THÈSE DE DOCTORAT
DE L'UNIVERSITÉ SORBONNE UNIVERSITÉ

Spécialité :

École doctorale n°130: EDITE de Paris

réalisée

au Laboratoire Physico-Chimie Curie

sous la direction de Hervé Isambert

présentée par

Vincent CABELI

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ SORBONNE UNIVERSITÉ

Sujet de la thèse :

Learning causal graphs from continuous or mixed datasets of biological or clinical interest

Date de soutenance le 15 Décembre 2021

devant le jury composé de :

Pr	**Philippe Leray**	**Rapporteur**
Dr	**Simon de Givry**	**Rapporteur**
Dre	**Chloé-Agathe Azencott**	**Examinatrice**
Dr	**Pierre-Henri Wuillemin**	**Examinateur**
Dre	**Simona Cocco**	**Examinatrice**
Dr	**Hervé Isambert**	**Directeur de thèse**

Contents

Résumé

La corrélation n'implique pas la causalité, une distinction importante à rappeler alors que les associations statistiques génèrent de plus en plus de discussions dans un monde toujours plus mesuré et documenté. C'est pourtant le but, avoué ou non, de la plupart des domaines scientifiques : définir les mécanismes de notre environnement qui ont produit ces observations. La nouvelle science de la causalité cherche à nous réconcilier avec ce concept en répondant à ces questions : comment formaliser les relations causales, comment nous les représenter, et quand peut-on les découvrir ? Les travaux de cette thèse s'inscrivent dans la théorie principalement développée par Judea Pearl sur les diagrammes causaux; des modèles graphiques qui permettent de dériver toutes les quantités causales d'intérêt (effet du traitement, contrefactuelles...) formellement et intuitivement. Nous traitons le problème de l'inférence de réseau causal à partir uniquement de données d'observation c'est-à-dire sans aucune intervention de la part de l'expérimentateur. En particulier, nous proposons d'améliorer les méthodes existantes pour les rendre plus aptes à analyser des données issues du monde réel, en nous affranchissant le plus possible des contraintes sur les distributions des données, et en les rendant plus interprétables.

Nous proposons une extension de MIIC, une approche basée sur les contraintes et la théorie de l'information pour retrouver la classe d'équivalence du graphe causal à partir d'observations. Notre contribution est un algorithme de discrétisation optimale pour simultanément estimer la valeur de l'information mutuelle (et multivariée) et évaluer sa significativité entre des échantillons de variables de n'importe quelle nature : continue, catégorique ou mixte. Cette discrétisation optimale est elle-même ancrée dans le principe de *longueur de description minimale* pour trouver les meilleures représentations des distributions jointes grâce à une estimation du maximum de vraisemblance normalisé. L'évaluation de la significativité de l'information au sens de la complexité stochastique est un dérivé de cette approche, et nous permet de reconstruire des graphes causaux de manière robuste sur des échantillons de taille finie. Nous proposons également des améliorations des algorithmes par contraintes pour s'assurer que le graphe final est plus cohérent avec les données, en modifiant les règles pour choisir les variables de conditionnements. Les outils d'inférence et de visualisation sont mis à disposition de la communauté pour permettre au plus grand nombre d'analyser leurs jeux de données.

Enfin, nous mettons à profit ces développements pour analyser des jeux de données mixtes, toujours en étroite collaboration avec les équipes responsables de la collecte des données. La première application majeure est l'analyse de dossiers médicaux de patients âgés atteints de troubles cognitifs en collaboration avec l'hôpital La Pitié-Salpêtrière. La seconde concerne les dossiers médicaux de patientes ayant entrepris une chimiothérapie néo-adjuvante contre le cancer du sein, avec le département de chirurgie oncologique de l'hôpital Curie. Enfin, nous présentons des résultats sur l'analyse de gènes métaboliques moteurs de la différentiation hématopoïétique sur des données de profil transcriptomiques de cellules précurseurs.

Abstract

Correlation does not imply causation, an important distinction to remember as statistical associations generate more and more discussion in an increasingly measured and documented world. It is, however, the goal of most of science, whether or not it is acknowledged: to define the mechanisms in our environment that produced these observations. The new science of causality seeks to reconcile us with this concept by answering these questions: how can we formalize causal relationships, how can we represent them, and when can we discover them? This thesis is grounded in the theory primarily developed by Judea Pearl on causal diagrams; graphical models that allow us to derive all causal quantities of interest (treatment effect, counterfactuals...) formally and intuitively. We address the problem of causal network inference from observational data only, *i.e.* without any intervention from the practitioner. In particular, we propose to improve existing methods to make them more suitable for analyzing real-world data, by dropping as much as possible any assumption about data distribution, and by making them more interpretable.

We propose an extension of MIIC, a constraint-based information-theoretic approach to recover the equivalence class of the causal graph from observations. Our contribution is an optimal discretization algorithm to simultaneously estimate the value of mutual (and multivariate) information and evaluate its significance between samples of variables of any nature: continuous, categorical or mixed. This optimal discretization is itself based on the principle of *minimum description length* to find the best representations of the joint distributions through a normalized maximum likelihood estimation. This discretization comes with an assessment of the significance of information in the sense of data complexity, which allows us to reconstruct causal graphs in a robust manner on finite sample sizes. We also propose improvements to constraint-based algorithms to ensure that the final graph is more consistent with the data, by modifying the rules for choosing the conditioning variables. Inference and visualization tools are also made available to the community.

Finally, we make use of these developments to analyze mixed datasets, always in close collaboration with the teams that were responsible for data collection. The first major application is the analysis of medical records of elderly patients with cognitive disorders in collaboration with La Pitié-Salpêtrière Hospital. The second concerns the medical records of patients undergoing neoadjuvant chemotherapy for breast cancer, in collaboration with the surgical oncology department of the Curie Hospital. Finally, we present results on the analysis of metabolic genes driving hematopoietic differentiation on transcriptomic profiles of precursor cells.

Chapter 1

Introduction

1.1 Scientific context

If correlation does not imply causation, then what does ? It is a good thing that this distinction has permeated modern scientific culture, but discovering the causal mechanisms remains the goal of most studies, and correlation their main tool to do so. The new science of causality is trying to reconcile us with this goal, formally defining how to represent causal relations, how to measure them, and most importantly, giving the necessary conditions to discover them.

The first question on how to represent causal relationships has perhaps found its best answer in the theory of causal diagrams mainly developed by Judea Pearl [1, 2, 3]. A causal diagram is a Bayesian network: a directed acyclic graph that encodes the conditional independences between random variables represented by the nodes; with an added causal dimension transcribed by the direction of the edges. From these graphs, one can derive answers to fundamentally causal questions like "what is the effect of this treatment on this population?", or even "what if this population had received this treatment?".

This thesis contributes to the field of causal graph discovery, which aims to reconstruct these graphical models from observational data only. The challenge of causal discovery lies in retaining the direct links that reflect some understanding of nature, the data generating process, and rejecting the spurious interactions that are indirect consequences of the meaningful relationships. In the right conditions, it is known that we can learn the causal graph up to an equivalence graph from only the pattern of dependencies and independencies found in the data, without any intervention.

In this work, we focus on constraint-based algorithms in general and MIIC specifically, an information-theoretic approach combining elements of both constraint-based and score-based methods. Where classical methods rely on frequentist tests of independence and a parameter α for the p-value threshold, MIIC estimates independence from data with the minimum description length principle and the normalized maximum likelihood distribution.

1.2 Contributions

The main objective of this thesis is to make MIIC and constraint-based methods more capable of handling real-world data. This class of algorithms relies entirely on conditional independence patterns on sampled data, which is notoriously hard to estimate without making assumption on the distributions.

We want to be able to use the data that is available to us under any form and making as little assumption as possible on its distribution. Concretely, we want to be able to estimate the conditional independence between two variables X, Y with a conditioning set Z regardless of the nature of the marginal distributions $(p(X), p(Y), p(Z))$ and of the joint distributions $(p(X,Y)...)$. This estimate must also be robust to small sample sizes while remaining computable when N is large, and ideally not favor any type of variable or interaction.

The method presented here is based on the *master* definition of mutual information:

$$\mathrm{I}(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} \mathrm{I}([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$$

where the supremum is on all finite partitions $\mathcal{P}$ and $\mathcal{Q}$ [4]. The developed approach consists in maximizing the value $\mathrm{I}'([X]_\Delta;[Y]_\Delta)$ corrected by the stochastic complexity associated with the discretization $[X]_\Delta;[Y]_\Delta$ to take into account the effects of the finite number of samples. Introducing the complexity also allows us to conclude on the independence on finite samples (for which the information estimate is always positive): $\mathrm{I}'([X]_\Delta;[Y]_\Delta) \leq 0$ implies independence between X and Y in the sense of the data complexity [5].

The other contributions of this thesis concern the operation of constraint-based methods. First, we propose an information theoretic test to perform test-wise omission in the case of missing data, avoiding as much as possible the spurious independencies brought by selection bias. In the same idea, we introduce a variant of constraint-based algorithms that guarantees that the conditioning sets used to remove edges are more consistent with the final graph $\mathcal{G}_{inf}$ and the data $\mathcal{D}$ [6].

We also propose a method to distinguish "genuine" from "putative" causal links returned by MIIC, by excluding the effect of an unobserved common cause for each predicted genuine causal link. It is achieved by evaluating the separate probabilities of the "head" and "tail" of directed links for all directed edges.

In addition to an open source library on R, we have also developed an online graphical interface to facilitate the exploration of MIIC results, available at `https://miic.curie.fr/`.

Finally, we show different applications of MIIC on real-life mixed datasets.

The first network is reconstructed from clinical data of the hospital La Pitié-Salpêtrière

of 1628 elderly patients with cognitive disorders. After processing the dataset, it contains 107 variables of different types (i.e. 19 continuous variables and 88 categorical variables) and of heterogeneous nature (i.e. variables related to medical history, comorbidities and comedications, results of cognitive tests, clinical, biological or radiological examinations, diagnoses and treatments).

The second application focuses on medical data from 1199 breast cancer patients who received neoadjuvant chemotherapy at the Curie Hospital over the last 20 years. The network approach allows for all variables to be considered together, distinguishing between indirect and direct relationships, and helps practitioners understand the mechanisms behind the creation of data, whether it is the way they are collected or the progression of the disease itself.

The third application of MIIC on mixed data concerns the discovery of driver genes that influence the differentiation of hematopoietic precursor cells, and the inference of the regulatory network of these genes.

1.3 Research articles

This thesis is a cumulative dissertation based on the research articles written during the PhD. After a general introduction on causal inference and information theory, the articles are either included verbatim or modified so as to give more context on related works and details on the implementations. Section 3.2 in particular gives a more detailed description of the methods introduced in [7], with the publication itself being more centered around benchmarks and an application on real data introduced in Section 5.1.

Article	Used in
Learning clinical networks from medical records based on information estimates in mixed-type data. V Cabeli, L Verny, N Sella, G Uguzzoni, M Verny, H Isambert. PLoS computational biology 2020	Sections 3.3, 5.1
Constraint-based Causal Structure Learning with Consistent Separating Sets. H Li, V Cabeli, N Sella, H Isambert. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)	Section 4.2.1
Reliable causal discovery based on mutual information supremum principle for finite datasets. H Li, V Cabeli, M Ribeiro Dantas, H Isambert. "Why-21" workshop at NeurIPS 2021	Section 4.3
A method to learn interpretable causal networks from very large datasets, application to 400,000 medical records of breast cancer patients. M Ribeiro Dantas, H Li, V Cabeli, H Isambert. In preparation	Section 4.1.2
Interactive data vizualisation and exploration tool of a global clinical network from a large dataset of breast cancer patients treated with neoadjuvant chemotherapy. N Sella*, A-S Hamy*, V Cabeli*, L Darrigues, B Grandal, M Laé, F Reyal, H Isambert. To be submitted	Section 5.2
Metabolic Heterogeneity in Hematopoietic Progenitors Fuels Innate Immunity. In preparation	Section 5.3

Chapter 2

Causal inference from observational data

2.1 Causal inference and causal structures

In this section we clarify the concept of using graphs to represent causality. We start by giving an intuitive example of a causal diagram, and we show that such graphs can describe quite naturally common ways of inferring causality, either through experiments or from observation data. Finally, we give the formal notations and definitions of the causal graph framework.

2.1.1 Intuition of causal graphs

Let us consider a familiar situation in which our intuition can be represented by a causal diagram (Fig 2.1). Assume that there are two causes that could be at the origin of a car breakdown, which we try to diagnose *before* intervening on the car. We consider the factors that could help us make the diagnosis, and decide to draw a diagram to see their relations visually. The two causes considered, low engine oil level or a flat battery, are represented as parents of the "Breakdown" node. There is no reason to think that the two are linked, which we represent by not drawing an edge between them. We include a fourth node that corresponds to another observation : the headlights do not turn on. We know that the headlights do not depend on the oil level but they need battery to run, and the "Lights" node is therefore linked to the "Battery" node only. We also know that usually, it is somehow related to a car breaking down : if the headlights do not turn on, the car will probably not start either. We represent this association with a dotted link. This link reflects a *correlation, not a causation* : the indirect interaction exists only because of the common ancestor "Battery" but does not inform us about a functional relationship (therefore it would not be included in the causal diagram). It can help us guess the origin of the failure (Battery or Oil), but fixing the headlights will not help to start the car.

Representing complex systems with a causal graph has two main advantages. First,

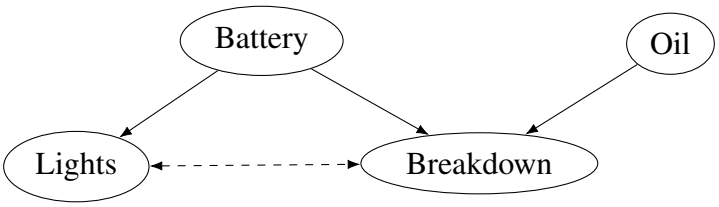

Figure 2.1: Car breakdown diagnosis with a causal diagram.

only the direct links, that correspond to functional relationships, are represented. Spurious correlations between two variables that are not causally linked can always be explained from a path of direct links in the graph. Secondly, the causal order can be *read off* the graph, via the direction of the edge $X \rightarrow Y$. If X is a parent of Y, then intervening on the distribution of X to give it an arbitrary value x will affect the distribution of Y, but the inverse is not true : intervening on Y will not affect its parent X. Such intervention is noted $p(Y|do(X = x)$ and is the basis of *do calculus*. Remark that $p(Y|X)$ and $p(Y|do(X = x))$ are not the same : the first is observational while the second is interventional. For example, let X be the reading on a barometer and Y the weather represented in a single variable. If we observe both every day and take note of their values, $p(Y|X)$ would show a strong relationship between the two : if the barometer measures low air pressure the weather is often bad, while if it measures high pressure the weather is better. The distribution $p(Y|do(X = x))$ however won't actually depend on the value x : we cannot change the weather by setting the reading on a barometer. Even though $p(Y|X)$ can be used for predicting the value of a variable by measuring the other one, it does not inform on the functional relationship between the two. For this, we need to go up a rung on the *ladder of causation*, by using do-calculus [3].

In the example of Fig 2.1, we draw the graph from pre-existing knowledge, but what can we do when such knowledge is not available ? This is the domain of *causal inference*, which aims to uncover causal effects either through experimentation or passive observation of the system.

2.1.2 Causal inference approaches as seen with causal graphs

The gold standard of causal inference is the randomized controlled trial, where an homogeneous population is randomly attributed either a treatment or a placebo. Let Y be the outcome of the trial for each patient, which can be positive or negative. We want to know the extent to which the outcome depends on the treatment, noted X, as opposed to other external factors which are all grouped in the node Z. Formally, we can answer this question by comparing $p(Y|do(X = treatment))$ and $p(Y|do(X = placebo))$. The causal graph of a truly random trial is shown in Figure 2.2.

We can see from the causal diagram that the causal effect of X on Y is direct, it is not affected by the rest of the graph. In this setting, random attribution of X is a kind of

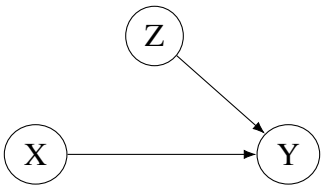

Figure 2.2: Randomized control trial where X is the treatment, Y the outcome and Z external factors.

intervention, and $p(Y|do(X))$ can be directly observed from the data as $p(Y|X)$. This type of experiment is generally reliable provided the treatment is assigned truly at random, but is has disadvantages. First, it needs to be conducted for each X for which we want to know the effect, and it may be too long or too difficult to enroll enough participants. Secondly, it is unethical when we suspect that the interaction is harmful, think for example of forcing test subjects to be exposed to carcinogens. Finally, it may be simply impossible to intervene on the potential cause, for example we can not randomize the genetic makeup of patients to study the prevalence of certain diseases.

For those cases, we can still perform causal inference by simply observing the potential cause, the outcome, and all of the confounding factors which affect both the cause and the effect (Fig 2.3). This graph summarizes the principle behind several approaches, even those that have not adopted the language of causal graphs and do-calculus. Matching procedures for example make exactly the same assumptions to estimate the effect of X on Y by taking the effect of Z into account. Their goal is to reduce the assignment bias for the "treatment" X and mimic a randomized controlled trial by creating samples that were matched on Z, essentially removing the edge $Z \rightarrow X$ [8, 9]. Fig 2.3 is also the typical setting where one can simply model Y from X while adjusting for Z. This is the approach taken by genome-wide association studies, which try to measure the effect of thousands of genes X on the apparition of a disease Y adjusting for some principal components Z to model ancestry differences between cases and controls [10].

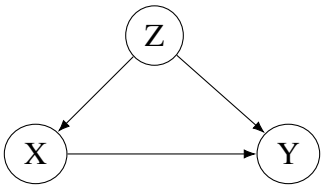

Figure 2.3: Observational studies require to adjust for the confounding variables Z.

All of these methods are theoretically sound but are often criticized for their predisposition to give biased results. According to Pearl, these shortcomings come mainly from the attitude that Z should contain as many covariates as possible, to adjust with all the information that is available [1, p. 350]. By doing so and ignoring the "strong ignorability conditions" for a variable to be included in Z, we will inevitably end up including variable that are not parents

but children of X and Y : $X \rightarrow Z \leftarrow Y$, violating the assumptions and the graph of Fig 2.3. Still according to Pearl, this kind of mistake is much less likely to occur when using the causal graph framework, as practitioners are forced to model the interactions first, thinking about the causal relationships between the treatment, the outcome, and the covariates.

As the last example, we will look at the case where we still cannot intervene on the potential cause X, and we know it is affected by common confounders of Y, but we cannot measure them, nor adjust for them. This was famously the defense of prominent statisticians employed by tobacco companies at the time of the first reports linking cigarettes with lung cancer. The association could not be denied, but they claimed it could be explained by a hidden common cause, some genetic factor for example, which caused a certain population to both want to smoke and develop more cancers than the general population. In 1964 we did not have access to sequencing technology, and since randomized controlled trials were out of the question, this argument was hard to disprove and supposedly delayed anti-smoking legislation [1, p. 83]. In this setting, we can still measure the effect of X on Y if we measure another "instrumental" variable I that we know to have an effect on X and to be independent of the latent confounders L (Fig 2.4).

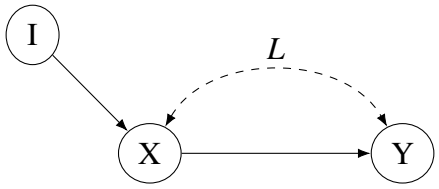

Figure 2.4: When the confounding variables are not observed, the effect for X on Y can be estimated from an instrumental variable I.

Indeed, we can see from the graph that any association measured between I and Y necessarily goes through X, proving the existence of the directed edge $X \rightarrow Y$. In the cigarettes cause lung cancer example, we can take I as the price of cigarette packs. Intuitively, if I is correlated with the number of lung cancers Y we can deduce the existence of a causal link between tobacco consummation and cancer [11]. Just as with the adjusting variables of the previous example, modelling the interactions in a causal graph is a way of making sure that I corresponds to an instrumental variable.

This brief discussion highlights the remarkable adaptability of causal graphs. These intuitive models are able to summarize most if not all approaches that aim to infer causality, federating all causal thinking into a single framework. Once the causal graph is known, one can then derive formal quantities of causal effect and counterfactual thinking using do-calculus.

2.1.3 Notations and definitions

The causal graph framework treats causality as a statistical property, it utilizes the languages of both graph theory and probability. We now take the time to review the definitions that are necessary to define a causal graph.

- Let $\mathcal{D}$ be a dataset comprised of V variables $X_1, ..., X_v$. For pairwise and conditional cases, we may use X, Y and Z as variables of $\mathcal{D}$ instead.
- Each variable has a distribution $p(X_i)$, and the joint distribution of $\mathcal{D}$ is $p(V)$.
- We note two variables that are independent as $X \perp\!\!\!\perp Y$, conditionally independent on Z as $X \perp\!\!\!\perp Y \mid Z$.
- $\mathcal{D}$ is represented by a graph $\mathcal{G}$ where each variable is a node. We note the true causal graph $\mathcal{G}_c$ and the graph inferred from the data $\mathcal{G}_{Inf}$.
- If variables X and Y of $\mathcal{D}$ are adjacent in $\mathcal{G}$, the edge between the two is either unoriented $X - Y$, oriented $X \rightarrow Y$, $X \leftarrow Y$ or bi-directed $X \leftrightarrow Y$.
- The variables that have an edge pointing towards X_i are its parents and are noted Pa_i. The variables that X_i points to are its children, noted Ch_i.
- The skeleton of $\mathcal{G}$ is the graph with same adjacencies and no oriented edges.
- A V-structure is a sub-graph of three nodes where $X \rightarrow Z \leftarrow Y$ where $X \not\sim Y$.
- The complete graph on V variables is the skeleton where all X_i, X_j are adjacent.

The true causal graph is, in short, the graph that describes the causal mechanisms that produce the data $\mathcal{D}$, as well as all possible randomized studies on the V variables.

Definition 2.1. The true causal graph $\mathcal{G}_c$ of given variables $X_1, ..., X_v$ with distribution $p(V)$ satisfies :

- $\mathcal{G}_c$ is a directed acyclic graph.
- $p(V)$ is *Markov with respect to* $\mathcal{G}_c$, *i.e.* if X and Y are *d-separated* by Z, then $X \perp\!\!\!\perp Y|Z$ in $p(V)$ (see Def 2.3).
- $p(V)$ satisfies causal minimality with respect to $\mathcal{G}_c$ (see Def 2.2).
- The distribution of the node X_i is a function of its parents Pa_i and some unique noise ε_i:

$$p(X_i) = f(p(Pa_i), \varepsilon_i)$$

and $\mathcal{G}_c$ is *compatible* [1] with the set of P_* of all possible interventional distributions $P(V|do(X = x))$ for all variable X and value x.

With the definition of causal minimality :

Definition 2.2. A distribution satisfies causal minimality with respect to $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$, but not to any sub-graph of $\mathcal{G}$.

Note that since $\mathcal{G}$ satisfies causal minimality, it is also unique.

The *d-separation* is the way to cut the flow of causality from one node to another :

Definition 2.3. Two nodes X and Y of a DAG $\mathcal{G}$ are d-separated by a set of nodes Z (which can be empty) if and only if :

- The path between X and Y contains a chain $i \to m \to j$ or $i \leftarrow m \to j$ with $m \in Z$, or
- The path between X and Y contains a *V-structure* $i \to m \leftarrow j$ such that $m \notin Z$ and no descendent of m is in Z.

The set Z d-separates X and Y if and only if Z blocks every path between X and Y in this way.

If X and Y are d-separated by Z in $\mathcal{G}$, then $X \perp\!\!\!\perp Y \mid Z$ in any distribution compatible with $\mathcal{G}$.

Observe that d-separations concern just one part of the true causal graph $\mathcal{G}_c$, they describe conditional independences but not the interventional distributions. Crucially, different DAGs may share the same d-separations. The class of graphs that are observationally equivalent is called the equivalence class of $\mathcal{G}$ [12] :

Definition 2.4. Two DAGs are in the same equivalence class if and only if they have the same skeletons and the same sets of V-structures.

As we will see in the next section, this is the limit that we can infer from dependencies and independences in $p(V)$ alone, without additional assumptions.

2.2 Causal structure learning

This thesis contributes to the field of causal structure learning from observational data, which aims to recover $\mathcal{G}_c$ from $\mathcal{D}$. Most of the methods that have been developed are divided in two groups : Bayesian scoring methods that assume $p(V)$ was generated from a Bayesian network and find the best fitting graph with likelihood scores, and constraint based approaches that try to reconstruct the graph from the data with iterative statistical tests.

2.2.1 Score-based approaches

The principle behind score-based approaches is fairly intuitive, but difficult to implement. Although there may be preliminary work pre-dating it, in this section we refer here the formal idea introduced by Geiger and Heckerman [13, 14], and Chickering [15]. Given the data $\mathcal{D}$ from a vector of V variables, find the graph $\hat{\mathcal{G}}$ that maximizes a likelihood score $S(\mathcal{D},\mathcal{G})$:

$$\hat{\mathcal{G}} = \text{argmax}_{\mathcal{G}} S(\mathcal{D},\mathcal{G}) \tag{2.2.1}$$

where $\mathcal{G}$ is searched over the space of DAGs.

We can think of several definitions for the scoring function. If the distribution $p(V)$ can be described with a parametric model (*e.g.* discrete multinomial distributions, linear Gaussian relationships), then we can define a set of parameters $\theta \in \Theta$. The Bayesian definition of $S(\mathcal{D},\mathcal{G})$ is the log posterior with prior beliefs $p_{pr}(\mathcal{G})$ and $p_{pr}(\theta)$ over DAGs and parameters respectively:

$$S(\mathcal{D},\mathcal{G}) = \log p_{pr}\mathcal{G} + \log p(\mathcal{D}|\mathcal{G}) \tag{2.2.2}$$

with $p(\mathcal{D}|\mathcal{G})$ the marginal likelihood

$$p(\mathcal{D}|\mathcal{G}) = \int_{\theta\in\Theta} p(\mathcal{D}|\mathcal{G},\theta)p_{pr}(\theta)$$

In this view, $\hat{\mathcal{G}}$ which maximizes the score is the maximum a posteriori estimator. In [16], Heckerman and Geiger discuss how to choose the priors accordingly.

Another way to define the scoring function is using the maximum likelihood estimator $\hat{\theta}$ from N observed samples, for each graph. We can then define the score function using the Bayesian Information Criterion (BIC) [17] :

$$S(\mathcal{D},\mathcal{G}) = \log p(\mathcal{D}|\hat{\theta},\mathcal{G}) - \frac{d}{2}\log N \tag{2.2.3}$$

which prevents overfitting by favoring models with fewer parameters d.

The space of all DAGs grows super-exponentially with V [15], so heuristics are needed to search it in practice. Greedy algorithms iterate over the set neighboring graphs, selecting the best candidate at each step and using it as a new reference point. Neighbors are usually defined as all DAGs that differ with at most one missing or extra edge from the reference graph. The greedy equivalence search (GES) [15] improves on this process by performing two phases : first adding edges up to a local maximum, then simplifying the model by removing edges, returning the pruned graph when a maximum is reached.

From our perspective, score-based methods suffer from the following drawbacks : (1) The score function requires simple modeling and parameters, which may destroy subtle

causality signals in real data. (2) The search space is limited to DAGs or their equivalence class, excluding bi-directed edges $X \leftrightarrow Y$. (3) The methods do not scale well with V and typically do not produce good results when $V > 50$.

2.2.2 Constraint-based approaches

Compared to the score-based methods, constraint-based algorithms have more of a local approach to graph reconstruction. They make two assumptions : d-separation in $\mathcal{G}_c$ implies conditional independence in $p(V)$ (the Markov condition), and all conditional independences in $p(V)$ correspond to d-separation in $\mathcal{G}_c$ (the faithfulness assumption). They will be discussed in more details in Sec 3.1.3, for now we simply assume that they hold for $p(V)$ and $\mathcal{G}_c$. Given both assumptions, constraint-based approaches are able to recover up to the equivalence class of $\mathcal{G}_c$ from the set of dependencies and conditional dependencies of $p(V)$.

We now describe the staple constraint-based approach, the PC algorithm named after Peter Spirtes and Clark Glymour [18], which is itself a refinement of the IC algorigthm [12]. It consists of three phases, as shown in Alg 1.

Algorithm 1 The PC Algorithm

Require: $\mathcal{D}$
Step 1: Find the graph skeleton and separating sets of removed edges
Step 2: Orient V-structures based on separating sets
Step 3: Propagate orientations of V-structures to as many remaining undirected edges as possible
return Output graph $\mathcal{G}_{Inf}$

It was proven to be consistent, returning the correct equivalence class of $\mathcal{G}_c$ if enough samples are observed. The skeleton reconstruction phase is an iterative process : starting from the complete graph, remove all edges $X - Y$ if $X \perp\!\!\!\perp Y$ or if $X \perp\!\!\!\perp Y|Z$ with Z a set of variables in the neighbors of X, Y. Colombo and Maathuis improved the original algorithm by making it order-independent, calling this version "PC-stable" [19]. This is the version detailed in Alg 2.

In the second step, we start orienting the edges of the resulting skeleton, V-structure by V-structure. For each triplet $X - Z - Y$ where $X \not- Z$, orient the edges $X \rightarrow Z \leftarrow Y$ if Z was not in the separating set to remove the edge $X - Z$. This orientation step also has two other variants, the conservative rule [20] and the majority rule [19]. Using the conservative rule, the V-structure is oriented only if Z is in none of the separating sets that satisfy $X \perp\!\!\!\perp Y|\{U_i\}$ (and accordingly, with majority rule orient if Z is less than 50 percent of those). Both give generally better results than the original scheme, although they require many more conditional independence tests.

Finally, these orientations are propagated to the rest of the graph following Meek's

Algorithm 2 Find skeleton and separating sets (Step 1 of PC-stable algorithm)

Require: Conditional independence test between all V variables

$\mathcal{G}_{Inf} \leftarrow$ the complete graph on V
$\ell \leftarrow -1$
repeat
  $\ell \leftarrow \ell + 1$
  for all vertices $X_i \in \mathcal{G}$ **do**
    $a(X_i) = \text{adj}(\mathcal{G}, X_i)$
  end for
  repeat
    select a new pair of vertices (X_i, X_j) adjacent in $\mathcal{G}$ and satisfying $|a(X_i) \setminus \{X_j\}| \geq \ell$
    repeat
      choose new $\boldsymbol{C} \subseteq a(X_i) \setminus \{X_j\}, |\boldsymbol{C}| = \ell$
      if $(X_i \perp\!\!\!\perp X_j | \boldsymbol{C})$ **then**
        Delete edge $X_i - X_j$ from $\mathcal{G}$
        $\text{Sepset}(X_i, X_j \mid \mathcal{G}) = \text{Sepset}(X_j, X_i \mid \mathcal{G}) \leftarrow \boldsymbol{C}$
      end if
    until X_i and X_j are no longer adjacent in $\mathcal{G}$ **or** all $\boldsymbol{C} \subseteq a(X_i) \setminus \{X_j\}$ with $|\boldsymbol{C}| = \ell$ have been considered
  until all pairs of adjacent vertices (X_i, X_j) in $\mathcal{G}$ with $|a(X_i) \setminus \{X_j\}| \geq \ell$ have been considered
until all pairs of adjacent vertices (X_i, X_j) in $\mathcal{G}$ satisfy $|a(X_i) \setminus \{X_j\}| \leq \ell$
return $\mathcal{G}$, sepsets

rules [21]. This step does not rely on independencies observed in the data, it is more of a convention to make the result $\mathcal{G}_{Inf}$ into a graph that is maximally oriented within the equivalence class [1]. As such, propagated orientations may be considered weaker causality signals than V-structures.

As opposed to score-based methods, constraint-based methods present many advantages. They can be used to infer larger networks provided the conditional independence test has a good time complexity. They also do not rely on a modeling of $p(V)$, only a conditional independence test, and are thus applicable to a much wider range of data. Additionally, modifications like the Fast Causal Inference (FCI) Algorithm [22] (and RFCI [23]) can make it tolerate and even discover unknown confounding variables.

2.2.3 (Conditional) independence tests

In Section 2.2.2, we purposefully introduced constraint-based methods with an "Oracle" independence test, *i.e.* able to infer $X \perp\!\!\!\perp Y$ or $X \not\perp\!\!\!\perp Y$ from $\mathcal{D}$ with no error. In practice, they also need a parameter α which sets the threshold for significance of a given dependence estimator. Few measures are able to reliably detect dependence between two random variables without being restricted to some type of interaction (see Fig 2.5 for examples of dependencies).

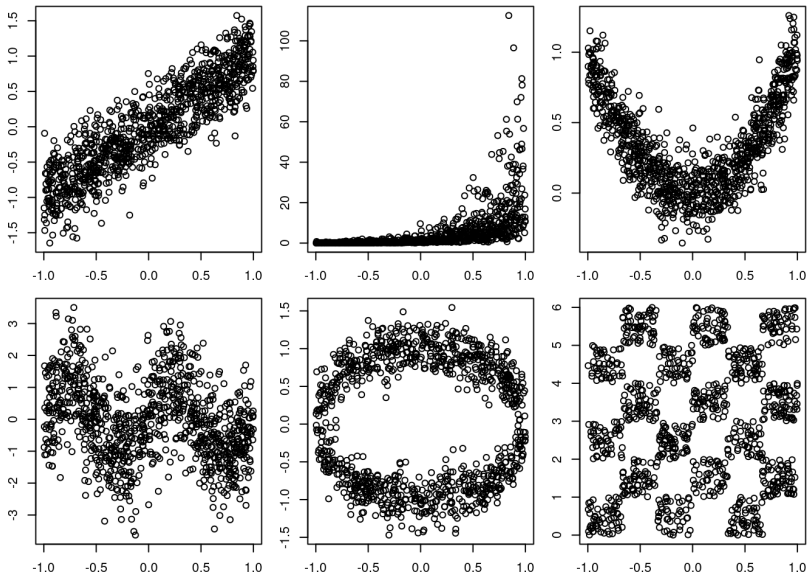

Figure 2.5: Various forms of dependencies, $X \not\perp\!\!\!\perp Y \Leftrightarrow p(x,y) \neq p(x)p(y)$

In the context of network inference, it is worth reviewing our options as the resulting graph will only contain the interactions in the class of models that are supported by the chosen dependence measure. I now give a brief overview of various dependence measures that can be used in constraint-based methods, going from the simplest to the most general.

Discrete independence tests

The problem of estimating (conditional) independence on discrete data has been studied for a long time. The χ^2 test was developed by Pearson to compare the observed joint frequencies of X and Y with those under the null hypothesis, where $X \perp\!\!\!\perp Y$:

$$\chi^2 = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - N \tag{2.2.4}$$

where x_i is the observed count, m_i the expected count and N the number of samples.

To assess the significance of χ^2, one can then compare it to the distribution under the null hypothesis, and accept the evidence for dependence if the cumulative density at the estimation (p-value) is lower than the specified $alpha$.

One can also compute the closely related G-statistic, of which the χ^2 is an approximation. Since it is related to information theoretic measures, the G-test will be discussed in Section 3.1.1.

Linear and nonlinear correlations

The most commonly used dependence measure is probably the linear correlation coefficient ρ, or the Pearson correlation coefficient (PCC). It is bounded in $[-1,1]$, it gives an estimation of both the strength and the direction of the covariance between two variables. As with covariance itself, it is only able to detect linear relationships and ignores many other types of dependencies. For the bivariate normal case, it implies $X \perp\!\!\!\perp Y$ if and only if $\rho_{X,Y} = 0$ (with sample size $N \to \infty$). In practice, $\rho_{X,Y}$ is never null and its significance is also assessed by comparing the estimated value to the distribution under the null hypothesis, for which we know the exact form.

It is generalized to non-linear relationships with Spearman's rank correlation coefficient, which measures the strength and direction of any monotonic function between X and Y. It is more general than the PCC, although still far from the strict equivalence that defines statistical independence : $X \perp\!\!\!\perp Y \Leftrightarrow p(x,y) = p(x)p(y)$.

Distance correlation

Distance correlation or distance covariance is a relatively new measure of dependence between two paired random vectors that is meant to be more universal that the product-moment covariance and correlation [24, 25].

The distance correlation R is obtained by normalizing the distance covariance, it behaves like the PCC but generalizes the idea of correlation in at least two fundamental ways. For all distributions with finite first moments, (1) $R(X,Y)$ is defined for X and Y of arbitrary dimensions (may not be equal) and (2) $R(X,Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$.

It is much more powerful than linear and non-linear correlations, for example it is typically able to detect all dependencies of Fig 2.5 except for the checkerboard pattern. In practice, it is however difficult to judge its significance as no closed form of the distribution under the null hypothesis is known. Instead, the null distribution is estimated by resampling the data many times and significance is assessed by comparing the empirical p-value to α.

Although they are unrelated, it has comparable power to Hoeffding's much older test of independence [26], based on the joint distribution's deviation from independence [27].

Kernel-based tests

Another popular approach to measure statistical dependence in the non-parametric case is to use kernel transformations of the data, and then measure the covariance in the kernel space [28, 29]. It was also adapted to independence testing in [30], giving the Hilbert-Schmidt independence criterion (HSIC). Significance is assessed by estimating the null distribution, either via Monte Carlo resampling or with a gamma distribution approximation.

[31] first proposed to adapt the HSIC to the conditional case, and infer causal structures with permutation-based significance testing via constraint-based algorithms. [32] also implemented kernel-based independence testing with the "Kernel PC" algorithm, although it also suffers from high complexity for independence testing. [33] improved on this idea and derived the null distribution for the conditional case, making the test less prone to type 1 and type 2 errors, and much more computationally efficient.

Kernel independence tests are very powerful to detect independencies in the non-parametric case, but they typically do not scale well with the number of samples.

Feature selection

Finally, we can mention feature selection as a whole, as it is strictly equivalent to finding the Markov blanket of the response variable (Fig 2.6). The statistical definition of the Markov blanket is the set of nodes B which can separate a variable Y from the rest of the dataset, $Y \perp\!\!\!\perp \mathcal{D}_{\setminus Y} | B$. In this sense, it is the optimal set that feature selection techniques try to recover. From the point of view of Bayesian networks, B is the set of all parents, children and spouses (other parents of its children) of Y.

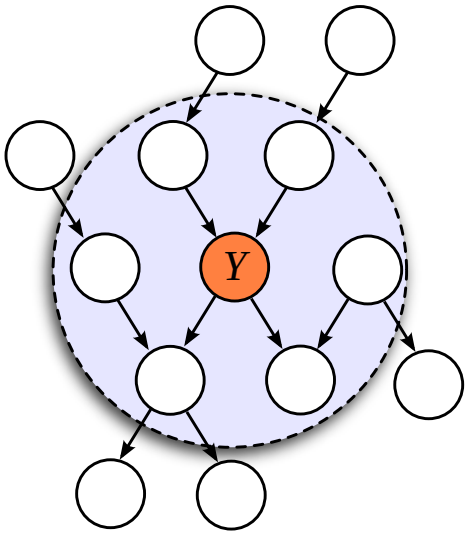

Figure 2.6: All nodes in the light blue shade are in the Markov blanket of the node Y.

Conceptually, this implies that any feature selection technique will in fact learn some local structure of the skeleton of $\mathcal{G}_c$, and we can learn the full skeleton by aggregating the results of feature selection for each node in the network. Such a method exists with GENIE 3 [34], which treats network reconstruction as several prediction problems, and combines random forests feature importance scores to recreate the graph.

However, one must pay particular attention to the process by which feature importance scores are computed. Typically, tree-based models subsample the feature space to avoid overfitting, but this implies that $\mathcal{G}$ may be different for each tree. As the simplest example, if

the true $\mathcal{G}_c$ is $X_1 \to X_2 \to Y$, the Markov blanket of Y consists of only $\{X_2\}$. By subsampling the feature space for each tree, we potentially remove X_2 and make the interaction $X_1 \to Y$ direct, which is not representative of $\mathcal{G}_c$.

2.2.4 Other graph reconstruction methods

The original PC algorithm is now more than thirty years old, and although it is consistent and theoretically sound, it has limitations (for example, it is not particularly robust to noise with finite sample size [5, 35]). In this section we give a very brief overview of other recent Bayesian structure learning and causal inference approaches that are neither score-based nor constraint-based.

A very different approach to causal graph reconstruction was introduced with the linear, non-Gaussian and acyclic model (LiNGAM) of Shimizu et al. [36]. Instead of examining only the dependencies between variables, in this framework the relationship between two variables X and Y is modeled with a structural equation :

$$Y = bX + \varepsilon \tag{2.2.5}$$

with b a factor, ε some noise such that $\varepsilon \perp\!\!\!\perp X$. The breakthrough of Shimizu et al. is the proof that $\mathcal{G}_c$ can be recovered from $\mathcal{D}$, in its entirety, whenever at most one ε is Gaussian. The intuition behind LiNGAM is that for non-Gaussian distributions, there is more information in the joint distribution than in the covariance matrix, which can be detected using independent component analysis. The key assumption here is the additive and independent noise model, which can be interpreted as the residual after predicting Y from its parents. DirectLiNGAM introduced another way to find the causal ordering by recursively performing regression and independence test between the predictor and residual [37].

In a similar idea, the Causal Additive Models (CAM) method aims to recover the underlying DAG by modeling the distribution $p(V)$ with additive structural equation models with Gaussian noise and non-parametric, non-linear relationships [38].

Another different strategy was introduced with NOTEARS, which formulates structure learning as a continuous optimization problem with a smooth function over real matrices [39], in opposition to the classical combinatorial score-based learning. The first version of this approach relied heavily on the linear parametrization in the weighted adjacency matrix, and was recently generalized to a larger class of models, that works without assuming any particular form of parametrization [40]. This general framework essentially makes the score-based method solvable using any of the existing generic solvers, such as multilayer perceptrons.

In a inspired paper, Wang and Blei proposed to take advantage of the multiplicity of

causes for a variable of interest by using unsupervised machine learning to estimate a latent confounding variable [41]. Having multiple causes is a blessing in this case : it gives more information to estimate a "substitute confounder" which in turn can prove or disprove causal relationships between causes and effect, with weaker assumptions than classical causal inference.

Finally, it was shown by [42] that causal structure can be learned by exploiting the invariance of functional relationships under a change of environments. First developed for linear models, [43] expands the framework of invariant causal prediction to the nonlinear and nonparametric case. This approach in particular will be discussed in Section 3.1.3 in relation with the "stability" or faithfulness assumption.

2.2.5 MIIC

MIIC (Multivariate Information-based Inductive Causation), combines constraint-based and information-theoretic frameworks to learn more robust causal graphical models. It was developed on the basis of the 3off2 algorithm by Affeldt and Isambert [5, 35], which takes advantages of multivariate information to reconstruct the skeleton and orient the edges. Corrected mutual information is described in more details in Section 3.1.1, for now we can simply treat it as a proxy for (conditional) independence between variables : $X \perp\!\!\!\perp Y \Leftrightarrow \mathrm{I}'(X;Y) < 0$ and $X \perp\!\!\!\perp Y|Z \Leftrightarrow \mathrm{I}'(X;Y|Z) < 0$.

Just like the PC algorithm it starts from a complete graph and prunes the edges to find the skeleton, then orients it (Alg 1) but there is a crucial difference in the way it chooses separating sets to remove edges. Where PC iterates over all the combinations of the neighbors of X and Y in order of increasing cardinality $\{U_i\}$ (until it can conclude conditional independence or it runs out of combinations), MIIC takes off the contributors one by one, using the chain rule of conditional information :

$$I(X;Y|\{U_i\},Z) = I(X;Y) - I(X;Y;U_1) - I(X;Y;U_2|U_1) - \cdots - I(X;Y;Z|\{U_i\}) \quad (2.2.6)$$

This allows to both speed up the process, removing the combinatorial search, and make it more robust to spurious independencies by removing the contributors in order of their information. The full algorithm is given in Alg 3.

Formally, the score $R(X,Y;Z|\{U_i\})$ is the minimum between the two conditions that Z indeed contributes to $I(X;Y|\{U_i\})$:

$$R(X,Y;Z|\{U_i\}) = \min\left(P_{\mathrm{nv}}(XYZ|\{U_i\}), P_b(XY|Z,\{u_i\})\right) \quad (2.2.7)$$

Algorithm 3 MIIC network reconstruction

Require: $\mathcal{D}$

- Skeleton reconstruction
$\mathcal{G} \leftarrow$ the complete graph on V
for all edges $X - Y \in \mathcal{G}$ **do**
  if $\mathrm{I}'(X;Y) < 0$ **then**
    Delete edge $X - Y$ from $\mathcal{G}$
    Sepset$\{X,Y\} \leftarrow \emptyset$
  else
    Find most contributing node $Z \in \{\mathrm{adj}(X) \cup \mathrm{adj}(Y)\}$ which maximizes $R(X,Y;Z|\emptyset)$
  end if
end for
while There is a link $X - Y$ with $R(X,Y;Z|\{U_i\}) > 1/2$ **do**
  for Top link $X - Y$ with highest rank $R(X,Y;Z|\{U_i\})$ **do**
    Expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z$
    if $\mathrm{I}'(X;Y|\{U_i\}) < 0$ **then**
      Delete edge $X - Y$ from $\mathcal{G}$
      Sepset$\{X,Y\} \leftarrow \{U_i\}$
    else
      Find next most contributing node $Z \in \{\mathrm{adj}(X) \cup \mathrm{adj}(Y)\}$ and compute $R(X,Y;Z|\{U_i\})$
    end if
    Sort the rank list $R(X,Y;Z|\{U_i\})$
  end for
end while

- Skeleton orientation
Sort list of unshielded triples $\mathcal{L}_c = \{(X,Z,Y)_{X \neq Y}\}$ in decreasing order of $|\mathrm{I}'(X;Y;Z|\{U_i\})|$
repeat
  Take $(X,Z,Y)_{X \neq Y} \in \mathcal{L}_c$ with highest $|\mathrm{I}'(X;Y;Z|\{U_i\})|$ on which R_0 or R_1 orientation rules can be applied
  if $\mathrm{I}'(X;Y;Z|\{U_i\}) < 0$ **then**
    if $(X,Z,Y)_{X \neq Y}$ has no diverging orientation, apply R_0 and orient $X \rightarrow Z \leftarrow Y$
  else
    if $(X,Z,Y)_{X \neq Y}$ has one converging orientation, apply R_1 and orient $X \rightarrow Z \rightarrow Y$
  end if
  Update all orientations of $(X,Z,Y)_{X \neq Y} \in \mathcal{L}_c$
until No additional orientation can be obtained

return $\mathcal{G}$

Where $P_{\mathrm{nv}}(XYZ|\{U_i\})$ is the probability that $X - Z - Y$ is not a V-structure :

$$P_{\mathrm{nv}}(XYZ|\{U_i\}) = \frac{1}{1 + e^{-N\mathrm{I}'(X;Y;Z|\{U_i\})}} \tag{2.2.8}$$

and $P_b(XY|Z,\{U_i\})$ the probability that the base is $X-Y$

$$P_b(XY|Z,\{U_i\}) = \frac{1}{1+\frac{e^{-N\mathrm{I}'(X;Z|\{U_i\})}}{e^{-N\mathrm{I}'(X;Y|\{U_i\})}}+\frac{e^{-N\mathrm{I}'(Y;Z|\{U_i\})}}{e^{-N\mathrm{I}'(X;Y|\{U_i\})}}} \tag{2.2.9}$$

The orientation rules are also based on information theoretic measures, and can even be expressed with probabilities (we refer the reader to [35] for the full derivations). This also makes it more robust than PC, even with majority or conservative rules.

Much like FCI, MIIC is also able to take into account and discover latent variables [44], making it more apt to analyze real-life datasets. It was however limited to discrete data, for which estimating I' is rather straightforward.

One of the main objectives of this thesis was to adapt MIIC to any distribution $p(V)$, which means developing a mutual information estimator for any type of variable : discrete, continuous or a mixture of both. In the next chapter, we formally define the mutual information and its connection to causal graphs, before introducing our general case estimator developed for general constraint-based reconstruction.

Chapter 3

Mutual information for general constraint-based causal inference

3.1 Mutual information and Conditional mutual information

The mutual information is a measure of the dependency between two random variables in the most general sense. It is agnostic to the nature of the random variables and of their relationship : noted $\mathrm{I}(X;Y)$, it simply defines the quantity of information one knows about X by knowing Y, and vice-versa. It was introduced by Claude Shannon in 1948 to characterize communication channels [45] but it has found success in a wide range of applications since. It is still seen by many as the ideal dependency measure, although it is difficult to use in practice as we will see in this chapter.

In this chapter, I review previous work and present how we developed a new general case (conditional) mutual information estimator for constraint-based causal discovery on mixed variables, introduced in [7]. The section is organized a follows : I first give the necessary definitions of information theoretic concepts, review the existing estimators for both the discrete and continuous case on finite data, and I introduce the concept of optimal discretization in terms of maximizing a penalized mutual information and detail our implementation. Then, I show qualitative and quantitative results on our discretization scheme to estimate the mutual and conditional information and assess its significance.

3.1.1 Definitions

Entropy and mutual information

Before giving the definition of the mutual information between two random variables, it is a good idea to start with the self-information contained in a single variable, called the entropy.

Let X be a discrete random variable with possible values in $\mathcal{X}$ and a probability mass function $p(x) = Pr\{X = x\}, x \in \mathcal{X}$. The entropy $\mathrm{H}(X)$ of X is defined by :

$$\mathrm{H}(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{3.1.1}$$

It is expressed in *bits* with a logarithm to the base 2, or *nats* with the base e, and it denotes the average information or "surprise" that is carried by a random variable. To get a better understanding of this concept, consider a game of chance where you try to predict the result of a coin flip. If the coin is balanced, each realisation has the same "surprise" as both outcomes, heads or tails, are equiprobable. When the coin is biased towards one outcome with probability p, the average surprise decreases as p approaches 0 or 1 at which point it becomes null and your willingness to bet on the outcome increases. Note that the entropy characterises the distribution of a random variable and not the surprise of one realisation.

This definition can be naturally extended to a pair of random variables X and Y (which can be thought of a single two-dimensional variable), giving the *joint entropy* :

$$\mathrm{H}(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \tag{3.1.2}$$

We can also define the *conditional entropy*, i.e. the expected "surprise" of the conditional distribution of a variable Y given X :

$$\mathrm{H}(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{3.1.3}$$

One desirable property of these information-theoretic values is that they can be combined in an intuitive manner with the "chain rule" :

$$\mathrm{H}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y|X) \tag{3.1.4}$$

$$\mathrm{H}(Y|X) = \mathrm{H}(X,Y) - \mathrm{H}(X) \tag{3.1.5}$$

Indeed, it comes easily to think of the joint entropy of X and Y as the sum of the information carried by X plus the residual information of Y after "removing" the knowledge of X, as some information may be redundant between the two.

So far we have only defined the entropy of discrete variables, but our ideal dependency measure should also include continuous or mixed (part discrete, part continuous) distributions which are present in real-life datasets. Continuous variables are defined by a probability density function $f(x)$ instead of a mass function, which was naturally considered by Shannon to be equivalent in the definition for their entropy. There are however subtle differences with

the discrete counterpart, which is why this value is called the *differential* entropy and is noted $h(X)$:

$$h(X) = -\int_S f(x) \log f(x)\, dx, \tag{3.1.6}$$

with S as the support set of X where $f(x) > 0$.

The source of the differences between differential and discrete entropies becomes evident with our previous example of predicting the value of a random variable : what is the surprise of a realisation of X given that there is an infinite number of possible values in any continuous interval, each with a probability that tends to 0 ?

It is better to think of the differential entropy as an estimate of the effective volume that a random variable occupies : a very focused distribution will have a low entropy as opposed to a more dispersed distribution with more *room* for randomness hence higher entropy. Formally, the differential entropy is the logarithm of the length of the smallest interval that contains most of the probability [4] : for example, the differential entropy of a uniform distribution on the interval $[0, a]$ is $\log(a)$.

Although the differences between entropy and differential entropy go beyond the scope of this thesis, it is still interesting to mention them to understand why discretization has been so popular for so long to estimate the entropy or mutual information between samples of continuous variables. Even though, as we will see in sections 3.1.4 and 3.2, one must be careful to discretize a continuous variable without introducing bias.

Entropy and differential entropy behave similarly and are interchangeable in the settings that interest us, namely for the joint and conditional differential entropy, the chain rule and especially the relationship to mutual information. For the rest of section 3.1.1, the probability mass function $p(x)$ can be replaced by the density function $f(x)$, and $H(X)$ by $h(X)$ to switch from discrete to continuous random variables. The special case of mixed variables with both continuous and discrete parts will be reviewed at a the end of this section.

We have established that information theory gives us the necessary tools to define the entropy of a random variable (which can be multidimensional or conditional), i.e. the amount of information needed on average to describe it. Next we show how we can also formalize how much information two variables have in common, giving a measure of how (in)dependent they are. For this we need to introduce the Kullback-Leibler divergence, also called the *relative entropy*. The relative entropy $D_{\mathrm{KL}}(p \parallel q)$ is a measure of the difference between two distributions p and q defined on the same space $\mathcal{X}$:

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{3.1.7}$$

Also called the Kullback-Leibler distance (although not a distance in the usual sense as it

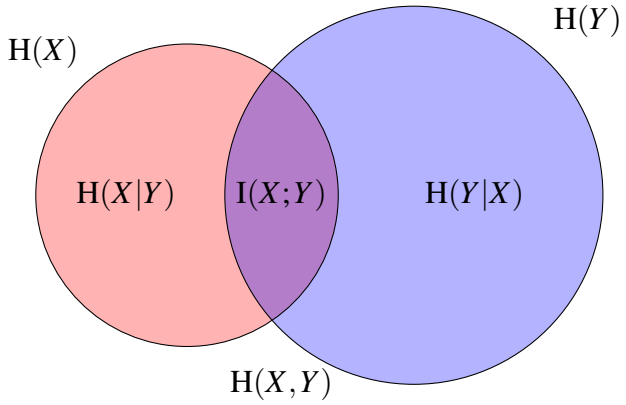

Figure 3.1: Relationship between the entropy, conditional entropy, joint entropy and mutual information between two variables

is not symmetric), it can thought of as the cost of describing the distribution p when using q as a reference model. As such, it is null if and only if $p = q$ and it is always non-negative.

Now we get back to our original goal which is to define the dependency between two random variables X and Y. In the most general sense, X and Y are independent if the realization of one does not affect the probability distribution of the other. Formally put, two random variables X and Y with marginal distributions $p(x)$, $p(y)$ and a joint distribution $p(x,y)$ are independent if and only if $p(x,y) = p(x)p(y)$. If these two quantities differ, some information is being shared between X and Y : knowing about X tells us *something* about Y and vice versa.

Using the measure of divergence we just introduced, it becomes natural to think of the divergence between the joint distribution and the product of marginals as a direct measure of the dependency. It is in fact the definition of the mutual information $\mathrm{I}(X;Y)$:

$$\mathrm{I}(X;Y) = D_{\mathrm{KL}}\left(p(x,y) \parallel p(x)p(y)\right) \tag{3.1.8}$$

$$= \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{3.1.9}$$

In agreement with our interpretation of the relative entropy, assuming the independence model where $p(x,y) = p(x)p(y)$ the mutual information is literally *the extra bits* that are required to encode the interaction between X and Y. It is always positive, or null if and only if X and Y are independent.

Just like the other measures, it fits naturally in the "chain rule" and can be expressed intuitively in terms of entropies (Fig 3.1):

$$\begin{aligned} \mathrm{I}(X;Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) && (3.1.10)\\ &= \mathrm{H}(Y) - \mathrm{H}(Y|X) && (3.1.11)\\ &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) && (3.1.12) \end{aligned}$$

What makes the mutual information a particularly interesting measure is its unique blend of desirable properties.

First, it satisfies the Data Processing Inequality (DPI) which states that one cannot increase the information content of a signal by processing it. Formally, if n variables form a Markov chain $X_0 \to X_1 \to \cdots \to X_n$, then $\mathrm{I}(X_i;X_j) \geq \mathrm{I}(X_i;X_k)$ with $i < j < k$.

In relation to the first point, mutual information is also widely considered to be equally sensitive to all types of relationships. This concept was termed "equitablity" by Reshef et al. [46] (although in a flawed form) and was then formally investigated by Kinney et Atwal [27]. Kinney et Atwal's "Self-Equitability" is defined to characterize a dependence measure $D[X;Y]$ if and only if it is symmetric between X and Y, and :

$$D[X;Y] = D[f(X);Y] \tag{3.1.13}$$

with f any deterministic function, $X \leftrightarrow f(X) \leftrightarrow Y$ forming a Markov chain. Put roughly, an equitable measure means that one can measure the strength of the signal (as compared to the noise) between Y and $f(X)$ without having to know the underlying function f.

Not only is it invariant to invertible transformations of X and Y, it is also invariant under any monotonic (i.e. rank preserving) transformations.

Put together, these three properties make the mutual information particularly interesting for general case causal discovery. True causal discovery should make no assumption of the natural mechanisms that produced the observed data, whether on the scale of the unit or shape of the joint distributions. As a simple example, a case can be made to measure the human weight in a logarithmic scale instead of a linear one : for most health related aspects, a difference of 30 kilograms is much more significant between 60 and 90kgs than between 120 and 150kgs. In an experimental context, we can think of the causal diagram as the natural laws that have produced the observations, which are themselves a function of the "observing" process. The self-equitability property and invariance under transformation go some way towards freeing ourselves from this observation process and our own biases.

Finally, as will be discussed later, these properties hold for any type of variable X and Y, be it continuous, discrete (ordinal or not), or a mixture of discrete and continuous parts.

A notable disadvantage of mutual information compared to other measures is that the *bit*, unit of information, is not commonly understood, and the fact that it is unbounded upwards.

One usually cannot easily derive a p-value from a mutual information estimation on sampled data, which makes it harder to communicate (although the benefits of standardising the p-value have been called into question [47, 48]). Different ideas to evaluate significativity will be presented in Section 3.1.4, but for now we are only interested in the "oracle" value, when the sample size N tends to infinity and the strict equivalence $X \perp\!\!\!\perp Y \leftrightarrow \mathrm{I}(X;Y)$ holds.

In the discrete case, its value is actually familiar as it is in fact the G-statistic multiplied by a factor of N. With O_i the number of observations in a contingency table between two categorical variables X and Y, with i joint levels, and E_i the expected counts under the null hypothesis $X \perp\!\!\!\perp Y$, the G-statistic is defined as :

$$G = 2\sum_i O_i \log\left(\frac{O_i}{E_i}\right) \tag{3.1.14}$$

Recall the definition of the KL divergence (Eq 3.1.7), the same formula as above except for the frequencies (noted o_i and e_i) instead of the *counts* O_i, E_i. Using the frequencies, the G-statistic becomes :

$$\begin{aligned} G &= 2N\sum_i o_i \log\left(\frac{o_i}{e_i}\right) \\ &= 2N \cdot D_{\mathrm{KL}}(o \parallel e) \\ &= 2N \cdot \mathrm{I}(X;Y) \end{aligned} \tag{3.1.15}$$

It also follows that the mutual information is related to the χ^2 test, as it is itself a second-order Taylor approximation of the G-statistic.

For two continuous variables, it may be harder to get a good intuition of what the mutual information measures. A first property that may seem odd is that if X is continuous and $Y = X$, then $\mathrm{I}(X;Y) = \infty$. This looks as though it contradicts the chain rule (Eq 3.1.10), since it implies that $\mathrm{H}(X) - \mathrm{H}(X|X) = \infty$. It is in fact one of the differences between entropy H and differential entropy h which is unbounded in the case of a singularity $h(X|X) = -\infty$, unlike H which is always finite. Thankfully, this theoretical property does not bleed into the real world for several reasons : first, even continuous distributions have always finite differential entropies since we actually treat real numbers up to a finite number of significant digits. Second, much like a correlation coefficient on observed data never reaches 1, we should not ever need to estimate $\mathrm{I}(X;Y)$ where $X = Y$. The analytical value of the mutual information on a bivariate Gaussian with correlation coefficient ρ is actually known :

$$\mathrm{I}(X;Y) = -\frac{1}{2}\log(1-\rho^2) \tag{3.1.16}$$

This equivalence is useful for practitioners who are unfamiliar with mutual information and wish to translate it to the better known dependence measure : thanks the self-equitability

property if one could transform two variables to a bivariate Gaussian distribution preserving the signal-to-noise ratio, using Eq 3.1.16 one could then get the corresponding correlation coefficient (see Fig 3.2).

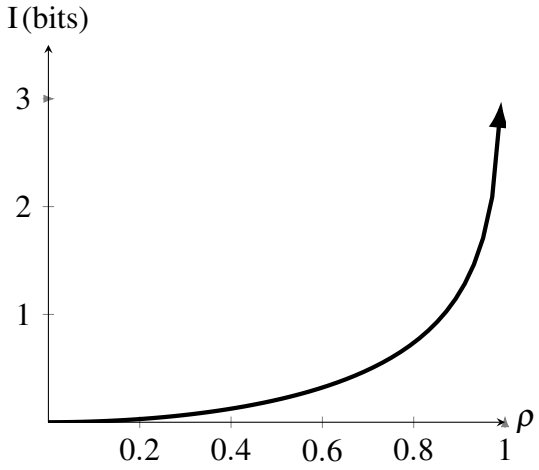

Figure 3.2: Value of the mutual information of a bivariate gaussian with correlation coefficient ρ

Conditional information and interaction information

Information theory also allows us to measure the conditional dependence between two variables X and Y given a third, Z. The conditional mutual information is defined as the expected value of the mutual information between X and Y given a third variable Z:

$$\mathrm{I}(X;Y|Z) = \mathbb{E}_z \left[D_{\mathrm{KL}} \left(p((x,y)|z) \parallel p(x|z)p(y|z) \right) \right] \tag{3.1.17}$$

It is symmetrically decomposable into two-points mutual informations :

$$\begin{aligned} \mathrm{I}(X;Y|Z) &= \mathrm{I}(X;Y,Z) - \mathrm{I}(X;Z) \\ &= \mathrm{I}(Y;X,Z) - \mathrm{I}(Y;Z) \end{aligned} \tag{3.1.18}$$

where X,Z and Y,Z are joint variables. The conditional mutual information can only be positive, or null if and only if $X \perp\!\!\!\perp Y|Z$.

