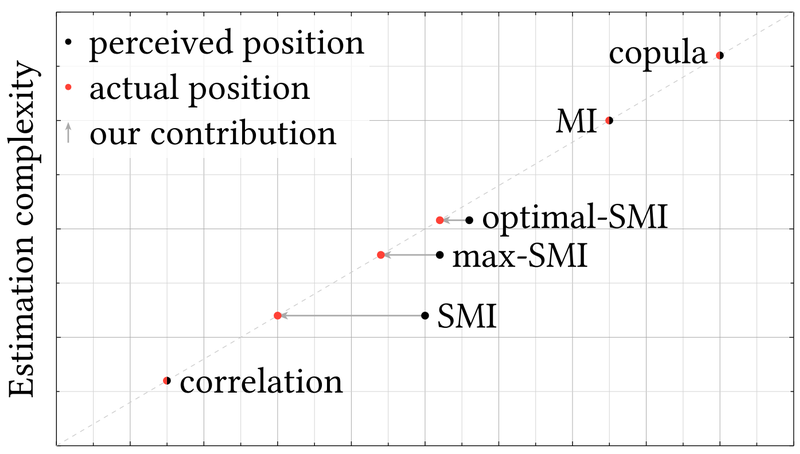
Curse of Slicing: Why Sliced Mutual Information is a Deceptive Measure of Statistical Dependence

Alexander Semenenko, Ivan Butakov, Alexey Frolov, Ivan Oseledets

Skoltech, MIPT, AIRI, Applied AI Institute, INM RAS

Skoltech
Wireless

Project Center for
Next Generation
Wireless and IoT



Ability to capture complex statistical dependencies

Introduction

Mutual Information (MI) is a universal measure of non-linear dependence between random X and Y.

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

Problem: MI is hard to estimate.

Solution (?): use random projections — **Sliced Mutual Information** (SMI).

Is this approach actually good?

Formal Definitions

Mutual Information:

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

- Equals zero iff X and Y are independent
- Invariant under bijections

Conditional Mutual Information:

$$\mathsf{I}(X;Y \mid Z) = \mathbb{E}_{z\sim Z}\,\mathsf{I}(X;Y \mid Z = z)$$

Sliced Mutual Information:

$$\mathsf{SI}_k(X;Y) = \mathsf{I}(\Theta^\mathsf{T}X;\Phi^\mathsf{T}Y \mid \Theta,\Phi),$$

where Θ,Φ are random uniformly distributed projectors on k-dimensional subspaces.

Key Findings

We analyzed SMI across diverse settings and demonstrated

1. **SMI *saturates* prematurely and fails to detect significant increases in statistical dependence.**
2. **SMI prioritizes information *redundancy* over information content.**
3. **SMI is, in fact, cursed with the curse of dimensionality.**
4. **SMI follows the universal scalability-utility trade-off in statistical dependence measures.**

Theoretical analysis

Lemma. For $(X,Y) \sim \mathcal{N}\left(0, \begin{pmatrix} \mathrm{I} & \rho\mathrm{I} \\ \rho\mathrm{I} & \mathrm{I} \end{pmatrix}\right)$, $\rho \in (-1;1)$, MI and SMI have analytical form

$$\mathsf{I}(X;Y) = -\frac{d}{2}\log(1-\rho^2),$$

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

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

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

$$\lim_{\rho^2\to 1}\mathsf{I}(X;Y) = +\infty \qquad \lim_{\rho^2\to 1}\mathsf{SI}(X;Y) \le \frac{3}{d-1}.$$

This example shows that

- SMI suffers the curse of dimensionality
- SMI rapidly saturates

Remarks.
1. k-SMI inherits the same issues.
2. Applying $\mathrm{V} = \mathbf{1}\cdot e_1^\mathsf{T}$ to the random vectors from Lemma, one can observe

