Curse of Slicing: Why Sliced Mutual Information is a Deceptive Measure of Statistical Dependence

The Fourteenth International Conference on Learning Representations

---

Alexander Semenenko, Ivan Butakov, Ivan Oseledets, Alexey Frolov

semenenko@applied-ai.ru, ivan.butakov@applied-ai.ru

Mutual Information

The **mutual information (MI)** between random vectors $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ with joint distribution $\mathbb{P}_{X,Y}$ and marginals $\mathbb{P}_X$ and $\mathbb{P}_Y$ is defined as

$$\mathsf{I}(X;Y) = \mathbb{E} \log \frac{\mathrm{d}\,\mathbb{P}_{X,Y}}{\mathrm{d}\,\mathbb{P}_X \otimes \mathbb{P}_Y} = \mathsf{KL}\left[\mathbb{P}_{X,Y} \,\middle\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\right].$$

Properties:

- Nullification Property: $\mathsf{I}(X;Y) = 0$ iff X and Y are independent
- Invariance to bijections, chain rule, etc.

Applications:

- Independence testing
- Feature selection
- Representation learning/disentanglement
- Analysis of DNNs

The **k-sliced mutual information (k-SMI)** (Goldfeld et al., 2022) between X and Y is defined as

$$\mathsf{SI}_k(X;Y) = \int_{\mathrm{St}(k,d_x)} \int_{\mathrm{St}(k,d_y)} \mathsf{I}(\Theta^\mathsf{T} X; \Phi^\mathsf{T} Y)\, \mathrm{d}\mu_{\mathrm{St}(k,d_x)}(\Theta)\, \mathrm{d}\mu_{\mathrm{St}(k,d_y)}(\Phi),$$

where $\mu_{\mathrm{St}(k,d_x)}$ denotes normalized Haar (uniform) probability measure on a Stiefel manifold $\mathrm{St}(k,d) = \{\mathrm{Q} \in \mathbb{R}^{d\times k} : \mathrm{Q}^\mathsf{T}\mathrm{Q} = \mathrm{I}\}$. Setting $k = 1$ recovers the SMI

Pros:

- Scalability
- Nullification Property: $\mathsf{SI}_k(X;Y) = 0$ iff X and Y are independent
- Monotonic: $\mathsf{SI}_{k_1}(X;Y) \leq \mathsf{SI}_{k_2}(X;Y), \quad k_1 < k_2$
- Chain rule

Cons (not well-covered in the literature):

- Asymptotics in high-dimensional regime
- Data Processing Inequality violation
- Suboptimality of random slicing

Lemma 1. Consider the following pair of jointly Gaussian d-dimensional random vectors: $(X, Y) \sim \mathcal{N}\left(0, \begin{pmatrix} \mathrm{I} & \rho\mathrm{I} \\ \rho\mathrm{I} & \mathrm{I} \end{pmatrix}\right), \quad \rho \in (-1; 1)$. In this setup, MI and SMI can be calculated analytically:

$$\mathsf{I}(X;Y) = -\frac{d}{2}\log(1-\rho^2), \qquad \mathsf{SI}(X;Y) = \frac{\rho^2}{2d}\,{}_3F_2\left(1, 1, \frac{3}{2}; \frac{d}{2}+1, 2; \rho^2\right),$$

where ${}_3F_2$ is the *generalized hypergeometric function*. Additionally, the following limits with the *digamma function* ψ hold:

$$\lim_{d\to\infty} \mathsf{I}(X;Y) = +\infty \qquad \lim_{d\to\infty} \mathsf{SI}(X;Y) = 0$$

$$\lim_{\rho^2\to 1} \mathsf{I}(X;Y) = +\infty \qquad \lim_{\rho^2\to 1} \mathsf{SI}(X;Y) = \psi(d-1) - \psi\left(\frac{d-1}{2}\right) - \log 2 \leq \frac{1}{d-1}.$$