Finally, the information between more than two variables is called the interaction information. We define it for three variables X,Y,Z and a conditioning set U_i :

$$\begin{aligned} \mathrm{I}(X;Y;Z|\{U_i\}) &= I(X;Y|\{U_i\}) - I(X;Y|\{U_i\},Z) \\ &= I(X;Z|\{U_i\}) - I(X;Z|\{U_i\},Y) \\ &= I(Y;Z|\{U_i\}) - I(Y;Z|\{U_i\},X) \end{aligned} \tag{3.1.19}$$

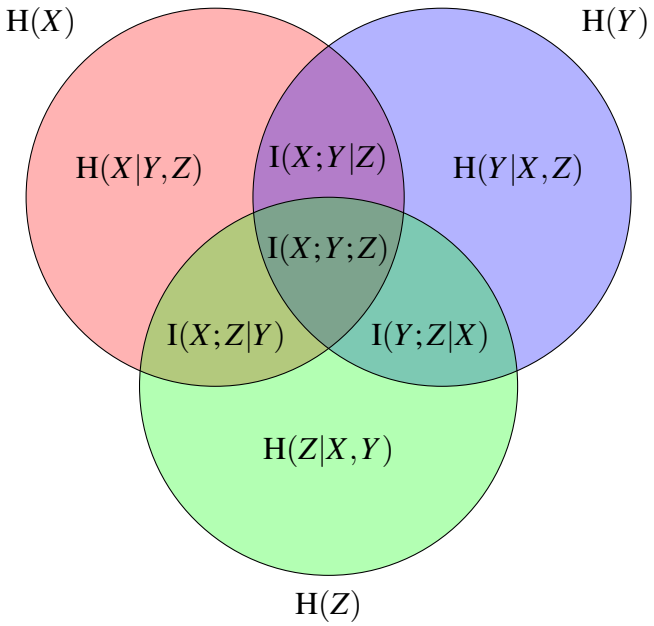

Figure 3.3: Relationship conditional mutual information and three-point information with three variables. Note that the three-point information can be negative when two variables are pairwise-independent but become dependent when conditioning on the third.

Unlike the other measures introduced so far, it can be both positive and negative. A positive interaction information indicates that the three variables share some common information. It is negative when there is more information when taking the three variables together than independently. To illustrate this property, we borrow the concept of V-structure from causal diagrams. Consider the 4 possible DAGs with 3 nodes and two edges, shown in Fig 3.4.

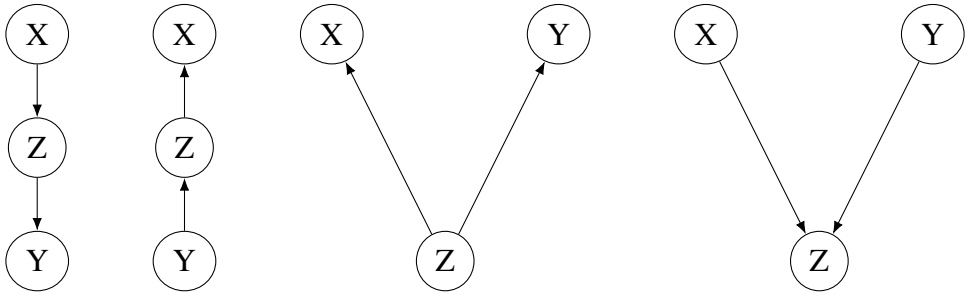

Figure 3.4: DAGs with 3 nodes and 2 edges.

As Bayesian networks, the first three graphs encode the same conditional dependencies : $X \not\perp\!\!\!\perp Y$, and $X \perp\!\!\!\perp Y|Z$. In terms of informations, $\mathrm{I}(X;Y) > 0$ and $\mathrm{I}(X;Y|Z) = 0$. They all share some information, either due to a common cause or having a continuous "flow" of information, so $\mathrm{I}(X;Y;Z)$ is also positive. Only the fourth graph on the right shows a different pattern : X and Y are marginally independent ($I(X;Y) = 0$), but become dependent when conditioning on Z ($\mathrm{I}(X;Y|Z) > 0$). This is the situation where we "create" information

by looking at the interaction of the three variable, and the three-point information $\mathrm{I}(X;Y;Z)$ is negative.

3.1.2 Mixed variables

In many real-life datasets, particularly in medical records of patients, we might encounter both discrete and continuous variables and want to measure their interactions, *without favoring one type of variable or the other*. In other words, the dependency measure should scale with the signal-to-noise ratio the same way for continuous-continuous, discrete-discrete or discrete-continuous combinations. In the context of constraint-based approaches, another layer of difficulty is added as the same applies for the variables of the conditioning set.

There also exists yet another type of variables that has the characteristics of both continuous and categorical variables, and fits neither definition. For example, think of the height measured in centimeters without decimals : it is not defined on a truly continuous interval as it has non-zero probability to take certain values, but it also has too many unique values, which are potentially infinite (but countable), to be considered discrete. Many estimators depend on one or the other of these properties (continuous density or finite number of levels) to measure the dependency between observations, and will likely struggle to give an unbiased estimation on this type of variable [49, 50, 51, 52]. Another problematic distribution is the *mixture random variable*, which is itself a mixture of discrete and continuous parts. A prominent example of this would be a distribution bounded by a minimum value before a certain threshold, and a continuous function of x after it. In such a case the "minimum value" x_{min} can be seen as the discrete part of the distribution as $p(x_{min}) > 0$, with the rest behaving like a continuous variable. Real-life examples include the values produced by real-time quantitative polymerase chain reaction (RT-qPCR), used to measure the levels of of messenger RNAs in a cell. One can view the data produced by RT-qPCR as $(A(x)$ if $x >$ threshold, else $0)$ with x the level of mRNA in the cell and A the amplification process. Another straightforward example is the ReLU activation function $f(x) = \max(0,x)$ widely used as an activation function in artificial neural networks (Fig 3.5).

The ReLU function was actually discussed recently in the context of mutual information, and the problems that classical estimators face with such zero-inflated distribution. Naftali Tishby was a prominent computer scientist and physicist, who also contributed to signal processing and tried to apply information-theoretic concepts to gain intuition on deep learning algorithms. With Pereira and Bialek, he proposed the Information Bottleneck framework, a self-described *surprisingly rich framework for discussing a variety of problems in signal processing and learning* with information theory [53]. Put simply, the idea of the Information Bottleneck is to "squeeze" a signal X to a compressed representation T while minimizing the

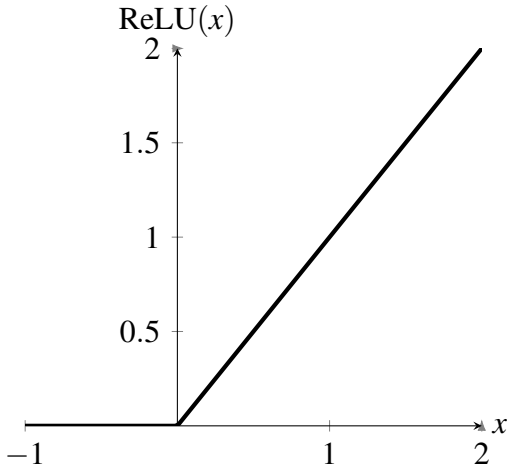

Figure 3.5: The ReLU activation function is mixed: it has $p(0) > 0$ and is continuous for $x > 0$.

loss of information with another relevant variable Y :

$$\min_{p(t|x)} \{\mathrm{I}(X;T) - \beta\, \mathrm{I}(T;Y)\}$$

With the Lagrange multiplier β for controlling how much loss we can tolerate when predicting Y from T as opposed to X (recall the DPI principle, since $Y \to X \to T$ is a Markov chain, $\mathrm{I}(X;Y) \geq \mathrm{I}(T;Y)$). By minimizing this difference, we want to reduce X to the only part that is relevant to Y, discarding the rest. This very general framework provides an elegant, if unpractical, solution to the majority of modern machine learning which has to learn which aspects of the input X is useful for predicting Y, and which are noise.

As deep learning models gathered success faster than a comprehensive theory could definitely explain why they work and how they can be further improved, Shwartz-Ziv and Tishby published new evidence that they claimed could explain the process of training a deep neural network [54]. In their experiments, they equated the noisy encoding T of the information bottleneck to the hidden layers of a deep neural network (DNN) and measured $\mathrm{I}(X;T)$ and $\mathrm{I}(T;Y)$ during the training process. Their results showed that the training process acts in two separate phases : first, the fitting phase in which the network maximizes $\mathrm{I}(T;Y)$, and then a compression phase that minimizes $\mathrm{I}(X;T)$. This was an unprecedented window inside the "black box" of deep learning and could potentially explain how they train, and most importantly how they are able to generalize. Later however, more studies were published and seemed to show that the two phases observed in the original experiment were not in fact an information-theoretic phenomena, but more of an artefact of how the mutual information is estimated between the hidden layers and Y. Saxe et al. could not replicate the two phases in other network architectures from the ones tested in the original study, and in particular no compression phase was observed when training with linear activation functions or ReLU

[55]. In response, Shwartz-Ziv and Tishby claimed that Saxe et al. had used a weak estimator of mutual information, and defended their general claim saying that "when properly done, there are essentially the same fitting and compression phases" on any network. There are however other reasons to believe the compression phase observed in the original study was more a result of geometric operations as the weights of the network are trained, and does not hold so much ground in information theory [56, 57]. Moreover, the simple DNNs are no longer used in practice, they are being replaced by extremely scaled up versions (with too many parameters in hidden layers for mutual information to ever be estimated) or more sophisticated architecture involving different training mechanisms like transfer learning, attention mechanisms etc... diverging from the simple picture of training that was examined.

We may not know the final word on the information bottleneck for deep learning, but it serves as a cautionary tale when we want to rely on mutual information estimates on big data (as the dimension of X gets large) and the distributions are unfamiliar. It is fortunately not the case for constraint based causal discovery approaches, where X and Y are usually one-dimensional, and the conditioning set Z *few-dimensional*. Moreover, recent advances were made to better understand mutual information estimators, including on such mixed distributions, as will be discussed in the next section.

It is not obvious if we are still allowed to swap differential entropy for entropy when considering the mixed case. Crucially, it is not well defined for mixture distribution which are defined neither by a probability density function nor a mass function alone.

Recent efforts to estimate the mutual information in this general setting have relied on the Radon–Nikodym theorem. With P_{XY} a probability measure on the space $\mathcal{X} \times \mathcal{Y}$, $\mathcal{X}$ and $\mathcal{Y}$ being Euclidean spaces. If P_{XY} is absolutely continuous with relation to $P_X P_Y$:

$$\mathrm{I}(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} log \frac{dP_{XY}}{dP_X P_Y} dP_{XY}, \tag{3.1.20}$$

where $\frac{dP_{XY}}{dP_X P_Y}$ is the Radon-Nikodym derivative. Note the only condition this definition is absolute continuity of P_{XY}, and if true it applies for all cases mentioned so far : X and Y are the same type of variable, X or Y is discrete and the other is continuous, or X, Y or the joint distribution is a mixture itself. Moreover, the Radon-Nikodym derivative is computable in practice [51].

Another way to deal with mixtures is to refer to the master definition of mutual information [4]. For two random variables X and Y discretized with partitions $\mathcal{P}$ and $\mathcal{Q}$:

$$\mathrm{I}(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} \mathrm{I}([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \tag{3.1.21}$$

where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$. It is called the master definition as it always applies, regardless of the nature of the marginal and joint distributions. For discrete

variables it is simply equivalent to the definition of mutual information (Eq 3.1.9), i.e. the partitions are fixed. For continuous variables, the supremum is obtained by refining $\mathcal{P}$ and $\mathcal{Q}$ into finer and finer bins, monotonically increasing $\mathrm{I}([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \nearrow$. When $N \to \infty$, this quantity tends to the real value of the mutual information (just as the entropy of a discretized variable is approached as the numbers of bins tends to infinity). On a finite sample size however, adding bins to $\mathcal{P}$ and $\mathcal{Q}$ will inevitably end up overestimating the mutual information, to the limit of having one unique value per bin for which (which results in $\mathrm{I}(X;Y) = log(N)$). In section 3.1.4, we review previous work on choosing the appropriate number of bins to estimate I on continuous data and in section 3.2 we introduce our solution based on the master definition.

As a general rule, methods that assume a continuous probability density function $p(x,y)$ over the domain of X and Y tend to not work well in the mixed case. Any dependence measure having this assumption will need to be adapted (with more or less difficulty), which may also affect the way we can evaluate its significativity for independence testing. On the other hand, one can still rely on the cumulative distribution function, which is well behaved even for mixture variables (although may not be smooth). For example, decision-tree-based algorithms like random forests and gradient boosting work well with such mixtures (although they are not adapted to all cases, for example they do not deal well with non-ordinal categorical variables with many levels).

Mutual information is one of the rare measures fit to deal with such distributions, all while keeping its desirable properties. Particularly, its strict equivalence with variable independence (and conditional independence), and its self-equitability property make it ideal for a general case constraint-based algorithm for causal inference. In the next subsection we show the equivalences between information theoretic measures and "constraints" in the causal diagrams.

3.1.3 Mutual information and causal graphs

In this part, we make the usual assumptions to bridge the gap between distributions and causal diagrams, namely the Faithfulness and the causal Markov condition. We now take the time to review and discuss these conditions as they are quite relevant to the topic of general-purpose conditional dependence measures.

The first is the causal Markov condition, defined as follows :

Definition 3.1. The causal Markov condition states that, given a set of variables V with joint probability $P(V)$ governed by the true causal graph $\mathcal{G}_c$, any given node X is conditionally independent from any other non-descendant node Y given its parents Pa_X : $X \perp\!\!\!\perp Y \mid Pa_X$.

Put roughly, this is equivalent to the Markov condition of Bayesian graphs with extra precautions on ancestor vs descendant nodes to avoid *collider bias*, which "opens up"

information flows by conditioning on a child (see Fig 3.4). Additionally, it means that a node is separated from the rest of the network, *i.e.* statistically independent from every other node, conditionally on its Markov blanket (Fig 2.6). This first assumption may seem specific to causal diagrams but it actually reflects common scientific reasoning : we assume that two variables that are correlated are involved at some point in the same causal mechanism, whether it be due to a common ancestor or one causing the other [1].

The converse to the Markov condition is the faithfulness assumption :

Definition 3.2. Given a set of variables V whose true causal graph is $\mathcal{G}_c$, their joint probability $P(V)$ is faithful to $\mathcal{G}_c$ if all conditional independences in $P(V)$ correspond to d-separation in $\mathcal{G}_c$.

It is also called the "stability" property by Pearl [1], referring to the process by which $P(V)$ results from $\mathcal{G}_c$. Indeed, in general we can think of two ways independencies are created in the data (recall that with a causal diagram, $p(X) = f(p(Pa_X), U_X)$). The first, intuitive way is to think that two variables are independent only if they can be separated in the causal diagram, either by removing the influence of a common cause or of the intermediate steps. These independencies are called "stable", they hold true under almost any parameterization of the causal mechanisms $f(p(Pa_X), U_X)$. But there is another way to create independencies in $P(V)$, by carefully choosing mechanisms that either cancel each other statistically, or are undetectable marginally and appear only when looking at interactions between nodes. For example, an XOR gate $X \rightarrow Y \leftarrow Z$ produces an unfaithful distribution as X and Z are marginally independent from Y, but are clearly implicated in the generation function for Y. This type of relation is considered "unstable" as it is strictly dependent on its parameterization, any deviation from this equilibrium would create a different pattern of dependencies. As explained by Pearl, the faithfulness or stability assumption can be thought of as a filter on the type of causal relations that we are considering [1]:

> *Any story that convincingly exemplifies a given pattern of dependencies must sustain that pattern regardless of the numerical values of the story parameters – a requirement we called "stability."*

This discussion on stable distributions prompts us to make another aside in deep learning, as an interesting parallel can be made with the recent theory of Causal Invariance advocated by Léon Bottou and Jonas Peters [58, 59, 42]. As we know, machine learning in its current form is entirely dependent on statistical dependencies in the training data and while it has found success with many applications it does not generate real knowledge on how the world functions. Invariant causal prediction aims to discover causal knowledge by observing correlations in many different environments (different interventions in a causal context, or different training datasets for machine learning) and positing that the true causal relationships

are the ones that hold true across the different conditions. It reflects the human thought process of observing our environment and detecting the patterns that stay true even in different situations.

We can see that this actually echoes the faithfulness assumption in the sense of stable distributions. Indeed, causal mechanisms produce stable distributions because the true relationships between two variables are invariant to external influence on other parts of the system, *i.e.* to changes in the environment. The main difference is that causal invariance is only interested in discovering the causes of Y as a set of variables, and not the full structure of the causal graph. This matters for example for the XOR relationship, where causal invariance would be able to detect both parents X and Y as causes of Z (provided they are always both included in the varying environments), whereas causal discovery would struggle to find the corresponding graph as the corresponding $P(V)$ is unfaithful in relation to the graph.

In other words, with the faithfulness assumption, we restrict our modelling of causal mechanisms to relationships $X - Y$ that stay invariant when modifying any other part of the causal graph. Crucially, this excludes interactions terms (like the XOR gate) and cancelling paths. Causal invariance aims for a larger class of mechanisms by looking for laws that stay invariant when changing the *environment* of the entire set $\{Pa_Y, Y\}$. This difference points to the inherent difficulty to represent interaction terms between nodes in Bayesian networks, as opposed to the multidimensional input in mainstream machine learning.

The Markov condition is generally considered as the less problematic assumption of the two. It is more of a convention, to limit the task of causal discovery to complete models, including latent variables (when would we want to discover incomplete models?). The faithfulness assumption on the other hand has generated discussion ever since the first constraint based method was published and an extensive body of work has aimed to remove or relax it [1, 20, 60, 61, 62]. With the previous discussion, we hope to have given enough reasons to consider it a fair assumption as we assume both are true going forward.

These assumptions are actually one step removed from the way constraint-based approaches work, as they have to estimate (conditional) independences from the data. As we have shown in Section 2.2.3, independence testing comes with a set of assumptions of its own, which narrows further the definition of causal mechanisms that most constraint-based approaches actually discover. This is not the case with mutual information, thanks to its strict equivalence with statistical independence and its self-equitability property. In the rest of this section, we show the direct equivalence between causal diagrams and information measures, and in the next section we discuss how they are estimated from finite data.

For a collection of variables which we note X, Y, Z..., in a causal graph $\mathcal{G}_c$:

- If X and Y are adjacent in $\mathcal{G}_c$, then $\mathrm{I}(X;Y) > 0$, and $\mathrm{I}(X;Y|Z) > 0$ for any set Z. This follows from the direct equivalence $X \perp\!\!\!\perp Y \Leftrightarrow \mathrm{I}(X;Y) = 0$

- Similarly, if X and Y are d-separated by Z in $\mathcal{G}_c$, then $\mathrm{I}(X;Y|Z) = 0$.
- If Y is a direct parent of X and Y_2 an ancestor of X only through Y, then $\mathrm{I}(X;Y) \geq \mathrm{I}(X;Y_2)$ (according to the DPI).
- If $X \to Z \leftarrow Y$ forms a V-structure and X and Y are d-separated by U_i, then $\mathrm{I}(X;Y;Z|U_i) < 0$. [35, 5]
- And the inverse, if $X - Y - Z$ does not form a V-structure and X and Y are d-separated by U_i, then $\mathrm{I}(X;Y;Z|U_i) > 0$

For constraint-based causal discovery, we can also exploit three-point informations to iteratively take off the best contributors, as was shown in [5, 35] to develop MIIC.

$$I(X;Y|\{U_i\},Z) = I(X;Y) - I(X;Y;U_1) - I(X;Y;U_2|U_1) - \cdots - I(X;Y;Z|\{U_i\}) \quad (3.1.22)$$

Another contribution of Affeldt et al. is the definition of orientation probability from three point information. Arrow head probabilities stem from v-structures like $X \to Z \leftarrow Y$, corresponding to a negative conditional 3-point information (Fig 3.4), $\mathrm{I}(X;Y;Z \mid \{U_i\}) < 0$, where $\{U_i\}$ separates X and Y, *i.e.* $X \perp\!\!\!\perp Y|\{U_i\}$ [5]. The head orientation probabilities can then be obtained through the probability decomposition formula as,

$$P(x \to \underline{z}) = P(x \to \underline{z}|\underline{z} \leftarrow y)P(\underline{z} \leftarrow y) + P(x \to \underline{z}|\underline{z} - y)P(\underline{z} - y) \quad (3.1.23)$$

or equivalently, writing the probability of a v-structure as

$$\begin{aligned} P(x \to \underline{z}, \underline{z} \leftarrow y) &= P(x \to \underline{z}|\underline{z} \leftarrow y)P(\underline{z} \leftarrow y) && (3.1.24)\\ \frac{P_{\to\leftarrow}}{P_{\to\leftarrow} + P_{\to -} + P_{-\leftarrow} + P_{--}} &= \frac{P_{\to|\leftarrow}}{P_{\to|\leftarrow} + P_{-|\leftarrow}} P(\underline{z} \leftarrow y) && (3.1.25)\\ \frac{1}{1 + 3e^{N\,\mathrm{I}(X;Y;Z \mid \{U_i\})}} &= \frac{1}{1 + e^{N\,\mathrm{I}(X;Y;Z \mid \{U_i\})}} P(\underline{z} \leftarrow y) && (3.1.26) \end{aligned}$$

which leads (by x/y symmetry) to

$$P(x \to \underline{z}) = P(\underline{z} \leftarrow y) = \frac{1 + e^{N\,\mathrm{I}(X;Y;Z \mid \{U_i\})}}{1 + 3e^{N\,\mathrm{I}(X;Y;Z \mid \{U_i\})}} \quad (3.1.27)$$

By default, we orient V-structures if the probability is larger than 0.5, but we can also be more or less strict and choose any arbitrary threshold. In Section 4.1.2, we introduce new probabilities that behave better numerically for large N.

Going beyond the problem of graph inference, Wieczorek et al. also showed how to derive causal effect quantities using only information theoretic terms [63]. The causal graph alone informs us of the presence or absence of interactions, but we have no way of comparing

them or even knowing if a given edge is a strong or a weak effect. These questions are answered by causal effect quantification, which has been studied in various forms. In the framework provided by Pearl's diagrams, the the causal effect of X on Y is simply described by the interventional distribution $P(Y|do(X))$. Other measures of causal strength (notably in Rubin's potential outcome framework) are the average treatment effect and the specific causal effect. The average treatment effect for binary variables is the expected value of the difference between $Y|do(X=1)$ and $Y|do(X=0)$:

$$ACE(X,Y) = \mathbb{E}[Y|do(X=1) - Y|do(X=0)] \tag{3.1.28}$$

And the specific causal effect is the average treatment effect conditional on a particular value of other variables Z (provided that Z are non-descendant of X in $\mathcal{G})c$). The contribution of [63] is to prove that the combination of conditional mutual information and conditional directed information also give a rigorous framework for causal effect quantification from causal diagrams. The directed information is typically used in time-series analysis, it measures the amount of information that flows from one process to the other. In the context of interventions, Wieczorek et al. define it as :

$$\mathrm{I}(X \to Y) = D_{\mathrm{KL}}\left(P(X|Y) \;\|\; P(X|do(Y))|P(Y)\right) \tag{3.1.29}$$

This line of work ties up the relationship between mutual information and causal diagrams. From graph inference with MIIC to the quantification of causal effects, all of the necessary concepts in causality can be expressed with information theoretic measures.

But can it be used in practice ? Earlier, we hinted at potential issues and shortcomings when estimating mutual information on sampled data. In the next section we present previous work on existing estimators before introducing our new method that can estimate mutual and conditional mutual information on any type of data and also assess its significance.

3.1.4 Existing estimators on finite data

The previous section has established *why* we would want to estimate information-theoretic quantities from data, and now we will study *how* and *how well* it can be done. Several decades of research later, and almost as many different estimators as there were applications, it may come as a surprise that many basic questions remain unanswered (although recent progress has been made, especially in the continuous case). To understand why, recall that mutual information $\mathrm{I}(X;Y)$ is defined for X and Y of any dimensions. For many applications in neuroscience, X may be the activation of hundreds or thousands or neurons, and Y a single-dimensional stimulus or response. Estimating $\mathrm{I}(X;Y)$ from sampled data in this setting is a very different problem than estimating it between two single-dimensional signal! In

this section we focus on the use of (conditional) mutual information for constraint-based algorithms, where X and Y are single-dimensional variables and Z may be multidimensional (but rarely very large).

Discrete estimators

Estimating $\mathrm{I}(X;Y)$ on discrete data is the most straightforward case. We can simply estimate the probability mass functions $\hat{p}(x)$, $\hat{p}(y)$ and $\hat{p}(x,y)$ from independently and identically distributed (i.i.d) data by counting how many times we observe each level. Using the chain rule (Eq 3.1.10), we actually only need an entropy estimator $\hat{H}$ to get an estimation $\hat{\mathrm{I}}$. Using the observed frequencies $\hat{p}_i$ with $i \in [1,m]$, we get what is called the "plug-in" or "naive" estimator :

$$\hat{H}_{\text{Naive}} = -\sum_{i=1}^{m} p_i \log p_i \tag{3.1.30}$$

Note that it is also the maximum likelihood estimator from the observed data. It is however suboptimal, it has long been known that it is negatively biased everywhere [64]. The short explanation is that while $\hat{p}_i$ is estimated with symmetric variance on either side of the true frequency p_i, the log transformation amplifies more variance towards 0 than towards 1, and the contribution of each $\hat{p}_i$ ends up being underestimated on average. To correct this shortcoming, a common fix is to add the Miller-Madow correction [65]:

$$\hat{H}_{\text{MM}} = \hat{H}_{\text{Naive}} + \frac{\hat{m}-1}{2N} \tag{3.1.31}$$

with $\hat{m}$ the number of categories with nonzero probability as estimated from the $\hat{p}_i$. This correction effectively reduces the bias of $\hat{H}_{\text{Naive}}$ without adding any complexity, and is preferred in many contexts.

Another popular idea is to use a jacknife resampling procedure, which trades lower bias for a slightly higher complexity [66] :

$$\hat{H}_{\text{JK}} = N\hat{H}_{\text{Naive}} - \frac{N-1}{N}\sum_{j=1}^{N} \hat{H}_{\text{Naive}-j} \tag{3.1.32}$$

where $\hat{H}_{\text{Naive}-j}$ is the naive estimator without the jth sample.

Finally, another way to correct the negative bias of the naive estimator is to act directly on the estimates $\hat{p}_i$ instead of applying a correction a posteriori. The Schurmann-Grassberger estimator does exactly that, by applying prior Bayesian belief that the samples follow a Dirichlet distribution (the multivariate generalization of the Beta distribution) [67]. It essentially "tricks" the estimator to think that more counts have been observed to compensate for the negative bias of the naive estimator, such that mN becomes the a priori sample size. The result is a less biased estimator, but the choice of the prior end up dominating the

estimation [68].

All of these improved estimators have been designed for the setting where $I(X;Y) >> 0$, as opposed to constraint-based discovery where we are more interested in the independence regime. Importantly, they all share another kind of bias : they overestimate dependencies on finite data. Without knowing the true distributions, any of these estimators will be positive $\hat{\mathrm{I}}(X;Y) > 0$ (resp. $\hat{\mathrm{I}}(X;Y|Z) > 0$) almost surely, even when $X \perp\!\!\!\perp Y$ (resp. $X \perp\!\!\!\perp Y|Z$). Several suggestions have been made, mostly based on fixed thresholds as a function of the sample size. A more inspired approach is to also take into account the distributions of the variables : indeed, we do not expect the same bias from sampling simple binary variables with balanced levels, versus more complicated variables with many unbalanced categories.

This is the route taken by MIIC, which corrects the naive estimate by subtracting a complexity cost that depends on X, Y and Z. It frames each test of independence in the context of graph reconstruction, favoring simpler models with fewer edges. Namely, it introduces a complexity cost for the edge $X - Y$ potentially separated by separating set U_i, noted $k_{X;Y|\{U_i\}}$. Then, the condition $\mathrm{I}(X;Y|U_i) < k_{X;Y|\{U_i\}}(N)/N$ to remove the edge $X - Y$ favors the simpler model compatible with the independencies in the sense of the model complexity, given the observed data. This replaces the strict equivalence $\mathrm{I}(X;Y|U_i) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|U_i$ which is only valid in the limit $N \to \infty$. The challenge now is to choose the form of $k_{X;Y|\{U_i\}(N)}$. A common complexity cost used in model selection would be the Bayesian Information Criterion :

$$k^{\mathrm{BIC}}_{X;Y|\{U_i\}}(N) = \frac{1}{2}(r_X - 1)(r_Y - 1)\prod_i r_{U_i}\log(N) \tag{3.1.33}$$

with r_X, r_Y, r_{U_i} the number of categories of each variable (U_i being a joint variable). This complexity cost can be improved by also taking into account the distributions of the variables, not only their number of levels [35]. Such a score will be discussed in Section 3.2.1 when introducing the new MDL-optimal discretization scheme.

Continuous estimators

Compared to the discrete case, estimating $\hat{\mathrm{I}}$ on continuous data is notoriously difficult. Historically, one of the most common way to deal with continuous data was to discretize them into bins, the same way we construct histograms. We note $[X]$ and $[Y]$ the quantized version of X and Y on finite data. This approach is conceptually straightforward, we can simply compute $\hat{\mathrm{I}}([X];[Y])$ with any discrete estimator and take it as an approximation of $\mathrm{I}(X;Y)$.

Perhaps because we are used to seeing histograms and picking the correct number of bins visually, surprisingly many applications perform this kind of naive discretization without much justification. In practice however, both the number of bins and their locations dominate

the estimation. Even for large N, $\hat{\mathrm{I}}([X],[Y])$ converges on some value that depends on the discretization parameters rather than $\mathrm{I}(X;Y)$, namely the number of bins $|\Delta_X|$ and $|\Delta_Y|$, as well as their size [69]. This bias was already documented as early as 1989, but was considered manageable if one chose a "reasonable number of cells" [70]. But the question of what is "reasonable" is more complicated than it appears. For example, Ross et al. note that there is no optimal value of $|\Delta|$ that works for all distributions : $N^{0.5}$ works well for the square wave distribution but $N^{0.7}$ is better for a Gaussian distribution [71]. Similarly, Seok et al. show that even for Gaussian bivariate distributions with the same marginals, the "correct" number of bins that gives the best approximation of $\mathrm{I}(X;Y)$ varies depending on the strength of the correlation ρ [49]. Note that the same applies for any estimator that takes a number of bins as parameter, regardless of how clever the discretization scheme is (for example, using B-splines [72]). Instead, it is essential to deduce the number of bins from the observations [73, 74]. Darbellay et al's recursive partitioning scheme [73] is conceptually one of the closest approach to the novel estimator introduced in Section 3.2, but it is limited in the placement of the bins.

Another common approach is to compute the mutual information using analytical formulas, having estimated $p(X)$, $p(Y)$ and $p(X,Y)$. It is only feasible for few applications with strong a priori on the data distribution, and even if we know the distributions the data is sampled from, only few analytical formulas for the information are known [75]. Instead of being imposed some priors, the density functions can also be estimated via the usual methods using e.g. kernel functions [76]. But, related to the problem of choosing the number of bins, one has to choose the type of kernel and its width, which has shown similar bias [70]. It is also exponentially more complex as the support's dimensions increase, limiting its use for conditional independence testing even with few variable Zs in the conditioning set.

Undoubtedly, the best results on continuous data are obtained with the "KSG" estimator from Kraskov, Stögbaueur and Grassberger [77]. We will also refer to this approach as the k-nn approach, as it employs a k-nearest neighbor estimation of the local entropy. It is based on earlier work by Kozachenko and Leonenko, who first derived an estimate of the entropy based on nearest-neighbor distances [78] :

$$\hat{H}_{\mathrm{KL}}(X) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{N c_{d,p}\rho_{k,i}}{k}\right) + \log(k) - \psi(k) \tag{3.1.34}$$

with $\rho_{k,i}$ the distance from the jth sample to its kth nearest neighbor, c_d the volume of the unit ball in d dimensions and $\psi(.)$ the digamma function. The original authors introduced this formula for a fixed $k = 1$, proving its consistency as N increases, and [79] proved it later for all k. Additionally, Jiao et al. derived an uniform upper bound on its performance proving its near optimality [80], a first for such estimators.

Given this strong estimator, a natural way to get to $\hat{\mathrm{I}}$ is to use the chain rule :

$$\hat{\mathrm{I}}_{3\mathrm{KL}} = \hat{H}_{\mathrm{KL}}(X) + \hat{H}_{\mathrm{KL}}(Y) - \hat{H}_{\mathrm{KL}}(X,Y) \tag{3.1.35}$$

This estimator is also consistent and performs fairly well in practice, but was shown to be uniformly inferior to the KSG estimator in many empirical settings. The KSG estimator is defined as :

$$\hat{\mathrm{I}}_{\mathrm{KSG}}(X;Y) = \psi(k) + \psi(N) - \left\langle \psi(n_{x,i}+1) + \psi(n_{y,i}+1) \right\rangle \tag{3.1.36}$$

with $n_{x,i}$ the number of points within an $\rho_{k,i}$ distance on the X dimension, and $\left\langle \psi(n_x+1) + \psi(n_y+1) \right\rangle$ the average taken on all samples. The ρ_{k_i} distance is usually taken with ℓ_∞ or ℓ^2 norm, see Fig 3.6. Since its introduction, no other estimator seems to be as performant in most settings and it has become the go-to solution to estimate $\hat{\mathrm{I}}$ on continuous data. The particularity behind the KSG estimator is to compare $H(X)$, $H(Y)$ and $H(X,Y)$ locally to estimate the mutual information directly, instead of having to estimate each of the three terms. Recently, Gao et al. revealed why this choice leads to uniformly better results than the $\hat{\mathrm{I}}_{3\mathrm{KL}}$ estimator. They have shown that the better performance stems from a *correlation boosting* effect, the bias of the joint entropy is positively correlated to the biases of the marginal entropies, which partly cancel each other when subtracting via the chain rule [81]. It makes no assumption on either the marginal or joint distributions, and seems to be equitable to all relationships [27]. Somewhat surprisingly, rank-ordering the variables still gives correct estimates (as it should), although it is not clear whether it should be preferred or not.

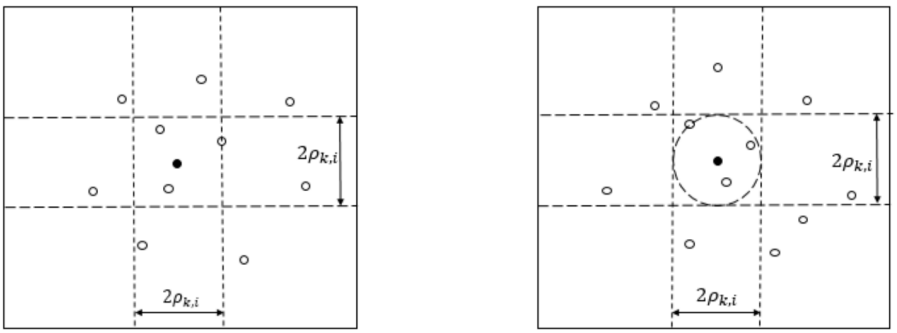

Figure 3.6: Choice of the $\rho_{k,i}$ distance with ℓ_∞ norm (left) or ℓ^2 norm (righ) for the KSG estimator. Figure taken from [81].

It was conveniently adapted to the conditional case, also using a direct formula instead of the chain rule [69, 82] :

$$\hat{\mathrm{I}}_{\mathrm{KSG}}(X;Y|Z) = \psi(k) + \left\langle \psi(n_{z,i}+1) - \psi(n_{xz,i}+1) - \psi(n_{yz,i}+1) \right\rangle \tag{3.1.37}$$

Still, we note a few disadvantages that discourage its use for general constraint-based

algorithms. First, the variance and bias of the estimation are tied to the choice of the parameter k [83]. The original authors themselves suggest a low k $(2-4)$ for good a estimation $\hat{\mathrm{I}}_{\mathrm{KSG}}$, and much larger for independence testing (up to $\simeq N/2$). In general, the trade-off is high variance and low bias for small values of k, and less variance but increased bias for large k [27, 84]. Secondly, as is the case with discrete estimators, the equivalence $\hat{\mathrm{I}}(X;Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$ is not respected, as variance still exists at independence. Crucially, there are currently no results on the distribution of the estimator, either exact nor asymptotically, and there is no easy way to test for independence [64, 83]. Runge proposed to test for conditional independence using a local permutations scheme, which reliably estimates the null distribution but requires significantly more computation [85]. Berrett and Samworth improved slightly on this idea, introducing an independence test based on either simulations when marginal distributions are known, or resampling when they are not [86].

Many other estimators exist, involving ensemble methods [87], copula transformations [88], dependence graphs [89], and even deep neural networks [90]. Overall, the KSG estimator has shown the best performance in the settings that interest us, and is the best understood. It has even been adapted to the mixed case and mixture variables, as we shall see now.

Mixed estimators

Compared to the discrete and the continuous case, relatively little work has been done on estimating mutual information in the mixed case, where X is discrete and Y is continuous, and even less in the case of mixture variables.

Ross et al. extended the KSG estimator to the mixed case, by counting the number of nearest neighbors in the continuous space Y on the subset of samples that share the same discrete value of X. More specifically, for each sample i, the method first finds the distance to the kth nearest neighbor which also share the same discrete value, and counts the number of neighbors within this distance in the full data, noted m. This estimator is given by :

$$\hat{\mathrm{I}}_{\mathrm{Ross}}(X;Y) = \psi(N) + \psi(k) - \langle \psi(N_X) - \psi(m+1) \rangle \tag{3.1.38}$$

with N_X the total number of data points that share the same discrete value on X.

It was then expanded by Gao et al. to mixture distributions by taking the average of the Radon-Nikodym derivative over all samples [51]. The way to estimate this derivative depends on each sample : plug-in estimator when the point is discrete (*i.e.* more than k point share the same value, so $\rho_{i,k} = 0$), and KSG estimator when there is a locally continuous joint density. The intuition behind this procedure is that the Radon-Nikodym derivative is well defined for all cases, and that it recovers either the plug-in estimator, the KSG estimator, or Ross's estimator depending on the local subspace. By then taking the average of all the

derivatives, this gives the value $\hat{\mathrm{I}}(X;Y)$ for any distributions X and Y. It was proven to be consistent, and has shown better results than binning procedures or noisy KSG on mixture variables. It shares however the same lack of significance test as the other k-nn estimators, which makes it less adapted to constraint-based algorithms.

Marx et al. also proposed a mixed estimator based on the Radon-Nikodym derivative and adaptive histogram models for the continuous parts of the mixture variables [91]. Just as our approach introduced in [7], they devised an heuristic to find the optimal discretization according to the Minimum Description Length (MDL) principle [92]. It also comes with easy independence testing with Normalised Maximum Likelihood (NML) correction on discrete data, as introduced in [35] using the factorized NML criteria [93] (which was later redefined by Marx et al., proving asymptotic behavior and consistency [94]). It is well adapted to constraint-based algorithms, however it considers mixture variable in a slightly different way from Gao et al (and [7]).

This difference is best explained through an example. Let (X,Y) be a mixture of one continuous and one discrete distribution. The continuous distribution is a bivariate Gaussian, with mean $\mu = 0$, marginal variance $\sigma = 1$ and correlation ρ. The discrete distribution is two binary variables, with probabilities $p(X = 1,Y = 1) = p(X = -1,Y = -1) = \beta$ and $p(X = 1,Y = -1) = p(X = -1,Y = 1) = \beta$. These two distributions are then mixed with probability p_{con} and p_{dis} respectively. The ground truth as derived by Gao et al. is given by :

$$\begin{aligned}\mathrm{I}(X;Y) = \frac{-p_{con}}{2} \times \log(1-\rho^2) + \frac{\beta}{2} \times \log \frac{\beta/2}{p_{dis}^2} + \frac{(1-\beta)}{2} \times \log \frac{(1-\beta)/2}{p_{dis}^2} \\ -p_{con} \times \log p_{con} - p_{dis} \times \log p_{dis}\end{aligned} \tag{3.1.39}$$

Marx et al. used a different ground truth for this distribution, without the last two terms of the sum, $-p_{con} \times \log p_{con} - p_{dis} \times \log p_{dis}$. In their framework, $X \perp\!\!\!\perp Y$ and $\mathrm{I}(X;Y) = 0$ if and only if $\rho = 0$ and $\beta = 0.5$. It is justified if one considers that the continuous and discrete parts do not share the same space, acting more like separate dimensions of the joint distribution. On the other hand, if we consider that all parts of X and Y share the same euclidean space, some information is "created" from the structure of the joint distribution, given by $-p_{con} \times \log p_{con} - p_{dis} \times \log p_{dis}$ (which equals to $\log 2$ when $p_{con} = p_{dis} = 0.5$). Indeed, even when $\rho = 0$ and $\beta = 0.5$, the distribution $p(x,y)$ is far from $p(x)p(y)$ due to the constraints imposed by sharing the same space (Fig 3.7).

The second view is closer to the master definition of mutual information (Eq 3.1.21) which implies that we can use any partitioning to discretize X and Y, potentially combining discrete and continuous parts in a single bin. This also corresponds to the approach taken to develop our own estimator based on optimal binning of X,Y, introduced in the next section.

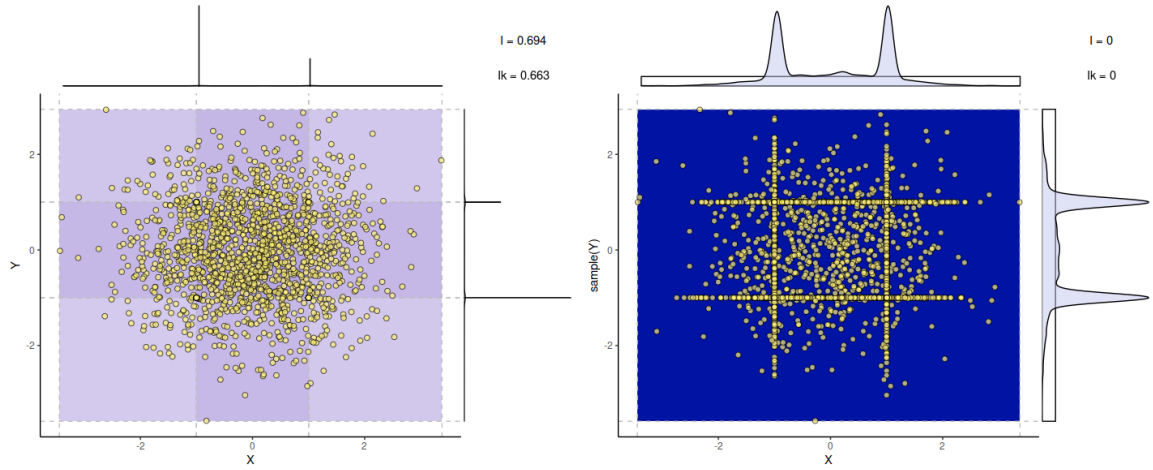

Figure 3.7: Two null hypotheses ($X \perp\!\!\!\perp Y$) of the mixture variable given as example, with $\rho = 0$, $\beta = 0.5$, $p_{con} = p_{dis} = 0.5$. Left : discrete and continuous parts are kept separated, as if in different dimensions. Right : all data points are on the same euclidean space, and the null hypothesis is $p(x,y) = p(x)p(y)$. The joint histogram corresponds to the optimal discretization of [7] for both cases.

3.2 Developing a new general case estimator

So far, we have rationalized the advantages of using information theoretic measures for causal inference, but our review of available methods showed a lack of a good estimator that would work on any type of data and provide a fair assessment of its significativity. In this section, I first give a brief review of the MDL and NML frameworks, and I introduce the method that was published in [7] to adapt MIIC to the general case, showing qualitative and quantitative results as a discretization scheme, an estimator of (conditional) mutual information and a test of (conditional) independence.

3.2.1 MDL-optimal histograms

The *Minimum description length* (MDL) principle is rooted in information theory, it was developed by Jorma Rissanen at the end of the 20th century [92, 95]. The fundamental idea behind it is that there exists a *code* that describes the data in a more succinct way than copying the data itself, and that the smaller the code, the more we have learned about the data. It provides a general and powerful framework for performing model selection given the available data by separating the signal from the noise. Specifically, the best model is the model with the smallest *stochastic complexity*, which is the shortest description length of a given data relative to a model class $\mathcal{M}$. This very abstract concept of minimizing the description length actually found an elegant solution via the *normalized maximum likelihood* (NML) distribution [96, 95], and found success in various applications, from data clustering [97] to image denoising [98]. We now define the NML density and show how it can be used

to find optimal discretizations on finite data, as developed in [99].

Let $x^n = (x_1, ..., x_n)$ be a data sample of n outcomes in the space $\mathcal{X}$, and $\hat{\theta}(x^n)$ its maximum likelihood estimate. The *normalized* maximum likelihood density is defined as :

$$f_{\text{NML}}(x^n|\mathcal{M}) = \frac{f(x^n \mid \hat{\theta}(x^n), \mathcal{M})}{\mathcal{C}_{\mathcal{M}}^n} \tag{3.2.1}$$

where $\mathcal{C}_{\mathcal{M}}^n$ is the universal normalizing constant, and is given by :

$$\mathcal{C}_{\mathcal{M}}^n = \sum_{x^n \in \mathcal{X}^n} f(x^n \mid \hat{\theta}(x^n), \mathcal{M}) \tag{3.2.2}$$

The stochastic complexity of the data x^n, the quantity to be minimized, is defined via the NML density :

$$SC(x^n \mid \mathcal{M}) = -\log f_{\text{NML}(x^n|\mathcal{M})} \tag{3.2.3}$$

$$= -\log f(x^n|\hat{\theta}(x^n), \mathcal{M}) + \log \mathcal{C}_{\mathcal{M}}^n \tag{3.2.4}$$

and $\log \mathcal{C}_{\mathcal{M}}^n$ is the *parametric complexity*. It acts as a normalizing constant, as it is related to the number of essentially different distributions in the model class with regards to x^n [99].

The NML distribution has several important properties. It is the unique solution to the minimax problem of [96], which essentially means that is is the optimal encoding of any observed x^n in the model class. Not only that, but it is also the optimal encoding for any data generating density, even outside the model class [100]. It automatically prevents any overfitting by learning both the model and the number of parameters of the model, using only the data at hand (as opposed to Bayesian priors). In most applications however, computing the NML density is intractable due to the sum (or integral for continuous data) in Eq 3.2.2.

Fortunately it not the case for choosing the cutpoints of a discretization where the model class is equivalent to that of multinomial distributions, for which the normalizing constant has a closed form and can actually be computed in linear time via recursion :

$$\mathcal{C}_n^r = \sum_{l_1+l_2+\cdots+l_r=n} \frac{n!}{l_1!l_2!\cdots l_r!} \prod_{k=1}^{r} \left(\frac{l_k}{n}\right)^{l_k} \tag{3.2.5}$$

$$= \mathcal{C}_n^{r-1} + \frac{n}{r-2}\mathcal{C}_n^{r-2} \tag{3.2.6}$$

From this result, [99] developed a dynamic programming scheme to find the MDL-optimal discretization of a sample, giving the best description possible without overfitting (Fig 3.8). This method essentially gives a solution to the problem of choosing a discretization for the naive estimator $\hat{H}_{\text{Naive}}$ which best describes the features of the sampled distribution, using the most complexity it can justify within the MDL framework. In the next section,

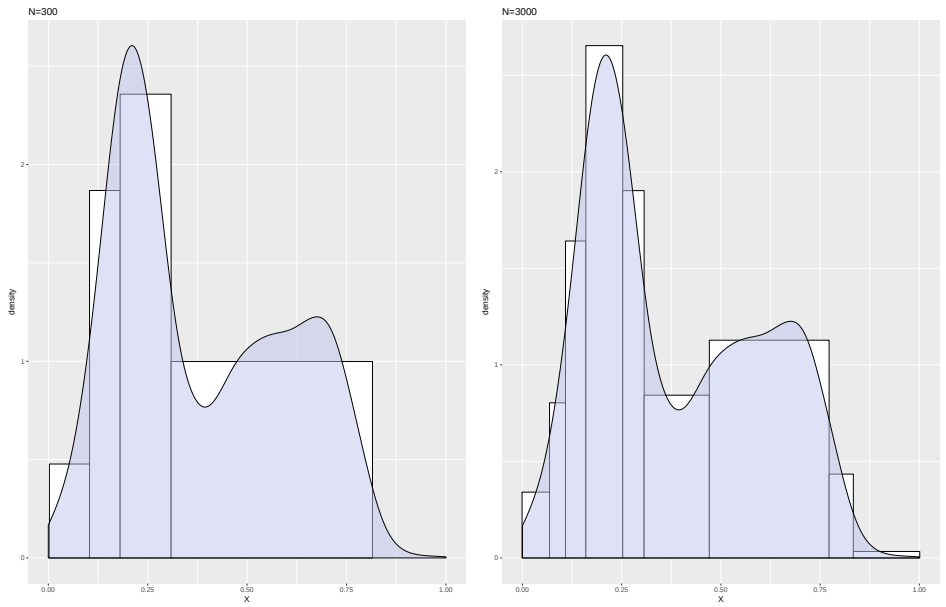

Figure 3.8: MDL-optimal histograms on a multimodal Gaussian distribution, with $N = 300$ samples (left) and $N = 3000$ (right) according to [99].

we will see how it is adapted to find the two-dimensional discretization, which is needed to estimate the interaction between two variables.

3.2.2 Pairwise mutual information estimation through optimal joint discretization

The proposed method starts from the *master* definition of the mutual information (Eq 3.1.21), which we redefine here for convenience. It consists in taking the supremum over all finite partitions, $\mathcal{P}$ and $\mathcal{Q}$, of variables, X and Y [4],

$$\mathrm{I}(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$$

which can be applied to continuous, discrete or mixture variables.

By continuing to refine some initial partitions through the addition of further cut points for continuous variable(s), one finds a monotonically increasing sequence [4], $\mathrm{I}([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$, as shown in Fig. 3.9. In practice, however, Eq. 3.1.21 cannot be used to estimate $\hat{\mathrm{I}}(X;Y)$ from sampled distributions, as the refinement of partitions eventually assigns each of the N different samples into N different bins. This leads to a shift of convergence towards $\log N$ instead of the theoretical limit, $\mathrm{I}(X;Y)$, which requires an infinite amount of data (dotted line in Fig. 3.9).

In [7], we proposed to adapt Eq. 3.1.21 to account for the finite number of samples in

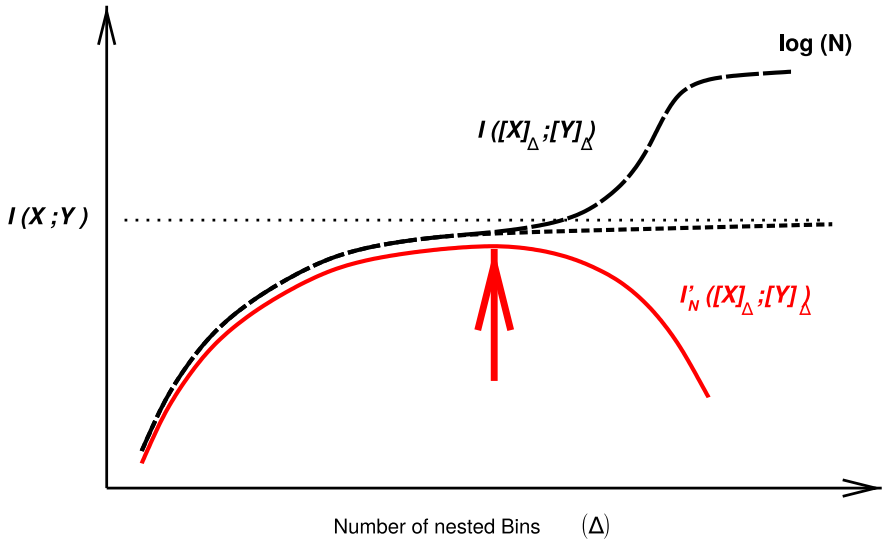

Figure 3.9: Outline of mutual information computation between continuous or mixed-type variables for a finite dataset of N samples. Theoretically, one approaches the true value $\mathrm{I}(X;Y)$ (dotted horizontal line) refining the discretization by adding more bins Δ (dotted curve). On finite sampling however, one inevitably overestimates $\mathrm{I}(X;Y)$ by adding too many bins for the sample size N, up to a maximum $\log N$ (dashed curve). Mutual information is estimated through an optimum partitioning of continuous variable(s) (solid red line and arrow) after introducing a complexity term to account for the finite size of the dataset.

actual datasets,

$$\mathrm{I}'_N(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} \mathrm{I}'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \tag{3.2.7}$$

by introducing a finite size correction to mutual information :

$$\mathrm{I}'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) = \mathrm{I}_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) - k'_{\mathcal{P};\mathcal{Q}}(N)\frac{1}{N} \tag{3.2.8}$$

where $k'_{\mathcal{P};\mathcal{Q}}(N)$ corresponds to a complexity term introduced in [5, 35] to discriminate between variable dependence (for $\mathrm{I}'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) > 0$) and variable independence (for $\mathrm{I}'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \leqslant 0$) given N samples. In the present context of finding an optimum discretization for continuous variables, this complexity term introduces a penalty which grows faster than the spurious information gained in refining bin partitions further, when there is not enough data to support such a refined model (Fig. 3.9). Conceptually, we can also think of this penalty as the uncertainty associated with estimating the frequency $\hat{p}$ of small bins compared to large bins when N is limited. But how do we choose $k'_{\mathcal{P};\mathcal{Q}}(N)$?

For discrete variables, typical complexity terms correspond to the Bayesian Information Criterion (BIC), $k^{\mathrm{BIC}}_{\mathcal{P};\mathcal{Q}}(N) = 1/2(r_x - 1)(r_y - 1)\log N$, where r_x and r_y are the number of bins for X and Y. Within the MDL framework, Roos et al. defined the X- and Y-Normalized

Maximum Likelihood (NML) criteria [93, 35] :

$$k_{\mathcal{P};\mathcal{Q}}^{X-\mathrm{NML}}(N) = \sum_{y}^{r_y} \log \mathcal{C}_{n_y}^{r_x} - \log \mathcal{C}_N^{r_x} \tag{3.2.9}$$

$$k_{\mathcal{P};\mathcal{Q}}^{Y-\mathrm{NML}}(N) = \sum_{x}^{r_x} \log \mathcal{C}_{n_x}^{r_y} - \log \mathcal{C}_N^{r_y} \tag{3.2.10}$$

where $\mathcal{C}_{n_y}^{r_x}$ is the parametric complexity associated with the yth bin of variable Y containing n_y samples, and similarly for $\mathcal{C}_{n_x}^{r_y}$ with the n_x-size bin of variable X in Eq. 3.2.10.

As mentioned, the parametric complexity or normalizing constant $\mathcal{C}_n^r$ is known for the domain of multinomial distributions. It is defined by summing a multinomial likelihood function over all possible partitions of n data points into a maximum of r bins :

$$\mathcal{C}_n^r = \sum_{\ell_1+\ell_2+\cdots+\ell_r=n}^{\ell_k \geqslant 0} \frac{n!}{\ell_1!\ell_2!\cdots\ell_r!} \prod_{k=1}^{r} \left(\frac{\ell_k}{n}\right)^{\ell_k} \tag{3.2.11}$$

which can in fact be computed recursively in linear-time [101]. For large n and r, inherent to large datasets with continuous or mixed-type variables, we found that $\mathcal{C}_n^r$ computation can be made numerically stable by implementing the recursion on parametric complexity ratios $\mathcal{D}_n^r = \mathcal{C}_n^r / \mathcal{C}_n^{r-1}$ rather than the parametric complexities themselves :

$$\mathcal{D}_n^r = 1 + \frac{n}{(r-2)\mathcal{D}_n^{r-1}} \tag{3.2.12}$$

$$\log \mathcal{C}_n^r = \sum_{k=2}^{r} \log \mathcal{D}_n^k \tag{3.2.13}$$

for $r \geqslant 3$, with $\mathcal{C}_n^1 = 1$ and $\mathcal{C}_n^2 = \mathcal{D}_n^2$, which can be computed directly with the general formula, Eq. 3.2.11, for $r = 2$,

$$\mathcal{C}_n^2 = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \tag{3.2.14}$$

or its Szpankowski approximation for large n (needed for $n > 1000$ in practice) [102, 103, 104],

$$\mathcal{C}_n^2 = \sqrt{\frac{n\pi}{2}} \left(1 + \frac{2}{3}\sqrt{\frac{2}{n\pi}} + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{3.2.15}$$

$$\simeq \sqrt{\frac{n\pi}{2}} \exp\left(\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}\right) \tag{3.2.16}$$

For continuous variables, however, the variable categories are not given *a priori* and need to be specified and thus encoded in the model complexity within the frame of the Minimum Description Length (MDL) principle [99]. In absence of priors for any specific partition with r bins, the model index should be encoded with a uniform distribution over all partitions with the same number of bins [99]. As there are $\binom{N-1}{r_x-1}$ ways to choose r_x-1 out of $N-1$ possible cut points, corresponding to a codelength of $\log\binom{N-1}{r_x-1}$ for a continuous variable X (and similarly for Y if it is continuous), the model complexity associated with the partitioning of continuous or mixed-type variables becomes,

$$k'_{\mathcal{P};\mathcal{Q}}(N) = k_{\mathcal{P};\mathcal{Q}}(N) + \log\binom{N-1}{r_x-1} + \log\binom{N-1}{r_y-1} \tag{3.2.17}$$

with $\log\binom{N-1}{r-1} = (r-1)\,C_{N,r}$, where $C_{N,r}$ corresponds to the encoding cost associated to each of the $r-1$ cut points with $r = r_x$ or r_y.

While finding the supremum of $\mathrm{I}'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$ over *all* possible partitions $\mathcal{P}$ and $\mathcal{Q}$ according to Eq. 3.2.7 seems intractable, it can be computed rather efficiently in practice.

The proposed approach is inspired by the computation of the MDL-optimal histogram for a single continuous variable of [99], which can be done exactly in $O(N^2 \times k)$ steps (with k the maximum number of bins). As the approach cannot be generalized to more than one variable, we implemented a local optimization heuristics, which finds the optimum cut points for a continuous variable X, maximizing its corrected information with a discrete variable $\mathrm{I}'(X;[Y])$. When both X and Y are continuous, we iteratively fix X and Y and compute $\mathrm{I}'([X];[Y])$ until a convergence is reached in the limit cycle, as will be detailed below.

In practice, for two variables variables we start from an initial (or optimized) $[Y]$ partition with r_y bins of various sizes and an estimate of the number of $[X]$ bins, $\hat{r_x}$ (before discretizing X). The sample-scaled mutual information with finite size correction, *i.e.*, $n\,\mathrm{I}'_n(X;Y)$, is then optimized iteratively for $n = 1,\cdots,N$ samples, over all X partitions, through the following $\mathcal{O}(N^2)$ dynamic programming scheme, using Eq. 3.2.9 as parametric complexity,

$$\begin{aligned} n\,\mathrm{I}'_n(X;[Y]) = \max_{0\leqslant j<n} \Big[j\,\mathrm{I}'_j(X;[Y]) + \sum_y^{r_y} n_{xy}\log n_{xy} - n_x \log n_x \\ - \log \mathcal{C}_{n_x}^{r_y} - C_{N,\hat{r_x}} \Big] \end{aligned} \tag{3.2.18}$$

where the last added bin on X, including the jth to nth samples distributed over the r_y bins of $[Y]$ (with $\sum_y^{r_y} n_{xy} = n_x$), comes with an independent mutual information contribution, $\sum_y^{r_y} n_{xy}\log n_{xy} - n_x \log n_x$, a parametric complexity, $\log \mathcal{C}_{n_x}^{r_y}$, and encoding cost, $C_{N,\hat{r_x}}$. The initial condition for $j = 0$ in (3.2.18) is set by convention to include all terms invariant under X-partitioning, *i.e.*, $-\sum_y^{r_y} n_y \log(n_y/N) + \log \mathcal{C}_N^{r_y} - (r_y-1)C_{N,r_y} + C_{N,\hat{r_x}}$. Equation 3.2.18 is illustrated by the pseudocode of Alg 4.

Algorithm 4 $Opt(\mathrm{I}(X;[Y]):$ MDL-Optimal discretization of X maximizing $\mathrm{I}(X;[Y])$

Require: Ranks of X, $[Y]$, coarse level c, $\hat{r_x}$
Define possible cutpoints C from N and c

for $j = 1$ to N in C **do**
  $\mathrm{I}'[j] \leftarrow \mathrm{I}'_{[0,j]}$
  for $k = 1$ to j in C **do**
    $\mathrm{I}'_{\text{new bin}} \leftarrow \mathrm{I}'[k] + \mathrm{I}'_{[k,j]}$
    if $\mathrm{I}'_{\text{new bin}} > \mathrm{I}'[j]$ **then**
      $\mathrm{I}'[j] \leftarrow \mathrm{I}'_{\text{new bin}}$
      Save $Cuts[j] \leftarrow k$
    end if
  end for
end for

Reconstruct $[X]$ from trace of best cutpoints, starting from $Cuts[N]$
return $[X]$

In this notation, the array $\mathrm{I}'[]$ saves the values of $\mathrm{I}'_j(X;[Y])$, the partial information taking first j samples corrected the full complexity term k including the combinatorial approximation (Eq 3.2.17). We can significantly speed up the computation at little cost by doing a coarse search of the partitions on C possible cutpoints, instead of all N samples. This allows the algorithm to run in $O(C^2)$ instead of $O(n^2)$, with C being typically a factor of $N^{1/3}$. Finally, $Cuts[j]$ corresponds to the location of the last cutpoint giving the best $\mathrm{I}'[j]$. The optimal partition of X can be retraced by following each cutpoint starting from $Cuts[N]$. Note the special case of independence, when $X \perp\!\!\!\perp [Y]$ no multi-bin partitioning creates a positive I', *i.e.* no information greater than its associated complexity cost can be found. In this case, the output is a single bin from 0 to N ($Cuts[N] = 0$).

