Assessing Causal Graph Uncertainty and Optimized Multivariate Discretization Strategy

Lucie Kunitomo-Jacquin[1][0000-0001-8961-3417], Aurore Lomet[2][0000-0002-3264-8050], and Geoffrey Daniel[2][0000-0001-8049-1433]

[1] National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2] Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191, Gif-sur-Yvette, France.
kunitomo-jacquin.lucie@aist.go.jp
aurore.lomet@cea.fr
geoffrey.daniel@cea.fr

Abstract. This study investigates the complexities of constructing causal graphs within discretization constraints, focusing on the uncertainty of causal link presence resulting from discretizing continuous variables. Despite discretization's significant influence on causal graph estimation, it may be necessary, prompting the introduction of a robust index to quantify uncertainty under various discretization strategies. In real-world scenarios, the approach uses an evaluation index tailored to align with estimation characteristics, providing a practical assessment of uncertainty in causal discovery. By examining simulation data and using the true positive mean of causal link presence, the study assesses the impact of discretization. The proposed index offers a realistic evaluation of uncertainty in real-world studies without known truth. The overarching goal is to improve the accuracy and reliability of causal graph discovery by systematically assessing the impact of discretization on Type I error rates. The study explores different discretizations and their effects on independence tests and inferred causal structures.

Keywords: Causal discovery · Discretization · Uncertainty · Sensitivity

1 Introduction

Over the last twenty years, the notion of causality has been increasingly integrated into machine learning methods to go beyond correlation [18] [15] and provide explainability to predictive models [17]. One of the most common frameworks for describing causal mechanisms is Structural Causal Models (SCMs) [16], which consist of structural equations specifying the causal effects of each variable and a causal graph consisting of a causal interpretation of a Bayesian network [7]. Causal discovery relies on the estimation of a directed-acyclic graph (DAG) representing the causal relationships among the variables. This graph is noted $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes, i.e., the set of variables involved in the causal mechanism, and $\mathcal{E}$ denotes the set of edges representing the causal links between these variables. Many real-world problems lead to dealing with continuous variables, which requires the development of methods for constructing a causal graph that considers these continuous quantities. While some methods can deal directly with continuous variables (Gaussian hypothesis [9], some specific heterogeneous data contexts [8]), various cases require the development of a first step of discretization of continuous variables. For instance, with heterogeneous (discrete and continuous) data, [12] suggests discretizing continuous variables when most other variables are discrete, cautioning against potential negative effects on independence tests and inferred causal structures. In approaches using (conditional) independence tests, such as estimating Mutual Information [10],[5], a bin-based partitioning approach is employed. Various studies, like [13] and [11], underscore the importance of discretization in addressing challenges associated with conditional independence tests and mitigating bias introduced by latent variables.

Using a discretization approach, which is necessary in some real-world cases, brings the challenge of applying the most relevant discretization to build the causal graph. Inherent uncertainty on the causal link discoveries arises when applying different discretization procedures. Therefore, it is necessary to assess the relevance of a discretization in the objective to optimize it, with the constraint that, in real-world application, the ground truth is unknown.

To address this, our study introduces an evaluation index to assess the reliability of inferred causal relationships in real-world scenarios. The index imitates the behavior of the true positive mean of causal link presence, using p-values from conditional independence tests to evaluate inferred causal relationships within a system that offers a practical and realistic evaluation of the impact of discretization on causal discovery.

This index is beneficial for comparing multiple discretization strategies, helping identify the optimal approach for causal graph construction. The aim is to improve accuracy and reliability by selecting the most suitable strategy based on data characteristics and the specific causal graph, focusing on evaluating the impact of discretization on Type I error rates. We apply the index with an optimization procedure using simulated data, demonstrating its effectiveness in determining the best discretization of continuous variables.