SMI's saturation: Gaussian Example

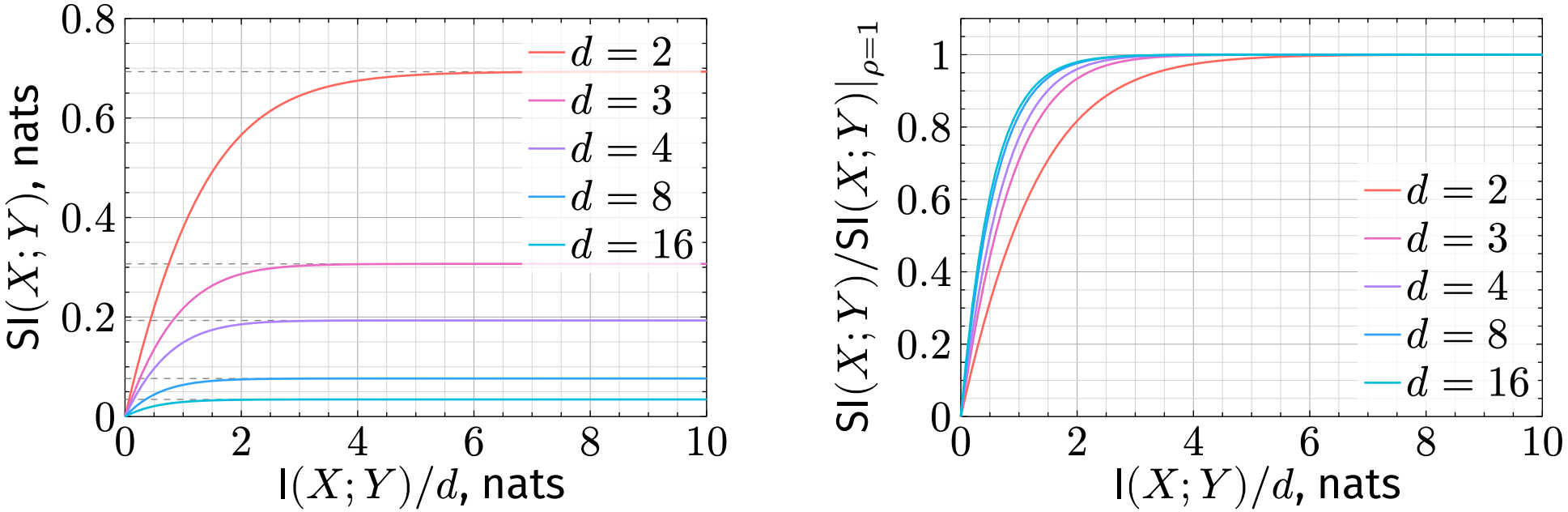
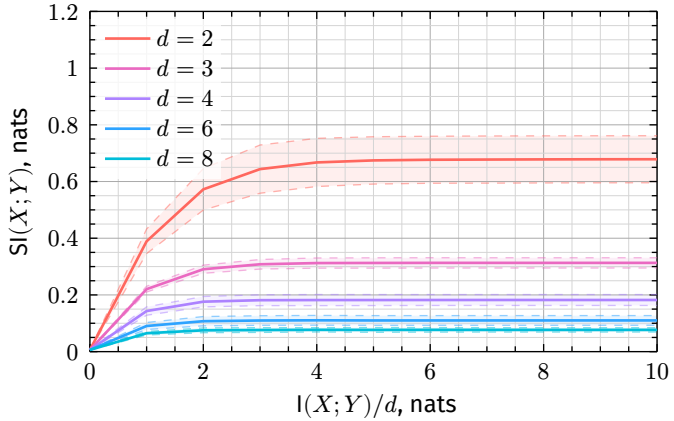
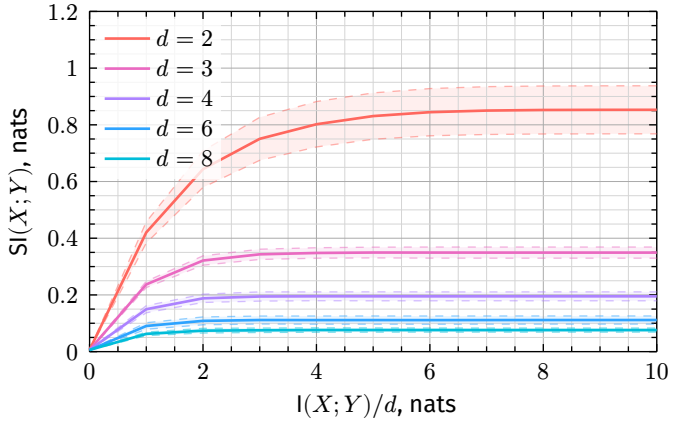


Figure 1: Saturation of $\mathsf{SI}(X;Y)$ as function of $\mathsf{I}(X;Y)/d$ for the example from Lemma 1, non-normalized (left) and normalized (right) versions. Note that the problem becomes more prominent in higher dimensions, both because of lower plateau and faster saturation.