Then, adopting this optimized partition for X, one can apply the same dynamic programming scheme for Y using Eq. 3.2.10 as parametric complexity and iterate the optimization of X and Y partitions until a stable two-state limit circle is reached. In practice, we set the initial partitioning over X and Y by testing equal-freq discretizations with $k = 2$ to $\lceil N^{1/3} \rceil$ bins and choosing the one which gives the highest $I'_N([X]^k_{ef}; Y^k_{ef})$. We found that while the convergence speed of the iterative dynamic programming is largely independent of these initial conditions, this scheme does improve it slightly. This leads after only a few iterations to a good estimate of mutual information (averaged over limit circle). The iterative process to compute $\mathrm{I}(X;Y)$ is shown in Alg 5.

Where r_x and r_y are the number of levels of $[X]$ and $[Y]$. Note that $[X]$ and $[Y]$ are not updated straight after the call of $Opt()$, to make the process symmetrical between X and Y ($\hat{\mathrm{I}}'(X;Y) = \hat{\mathrm{I}}'(Y;X)$).

We will now analyze this estimator in empirical situations, first qualitatively and then by

Algorithm 5 $\hat{\mathrm{I}}'(X;Y)$ heuristic

Require: Ranks of X, Y, coarse level c

$\mathrm{I}'_{init} = 0$

for $k = 2$ to $\lceil N^{1/3} \rceil$ **do**

if $\mathrm{I}'([X]^k_{ef};[Y]^k_{ef}) > \mathrm{I}'_{init}$ **then**

$\mathrm{I}'_{init} \leftarrow \mathrm{I}'([X]^k_{ef};[Y]^k_{ef})$

$[X] \leftarrow [X]^k_{ef}$, $[Y] \leftarrow [Y]^k_{ef}$

end if

end for

repeat

$[X]_{new} \leftarrow Opt(\mathrm{I}(X;[Y]),c,r_x)$

$[Y]_{new} \leftarrow Opt(\mathrm{I}(Y;[X]),c,r_y)$

Update $[X] \leftarrow [X]_{new}$, $[Y] \leftarrow [Y]_{new}$

$\mathrm{I}' \leftarrow \mathrm{I}'([X],[Y])$

until Max iteration reached or limit cycle convergence

return I', $[X]$, $[Y]$

quantitatively comparing $\hat{\mathrm{I}}'(Y;X)$ to other estimators in the discrete and in the mixed case. Perhaps the most noticeable result of this approach is that the optimal discretization $[X]$ of any variable depends on the joint distribution X,Y, no $[X]$ can be MDL-optimal with regards to all joint distributions (Fig 3.10). This implies that we need to run Alg 5 for each pair of variable to estimate $\hat{\mathrm{I}}'(X;Y)$ correctly, we cannot reuse the same cutpoints. Importantly, even if all variables are jointly Gaussian, the number of bins still depends on the amount of information, scaling monotonically with the strength of the interaction (Fig 3.9).

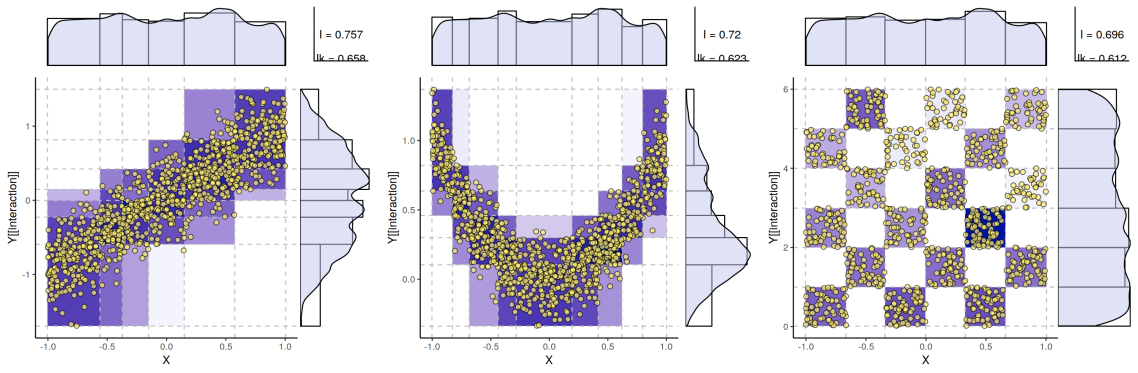

Figure 3.10: Optimal discretization of three joint distributions with the same marginal X, as found by maximizing $\hat{\mathrm{I}}'(X;Y)$.

This concept of context-dependent discretization perhaps seems fundamentally incompatible with Bayesian networks, for which each node must have a marginal probability distribution defined independently of the rest of the network. In our case, the optimal discretizations $[X]$ and $[Y]$ must be considered in pairs, they inform us of the *edge* between X

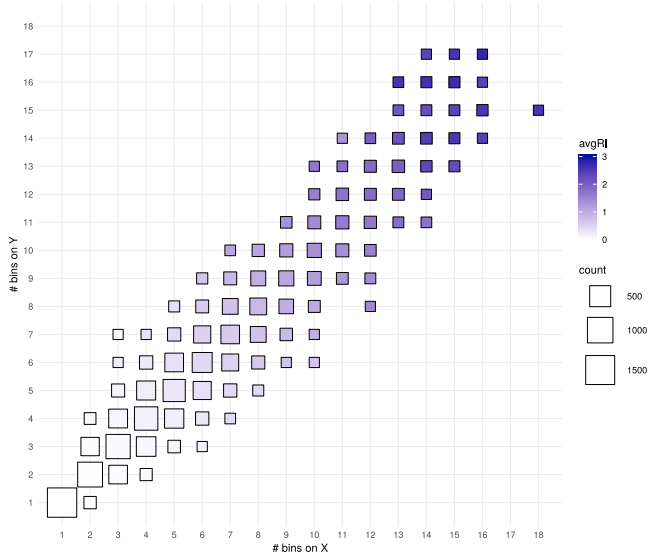

Figure 3.11: Adaptive information-maximizing partitions depending on interaction strength. 10,000 Gaussian bivairate distributions with $N = 1,000$ samples were generated with uniformly distributed correlation coefficients ρ in $[-1,1]$, and discretized using Alg 5. The real mutual information (RI, shades of blue) of Gaussian bivariate distributions can be computed directly with Eq 3.1.16.

and Y, and not of the *nodes* X and Y. Alternatively, one can think of the optimal discretization scheme only as a proxy to measure $\hat{\mathrm{I}}'(X;Y)$ (of which $[X]$ and $[Y]$ are by-products), which is indeed linked to the edge $X-Y$.

There are other discretization schemes for mixed Bayesian networks, for example the work done by Neil et al. on dynamic discretization [105, 106]. In this setting, continuous nodes are discretized so as to give the best *inference*, *i.e.* the best distribution $P(v)$ as reconstructed from the Bayesian network parameters. From the inference point of view, one needs to find $[X]$ not only in relation to a single variable Y, but to all of its neighbors (as well as all of the other parents of its children, to take into account interaction effects). To re-use the Information Bottleneck terminology, if one is looking for the best inference, one wants to find an encoding $[X]$ of a continuous variable X that maximizes the information between X and all of its neighbors : $\mathrm{I}([X];Adj(X))$. On the other hand, for constraint-based graph discovery we want to perform (conditional) independence testing for each edge $X-Y$, which only requires to look at the (conditional) interaction between two nodes (the conditional estimator is introduced in the next section).

Next, we assess how good is the estimation of $\hat{\mathrm{I}}'(Y;X)$ on known distributions. On bivariate Gaussian distributions with correlation ρ ranging from 0.01 to 0.9, our estimator is competitive with the KSG estimator (as implemented in JIDT [107]). We note a particularly desirable property of the miic estimation : its error and variance tend to zero as the signal

disappears ($\rho \to 0$) and as the complexity cost is greater than any information coming from the joint discretization of the data. This results in few false positives when doing the graph reconstruction while still having decent power.

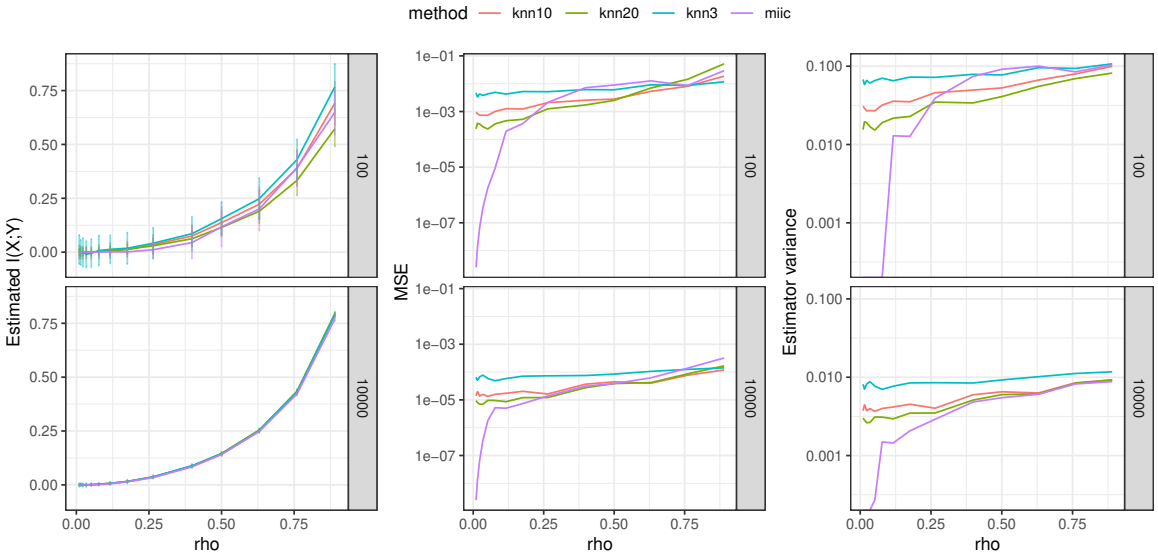

Figure 3.12: Mutual information estimation for 100 Gaussian bivariate distributions. The mean squared error (center graphs) was calculated thanks to the analytical result of the mutual information of the bivariate Gaussian (Eq 3.1.16). The standard deviation of each estimator over the 100 replications was also plotted against the correlation coefficient (right).

We also assessed its performance on the mixed case, by conducting the numerical experiments proposed in [51]. Our approach fared similarly or better compared to a naive equal-frequency discretization with $N^{1/3}$ bins, a kernel estimator and a noisy KSG estimator, as well $\hat{\mathrm{I}}_{\mathrm{Ross}}$ [71] and $\hat{\mathrm{I}}_{\mathrm{Gao}}$ [51]. Particularly, our estimator gives the best estimation for the mixture distribution of Fig 3.7. It converges at the ground truth value given by Eq 3.1.39, in accordance with the master definition of mutual information. For details on benchmark settings and other results, see Supplementary materials of [7].

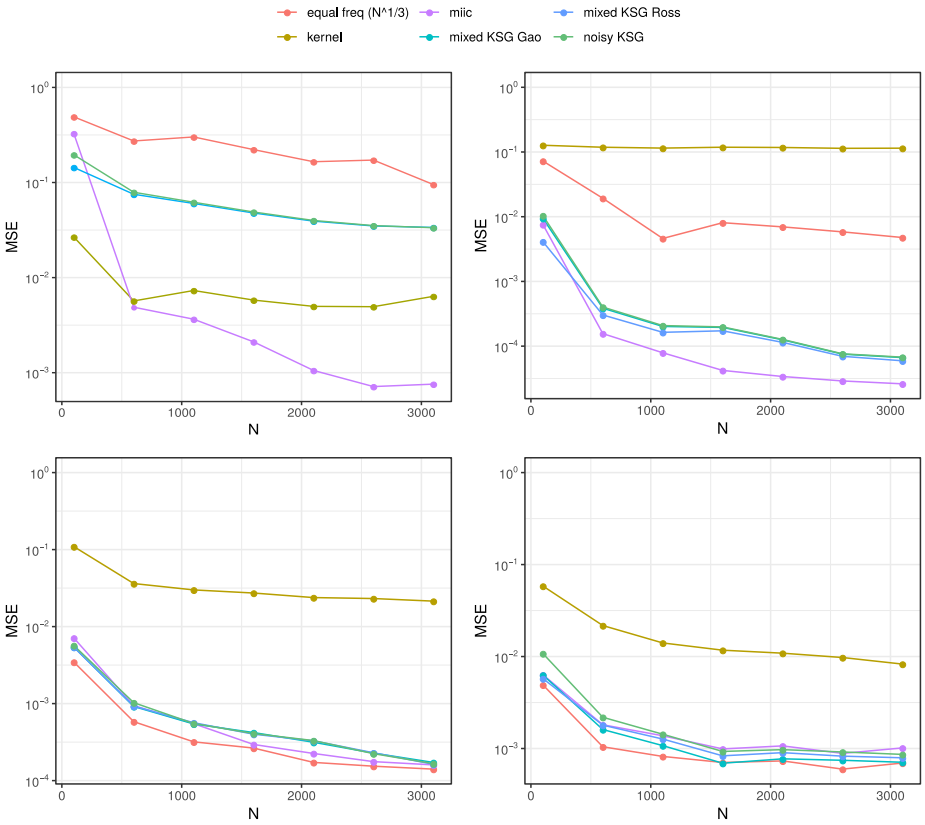

Figure 3.13: Mutual information estimation of mixed variables. Experiment set-ups and analytical values for the mutual information were taken fom [51] and 50 runs were performed for each sample size N. From left to right, top to bottom, the simulations are devised after experiment I, experiment II, experiment IV with $p = 0$ and experiment IV with $p = 0.15$, from [51]. The top left experiment corresponds to the distribution of Fig 3.7 with $\rho = 0.9$ and $\beta = 0.9$ and $p_{con} = p_{dis} = 0.5$

3.2.3 Conditional case

This optimization scheme, Alg 4, and its iterative dynamic programming computation, Alg 5, can also be adapted to compute mutual information involving joined variables, such as $\mathrm{I}'_N(X;\{U_i\})$, with corresponding finite size corrections and cut point encoding costs extended from Eqs. 3.2.8−3.2.17. Similarly, the approach can compute the conditional mutual information $\mathrm{I}'_N(X;Y|\{U_i\})$, involving continuous, discrete or mixture variables. Remark that we do not want to maximize $\mathrm{I}'(X;Y|\{U_i\})$ directly, as two dependent variable $X \not\perp\!\!\!\perp Y$ that are conditionally independent $X \perp\!\!\!\perp Y|\{U_i\}$ can always have positive conditional information $\mathrm{I}(X;Y|[U_i])$ if the $[U_i]$ encoding is chosen so that $\mathrm{I}(X;[U_i])$ or $\mathrm{I}(Y;[U_i])$ is null (*e.g.* imposing one single bin for all U). Instead, we want an estimation that converges towards 0 for conditional independence, and a positive value otherwise.

To this end, we can define $\hat{\mathrm{I}}'_N(X;Y|\{U_i\})$ using the chain rule 3.1.19, as the *difference* between maximized mutual information terms involving either $\{Y,\{A_i\}\}$ and $\{A_i\}$ (Eq. 3.2.19) or $\{X,\{A_i\}\}$ and $\{A_i\}$ (Eq. 3.2.20) as joined variables,

$$\hat{\mathrm{I}}'_N(X;Y|\{U_i\}) = \hat{\mathrm{I}}'_N(X;Y,\{U_i\}) - \hat{\mathrm{I}}'_N(X;\{U_i\}) \tag{3.2.19}$$

$$= \hat{\mathrm{I}}'_N(Y;X,\{U_i\}) - \hat{\mathrm{I}}'_N(Y;\{U_i\}) \tag{3.2.20}$$

Starting from an initial (or optimized) partition $[Y]$, each term of Eq. 3.2.19 is optimized with respect to X and $\{U_i\}$ partitions using Eq. 3.2.9 as parametric complexity extended to multivariate categories, $n_{x,\{u_i\}}$ and $n_{\{u_i\}}$. Then, in turn, each term of Eq. 3.2.20 is optimized with respect to Y and $\{U_i\}$ partitions using Eq. 3.2.10 as parametric complexity extended to multivariate categories, $n_{y,\{u_i\}}$ and $n_{\{u_i\}}$. Note, in particular, that $\{U_i\}$ partitions are optimized *separately* for each of the four terms in Eqs. 3.2.19 & 3.2.20, before taking their differences, as these optimized $\{U_i\}$ partitions might be different in general. This process is detailed by Alg 6.

The $[U]$ optimization routine is a small loop of 3 iterations maximizing the relevant information, where one U_i is discretized while the rest are fixed. As [91] remarked in their own benchmarks, treating each U_i independently gives MIIC a unique advantage compared to other methods as it discards non-informative U_is by discretizing them in a single bin, effectively removing one dimension. This is particularly interesting for constraint-based learning, as it implies that adding irrelevant U_is to the conditioning set does not change the result of the conditional test. In practice, it can sometimes lead to situations where I'_1 or I'_2 is negative because the $[U]$ optimization gets stuck in a local optimum. This can be fixed by re-using the corresponding $[U]$ cutpoints for the next iteration, ensuring that the differences $\mathrm{I}'([X],[Y,U]) - \mathrm{I}'([X],[U])$ and $\mathrm{I}'([Y],[X,U]) - \mathrm{I}'([Y],[U])$ are positive. This recycling scheme is described in Alg 7.

Algorithm 6 $\hat{\mathrm{I}}'(X;Y|U)$ heuristic

Require: Ranks of X, Y, U, coarse level c

$\mathrm{I}'_{init} = 0$
for $k = 2$ to $init\_bins_{max}$ **do**
  if $\mathrm{I}'([X]^k_{ef};[Y,U]^k_{ef}) > \mathrm{I}'_{init}$ **then**
    $\mathrm{I}'_{init} \leftarrow \mathrm{I}'([X]^k_{ef};[Y,U]^k_{ef}) + \mathrm{I}'([Y]^k_{ef};[X,U]^k_{ef})$
    $[X] \leftarrow [X]^k_{ef}$, $[Y] \leftarrow [Y]^k_{ef}$
  end if
end for

repeat
  $[U]$ optimization on $\mathrm{I}'([Y];[X],U)$
  Compute and save $\mathrm{I}'([Y];[X,U])$
  $[X]_{new} \leftarrow Opt(\mathrm{I}'([Y],X[U]),c)$

  $[U]$ optimization on $\mathrm{I}([X];[Y],U)$
  Compute and save $\mathrm{I}'([X];[Y,U])$
  $[X]_{new} \leftarrow Opt(\mathrm{I}'([X],Y[U]),c)$

  $[U]$ optimization on $\mathrm{I}'([X];U)$
  Compute and save $\mathrm{I}'([X];[U])$

  $[U]$ optimization on $\mathrm{I}'([Y];U)$
  Compute and save $\mathrm{I}'([Y];[U])$

  Update $[X] \leftarrow [X]_{new}$, $[Y] \leftarrow [Y]_{new}$
  $\mathrm{I}'_1 \leftarrow \mathrm{I}'([X],[Y,U]) - \mathrm{I}'([X],[U])$
  $\mathrm{I}'_2 \leftarrow \mathrm{I}'([Y],[X,U]) - \mathrm{I}'([Y],[U])$
  $\mathrm{I}' \leftarrow 0.5(\mathrm{I}'_1 + \mathrm{I}'_2)$
until Max iteration reached or limit cycle convergence on I'

return I', $[X]$, $[Y]$

To benchmark the conditional estimator, four-dimensional normal distributions $P(X,Y,Z_1,Z_2)$ were sampled for $N = 100$ and $10,000$ samples 100 times for each correlation coefficient $\rho = \rho_{XY}$ between 0.05 and 0.95. The other pairwise correlation coefficients are fixed as $\rho_{XZ_1} = \rho_{XZ_2} = \rho_{YZ_1} = \rho_{YZ_2} = \lambda = 0.7$ and $\rho_{Z_1Z_2} = 0.9$. The conditional mutual information $\mathrm{I}(X;Y|Z_1,Z_2)$ was then estimated using the proposed optimum partitioning scheme as well as with k-nn conditional information estimates as in Fig 3.12. In this experiment, ρ values closed to zero, mimick "V-structures" as they correspond to pairwise independence but conditional dependence; by constrast $\rho = 2\lambda^2/(1+\rho_{Z_1Z_2}) \simeq 0.5158$ corresponds to conditional independence, while $\rho > 0.5158$ implies that X and Y share more information than the indirect flow through Z_1 and Z_2. The analytical value of the conditional mutual information is derived as follows : given the 4×4 covariance matrix $\mathbf{\Sigma}$ and its four 2×2 partitions $\mathbf{\Sigma}_{ij}$, we first compute the conditional covariance matrix $\overline{\mathbf{\Sigma}} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ where $\mathbf{\Sigma}_{22}^{-1}$ is the generalized inverse of $\mathbf{\Sigma}_{22}$. The partial correlation between X and Y is obtained as $\rho_{XY\cdot Z_1Z_2} = \overline{\mathbf{\Sigma}}_{12}/\sqrt{\overline{\mathbf{\Sigma}}_{11} * \overline{\mathbf{\Sigma}}_{22}}$,

Algorithm 7 $\hat{\mathrm{I}}'(X;Y|U)$ heuristic, re-using cutpoints

Require: Ranks of X, Y, U, coarse level c

Initialize $[X]$, $[Y]$ with best equal freq, as Alg 6

$Reuse\_X\_cuts \leftarrow$ false
$Reuse\_Y\_cuts \leftarrow$ false
repeat
  if $Reuse\_Y\_cuts$ **then**
    $[U] \leftarrow [U]_Y$
  else
    $[U]$ optimization on $\mathrm{I}'([Y];[X],U)$
  end if
  Compute and save $\mathrm{I}'([Y];[X,U])$
  $[X]_{new} \leftarrow Opt(\mathrm{I}'([Y],X[U]),c)$
  if $Reuse\_X\_cuts$ **then**
    $[U] \leftarrow [U]_X$
  else
    $[U]$ optimization on $\mathrm{I}'([X];[Y],U)$
  end if
  Compute and save $\mathrm{I}'([X];[Y,U])$
  $[Y]_{new} \leftarrow Opt(\mathrm{I}'([X],Y[U]),c)$
  if $Reuse\_X\_cuts$ **then**
    $[U] \leftarrow [U]_X$
  else
    $[U]$ optimization on $\mathrm{I}'([X];U)$
    Save $[U]_X$ cutpoints $\leftarrow [U]$
  end if
  Compute and save $\mathrm{I}'([X];[U])$
  if $Reuse\_Y\_cuts$ **then**
    $[U] \leftarrow [U]_Y$
  else
    $[U]$ optimization on $\mathrm{I}'([Y];U)$
    Save $[U]_Y$ cutpoints $\leftarrow [U]$
  end if
  Compute and save $\mathrm{I}'([Y];[U])$
  Update $[X] \leftarrow [X]_{new}$, $[Y] \leftarrow [Y]_{new}$
  $\mathrm{I}'_1 \leftarrow \mathrm{I}'([X],[Y,U]) - \mathrm{I}'([X],[U])$
  $\mathrm{I}'_2 \leftarrow \mathrm{I}'([Y],[X,U]) - \mathrm{I}'([Y],[U])$
  Assign $Reuse\_X\_cuts \leftarrow (\mathrm{I}'_1 < 0)$, $Reuse\_Y\_cuts \leftarrow (\mathrm{I}'_2 < 0)$
  $\mathrm{I}' \leftarrow 0.5(\mathrm{I}'_1 + \mathrm{I}'_2)$
until Max iteration reached or limit cycle convergence on I'
return I', $[X]$, $[Y]$

and the analytical conditional mutual information for a multivariate normal distribution is given by $I(X;Y|Z_1,Z_2) = -\log(1-\rho^2_{XY\cdot Z_1Z_2})/2$. The results, shown in Figure 3.14, suggest

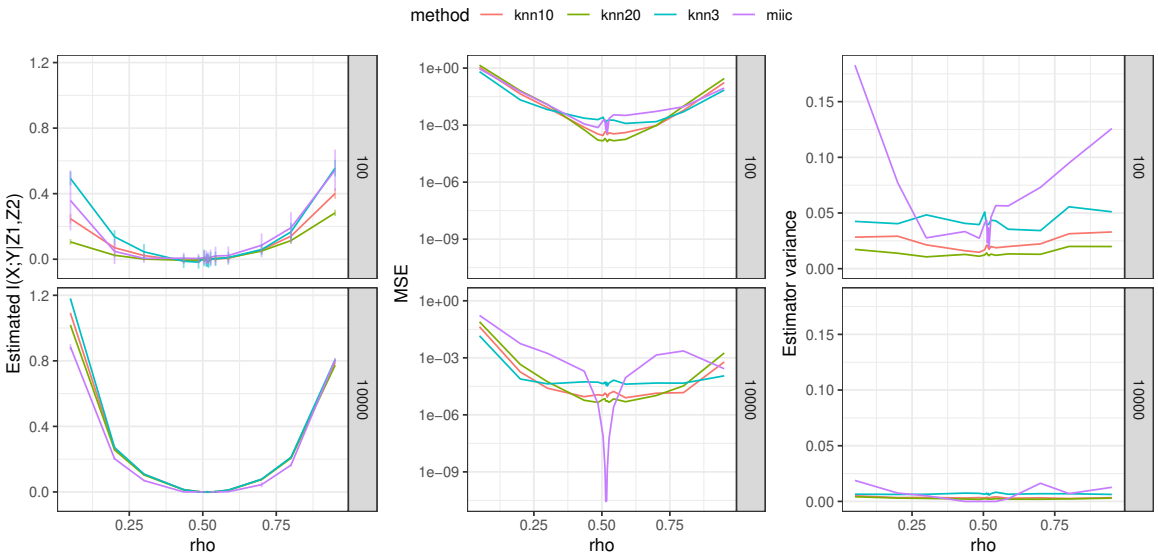

Figure 3.14: Conditional mutual information estimation for multivariate Gaussian distributions Four-dimensional normal distributions $P(X,Y,Z_1,Z_2)$ were sampled for $N = 100$ and $10,000$ samples 100 times for each correlation coefficient $\rho = \rho_{XY}$, chosen between 0.05 and 0.95.

that our estimator is also adequate to measure $\hat{\mathrm{I}}(X;Y|\{U_i\})$, even though it is a significantly harder problem (even for the KSG estimator). As is the case for the pairwise information, it also seems to converge towards zero at the independence regime, contrary to the k-nn approaches that always give noisy estimates.

Finally, we tested the sensitivity and power of our estimator to detect (conditional) independence. We reproduced the tests for mixed conditional independence test by [50] based around the "Local Causal Discovery" algorithm [108]. In the original article, independences tests are either frequentist or bayesian, and are compared using different detection thresholds to compute the ROC curves and AUCs. Our estimator $I'_N(X;Y)$ cannot be readily compared in this way since it is unbounded and it behaves the opposite way of these other tests (dependence implies a large positive value, independence gives a null estimation). For many estimators one can always get an "empirical p-value" without knowing the standard asymptotic distribution by running permutations on the observed data. In our case however, it would not be efficient as the optimal discretization for shuffled data without information is one single bin, and $I'_N(X;Y)$ is strictly 0. Instead, to obtain a value between $]0,1]$ that behaves the same way as the other tests, we computed the following :

$$I'_{pval}(X;Y) = 1 - \frac{I'_N(X;Y)}{\min(I'_N(X;X), I'_N(Y;Y))} \tag{3.2.21}$$

$$I'_{pval}(X;Y|Z) = 1 - \frac{I'_N(X;Y|Z)}{\min(I'_N(X;X), I'_N(Y;Y))} \tag{3.2.22}$$

Where $\min(I'_N(X;X), I'_N(Y;Y))$ can be thought of as the maximum value $I'_N(X;Y)$ or $I'_N(X;Y|Z)$ can have in this setting. We can then compare I'_{pval} with different marginals X, Y and Z, and compute ROC curves and the area under them by setting different thresholds in $]0,1]$. The results show that our proposed estimator has the best overall AUC when combining the three independence tests $C \not\perp\!\!\!\perp X$, $X \not\perp\!\!\!\perp Y$ and $C \perp\!\!\!\perp Y|X$ on mixed data (Fig 3.15). It means that even though the estimator essentially filters out the very weak interactions by setting 1-bin discretization, we are able to compare and rank the estimates $\hat{I}$ better than any other test in these settings. This is the closest experiment to causal graph inference, which essentially consists of serial (conditional) independence tests for the skeleton discovery. For details of the different simulations used to benchmark independence testing, I refer the reader to the original study [50].

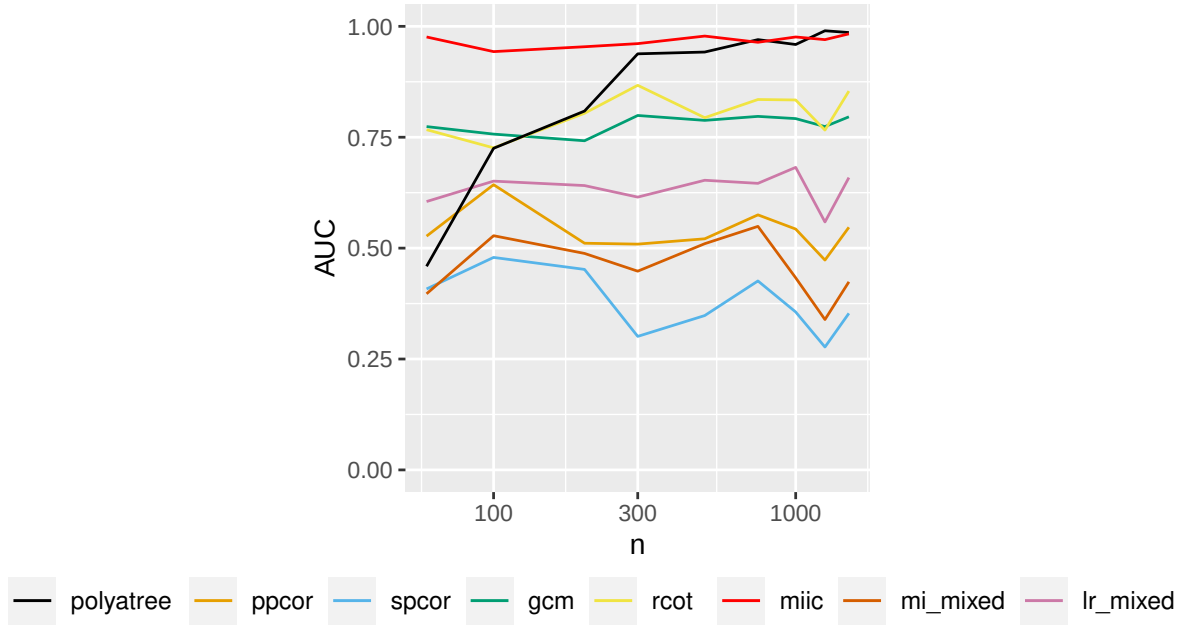

Figure 3.15: Conditional independence tests on mixed variables. Mean Area Under the Curve of ROC curves from 200 rounds of simulation at each sample size n for the LCD triple [50]. The triple is scored according to a combination of three p-values for three independence tests : $C \not\perp\!\!\!\perp X$, $X \not\perp\!\!\!\perp Y$ and $C \perp\!\!\!\perp Y|X$, and is given a true 'positive' label if the data is simulated according to the relationship $C \to X \to Y$, 'negative' otherwise.

3.3 Publication in PLoS Computation Biology

Our publication in the Public Library of Science Computational Biology journal [7] introduced this joint optimal discretization algorithm, showing more examples of discretizations and comparisons with other methods.

Importantly, it evaluated the performance of MIIC using this estimator on continuous and mixed data against other state of the art causal discovery approaches. It was shown to have the best overall performance even when testing over the full range of parameters of the other methods (whereas MIIC stays parameter free). Maybe surprisingly, it was even shown to outperform CAM [38], which makes explicit assumptions that give it an edge in a simulations setting, and kPC [33] based on the HSIC, which is known to be one of the most powerful methods for non-parametric conditional independence testing. On mixed datasets, it fared better than either CausalMGM [109] and MXM [110] (also on their full range of parameters), the only two known methods that deal with mixed data at the time of writing the article.

It also presented and analyzed the network inferred by MIIC on a mixed dataset of medical records of elderly patients, which will be introduced in Section 5.1.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Learning clinical networks from medical records based on information estimates in mixed-type data

Vincent Cabeli[1,2☯], **Louis Verny**[1,2☯], **Nadir Sella**[1,2,3☯], **Guido Uguzzoni**[1,2], **Marc Verny**[2,4]*, **Hervé Isambert**[1,2]*

1 Institut Curie, PSL Research University, CNRS, UMR168, 26 rue d'Ulm, 75005 Paris, France, **2** Sorbonne Université, 4, place Jussieu, 75005 Paris, France, **3** LIMICS, UMRS 1142, 15 rue de l'école de médecine, 75006 Paris, France, **4** Hôpital La Pitié-Salpêtrière, 47-83 boulevard de l'Hôpital, 75013 Paris, France

☯ These authors contributed equally to this work.
* marc.verny@aphp.fr (MV); herve.isambert@curie.fr (HI)

OPEN ACCESS

Citation: Cabeli V, Verny L, Sella N, Uguzzoni G, Verny M, Isambert H (2020) Learning clinical networks from medical records based on information estimates in mixed-type data. PLoS Comput Biol 16(5): e1007866. https://doi.org/10.1371/journal.pcbi.1007866

Editor: Sushmita Roy, University of Wisconsin, Madison, UNITED STATES

Received: September 9, 2019

Accepted: April 10, 2020

Published: May 18, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1007866

Data Availability Statement: All relevant data, properly de-identified to guarantee the anonymity of patients, is provided as a supplementary table (S1 Table).

Abstract

The precise diagnostics of complex diseases require to integrate a large amount of information from heterogeneous clinical and biomedical data, whose direct and indirect interdependences are notoriously difficult to assess. To this end, we propose an efficient computational approach to simultaneously compute and assess the significance of multivariate information between any combination of mixed-type (continuous/categorical) variables. The method is then used to uncover direct, indirect and possibly causal relationships between mixed-type data from medical records, by extending a recent machine learning method to reconstruct graphical models beyond simple categorical datasets. The method is shown to outperform existing tools on benchmark mixed-type datasets, before being applied to analyze the medical records of eldery patients with cognitive disorders from La Pitié-Salpêtrière Hospital, Paris. The resulting clinical network visually captures the global interdependences in these medical records and some facets of clinical diagnosis practice, without specific hypothesis nor prior knowledge on any clinically relevant information. In particular, it provides some physiological insights linking the consequence of cerebrovascular accidents to the atrophy of important brain structures associated to cognitive impairment.

Author summary

We developed a machine learning approach to analyze medical records and help clinicians visualize the direct and indirect interrelations between clinical examinations and the variety of syndromes implicated in complex diseases. The reconstruction of such clinical networks is illustrated on the spectrum of cognitive disorders, originating from either neurodegenerative, cerebrovascular or psychiatric dementias. This global network analysis is also shown to uncover novel direct associations and possible cause-effect relationships between clinically relevant information, such as medical examinations, diagnoses, treatments and personal data from patients' medical records.

PLOS COMPUTATIONAL BIOLOGY

Funding: HI received funding from IRIS data science program of PSL university, DIM program from Region Ile-de-France and Labex celtisphybio. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The precise diagnostics of neurological disorders require to integrate a large amount of information from a variety of biomedical tests and clinical examinations. These diagnostics must also take into account age-related comorbid medical conditions, such as diabetes and cardiovascular diseases, which concern a large fraction of patients, as the incidence of neurodegenerative diseases increases with age. Such comorbid medical conditions influence neuropathology treatment decisions as well as short- and long-term survival of patients but are often overlooked in clinical trials. This situation underlines the need to directly analyze real life medical records to learn *clinical networks*, that are graphical models highlighting direct, indirect and possibly causal associations between clinically relevant information in patients' medical records.

Medical records contain, however, mixed types of data from simple binary or nominal variables (*i.e.*, with multiple unordered categories) to ordinal (*e.g.* neuropsychological test scales) or continuous (*e.g.* age, body mass index) variables, whose interdependences are not readily assessed within a unified information-theoretic framework. As mutual information is primarily defined between nominal variables, its estimation for continuous or mixed-type variables is notoriously difficult beyond the gaussian approximation of continuous distributions, for which a simple relation exists with correlation coefficients [1]. In particular, arbitrary discretization of continuous variables tends to underestimate mutual information for small number of bins, while overestimating it for large number of bins due to finite numbers of patients, as sketched in Fig 1. Moreover, so far, no rationale provides optimum bin partitions to estimate mutual information, for typical cohort size of patients. Alternatively, local metric approaches have been proposed to estimate mutual information [2] and conditional information [3–5], including between mixed-type variables [6–8], based on k-nearest neighbor (kNN) statistics. However, the statistical significance of kNN information estimates remains difficult to assess in practice [2, 9], thereby limiting their use to uncover (conditional) independences between continuous or mixed-type variables from real-life datasets.

In this paper, we first develop and implement an optimum binning method to simultaneously compute and assess the significance of mutual information, as well as conditional multivariate information, between any combination of continuous or mixed-type variables. The method is based on minimum description length principles [10, 11] and finds optimum bin

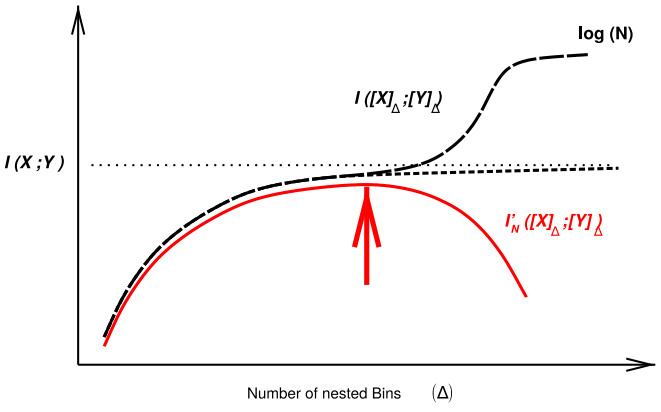

Fig 1. Mutual information computation between continuous or mixed-type variables. Outline of mutual information computation between continuous or mixed-type variables for a finite dataset of *N* samples. Mutual information is estimated through an optimum partitioning of continuous variable(s) (solid red line and arrow) after introducing a complexity term to account for the finite size of the dataset, see main text.

https://doi.org/10.1371/journal.pcbi.1007866.g001

PLOS COMPUTATIONAL BIOLOGY Learning clinical networks based on information estimates in mixed-type data

partitions, iteratively for each continuous variable, through an efficient dynamic programming scheme with quadratic complexity, $\mathcal{O}(N^2)$, where N is the number of patients in the dataset. This efficient approach is then used to assess direct *versus* indirectcause-effect relationships between mixed-type data from medical records, by extending a recent network learning method [12, 13] to recontruct graphical models beyond simple categorical datasets.

The method is shown to outperform existing tools on benchmark mixed-type datasets, before being applied to analyze the medical records of eldery patients with cognitive disorders from La Pitié-Salpêtrière Hospital, Paris. The resulting clinical network visually captures the global interdependences in these medical records and some facets of clinical diagnosis practice, without specific hypothesis nor prior knowledge on any clinically relevant information. The reconstructed clinical network recovers well known as well as novel direct and indirect relations between medically relevant variables. In particular, it provides some physiological insights linking the consequence of cerebrovascular accidents to the atrophy of important brain structures associated to cognitive impairment.

Methods

Assessing information in continuous or mixed-type data

Information-maximizing discretization of continous data. While mutual information is usually defined as a discrete summation over nominal variables, *i.e.*, $I(X;Y) = \Sigma_{x,y}\, p_{x,y} \log(p_{x,y}/p_x\, p_y)$, its most general definition consists in taking the supremum over all finite partitions, $\mathcal{P}$ and $\mathcal{Q}$, of variables, X and Y [1],

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \tag{1}$$

which can be applied to continuous or mixed-type variables. Moreover, by continuing to refine some initial partitions through the addition of further cut points for continuous variable(s), one finds a monotonically increasing sequence [1], $I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$, as depicted on Fig 1. In practice, however, Eq 1 cannot be used to estimate $I(X;Y)$ from an actual dataset with finite sample size, as the refinement of partitions eventually assigns each of the N different samples into N different bins. This leads to a shift of convergence towards log$N$ instead of the theoretical limit, $I(X;Y)$, which requires an infinite amount of data (dotted line in Fig 1).

In this paper, we propose to adapt Eq 1 to account for the finite number of samples in actual datasets,

$$I'_N(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \tag{2}$$

by introducing a finite size correction to mutual information,

$$I'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) = I_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) - k'_{\mathcal{P},\mathcal{Q}}(N)\frac{1}{N} \tag{3}$$

where $k'_{\mathcal{P},\mathcal{Q}}(N)$ corresponds to a complexity term introduced in [14, 15] to discriminate between variable dependence (for $I'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) > 0$) and variable independence (for $I'_N([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \leqslant 0$) given a finite dataset of size N. In the present context of finding an optimum discretization for continuous variables, this complexity term introduces a penalty which eventually outweights the information gain in refining bin partitions further, when there is not enough data to support such a refined model, as depicted on Fig 1.

For discrete variables, typical complexity terms correspond to the Bayesian Information Criterion (BIC), $k^{\mathrm{BIC}}_{\mathcal{P},\mathcal{Q}}(N) = 1/2(r_x - 1)(r_y - 1)\, \log N$, where r_x and r_y are the number of bins

for X and Y, or the X- and Y-Normalized Maximum Likelihood (NML) criteria [14–16], defined as,

$$k_{\mathcal{P};\mathcal{Q}}^{X-\mathrm{NML}}(N) = \sum_{y}^{r_y} \log\, \mathcal{C}_{n_y}^{r_x} - \log\, \mathcal{C}_{N}^{r_x} \tag{4}$$

$$k_{\mathcal{P};\mathcal{Q}}^{Y-\mathrm{NML}}(N) = \sum_{x}^{r_x} \log\, \mathcal{C}_{n_x}^{r_y} - \log\, \mathcal{C}_{N}^{r_y} \tag{5}$$

where $\mathcal{C}_{n_y}^{r_x}$ is the parametric complexity associated with the yth bin of variable Y containing n_y samples, and similarly for $\mathcal{C}_{n_x}^{r_y}$ with the n_x-size bin of variable X in Eq 5.

Parametric complexities $\mathcal{C}_n^r$ are defined by summing a multinomial likelihood function over all possible partitions of n data points into a maximum of r bins as,

$$\mathcal{C}_n^r = \sum_{\ell_1+\ell_2+\cdots+\ell_r=n}^{\ell_k \geqslant 0} \frac{n!}{\ell_1!\ell_2!\cdots\ell_r!} \prod_{k=1}^{r} \left(\frac{\ell_k}{n}\right)^{\ell_k} \tag{6}$$

which can in fact be computed recursively in linear-time [17]. For large n and r, inherent to large datasets with continuous or mixed-type variables, we found that $\mathcal{C}_n^r$ computation can be made numerically stable by implementing the recursion on parametric complexity ratios $\mathcal{D}_n^r = \mathcal{C}_n^r/\mathcal{C}_n^{r-1}$ rather than parametric complexities themselves as,

$$\mathcal{D}_n^r = 1 + \frac{n}{(r-2)\mathcal{D}_n^{r-1}} \tag{7}$$

$$\log\, \mathcal{C}_n^r = \sum_{k=2}^{r} \log\, \mathcal{D}_n^k \tag{8}$$

for $r \geqslant 3$, with $\mathcal{C}_n^1 = 1$ and $\mathcal{C}_n^2 = \mathcal{D}_n^2$, which can be computed directly with the general formula, Eq 6, for $r = 2$,

$$\mathcal{C}_n^2 = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \tag{9}$$

or its Szpankowski approximation for large n (needed for $n > 1000$ in practice) [18–20],

$$\mathcal{C}_n^2 = \sqrt{\frac{n\pi}{2}}\left(1 + \frac{2}{3}\sqrt{\frac{2}{n\pi}} + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{10}$$

$$\simeq \sqrt{\frac{n\pi}{2}} \exp\left(\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}\right) \tag{11}$$

For continuous variables, however, the variable categories are not given *a priori* and need to be specified and thus encoded in the model complexity within the frame of the Minimum Description Length (MDL) principle [11]. In absence of priors for any specific partition with r bins, the model index should be encoded with a uniform distribution over all partitions with the same number of bins [11]. As there are $\binom{N-1}{r_x-1}$ ways to choose $r_x - 1$ out of $N - 1$ possible cut points, corresponding to a codelength of $\log\, \binom{N-1}{r_x-1}$ for a continuous variable X (and

similarly for Y if it is continuous), the model complexity associated with the partitioning of continuous or mixed-type variablesbecomes,

$$k'_{\mathcal{P};\mathcal{Q}}(N) = k_{\mathcal{P};\mathcal{Q}}(N) + \log\binom{N-1}{r_x-1} + \log\binom{N-1}{r_y-1} \tag{12}$$

with $\log\binom{N-1}{r-1} = (r-1)\, C_{N,r}$, where $C_{N,r}$ corresponds to theencoding cost associated to each of the $r-1$ cut points with $r = r_x$ or r_y.

While finding the supremum of $I'_N([X]_{\mathcal{P}}; [Y]_{\mathcal{Q}})$ over *all* possible partitions $\mathcal{P}$ and $\mathcal{Q}$ according to Eq 2 seems intractable, it can be computed rather efficiently in practice.

The approach is inspired by the computation of an MDL-optimal histogram for a single continuous variable [11], which can be done exactly in $\mathcal{O}(N^3)$ steps. As the approach cannot be generalized to more than one variable, we implemented a local optimization heuristics, which finds the optimum cut points for each continuous variable, iteratively, keeping the partitions of the other continuous variable(s) fixed. This enables to gain an order of magnitude in the optimization running time at each iteration, which scales as $\mathcal{O}(N^2)$, as detailed below.

In practice for two variables, we start from an initial (or optimized) X partition with r_x bins of various sizes and an estimate of the number of Y bins, r_y°. The sample-scaled mutual information with finite size correction, *i.e.*, $nI'_n(X;Y)$, is then optimized iteratively for $n = 1, \cdots, N$ samples, over all Y partitions, through the following $\mathcal{O}(N^2)$ dynamic programming scheme, using Eq 4 as parametric complexity,

$$nI'_n(X;Y) = \max_{0\leq j<n} \left[jI'_j(X;Y) + \sum_x^{r_x} n_{xy} \log n_{xy} - n_y \log n_y - \log \mathcal{C}_{n_y}^{r_x} - C_{N,r_y^\circ} \right] \tag{13}$$

where the last added Y bin, including $n_y = n - j$ samples distributed over the r_x bins of X (with $\sum_x^{r_x} n_{xy} = n_y$), comes with an independent mutual information contribution, $\sum_x^{r_x} n_{xy} \log n_{xy} - n_y \log n_y$, a parametric complexity, $\log \mathcal{C}_{n_y}^{r_x}$, and encoding cost, C_{N,r_y°. The initial condition for $j = 0$ in (13) is set by convention to include all terms invariant under Y-partitioning, *i.e.*, $-\sum_x^{r_x} n_x \log(n_x/N) + \log \mathcal{C}_N^{r_x} - (r_x - 1) C_{N,r_x} + C_{N,r_y^\circ}$.

Then, adopting this optimized partition for Y, one can apply the same dynamic programming scheme for X using Eq 5 as parametric complexity and iterate the optimization of X and Y partitions until a stable two-state limit circle is reached. In practice, we set the initial partitioning over X and Y by testing equal-freq discretizations with 2 to $\lceil N^{1/3} \rceil$ bins and choosing the one which gives thehighest $I'_N(X;Y)$. We found that while the convergence speed of the iterative dynamic programming is largely independent of these initial conditions, this scheme does improve it slightly. This leads after only a few iterations to a good estimate of mutual information (averaged over limit circle) that is comparable to the existing state of the art, for both continuous and mixed-type variables, as shown below.

This optimization scheme, Eq 2, and its iterative dynamic programming computation, Eq 13, can also be adapted to compute mutual information involving joined variables, such as $I'_N(X; \{A_i\})$, with corresponding finite size correctionsand cut point encoding costs extended from Eqs 3–12. Similarly, the approach can compute conditional mutual information, such as $I'_N(X; Y|\{A_i\})$, involving continuous or mixed-type variables. To this end, $I'_N(X; Y|\{A_i\})$ needs to be defined, using the chain rule [1], as the *difference* between maximized mutual information terms involving either $\{Y, \{A_i\}\}$ and $\{A_i\}$ (Eq 14) or $\{X, \{A_i\}\}$ and $\{A_i\}$ (Eq 15) as joined

variables,

$$I'_N(X;Y|\{A_i\}) = I'_N(X;Y,\{A_i\}) - I'_N(X;\{A_i\}) \quad (14)$$

$$= I'_N(Y;X,\{A_i\}) - I'_N(Y;\{A_i\}) \quad (15)$$

Thus, starting from an initial (or optimized) partition for X, each term of Eq 14 is optimized with respect to Y and $\{A_i\}$ partitions using Eq 4 as parametric complexity extended to multivariate categories, $n_{y,\{ai\}}$ and $n_{\{ai\}}$. Then, in turn, each term of Eq 15 is optimized with respect to X and $\{A_i\}$ partitions using Eq 5 as parametric complexity extended to multivariate categories, $n_{x,\{ai\}}$ and $n_{\{ai\}}$. Note, in particular, that $\{A_i\}$ partitions are optimized *separately* for each of the four terms in Eqs 14 & 15, before taking their differences, as these optimized $\{A_i\}$ partitions might be different in general.

Learning networks from continuous or mixed-type data

The above information maximization scheme to estimate (conditional) mutual information between continuous or mixed-type variables can then be used to extend our recent network learning algorithm MIIC [12] beyond simple categorical datasets.

Outline of MIIC algorithm. MIIC combines constraint-based approach and information-theoretic framework to robustly learn a broad class of causal or non-causal networks including possible latent variables [12, 13]. MIIC proceeds in three steps:

i). *Edge pruning*. Starting from a fully connected network, MIIC first removes dispensable edges by iteratively subtracting the most significant information contributions from indirect paths between each pair of variables. Significant contributors are collected based on the *3off2* score [14, 15] maximizing conditional three-point information while minimizing conditional two-point (mutual) information, which reliably assesses conditional independence, even in the presence of strongly linked variables [21]. The residual (conditional) mutual information including finite size corrections, $I'_N(X;Y|\{A_i\})$ (*i.e.* after indirect effects of significant contributors, $\{A_i\}$, have been subtracted from $I'_N(X;Y)$), is related to the removal probability of each edge, $P_{XY} = \exp(-NI'_N(X;Y|\{A_i\}))$, where $NI'_N(X;Y|\{A_i\}) > 0$ corresponds to the strength of the retained edge, as visualized by its width in MIIC graphical models [12].

ii). *Edge filtering (optional)*. The remaining edges can be further filtered based on confidence ratio assessment [12],$C_{XY} = P_{XY}/\langle P_{XY}^{\text{rand}}\rangle$, where P_{XY}^{rand} is the average of the probability to remove the XY edge after randomly permutating the dataset for each variable. Hence, the lower C_{XY}, the higher the confidence on the XY edge. In practice, filtering edges with $C_{XY} > 0.1$ or 0.01 limits the false discovery rates with small datasets, while maintaining satisfactory true positive rates [12].

iii). *Edge orientation*. Retained edges are then oriented based on the signature of causality in observational data given by the sign of (conditional) three-point information [14, 15]. The final network contains up to three types of edges [12]: undirected, directed, as well as, bidirected edges, which originate from a latent variable, L, unobserved in the dataset but predicted to be a common cause of X and Y, *i.e.* $X \dashleftarrow (L) \dashrightarrow Y$. For clarity, bidirected edges are represented with dashed lines in MIIC networks.

An important aspect of MIIC algorithm is its ability to take into account datasets with missing values, which are frequent in heterogeneous clinical datasets. In practice, MIIC computes multivariate information estimates (such as $I'_N(X;Y|\{A_i\})$) on sub-datasets for which X, Y and

PLOS COMPUTATIONAL BIOLOGY

$\{A_i\}$ do not have missing values. While including iteratively additional conditioning variables A_i might further restrict the size of the sub-dataset without missing value, we only consider variables A_i if their missing values are missing at random (checking Kullback Leibler divergence between distributions of decreasing supports). If some data is not missing at random, the *3off2* scheme [14, 15], $I(X; Y|\{A_i\}_n) = I(X; Y) - I(X; Y; A_1) - I(X; Y; A_2|A_1) - \cdots - I(X; Y; A_n|\{A_i\}_{n-1})$, might end without finding conditional independence, ie $I(X;Y|\{A_i\}_n)>0$, and MIIC edge pruning step is conservative by retaining the corresponding edge X-Y due to possible bias in the dataset.

MIIC's extension to continuous or mixed-type data has been implemented in MIIC online server and R package, see SI.

Results

Application to benchmark synthetic data

Optimum discretization and mutual information estimates for continuous or mixed-type data. The multivariate discretization scheme and resulting estimates of (conditional) mutual information were first benchmarked using synthetic data from known mixed or continuous probability distributions for which (conditional) mutual information can be obtained either analytically or through numerical integration. Examples of bivariate information-maximizing discretizations are shown in Fig 2 and S1 Fig for increasing sample size. The number of bins increases both with the number of samples, S1 Fig, and the magnitude of mutual information, $I_N(X; Y)$, S2A Fig. These tendencies have intuitive explanations: first, more samples means that we can assign smaller bins (width-wise) with more certainty; and second, more information means that more bins are needed to describe the interaction between the variales. We note that no single discretization of a variable X can be optimal with regards to every joint distribution, see S3 Fig. While the precise cut points of variable X actually depend on the

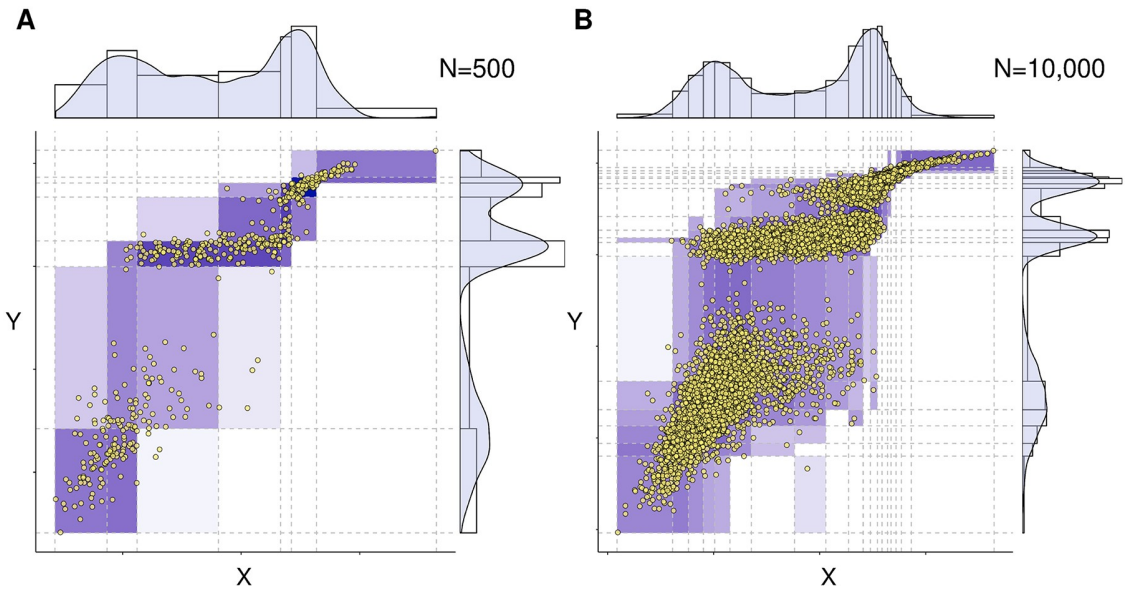

Fig 2. Optimum bivariate discretization for mutual information estimate. The proposed information-maximizing discretization scheme is illustrated for a joint distribution defined as a Gumbel bivariate copula with parameter $\theta = 5$ and marginal distributions chosen as Gaussian mixtures with three equiprobable peaks and respective means and variances, $\mu_X = \{0, 4, 6\}$, $\sigma_X = \{1, 2, 0.7\}$ and $\mu_Y = \{-3, 6, 9\}$, $\sigma_Y = \{2, 0.5, 0.5\}$. The information-maximizing partition yields (**A**) $I_N(X; Y) = 1.04$ for $N = 500$ samples and (**B**) $I_N(X; Y) = 1.142$ for $N = 10,000$ samples, as compared to the exact expected value $I(X; Y) = 1.205$ computed with numerical integration. See S1 Fig for additional results. Codes are provided at https://github.com/vcabeli/miic_PLoS.

https://doi.org/10.1371/journal.pcbi.1007866.g002

variable Y of interest, the number of X and Y bins are roughly similar(for the chosen test settings), S2A Fig, unlike found with information-maximization discretization methods lacking complexity terms [22], S2B Fig.

Next, we compared our estimation of $I_N(X; Y)$ by optimal discretization to the state of the art Kraskov–Stögbauer–Grassberger (KSG) estimator [2] for continuous distributions, specifically bivariate Gaussian distributions S4 Fig. Like otherinformation estimators based on kNN statistics, the KSG approach has a tunable parameter k which will typically scale with the sample size N, and has to be chosen depending on the objective: the original authors recommend $k = 2$ to 4 for the best estimation, and up to $N/2$ if one is more interested in independence testing. We found that our optimal discretization with the NML complexity does indeed give a correct estimation of $I_N(X; Y)$ for all sample sizes and correlation strengths. Our approach also natively deals with categorical and mixed (*i.e.* part categorical and part continuous) variables, as the master definition of the mutual information, Eq 1, can be applied to variables of any type. Recent efforts were made to extend the KSG estimator to such cases [6–8] which are frequently encountered in real-life data, and specifically in clinical datasets. We compared the mixed-type information estimates of our method to other existing methods for varying sample sizes and found its performance to be similar or superior, S5 Fig. In addition, our information-maximizing discretization approach facilitates the interpretation of the dependences between continuous or mixed-type variables by returning their most informative categories.

Information-maximizing discretization and corresponding (conditional) mutual information estimates can be computed for any continuous or mixed-type dataset using the `discretizeMutual` function from the MIIC R package.

Optimum discretization as an independence test between continuous or mixed-type variables. Most importantly,our optimum discretization scheme also acts as an independence test by allowing for single bin partitions whenever no multiple-bin partitioning can glean information that is greater than its associated complexity cost. In such cases, our estimator implies variable independence, *i.e.* $I_N(X; Y) = 0$, with drastically reduced sampling error and variance, S4 Fig, as compared to other direct estimators such as KSG, which always give noisy information estimates even for vanishing mutual information between nearly independent variables and need additional hypothesis testing to be used as independence test.

Similarly, our approach robustly learns conditional independence,given a set of separating variables, $\{Z_i\}$, *i.e.*, $I_N(X; Y|\{Z_i\}) = 0$, S6 Fig, as in the case of a single common ancestor Z of X and Y, *i.e.*, $X \leftarrow Z \rightarrow Y$, with concomitant changes in optimum X and Y partitionings from multiple to single bins under conditioning over a continuous (S7 Fig) or categorical (S8 Fig) variable Z. By contrast, spurious dependency between independent variables, X and Y, can be induced, as expected [23], by conditioning over a common descendent Z, as in the case of a "v-structure", $X \rightarrow Z \leftarrow Y$, S9 Fig.

Hence, the intrinsic robustness of the present optimum discretization scheme in inferring (conditional) independence and dependency is an important feature of the method as compared to kNN (conditional) information estimates, whose statistical significance remains difficult to assess in practice [2, 9].

Reconstruction of benchmark graphical models. We first tested the mixed-type data extension of MIIC network reconstruction method on benchmark mixed-type data. Datasets were generated based on non-linear bayesian rules using the R script provided as Supplementary code; an example of non-Gaussian mixed-type distribution dataset is shown in S10 Fig. The resulting reconstructed network F-scores are shown in Fig 3 for an increasing proportion of continuous variables over discrete variables and compared to the recent alternative methods, CausalMGM [24] and MXM [25], also designed to analyze mixed-type data. Precision, Recall and F-scores are shown for both skeleton and CPDAG in S11 and S12 Figs, respectively.

