Learning causal networks from mixed type variables with (conditional) mutual information estimation

Vincent Cabeli,[1]
Hervé Isambert[1]
[1]Laboratoire Physico-Chimie Curie,
Institut Curie, Paris
vincent.cabeli@curie.fr, herve.isambert@curie.fr

Abstract

Causal network inference can be a very powerful method to distinguish direct and indirect relationships in observed data, and when the right conditions are met one can even hope to uncover the real causal ordering between the variables. However, this often proves difficult to do on real-world datasets due to incomplete data, mixed-type variables (i.e. categorical, ordinal as well as continuous observations), unobserved confounders etc... To this end we present the latest improvements of the miic algorithm, an information-theoretic approach that learns causal or non-causal graphical models on any type of data, also taking into account the effects of unobserved latent variables. The main talking point of this article is an efficient optimal discretization scheme that simultaneously estimates and assesses the significance of the (conditional) mutual information between any mixed variables. With this new (conditional) independence test, the method is shown to outperform existing tools on mixed-type benchmarks, before being applied to analyze the medical records of elderly patients with cognitive disorders from La Pitié-Salpêtrière Hospital, Paris. This article is an abridged version of (Cabeli et al. 2020).

Virtually all of science is concerned with understanding the cause-effect relationships between events, whether they be weather patterns or the human's body response to treatment. When possible, the gold standard for discovering causal relationships is the case-control studies where the experimentalist can control all variables except for the studied intervention. However, such randomized experiments can often prove to be either too costly, unethical or simply impossible. Designed for these cases, causal inference methods have steadily improved during the last decades benefiting from more computing power and better data collection practices. There are two main families of methods to formally define causal effects, with similar capabilities but different formulations : Rubin's potential outcome framework and Pearl's graph-based do calculus.

Pearl's graphs are a simple and powerful way to represent causal links between many variables in a causal network, a sub-class of Bayesian networks with the added requirement that the relationships be causal i.e. intervening on x will change the probability of y in the graph $x \longrightarrow y$, but the inverse is not true. This kind of representation makes visually clear if a given node is a confounder or a collider and whether or not one should include it to compute causal effect such as the Average Treatment Effect (Pearl 2009; Pearl, Glymour, and Jewell 2016). Many algorithms have been proposed to try to infer such networks from observations, ranging from Bayesian score-based approaches (Heckerman, Meek, and Cooper 2006), constraint-based approaches based on (conditional) independence tests (Spirtes and Glymour 1991), graphical lasso (Meinshausen and Bühlmann 2006), continuous optimization over matrices (Zheng et al. 2018, 2020), independent component analysis (Shimizu et al. 2006) or even using random forests' feature importance scores (Irrthum, Wehenkel, and Geurts 2010).Naturally, these methods rely on several assumptions to retrieve the causal signal from pure observational data. Most methods that output a causal graph rely on both the *Causal Faithfulness Assumption* (i.e. the data distribution holds no extra conditional independence that what is described by its corresponding graph, it is *faithful to the graph*) and the *Causal Markov Assumption* (an extension of the Markov property of Bayesian graphs that takes into account the causal hierarchy) (Glymour, Zhang, and Spirtes 2019). Some methods will then add extra assumptions on either the data distribution (continuous vs discrete, Gaussian vs non-Gaussian), the relationships between variables (linear vs non-linear) or the nature of residuals when regressing a node on its parents. This is where constraint-based approaches shine and why they tend to be more usable on real-world data, since except for the two base assumptions already mentioned they are only limited by their conditional independence test. When needed, they can be tailored for a specific problem (i.e. Gaussian distributions and linear relationships for normalized gene expressions (Spirtes et al. 2000)) or on the contrary they can aim to be as general as possible with no assumption on either data distribution or relationships (Azadkia and Chatterjee 2021).

Despite a large community effort, conditional independence testing for the general case remains a difficult problem. Existing methods tend to be designed for specific cases in which they can work well (Pfister et al. 2016; Shah and Peters 2020) but we are still lacking a test that would enable truly general purpose constraint-based causal discovery. This article focuses on the mixed case, for which we want to take into account both discrete (ordinal or not) and

continuous variables (of any distribution) in a way that does not favor one or the other. Few methods have been proposed to deal with this type of data, and existing examples simply rely on using different tests depending on the combination of nodes being considered (Sedgewick et al. 2018; Tsagris et al. 2018). Crucially, one would like the output of the algorithm to be independent of the nature in which the data comes, which is difficult to guarantee when one uses different tests with different sensitivities and power. An ideal way to measure dependence between variables of any type is the (conditional) mutual information, as it is defined in both the discrete and the continuous case. Regardless of data distribution, mutual information has the desirable property that $I(X;Y) = 0 \leftrightarrow X \perp\!\!\!\perp Y$ and $I(X;Y|Z) = 0 \leftrightarrow X \perp\!\!\!\perp Y|Z$. However, this quantity is difficult to estimate on finite data, particularly in the regime that interests us at (conditional) independence where estimators give a small but non-null value (Kraskov, Stögbauer, and Grassberger 2004; Belghazi et al. 2018). One way that has been proposed to assess the significance of near-null estimates of mutual information is by running permutation tests, which was then adapted to the conditional case (Runge 2017) but it can be slow and impractical.

