SUPPORTING INFORMATION

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 [7] and MXM [8] 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` [2,3]. 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 [4,5]) 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. 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.

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

4. 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.

5. 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.

6. 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. doi:10.1093/bioinformatics/bts707.

7. 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.

8. 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.

9. 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. doi:10.1126/science.1205438.

10. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences. 2014;111(9):3354–3359.

11. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Wiley; 2006.

12. 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.

13. Lizier JT. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. Frontiers in Robotics and AI. 2014;1:11.

14. Ross BC. Mutual information between discrete and continuous data sets. PloS one. 2014;9(2):e87357.

15. 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. doi:10.18637/jss.v047.i11.

16. 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. doi:10.1214/14-aos1260.

Supplementary Figures

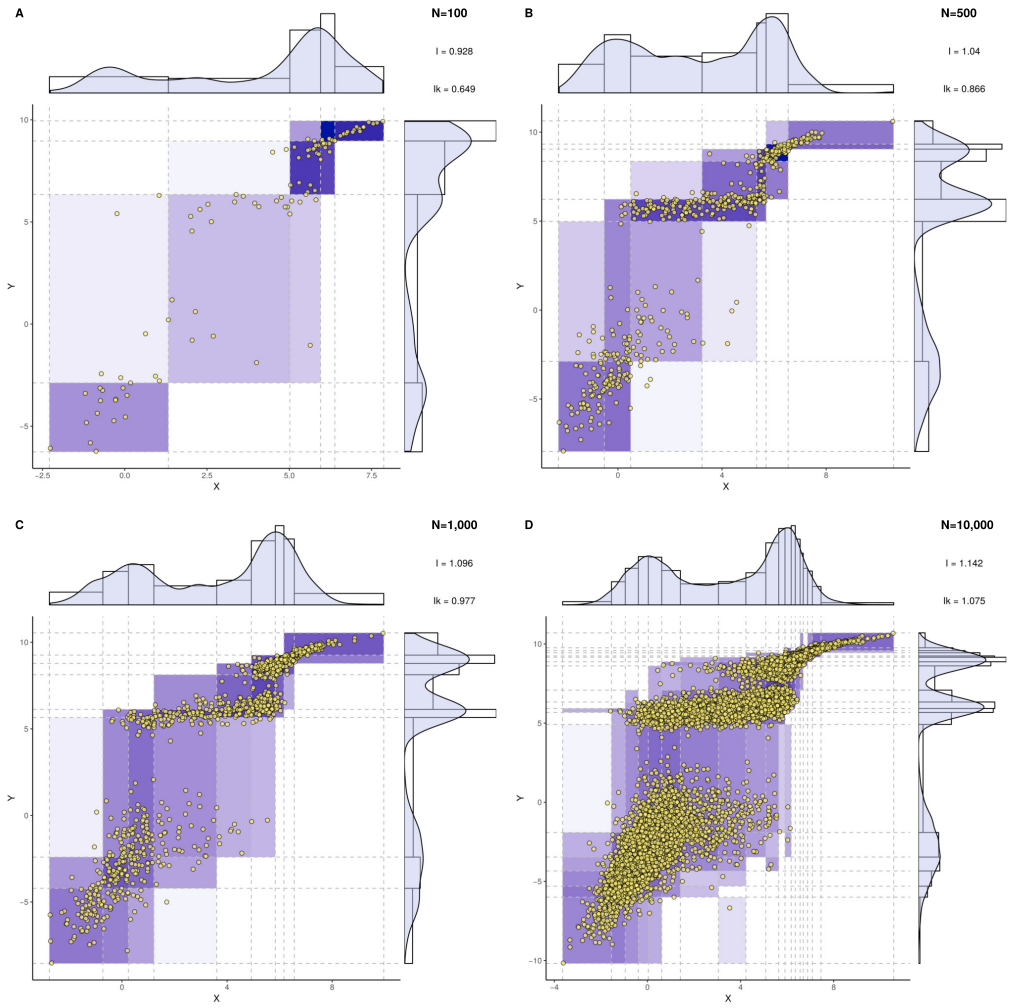

Fig S1. Optimum bivariate discretization for mutual information estimation. The proposed information-maximizing discretization scheme is illustrated for a joint distribution defined as a Gumbel bivariate copula with parameter $\theta = 5$ and univariate marginal-distribution functions chosen as Gaussian 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 integration of the marginals and the joint probability distribution and yields, $I(X;Y) = 1.205$, in good agreement with the obtained estimates for large N.

