Identifying metric structures of deep latent variable models

Stas Syrota♣, Yevgen Zainchkovskyy♣, Johnny Xi♦, Benjamin Bloem-Reddy♦, Søren Hauberg♣

♣ Technical University of Denmark, ♦ University of British Columbia

Blogpost
Paper
Code

ICML International Conference On Machine Learning

1. Motivation

Applied scientists care about the latent space to make scientific discoveries

Latent space is not identifiable (VAEs, NFs, CNFs)

Problem

Scientific discoveries and downstream applications are unreliable

Current solutions

Disregard geometry, focus on parameters, constrain the model, require extra labeled data

Key observation

Exploration of the latent space is often relational (based on distances, angles, etc.)

OUR SOLUTION

IDENTIFIABLE GEOMETRIC RELATIONS IN THE LATENT SPACE via pullback geometry

2. Problem setup

a Deep latent variable models and Identifiability

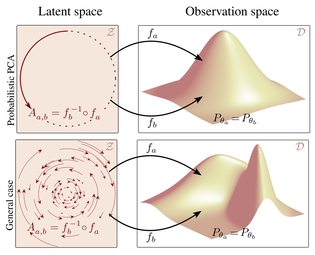
- **Deep latent variable models** learn densities of data $\mathrm{X} \in \mathscr{D}$ induced by latent variables $\mathrm{Z} \in \mathscr{Z}$.
- An **identifiable** model means that we can uniquely determine the latent variables Z from data.
- **Indeterminacy transformations** are maps between any pair of latent spaces of equivalent models. They are the underlying causes of non-identifiability of our models.

TL;DR | **Definitions**

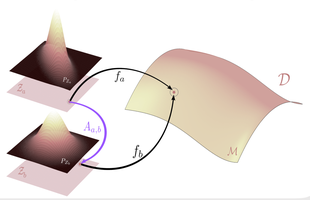
Model parameters $\theta = (f, P_Z)$ define the density of data: $P_\theta(\mathrm{X}) = \int P(\mathrm{X} \mid f(\mathrm{Z}))P_Z d\mathrm{Z}$
by a generator (decoder) function $f : \mathscr{Z} \to \mathscr{D}$ and the distribution of the latent variables P_Z

Two parameterizations $\theta_a = (f_a, P_{Z_a})$ and $\theta_b = (f_b, P_{Z_b})$ are equivalent if $P_{\theta_a} = P_{\theta_b}$ and lead to the equivalence class $[\theta] = \{\theta' : P_\theta = P_{\theta'}\}$. The model is **identifiable** if $[\theta]$ is a singleton. ($P_\theta(X) := P_\theta$ for shorthand).

Given parameterizations $\theta_a = (f_a, P_{Z_a})$ and $\theta_b = (f_b, P_{Z_b})$ with $P_{\theta_a} = P_{\theta_b}$, an **indeterminacy transformation** at (θ_a, θ_b) is a function $A_{a,b} : Z_a \to Z_b$ s.t. $f_a \circ A_{a,b}^{-1} = f_b$.



All indeterminacy transformations $A_{a,b} : \mathscr{Z}_a \to \mathscr{Z}_b$ of a generative model are a.e. equal to $A_{a,b}(z) = f_b^{-1} \circ f_a(z)$ [1].



Assumptions. All decoder functions f are:
A2 injective; **A3** have full rank Jacobian;
A4 have the same image $\mathscr{M} \in \mathscr{D}$

2. Problem setup (cont'd)

b Latent space geometry

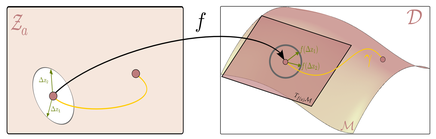
Pullback metric represents the metric structure of the manifold $\mathscr{M} \in \mathscr{D}$ wrt. local coordinates in latent space $\mathscr{Z}$ by considering local neighborhoods of a point.

In a local neighborhood of $\mathrm{z} \in \mathscr{Z}$, we can approximate f using Taylor,
$f(\mathrm{z} + \Delta\mathrm{z}) \approx f(\mathrm{z}) + \mathbf{J}_z \Delta\mathrm{z}$ with Δz denoting a small perturbation.