In this article we describe the discretization scheme that has been developed to take advantage of the significance test for the mutual information introduced with the *miic* algorithm (Affeldt and Isambert 2015; Affeldt, Verny, and Isambert 2016), based on the stochastic complexity of discrete data and proven to be asymptotically correct (Marx and Vreeken 2019). Our method is tailored to work in the constraint-based setting, i.e. univariate X and Y and a potentially high dimension conditioning set Z, but makes no assumption on the data distribution or the nature of the relationships. We base our estimator on a Minimum Description Length (MDL)-optimal binning scheme for univariate variables (Kontkanen and Myllymäki 2007b), adapted to the joint case to give at the same time an estimation of the information and a test of its significance. Before being plugged into *miic*, our proposed discretization approach is shown to both give a correct estimation of $I(X;Y)$ and $I(X;Y|Z)$ on any X, Y, Z, and perform favorably compared to state-of-the-art mixed case non-parametric (conditional) independence test. We then compare our performance on full-scale causal inference benchmarks on synthetic dataset for both the continuous and mixed case, once again showing that our optimal discretization scheme is able to keep up or outperform competing methods. Finally, we use *miic* on medical records from elderly patients presenting signs of mental disorders in collaboration with La Pitié-Salpêtrière Hospital, Paris.

Methods

Mutual information testing

While mutual information is usually defined as a discrete summation over nominal variables, *i.e.*, $I(X;Y) = \sum_{x,y} p_{x,y} \log\left(p_{x,y}/p_x p_y\right)$, its most general definition consists in taking the supremum over all finite partitions, $\mathcal{P}$ and $\mathcal{Q}$, of variables, X and Y (Cover and Thomas 2012),

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

which can be applied to continuous or mixed-type variables.

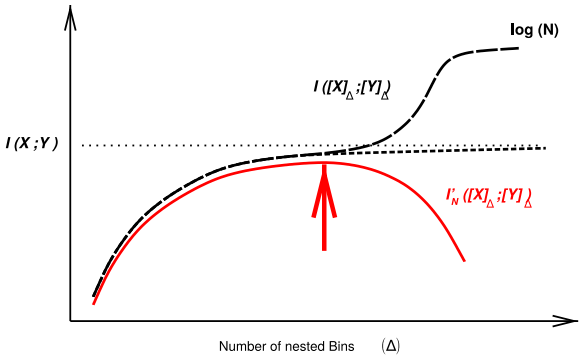

Figure 1: **Mutual information computation between continuous or mixed-type variables.** Outline of mutual information computation between continuous or mixed-type variables for a finite dataset of N samples. Mutual information is estimated through an optimum partitioning of continuous variable(s) (solid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

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 (Cover and Thomas 2012), $I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}})$, 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 (Affeldt and Isambert 2015; Affeldt, Verny, and Isambert 2016) 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 (Roos et al. 2008; Affeldt and Isambert 2015; Affeldt, Verny, and Isambert 2016), defined as,

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

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

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

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

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

which can in fact be computed recursively in linear-time (Kontkanen and Myllymäki 2007a). For large n and r, inherent to large datasets with continuous or mixed-type variables, we found that $\mathcal{C}^r_n$ computation can be made numerically stable by implementing the recursion on parametric complexity ratios $\mathcal{D}^r_n = \mathcal{C}^r_n/\mathcal{C}^{r-1}_n$ rather than parametric complexities themselves as,

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

$$\log \mathcal{C}^r_n = \sum_{k=2}^{r} \log \mathcal{D}^k_n \quad (8)$$

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

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

or its Szpankowski approximation for large n (needed for $n > 1000$ in practice) (Szpankowski 2011; Kontkanen et al. 2003; Kontkanen 2009),

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

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

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 (Kontkanen and Myllymäki 2007b). 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 (Kontkanen and Myllymäki 2007b). 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,

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

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 $I'_N([X]_{\mathcal{P}}; [Y]_{\mathcal{Q}})$ 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 (Kontkanen and Myllymäki 2007b), which can be done exactly in $O(N^3)$ 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 $O(N^2)$, 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,

$$nI'_n(X;Y) = \max_{0 \leqslant j < n} \Big[jI'_j(X;Y) + \sum_{x}^{r_x} n_{xy} \log n_{xy} - n_y \log n_y - \log \mathcal{C}^{r_x}_{n_y} - C_{N,r^\circ_y} \Big] \quad (13)$$

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, $\log \mathcal{C}^{r_x}_{n_y}$, 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.*, $-\sum_x^{r_x} n_x \log(n_x/N) + \log \mathcal{C}^{r_x}_N - (r_x - 1)C_{N,r_x} + C_{N,r^\circ_y}$.

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 the highest $I'_N(X;Y)$. 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 corrections and 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 (Cover and Thomas 2012), 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, $n_{y,\{a_i\}}$ and $n_{\{a_i\}}$. 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, $n_{x,\{a_i\}}$ and $n_{\{a_i\}}$. 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 (Verny et al. 2017) 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 (Verny et al. 2017; Sella et al. 2017). 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 (Affeldt and Isambert 2015; Affeldt, Verny, and Isambert 2016) 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 (Zhao et al. 2016). 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 (Verny et al. 2017)

ii) *Edge filtering (optional).* The remaining edges can be further filtered based on confidence ratio assessment (Verny et al. 2017), $C_{XY} = P_{XY}/\langle P_{XY}^{\text{rand}}\rangle$, 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 (Verny et al. 2017).

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 (Affeldt and Isambert 2015; Affeldt, Verny, and Isambert 2016). The final network contains up to three types of edges (Verny et al. 2017): 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.

The full source code is available as an R package published on CRAN (package "miic") or at https://github.com/miicTeam/miic_R_package. An online server is also available at https://miic.curie.fr.

