A Gentle Introduction to Identifiable Generative Models

Erdun Gao

The University of Melbourne

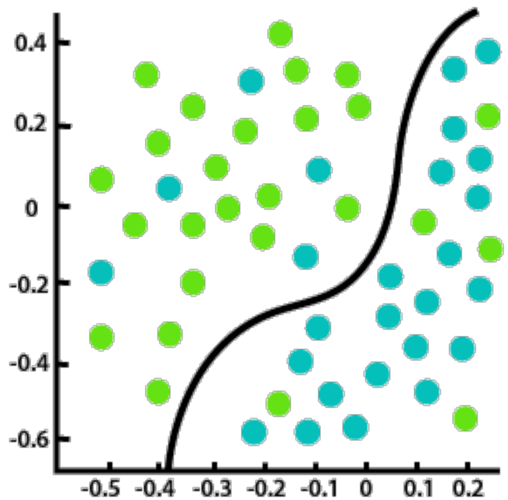
24 February 2022

Some of the contents/viewpoints are cited from *Nonlinear Independent Component Analysis*

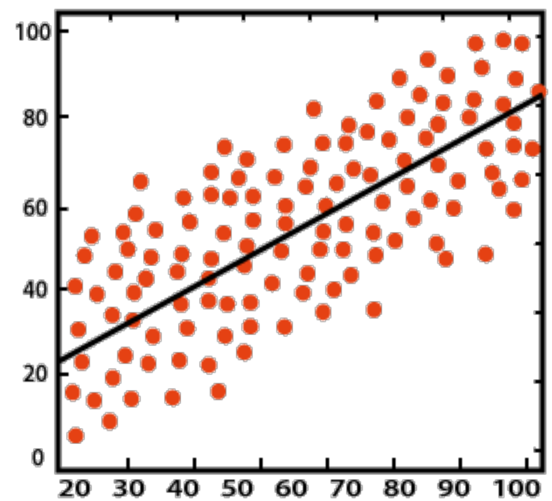
Table of Contents

Unsupervised Learning

- *Supervised Learning*



Classification



Regression

With N examples sampled from $P(X, Y)$, seek a function $g: X \to Y$ using a score function $f: X \times Y \to \mathbb{R}$ so that

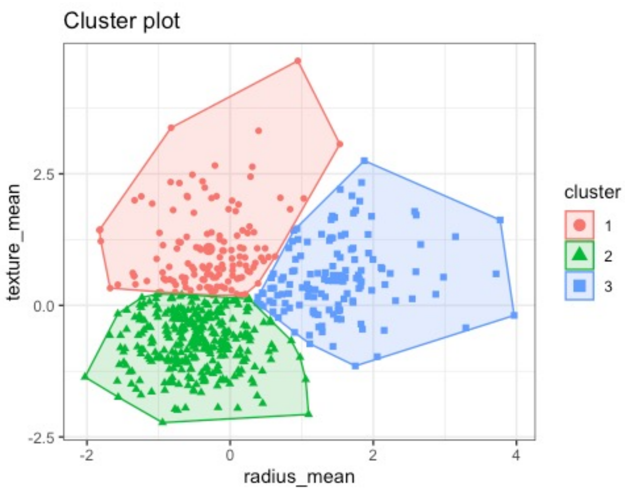
$$g(x) = \underset{y}{\operatorname{argmax}}\, f(x, y)$$

Unsupervised Learning

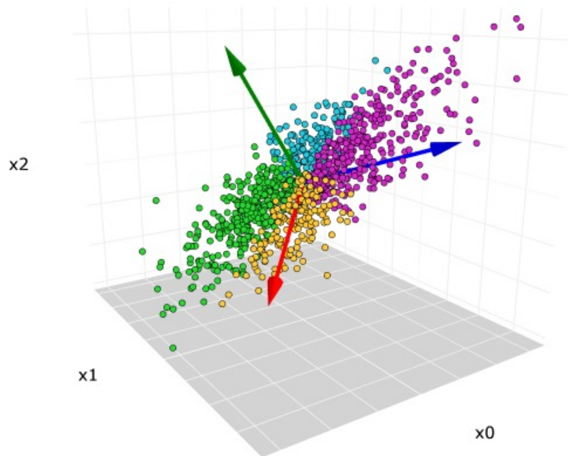
Problems:

😭 Labels may be difficult to obtain
😭 Human annotation may be required
😭 Labels may not be informative

$\Longrightarrow$ With only $P(X)$, what to learn?



