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

Vincent Cabeli, Honghao Li, Marcel da Câmara Ribeiro-Dantas, Franck Simon, Hervé Isambert
Institut Curie, Université PSL, Sorbonne Université,
CNRS UMR168, 75005 Paris, France
first-name.last-name@curie.fr

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

WHY-21 @ 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

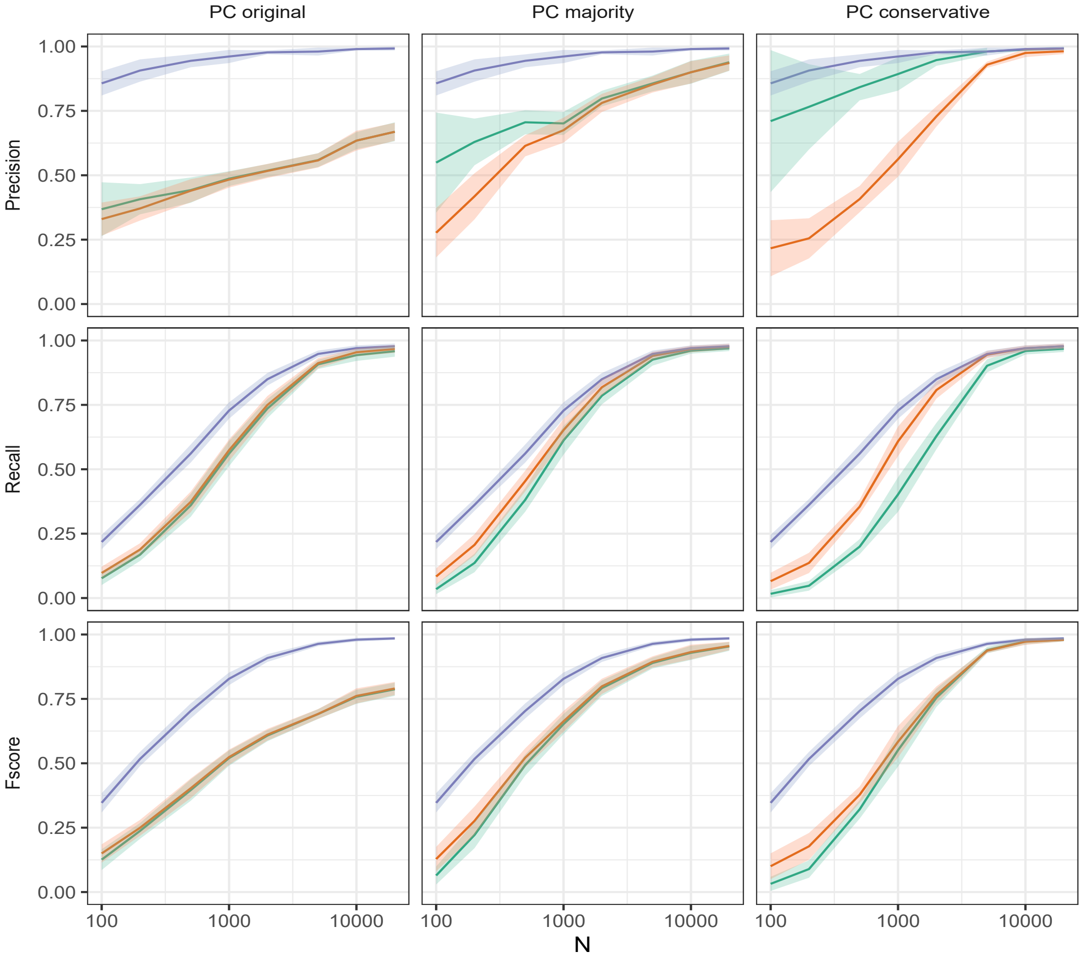

Figure 1: **PC original, majority and conservative orientation rules on discrete datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 3.8 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured 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.

