Analyzing Generative Models by Manifold Entropic Metrics

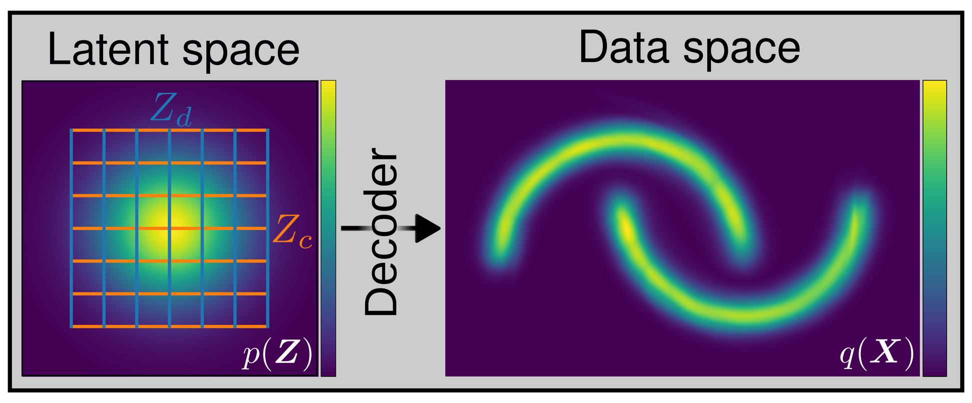
Daniel Galperin, Ullrich Köthe
Computer Vision and Learning Lab, Heidelberg University, Germany

arXiv [2410.19426]

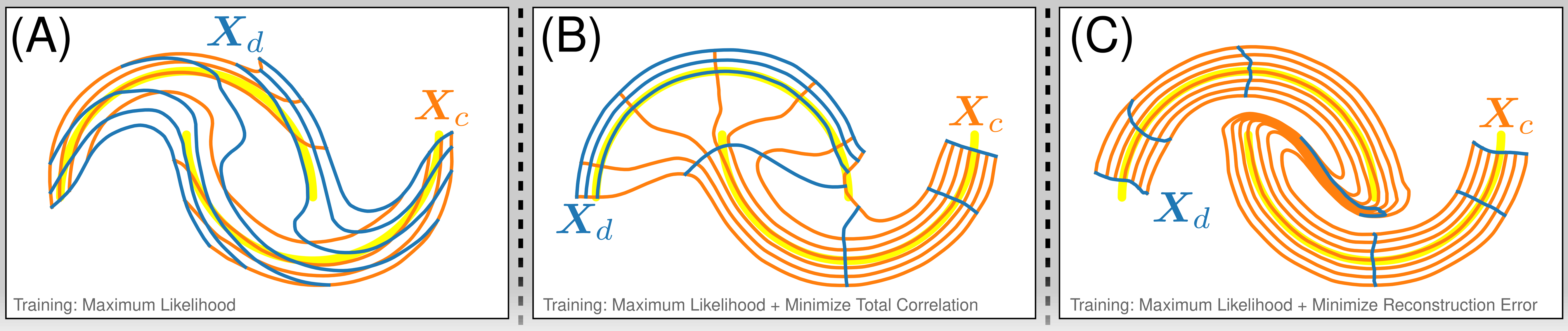
STRUCTURES CLUSTER OF EXCELLENCE

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Which learned representation is better?



Competing generative models for the two moons dataset:

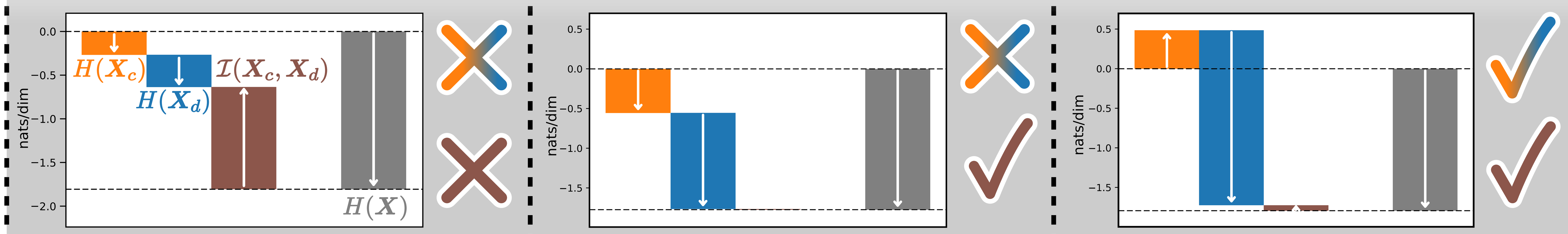


We can quantify that by Manifold Entropic Metrics in terms of

$H(X_c) \gg H(X_d)$ **Alignment** and **Disentanglement** $\mathcal{I}(X_c, X_d) \approx 0$