PLOS COMPUTATIONAL BIOLOGY Learning clinical networks based on information estimates in mixed-type data

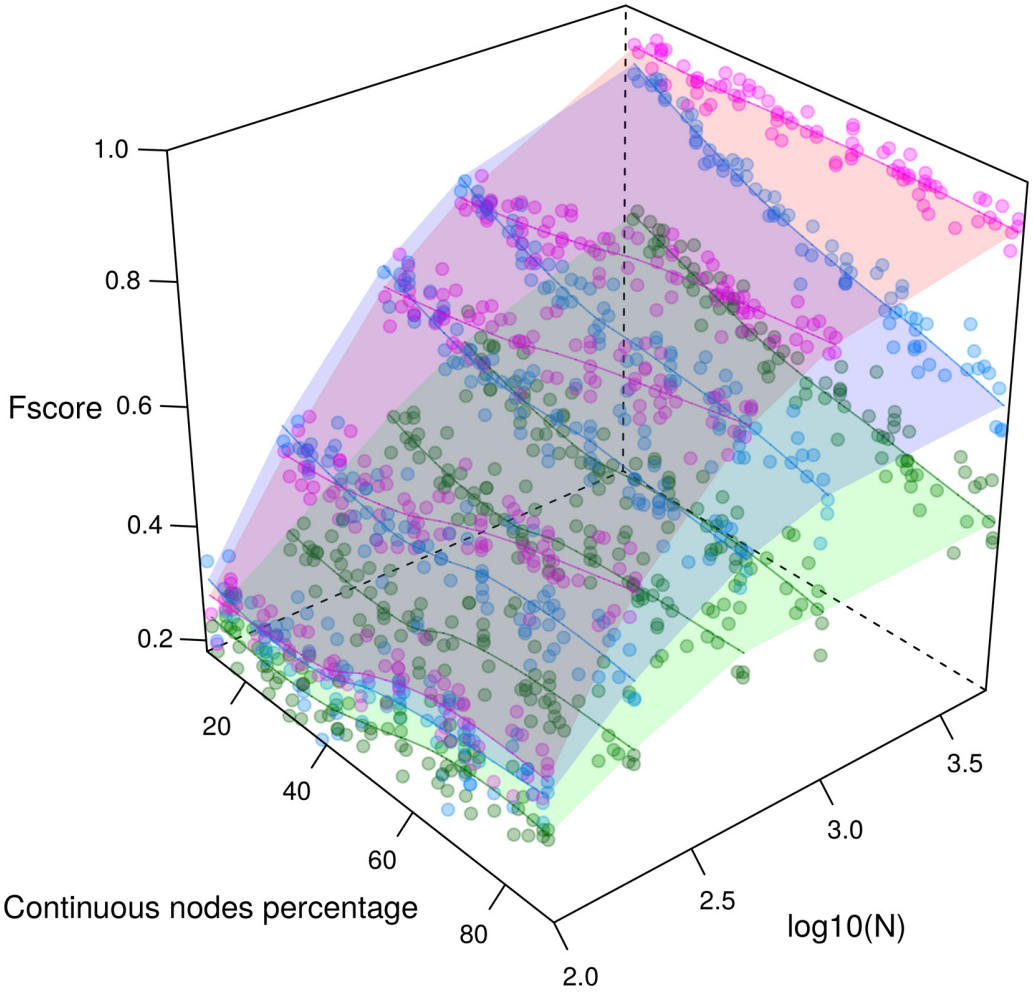

Fig 3. Reconstruction of benchmark networks for mixed-type, non-linear, non-Gaussian datasets. CPDAG F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100$–5,000 samples (see histogram example S10 Fig). F-scores obtained with our parameter-free information-theoretic approach MIIC (magenta, upper surface) are compared to the best results obtained with alternative mixed-type data methods, CausalMGM [24] (blue, middle surface) and MXM [25] (green, lower surface), by optimizing CausalMGM regularization parameters (λ) and MXM significance parameter (α), for each sample size N. See additional results in S11–S15 Figs. Codes are provided at https://github.com/vcabeli/miic_PLoS.

https://doi.org/10.1371/journal.pcbi.1007866.g003

Comparisons with fully continuous datasets, S13 Fig, were also performed with additional methods, CAM [26], kPC, rank-PC and rank-FCI [27] algorithms, S14 and S15 Figs, and confirm the better performance of MIIC over alternative continuous or mixed-type network learning methods.

Application to medical records of eldery patients with cognitive disorders

We applied this information maximization analysis for mixed-type data to reconstruct a clinical network from the medical records of 1,628 eldery patients consulting for cognitive disorders at La Pitié-Salpêtrière hospital, Paris. The dataset,provided as S1 Table, contains 107 variables of different types (namely, 19 continuous and 88 categorical variables) and heterogeneous nature (*i.e.*, variables related to previous medical history, comorbidities and

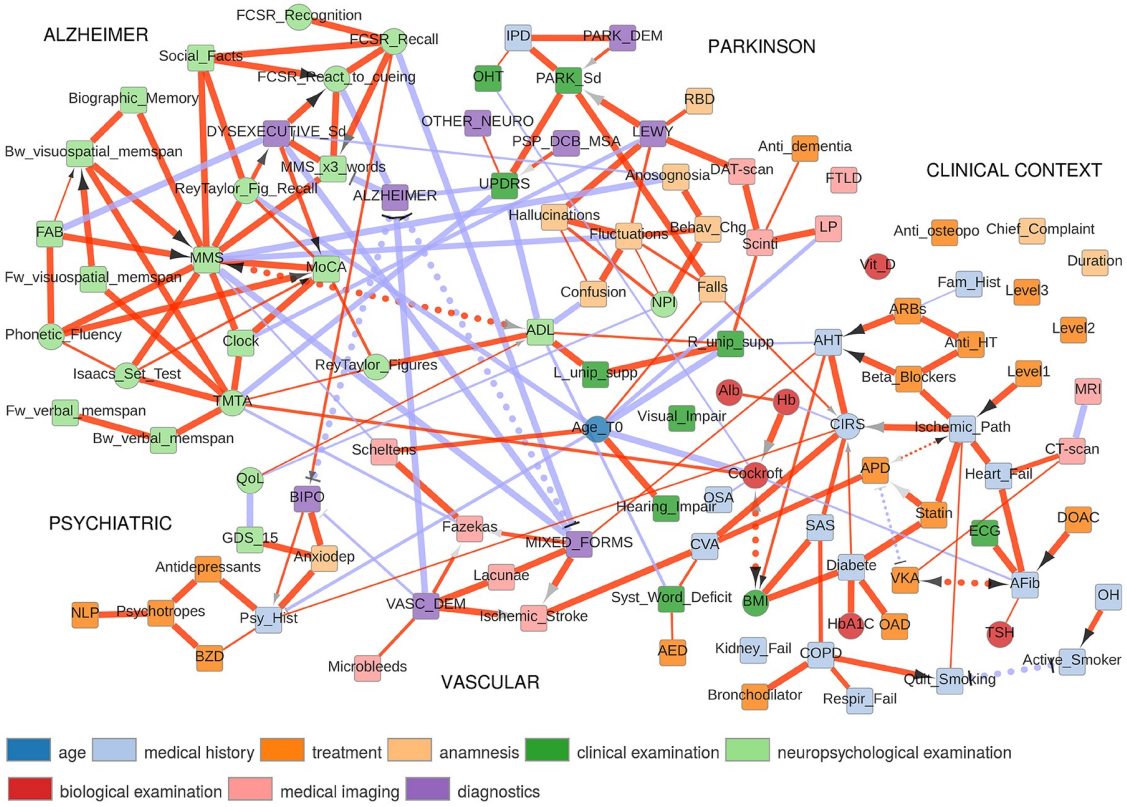

Fig 4. Network reconstructed from medical records of 1,628 eldery patients with cognitive disorders. Square (resp. circle) nodes correspond to discrete (resp. continuous) variables. Red (resp. blue) edges correspond to correlation (resp. anticorrelation) between variables. Dotted edges reflect latent variables, see Discussion.

https://doi.org/10.1371/journal.pcbi.1007866.g004

comedications, scores from cognitive tests, clinical, biological or radiological examinations, diagnostics and treatments). Beyond the different types and heterogeneous nature of the recorded data, nodes of the clinical network, Fig 4, can be partitioned into groups associated to specific dementia disorders and patient clinical context, including comorbidities (diabetes, hypertension, etc) and related comedications.

Parkisonian syndromes. The first group of nodes contains variables classically linked to primary degenerative dementias associated to parkinsonian syndromes (Park_Sd), notably the rarity and slowness of movements, tremor at rest and muscle stiffness, caused either by a parkinsonian dementia (PARK_DEM, 80% of cases) or a dementia with Lewy bodies (LEWY, 15% of cases). Park_Sd are identified with the Unified Parkinson Disease Rating Scale (UPDRS) which distinguishes them from Parkinson plus syndromes such as Progressive Supranuclear Palsy (PSP), Cortico Basal Degeneration (CBD) or Multiple System Atrophy (MSA). Parkinsonian syndromes are also linked to more frequent falls, idiopathic Parkinson's disease (IPD) and associated to orthostatic hypotension (OHT), in agreement with previous studies [28]. By contrast, dementia with Lewy bodies (LEWY) is found to be directly associated to cognitive fluctuations, halluciations and Rapid eye movement sleep Behavior Disorder (RBD) as well as indirectly connected (2nd neighbor) to confusions and behavioural changes assessed through the Neuro Psychiatric Inventory (NPI) score and with a deficit of self-awareness (Anosognosia). LEWY diagnoses are also correctly associated with dopamine transporter imaging (DAT-scan) examination [29].

Alzheimer's versus dysexecutive syndromes. The second and largest group of nodes mostly consists of the results from neuropsychologic tests used to assess the cognitive

functions of patients and diagnose Alzheimer's disease *versus* dysexecutive syndromes. Two types of tests can be distinguished: simple tests probing a precise cerebral function and composite tests combining the results of multiple simple tests to explore more global cognitive processes. The Trail Making Test part A (TMTA) is a simple test primarily used to examine cognitive processing speed (continuous score) by recording the time needed by the patient to connect ordered nodes (from 1 to 25) randomly placed on a sheet of paper. Our network analysis shows that TMTA is directly connected to a number of other simple tests, such as forward memory spans probing attentional capacity, backward memory spans probing immediate working memory, immediate recall of Taylor or Rey complex figures, verbal semantic fluency (Issacs set test) and the clock-drawing test. This highlights the rationale of neuropsychology in combining simple tests into more informative composite tests. Three composite tests are included in the clinical network, the Mini Mental State (MMS), the Frontal Assessment Battery (FAB) and the Montreal Cognitive Association (MoCA) tests.

- The Mini Mental State (MMS) test assesses cognitive functions related to memory, spacial and temporal orientations but not to executive functions, which require to integrate multiple information sources. MMS is found to be the main hub (with 15 neighbors) of the reconstructed network, as it is directly connected, as expected, to most of the memory test results (forward/backward verbal and visuospatial memory spans, biographic memory and delayed recalls of Taylor or Rey–Osterrieth complex figures). By constrast, MMS is found to be negatively correlated to the Alzheimer's diagnostic, through the MMS 3 word memory test, which is known to be one of the most specific tests for Alzheimer's disease, together with the Free and Cued Selective Reminding (FCSR) test. Interestingly, our network analysis shows that the Alzheimer's disease diagnostic is directly connected to the FCSR test through the low percent reactivity to cueing, which identifies genuine storage deficits (not facilitated by cueing) due to amnesic syndrome of the hippocampal type known to be characteristic of Alzheimer's disease [30].
- The Frontal Assessment Battery (FAB) test is complementary to MMS, as it is entirely focussed on executive functions, centralized in the frontal cortex; it is thus very consistent that FAB is found to be directly connected and negatively correlated to dysexecutive syndrome. Note, however, that patients suffering from dysexecutive syndrome do not typically show poor FCSR scores unlike Alzheimer patients. This confirms the specificity and sensibility of the FCSR test to Alzheimer's disease [31].
- Finally, the Montreal Cognitive Association (MoCA) composite test integrates a variety of other tests such as the clock-drawing test, the phonetic fluency test as well as semantic fluency test (Isaacs Set Test), which is consistent with the direct connections recovered between MoCA and these three individual tests in the inferred network.

Psychiatric conditions. The third group of nodes concerns variables associated with the psychiatric conditions of patients. It includes their past psychiatric history (Psy_Hist) and present psychiatric conditions, *i.e.*, anxio-depressive or bipolar (BIPO) syndromes, associated treatments (antidepressants, psychotropes, benzodiazepine BZD and neuroleptics NLP) and finally scores used to diagnose depression (GDS_15) and a deterioration in the quality of life (QoL). The analysis of all the links between these variables confirms the overall consistency of this psychiatric cluster: a good quality of life is closely associated with a low GDS_15 score (corresponding to a low probability of depression). Note, however, that psychiatric pathologies are all linked to each other, underlying the difficulty to distinguish them accurately. Yet, our

network analysis shows that patients with bipolar syndrome (BIPO) tend to show better scores at the FCSR recall test.

Vascular versus mixed forms of dementias. The fourth group of nodes of the clinical network is associated with variables implicated in vascular dementias (VASC_DEM) originating from cerebral vascular accidents (CVA) which damage brain regions essential for cognitive processes. Different types and sizes of vascular accidents are distinguished from microbleeds to ischemic stroke (clot) and lacunae (empty spaces in the deep brain structures). These more severe vascular accidents may also lead to degenerative dementia syndromes, corresponding to a mixed form of dementia (MIXED_FORMS), which is inferred to be directly associated to low MMS scores and poor scores at the FCSR Recall test (*i.e.*, negative direct links). VASC_DEM and MIXED_FORMS are also found to be connected to the Fazekas scale [32], which detects and quantifies white matter hyperintensities in the brain that are the consequence of cerebral small vessel disease including demyelination and axonal loss of neuronal cells. The Fazekas scale is found to be directly associated to low cognitive processing speed (TMTA) and also strongly correlated to the Scheltens scale [33] quantifying the severity of hippocampal atrophy, in agreement with a recent independent report [34]. The hippocampus is a brain structure involved in memory and space navigation, which is consistent with our finding of a direct negative association between Scheltens scale and MMS score. Interestingly, this predicted association between the Fazekas and the Scheltens scales, inferred from our unsupervised global network analysis, provides some physiological insights linking the consequence of vascular accidents (Fazekas scale) to the atrophy of important brain structures (Scheltens scale) and, thereby, to cognitive and functional impairments, as reported in clinical studies linking white matter hyperintensities (Fazekas scale) to cognitive impairment [35].

Patient clinical context. The last important group of nodes of the clinical network includes variables associated with the patient clinical context including comorbidities, related examinations and treatments. These are different anterior chronic diseases, such as arterial hypertension (AHT), diabetes, chronic obstructive pulmonary disease (COPD), atrial fibrillation (AFib), that might have an impact on the patient's vital prognosis. All the links within this comorbidity cluster are very consistent, each pathology being directly associated with its known risk and predisposition factors, biological markers, specific examinations and treatments. In particular, diabetes is associated with a high body mass index (BMI), glycated hemoglobin blood test (HbA1c), treatment by oral antidiabetic (OAD) drugs and statin; COPD is associated with sleep apnea syndrome (SAS) and the risk of respiratory failure, the use of bronchiodilator drugs and the necessity to quit smoking; AHT is associated with an increase risk of mixed form dementia and treatments by angiotensin receptor blockers (ARBs), beta-blockers and other anti-hypertension (Anti HT) drugs; Finally, AFib, detected by electrocardiogram (ECG), is associated with an increased risk of heart failure and high levels of thyroid-stimulating hormone (TSH) and treated with vitamine K antagonist (VKA) and direct oral anticoagulants (DOAC).

Discussion

We report in this paper a novel optimal discretization method to simultaneously compute and assess the significance of mutual information, as well as conditional multivariate information, between any combination of continuous or mixed-type variables. The approach is used to reconstruct graphical models from mixed-type datasets by uncovering direct, indirect and possibly causal relationships in complex heterogenous data. The method is shown to outperform state-of-the-art approaches on benchmark mixed-type datasets, before being applied to analyze the medical records of eldery patients with cognitive disorders from La Pitié-Salpêtrière Hospital, Paris.

From a methodological perspective, this information-maximizing discretization approach facilitates the interpretation of either the dependences or the independencies between continuous or mixed-type variables. First, obtaining optimal discretization helps explain the dependences in terms of the most informative categories of continuous variables. Second, and most importantly, optimal discretization also acts as an independence test by allowing for single bin partitions whenever multiple-bin partitioning provides less information than its associated complexity cost.

From the perspective of clinical applications, the method is able to globally uncover interdependences within complex heterogeneous data from medical records without specific hypothesis nor prior knowledge on any clinically relevant information. The reconstructed clinical network from cognitive disorder patients (Fig 4) recovers well known as well as novel direct and indirect relations between medically relevant variables.

In addition, we found that this reconstructed clinical network captures also some facets of the neurologist's reasoning behind the diagnoses of distinct dementias. In particular, diagnosis nodes can be interpreted as "explanatory" variables associated to a number of "explaining-away effects" [23] in the form of "v-structures", *i.e.*, $D_1 \rightarrow S/E \leftarrow D_2$, whenever alternative diagnoses, D_1 or D_2, can independently explain a given syndrome, S, or the result of a specific examination, E. Examples discussed in more details above are PARK_DEM → PARK_Sd ← LEWY, VASC_DEM → Fazekas ← MIXED_FORMS and VASC_DEM → Ischemic_Stroke ← MIXED_FORMS. In addition, anticorrelations between different diagnostic nodes reflect the alternative choices of diagnosis by the neurologist, either in the form of "differential diagnoses" through a reasoning by elimination, in particular, to diagnose Alzheimer's disease, *i.e.*, VASC_DEM ⊣ ALZHEIMER, or in the form of a latent variable, visualized as bidirected dotted edges and corresponding to alternative diagnoses by the neurologist, *i.e.*, ALZHEIMER⇠*diagnosis*⇢MIXED_FORMS or ALZHEIMER⇠*diagnosis*⇢BIPO. Latent variables may also represent the clinician's decisions between alternative treatments, *e.g.*, APD⇠*clinician_decision*⇢VKA or a nonrecorded or implicite information in the patient personal or medical history, *e.g.*, active_smoker⇠*ever_smoked*⇢quit_smoking, Fig 4.

The main strengths of our clinical network reconstruction method are three-fold. First, it performs an unbiased check on the database content (expected, yet missing direct links in the reconstructed network hint to likely problems in the database *e.g.*, erroneous or missing data). Second, it does not need any expert-informed hypothesis and provides, without prior knowledge in the field, graphical models complementing analyses by experts. Finally, it can discover novel unexpected direct interdependencies between clinically relevant information, such as the direct connection between Fazekas and Scheltens scales, Fig 4, which may provide some physiological insights and suggest new research directions for further investigation.

Hence, beyond the challenge of learning clinical networks from mixed-type data, our method offers a user-friendly global visualisation tool of complex, heterogeneous clinical data which could help other practitioners visualize and analyze direct, indirect and possibly causal effects from patient medical records.

Supporting information

S1 File. Supplementary Materials and Methods. Benchmark data generation (continuous and discrete variables). Performance measures. Benchmark parameter tuning. Resource availability.
(PDF)

S1 Table. Dataset from 1,628 eldery patients with cognitive disorders from La Pitié-Salpêtrière hospital, Paris. The dataset, fully deidentified, contains 107 variables of different types

(namely, 19 continuous and 88 categorical variables) and heterogeneous nature (*i.e.*, variables related to previous medical history, comorbidities and comedications, scores from cognitive tests, clinical, biological or radiological examinations, diagnostics and treatments).
(XLSX)

S1 Fig. Optimum bivariate discretization for mutual information estimation. The proposed information-maximizing discretization scheme is illustrated for a joint distribution defined as a Gumbel bivariate copula with parameter $\theta = 5$ and univariate marginal-distribution functions chosen as Gaussian mixtures with three equiprobable peaks and respective means and variances, $\mu_X = \{0, 4, 6\}$, $\sigma_X = \{1, 2, 0.7\}$ and $\mu_Y = \{-3, 6, 9\}$, $\sigma_Y = \{2, 0.5, 0.5\}$. Information-maximizing partitions are displayed for different sample sizes with corresponding mutual information estimates: **(A)** $N = 100$ samples, $I_N(X; Y) = 0.928$ (and $I'_N(X; Y) = 0.649$); **(B)** $N = 500$ samples, $I_N(X; Y) = 1.040$ (and $I'_N(X; Y) = 0.866$); **(C)** $N = 1,000$ samples, $I_N(X; Y) = 1.096$ (and $I'_N(X; Y) = 0.977$); **(D)** $N = 10,000$ samples, $I_N(X; Y) = 1.142$ (and $I'_N(X; Y) = 1.075$). The actual mutual information value was computed through numerical integrationof the marginals and the joint probability distribution and yields, $I(X; Y) = 1.205$, in good agreement with the obtained estimates for large N.
(EPS)

S2 Fig. Adaptive information-maximizing partitions depending on interaction strength. To assess the range in bin numbers depending on the strength of interaction between variables, we generated $N = 1,000$ independent samples for 10,000 Gaussian bivariate distributions with a uniformly distributed correlation coefficient ρ in $[-1, 1]$. The real mutual information (RI) of Gaussian bivariate distributions can be computed directly [1], as $RI(X; Y) = -\log(1 - \rho^2)/2$. For each pair (X, Y), we estimated the mutual information with the proposed optimum bivariate discretization as well as the Maximal Information Coefficient [22] using the `minepy` package [36] **(A)** The information-maximizing partition proposed in the present paper behaves as expected: the number of bins on each variable is roughly similar and scales monotonically with the strength of the interaction between variables. This implies that additional bins are only introduced when their associated complexity cost is justified by a larger gain in mutual information. Conversely, when the information between X and Y approaches zero, both variables are partitioned into fewer and fewer bins until a single bin is selected for each variable, when they are inferred to be independent, given the available data. **(B)** The partition chosen to estimate the Maximal Information Coefficient is very different, regardless of the interaction strength, as it systematically corresponds to an unbalanced distribution of bins between the two variables, with one variable usually partitioned into the maximum number of bins(set by default to $\text{floor}(N^{0.6}/2) = 31$) while the other is discretized into two levels only. This result is not unexpected, however, as the Maximal Information Coefficient [22] is defined by maximizing the mutual information of the discretized variables over the grid, $I([X]_{\Delta_x}; [Y]_{\Delta_y})$, *normalized by the minimum* of $\log \Delta_x$ and $\log \Delta_y$. Indeed, maximizing the normalized mutual information is done by partitioning as few samples as possible into the maximum number of bins in one dimension (as sketched in Fig 1), while simultaneously minimizing the number of bins, and thus $\log \Delta_i$, in the other dimension. See further discussion in [37].
(EPS)

S3 Fig. Interaction-dependent optimum discretization. Optimum bivariate partitions obtained from $N = 1,000$ samples of two different joint distributions $P(X, Y)$ sharing the same sampling of X taken from a uniform distribution on $[0, 0.3]$, but with different dependences for Y. **(A)** Y is defined as $\log(X) + \epsilon_1$, and **(B)** Y is defined as $X^5 + \epsilon_2$, where ϵ_1 and ϵ_2 are Gaussian noise terms chosen so that the mutual informations of both examples are

comparable,$I(X;Y) \simeq 0.75$. This example shows that the optimum partition for X depends on its specific relation with Y and needs to be discretized with finer partitions in **(A)** at low X values for which $Y \simeq \log X$ varies the most and in **(B)** at higher X values for $Y \simeq X^5$.
(EPS)

S4 Fig. Mutual information estimation for Gaussian bivariate distributions. 100 bivariate normal distributions were sampled for varying sample sizes, increasing from top to bottom, and correlation coefficients ρ ranging from 0.01 to 0.9. The mutual information was estimated with the proposed optimum discretization scheme and the KSG estimator with different parameters k. The mean squared error (center graphs) was calculated thanks to the analytical result of the mutual information of the bivariate Gaussian: $I(X; Y) = -\log(1 - \rho^2)/2$. The standard deviation of each estimator over the 100 replications was also plotted against the correlation coefficient (right).
(EPS)

S5 Fig. Mutual information estimation of mixed variables. Experiment set-ups and analytical values for the mutual information were taken fom [7] and 50 runs were performed for each sample size N. Our proposed approach is compared to a naive equal-frequency discretization with $N^{1/3}$ bins, a kernel and a noisy KSG estimator as implemented in JIDT [38], as well as the recent KSG extensions for estimating the mutual informmation between a categorical and a continuous variable (mixed KSG Ross [6]), and between mixed-type variables (mixed KSG Gao [7]). For all nearest-neighbour based approaches, the number of nearest neighbours was set to $k = 5$. From left to right, top to bottom, the simulations are devised after experiment I, experiment II, experiment IV with $p = 0$ and experiment IV with $p = 0.15$, from [7].
(EPS)

S6 Fig. Conditional mutual information estimation for multivariate Gaussian distributions. Four-dimensional normal distributions $P(X, Y, Z_1, Z_2)$ were sampled for $N = 100$ to 5, 000 samples 100 times for each correlation coefficient $\rho = \rho_{XY}$, chosen between 0.05 and 0.95. The other pairwise correlation coefficients were fixed as $\rho_{XZ_1} = \rho_{XZ_2} = \rho_{YZ_1} = \rho_{YZ_2} = \lambda = 0.7$ and $\rho_{Z_1Z_2} = 0.9$. The conditional mutual information $I(X; Y|Z_1, Z_2)$ was then estimated using the proposed optimum partitioning scheme as well aswith kNN conditional information estimates as in S4 Fig. ρ values closed to zero, mimick "V-structures" as they correspond to pairwise independence but conditional dependence; by constrast $\rho = 2\lambda^2/(1 + \rho_{Z_1Z_2}) \simeq 0.5158$ corresponds to conditional independence, while $\rho > 0.5158$ impliesthat X and Y share more information than the indirect flow through Z_1 and Z_2. The analytical value of the conditional mutual information is derived as follows; given the 4×4 covariance matrix Σ and its four 2×2 partitions Σ_{ij}, we first compute the conditional covariance matrix $\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ where Σ_{22}^{-1} is the generalized inverse of Σ_{22}. The partial correlation between X and Y is obtained as $\rho_{XY \cdot Z_1Z_2} = \bar{\Sigma}_{12}/\sqrt{\bar{\Sigma}_{11} * \bar{\Sigma}_{22}}$, and the analytical conditional mutual information for a multivariate normal distribution is given by $I(X; Y|Z_1, Z_2) = -\log(1 - \rho^2_{XY \cdot Z_1Z_2})/2$.
(EPS)

S7 Fig. Pairwise dependence and conditional independence between *X* and *Y* sharing a common cause *Z*. This example illustrates the (conditional) correlation patterns emerging from the presence of a confounding variable, as depicted by the causal diagram $X \leftarrow Z \rightarrow Y$. Z is generated with a uniform law $U(0, 1)$ for $N = 1, 000$ observations and X, Y are both defined as $2Z + \epsilon$ with independent normal noise $\epsilon \sim \mathcal{N}(0, 0.2)$. **(A)** optimum discretization maximizing $I'_N(X; Y)$ with a strong pairwise correlation, and **(B)** optimum discretization which maximizes the conditional mutual information with finite size correction, $I'_N(X; Y|Z)$. In the latter

case, the optimum discretization scheme results in a single bin on both variables as the flow information between X and Y is blocked by conditioning on the common cause Z.
(EPS)

S8 Fig. Pairwise dependence and conditional independence between non Gaussian X and Y sharing a common categorical cause. Another confounding example, $X \leftarrow Z \rightarrow Y$, taken from [25] with a uniform categorical Z with three levels, X and Y being continuous, for $N = 1,000$ observations. With Z_i the binary variable corresponding to the i-th dummy variable of Z, we defined $X = -Z_1 + Z_2 + 0.2\epsilon_X$ which is centered around either -1 if $Z = 1$, 0 if $Z = 3$ or 1 if $Z = 2$; and $Y = Z_1 + Z_2 + 0.2\epsilon_Y$, $\epsilon \sim \mathcal{N}(0, 1)$ which is centered around either 0 if $Z = 3$ or 1 if $Z = 1$ or $Z = 2$. As for continuous common cause in S7 Fig, there is **(A)** some non-zero mutual information between X and Y corresponding to an optimum discretization, while **(B)** conditional mutual information vanishes when conditioning on the categorial common cause, Z, with the partitions of both X and Y variables consisting in a single bin.
(EPS)

S9 Fig. Pairwise independence and conditional dependence with a v-structure. Example of two independent variables X, Y both causing a third variable Z as: $X \rightarrow Z \leftarrow Y$. $N = 1,000$ observations are drawn for $X, Y \sim \mathcal{N}(0, 1)$ and $Z = X + Y$. **(A)** The two variables X and Y being independent, no multi-bin discretization can be found to yield an information estimate that is greater than the corresponding complexity cost. However, **(B)** conditioning on the common effect Z 'activates' the v-structure path generating a spurious relationship between X and Y. This is reflected in the fact that the induced interaction between X and Y requires a multiple bin optimum discretization to estimate $I_N(X; Y|Z) = 1.188$ (with $I'_N(X; Y|Z) = 0.745$).
(EPS)

S10 Fig. Example of dataset generated for mixed-type, non-linear, non-Gaussian benchmarking with 69 continuous and 31 categorical variables. Each plot represents the observed density or histogram ($N = 1,000$) of the continuous or categorical variable X_i, constructed by structural equation models given its parents' distributions (see Supporting Information).
(EPS)

S11 Fig. Skeleton assessment of benchmark networks for mixed-type, non-linear, non-Gaussian datasets. Skeleton Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100$–5,000 samples (see histogram example Fig. S11). Performances obtained with our parameter-free information-theoretic approach MIIC (magenta) are compared to the results obtained with the best parameterization (maximizing the skeleton F-score) of CausalMGM [24] (blue) and MXM [25] (green). See Supporting Information.
(EPS)

S12 Fig. CPDAG assessment of benchmark networks for mixed-type, non-linear, non-Gaussian datasets. CPDAG Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100$–5,000 samples (see histogram example S11 Fig). Performances obtained with our parameter-free information-theoretic approach MIIC (magenta) are compared to the results obtained with the best parameterization (maximizing the CPDAG F-score) of CausalMGM [24] (blue) and MXM [25] (green). See Supporting Information.
(EPS)

S13 Fig. Example of dataset used for continuous, non-linear, non-Gaussian benchmarking with 100 continuous variables.
(EPS)

S14 Fig. Skeleton assessment of benchmark networks for continuous, non-linear, non-Gaussian datasets. Skeleton Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100 - 10,000$ samples (see histogram example Fig. S14). Results obtained with our parameter-free information-theoretic approach MIIC are compared for optimum non-uniform bin sizes and for equal frequency bin sizes (with $N^{1/3}$ bins) as well as to the best results obtained with alternative continuous data methods: PC with Gaussian conditional independence test, rankPC and rankFCI from the `pcalg` package [27], kPC with theHelbert-Schmidt Independence Criterion [39, 40] and CAM [26] algorithms, after optimizing their respective parameter (α) for each sample size N. See Supporting Information.
(EPS)

S15 Fig. CPDAG assessment of benchmark networks for continuous, non-linear, non-Gaussian datasets. CPDAG Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100 - 10,000$ samples (same simulation settings as in Fig. S15).
(EPS)

Acknowledgments

We thank Etienne Birmelé, Pierre Charbord, Eric Gaussier, Gregory Nuel, Elisabeth Remy, Denis Thieffry for discussions.

Author Contributions

Conceptualization: Hervé Isambert.

Data curation: Louis Verny, Nadir Sella, Marc Verny.

Formal analysis: Hervé Isambert.

Funding acquisition: Hervé Isambert.

Investigation: Vincent Cabeli, Louis Verny, Nadir Sella, Guido Uguzzoni, Marc Verny, Hervé Isambert.

Methodology: Vincent Cabeli, Guido Uguzzoni, Hervé Isambert.

Project administration: Hervé Isambert.

Resources: Vincent Cabeli, Louis Verny, Nadir Sella, Marc Verny.

Software: Vincent Cabeli, Nadir Sella.

Supervision: Marc Verny, Hervé Isambert.

Validation: Vincent Cabeli, Louis Verny, Guido Uguzzoni, Marc Verny.

Visualization: Vincent Cabeli, Nadir Sella.

Writing – original draft: Vincent Cabeli, Louis Verny, Hervé Isambert.

Writing – review & editing: Hervé Isambert.

References

1. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Wiley; 2006.
2. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E. 2004; 69:066138. https://doi.org/10.1103/PhysRevE.69.066138
3. Frenzel S, Pompe B. Partial Mutual Information for Coupling Analysis of Multivariate Time Series. Phys Rev Lett. 2007; 99:204101. https://doi.org/10.1103/PhysRevLett.99.204101 PMID: 18233144
4. Vejmelka M, Paluš M. Inferring the directionality of coupling with conditional mutual information. Phys Rev E. 2008; 77:026214. https://doi.org/10.1103/PhysRevE.77.026214
5. Tsimpiris A, Vlachos I, Kugiumtzis D. Nearest neighbor estimate of conditional mutual information in feature selection. Expert Systems with Applications. 2012; 39(16):12697–12708. https://doi.org/10.1016/j.eswa.2012.05.014
6. Ross BC. Mutual information between discrete and continuous data sets. PloS one. 2014; 9(2):e87357. https://doi.org/10.1371/journal.pone.0087357 PMID: 24586270
7. Gao W, Kannan S, Oh S, Viswanath P. Estimating mutual information for discrete-continuous mixtures. In: Advances in neural information processing systems; 2017. p. 5986–5997.
8. Zeng X, Xia Y, Tong H. Jackknife approach to the estimation of mutual information. Proceedings of the National Academy of Sciences. 2018; 115(40):9956–9961. https://doi.org/10.1073/pnas.1715593115
9. Runge J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: Storkey A, Perez-Cruz F, editors. Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. vol. 84 of Proceedings of Machine Learning Research. Playa Blanca, Lanzarote, Canary Islands: PMLR; 2018. p. 938–947.
10. Rissanen J. Modeling by shortest data description. Automatica. 1978; vol. 14:465–471. https://doi.org/10.1016/0005-1098(78)90005-5
11. Kontkanen P, Myllymäki P. MDL Histogram Density Estimation. Journal of Machine Learning Research. 2007; 2:219–226.
12. Verny L, Sella N, Affeldt S, Singh PP, Isambert H. Learning causal networks with latent variables from multivariate information in genomic data. PLoS Comput Biol. 2017; 13(10):e1005662. https://doi.org/10.1371/journal.pcbi.1005662 PMID: 28968390
13. Sella N, Verny L, Uguzzoni G, Affeldt S, Isambert H. MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. Bioinformatics. 2018; 34(13):2311–2313. https://doi.org/10.1093/bioinformatics/btx844 PMID: 29300827
14. Affeldt S, Isambert H. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015; 2015. p. 42–51.
15. Affeldt S, Verny L, Isambert H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. BMC Bioinformatics. 2016; 17(S2):12. https://doi.org/10.1186/s12859-015-0856-x PMID: 26823190
16. Roos T, Silander T, Kontkanen P, Myllymäki P. Bayesian Network Structure Learning using Factorized NML Universal Models. In Proc ITA'08. 2008;.
17. Kontkanen P, Myllymäki P. A linear-time algorithm for computing the multinomial stochastic complexity. Inf Process Lett. 2007; 103(6):227–233. https://doi.org/10.1016/j.ipl.2007.04.003
18. Szpankowski W. Average case analysis of algorithms on sequences.: John Wiley & Sons; 2001.
19. Kontkanen P, Buntine W, Myllymäki P, Rissanen J, Tirri H. Efficient computation of stochastic complexity. in: C Bishop, B Frey (Eds) Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics. 2003;103:233–238.
20. Kontkanen P. Computationally Efficient Methods for MDL-Optimal Density Estimation and Data Clustering; 2009.
21. Zhao J, Zhou Y, Zhang X, Chen L. Part mutual information for quantifying direct associations in networks. Proceedings of the National Academy of Sciences. 2016; 113(18):5130–5135. https://doi.org/10.1073/pnas.1522586113
22. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting Novel Associations in Large Data Sets. Science. 2011; 334(6062):1518–1524. https://doi.org/10.1126/science.1205438 PMID: 22174245
23. Pearl J. Causality: models, reasoning and inference. 2nd ed. Cambridge University Press; 2009.

24. Sedgewick AJ, Buschur K, Shi I, Ramsey JD, Raghu VK, Manatakis DV, et al. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. Bioinformatics. 2018.

25. Tsagris M, Borboudakis G, Lagani V, Tsamardinos I. Constraint-based causal discovery with mixed data. International Journal of Data Science and Analytics. 2018; 6(1):19–30. https://doi.org/10.1007/s41060-018-0097-y PMID: 30957008

26. Bühlmann P, Peters J, Ernest J. CAM: Causal additive models, high-dimensional order search and penalized regression. The Annals of Statistics. 2014; 42(6):2526–2556. https://doi.org/10.1214/14-AOS1260

27. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the R package pcalg. J Stat Softw. 2012; 47(11):1–26. https://doi.org/10.18637/jss.v047.i11

28. Senard JM, Raï S, Lapeyre-Mestre M, Brefel C, Rascol O, Rascol A, et al. Prevalence of orthostatic hypotension in Parkinson's disease. Journal of Neurology, Neurosurgery & Psychiatry. 1997; 63(5):584–589. https://doi.org/10.1136/jnnp.63.5.584

29. Papathanasiou ND, Boutsiadis A, Dickson J, Bomanji JB. Diagnostic accuracy of ^{123}I-FP-CIT (DaTSCAN) in dementia with Lewy bodies: a meta-analysis of published studies. Parkinsonism & related disorders. 2012; 18(3):225–229. https://doi.org/10.1016/j.parkreldis.2011.09.015

30. Tounsi H, Deweer B, Ergis AM, Van der Linden M, Pillon B, Michon A, et al. Sensitivity to semantic cuing: an index of episodic memory dysfunction in early Alzheimer disease. Alzheimer Dis Assoc Disord. 1999; 13(1):38–46. https://doi.org/10.1097/00002093-199903000-00006 PMID: 10192641

31. Teichmann M, Epelbaum S, Samri D, Nogueira ML, Michon A, Hampel H, et al. Free and Cued Selective Reminding Test—accuracy for the differential diagnosis of Alzheimer's and neurodegenerative diseases: A large-scale biomarker-characterized monocenter cohort study (ClinAD). Alzheimer's & Dementia. 2017; 13(8):913–923. https://doi.org/10.1016/j.jalz.2016.12.014

32. Fazekas F, Chawluk J, Alavi A, Hurtig H, Zimmerman R. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. Am J Roentgenology. 1987; 149(2):351–356. https://doi.org/10.2214/ajr.149.2.351

33. Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, et al. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. Journal of Neurology, Neurosurgery & Psychiatry. 1992; 55(10):967–972. https://doi.org/10.1136/jnnp.55.10.967

34. Fiford CM, Manning EN, Bartlett JW, Cash DM, Malone IB, Ridgway GR, et al. White matter hyperintensities are associated with disproportionate progressive hippocampal atrophy. Hippocampus. 2017; 27(3):249–262. https://doi.org/10.1002/hipo.22690 PMID: 27933676

35. Prins ND, Scheltens P. White matter hyperintensities, cognitive impairment and dementia: an update. Nature Reviews Neurology. 2015; 11(3):157–165. https://doi.org/10.1038/nrneurol.2015.10 PMID: 25686760

36. Furlanello C, Albanese D, Jurman G, Filosi M, Visintainer R, Riccadonna S. minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics. 2012; 29(3):407–408. https://doi.org/10.1093/bioinformatics/bts707 PMID: 23242262

37. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences. 2014; 111(9):3354–3359. https://doi.org/10.1073/pnas.1309933111

38. Lizier JT. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. Frontiers in Robotics and AI. 2014; 1:11. https://doi.org/10.3389/frobt.2014.00011

39. Gretton A, Herbrich R, Smola A, Bousquet O, Schölkopf B. Kernel methods for measuring independence. Journal of Machine Learning Research. 2005; 6(Dec):2075–2129.

40. Gretton A, Spirtes P, Tillman RE. Nonlinear directed acyclic structure learning with weakly additive noise models. In: Advances in neural information processing systems; 2009. p. 1847–1855.

SUPPORTING INFORMATION

for manuscript

Learning clinical networks from medical records based on information estimates in mixed-type data

Vincent Cabeli, Louis Verny, Nadir Sella, Guido Uguzzoni, Marc Verny, Hervé Isambert

Supplementary Materials and Methods

Benchmark data generation

In this section we describe the generation of datasets used for the mixed-type (Fig. 3, Figs. S11 and S12) and continuous (Figs. S14 and S15) benchmarks, and implemented in the R script provided as supplementary material. First, the underlying DAG models were randomly drawn from the space of all possible DAGs [1], allowing for a maximum degree of 4 neighbours. Datasets were generated following the causal order of the generated DAG using non-linear structural equations models (SEMs), as outlined below.

The first nodes in the causal order have no parents, their distributions are sampled either from Gaussian mixtures of 1 to 5 modes (with equal σ) for continuous nodes or with a uniform random sampling of 2 to 4 categorical levels. The distribution of every other node X was generated as a function of its parents $\mathrm{Pa}(X)$ plus some Gaussian noise as, $X = f(\mathrm{Pa}(X)) + \epsilon$. Depending on whether X and its parents are continuous or categorical, different models were used:

- **Continuous variable** X

  The causal relationship between a continuous node X and its continuous parents $\mathrm{Pa}_c(X)$ plus their pairwise interaction products $I(\mathrm{Pa}_c(X))$ was modeled using polynomials: $X = R(\sum_{Y_i \in \{\mathrm{Pa}_c(X) \cup I(\mathrm{Pa}_c(X))\}} R(Y_i, -1, 1)^{c_i} + \epsilon, 0, 1)$ with c_i chosen in $[1, 3]$, ϵ some Gaussian noise with variance depending on the number of parents and c_i, and $R(X, \min, \max)$ a re-scaling function so that the distribution X is in the range $[\min, \max]$. In the case of mixed-type parents, *i.e.* with some continuous and some discrete parent variables, sets of c_i were drawn for each combination of the discrete parents $\mathrm{Pa}_d(X)$. If all its parents are categorical, a child node is categorical as well. Finally, the distribution of a continuous node has an equal probability to be transformed with a non-linear function, e^X, $\sin(X)$ or $\cos(X)$, or to be retained as is.

- **Discrete variable** X

  The continuous parents of a discrete node are first discretized by attributing categorical levels to the distinct peaks if there are any (see Fig. S13), or using equal frequency binning with $\log(N)$ bins otherwise. The discrete distribution of the node X is then drawn from random sampling with probability w_i for the ith level of X, where each combination of the levels of $\mathrm{Pa}_d(X)$ are associated to a different set of probabilities $\{w_i\}$.

Performance measures

For the evaluation, the network reconstruction was treated as a binary classification task and classical performance measures, precision, recall and F-score, were used, based

on the numbers of true *versus* false positive (TP *vs* FP) edges and true *versus* false negative (TN *vs* FN) edges. The precision $Prec = TP/(TP + FP)$ indicates how reliable the edges of the reconstructed network are. This measure does not indicate, however, which fraction of the true edges are detected, which corresponds to the sensitivity or recall of the reconstruction, $Rec = TP/(TP + FN)$. Finally, the F-score is a global performance measure, which is defined as the harmonic mean of precision and recall measures: $Fscore = 2Prec \times Rec/(Prec + Rec)$. In particular, a Fscore of 1 implies a perfect reconstruction without FP nor FN edges.

In order to measure how well the orientations of the edges match those of the true DAG, we also define the orientation-dependent counts $TP' = TP - TP_{misorient}$ and $FP' = FP + TP_{misorient}$ with $TP_{misorient}$ corresponding to all true positive edges of the skeleton with different orientation/non-orientation status as in the true Complete Partially Directed Acyclic Graph (CPDAG). Here, CPDAG refers to the equivalence class of the true DAG, which is taken as the benchmark reference since different DAGs might be equivalent from the data point of view (*i.e.* if and only if they have the same skeleton and the same v-structures). The CPDAG precision, recall and F-score were then computed with the orientation-dependent TP' and FP'.

Benchmark parameter tuning

The performances of some methods rely on tunable parameters which typically determine the sparsity of the inferred graph. In contrast, miic uses a complexity term derived from the normalised maximum likelihood and is essentially parameter-free. Although in real world applications the best settings cannot be known for certain, meaningful comparisons can only be done after each method has been properly parameterized. Here we detail the steps taken to find the best parameters for each benchmark setting.

For the mixed-type benchmarks, ranges of parameters for both CausalMGM [2] and MXM [3] methods were tested, and their best results (*i.e.* best F-scores) obtained for a given sample size (N) and percentage of continuous node (p_c) were compared to miic results. For CausalMGM, the λ sparsity parameter for all edge types (discrete-discrete, continuous-continuous, discrete-continuous) was tested in $\{0.050, 0.073, 0.108, 0.158, 0.232, 0.341, 0.500\}$. For MXM, the significance threshold α used for the various independence tests was tested in $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.2\}$.

For the continuous benchmarks, we first optimized each method on separate simulations to find a good approximating function for the best parameter $\hat{\alpha} = f_p(N)$. The best values for the α_N parameter of PC gaussian, PC rank, CAM for sample sizes N spaced evenly on a log scale between 100 and $10,000$ were first found using a zeroth order parameter optimization implemented in `dlib` [4,5]. Then, the function f_p was fitted as a second order polynomial over all values of N and α_N. kPC (using the Hilbert-Schmidt independence criterion with gamma approximation [6,7]) was not optimized so extensively, due to its much longer execution time, and was only tested for the conservative values of α: 0.05 and 0.15.

Resource availability

- **MIIC R package** for ***mixed-type data*** is available at this URL: https://miic.curie.fr/download/miic_mixed.tar.gz
- **MIIC online server** for ***mixed-type data*** is accessible here: https://miic.curie.fr/workbench_mixed.php

References

1. Melançon G, Philippe F. Generating connected acyclic digraphs uniformly at random. Information Processing Letters. 2004;90(4):209–213.

2. Sedgewick AJ, Buschur K, Shi I, Ramsey JD, Raghu VK, Manatakis DV, et al. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. Bioinformatics. 2018;doi:10.1093/bioinformatics/bty769.

3. Tsagris M, Borboudakis G, Lagani V, Tsamardinos I. Constraint-based causal discovery with mixed data. International Journal of Data Science and Analytics. 2018;6(1):19–30.

4. Malherbe C, Vayatis N. Global optimization of lipschitz functions. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. p. 2314–2323.

5. King DE. Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research. 2009;10:1755–1758.

6. Gretton A, Herbrich R, Smola A, Bousquet O, Schölkopf B. Kernel methods for measuring independence. Journal of Machine Learning Research. 2005;6(Dec):2075–2129.

7. Gretton A, Spirtes P, Tillman RE. Nonlinear directed acyclic structure learning with weakly additive noise models. In: Advances in neural information processing systems; 2009. p. 1847–1855.

Chapter 4

Improvements to constraint-based algorithms and MIIC

In this chapter, we will discuss other contributions to both constraint-based methods and MIIC specifically to be able to deal with imperfect real-life situations.

In the first section, we briefly discuss the presence of missing data and its implications in causal graph reconstruction with constraint-based methods. We then propose our solution to deal with missing data, based on test-wise deletion and an information-theoretic test to accept or reject potential conditioning nodes. Next, we introduce improved orientation scores for MIIC, adapted for larger sample sizes, and the concept of "putative" versus "genuine" causal orientations, based on the orientation probabilities of both the head and the tail of the edge $X \to Y$. We also mention the advantages of interactive visualisation of the results, by presenting the updated MIIC webserver.

Finally, we introduce two papers, [6] published at NeurIPS 2019 in which we aim to make constraint-based methods more interpretable with regard to the choice of separating nodes, making them more *consistent* with the final graph. Another paper accepted for publication at the Why-21 conference introduces conservative MIIC, which infers more reliable orientations thanks to modifications to the mutual information estimation.

4.1 Improvements

4.1.1 Handling missing data

One dimension of observational data that has not been discussed yet is the problem of missing data. Because the data collection was unreliable, or because the variable we are interested is note defined for all cases, some samples may have undefined values. There are typically two ways to deal with those missing values in a dataset. First, one can omit all data affected by missingness and perform the analysis on the subset of complete samples. Using this approach

we lose a lot of partial but valuable information, it is not ideal on *e.g.* bio-medical datasets, often small or medium size, and for which having a reliable and systematic observation of all variables is difficult.

The second way to deal with missing data is to impute them using the rest of the observations. If the missing values can be estimated correctly, then one can analyse the data with the same power as the full data analysis would have had. This is the preferred way in many applications, but must make assumptions on the distribution of the missing data $P(X_{Mis})$. We can distinguish 3 mechanisms by which the data is missing, identified by Rubin in his seminal work [111] :

- Missing Completely at Random (MCAR) : the missingness mechanism is completely random, it does not depend on any other value : $P(X_{Obs}) = P(X_{Mis})$ and $P(X_{Mis}) \perp\!\!\!\perp P(V)$.
- Missing at Random (MAR) : unlike the name suggests, the missing mechanism can be biased ($p(X_{Mis}) \not= p(X_{Obs})$) on the condition that it can be explained by the observed data.
- Missing Not at Random (MNAR) : the probability of the missing data cannot be deduced from observations.

As Little and Rubin observe in [112], essentially all work on multivariate incomplete data, including imputation, makes at least the MAR assumption. It is however often problematic as there is no recognizable criterion for MAR from a dataset alone [113]. Another issue with data imputation is the prediction of the values itself, which depends on the amount of information $\mathrm{I}(X; V_{\setminus X})$ and is a difficult problem of its own, especially for continuous distributions. As is often the case when analyzing real data, *there is no free lunch* when dealing with missing samples. Rubin himself concluded his article [111] saying that the only correct solution is to act on a case-by-case basis, preferably modeling explicitly the missingness mechanism.

Interesting work has been done in this direction by Mohan and Pearl [114] using graphical models called missingness graphs. As suggested by Rubin, the idea is to model the missingness processes explicitly, using Pearl's causal graphs, from which one can easily deduce whether the data is MCAR, MAR or MNAR. Later, the authors also showed that certain properties of the distribution may be recoverable when data is partially observed, including conditional independence relations [113].

But how can we apply those findings to improve graph reconstruction with incomplete data ? This topic has been investigated for some time, for example [115] introduced a variant of PC using a pseudo-Bayesian test of independence which relies on the local graph and its parameters. However, it needs to represent the data with contingency tables, making it unusable for the mixed case (generally, methods relying on data inference through the

Bayesian graph are difficult to implement in the non-parametric setting). Similarly, [116] uses the expectation maximisation principle to learn both the graph and its parameters with incomplete data.

These solutions are much harder to implement for the general case, *i.e.* with as little restriction as possible on the distribution $P(V)$. We will instead focus on *test-wise* omission of missing samples, removing only what is necessary to have complete support for each conditional independence test. Strobl et al. developed a scheme combining FCI and test-wise deletion that is still able to recover the PAG from incomplete data, even when MNAR holds, provided that no missingness mechanisms causally affect each other [117]. This last condition is discussed in more details in [118], which focuses on rectifying erroneous edges produced by test-wise deletion PC when MAR or MNAR holds. However, their method is not very well suited to our setting as it needs either the residuals of a linear regression model or estimates of the full data distribution via kernel density estimation.

What we propose here is a simple information-theoretic rule for rejecting or accepting potential conditioning nodes during the skeleton phase of constraint-based algorithms. Specifically, we want to avoid removing edges because of selection bias as opposed to "explaining away" the direct correlation via other information flow in the causal graph. Our reasoning is that as much as possible, conditional independencies should be read off the inferred graph $\mathcal{G}_{inf}$, *i.e.* X and Y are not adjacent if and only if $X \perp\!\!\!\perp Y|Z$ (regardless of missingness). We compare joint the distributions X,Y before and after removing the samples that are missing for a potential separating node Z, respectively noted (X,Y) and $(X,Y|Z_{Obs})$. If $(X,Y|Z_{Obs})$ is too different from (X,Y), then we do not accept Z in the conditioning set of $X-Y$, as the observed interaction on the reduced support is not representative of the full data. Note that in contrast with [118], this scheme can only add back edges in $\mathcal{G}_{inf}$. We argue that it makes the result more interpretable in relation to the way constraint-based methods operate : starting from the complete graph, we remove the edge $X-Y$ only if there is evidence in the data that the link is either non existent ($X \perp\!\!\!\perp Y$) or indirect ($X \perp\!\!\!\perp Y|Z$). If there is no such evidence, or if we cannot accept it due to potential selection bias, the edge $X-Y$ stays in $\mathcal{G}_{inf}$. Moreover, it is a simple rule that does not require any assumption about the data distribution, using an information-theoretic measure that fits well with the rest of the MIIC algorithm.

This problem is a version of the two-sample test, which aims to determine if two samples come from the same population. It can be naturally approached using the KL divergence (Eq 3.1.7) [119], which we redefine here for the discrete and continuous cases :

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{4.1.1}$$

$$= \int_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \tag{4.1.2}$$

where P and Q are two distributions defined on the same space $\mathcal{X}$.

We can then restate our goal using this divergence : during the skeleton phase, test for conditional independence $X \perp\!\!\!\perp Y|Z$ only if

$$D_{\mathrm{KL}}((X,Y \mid Z_{Obs}) \parallel (X,Y)) < t_{\mathrm{KL}} \tag{4.1.3}$$

where t_{KL} is the threshold for how much divergence we tolerate.

Unsurprisingly, the approaches to estimate this divergence from samples are similar to the ones introduced for entropy and mutual information. On discrete data, one can simply use the observed frequencies and use a plug-in estimator with Eq 3.1.7 [64]. It may be tempting to re-use the optimal discretization found by optimizing $\hat{\mathrm{I}}'(X;Y)$, and simply use the plug-in estimator for Eq 4.1.3. But optimizing $\hat{\mathrm{I}}'(X;Y|Z)$ may give a very different discretization, depending on the interaction with Z which is not represented by optimizing on (X,Y) alone. Additionally, MIIC takes off the contributors one by one for each edge $X-Y$, so we do not know the full set U_i in advance. For these reasons, Eq 4.1.3 should be computed with the full data, not on reduced discretized versions, to be able to account for future discretizations. However, much like for the mutual information, estimating D_{KL} on continuous and mixture variables is a challenge.

On continuous data, the best-behaved estimator seems to be a k-nn scheme introduced in [120] : with P and Q two samples respectively defined on the spaces $\mathcal{X}$ and $\mathcal{X}'$, of length n and m samples on d dimensions :

$$\hat{D}_{\mathrm{KL}}(P \parallel Q) = d \left\langle \log \frac{r_k(x_i)}{s_k(x_i)} \right\rangle_n + \log \frac{m}{n-1} \tag{4.1.4}$$

where $r_k(x_i)$ and $s_k(x_i)$ are respectively the distance to the kth nearest neighbor from the point x_i in $\mathcal{X}$ and $\mathcal{X}'$, and $\left\langle \log \frac{r_k(x_i)}{s_k(x_i)} \right\rangle_n$ is the average taken over all n samples i.

Our proposed estimator is inspired by mixed estimator of mutual information [71, 51] and treats each case differently. When X and Y are discrete, the plug-in estimator is used with the observed joint counts X,Y. On two continuous samples, we use the k-nn estimator $\hat{D}_{\mathrm{KL}}$ of Eq 4.1.4 with a fixed $k=5$. When X is discrete and Y is continuous, the estimator needs to be a bit more involved. Just like [71], we sum partial terms over all levels r_x of the discrete variable X :

$$\hat{D}_{\mathrm{KL}}(X,Y|Z_{Obs} \parallel X,Y) = \frac{d}{n} \sum_{r_x \in X} n_{r_x} \left\langle \log \frac{r_k(x_i)_{r_x}}{s_k(x_i)_{r_x}} + \log \frac{m_{r_x}}{n_{r_x}-1} \right\rangle_{n_{r_x}} + \hat{D}_{\mathrm{KL}}(X|Z_{Obs} \parallel X) \tag{4.1.5}$$

with $r_k(x_i)_{r_x}$ and $s_k(x_i)_{r_x}$ respectively the distance to the kth nearest neigbhor in the space of Y in the subsample Z_{Obs} and in the full data, m_{r_x} and n_{r_x} the total number of points with the same discrete value r_x in the full data and in the subsample. Correspondingly, the average is

taken over the samples n_{r_x}. $\hat{D}_{\mathrm{KL}}(Y|Z_{Obs} \parallel Y)$ is the divergence between the frequencies of the discrete levels X.

The implementation in MIIC uses an efficient k nearest neighbor scheme using KD-tree index, implemented in nanoflann [121]. The complexity of finding the nearest neighbor using KD-tree is $O(\log n)$ on average and $O(n)$ in the worst case. Finally, to deal with mixture variables, it adds random uniform noise to break up discrete points (just like the noisy KSG estimator). The distances r_k and s_k are in the ℓ^2 norm, which was shown to improve k-nn estimates [51].

Having defined an estimator for the general case, all that remains is to choose divergence threshold t_{KL} in Eq 4.1.3. Of course, the ideal threshold would be adapted to the data in order to control for false-postive and false-negative rates, which can be done for information theoretic values [122]. We propose a simple heuristic based on the MDL principle and the BIC, and compare $|Z_{Obs}| \cdot \hat{D}_{\mathrm{KL}}(X,Y|Z_{Obs} \parallel X,Y)$ to $t_{\mathrm{KL}} = \log|Z_{Obs}|$. Note that it does not behave like a p-value : a smaller, noisier subsample will necessarily create more diverging distributions even under the null hypothesis (MAR), and a fixed p-value will correspond to higher values of D_{KL}; whereas the threshold $\log|Z_{Obs}|$ becomes more stringent as the subsample decreases in size. This reflects the fact that smaller samples are less representative of the full data, and so contain less information to remove the edge $X - Y$. The choice of a better threshold is left as a perspective for future research in the group, using this heuristic the goal is to avoid the worst cases of selection bias.

4.1.2 Orientation probability for large N, putative versus genuine orientations

While constraint-based methods can in principle learn the presence or absence of orientation of individual edges from the available data, the orientation of a V-structure, $X \to Z \leftarrow Y$ (and $X \nsim Y$) corresponds in fact to the discovery of 'putative causality' as one cannot rule out *a priori* that the edge between X (or Y) and Z is not due to the effect of a latent common cause, L, unobserved in the dataset, $X \dashleftarrow L \dashrightarrow Z$. In order to discover a genuine cause-effect relations explaining at least part of the association between X and Z, $X \to Z$, one needs to exclude the possibility of such a latent variable, L, between X and Z. We can think of "genuine" orientation in constraint-based methods as setting both the "head" of the edge $X \to Y$ and its "tail", excluding latent common causes. Using the MIIC framework, genuine causal edges are then predicted if the head and tail probabilities are statistically significant, while causal edges remain "putative" if their tail probability is not statistically significant or cannot be determined from purely observational data (i.e., undirected links in the $\mathcal{G}_c$ equivalence class). This gives a better interpretation of constraint methods on real data, for which it is difficult to ensure with certainty that all variables in the system are observed, and

thus that the directed links of $\mathcal{G}_{inf}$ are "genuine".

We now outline the principles to uncover cause-effect relations and distinguish genuine from putative causes through an intuitive toy example of an imaginary dataset of old cars (Fig 4.1). (**A**) The signature of causality in such observational datasets corresponds to 3-variable V-structure motifs involving two *independent* and thus *unconnected* possible causes, "Broken fuel pump?" and "Discharged battery?", and a resulting effect, "Broken down car?". The converging orientations of this v-structure towards its middle variable, "Broken down car?", stem from the fact that these two edges cannot be undirected, nor can they point towards either "Broken fuel pump?" or "Discharged battery?", as these alternative graphical models would imply correlations contradicting the independence between "Broken fuel pump?" and "Discharged battery?". In practice, such independences between possible causes might in fact be conditional on other variable(s), not considered here. (**B**) Note, however, that v-structures only identify, "putative" causes, which might not be "genuine" causes; for instance, the variable "Clock stopped?", which can be used as a proxy for the variable "Discharged battery?", also forms a v-structure with the other independent putative cause "Broken fuel pump?". Yet, we know that "Clock stopped?" cannot be a genuine cause of "Broken down car?", as tampering with a car's clock cannot actually cause a car to break down. (**C**) In absence of background knowledge, showing that "Discharged battery?" is actually a genuine cause of "Broken down car?" (displayed with a green arrowhead) requires to find another v-structure upstream of "Discharged battery?" (*e.g.* "Lights left on?"$\rightarrow$"Discharged battery?"$\leftarrow$"Old battery?") or to have prior knowledge about an upstream (putative) cause and to show that the effect of these upstream variables on the downstream variable "Broken down car?" is entirely *indirect* and mediated (at least in part) by the intermediary variable "Discharged battery?". This requires to find a conditional independence between upstream and downstream variables conditioned on a separating set including the intermediary variable "Discharged battery". These conditions are needed to exclude the possibility of an unobserved common cause between the intermediary variable ("Discharged battery") and the downstream variable ("Broken down car?"), as illustrated in (D). (**D**) Ruling out a putative cause as genuine cause is done by finding a fourth variable (*e.g.* "Out-of-order clock?") defining another v-structure sharing the edge between "Broken down car?" and "Clock stopped?" with the v-structure in (B). It implies that the relation between these two variables is actually due to a latent common cause unobserved in the dataset (here "Discharged battery?") and represented with a bidirected edge.