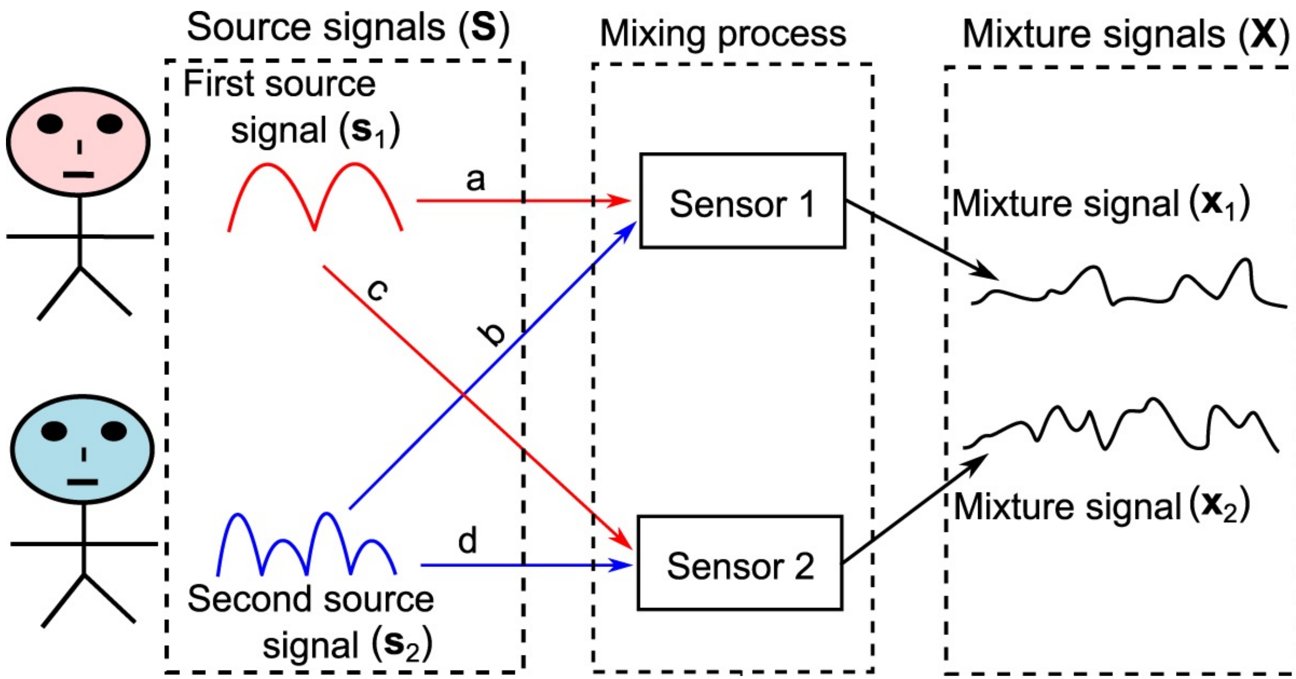
Clustering



Dimensions Reduction

Unsupervised Learning

Independent Component Analysis



Assume observed signals $x = (x_1, x_2, \cdots, x_d)$ are generated as a transformation $f = (f_1, f_2, \cdots, f_d)$ of d independent source signals $s = (s_1, s_2, \cdots, s_d)$:

$$x_i = f_i(s)$$

Identifiable Models

A model is identifiable if it is theoretically possible to learn the ***true model's parameters*** after obtaining an infinite observations. Or, different parameters must induce different probability distributions.

Mathmatically, let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model where the parameter space Θ is either finite or infinite dimensional. We say that $\mathcal{P}$ is identifiable if the mapping $\theta \mapsto P_\theta$ is one-to-one:

$$P_{\theta_1} = P_{\theta_2} \Longrightarrow \theta_1 = \theta_2 \text{ for all } \theta_1, \theta_2 \in \Theta$$

In the ICA problem, the identifiability is defined to recover the parameters of all functions in $\mathcal{F}$.

Theoretical results on ICA

- *Linear independent component analysis (ICA)*

$$x_i = \sum_{j=1}^{d} A_{ij} s_j \quad \text{for all} \quad i = 1,2,\cdots,d$$

A_{ij}constant parameters describing “mixing” and A is the mixing matrix.

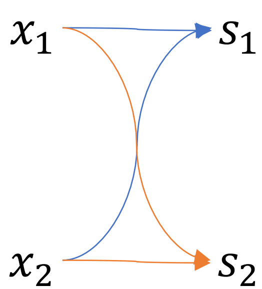
Theorem: Assume that (1) all the independent components s_i must be non-Gaussian. (2) the number of observed linear mixtures must at least as large as the number of independent components. (3) A must be of full column rank. Then, observing only x we can recover both A and s.