Manifold Entropic Metrics:

Entropic decomposition: $X = \{X_c, X_d\} \longrightarrow H(X) = H(X_c) + H(X_d) - \mathcal{I}(X_c, X_d)$



Background

How to invert an unknown data-generating process **DGP**:

$$x = \Phi(s)$$

x generated data, Φ mixing function, s (semantic) source vectors

Independent Component Analysis **ICA**:
Assume statistically indepenent sources s_i
→ non-Gaussian latent distribution
→ Mixing function is unaffected

⟷

Independent Mechanism Analysis **IMA**:
Assume causally independent source contributions $\partial\Phi/\partial s_i$
→ Jacobian of mixing function becomes orthogonal

Approach

Disentangled Representation Learning **DRL**: Varying a single feature Δs_i only varies a single semantically meaningful and isolated variation in the data Δx_i

Quantify success of a generative model on **DRL** given sources:
- One latent variable should model the same source feature globally
- Different latent variables should not model the same source feature locally

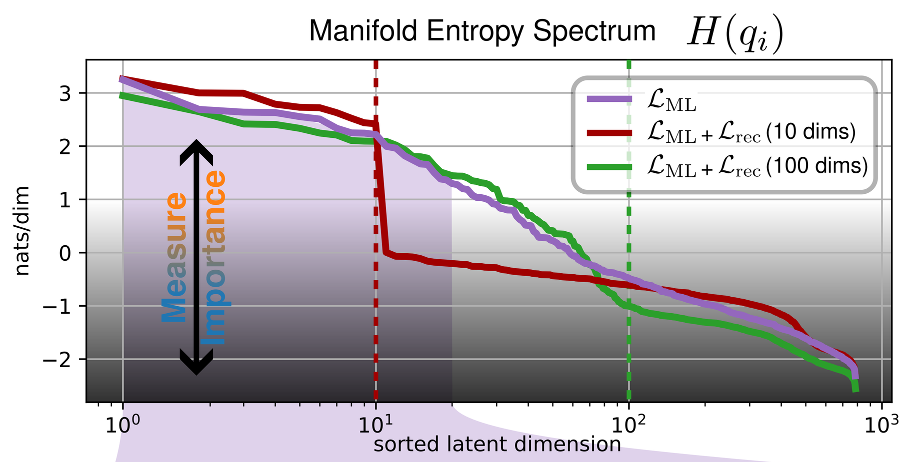
Make two assumptions about the **DGP** to reformulate *supervised* necessary conditions into *unsupervised* desiderata using **IMA** principle:

Alignment: The importance of different semantic features varies greatly
→ Sorting of Manifold Entropy
(In PCA → Sorting of eigenvalues)

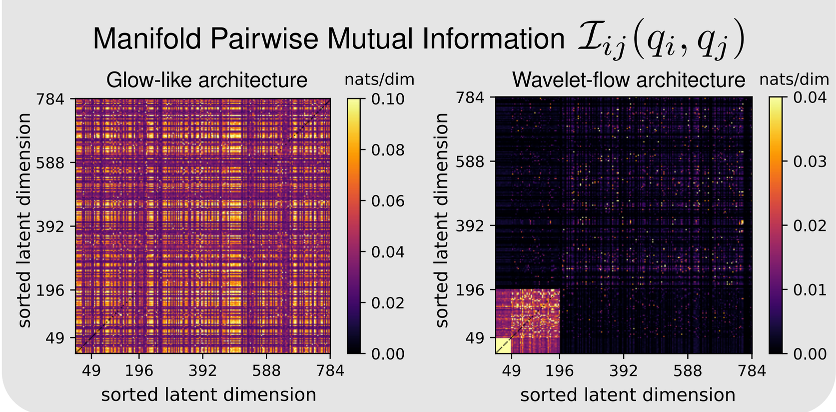
Disentanglement: Semantic features mostly model independent variations in data
→ Vanishing Manifold Mutual Information
(In PCA → Orthogonality between eigenvectors)