$$\mathsf{SI}_k(\mathrm{V}X;\mathrm{V}Y) > \mathsf{SI}_k(X;Y).$$

In contrast,

$$\mathsf{I}(\mathrm{V}X;\mathrm{V}Y) < \mathsf{I}(X;Y).$$

Synthetic Experiments

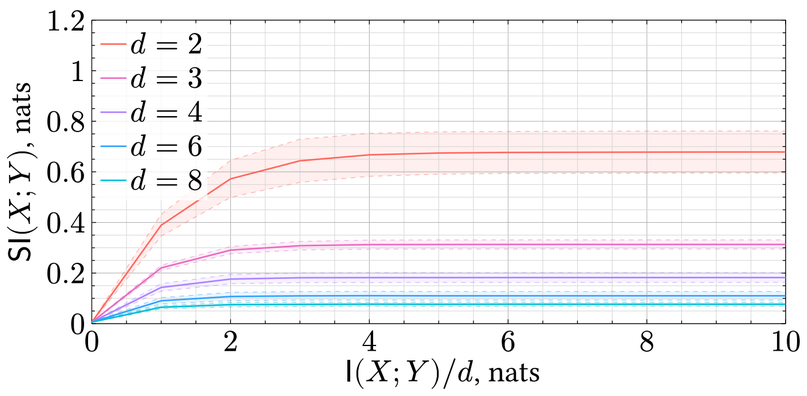


Fig. 2: Correlated Normal

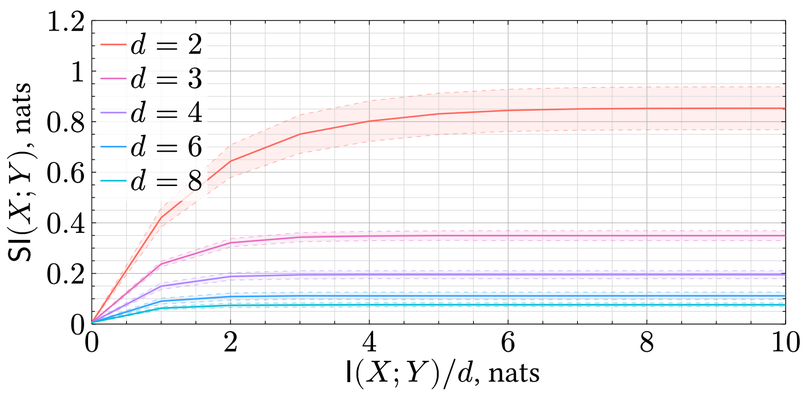
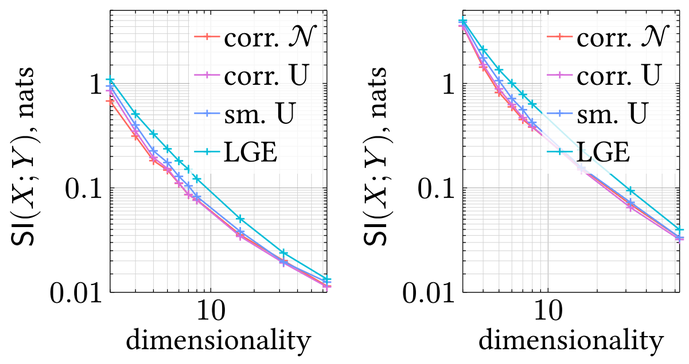


Fig. 3: Correlated Uniform



(a) $k = 1$ (b) $k = 2$

Fig. 4: Decaying trends of k-SMI. We plot saturated values of k-SMI against data dim d.

SMI: Gaussian channel

Let X be a zero-mean d-dimensional random vector, and let $Z \sim \mathcal{N}(0,\sigma\mathrm{I})$ be an independent noise. Consider AWGN channel $X \to X+Z$ and the problem,

$$\sup_{\mathbb{E}\,X_i^2=1}\mathsf{I}(X;X+Z) = \frac{d}{2}\log\left(1+\frac{1}{\sigma^2}\right),$$

where $X_{\text{opt}} \sim \mathcal{N}(0,\mathrm{I})$.

1. *Whitening*: $\Sigma^{-1/2}X$;
2. *Standardization*: $\mathrm{D}^{-1/2}X$, $\mathrm{D} = \mathrm{diag}(\Sigma)$.

	MI		2-SMI	
	$\Sigma^{-1/2}$	$\mathrm{D}^{-1/2}$	$\Sigma^{-1/2}$	$\mathrm{D}^{-1/2}$
X'	7.48 ± 0.01	3.04 ± 0.01	0.96 ± 0.04	2.46 ± 0.03
X''	7.49 ± 0.02	3.04 ± 0.01	0.82 ± 0.05	2.49 ± 0.05

SMI: Representation Learning

Let X be a random vector and f be an encoder (modeled by a NN).

■ In Deep InfoMax we learn most informative embeddings by

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

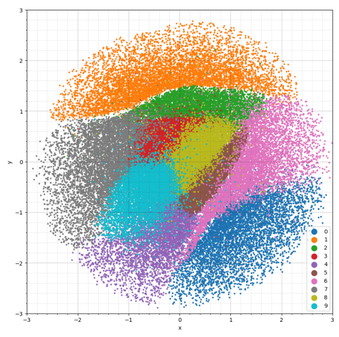
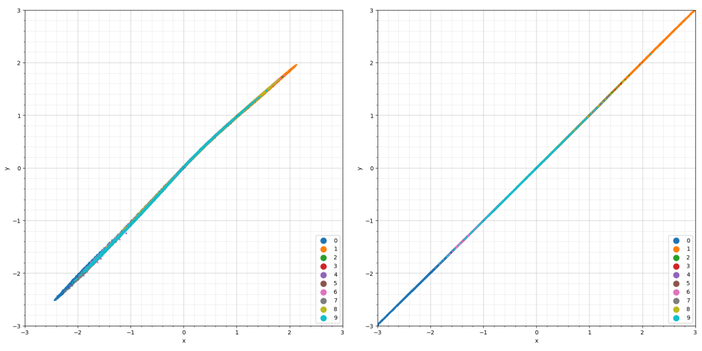


Fig. 5: MI $\to$ max, 2000 epochs.

■ Replace MI with SMI

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



(a) SMI $\to$ max, 10 epochs.

(b) SMI $\to$ max, 2000 epochs.

▸ MI produces clustered low-redundancy representations

▸ SMI maximization results in immediate (after 10 epochs) collapse

Takeaways

1. **No Free Lunch.**
2. **Existing works that utilize SMI should be reevaluated to account for the biases uncovered.**
3. **Similar biases might arise in other slicing-based methods.**

Fall Into ML 2025, HSE, Moscow