Benchmarks

(Conditional) Mutual information estimation In order to benchmark the accuracy of the mutual information estimation given by our optimal discretization scheme, we used various joint distributions for which we know the value of $I(X;Y)$ analytically. We include three settings : mutual information in the bivariate gaussian case, mutual information for 4 mixed cases directly taken from another estimator (Gao et al. 2017), and the conditional mutual information with a 4-variables setup simulating "V-structures", conditional independences and neutral conditioning nodes. For each experiment we compare the "miic" estimation obtained with optimal discretization $I'_N(X;Y)$ to the value returned by state-of-the art estimators with varying parameter if necessary, as well as the Mean Squared Error, and the estimators' variances. The code to reproduce all of these experiments is available at https://github.com/vcabeli/miicPLoS.

(Conditional) Independence testing Next, we tested the sensitivity and power of our estimator to detect (conditional) independence. We reproduced the newly published tests for mixed conditional independence test by (Boeken and Mooij 2020) based around the "Local Causal Discovery" algorithm (Mooij, Magliacane, and Claassen 2016). 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 any estimator 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 :

$$I'_{pval}(X;Y) = 1 - \frac{I'_N(X;Y)}{min(I'_N(X;X), I'_N(Y;Y))} \quad (16)$$

$$I'_{pval}(X;Y|Z) = 1 - \frac{I'_N(X;Y|Z)}{min(I'_N(X;X), I'_N(Y;Y))} \quad (17)$$

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 $]0,1]$.

For details of the different simulations used to benchmark independence testing, we refer the reader to (Boeken and Mooij 2020). The code for this part is available at https://github.com/vcabeli/PTTests.

Network reconstruction benchmarks In order to test MIIC with the new optimal discretization scheme, we simulated observational data from known data-generating graphs and compared how well different causal reconstruction methods are able to recover the true DAGs from the observations. First, the underlying DAG models were randomly drawn from the space of all possible DAGs (Melancon and Philippe 2004), 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 in (Cabeli et al. 2020).

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.

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'.

Methods which have a tunable parameter (such as the α level for significance, typically controlling the false positive rate) were tuned for each sample size N and network type to produce the best F-score, and were then compared to MIIC with default settings. See (Cabeli et al. 2020) for details on parameter tuning.

Results

We will present three different types of benchmark before showing the network obtained on real data from medical records of La Pitié-Salpêtrière patients. First, we show that our optimal discretization scheme gives a correct (conditional) mutual information estimation between variables of any type, comparable to state of the art estimators. Second, we focus on the performance of miic's discretization when used as a (conditional) independence test in difficult settings with mixed variables and nonlinear relationships. Finally, we run full benchmarks on continuous as well as mixed datasets to see how well our method is able to recover the true data-generating CPDAG.

Pairwise and conditional mutual information estimation

Before testing the accuracy of the estimation, we first make qualitative observations on the optimal discretization returned by our method. We note that number of bins increases both with the number of samples and the magnitude of mutual information, $I_N(X;Y)$, Fig. S1. 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 variables. We also note that no single discretization of a variable X can be optimal with regards to every joint distribution, Fig. S1(D-F). See (Cabeli et al. 2020) for further discussion.

We compared our estimation of $I'_N(X;Y)$ by optimal discretization to the state of the art Kraskov–Stögbauer–Grassberger (KSG) estimator (Kraskov, Stögbauer, and Grassberger 2004) for continuous distributions, specifically bivariate Gaussian distributions with more or less covariance Fig. 2. Like other information 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 (Ross 2014; Gao et al. 2017; Zeng, Xia, and Tong 2018) which are frequently encountered in real-life data. We compared the mixed-type information estimates of our method to other existing meth-

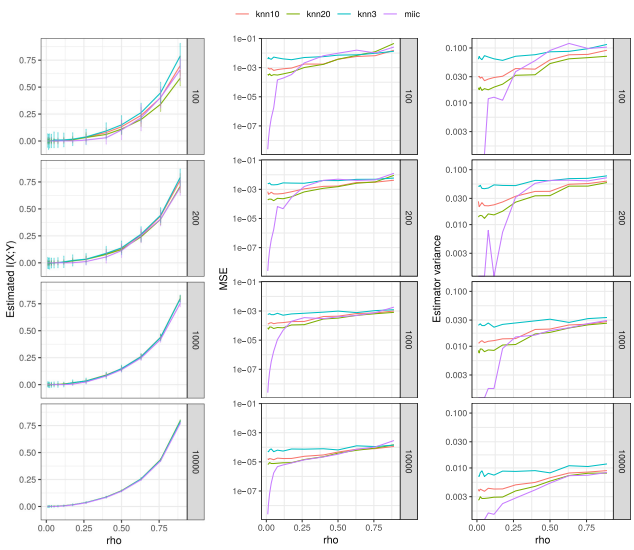

Figure 2: **Mutual information estimation benchmarks on bivariate gaussian distributions.** 100 bivariate normal distributions were sampled for varying sample sizes, increasing from top to bottom, and correlation coefficients ρ ranging from 0.01 to 0.9. The mutual information was estimated with the proposed optimum discretization scheme and the KSG estimator with different parameters k (left). The mean squared error (MSE, middle) 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). Code is available at https://github.com/vcabeli/miic_PLoS

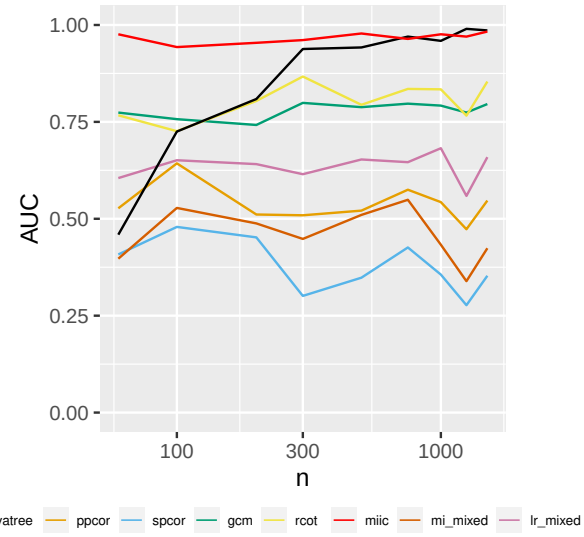

Figure 3: **Conditional independence tests on mixed variables**. Mean Area Under the Curve of ROC curves from 200 rounds of simulation at each sample size n for the LCD triple (Boeken and Mooij 2020). The triple is scored according to a combination of three p-values for three independence tests : $C \not\perp\!\!\!\perp X$, $X \not\perp\!\!\!\perp Y$ 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. Code and detailed outputs are available at https://github.com/vcabeli/PTTests

ods for varying sample sizes and found its performance to be similar or superior, Fig. S2.