Given two small perturbations around z, $\Delta z_1, \Delta z_2$, we can compute the inner product by:

$$\begin{aligned} \|f(\mathrm{z} + \Delta\mathbf{z}_1) - f(\mathrm{z} + \Delta\mathbf{z}_2)\|^2 &= \|f(\mathrm{z}) + \mathbf{J}_z\Delta\mathbf{z}_1 - f(\mathrm{z}) - \mathbf{J}_z\Delta\mathbf{z}_2\|^2 \\ &= (\Delta\mathbf{z}_1 - \Delta\mathbf{z}_2)^\top \mathbf{J}_z^\top \mathbf{J}_z (\Delta\mathbf{z}_1 - \Delta\mathbf{z}_2) \\ &= (\Delta\mathbf{z}_1 - \Delta\mathbf{z}_2)^\top \mathbf{g}(\mathrm{z}) (\Delta\mathbf{z}_1 - \Delta\mathbf{z}_2) \end{aligned}$$

where $\mathbf{g}(\mathrm{z})$ denotes the pullback metric that at each $\mathrm{z} \in \mathscr{Z}$ assigns a symmetric positive definite matrix defining an inner product.



3. Main results

a Pullback metric is identifiable

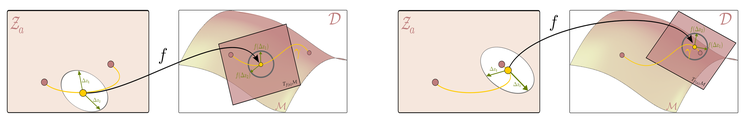
Theorem 1: Let $\theta_a = (f_a, P_{Z_a})$ and $\theta_b = (f_b, P_{Z_b})$ be equivalent models ($P_{\theta_a} = P_{\theta_b}$) with the associated pullback metrics $\mathbf{g}_a$ and $\mathbf{g}_b$. Then all the possible indeterminacy transformations are isometries. I.e.:

$$A_{a,b}{}^{*}\mathbf{g}_a = \mathbf{g}_b$$

* denotes the pushforward

This makes **lengths of curves, angles, volumes, Ricci curvature tensor, optimal transport, geodesics, logarithmic and exponential maps identifiable.**

In particular, identifiable distances for downstream tasks



Corollary 1: Let $\theta_a = (f_a, P_{Z_a})$ and $\theta_b = (f_b, P_{Z_b})$ be equivalent models ($P_{\theta_a} = P_{\theta_b}$) with the associated pullback metrics $\mathbf{g}_a$ and $\mathbf{g}_b$. Pick x_1 and $\mathrm{x}_2 \in \mathscr{M}$, then the **geodesic distance between the latent codes**, $\mathrm{z}_1^a = f_a^{-1}(\mathrm{x}_1)$ and $\mathrm{z}_2^a = f_a^{-1}(\mathrm{x}_2)$ is **identifiable**. I.e.:

$$d_{\mathbf{g}_a}(\mathbf{z}_1^a, \mathbf{z}_2^a) = d_{\mathbf{g}_b}(A_{a,b}(\mathbf{z}_1^a), A_{a,b}(\mathbf{z}_2^a)) = d_{\mathbf{g}_b}(\mathbf{z}_1^b, \mathbf{z}_2^b)$$

where **geodesic distance** is defined to be:

$$d_{\mathbf{g}_a}(\mathbf{z}_1^a, \mathbf{z}_2^a) = \inf_\gamma \int_0^1 |\gamma'(t)|_{\mathbf{g}_a} \mathrm{d}t$$

that is the curve $\gamma \in \mathscr{Z}_a$ with the lowest **energy** satisfying $\gamma(0) = \mathrm{z}_1^a$ and $\gamma(1) = \mathrm{z}_1^b$.

b Euclidian means flat, no matter how we get it

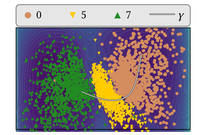
Corollary 2: Let $\mathscr{Z}_a$ be a latent space, f_a a decoder and $\mathrm{g}_\mathbb{E}$ a metric on $\mathscr{Z}_a$ that is (proportionally) Euclidean. If we set $\mathrm{g}_a = \mathrm{g}_\mathbb{E}$, then g_a can only be identifiable if $f_a(\mathscr{Z}_a) = \mathscr{M}$ is a flat manifold.