2 Related Works

2.1 Constraints-based methods

The constraints-based methods perform causal discovery by exploiting the conditional independence relationships in the data. These methods make the so-called *Causal Faithfulness Assumption*, i.e., the reciprocal of Markov condition: if $A \perp\!\!\!\perp B|\mathcal{S}$ in D, i.e., A and B are independent given $\mathcal{S}$, then A and B are d-separated by $\mathcal{S}$ in G. With the Markov and *faithfulness* assumptions, we have a one-to-one correspondence between the d-separations in the graph and the conditional independences in the data distribution. In this paper, we focus our

Table 1: Confusion matrix

	$p''_{A-B} < \alpha$	$p''_{A-B} \geq \alpha$
$A - B$ truly present	true positive	false negative
$A - B$ truly absent	false positive	true negative

approach in Pearl context [16] with PC algorithm [19] that starts by estimating the non-oriented graph according to d-separation and then orienting the edges based on the identification of V-structures [3], assuming that there are no unmeasured common causes and no selection variables.

2.2 Edge-wise p-value

The problem of testing for the absence or presence of an edge in the skeleton has been analyzed by Strobl et al. [20], where it is demonstrated how to construct such a test. They address the inherent problem of the PC algorithm about the need for more confidence level information for each edge. For two nodes, A and B, the null hypothesis is defined by H_0: the link between A and B noted $A - B$, is absent, and its alternative H_1: $A - B$ is present. The authors assume a zero Type II error rate under the condition of faithfulness.

In this case, assuming a zero Type II error rate means the method is expected to miss no true causal of all true causal relationships in the data. However, it is essential to note that in practical scenarios, achieving a zero Type II error rate is usually unrealistic, and the authors acknowledge this by mentioning that the assumption may not hold in real-world situations. This work explores strategies to enhance reliability under this assumption and discusses potential heuristic criteria to reduce the Type II error rate in the context of discrete data.

In this context, Strobl shows we can bound the p-value associated to H_0 by the quantity $p'_{A-B} = \max_{i=1,\dots,q'} p_{A \perp\!\!\!\perp B | R_i}$ where $p_{A \perp\!\!\!\perp B | R_i}$ is the p-value returned by the independence test between A and B conditionally to R_i, $R_i \subseteq \{Pa(A) \backslash B\}$ or $R_i \subseteq \{Pa(B) \backslash A\}$ and q' denotes the total number of such subsets, and $Pa(A)$ is the set of parent nodes of A. For practical estimation, Strobl also demonstrates that we can further bound this quantity assuming zero type II error rate of the conditional independence test, in other words, assuming that conditional independence is always correctly rejected when it should. In this case, we get:

$$p'_{A-B} \leq \max_{i=1,\dots,q''} p_{A \perp\!\!\!\perp B | S_i} = p''_{A-B}, \tag{1}$$

where q'' denotes the number of conditional independence tests between A and B during PC algorithm estimation, and S_i the respective conditioning sets.

2.3 Multivariate Discretization

In the context of continuous variables, Bay [1] proposes a discretization strategy by partitioning into n intervals and then merging adjacent intervals with similar multivariate distributions. The author proposes a shift from univariate discretization, which considers a single variable at a time, to multivariate discretization to take into account interactions between variables before deciding on discretized intervals, potentially providing a more comprehensive and meaningful approach for knowledge discovery. The method aims to automatically determine coherent regions (by merging adjacent intervals based on similar distributions), identify irrelevant attributes, and efficiently handle large databases for more meaningful knowledge discovery. Testing the similarity of interval distributions using the STUCCO test [2], which considers multivariate distributions, can preserve hidden patterns and ensure coherence among different variables, thereby preventing the loss of valuable information in the process. This process is modified in our approach to propose a causal discretization described in the next section.

3 Proposed approach

3.1 Definition of a causal relevance index

Our approach aligns with foundational assumptions from Strobl [20], using the PC algorithm to learn a Directed Acyclic Graph (DAG) following the global directed Markov property. We adopt causal sufficiency, d-separation faithfulness, and a zero Type II error rate. The goal is to quantify uncertainty in constructing causal graphs within discretization, focusing on leveraging p-values from conditional independence tests. This addresses the impact of discretizing continuous variables, including the necessity of tests like Conditional Mutual Information (CMI) estimation with bin-based partitioning.

