UNIVERSITY OF CAMBRIDGE

NEURAL INFORMATION PROCESSING SYSTEMS

Neural Mutual Information Estimation with Vector Copulas

Yanzhi Chen, Zijing Ou, Adrian Weller, Michael Gutmann

Background

Information-theoretic measures

- **Mutual information (MI)**

$$\mathbb{I}(X;Y) = \int p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x} d\mathbf{y}$$

- **Differential entropy**

$$\mathbb{H}[X] = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

Information-theoretic measures

- **Mutual information (MI)**

$$\mathbb{I}(X;Y) = \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} d\mathbf{x} d\mathbf{y}$$

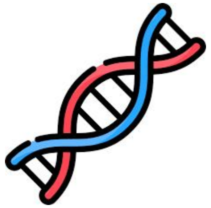
A fundamental metric for measuring statistical dependence

- **Differential entropy**

$$\mathbb{H}[X] = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

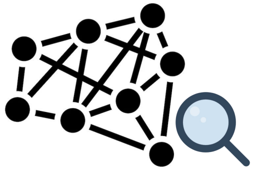
Applications of MI

Identifying co-expressed genes



Bayesian experimental design

Explainability & Interpretability

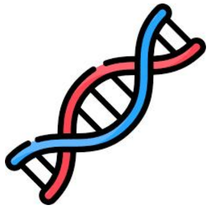


Text-to-image alignment

“blue car”

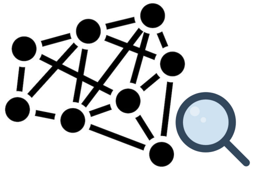
Applications of MI

Identifying co-expressed genes



Bayesian experimental design

Explainability & Interpretability



Text-to-image alignment

"blue car"

*However, MI remains very challenging to estimate in moderate dimensional spaces ($d > 20$)

Divide-and-conquer MI estimation

Divide-and-conquer MI estimation

MI estimation

(hard)

{Marginal distribution learning}

(easier)

{Dependence structure modeling}

(easier)

Divide-and-conquer MI estimation

① Marginal distribution learning

use two **flows** to transform the marginal distributions to **uniform** distributions

Divide-and-conquer MI estimation

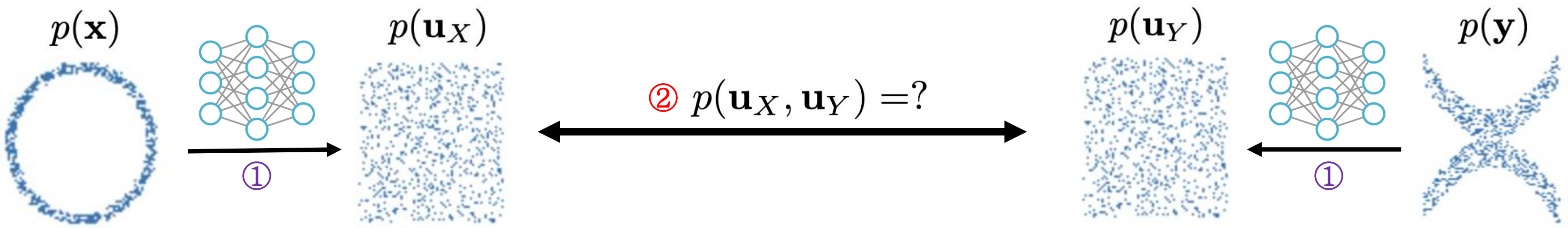
① **Marginal distribution learning**

use two **flows** to transform the marginal distributions to **uniform** distributions

(much easier due to reduced dimensionality)

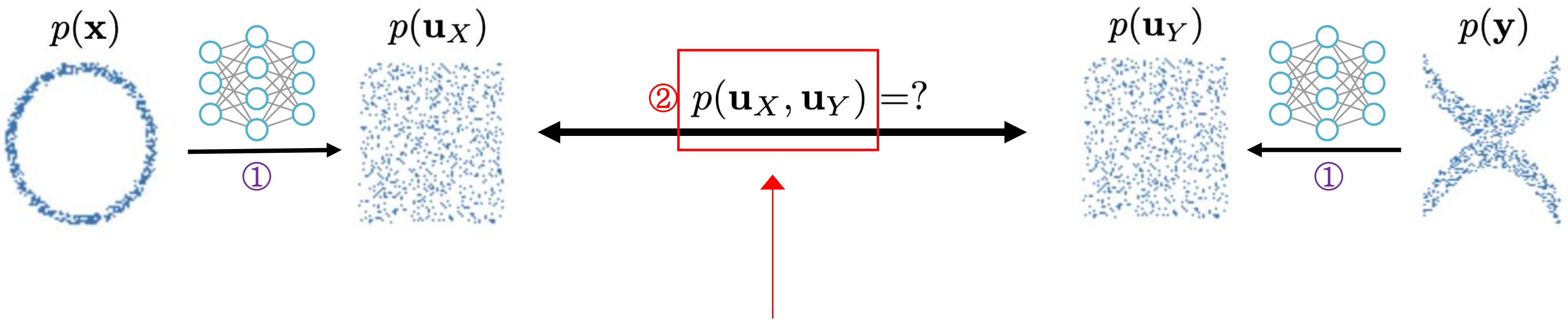
Divide-and-conquer MI estimation

② Dependence structure modeling



Divide-and-conquer MI estimation

② **Dependence structure modeling**

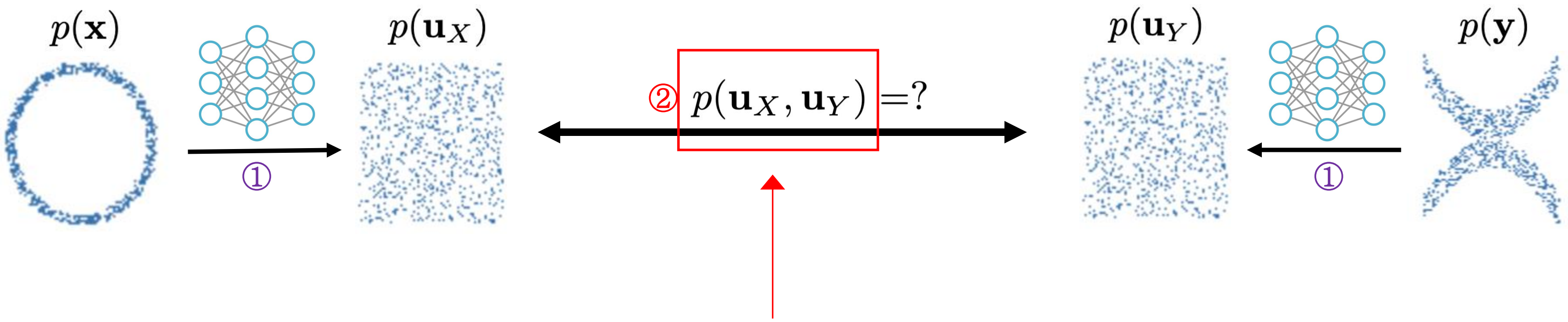


the **vector copula**

it fully captures the **dependence structure** between *X* and *Y*

Divide-and-conquer MI estimation

② Dependence structure modeling



the **vector copula**
it fully captures the **dependence structure** between *X* and *Y*

$$\mathbb{I}(X;Y) = -\mathbb{H}[p(\mathbf{u}_X, \mathbf{u}_Y)]$$

Experiments

Experiments

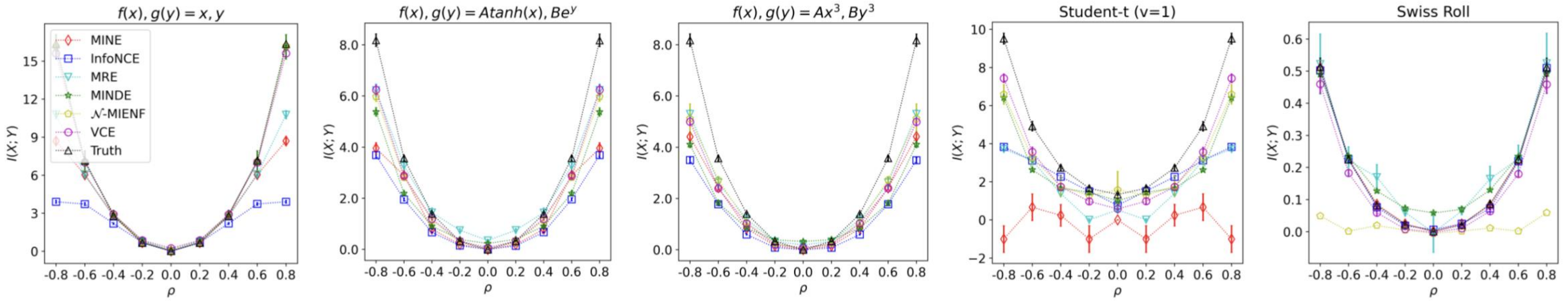
Benchmarks

- Extended **BMI benchmarks**
- High-dimensional correlated **images**
- Resampled **text** embeddings

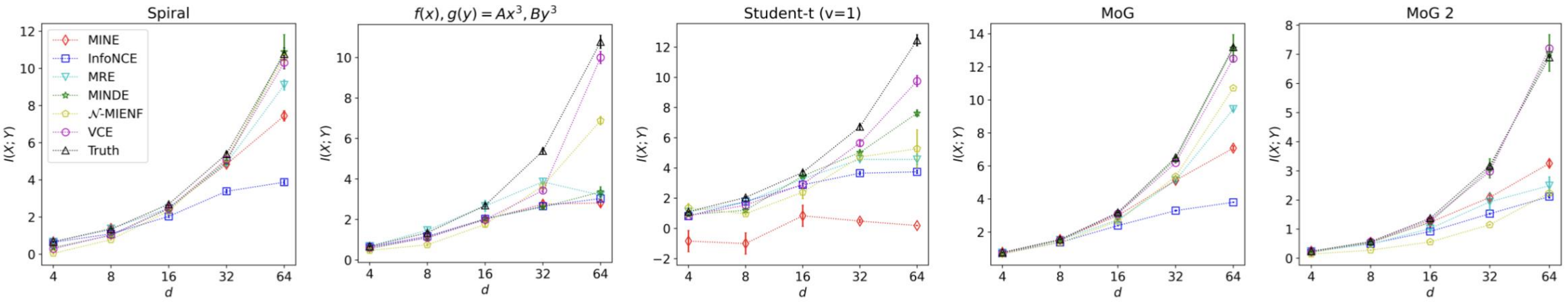
Baselines

- MINE
- InfoNCE
- MRE (*state-of-the-art discriminative estimator*)
- *N*-MIENF
- MINDE (*state-of-the-art generative estimator*)
- **VCE** (*proposed method*)

Experiments - Extended BMI benchmarks



(a) $X, Y \sim \mathcal{N}$ (b) $\mathbf{A}\tanh(X), \mathbf{B}e^{Y}$ (c) $\mathbf{A}X^3, \mathbf{B}Y^3$ (d) Student-t (e) Swiss Roll



(a) Spiral (b) $\mathbf{A}X^3, \mathbf{B}Y^3$ (c) Student-t (d) MoG 1 (e) MoG 2

*VCE offers significant advantage in **high-MI** and **high-dimensional** settings

Experiments - Correlated images (256D)

(a) **Gaussian plates** (b) **Rectangles**

Method	Gaussian Plates			Rectangles		
	$I(X;Y)=1$	$I(X;Y)=3$	$I(X;Y)=7$	$I(X;Y)=1$	$I(X;Y)=3$	$I(X;Y)=7$
MINE	0.89 ± 0.07	2.86 ± 0.24	5.46 ± 0.27	0.81 ± 0.13	$\mathbf{2.57 \pm 0.26}$	5.39 ± 0.23
InfoNCE	0.86 ± 0.14	2.63 ± 0.13	3.83 ± 0.12	0.78 ± 0.17	2.49 ± 0.28	3.86 ± 0.15
MRE	1.23 ± 0.16	2.85 ± 0.21	5.91 ± 0.28	0.82 ± 0.24	2.56 ± 0.48	$\mathbf{5.45 \pm 0.31}$
$\mathcal{N}$-MIENF	0.74 ± 0.12	2.42 ± 0.16	3.85 ± 0.22	0.54 ± 0.13	0.76 ± 0.14	1.54 ± 0.11
VCE	$\mathbf{0.92 \pm 0.04}$	$\mathbf{2.93 \pm 0.12}$	$\mathbf{6.53 \pm 0.36}$	$\mathbf{0.83 \pm 0.12}$	2.27 ± 0.23	5.02 ± 0.14

*VCE is either **the best** or **very close to the champion**

Experiments - Text embeddings (512D+)

	X	Y
1	(positive) I thought this was a wonderful way to spend time on ...	(positive) If you like original gut wrenching laughter you will like ...
2	(negative) So im not a big fan of Boll's work but then ...	(positive) This a fantastic movie of three prisoners who become famous...

Method	$I(X;Y) \approx 2.1$	$I(X;Y) \approx 0.9$
MINE	1.83 ± 0.04	0.71 ± 0.05
InfoNCE	1.64 ± 0.09	0.70 ± 0.06
MRE	1.72 ± 0.07	1.23 ± 0.02
$\mathcal{N}$-MIENF	0.91 ± 0.05	0.43 ± 0.03
VCE	$\mathbf{2.01 \pm 0.04}$	$\mathbf{0.83 \pm 0.01}$

(a) Llama-3 13B

Method	$I(X;Y) \approx 1.5$	$I(X;Y) \approx 0.2$
MINE	$\mathbf{1.42 \pm 0.04}$	0.18 ± 0.02
InfoNCE	1.41 ± 0.03	0.19 ± 0.04
MRE	1.23 ± 0.09	0.31 ± 0.09
$\mathcal{N}$-MIENF	0.73 ± 0.03	0.11 ± 0.02
VCE	1.22 ± 0.02	$\mathbf{0.19 \pm 0.02}$

(b) BERT

*VCE offers less advantanges but it still provides very **robust estimates**

Summary

Summary

- A new MI estimator based on the divide-and-conquer principle

- Disentangling marginal distribution and dependence structure (i.e. the vector copula)

- Achieving accurate and robust estimation across diverse data patterns and modalities

code ->