MLBBQ

VARIATIONAL AUTOENCODERS AND NONLINEAR ICA

A Unifying Framework

based on I. Khemakhem’s paper: https://arxiv.org/pdf/1907.04809.pdf

Rogers F. Silva, Ph.D.

TReNDS Center, GSU/Emory/GATech, Jun/11/2021

TReNDS Translational Research in Neuroimaging & Data Science

Georgia State University EMORY UNIVERSITY Georgia Tech

OUTLINE

- PART I – Background
    - Why Independence?
    - The linear success
    - The non-linear failure
    - Picking up the slack... and some auxiliary signals

- PART II – Unifying VAEs and Nonlinear ICA
    - Idea
    - Theory
    - Proofs… ya know.
    - Results

BACKGROUND

PART I

WHY INDEPENDENCE?

- Basically, it's for interpretability
  - Discuss whatever about latent variable (or source/component) WITHOUT regard for remaining latents
  - Decluttering
  - Caveat: Dependent sources (subspaces)

THE LINEAR SUCCESS

- Identifiability
  - Linear data generation model: independent sources were linearly mixed, yielding your data
  - Sources are provably recoverable: unique solution
  - Caveat: Gaussian sources, ambiguities (scale, order)

- MLE link
  - Grounded statistical framework
  - CRLB bounds, unbiased estimators, etc.

THE NON-LINEAR FAILURE

- NOT identifiable
  - Multiple (nonlinear) transformations yield independent latents

Hyvärinen and Pajunen (1999)

  - Which one is the right one? Non-unique

  - Independence alone is not enough.

  - Can’t learn structure behind the data

PICKING UP THE SLACK…
AND SOME AUXILIARY SIGNALS

- New theory:
  - Bring in some additional information (aka, auxiliary variables)

  - Then, CONDITIONED on those variables, independence does the trick!

- Examples:
  - labels
  - temporal/spatial structure (conditional sampling)
  - (non)stationarity (distribution changes)

QUESTIONS

… or curiosities?

UNIFYING VAES and NONLINEAR ICA

PART II

IDEA

- VAEs are efficient (but I think they mean effective)
  - learns latents **z** s.t. $$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} \approx p(\mathbf{x}) \qquad (2)$$

    marginal model data (true/unknown)
  - But learning prior and joint is impossible (non-identifiable) - non-linear case

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

  - This paper: show it is possible for a broad class of deep-latent models.
  - Fancy word of the day: disentanglement (spoiler: it's just source separation)
  - REQUIREMENT: prior factorizes when conditioned on auxiliary variable
  - Includes: undercomplete (#mix > #source), noise, MLE. Special case: Flow model

IDEA

- Learning joint implies learned latent’s true prior and posterior
- Only if model is identifiable
- Original VAE:
  - theory is insufficient to determine identifiability conditions
  - does suffice to enable parameter optimization s.t. $p_\theta(\boldsymbol{x}) \approx p(x)$
  - no guarantee that $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ is correctly estimated
- Disentanglement: no proofs. β-VAE: hyperparam. encourage disentag.
- GAN+independence: non-identifiable bc no aux. vars.
- Nonlinear ICA: invertible nonlin. transfo. BUT NO data distn. model and NO data synthetization

IDEA

- Show that VAE joint is identifiable and learnable
- Bridges gap between VAE and nonlinear ICA
- Unified view of two complementary unsup. representation learning methods
- How:
    - Latent prior factorizes conditioned on aux. vars.
- Not limited to VAE (but VAE allows efficient latent inference and scales well)
- Beats other models in simulations

QUESTIONS

THEORY

- Unidentifiability of Deep Latent Models
- An identifiable model based on <u>conditionally factorial priors</u>
- Identifiability Theory

THEORY UNIDENTIFIABILITY

- θ are the decoder parameters
- Data are observations of x according to:

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

- Equivalently, $\mathbf{x}^{(i)} \sim p_{\boldsymbol{\theta}^*}(\mathbf{x})$
- VAE: MLE of marginal, yielding $p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x})$
- VAE learns
  - Full generative model: $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z})$, <u>joint = decoder*prior</u>
  - Inference model (posterior proxy): $q_\phi(z|x) \approx p_\theta(z|x)$, <u>encoder</u>
- EXCEPT FOR $p_\theta(\boldsymbol{x})$, it's all <u>MEANINGLESS</u> bc:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \quad \not\Longrightarrow \quad \boldsymbol{\theta} = \boldsymbol{\theta}' \quad \forall(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

same p from different θ,
thus, different joints.
want opposite ($\theta' = \theta^*$)

THEORY UNIDENTIFIABILITY

- Key reason: $p_{\boldsymbol{\theta}}(\mathbf{z})$ is unconditional

- Toy example: spherical Gaussian p(z)
  - rotation does NOT change p(z) or p(x)
  - DOES change z → change x → changes p(x|z)

- Proofs in Supplement D. Sketch:
  - z of any distn. → sequentially to Gaussian, indep. of previous → 1st still unmixed
  - transform to spherical Gaussian, rotate, transform back
  - p(z) is unchanged, but p(x|z) change
  - Extreme case: transform x and, oddly, x_i is component, but still unmixed.

QUESTIONS

THEORY
CONDITIONALLY FACTORIAL PRIORS
MODEL DEFINITION

- Identifiable VAE (iVAE)
- Conditional generative (decoder) model: $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})$
  $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$
- Independent noise: $p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) = p_{\boldsymbol{\varepsilon}}(\mathbf{x} - \mathbf{f}(\mathbf{z}))$
  $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}$

- NN to approx. f
- SUPPLEMENT C for discrete variables.

- Noiseless case when noise variance → 0 (link to flow models: f is invertible flow)

THEORY CONDITIONALLY FACTORIAL PRIORS MODEL DEFINITION

- p(z|u) factorizes, not p(z): each z_i ~ exponential family distn. (univ. approx.)
  - Qi: base measure
  - Zi(u) normalizing cst
  - Ti: sufficient stats.
  - $\lambda_i(\mathbf{u})$: k-dim exp. fam. params. (arbitrary: look-up, NN, etc)
  - k: fixed, not estimated

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

- Univariate Gaussian in exponential family form:

$$p_{\mathbf{T},\boldsymbol{\lambda}}(z|\mathbf{u}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{z^2}{2\sigma^2})\exp(-\frac{\mu^2}{2\sigma^2})\exp(-\frac{z\mu}{\sigma^2}), \mathbf{u} = (\mu, \sigma)^{\mathsf{T}}$$

$$Q(z) = \frac{1}{\sqrt{2\pi}}, Z(\mathbf{u}) = \sigma\ \exp(\frac{\mu^2}{2\sigma^2})$$

$$\textstyle\sum_{j=1}^{2} T_j(z)\lambda_j(\mathbf{u}) = \left(-\frac{z^2}{2}\frac{1}{\sigma^2}\right)_{j=1} + \left(z\frac{\mu}{\sigma^2}\right)_{j=2}$$

THEORY
CONDITIONALLY FACTORIAL PRIORS
VAE ESTIMATION

- Simultaneously learn:
  - deep latent generative model (decoder)
  - variational approximation $q_\varphi(z|x, u)$ of its true posterior $p_\theta(z|x, u)$ (encoder)
- $p_\theta(x|u) = \int p_\theta(x,z,|u)dz$ the conditional marginal → Gaussian(mu,var)
- $q_D(x,u)$: the empirical data distribution given by dataset D
- Maximize L(θ, φ), a lower bound on the data log-likelihood :

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

- Reparameterization trick (Kingma and Welling, 2013) to sample from $q_\varphi(z|x,u)$
  - low-variance stochastic estimator for gradients wrt φ
- Latent estimates: sample from variational posterior

THEORY
CONDITIONALLY FACTORIAL PRIORS
VAE ESTIMATION

[2] 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.