Theoretical results on ICA

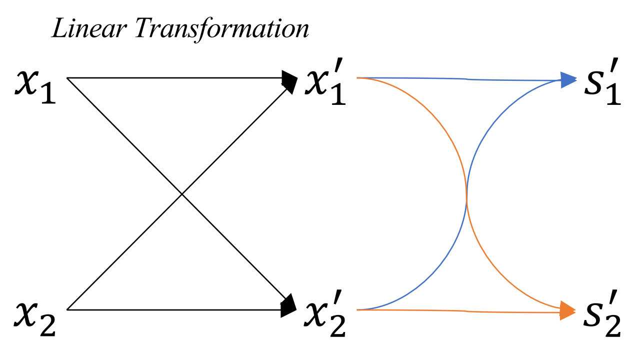
- *Non-Linear ICA*

    😀 Extending ICA to the non-linear case can get general disentanglement.

    😌 The general nonlinear ICA is not identifiable. That is to say, we cannot recover the origianl sources.



Ground-truth models



Hyvärinen A, Pajunen P. Nonlinear independent component analysis: Existence and uniqueness results[J]. Neural networks, 1999, 12(3): 429-439.

Nonlinear ICA

- *The final task of Nonlinear ICA*

$$x = f(s) \qquad \Rightarrow \qquad s = g(x)$$

Generally, we do not put many constraints on f but usually smooth and invertible. Then, we need to constraint the distribution of s.

With:

$$p(x)|\det(J_f)| = p(s) \quad \Leftrightarrow \quad p(x) = p(s)|\det(J_g)|$$

$$\log p(x) = \sum_{i=1}^{d} \log p(s_i) + \log|\det(J_g)|$$

$$\log p(s_i) = q_{i,0}(s_i) + \sum_{v=1}^{V} \lambda_{i,v}(\tau) q_{i,v}(s_i) - \log Z(\lambda_{i,1}(\tau), \cdots, \lambda_{i,V}(\tau))$$

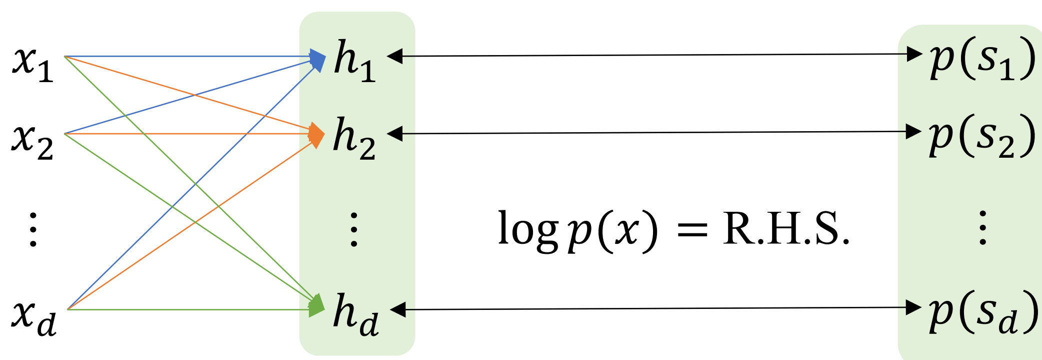
Nonlinear ICA

$$p(x)|\det(J_f)| = p(s) \quad \Leftrightarrow \quad p(x) = p(s)|\det(J_g)|$$

$$\log p(x) = \sum_{i=1}^{d} \log p(s_i) + \log|\det(J_g)|$$

One straightforward way:

😎 Notice that R.H.S. is the sum of d independent components. That is to say, it can expand a d dimensional vector space. More complex, easier to be identified (no free lunch).



Nonlinear ICA

$$p(x)|\det(J_f)| = p(s) \quad \Leftrightarrow \quad p(x) = p(s)|\det(J_g)|$$

$$\log p(x) = \sum_{i=1}^{d} \log p(s_i) + \log|\det(J_g)|$$

Two more easy problems:

😌 How to connect $\{h_1(x), h_2(x), \cdots, h_d(x)\}$ with $\log p(x)$?

😌 How to elimate the Jacobian term ? (Keep in mind that it will magically disappear.)

One terrible problem:

😵‍💫 How to identify $p(s_1), p(s_2), \cdots, p(s_d)$ from $\{h_1(x), h_2(x), \cdots, h_d(x)\}$?

Nonlinear ICA

Two more easy problems:

😌 How to connect $\{h_1(x), h_2(x), \cdots, h_d(x)\}$ with $\log p(x)$?

