SΩI: Score-based O-INFORMATION Estimation

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

Abstract

The analysis of scientific data and complex multivariate systems requires information quantities that capture relationships among multiple random variables. Recently, new information-theoretic measures have been developed to overcome the shortcomings of classical ones, such as mutual information, that are restricted to considering pairwise interactions. Among them, the concept of information synergy and redundancy is crucial for understanding the high-order dependencies between variables. One of the most prominent and versatile measures based on this concept is O-INFORMATION, which provides a clear and scalable way to quantify the synergy-redundancy balance in multivariate systems. However, its practical application is limited to simplified cases. In this work, we introduce SΩI, which allows to compute O-INFORMATION without restrictive assumptions about the system while leveraging a unique model. Our experiments validate our approach on synthetic data, and demonstrate the effectiveness of SΩI in the context of a real-world use case.

1. Introduction

Mutual Information (MI) is a fundamental measure which allows investigation of the non-linear dependence between random variables (Shannon, 1948; MacKay, 2003). Despite its success in various domains, classical MI suffers from limitations when analyzing systems composed by more than two variables. This constitutes an important limitation, considering that many scientific endeavors aim at an accurate statistical characterization of systems which are composed of many random variables. Examples includes neuroscience (Latham & Nirenberg, 2005; Ganmor et al., 2011; Gat & Tishby, 1998), climate models (Runge et al., 2019), econometrics (Dosi & Roventini, 2019), and machine learning (Tax et al., 2017), to name a few.

A recent attempt to overcome such limitations, and to extend the applicability of information-theoretic tools to multivariate systems, is represented by Partial Information Decomposition (PID) (Williams & Beer, 2010). The key idea behind such method is the *decomposition* of the overall MI between a set of source variables and a given target variable into non-negative constituents. In particular, PID quantifies how much of the total information about the target variable is encoded redundantly, synergistically or uniquely into given subsets of variables. *Redundancy* quantifies information that is shared between subsets of the partition, *synergy* describes the additional information that is endowed to all subsets observed jointly but that is not available from individual constituents of the partition, and *uniqueness* quantifies the information that is lost when a given subset is not observed, removing the amount of redundant and synergistic information associated to that subset. The PID method requires partitioning the source system into all its possible subsets and computing the information decomposition of all constituents with respect to the target variable.

Despite its elegance, this measure is not without drawbacks. Indeed, there is no consensus on the best way to define and compute PID, and several variants have emerged, including (Barrett, 2014), who reformulate synergy and redundancy for Gaussian systems (but that has been judged as poorly motivated by (Venkatesh et al., 2023)), (Finn & Lizier, 2020), who use the algebraic structure of information sharing, (Ay et al., 2019), who rely on cooperative game theory, (Rosas et al., 2020), who build on concepts related to data privacy and disclosure, (Kolchinsky, 2019), who use set theory, (van Enk, 2023), who deal with scalability issues by pooling probabilities, (Gutknecht et al., 2023), who use a mereological formulation, and (Makkeh et al., 2021; Ehrlich et al., 2023), who advocate for methods based on the exclusions of probability mass. Nevertheless, the main limitations of PID persist in all variants. Indeed, computational complexity grows extremely fast, precisely as the Dedekind number of variables (which is more than 10^{31} for 9 variables). Moreover, PID computation relies on a partition of the system into a set of sources and a unique target. This can be an artificial distinction which limits usability and interpretability of the results. This latter problem is

[1]Ampere Software Technology, France [2]Department of Data Science, Eurecom, France. Correspondence to: <mustapha.bounoua@eurecom.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

partially addressed in (Varley et al., 2023), who introduce Partial Entropy Decomposition (PED).

Motivated by these limitations, (Rosas et al., 2019) introduce the concept of O-INFORMATION, a measure which captures the synergy-redundancy dominance in multivariate systems. In contrast to PID, this measure does not require the system to be partitioned into sources and a target, and gracefully scales in the number of its components (Martinez Mediano, 2022). Furthermore, recent extensions such as O-INFORMATION locality (Scagliarini et al., 2021) and gradient computation (Scagliarini et al., 2023) allow a fine-grained analysis of system behavior. However, O-INFORMATION measures are accessible only in restricted scenarios. Indeed, existing methods rely on estimation techniques that requires either i) discrete distributions (or binning of continuous ones) or ii) Gaussian distributions. In this work, we show that such limitations can be lifted by using and extending recent methods to estimate MI (Franzese et al., 2024; Kong et al., 2024).

Our work is organized as follows: § 2 introduces the high-dimensional interaction measures which we investigate in this work, while § 3 proposes Score-based O-Information estimation (SΩI), our novel methodology which allows scalable and flexible O-INFORMATION estimation. § 4 validates experimentally our proposed method, where we report a series of compelling results on various synthetic systems, for which ground truth values are known and accessible analytically. Furthermore, we consider a realistic endeavor by revisiting previous studies (Venkatesh et al., 2023) that focus on the analysis of brain activity in mice. Our method allows lifting previous limiting assumptions, and allow synergy-redundancy characterizations that are compatible with observations made by domain experts. Finally, we summarize our findings in § 5.

2. High dimensional interaction measures

Consider the continuous **multivariate** random variable $X = \{X^1, \dots, X^N\} \sim p(x^1, \dots, x^N)$. We indicate the collection of all but the i_{th} random variable with the symbol $X^{\setminus i} \stackrel{\text{def}}{=} \{X^1, .., X^{i-1}, X^{i+1}, .., X^N\}$. When necessary, we indicate marginal and conditional distributions by properly specifying the arguments of the distribution, e.g. $X^i \sim p(x^i)$ or $X^{\setminus i} \,|\, X^i \sim p(x^1, \dots, x^{i-1}, x^i, \dots, x^M \,|\, x^i)$.

A central quantity in this work is the Shannon entropy associated to a given random variable $\mathcal{H}(X) \stackrel{\text{def}}{=} \mathbb{E}\left[-\log p(X)\right]$ (Cover et al., 1991). Considering the case of bi-variate (i.e. $N = 2$) random variable X, entropy and conditional entropy allow computation of the mutual information (MI) flow $\mathcal{I}$ between the two random variables X^1, X^2: $\mathcal{I}(X^1; X^2) = \mathcal{H}(X^1) - \mathcal{H}(X^1|X^2)$, where $\mathcal{H}(X^1|X^2) = \mathbb{E}\left[-\log p(X^1 \,|\, X^2)\right]$. Importantly, such quantity can also be expressed as the Kullback-Leibler (KL) divergence (Cover et al., 1991) between the joint and the product of marginal distributions: $\mathcal{I}(X^1; X^2) = \text{KL}\left[p(x^1, x^2) \,\|\, p(x^1)p(x^2)\right]$. For the case of $N = 3$, it is possible to define the MI as $\mathcal{I}(X^1; X^2; X^3) = \mathcal{I}(X^1; X^2) - \mathcal{I}(X^1; X^2|X^3)$, where $\mathcal{I}(X^1; X^2|X^3) = \mathcal{H}(X^1 \,|\, X^3) - \mathcal{H}(X^1 \,|\, X^2, X^3)$. This quantity, also known as co-information or interaction information, can counter-intuitively result in a negative value, and measures the difference between synergistic and redundant interactions (Rosas et al., 2019).

Since, for $N > 3$, interaction information becomes difficult to grasp (Williams & Beer, 2010; Rosas et al., 2019), our goal in this work is to consider extensions to MI, while preserving interpretability. In particular, a measure of the interaction strengths in a system with $N > 3$ can be obtained by studying the summand mutual information between one variable and the rest of the system:

$$\mathcal{S}(X) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \mathcal{I}(X^i; X^{\setminus i}). \tag{1}$$

This quantity, named S-INFORMATION, can be decomposed into the redundant and synergistic components of the considered multivariate system. In particular, since $X^{\setminus i} = \{X^{<i}, X^{>i}\}$, where $X^{<i} = \{X^1, \dots, X^{i-1}\}$ and $X^{>i} = \{X^{i+1}, \dots, X^N\}$ (with $X^{>N} = \emptyset$), we can use the conditional mutual information laws (Cover et al., 1991) and rewrite $\mathcal{S}(X)$ as:

$$\mathcal{S}(X) = \sum_{i=1}^{N} \mathcal{I}(X^i; X^{>i}) + \sum_{i=1}^{N} \mathcal{I}(X^i; X^{<i} \,|\, X^{>i}). \tag{2}$$

The two **positive** series which constitute $\mathcal{S}(X)$ are equivalent to the Total Correlation (TC) (Sun, 1975) and the Dual Total Correlation (DTC) (Sun Han, 1980) denoted by $\mathcal{T}(.)$ and $\mathcal{D}(.)$ respectively. Then, $\mathcal{S}(X) = \mathcal{T}(X) + \mathcal{D}(X)$, where (proof in Appendix A)

$$\mathcal{T}(X) = \sum_{i=1}^{N} \mathcal{H}(X^i) - \mathcal{H}(X), \tag{3}$$

$$\mathcal{D}(X) = \mathcal{H}(X) - \sum_{i=1}^{N} \mathcal{H}(X^i|X^{\setminus i}). \tag{4}$$

TC is high in cases where, for each variable X^i, at least one of its "*children*" (variables in $X^{>i}$) carries information about it. Importantly, the number of children conveying information (whether 1, 2, or $N-1$) is irrelevant. Since $\mathcal{T}(X)$ is permutation invariant, a high value implies that for every ordering of the variables, and hence for all possible combinations of children of a given variable, the summand mutual information between variables and their children remains high.

This intuition, which suggests *redundancy*, can similarly be obtained by considering the entropic formulation. Indeed, whenever a system is composed of perfectly independent variables ($X^i \perp X^j, i \neq j$) $\mathcal{H}(X) = \sum_{i=1}^N \mathcal{H}(X^i)$ and consequently $\mathcal{T}(X) = 0$. On the other hand, a *copy* system ($X^i = X^j, \forall i, j$) achieves infinite $\mathcal{T}(X)$, as $\mathcal{H}(X) = -\infty$, since the support of the joint distribution is on a lower than $N-$dimensional space. TC also admits a representation in terms of KL divergences, $\mathcal{T}(X) = \text{KL}\left[p(x) \parallel \prod_{i=1}^N p(x^i)\right]$, which we will exploit later in our proposed methodology.

Similar considerations can be carried out for the DTC. Consider a single MI term $I(X^i; X^{<i}|X^{>i})$. The focus of this conditioning is about quantifying how much **additional** information the variables $X^{<i}$ carry about X^i if we are also given access $X^{>i}$. Whenever the variables are independent or redundant (the copy system), this value is identically zero. However, whenever the aid of the *extra* measurements unlocks new bits of information, which suggests a *synergistic* scenario, its value is positive.

Having recognized that $\mathcal{S}(X)$ in a multivariate system can be decomposed into measures of redundancy $\mathcal{T}(X)$ and synergy $\mathcal{D}(X)$, we can introduce a new information theoretic measure which quantifies the difference between the two behaviours. This quantity, named O-INFORMATION (Rosas et al., 2019), is defined as

$$\Omega(X) = \mathcal{T}(X) - \mathcal{D}(X). \tag{5}$$

In summary, while S-INFORMATION only quantifies the strength of interactions in a system, O-INFORMATION also determines the *nature* of these interactions, being them redundant or synergistic. Intuitively, a redundancy-dominated system is the most parsimonious explanation — in an Occam's razor sense — whenever $\Omega(X) > 0$. Conversely, a negative value $\Omega(X) < 0$ is associated with a synergy-dominated system. O-INFORMATION is a natural generalization of MI for more than 3 variables: indeed, it is equal to the co-information for $N = 3$, and is a measure which preserves interpretability for any positive N.

One important property of O-INFORMATION is that it gracefully scales with the number of random variables composing a system, as opposed to, e.g. the PID measure, which has much worse scalability.

Since O-INFORMATION measures the *overall* information dynamics among variables, recent work focus on ways to study the *individual* influence of variables to the high-order interactions, and capture the interaction structure of a multivariate system (Scagliarini et al., 2023). The first order difference, called the *gradient* of O-INFORMATION, captures how much O-INFORMATION changes when adding or removing a given system variable i:

$$\partial_i \Omega(X) = \Omega(X) - \Omega(X^{\setminus i}). \tag{6}$$

A positive value implies that X^i provides redundant information to the system, while a negative one suggests that its interaction with other variables is mainly synergistic.

3. Score-based O-INFORMATION estimation

O-INFORMATION and its gradient represent extremely useful information theoretic measures to study multivariate systems. However, as it is clear from Equations (3) to (5), their estimation requires access to entropies, conditional entropies and KL divergence measures. When strict assumptions about the distribution of variables composing the system are possible, such as discrete or Gaussian distributions, existing implementations of O-INFORMATION estimators have been used successfully in a number of application domains (Varley et al., 2022; Sparacino et al., 2023; Stramaglia et al., 2021; Chiarion et al., 2023). However, in more realistic cases where such assumptions are not valid, there currently does not exist a method to estimate the constituents of O-INFORMATION in a reliable and scalable manner. In this work, we present the first methodology allowing estimation of O-INFORMATION for more general scenarios. Our method unfolds according to the observation that all quantities of interest can be expressed in terms of KL divergences, and relies on a technique to estimate such divergences which scales gracefully with the system size. Our key ingredient is the score function associated to data distributions (Vincent, 2011; Song & Ermon, 2019) and the method we present leverages recent advances in the field of MI estimation (Franzese et al., 2024; Kong et al., 2024).