THEORY
CONDITIONALLY FACTORIAL PRIORS
INDETERMINACY - WEAK

- k = 1

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

- point-wise (component-wise) transformations
- linear relation to original z*

- excluding families with location-only changes: A is a permutation matrix (i.e., inconsequential)

QUESTIONS

THEORY
CONDITIONALLY FACTORIAL PRIORS
INTERPRETATION AS NONLINEAR ICA

- Noiseless nonlinear-ICA with same dimensions:
  - $x = f(z)$, factorial $p(z)$
  - deep generative model (decoder)
  - degenerate posteriors

- Identify f^{-1} based on x alone
  - not attainable by deep latent models

- Identifiability requires:
  - restricting f (e.g., linear)
  - constrain $p(z)$ (e.g., TCL, PCL, GCL)

THEORY
CONDITIONALLY FACTORIAL PRIORS
INTERPRETATION AS NONLINEAR ICA

- Differences wrt GCL:
  - posteriors are not degenerative (noisy case) → VAE connection
  - principled: MLE in terms of ELBO (GCL used self-supervision heuristics)
  - ELBO is useful for model selection and validation
  - SUPPLEMENT F: links MLE to maximizing latent independence
  - Learn both FW and BW models: can recover latents from data, and generate new data
  - FW model: study meaning of latents
  - stronger identifiability theory ($n < d$, noise)
  - SUPPLEMENT G: further discussion

THOERY IDENTIFIABILITY

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).

THOERY IDENTIFIABILITY

- Identifiability up to an equivalence:

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

- Define:

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

THOERY IDENTIFIABILITY

Linear Indeterminacy

AFTER iVAE

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 $\mathbf{u}^0, \ldots, \mathbf{u}^{nk}$ such that the matrix

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

of size $nk \times nk$ is invertible.[5]

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

THOERY IDENTIFIABILITY

Reduction to

Permutation Indeterminacy

AFTER iVAE

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*[6].

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

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

THOERY IDENTIFIABILITY

Gaussian case is HOPELESS.

Cannot reduce to

Permutation Indeterminacy

AFTER iVAE

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.*

Univariate Gaussian in exponential family form:

$$p_{\mathbf{T},\boldsymbol{\lambda}}(z|\mathbf{u}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{z^2}{2\sigma^2})\exp(-\frac{\mu^2}{2\sigma^2})\exp(-\frac{z\mu}{\sigma^2}), \mathbf{u} = (\mu, \sigma)^{\mathsf{T}}$$

$$Q(z) = \frac{1}{\sqrt{2\pi}}, Z(\mathbf{u}) = \sigma \exp(\frac{\mu^2}{2\sigma^2})$$

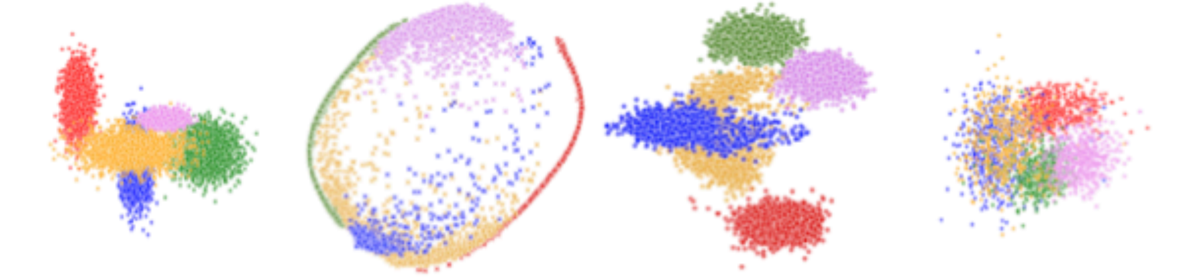
$$\textstyle\sum_{j=1}^{2} T_j(z)\lambda_j(\mathbf{u}) = \left(-\frac{z^2}{2}\frac{1}{\sigma^2}\right)_{j=1} + \left(z\,\frac{\mu}{\sigma^2}\right)_{j=2}$$

PROOFS… YA KNOW.

- SOME OTHER TIME…

RESULTS

- SIMULATIONS: like TCL
  - time course sources with changing distn. params over segments(windows), i.e., non-stationary sources.
  - MLP to mix sources
  - + sensor noise



(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) $p_{\mathrm{VAE}}(\mathbf{z}|\mathbf{x})$

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.

END

prematurely…