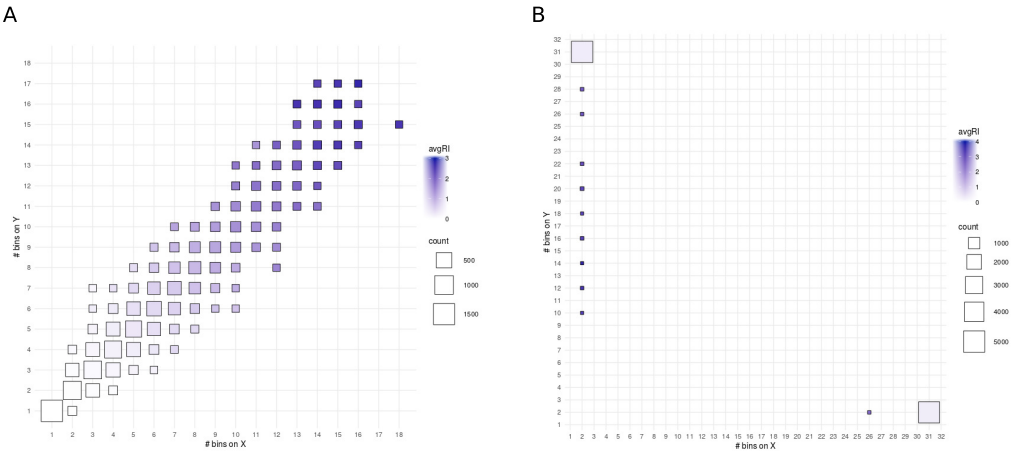

Fig S2. Adaptive information-maximizing partitions depending on interaction strength. To assess the range in bin numbers depending on the strength of interaction between variables, we generated $N = 1,000$ independent samples for 10,000 Gaussian bivariate distributions with a uniformly distributed correlation coefficient ρ in $[-1, 1]$. The real mutual information (RI) of Gaussian bivariate distributions can be computed directly [11], 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 [9] using the `minepy` package [6]. **(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 [9] 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 [10].

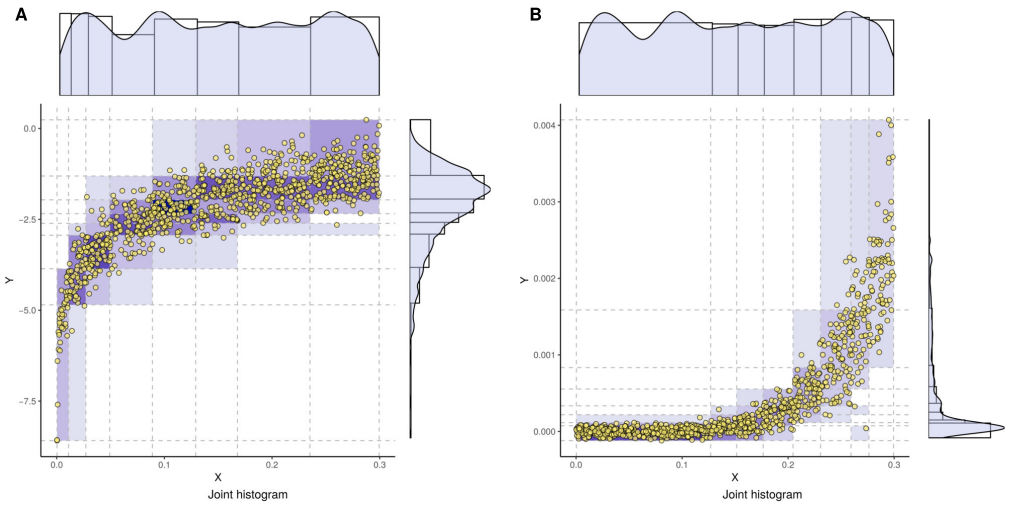

Fig S3. Interaction-dependent optimum discretization. Optimum bivariate partitions obtained from $N = 1,000$ samples of two different joint distributions $P(X, Y)$ sharing the same sampling of X taken from a uniform distribution on $[0, 0.3]$, but with different dependences for Y. **(A)** Y is defined 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 $Y \simeq X^5$.