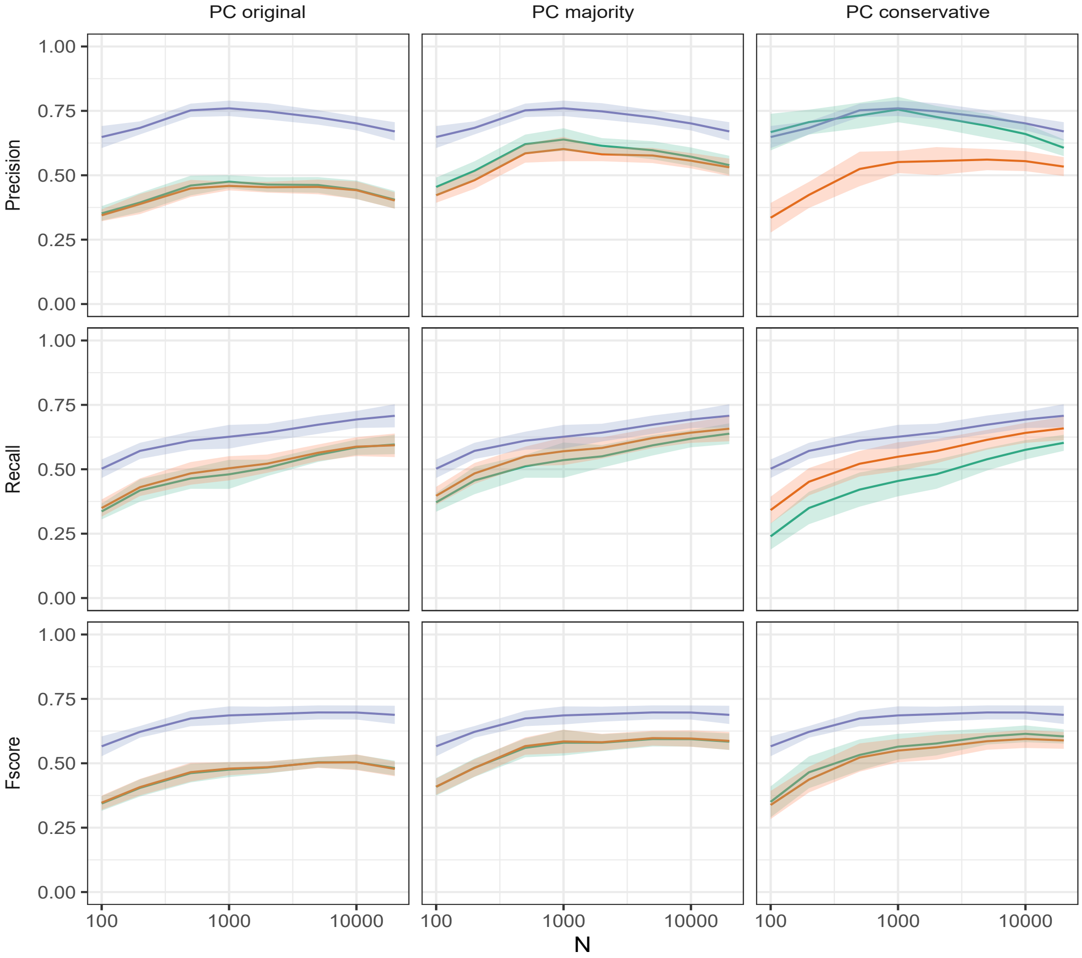

Figure 2: **PC original, majority and conservative orientation rules on continuous datasets**. Benchmark datasets are generated from random 100-node DAGs with average degree 3.8 and maximum degree 4 (See Data generation and benchmarks section for details). PC structure learning performance is measured 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, $k_N^{\text{NML}}(X;Y|\{A_i\})/N$, 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,

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

where $k_N^{\text{NML}}(X;Y|\{A_i\})$ 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 $k_N^{\text{NML}}(X;Y|\{A_i\}) = 0$ 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, $X \rightarrow Z \leftarrow Y$, 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 $X \rightarrow Z - Y$ as $X \rightarrow Z \rightarrow Y$ (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, $X \rightarrow Z \leftarrow Y$, 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, $X \rightarrow Z \leftarrow Y$.

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, $X - Z - Y$.

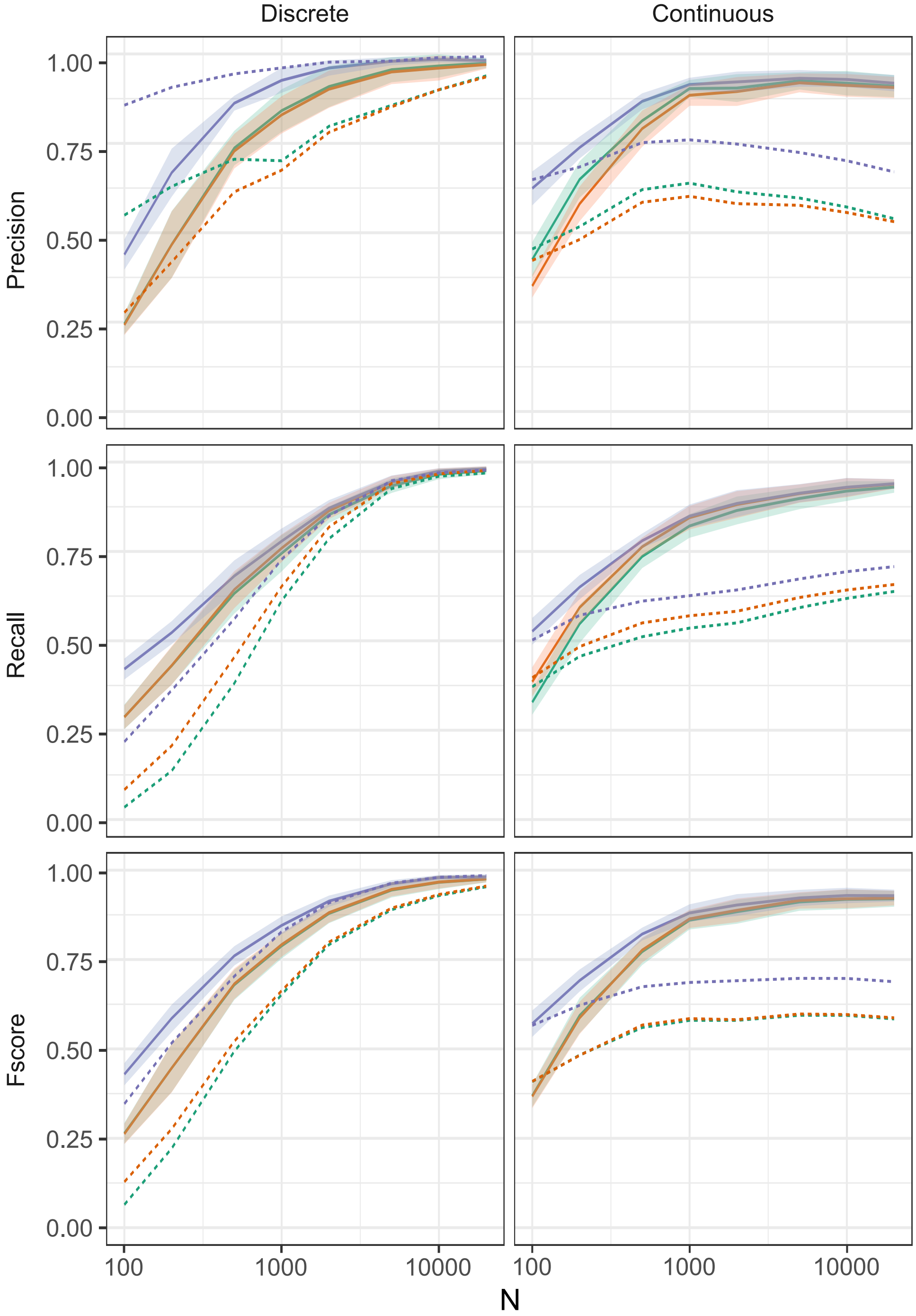

Figure 3: **Original MIIC with orientation rules allowing for negative NML-regularized MI & CMI on discrete data (left) and negative NML-regularized CMI on continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. MIIC structure learning performance is measured in terms of Precision, 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 $k_N^{\text{NML}}(X;Y) = 0$ 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 3.8. Further details and dataset examples can be found in Cabeli et al. (2020).