Formally, we implement the idea of separate likelihood-based estimation of orientation probability. For an edge $X - Z$, each end point of which is either an arrow head or tail, we denote by p_x (p_z) the probability of the end point at X (Z) being an arrowhead $X \leftarrow Z$ ($X \rightarrow Z$), and by $1 - p_x$ ($1 - p_z$) the probability of the that end being a tail. Undecided head or tail orientations thus correspond to $p = 1 - p = 0.5$. With this notation, we predict

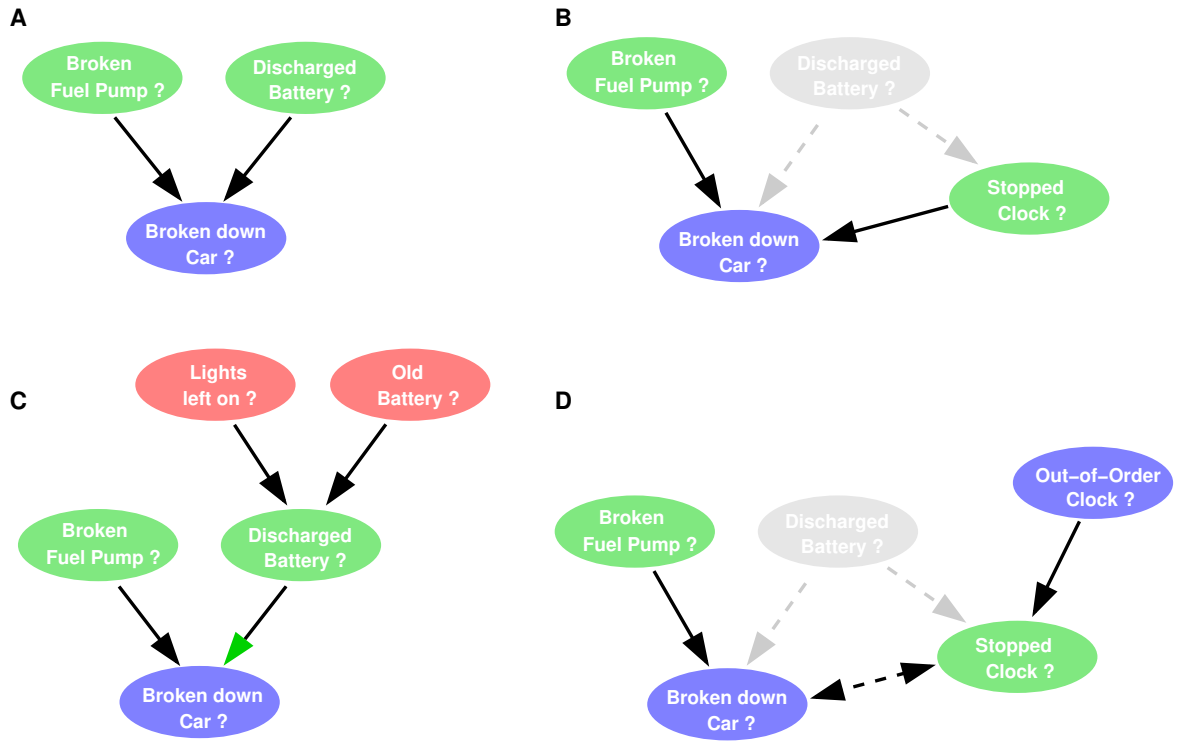

Figure 4.1: Toy example of putative and genuine causal relations.

genuine causal edges, $X \to Z$, highlighted with green arrow head in Fig 4.1, if $p_z > 1/2$ and $p_x < 1/2$, whereas putative causal edges are predicted for $p_z > 1/2$ and $p_x = 1/2$, that is when p_x cannot be decided with the available data. By contrast, undirected (or undecided end) edges are expected for $p_z \leqslant 1/2$ and $p_x \leqslant 1/2$, while bidirected edges, corresponding to the presence of a latent common cause, are predicted for $p_z > 1/2$ and $p_x > 1/2$. Orientation probability estimates are computed below, together with the introduction of an orientation confidence threshold, β, enhancing the precision on arrow head prediction. The present probabilistic framework of edge orientation also allows for enforcing prior knowledge about certain orientations, in particular, when a variable is not freely varying like other variables of the dataset as it corresponds to a control parameter or experimental condition. Such contextual variables will have all their edges without incoming arrow head, *i.e.*, $p_{\text{in}} = 0$, by assumption. This expresses our prior knowledge that contextual variables cannot be the consequence of other observed or non-observed variables as they actually correspond to manually set external parameters or experimental conditions.

Using the orientation threshold $1 > \beta > 0.5$, we can enhance the precision of arrowheads. The condition for predicting genuine causal edges then becomes $X \to Z$, if $p_z > \beta$ and $p_x < 1 - \beta$. By contrast, putative causal edges are predicted for $p_z > \beta$ and $\beta \geqslant p_x \geqslant 1 - \beta$, while undirected edges are expected for $p_z \leqslant \beta$ and $p_x \leqslant \beta$, and bidirected edges, corresponding to the presence of a latent common cause, for $p_z > \beta$ and $p_x > \beta$ (Fig 4.2).

The way to compute the probabilities p_x, p_y and p_z builds on the approach introduced in

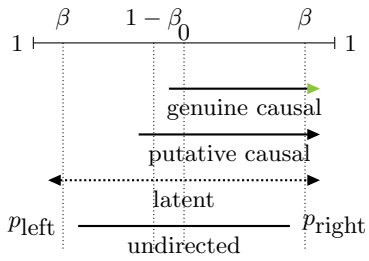

Figure 4.2: Orientation of putative, genuine or latent edges based on an orientation threshold β.

[5], deducing tail / head orientation probabilities from an existing arrowhead $\underline{z} \leftarrow y$:

$$P(x -\!\!* \underline{z}) = P(x -\!\!* \underline{z} | \underline{z} \leftarrow y) P(\underline{z} \leftarrow y) + P(x -\!\!* \underline{z} | \underline{z} - y) P(\underline{z} - y) \tag{4.1.6}$$

where $*$ stands for a tail [resp. head] depending on the positivity [resp. negativity] of $\mathrm{I}'(X;Y;Z|\{A_i\})$ with $X \perp\!\!\!\perp Y | \{A_i\} \cup Z$ [resp. $X \perp\!\!\!\perp Y | \{A_i\}$].

However, using the full probability decomposition above can lead to a higher confidence in tail or head induced probabilities than in the head probabilities they derive from, due to the Markov equivalence of non-V-structures. In addition, induced tail / head probabilities can be numerically difficult to compare for large N. To circumvent these issues and capture the rationale that our confidence in induced tail / head orientations can only be lower than our confidence in the arrowhead from which they derive, we propose to redefine the tail / head induced probabilities by retaining only the first term in the probability decomposition above, that is, by assuming that the arrowhead $\underline{z} \leftarrow y$ exists,

$$\begin{aligned}
P(x -\!\!* \underline{z}) &= P(x -\!\!* \underline{z} | \underline{z} \leftarrow y) P(\underline{z} \leftarrow y) \\
&= \frac{1}{1 + e^{-N|\mathrm{I}'(X;Y;Z\|\{A_i\})|}} P(\underline{z} \leftarrow y) \\
&= \frac{1}{1 + e^{-N|\mathrm{I}'(X;Y;Z\|\{A_i\})|}} \times \frac{1}{1 + e^{-\mathrm{score_v}}} \\
&= \frac{1}{1 + e^{-N|\mathrm{I}'(X;Y;Z\|\{A_i\})|} + e^{-\mathrm{score_v}} + e^{-N|\mathrm{I}'(X;Y;Z\|\{A_i\})| - \mathrm{score_v}}} \\
&= \frac{1}{1 + e^{-\min} + e^{-\max} + e^{-\min-\max}} \\
&= \frac{1}{1 + e^{-\min}(1 + e^{-\max+\min} + e^{-\max})} \\
&= \frac{1}{1 + e^{-\mathrm{score_i}}}
\end{aligned}$$

where we introduce $\mathrm{score_i}$ to enable numerical ordering of orientation probabilities for large N,

$$\text{score}_\text{i} = \min - \log\left(1 + e^{-\max+\min} + e^{-\max}\right) \tag{4.1.7}$$

$$\min = \min\left(N|\mathrm{I}'(X;Y;Z|\{A_i\})|, \text{score}_\text{v}\right) \tag{4.1.8}$$

$$\max = \max\left(N|\mathrm{I}'(X;Y;Z|\{A_i\})|, \text{score}_\text{v}\right) \tag{4.1.9}$$

Hence $0.5 \leqslant p_1 < p_2 < 1$ is equivalent to $0 \leqslant \text{score}_1 < \text{score}_2 < \infty$, where score_1 and score_2 can be numerically ordered even for very large N, unlike p_1 and p_2.

4.1.3 Webserver and interactive visualisation

In addition to an open-source R package, we have developed a web interface to analyze visually the results of MIIC. Constraint-based approaches, and MIIC, being entirely non-parametric, it can help to be able to visualize the joint distributions of the inferred direct links. We implemented various plots for the continuous, discrete and mixed case using the D3 JavaScript library and plotly (Fig A.6). When applicable, it also shows the result of the optimal discretization for a given edge. More examples are shown with the application on medical record of breast cancer patients in Section 5.2. Related to the next section, we made it also intuitive to visually control the validity of separating sets found by MIIC. From the inferred graph, one can easily see if they satisfy d-separation or if they violate *consistency* with respect to $\mathcal{G}_{Inf}$.

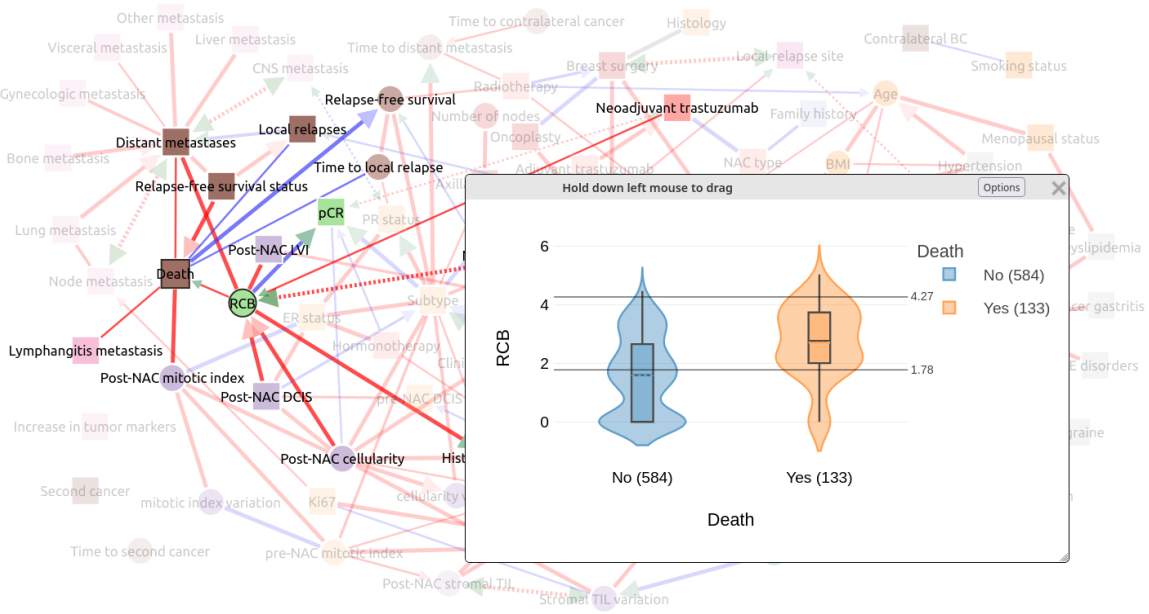

Figure 4.3: Online network inteface. In this example, a violin plot describes the joint distribution between the continuous variable "RCB", and "Death" which is discrete. The horizontal bold lines inform on the optimal discretization found to infer the edge.

4.2 Consistent separating sets

In the same idea as the missing data test, we worked on a variant of constraint-based algorithms that guarantees that the conditioning sets used to remove links are more consistent with the final graph $\mathcal{G}_{inf}$ and the data $\mathcal{D}$ [6]. In their original form, these methods rely only on the conditional independences in $\mathcal{D}$ and offer no guarantee that the separating sets correspond to d-separations sets in the final graph. In fact, they do not even guarantee that they are still in the same connected component in $\mathcal{G}_{inf}$ after the iterative removal of edges.

This flaw not only makes the result less interpretable but also makes the performance worse. These inconsistent conditioning sets tend to come from sampling noise rather than from functional realities, and graphs reconstructed on complex data are typically very sparse. The consistent version of constraint-based algorithms produces a $\mathcal{G}_{inf}$ graph that is less prone to spurious independencies and from which it is easier to infer the sets of condensations used, making the method more interpretable. This extension is particularly adapted to MIIC which removes the contributors in order starting with the best score, compared to reference methods which try all possible combinations until significance is found.

4.2.1 Publication at NeurIPS 2019

Constraint-based Causal Structure Learning with Consistent Separating Sets

Honghao Li, Vincent Cabeli, Nadir Sella, Hervé Isambert*
Institut Curie, PSL Research University, CNRS UMR168, Paris
{honghao.li, vincent.cabeli, nadir.sella, herve.isambert}@curie.fr

Abstract

We consider constraint-based methods for causal structure learning, such as the PC algorithm or any PC-derived algorithms whose first step consists in pruning a complete graph to obtain an undirected graph skeleton, which is subsequently oriented. All constraint-based methods perform this first step of removing dispensable edges, iteratively, whenever a separating set and corresponding conditional independence can be found. Yet, constraint-based methods lack robustness over sampling noise and are prone to uncover spurious conditional independences in finite datasets. In particular, there is no guarantee that the separating sets identified during the iterative pruning step remain consistent with the final graph. In this paper, we propose a simple modification of PC and PC-derived algorithms so as to ensure that all separating sets identified to remove dispensable edges are consistent with the final graph, thus enhancing the explainability of constraint-based methods. It is achieved by repeating the constraint-based causal structure learning scheme, iteratively, while searching for separating sets that are consistent with the graph obtained at the previous iteration. Ensuring the consistency of separating sets can be done at a limited complexity cost, through the use of block-cut tree decomposition of graph skeletons, and is found to increase their validity in terms of actual d-separation. It also significantly improves the sensitivity of constraint-based methods while retaining good overall structure learning performance. Finally and foremost, ensuring sepset consistency improves the interpretability of constraint-based models for real-life applications.

1 Introduction

While the oracle versions of constraint-based methods have been demonstrated to be sound and complete (Zhang, 2008; Spirtes, Glymour, and Scheines, 2000; Pearl, 2009), a major limitation of these methods is their lack of robustness with respect to sampling noise for finite datasets. This has largely limited their use to analyze real-life data so far, although important advances have been made lately, in particular, to limit the order-dependency of constraint-based methods (Colombo and Maathuis, 2014) or to improve their robustness to sampling noise by recasting them within a maximum likelihood framework (Affeldt and Isambert, 2015; Affeldt, Verny, and Isambert, 2016).

However, it remains that constraint-based methods still lack graph consistency, in practice, as they do not guarantee that the learnt structures belong to their presumed class of graphical models, such as a completed partially directed acyclic graph (CPDAG) model for the PC (Spirtes and Glymour, 1991; Kalisch and Bühlmann, 2008; Kalisch et al., 2012) or IC (Pearl and Verma, 1991) algorithms, or a partial ancestral graph (PAG) for FCI or related constraint-based algorithms allowing for unobserved latent variables (Spirtes, Meek, and Richardson, 1999; Richardson and Spirtes, 2002; Colombo et al., 2012; Verny et al., 2017; Sella et al., 2018). By contrast, search-and-score structure learning

*corresponding author

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

methods (Koller and Friedman, 2009) inherently enforce graph consistency by searching structures within the assumed class of graphs, *e.g.*, within the class of directed acyclic graphs (DAG). Similarly, hybrid methods such as MMHC (Tsamardinos, Brown, and Aliferis, 2006) can also ensure graph class consistency by maximizing the likelihood of edge orientation within the class of DAGs.

This paper concerns, more specifically, the inconsistency of separating sets used to remove dispensable edges, iteratively, based on conditional independence tests. This inconsistency arises as some separating sets might no longer be compatible with the final graph, if they were not already incompatible with the current skeleton, when testing for conditional independence during the pruning process. It occurs, for instance, when a node in a separating set is not on any indirect path linking the extremities of a removed edge, as noted in (Spirtes, Glymour, and Scheines, 2000). Such inconsistencies can be seen as a major shortcoming of constraint-based methods, as the primary motivation to learn and visualize graphical models is arguably to be able to read off conditional independences directly from the graph structure (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009).

In the following, we propose a simple modification of PC or PC-derived algorithms so as to ensure that all conditional independences identified and used to remove dispensable edges are consistent with the final graph. It is achieved by repeating the constraint-based causal structure learning scheme, iteratively, while searching for separating sets that are consistent with the graph obtained at the previous iteration, until a limit cycle of successive graphs is reached. The union of the graphs over this limit cycle is then guaranteed to be consistent with the separating sets and corresponding conditional independences used to remove all dispensable edges from the initial complete graph. Enforcing sepset consistency of constraint-based methods is found to limit their tendency to uncover spurious conditional independences early on in the pruning process when the combinatorial space of possible separating sets is still large. As a result, enforcing sepset consistency reduces the large number of false negative edges usually predicted by constraint-based methods (Colombo and Maathuis, 2014) and, thereby, achieve a better balance between their sensitivity and precision. Ensuring the consistency of separating sets is also found to increase their validity in terms of actual d-separation and, therefore, to improve the interpretability of constraint-based models for real-life applications. Moreover, ensuring the consistency of separating sets can be done at a limited complexity cost, through the use of block-cut tree decomposition of graph skeletons, which enables to learn causal structures with consistent separating sets for a few hundred nodes. By contrast, earlier methods aiming at reducing the number of d-separation conflicts or other structural inconsistencies through SAT-based approaches, *e.g.* (Hyttinen et al., 2013), have a much larger complexity burden, which limits their applications to very small networks in practice.

2 Result

2.1 Background

2.1.1 Terminology

A **graph** $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ consists of a **vertex set** $\boldsymbol{V} = \{X_1, \cdots, X_p\}$ and an **edge set** $\boldsymbol{E}$. All graphs considered here have at most one edge between any pair of vertices. A **walk** is a sequence of edges joining a sequence of vertices. A **trail** is a walk without repeated edge. A **path** is a trail without repeated vertices. A **cycle** is a trail in which the only repeated vertices are the first and last vertices. Vertices are said to be **adjacent** if there is an edge between them. If all pairs of vertices in a graph are adjacent, it is called a **complete graph** and is denoted by $\mathcal{G}_c$. By constrast, an **empty graph**, denoted by $\mathcal{G}_\emptyset$, consists of isolated vertices with no edges. The **adjacency set** of a vertex X_i in a graph $\mathcal{G}$, denoted by $\mathrm{adj}(\mathcal{G}, X_i)$, is the set of all vertices in $\boldsymbol{V}$ that are adjacent to X_i in $\mathcal{G}$. If an edge is directed, as $X \to Y$, X is a parent of Y and Y a child of X. A **collider** is a triple (X_i, X_j, X_k) in a graph where the edges are oriented as $X_i \to X_k \leftarrow X_j$. A **v-structure** is a **collider** for which X_i and X_j are not adjacent. Given a statistical significance level α, the **conditional independence** of a pair of variables (X_i, X_j) given a set of variables $\boldsymbol{C}$, is denoted by $(X_i \perp\!\!\!\perp X_j | \boldsymbol{C})_\alpha$, where $\boldsymbol{C}$ is called a **separating set** or "**sepset**" for (X_i, X_j).

2.1.2 The PC and PC-stable Algorithms

The PC algorithm (Spirtes and Glymour, 1991), outlined in algorithm 1, is the archetype of constraint-based structure learning methods (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009), as illustrated

in Figure 1. Given a dataset over a set of variables (vertices), it starts from a complete graph $\mathcal{G}$. By a series of statistical tests on each pair of variables, all dispensable edges $X - Y$ are removed if a (conditional) independence and separating set $\boldsymbol{C}$ can be found, *i.e.* $(X \perp\!\!\!\perp Y \mid \boldsymbol{C})$ (step 1). The resulting undirected graph is called the **skeleton**. V-structures are then identified, $X \to Z \leftarrow Y$, if $(X \perp\!\!\!\perp Y \mid \boldsymbol{C})$ and $Z \notin \boldsymbol{C}$ (step 2). Additional assumptions (e.g., acyclicity) allow for the propagation of v-structure orientations to some of the remaining undirected edges (Zhang, 2008) (step 3).

Algorithm 1 The PC Algorithm

Require: $\boldsymbol{V}, \mathcal{D}(\boldsymbol{V})$, significance level α
Step 1: Find the graph skeleton and separating sets of removed edges
Step 2: Orient v-structures based on separating sets
Step 3: Propagate orientations of v-structures to as many remaining undirected edges as possible
return Output graph

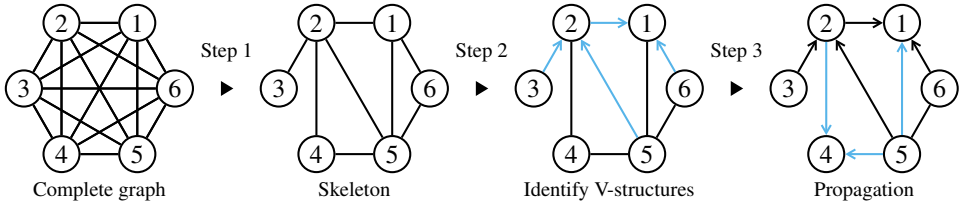

Figure 1: General procedure of constraint-based structure learning.

While the oracle version of the PC-algorithm has been shown to be sound and complete, its application is known to be sensitive to the finite size of real life datasets. In particular, the PC-algorithm in its original implementation (Spirtes, Glymour, and Scheines, 2000) is known to be order-dependent, in the sense that the output depends on the lexicographic order of the variables. This issue can be circumvented, however, for the first step of algorithm 1 with a simple modification given in algorithm 2 and referred to as Step 1 of PC-stable algorithm (Colombo and Maathuis, 2014).

Algorithm 2 Find skeleton and separating sets (Step 1 of PC-stable algorithm)

Require: Conditional independence assessment between all variables $\boldsymbol{V}$ with significance level α
$\mathcal{G} \leftarrow \mathcal{G}_c$
$\ell \leftarrow -1$
repeat
  $\ell \leftarrow \ell + 1$
  for all vertices $X_i \in \mathcal{G}$ **do**
  end for
  $a(X_i) = \mathrm{adj}(\mathcal{G}, X_i)$
  repeat
    select a new pair of vertices (X_i, X_j) adjacent in $\mathcal{G}$ and satisfying $|a(X_i) \backslash \{X_j\}| \geq \ell$
    repeat
      choose new $\boldsymbol{C} \subseteq a(X_i) \backslash \{X_j\}, |\boldsymbol{C}| = \ell$
      if $(X_i \perp\!\!\!\perp X_j | \boldsymbol{C})_\alpha$ **then**
        Delete edge $X_i - X_j$ from $\mathcal{G}$
        Sepset$(X_i, X_j \mid \mathcal{G}) =$ Sepset$(X_j, X_i \mid \mathcal{G}) \leftarrow \boldsymbol{C}$
      end if
    until X_i and X_j are no longer adjacent in $\mathcal{G}$ or all $\boldsymbol{C} \subseteq a(X_i) \backslash \{X_j\}$ with $|\boldsymbol{C}| = \ell$ have been considered
  until all pairs of adjacent vertices (X_i, X_j) in $\mathcal{G}$ with $|a(X_i) \backslash \{X_j\}| \geq \ell$ have been considered
until all pairs of adjacent vertices (X_i, X_j) in $\mathcal{G}$ satisfy $|a(X_i) \backslash \{X_j\}| \leq \ell$
return $\mathcal{G}$, sepsets

2.2 The Consistent PC Algorithm

2.2.1 Lack of Robustness and Consistency of Constraint-based Methods

Beyond the order-dependence of the PC Algorithm, the general lack of robustness of constraint-based methods stems from their tendency to uncover spurious conditional independences (false negatives) between variables. This trend originates from the fact that conditioning on other variables amounts to "slicing" the available data into smaller and smaller subsets, corresponding to different combinations of categories or discrete values of the conditioning variables, over which independence tests are essentially "averaged" to assess conditional independence.

Hence, by making sure that all separating sets are actually consistent with the final graph, one expects to reduce the number of false negative edges due to spurious conditional independences inferred during the edge pruning process and, thereby, to improve the sensitivity (or recall) of the PC or PC-stable algorithms.

The inconsistency of separating sets can be of different forms, regarding either the skeleton (type I) or the final (partially) oriented graph (type II), as illustrated on Figure 1.

A type I inconsistency corresponds to a conditional independence relation such as $(2 \perp\!\!\!\perp 6 \mid 3)$ in Figure 1, for which there is no path between vertex 2 and 6 that passes through 3. This type of inconsistency often involves edges evaluated early on in the pruning process when few edges have been removed, and thus the combinatorial space of possible separating sets is still large. In particular, edge $3 — 6$, which is eventually removed in the final graph, may still exist when the edge $2 — 6$ is under consideration.

A type II inconsistency is a different kind of incompatibility originating from the orientation of the skeleton. It occurs, in particular, when a conditional independence relation is conditioned on at least one common descendant of the pair of interest in the final graph, *e.g.* $(3 \perp\!\!\!\perp 6 \mid 1)$ in Figure 1. Since it stems from the orientation of edges (steps 2&3), the origin of type II inconsistencies is generally more complex and results from a cascade of errors in both conditional independence tests and orientation.

These two types of inconsistency help define the following consistent set for candidate nodes of separating sets in absence of latent variables:

Definition 1 (Consistent set)**.** Given a graph $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ and a set of variables $\{ X, Y, Z \} \subseteq \boldsymbol{V}$,

$$\text{Consist}(X, Y \mid \mathcal{G}) = \{ Z \in \text{adj}(X) \setminus \{ Y \} \mid 1.\ \text{at least one path } \gamma^Z_{XY} \text{ exists in } \mathcal{G};\ 2.\ Z \text{ is not a child of } X \text{ in } \mathcal{G} \}$$

where γ^Z_{XY} is a path from X to Y passing through Z. Note that for an undirected graph, the second condition is always satisfied.

2.2.2 Consistent PC Pseudocodes

Definition 2. NewStep1$(\mathcal{G}_1|\mathcal{G}_2)$ is a modified version of PC-stable step 1 (algorithm 2) where,

1. $\mathcal{G}_c$ is replaced by $\mathcal{G}_1$, and
2. $a(X_i) \setminus \{X_j\}$ is replaced by $a(X_i) \setminus \{X_j\} \cap \text{Consist}(X_i, X_j \mid \mathcal{G}_2)$

Note that algorithm NewStep1$(\mathcal{G}_c|\mathcal{G}_c)$ corresponds to the unmodified step 1 of original PC-stable algorithm 2. By constrast, algorithm NewStep1$(\mathcal{G}_c|\mathcal{G}_\emptyset)$ removes all edges corresponding to independence without conditioning, as no separating set is involved. This unconditional independence search will be noted **step 1a**, while the subsequent conditional independence search will be referred to as **step 1b**, thereafter.

Definition 3. $S(\mathcal{G}_1|\mathcal{G}_2)$ is a modified version of the PC-stable algorithm, where step 1 in algorithm 1 is replaced by NewStep1$(\mathcal{G}_1|\mathcal{G}_2)$ from definition 2.

Then, definition 3 allows to define algorithm 3, which ensures a consistent constraint-based algorithm through an iterative call of S algorithms, $(S_k)_{k\in\mathbb{N}^*}$, following an initial **step 1a**, NewStep1$(\mathcal{G}_c|\mathcal{G}_\emptyset)$. As illustrated on Figure 2 and proved below, algorithm 3 achieves separating set consistency by repeating **step 1b** and **step 2&3**, iteratively, while searching for separating sets that are consistent with the graph obtained at the previous iteration, until a limit cycle of successive graphs is reached.

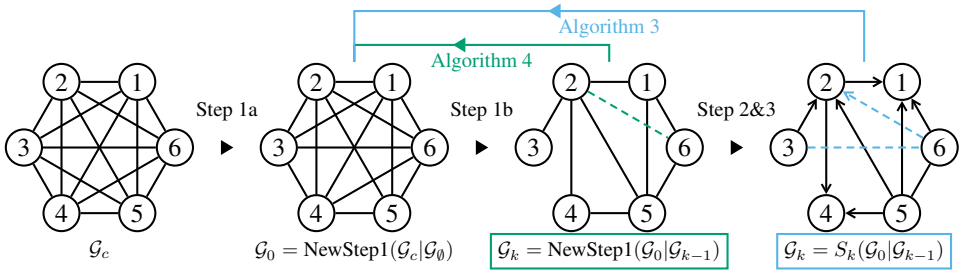

Figure 2: Illustration of the iterative procedure to learn graphical models with orientation-consistent (algorithm 3) or skeleton-consistent (algorithm 4) separating sets. Dashed edges mark the difference between two successive iterations. Proof of separating set consistency is given in theorem 4.

Algorithm 3 Sepset consistent PC algorithm (1st version, orientation consistency)

Require: $\boldsymbol{V}, \mathcal{D}(\boldsymbol{V})$, significance level α
Ensure: $\mathcal{G}$ with consistent separating sets
  $\mathcal{G}_0 \leftarrow \text{NewStep1}(\mathcal{G}_c|\mathcal{G}_\emptyset)$
  $k \leftarrow 0$
  repeat
    $k \leftarrow k + 1$
    $\mathcal{G}_k \leftarrow S_k(\mathcal{G}_0|\mathcal{G}_{k-1})$
  until loop detected, *i.e.*, $\exists n > 0, \mathcal{G}_{k-n} = \mathcal{G}_k$
  $\mathcal{G} \leftarrow \bigcup (\mathcal{G}_j)_{j=k-n}^{k}$, with discarded conflicting orientations
  return $\mathcal{G}$ and consistent separating sets

Alternatively, one may require a separating set consistency at the level of the skeleton only, *i.e.*, before the orientation steps, which corresponds to algorithm 4, below. Indeed, early sepset inconsistencies at the level of the skeleton might cause orientation errors, which in turn can lead to the rejection of valid consistent separating sets in algorithm 3. As outlined in Figure 2, the modification of algorithm 4 only concerns step 1b, which is called iteratively until a limit cycle is reached. Then, the orientation steps 2&3 are performed as for classical PC or PC-derived algorithms, but using consistent separating sets with respect to the union of skeletons returned by the iterative call of step 1b in algorithm 4. However, as the orientation steps 2&3 might induce additional type II inconsistencies, algorithm 4 requires a final consistency check for all separating sets with respect to the final graph $\mathcal{G}$.

Algorithm 4 Sepset consistent PC algorithm (2nd version, skeleton consistency)

Require: $\boldsymbol{V}, \mathcal{D}(\boldsymbol{V})$, significance level α
Ensure: $\mathcal{G}$ with consistent separating sets
  $\mathcal{G}_0 \leftarrow \text{NewStep1}(\mathcal{G}_c|\mathcal{G}_\emptyset)$
  $k \leftarrow 0$
  repeat
    $k \leftarrow k + 1$
    $\mathcal{G}_k \leftarrow \text{NewStep1}(\mathcal{G}_0|\mathcal{G}_{k-1})$
  until loop detected, *i.e.*, $\exists n > 0, \mathcal{G}_{k-n} = \mathcal{G}_k$
  $\mathcal{G} \leftarrow \bigcup (\mathcal{G}_j)_{j=k-n}^{k}$ and consistent separating sets with respect to the graph skeleton $\mathcal{G}$
  Step 2 (orientation of v-structures in $\mathcal{G}$)
  Step 3 (propagation of orientations in $\mathcal{G}$)
  for all removed edges (X, Y) in $\mathcal{G}$ **do**
    $\text{Sepset}(X, Y \mid \mathcal{G}) \leftarrow \text{Sepset}(X, Y \mid \mathcal{G}_k)$
    if $\text{Sepset}(X, Y \mid \mathcal{G}) \notin \text{Consist}(X, Y \mid \mathcal{G})$ and $\text{Sepset}(X, Y \mid \mathcal{G}) \notin \text{Consist}(Y, X \mid \mathcal{G})$ **then**
      Add undirected edge (X, Y) to $\mathcal{G}$
    end if
  end for
  return $\mathcal{G}$ and consistent separating sets

Theorem 4. *The separating sets returned by algorithms 3 and 4 are consistent with respect to the final graph $\mathcal{G}$.*

Proof. Firstly, the limit cycles in algorithms 3 and 4 are warranted to be finite by the deterministic nature of these algorithms and the finite set of graphs $\mathcal{G}_j$.

In algorithm 3, as the union of graphs $\bigcup (\mathcal{G}_j)_{j=k-n}^{k}$ does not remove any edge from the last graph $\mathcal{G}_k$ and discards all conflicting orientations with previous graphs $\mathcal{G}_j$, $j \in \{ k-n, k-1 \}$, taking the union of graphs does not create any new conditional independence relation, nor any inconsistency regarding the final separating sets. More precisely, all removed edges in $\mathcal{G}_k$ have separating sets consistent with respect to at least one graph in the union ($\mathcal{G}_{k-1}$), which is thus also consistent with respect to the union of graphs $\mathcal{G}$.

In algorithm 4, the consistency of separating sets is guaranteed by similar arguments, but only with respect to the skeleton. As the orientation and propagation steps 2&3 might induce additional type II inconsistencies, algorithm 4 requires a final consistency check for all separating sets. Adding back edges with inconsistent separating sets in the final graph $\mathcal{G}$ then guarantees that all the separating sets are consistent with respect to definition 1. □

2.2.3 Tests of Consistency

A unitary operation of algorithms 3 and 4 is to test, for a vertex $Z \in \text{adj}(X) \setminus \{ Y \}$ in $\mathcal{G}$, if $Z \in \text{Consist}(X, Y \mid \mathcal{G})$, which requires that 1) at least one path from X to Y passing through Z (*i.e.* γ_{XY}^{Z}) exists in $\mathcal{G}$ and 2) Z is not a child of X in $\mathcal{G}$ (definition 1).

To test the first condition, it is conceptually simple to first get all paths between X and Y, then check if Z lies in at least one of them, This is however unfeasible as the complexity of getting all paths between two vertices can be large, depending on the edge density of the graph. Fortunately, it is possible to get directly the set of all Z for which at least on path γ_{XY}^{Z} exists. This can be done very efficiently with the help of biconnected component analysis based on block-cut tree decomposition, as detailed in Supplementary Material.

The second condition assumes the absence of latent variables, which allows for condition independence tests on adjacent nodes only in algorithm 2. It is thus straightforward to test without additional complexity burden.

Hence, the overall complexity of the consistency tests of separating sets relies on the block-cut tree decomposition, which can be done beforehand within a single depth first search with complexity $\mathcal{O}(|\boldsymbol{V}| + |\boldsymbol{E}|)$. Thus for each pair (X, Y), the complexity of finding all candidate Z depends on the size of the block-cut tree, which is in the worst case (when the underlying skeleton is a forest) linear in the size of the graph, $\mathcal{O}(|\boldsymbol{V}| + |\boldsymbol{E}|)$, see Supplementary Material.

2.3 Empirical Evaluation

We conducted a series of benchmark structure learning simulations to study the differences between the original PC-stable algorithm and the proposed modifications ensuring consistent separating sets.

For each simulation setting, we first quantified the fraction of inconsistent separating sets predicted by the original PC-stable algorithm, Figure 3. We then compared the performance of the original PC-stable (algorithm 1 and algorithm 2), orientation-consistent PC-stable (algorithm 3) and skeleton-consistent PC-stable (algorithm 4), for different significance levels α, in terms of the precision and recall of the adjacencies found in the inferred graph with respect to the true skeleton, Figures 4 and 5. Figure 4 highlights situations for which the original PC manages to recover a DAG that is already closely related to the ground truth but produces inconsistent separating sets, as shown in Figure 3. By constrast, Figure 5 highlights standard benchmarks from the BNlearn repository (Scutari, 2010) for which the original PC show a poor Recall due to too many spurious conditional independences, and ultimately outputs a graph with only a few obvious edges. Finally, we also measured the fraction of the separating sets used for discarding edges by the three approaches that correspond to true D-separation in the ground-truth DAG, Figure 6.

2.3.1 Data generation and benchmarks

The data-sets used for the numerical experiments were generated with the following scheme. The underlying DAGs were generated with TETRAD (Scheines et al., 1998) as scale-free DAGs with 50 nodes ($\alpha = 0.05$, $\beta = 0.4$, average total degree $d(G) = 1.6$) using a preferential attachment model and orienting its edges based on a random topological ordering of the vertices. Data-sets were simulated with linear structural equation models for three settings : strong, medium and weak interactions (with respective coefficient ranges $[0.2, 0.7]$, $[0.1, 0.5]$, and $[0, 0.3]$ and covariance ranges $[0.5, 1.5]$, $[0.5, 1]$, and $[0.2, 0.7]$). In addition, we also generated data-sets for the classical benchmarks Insurance (27 nodes, 52 links, 984 paramaters), Hepar2 (70 nodes, 123 links, 1453 paramaters) and Barley (48 nodes, 84 links, 114005 paramaters) networks from the Bayesian Network repository (Scutari, 2010).

Reconstruction benchmarks were performed with pcalg's (Kalisch et al., 2012) PC-stable implementation, modified for enforcing separating set consistency either taking into account orientations (algorithm 3) or at the level of the skeleton (algorithm 4). The (conditional) independence test used in all simulations is a linear (partial) correlation with Fisher's z-transformation. Performances are obtained with relation to the true skeleton by measuring the Precision (positive predictive value), $Prec = TP/(TP + FP)$ and Recall or Sensitivity (true positive rate), $Rec = TP/(TP + FN)$ where TP is a correctly predicted adjacency, FP an incorrectly predicted adjacency and FN an incorrectly discarded adjacency.

2.3.2 Benchmark Results

The fraction of inconsistent separating sets that were used to remove edges was first estimated for increasing sample size and varying parent-child interaction strength, using the original PC-stable algorithm for random and scale-free DAGs of 50 nodes, Figure 3. We note that in typical settings, a significant fraction of the separating sets that were used to remove edges during Step 1 of the PC-stable algorithm cannot be "read off" the returned graph, either because there is no path containing Z that connects X and Y (skeleton inconsistency, green in Figure 3) or because there is a conditioning on an invalid child node (orientation inconsistency, *i.e.*, difference between blue and green inconsistencies in Figure 3). Both increasing the sample size and increasing the interaction strength reduces the number of inconsistent sepsets. We attribute this in part to the severity of the PC-stable algorithm which tends to remove to many false negative edges because of spurious inconsistencies. With a larger sample size N and stronger interactions, consistent separating sets are still not guaranteed by the original algorithm but these settings decrease the number of spurious independencies and leads to denser reconstructed graphs, thus making it more likely for potential separating sets to be consistent. Orientation consistency is particularly difficult to obtain with respect to the returned CPDAG, as orientation and propagation steps generally suffer even more from sampling noise and previous mistakes than the skeleton reconstruction (Step 1). Notably, the orientation depends on the order in which separating sets are tested in PC-stable (in pcalg it depends on the ordering of the variables in the data-set).

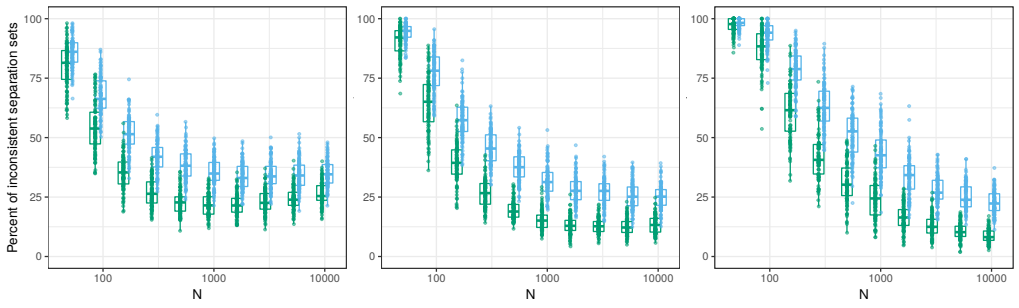

Figure 3: **Sepset inconsistency of the original PC-stable algorithm.** In each subplot the fraction of inconsistent separating sets with respect to the skeleton (green) or CPDAG (blue) obtained with the original PC-stable algorithm with a fixed $\alpha = 0.05$ is displayed for increasing sample size N. Data-sets were generated from 100 scale-free graphs of 50 nodes and $d(G) = 1.6$ with different parent-child interaction strengths : strong (left), medium (middle) and weak (right).

We then compared the performance of the original PC-stable (algorithm 1 and algorithm 2), orientation-consistent PC-stable (algorithm 3) and skeleton-consistent PC-stable (algorithm 4), for different significance levels α, in terms of the precision and recall of the adjacencies found in the inferred graph with respect to the true skeleton, Figures 4, 5 and S1. Enforcing the sepset consistency is shown to significantly improve the sensitivity of constraint-based methods, for a given α, while achieving equivalent or better overall structure learning performance.

It is particularly the case for standard benchmark networks from the BNlearn repository (Scutari, 2010), Figure 5, for which the original PC-stable algorithm shows good precision but poor recall (Rec<0.15-0.35 and Prec>0.65 at maximum Fscore, see iso-Fscore dotted lines in Figure 5), while consistent PC-stable achieves a better balance between precision and recall (Rec$\simeq$0.5 and Prec$\simeq$0.5-0.6 at maximum Fscore, Figure 5).

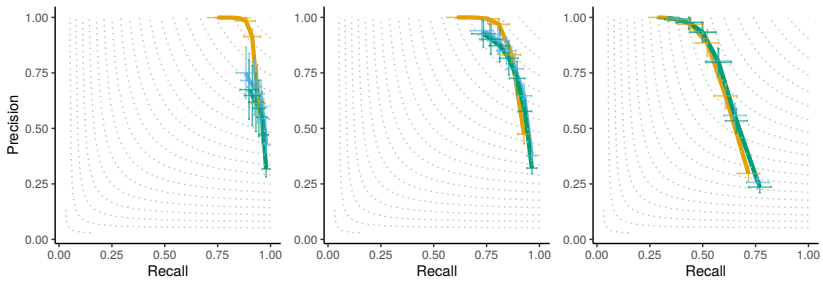

Figure 4: **Precision-recall curves for the original PC-stable (yellow), skeleton-consistent PC-stable (green) and orientation-consistent PC-stable (blue)**. The mean performances and standard deviations (error bars) obtained over 100 networks are shown for 7 values of the (conditional) independence significance threshold α between 10^{-5} and 0.2 Data-sets with N=500 samples were generated from the same graphs as in Figure 3 with strong (left), medium (middle) and weak (right) interactions. See Figure S1 for N=100, 1000.

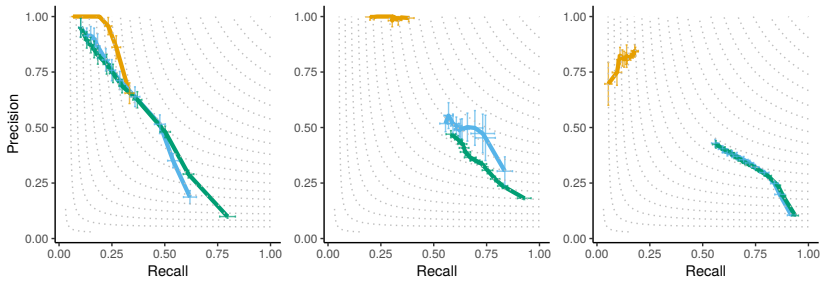

Figure 5: **Precision-recall curves for the original PC-stable (yellow), skeleton-consistent PC-stable (green) and orientation-consistent PC-stable (blue)**. The mean performances and standard deviations (error bars) obtained over 100 networks are shown for 12 values of the (conditional) independence significance threshold α between 10^{-25} and 0.5 (1e-25 1e-20 1e-17 1.0e-15 1.0e-13 1.0e-10 8.7e-09 7.6e-07 6.6e-05 5.7e-03 5.0e-02 5.0e-01). Data-sets with N=1000 samples were generated for the standard benchmarks Hepar2 (left), Insurance (middle) and Barley (right) networks from the BNlearn repository (Scutari, 2010).

Finally, we also compared the fraction of valid separating sets used for discarding edges, which entail true d-separation in the ground-truth DAG, Figures 6 and S2. Ensuring the consistency of separating sets tends to increase, although not guarantee, their validity in terms of actual d-separation. Consistent sepsets with invalid d-separation are primarily caused by edge mis-orientations rather than skeleton errors. In particular, skeleton-consistent separating sets yield better performance in terms of valid d-separation than orientation-consistent separating sets with the setting of the PC-stable algorithm used here. This is, however, expected to depend on the specific settings for conditional independence test, orientation and propagation rules, used in different constraint-based methods.

3 Conclusion

In this paper, we propose and implement simple modifications of the PC algorithm also applicable to any PC-derived constraint-based methods, in order to enforce the consistency of the separating sets

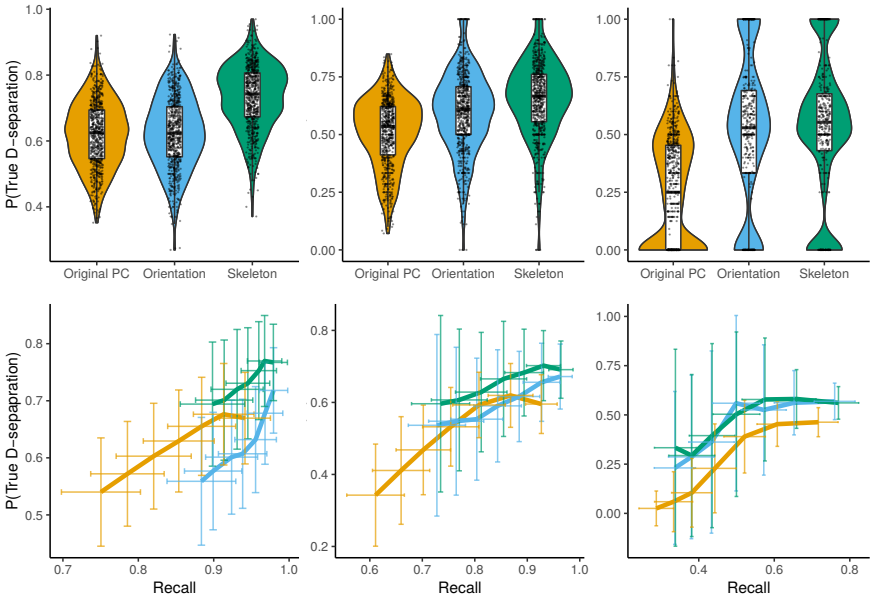

Figure 6: **Proportion of valid d-separation sepsets among edge-removing sepsets.** Top row shows the proportion of sepsets that correspond to a valid d-separation in the true DAG that were used for removing edges during Step 1 of original, orientation-consistent and skeleton-consistent PC-stable algorithms for all tested α. Bottom row shows the average proportion of valid d-separation for a given average recall over all tested values of α. Data-sets with N=500 were generated from 100 DAGs with linear SEMs with strong (left), medium (middle) and weak (right) interaction (see Figure S2 for N=100, 1000).

of discarded edges with respect to the final graph, which is an actual shortcoming of constraint-based approaches, Figure 3. Enforcing sepset consistency is shown to significantly improve the sensitivity of constraint-based methods, while achieving equivalent or better overall structure learning performance, Figures 4, 5 and S1. In addition, ensuring the consistency of separating sets increases also their validity in terms of actual d-separation, Figures 6 and S2.

The existence of sepset inconsistencies with constraint-based methods originates from their tendency to uncover spurious conditional independences early on in the pruning process when the combinatorial space of possible separating sets is still large, unlike in the final typically sparse skeleton. Such spurious conditional independences are responsible, in particular, for the large number of false negative edges and, therefore, frequently poor sensitivity of constraint-based methods (Colombo and Maathuis, 2014). By contrast, enforcing sepset consistency enables to achieve a better balance between sensitivity and precision.

To circumvent this inconsistency issue during the skeleton step, we have shown that one can either use sepset consistency taking into account orientations to help reject inconsistent sepsets (algorithm 3) or use sepset consistency of the skeleton to help determine the orientations (algorithm 4). The later approach tends to yield slightly better performance with the setting of the PC-stable algorithm used here but this is expected to be dependent on the specific settings used, for conditional independence test, orientation and propagation rules, in different constraint-based methods.

Indeed, the methods and algorithmic implementations presented here are not primarily meant to outcompete a specific PC or PC-derived algorithm but rather to improve the explainability of constraint-based methods, by ensuring the consistency of all separating sets in the final causal graphs.

The approach is very general and applicable to the large variety of constraint-based methods, starting with a complete graph and discarding dispensable edges iteratively based on conditional independence search. Beyond the formal interest of guaranteeing sepset consistency, this is also especially important, in practice, for the interpretability of constraint-based models for real-life applications.

Acknowledgements
The authors acknowledge financial support from the French Ministry of Higher Education and Research, PSL Research University and Sorbonne University.

References

Affeldt, S., and Isambert, H. 2015. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015*, 42–51.

Affeldt, S.; Verny, L.; and Isambert, H. 2016. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* 17(S2).

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* 15:3741–3782.

Colombo, D.; Maathuis, M. H.; Kalisch, M.; and Richardson, T. S. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* 40(1):294–321.

Hyttinen, A.; Hoyer, P. O.; Eberhardt, F.; and Järvisalo, M. 2013. Discovering cyclic causal models with latent variables: A general sat-based procedure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, 301–310. Arlington, Virginia, United States: AUAI Press.

Kalisch, M., and Bühlmann, P. 2008. Robustification of the pc-algorithm for directed acyclic graphs. *Journal Of Computational And Graphical Statistics* 17(4):773–789.

Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M. H.; and Bühlmann, P. 2012. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* 47(11):1–26.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, 441–452. Morgan Kaufmann Publishers Inc.

Pearl, J. 2009. *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition.

Richardson, T., and Spirtes, P. 2002. Ancestral graph markov models. *Ann. Statist.* 30(4):962–1030.

Scheines, R.; Spirtes, P.; Glymour, C.; Meek, C.; and Richardson, T. 1998. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research* 33(1):65–117.

Scutari, M. 2010. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35(3):1–22.

Sella, N.; Verny, L.; Uguzzoni, G.; Affeldt, S.; and Isambert, H. 2018. Miic online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* 34(13):2311–2313.

Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9:62–72.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2nd edition.

Spirtes, P.; Meek, C.; and Richardson, T. 1999. An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery*. Menlo Park, CA: AAAI Press. 211–252.

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning* 65(1):31–78.

Verny, L.; Sella, N.; Affeldt, S.; Singh, P. P.; and Isambert, H. 2017. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* 13(10):e1005662.

Zhang, J. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* 172(16-17):1873–1896.

SUPPLEMENTARY MATERIAL

on NIPS 2019 paper

Constraint-based Causal Structure Learning with Consistent Separating Sets

Honghao Li, Vincent Cabeli, Nadir Sella, Hervé Isambert
Institut Curie, PSL Research University, CNRS UMR168, Paris
{honghao.li, vincent.cabeli, nadir.sella, herve.isambert}@curie.fr

An R implementation of the methods in the case of the PC-stable algorithm is available with examples at `https://github.com/honghaoli42/consistent_pcalg`.

A Test of Consistency

A.1 Terminology

A **connected graph** $\mathcal{G}$ is such that there is a path between each pair of vertices of $\mathcal{G}$. A **connected component** of a graph is a maximal connected subgraph. An **articulation point** (or **cut point**) is a vertex in a connected graph whose removal would disconnect the graph and thus increase its number of connected components. A **biconnected graph** is a connected graph without articulation point. A **biconnected component** (or **block**) is a maximal biconnected subgraph.

A.2 Biconnected Component Analysis

For a pair (X, Y) in a graph $\mathcal{G}$, one of the necessary conditions for its separating set to be consistent, as stated in definition 1, is that for each vertex Z in the separating set, Z lies on a path γ_{XY}^{Z} between X and Y in the skeleton of $\mathcal{G}$. For one pair of vertices, checking the existence of a path for all Z can already be time consuming if the degrees of the vertices are large. In addition, the complexity will be further multiplied by the number of pairs to be considered. Fortunately, it is possible to avoid this high complexity with the help of the biconnected component analysis based on block-cut tree decomposition, and thus to limit the search of consistent separating vertices within those that are consistent with respect to the skeleton.

Definition 5 (Block-cut tree)**.** $\mathcal{G}$ a connected (sub)graph. The block-cut tree decomposition of $\mathcal{G}$ is denoted by $\mathcal{T}(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{Br})$ where $\boldsymbol{B} = \{ b_i \}_{i=1}^{m}$ is the set of biconnected components (or blocks) of $\mathcal{G}$, $\boldsymbol{C} = \{ c_j \}_{j=1}^{n}$ is the set of articulation points (or cut points) and $\boldsymbol{Br} = \{ (b_i, c_j) \mid b_i \in \boldsymbol{B}, c_j \in \boldsymbol{C}, b_i \text{ and } c_j \text{ are adjacent in } \mathcal{T} \}$ is the set of connections between $\boldsymbol{B}$ and $\boldsymbol{C}$.

In the following we establish a relation between biconnected components and the path existence problem.

Lemma 6 (Menger's theorem for biconnected graph)**.** *Let $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ be a biconnected graph, $\{ X, Y \} \subseteq \boldsymbol{V}$ a pair of vertices. There is a cycle in $\mathcal{G}$ that contains X and Y.*

Theorem 7. *Let $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ be an undirected graph, $\mathcal{H}(\boldsymbol{V}_{\mathcal{H}}, \boldsymbol{E}_{\mathcal{H}}) \subseteq \mathcal{G}$ a biconnected component of $\mathcal{G}$, $\{ X, Y \} \subseteq \boldsymbol{V}_{\mathcal{H}}$ a pair of vertices, and $Z \in \boldsymbol{V}_{\mathcal{G}}$ a third vertex. There is a path γ_{XY}^{Z} if and only if $Z \in \boldsymbol{V}_{\mathcal{H}}$.*

Proof. If there is a path γ_{XY}^{Z}, suppose that $Z \notin \boldsymbol{V}_{\mathcal{H}}$, then the subgraph $\mathcal{H}'$ of $\mathcal{G}$ over $\boldsymbol{V}_{\mathcal{H}} \cup \{ Z \}$ is biconnected thanks to γ_{XY}^{Z}, and $\mathcal{H} \subset \mathcal{H}'$ is not a biconnected component of $\mathcal{G}$ as it is not maximal. Therefore we must have $Z \in \boldsymbol{V}_{\mathcal{H}}$.

If $\{ X, Y, Z \} \subseteq \boldsymbol{V}_{\mathcal{H}}$, then lemma 6 guarantees a cycle that contains Z and Y. Since $\boldsymbol{V}_{\mathcal{H}}$ contains at least three vertices, such a cycle contains $n \geq 1$ vertices other than Z and Y, and can be represented by two edge-distinct paths between Z and Y:

$$\gamma_{ZY}^{(1)} = ZU_1U_2 \cdots U_kY, \quad \gamma_{ZY}^{(2)} = ZU_{k+1}U_{k+2} \cdots U_nY$$

where $k \in \mathbb{Z}^{\geq 0}$ (with $k = 0$ indicating a direct edge between Z and Y), $n \in \mathbb{Z}^{+}$, $k < n$ and $\{ U_i \}_{i=1}^{n}$ are distinct vertices. Since Y is not an articulation point, there is a path γ_{XZ} that does not contain Y:

$$\gamma_{XZ} = XD_1D_2 \cdots D_mZ$$

where $m \in \mathbb{Z}^{\geq 0}$ and $\{ D_j \}_{j=1}^m$ are distinct vertices. If $\{ U_i \}_{i=1}^n \cap \{ D_j \}_{j=1}^m = \emptyset$, then there is a path

$$\gamma_{XY}^Z = \gamma_{XZ}\gamma_{ZY}^{(i)}, i \in \{ 1, 2 \}.$$

Otherwise, suppose $\{ U_i \}_{i=1}^n \cap \{ D_j \}_{j=1}^m = \{ D_{p_1}, D_{p_2}, \ldots, D_{p_t} \}$ where $t \in \mathbb{Z}^+$ and $p_1 < p_2 < \cdots < p_t$, and suppose $D_{p_1} = U_l$. If $l \leq k$, then there is a path

$$\gamma_{XY}^Z = XD_1D_2 \cdots D_{p_1}(U_l)U_{l-1} \cdots U_1\gamma_{ZY}^{(2)},$$

if $l > k$, then there is a path

$$\gamma_{XY}^Z = XD_1D_2 \cdots D_{p_1}(U_l)U_{l-1} \cdots U_{k+1}\gamma_{ZY}^{(1)}.$$

As a result, if $\{ X, Y, Z \} \subseteq \boldsymbol{V}_{\mathcal{H}}$, then there is always a path γ_{XY}^Z. □

Corollary 8. *Let $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ be a connected graph, $\mathcal{T}(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{Br})$ the block-cut tree decomposition of $\mathcal{G}$, $\{ X, Y \} \subseteq \boldsymbol{V}$ a pair of vertices, n_X, n_Y the corresponding nodes of X and Y in $\mathcal{T}$, and $\boldsymbol{S} = \{ Z \in \boldsymbol{V} \setminus \{ X, Y \} \mid$ at least one path γ_{XY}^Z exists. $\}$*

1. *If $n_X = n_Y = b_i \in \boldsymbol{B}$, then $\boldsymbol{S} = \boldsymbol{V}(b_i) \setminus \{ X, Y \}$.*
2. *If $n_X \neq n_Y$, let $\nu_{XY} = w_1w_2 \cdots w_k, w_1 = n_X, w_k = n_Y$ be the path between n_X and n_Y where each w_i belongs to $\boldsymbol{B}$ or $\boldsymbol{C}$, then $\boldsymbol{S} = \bigcup (\boldsymbol{V}(w_i))_{i=1}^k \setminus \{ X, Y \}$.*

The first case is a direct result of theorem 7. The second case is not difficult to prove once we notice the fact that ν_{XY} is the unique path between n_X and n_Y in $\mathcal{T}$, and that every γ_{XY} must contain all the cut points in ν_{XY}, and thus can be decomposed into segments of paths between these cut points.

Each undirected graph $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$ can be decomposed into a set of single vertices and a set of connected subgraphs, where each subgraph can be represented by a block-cut tree. Based on this decomposition, algorithm 5 gives the consistent candidate vertices for separating set for a pair of vertices as described in definition 1.

Algorithm 5 Consistent candidates

Require: (Partially directed) graph $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$, its block-cut tree decomposition for each connected component (with respect to its skeleton) $\{ \mathcal{T}_i(\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{Br}) \}$, two vertices $\{ X, Y \} \subseteq \boldsymbol{V}$
Ensure: Set of all candidate vertices Consist$(X, Y \mid \mathcal{G})$.
  if X and Y do not belong to the same block-cut tree $\mathcal{T}_i$ **then**
    return $\emptyset$
  end if
  if X and Y belong to the same block $b_i \in \boldsymbol{B}$ **then**
    return $(\text{Ne}(X) \setminus \text{Child}(X)) \cap (\boldsymbol{V}(b_i) \setminus \{ X, Y \})$
  else
    $\nu_{XY} \leftarrow \text{TreePath}(n_X, n_Y) = w_1w_2 \cdots w_k$
    return $(\text{Ne}(X) \setminus \text{Child}(X)) \cap (\bigcup (\boldsymbol{V}(w_i))_{i=1}^k \setminus \{ X, Y \})$
  end if

The block-cut tree decomposition can be done beforehand within a single depth first search with complexity $\mathcal{O}(|\boldsymbol{V}| + |\boldsymbol{E}|)$. Thus for each pair (X, Y), the complexity of finding all candidate Z depends on the size of the block-cut tree. In the worst case where $\mathcal{G}$ is a forest with only bridges (edges, the removal of each bridge increases the number of connected components of $\mathcal{G}$), the number of nodes and branches in the block-cut tree $\mathcal{T}$ of $\mathcal{G}$ is of the same order of $|\boldsymbol{V}|$ and $|\boldsymbol{E}|$, and for all pair of vertices $\{ X, Y \} \subseteq \boldsymbol{V}$ we need to perform a path search in $\mathcal{T}$ of complexity $\mathcal{O}(|\boldsymbol{V}| + |\boldsymbol{E}|)$ to get $\boldsymbol{S}$. In the best scenario where $\mathcal{G}$ is biconnected, $\boldsymbol{S} = \boldsymbol{V} \setminus \{ X, Y \}$ for all pairs. Then, an operation of set intersection $(\text{Ne}(X) \setminus \text{Child}(X)) \cap \boldsymbol{S}$ with linear complexity $\mathcal{O}(|\text{Ne}(X)| + |\boldsymbol{S}|)$ will give the result.

B Supplementary Figures

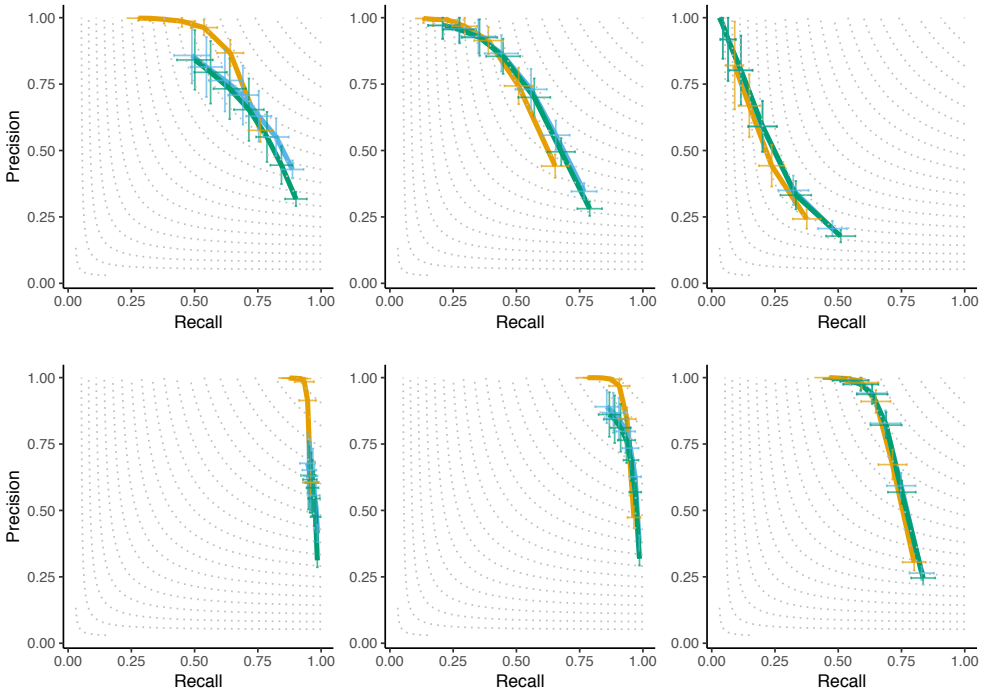

Figure S1: **Precision-recall curves for the original PC-stable (yellow), orientation-consistent PC-stable (blue) and skeleton-consistent PC-stable (green)**. Data-sets of N=100 samples (top row) or of N=1000 (bottom row), with strong (left), medium (middle) and weak (right) interactions. See Figure 4 for more information.