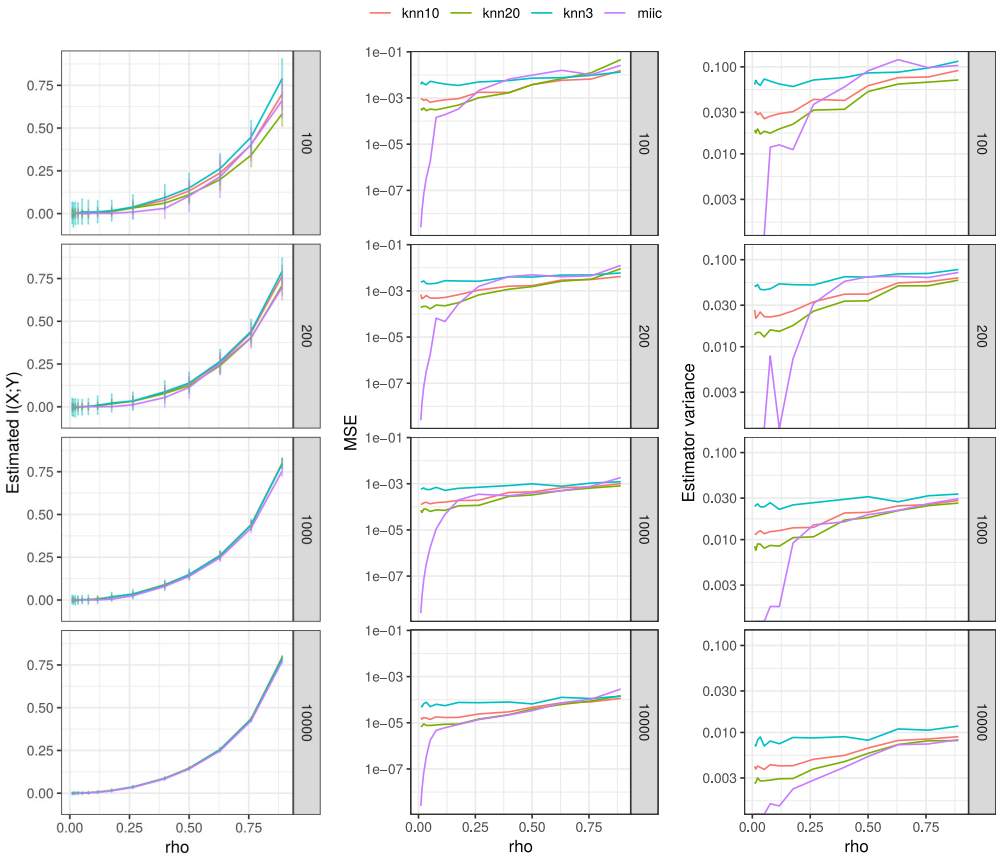

Fig S4. Mutual information estimation for Gaussian bivariate distributions. 100 bivariate normal distributions were sampled for varying sample sizes, increasing from top to bottom, and correlation coefficients ρ ranging from 0.01 to 0.9. The mutual information was estimated with the proposed optimum 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).

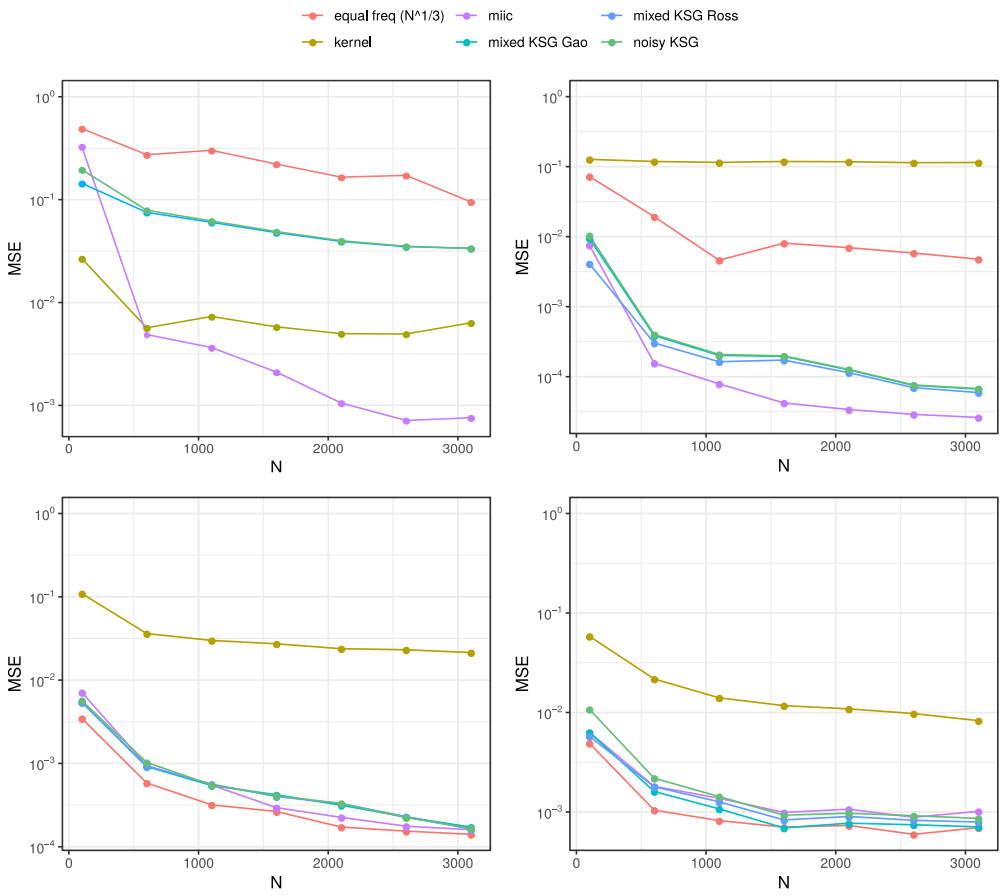

Fig S5. Mutual information estimation of mixed variables. Experiment set-ups and analytical values for the mutual information were taken fom [12] and 50 runs were performed for each sample size N. Our proposed approach is compared to a naive equal-frequency discretization with $N^{1/3}$ bins, a kernel and a noisy KSG estimator as implemented in JIDT [13], as well as the recent KSG extensions for estimating the mutual information between a categorical and a continuous variable (mixed KSG Ross [14]), and between mixed-type variables (mixed KSG Gao [12]). 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 [12].