Similarly, our approach gives a robust estimation of the conditional mutual information, given a set of separating variables, $\{Z_i\}$, Fig. S3. The experiment shows that our estimator works in all conditioning regimes : first, spurious dependency can be induced between independent X and Y by conditioning over a common descendent Z, as in the case of a "v-structure", $X \to Z \leftarrow Y$, Fig. S3 (Pearl 2009). Conditional independence, i.e. $I'_N(X;Y|Z) = 0$, can also be obtained as in the case of a single common ancestor Z of X and Y, *i.e.*, $X \leftarrow Z \to Y$, with concomitant changes in optimum X and Y partitionings from multiple to single bins under conditioning over a continuous or categorical variable Z. Finally, our approach also gives a correct estimation when X and Y share more information that the indirect flow that goes through the conditioning set Z, in which case $I'_N(X;Y|Z)$ is the residual information once taking into account Z.

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.* $I_N(X;Y) = 0 \leftrightarrow X \perp\!\!\!\perp Y$, with drastically reduced sampling error and variance, Fig. 2, 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 (Kraskov, Stögbauer, and Grassberger 2004; Runge 2018).

When converted to a p-value-like value with Eq. 17, we can also compare our estimator over the full range of significance levels and between different marginal distributions, and compute ROC curves along with their Area Under the Curve (AUC), Fig. 3. In difficult settings with non-linear relationships and possibly discrete variables, we found that our approach is similar or superior to all of the standard methods tested, such as the generalized covariance measure (gcm, (Shah and Peters 2020)), the partial correlation test (ppcor, implemented in (Kim 2015)), the Spearman correlation test (spcor, promoted by (Harris and Drton 2013)) and Randomised Conditional Correlation Test (rcot, (Strobl, Zhang, and Visweswaran 2019)). We found that it is also superior to methods specifically designed for the mixed case such as the mixed mutual information estimation by (Scutari, Scutari, and MMPC 2017) (mi_mixed), the likelihood ratio test by (Sedgewick et al. 2018) (lr_mixed) and the recent Pólya tree-based Bayesian nonparametric test (Boeken and Mooij 2020) (polyatree).

The intrinsic robustness of the present optimum dis-

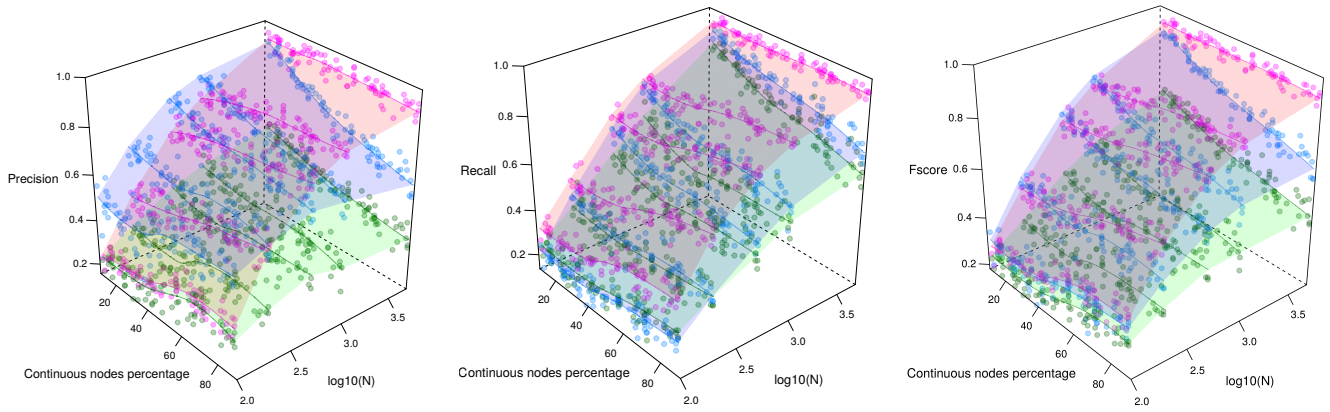

Figure 4: **CPDAG benchmark resuls Reconstruction of benchmark networks for mixed-type, non-linear, non-Gaussian datasets**. CPDAG F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from N=100–5,000 samples. F-scores obtained with our parameter-free information-theoretic approach MIIC (magenta) are compared to the best results obtained with alternative mixed-type data methods, CausalMGM (Sedgewick et al. 2018) (blue) and MXM (Tsagris et al. 2018) (green), by optimizing CausalMGM regularization parameters (λ) and MXM significance parameter (α), for each sample size N.

cretization scheme in inferring (conditional) independence and dependency is a central feature of MIIC as a causal network inference algorithm.