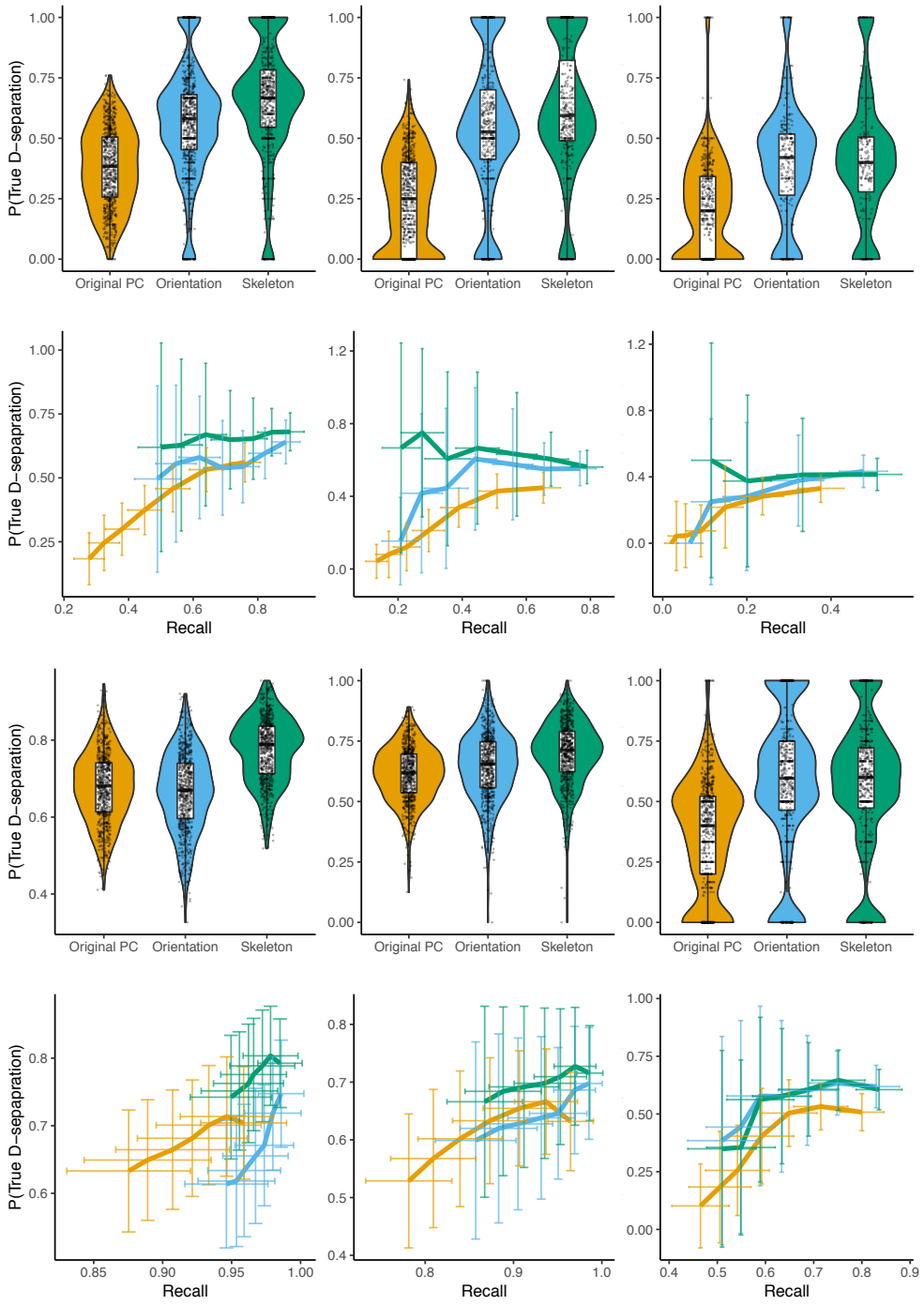

Figure S2: **Proportion of valid d-separation sepsets among edge-removing sepsets found during reconstruction.** Data-sets of N=100 samples (top two rows) or of N=1000 (bottom two rows), with strong (left), medium (middle) and weak (right) interactions. See Figure 6 for more information.

4.3 Reliable orientations with mutual information supremum

Finally, we propose the equivalent of conservative PC for MIIC, which improves the reliability of inferred orientations at only a small cost to sensitivity. It does not rely on an orientation cut β as introduced in Section 4.1.2, but rather on a formulation of information supremum, for both the continuous and the discrete case.

The publication accepted at the Why21 workshop contextualizes MIIC, proposes the simple change to the orientation rules enforcing non-negative regularized information terms, and compares the performance between the old and the new rules on simulated networks.

4.3.1 Publication at Why21 workshop, NeurIPS 2021

Reliable causal discovery based on mutual information supremum principle for finite datasets

Anonymous Author(s)
Affiliation
Address
email

Abstract

The recent method, MIIC (Multivariate Information-based Inductive Causation), combining constraint-based and information-theoretic frameworks, has been shown to significantly improve causal discovery from purely observational data. Yet, a substantial loss in precision has remained between skeleton and oriented graph predictions for small datasets. Here, we propose and implement a simple modification, named conservative MIIC, based on a general mutual information supremum principle regularized for finite datasets. In practice, conservative MIIC rectifies the negative values of regularized (conditional) mutual information used by MIIC to identify (conditional) independence between discrete, continuous or mixed-type variables. This modification is shown to greatly enhance the reliability of predicted orientations, for all sample sizes, with only a small sensitivity loss compared to MIIC original orientation rules. Conservative MIIC is especially interesting to improve the reliability of causal discovery for real-life observational data applications.

1 Background

Constraint-based structure learning methods can, in principle, discover causal relations in purely observational data (Pearl, 2009; Spirtes, Glymour, and Scheines, 2000). This is theoretically feasible up to some independence equivalence classes, as the orientations of certain edges may only be uncovered through perturbative data and remain undetermined if only observational data is available.

Yet, regardless of this theoretical limitation, it has long been recognized (Ramsey, Spirtes, and Zhang, 2006; Colombo and Maathuis, 2014) that orientations predicted by constraint-based methods are often unreliable, which has largely limited, in practice, the application of constraint-based methods to uncover causal relations in real-life observational data.

This causal uncertainty originates from the extensive number of steps and conditions that constraint-based methods, such as the original IC (Pearl and Verma, 1991) and PC (Spirtes and Glymour, 1991) algorithms, have to meet before they can infer edge orientation. Indeed, they must first learn an undirected skeleton, by uncovering (conditional) independences between all pairs of variables, before inferring the orientation of v-structures and finally propagating these orientations to other undirected edges. This long chain of uncertain computational decisions leads to the accumulation of errors which ultimately limit the accuracy of the final orientation and propagation steps of constraint-based methods. As a result, edge orientations significantly reduce the precision (or positive predicted value) of inferred causal graphs compared to their undirected skeleton. In addition, constraint-based methods are known to suffer from much lower sensitivity or recall (*i.e.*, true positive rate) than precision scores, in general (Colombo and Maathuis, 2014; Li et al., 2019). This is related to the fact that separating sets used to remove edges in the (early) steps of constraint-based methods are frequently not consistent with the final skeleton and oriented graphs (Li et al., 2019). They correspond to

Submitted to WHY-21 workshop @ NeurIPS 2021. Do not distribute.

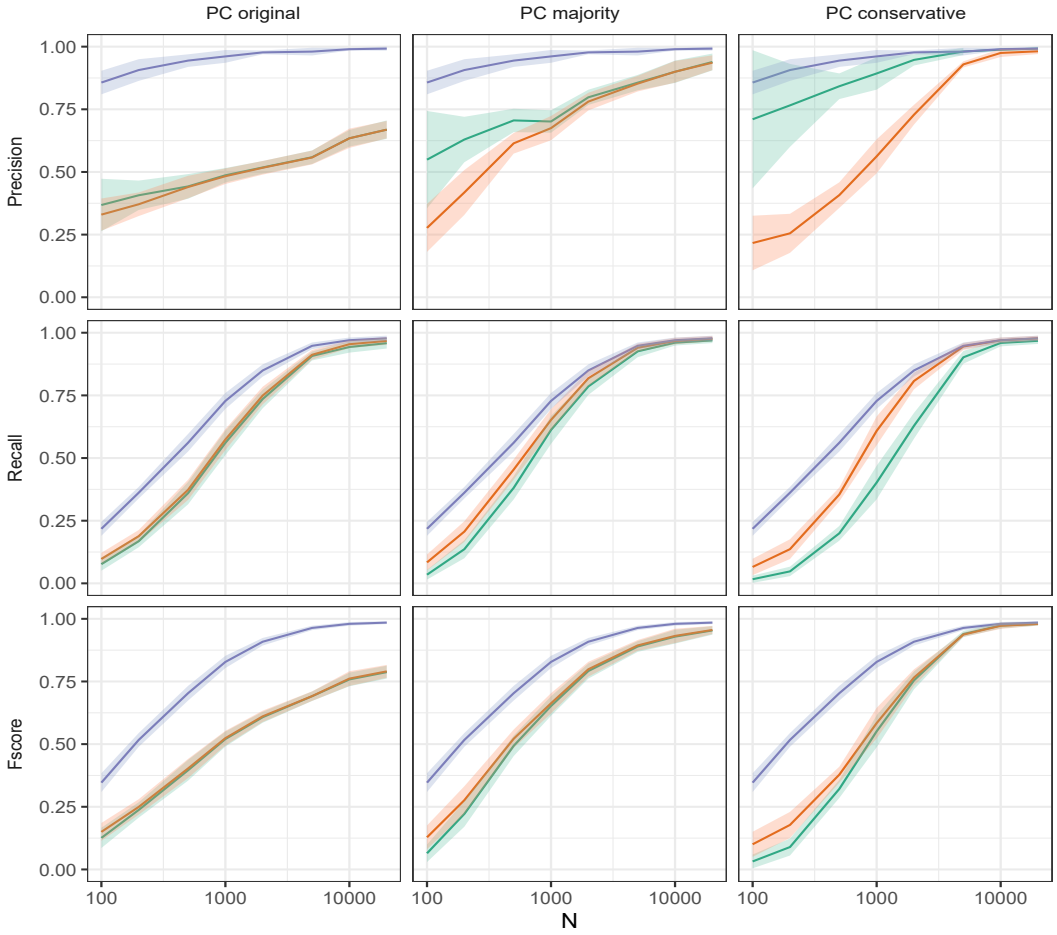

Figure 1: **PC original, majority and conservative orientation rules on discrete datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 2.7 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green).

spurious conditional independences responsible for the large number of false negative edges and, therefore, low sensitivity of constraint-based methods.

While successive refinements of orientation rules, such as conservative rules (Ramsey, Spirtes, and Zhang, 2006) and majority rules (Colombo and Maathuis, 2014), have helped improve the average precision of orientations, they also lead to large precision variance and further aggravate the poor recall of edge orientations at small sample sizes. This is illustrated here for both discrete (Fig. 1) and continuous (Fig. 2) benchmark datasets generated by random Bayesian networks using the available codes from (Cabeli et al., 2020), see section on Data generation and benchmarks, below.

The recently developed method, MIIC, combining constraint-based and maximum likelihood frameworks, has been shown to significantly improve the situation by greatly reducing the imbalance between precision and recall, for all sample sizes (Verny et al., 2017; Cabeli et al., 2020). Compared to traditional constraint-based methods, MIIC also significantly reduces the precision gap between skeleton and oriented graphs for large enough datasets, as discussed below. However, a substantial loss in precision remains between skeleton and oriented graphs for smaller datasets.

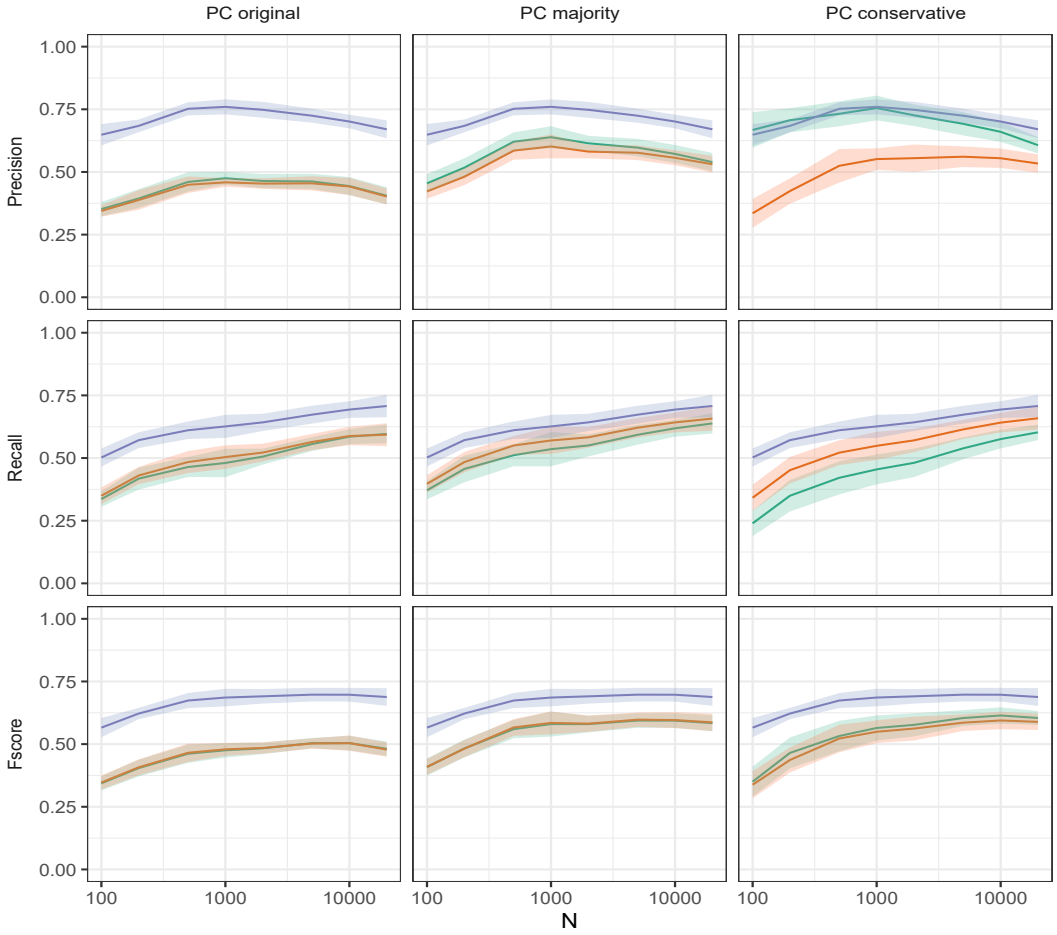

Figure 2: **PC original, majority and conservative orientation rules on continuous datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 2.7 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green).

In this paper, we propose and implement a simple modification of MIIC algorithm, which is found to greatly improve the precision of predicted orientations even for relatively small datasets. It is achieved at the expense of a small loss of orientation recall but significantly enhances the reliability of predicted orientations for all sample sizes. This simple modification, referred to as conservative MIIC, is especially interesting, in practice, to improve the reliability of causal discovery for real-life observational data applications.

2 Results

2.1 MIIC outline

MIIC (Multivariate Information-based Inductive Causation) is a novel structure learning method (Verny et al., 2017; Cabeli et al., 2020) and online server (Sella et al., 2018), combining constraint-based and information-theoretic frameworks. Starting from a fully connected graph, MIIC iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths based on the "3off2" scheme (Affeldt and Isambert, 2015; Affeldt, Verny, and Isambert,

2016). This amounts to progressively uncover the best supported conditional independencies, *i.e.* $I(X;Y|\{A_i\}_n) \simeq 0$, by iteratively "taking off" the most significant indirect contributions of *positive* conditional 3-point information, $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$, from every 2-point (mutual) information, $I(X;Y)$, as,

$$I(X;Y|\{A_i\}_n) = I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \cdots - I(X;Y;A_n|\{A_i\}_{n-1}) \quad (1)$$

In practice, (conditional) independence is established by comparing mutual information (MI) or conditional mutual information (CMI) to a universal Normalized Maximum Likelihood (NML) complexity term, $k_N^{\mathrm{NML}}(X;Y|\{A_i\})/N$, computed over all datasets of the same size N and marginal distributions $p(X,\{A_i\})$ and $p(Y,\{A_i\})$ (Affeldt and Isambert, 2015). This can be seen as a NML-regularization of MI and CMI for datasets of finite sample size N as,

$$I'_N(X;Y|\{A_i\}) = I_N(X;Y|\{A_i\}) - \frac{1}{N}\, k_N^{\mathrm{NML}}(X;Y|\{A_i\}) \quad (2)$$

where $k_N^{\mathrm{NML}}(X;Y|\{A_i\})$ is computed iteratively in linear time (Kontkanen and Myllymäki, 2007; Roos et al., 2008) for increasing numbers of X and Y partitions, r_x and r_y, starting with $k_N^{\mathrm{NML}}(X;Y|\{A_i\}) = 0$ for $r_x = r_y = 1$ (Affeldt and Isambert, 2015; Cabeli et al., 2020).

Hence, (conditional) independence is established for $I'_N(X;Y|\{A_i\}) \leqslant 0$, whenever sufficient and significant indirect positive contributions could be iteratively collected in Eq. 1 to warrant the removal of the XY edge.

This leads to an undirected skeleton, which MIIC then (partially) orients based on the sign and amplitude of the NML-regularized conditional 3-point information terms (Affeldt and Isambert, 2015; Verny et al., 2017), corresponding to the difference between NML-regularized (C)MI terms.

$$I'_N(X;Y;Z|\{A_i\}) = I'_N(X;Y|\{A_i\}) - I'_N(X;Y|\{A_i\},Z) \quad (3)$$

In particular, negative NML-regularized conditional 3-point information terms, $I'_N(X;Y;Z|\{A_i\}) < 0$, correspond to the signature of causality in observational data (Affeldt and Isambert, 2015) and lead to the prediction of a v-structure, $X \to Z \leftarrow Y$, if $X - Z - Y$ is an unshielded triple in the skeleton (with $I'_N(X;Y|\{A_i\}) \leqslant 0$). By contrast, a positive NML-regularized conditional 3-point information term, $I'_N(X;Y;Z|\{A_i\}) > 0$, suggests to propagate the orientation of a previously directed edge $X \to Z - Y$ as $X \to Z \to Y$ (with $I'_N(X;Y|\{A_i\},Z) \leqslant 0$), to fulfill the assumptions of the underlying graphical model class.

2.2 MIIC performance on discrete data, allowing for negative NML-regularized MI & CMI

MIIC was originally developed for discrete variables only, for which MI and CMI are straightforward to compute. Compared to traditional constraint-based methods on discrete data, MIIC greatly reduces the imbalance between precision and recall, for all sample sizes, Fig. 3. MIIC also significantly reduces the precision gap between skeleton and oriented graphs, for large enough datasets. However, a substantial loss in precision remains between skeleton and oriented graphs, for small datasets, irrespective of the CPDAG or oriented-edge-only subgraph scores used for the comparison, Fig. 3.

These results illustrate the interest in integrating multivariate information criteria into constraint-based methods. However, for small datasets or datasets including variables with many discrete levels, NML complexities can easily out-weight MI and CMI terms for weakly dependent variables. As a result, MIIC tends to infer some v-structure orientations, $X \to Z \leftarrow Y$, for which both NML-regularized (C)MI terms in Eq. 3 are negative, *i.e.* $I'_N(X;Y|\{A_i\}) < I'_N(X;Y|\{A_i\},Z) < 0$, suggesting that Z could in fact be included in a separating set of the $\{X,Y\}$ pair, in contradiction with the inferred v-structure, $X \to Z \leftarrow Y$.

Note that such a v-structure would be excluded from the final graph in the frame of traditional constraint-based methods implementing conservative orientation rules, which check that Z is not included in any separating set of the $\{X,Y\}$ pair (Ramsey, Spirtes, and Zhang, 2006). Similarly, rectifying all negative NML-regularized (C)MI values into null values, as proposed and implemented in the present paper below, leads to a vanishing NML-regularized (conditional) 3-point information term in Eq. 3, *i.e.* $I'_N(X;Y;Z|\{A_i\}) = 0$, which precludes the orientation of the unshielded triple, $X - Z - Y$.

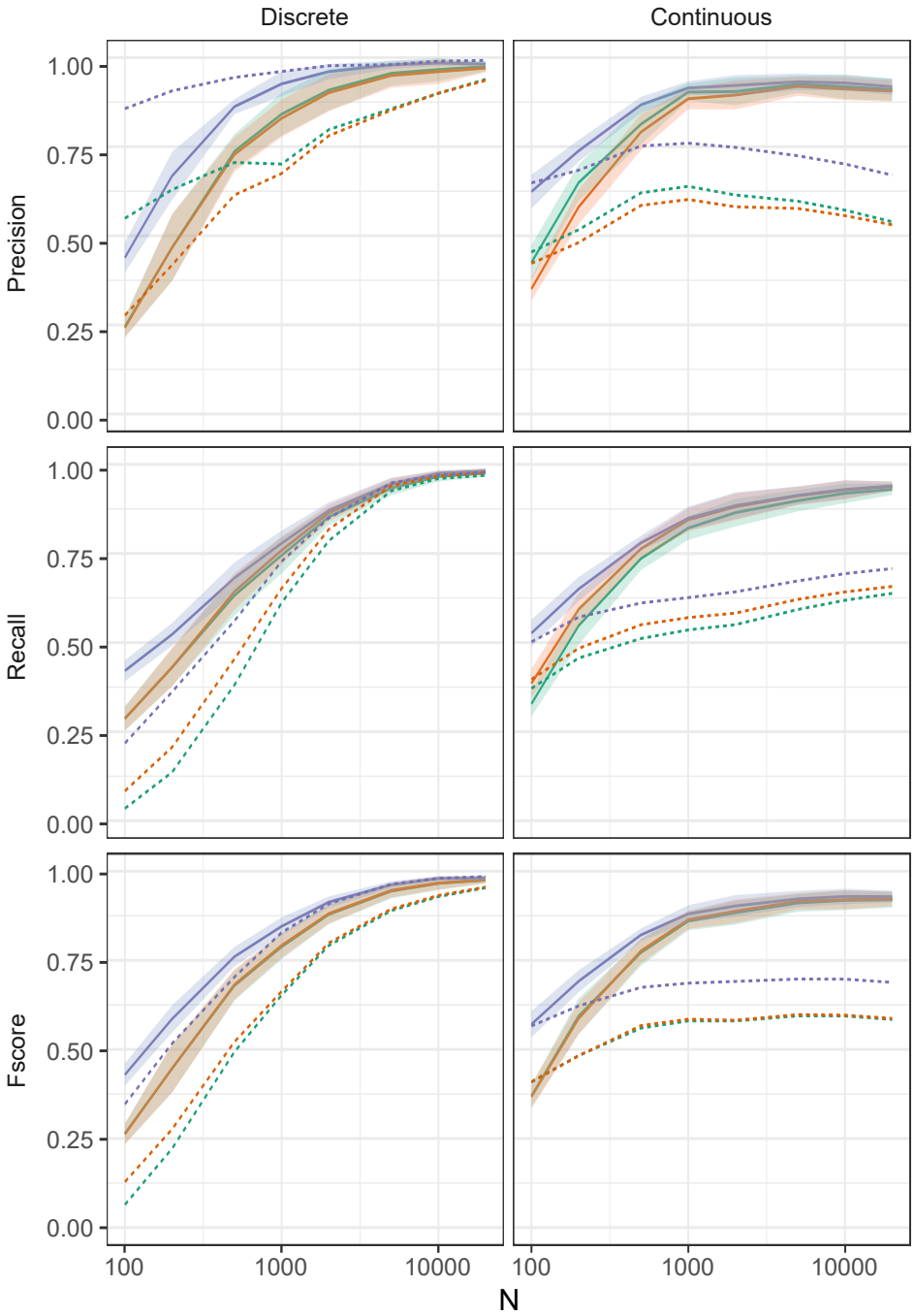

Figure 3: **Original MIIC with orientation rules allowing for negative NML-regularized MI & CMI on discrete data (left) and negative NML-regularized CMI on continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. MIIC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green). PC average scores for majority orientation rules are shown as dashed lines for comparison.

2.3 MIIC performance on continuous data, allowing for negative NML-regularized CMI

More recently MIIC was extended to handle continuous as well as mixed-type variables (either combination of discrete and continous variables or variables with both continuous and discrete ranges of values), for which MI & CMI are notoriously more difficult to estimate (Cabeli et al., 2020).

While distance-based k-nearest neighbor (kNN) estimates of MI and CMI are often used for continuous variables (Kraskov, Stögbauer, and Grassberger, 2004; Frenzel and Pompe, 2007), MIIC's MI and CMI estimates are instead computed through an approximate optimum discretization scheme, based on a general MI supremum principle (Cover and Thomas, 2006) regularized for finite datasets and using an efficient $\mathcal{O}(N^2)$ dynamic programming algorithm (Cabeli et al., 2020). This approach finds optimum partitions, $\mathcal{P}$ and $\mathcal{Q}$, specifying the number and positions of cut-points of each continuous variable, X and Y, to maximize the NML-regularized MI between them,

$$I'_N(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I'_N([X]_\mathcal{P};[Y]_\mathcal{Q}) \tag{4}$$

The NML regularization term, introduced in $I'_N([X]_\mathcal{P};[Y]_\mathcal{Q})$, is necessary for finite datasets and amounts to a model complexity cost, which eventually out-weights the information gain in refining bin partitions further, when there is not enough data to support such a refined model (Cabeli et al., 2020).

Such optimization-based estimates of MI are at par with alternative distance-based kNN approaches but have also the unique advantage of providing an effective independence test to identify independent continuous or mixed-type variables (Cabeli et al., 2020). This is achieved when partitioning X and Y into single bins maximizes the NML-regularized MI in Eq. 4, which vanishes exactly, in this case, with dramatic reductions in sampling error and variance (Cabeli et al., 2020). By contrast, kNN-MI estimates still need an actual independence test to decide whether some variables are effectively independent or not, as kNN MI estimates are never exactly null.

MIIC Precision, Recall and F-score on continuous data are comparable to those on discrete data, Fig. 3, and typically much better than the results obtained with traditional constraint-based methods, which, unlike MIIC, need to rely on independence tests, that are notoriously difficult for continuous data.

However, by contrast with discrete data, the remaining loss between skeleton and oriented graph precisions appears to differ between the CPDAG score and the oriented-edge-only subgraph score used for the comparison, Fig. 3. It indicates that the precision of the oriented-edge-only subgraph is slightly though significantly better than for the overall partially oriented graph, with a small concomitant loss of orientation recall, at small sample sizes, Fig. 3. This trend is due to the more stringent condition for v-structure orientation brought by the non-negative NML-regularized MI estimates obtained by MIIC for continuous variables. Yet, the optimum partitioning principle only applies to MI (Cover and Thomas, 2006), not CMI, which need to be estimated through the *difference* between optimum NML-regularized MI terms, as $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$ (Cabeli et al., 2020). As a result, the approximate NML-regularized CMI estimates between conditionally independent variables can sometime be negative and lead to v-structure orientations contradicting conditional independence, as discussed for discrete data above.

2.4 Improving MIIC causal discovery by rectifying negative NML-regularized MI & CMI

The general MI supremum principle (Cover and Thomas, 2006), regularized in Eq. 4 for finite datasets, is theoretically valid for any type of variables, not just continuous variables. In particular, it could be applied to small size datasets with discrete or categorical variables with many levels. It would result in the merging of rare levels to better estimate MI and CMI between weakly dependent discrete variables. Ultimately, MI estimates between independent discrete variables should lead to the merging of each variable into a single bin, thereby, resulting in NML-regulatized MI estimates to vanish exactly in this case, as already observed for continuous variables (Cabeli et al., 2020). As a result, optimum NML-regulatized MI should be non-negative as well as, by extension, NML-regulatized CMI, as shown now.

Theorem 1. *Optimum NML-regulatized MI and NML-regulatized CMI are non-negative.*

Proof. We first address optimum NML-regularized MI, noting that $I'_N(X;Y) \geqslant I'_N([X]_1;[Y]_1) = 0$, where $[X]_1$ and $[Y]_1$ are the X and Y variables partitioned into single bins, which leads to a vanishing

NML-regularized MI, as both MI and NML complexity cost are null, in this case, as $k_N^{\mathrm{NML}}(X;Y) = 0$ for $r_x = r_y = 1$ (Affeldt and Isambert, 2015).

Then, NML-regularized CMI is defined as the *difference* between optimum NML-regularized MI terms as, $I'_N(X;Y|U) = I'_N(Y;\{X,U\}) - I'_N(Y;U) = I'_N(X;\{Y,U\}) - I'_N(X;U)$. However, partitioning X and Y into a single bin leads to $I'_N(Y;\{X,U\}) \geqslant I'_N(Y;\{[X]_1,U\}) = I'_N(Y;U)$ and $I'_N(X;\{Y,U\}) \geqslant I'_N(X;\{[Y]_1,U\}) = I'_N(X;U)$ thus implying $I'_N(X;Y|U) \geqslant 0$ □

Following these considerations on the negativity of NML-regularized (C)MI with MIIC original orientation implementation, we propose a small modification, based on Theorem 1 and referred to as conservative MIIC, by analogy to the conservative orientation rules of traditional constraint-based methods (Ramsey, Spirtes, and Zhang, 2006), as noted above.

Proposition 2. *Conservative MIIC rectifies negative values of NML-regularized (C)MI, indicating (conditional) independence, to null values instead.*

The effects on this modification on discrete and continuous benchmark data are show in Fig. 4. While conservative MIIC hardly affects skeleton scores, it clearly has an impact on CPDAG and oriented-edge-only subgraph scores, which exhibit different trends relative to their original MIIC values.

CPDAG Precision, Recall and, hence, F-scores appear to be slightly lower under conservative MIIC (Fig. 4) than with original MIIC (Fig. 3), for discrete data. This illustrates the overall "better" orientation/non-orientation scores of the original MIIC against the theoretical CPDAG objective. Indeed, allowing for negative NML-regularized MI enables to infer weakly supported v-structures at small sample sizes. Besides, no significant difference is observed for CPDAG scores on continuous data, as original MIIC already enforces non-negative NML-regularized MI through optimization for continuous data (Cabeli et al., 2020), suggesting that enforcing also non-negative NML-regularized CMI with conservative MIIC has little impact on the reliability of CPDAG scores for continuous data, at least for the benchmarks tested here.

By contrast, conservative MIIC is found to greatly improve the precision of oriented-edge-only subgraphs, on discrete datasets, even for relatively small sample sizes, Fig. 4. This large increase in orientation precision is achieved at the expense of a relatively small loss of orientation recall. Hence, conservative MIIC significantly enhances the reliability and sensitivity of predicted orientations for all sample sizes, as compared to traditional constraint-based methods with conservative orientation rules, Fig. 4. For instance, conservative MIIC already reaches nearly 90% orientation precision with 25% orientation recall for $N \simeq 250$ (against about 80% orientation precision with only 5% orientation recall for conservative PC). While, by the time conservative PC reaches 90% orientation precision with 25% orientation recall for $N \simeq 700$, conservative MIIC achieves nearly 100% orientation precision with 50% orientation recall, Fig. 4. In addition, while original MIIC achieves a significantly better 65% orientation recall for $N \simeq 700$, Fig. 3, its orientation precision simultaneously drops to about 75%, which clearly impacts its reliability for causal discovery.

On continuous data, conservative MIIC also achieves a large increase in orientation precision, which becomes at par with skeleton precision, even for small datasets, and clearly much better than the corresponding scores obtained with traditional constraint-based methods for large datasets, Fig. 4. For instance, conservative MIIC reaches nearly 75% orientation precision with 50% orientation recall for $N \simeq 200$ (against about 70% orientation precision with 35% orientation recall for conservative PC). While, by the time conservative PC reaches 75% orientation precision with 45% orientation recall for $N \simeq 1,000$, conservative MIIC achieves more than 90% orientation precision with 80% orientation recall, Fig. 4.

3 Data generation and benchmarks

Datasets were simulated using structural equations models (SEMs) following the causal order of randomly generated DAGs. Continuous examples were constructed using linear and non-linear functions, and discrete datasets using unique state probabilities for each of the parents' combinations. The DAGs themselves were randomly drawn from the space of all possible 100 node DAGs (Melancon and Philippe, 2004) allowing for a maximum degree of 4 neighbors, resulting in an average degree of 2.7. Further details and dataset examples can be found in Cabeli et al. (2020).

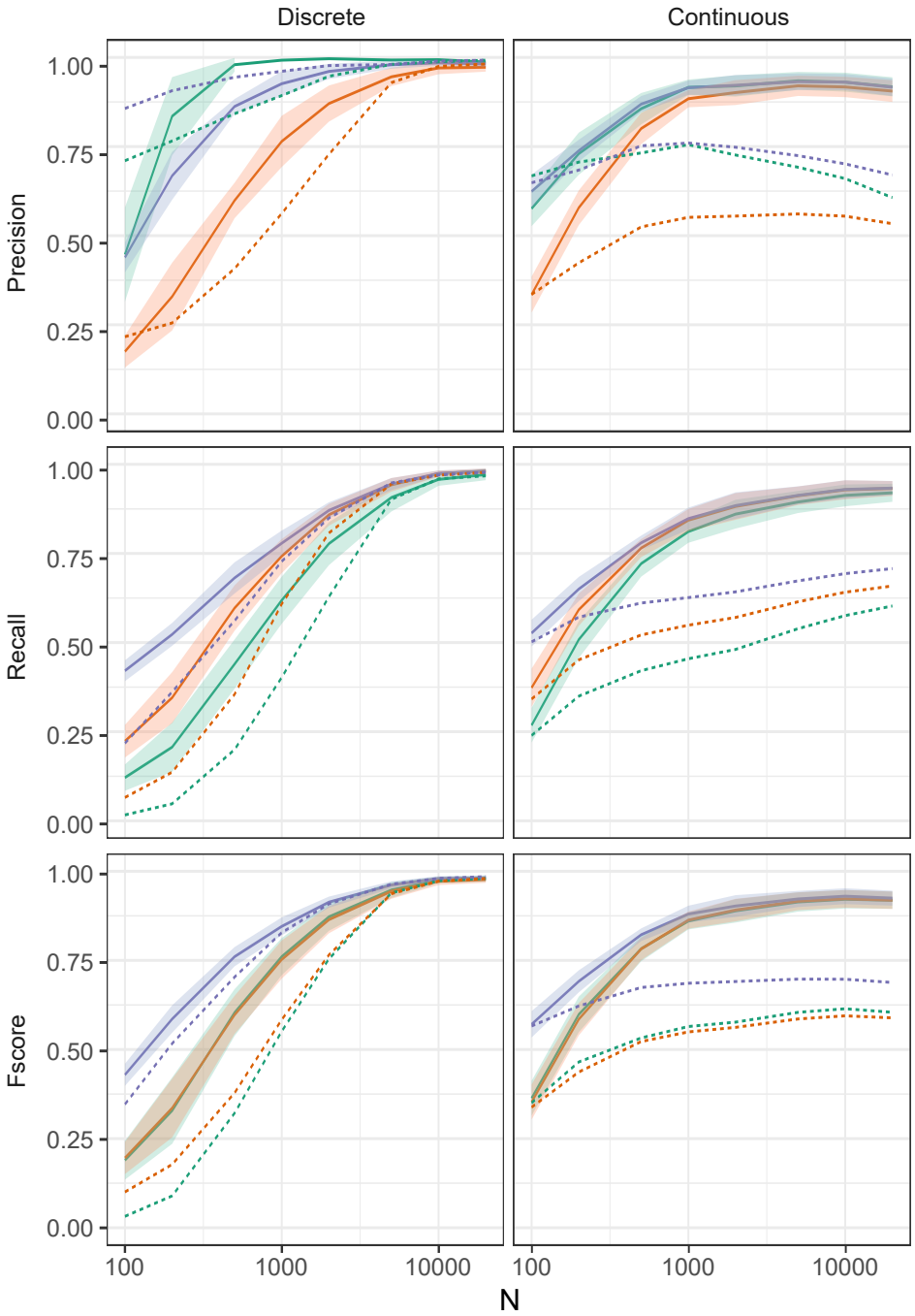

Figure 4: **Conservative MIIC with new orientation rules enforcing non-negative NML-regularized MI & CMI on discrete data (left) as well as continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. Conservative MIIC structure learning performance is measured in terms of Precision, Recall and F-scores ($\pm\sigma$) for skeleton (blue), CPDAG (red) and oriented-edge-only subgraph (green). PC average scores for conservative orientation rules are shown as dashed lines for comparison.

For evaluation purposes, network reconstruction was treated as a binary classification task and classical performance measures, Precision, Recall and F-score, were first used to evaluate skeleton reconstruction, based on the numbers of true *versus* false positive (TP *vs* FP) edges and true *versus* false negative (TN *vs* FN) edges, irrespective of their orientation.

Then, in order to evaluate edge orientations, we also define two orientation-dependent measures.

The first measure, referred to as the "CPDAG" score, aims to score the overall reconstruction with regards to the equivalence class of the true DAG. Edge types are used to redefine the orientation-dependent counts as, $TP' = TP - TP_{misorient}$ and $FP' = FP + TP_{misorient}$ with $TP_{misorient}$ corresponding to all true positive edges of the skeleton with a different orientation/non-orientation status as in the true CPDAG. The CPDAG precision, recall and F-score were then computed with the orientation-dependent TP' and FP'. In particular, the CPDAG score equivalently rates as "false positive" the erroneous orientation of an non-oriented edge in the CPDAG and the erroneous non-orientation of an oriented edge in the CPDAG. However, these errors are not equivalent from a causal discovery perspective.

The second measure, referred to as oriented-edge-only score, uses the same metrics but is restricted to the subgraphs of the CPDAG and the inferred graph containing oriented edges only. It is designed to specifically assess the method performance with regards to causal discovery, that is, on the oriented edges which can in principle be learnt from observational data *versus* those effectively predicted by the causal structure learning method.

MIIC was run with default parameters for all settings on the latest version (available at `https://github.com/miicTeam/miic_R_package`), and PC with the `pcalg` package (Kalisch et al., 2012) using `bnlearn`'s (Scutari, 2010) mutual information test for discrete datasets and rank correlation for continuous ones. For PC, the α threshold for significance testing was tuned for each sample size N and network type to produce the best average between skeleton and "CPDAG" F-scores using a zeroth order optimization implemented in `dlib` (King, 2009).

4 Conclusion

Causal uncertainty and limited sensitivity of traditional constraint-based methods have so far hampered their dissemination for a wide range of possible causal discovery applications on real-life observational datasets. Hence, fulfilling the promise of causal discovery methods in the new data analysis area requires to improve their reliability as well as scalability.

We propose and implement, in this paper, a simple modification of the recent causal discovery method, MIIC, which greatly enhances the reliability of predicted orientations, for all sample sizes, with only a small sensitivity loss compared to MIIC original orientation rules. This conservative MIIC approach is especially interesting, in practice, to improve the reliability of cause-effect discovery for real-life observational data applications.

References

Affeldt, S., and Isambert, H. 2015. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, 42–51.

Affeldt, S.; Verny, L.; and Isambert, H. 2016. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* 17(S2):12.

Cabeli, V.; Verny, L.; Sella, N.; Uguzzoni, G.; Verny, M.; and Isambert, H. 2020. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLOS Computational Biology* 16(5):e1007866.

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15:3741–3782.

Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory*. Wiley, 2nd edition.

Frenzel, S., and Pompe, B. 2007. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.* 99:204101.

Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M. H.; and Bühlmann, P. 2012. Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* 47(11):1–26.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10:1755–1758.

Kontkanen, P., and Myllymäki, P. 2007. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.* 103(6):227–233.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Phys. Rev. E* 69:066138.

Li, H.; Cabeli, V.; Sella, N.; and Isambert, H. 2019. Constraint-based Causal Structure Learning with Consistent Separating Sets. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 14257–14266.

Melancon, G., and Philippe, F. 2004. Generating connected acyclic digraphs uniformly at random. *arXiv:cs/0403040*. arXiv: cs/0403040.

Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.* 441–452.

Pearl, J. 2009. *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition.

Ramsey, J.; Spirtes, P.; and Zhang, J. 2006. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, UAI, 401–408. Oregon, USA: AUAI Press.

Roos, T.; Silander, T.; Kontkanen, P.; and Myllymäki, P. 2008. Bayesian network structure learning using factorized nml universal models. In *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)*. IEEE Press.

Scutari, M. 2010. Learning bayesian networks with the bnlearn r package.

Sella, N.; Verny, L.; Uguzzoni, G.; Affeldt, S.; and Isambert, H. 2018. Miic online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* 34(13):2311–2313.

Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9:62–72.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2nd edition.

Verny, L.; Sella, N.; Affeldt, S.; Singh, P. P.; and Isambert, H. 2017. Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* 13(10):e1005662.

Chapter 5

Applications

In this chapter, we take advantage of the various improvements to MIIC and infer causal graphs from real-life data. In each case, we pre-processed the data in close collaboration with the teams that were responsible for their collection, and benefited from their insight again when analyzing the results.

5.1 Learning causal graphs from medical records of patients with cognitive disorders

The first application of MIIC with the mixed case estimator was on medical records of elderly patients from La Pitié-Salpêtrière Hospital. In the final network, 107 variables of 1,628 patients admitted for cognitive disorder were used for the network reconstruction.

Graphical model reconstruction methods are not commonly used on clinical databases; however, these techniques can be of great help in understanding the structure of the data. Two main benefits emerged from this application on medical records. First, it helps in understanding the relationship between variables : knowing which aspects of cognition are correlated is fairly intuitive for experienced physicians, but it is much harder to conceptualize mediating effects and discerning direct from indirect relations. Analyzing this kind of data also performs quality control in a way : if a link is missing, we can easily understand why by analyzing the detailed information of the network reconstruction, but also by directly observing the dataset. Secondly, it is able to infer previously ignored associations, which could also be a considerable advantage in the initial analysis of this type of varied and complex data set. In this network for example, an unexpected direct edge was inferred between the Fazekas (which measures the amount of hyperintensity in white matter attributed to chronic small vessel ischemia) and Scheltens (medial temporal lobe atrophy) scales, which was indepedently reported recently [123].

The full discussion of the network is published in [7], which is included in Section 3.3.

5.2 NEOREP study on breast cancer patients

The second application, also on medical records of patients, concerns breast cancer patients that have undergone neoadjuvant (*i.e.* before the main treatment, usually surgery) chemotherapy.

In this article, we emphasize the benefits of visually inspecting the dataset after having removed the indirect relationships with MIIC. We note at least three ways in which we benefit from the network approach. First, as it was highlighted in the previous application, it performs an unbiased check of the dataset. For example, it was able to identify differences in clinical practices between the two treatment centers of the cohort which may create instances of Simpson's paradox if not accounted for. By including the node in the dataset, it is a potential contributing node to any relationship and we make sure we remove the effect of the center for each remaining edge. Secondly, MIIC traces the natural course of the disease by having access to patients that are both early in the diagnosis or long after initial treatment. This allows to predict statistically the course of the disease, based on this specific population of patients. Finally, MIIC identifies factors likely to improve prediction or prognosis. Specifically, the direct neighbors of the node reporting the vital status of the patient give some unique information about the outcome that could not be removed due to indirect effects, and merit particular attention. Moreover, the optimal discretization scheme informs us of the most informative cutpoints with regards to the outcome of the disease for the observed population. In this application specifically, we observe that the discretized version of a composite score, which is usually used for prognosis, is less informative of the vital status than its continuous variable. This highlights the fact that many bio-medical discrete variables are discretized using a priori bins and may not be adapted for all situations.

The manuscript was written in collaboration with the Residual Tumor and Response to Treatment Laboratory and is yet to be submitted, it is included in the next subsection.

5.2.1 Manuscript

Interactive exploration of a global clinical network from a large breast cancer cohort

Nadir Sella[a,b,1], Anne-Sophie Hamy[a,c,d,2], Vincent Cabeli[b,3], Lauren Darrigues[d], Marick Laé[e,f], Fabien Reyal[a,d], Hervé Isambert[b]

a. Residual Tumor & Response to Treatment Laboratory, RT2Lab, INSERM, U932 Immunity and Cancer, Institut Curie, Paris, F-75248, France.

b. Laboratoire Physico Chimie Curie, Institut Curie, PSL Research University, CNRS UMR168, 75005, Paris, France

c. Department of Medical Oncology, Institut Curie, Saint-Cloud, F-92230, France.

d. Department of Surgery, Institut Curie, Université Paris, Paris, F-75248, France.

e. Department of Tumor Biology, Institut Curie, Paris, F-75248, France.

f. Department of Pathology, Henri Becquerel Cancer Center, INSERM U1245, UniRouen Normandy University

Corresponding Author information: Hervé Isambert, herve.isambert@curie.fr, +33 1 56 24 64 74

Co-corresponding Author information : Pr Fabien REYAL, 0033144324087, 0033615271980, fabien.reyal@curie.fr

Author contributions: NS, VC and HI designed the analytic tool; ASH, LD, ML performed research; NS, VC, HI, LD, FR and ASH analyzed data; LD and ASH wrote the paper.

Competing Interest Statement: The authors have no potential conflicts of interest to declare.

[1]N.S., [2]A-S.H. and [3]V.C. **contributed equally** to this work.

Dual Classification (Major / Minor): Biological Sciences / Medical Sciences & Physical Sciences / Computer Sciences

Keywords: Machine learning - data visualization- residual cancer burden - neoadjuvant chemotherapy – breast cancer.

Abstract

With the rapid accumulation of information from medical records in health databases, there is an urgent need for innovative interactive tools specifically designed for the exploration of these data by medical practitioners. Here, we report a novel interactive graphical interface for use as the front end of a machine learning causal inference server (MIIC), to facilitate the visualization and comprehension by clinicians of relationships between clinically relevant variables. We demonstrate the utility of the MIIC interactive interface, by exploring the clinical network of a large cohort of breast cancer patients treated with neoadjuvant chemotherapy (NAC). This example highlights, in particular, the direct and indirect links between post-NAC clinical responses and patient survival. The MIIC interactive graphical interface has the potential to help clinicians to identify actionable nodes and edges in clinical networks, thereby ultimately improving the patient care pathway.

Significant statement:

Despite unprecedented amount of information now available in medical records, health data remain underexploited due to their heterogeneity and complexity. Simple charts and hypothesis-driven statistics can no longer apprehend the content of information-rich clinical data. There is, therefore, a clear need for powerful interactive visualization tools enabling medical practitioners to perceive the patterns and insights gained by state-of-the-art machine learning algorithms. We report here an exploratory analysis of a global clinical network from a large breast cancer cohort, with a novel interactive graphical interface for the exploration of health data. The widespread use of such tools, facilitating the interactive exploration of datasets, is crucial both for data visualization and for the generation of research hypotheses.

INTRODUCTION

The availability of health data from patient medical records is increasing, and these data constitute, in theory, a rich resource for research purposes. However, despite the unprecedented amount of information now available, health data remain underexploited due to their heterogeneity and complexity. There is, therefore, an urgent need for innovative tools, based on intuitive and interactive graphical interfaces, specifically designed for the exploration of health data by medical practitioners. Data visualization is gradually emerging as a new field of research, and graphical representations are used for two main purposes: (i) explanatory illustration, to highlight novel scientific insights graphically and to ensure efficient communication between scientists [1–4]; and (ii) exploratory analysis, searching for relationships previously overlooked and leading to new discoveries, thereby maximizing the potential of information-rich databases. We present here an *exploratory analysis* of a global clinical network from a large breast cancer cohort, with a novel interactive graphical interface for the exploration of health data.

We previously developed an advanced computational method for graphical analyses, including causal relationships, from multivariate data[5]. The underlying MIIC (multivariate information inductive causation) algorithm, which was released as an online server[6], uses a machine learning method combining constraint-based and information theory approaches to reconstruct causal, non-causal or mixed networks from large datasets. The MIIC algorithm was first developed to analyze categorical genomic data[5, 6] and has recently been extended to the analysis of more challenging heterogeneous datasets, such as medical records, combining both categorical and continuous variables, in which interdependence is notoriously difficult to assess[7].

Breast cancer (BC) clinical datasets are particularly suitable for the type of exploratory analysis presented here, as BC is a complex heterogeneous disease highly variable in its aggressiveness and prognosis. BC remains one of the leading causes of cancer-related death among women. The BC patients included in the cohort analyzed here were treated with neoadjuvant (or

preoperative) chemotherapy (NAC). NAC was originally restricted to patients with inflammatory or locally advanced BC, but is now the standard care for aggressive early-stage breast cancers, *i.e.* triple-negative (TNBC) and *HER2*-positive BCs[8, 9]. From the patient's viewpoint, the benefits of the neoadjuvant strategy include a greater feasibility of breast-conserving surgery and the prognostic stratification of risk obtained after analyses of the residual tumor burden at surgery. From the research and development standpoint, the neoadjuvant setting makes it possible to monitor the chemosensitivity of the tumor *in vivo*, and provides an opportunity for the rapid validation of research hypotheses and the acceleration of drug approval.

The novel interactive graphical interface described here, designed as a front-end for the MIIC server, should enable clinicians to visualize and understand the relationships between clinically relevant variables, such as post-NAC clinical responses and survival. In addition, the MIIC interactive graphical interface should help practitioners to identify actionable nodes and edges in clinical networks, potentially leading to improvements in the patient care pathway.

MATERIALS AND METHODS

Patients and treatment

We analyzed a cohort of 1197 patients with non-metastatic BC treated by NAC, with or without trastuzumab, followed by surgery, at either of the two Institut Curie sites (Paris and Saint Cloud) between 2002 and 2012 (NEOREP Cohort, CNIL declaration number 1547270). We included unilateral, non-recurrent, non-inflammatory, non-metastatic tumors, and excluded T4 tumors. This study was conducted in accordance with institutional and ethical rules regarding research on tissue specimens and patients. Information on family history, clinical characteristics (age; menopausal status; body mass index) and tumor characteristics (clinical tumor stage and grade; histology; clinical nodal status; ER, PR and *HER2* status; BC subtype; mitotic index; Ki67; lymphovascular invasion) was retrieved from electronic medical records. All the patients of the cohort received NAC, and additional treatments were decided in accordance with national guidelines.

Tumor samples and pathological review

In accordance with the guidelines used in France (Group for Evaluation of Prognostic Factors using Immunohistochemistry in Breast Cancer[10], cases were considered estrogen receptor (ER)-positive or progesterone receptor (PR)-positive if at least 10% of the tumor cells expressed estrogen and/or progesterone receptors (ER/PR). Endocrine therapy was prescribed when this threshold was exceeded. *HER2*-negative status was defined as a score of 0 or 1+ for the tissue section stained by immunohistochemistry (IHC). Tissue sections with scores of IHC 2+ or IHC 3+ were then analyzed by fluorescence *in situ* hybridization (FISH) to confirm *HER2* positivity. BC tumors were classified into subtypes (TNBC, *HER2*-positive, and luminal *HER2*-negative [referred to hereafter as "luminal"]). BC subtypes were defined as follows: luminal, ER^+ or PR^+/ $HER2^-$; TNBC, $ER^-/PR^-/HER2^-$; *HER2*-positive BC, $HER2^+$. Pretreatment core needle biopsy specimens and/or the corresponding post-NAC surgical specimens were reviewed independently by breast disease experts for research purposes, to assess residual cancer burden index, and the levels of tumor-infiltrating lymphocytes. The pathological reviews of these specimens are described in detail elsewhere[11–13]. Pathological complete response (pCR) was defined as the absence of residual invasive cancer cells in the breast and axillary lymph nodes (ypT0/is þ/ypN0).

Survival endpoints

Relapse-free survival (RFS) was defined as the time from surgery to death, locoregional recurrence or distant recurrence, whichever occurred first. Overall survival (OS) was defined as the time from surgery to death. The date of last known contact was retained for patients for whom none of these events were recorded. The cutoff date for survival analysis was March, 13th, 2019.

Variables of interest

The care pathway of BC patients eligible for neoadjuvant chemotherapy can be summarized as follows: *i)* pretreatment biopsy for BC diagnosis; *ii)* administration of chemotherapy as the first-line treatment; *iii)* removal of the tumor by surgery; *iv)* histological analysis of the specimens obtained; *v)* prescription of adjuvant treatments, if indicated (radiotherapy, hormonotherapy, chemotherapy); (*vi*) patient follow-up to monitor for relapse or death. We identified 94 clinically relevant variables from clinical, radiological, pathological and outcome data, which we grouped into 14 categories (hospital, history, co-medication, comorbidities, clinical baseline, baseline histology, pre-NAC pathology, treatment response, surgery, treatment, changes during NAC, post-NAC pathology, delayed relapse/survival, metastasis). For composite variables derived from raw variables (e.g. BC subtype, constructed from a combination of ER status, PR status, *HER2* status), both the derived and raw variables were represented on the network.

MIIC algorithm

The functioning of the algorithm has been described in detail elsewhere[5, 7]. Briefly, starting from a fully connected network, the MIIC algorithm first removes dispensable edges by iteratively subtracting the most significant information contributions from indirect paths between each pair of variables. The remaining edges, the underlying effect of which cannot be explained by indirect paths, are then oriented based on the causality signature in the data.

The original algorithm was restricted to categorical variables[5], but MIIC has recently been extended to include continuous variables, the values of which are partitioned into optimal bins, maximizing mutual information with another (continuous or categorical) variable of interest, while preventing the overfitting of datasets of finite size due to the use of too many bins[7]. In particular, each continuous variable may have different information-maximizing partitions depending on the associated variable of interest. For instance, MIIC finds three maximally informative bins for the residual cancer burden (RCB) score in association with patient survival status (**Fig. S1A**), whereas

eight RCB bins are required to estimate its mutual information with post-NAC cellularity correctly (**Fig. S1B).**

MIIC online server

The MIIC online server is freely accessible at https://miic.curie.fr/index.php, and can be used with the Google Chrome, Mozilla Firefox, Edge, and Safari browsers. The user guide summarizing the main steps for running the MIIC algorithm is accessible at https://miic.curie.fr/user_guide.php, and an online video tutorial is available from: https://miic.curie.fr/tutorial.php. The workbench is available from https://miic.curie.fr/workbench.php. As input data, the user can upload a dataset formatted as a table with commas, semicolons, tabs, pipes or colons, as eld separators, without sample ID. Each variable can be either categorical or quantitative (discrete or continuous). Variables can be grouped into families, identified with different colors on the network. Missing values are allowed in the dataset and their possible statistical biases are taken into account by MIIC[7]. They should be indicated as "NA" in the dataset table. Once the dataset has been prepared, the user runs the algorithm, and an e-mail is sent when the job is completed.

MIIC output

The MIIC online server generates a visualization of the global network of the dataset. An example based on the NEOREP dataset is displayed in Fig. 1, and is accessible as an interactive network at https://miic.curie.fr/job_results_showcase.php?id=NEOREP.

Each node corresponds to a variable of the dataset, with continuous variables displayed as circles, and categorical variables displayed as squares. The color of the node indicates the group to which the variable belongs, as specified by the user.

Each edge corresponds to a "direct" association between two variables, that is, a statistical association that cannot be entirely explained by indirect effects involving other variables. Red and blue edges correspond to positive and negative (*i.e.* anti-correlated) associations, respectively. Four types of

edge orientations are distinguished by the MIIC online server: *i)* directed edges with a gray arrowhead represent inferred causal relationships; *ii)* bidirected edges (drawn with dashed lines) reflect the presence of a latent common cause (L) unobserved in the available dataset, *i.e.* $X \leftarrow (L) \rightarrow Y$; *iii)* directed edges with a colored (red or blue) arrowhead consistent with either a causal or a latent common cause relationship; and *iv)* undirected edges, the orientation of which, if indeed there is one, cannot be inferred from non-perturbative data.

Interactive exploration of the network

The distributions and neighborhoods of each node and edge of the inferred network can be explored through an interactive interface, through the mouse-over right- or left-click buttons on the browser page, as detailed in the online tutorials. Briefly, any variable can be highlighted by clicking on the network or through the "Search" toolbox (Fig. S2A). The corresponding plots can be downloaded as .png or .svg images. Each node can be explored individually in terms of counts (categorical variables, Fig. S2B-C) or distribution (continuous variables Fig. S2D-E). Each edge can be explored by a right click and the choice of "plot join distribution" or "plot discretization". The resulting plots are (i) proportion plots, with the edge representing the association between two categorical variables (Fig. S2F); (ii) distribution histograms (Fig. S2G) or boxplots (Fig. S2H), in which the edge represents the association between a categorical and a continuous variable or (iii) scatter plots (Fig. S2I), in which the edge represents the association between two continuous variables. All the plots can be customized by zooming in and out. Additional options include inverting the *x* and *y* axes, the choice of frequency or absolute counts, or NA removal (proportion plots), and faceting or superimposing the variables (distribution histograms). All the figures presented here were generated with the MIIC online interactive visualization tool.

RESULTS

The global network displayed in Fig. 1 is accessible at

https://miic.curie.fr/job_results_showcase.php?id=NEOREP.

We discuss below some of the links inferred in the NEOREP network after grouping according to several clinically relevant concepts identified from published studies on BC.

MIIC performs quality control

MIIC identifies inherent associations between variables

The duration of neoadjuvant treatment is directly linked to the type of NAC regimen delivered (Fig. 2A) reflecting the fact that anthracycline-based (AC) regimens usually include four cycles (median of 106 days, Fig. 2B), whereas sequential regimens in which anthracyclines are followed by taxanes are generally administered over six or eight cycles (median of 147 days, Fig. 2B). The number of nodes retrieved is associated with the type of axillary surgery (Fig. 2C), consistent with the fact that sentinel node (SLN) biopsy procedures were developed to reduce the number of lymph nodes removed during dissection (LND) (Fig. 2D)[14]. MIIC correctly represents the direct links between residual cancer burden (RCB) (Fig.2E) and the patterns making up this score, derived from measurements on the primary tumor bed (size, fraction of invasive cancer, cellularity) and the regional lymph nodes (number of positive lymph nodes).

MIIC first identifies relationships between a disease and the corresponding treatment. ER positivity — which is predictive of efficacy for anti-hormonal treatment[15]— is associated with the use of endocrine therapy (Fig. S3A), and a similar association is observed for *HER2*-positivity and trastuzumab use (Fig. S3B)[16]. Beyond cancer, significant associations are also found between depression and the use of psycholeptics (Fig. S3C), between thyroid disorders and thyroid hormone use (Fig. S3D), and between hypertension and drugs for the treatment of cardiovascular diseases (Fig. S3E). More generally, comedication use is associated with the type of NAC (Fig. S3F), reflecting the greater likelihood of less toxic regimens being prescribed to fragile patients (patients on other types of medication) than to patients without comedication[17–19].

MIIC then identifies clinical factors known to be epidemiologically related (Fig. S4A). Menopause, a process occurring in older women, is directly linked to age (Fig. S4B) (median age: 43 years for premenopausal, versus 58 years for postmenopausal women). Postmenopausal status is associated with dyslipidemia (Fig. S4C)[20]. Consistent with these associations, body mass index (BMI) increases with age (Fig. S4A, S4D) and both factors, which have been reported to increase cardiovascular risks, are linked to hypertension (Fig. S4A, S4E). The number of drugs taken by a patient (comedication) increases with the number of comorbidities (Fig. S4A, S4F).

MIIC identifies intra- and inter-modality associations

For the variables derived from pathology records, MIIC found associations between tumor grade, Ki67, and mitotic index (Fig. S5A-B-C), all of which are markers of tumor proliferation[21]. MIIC can also visualize links between patterns assessed in different ways. Measurements of pre-NAC tumor size evaluated clinically, by mammography and by MRI, were found to be closely related (Fig. S5C-E) as previously reported[22, 23]. Similarly, the response to treatment assessed clinically at NAC completion was found to be associated with histological size based on the surgical specimen (Fig. S5F).

MIIC provides insight into tumor biology and response to treatment

The presence of lymphovascular invasion (LVI) in the post-NAC specimen is associated with a higher RCB index, consistent with the strong resistance to chemotherapy of these tumors[12] (Fig. S6A). TNBCs and *HER2*-positive tumors have a higher pre-NAC mitotic index and more stromal TIL infiltration (Fig. S6B-C) than luminal BCs[24, 25]. Consistently, high TIL levels are significantly associated with histological grade 3 tumors (Fig. S6D).

MIIC reflects clinical practice

Several associations highlighted in the network reflect clinical practice decisions applied throughout BC centers. For example, the likelihood of performing conservative breast surgery depends on tumor histology (higher rates of mastectomy have been reported for patients with lobular or other histological types of tumor less likely to respond to NAC) [26, 27] (Fig. S7A) and is positively associated with the practice of oncoplastic surgery[28] (Fig. S7B). Similarly, lumpectomy is more frequently associated with radiation therapy than with mastectomy (Fig. S7C)[29–32]. After surgery, the addition of a second line of treatment by adjuvant chemotherapy, to decrease the risk of relapse, is driven by the identification of factors associated with a poor prognosis[33], such as high levels of lymph node involvement (Fig. S7D).

