywhere and $\frac{\partial T_{i,j}}{\partial z}(z) \neq 0$ almost surely for $z \in \mathcal{Z}_i$ and for all (i, j).

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

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

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.

Understanding assumption (iii) in Theorem 1

Let $\mathbf{u}^0$ be an arbitrary point in its support $\mathcal{U}$, and $h(\mathbf{u}) = \big(\lambda_{1,1}(\mathbf{u}) - \lambda_{1,1}(\mathbf{u}^0), \ldots, \lambda_{n,k}(\mathbf{u}) - \lambda_{n,k}(\mathbf{u}^0)\big)^T \in \mathbb{R}^{nk}$.

Let's suppose for a second that for any choice of points, the vectors $(h(\mathbf{u}^1), \ldots, h(\mathbf{u}^{nk}))$ 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.

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.

Identifiability result 3

Theorem 3

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.