MI is All You Need

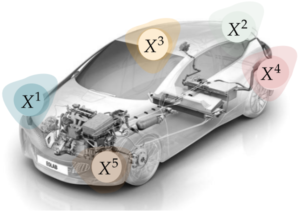
Understanding Complex Multivariate Systems Through the Lenses of GenAI

Prof. Pietro Michiardi

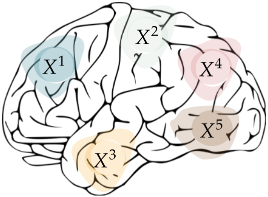
EURECOM
Sophia Antipolis

Introduction

Complex systems are often described by **multivariate** information



Sensors



Brain regions

Understanding the **relationship** among multiple random variables is crucial to analyse **information content and flow** in these systems

What do we use to study information?

- Shannon's Mutual Information (MI): $\mathcal{I}(X^1; X^2)$
- Not interpretable for large systems $X = \{X^1, \cdots, X^N\},\ N > 3$

PID

- Requires a partition into sources and one target
- Not scalable

O-information

- No partition needed
- Scalable

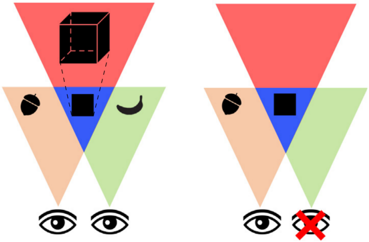
SOTA is limited to discrete or Gaussian distributions

Our methods, MINDE and SΩI, estimate MI and O-information on **arbitrary continuous** systems of **any number** of variables

Multivariate interactions

Redundancy : The **shared** information between variables, which can be recovered from variables or subset of variables

Synergy : The information that arises from **jointly** observing the variables but not accessible from individual variables alone



O-information: a system-wide global measure

$$\boxed{\Omega(X) = \mathcal{T}(X) - \mathcal{D}(X)} \quad \begin{cases} \Omega(X) > 0 & \textit{Redundancy} \\ \Omega(X) < 0 & \textit{Synergy} \end{cases}$$

$\mathcal{T}(X) \Rightarrow$ Information each variable X^i **shares** with others

$\mathcal{D}(X) \Rightarrow$ **Additional** information the variables X^i carry about part of the system, when the remaining part is known

Functions of **Mutual Information** (or Entropy)

Gradient of $\Omega(X)$ captures **individual** flow of information

$$\boxed{\partial_i \Omega(X) = \Omega(X) - \Omega(X^{\setminus i})}$$

Hard for **high dimensional** and **continuous** distributions

How to estimate MI? GenAI to the rescue!

Consider a **joint** generative diffusion model for two variables X^1 and X^2:

$$\begin{cases} \mathrm{d}\left[X_t^1, X_t^2\right]^\top = [X_t^1, X_t^2]^\top \mathrm{d}t + \sqrt{2}\left[\mathrm{d}W_t^1, \mathrm{d}W_t^2\right]^\top, \\ [X_0^1, X_0^2] \sim p^{X^1, X^2} \end{cases}$$

In our work we show that the following holds:

$$\mathrm{KL}\left[p^{X^1} \parallel p^{X^2}\right] = \mathbb{E}_{x \sim p^{X^1}}\left[\int_0^T \underbrace{\left\|\nabla \log p_t^{X^1}(x) - \nabla \log p_t^{X^2}(x)\right\|^2}_{\textit{Difference of score functions}} \mathrm{d}t\right]$$

Approximate the KL by learning to denoise X_t^i :

$$\nabla \log p_t^{X^i}(x) = \frac{1}{2t}\Big(\underbrace{\mathbb{E}[X^i \mid X_t^i]}_{\textit{Denoiser} \approx \epsilon_\theta(\cdot)} - x\Big)$$

MI estimation

So what? We have a way to estimate MI

$$\begin{aligned}\mathcal{I}(X^1; X^2) =& \mathrm{KL}\left[p(X^1, X^2) \parallel p(X^1)p(X^2)\right] \\ =& \mathbb{E}_{p(X^2)}\left[\mathrm{KL}\left[p(X^1 \mid X^2) \parallel p(X^1)\right]\right]\end{aligned}$$

MINDE: mutual information neural diffusion estimation [ICLR 2024][1]:

$$\mathcal{I}(X^1, X^2) = \int \frac{1}{4t^2} \mathbb{E}\left\| \mathbb{E}[X^1 \mid X_t^1] - \mathbb{E}[X^1 \mid X_t^1, X^2] \right\|^2 \mathrm{d}t$$

Just compare the denoiser output when the variable X^1 is denoised **alone** or **conditioned** on X^2 !

[1] We have a discrete version of our MI estimator, preview at Delta Workshop, ICLR 2025

SΩI: Score-based O-information estimation [ICML 2024]

Rewrite $\mathcal{T}(X)$ and $\mathcal{D}(X)$ in terms of KL divergence, apply previous results:

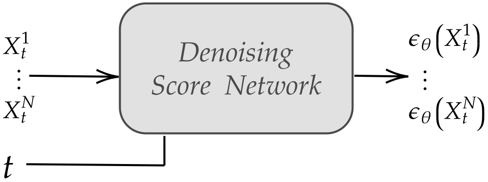
$$\mathrm{T}(X) = \int \frac{1}{4t^2}\mathbb{E}\left\| \mathbb{E}[X \mid X_t] - \left[\mathbb{E}[X^i \mid X_t^i]\right]_{i=1}^{N} \right\|^2 \mathrm{d}t$$

Compare denoiser output when all the variables are denoised **jointly** or **marginally**

$$\mathcal{D}(X) = \int \frac{1}{4t^2}\mathbb{E}\left\| \mathbb{E}[X \mid X_t] - \left[\mathbb{E}[X^i \mid X_t^i, X^{\setminus i}]\right]_{i=1}^{N} \right\|^2 \mathrm{d}t$$

Compare the denoiser output when all the variables are denoised **jointly** or **conditionally** on the remaining clean variables

Amortized approach using a unique network



Algorithm 1: SΩI O-information estimation

Input: $X = \{X^i\}_{i=1}^N, t \sim \mathcal{U}[0,T], \quad X_t = X + \sqrt{2t}W$

$\epsilon(X_t) \leftarrow \epsilon_\theta([X_t^1, \ldots, X_t^N], t)$ // Joint

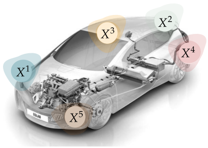
for $i = 1$ **to** N **do**

$\quad \epsilon(X_t^i | X^{\setminus i}) \leftarrow \epsilon_\theta\left([X^1, .., X_t^i, .., X^N], t\right)$ // Conditional

$\quad \epsilon(X_t^i) \leftarrow \epsilon_\theta(X_t^i, t)$ // Marginal

Return $\underbrace{\frac{1}{4t^2}\left\|\epsilon(X_t) - \left[\epsilon(X_t^i)\right]_{i=1}^N\right\|^2}_{\mathcal{T}(X)} - \underbrace{\frac{1}{4t^2}\left\|\epsilon(X_t) - \left[\epsilon(X_t^i|X^{\setminus i})\right]_{i=1}^N\right\|^2}_{\mathcal{D}(X)}$

Applications: Automotive (1)



- Sensors are **unreliable**: can we exploit redundancy?
- Objective: cross generation

How? MLD: multivariate generative model [ENTROPY 2024]

$$\begin{cases} \mathrm{d}\left[X_t^1, X_t^2\right]^\top = [X_t^1, X_t^2]^\top \mathrm{d}t + \sqrt{2}\left[\mathrm{d}W_t^1, \mathrm{d}W_t^2\right]^\top, \\ [X_0^1, X_0^2]^\top \sim p^{A,B}, \end{cases}$$

where X^1 and X^2 correspond to two **modalities** (for simplicity)

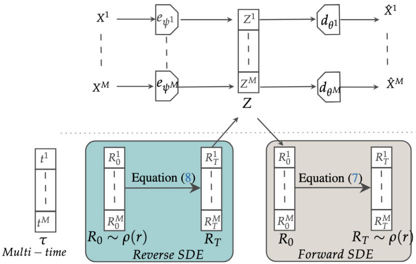
Applications: Automotive (2)

Implementation

- Project inputs to latent variables
- Learn a joint diffusion model, use “multiple arrows of time”

Results

- Dataset: 3 modalities (text, segmentation map, image)
- Any-to-any generation, **coherence is defined by the joint score**



Applications: Alignment (1)

- Common alignment issues: Catastrophic neglecting, Incorrect attribute binding, Incorrect spatial layout
- Existing solutions in the literature:
  - Test time: linguistic steering of generative pathways
  - Fine-tuning: ask GPT for help
- All methods require auxiliary LLM models

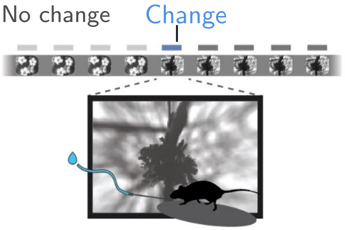
Applications: Alignment (2)

Our method: MITUNE [ICLR 2025][2]

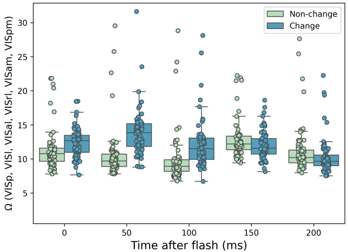
- Self-supervised fine-tuning: all is done with the generative model
- Steps:
  - Generate synthetic data using the pre-trained model
  - Compute **point-wise** MI for each prompt-image pair
  - Select top-k pairs with the highest MI
  - Fine-tune with adapters

[2] We also have a version for Rectified Flows, preview at Delta Workshop, ICLR 2025.

Applications: Neuroscience



O-information in the mice brain



6 brain regions

SΩI is used to estimate O-information for each 50*ms* bin of spikes recording after the stimulus flash

Higher **redundant** information in the visual cortex regions is transmitted in case of a flash with new scene

Conclusion

- GenAI-based extension to Information Theory!
- New information measures unlock a wide array of applications
  - Neuroscience, Cellular development studies
  - Learning from unpaired data
  - Synthetic data augmentation, compression, ...
- Several fundamental problems related to neural estimation
  - Sample efficiency
  - Computational scalability
  - Interpretability and explainability

Thank you !

Joint work with Prof. Giulio Franzese (EURECOM)

Published in ICLR 2024, ICML 2024, ENTROPY 2024, ICLR 2025

Supported by the Huawei Labs Paris, MUSE-COM2 CHIST-ERA project
https://musecom2.eu/

Backup Slides

Why we need Mutual Information

Here's our complex, multivariate system, in abstract terms:

$$X = \{\underbrace{X^1, \dots, X^{i-1}}_{X^{<i}}, X^i, \underbrace{X^{i+1}, \dots, X^N}_{X^{>i}}\},\ X^{\setminus i} = \{X^{<i}, X^{>i}\}$$

- **Total correlation**: $\mathcal{T}(X) = \sum \mathcal{I}(X^i; X^{>i})$
  How much information each variable X^i, **shares** with $X^{>i}$, which suggests a *redundant* scenario

- **Dual total correlation**: $\mathcal{D}(X) = \sum \mathcal{I}(X^i; X^{<i} \mid X^{>i})$
  How much **additional** information the variables X^i carry about $X^{<i}$ if $X^{>i}$ is also available which suggests a *synergistic* scenario

Technical Details I

- Consider random variable A with probability measure $p^A(x)\mathrm{d}x$
- Build a simple SDE in $[0, T]$ with initial conditions $\sim p^A$

$$\begin{cases} \mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \\ X_0 \sim p^A \end{cases} \tag{1}$$

- This SDE corresponds to a path measure $\mathbb{P}^A$
- It is possible to show that two SDEs which differ only by initial conditions have KL divergence

$$\mathrm{KL}\left[\mathbb{P}^A \parallel \mathbb{P}^B\right] = \mathbb{E}_{\mathbb{P}^A}\left[\log \frac{\mathrm{d}\mathbb{P}^A}{\mathrm{d}\mathbb{P}^B}\right] = \mathbb{E}_{\mathbb{P}^A}\left[\log \frac{\mathrm{d}p^A}{\mathrm{d}p^B}\right] = \mathrm{KL}\left[p^A \parallel p^B\right] \tag{2}$$

Technical Details II

Time-reversal $\hat{X}_t$:

- KL between path measures is invariant to time-reversal $\mathrm{KL}\left[\mathbb{P}^A \parallel \mathbb{P}^B\right] = \mathrm{KL}\left[\hat{\mathbb{P}}^A \parallel \hat{\mathbb{P}}^B\right]$
- Time reversal of SDE is again an SDE

$$\mathrm{d}\hat{X}_t = \hat{X}_t + 2\underbrace{\nabla \log p^A_{T-t}(\hat{X}_t)}_{\text{score function!}}\mathrm{d}t + \sqrt{2}\mathrm{d}\hat{W}_t \tag{3}$$

Girsanov theorem: (informal) exrpess the KL between path measures corresponding to two SDEs with different drifts.

$$\mathrm{KL}\left[\hat{\mathbb{P}}^{\mu^A} \parallel \hat{\mathbb{P}}^{\mu^B}\right] \simeq \mathbb{E}_{\mathbb{P}^{\mu^A}}\left[\int_0^T \left\|\nabla \log p^A_t(X_t) - \nabla \log p^B_t(X_t)\right\|^2 \mathrm{d}t\right] \tag{4}$$

Combining with $\mathrm{KL}\left[\mathbb{P}^A \parallel \mathbb{P}^B\right] = \mathrm{KL}\left[\hat{\mathbb{P}}^A \parallel \hat{\mathbb{P}}^B\right]$ we can obtain a KL-estimator!

Mutual Information Neural Diffusion Estimation

Mutual Information between two random variables A, B (many equivalent formulations):

$$\mathrm{I}(A, B) = \mathrm{KL}\left[p^{A,B} \parallel p^{A} p^{B}\right] \tag{5}$$

Idea: estimation using score functions! Two families diffusion processes: joint (J) and conditional (C)

$$\begin{cases} \mathrm{d}\left[X_t, Y_t\right]^\top = -[X_t, Y_t]^\top \mathrm{d}t + \sqrt{2}\left[\mathrm{d}W_t, \mathrm{d}W'_t\right]^\top \\ [X_0, Y_0]^\top \sim p^{A,B} \end{cases}$$

$$\begin{cases} \mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}W_t \\ X_0 \sim p^{A|B} \end{cases}$$