➢ *Logistic regression*

First, let us take the $t = 1$ (T classes totally) as pivot. Then, we have

$$p(t=1|x;\theta,w) = \frac{1}{1+\sum_{t=2}^{T} e^{w_t h(x;\theta)}} \quad p(t=\tau|x;\theta,w) = \frac{e^{w_\tau h(x;\theta)}}{1+\sum_{t=1}^{T} e^{w_t h(x;\theta)}}$$

$$p(t=\tau|x) = \frac{p_\tau(x)p(t=\tau)}{\sum_{t=1}^{T} p_t(x)p(t=\tau)} \qquad \text{where } p_\tau(x) = p(x|t=\tau)$$

With infinite examples, the above things lead to the relationship

$$\boxed{w_\tau h(x;\theta) = \log p_\tau(x)} - \log p_1(x) + \log\frac{p(t=\tau)}{p(t=1)}$$

😎

Nonlinear ICA

Two more easy problems:

😌 How to connect $\{h_1(x), h_2(x), \cdots, h_d(x)\}$ with $\log p(x)$?

The remaining question is:

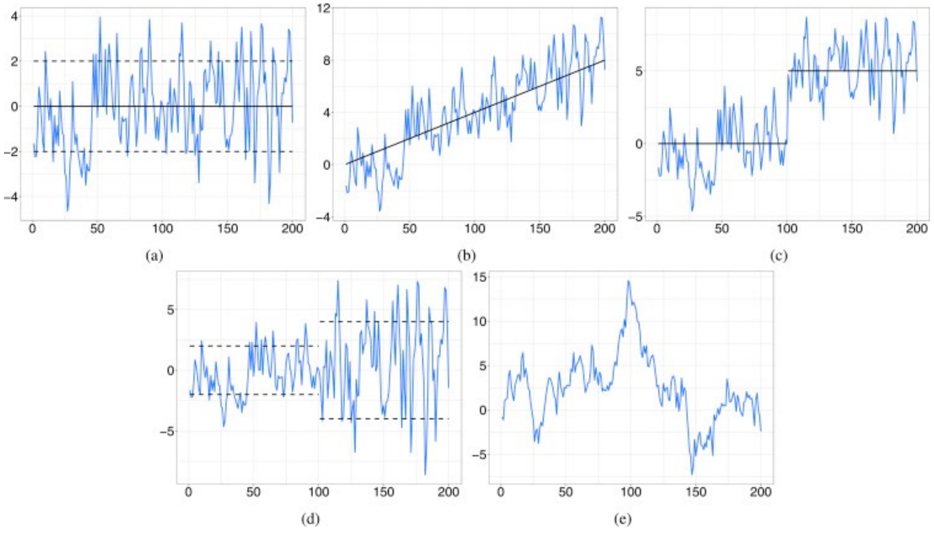
How can we constructure the (multinomial) logistic regression things?

Constrcure some classification tasks? 😄

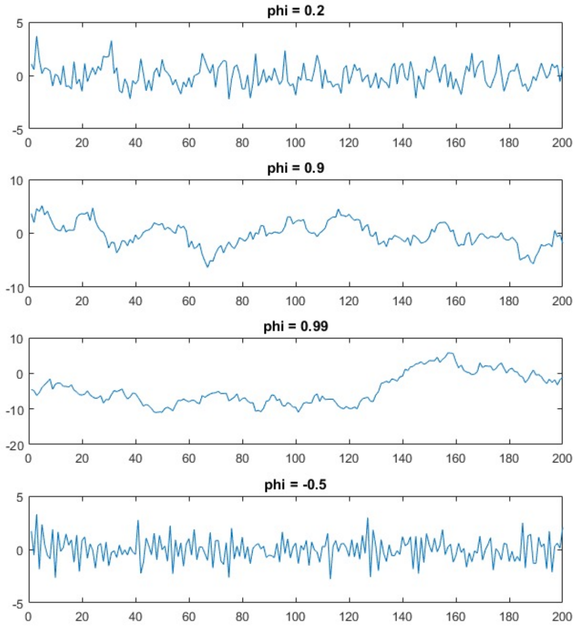
Unsupervised task 😔 Self-supervised? Yes! 😄

Nonlinear ICA

Temporal structural data:

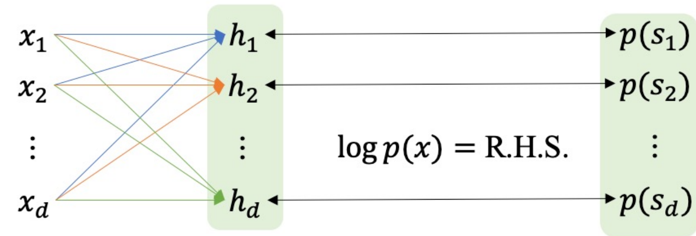


Nonstationary
(TCL, NIPS16)