3.1. Score-based divergence estimation

Consider the generic multivariate random variable X with associated distribution $p(x)$. Provided that certain minimal regularity assumptions are met (Vincent, 2011), it is always possible to associate the distribution $p(x)$ to its *score function*, defined as the gradient of its logarithm, $\nabla \log p(x)$.

Recently, the community has showed tremendous interest (Song & Ermon, 2019; Song et al., 2021) in a generalization of such concept, which involves computing the score function of a *noised* version of the variable X, due to the possibility of adopting such concept for generative modelling purposes. Accordingly, in this work we define a noised version of the variable X with corresponding intensity indexed by $t \in [0, \infty)$. Then, the new variable is constructed as $X_t = X + \sqrt{2t}W$, where W is a Gaussian random vector with the same dimension of X, zero mean and identity covariance matrix.

This new random variable can be associated to its *time-*

varying score function $\nabla \log p_t(x)$. In particular the analytic expression of $p_t(x)$ can be obtained as the solution of the Partial Differential Equation (PDE) $\frac{\mathrm{d}p_t(x)}{\mathrm{d}t} = \Delta p_t(x)$, with initial conditions given by $p_0(x) = p(x)$.

Next, we consider the KL divergence between two generic distributions and define how it can be computed using score functions, a result which we will use later for computing O-INFORMATION.

Proposition 1. *(Franzese et al., 2024; Kong et al., 2024) The* KL *divergence between two generic distributions $p(x)$ and $q(x)$, defined as*

$$\mathrm{KL}\left[p(x) \parallel q(x)\right] = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x,$$

can be computed considering the time-varying score functions $\nabla \log(p_t)$ and $\nabla \log(q_t)$, according to the following expression:

$$\mathrm{KL}\left[p(x) \parallel q(x)\right] = \int p_t(x) \left\| \nabla \log \left(\frac{p_t(x)}{q_t(x)} \right) \right\|^2 \mathrm{d}x\mathrm{d}t.$$

Proof sketch. To avoid clutter, we drop the dependence on x of the distributions. Let's define $r_t \stackrel{\text{def}}{=} \int p_t \log \frac{p_t}{q_t} \mathrm{d}x$.

Since it holds that $r_\infty - \mathrm{KL}\left[p \parallel q\right] = \int_0^\infty \frac{\mathrm{d}r_t}{\mathrm{d}t} \mathrm{d}t$, we need

$$\int \frac{\mathrm{d}r_t}{\mathrm{d}t} \mathrm{d}t = \int \frac{\mathrm{d}p_t}{\mathrm{d}t} \log \left(\frac{p_t}{q_t} \right) + p_t \frac{\mathrm{d}}{\mathrm{d}t} \log \left(\frac{p_t}{q_t} \right) \mathrm{d}x\mathrm{d}t.$$

Note that $\int p_t \frac{\mathrm{d}}{\mathrm{d}t} \log \left(\frac{p_t}{q_t} \right) \mathrm{d}x\mathrm{d}t = \int \frac{\mathrm{d}}{\mathrm{d}t} p_t - \frac{p_t}{q_t} \Delta q_t \mathrm{d}x\mathrm{d}t$, and $\int \frac{\mathrm{d}}{\mathrm{d}t} p_t \mathrm{d}x\mathrm{d}t = 0$ (See Appendix A.1 for detailed proof). Then, the expression above can be rewritten as $\int p_t \Delta \log \left(\frac{p_t}{q_t} \right) - \frac{p_t}{q_t} \Delta q_t \mathrm{d}x\mathrm{d}t$. Integrating by parts we obtain $\int -\nabla p_t \nabla \log \left(\frac{p_t}{q_t} \right) + \nabla (\frac{p_t}{q_t}) \nabla q_t \mathrm{d}x\mathrm{d}t$. Since $\nabla p_t = p_t \nabla \log p_t$ and $\nabla (\frac{p_t}{q_t}) \nabla q_t = p_t \nabla \log q_t \nabla \log \left(\frac{p_t}{q_t} \right)$, and $r_\infty = 0$ (Franzese et al., 2023; Villani, 2009; Collet & Malrieu, 2008), the proposition follows. □

The result in Proposition 1 allows, in principle, the exact computation of KL divergences, provided knowledge of the score functions $\nabla \log p_t, \nabla \log q_t$. Such knowledge is however out of reach in practical cases, which is why in this work we consider a *parametric* approximation of such vector fields, leading to a KL divergence *estimator*. In particular, we leverage the methodology considered in (Song & Ermon, 2019; Song et al., 2021) where the parametric score s_t is obtained by minimizing the so called *denoising score-matching* loss

$$\int p(x) p_{0t}(\tilde{x} \mid x) \| s_t(\tilde{x}) - \nabla \log(p_{0t}(\tilde{x} \mid x)) \|^2 \mathrm{d}x\mathrm{d}\tilde{x}\mathrm{d}t,$$

where $p_{0t}(\tilde{x} \mid x)$ is the conditional distribution of the noised random variable *given* initial conditions $X = x$, i.e. $p_t(\tilde{x}) = \int p_{0t}(\tilde{x} \mid x) p(x) \mathrm{d}x$. Note that p_{0t} has known Gaussian distribution with mean x and variance $2t$. This allows, together with the knowledge of the score functions, the implementation of an estimator for the KL divergence.

Informally, learning the score can be understood as learning to *denoise* the variable X_t to obtain X. Indeed, the score functions have analytic expression $\nabla \log p_t(x) = \frac{\mathbb{E}[X \mid X_t = x] - x}{2t}$, where the only unknown is $\mathbb{E}[X \mid X_t = x]$. An alternative, but equivalent parametrization of the problem, consists in estimating the noise W, given X_t. We use this approach in our work since is considered to be more stable numerically (Ho et al., 2020). In practice, the VP-SDE (Song et al., 2021) framework is adopted as the noising process. With such a schedule varying between $[0, T]$, it's valid to assume that X_T is practically indistinguishable from pure noise (More details in Appendix B).

3.2. Estimating O-INFORMATION

Armed with Proposition 1, we can leverage score functions to estimate the information-theoretic quantities introduced in § 2. Here we consider an extension of the simple noising process described in § 3.1, where we allow i) noising of only certain subsets of the variables or ii) deletion of subset of variables. In practice, the first case corresponds to learning to denoise a portion of the variables, given auxiliary information about the other (noiseless) variables, e.g. to learn $\mathbb{E}[X^i \mid X_t^i = \tilde{x}^i, X^{\setminus i} = x^{\setminus i}]$. Instead, the second case amounts to denoising problems akin to $\mathbb{E}[X^i \mid X_t^i = \tilde{x}^i]$. In our implementation, we follow the approach proposed in (Bounoua et al., 2024) (See Appendix B). Next, we use such an intuition to derive a series of propositions that pave the way to O-INFORMATION computation.

In what follows, we use the compact notation $\left[(\cdot)^i\right]_{i=1}^N$, to indicate a concatenation of N elements in a column vector.

Proposition 2. *Given a multivariate random variable $X = \{X^1, \ldots, X^N\} \sim p(x^1, \ldots, x^N)$, and its corresponding noised version, the Total Correlation $\mathcal{T}(X)$ is equal to:*

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

Proof Sketch. Recall that $\mathcal{T}(X) = \mathrm{KL}\left[p(x) \parallel \prod_{i=1}^N p(x^i)\right]$. Then, by virtue of Proposition 1, we have that $\mathcal{T}(X)$ equals

$$\int p_t(x) \left\| \nabla \log p_t(x) - \left[\frac{\partial}{\partial x^i} \log p_t(x^i)\right]_{i=1}^N \right\|^2 \mathrm{d}x\mathrm{d}t.$$

The terms $\frac{\partial}{\partial x^i} \log p_t(x^i)$ correspond to $1/2t(\mathbb{E}[X^i \mid X_t^i = x^i] - x^i)$. Then, the proposition follows. □

Proposition 3. *Given a multivariate random variable $X = \{X^1, \dots, X^N\} \sim p(x^1, \dots, x^N)$, and its corresponding noised version, the* S-INFORMATION $\mathcal{S}(X)$ *is equal to:*

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

Proof Sketch. In light of Equation (1), it holds that

$$\mathcal{S}(X) = \sum_{i=1}^N \int p(x^{\setminus i}) \mathrm{KL}\left[p(x^i \mid x^{\setminus i}) \parallel p(x^i)\right] \mathrm{d}x^{\setminus i},$$

where the i_{th} KL term of the sum is equal to (Proposition 1)

$$\int p(x^i \mid x^{\setminus i}) p_{0t}(\tilde{x}^i \mid x^i)$$
$$\left\| \frac{\partial}{\partial \tilde{x}^i} \log\left(\frac{p_t(\tilde{x}^i)}{\hat{p}_{0t}(\tilde{x}^i \mid x^{\setminus i})}\right)\right\|^2 \mathrm{d}\tilde{x}^i \mathrm{d}x^i \mathrm{d}t.$$

Now, we can move the terms $p(x^{\setminus i})$ inside the KL computation integrals and write the sum of the norms as the norm of a vector, which allows computing S-INFORMATION as

$$\mathcal{S}(X) = \int p(x) p_{0t}(\tilde{x} \mid x)$$
$$\left\| \left[\tfrac{\partial}{\partial \tilde{x}^i} \log p_t(\tilde{x}^i)\right]_{i=1}^N - \left[\tfrac{\partial}{\partial \tilde{x}^i} \log p_t(\tilde{x}^i \mid x^{\setminus i})\right]_{i=1}^N \right\|^2 \mathrm{d}\tilde{x} \mathrm{d}x \mathrm{d}t,$$

where $p_t(\tilde{x}^i \mid x^{\setminus i}) = \int p_{0t}(\tilde{x}^i \mid x^i) p(x^i \mid x^{\setminus i}) \mathrm{d}x^i$.

Finally, the proposition follows since we can interpret the elements inside the square norm in terms of denoisers, with $\frac{\partial}{\partial \tilde{x}^i} \log p_t(\tilde{x}^i \mid x^{\setminus i}) = 1/2t(\mathbb{E}[X^i \mid X_t^i = \tilde{x}^i, X^{\setminus i} = x^{\setminus i}] - \tilde{x}^i)$ □

Proposition 4. *Given a multivariate random variable $X = \{X^1, \dots, X^N\} \sim p(x^1, \dots, x^N)$, and its corresponding noised version, the Dual Total Correlation $\mathcal{D}(X)$ equals:*

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

Proof Sketch. The starting point to obtain DTC is to recall that $\mathcal{D}(X) = \mathcal{S}(X) - \mathcal{T}(X)$. Then, it is sufficient to expand the square norms of $\mathcal{S}(X)$ and $\mathcal{T}(X)$ and combine the different terms, to state that $\mathcal{D}(X)$ equals:

$$\int p(x) p_{0t}(\tilde{x} \mid x)$$
$$\left\| \nabla \log p_t(\tilde{x}) - \left[\tfrac{\partial}{\partial \tilde{x}^i} \log p_t(\tilde{x}^i \mid x^{\setminus i})\right]_{i=1}^N \right\|^2 \mathrm{d}\tilde{x} \mathrm{d}x \mathrm{d}t.$$

This can be proven considering that i) $\mathbb{E}\left[\mathbb{E}[X^i \mid X_t]\mathbb{E}[X^i \mid X_t^i]\right] = \mathbb{E}[(\mathbb{E}[X^i \mid X_t^i])^2]$ ii) $\mathbb{E}\left[\mathbb{E}[X^i \mid X_t^i]\mathbb{E}[X^i \mid X_t^i, X^{\setminus i}]\right] = \mathbb{E}\left[(\mathbb{E}[X^i \mid X_t^i])^2\right]$ and iii) $\mathbb{E}\left[\mathbb{E}[X^i \mid X_t]\mathbb{E}[X^i \mid X_t^i, X^{\setminus i}]\right] = \mathbb{E}[(\mathbb{E}[X^i \mid X_t^i])^2]$. Then, the proposition follows. □

Finally, to estimate O-INFORMATION, it is sufficient to combine Proposition 2 and Proposition 4, and apply Equation (5). In practical terms, our method requires access to *denoisers* for the three following scenarios: i) given X_t estimate X ii) given X_t^i estimate X^i iii) given X_t^i and $X^{\setminus i}$ estimate X^i. To achieve this, we extend the methodology proposed in (Bounoua et al., 2024), and amortize the three different scenarios with a *unique denoising network*, which takes as input the concatenation of noised and clean variables and outputs the corresponding estimates (see Appendix B). Additionally, the estimation of the gradients of O-information requires approximating additional denoising score functions to access Equation (9) (More details in Appendix B.2).

4. Experimental validation

We evaluate our method according to two strategies. First, we focus on a synthetic setup that allows analytic computation of O-INFORMATION and full control on system scale. Then, we consider real data collected in a study of brain activity in mice, to demonstrate how SΩI unlocks new avenues in the application of information measures in real systems without the need for restrictive assumptions.

4.1. Synthetic benchmark

We consider a canonical Gaussian system, whereby we control the number of variables describing the system N, the dimension of each variable (**Dim**), the inter-dependencies between variables describing how they interact, and the strength of interaction (More details in Appendix B). Inspired by (Czyż et al., 2023), we consider more challenging distribution going beyond the Gaussian setting (Please refer to Appendix E). No other neural estimator capable of estimating O-INFORMATION was explored in the literature. Next, we construct an original baseline that relies on neural estimation of MI to access the MI decomposition of O-INFORMATION.

