# SΩI: Score-based O-Information Estimation

Mustapha Bounoua [1,2] Giulio Franzese [1] Pietro Michiardi [1]

[1]EURECOM [2]Ampere Software Technology , France

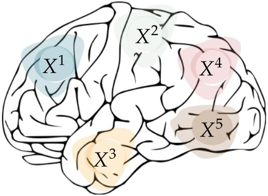
EURECOM Sophia Antipolis

AMPERE

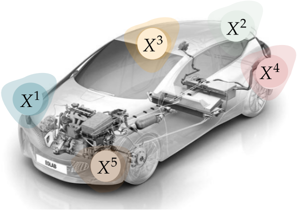
ICML International Conference On Machine Learning

# Introduction

- Complex systems are often described by **multivariate** information
- Understanding the **relationships** among multiple random variables is crucial to analyse these systems




Brain regions




Sensors

How do the system components interact ?

# Extensions of Shannon Mutual information

- Shannon Mutual information: $\mathcal{I}(X^1; X^2)$
- Not interpretable for large systems $N > 3$

**PID**[1]:

- Requires a partition into sources and target
- Not scalable

**O-information**[2]:

- No partition needed
- Scalable

*SOTA is limited to discrete or Gaussian distribution*

- SΩI estimates O-information **without restrictions** on the data type or **number** of variables

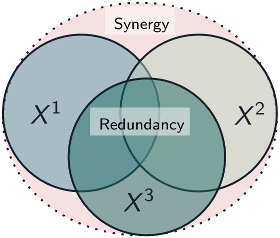
[1] P. L. Williams and R. D. Beer (2010). *Nonnegative Decomposition of Multivariate Information*. arXiv: 1004.2515 [cs.IT]

[2] F. E. Rosas et al. (2019). "Quantifying High-order Interdependencies via Multivariate Extensions of the Mutual Information". In: *Physical review. E* 100 3-1, p. 032305

# 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 and not accessible from individuals



# High dimensional interaction measures

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

- Total correlation: $\mathcal{T}(X) = \sum_{i=1}^{N} \mathcal{I}(X^i; X^{>i})$
  How much information each variable $X^i$, **shares** with $X^{>i}$ which suggests *redundancy*
- Dual total correlation: $\mathcal{D}(X) = \sum_{i=1}^{N} \mathcal{I}(X^i; X^{<i} \mid X^{>i})$
  How much **additional** information the variables $X^i$ carry about $X^{<i}$ if $X^{>i}$ is also available which suggests a *synergistic* scenario

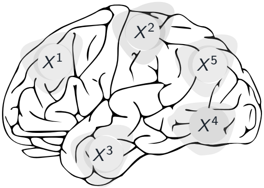
# O-information

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

Gradient of O-information [3]

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

Captures the individual influence of each variable




- Hard to estimate : **high dimensional** and arbitrary data **distribution**

[3] T. Scagliarini et al. (2023). "Gradients of O-information: Low-order descriptors of high-order dependencies". In: *Physical Review Research* 5

# Score-based KL Divergence estimation

Let $X$ a random variable and $X_t = X + \sqrt{2t}W$ its noised version with an intensity indexed by $t \in [0, \infty)$. Using results from[4]:

$$\mathrm{KL}\left[p(x) \parallel q(x)\right] = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x.$$
$$= \int p_t(x) \underbrace{\left\| \nabla \log p_t(x) - \nabla \log q_t(x) \right\|^2}_{\textit{Difference of score functions}} \mathrm{d}x\mathrm{d}t$$

- Learning the score function $\nabla \log p_t(.)$ by learning to denoise $X_t$ : $\nabla \log p_t(x) = \frac{1}{2t}(\underbrace{\mathbb{E}[X \mid X_t]}_{\textit{Denoiser}} - x)$

[4] G. Franzese, M. BOUNOUA, and P. Michiardi (2024). "MINDE: Mutual Information Neural Diffusion Estimation". In: *ICLR*

# Score-based O-information estimation

Consider a multivariate random variable $X \sim p(x^1, \dots, x^N)$:

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

- Comparing the denoiser output when all the variables are denoised **together** (*Joint*) or **separately** (*Marginals*)

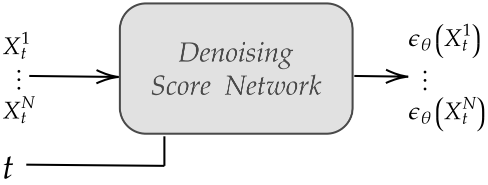
# Score-based O-information estimation

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

- Comparing the denoising process when all the variables are denoised **together** *(joint)* or the individual denoising **conditioned** (*Conditionals*) on the remaining clean variable

$$\Omega(X) = \mathcal{T}(X) - \mathcal{D}(X)$$

# 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`

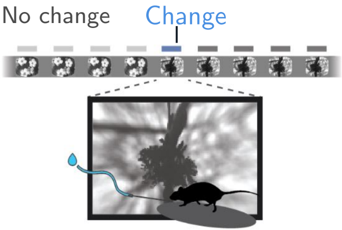
$\quad \epsilon(X_t^i) \leftarrow \epsilon_\theta(X_t^i, t)$ `// Marginal`

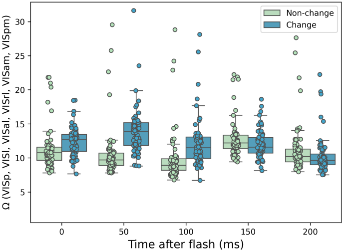
**Return** $\underbrace{\frac{1}{4t^2}\left\| \epsilon(X_t) - \left[\epsilon(X_t^i)\right]_{i=1}^N \right\|^2}_{\mathcal{T}(X)} - \underbrace{\frac{1}{4t^2}\left\| \epsilon(X_t) - \left[\epsilon(X_t^i | X^{\setminus i})\right]_{i=1}^N \right\|^2}_{\mathcal{D}(X)}$

# Experimental validation

- **Synthetic benchmark**
  - Multivariate Gaussian with/without transformation
  - Redundancy, Synergy or a mix
  - Number of variables
  - Dimension of each variable
  - Strength of the interaction
- **Baseline**: We use MI neural estimators to build baselines
- **Implementation**: *MLP* with skip connections is enough for simple settings, while more capacity (*Transformer*) is needed for complex ones (gradient of O-information)
- **SΩI** efficiently estimates O-information across all the challenging settings

# O-information in the mice brain [6]




6 brain regions

SΩI is used to estimate O-information for each 50*ms* bin of spikes recording after the stimulus flash[5]

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

[5] Allen-Institute (2022). "Visual behavior neuropixels dataset overview". In: *https://portal.brain-map.org/explore/circuits/visual-behavior-neuropixels*

[6] P. Venkatesh et al. (2023). "Gaussian partial information decomposition: Bias correction and application to high-dimensional data". In: *Neurips*

# Conclusion

- SΩI can capture the multivariate interactions for **any data distribution** and **large number of variables** whereas classical tools are restricted to discrete or Gaussian data distributions
- The only needed ingredient is access to **the score functions**, which can also be applied to other kind of data: Image, Audio, fMRI, multimodal, . . . etc
- SΩI opens the door for many **scientific applications** ( We're open for collaboration ! )

Project repo !

# Thank you !

See you at the poster session (Number: **1702** )