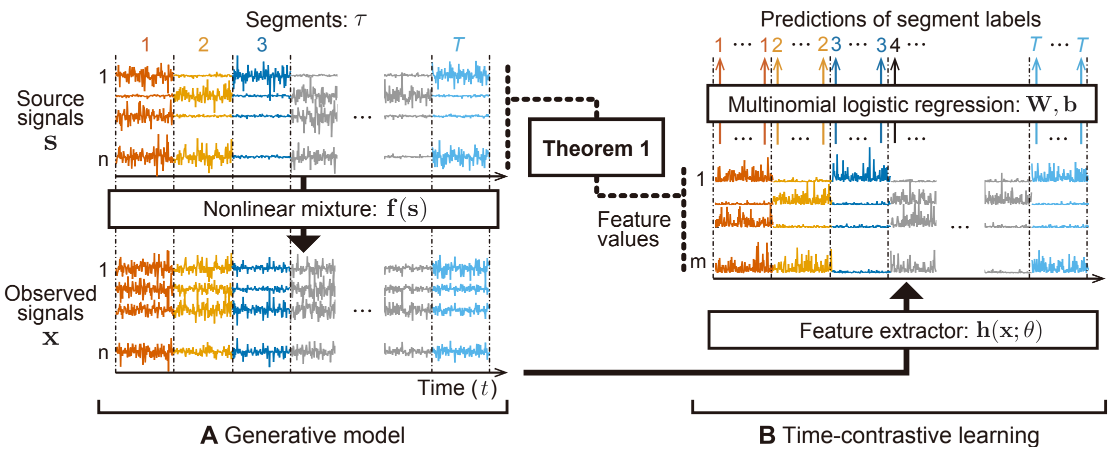


Autocorrelations
(PCL, AISTATS17)

Nonlinear ICA



- Temporal Contrastive Learning (TCL), NIPS16



The temporal data are assumed to be nonstationary and segmented to T parts, which leads to a T-classification task.

$$\log p(s_i) = q_{i,0}(s_i) + \sum_{v=1}^{V} \lambda_{i,v}(\tau) q_{i,v}(s_i) - \log Z(\lambda_{i,1}(\tau), \cdots, \lambda_{i,V}(\tau))$$

Theorem: The modulation parameter matrix $L_{\tau,i} = \lambda_{i,1}(\tau) - \lambda_{i,1}(1), \tau = 1, \cdots, T; i = 1, \cdots, N$ has full column rank N. Then, $q(s)$ can be identified up to an invertible linear transformation.

$$q(s) = Ah(x;\theta) + d$$

Nonlinear ICA

- Temporal Contrastive Learning (TCL), NIPS16

$$q(s) = Ah(x;\theta) + d$$

Proof Sketch:

GT: $$\log p_\tau(x) = \sum_{i=1}^{d} \lambda_{\tau,i} q(g_i(x)) + \log|\det(J_g)| - \log Z(\lambda_\tau)$$

Learning: $$\log p_\tau(x) = \sum_{i=1}^{d} w_{\tau,i} h_i(x) + \log p_1(x) - \log \frac{p(t=\tau)}{p(t=1)}$$

$$= \sum_{i=1}^{d} (w_{\tau,i} h_i(x) + \lambda_{1,i} q(g_i(x))) + \log|\det(J_g)| - \log Z(\lambda_1) - \log \frac{p(t=\tau)}{p(t=1)}$$

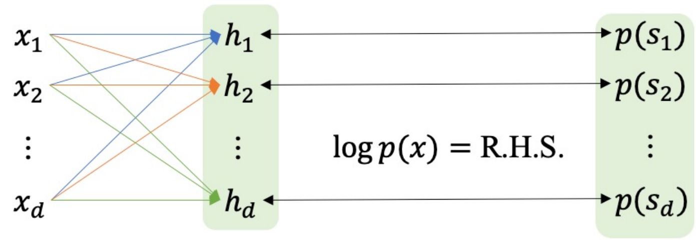
Then: $$\sum_{i=1}^{d} \boxed{(\lambda_{\tau,i} - \lambda_{1,i})} q(g_i(x)) = \sum_{i=1}^{d} w_{\tau,i} h_i(x) + \boxed{\log \frac{Z(\lambda_\tau)}{Z(\lambda_1)} - \log \frac{p(t=\tau)}{p(t=1)}}$$

$$Lq(s) = Wh(x) + \beta$$

We have $L^+ L = I$:

$$q(s) = L^+ W h(x) + L^+ \beta$$

Nonlinear ICA



- Permutation Contrastive Learning (PCL), AISTATS17

Autocorrelation describes the correlations encoded in the temporal structure.