Baseline. Recent work (Bai et al., 2023) describes a method to compute TC by leveraging a decomposition into pairwise MI terms. Clearly, DTC can also be decomposed into MI terms. Therefore, we extend (Bai et al., 2023) such that it can be used as a baseline to compute O-INFORMATION. The main limitation of such a baseline is poor scalability: it requires training an individual model for each MI term in which TC and DTC are decomposed in. We adopt the linear-decomposition method (Bai et al., 2023), which results in $2(N-1)$ MI terms (see Appendix C), and propose four variants to estimate MI based on (Belghazi et al., 2018; Nguyen et al., 2007; Oord et al., 2018; Cheng et al., 2020). We label this baseline approach according to the MI estimators: MINE ,NWJ, INFONCE, and CLUB.

Experimental protocol For each experiment, we use $100k$ samples for training the various neural estimators, and $10k$ samples at inference time, to estimate O-INFORMATION. For our method SΩI, we use the VP-SDE formulation (Song et al., 2021) and learn a *unique* denoising network to estimate the various score terms. The denoiser is a simple, stacked multilayer perceptron (MLP) with skip connections, adapted to the input dimension. We apply importance sampling (Huang et al., 2021; Song et al., 2021) at both training and inference time. Finally, we use 10-sample Monte Carlo estimates for computing integrals. More details about the implementation are included in Appendix C. For the baseline variants, for each MI term we use an MLP that is sufficiently expressive given the data dimension. All results are averaged over 5 seeds. Additional results are included in Appendix F.

Our experiments unfold according to three inter-dependency scenarios, for systems characterized by either redundancy, synergy or a mix of both interactions.

Redundancy benchmark. We consider $R = \mathcal{N}(0, \mathbb{I})$ as the redundant information component in the system. All system variables are of the form $X = \{X^1, \dots, X^N\} = \{R + \epsilon_i, \dots, R + \epsilon_N\}$, where $\epsilon_i \sim \mathcal{N}(0, \sigma\mathbb{I})$ are mutually independent random noise samples with standard deviation σ. We use σ to modulate the redundancy level: higher noise levels decrease the strength of redundant interaction, and this has an impact on the value of O-INFORMATION.

Next, we discuss results for a system with $N = 10$ variables, organized as 3 redundant subsystem, each defined as described above. Figure 1 illustrates, for various variable dimension, ranging from 5 to 20 dimensional Gaussians, the ground-truth and the estimated O-INFORMATION, for SΩI and the various baselines. In this scenario, SΩI and baseline competitors produce fairly accurate O-INFORMATION estimates, when the dimensionality of each random variable is small. When the dimension of systems variables grows, however, the performance of the baseline methods degrades considerably. This is due to the inherent limitations of the pairwise neural MI estimators, that struggle with high dimensional data (Czyż et al., 2023). Instead, the performance of SΩI remains stable when increasing variable dimension, and O-INFORMATION estimates are accurate, even when interaction strength is high.

Synergy benchmark. In this case, we synthesize synergistic inter-dependency among system variables by considering the following setup. For simplicity, consider three random variables that behave as follows:

$$X^1 \sim \mathcal{N}(0, \mathbb{I}), \quad X^2 = X^1 + S$$
$$X^3 = S + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma) \text{ and } X^1 \perp\!\!\!\perp X^3$$

with $S \sim \mathcal{N}(0, \mathbb{I})$.

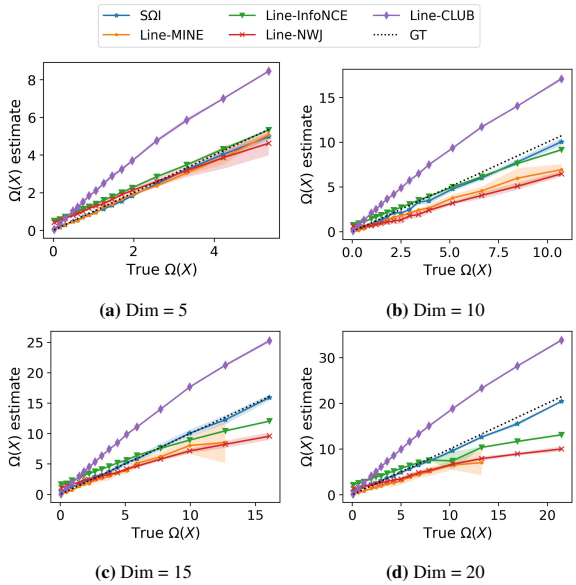


Figure 1. Redundant system with N =10 variables, organized into subsets of sizes $\{3, 3, 4\}$ and increasing interaction strength.

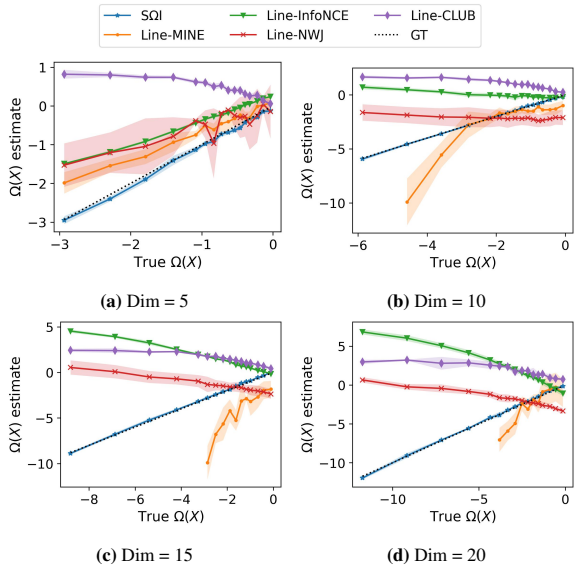
When $\sigma = 0$, the synergy emerges through the Markov chain $\{X^2, X^3\} - X^1$, $\{X^1, X^3\} - X^2$ and $\{X^1, X^2\} - X^3$, since no element alone is sufficient to recover the remaining variables. We modulate σ to achieve different synergistic strengths. More generally, we simulate N synergistic variables as: $X^1 \sim \mathcal{N}(0, \mathbb{I})$, $X^2 = X^1 + S_1 + ... + S_{N-2}$ and $X^i = S_{i-2} + \epsilon_{i-2} \forall i \in \{3, .., N\}$.

Results in Fig. 2 show that SΩI achieves consistent results in all scenarios, whereas the baselines behave poorly. Indeed, a synergy-only setting is challenging, as it's dominated by high DTC values required to capture high-order interactions, on which the baselines based on pairwise MI estimator fail.

Mixed benchmark. In general, systems components are characterized by a mix of redundant and synergistic interactions. Then, we synthesize such a system by creating subgroups dominated by redundancy and synergy, respectively, following the procedures defined above.

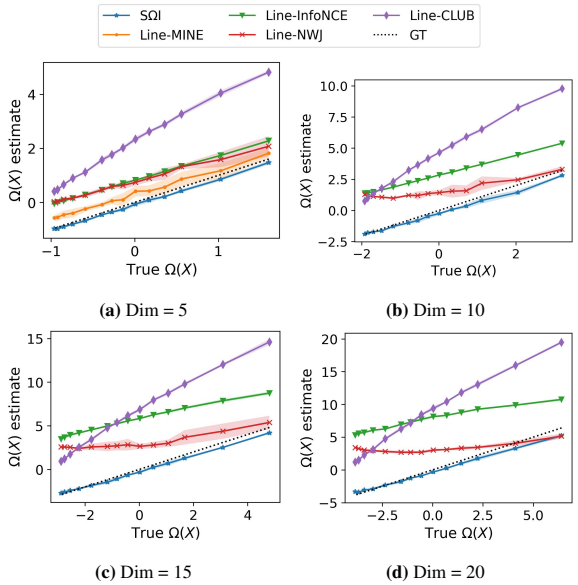
Results in Fig. 3, demonstrate that our method SΩI stands out as the best estimator in this challenging scenario. Baseline methods produce poor estimates, especially when the synergistic interaction is dominant. Note that SΩI reports a negative O-INFORMATION whenever the system is synergy-dominant and also succeeds in capturing interaction strengths, when the system equilibrium changes in favor redundant interactions, by estimating correctly a positive O-INFORMATION.

Discussion. We attribute the superior performance of SΩI, compared to the baselines, to several factors. Score-based



(a) Dim = 5 **(b)** Dim = 10 **(c)** Dim = 15 **(d)** Dim = 20

Figure 2. Synergistic system with $N = 10$ variables, organized into subsets of sizes $\{3, 3, 4\}$ and increasing interaction strength.



(a) Dim = 5 **(b)** Dim = 10 **(c)** Dim = 15 **(d)** Dim = 20

Figure 3. Mixed-interaction system with $N = 10$ variables, organized into 2 redundancy-dominant subsets of size $\{3, 4\}$ variables and one synergy-dominant subset with 3 variables. O-INFORMATION is modulated by fixing the synergy interdependency and increasing the redundancy.

estimators have shown to be extremely successful in fitting complex distributions, for example in the context of generative modeling (Song & Ermon, 2020; Song et al., 2021). Moreover, our technique relies on Proposition 1, whereby the difference of score functions has been shown to produce an accurate estimate of KL divergences, due to canceling effects of estimation errors (Franzese et al., 2024). Note also that the baselines we adopt in our work use MI estimators that produce a bound only. Moreover, using individual models to estimate several MI terms can naturally suffer from cumulative bias, which is avoided in our case by amortizing computation with a unique neural network.

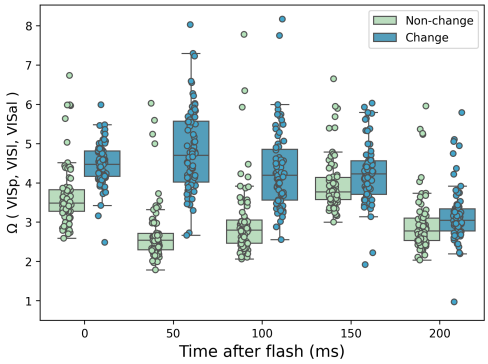
Gradient of O-information While O-INFORMATION provides global information about dominance of either synergy or redundancy, the contribution of individual variables to either effects is not available. Next, we rely on the gradient of O-INFORMATION to study individual system components, as introduced in § 2. Indeed, our method SΩI can be easily extended to output such gradients, by estimating additional score functions, as described in Appendix B. In Figure 4, we illustrate gradients of O-INFORMATION applied to the mixed benchmark scenario discussed above. While O-INFORMATION of the whole system can be positive due to the redundancy strength of some subgroup of variables, we notice that three variables report a negative gradient, which is indicative of their synergistic interaction. In Figure 4, ground truth gradient values are showed using a diamond marker. Our estimator, despite suffering from some bias, correctly attributes the role and interaction type of each system constituent.

4.2. Application to a real system

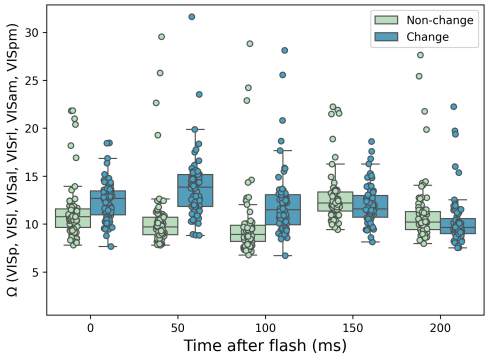
Multivariate analysis is a powerful tool for the field of neuroscience, as it allows scientists to analyze activity patterns of different brain regions. Understanding how the brain processes and transmits information during different stimulus requires analysing the underlying inter-dependencies between different brain regions. To show that SΩI is an effective tool also in practical use cases, we now consider the Visual Behavior project, which used the Allen Brain Observatory to collect a highly standardized dataset consisting of recordings of neural activity in mice that have learned to perform a visually guided task (Allen-Institute, 2022).

A visual change-detection task experiment was conducted on 80 mice using six neuropixels probes tasked to report the activity of different regions of the visual cortex. During the recordings, a set of 8 natural scenes were presented in 250 ms flashes, at intervals of 750 ms. The same image was shown during several flashes before a change to a new image. The mouse had to perform an action to receive a water reward when the image changed. Ultimately, the purpose of this experiment is to investigate how the different brain region of the mice react to different types of stimulus, such as detecting a new image (*change*) or not (*no change*).

In this work, we follow the prepossessing procedure described by (Venkatesh et al., 2023), where in each experimental session, good quality units from each area are chosen (See Appendix C). For each trial, the recorded spikes are

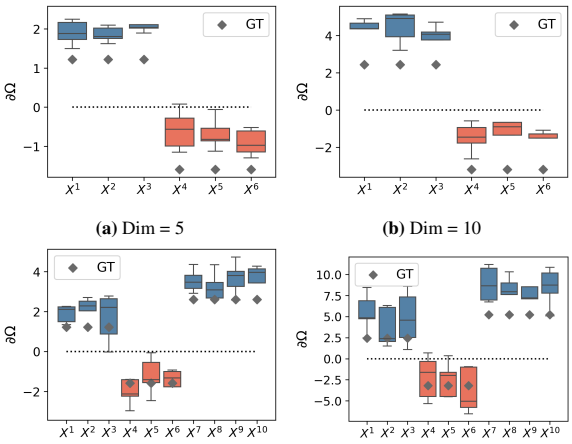


(a) 3 areas



(b) 6 areas

Figure 5. O-INFORMATION estimate in the visual cortex region activity after two types of stimulus flash across 72 trial sessions. Top: Analysis using three brain region areas, Bottom: Extended analysis using six brain region areas. The step size is set to $2ms$ which results in 25 dimensional data for each bin per area. Different step sizes led to the same behavior (see Appendix F).



