Distributional Autoencoders Know the Score

Andrej Leban
Department of Statistics, University of Michigan

"Nonlinear PCA that learns the data score":

Result 1 — Geometry: The level sets align *exactly* with the score in the normal directions.

Result 2 — Dimensionality: Encoding dimensions beyond the data manifold are *completely* uninformative.

Motivation

Unsupervised learning method card

1) **PCA:** linear, ordered components, mean reconstructions only.
2) **Autoencoder:** non-linear encodings, no ordering, mean *tendency* for reconstruction.
3) **Distributional Principal Autoencoder (DPA)** [2]: non-linear, ordered components **and** distribution-faithful reconstructions.

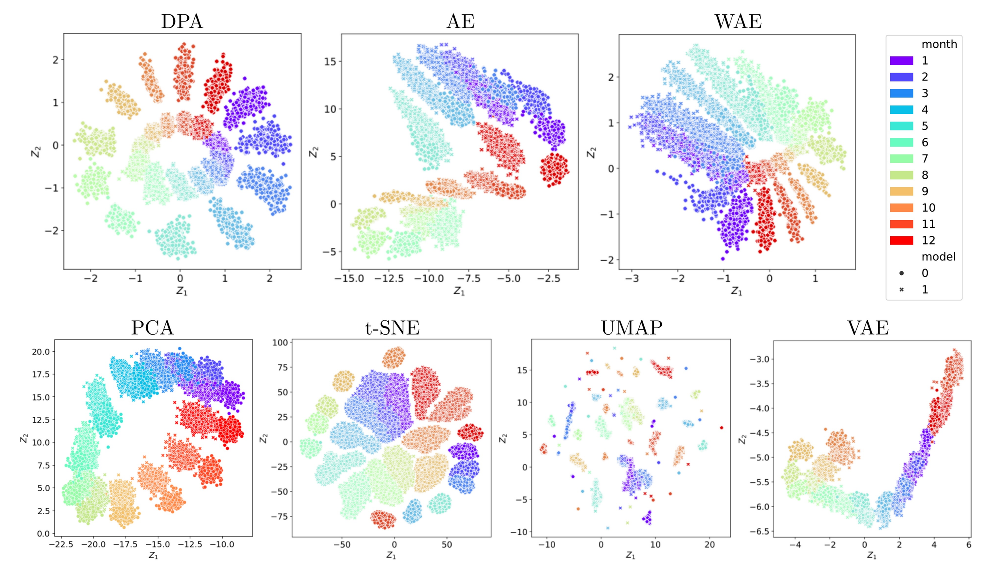
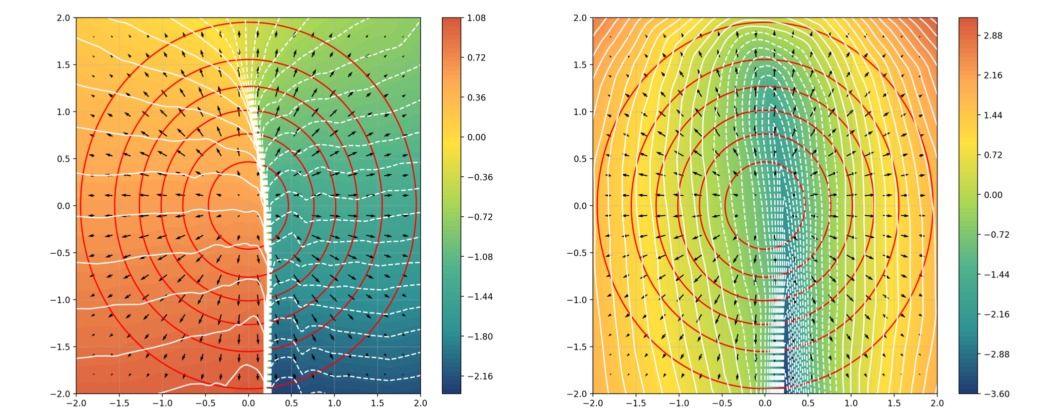


Figure 9 in [2]: DPA disentangles spatial and temporal dimensions for monthly (i.e., *periodic*) global precipitation data.