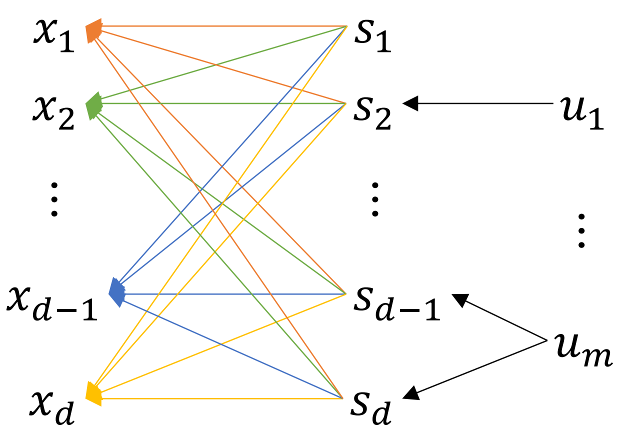
$$y(t) = \begin{pmatrix} x(t) \\ x(t-1) \end{pmatrix} \qquad y^*(t) = \begin{pmatrix} x(t) \\ x(t^*) \end{pmatrix}$$

$$r(y) = \sum_{i=1}^{d} B_i \left(h_i(y^1), h_i(y^2)\right)$$

In $y(t)$, $x(t)$ and $x(t-1)$ are correlated, so $\log p_1(y)$ models the joint distribution.
In $y^*(t)$, $x(t)$ and $x(t^*)$ are independent , so $\log p_2(y)$ can be decomposited.

Nonlinear ICA

➢ Generalized Contrastive Learning (GCL), AISTATS19



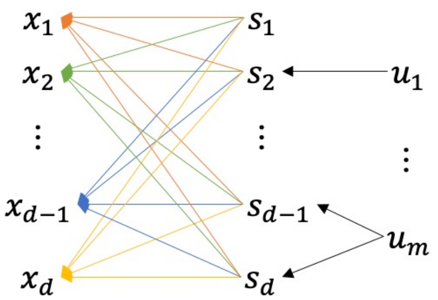
Assume that each s_i is statistically dependent on u, but conditionally independent of the other s_j:

$$\log p(s|u) = \sum_{i=1}^{d} q_i(s_i, u)$$

$$p(s_i|u) = \frac{Q_i(s_i)}{Z_i(u)} \exp\left[\sum_{i=1}^{d} \widetilde{q_{ij}}\,(s_i)\lambda_{ij}(u)\right]$$

the sufficient statistics $\widetilde{q_{ij}}$ are assumed linearly independent (over j for each i)

Nonlinear ICA

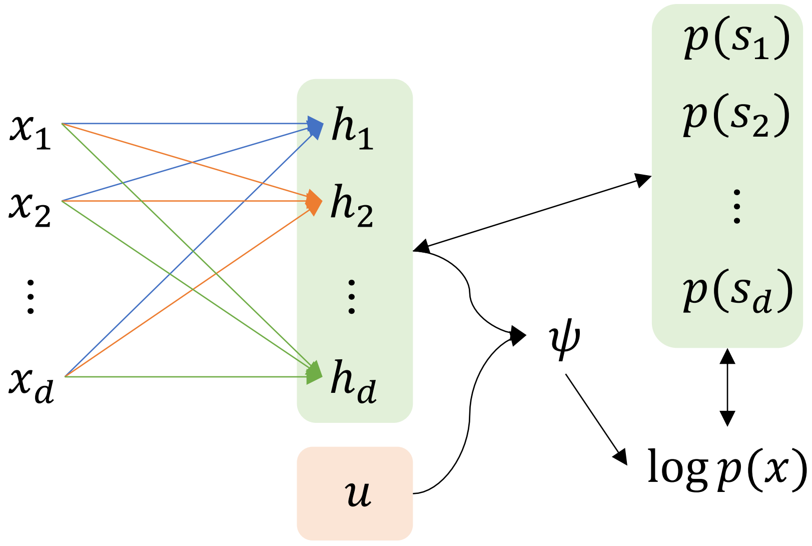


- Generalized Contrastive Learning (GCL), AISTATS19

Learning Algorithm:

$$\tilde{x} = (x, u) \qquad \tilde{x}^* = (x, u^*)$$

where u^* is a random value from the distribution of the u, but independent of x



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

Nonlinear ICA

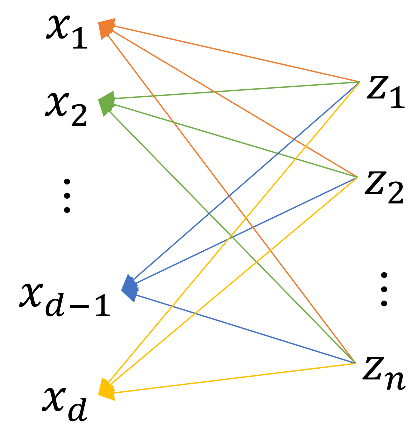
- Generalized Contrastive Learning (GCL), AISTATS19

Assumptions for identifiability:

1. The conditional log-pdf q_i is sufficiently smooth as a function of s_i for any fixed u.
2. [Assumption of Variability] For any $y \in \mathbb{R}^n$, there exists $2n + 1$ values for u, denoted by $u_j, j = 0, \cdots, 2n$ such that the $2n$ vectors in $\mathbb{R}^{2n}$ given by
   $$\big(\big(w(y, u_1) - w(y, u_0)\big), \big(w(y, u_2) - w(y, u_0)\big), \cdots, \big(w(y, u_{2n}) - w(y, u_0)\big)\big)$$
   with
   $$w(y, u) = \left(\frac{\partial q_1(y_1, u)}{\partial y_1}, \cdots, \frac{\partial q_n(y_n, u)}{\partial y_n}, \frac{\partial^2 q_1(y_1, u)}{\partial y_1^2}, \cdots, \frac{\partial^2 q_n(y_n, u)}{\partial y_n^2}\right)$$
   are linearly independent.

The functions $h_i(x)$ give the independent components, up to scalar (component-wise) invertible transformations.

Identifiable DGM



- Variational AutoEncoder (VAE)

$$\log p_\theta(x) = \log \int p_\theta(x,z)dz$$

$$= \log \int q_\phi(z|x)\frac{p_\theta(x,z)}{q_\phi(z|x)}dz = \log E_{z\sim q_\phi(z|x)}\frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$$\geq E_{z\sim q_\phi(z|x)}\log\frac{p_\theta(x,z)}{q_\phi(z|x)} = E_{z\sim q_\phi(z|x)}\log\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}$$

$$= E_{z\sim q_\phi(z|x)}\log p_\theta(x|z) - E_{z\sim q_\phi(z|x)}\log\frac{q_\phi(z|x)}{p_\theta(z)}$$

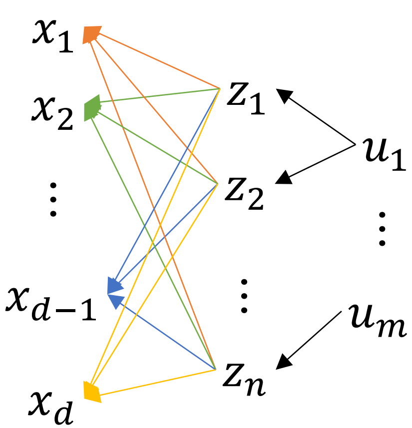
$$KL(q_\phi(z|x)||p_\theta(z))$$

Learning Objective:

$$\arg\max_{\phi,\theta}\mathbb{E}_{x\sim p(x)}[E_{z\sim q_\phi(z|x)}\log p_\theta(x|z) - KL(q_\phi(z|x)||p_\theta(z))]$$

Identifiable DGM

- Identifiable VAE(iVAE)



The conditional generative model is assume to be

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

$$p_f(x|z) = p_\epsilon(x - f(z))$$

The conditional pdf is thus given by:

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

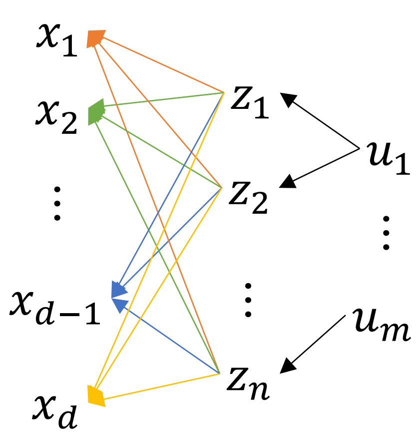
Then the ELBO for data log-likelihood is defined by:

$$\mathbb{E}_{\mathcal{D}}[\log p_\theta(x|u) \geq \mathcal{L}(\theta, \phi)] \coloneqq$$

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

Identifiable DGM

➢ Identifiable VAE(iVAE)



Definition:

$(f, T, \lambda) \sim (\tilde{f}, \tilde{T}, \tilde{\lambda}) \Leftrightarrow.$

$\exists\, A, c \mid T\left(f^{-1}(x)\right) = A\tilde{T}\left(\tilde{f}^{-1}(x)\right) + c, \forall x \in \mathcal{X}$

if A is invertible, we denote $\sim_A$-identifiable. If A is a block permutation matrix, we denote it by $\sim P$