(a) Dim = 5 **(b)** Dim = 10

(c) Dim = 5 **(d)** Dim = 10

Figure 4. Gradient of O-INFORMATION for the mixed benchmark, for a system of N =6 variables, and a system of N =10 variables, and different dimension of variables.

binned in 50 ms intervals, starting from the stimulus flash. We consider two types of flashes: *change* and *no change*. For both cases, SΩI is used for each time bin to estimate O-information (O-INFORMATION). The reported estimation is done using 10 Monte Carlo integration steps and averaged over multiple seeds. We first consider three visual cortex regions VISP, VISL and VISAL, as done in (Venkatesh et al., 2023). We then extend the experiment to six brain regions by including VISRL, VISAM and VISPM.

We show our results in Figure 5, where the distribution of O-INFORMATION values are reported as box-plots for each bin. We remark that values of O-INFORMATION are higher in cases of *change* stimulus, and lower for the *no change* stimulus. This suggests that higher amount of redundant information in the visual cortex regions is transmitted in case of a flash with new scene. Interestingly, when considering six areas of the visual cortex, our observations remain valid, suggesting that the measured behaviour is common to these other brain areas as well. Our results are aligned with (Venkatesh et al., 2023). However, prior work rely on the PID measure, which requires the brain regions to be artificially organized into two areas and a target variable, due to scalability issues affecting PID. Our work confirms that SΩI does not have such a limitation and allows a single estimation procedure to obtain the same conclusions.

5. Conclusion

We addressed the problem of analyzing multivariate systems, whereby the essence of complexity does not only lie in the nature of the individual system components, but also in the structure of their inter-dependencies. Indeed, the analysis of high-order interaction among variables has emerged as an important tool to deepen our understanding of such complex systems, with application domains including machine learning, neuroscience, climate modeling, and many more.

Recently, the scientific community has spent considerable effort on extending information theory to allow the study of complex, multivariate systems according to notions of uniqueness, redundancy and synergy. While no consensus exists yet, on a information measure that can fully and reliably characterize high-order interactions, in this work we focused on O-INFORMATION, which has desirable properties such as interpretability and scalability in number of variables. The current state of the art is however at a roadblock. The existing techniques rely on strong assumptions on the data distribution. Additionally, we explore an exhaustive use of the neural MI estimators to access the O-INFORMATION which resulted in sub-optimal performance and scalability issues. Then, the endeavour of our work was to present a method to lift such limitations, and endow practitioners and scientists with a flexible and reliable tool to study complex systems associated to natural phenomena.

In this paper, we proposed SΩI, a novel technique that leverages recent neural estimators of mutual information and uses score functions of joint and conditional distributions to compute divergences. We showed that SΩI can compute O-INFORMATION by training a unique parametric model, which is efficient and flexible. We validated our technique with a comprehensive experimental protocol, both in synthetic and realistic settings. We demonstrated that SΩI is accurate and robust across different system configurations and complexities. We also applied SΩI to a case study of mice brain activity, where we obtained plausible and interpretable results, and showcased the scalability of SΩI to handle larger systems than previously possible. We believe that our work contributes to a substantial advancement of information measures computation and their applications to real-world, complex systems.

Acknowledgment

Pietro Michiardi was partially funded by project MUSE-COM^2 - AI-enabled MUltimodal SEmantic COMmunications and COMputing, in the Machine Learning-based Communication Systems, towards Wireless AI (WAI), Call 2022, ChistERA.

Impact Statement

This paper presents work to improve current methods to compute information measures of complex systems, modeled as ensembles of multiple random variables. Such information measures have been recently brought to the attention of the scientific community, for their potential in explaining the high-order interactions between systems part, and specifically to understand information redundancy, uniqueness and synergy. Applications of such measures range from multi-modal machine learning, neuroscience, climate modeling and many more. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Allen-Institute. Visual behavior neuropixels dataset overview. 2022. URL https://portal.brain-map.org/explore/circuits/visual-behavior-neuropixels.

Ay, N., Polani, D., and Virgo, N. Information decomposition based on cooperative game theory. *ArXiv*, abs/1910.05979, 2019. URL https://api.semanticscholar.org/CorpusID:204512236.

Bai, K., Cheng, P., Hao, W., Henao, R., and Carin, L. Estimating total correlation with mutual information estimators. In *International Conference on Artificial Intelligence and Statistics*, pp. 2147–2164. PMLR, 2023.

Barrett, A. B. An exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *CoRR*, abs/1411.2832, 2014. URL http://arxiv.org/abs/1411.2832.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Bounoua, M., Franzese, G., and Michiardi, P. Multi-modal latent diffusion. *Entropy*, 26(4), 2024. ISSN 1099-4300. doi: 10.3390/e26040320. URL https://www.mdpi.com/1099-4300/26/4/320.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

Chiarion, G., Sparacino, L., Antonacci, Y., Faes, L., and Mesin, L. Connectivity analysis in eeg data: A tutorial review of the state of the art and emerging trends. *Bioengineering*, 10(3), 2023. ISSN 2306-5354. doi: 10.3390/bioengineering10030372. URL https://www.mdpi.com/2306-5354/10/3/372.

Collet, J.-F. and Malrieu, F. Logarithmic sobolev inequalities for inhomogeneous markov semigroups. *ESAIM: Probability and Statistics*, 12:492–504, 2008.

Cover, T. M., Thomas, J. A., et al. Entropy, relative entropy and mutual information. *Elements of information theory*, 2(1):12–13, 1991.

Czyż, P., Grabowski, F., Vogt, J. E., Beerenwinkel, N., and Marx, A. Beyond normal: On the evaluation of mutual information estimators. *Advances in Neural Information Processing Systems*, 2023.

Dosi, G. and Roventini, A. More is different... and complex! the case for agent-based macroeconomics. *Journal of Evolutionary Economics*, 29:1–37, 2019.

Ehrlich, D. A., Schick-Poland, K., Makkeh, A., Lanfermann, F., Wollstadt, P., and Wibral, M. Partial information decomposition for continuous variables based on shared exclusions: Analytical formulation and estimation. *arXiv preprint arXiv:2311.06373*, 2023.

Finn, C. and Lizier, J. T. Generalised measures of multivariate information content. *Entropy*, 22(2), 2020. ISSN 1099-4300. doi: 10.3390/e22020216. URL https://www.mdpi.com/1099-4300/22/2/216.

Franzese, G., Rossi, S., Yang, L., Finamore, A., Rossi, D., Filippone, M., and Michiardi, P. How much is enough? a study on diffusion times in score-based generative models. *Entropy*, 2023.

Franzese, G., BOUNOUA, M., and Michiardi, P. MINDE: Mutual information neural diffusion estimation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0kWd8SJq8d.

Ganmor, E., Segev, R., and Schneidman, E. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of sciences*, 108(23):9679–9684, 2011.

Gat, I. and Tishby, N. Synergy and redundancy among brain cells of behaving monkeys. *Advances in neural information processing systems*, 11, 1998.

Gutknecht, A. J., Makkeh, A., and Wibral, M. From babel to boole: The logical organization of information decompositions. *ArXiv*, abs/2306.00734, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Huang, C.-W., Lim, J. H., and Courville, A. C. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34:22863–22876, 2021.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Kolchinsky, A. A novel approach to multivariate redundancy and synergy. *CoRR*, abs/1908.08642, 2019. URL http://arxiv.org/abs/1908.08642.

Kong, X., Liu, O., Li, H., Yogatama, D., and Steeg, G. V. Interpretable diffusion via information decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=X6tNkN6ate.

Latham, P. E. and Nirenberg, S. Synergy, redundancy, and independence in population codes, revisited. *Journal of Neuroscience*, 25(21):5195–5206, 2005.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Makkeh, A., Gutknecht, A. J., and Wibral, M. Introducing a differentiable measure of pointwise shared information. *Physical Review E*, 103(3):032149, 2021.

Martinez Mediano, P. A. Integrated information theory in complex neural systems. 2022.

Nguyen, X., Wainwright, M. J., and Jordan, M. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2007.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *Advances in neural information processing systems*, 2018.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Rosas, F. E., Mediano, P. A. M., Gastpar, M., and Jensen, H. J. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical review. E*, 100 3-1:032305, 2019. URL https://api.semanticscholar.org/CorpusID:67855406.

Rosas, F. E., Mediano, P. A. M., Rassouli, B., and Barrett, A. An operational information decomposition via synergistic disclosure. *Journal of Physics A: Mathematical and Theoretical*, 53, 2020. URL https://api.semanticscholar.org/CorpusID:210932609.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

Scagliarini, T., Marinazzo, D., Guo, Y., Stramaglia, S., and Rosas, F. E. Quantifying high-order interdependencies on individual patterns via the local o-information: Theory and applications to music analysis. *Physical Review Research*, 2021. URL https://api.semanticscholar.org/CorpusID:237303787.

Scagliarini, T., Nuzzi, D., Antonacci, Y., Faes, L., Rosas, F., Marinazzo, D., and Stramaglia, S. Gradients of o-information: Low-order descriptors of high-order dependencies. *Physical Review Research*, 5, 01 2023. doi: 10.1103/PhysRevResearch.5.013025.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448. Curran Associates, Inc., 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Sparacino, L., Faes, L., Mijatović, G., Parla, G., Re, V. L., Miraglia, R., de Ville de Goyet, J., and Sparacia, G. Statistical approaches to identify pairwise and high-order brain functional connectivity signatures on a single-subject basis. *Life*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:264314627.

Stramaglia, S., Scagliarini, T., Daniels, B. C., and Marinazzo, D. Quantifying dynamical high-order interdependencies from the o-information: An application to neural spiking dynamics. *Frontiers in Physiology*, 11, 2021. ISSN 1664-042X. doi: 10.3389/fphys.2020.595736. URL https://www.frontiersin.org/articles/10.3389/fphys.2020.595736.

Sun, T. Linear dependence structure of the entropy space. *Inf Control*, 29(4):337–68, 1975.

Sun Han, T. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1):26–45, 1980. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(80)90478-7. URL https://www.sciencedirect.com/science/article/pii/S0019995880904787.

Tax, T. M., Mediano, P. A., and Shanahan, M. The partial information decomposition of generative neural network models. *Entropy*, 19(9), 2017. ISSN 1099-4300. doi: 10.3390/e19090474. URL https://www.mdpi.com/1099-4300/19/9/474.

van Enk, S. J. Pooling probability distributions and partial information decomposition. *Physical review. E*, 107 5-1:054133, 2023. URL https://api.semanticscholar.org/CorpusID:256615444.

Varley, T. F., Pope, M., Faskowitz, J., and Sporns, O. Multivariate information theory uncovers synergistic subsystems of the human cerebral cortex. *Communications Biology*, 6, 2022. URL https://api.semanticscholar.org/CorpusID:249642639.

Varley, T. F., Pope, M., Puxeddu, M. G., Faskowitz, J., and Sporns, O. Partial entropy decomposition reveals higher-order information structures in human brain activity. *Proceedings of the National Academy of Sciences of the United States of America*, 120, 2023. URL https://api.semanticscholar.org/CorpusID:255825886.

Venkatesh, P., Bennett, C., Gale, S., Ramirez, T. K., Heller, G., Durand, S., Olsen, S. R., and Mihalas, S. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1PnSOKQKvq.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Williams, P. L. and Beer, R. D. Nonnegative decomposition of multivariate information, 2010.

Score-based O-INFORMATION Estimation — Supplementary material

A. Proofs

A.1. Detailed proof of Proposition 1

Here we provide the full proof for Proposition 1 (to avoid unnecessary complications, we assume the 1-d case, the vector proof is identical). Starting from the equation :

$$C = \int \frac{dp_t}{dt} \log\left(\frac{p_t}{q_t}\right) + p_t \frac{d}{dt} \log\left(\frac{p_t}{q_t}\right) dxdt$$

Concerning the first part of the integral:

$$\int \frac{dp_t}{dt} \log\left(\frac{p_t}{q_t}\right) dxdt = \int \Delta(p_t) \log\left(\frac{p_t}{q_t}\right) dxdt = \int p_t \Delta(\log\left(\frac{p_t}{q_t}\right)) dxdt,$$

Where the first equality is simply due to $\frac{dp_t}{dt} = \Delta p_t$, and the second is obtained by properties of the adjoint of the Δ operator. In particular, we need to perform a double application of integration by parts, where we should remember that densities p_t, q_t are equal to zero at infinite values of x and that $\Delta = \nabla\nabla$.

Focusing on the second part of the integral:

$$\int p_t \frac{d}{dt} \log\left(\frac{p_t}{q_t}\right) dxdt = \int p_t (\frac{d \log p_t}{dt} - \frac{d \log q_t}{dt}) dxdt = \int p_t (\frac{\frac{dp_t}{dt}}{p_t} - \frac{\frac{dq_t}{dt}}{q_t}) dxdt$$

The first summand $p_t \frac{\frac{dp_t}{dt}}{p_t}$ simplifies to $\frac{dp_t}{dt}$.

Since $\int \frac{dp_t}{dt} dxdt = \int \frac{d}{dt}(\int p_t dx) dt = \int \frac{d}{dt}(1) dt = 0$, this term is cancelled.

The second is transformed as :

$p_t \frac{\frac{dq_t}{dt}}{q_t} = \frac{p_t}{q_t} \frac{dq_t}{dt} = \frac{p_t}{q_t} \Delta q_t$ where again we leveraged $\frac{dq_t}{dt} = \Delta q_t$.

Consequently, we obtain:

$$C = \int p_t \Delta \log\left(\frac{p_t}{q_t}\right) - \frac{p_t}{q_t} \Delta q_t dxdt$$

We apply one step of integration by parts on both Δ operators and obtain :

$$\int -\nabla p_t \nabla \log\left(\frac{p_t}{q_t}\right) + \nabla(\frac{p_t}{q_t}) \nabla q_t dxdt$$

The remaining missing clarification in the sketch proof of Proposition 1 is that :

$$\nabla(\frac{p_t}{q_t}) \nabla(q_t) = \frac{\nabla(p_t) q_t - \nabla(q_t) p_t}{q_t^2} \nabla(q_t) =$$
$$\frac{\nabla(p_t)}{q_t} \nabla(q_t) - p_t (\frac{\nabla q_t}{q_t})^2 = \nabla p_t \nabla(\log(q_t)) - p_t (\nabla(\log q_t))^2 =$$
$$p_t \nabla(\log p_t) \nabla(\log(q_t)) - p_t (\nabla(\log q_t))^2 = p_t \nabla(\log q_t)(\nabla(\log p_t) - \nabla(\log q_t)) = p_t \nabla(\log q_t)(\nabla(\log \frac{p_t}{q_t}))$$

A.2. TC and DTC equivalences

We here prove the equivalences about TC and DTC. Starting from TC :

$$\sum_{i=1}^{N} \mathcal{H}(X^i) - \mathcal{H}(X) = \sum_{i=1}^{N} \mathcal{H}(X^i) - \sum_{i=1}^{N} \mathcal{H}(X^i \,|\, X^{>i}) = \sum_{i=1}^{N-1} \mathcal{I}(X^i; X^{>i}) = \mathcal{T}(X)$$

Concerning DTC

$$\mathcal{H}(X) - \sum_{i=1}^{N} \mathcal{H}(X^i|X^{\setminus i}) = \mathcal{H}(X^1) + \mathcal{H}(X^{\setminus 1} \,|\, X^1) - \mathcal{H}(X^1|X^{\setminus 1}) - \sum_{i=2}^{N} \mathcal{H}(X^i|X^{\setminus i}) =$$

$$\mathcal{I}(X^1; X^{\setminus 1}) + \mathcal{H}(X^{\setminus 1} \,|\, X^1) - \sum_{i=2}^{N} \mathcal{H}(X^i|X^{\setminus i}) =$$

$$\mathcal{I}(X^1; X^{\setminus 1}) + \mathcal{H}(X^2 \,|\, X^1) + \mathcal{H}(X^{\setminus 1,2} \,|\, X^1, X^2) - \mathcal{H}(X^2|X^{\setminus 2}) - \sum_{i=3}^{N} \mathcal{H}(X^i|X^{\setminus i}) =$$

$$\mathcal{I}(X^1; X^{\setminus 1}) + \mathcal{I}(X^2; X^{>2}|X^1) + \mathcal{H}(X^{\setminus 1,2} \,|\, X^1, X^2) - \sum_{i=3}^{N} \mathcal{H}(X^i|X^{\setminus i}) = \dots$$

$$\sum_{i=1}^{N-1} \mathcal{I}(X^i; X^{>i} \,|\, X^{<i}) = \mathcal{D}(X)$$

Where for the last equality it suffices to consider trivial reordering arguments, $\sum_{i=2}^{N} \mathcal{I}(X^i; X^{<i} \,|\, X^{>i}) = \sum_{i=1}^{N-1} \mathcal{I}(X^i; X^{>i} \,|\, X^{<i})$.

B. Details of SΩI

In the section we provide additional implementation details about SΩI.

B.1. Computing O-INFORMATION

In § 3.1, we presented how TC and DTC can be estimated using denoising score functions. Our estimators requires different score functions which can be obtained by learning different denoisers. More particularly, TC requires the joint denoiser $\mathbb{E}[X \mid X_t]$ and the marginals $\mathbb{E}[X^i \mid X_t^i]$ for $i \in \{1, \dots, N\}$. DTC estimation is obtained using the joint and the following conditional terms $\mathbb{E}[X^i \mid X_t^i, X^{\setminus i}]$ for $i \in \{1, \dots, N\}$. Our formulation in § 3.1 is general and can be applied to a wide range of denoising score learning techniques. For the implementation of SΩI, we adopt VP-Stochastic Differential Equation (SDE) framework (Song & Ermon, 2019). The latter perturbs the data using an SDE parameterized by a drift f_t and a diffusion coefficient g_t.

Muti-variate denoising score network. We extend the work from (Bounoua et al., 2024) to amortize the learning of all the required terms using a **unique** denoising score network. The denoising score network ϵ_θ accepts as input the concatenation of the variables each perturbed at different times. The second input is a vector of size N which describes the state of each variable and allows a parametrization of different denoising score functions.

The joint term corresponds to the case where all the variables are perturbed with the same intensity t and all the elements of the vector $\tau = [t, \dots, t]$ are set equivalently to t. The conditional terms correspond to the case where only the conditioned variable i is perturbed with intensity t whereas the remaining conditioning variables $\setminus i_{\text{th}}$ are kept unperturbed at $t = 0$. Consequently the parameter describing this case is of the form $[0, \dots, t, \dots, 0]$.

While (Bounoua et al., 2024) framework is not able to learn the marginal denoising score, it's possible via an additional parameterization to include this configuration. This corresponds to the case where the marginal variable i is perturbed with intensity t while all the other variables are made uninformative. The non marginal variables $\setminus i_{\text{th}}$ are replaced with pure noise corresponding to a maximal perturbation at $t = T$. Consequently the parameter describing this case is of the form $[T, \dots, t, \dots, T]$.

Training. The training is carried out through a randomized procedure. At each training step, we select randomly a set of the denoising score functions required for the O-INFORMATION estimation (joint, conditional or marginals). These denoising scores function are learned by the unique network following Algorithm 1. In total, estimating O-INFORMATION requires calling $2N + 1$ denoising score functions which we learn using a unique denoising network.

Algorithm 1 SΩI Training step

Data: $X = \{X^i\}_{i=1}^N$
$t \sim \mathcal{U}[0, T]$ // Importance sampling schemes (Huang et al., 2021; Song et al., 2021) can be adopted to reduce variance
if *Joint* **then**
  $X_t \sim p_t$ // Obtain noisy version of all the variables using VPSDE (Song & Ermon, 2019) with drift f_t and diffusion coefficient g_t.
  $s_t(X_t) = \epsilon_\theta([X_t^1, \dots, X_t^N], \tau = [t, \dots, t, \dots, t])$
  Return $\nabla_\theta \| s_t(X_t) - \nabla \log p_t(X_t|X) \|$ // Denoising score matching of all the variables
if *Conditional* **then**
  $X_t^i \sim p_t$ // Obtain noisy version of the variable i while the remaining variables are kept unperturbed at $(t = 0)$
  $s_t(X_t^i | X^{\setminus i}) = \epsilon_\theta \left([X^1, \dots, X^{i-1}, X_t^i, X^{i+1}, \dots, X^N], \tau = [0, \dots, t, \dots, 0]\right)$
  Return $\nabla_\theta \left\| s_t(X_t^i | X^{\setminus i}) - \nabla \log p_t(X_t^i | X^i) \right\|$ // Denoising score matching of the conditioning variable i
if *Marginal* **then**
  $X_t^i \sim p_t$
  $X_T^{\setminus i} \leftarrow p_T = \mathcal{N}(0, \mathbb{I})$ // Obtain noisy version of the variable i while the remaining variables are replaced with pure noise $(t = T)$.
  $s_t(X_t^i) = \epsilon_\theta([X_T^1, \dots, X_T^{i-1}, X_t^i, X_T^{i+1}, \dots, X_T^N], \tau = [T, \dots, t, \dots, T])$
  Return $\nabla_\theta \left\| s_t(X_t^i) - \nabla \log p_t(X_t^i | X^i) \right\|$ // Denoising score matching of the marginal variable i

Inference. Once all the denoising score functions are learned, it's possible to estimate TC and DTC via a Monte Carlo estimation of the integral over t in Proposition 2 and Proposition 4 . The outer integration w.r.t. to the time instant is possible by sampling $t \sim \mathcal{U}(0, T)$, and then using the estimation $\int_0^T (\cdot) \mathrm{d}t = T \mathbb{E}_{t \sim \mathcal{U}(0,T)}[(\cdot)]$. In practice we adopt 10 steps for the computation of the expectation. The procedure to estimate O-INFORMATION is described in algorithm 2. First, samples from $x \sim p(x)$ are considered, then sampling the time $t \sim \mathcal{U}[0, T]$. A perturbed version of the variables X_t is computed using the Variance preserving SDE (VPSDE). The joint, conditional and marginal denoising scores are computed leveraging the unique denoising score network. This is possible by choosing different perturbation times and manipulating the vector τ as described earlier. Computing the difference of the denoising scores functions (see Proposition 2 and Proposition 2) allows the computation of TC and DTC respectively. Please note that it is possible to implement importance sampling schemes to reduce the variance, along the lines of what described by Huang et al. (2021).

Algorithm 2 SΩI inference time

Data: $X = \{X^i\}_{i=1}^N$
$t \sim \mathcal{U}[0, T]$ // Importance sampling scheme can also be adopted
$X_t \sim p_t$ // Obtain the noisy version of all the variables using VPSDE (Song & Ermon, 2019) with drift f_t and diffusion coefficient g_t.

$s_t(X_t) \leftarrow \epsilon_\theta([X_t^1, \dots, X_t^N], \tau = [t, \dots, t, \dots, t])$ // Compute the joint score

for *i = 1 to N* // Compute the conditional and marginal terms
do
$s_t(X_t^i | X^{\setminus i}) \leftarrow \epsilon_\theta\left([X^1, \dots, X^{i-1}, X_t^i, X^{i+1}, \dots, X^N], \tau = [0, \dots, t, \dots, 0]\right)$

$s_t(X_t^i) \leftarrow \epsilon_\theta([X_T^1, \dots, X_T^{i-1}, X_t^i, X_T^{i+1}, \dots, X_T^N], \tau = [T, \dots, t, \dots, T])$ // Similarly to Algorithm 1 the non marginal variables are replaced with pure noise $X_T^{\setminus i} \sim \mathcal{N}(0, \mathbb{I})$
end

$\hat{\mathcal{T}}(X) \leftarrow \frac{g_t^2}{2} \left\| s_t(X_t) - \left[s_t(X_t^i)\right]_{i=1}^N \right\|^2$ // See Proposition 2
$\hat{\mathcal{D}}(X) \leftarrow \frac{g_t^2}{2} \left\| s_t(X_t) - \left[s_t(X_t^i | X^{\setminus i})\right]_{i=1}^N \right\|^2$ // See Proposition 4
$\hat{\Omega}(X) \leftarrow \hat{\mathcal{T}}(X) - \hat{\mathcal{D}}(X)$
Return $\hat{\Omega}(X)$

B.2. Computing gradient of O-INFORMATION

To compute the gradient of O-INFORMATION recall that $\partial_i \Omega(X) = \Omega(X) - \Omega(X^{\setminus i})$. The first order gradient of O-INFORMATION requires the estimation of O-INFORMATION of all the subsystems of size $N - 1$.

$$\Omega(X^{\setminus i}) = \mathcal{T}(X^{\setminus i}) - \mathcal{D}(X^{\setminus i}) \tag{7}$$

$$= \sum_{j=1, j \neq i}^{N} \mathcal{H}(X^j) - \mathcal{H}(X^{\setminus i}) \tag{8}$$

$$- (\mathcal{H}(X^{\setminus i}) - \sum_{j=1, j \neq i}^{N} \mathcal{H}(X^j | X^{\setminus \{i,j\}})) \tag{9}$$

It's possible to use an alternative formulation to estimate the gradient of O-INFORMATION based on MI terms:

$$\partial_i \Omega(X) = (2 - N) \mathcal{I}(X^i, X^{\setminus i}) + \sum_{j=1, j \neq i}^{N} \mathcal{I}(X^i, X^{\setminus \{i,j\}}) \tag{10}$$

$$= (2 - N) \left[\mathcal{H}(X^i) - \mathcal{H}(X^i | X^{\setminus i})\right] + \sum_{j=1, j \neq i}^{N} \mathcal{H}(X^i) - \mathcal{H}(X^i | X^{\setminus \{i,j\}}) \tag{11}$$

Many denoising score functions in Equation (9) were also used to estimate the global O-INFORMATION. To learn the additional necessary terms to compute $\Omega(X^{\setminus i})$, the randomized set of scores adopted during the training step (see Appendix B.1) is extended to account for the new requirements. Please note that we still use a unique denoising network that considers all the terms necessary to compute O-INFORMATION and its gradient. A large number of learned denoising score functions is a potential reason for the bias observed in our experiment Figure 4. A highly flexible architecture capable of fitting large number of scores may be needed to infer gradient of O-INFORMATION.

C. Experimental settings

C.1. Canonical multivariate Gaussian system

In this section we provide additional details about the construction of the synthetic benchmark § 4.1.

Redundancy benchmark. All the variable of the system are composed of a redundant component and unique information specific to each variable.

We modulate the redundant inter-dependency strength by setting different values for σ. We consider a standardized system where all the variables mean is 0 and standard deviation equal to $\mathbb{I}$. This results in the following covariance matrix:

$$\begin{bmatrix} \mathbb{I} & \rho\mathbb{I} & \vdots & \rho\mathbb{I} \\ \rho\mathbb{I} & \mathbb{I} & \dots & \rho\mathbb{I} \\ \vdots & \vdots & \ddots & \rho\mathbb{I} \\ \rho\mathbb{I} & \rho\mathbb{I} & \dots & \mathbb{I} \end{bmatrix} \tag{12}$$

With $\rho = \frac{1}{1+\sigma^2}$ which modulates the interactions strength in the system.

Synergy benchmark. We consider a standardized system where all the variables mean is 0 and standard deviation equal to $\mathbb{I}$. This results in the following covariance matrix :

$$\begin{bmatrix} \mathbb{I} & \frac{1}{\sqrt{N-1}}\mathbb{I} & 0 & \dots & 0 \\ \frac{1}{\sqrt{N-1}}\mathbb{I} & \mathbb{I} & \frac{\rho}{\sqrt{N-1}}\mathbb{I} & \dots & \frac{\rho}{\sqrt{N-1}}\mathbb{I} \\ 0 & \frac{\rho}{\sqrt{N-1}}\mathbb{I} & \mathbb{I} & \dots & 0 \\ 0 & \vdots & 0 & \ddots & 0 \\ 0 & \frac{\rho}{\sqrt{N-1}} & 0 & \dots & \mathbb{I} \end{bmatrix} \tag{13}$$

Where $\rho = \frac{1}{\sqrt{1+\sigma^2}}$ modulates the interactions strength in the system.

Mixed benchmark. The covariance matrix is easy to obtain as the mixed benchmark is made of independent subsystems.

Ground Truth. Having access to the covariance matrix of the system, computing entropy in close form for Gaussian distribution is possible. For $X \sim \mathcal{N}(\mu, \sigma)$:

$$\mathcal{H}(X) = \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2} \tag{14}$$

For a multivariate Gaussian distribution $X^d \sim \mathcal{N}_d(\mu, \Sigma)$:

$$\mathcal{H}(X) = \frac{D}{2}(1 + \log(2\pi)) + \frac{1}{2}\log det(\Sigma) \tag{15}$$

C.2. SΩI implementation details

We provide code-base for SΩI implementation at [1]. The training of SΩI is carried out using *Adam optimizer* (Kingma & Ba, 2015). We use Exponential moving average (EMA) with a momentum parameter $m = 0.999$. Importance sampling (Huang et al., 2021) ([2]) at train and test-time. The hyper-parameters are presented in Table 1. To estimate the gradient of O-INFORMATION (Figure 4) the model width is double the one presented in Table 1 to account for the additional necessary terms to learn. Concerning the experiments in Figure 5 , we use the same architecture used for the canonical examples and follow the same procedure to choose the model capacity(see Table 1 for the hyper-parameters details).

[1] https://github.com/MustaphaBounoua/soi

[2] https://github.com/CW-Huang/sdeflow-light

Table 1. SΩI network training details. Dim of the task correspond the sum of the dimensions of all variables of the system. For the neural data application we report the number of training iterations (.,.) corresponding the "change" case and "No change" case. The number of iteration used for the "No change" is higher since the dataset contains more "no change" flashes compared to "change" flashes.

	Width	Time embed	Batch size	Lr	Iterations	Number of params
$(Dim \leq 50)$	128	128	256	1e-2	195k	320k
$(Dim \leq 100)$	192	192	256	1e-2	195k	747k
$(Dim \geq 100)$	256	256	256	1e-2	195k	1003k
Neural application						
$(Dim \leq 30)$	128	128	256	1e-2	(100k,160k)	320k
$(Dim \leq 75)$	192	192	256	1e-2	(100k,160k)	737k
$(Dim \leq 150)$	256	256	256	1e-2	(100k,160k)	1300k
$(Dim \geq 150)$	384	384	256	1e-2	(100k,160k)	3000k

C.3. Baselines

(Bai et al., 2023) decomposes TC into $N-1$ MI terms which are estimated using pairwise neural MI estimator. Similarly by leveraging Equation (18) DTC can also be retrieved by estimating $N-1$ additional MI terms.

$$\mathcal{T}(X) = \sum_{i=1}^{N-1} \mathcal{I}(X^i; X^{>i}) \tag{16}$$

$$\mathcal{D}(X) = \mathcal{S}(X) - \mathcal{T}(X) = \sum_{i=1}^{N} \mathcal{I}(X^i; X^{\setminus i}) - \mathcal{T}(X) \tag{17}$$

$$\mathcal{D}(X) = \sum_{i=2}^{N} \mathcal{I}(X^i; X^{\setminus i}) - \sum_{i=2}^{N-1} \mathcal{I}(X^i; X^{>i}) \tag{18}$$

Our implementation in based on the the official codebase [3] of (Bai et al., 2023). We use the same architecture and hyper parameters from (Bai et al., 2023): LR $= 1e-3$, Batch size $= 64$. We use an MLP architecture for all the variants of the baseline with 3 linear layers with varying width. For each MI term, the capacity of the neural network is aligned to the input dimension. *Adam optimizer* (Kingma & Ba, 2015) is used for training. We increase the width of the hidden layer to accommodate the data dimension. For the variant of the baseline implemented with MINE, we used smaller layer size as large capacity led to divergence during training. To ensure the best performance, we train each MI estimator model for $80k$ steps for a number of variables $N = 10$ and $40k$ for number of variables $N = 6$. In the different experiments, we reported the performance results averaged over 5 seeds and dropped the baseline in case of divergence during training.

Limitations of the baseline in computing gradients of O-INFORMATION It's possible to leverage the decomposition of (Bai et al., 2023), using the compact gradient of O-information formulation Figure 6.

This will require N MI term for each $\partial_i \Omega(X)$. Consequently to compute all the terms, it's required to train $N * N$ pairwise MI models. While it's possible to leverage some MI terms, if already estimated for the computation of O-information, the overall complexity remains of order $\mathcal{O}(N^2)$.

This naturally raises a scalability problem in training a large number of neural estimator models. Moreover, as the number of MI terms increases, this approach is likely to suffer from cumulative errors observed when estimating O-information.

To compute the gradient of O-information with SΩI, we are instead required to approximate an additional number of denoising score functions. However, our method SΩI amortizes the training costs : we use a unique score network to approximate all the required score functions.

[3] https://github.com/Linear95/TC-estimation

C.4. The Visual Behavior Neuropixels

Hereafter we describe the different pre-processing steps applied on the Visual Behavior Neuropixels in § 4.2. We follow the same procedure described by (Venkatesh et al., 2023). The selected mice are the ones with both familiar and novel sessions and a minimum number of 20 units in each of the six brain regions: VISP, VISL, VISAL ,VISRL, VISAM and VISPM. Only the units of good quality are kept. The selection criteria was based on an SNR at least 1, and with fewer than 1 inter-spike interval violations. The non-change flashes correspond to the ones where the image does not change and happen between 4 and 10 flashes after the trial start. Trials corresponding to a change are naturally the ones when the image has changed. Only flashes that occurred while the animal was engaged (based on the reward information) is kept, while the ones corresponding to an omission, or after an omission, and flashes during which the animal licked, were all removed.

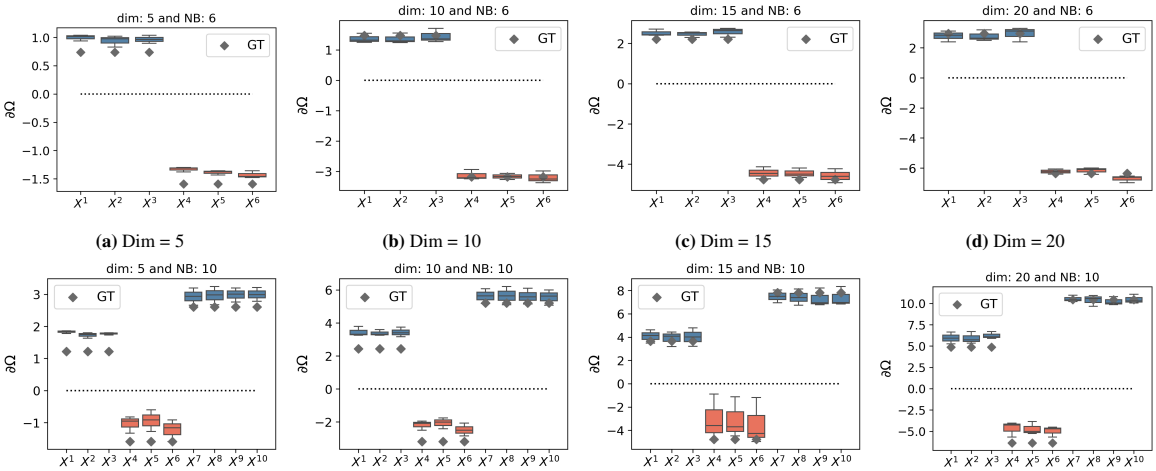
The trials were aligned to the start of each stimulus flash, and the 250ms recordings were divided into 5 bins of 50 ms duration averaged over the units of the same region. We use different step sizes to count the spikes which resulted in different dimensional representation but resulted in the same intuition (See Figure 22,Figure 21 and Figure 20). Please note that unlike (Venkatesh et al., 2023), we don't use PCA to reduce the dimension of the data, and count the number of spikes per unit by averaging the activity over the units of the same region indexed by time.

D. A transformer based SΩI

Throughout our experimental campaign as referenced in § 4, we employed an MLP structure enhanced with skip connections. While this setup reliably estimated O-INFORMATION, it produced perfectible gradient of O-INFORMATION estimation. We address this shortcoming by integrating a more robust architecture capable of scaling with an increased number of denoising score functions. Our approach is based on the latest developments in denoising score matching, incorporating a transformer-based model.

Our method is simple: we adopt the architecture from (Peebles & Xie, 2023) to learn the denoising score functions, treating each modality as a distinct token, while substituting any non-marginal modality with a NULL token (a token with zero value). A transformer block is employed to learn the conditional signal, which is subsequently merged with the temporal signal. This conditioning employs the adaLN-Zero configuration. Our model consists of 4 Blocks, each with 6 attention heads, and the width of the transformer's linear layers is scaled according to the dimension size of the benchmark.The training follows a randomized approach akin to that detailed in § 5 eliminating the need for a multi-time vector. To compute gradient of O-INFORMATION, we utilize the formulation presented in Equation (11).

The results presented in Figure 6 demonstrate the ability of SΩI to accurately estimate the gradients of O-INFORMATION, provided that the denoising network has sufficient capacity to approximate all the denoising score functions.



(a) Dim = 5 **(b)** Dim = 10 **(c)** Dim = 15 **(d)** Dim = 20

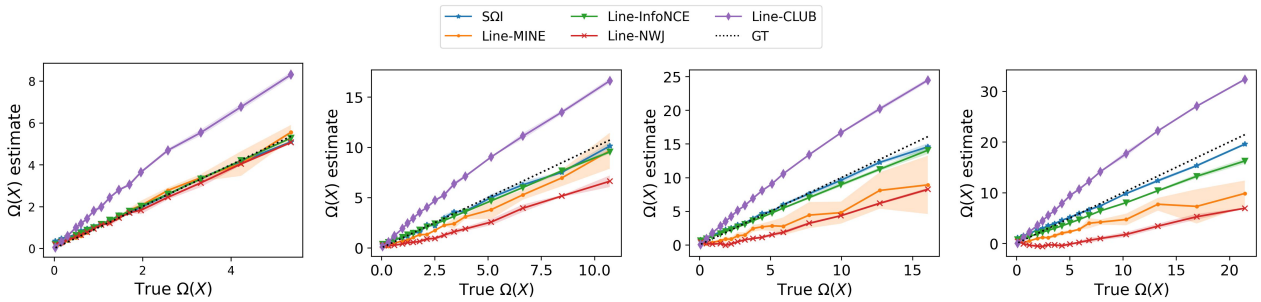
(e) Dim = 5 **(f)** Dim = 10 **(g)** Dim = 15 **(h)** Dim = 20

Figure 6. Gradient of O-INFORMATION using a **transformer based architecture** for the mixed benchmark, for a system of 6 variables, and a system of 10 variables, and different dimension of variables.

E. Beyond Normal Benchmarks

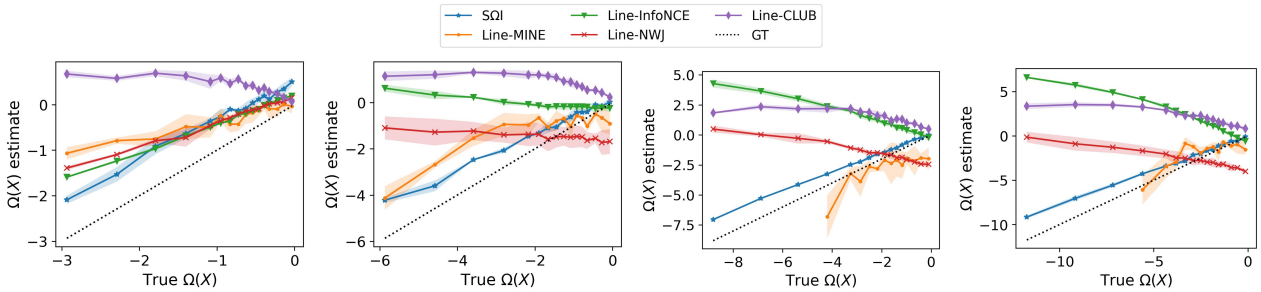
In this section, we evaluate SΩI and alternatives across more challenging distributions. To construct such settings we apply MI-invariant transformations to the benchmarks established in Section § 4. Since TC and DTC can be written in terms of MI terms, the in-variance of O-INFORMATION to MI invariant transformations is self-evident.

Half-cube $x \to x\sqrt{|x|}$ is recognized as an MI invariant transformation, which serves to lengthen the tail of the distribution. Addressing the long tail distribution poses a significant challenge for neural MI estimators, as highlighted in recent studies by(Franzese et al., 2024; Czyż et al., 2023). InFigure 7,Figure 8 and Figure 9, we showcase the performance outcomes of SΩI and other baselines on half-Cube transformed benchmarks that exhibit similar interactions as detailed in § 4. Our approach stands out by delivering superior performance. Notably, the synergistic transformed benchmark emerges as the most demanding scenario: competitors suffer particularly with high-dimensional variables, while SΩI shows bias, especially in cases of high synergistic interactions, indicated by very low O-INFORMATION values.



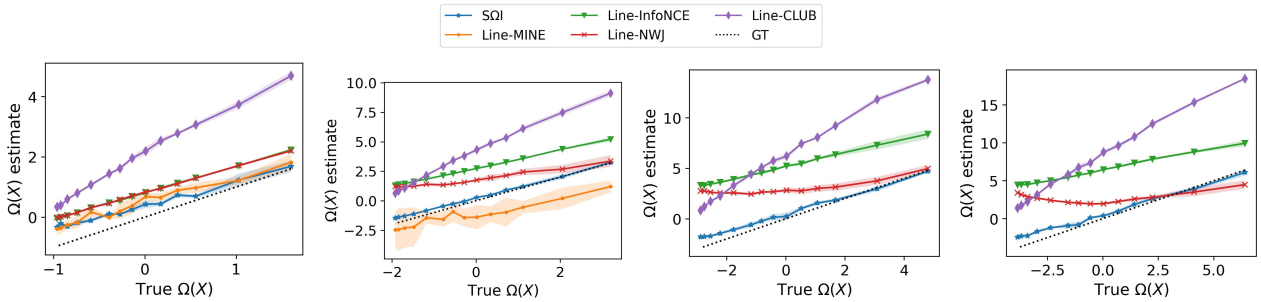
(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 7. Redundant system with 10 variables, organized into subsets of sizes $\{3, 3, 4\}$ and increasing interaction strength. A **half-cube** transformation is applied on-top of the multi-normal distribution



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

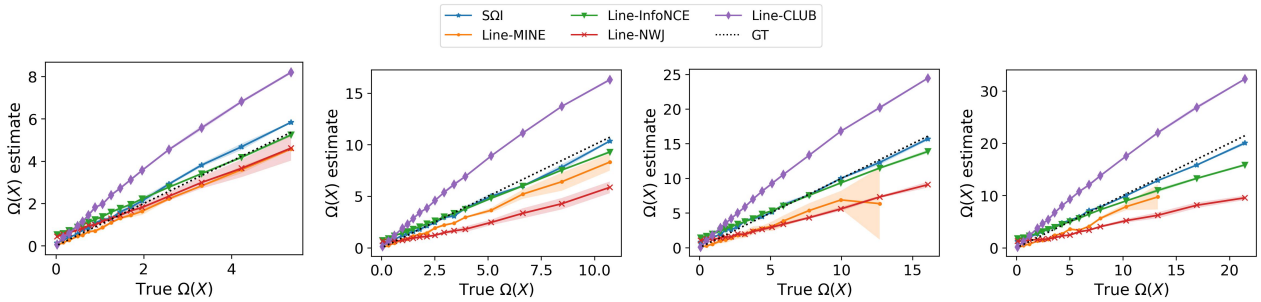
Figure 8. Synergistic system with 10 variables, organized into subsets of sizes $\{3, 3, 4\}$ and increasing interaction strength. A **half-cube** transformation is applied on-top of the multi-normal distribution.



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

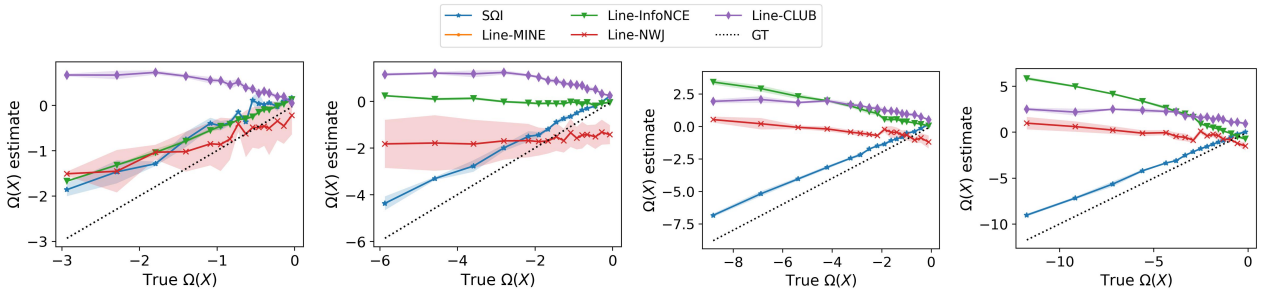
Figure 9. Mixed-interaction system with 10 variables, organized into 2 redundancy-dominant subsets of size $\{3, 4\}$ variables and one synergy-dominant subset with 3 variables. O-INFORMATION is modulated by fixing the synergy inter-dependency and increasing the redundancy. A **half-cube** transformation is applied on-top of the multivariate-normal distribution.

CDF The second transformation we consider is the application of a normal cumulative distribution function (CDF), which uniformizes the distribution margins (See (Czyż et al., 2023)). InFigure 10,Figure 11 and Figure 12, we present the performance results of SΩI and alternatives on CDF-transformed benchmarks with a similar configuration used in § 4. Our method outperforms competitors, especially for high-dimensional variables. On the challenging synergistic benchmark, SΩI shows perfectible performance for very low O-INFORMATION, while competitors fail completely in this setting.



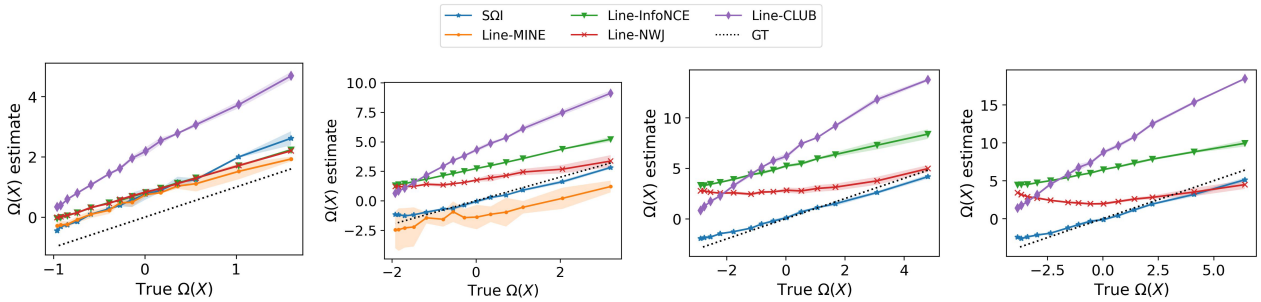
(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 10. Redundant system with 10 variables, organized into subsets of sizes $\{3, 3, 4\}$ and increasing interaction strength. A **CDF** transformation is applied on-top of the multi-normal distribution



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 11. Synergistic system with 10 variables, organized into subsets of sizes $\{3, 3, 4\}$ and increasing interaction strength. A **CDF** transformation is applied on-top of the multi-normal distribution



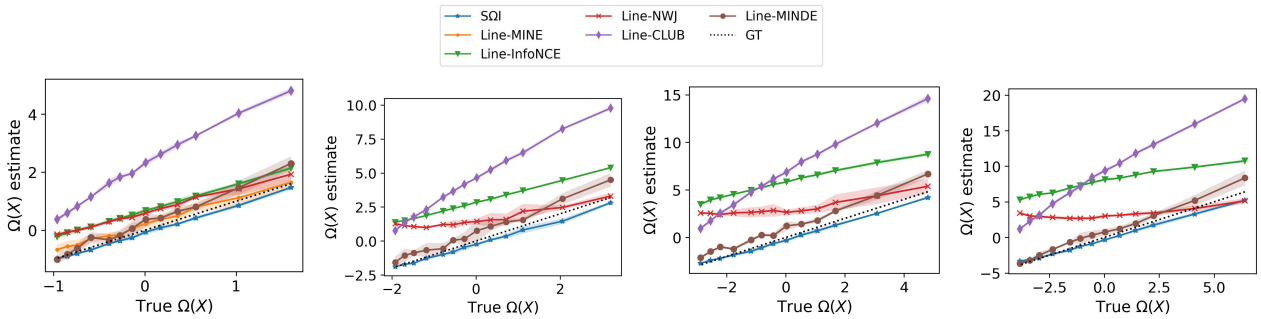
(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 12. Mixed-interaction system with 10 variables, organized into 2 redundancy-dominant subsets of size $\{3, 4\}$ variables and one synergy-dominant subset with 3 variables. O-INFORMATION is modulated by fixing the synergy inter-dependency and increasing the redundancy. A **CDF** transformation is applied on-top of the multivariate-normal distribution.

F. Additional results

F.1. Additional baseline

(Franzese et al., 2024) have shown that the KL divergence between two distributions can be computed using the denoising score function enabling the proposition of an MI estimator. In Figure 13, we present results on the mixed benchmark (redundancy and synergy) extended with the new baseline called Line-MINDE, that computes O-information using the MI estimator from (Franzese et al., 2024). Note that this approach requires learning a set of independent score models, one for each MI term: this increases the total number of parameters to learn, resulting in a more computationally heavy training process compared to our proposed method. In these new experiments, we follow the authors hyper-parameters and score network architecture. We observe that while Line-MINDE outperforms other pairwise MI based estimators, SΩI stands out with the best performance. Our findings indicate that the superiority of SΩI is due to efficiency of score based models in estimating information theoretic measures, which explains the superiority of SΩI and Line-MINDE against other neural estimators. Secondly, the direct estimation of TC and DTC and the amortized training using a unique network is more efficient which explains why SΩI outperforms Line-MINDE.



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 13. **Additional Line-MINDE (Franzese et al., 2024) baseline**. Mixed-interaction system with 10 variables, organized into a redundancy-dominant subsets of size 3, 4 variables and one synergy-dominant subset with 3 variables. O-INFORMATION is modulated by fixing the synergy inter-dependency and increasing the redundancy.

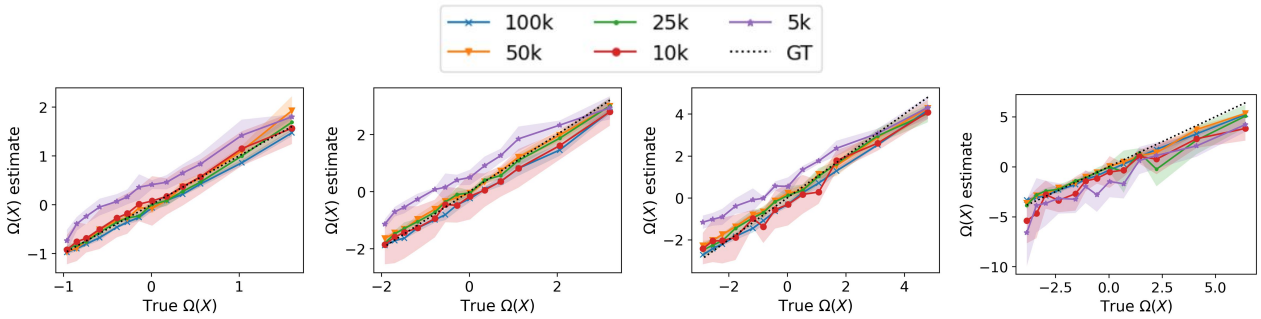
F.2. Ablation study

F.2.1. Data size

In Figure 14, we present a training size ablation study on the mixed benchmark. The considered number of training samples are of 5k,10k,25k,50k,100k samples. We fix the testset to 10k samples, except when the training size is 5k, for which we use 5k test samples. We observe that for data size superior to 10k, SΩI obtains very good estimates in terms of bias and variance; when the training size has 10k samples, SΩI estimates have increased variance; when we use only 5k training samples, SΩI have increased bias. These results are to be expected, since neural estimators, in general, require sufficient training data to shine.

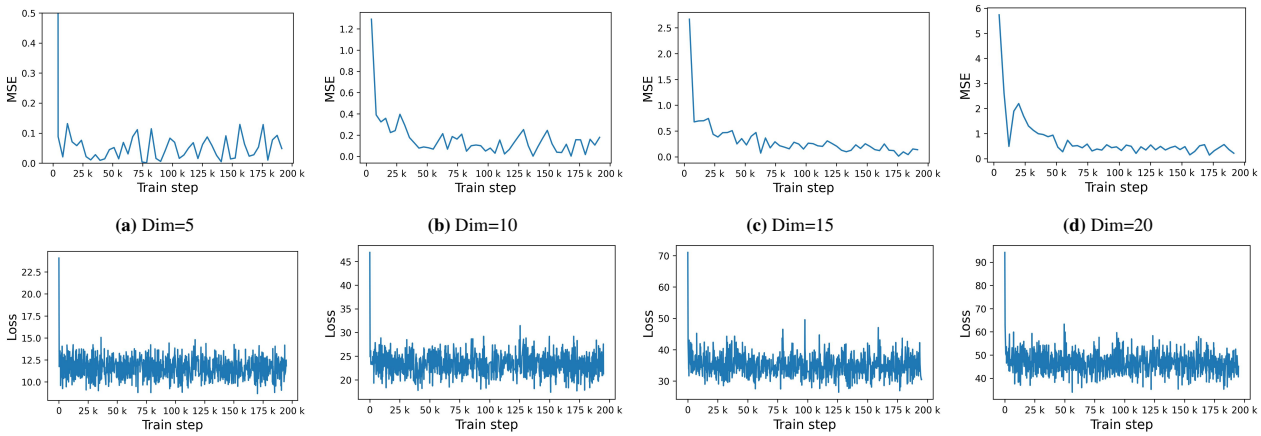
F.2.2. Number of training iterations

In Figure 15, we present the training curves contrasted with MI estimate mean squared error. Clearly, the number of iterations required to achieve satisfactory results depends on the dataset complexity.