Euclidean identifiability through multiple views [4,5] and model restrictions [6] only works if the target manifold is assumed flat.

4. Experiments

Demonstrate reliable distances in the latent space without model restrictions or extra data

- For each **dataset**:
  1. Randomly pick **100 point pairs** from the test set
  2. Train **30 models** with different initializations
  - For each **point pair:**
    - For each **model:**
      1. Encode the points in the latent space
      2. Measure the Euclidean distance between the points
      3. Measure the geodesic distance between the points
    - For each **distance:**
      - Compute coefficient of variation: $\mathrm{CV}(\text{point pair}) = \dfrac{\text{mean over 30 measurements}}{\text{std. deviation over 30 measurements}}$
  - Return coefficients of variation for the Euclidean and geodesic distances for 100 point pairs

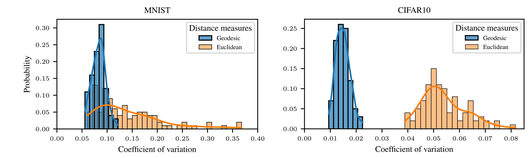


Geodesic between two points in 2D latent space for MNIST model

Theory validated if geodesic distances show lower coefficient of variation

a Injective decoder models (MNIST & CIFAR10)

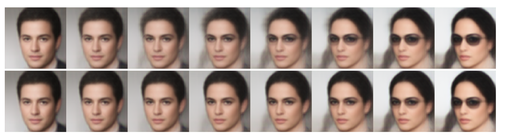
- An injective decoder/encoder inspired by the $\mathscr{M}$-flow architecture using Normalizing Flows [3].
- Geodesics efficiently parameterized by a natural splines with parameters trained by optimizing discretized curve energy using gradient descent.
- To account for stochasticity in manifold estimation we use an ensemble of decoders to compute geodesics wrt. following [2].



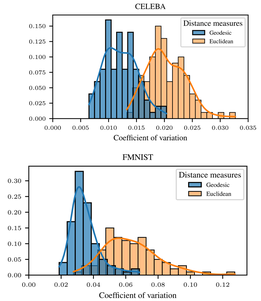
Histograms of coefficients of variation for the two datasets. Geodesic distance measure shows a narrower distribution with lower mean.

b Non-injective decoder models (FMNIST & CelebA)

- Non-injective deep CNN decoder architecture
- Separate deep CNN encoder
- Verifiable **A3** assumption (full rank Jacobian)
- Geodesics computation as in (a) above



(Left): Interpolation between two CelebA datapoints. Top: following the geodesic: bottom: following the straight line in the latent space.
(Right): Histograms of coefficients of variation for the two datasets. Geodesic distance measure shows a narrower distribution with lower mean.



5. Conclusions

- **Strong theoretical identifiability guarantees the pullback metric: distances, angles, volumes, logarithmic and exponential maps, etc.**
- **Does not require extra data, model restrictions or special training procedures (post hoc.).**
- **Fully compatible with domain specific metrics and modern architectures.**

References:
[1] Xi, Q. & Bloem-Reddy, B.. (2023). Indeterminacy in Generative Models: Characterization and Strong Identifiability. Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 206:6912-6939 Available from https://proceedings.mlr.press/v206/xi23a.html.
[2] Syrota, S., Moreno-Muñoz, P. & Hauberg, S.. (2024). Decoder ensembling for learned latent geometries. Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM), in Proceedings of Machine Learning Research 251:277-285 Available from https://proceedings.mlr.press/v251/syrota24a.html.
[3] Brehmer, J., & Cranmer, K. (2020). Flows for simultaneous manifold learning and density estimation. *Advances in neural information processing systems*, *33*, 442-453.
[4]Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., & Schölkopf, B. (2020, August). The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence* (pp. 217-227). PMLR.
[5] Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., & Tschannen, M. (2020, November). Weakly-supervised disentanglement without compromises. In *International conference on machine learning* (pp. 6348-6359). PMLR.
[6] Khemakhem, I., Kingma, D., Monti, R., & Hyvarinen, A. (2020, June). Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics* (pp. 2207-2217). PMLR.