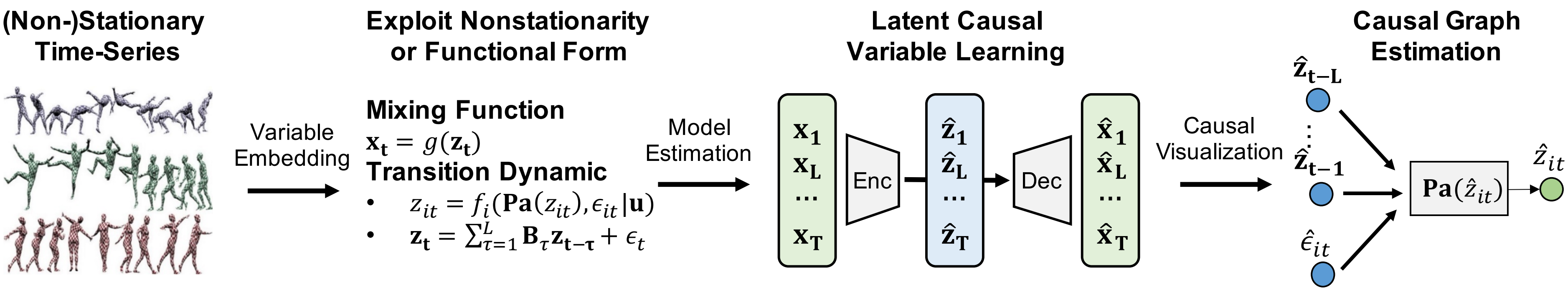
Assumptions for identifiability:

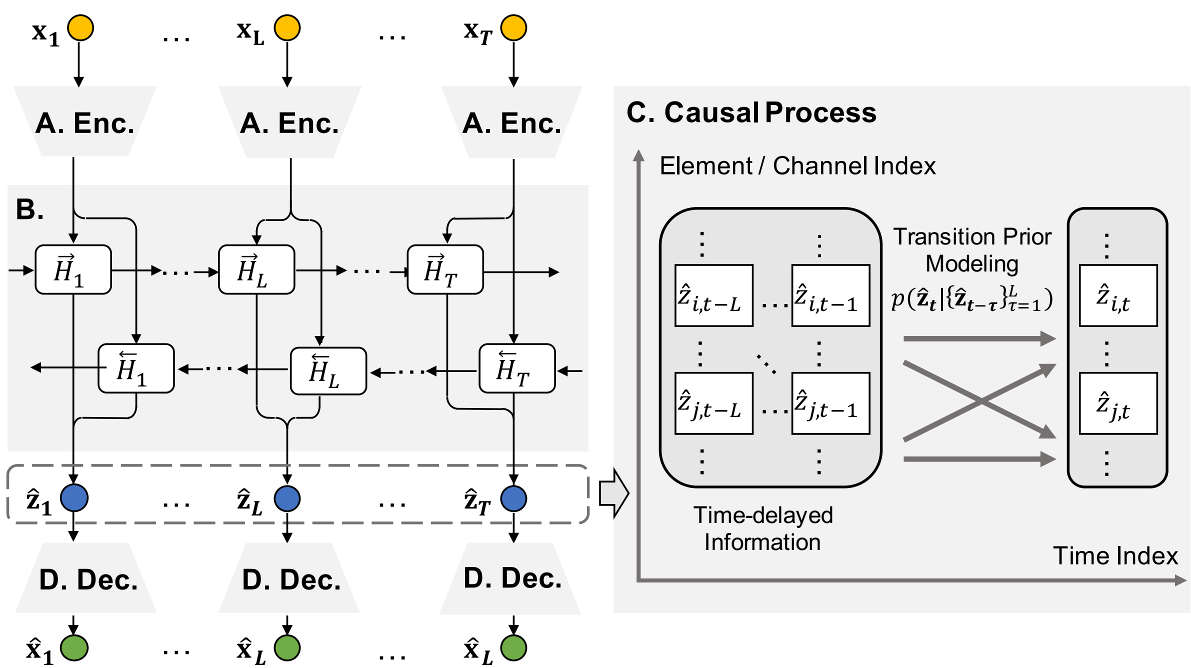
1. The sufficient statistics T_{ij} are differentiable almost everywhere and $\left(T_{ij}\right)_{i \le j \le k}$ are linearly independent.
2. There exists $nk + 1$ distinct points $u^0, \cdots, u^{nk}$ such that the matrix
   $L = (\lambda(u_1) - \lambda(u_0), \cdots, \lambda(u_{nk}) - \lambda(u_0))$
   of size $nk \times nk$ is invertible.

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

Identifiable DGM

➢ Causality + Nonlinear ICA





Thank you!