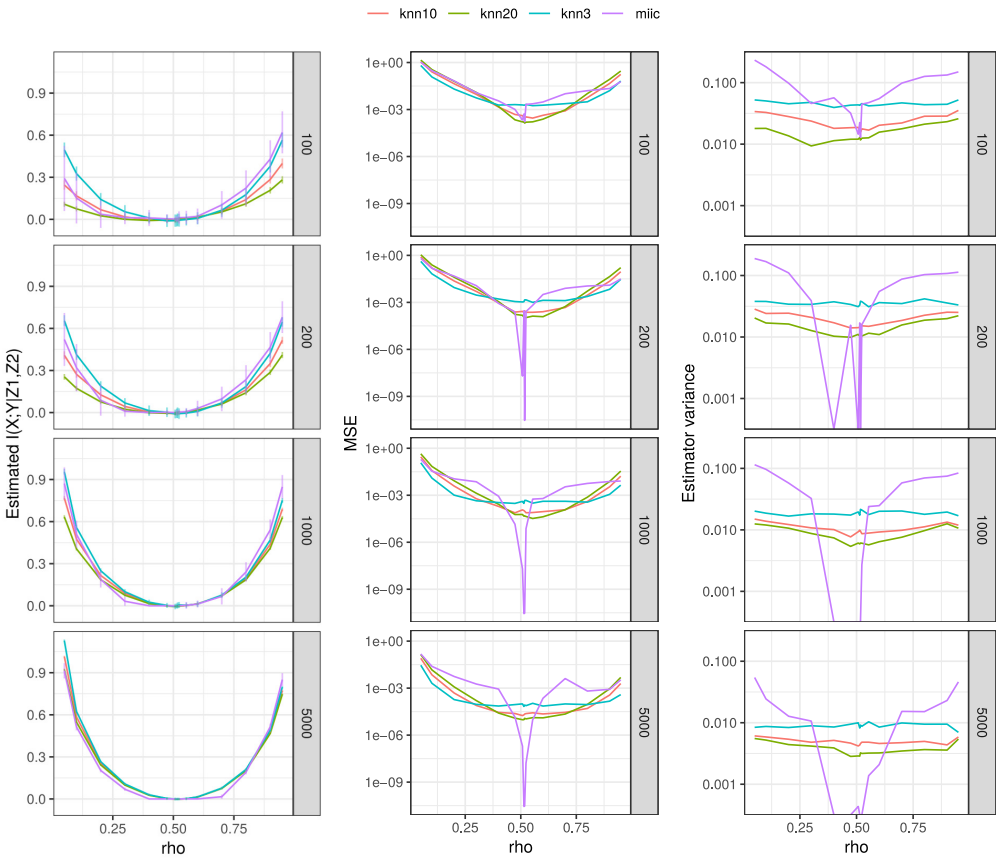

Fig S6. Conditional mutual information estimation for multivariate Gaussian distributions. Four-dimensional normal distributions $P(X, Y, Z_1, Z_2)$ were sampled for $N = 100$ to $5,000$ samples 100 times for each correlation coefficient $\rho = \rho_{XY}$, chosen between 0.05 and 0.95. The other pairwise 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. S4. 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 $I(X; Y|Z_1, Z_2) = -\log(1 - \rho^2_{XY \cdot Z_1Z_2})/2$.

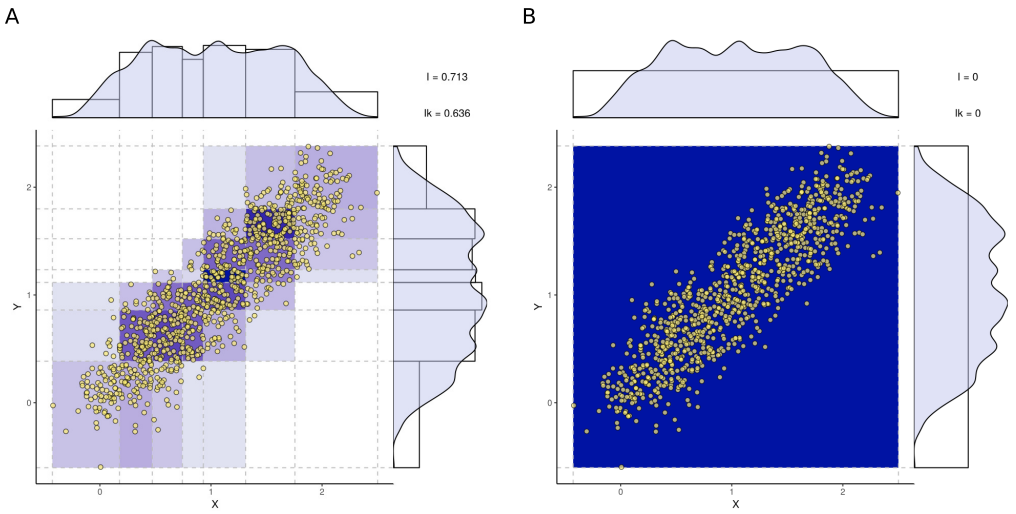

Fig S7. Pairwise dependence and conditional independence between X and Y sharing a common cause Z. This example illustrates the (conditional) correlation patterns emerging from the presence of a confounding variable, as depicted by the causal diagram $X \leftarrow Z \rightarrow Y$. Z is generated with 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.