To control Type I error, we propose a robust index to compare discretization strategies, guiding selection based on data characteristics and specific causal graphs. This aims to enhance the accuracy and reliability of causal graph discovery by evaluating the impact of discretization on Type I error rates.

In simulation data, we assess the discretization impact using the true positive mean. In real-world studies, lacking a known truth, our proposed evaluation index mimics the true positive mean's behavior for practical assessment.

Our approach relies on the definition of an index reflecting the causal relevance of a certain partition denoted D of the continuous multivariate data space $\mathbb{R}^{|\mathcal{V}|}$. We evaluate partition D through its associated estimation of a causal graph G via the PC algorithm [4]. Our idea is to exploit the edge-wise p-value associated with each of the $|\mathcal{V}|^2 - |\mathcal{V}|/2$ possible edges. We build such an index by using the following heuristic.

Let define Da the variable of the observing data and the test to assess the absence of an edge in the skeleton of a connected undirected graph as: $H_0 : A - B$ is absent and $H_1 : A - B$ is present. In an ideal situation, we want to evaluate $P(H_1|Da)$ and compare it to an acceptance threshold β to accept H_1 or not. However, in practice, this probability is not observed. By using the assumptions and the method described in section 2.2, we estimate $P(Da|H_0)$ by p''_{A-B} in Equation 1 and compare it to the threshold α to reject H_0 or not. By the Bayes theorem, we obtain the link between these quantities:

$$P(H_1|Da) = 1 - P(H_0|Da) = 1 - p''_{A-B}\frac{P(H_0)}{P(Da)} \quad . \tag{2}$$

In Equation 2, $P(H_0)$ and $P(Da)$ are unknown and can be defined as a prior with the following constraints so that the acceptance thresholds β and α match each other:

$$\begin{aligned} &P(H_1|Da) = \beta \text{ when } p''_{A-B} = \alpha, \ \ P(H_1|Da) = 1 \text{ when } p''_{A-B} = 0 \ , \\ &P(H_0|Da) = 1 - \beta \text{ when } p''_{A-B} = 1 \ , P(H_0|Da) = 1 \text{ when } p''_{A-B} = 1 \ . \end{aligned} \tag{3}$$

Without more information on $P(H_0), P(H_1)$ and $P(Da)$, we derive a linear transformation to compute $P(H_1|Da)$ and $P(H_0|Da)$ from the only known quantity $p''_{A,B}$ and that enforces the constraints given in Equation 3:

$$\begin{aligned} P(H_1|Da) &\approx 1 - \frac{1-\beta}{\alpha}p''_{A-B} \ , \\ P(H_0|Da) &\approx 1 - \frac{\beta}{1-\alpha}(1 - p''_{A-B}) \ . \end{aligned}$$

Considering $(|\mathcal{V}|^2 - |\mathcal{V}|)/2$ possible edges, with the strong assumption that $P(H_1|Da)$ is correctly calibrated for all nodes $\{A, B\}$, we can approximate the mean number of True Positive causal links by:

$$\begin{aligned} \frac{TP}{N_{edges}} &\approx \frac{2}{(|\mathcal{V}|^2 - |\mathcal{V}|)}[\sum_{\{A,B\}\in\mathcal{V}} P(H_1|Da)] \\ &= 1 - \frac{2}{(|\mathcal{V}|^2 - |\mathcal{V}|)}[\sum_{\{A,B\}\in\mathcal{V}} \frac{1-\beta}{\alpha}p''_{A-B}] \end{aligned} \tag{4}$$

Finally, the constraint $\forall p''_{A-B}, P(H_0|Da) + P(H_1|Da) = 1$ is ensured if $\beta = 1-\alpha$. Injecting this constraint in Equation 4 gives the definition of our proposed **causal relevance index** (cri) that we build to evaluate the causal relevance of a causal graph G:

$$cri(G) = 1 - \frac{2}{|\mathcal{V}|^2 - |\mathcal{V}|}\sum_{\{A,B\}\in\mathcal{V}} p''_{A-B} \quad . \tag{5}$$

This index has to be maximized to obtain the most relevant graph G, i.e. we consider the ideally causally relevant discretization to be the one for which the best conditions for detection are met. Table 1 reminds us that depending on the value of p''_{A-B}, and the true status of the edge, we can consider four scenarios. Furthermore, decreasing p''_{A-B} mechanically reduces false positive. In other words, maximizing link discovery reduces to jointly minimizing the set of edges-wise p-values. Let us note that our approach relies strongly on the assumption of zero type II error rate of the conditional independence test used in the PC algorithm as in [20]. Without this assumption, the detection of false positives cannot be controlled. By construction, our index aims to quantify precision (TP) in the edge discovery task, which quantifies the causal graph uncertainty. We then use it to quantify the impact of discretization on the causal graph relevance. Note that discretization does not change the graph dimension (number of variables), but our index is sufficiently general to quantify uncertainty in other use cases with varying numbers of variables.

3.2 Optimization procedure

The search for the most causally relevant partition from the initial continuous data is described in the following. Our procedure structure is inspired by Bay's multivariate discretization, except that in our case, the criterion for merging intervals is based on the *cri* index derived in Equation 5. The proposed approach operates as follows:

1. Finely partition all continuous attributes into m intervals using a random discretization.
2. Estimate G and compute $cri(G)$ from this initial data partition.
3. Select the two adjacent intervals (of any variable) with the minimum combined supports.
4. Temporarily merge these two intervals, estimate G', and compute $cri(G')$ from the resulting data.
5. If merging these two intervals makes the causal graph more relevant, i.e., $cri(G') - cri(G) > 0$, then the merging is adopted, and G takes G' value. Otherwise, set a definitive border between the two intervals.
6. If there are no more candidate intervals, stop. Otherwise, return to step 3.

By pursuing this procedure, we will progressively increase the causal relevance of the partition as we can only increase the causal relevance index *cri* at each interval merging. However, optimization may be halted due to local maximums. We, therefore, perform several optimization trajectories starting from different random initial partitions and keep the graph that provides the highest *cri*.

4 Problem modelling

We aim to build an elementary model of the considered problem of multivariate discretization for causal discovery. We assume that the continuous variables derive from a discrete underlying causal phenomenon. This assumption allows us to acknowledge the perfect multivariate discretization as well as the true causal graph matching the simulated data in our experiments and thus carry out some investigations on the ability of our defined index *cri* to reflect the true positive mean and the proposed method to recover the true causal graph.

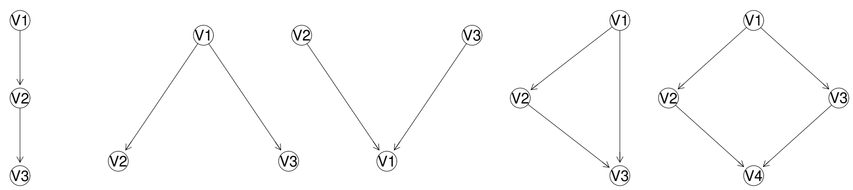
Consider a set of $|\mathcal{V}|$ random variables $V_1, \ldots, V_{|\mathcal{V}|}$ defined on $[0,1]^{|\mathcal{V}|}$, for which the causal graph G is known and such that the relationships between the variables are reduced to partially determined functions on intervals. For two variables V_i and V_j, $i, j \in \{1, \dots, |\mathcal{V}|\}$, such that $V_i \to V_j$, we consider the joint realization (v_i, v_j) of (V_i, V_j) is in the form:

$$(v_1 \in I^a_{V_1} \implies v_2 \in I^b_{V_2}) \wedge (v_1 \notin I^a_{V_1} \implies v_2 \in [0, 1]), \tag{6}$$

where $I^a_{V_1}$ and $I^b_{V_2}$ are intervals in $[0, 1]$. We restrict the study of the problem to the five classical causal graph structures represented in Figure 1.