Rotationally symmetric data: *Left* component has level sets tangential; *right* component has level sets normal to the score. Together, they recover polar coordinates.

Key Results

The Distributional Principal Autoencoder (DPA) is an autoencoder variant recently introduced in Shen and Meinshausen [2] whose goal is **distributionally-correct reconstruction** of *all the data mapped to a single value by the encoder* – the encoder *level set*. Thus, for a sample $x \sim P_{\text{data}}$, an optimal encoder $e^*: \mathbb{R}^p \to \mathbb{R}^k$, and an optimal *stochastic* decoder d^* padding the extra dimensions with noise $\epsilon \sim \mathcal{N}(0, I_{p-k})$, the following must hold:

$$d^*\big(e^*(x), \epsilon\big) \stackrel{d}{=} \text{Law}(X \mid e^*(X) = e^*(x))$$

The encoder–decoder optimization objective is based on the *energy score*:

$$(e^*, d^*) \in \underset{e,d}{\arg\min} \sum_{k=0}^{p} \mathbb{E}_X\Big[\mathbb{E}_{Y \sim P_{d,e_{1:k}(X)}}[\|X-Y\|^\beta]\Big] - \tfrac{1}{2}\mathbb{E}_X\Big[\mathbb{E}_{Y,Y' \stackrel{\text{iid}}{\sim} P_{d,e_{1:k}(X)}}[\|Y-Y'\|^\beta]\Big] \triangleq \sum_{k=0}^{p} L_k[e,d],$$

where $P_{d,e_{1:k}(X)}$ is the reconstructed distribution when using (only) the first k components of e.

Result 1: Geometry aligns exactly with the data score

Theorem

For $\beta = 2$ and under relatively mild assumptions we have, for almost every sample $X \sim P_{\text{data}}$ and encoder level set $\mathcal{L}_{e^*(X)}$, the following balance equation for almost every $y \in \mathcal{L}_{e^*(X)}$:

$$\boxed{\frac{2\big(y - c(X)\big)}{\frac{V(X)}{Z(X)} - \|y - c(X)\|^2} D_{e^*}^\top(y) \;=\; s_{\text{data}}(y)\, D_{e^*}^\top(y),}$$

*where $s_{\text{data}}(y) \triangleq \nabla_y \log P_{\text{data}}(y)$ is the Stein score and $D_{e^*}(y)$ the encoder Jacobian at y, whenever the following quantities: the **level-set center-of-mass**:*

$$c(X) = \frac{1}{Z(X)} \int y\, P_{data}(y)\, \delta(e(y) - e(X))\, dy,$$

*and the **level-set variance**:*

$$V(X) = \int \|y - c(X)\|^2\, P_{data}(y)\, \delta(e(y) - e(X))\, dy$$

*are finite, and the **level-set mass** $Z(X) = \int P_{data}(z)\, \delta(e(z) - e(X))\, dz > 0$.*

Result 2: Extra dimensions are completely uninformative

Definition (K'-parameterizable manifold & K'-best-approximating encoder)

A K-dimensional manifold is K'-parameterizable, $K' \geq K$, if for the optimal encoder/decoder, the K'-term in the loss is globally the smallest among all terms and among all encoder/decoder pairs:

$$L_{K'}[(e^*, d^*)] = \min_{e,d,k} L_k[e,d]$$

If a solution (e^*, d^*) satisfying the above is also optimal among all dimension-K' encoders:

$$(e^*, d^*) \in \underset{e,d}{\operatorname{argmin}} \sum_{k=0}^{K'} L_k[e,d],$$

we denote it as the K'-best-approximating encoder.