Causal network inference benchmarks

We then tested the mixed-type data extension of MIIC network reconstruction method on benchmark mixed-type data. The performance at reconstructing directed networks are shown as recall, precision and F-score in Fig. S4 for fully continuous networks compared to state of the art methods for non-linear relationships, and in Fig. 4 for mixed networks with varying proportion of continuous variables over discrete variables and compared to the recent alternative methods, CausalMGM(Sedgewick et al. 2018) and MXM(Tsagris et al. 2018), also designed to analyze mixed-type data.

We note that for continuous networks, MIIC performs better than competing methods even after careful parameter optimization. In particular, introducing interaction terms when a node has several parents means that the Causal Additive Model (Bühlmann, Peters, and Ernest 2014) does not perform as well as it should, given that the simulated datasets were generated using additive models. The well-studied kernel-based Helbert-Schmidt Independence Criterion (Gretton et al. 2005; Gretton, Spirtes, and Tillman 2009) also seems to perform well, but its complexity scales badly with sample size N. Being based on the mutual information, our method is sensitive to all types of (in)dependencies present in the data, hence its better performance on recovering CPDAGs.

MIIC also seems to outperform competing mixed methods, and in particular has the same performance on mostly discrete versus mostly continuous datasets, which suggests better adaptability and stability against the various forms in which the data may be collected.

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 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). Beyond the different types and heterogeneous nature of the recorded data, nodes of the clinical network, Fig. 5, can be partitioned into groups associated to specific dementia disorders and patient clinical context, including comorbidities (diabetes, hypertension, etc) and related comedications. See (Cabeli et al. 2020) for additional information on each variable and further discussion.

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 (in)dependencies between continuous or

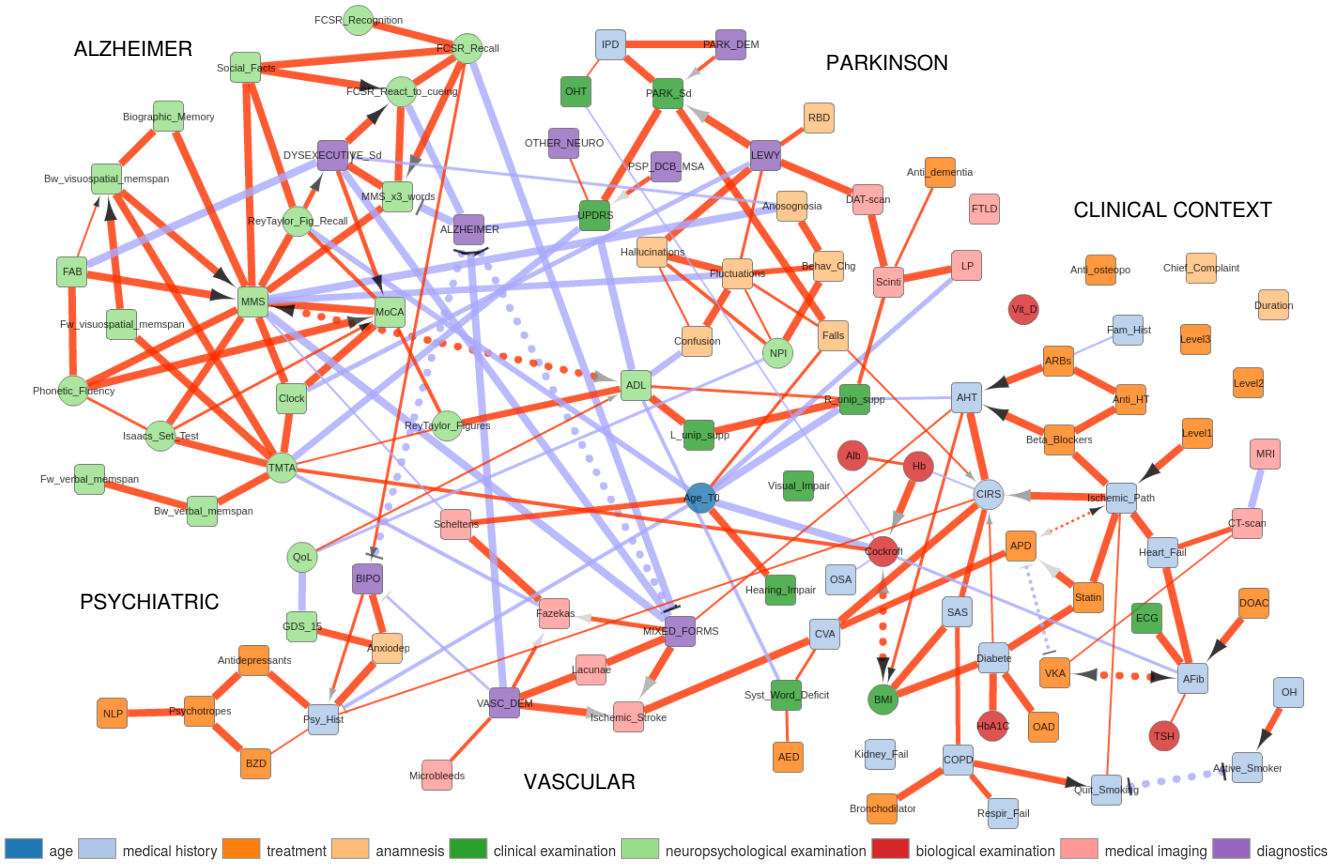

Figure 5: **Network reconstructed from medical records of 1,628 eldelry patients with cognitive disorders**. Square (resp. circle) nodes correspond to discrete (resp. continuous) variables. Red (resp. blue) edges correspond to correlation (resp. anticorrelation) between variables. Dotted edges reflect latent variables, see Discussion and (Cabeli et al. 2020) for additional information.