Experiments on EMNIST

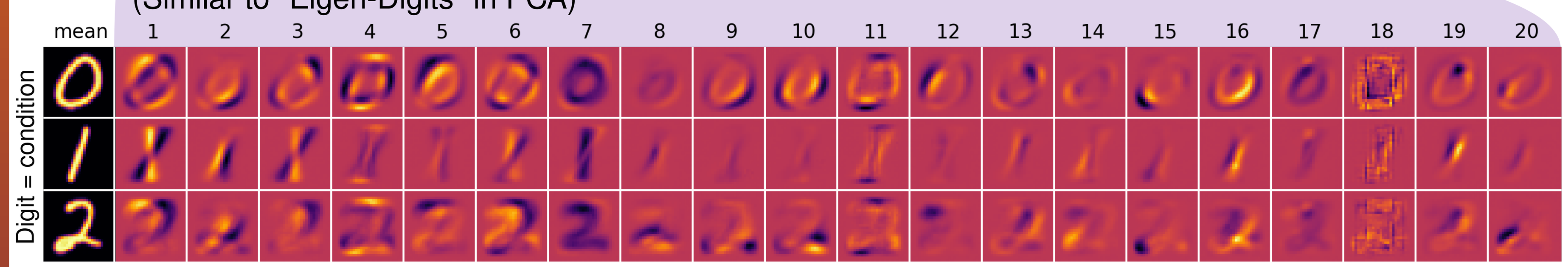
Analyze class-conditioned Normalizing Flows (cINN) by Manifold Entropic Metrics:



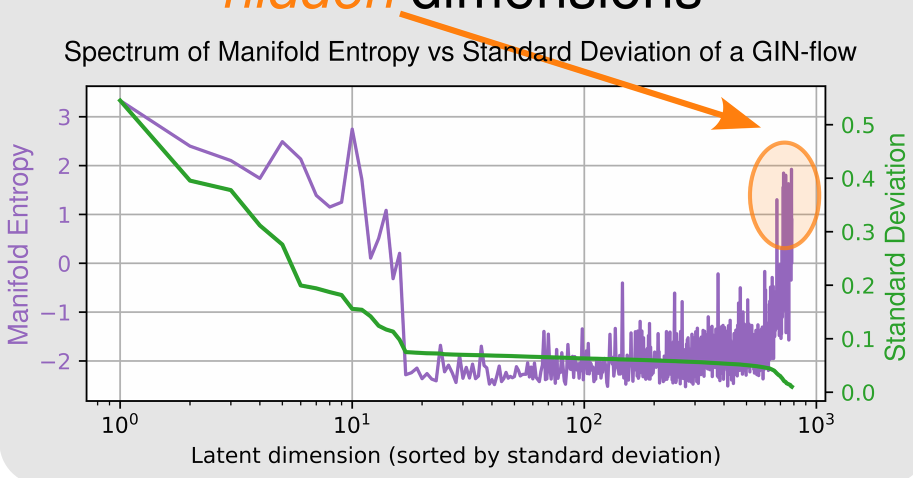
We can *reveal biases* between different architectures



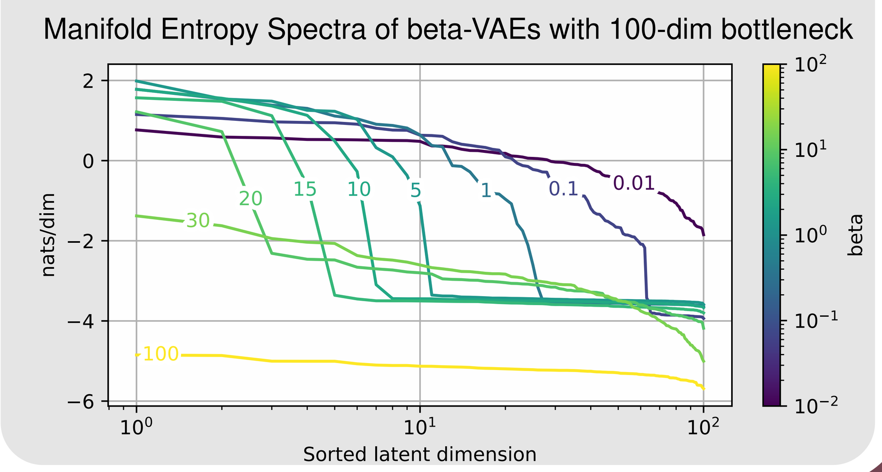
Visualize high-dimensional ***Latent Manifolds*** via Averaged ***Jacobian columns*** (Similar to "Eigen-Digits" in PCA)