Theorem (Extra dimensions are completely uninformative)

For a K'-parameterizable manifold, the dimensions $(K'+1, \cdots, p)$ of the K'-best-approximating encoder obey:

$$P_{d^*, e^*_{1:k}(X)} = P_{d^*, e^*_{1:K'}(X)}, \quad \text{for } k \in [K'+1, \ldots, p].$$

Furthermore, these dimensions are conditionally independent of the data X, given the relevant components $(e_1^*, \cdots, e_{K'}^*)$,

$$\boxed{X \perp\!\!\!\perp e^*_{K'+i}(X) \mid e^*_{1:K'}(X), \qquad \forall i \in [1, \ldots, p-K'].}$$

or equivalently, they carry no additional information about the data distribution:

$$I\left(X; e^*_{K'+i}(X) \mid e^*_{1:K'}(X)\right) = 0, \qquad \forall i \in [1, \ldots, p-K'],$$

Consequences

1. **Free lunch:** Typically, *data approximation/reconstruction* and *dimensionality reduction/disentanglement* represent a **tradeoff**. Example – β-VAE:

$$\underset{\theta,\varphi}{\arg\min}\ \mathbb{E}_{p_{\text{data}}(x)}\Big[\underbrace{\mathbb{E}_{q_\varphi(z|x)}[-\log p_\theta(x \mid z)]}_{\text{reconstruction}} + \underbrace{\beta}_{\text{tradeoff}} \underbrace{\text{KL}\Big(q_\varphi(z \mid x) \,\|\, \textstyle\prod_j p(z_j)\Big)}_{\text{disentanglement}}\Big]$$

Here, both main results hold **simultaneously**.

2. **Immediate Scientific impact:** In chemical applications, the data is typically distributed by the *Boltzmann distribution*, whose score is proportional to the force, which is recovered by the encoding:

$$\vec{F}(y)\, D_{e^*}^\top \;=\; 2\, k_B T\, \frac{y - c(X)}{\frac{V(X)}{Z(X)} - \|y - c(X)\|^2} D_{e^*}^\top(y),$$

After encoding data from, e.g., the Müller-Brown potential, if one starts in a potential minimum and moves with the encoding, one recovers the *Minimum Free Energy Path* (MFEP) – "least-energy-costly" transition between states. The encoding from raw data can hence be used to "guide" subsequent simulations for, e.g., *protein folding*.

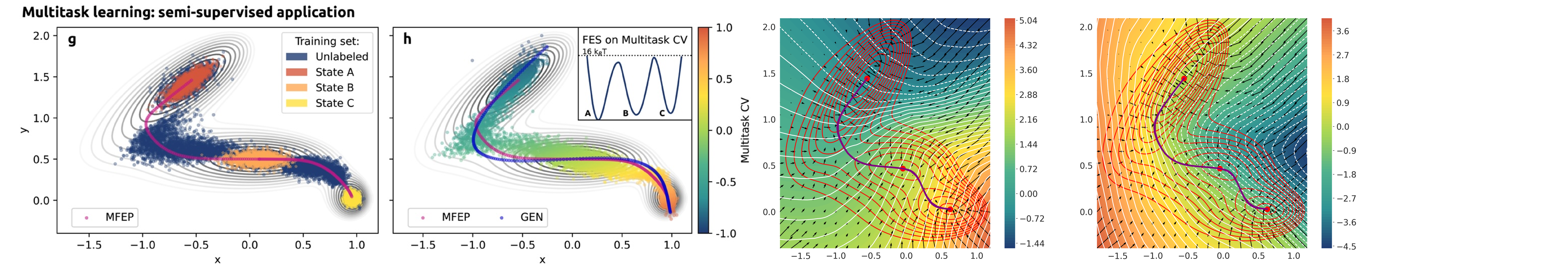


Figure 1. Müller-Brown potential. **Left:** the MFEP approximation from using **labeled** data in [1]. **Right:** the first two components of the DPA – **single** encoding of **unlabeled** trajectories.

References

[1] L. Bonati, E. Trizio, A. Rizzi, and M. Parrinello. A unified framework for machine learning collective variables for enhanced sampling simulations. *The Journal of Chemical Physics*, 159.

[2] X. Shen and N. Meinshausen. Distributional Principal Autoencoders, Apr. 2024.