mixed-type variables. First, obtaining optimal discretization helps explain the dependencies 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. 5) 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" (Pearl 2009) 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 in (Cabeli et al. 2020) are PARK_DEM $\rightarrow$ PARK_Sd $\leftarrow$ LEWY, VASC_DEM $\rightarrow$ Fazekas $\leftarrow$ MIXED_FORMS and VASC_DEM $\rightarrow$ Ischemic_Stroke $\leftarrow$ 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 $\dashv$ 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$\leftarrow\!\!-\!\!-$ *diagnosis* $-\!\!-\!\!\rightarrow$BIPO, or, ALZHEIMER$\leftarrow\!\!-\!\!-$ *diagnosis* $-\!\!-\!\!\rightarrow$MIXED_FORMS. Latent variables may also represent the clinician's decisions between alternative treatments, *e.g.*, APD$\leftarrow\!\!-\!\!-$ *clinician_decision* $-\!\!-\!\!\rightarrow$VKA or a nonrecorded or implicit information in the patient personal or medical history, *e.g.*, active_smoker$\leftarrow\!\!-\!\!-$ *ever_smoked* $-\!\!-\!\!\rightarrow$quit_smoking, Fig. 5.

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. 5, 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.

Acknowledgements

VC was supported by a PhD fellowship from IRIS data science program of PSL and by the ARC foundation, and HI by the Labex Cell(n)Scale.

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 UAI 2015 Conference on Advances in Causal Inference-Volume 1504*, 1–29. CEUR-WS. org.

Affeldt, S.; Verny, L.; and Isambert, H. 2016. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. In *BMC bioinformatics*, volume 17, 12. BioMed Central Ltd.

Azadkia, M.; and Chatterjee, S. 2021. A simple measure of conditional dependence. *arXiv:1910.12327 [cs, math, stat]* URL http://arxiv.org/abs/1910.12327. ArXiv: 1910.12327.

Belghazi, I.; Rajeswar, S.; Baratin, A.; Hjelm, R. D.; and Courville, A. 2018. MINE: mutual information neural estimation. *arXiv preprint arXiv:1801.04062* .

Boeken, P. A.; and Mooij, J. M. 2020. A bayesian nonparametric conditional two-sample test with an application to local causal discovery. *arXiv preprint arXiv:2008.07382* .

Bühlmann, P.; Peters, J.; and Ernest, J. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* 42(6): 2526–2556. ISSN 0090-5364. doi:10.1214/14-AOS1260. URL http://arxiv.org/abs/1310.1533. ArXiv: 1310.1533.

Cabeli, V.; Verny, L.; Sella, N.; Uguzzoni, G.; Verny, M.; and Isambert, H. 2020. Learning clinical networks from medical recordsbased on information estimates in mixed-type data. *PLoS Computational Biology* .

Cover, T. M.; and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Gao, W.; Kannan, S.; Oh, S.; and Viswanath, P. 2017. Estimating Mutual Information for Discrete-Continuous Mixtures. *arXiv:1709.06212 [cs, math]* URL http://arxiv.org/abs/1709.06212. ArXiv: 1709.06212.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics* 10. ISSN 1664-8021. doi:10.3389/fgene.2019.00524. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00524/full. Publisher: Frontiers.

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

Gretton, A.; Spirtes, P.; and Tillman, R. E. 2009. Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in neural information processing systems*, 1847–1855.

Harris, N.; and Drton, M. 2013. PC Algorithm for Nonparanormal Graphical Models. *Journal of Machine Learning Research* 14: 3365–3383. URL http://www.jmlr.org/papers/v14/harris13a.html.

Heckerman, D.; Meek, C.; and Cooper, G. 2006. A Bayesian Approach to Causal Discovery. In *Innovations in Machine Learning*, 1–28. Springer, Berlin, Heidelberg. ISBN 978-3-540-30609-2 978-3-540-33486-6. doi:10.1007/3-540-33486-6_1. URL https://link.springer.com/chapter/10.1007/3-540-33486-6_1.

Irrthum, A.; Wehenkel, L.; and Geurts, P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PloS one* 5(9): e12776.

Kalisch, M.; Hauser, A.; Maechler, M.; Colombo, D.; Entner, D.; Hoyer, P.; Hyttinen, A.; Peters, J.; Andri, N.; and Perkovic, E. 2017. Package 'pcalg' .

Kim, S. 2015. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods* 22(6): 665.

Kontkanen, P. 2009. Computationally efficient methods for MDL-optimal density estimation and data clustering .

Kontkanen, P.; Buntine, W.; Myllymäki, P.; Rissanen, J.; and Tirri, H. 2003. Efficient computation of stochastic complexity. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*. Citeseer.

Kontkanen, P.; and Myllymäki, P. 2007a. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters* 103(6): 227–233.

Kontkanen, P.; and Myllymäki, P. 2007b. MDL histogram density estimation. In *Artificial Intelligence and Statistics*, 219–226.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical review E* 69(6): 066138.

Lizier, J. T. 2014. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI* 1: 11.

Marx, A.; and Vreeken, J. 2019. Testing Conditional Independence on Discrete Data using Stochastic Complexity. *arXiv preprint arXiv:1903.04829* .

Meinshausen, N.; and Bühlmann, P. 2006. High-dimensional graphs and variable selection with the lasso. *The annals of statistics* 34(3): 1436–1462.

Melancon, G.; and Philippe, F. 2004. Generating connected acyclic digraphs uniformly at random. *arXiv:cs/0403040* URL http://arxiv.org/abs/cs/0403040. ArXiv: cs/0403040.

Mooij, J. M.; Magliacane, S.; and Claassen, T. 2016. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351* .

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: a primer*. John Wiley & Sons.