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 14. SΩI training size ablation study : 100k,50k,25k,10k,5k. We use a test size of 10k for all the settings except when the train set size is equal to 5k where we use test size of similar size. The considered benchmark is a mixed-interaction system with 10 variables, organized into a redundancy-dominant subsets of size 3, 4 variables and one synergy-dominant subset with 3 variables. O-INFORMATION is modulated by fixing the synergy inter-dependency and increasing the redundancy.



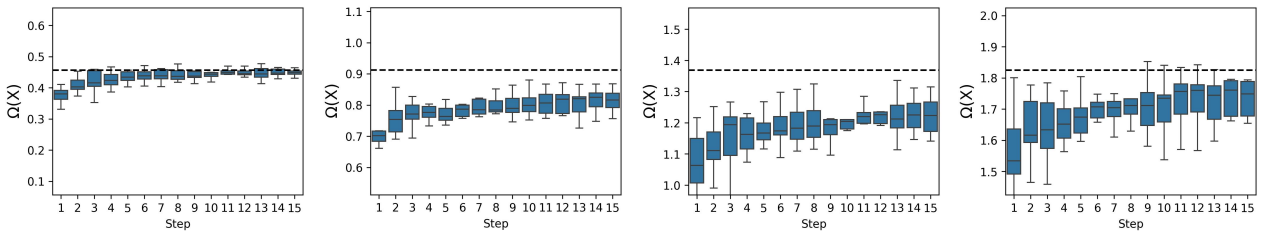
(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

(e) Dim=5 **(f)** Dim=10 **(g)** Dim=15 **(h)** Dim=20

Figure 15. **Training Loss curve Vs Estimation of O-INFORMATION MSE**. Mixed-interaction system with 10 variables, organized into a redundancy-dominant subsets of size 3, 4 variables and one synergy-dominant subset with 3 variables. For different benchmark dimensions, we report: **Top:** O-INFORMATION estimation mean square error as a function of the training iterations. **Bottom:** Training loss curve.

F.2.3. MONTE CARLO INTEGRATION STEPS

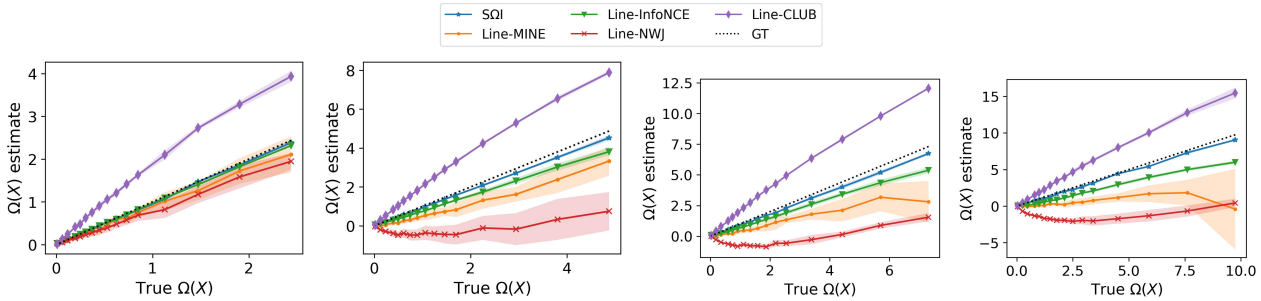
In Figure 16, we present an ablation on the number of Monte Carlo steps, for the case of a mixed (redundancy and synergy) benchmark with $N = 10$ random variables. We notice that an increased number of steps improves the estimation variance and bias. Naturally, this depends on the data dimension and complexity.



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

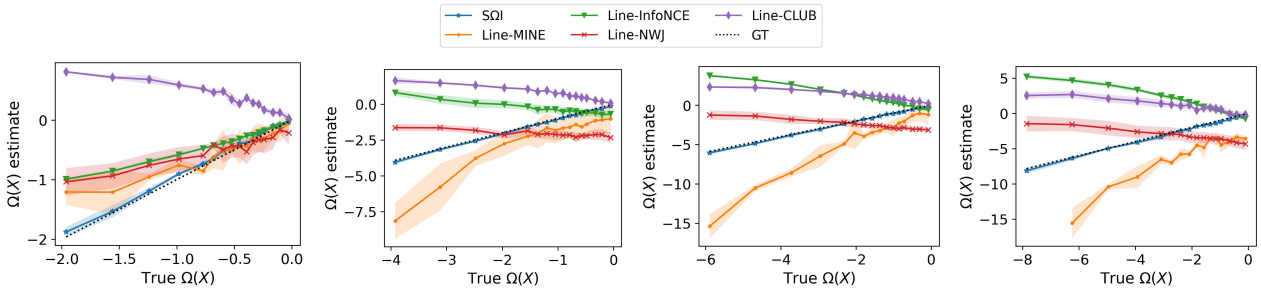
Figure 16. **Estimation of O-INFORMATION as a function of Monte Carlo Averaging steps run over 10 seeds**. Mixed-interaction system with 10 variables, organized into a redundancy-dominant subsets of size $3, 4$ variables and one synergy-dominant subset with 3 variables. Dashed line represents ground truth O-INFORMATION.

F.3. Additional synthetic experiments



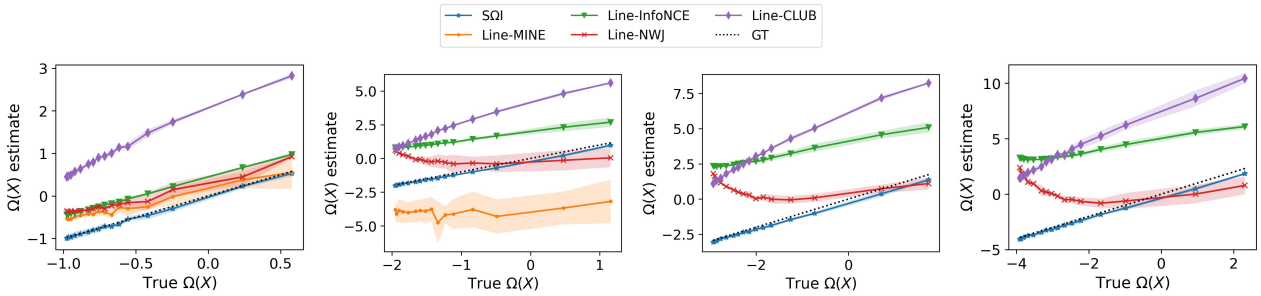
(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 17. Redundant system with 6 variables, organized into subsets of sizes $\{3, 3\}$ and increasing interaction strength.



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

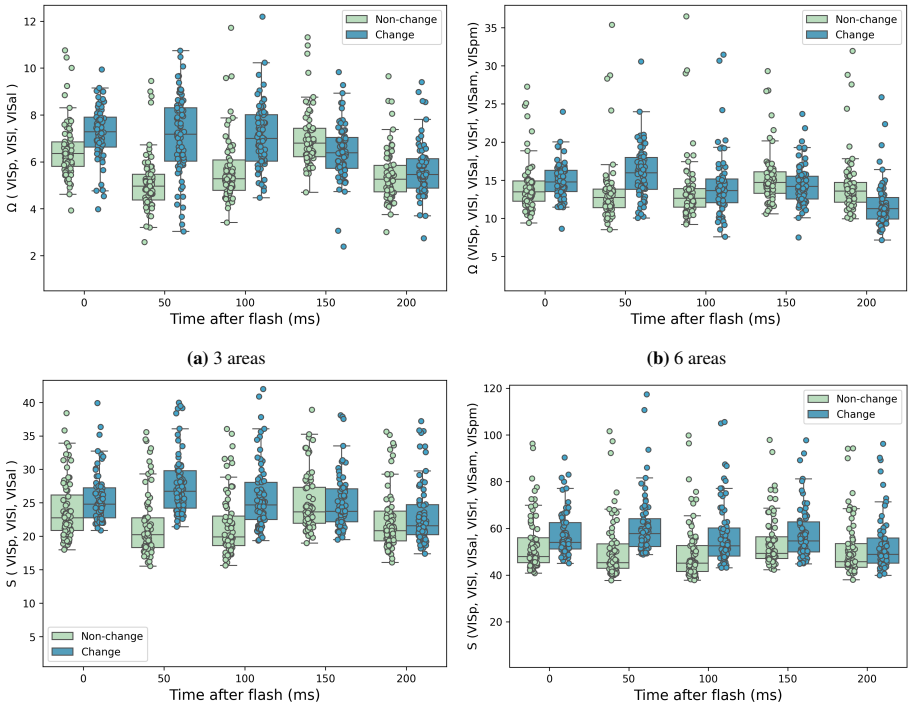
Figure 18. Synergistic system with 6 variables, organized into subsets of sizes $\{3, 3\}$ and increasing interaction strength.



(a) Dim=5 **(b)** Dim=10 **(c)** Dim=15 **(d)** Dim=20

Figure 19. Mixed-interaction system with 6 variables, organized into a redundancy-dominant subsets of size 3 variables and one synergy-dominant subset with 3 variables. O-INFORMATION is modulated by fixing the synergy inter-dependency and increasing the redundancy.

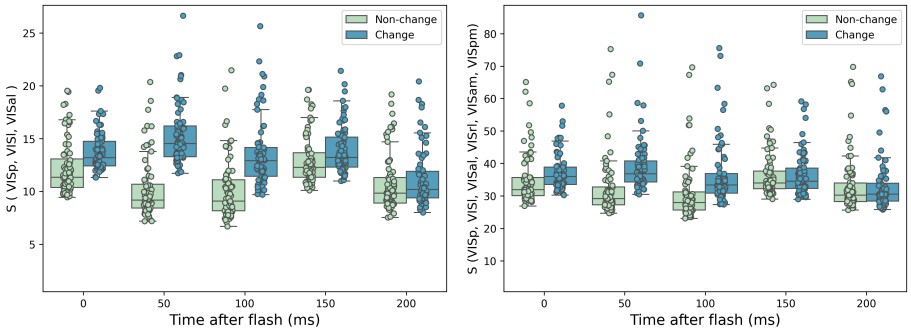
F.4. The neural application additional experiments



(a) 3 areas **(b)** 6 areas

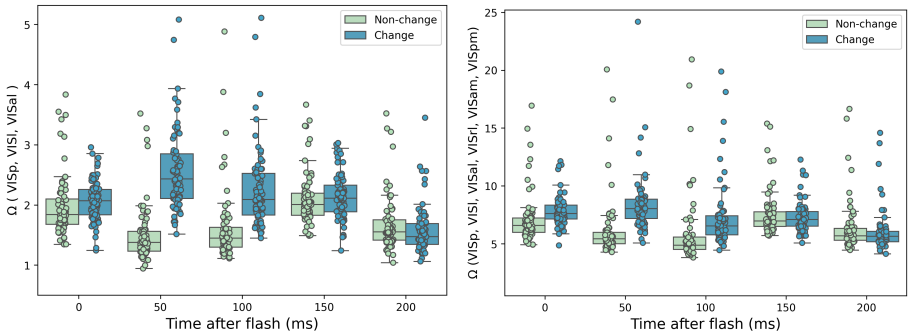
(c) 3 areas **(d)** 6 areas

Figure 20. O-INFORMATION and S-INFORMATION estimate in the visual cortex region activity after two types of stimulus flash across 72 trial sessions. **Left**: Analysis using three brain region areas, **Right**: Extended analysis using six brain region areas. The step size is set to $1ms$ which results in **50** dimensional data for each bin per area.

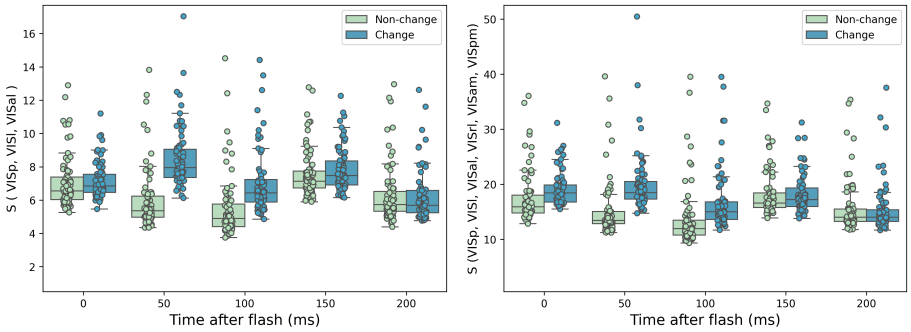


(a) 3 areas **(b)** 6 areas

Figure 21. S-INFORMATION estimate in the visual cortex region activity after two types of stimulus flash across 72 trial sessions. **Left**: Analysis using three brain region areas, **Right**: Extended analysis using six brain region areas. The step size is set to $2ms$ which results in **25** dimensional data for each bin per area.



(a) 3 areas **(b)** 6 areas



(c) 3 areas **(d)** 6 areas

Figure 22. O-INFORMATION and S-INFORMATION estimate in the visual cortex region activity after two types of stimulus flash across 72 trial sessions. **Left**: Analysis using three brain region areas, **Right**: Extended analysis using six brain region areas. The step size is set to $5ms$ which results in **10** dimensional data for each bin per area.