Distributional Autoencoders Know the Score

(NeurIPS 2025)

Andrej Leban

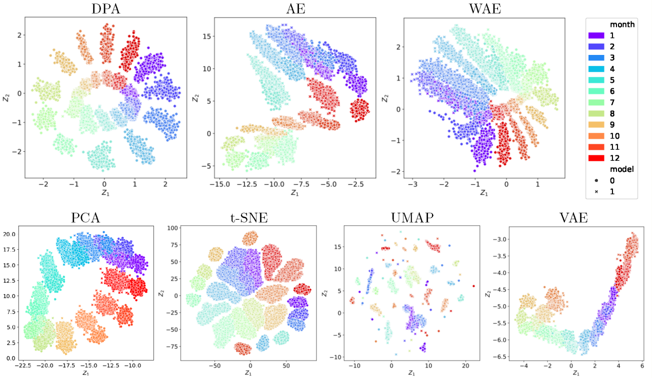
leban@umich.edu

Department of Statistics,
University of Michigan

"Nonlinear PCA that learns the data score."

Unsupervised learning method card

1) **PCA:** linear, ordered components, mean reconstructions only.
2) **AE:** non-linear encodings, no ordering, mean *tendency* for reconstruction.
3) **DPA:** non-linear, ordered components **and** distribution-faithful reconstructions.



Introduction to DPA

The goal of the *Distributional Principal Autoencoder* (DPA) [Shen and Meinshausen, 2024] is distributionally-faithful reconstruction of *all data (X) mapped to the same value by the encoder (e):*

$$P^*_{e,x} = \mathrm{Law}(X \mid e(X) = e(x)).$$

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

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

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

First main result: geometry aligns exactly with the data score ($\beta = 2$)

Theorem 1

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 $y \in \mathcal{L}_{e^*(X)}$:

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

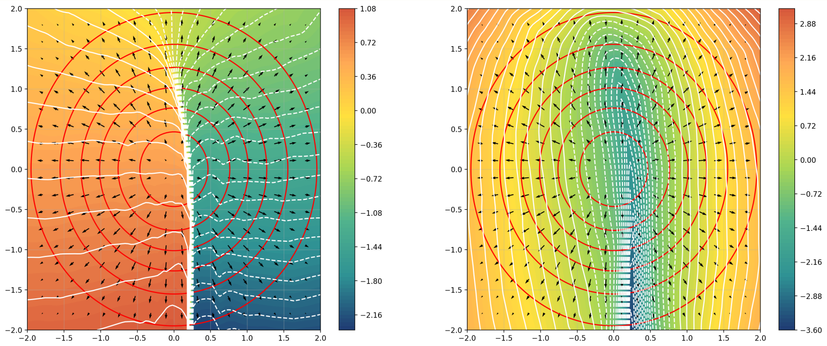
*where $D_{e^*}(y)$ is the encoder Jacobian at y, whenever the following quantities: the **level-set center-of-mass**:*

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

*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** $Z(X) = \int P_{data}(z)\, \delta(e(z) - e(X))\, dz > 0$.*



Rotational symmetry:

One component is tangential (both sides ≈ 0).

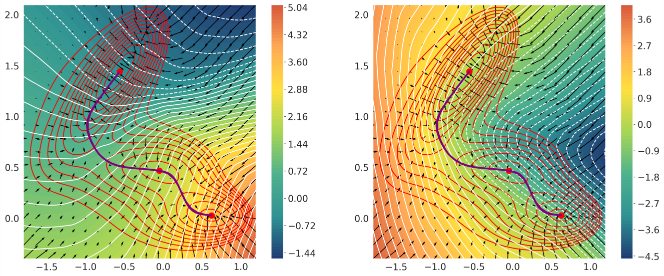
The other is normal (level sets orthogonal to the score); together they recover polar coordinates.

Boltzmann data: forces & Minimum Free-Energy Path (MFEP)

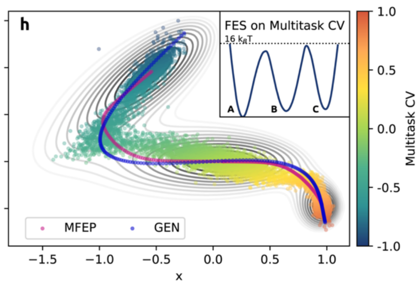
For Boltzmann distributed data, we recover the normal force field:

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

Encoding of molecular simulation trajectories thus reveals the *Minimum free energy path* (MFEP):



DPA components trace the MFEP in a single fit.



Existing methods require iteration/supervision [Bonati et al., 2023].

Second main result: encoding dimensions beyond the manifold are uninformative

Theorem (Extra dimensions are completely uninformative)

For a manifold that can be approximated in K' dimensions by the encoder, the dimensions $(K'+1, \cdots, p)$ of such optimal encoder obey:

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

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:

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

So What?

Distribution approximation / reconstruction and **dimensionality reduction / disentanglement** almost always present a **trade-off**. For example, this is what the β in β-VAE does:

$$\arg\min_{\theta,\phi} \; \mathbb{E}_{p_{\mathrm{data}}(x)} \left[\underbrace{\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x \mid z)]}_{\text{reconstruction}} + \beta \; \underbrace{\mathrm{KL}\Big(q_\phi(z \mid x) \,\Big\|\, \textstyle\prod_j p(z_j)\Big)}_{\text{disentanglement}} \right]$$

https://arxiv.org/abs/2502.11583

Thank you!

References

Luigi Bonati, Enrico Trizio, Andrea Rizzi, and Michele Parrinello. A unified framework for machine learning collective variables for enhanced sampling simulations: mlcolvar. *The Journal of Chemical Physics*, 159(1):014801, July 2023. ISSN 0021-9606, 1089-7690. doi: 10.1063/5.0156343. URL https://doi.org/10.1063/5.0156343.

Xinwei Shen and Nicolai Meinshausen. Distributional Principal Autoencoders, April 2024. URL http://arxiv.org/abs/2404.13649. arXiv:2404.13649 [cs, stat].