Pfister, N.; Bühlmann, P.; Schölkopf, B.; and Peters, J. 2016. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Roos, T.; Silander, T.; Kontkanen, P.; and Myllymaki, P. 2008. Bayesian network structure learning using factorized NML universal models. In *2008 Information Theory and Applications Workshop*, 272–276. IEEE.

Ross, B. C. 2014. Mutual Information between Discrete and Continuous Data Sets. *PLOS ONE* 9(2): e87357. ISSN 1932-6203. doi:10.1371/journal.pone.0087357. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0087357.

Runge, J. 2017. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. *arXiv preprint arXiv:1709.01447* .

Runge, J. 2018. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, 938–947. URL http://proceedings.mlr.press/v84/runge18a.html.

Scutari, M.; Scutari, M. M.; and MMPC, H.-P. 2017. *Package 'bnlearn'*.

Sedgewick, A. J.; Buschur, K.; Shi, I.; Ramsey, J. D.; Raghu, V. K.; Manatakis, D. V.; Zhang, Y.; Bon, J.; Chandra, D.; and Karoleski, C. 2018. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics* .

Sella, N.; Verny, L.; Uguzzoni, G.; Affeldt, S.; and Isambert, H. 2017. MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. *Bioinformatics* .

Shah, R. D.; and Peters, J. 2020. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics* 48(3). ISSN 0090-5364. doi:10.1214/19-AOS1857. URL http://arxiv.org/abs/1804.07203. ArXiv: 1804.07203.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7(Oct): 2003–2030.

Spirtes, P.; and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review* 9(1): 62–72.

Spirtes, P.; Glymour, C.; Scheines, R.; Kauffman, S.; Aimale, V.; and Wimberly, F. 2000. Constructing Bayesian network models of gene expression networks from microarray data .

Strobl, E. V.; Zhang, K.; and Visweswaran, S. 2019. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference* 7(1).

Szpankowski, W. 2011. *Average case analysis of algorithms on sequences*, volume 50. John Wiley & Sons.

Tsagris, M.; Borboudakis, G.; Lagani, V.; and Tsamardinos, I. 2018. Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics* 1–12.

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 Computational Biology* 13(10): e1005662.

Zeng, X.; Xia, Y.; and Tong, H. 2018. Jackknife approach to the estimation of mutual information. *Proceedings of the National Academy of Sciences* 115(40): 9956–9961.

Zhao, J.; Zhou, Y.; Zhang, X.; and Chen, L. 2016. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences* 113(18): 5130–5135. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1522586113. URL http://www.pnas.org/content/113/18/5130.

Zheng, X.; Aragam, B.; Ravikumar, P.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422* .

Zheng, X.; Dan, C.; Aragam, B.; Ravikumar, P.; and Xing, E. 2020. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, 3414–3425. PMLR.

Supplementary information

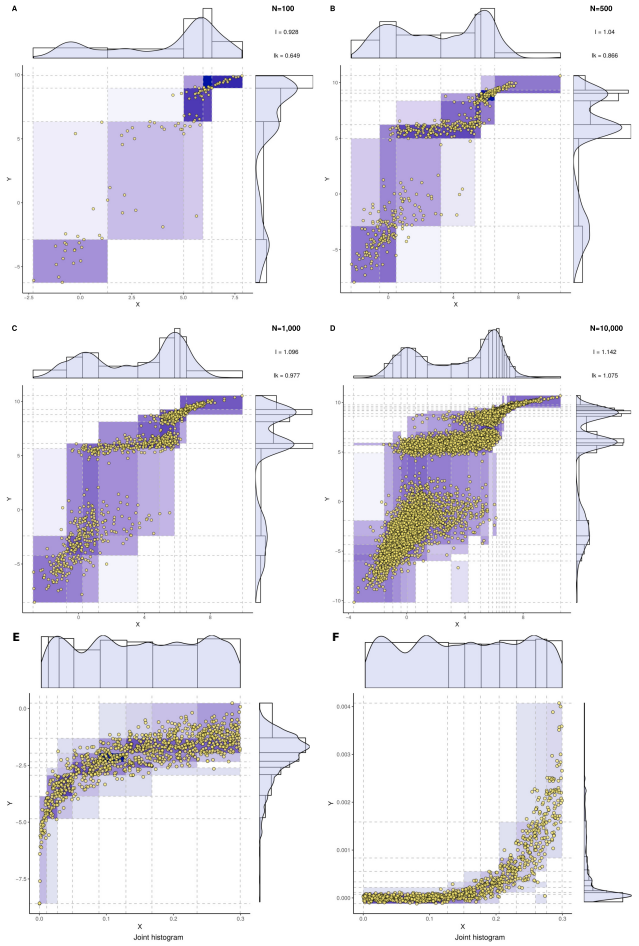

Figure S1: **Interaction-dependent optimum discretization**. **A - D** The proposed information-maximizing discretization scheme is illustrated for a joint distribution defined as a Gumbel bivariate copula with parameter $\theta = 5$ and univariate marginal-distribution functions chosen as Gaussian mixtures 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 N with corresponding mutual information estimates. **E - F** 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 dependencies for Y. **(E)** Y is defined as $\log(X) + \epsilon_1$, and **(F)** 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, $I(X; Y) \simeq 0.75$.