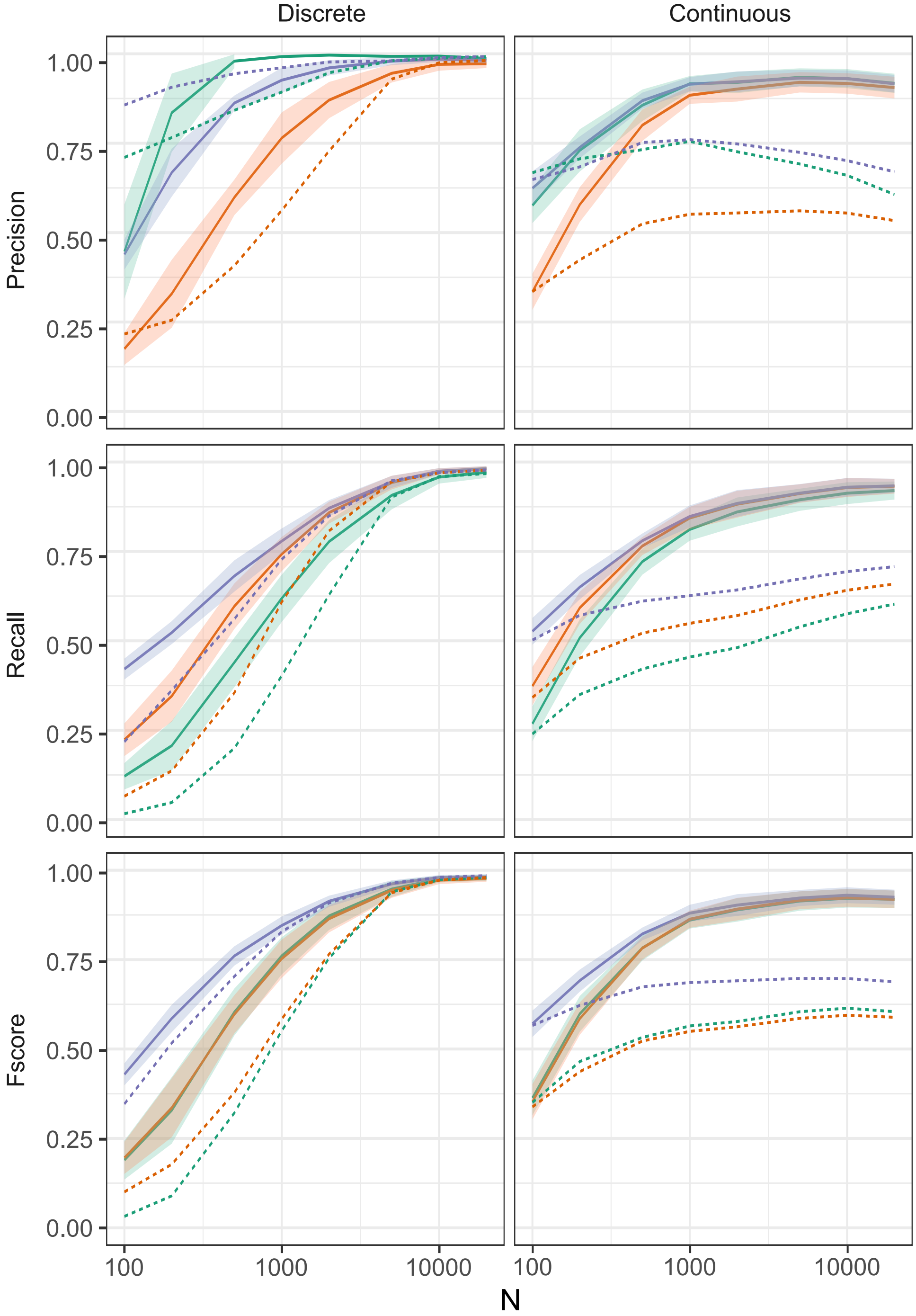

Figure 4: **Conservative MIIC with new orientation rules enforcing non-negative NML-regularized MI & CMI on discrete data (left) as well as continuous data (right)**. Benchmark datasets are the same as in Figs. 1 & 2. Conservative MIIC structure learning performance is measured in terms of Precision, 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.

5 Acknowledgements

We would like to acknowledge the supports of the following funding agencies: Fondation ARC (VC), French Ministery of Higher Education, Research and Innovation (HL), EU IC-3i cofund PhD programme (MCRD), INSERM ITMO Cancer (FS, HI), CNRS and Institut Curie (HI).

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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In main text and benchmark figures
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See main text
   (b) Did you include complete proofs of all theoretical results? [Yes] See proof
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The codes for data generation and benchmarks are accessible on github
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Data generation and benchmarks section
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figures
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Data generation and benchmarks section
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] (Cabeli et al., 2020)
   (b) Did you mention the license of the assets? [Yes] See Data generation and benchmarks section
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Data generation and benchmarks section
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]