Beyond these well-established practices, MIIC also identified differences in clinical practices between the two centers of the cohort (Fig. 3A). For example, oncoplastic surgery and adjuvant chemotherapy were performed at only one of the two centers (Fig.3B-C); the NAC regimen also differed between centers, with the Curie St Cloud center using more AC regimens than AC-taxane combinations, resulting in a shorter duration of NAC treatment (Fig. 3D-E).

MIIC traces the natural course of the disease

The natural course of BC may include local relapse, possibly followed by distant metastases, the trigger events leading to death[34–38] (Fig. 4A-C). Contralateral BC is often used in composite survival endpoints, such as distant relapse-free survival[39], but MIIC clearly identifies contralateral BC as an event being independent of other oncologic events and almost totally isolated from the rest of the network (Fig.1).

Luminal BC is known to recur and develop metastases later than *HER2*-positive BC and TNBC (Fig. 4D)[24, 25, 40, 41]. The link between has also been found between PR negativity and a higher risk of brain metastasis[42–46] (Fig. 4E).

MIIC identifies factors likely to improve prediction or prognosis

MIIC identifies unexpected associations, leading to new discoveries

With more than 15 associations involving treatment center (Fig. 3A), MIIC unmasked an unexpected “batch” effect relating to the site of BC treatment in this cohort. The observed differences reflect not only differences in therapeutic practice, but also differences in the characteristics of the population (differences in the proportion of women with psychological disorders, difference in incomes), differences in tumor presentation (tumor size), differences in pathological variable scoring (grade, presence of pre-NAC LVI, tumor cellularity, TILs), and differences in time to treatment within the care pathway.

MIIC also favors new discoveries. For example, comedication appears to protect against local relapse (Fig. 5A). Several retrospective studies have reported this association, with the use of statins[47], NSAIDs[48], or beta-blockers[49] found to have indirect anticarcinogenic effects. It has recently been suggested that these non-oncological treatments may have immunomodulatory and chemosensitizing effects[50].

MIIC suggests relevant combinations of predictive of prognostic biomarkers

MIIC may provide clues to combinations of new prognostic biomarkers likely to improve the prediction of response to chemotherapy, or post-NAC prognosis. Pre-NAC lymphovascular invasion (LVI) was found to be associated with both lower rates of clinical response (Fig. 5B) and shorter relapse-free survival (Fig. 5C). Both RCB (Fig. 5D-E) and post-NAC mitotic index (Fig.5D-F), a parameter rarely used in practice but nevertheless reported to be a predictor of BC recurrence[51, 52], appear to be strongly associated with the risk of death. MIIC may, therefore, be an efficient tool for identifying features likely to improve prognosis, by combining gold standard indicators with other parameters, such as post-NAC mitotic index, and post-NAC LVI, for example. Finally, MIIC also makes it possible to optimize the binning of residual cancer burden (RCB). RCB is a post-NAC histological score calculated as an increasing continuous index, and then subdivided into four classes

(0, I, II, and III)[53]. Our analysis based on information maximization principles suggested a new unsupervised classification of RCB scores into three categories (Fig. 5E), with RCB=0 with low RCB values merged, in particular, into a single class associated with a good prognosis.

DISCUSSION

When applied to a large cohort of BC patients, the MIIC algorithm successfully (i) performed quality controls; (ii) identified intra- and inter-modality correlations; (iii) highlighted differences in clinical practice, including center specificities; (iv) traced the natural course of the disease; (v) highlighted unsuspected and hidden associations, leading to new discoveries. The interactive visualization and causal analyses provided by this algorithm make it a promising tool for fast and effective explorations of the increasing amount of available health data.

The amount of exploitable health data is increasing exponentially. The best known health data resource for cancer studies remains the SEER (Surveillance, Epidemiology, and End Results) database, which collects data from population-based cancer registries covering approximately 34.6% of the US population[54, 55]. By 2016, the National Cancer Database (NCDB) had amassed more than 34 million hospital records from cancer patients (almost four times the size of the SEER database), to become the largest clinical cancer registry in the world[56]. In France, the French administrative health care database, the SNDS (*Système National des Données de Santé*), is one of the largest administrative databases in the domain of medicine, providing many opportunities for medical research[57, 58], as it covers 99% of the French population (about 66 million people). The French government is planning to ease access to this almost exhaustive population research resource, through release as part of the "Health data hub" project. Finally, beyond these structured databases, the largest mine of untapped data worldwide remains the content of electronic health records (EHRs), encompassing a full range of data (clinical notes, laboratory results, imaging, genetic data, etc.) relating to patient care. Recent advances in information technology have made it easier for both hospitals and healthcare institutions to collect large amounts of healthcare data.

Biomedical scientists are now facing new challenges in the management and analysis of massive, heterogeneous datasets[59]. These challenges include the development of tools for exploration and visualization, analytical methods, integration into a comprehensive overview, and translation of the findings into public health impact. The visualization of information makes it possible for users to find profound patterns in clinical data, through visual recognition. Simple charts cannot represent the complexity of big data analyses and fail to support multifaceted tasks effectively[3, 4]. There is, therefore, a need for sophisticated visualization tools dealing with many elements simultaneously and enabling users to perceive the patterns and insight generated by the algorithm[60]. Supplementary Table 1 shows the main data visualization tools used to present medical data. Many of the visual methods have been adopted directly from the field of data mining, but others, specific to the healthcare domain, have also been designed (Supplementary Table 2). For example, Happe and Drezen built the ePEPs toolbox, which displays relevant patterns extracted by eye from patient reimbursement data in the SNDS database, and supporting interactive exploration by researchers[61]. CARRE provides web-based components for interactive health data (fitness and biomarkers) visualization and risk analysis for the management of cardiorenal diseases[62]. The MITRE Corporation has also developed a web-based solution that provides an overview of an individual's health through graphical representations of EHR data, highlighting abnormal values[63]. None of these visualization programs has yet managed to bridge the gap between of the large amounts of clinical data available and the discovery of clinical knowledge or paths for scientific research. By processing large heterogeneous sets of variables inherent to clinical records, MIIC provides physicians with a full picture of BC disease.

In addition to this use for visualization, the MIIC algorithm presents several other advantages for analyses, including its unsupervised nature, overcoming the need for training or human involvement. This feature makes it possible to obtain new knowledge through the automatic identification of patterns and dependences in the data, highlighting new interactions, and it may be of use for feature selection in machine learning models.

In conclusion, MIIC, an open-access, interactive, multitask tool, is designed to visualize datasets to help clinicians and researchers to understand the relationships between the variables within them. It opens up promising perspectives for guiding the generation of new hypotheses, helping clinicians to identify actionable nodes and edges in clinical networks, and revealing new clues to relationships of interest for research purposes. Its widespread use in the field of health data could increase the accuracy of prediction for treatment responses and prognosis. This tool has the potential to improve the care pathway and, ultimately, patient survival.

References

1. Bärtschi M: Health Data Visualization-A review ∗ Seminar Collaborative Data Visualization, in 2015

2. Luo J, Wu M, Gopukumar D, et al: Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Inform Insights 8:1–10, 2016

3. Ola O, Sedig K: Beyond simple charts: Design of visualizations for big health data [Internet]. Online J Public Health Inform 8, 2016[cited 2019 Aug 14] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5302463/

4. Shneiderman B, Plaisant C, Hesse BW: Improving Healthcare with Interactive Visualization. Computer 46:58–66, 2013

5. Verny L, Sella N, Affeldt S, et al: Learning causal networks with latent variables from multivariate information in genomic data. PLoS Comput Biol 13:e1005662, 2017

6. Sella N, Verny L, Uguzzoni G, et al: MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. Bioinformatics 34:2311–2313, 2018

7. Cabeli V, Verny L, Sella N, et al: Learning clinical networks from medical records based on information estimates in mixed-type data [Internet]. PLoS Comput Biol 16, 2020[cited 2021 Feb 4] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7259796/

8. Brandão M, Reyal F, Hamy A-S, et al: Neoadjuvant treatment for intermediate/high-risk HER2-positive and triple-negative breast cancers: no longer an “option” but an ethical obligation. ESMO Open 4:e000515, 2019

9. Reyal F, Hamy AS, Piccart MJ: Neoadjuvant treatment: the future of patients with breast cancer. ESMO Open 3:e000371, 2018

10. [Recommendations for the immunohistochemistry of the hormonal receptors on paraffin sections in breast cancer. Update 1999. Group for Evaluation of Prognostic Factors using Immunohistochemistry in Breast Cancer (GEFPICS-FNCLCC)]. Ann Pathol 19:336–343, 1999

11. Hamy A-S, Pierga J-Y, Sabaila A, et al: Stromal lymphocyte infiltration after neoadjuvant chemotherapy is associated with aggressive residual disease and lower disease-free survival in HER2-positive breast cancer. Ann Oncol 28:2233–2240, 2017

12. Hamy A-S, Lam G-T, Laas E, et al: Lymphovascular invasion after neoadjuvant chemotherapy is strongly associated with poor prognosis in breast carcinoma. Breast Cancer Res Treat 169:295–304, 2018

13. Hamy-Petit A-S, Belin L, Bonsang-Kitzis H, et al: Pathological complete response and prognosis after neoadjuvant chemotherapy for HER2-positive breast cancers before and after trastuzumab era: results from a real-life cohort. Br J Cancer 114:44–52, 2016

14. Veronesi U, Paganelli G, Viale G, et al: Sentinel-lymph-node biopsy as a staging procedure in breast cancer: update of a randomised controlled study. Lancet Oncol 7:983–990, 2006

15. Burstein HJ, Temin S, Anderson H, et al: Adjuvant Endocrine Therapy for Women With Hormone Receptor–Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. J Clin Oncol 32:2255–2269, 2014

16. Wilson FR, Coombes ME, Wylie Q, et al: Herceptin® (trastuzumab) in HER2-positive early breast cancer: protocol for a systematic review and cumulative network meta-analysis. Syst Rev 6:196, 2017

17. Aaldriks AA, Maartense E, Nortier HJWR, et al: Prognostic factors for the feasibility of chemotherapy and the Geriatric Prognostic Index (GPI) as risk profile for mortality before chemotherapy in the elderly. Acta Oncol 55:15–23, 2016

18. van Leeuwen RWF, Swart EL, Boven E, et al: Potential drug interactions in cancer therapy: a prevalence study using an advanced screening method. Ann Oncol 22:2334–2341, 2011

19. Popa MA, Wallace KJ, Brunello A, et al: Potential drug interactions and chemotoxicity in older patients with cancer receiving chemotherapy. J Geriatr Oncol 5:307–314, 2014

20. Wang N, Qin MZ, Cui J: [Lipid profile comparison between pre- and post-menopausal women]. Zhonghua Xin Xue Guan Bing Za Zhi 44:799–804, 2016

21. Weidner N, Moore DH, Vartanian R: Correlation of Ki-67 antigen expression with mitotic figure index and tumor grade in breast carcinomas using the novel “paraffin”-reactive MIB1 antibody. Hum Pathol 25:337–342, 1994

22. Cortadellas T, Argacha P, Acosta J, et al: Estimation of tumor size in breast cancer comparing clinical examination, mammography, ultrasound and MRI—correlation with the pathological analysis of the surgical specimen. Gland Surg 6:330–335, 2017

23. Berg WA, Gutierrez L, NessAiver MS, et al: Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. Radiology 233:830–849, 2004

24. Meyers MO, Klauber-Demore N, Ollila DW, et al: Impact of breast cancer molecular subtypes on locoregional recurrence in patients treated with neoadjuvant chemotherapy for locally advanced breast cancer. Ann Surg Oncol 18:2851–2857, 2011

25. Lowery AJ, Kell MR, Glynn RW, et al: Locoregional recurrence after breast cancer surgery: a systematic review by receptor phenotype. Breast Cancer Res Treat 133:831–841, 2012

26. Waljee JF, Hu ES, Newman LA, et al: Predictors of re-excision among women undergoing breast-conserving surgery for cancer. Ann Surg Oncol 15:1297–1303, 2008

27. Truin W, Vugts G, Roumen RMH, et al: Differences in Response and Surgical Management with Neoadjuvant Chemotherapy in Invasive Lobular Versus Ductal Breast Cancer. Ann Surg Oncol 23:51–57, 2016

28. Munhoz AM, Montag E, Gemperli R: Oncoplastic breast surgery: indications, techniques and perspectives. Gland Surg 2:143–157, 2013

29. Buchholz TA: Radiation Therapy for Early-Stage Breast Cancer after Breast-Conserving Surgery. New England Journal of Medicine 360:63–70, 2009

30. Carlson RW, Allred DC, Anderson BO, et al: Invasive breast cancer. J Natl Compr Canc Netw 9:136–222, 2011

31. Eifel P, Axelson JA, Costa J, et al: National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. J Natl Cancer Inst 93:979–989, 2001

32. Halberg FE, Shank BM, Haffty BG, et al: Conservative surgery and radiation in the treatment of stage I and II carcinoma of the breast. American College of Radiology. ACR Appropriateness Criteria. Radiology 215 Suppl:1193–1205, 2000

33. Masuda N, Lee S-J, Ohtani S, et al: Adjuvant Capecitabine for Breast Cancer after Preoperative Chemotherapy. N Engl J Med 376:2147–2159, 2017

34. Dent R, Valentini A, Hanna W, et al: Factors associated with breast cancer mortality after local recurrence. Curr Oncol 21:e418–e425, 2014

35. Whelan T, Clark R, Roberts R, et al: Ipsilateral breast tumor recurrence postlumpectomy is predictive of subsequent mortality: results from a randomized trial. Investigators of the Ontario Clinical Oncology Group. Int J Radiat Oncol Biol Phys 30:11–16, 1994

36. Kurtz JM, Spitalier JM, Amalric R, et al: The prognostic significance of late local recurrence after breast-conserving therapy. Int J Radiat Oncol Biol Phys 18:87–93, 1990

37. Sopik V, Nofech-Mozes S, Sun P, et al: The relationship between local recurrence and death in early-stage breast cancer. Breast Cancer Res Treat 155:175–185, 2016

38. Witteveen A, Kwast ABG, Sonke GS, et al: Survival after Locoregional Recurrence or Second Primary Breast Cancer: Impact of the Disease-Free Interval. PLOS ONE 10:e0120832, 2015

39. Hudis CA, Barlow WE, Costantino JP, et al: Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. J Clin Oncol 25:2127–2132, 2007

40. Voduc KD, Cheang MCU, Tyldesley S, et al: Breast Cancer Subtypes and the Risk of Local and Regional Relapse. JCO 28:1684–1691, 2010

41. Wu X, Baig A, Kasymjanova G, et al: Pattern of Local Recurrence and Distant Metastasis in Breast Cancer By Molecular Subtype [Internet]. Cureus 8, 2016[cited 2021 Feb 4] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5222631/

42. Snell CE, Gough M, Middleton K, et al: Absent progesterone receptor expression in the lymph node metastases of ER-positive, HER2-negative breast cancer is associated with relapse on tamoxifen. J Clin Pathol 70:954–960, 2017

43. Nishimura R, Osako T, Okumura Y, et al: Changes in the ER, PgR, HER2, p53 and Ki-67 biological markers between primary and recurrent breast cancer: discordance rates and prognosis. World J Surg Oncol 9:131, 2011

44. Nishimura R, Osako T, Nishiyama Y, et al: Evaluation of factors related to late recurrence--later than 10 years after the initial treatment--in primary breast cancer. Oncology 85:100–110, 2013

45. Darlix A, Griguolo G, Thezenas S, et al: Hormone receptors status: a strong determinant of the kinetics of brain metastases occurrence compared with HER2 status in breast cancer. J Neurooncol 138:369–382, 2018

46. Zhou L, Zhou W, Zhang H, et al: Progesterone suppresses triple-negative breast cancer growth and metastasis to the brain via membrane progesterone receptor α. Int J Mol Med 40:755–761, 2017

47. Ahern TP, Pedersen L, Tarp M, et al: Statin prescriptions and breast cancer recurrence risk: a Danish nationwide prospective cohort study. J Natl Cancer Inst 103:1461–1468, 2011

48. Kwan ML, Habel LA, Slattery ML, et al: NSAIDs and Breast Cancer Recurrence in a Prospective Cohort Study. Cancer Causes Control 18:613–620, 2007

49. Powe DG, Voss MJ, Zänker KS, et al: Beta-Blocker Drug Therapy Reduces Secondary Cancer Formation in Breast Cancer and Improves Cancer Specific Survival. Oncotarget 1:628–638, 2010

50. Hamy A-S, Derosa L, Valdelièvre C, et al: Comedications influence immune infiltration and pathological response to neoadjuvant chemotherapy in breast cancer. OncoImmunology 9:1677427, 2020

51. Farrugia DJ, Landmann A, Diego E, et al: Mitotic index to predict breast cancer recurrence after neoadjuvant systemic therapy. JCO 34:e23265–e23265, 2016

52. Pattali S, Harding N, Visotcky A, et al: Value of mitotic index in residual tumors following neoadjuvant therapy for breast cancer: Single institution experience. JCO 34:548–548, 2016

53. Symmans WF, Peintinger F, Hatzis C, et al: Measurement of Residual Breast Cancer Burden to Predict Survival After Neoadjuvant Chemotherapy. Journal of Clinical Oncology 25:4414–4422, 2007

54. Duggan MA, Anderson WF, Altekruse S, et al: The Surveillance, Epidemiology and End Results (SEER) Program and Pathology: Towards Strengthening the Critical Relationship. Am J Surg Pathol 40:e94–e102, 2016

55. James B. Yu MD: NCI SEER Public-Use Data: Applications and Limitations in Oncology Research [Internet]. Cancer Network , 2009[cited 2019 Aug 27] Available from: https://www.cancernetwork.com/oncology-journal/nci-seer-public-use-data-applications-and-limitations-oncology-research

56. Boffa DJ, Rosen JE, Mallin K, et al: Using the National Cancer Database for Outcomes Research: A Review. JAMA Oncol 3:1722–1728, 2017

57. Bezin J, Duong M, Lassalle R, et al: The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf 26:954–962, 2017

58. Tuppin P, Rudant J, Constantinou P, et al: Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France [Internet]. /data/revues/03987620/v65sS4/S0398762017304315/ , 2017[cited 2019 Aug 13] Available from: https://www.em-consulte.com/en/article/1140905

59. Margolis R, Derr L, Dunn M, et al: The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc 21:957–958, 2014

60. Keim D, Andrienko G, Fekete J-D, et al: Visual Analytics: Definition, Process, and Challenges [Internet], in Kerren A, Stasko JT, Fekete J-D, et al (eds): Information Visualization. Berlin, Heidelberg, Springer Berlin Heidelberg, 2008, pp 154–175[cited 2019 Aug 14] Available from: http://link.springer.com/10.1007/978-3-540-70956-5_7

61. Happe A, Drezen E: A visual approach of care pathways from the French nationwide SNDS database - from population to individual records: the ePEPS toolbox [Internet], 2018[cited 2019 Aug 18] Available from: https://hal-univ-rennes1.archives-ouvertes.fr/hal-01697626

62. Zhao Y, Parvinzamir F, Wei H, et al: Visual Analytics for Health Monitoring and Risk Management in CARRE. E-Learning and Games; 10th International Conference, Edutainment 2016, Hangzhou, China, April 14-16, 2016, Revised Selected Papers 9654:380–391, 2016

63. Ledesma A, Al-Musawi M, Nieminen H: Health figures: an open source JavaScript library for health data visualization [Internet]. BMC Med Inform Decis Mak 16, 2016[cited 2019 Aug 14] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4802654/

Figures and Tables

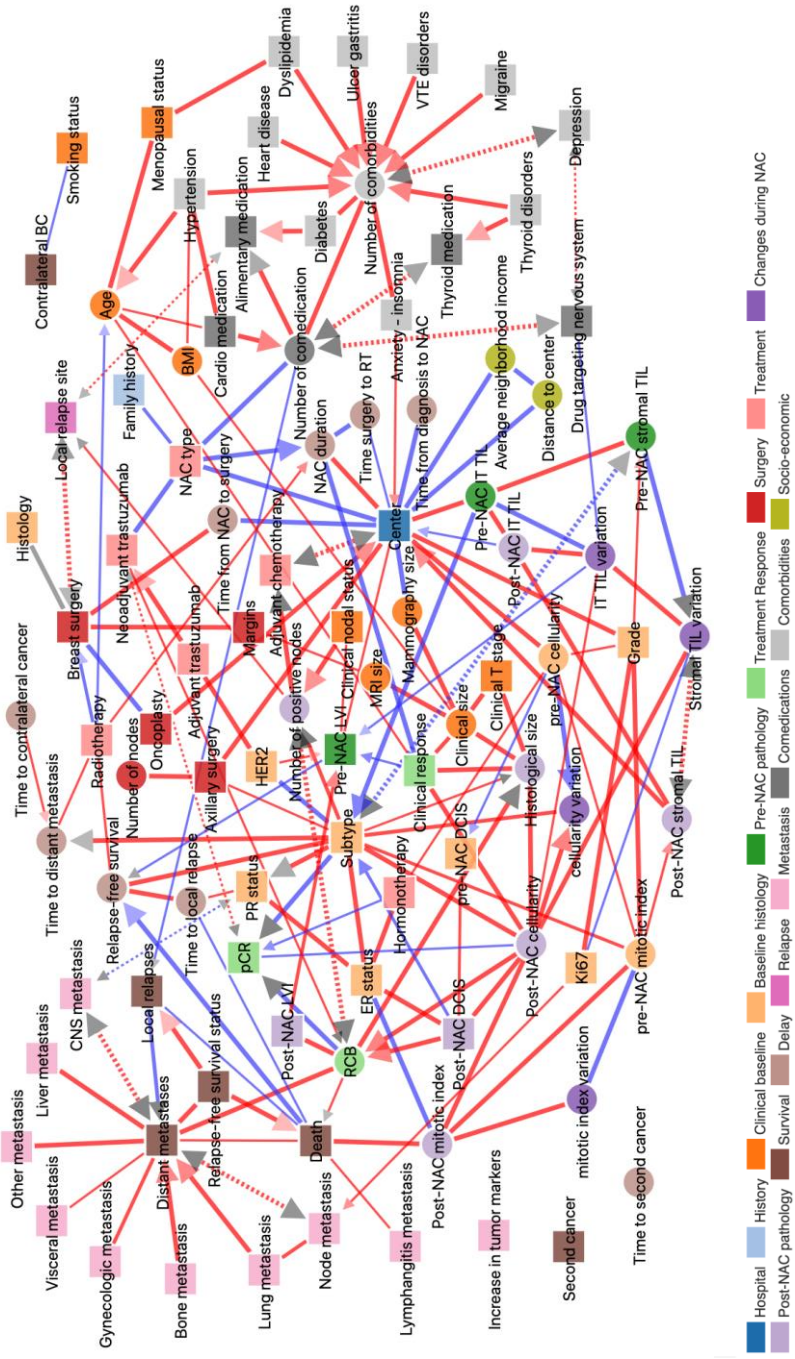

Figure 1: MIIC global network for the NEOREP breast cancer cohort. Each node corresponds to a variable of the dataset, with circles indicating continuous variables and squares indicating categorical variables. The colors define a category of variables, as detailed under the figure. Each edge corresponds to a "direct" association between two variables. *BC= breast cancer, BMI= body mass index, DCIS=ductal carcinoma* in situ, *ER= estrogen receptor status, LVI= lymphovascular invasion, NAC= neoadjuvant chemotherapy, CNS= central nervous system, pCR= pathological complete response, PR=progesterone receptor status, RCB= residual cancer burden, TILs= tumor-infiltrating lymphocytes*. Blue edges indicate negative partial correlations, red edges indicate positive partial correlations. Squares represent categorical variables, circles represent continuous variables.

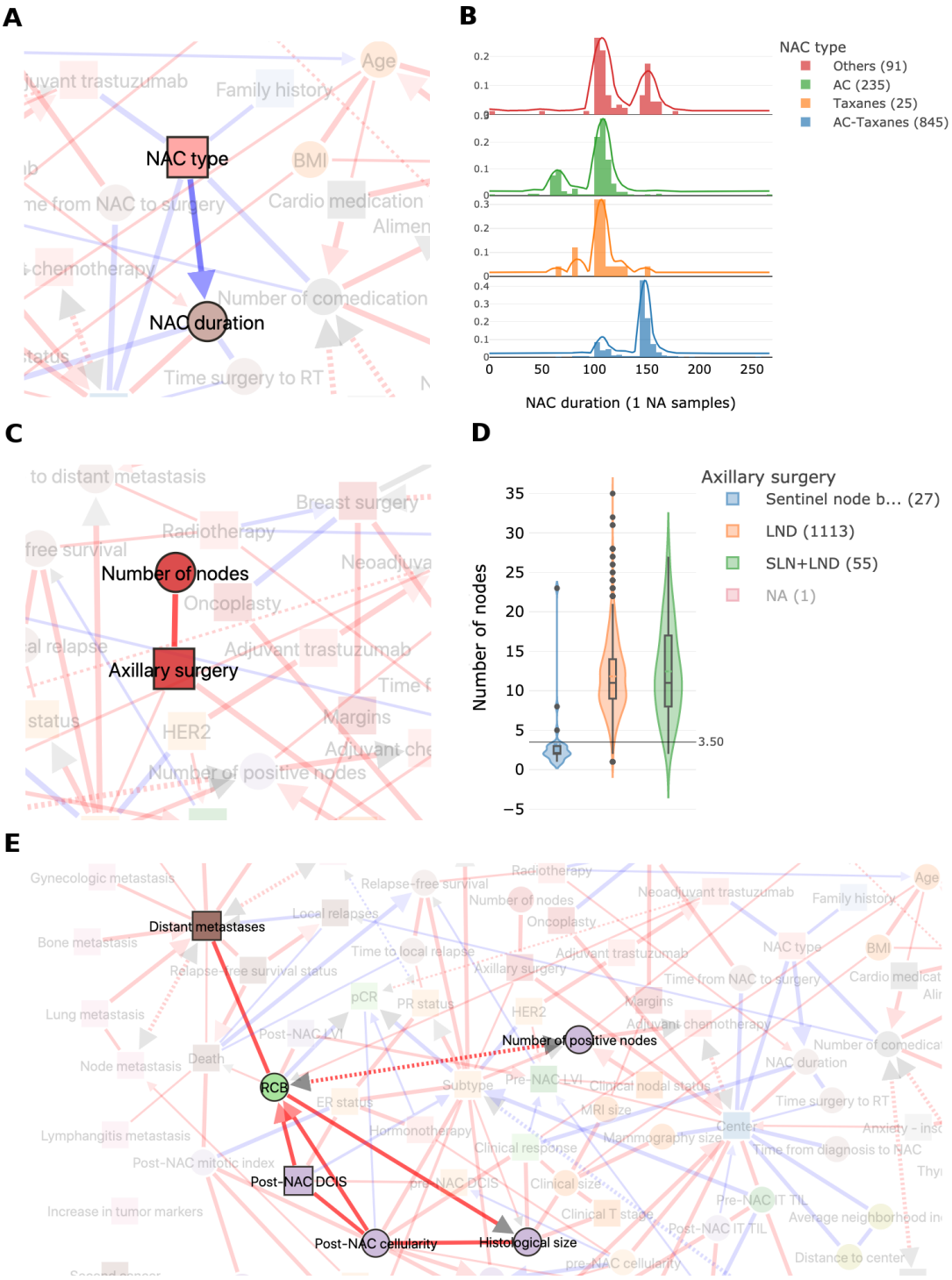

Figure 2. The MIIC interactive online interface identifies inherent associations between variables.
A) NAC type is directly correlated with NAC duration. NAC=neoadjuvant chemotherapy B) Distribution of neoadjuvant chemotherapy (NAC) duration (in days) according to the NAC regimen administered: anthracyclines (AC), taxanes or sequential AC-taxanes C) The number of axillary nodes in the histological specimen depends on the type of axillary surgery performed D) Boxplot showing the number of axillary nodes removed according to the type of surgery performed: lymph node dissection (LND), sentinel lymph node biopsy (SLN) or both E) Network interactions of the RCB node with the five patterns making up the RCB score.

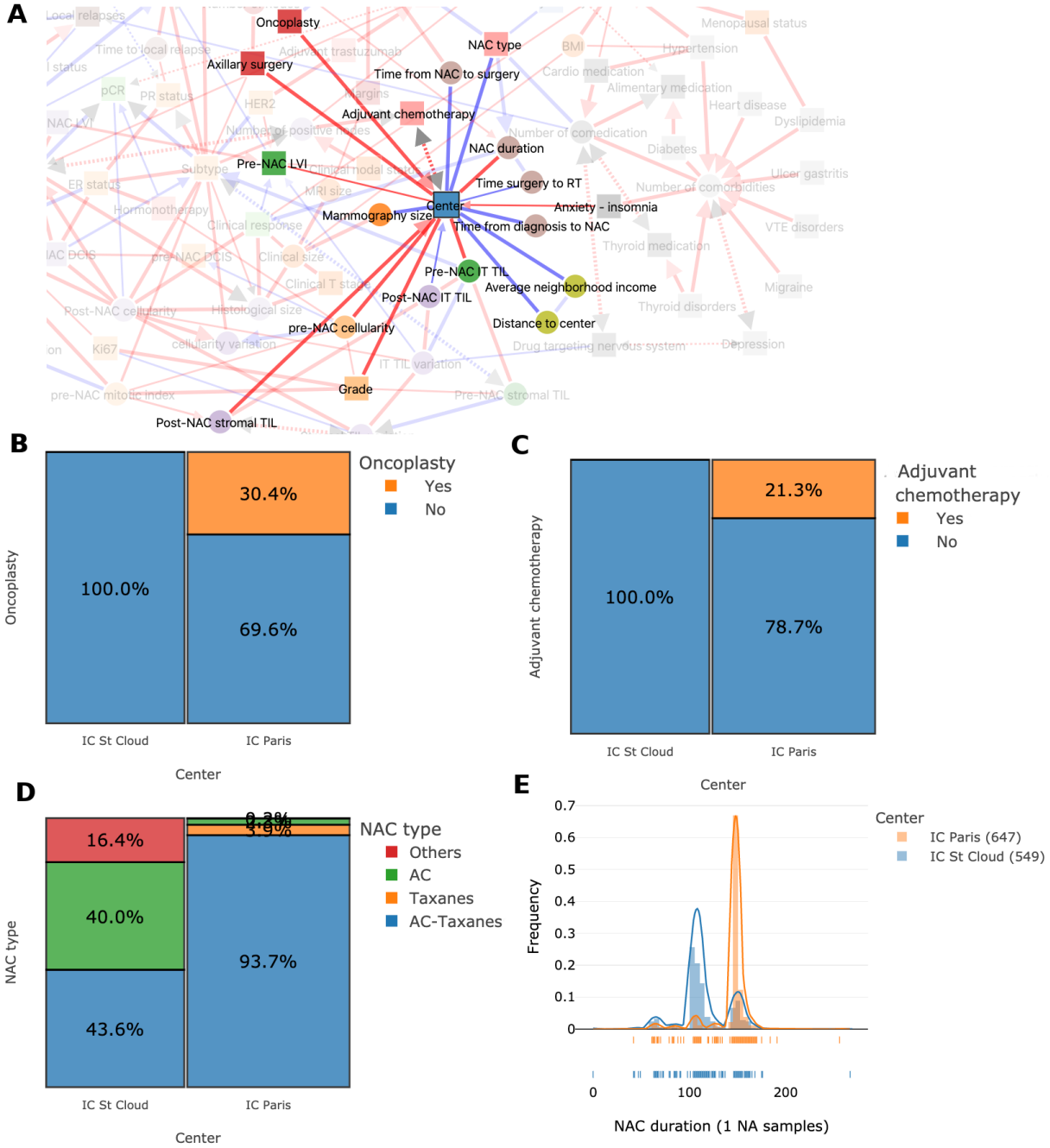

Figure 3. MIIC identifies differences in clinical practices between the two centers of the cohort
A) Network interactions around the node "center" of treatment. B) Proportion of patients undergoing oncoplastic surgery, according to treatment center: Paris or St Cloud C) Proportion of patients receiving adjuvant chemotherapy according to treatment center: Paris or St Cloud. D) Proportion of the various NAC regimens according to treatment center. E) Distribution plot for NAC duration in days, according to treatment center.

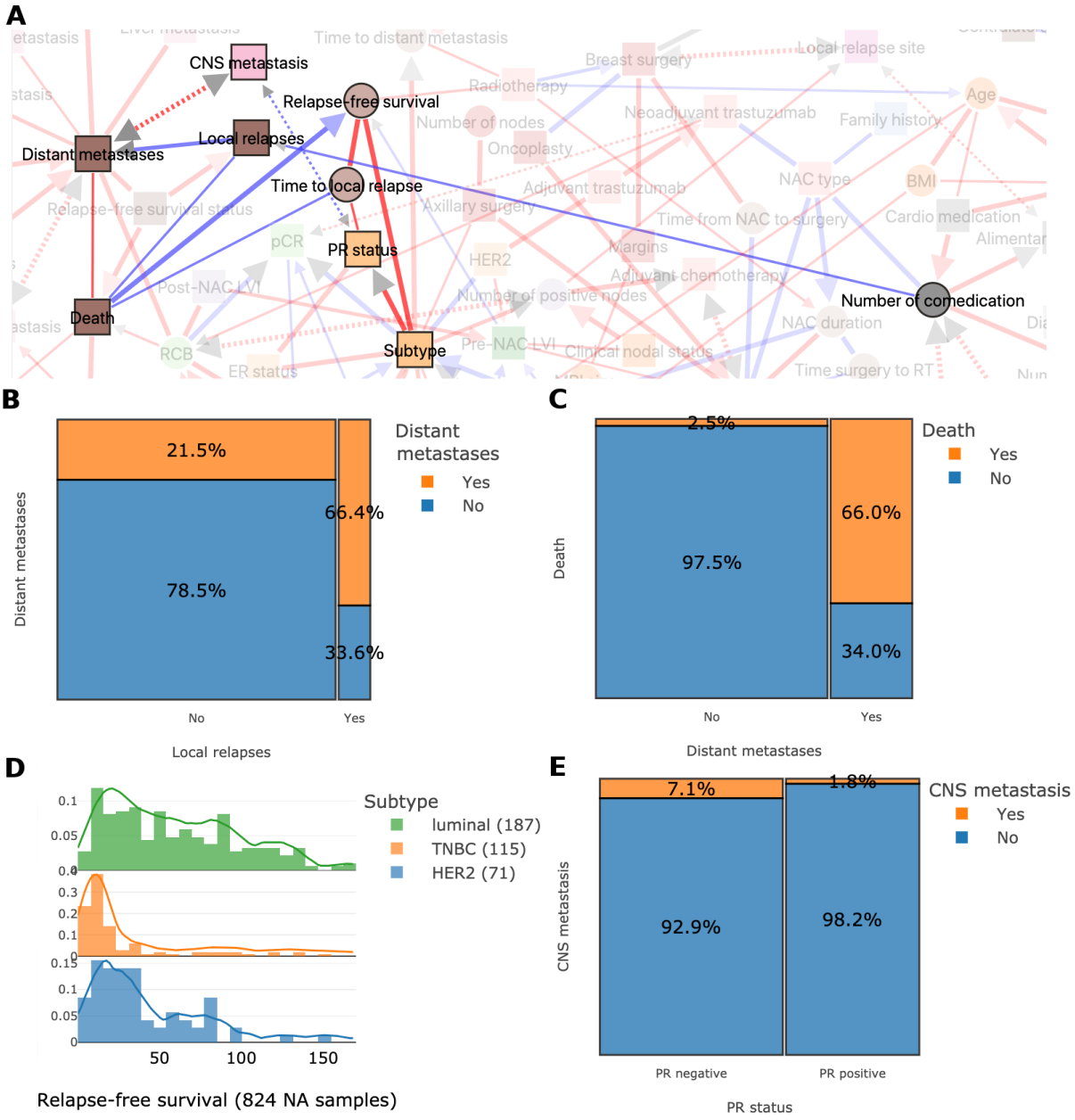

Figure 4. MIIC traces the natural course of the disease
A) Network interactions showing links between relapses, metastases and death in breast cancer. B) Proportion of distant metastases according to the occurrence or absence of local relapses. C) Proportion of deaths according to distant metastasis status. D) Distribution plot for relapse-free survival (in months) according to breast cancer subtype. E) Proportion plot displaying the relationship between central nervous system (CNS) metastasis and progesterone receptor (PR) status.

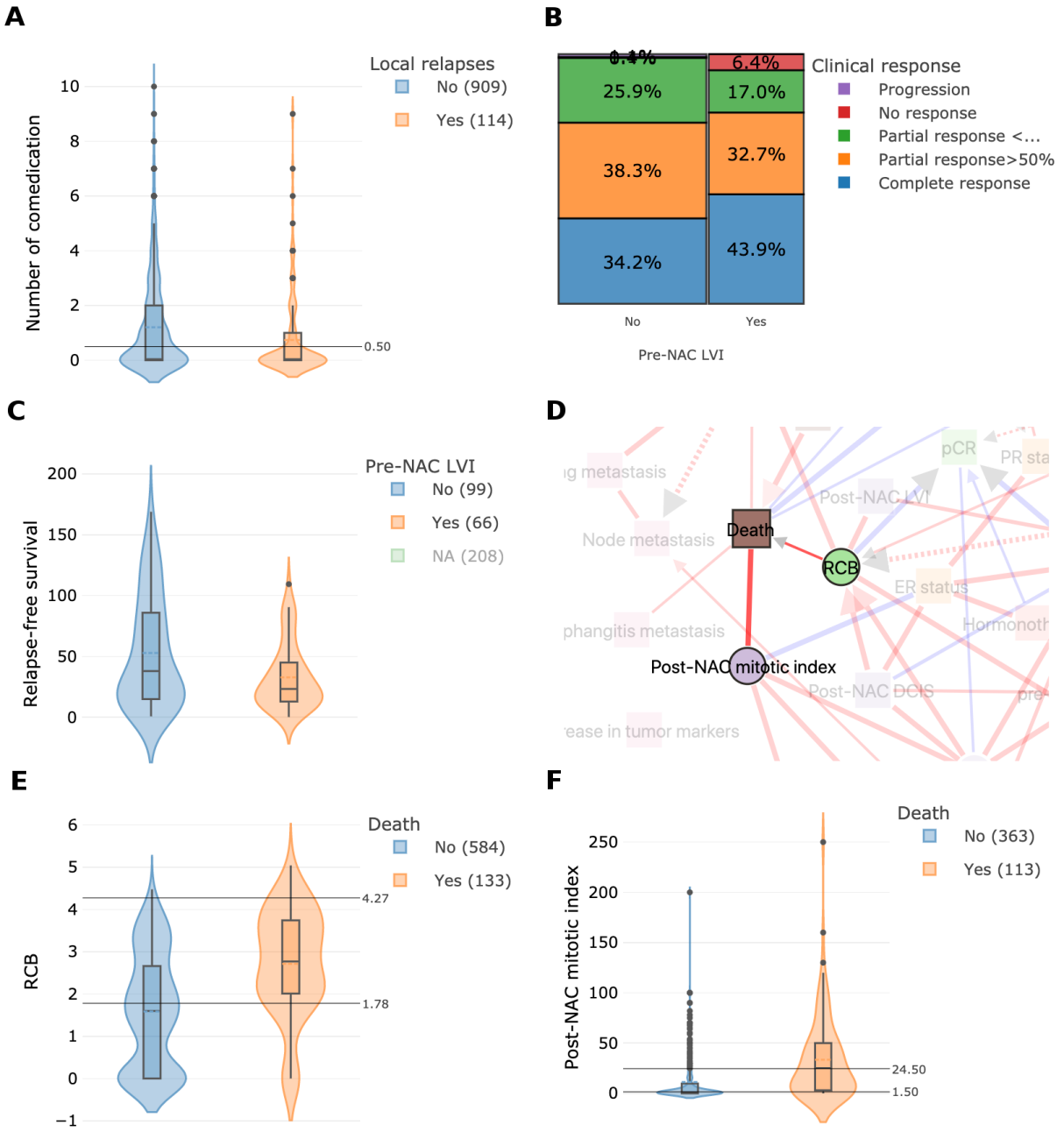

Figure 5 MIIC identifies factors likely to improve prediction or prognosis.
A) Network interaction displaying the link between local relapse occurrence and the number of drugs taken (comedication). B) Proportion plot showing the percentage of different clinical responses according to the presence or absence of pre-NAC lymphovascular invasion. C) Boxplot of relapse-free survival according to the presence or absence of pre-NAC lymphovascular invasion. D) Network interaction displaying the link between death, RCB and post-NAC mitotic index. E) Boxplot of RCB values according to vital status. F) Boxplot of post-NAC mitotic index according to vital status.

5.3 Metabolic drivers of hematopoietic differentiation

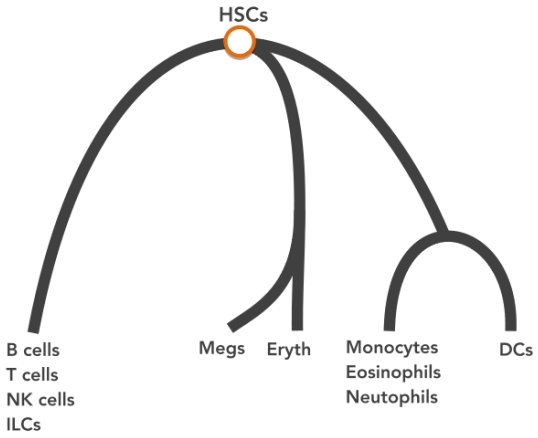

Figure 5.1: Haematopoietic stem cells differentiation.

Recent advances in -omics technologies have provided many valuable insights into complex biological systems. However, the analysis of -omics data is still a challenging frontier, with datasets characterised by high variability, sparsity and technical noise. These complex features make it difficult to discern causal relationships from spurious associations, limiting our ability to obtain novel mechanistic insights, and to optimise the design of resource-intensive downstream experiments. In this application in collaboration with the Perié team, we combine causal network reconstruction, machine learning, and experimental approaches to identify molecular drivers of fate decisions in hematopoietic stem and progenitor cells (HSPCs) which differentiate into all blood cells.

We focus on HSPC differentiation towards the erythroid, myeloid, and lymphoid lineages. Recent scRNAseq studies show that HSPCs do not form distinct subgroups but rather a transcriptomic continuum (Fig 5.1). This complex feature of the data makes it difficult to predict whether an individual progenitor will give rise to a specialised blood cell subset, and consequently, the molecular drivers of fate decisions in HSPCs are poorly understood.

Additionally, recently developed metabolomics technologies and small molecule inhibitors have permitted bulk-level analyses of hematopoietic cell-types, showing that metabolism actively regulates hematopoiesis. These studies show that metabolism influences a range of HSC behaviours, modulating not only bioenergetics, but also epigenetic state and signalling pathways [124, 125, 126]. Despite this progress, the role of metabolism in other progenitor subsets is poorly understood. To assess metabolic heterogeneity across the hematopoietic system, we first analysed published bulk and single cell transcriptomic datasets of all major hematopoietic cell types. Using a supervised learning approach we first constructed a classifier model capable of accurately predicting mature lineage identity using the expression of metabolic genes that are variably expressed within HSPCs. This result shows that a subset of

variably expressed genes in the HSPC compartment are predictive of lineage fate, but this classifier model cannot assess whether they are functionally linked or are merely associated with the differentiation.

After having established that HSPCs are metabolically heterogeneous, and differences in metabolism can influence cell fate, we wanted to know if targeting HSPC metabolism can be used to regulate myelopoiesis. To predict putative molecular drivers of differentiation we construct a causal gene regulatory network in mature cells from the haemopedia database [127], using genes that are variably expressed within the progenitor continuum and including the cell-lineage annotation as an additional variable. There are too many differentially expressed genes for them to be all included in MIIC, so we first select the 200 genes with the highest feature importance when predicting lineage from the full transcriptomic profile, using SHAP scores [128, 129]. We infer the network from the dependencies and independencies between those genes, including the categorical node "lineage" with three values : erythroid, myeloid or neutrophil. We use the resulting network for feature selection, selecting the first order neighborhood of the lineage node as these genes are inferred to have direct relationship to cell differentiation, and paying particular attention to the ancestors of lineage (Fig 5.2).

This analysis predicted genes relating to glycolysis (Pkm) and the Pentose Phosphate Pathway (G6pdx) as key drivers of the myeloid metabolic program, while mitochondrial metabolism (Atpif1, Uqcr11) membrane transporters (Slc14a1, Abcb10) were key drivers of the erythroid program. Interestingly, different genes relating to glutathione metabolism were found as high causality markers of both the myeloid (Gsr), and erythroid lineages (Gpx1, Gstm5). Both Slc14a1 and Gsr have been independently reported in the literature as early fate markers, as determined through RNA state-fate analyses of hematopoiesis [130, 131], supporting the validity of our causal inference predictions.

We propose that network reconstruction approaches are particularly suited to tackle feature selection problems in the context of -omics analysis. In particular, by modeling genes that drive biological processes rather than simply accompany them, these methods can guide downstream experimental efforts more efficiently.

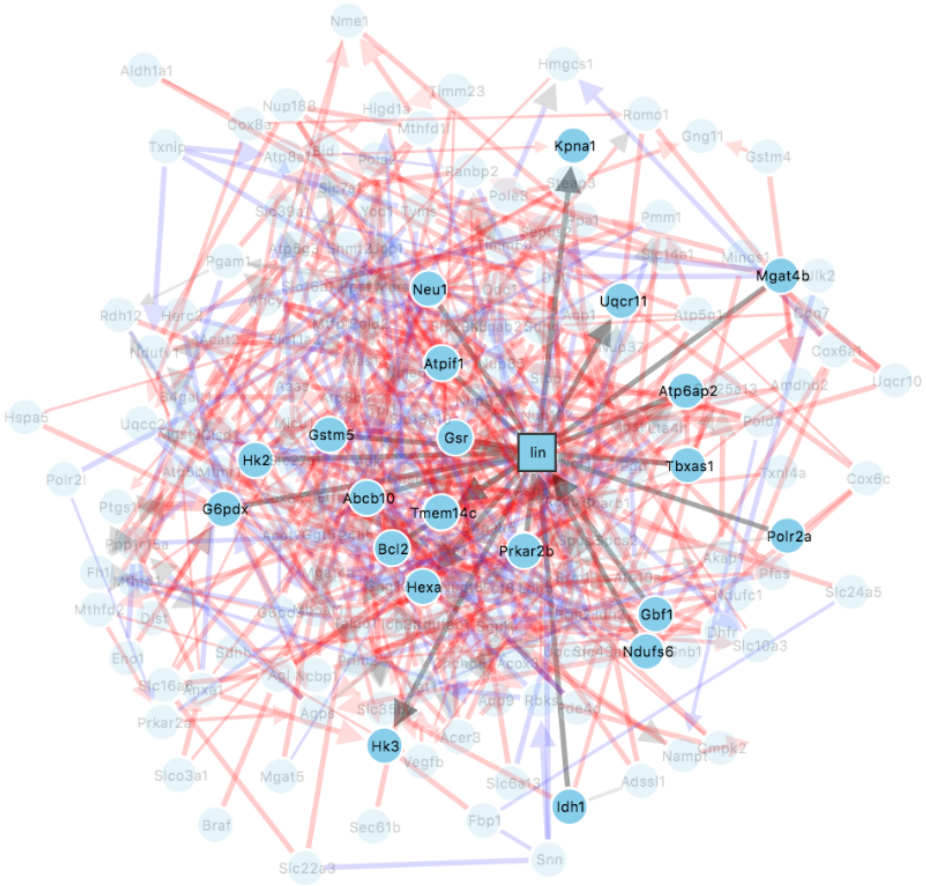

Figure 5.2: MIIC network inferred from candidate driver genes, centered on the lineage node. Highlighted nodes are directed neighbors of lineage.

Chapter 6

Conclusion

The ability to infer causality from observation is extremely powerful in the right conditions. The formalism introduced by causal inference theory is a way to tap into the vast, ever increasing amount of data and gain knowledge about our world without the need for additional costly experiments. It also opens the way to more avenues of research, bringing back causality into the domain of what is attainable when interventions are impossible.

In this thesis, we contributed to make constraint based methods, and MIIC in particular, more apt to deal with real-life datasets. We devised a general case (conditional) independence test based on the mutual information and the stochastic complexity of the data, binning continuous distributions into MDL-optimal discretizations. Using this estimator, MIIC is able to infer causal graphs from any type of data, which means that there is no restriction on the type of "causality" it is able to discover, besides the faithfulness and Markov condition. We also proposed some modifications to make the inference process more consistent with the resulting graph, ensuring that the separating sets can be *read off* the graph. This includes both modifications to the skeleton inference algorithm and a rule for test-wise omission of missing data. Other contributions include more reliable orientation of the edges, and the ability to tell apart putative from genuine causes.

Applications on real data showed the advantages of using this network-oriented approach, with uses ranging from data quality control to decision making for future experiments.

Perspectives and future research

As future research directions, I would like to adapt MIIC to temporal data, which carry intrinsically the causality signature but are harder to analyze using traditional methods. On one hand, information theory is already deeply linked with time-series data, with methods such as Granger causality [132] and the transfer entropy [133]. On the other, recent work has shown that the temporal information can be formally included in the theory of causal graphs [134, 135]. A MIIC extension to temporal series would benefit from previous work in both

domains, bridging the gap between information theory and causal graph theory for temporal data.

I would also like to investigate whether the optimal discretization found for a single edge can be used for Bayesian inference, where one tries to reconstruct $p(V)$ from $\mathcal{G}_{Inf}$. The discretization is only optimal in relation to each edge, but there may be a way to combine those results to get an MDL-optimal encoding of the variable in relation to either all of its neighbors or its parents, making it usable for Bayesian inference.

As mentioned in the introduction, feature selection and the causal structure are closely related. I would want to investigate how we can infer causal structure from the knockoffs framework, which perform feature selection while rigorously controlling the expected fraction of false positives [136, 137]. The biggest drawback of this method is the generation of knockoffs, which could perhaps be simplified using the MDL principle and the NML distribution.

Chapter A

Résumé long en français

Contexte scientifique

La corrélation n'implique pas la causalité, une distinction importante à se rappeler alors que les associations statistiques génèrent de plus en plus de discussions dans un monde toujours plus mesuré et documenté. C'est pourtant le but, avoué ou non, de la plupart des domaines scientifiques : définir les mécanismes de notre environnement qui ont produit ces observations. La nouvelle science de la causalité cherche à nous réconcilier avec ce concept en répondant à ces questions : comment formaliser les relations causales, comment nous les représenter, et quand peut-on les découvrir ? En particulier, les travaux de cette thèse contribuent aux méthodes d'inférence de causalité à partir uniquement de données d'observation. Si la corrélation seule ne suffit pas à inférer une causalité il est en effet possible d'arriver à ce genre de conclusion sans aucune intervention de la part de l'expérimentateur, en observant les bonnes données dans les bonnes conditions.

Les travaux de cette thèse s'inscrivent dans la théorie principalement développée par Judea Pearl sur les diagrammes causaux; des modèles graphiques qui permettent de dériver toutes les quantités causales d'intérêt (effet du traitement, contrefactuelles...) formellement et intuitivement [1, 2, 3]. Un diagramme causal est un réseau bayésien : un graphe dirigé et acyclique qui encode les indépendances conditionnelles entre les distributions de variables aléatoires représentées par les noeuds; avec une dimension causale retranscrite par la direction des arêtes. Ainsi, si X est un parent de Y, alors nous savons qu'une intervention sur la variable X pour lui donner une distribution arbitraire $\widetilde{p}(x)$, notée $do(X = \widetilde{p}(x))$, changera la distribution $p(Y|(do(X = \widetilde{p}(x))))$, mais intervenir sur Y ne changera pas la distribution de son parent.

Considérons une situation familière dans laquelle notre intuition peut être assez naturellement représentée par un diagramme causal (Fig A.1). Admettons qu'il y ait deux causes qui puissent être à l'origine d'une panne de voiture, que nous essayons de diagnostiquer *avant* d'intervenir sur la voiture. Les deux causes considérées, un niveau d'huile trop bas ou

une batterie vide, seraient donc représentées comme des parents du noeud "Panne". Après observation, nous disposons d'une autre information : les phares ne s'allument pas. Nous savons que les phares ne dépendent pas du niveau d'huile mais ont besoin de batterie, et sont donc reliés au noeud "Batterie" uniquement. Notez la présence du lien en pointillés entre "Phares" et "Panne", qui traduit l'idée que si les phares ne s'allument pas, la voiture ne va probablement pas démarrer. Ce lien retranscrit une *corrélation, et pas une causalité* : cette interaction indirecte existe seulement à cause de l'ancêtre commun "Batterie" mais ne nous informe pas sur une relation fonctionnelle. Il peut nous renseigner sur l'origine de la panne (Batterie ou Huile), mais réparer les phares n'aidera pas à faire démarrer la voiture.

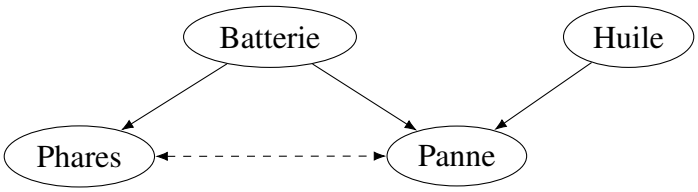

Figure A.1: Diagnostique d'une panne de voiture par diagramme causal.

Dans cet exemple le diagramme nous est déjà connu, mais comment faire quand nous n'avons aucun modèle pré-établi, par exemple est-ce qu'un nouveau traitement est efficace pour soigner une maladie ? Dans ce cas, la manière préférée pour établir un lien est de procéder à une essai cas-contrôle où deux groupes sont attribués au hasard soit un traitement soit un placebo. Si le traitement est réellement attribué au hasard, cela revient à intervenir sur sa distribution et son effet causal peut être simplement mesuré en comparant les distributions $p(\text{Rétablissement}|do(\text{Traitement} = \text{placebo})$ et $p(\text{Rétablissement}$ $|do(\text{Traitement} = \text{traitement})$. Ce protocole n'est pas toujours fiable, peut être trop long et difficile à mettre en place, voire il peut être immoral ou impossible d'intervenir sur certaines variables. Par ailleurs, nous avons à disposition toujours plus de données d'observation. Nous traitons ici le problème d'inférence de diagrammes causaux à partir de ces données collectées passivement en analysant les dépendances et indépendances statistiques. Le but de cette méthode est double : retenir seulement les interactions directes qui reflètent une relation fonctionnelle (Batterie $-$ Panne) en rejetant les corrélations indirectes (Phares $-$ Panne); et pouvoir inférer la direction de la causalité.

Soit $\mathcal{D}$ une collection de v variables $X_1, \cdots, X_v$ avec une distribution jointe $P(v)$ et N échantillons indépendants.

Definition A.1. Le vrai graphe causal $\mathcal{G}_c$ qui correspond aux données $\mathcal{D}$ satisfait :

1. $\mathcal{G}_c$ est un réseau dirigé et acyclique.
2. La distribution de chaque noeud X_i peut être exprimé en fonction de celle de ses parents

pa_i, plus un terme de bruit U_i :

$$P(X_i) = f(P(pa_i), U_i)$$

et $\mathcal{G}_c$ est *compatible* [1] avec le set $\mathbf{P}_*$ de toutes des distributions interventionnelles $P(v|do(X = x)), X \subseteq V$.

3. Les indépendances conditionnelles observées dans $\mathcal{D}$ correspondent à des *d-séparation* dans $\mathcal{G}_c$.
4. Les *d-séparation* observées dans $\mathcal{G}_c$ correspondent à des indépendances conditionnelles dans $\mathcal{D}$.

Definition A.2. Deux noeuds X et Y d'un graphe acyclique dirigé $\mathcal{G}$ sont d-séparés par un ensemble de noeuds Z si et seulement si :

- Le chemin entre X et Y contient une chaîne $i \to m \to j$ ou $i \leftarrow m \to j$ avec $m \in Z$, ou
- Le chemin entre X et Y contient une *V-structure* $i \to m \leftarrow j$ telle que $m \notin Z$ et aucun descendant de m n'est dans Z.

Le set Z d-sépare X et Y si et seulement si Z bloque tous les chemins de X vers Y de cette façon.

Sans la possibilité d'intervenir sur $P(v)$, seuls les points 1, 3 et 4 de la définition A.1 sont exploitables. En se reposant uniquement sur les observations, on peut identifier jusqu'à la *classe d'équivalence* de $\mathcal{G}_c$: tous les graphes qui partagent la même structure non orientée ainsi que ses *V-structures* [1, 18]. Une V-structure est un sous-graphe composée de trois noeuds $X \to Z \leftarrow Y$ avec X et Y non-adjacent. Cette structure locale porte le signe de causalité, c'est le seul graphe causal à trois noeuds et deux liens qui encode une indépendance $X \perp\!\!\!\perp Y$ et une dépendance conditionnelle $X \not\perp\!\!\!\perp Y|Z$.

Definition A.3. La classe d'équivalence du graphe $\mathcal{G}_c$ est l'ensemble des graphes qui partagent le même squelette ($\mathcal{G}_c$ non dirigé) et les même V-structures, $X \to Z \leftarrow Y$ avec X et Y non-adjacent.

Theorem A.1. *Si toutes les variables de $\mathcal{D}$ sont observées, alors la classe d'équivalence de $\mathcal{G}_c$ est identifiable à partir des indépendances conditionnelles dans $\mathcal{D}$.*

Parmi les méthodes pour inférer un graphe $\mathcal{G}_{inf}$ à partir de $\mathcal{D}$, les plus plus utilisées sont appelées méthodes basées sur les contraintes. En faisant l'hypothèse que *toutes* les indépendances conditionnelles correspondent à d-séparations (et inversement), les "contraintes" font référence aux équivalences entre les indépendances dans $\mathcal{D}$ et la structure de $\mathcal{G}_{inf}$. Par exemple, si deux variables sont toujours dépendantes peu importe le set de conditionnement $X \not\perp\!\!\!\perp Y \mid Z,\ Z \subseteq V \setminus \{X, Y\}$ alors elles doivent être adjacentes dans $\mathcal{G}_{inf}$. A l'inverse, on

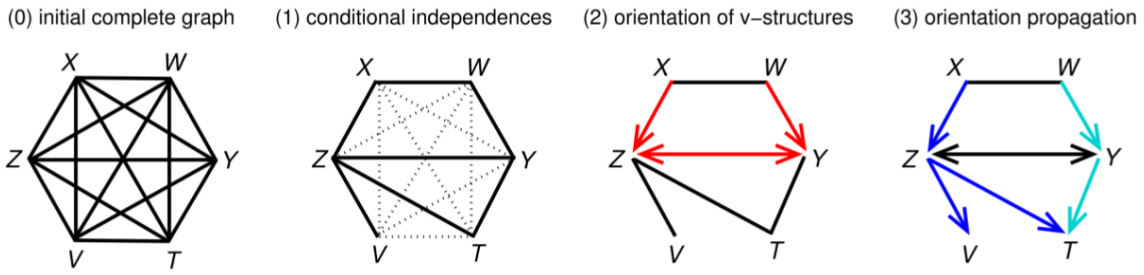

Figure A.2: Inférence de graphe causal par méthode basée sur les contraintes

s'attend à ce qu'une variable soit indépendante de tous ses ancêtres après avoir conditionné sur ses parents directs dans $\mathcal{G}_c$; ou encore que deux variables qui n'ont pas de lien direct soient indépendantes en conditionnant sur leurs ancêtres communs. La première étape des méthodes par contraintes consiste à déterminer quelles variables sont adjacentes dans $\mathcal{G}_{inf}$ en partant du graphe complet et en testant une à une les indépendances conditionnelle de $\mathcal{D}$ (Fig A.2). Le lien $X - Y$ est retiré si on peut conclure $X \perp\!\!\!\perp Y \mid Z$, avec Z un sous-ensemble des voisins de X et Y. Une fois le squelette obtenu, il est possible d'orienter les liens inférés jusqu'à la classe d'équivalence du graphe $\mathcal{G}_c$ en orientant les V-structures et en appliquant les règles de propagation [18].

L'algorithme MIIC (multivariate information-based inductive causation) partage le fonctionnement général des algorithmes basés sur les contraintes, avec des modifications qui le rendent plus robustes au bruit et plus efficace [35, 44] :

- L'indépendance (conditionnelle) est inférée à partir de l'estimation de l'information mutuelle (conditionnelle) corrigée avec la complexité stochastique, qui présente certains avantages par rapport aux test fréquentistes habituels en particulier pour une taille d'échantillon N réduite [5, 138].
- La recherche des indépendances conditionnelles se fait itérativement en enlevant les meilleurs contributeurs grâce à la règle de la chaîne de l'information mutuelle conditionelle :

$$\mathrm{I}(X;Y|\{U_i\},Z) \;=\; \mathrm{I}(X;Y) - \mathrm{I}(X;Y;u_1) - \mathrm{I}(X;Y;u_2|u_1) - \cdots - \mathrm{I}(X;Y;z|\{U_i\})$$

  Dans la méthode de référence, tous les sets de séparation $\{U_i\}$ sont essayés jusqu'à ce qu'une indépendance soit trouvée $X \perp\!\!\!\perp Y \mid \{U_i\}$, et les liens $X - Y$ sont testés dans un ordre arbitraire. En commençant par les meilleurs contributeurs, MIIC est moins sensible aux fausses indépendances dues au bruit d'échantillonage.
- L'orientation des V-structures et leur propagation se basent sur des probabilités calculées avec les informations mutuelles multivariées et ressemble davantage à de l'inférence Bayésienne, donnant en général une meilleure orientation du squelette

[44].