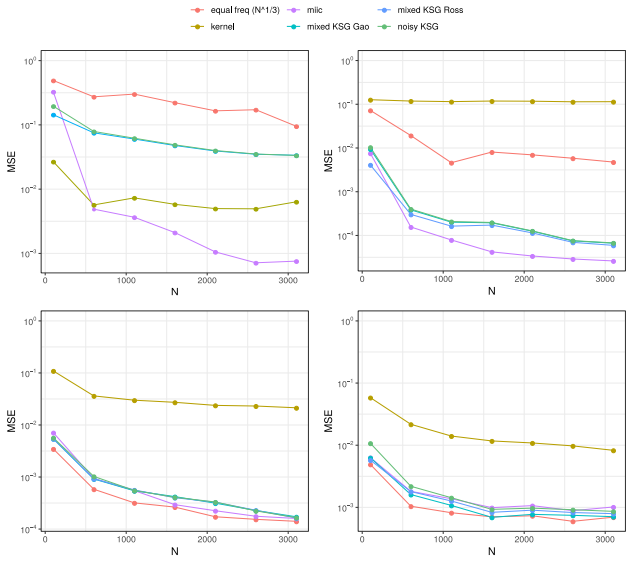

Figure S2: **Mutual information estimation of mixed variables.** Experiment set-ups and analytical values for the mutual information were taken fom (Gao et al. 2017) and 50 runs were performed for each sample size N. Our proposed approach is compared to a naive equal-frequency discretization with $N^{1/3}$ bins, a kernel and a noisy KSG estimator as implemented in JIDT (Lizier 2014), as well as the recent KSG extensions for estimating the mutual informmation between a categorical and a continuous variable (mixed KSG Ross (Ross 2014)), and between mixed-type variables (mixed KSG Gao (Gao et al. 2017)). 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 (Gao et al. 2017). Code is available at https://github.com/vcabeli/miic_PLoS

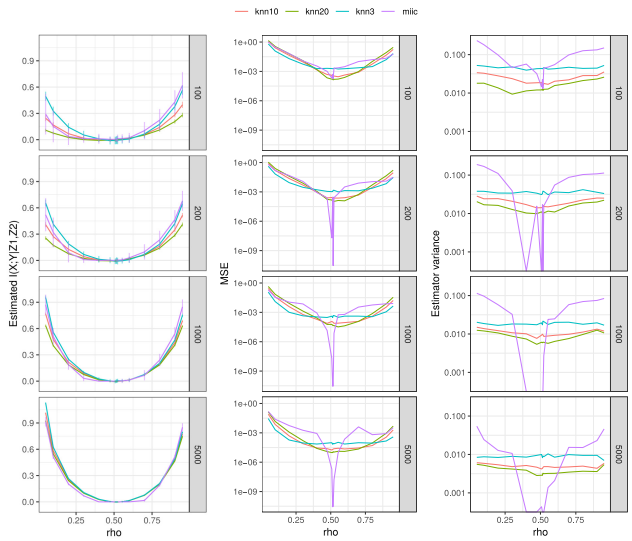

Figure S3: **Conditional mutual information estimation for multivariate Gaussian distributions**. Four-dimensional normal distributions $P(X, Y, Z_1, Z_2)$ were sampled for $N = 100$ to $5,000$ samples 100 times for each correlation coefficient $\rho = \rho_{XY}$, chosen between 0.05 and 0.95. The other 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 as with kNN conditional information estimates as in Fig. 2. 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 $\rho_{XY\cdot Z_1Z_2} = \overline{\mathbf{\Sigma}}_{12}/\sqrt{\overline{\mathbf{\Sigma}}_{11} * \overline{\mathbf{\Sigma}}_{22}}$, and the analytical conditional mutual information for a multivariate normal distribution is given by $I(X;Y|Z_1,Z_2) = -\log(1-\rho^2_{XY\cdot Z_1Z_2})/2$.

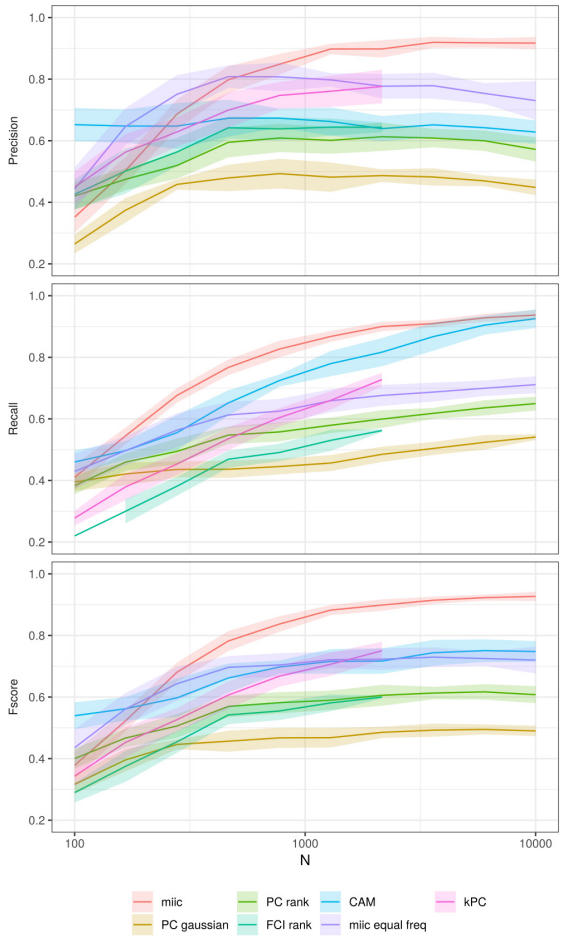

Figure S4: **CPDAG assessment of benchmark networks for continuous, non-linear, non-Gaussian datasets**. Skeleton Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100 - 10,000$ samples. Results obtained with our parameter-free 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 (Kalisch et al. 2017), kPC with the Helbert-Schmidt Independence Criterion (Gretton et al. 2005; Gretton, Spirtes, and Tillman 2009) and CAM (Bühlmann, Peters, and Ernest 2014) algorithms, after optimizing their respective parameter (α) for each sample size N.