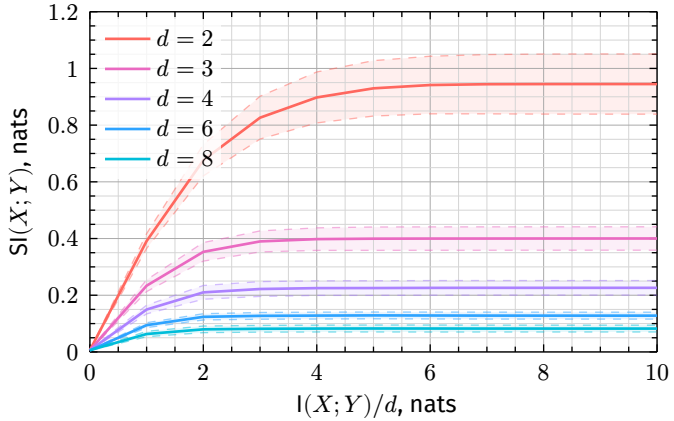
SMI saturates: Synthetic Benchmarks



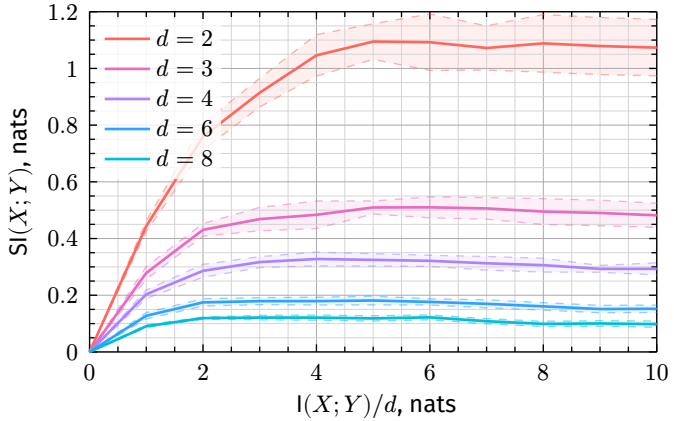
(a) Correlated Normal



(b) Correlated Uniform

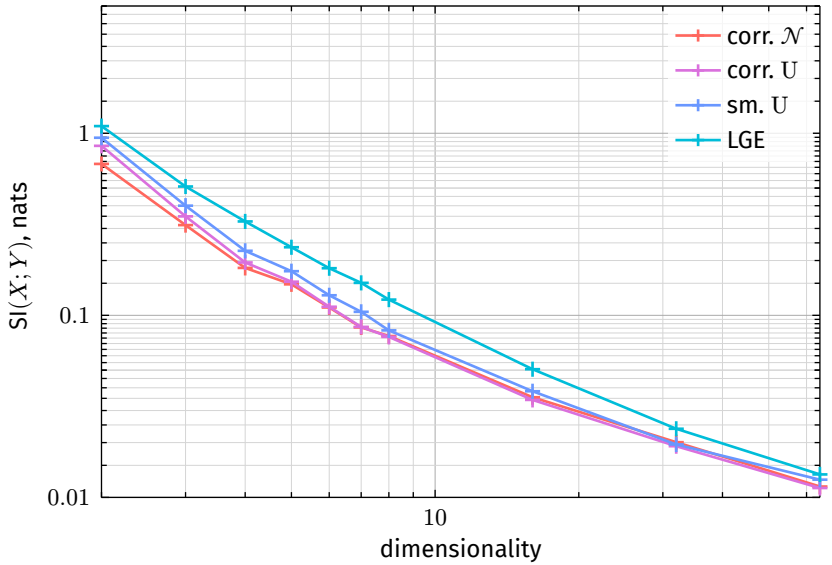


(c) Smoothed Uniform

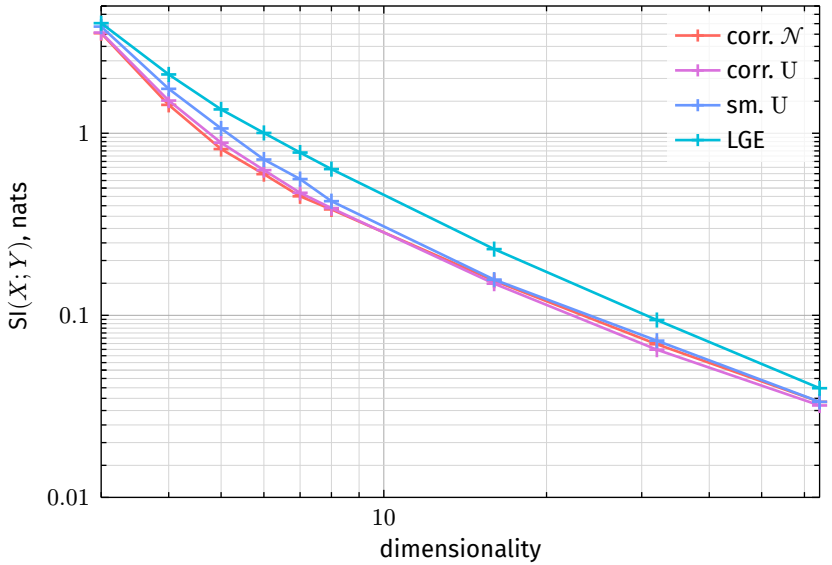


(d) Log-Gamma-Exponential

SMI decays to zero: Synthetic Benchmarks



(a) $k = 1$



(b) $k = 2$

Figure 3: Decaying trends of k-SMI for *correlated normal* (corr. $\mathcal{N}$), *correlated uniform* (corr. U), *smoothed uniform* (sm. U) and *log-gamma-exponential* (LGE).

Proposition 1. Let X and Y be d_x, d_y-dimensional random vectors correspondingly, with $d_x, d_y < k$. Let $\mathrm{A} \in \mathbb{R}^{m_x \times d_x}$ and $\mathrm{B} \in \mathbb{R}^{m_y \times d_y}$ be matrices of ranks d_x, d_y. Then $\mathsf{SI}_k(\mathrm{A}X; \mathrm{B}Y) = \mathsf{I}(X; Y)$.

Corollary 1. Consider the following pair of jointly Gaussian d-dimensional random vectors:

$$(X, Y) \sim \mathcal{N}\left(0, \begin{pmatrix} \mathrm{J} & \rho\mathrm{J} \\ \rho\mathrm{J} & \mathrm{J} \end{pmatrix}\right), \quad \rho \in (-1; 1),$$

where $\mathrm{J} = \mathbf{1} \cdot \mathbf{1}^\mathsf{T}$ with $\mathbf{1}^\mathsf{T} = (1, \ldots, 1)$. Then $\mathsf{SI}_k(X; Y) = \mathsf{I}(X; Y) = -\frac{1}{2}\log(1 - \rho^2)$.

***Remark.* 1.** Applying $\mathrm{V} = \mathbf{1} \cdot e_1^\mathsf{T}$ to the random vectors from Lemma 1 individually yields the example from Corollary 1. Therefore,

$$\mathsf{SI}_k(\mathrm{V}X; \mathrm{V}Y) > \mathsf{SI}_k(X; Y), \text{ but } \mathsf{I}(\mathrm{V}X; \mathrm{V}Y) < \mathsf{I}(X; Y).$$

SMI for Representation Learning

- Let X be a high-dimensional random vector and f be an encoder (approximated by a neural network). We aim to obtain low-dimensional representation $f(X)$ by maximizing MI (Hjelm et al., 2019):