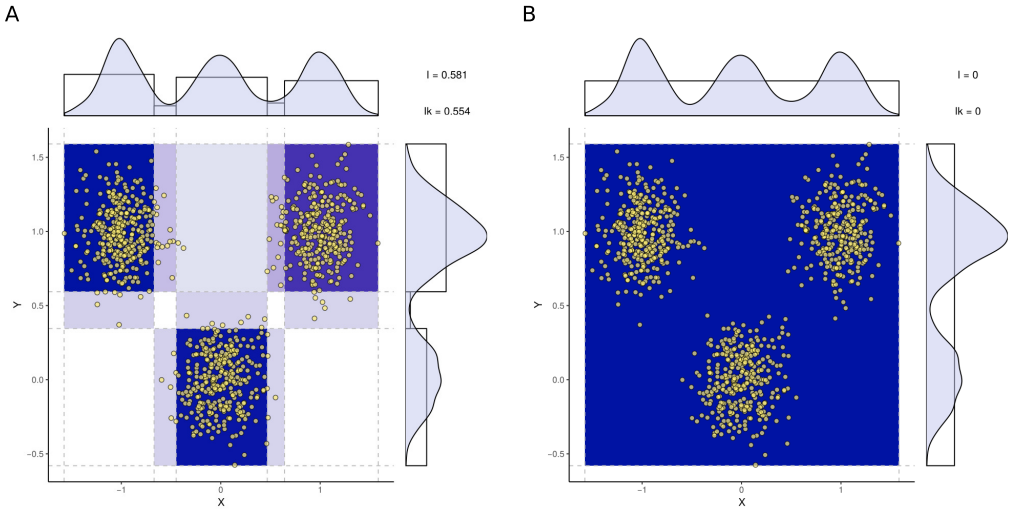

Fig S8. Pairwise dependence and conditional independence between non Gaussian X and Y sharing a common categorical cause. Another confounding example, $X \leftarrow Z \rightarrow Y$, taken from [8] with a uniform categorical Z with three levels, X and Y being continuous, for $N = 1,000$ observations. 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 Fig. S7, there is some non-zero mutual information **(A)** between X and Y corresponding to an optimum discretization, while the conditional mutual information **(B)** vanishes when conditioning on the categorial common cause, Z.

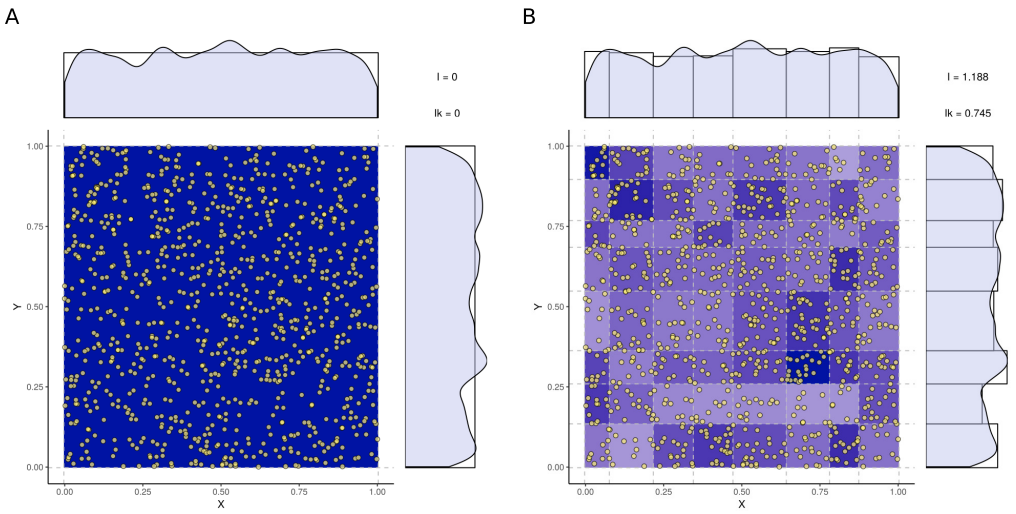

Fig S9. Pairwise independence and conditional dependence with a v-structure. Example of two independent variables X, Y both causing a third variable Z as: $X \rightarrow Z \leftarrow Y$. $N = 1,000$ observations are drawn for $X, Y \sim \mathcal{N}(0, 1)$ and $Z = X + Y + \epsilon$. **(A)** The two variables 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. **(B)** However, conditioning on the common effect Z 'activates' the v-structure path generating some information between X and Y. This is reflected in the fact that the induced interaction between X and Y conditioned on Z requires a multiple bin optimum discretization to estimate $I_N(X; Y|Z) = 1.188$ (with $I'_N(X; Y|Z) = 0.745$).

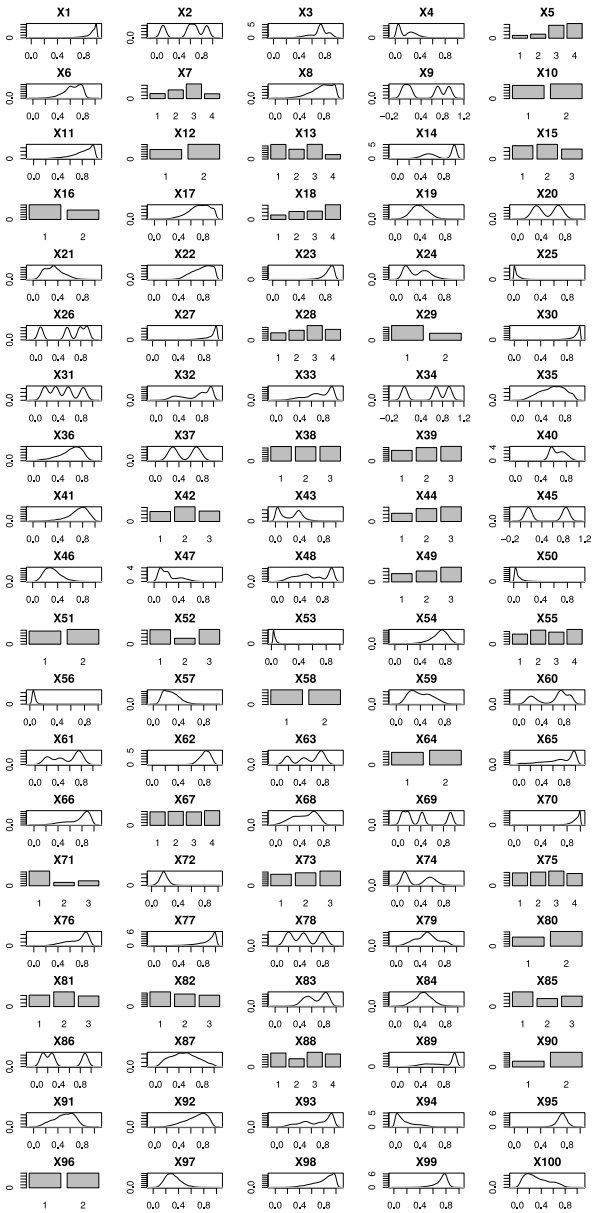

Fig S10. Example of dataset generated for mixed-type, non-linear, non-Gaussian benchmarking with 69 continuous and 31 categorical variables. Each plot represents the observed density or histogram ($N = 1,000$) of the continuous or categorical variable X_i, constructed by structural equation models given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