Les références [44, 139] montrent des cas d'utilisation à échelles variées : de trajectoires de différentiation à partir de données single-cell à l'étude de la valeur adaptative de différentes caractéristiques de gènes dans un jeu de données phylogénétique. MIIC est cependant limité aux valeurs discrètes pour lesquelles l'information mutuelle et la complexité stochastique peuvent être facilement estimées.

Contributions

L'objectif de cette thèse est d'améliorer MIIC pour le rendre plus apte à traiter des données issues du monde réel. Premièrement, s'affranchir le plus possible des conditions imposées sur la distribution $P(v)$ permettrait d'inclure toutes les données à disposition pour reconstruire $\mathcal{G}_{inf}$. Concrètement, nous voulons pouvoir estimer l'indépendance conditionnelle entre deux variables X, Y avec un set de conditionnement Z peu importe la nature des distributions marginales ($p(X), p(Y), p(Z)$) et des distributions jointes ($p(X,Y)$ etc...). Cette estimation doit aussi être robuste à des petite taille d'échantillon tout en restant calculable quand N est grand, et idéalement ne favorise aucun type de variable ou d'interaction.

L'information mutuelle est une quantité idéale pour accomplir ces objectifs : elle mesure la dépendance entre deux variables aléatoires au sens le plus général. Elle est définie pour tout type de variables et tout type de relation : notée $\mathrm{I}(X;Y)$, elle donne simplement la quantité d'information que l'on a sur X en connaissant Y, et vice-versa. Introduite par Claude Shannon en 1948 pour caractériser les canaux de communication [45], elle a trouvé son succès dans de nombreux domaines grâce à une unique combinaison de propriétés désirables. Premièrement, elle est strictement équivalente à l'indépendance statistique : $\mathrm{I}(X;Y) \leftrightarrow X \perp\!\!\!\perp Y$, peu importe les distributions $p(X)$, $p(Y)$ et $p(X,Y)$. Elle est aussi décomposable grâce à la *chain rule*, statisfait le principe de *data processing inequality*, et est invariable aux transformations sur X et Y qui conservent les rangs [4]. Elle est aussi considérée comme équitable : elle détecte avec la même puissance tout type d'interaction du moment qu'elles ont le même rapport signal sur bruit [27].

La mesure de dépendance idéale en théorie, son estimation sur des échantillons finis est notoirement difficile si X, Y a des composantes continues. Les approches par estimation locale de l'entropie en regardant les k plus proches voisins donnent de bons résultats empiriques [77, 140], mais leur significativité est difficile à évaluer quand $X \perp\!\!\!\perp Y$ ou que le signal est très faible [77, 85, 51], ce qui complique leur usage pour les méthodes par contraintes. Une autre façon d'estimer l'information mutuelle sur des variables continues est de les discrétiser dans des partitions, à la manière d'un histogramme. Cependant, le résultat est alors dépendant de la discrétisation de chaque variable X_v : l'estimation à partir des versions discrétisées

$\mathrm{I}([X_1]_\Delta;[X_2]_\Delta)$ ne converge pas vers la vraie valeur $\mathrm{I}(X_1;X_2)$ mais vers une valeur qui dépend du nombre et de la taille des partitions [69, 49].

La méthode présentée ici repose sur la définition *maître* de l'information mutuelle :

$$\mathrm{I}(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} \mathrm{I}([X]_\mathcal{P};[Y]_\mathcal{Q}) \tag{A.0.1}$$

où le supremum est sur toutes les partitions finies $\mathcal{P}$ et $\mathcal{Q}$ [4]. Cette définition est correcte sur les populations X et Y, mais quand la taille de l'échantillon N est finie, augmenter le nombre de partitions dans la discrétisation $[X]_\Delta$ ou $[Y]_\Delta$ finit inévitablement par surestimer l'information mutuelle $\mathrm{I}([X]_\Delta;[Y]_\Delta)$ (jusqu'au maximum $\log(N)$ avec une partition pour chaque valeur unique observée).

L'approche développée consiste à maximiser la valeur $\mathrm{I}'([X]_\Delta;[Y]_\Delta)$ corrigée par la complexité stochastique associée à la discrétisation $[X]_\Delta;[Y]_\Delta$ pour prendre en compte les effets du nombre fini d'échantillons (Fig A.3) :

$$\mathrm{I}'([X]_\Delta;[Y]_\Delta) = \mathrm{I}([X]_\Delta;[Y]_\Delta) - k_{X_\Delta;Y_\Delta}(N)\frac{1}{N} \tag{A.0.2}$$

où $k_{X_\Delta;Y_\Delta}(N)$ est le terme de complexité, par exemple $k^{BIC}_{X_\Delta;Y_\Delta}(N) = \frac{1}{2}(\Delta_X - 1)(\Delta_Y - 1)\log(N)$ pour le *Bayesian Information Criterion*. Introduire la complexité permet aussi de conclure sur l'indépendance sur des échantillons finis (pour lesquels l'estimation de l'information est toujours positive) : $\mathrm{I}'([X]_\Delta;[Y]_\Delta) \leq 0$ implique l'indépendance entre X et Y au sens de la complexité des données [5].

La maximisation de l'information mutuelle est calculée par programmation dynamique et est inspirée de Kontkanen et al. [99]. Dans cette étude, les auteurs proposent un algorithme pour trouver la discrétisation optimale d'un échantillon de variable aléatoire en maximisant un score de vraisemblance normalisé dérivé selon le principe de longueur de description minimale. La méthode est adaptée à deux dimensions pour trouver la discrétisation d'une variable X qui maximise l'information corrigée $\mathrm{I}'([X]_\Delta;Y)$ avec une variable discrète Y. Le résultat est un algorithme qui permet simultanément d'estimer la valeur de l'information mutuelle et d'évaluer sa significativité au sens de la complexité stochastique, peu importe la nature des variables étudiées qui peuvent être continues, discrètes ou une mixture des deux. Les partitions trouvées de cette manière satisfont le principe de description minimale, et encodent les données non pas pour décrire les distributions marginales comme dans [99] mais la distribution jointe (Fig A.4). En pratique, il faut donc discrétiser chaque distribution jointe pour conclure sur la dépendance (conditionnelle) entre deux variables : on ne peut pas discrétiser chaque variable une à une et espérer une estimation non biaisée de leurs informations mutuelles. Nous améliorons la complexité de l'approche originale de Kontkanen et al. en limitant le nombre de *cutpoints* possibles à une valeur $c << N$ (typiquement, un

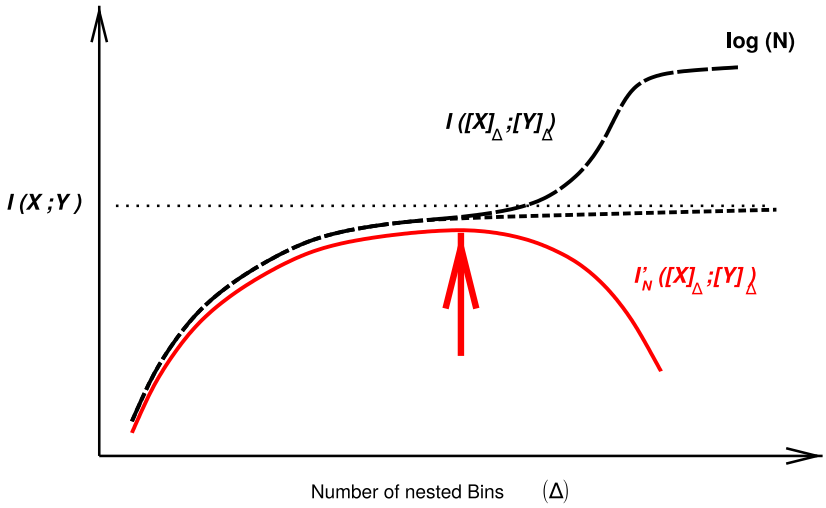

Figure A.3: Estimation de l'information mutuelle par discrétisation. La vraie valeur $\mathrm{I}(X;Y)$ (pointillés horizontaux) est théoriquement obtenue en raffinant la discrétisation Δ (courbe pointillée) mais est inévitablement surestimée quand la taille de l'échantillon N est finie (courbe en tirets) jusqu'à un maximum $\log(N)$. L'information corrigée (courbe rouge) approche la vraie valeur à son maximum.

facteur de $N^{1/3}$), de $O(N^2 \times k)$ à $O(c^2)$. Cette approche est également utilisée pour tester les indépendances conditionnelles grâce à la *chain rule* : $\mathrm{I}(X;Y|Z) = \mathrm{I}(X;YZ) - \mathrm{I}(X;Z) = \mathrm{I}(Y;XZ) - \mathrm{I}(Y;Z)$, et les informations mutuelles multivariées.

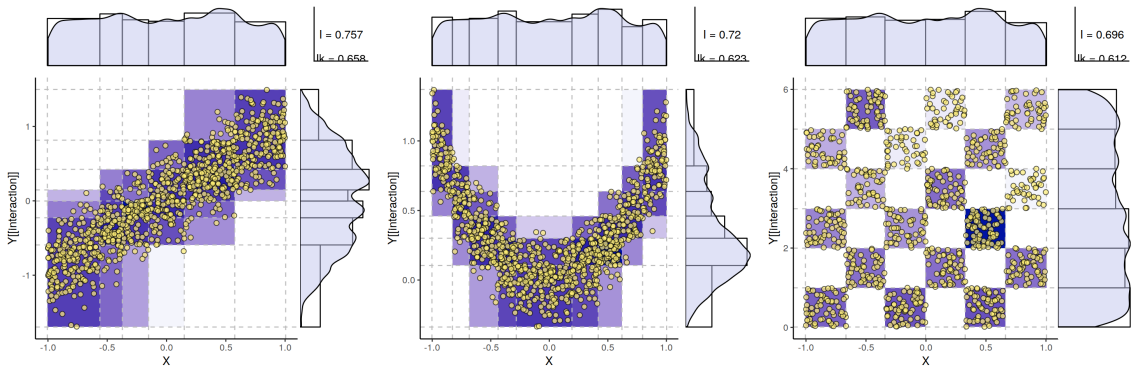

Figure A.4: Discrétisation optimale de trois distributions jointes X, Yi avec la même variable X et trois Y_i différents. La même distribution marginale $p(X)$ a des partitions optimales différentes selon la distribution jointe.

La discrétisation optimale est évaluée d'abord comme estimateur de l'information mutuelle (conditionelle) sur variables continues, et se compare favorablement à l'état de l'art en particulier quand le signal se rapproche de l'indépendance $X \perp\!\!\!\perp Y$ et $X \perp\!\!\!\perp Y \mid Z$ [7]. Elle a également de bon résultats sur les variables mixtes, en accord avec la définition maître de l'information mutuelle, comparée aux autres approches conçues spécialement pour ce cas [71, 51, 91]. Nous notons plusieurs avantages par rapport aux autres méthodes : le résultat ne dépend pas du choix d'un paramètre (par exemple le nombre de plus proches voisins k, le type

de kernel etc...), et la significativité est évaluée de manière strictement identique pour tous les cas de figures : continu-continu, discret-discret ou mixte. L'intégration de cet estimateur à MIIC permet de reconstruire le squelette du graphe en estimant les informations mutuelles conditionnelles, puis d'orienter les liens grâce aux informations mutuelles multivariées; pour *tout* type de variable. Nous comparons aussi ses performances pour reconstruire des graphes causaux à partir de données simulées, et trouvons des résultats similaires ou supérieurs aux méthodes existantes [109, 110] (Fig A.5). En particulier, notre approche est la seule qui semble être non biaisée envers certains types de variable ou d'interaction.

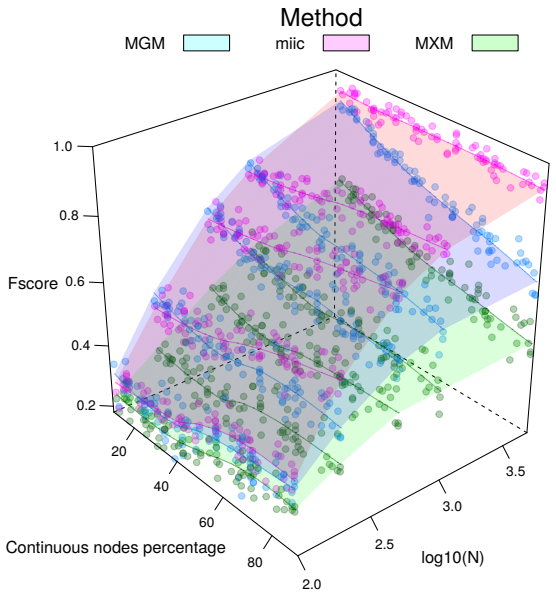

Figure A.5: Benchmarks de reconstruction de réseaux causaux. Résultat sur 50 simulation de 100 noeuds et un pourcentage de noeuds continus entre 10 et 90%. Le F score est calculé comme $F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Les autres contributions de cette thèse portent sur le fonctionnement méthodes basées sur les contraintes. Nous traitons la gestion des données manquantes, qui peuvent fausser le résultat des test d'indépendance conditionnelle par biais de sélection. Dans certains cas de figure, une interaction entre X et Y peut disparaître en conditionnant sur Z non pas parce que Z permet d'expliquer l'interaction indirecte comme cause commune par exemple, mais parce que X et Y deviennent indépendants, même sans conditionner, après avoir filtré sur les échantillons définis (sans valeur manquante) sur X, Y, Z. Par exemple, les jeux de données bio-médicales peuvent contenir des valeurs qui dépendent d'autres pour être définies : la taille d'une tumeur post-exérèse ne peut être mesurée que si le ou la patient(e) a subi une chirurgie. Dans ce cas, inclure la taille de la tumeur dans un set de conditionnement filtre automatiquement les échantillons sur les patients ayant eu une chirurgie ce qui peut créer des biais de sélection. Pour détecter cette situation, nous calculons la divergence de Kullback-

Leibler sur la distribution jointe X, Y entre le support avec des données complètes sur X, Y et celui avec les données complètes sur X, Y, Z :

$$D_{KL}((XY \mid Z_{notNA}) \mid\mid (XY))$$

Et nous acceptons Z seulement si cette divergence ne dépasse pas un certain seuil, dérivé de la théorie de l'information.

Dans la même idée, nous avons travaillé sur une variante des algorithmes par contraintes qui garantit que les sets de conditionnement utilisés pour retirer les liens soient plus *cohérents* avec le graphe final $\mathcal{G}_{inf}$ et les données $\mathcal{D}$ [6]. En effet, ces méthodes se basent uniquement sur les indépendances conditionnelles dans $\mathcal{D}$ mais n'offrent aucune garantie que les sets de conditionnement utilisés pour retirer les liens correspondent à des sets de *d-séparation* dans le graphe final. En fait, leur fonctionnement ne garantit même pas qu'elles soient toujours dans la même composante connexe dans $\mathcal{G}_{inf}$. Ce défaut rend non seulement le résultat peu interprétable mais cause aussi des problèmes de performance. Ces sets de conditionnement *incohérents* ont tendance à venir du bruit d'échantillonage plutôt que de réalités fonctionnelles, et les graphes reconstruits sur des données complexes sont typiquement très peu connectés. La version cohérente des algorithmes par contraintes produit un graphe $\mathcal{G}_{inf}$ moins sujet aux indépendances bruitées et duquel il est plus facile de déduire les sets de condionnement utilisés, ce qui rend la méthode plus interprétable. Cette variante est particulièrement adaptée à MIIC qui retire les contributeurs dans l'ordre en commençant par le meilleur score, par rapport aux méthodes de référence qui essayent toutes les combinaisons possibles jusqu'à trouver une significativité.

Les méthodes traditionnelles basées sur les contraintes (MIIC compris), ne font que découvrir des relations causales "putatives", en découvrant les orientations des V-structures, qui sont en fait compatibles à la fois avec une relation cause-effet réelle et avec un lien bi-directionnel provenant d'une cause commune non observée. Nous contribuons aussi aux méthodes par contraintes en montrant comment distinguer les liens de causalité "authentiques" des liens "putatifs" en excluant l'effet d'une cause commune non observée pour chaque lien de causalité authentique prédit. Nous y parvenons en évaluant les probabilités séparées de la "tête" et de la "queue" des liens dirigés pour toutes les arêtes orientées. Les arêtes causales authentiques sont alors prédites si les probabilités de la tête et de la queue sont statistiquement significatives, tandis que les arêtes causales restent "putatives" si leur probabilité de queue n'est pas statistiquement significative ou ne peut être déterminée à partir de données purement observationnelles (c'est à dire, liens non dirigés dans la classe d'équivalence de $\mathcal{G}_c$). Cela donne une meilleure interprétation des méthodes par contraintes sur des données réelles, pour lesquelles il est difficile d'assurer avec certitude que toutes les variables du système sont observées, et donc que les liens dirigés de $\mathcal{G}_{inf}$ soient "authentiques".

En plus d'une librairie open source sur R, avons aussi développé une interface graphique en ligne pour faciliter l'exploration des résultats de MIIC, notamment en proposant de voir directement les distributions jointes des relations prédites comme étant des relations directes (Fig A.6).

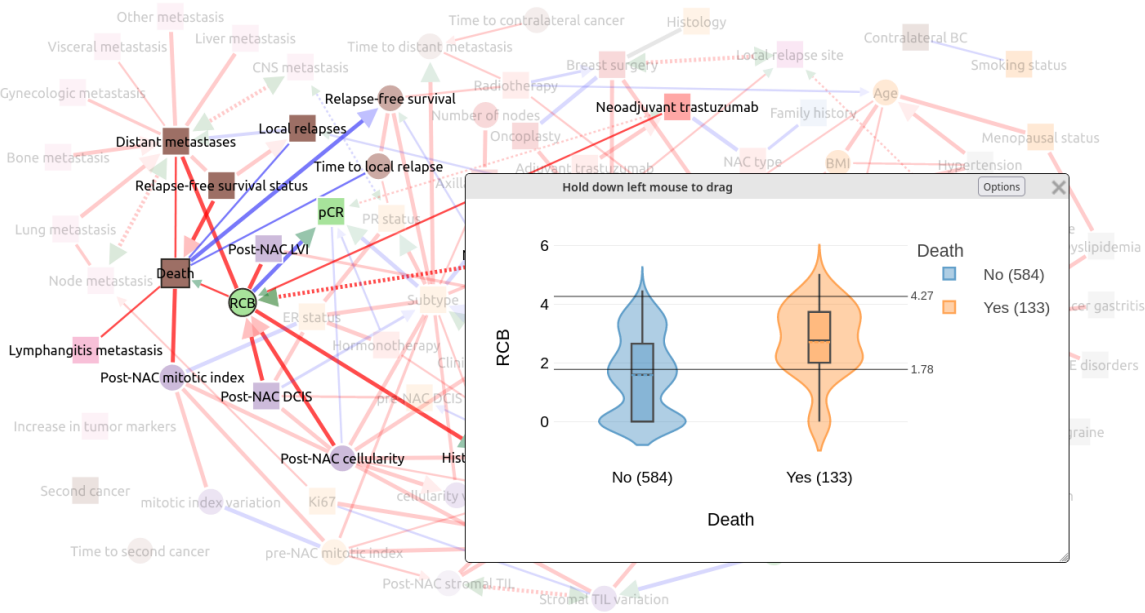

Figure A.6: Nouvelle interface de visualisation avec plot de distribution jointe entre une variable continue, "RCB", et la variable discrète "Death". Les lignes noires horizontales correspondent à la discrétisation optimale trouvée pour "RCB".

Applications

Enfin, nous montrons différentes applications de MIIC sur des données mixtes, en collaboration avec les différentes équipes responsables de la collecte des données.

Le premier réseau est reconstruit à partir de données cliniques de l'hôpital La Pitié-Salpêtrière de 1628 patients âgés atteints de troubles cognitifs. Après traitement du jeu de données, il contient 107 variables de différents types (à savoir 19 variables continues et 88 variables catégorielles) et de nature hétérogène (c'est-à-dire des variables liées aux antécédents médicaux, aux comorbidités et comédications, aux résultats des tests cognitifs, aux examens cliniques, biologiques ou radiologiques, aux diagnostics et aux traitements). Au-delà des différents types et de la nature hétérogène des données enregistrées, les noeuds du réseau clinique (Fig A.7) peuvent être divisés en groupes associés à des troubles spécifiques de la démence et au contexte clinique du patient, y compris les comorbidités (diabète, hypertension, etc.) et les médicaments associés. Le résultat est un réseau créé sans aucune connaissance préalable sur le domaine, que ce soit sur la distinction entre relations directes

et indirectes, ou la direction de la causalité. Il permet de vérifier le contenu du jeu de données, en mettant en évidence des liens inattendus ou au contraire des indépendances qui pourraient indiquer des biais dans la collecte des données. Le réseau capture également certaines facettes du raisonnement du neurologue derrière les diagnostics de différentes formes de démences. En particulier, les nœuds de diagnostic peuvent être interprétés comme des variables "explicatives" associées à un certain nombre d'effets "explaining-away" [1] sous la forme de V-structure. Nous notons aussi des liens directs inattendus entre des informations cliniquement pertinentes, telles que la connexion directe entre les échelles de Fazekas (qui mesure la quantité d'hyperintensité dans la substance blanche, attribuées à l'ischémie chronique des petits vaisseaux) et de Scheltens (atrophie du lobe temporal médian), qui peuvent fournir des informations physiologiques et suggérer de nouvelles directions de recherche [123].

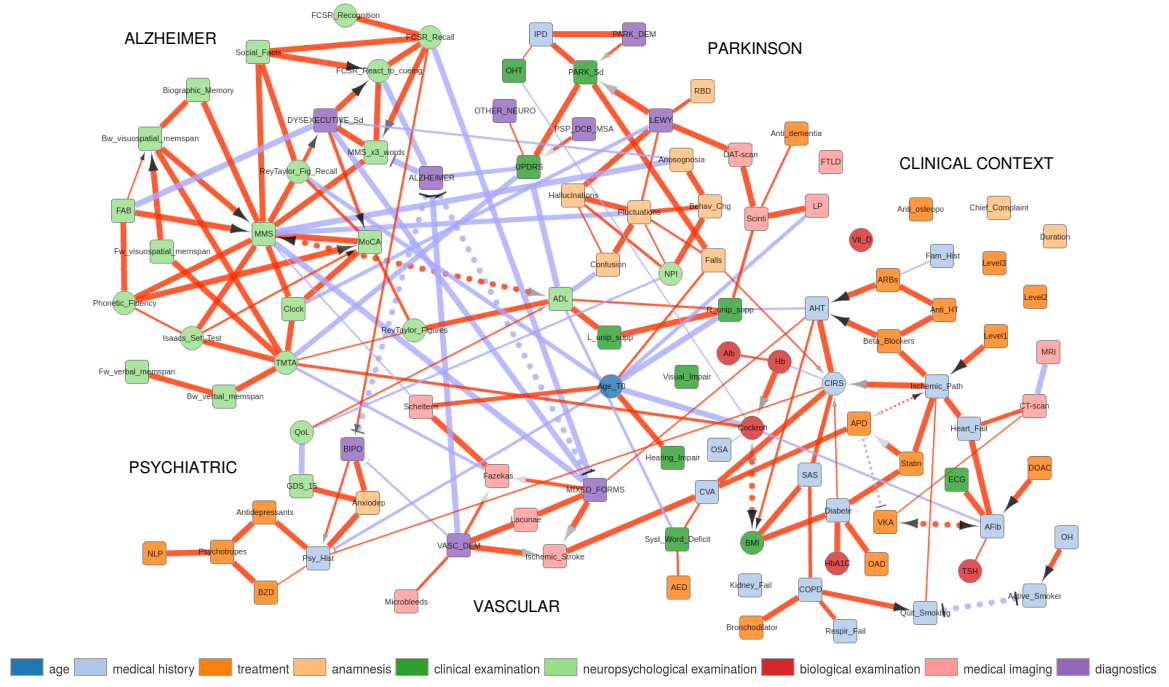

Figure A.7: Réseau reconstruit de dossiers médicaux de patients atteints de troubles cognitifs. Les nœuds carrés (respectivement cercles) correspondent à des variables discrètes (respectivement continues). Les arêtes rouges (bleues) correspondent à la corrélation (anticorrélation) entre les variables. Les liens en pointillés reflètent les variables latentes.

La deuxième application porte sur des données médicales de 1199 patientes atteintes du cancer du sein et ayant reçu une chimiothérapie néo-adjuvante à l'hôpital Curie sur les vingt dernières années (Fig A.8). L'approche systémique permet de mettre en relation toutes les variables en faisant la distinction entre relations indirectes et directes, et aide les praticiens à comprendre les mécanismes derrière la création des données, que ce soit la manière dont elles sont collectées ou la progression de la maladie elle même. Par exemple, notre approche

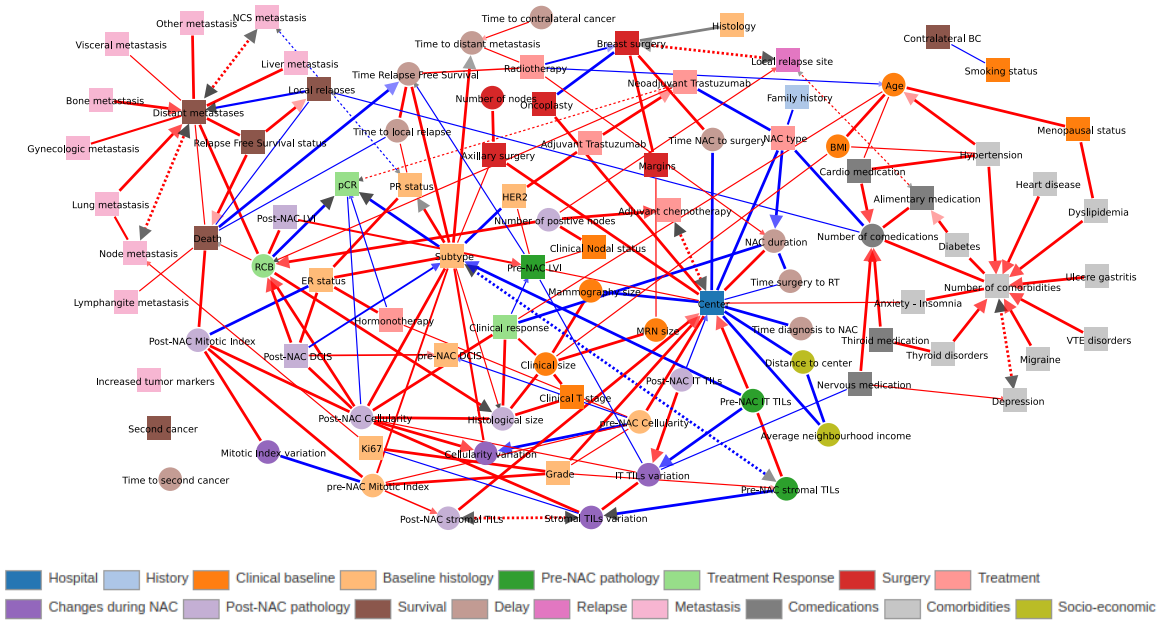

Figure A.8: Reconstruction causale sur les données cliniques de patients atteins du cancer du sein.

a mis en évidence la centralité du noeud "centre de recherche" qui correspond au lieu de prise en charge des patients, Paris ou Saint-Cloud. Il y a au moins deux explications à la présence des liens directes avec le noeud "centre" : les populations de patients traités aux deux lieux sont différentes, ce qui peut causer des biais de sélection ailleurs si ce n'est pas correctement pris en compte; et les deux équipes médicales ont des pratiques et du matériel disponible différents ce qui ne donne pas les mêmes options thérapeutiques. Un autre résultat qui mérite une mention est le voisinage du noeud "Death" qui indique l'état vital de la patiente à l'issue de sa prise en charge. Toutes les variables qui y sont directement liées n'ont pas pu être expliquée par d'autres conditionnements, donc en théorie elles donnent une information unique sur le pronostic final des patients. L'équipe de cliniciens du département de chirurgie est particulièrement intéressée par le lien "Death" - "RCB", *Residual Cancer Bruden* un indice composite généralement binairisé pour donner la variable "pCR", *Pathological Complete Response*, qui est utilisé comme facteur de risque pour prédire la mortalité [141]. Or le résultat de MIIC met en évidence un lien direct avec "RCB" et non "pCR", ce qui indique qu'une partie de l'information est perdue en passant par la dichotomisation.

Notre collaboration avec l'équipe Périé nous a permis de valider en partie les prédictions de MIIC avec des expériences in vitro. La troisième application de MIIC sur données mixtes porte sur la découverte de gènes qui influencent la différentiation de cellules précurseur hématopoïétiques, et l'inférence du réseau de régulation de ces gènes. L'équipe est particulièrement intéressée par les gènes impliqués dans le métabolisme qui sont moins étudiés que les facteurs de transcription, et a produit des jeux de données d'expression single-cell et

bulk pour étudier la différenciation de cellules précurseurs vers les branches myéloïde ou érytroïde. MIIC est utilisé pour trouver les gènes dont l'expression est directement liée avec la lignée cellulaire dans le graphe final, en complément avec d'autres méthodes de sélection de variables. Une attention particulière est aussi donnée aux gènes dont l'orientation les place en amont de la lignée dans l'ordre causal, puisque le but est de trouver les gènes non seulement prédictifs mais qui causent la différenciation. En partant du jeu de données complet, des expériences *in vitro* ont confirmé le rôle de certaines familles de gènes identifiées par MIIC comme jouant un rôle dans la différenciation : en particulier, des gènes impliqués dans la glycolyse et la voie des pentoses phosphates semblent impliqués dans la spécialisation en cellules myéloïdes, alors que l'expression des gènes du métabolisme mitochondrien dirigent vers le programme érythroïde.

Bibliography

[1] J. PEARL; *Causality* (Cambridge university press) (2009).

[2] J. PEARL, M. GLYMOUR & N. P. JEWELL; *Causal inference in statistics: a primer* (John Wiley & Sons) (2016).

[3] J. PEARL & D. MACKENZIE; *The Book of Why: The New Science of Cause and Effect* (Basic Books) (2018).

[4] T. M. COVER & J. A. THOMAS; *Elements of information theory* (John Wiley & Sons) (2012).

[5] S. AFFELDT & H. ISAMBERT; "Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information"; in "Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504," pp. 1–29 (CEUR-WS. org) (2015).

[6] H. LI, V. CABELI, N. SELLA & H. ISAMBERT; "Constraint-based Causal Structure Learning with Consistent Separating Sets"; in "Advances in Neural Information Processing Systems," pp. 14257–14266 (2019).

[7] V. CABELI, L. VERNY, N. SELLA, G. UGUZZONI, M. VERNY & H. ISAMBERT; "Learning clinical networks from medical recordsbased on information estimates in mixed-type data"; PLoS Computational Biology (2020).

[8] D. B. RUBIN; "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials"; Statistics in Medicine **26**, pp. 20–36 (2007). ISSN 1097-0258. `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2739`; _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2739.

[9] P. R. ROSENBAUM & D. B. RUBIN; "The central role of the propensity score in observational studies for causal effects"; Biometrika **70**, pp. 41–55 (1983). ISSN 0006-3444. `https://doi.org/10.1093/biomet/70.1.41`.

[10] A. L. PRICE, N. J. PATTERSON, R. M. PLENGE, M. E. WEINBLATT, N. A. SHADICK & D. REICH; "Principal components analysis corrects for stratification in genome-wide association studies"; Nature Genetics **38**, pp. 904–909 (2006). ISSN 1546-1718. `https://www.nature.com/articles/ng1847`; bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group.

[11] E. P. MARTENS, W. R. PESTMAN, A. DE BOER, S. V. BELITSER & O. H. KLUNGEL; "Instrumental variables: application and limitations"; Epidemiology pp. 260–267 (2006)Publisher: JSTOR.

[12] T. VERMA & J. PEARL; "Equivalence and synthesis of causal models"; (1991)Publisher: UCLA, Computer Science Department Los Angeles, CA.

[13] D. GEIGER & D. HECKERMAN; "Learning gaussian networks"; in "Uncertainty Proceedings 1994," pp. 235–243 (Elsevier) (1994).

[14] D. HECKERMAN, C. MEEK & G. COOPER; "A Bayesian Approach to Causal Discovery"; in "Innovations in Machine Learning," pp. 1–28 (Springer, Berlin, Heidelberg) (2006); ISBN 978-3-540-30609-2 978-3-540-33486-6. https://link.springer.com/chapter/10.1007/3-540-33486-6_1.

[15] D. M. CHICKERING; "Optimal structure identification with greedy search"; Journal of machine learning research **3**, pp. 507–554 (2002).

[16] D. HECKERMAN & D. GEIGER; "Likelihoods and parameter priors for Bayesian networks"; arXiv preprint arXiv:2105.06241 (2021).

[17] J. M. PETERS; *Restricted structural equation models for causal inference*; PhD Thesis; ETH Zurich (2012).

[18] P. SPIRTES & C. GLYMOUR; "An algorithm for fast recovery of sparse causal graphs"; Social science computer review **9**, pp. 62–72 (1991).

[19] D. COLOMBO & M. H. MAATHUIS; "Order-independent constraint-based causal structure learning"; The Journal of Machine Learning Research **15**, pp. 3741–3782 (2014).

[20] J. RAMSEY, J. ZHANG & P. L. SPIRTES; "Adjacency-faithfulness and conservative causal inference"; arXiv preprint arXiv:1206.6843 (2012).

[21] C. MEEK; "Causal inference and causal explanation with background knowledge"; in "Proceedings of the Eleventh conference on Uncertainty in artificial intelligence," pp. 403–410 (Morgan Kaufmann Publishers Inc.) (1995).

[22] P. SPIRTES, C. N. GLYMOUR, R. SCHEINES, D. HECKERMAN, C. MEEK, G. COOPER & T. RICHARDSON; *Causation, prediction, and search* (MIT press) (2000).

[23] D. COLOMBO, M. H. MAATHUIS, M. KALISCH & T. S. RICHARDSON; "Learning high-dimensional directed acyclic graphs with latent and selection variables"; The Annals of Statistics pp. 294–321 (2012).

[24] G. J. SZÉKELY, M. L. RIZZO & N. K. BAKIROV; "Measuring and testing dependence by correlation of distances"; The Annals of Statistics **35**, pp. 2769–2794 (2007).

[25] G. J. SZÉKELY & M. L. RIZZO; "Brownian distance covariance"; The Annals of Applied Statistics **3**, pp. 1236–1265 (2009). ISSN 1932-6157, 1941-7330. https://projecteuclid.org/euclid.aoas/1267453933.

[26] W. HOEFFDING; "A non-parametric test of independence"; The annals of mathematical statistics pp. 546–557 (1948)Publisher: JSTOR.

[27] J. B. KINNEY & G. S. ATWAL; "Equitability, mutual information, and the maximal information coefficient"; Proceedings of the National Academy of Sciences **111**, pp. 3354–3359 (2014). ISSN 0027-8424, 1091-6490. http://www.pnas.org/content/111/9/3354.

[28] A. GRETTON, R. HERBRICH, A. SMOLA, O. BOUSQUET & B. SCHÖLKOPF; "Kernel methods for measuring independence"; Journal of Machine Learning Research **6**, pp. 2075–2129 (2005).

[29] A. GRETTON, O. BOUSQUET, A. SMOLA & B. SCHOLKOPF; "Measuring Statistical Dependence with Hilbert-Schmidt Norms"; (2005).

[30] A. GRETTON, K. FUKUMIZU, C. H. TEO, L. SONG, B. SCHÖLKOPF & A. J. SMOLA; "A kernel statistical test of independence." in "Nips," , volume 20pp. 585–592 (Citeseer) (2007).

[31] X. SUN, D. JANZING, B. SCHÖLKOPF & K. FUKUMIZU; "A kernel-based causal learning algorithm"; in "Proceedings of the 24th international conference on Machine learning," pp. 855–862 (ACM) (2007).

[32] A. GRETTON, P. SPIRTES & R. E. TILLMAN; "Nonlinear directed acyclic structure learning with weakly additive noise models"; in "Advances in neural information processing systems," pp. 1847–1855 (2009).

[33] K. ZHANG, J. PETERS, D. JANZING & B. SCHÖLKOPF; "Kernel-based conditional independence test and application in causal discovery"; arXiv preprint arXiv:1202.3775 (2012).

[34] V. A. HUYNH-THU, A. IRRTHUM, L. WEHENKEL & P. GEURTS; "Inferring regulatory networks from expression data using tree-based methods"; PloS one **5**, p. e12776 (2010)Publisher: Public Library of Science San Francisco, USA.

[35] S. AFFELDT, L. VERNY & H. ISAMBERT; "3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics"; in "BMC bioinformatics," , volume 17p. 12 (BioMed Central Ltd) (2016).

[36] S. SHIMIZU, P. O. HOYER, A. HYVÄRINEN & A. KERMINEN; "A linear non-Gaussian acyclic model for causal discovery"; Journal of Machine Learning Research **7**, pp. 2003–2030 (2006).

[37] S. SHIMIZU, T. INAZUMI, Y. SOGAWA, A. HYVÄRINEN, Y. KAWAHARA, T. WASHIO, P. O. HOYER & K. BOLLEN; "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model"; Journal of Machine Learning Research **12**, pp. 1225–1248 (2011).

[38] P. BÜHLMANN, J. PETERS & J. ERNEST; "CAM: Causal additive models, high-dimensional order search and penalized regression"; The Annals of Statistics **42**, pp. 2526–2556 (2014). ISSN 0090-5364. `http://arxiv.org/abs/1310.1533`; arXiv: 1310.1533.

[39] X. ZHENG, B. ARAGAM, P. RAVIKUMAR & E. P. XING; "Dags with no tears: Continuous optimization for structure learning"; arXiv preprint arXiv:1803.01422 (2018).

[40] X. ZHENG, C. DAN, B. ARAGAM, P. RAVIKUMAR & E. XING; "Learning sparse nonparametric DAGs"; in "International Conference on Artificial Intelligence and Statistics," pp. 3414–3425 (PMLR) (2020).

[41] Y. WANG & D. M. BLEI; "The Blessings of Multiple Causes"; arXiv:1805.06826 [cs, stat] (2018)`http://arxiv.org/abs/1805.06826`; arXiv: 1805.06826.

[42] J. PETERS, P. BÜHLMANN & N. MEINSHAUSEN; "Causal inference by using invariant prediction: identification and confidence intervals"; Journal of the Royal Statistical Society: Series B (Statistical Methodology) **78**, pp. 947–1012 (2016). ISSN 1467-9868. `http://onlinelibrary.wiley.com/doi/10.1111/rssb.12167/abstract`.

[43] C. HEINZE-DEML, J. PETERS & N. MEINSHAUSEN; "Invariant Causal Prediction for Nonlinear Models"; Journal of Causal Inference **6** (2018). `https://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2017-0016/jci-2017-0016.xml`.

[44] L. VERNY, N. SELLA, S. AFFELDT, P. P. SINGH & H. ISAMBERT; "Learning causal networks with latent variables from multivariate information in genomic data"; PLoS Computational Biology **13**, p. e1005662 (2017).

[45] C. E. SHANNON; "A mathematical theory of communication"; The Bell system technical journal **27**, pp. 379–423 (1948)Publisher: Nokia Bell Labs.

[46] D. N. RESHEF, Y. A. RESHEF, H. K. FINUCANE, S. R. GROSSMAN, G. MCVEAN, P. J. TURNBAUGH, E. S. LANDER, M. MITZENMACHER & P. C. SABETI; "Detecting Novel Associations in Large Datasets"; Science (New York, N.y.) **334**, pp. 1518–1524 (2011). ISSN 0036-8075. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3325791/.

[47] V. AMRHEIN, S. GREENLAND & B. MCSHANE; "Scientists rise up against statistical significance"; Nature **567**, pp. 305–307 (2019). https://www.nature.com/articles/d41586-019-00857-9; bandiera_abtest: a Cg_type: Comment Number: 7748 Publisher: Nature Publishing Group Subject_term: Research data, Research management.

[48] J. LEEK, B. B. MCSHANE, A. GELMAN, D. COLQUHOUN, M. B. NUIJTEN & S. N. GOODMAN; "Five ways to fix statistics"; Nature **551**, pp. 557–559 (2017). https://www.nature.com/articles/d41586-017-07522-z; bandiera_abtest: a Cg_type: Comment Number: 7682 Publisher: Nature Publishing Group Subject_term: Research data, Lab life, Mathematics and computing.

[49] J. SEOK & Y. S. KANG; "Mutual information between discrete variables with many categories using recursive adaptive partitioning"; Scientific Reports **5**, p. 10981 (2015).

[50] P. A. BOEKEN & J. M. MOOIJ; "A bayesian nonparametric conditional two-sample test with an application to local causal discovery"; arXiv preprint arXiv:2008.07382 (2020).

[51] W. GAO, S. KANNAN, S. OH & P. VISWANATH; "Estimating Mutual Information for Discrete-Continuous Mixtures"; arXiv:1709.06212 [cs, math] (2017)http://arxiv.org/abs/1709.06212; arXiv: 1709.06212.

[52] X. ZENG, Y. XIA & H. TONG; "Jackknife approach to the estimation of mutual information"; Proceedings of the National Academy of Sciences **115**, pp. 9956–9961 (2018).

[53] N. TISHBY, F. C. PEREIRA & W. BIALEK; "The information bottleneck method"; arXiv:physics/0004057 (2000)http://arxiv.org/abs/physics/0004057; arXiv: physics/0004057.

[54] N. TISHBY & N. ZASLAVSKY; "Deep Learning and the Information Bottleneck Principle"; arXiv:1503.02406 [cs] (2015)http://arxiv.org/abs/1503.02406; arXiv: 1503.02406.

[55] A. M. SAXE, Y. BANSAL, J. DAPELLO, M. ADVANI, A. KOLCHINSKY, B. D. TRACEY & D. D. COX; "On the Information Bottleneck Theory of Deep Learning"; (2018). https://openreview.net/forum?id=ry_WPG-A-.

[56] Z. GOLDFELD, E. V. D. BERG, K. GREENEWALD, I. MELNYK, N. NGUYEN, B. KINGSBURY & Y. POLYANSKIY; "Estimating Information Flow in Deep Neural Networks"; in "Proceedings of the 36th International Conference on Machine Learning," pp. 2299–2308 (PMLR) (2019). https://proceedings.mlr.press/v97/goldfeld19a.html; iSSN: 2640-3498.

[57] B. C. GEIGER; "On Information Plane Analyses of Neural Network Classifiers – A Review"; arXiv:2003.09671 [cs, math, stat] (2021)http://arxiv.org/abs/2003.09671; arXiv: 2003.09671.

[58] D. LOPEZ-PAZ, R. NISHIHARA, S. CHINTALA, B. SCHOLKOPF & L. BOTTOU; "Discovering causal signals in images"; in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition," pp. 6979–6987 (2017).

[59] M. ARJOVSKY, L. BOTTOU, I. GULRAJANI & D. LOPEZ-PAZ; "Invariant risk minimization"; arXiv preprint arXiv:1907.02893 (2019).

[60] J. ZHANG & P. SPIRTES; "Detection of Unfaithfulness and Robust Causal Inference"; Minds and Machines **18**, pp. 239–271 (2008). ISSN 1572-8641. https://doi.org/10.1007/s11023-008-9096-4.

[61] J. PETERS, J. M. MOOIJ, D. JANZING & B. SCHÖLKOPF; "Causal Discovery with Continuous Additive Noise Models"; Journal of Machine Learning Research **15**, pp. 2009–2053 (2014).

[62] A. MARX, A. GRETTON & J. M. MOOIJ; "A Weaker Faithfulness Assumption based on Triple Interactions"; arXiv preprint arXiv:2010.14265 (2020).

[63] A. WIECZOREK & V. ROTH; "Information Theoretic Causal Effect Quantification"; Entropy **21**, p. 975 (2019). https://www.mdpi.com/1099-4300/21/10/975; number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[64] L. PANINSKI; "Estimation of entropy and mutual information"; Neural computation **15**, pp. 1191–1253 (2003).

[65] G. MILLER; "Note on the bias of information estimates"; Information theory in psychology (1955).

[66] B. EFRON & C. STEIN; "The Jackknife Estimate of Variance"; The Annals of Statistics **9**, pp. 586–596 (1981). ISSN 0090-5364, 2168-8966. https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-3/The-Jackknife-Estimate-of-Variance/10.1214/aos/1176345462.full; publisher: Institute of Mathematical Statistics.

[67] T. SCHÜRMANN & P. GRASSBERGER; "Entropy estimation of symbol sequences"; Chaos: An Interdisciplinary Journal of Nonlinear Science **6**, pp. 414–427 (1996). ISSN 1054-1500, 1089-7682. http://arxiv.org/abs/cond-mat/0203436; arXiv: cond-mat/0203436.

[68] I. NEMENMAN, W. BIALEK & R. D. R. VAN STEVENINCK; "Entropy and information in neural spike trains: Progress on the sampling problem"; Physical Review E **69**, p. 056111 (2004)Publisher: APS.

[69] M. VEJMELKA & M. PALUŠ; "Inferring the directionality of coupling with conditional mutual information"; Physical Review E **77**, p. 026214 (2008).

[70] R. MODDEMEIJER; "On estimation of entropy and mutual information of continuous distributions"; Signal processing **16**, pp. 233–248 (1989).

[71] B. C. ROSS; "Mutual Information between Discrete and Continuous Data Sets"; PLOS ONE **9**, p. e87357 (2014). ISSN 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0087357.

[72] C. O. DAUB, R. STEUER, J. SELBIG & S. KLOSKA; "Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data"; BMC bioinformatics **5**, p. 118 (2004).

[73] G. A. DARBELLAY & I. VAJDA; "Estimation of the information by an adaptive partitioning of the observation space"; IEEE Transactions on Information Theory **45**, pp. 1315–1321 (1999).

[74] Q. WANG, S. R. KULKARNI & S. VERDÚ; "Divergence estimation of continuous distributions based on data-dependent partitions"; IEEE Transactions on Information Theory **51**, pp. 3064–3074 (2005).

[75] G. A. DARBELLAY & I. VAJDA; "Entropy expressions for multivariate continuous distributions"; IEEE Transactions on Information Theory **46**, pp. 709–712 (2000). ISSN 0018-9448.

[76] Y.-I. MOON, B. RAJAGOPALAN & U. LALL; "Estimation of mutual information using kernel density estimators"; Physical Review E **52**, pp. 2318–2321 (1995). https://link.aps.org/doi/10.1103/PhysRevE.52.2318.

[77] A. KRASKOV, H. STÖGBAUER & P. GRASSBERGER; "Estimating mutual information"; Physical review E **69**, p. 066138 (2004).

[78] L. F. KOZACHENKO & N. N. LEONENKO; "Sample estimate of the entropy of a random vector"; Problemy Peredachi Informatsii **23**, pp. 9–16 (1987)Publisher: Russian Academy of Sciences, Branch of Informatics, Computer Equipment and ....

[79] H. SINGH, N. MISRA, V. HNIZDO, A. FEDOROWICZ & E. DEMCHUK; "Nearest neighbor estimates of entropy"; American journal of mathematical and management sciences **23**, pp. 301–321 (2003)Publisher: Taylor & Francis.

[80] J. JIAO, W. GAO & Y. HAN; "The Nearest Neighbor Information Estimator is Adaptively Near Minimax Rate-Optimal"; arXiv:1711.08824 [cs, math, stat] (2018)http://arxiv.org/abs/1711.08824; arXiv: 1711.08824.

[81] W. GAO, S. OH & P. VISWANATH; "Demystifying Fixed k-Nearest Neighbor Information Estimators"; arXiv:1604.03006 [cs, math, stat] (2016)http://arxiv.org/abs/1604.03006; arXiv: 1604.03006.

[82] A. TSIMPIRIS, I. VLACHOS & D. KUGIUMTZIS; "Nearest neighbor estimate of conditional mutual information in feature selection"; Expert Systems with Applications **39**, pp. 12697–12708 (2012).

[83] F. PÉREZ-CRUZ; "Estimation of Information Theoretic Measures for Continuous Random Variables"; in "Advances in Neural Information Processing Systems," , volume 21 (Curran Associates, Inc.) (2009). https://proceedings.neurips.cc/paper/2008/hash/ccb0989662211f61edae2e26d58ea92f-Abstract.html.

[84] C. M. HOLMES & I. NEMENMAN; "Estimation of mutual information for real-valued data with error bars and controlled bias"; bioRxiv p. 589929 (2019). https://www.biorxiv.org/content/10.1101/589929v2-0.

[85] J. RUNGE; "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information"; in "International Conference on Artificial Intelligence and Statistics," pp. 938–947 (2018). http://proceedings.mlr.press/v84/runge18a.html.

[86] T. B. BERRETT & R. J. SAMWORTH; "Nonparametric independence testing via mutual information"; Biometrika **106**, pp. 547–566 (2019). ISSN 0006-3444. https://academic.oup.com/biomet/article/106/3/547/5511208.

[87] K. R. MOON, K. SRICHARAN & A. O. HERO; "Ensemble estimation of mutual information"; in "2017 IEEE International Symposium on Information Theory (ISIT)," pp. 3030–3034 (2017); iSSN: 2157-8117.

[88] R. A. INCE, B. L. GIORDANO, C. KAYSER, G. A. ROUSSELET, J. GROSS & P. G. SCHYNS; "A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula"; Human brain mapping **38**, pp. 1541–1573 (2017).

[89] M. NOSHAD, Y. ZENG & A. O. HERO III; "Scalable Mutual Information Estimation using Dependence Graphs"; arXiv:1801.09125 [cs, math, stat] (2018)`http://arxiv.org/abs/1801.09125`; arXiv: 1801.09125.

[90] I. BELGHAZI, S. RAJESWAR, A. BARATIN, R. D. HJELM & A. COURVILLE; "MINE: mutual information neural estimation"; arXiv preprint arXiv:1801.04062 (2018).

[91] A. MARX, L. YANG & M. VAN LEEUWEN; "Estimating Conditional Mutual Information for Discrete-Continuous Mixtures using Multi-Dimensional Adaptive Histograms"; in "Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)," pp. 387–395 (SIAM) (2021).

[92] J. RISSANEN; "Modeling by shortest data description"; Automatica **14**, pp. 465–471 (1978).

[93] T. ROOS, T. SILANDER, P. KONTKANEN & P. MYLLYMAKI; "Bayesian network structure learning using factorized NML universal models"; in "2008 Information Theory and Applications Workshop," pp. 272–276 (IEEE) (2008).

[94] A. MARX & J. VREEKEN; "Stochastic Complexity for Testing Conditional Independence on Discrete Data"; in "NeurIPS 2018 Workshop on Causal Learning," (2018).

[95] J. RISSANEN; "Fisher information and stochastic complexity"; IEEE Transactions on Information Theory **42**, pp. 40–47 (1996). ISSN 1557-9654; conference Name: IEEE Transactions on Information Theory.

[96] Y. M. SHTAR'KOV; "Universal sequential coding of single messages"; Problemy Peredachi Informatsii **23**, pp. 3–17 (1987)Publisher: Russian Academy of Sciences, Branch of Informatics, Computer Equipment and ....

[97] P. KONTKANEN, P. MYLLYMÄKI, W. BUNTINE, J. RISSANEN & H. TIRRI; "An MDL Framework for Data Clustering"; in "Advances in Minimum Description Length," p. 323 (The MIT press) (2005).

[98] T. ROOS, P. MYLLYMÄKI & H. TIRRI; "On the Behavior of MDL Denoising"; in "International Workshop on Artificial Intelligence and Statistics," pp. 309–316 (PMLR) (2005). `https://proceedings.mlr.press/r5/roos05a.html`; iSSN: 2640-3498.

[99] P. KONTKANEN & P. MYLLYMÄKI; "MDL histogram density estimation"; in "Artificial Intelligence and Statistics," pp. 219–226 (2007).

[100] J. RISSANEN; "Strong optimality of the normalized ML models as universal codes and information in data"; IEEE Transactions on Information Theory **47**, pp. 1712–1717 (2001)Publisher: IEEE.

[101] P. KONTKANEN & P. MYLLYMÄKI; "A linear-time algorithm for computing the multinomial stochastic complexity"; Information Processing Letters **103**, pp. 227–233 (2007).

[102] W. SZPANKOWSKI; *Average case analysis of algorithms on sequences*; volume 50 (John Wiley & Sons) (2011).

[103] P. KONTKANEN, W. BUNTINE, P. MYLLYMÄKI, J. RISSANEN & H. TIRRI; "Efficient computation of stochastic complexity"; in "Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics," (Citeseer) (2003).

[104] P. KONTKANEN; "Computationally efficient methods for MDL-optimal density estimation and data clustering"; (2009).

[105] M. NEIL, M. TAILOR & D. MARQUEZ; "Inference in hybrid Bayesian networks using dynamic discretization"; Statistics and Computing **17**, pp. 219–233 (2007). ISSN 1573-1375. `https://doi.org/10.1007/s11222-007-9018-y`.

[106] D. MARQUEZ, M. NEIL & N. FENTON; "Improved reliability modeling using Bayesian networks and dynamic discretization"; Reliability Engineering & System Safety **95**, pp. 412–425 (2010). ISSN 0951-8320. `https://www.sciencedirect.com/science/article/pii/S0951832009002646`.

[107] J. T. LIZIER; "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems"; Frontiers in Robotics and AI **1**, p. 11 (2014).

[108] J. M. MOOIJ, S. MAGLIACANE & T. CLAASSEN; "Joint causal inference from multiple contexts"; arXiv preprint arXiv:1611.10351 (2016).

[109] A. J. SEDGEWICK, K. BUSCHUR, I. SHI, J. D. RAMSEY, V. K. RAGHU, D. V. MANATAKIS, Y. ZHANG, J. BON, D. CHANDRA & C. KAROLESKI; "Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis"; Bioinformatics (2018).

[110] M. TSAGRIS, G. BORBOUDAKIS, V. LAGANI & I. TSAMARDINOS; "Constraint-based causal discovery with mixed data"; International Journal of Data Science and Analytics pp. 1–12 (2018).

[111] D. B. RUBIN; "Inference and missing data"; Biometrika **63**, pp. 581–592 (1976)Publisher: Oxford University Press.

[112] R. J. LITTLE & D. B. RUBIN; *Statistical analysis with missing data*; volume 793 (John Wiley & Sons) (2019).

[113] K. MOHAN & J. PEARL; "Graphical Models for Processing Missing Data"; Forthcoming, Journal of American Statistical Association (JASA) (2019).

[114] K. MOHAN, J. PEARL & J. TIAN; "Graphical Models for Inference with Missing Data"; Advances in Neural Information Processing Systems 26 pp. 1277–1285 (2013)Publisher: Citeseer.

[115] D. DASH & M. J. DRUZDZEL; "Robust Independence Testing for Constraint-Based Learning of Causal Structure." in "UAI," , volume 3pp. 167–174 (2003).

[116] P. LERAY & O. FRANÇOIS; "Bayesian network structural learning and incomplete data"; in "Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005), Espoo, Finland," pp. 33–40 (2005).

[117] E. V. STROBL, K. ZHANG & S. VISWESWARAN; "Approximate Kernel-Based Conditional Independence Tests for Fast Non-Parametric Causal Discovery"; Journal of Causal Inference **7** (2017). ISSN 2193-3685. https://www.degruyter.com/view/j/jci.2019.7.issue-1/jci-2018-0017/jci-2018-0017.xml?format=INT.

[118] R. TU, K. ZHANG, P. ACKERMANN, B. C. BERTILSON, C. GLYMOUR, H. KJELLSTRÖM & C. ZHANG; "Causal Discovery in the Presence of Missing Data"; arXiv:1807.04010 [cs, stat] (2020)http://arxiv.org/abs/1807.04010; arXiv: 1807.04010.

[119] N. H. ANDERSON, P. HALL & D. M. TITTERINGTON; "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates"; Journal of Multivariate Analysis **50**, pp. 41–54 (1994)Publisher: Elsevier.

[120] F. PÉREZ-CRUZ; "Kullback-Leibler divergence estimation of continuous distributions"; in "2008 IEEE international symposium on information theory," pp. 1666–1670 (IEEE) (2008).

[121] J. L. BLANCO & P. K. RAI; "nanoflann: a {C}++ header-only fork of {FLANN}, a library for Nearest Neighbor ({NN}) with KD-trees"; (2014). https://github.com/jlblancoc/nanoflann.

[122] S. A. BHAVE, R. RANGANATH & A. PEROTTE; "Information Theoretic Approaches for Testing Missingness in Predictive Models"; (2020)https://openreview.net/forum?id=6Y05VJfGlFM.

[123] C. M. FIFORD, E. N. MANNING, J. W. BARTLETT, D. M. CASH, I. B. MALONE, G. R. RIDGWAY, M. LEHMANN, K. K. LEUNG, C. H. SUDRE, S. OURSELIN, G. J. BIESSELS, O. T. CARMICHAEL, N. C. FOX, M. J. CARDOSO, J. BARNES & ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE; "White matter hyperintensities are associated with disproportionate progressive hippocampal atrophy"; Hippocampus **27**, pp. 249–262 (2017). ISSN 1098-1063.

[124] K. ITO & K. ITO; "Hematopoietic stem cell fate through metabolic control"; Experimental hematology **64**, pp. 1–11 (2018)Publisher: Elsevier.

[125] A. NAKAMURA-ISHIZU, K. ITO & T. SUDA; "Hematopoietic stem cell metabolism during development and aging"; Developmental cell **54**, pp. 239–255 (2020)Publisher: Elsevier.

[126] L. OBUROGLU, S. TARDITO, V. FRITZ, S. C. DE BARROS, P. MERIDA, M. CRAVEIRO, J. MAMEDE, G. CRETENET, C. MONGELLAZ & X. AN; "Glucose and glutamine metabolism regulate human hematopoietic stem cell lineage specification"; Cell stem cell **15**, pp. 169–184 (2014)Publisher: Elsevier.

[127] J. CHOI, T. M. BALDWIN, M. WONG, J. E. BOLDEN, K. A. FAIRFAX, E. C. LUCAS, R. COLE, C. BIBEN, C. MORGAN, K. A. RAMSAY, A. P. NG, M. KAUPPI, L. M. CORCORAN, W. SHI, N. WILSON, M. J. WILSON, W. S. ALEXANDER, D. J. HILTON & C. A. DE GRAAF; "Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans"; Nucleic Acids Research **47**, pp. D780–D785 (2019). ISSN 0305-1048. https://doi.org/10.1093/nar/gky1020.

[128] S. M. LUNDBERG, B. NAIR, M. S. VAVILALA, M. HORIBE, M. J. EISSES, T. ADAMS, D. E. LISTON, D. K.-W. LOW, S.-F. NEWMAN, J. KIM & S.-I. LEE; "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery"; Nature Biomedical Engineering **2**, p. 749 (2018). ISSN 2157-846X. https://www.nature.com/articles/s41551-018-0304-0.

[129] S. M. LUNDBERG, G. G. ERION & S.-I. LEE; "Consistent Individualized Feature Attribution for Tree Ensembles"; arXiv:1802.03888 [cs, stat] (2018)http://arxiv.org/abs/1802.03888; arXiv: 1802.03888.

[130] S.-W. WANG & A. M. KLEIN; "Learning dynamics by computational integration of single cell genomic and lineage information"; Technical report (2021). https://www.biorxiv.org/content/10.1101/2021.05.06.443026v1; company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[131] C. WEINREB, A. RODRIGUEZ-FRATICELLI, F. D. CAMARGO & A. M. KLEIN; "Lineage tracing on transcriptional landscapes links state to fate during differentiation"; Science **367** (2020). ISSN 0036-8075, 1095-9203. https://science.sciencemag.org/content/367/6479/eaaw3381; publisher: American Association for the Advancement of Science Section: Research Article.

[132] C. W. GRANGER; "Investigating causal relations by econometric models and cross-spectral methods"; Econometrica: journal of the Econometric Society pp. 424–438 (1969).

[133] T. SCHREIBER; "Measuring information transfer"; Physical review letters **85**, p. 461 (2000).

[134] A. A. MASTAKOURI, B. SCHÖLKOPF & D. JANZING; "Necessary and sufficient conditions for causal feature selection in time series with latent common causes"; arXiv:2005.08543 [stat] (2020)http://arxiv.org/abs/2005.08543; arXiv: 2005.08543.

[135] J. RUNGE; "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets"; in "Conference on Uncertainty in Artificial Intelligence," pp. 1388–1397 (PMLR) (2020). http://proceedings.mlr.press/v124/runge20a.html; iSSN: 2640-3498.

[136] E. CANDES, Y. FAN, L. JANSON & J. LV; "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection"; Journal of the Royal Statistical Society: Series B (Statistical Methodology) **80**, pp. 551–577 (2018).

[137] S. BATES, E. CANDÈS, L. JANSON & W. WANG; "Metropolized Knockoff Sampling"; arXiv:1903.00434 [stat] (2019)http://arxiv.org/abs/1903.00434; arXiv: 1903.00434.

[138] A. MARX & J. VREEKEN; "Testing Conditional Independence on Discrete Data using Stochastic Complexity"; arXiv preprint arXiv:1903.04829 (2019).

[139] N. SELLA, L. VERNY, G. UGUZZONI, S. AFFELDT & H. ISAMBERT; "MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data"; Bioinformatics (2017).

[140] T. TASHIRO, S. SHIMIZU, A. HYVÄRINEN & T. WASHIO; "Estimation of causal orders in a linear non-Gaussian acyclic model: a method robust against latent confounders"; Artificial Neural Networks and Machine Learning–ICANN 2012 pp. 491–498 (2012).

[141] W. F. SYMMANS, F. PEINTINGER, C. HATZIS, R. RAJAN, H. KUERER, V. VALERO, L. ASSAD, A. PONIECKA, B. HENNESSY & M. GREEN; "Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy"; Journal of Clinical Oncology **25**, pp. 4414–4422 (2007)Publisher: American Society of Clinical Oncology.