We can reveal previously *hidden* dimensions



We can distinguish *active* from *inactive* dimensions in beta-VAEs



Derivation

Normalizing Flows:
Encoder: $z = f(x) \in \mathbb{R}^D$ Decoder: $x = f^{-1}(z) =: g(z) \in \mathbb{R}^D$
Prior latent distribution: $p(Z = z) = \mathcal{N}(z \,|\, 0, I_D)$

→ Select latent dimensions via index set $\mathbb{S} \subseteq \{1, ..., D\}$ and its complement $\overline{\mathbb{S}}$ to split latent vector $z = [z_\mathbb{S}, z_{\overline{\mathbb{S}}}]$

Define ***Manifold random variable*** and ***Latent manifold***

$$X_\mathbb{S} := g([Z_\mathbb{S}, z_{\overline{\mathbb{S}}}])$$

$$\mathcal{M}_\mathbb{S}(z_{\overline{\mathbb{S}}}) := \{x = g([z_\mathbb{S}, z_{\overline{\mathbb{S}}}]) : z_\mathbb{S} \in \mathbb{R}^{|\mathbb{S}|}\}$$

(Sub-)Jacobian:

$$J_\mathbb{S} := \frac{\partial g(z')}{\partial z'_\mathbb{S}}\bigg|_{z'=z} = \begin{pmatrix} \frac{\partial g_1}{\partial z_{\mathbb{S}_1}} & \cdots & \frac{\partial g_1}{\partial z_{\mathbb{S}_{|\mathbb{S}|}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_D}{\partial z_{\mathbb{S}_1}} & \cdots & \frac{\partial g_D}{\partial z_{\mathbb{S}_{|\mathbb{S}|}}} \end{pmatrix}$$

Matrix Volume:

$$|J_\mathbb{S}| := \det\left(J_\mathbb{S}(z)^T J_\mathbb{S}(z)\right)^{\frac{1}{2}}$$

Derive ***Manifold pdf*** via (injective) change of variables

$$q_\mathbb{S}(X_\mathbb{S}) = p_\mathbb{S}(Z_\mathbb{S} = z_\mathbb{S}) \left|J_\mathbb{S}(z)\right|^{-1}$$

Derive ***Manifold Entropy*** as Differential Entropy of $X_\mathbb{S}$

$$H(q_\mathbb{S}) = \mathbb{E}_z[-\log(q_\mathbb{S}(x_\mathbb{S}))] = \mathbb{E}_z[-\log(p_\mathbb{S}(z_\mathbb{S})) + \log|J_\mathbb{S}(z)|]$$

Derive ***Manifold Mutual Information*** analogously between $X_\mathbb{S}$ and $X_\mathbb{T}$

$$\mathcal{I}(q_\mathbb{S}, q_\mathbb{T}) = \mathbb{E}_z\left[\log\left(\frac{q_{\mathbb{ST}}(x_{\mathbb{ST}})}{q_\mathbb{S}(x_\mathbb{S})\,q_\mathbb{T}(x_\mathbb{T})}\right)\right] = \mathbb{E}_z\left[\log|J_\mathbb{S}(z)| + \log|J_\mathbb{T}(z)| - \log|J_{\mathbb{ST}}(z)|\right]$$

Derive application-specific metrics for arbitrary $\mathbb{S}$ via decomposition rule:

$$X_{\mathbb{S},\mathbb{T}} = \{X_\mathbb{S}, X_\mathbb{T}\} \qquad H(q_{\mathbb{S},\mathbb{T}}) = H(q_\mathbb{S}) + H(q_\mathbb{T}) - \mathcal{I}(q_\mathbb{S}, q_\mathbb{T})$$

Empirically useful metrics for $\mathbb{S} = \{i\}$:

– ***Manifold Entropy*** (dimensionwise) $H(q_i)$
Marginal information in X_i: "non-linear std" → Measure importance per dimension
Use: Plot spectrum and identify (un)important manifolds random variables

– ***Manifold Pairwise Mutual Information*** $\mathcal{I}_{ij}(q_i, q_j)$
Shared information between X_i and X_j: "non-linear cosine similarity" → Measure orthogonality between individual latent dimensions
Use: Plot matrix and identify (dis)entangled manifold random variables

– ***Manifold Total Correlation*** $\mathcal{I}$
Residual information between all X_m, $m \in \{1, \ldots, D\}$: "Global orthogonality"

– ***Manifold Cross-Pairwise Mutual Information*** $\mathcal{I}^{ab}_{ij}(q^a_i, q^b_j)$
Shared information between X^a_i and X^b_j coming from two different models:
Compare two models by identifying (dis)similar manifold random variables