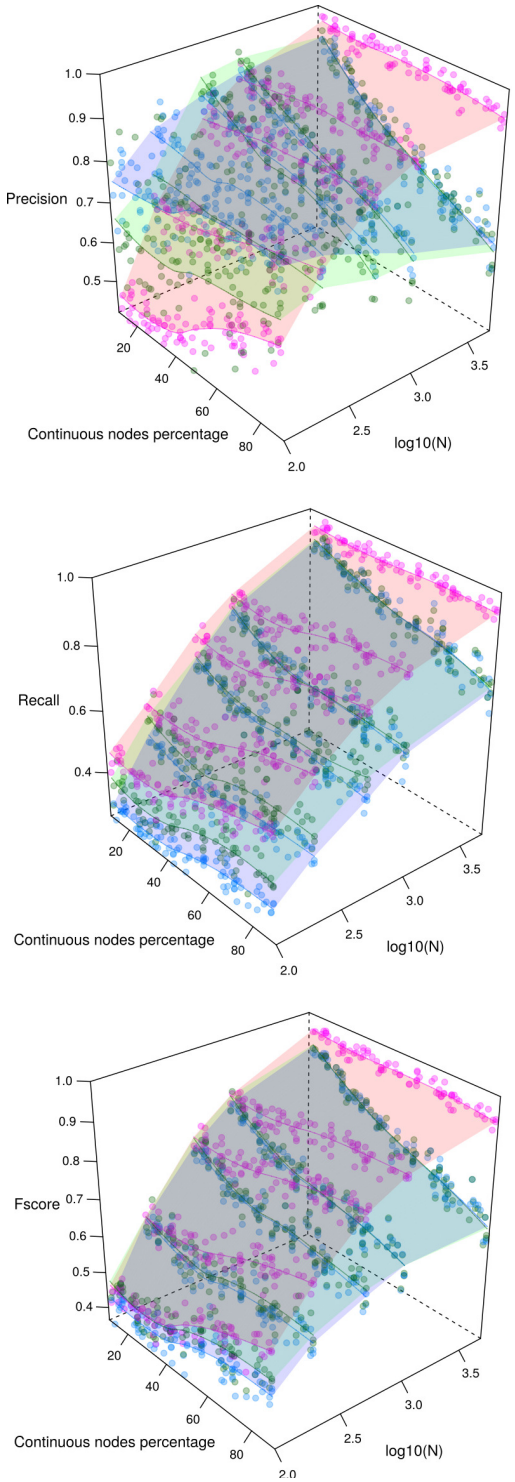

Fig S11. Skeleton assessment of benchmark networks for mixed-type, non-linear, non-Gaussian datasets. Skeleton Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from N=100–5,000 samples (see histogram example Fig. S10). Performances 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 [7] (blue) and MXM [8] (green). See Supporting Information.

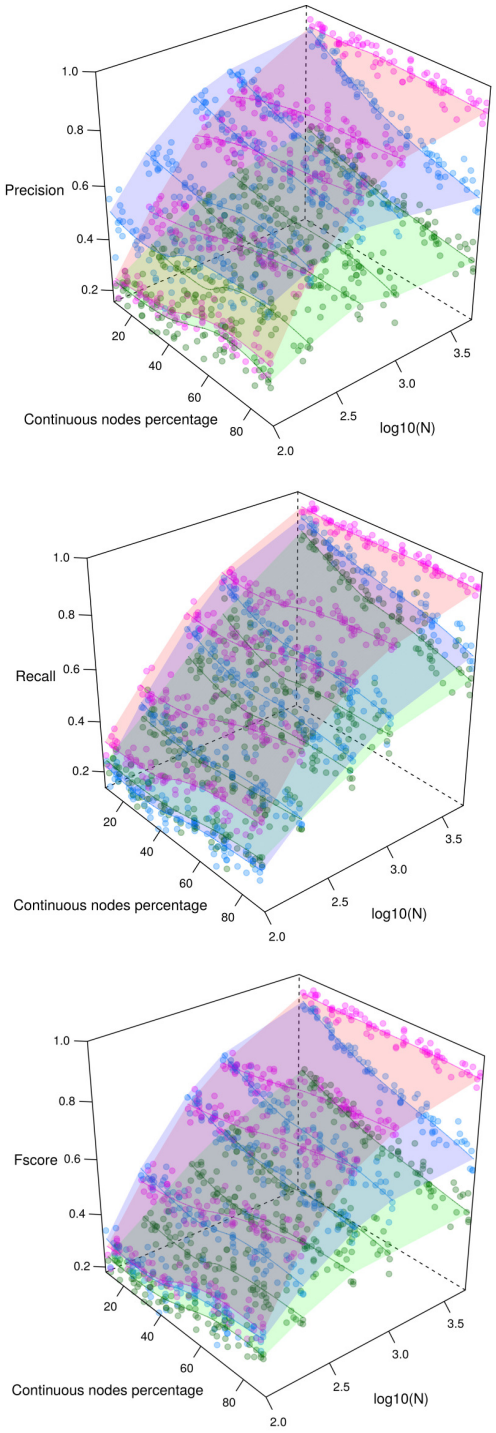

Fig S12. CPDAG assessment of benchmark networks for mixed-type, non-linear, non-Gaussian datasets. CPDAG Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from N=100–5,000 samples (see histogram example Fig. S10). Performances obtained 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 [7] (blue) and MXM [8] (green). See Supporting Information.

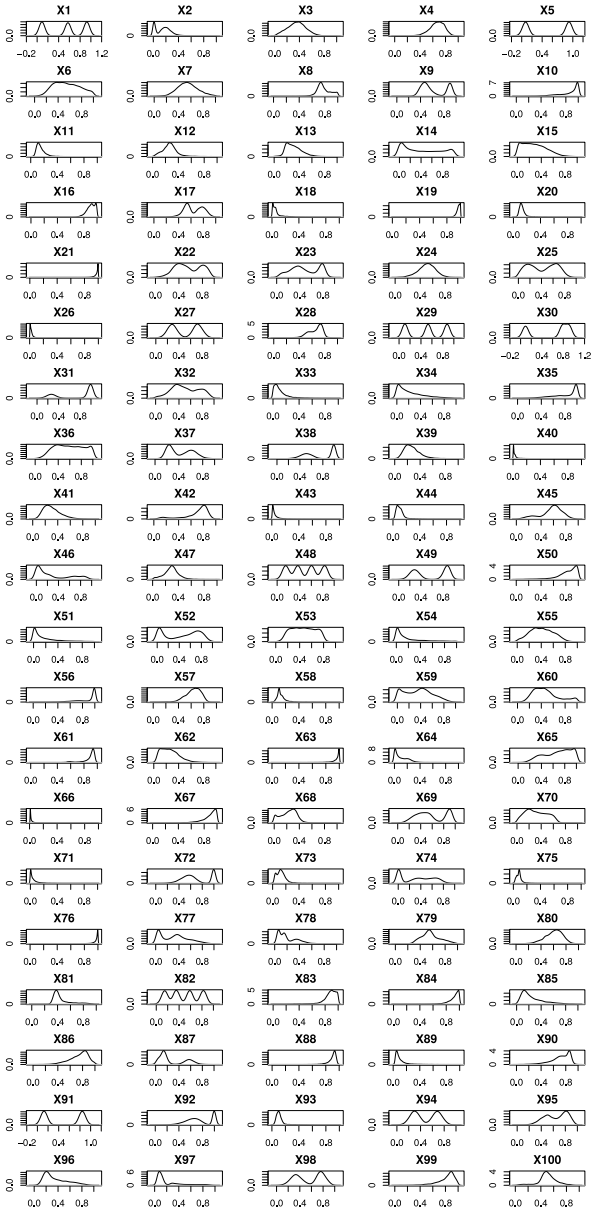

Fig S13. Example of dataset used for continuous, non-linear, non-Gaussian benchmarking with 100 continuous variables.

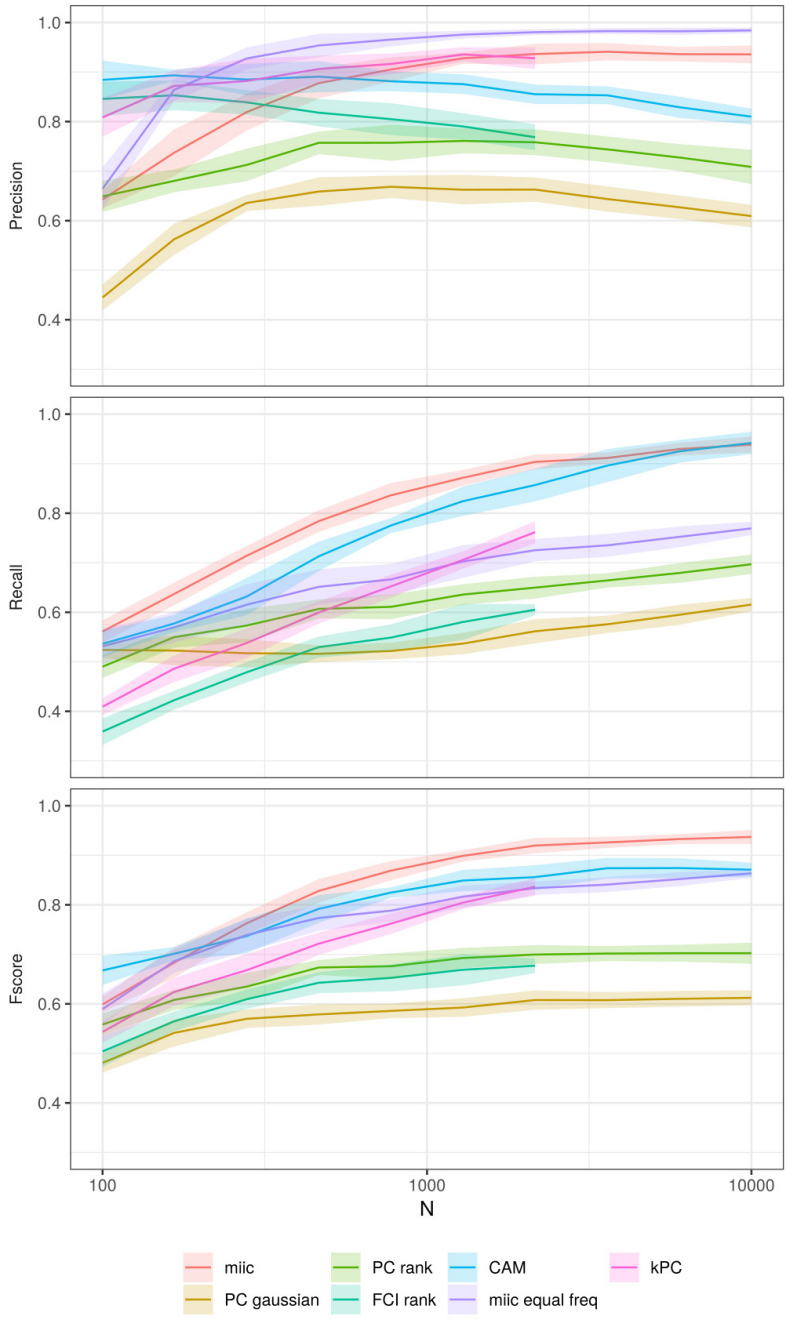

Fig S14. Skeleton assessment of benchmark networks for continuous, non-linear, non-Gaussian datasets. Skeleton Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100 - 10,000$ samples (see histogram example Fig. S13). Results 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 [15], kPC with the Helbert-Schmidt Independence Criterion [4,5] and CAM [16] algorithms, after optimizing their respective parameter (α) for each sample size N. See Supporting Information.

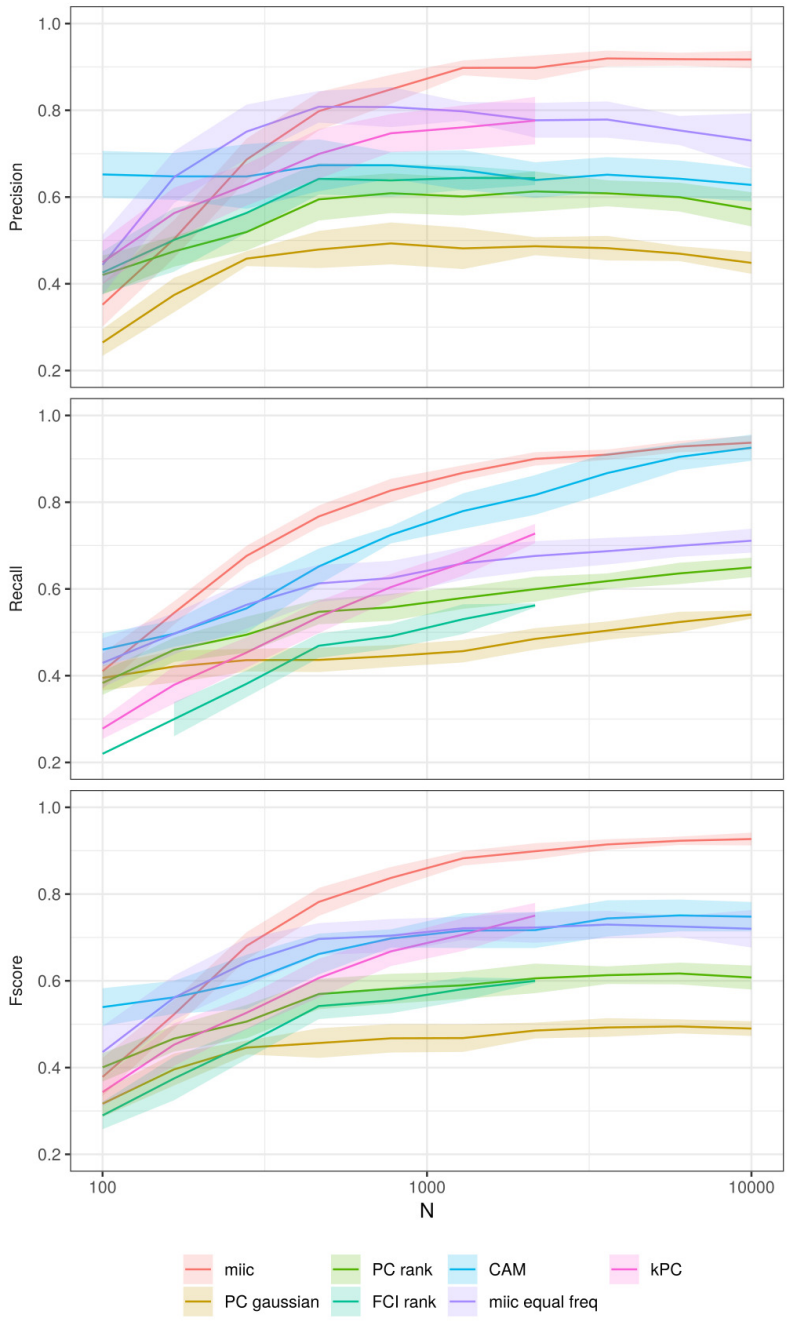

Fig S15. CPDAG assessment of benchmark networks for continuous, non-linear, non-Gaussian datasets. CPDAG Precision, Recall and F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from $N = 100 - 10,000$ samples (same simulation settings as in Fig. S14).