$$\mathsf{I}(X; f(X)) \to \max.$$

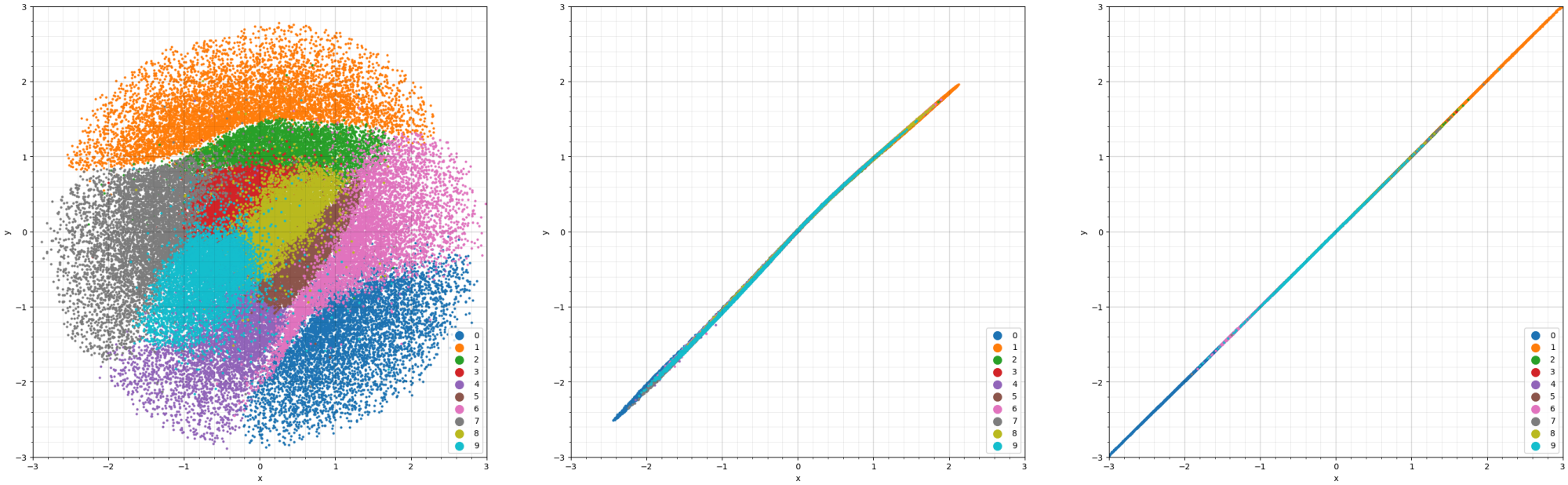
- We replace MI with SMI in Deep InfoMax framework

$$\mathsf{SI}_k(X; f(X)) \to \max.$$

This substitution is easy since both MI and SMI admit Donsker-Varadhan bounds:

$$\mathsf{I}(X;Y) = \sup_{T:\Omega\to\mathbb{R}} \Big[\mathbb{E}_{\mathbb{P}_{X,Y}} T(X,Y) - \log\Big(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Y} e^{T(X,Y)}\Big)\Big],$$

$$\mathsf{SI}_k(X;Y) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\Theta,\Phi}\Big[\mathbb{E}_{\mathbb{P}_{X,Y}} T(\Theta^\mathsf{T}X, \Phi^\mathsf{T}Y, \Theta, \Phi) - \log\Big(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Y} e^{T(\Theta^\mathsf{T}X,\Phi^\mathsf{T}Y,\Theta,\Phi)}\Big)\Big].$$



(a) MI $\rightarrow \max$, 2000 epochs. (b) SMI $\rightarrow \max$, 10 epochs. (c) SMI $\rightarrow \max$, 2000 epochs.

Figure 4: Visualizations of embeddings from the representation learning experiments, with points colored by class. Note that MI maximization (left) produces clustered low-redundancy representations, while SMI maximization results in immediate (after 10 epochs) collapse.

What the talk left out

- **Replication of original SMI experiments:** feature extraction and independence testing. With minor modifications SMI fails where MI succeeds, contradicting earlier claims.
- **Theoretical analysis beyond Gaussians:** saturation and decay hold for general distributions with independent components.
- **k-SMI and optimal slicing (mSMI, oSMI)** do not escape the redundancy bias.
- **Gaussian channel & feature extraction:** SMI prefers ill-conditioned transformations (redundancy bias) and collapses to degenerate solutions in InfoMax tasks.

Thank you for your attention!

References

[1] Z. Goldfeld, K. Greenewald, T. Nuradha, and G. Reeves, "k-Sliced Mutual Information: A Quantitative Study of Scalability with Dimension," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=L-ceBdl2DPb

[2] R. D. Hjelm *et al.*, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX