Nonlinear ICA

Presenter: Zhe Wang

https://qdata.github.io/deep2Read

Zhe Wang

201909

Content

Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, Aapo Hyvärinen

UCL, Google Brain, Inria, University of Helsinki

VAE

Suppose the ground truth data distribution is $P_{\theta*}(x), x \in \mathrm{R}^d$.

Data generating process:

- Sample a latent variable z from $P_{\theta*}(z), z \in \mathrm{R}^n, n \leqslant d$.
- Sample x from the conditional distribution $P_{\theta*}(x|z)$.
- The marginal distribution $p_{\theta*}(x) = \int p_{\theta*}(x|z)p_{\theta*}(z)dz$

VAE is a latent-variable model, which learns a full generative model $p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$ and an inference model $q_\phi(z|x)$ that approximates its posterior $p_\theta(z|x)$.

VAE allows us to approximate the true (but unknown) marginal distribution over the observed variables:

$$p_\theta(x) = p_{\theta*}(x)$$

VAE models learn the full data generation process $P_\theta(Z), P_\theta(X|Z), P_\theta(Z|X)$. All we know is $P_\theta(X) = P_{\theta*}(X)$, we don't know what the remaining are.

Goal: under what condition: $P_\theta(X) = P_{\theta*}(X)$ can guarantee $P_\theta(Z|X) = P_{\theta*}(Z), P_\theta(X|Z) = P_{\theta*}(X|Z)$

Identifiability of the deep latent-variable models.

$$p_\theta(x) = p_{\theta*}(x) \longrightarrow \theta* = \theta \longrightarrow p_{\theta*}(x, z) = p_\theta(x, z)$$

In causality and ICA literature:

After observing x, we can construct infinitely many generative models which have the same marginal distribution of x. Any one of these models could be the true causal generative model for the data, and the right model cannot be identified given only the distribution of x

Nonlinear ICA:

In nonlinear ICA, we assume observations $x \in \mathrm{R}^d$, which are the result of an unknown (but invertible) transformation f of latent variables $z \in \mathrm{R}^d$:

Difference of VAE with nonlinear ICA:

- The dimension of observation and dimension of hidden variables are the same
- The transformation function f is deterministic.

The goal of nonlinear ICA is to recover f^{-1} and find the independent variables $z = f^{-1}(x)$. Thus, the goal of nonlinear ICA was always **identifiability**,

Main Assumption: A conditionally factorized prior distribution over the latent variables $p_\theta(z|u)$, where u is an additionally observed variable
And the data generation stage is a additive noise model $x = f(z) + \epsilon$

$p(z|u)$ is conditionally factorial

$$p(z|u) = \prod_{i=1}^{n} p(z_i|u),$$

The prior on the latent variables $p_\theta(z|u)$ is assumed to be conditionally factorial, where each element of $z_i \in z$ has a univariate exponential family distribution given conditioning variable u.

$$p_{T,\lambda}(z_i|u) = \frac{Q_i(z_i)}{Z_i(u)} \exp[\sum_{j=1}^{k} T_{ij}(z_i)\lambda_{ij}(u)]$$

Q_i is the base measure, $Z_i(u)$ is the normalizing constant and
$T_i = (T_{i,1}, \cdots, T_{i,k})$ are the sufficient statistics and
$\lambda_i(u) = (\lambda_{i,1}(u), \cdots, \lambda_{i,k}(u))$ the corresponding parameters, crucially depending on u.

Estimate the latent variable from a dataset $D = \{(x^1, u^1), \cdots, (x^N, u^N)\}$ generated by the introduced process. The paper propose to use VAE as a means of learning the true generating parameters $\theta^* = (f^*, T^*, \lambda^*)$

$$ELBO = E_D(E_{q_\phi(z|x,u)} \log p_\theta(x|z,u) - KL(q_\phi(z|x,u)||p(z|u))$$

Identifiability theory:

We denote:
$T(z) = [T_{1,1}(z_1), \cdots, T_{1,k}(z_1), \cdots, T_{n,1}(z_n), \cdots, T_{n,k}(z_n)] \in \mathrm{R}^{nk}$
$\lambda(u) = [\lambda_{1,1}(u_1), \cdots, \lambda_{1,k}(u_1), \cdots, \lambda_{n,1}(u_n), \cdots, \lambda_{n,k}(u_n)] \in \mathrm{R}^{nk}$
$\theta = (f, T, \lambda)$

Definition 1:
Let $\sim$ be an equivalence relation on θ. We say VAE is identifiable up to $\sim$ if:

$$p_\theta(x) = p_{\theta*}(x) \longrightarrow \theta \sim \theta^*$$

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

$$(f, T, \lambda) \sim (f^*, T^*, \lambda^*) \Longleftrightarrow A, c | \, T^*(f^{*-1}(x)) = AT(f^{-1}(x)) + c \quad \forall x$$

If A is invertible, we note it as $\sim_A$, if A is a permutation matrix, we note it as $\sim_p$

Main theorem:

Theorem 1:
Suppose the true data generating process is defined as above, and assume the following holds:

- f^* is injective,
- $T^*_{i,j}$ in are differentiable almost everywhere,
- There exist $nk + 1$ distinct points $u_0, ..., u_{nk}$ such that the matrix $L = (\lambda^*(u_1) - \lambda^*(u_0), \cdots, \lambda^*(u_{nk}) - \lambda^*(u_0))$ of size $nk \times nk$ is invertible

Then the parameters (f^*, T^*, λ^*) are $\sim_A$-identifiable.

Theorem 2: If $k \geqslant 2$, the assmption in theorem 1 holds, also

- The sufficient statistics $T^*_{i,j}$ in are twice differentiable.
- The mixing function f^* has all second order cross derivatives.

Then the parameters (f^*, T^*, λ^*) are $\sim_P$-identifiable.

Theorem 3: If $k = 1$, the assmption in theorem 1 holds, also

- The sufficient statistics $T^*_{i,1}$ are not monotonic
- All partial derivatives of f^* are continuous.

Then the parameters (f^*, T^*, λ^*) are $\sim_P$-identifiable.

Theorem 4: Assume the following:

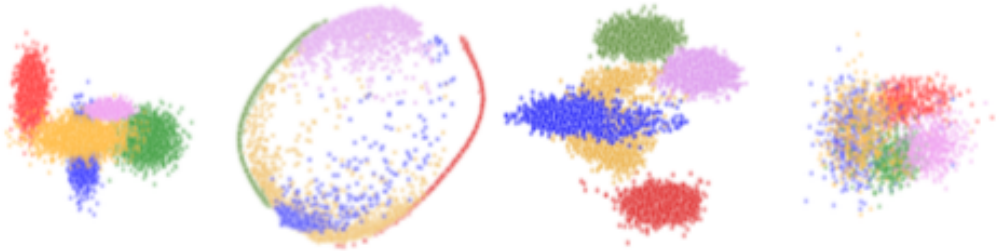
- The family of distributions $q_\phi(z|x, u)$ contains $p_{\theta*}(z|x, u)$.
- We maximize $L(\theta, \phi)$ with respect to both θ and ϕ.

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

Experiments

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 u is the segment label, and its distribution is uniform on the integer set $[1, M]$.

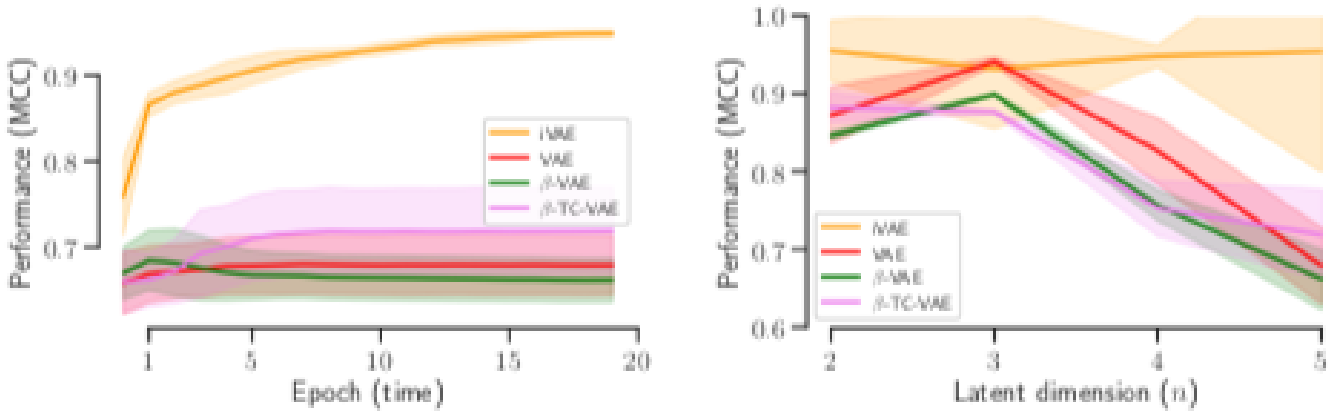
Experiments on 2D dataset.



(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.

Metric: mean correlation coefficient (MCC) between the original sources and recovered latent space.

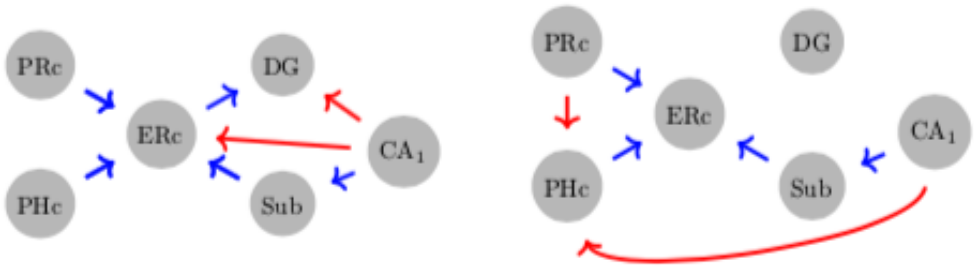


(a) Training dynamics (b) Changing n

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)).

Causal analysis

Consider data $x = (x_1, x_2)$. The goal is to establish if the causal direction is $x_1 \to x_2$, or $x_2 \to x_1$.
Assume the data generation process is $x_1 = f_1(n_1), x_2 = f_2(x_1, n_2)$ where $f = (f_1, f_2)$ is a (possibly nonlinear) mapping. Then, we can recover the distribution of n_1, n_2. Then we can perform independence analysis to find the causal direction.
If $x_1 \to x_2$, 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 $x_2 \not\perp\!\!\!\perp n_2$.



(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.

Summary:
Given an additional observable variable u, such as class labels and time indices for times series data, we can discover the true latent distributions $p(z)$, data generating process $f(x|z)$ and the inverse.

Limitations:
The dimension of latent variables is required as a prior knowledge.

Some recent work: Some recent results show if the true latent distribution $p(z)$ is multi-Gaussian, their encoder can better reconstruct the latent distribution and discover the dimension of hidden space automatically.

Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning

Aapo Hyvärinen, Hiroaki Sasaki, Richard E. Turner

UCL, University of Helsinki, NAIST, Microsoft Research

This paper focus on the identifiability of nonlinear ICA.
Suppose the observed variable $x \in \mathbb{R}^n$, which is generated from n independent latent variables called independent components:$\{z_i\}_{i=1}^n$

In nonlinear ICA, the observation x is an nonlinear transformation of independent components $z = (z_1, z_2, \cdots, z_n)$.

$$x = f(z),$$

where f is a smooth and invertible function.

Goal:
Recover f^{-1} and independent components z given the observation x.

The key assumption is the independent components is factorial conditioned on an auxiliary variable u:

$$p(z|u) = \prod_{i=1}^{n} p(z_i|u)$$

$$p(z_i|u) = \frac{Q_i(z_i)}{Z_i(u)} \exp[\sum_{j=1}^{k} T_{ij}(z_i)\lambda_{ij}(u)]$$

Learning algorithm: Contrastive learning, inspired by the idea of transforming unsupervised learning to supervised learning

Define two datasets:

$$\tilde{x} = (x, u) \quad vs. \quad \tilde{x}^* = (x, u^*)$$

where u^* is a random value from the distribution of the u, but independent of x, created in practice by random permutation of the empirical sample of the u.

Algorithm:
Learn a nonlinear logistic regression system using a regression of the form

$$r(x, u) = \sum_{i=1}^{n} \phi_i(h_i(x), u)$$

which then gives the posterior probability of the first class as $1/(1 + exp(-r(x, u))$. The scalar features h_i would typically be computed by hidden units in a neural network.

Theory:

Suppose the true data generating process is defined as above, and assume the following holds:

- There exist $nk + 1$ distinct points $u_0, ..., u_{nk}$ such that the matrix $L = (\lambda^*(u_1) - \lambda^*(u_0), \cdots, \lambda^*(u_{nk}) - \lambda^*(u_0))$ of size $nk \times nk$ is invertible.
- We train a nonlinear logistic regression system with universal approximation capability to discriminate between $\tilde{x}$ and $\tilde{x}^*$ with regression function $r(x, u)$.
- In the regression function $r(x, u)$ we constrain $h = (h_1, \cdots, h_n)$ to be invertible, as well as smooth, and constrain the inverse to be smooth as well.

In the limit of infinite data, $h(x) = [h_1(x), \cdots, h_n(x)]$ provides a consistent estimator of the independent components, up to a linear transformation of point-wise scalar functions of the independent components.

Different possible choices of u:

A fundamental case is to consider time series, where $x(t) = f(z(t))$

Using time as auxiliary variable:

In the case of nonstationary data.
Assume we observe a time series $x(t)$.
Assume the n independent components are nonstationary, with densities $p(z_i|t)$.
For analysing such nonstationary data in our framework, define $x = x(t)$ and $u = t$

Using history as auxiliary variables:

we consider the theory in the case where u is the history of each variable. For the purposes of our present theory, we define $x = x(t)$ and $u = x(t-1)$ based on a time-series model

Combining time and history:

Clearly, we can combine these two by defining $u = (x(t1), t)$, and thus discriminating between

$$\tilde{x}(t) = (x(t), x(t-1), t) \quad vs. \quad \tilde{x}^*(t) = (x(t), x(t^*-1), t^*),$$

where t^* is a random time index.

Using class label as auxiliary variable

Denote by $c \in \{1, \cdots, k\}$ the class label with k different classes. As a straight-forward application of the theory above, we learn to discriminate between

$$\tilde{x} = (x, c) \quad vs. \quad \tilde{x}^* = (x, c^*)$$

where c is the class label of x, and c^* is a randomized class label.