Fig. 1: Structures under study

(a) Chain (b) Fork (c) V-structure (d) Mediator (e) Diamond



4.1 Simulations

Our simulation procedure consists in generating the triplet, composed by a ground truth multivariate discretization, its corresponding true causal graph, and a sample of corresponding continuous multivariate data of size n. We obtain these three elements by successively completing the values of the n joint realizations of the $|\mathcal{V}|$ variables:

1. Choose a causal graph G from the considered structures.
2. For each level 1 variable, i.e. nodes that are effects of no nodes, independently generate n realizations of $\mathcal{U}([0, 1])$.
3. For each unprocessed edge coming from nodes with no missing realizations (on the first pass, these are level 1 nodes), randomly create a non-empty I_a interval of the cause and an I_b interval of the effect. For each realization of the cause falling in I_a, generate[3] the effect realisation by a uniform distribution on the I_b interval, $\mathcal{U}(I_b)$
4. For nodes where all causes have been taken into account, complete the missing z joint realizations by generating z realizations from $\mathcal{U}([0, 1])$.
5. Return to step 3 for as long as some node realisations are still missing.

[3] It is possible that the effects of one cause will overwrite those of another. To deal with this, we randomly select one of the realizations from each cause. If the two cause variables are independent, the expected apparition value of such conflict is $1/4$, since we picked at random the effect to consider, the causal condition is ignored with a probability $1/8$.

4.2 Example of simulation (V-structure)

We provide details on the generation procedure for a simulation of the problem with the V-structure. First, the variables V_2 and V_3 (level 1 variables) are generated by a uniform distribution $\mathcal{U}([0,1])$. Then, we simulate the links corresponding to the two edges coming to V_1. We generate the intervals shown in Table 2. Then, we apply the conditions. For the first condition, for all realizations of V_2 in $[0.761, 0.939]$, we generate a joint realization for V_1 by $\mathcal{U}([0.417, 0.512])$. We proceed similarly for the second condition. If the two conditions may apply, we randomly select one condition with equiprobability. At this stage, the vector of V_1 variable is partially completed. We complete the missing realizations of V_1 (which correspond to the cases $V_2 \notin [0.761, 0.939]$ and $V_3 \notin [0.34, 0.421]$) by a $\mathcal{U}([0,1])$.

Table 2: Explicit causal links for the considered example of V-structure and, for the detailed chosen trajectory (Figure 2), rate r of realizations subject to the condition.

causal link	condition	r
$V_2 \to V_1$	$v_2 \in [0.761, 0.939] \implies v_1 \in [0.417, 0.512]$	0.17
$V_3 \to V_1$	$v_3 \in [0.34, 0.421] \implies v_1 \in [0.334, 0.351]$	0.11

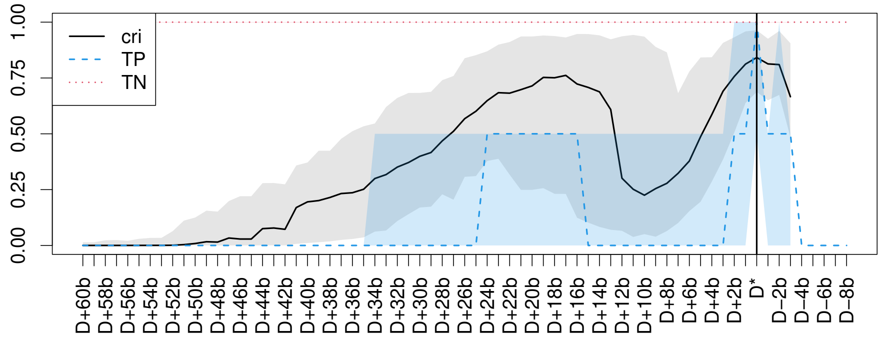
5 Experiments

Since we are only considering the presence or absence of the causal links, not their orientation, we chose to evaluate the performances with the rate of true positives, i.e., proportion of true links in all found links, and true negatives, i.e., proportions of not found links in all absent links. In our experiments, we considered the following conditional independence testing:

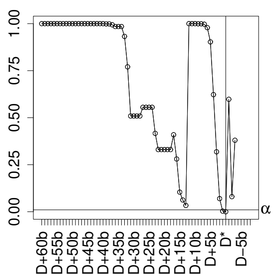
- Fisher z-transformation Conditional Independence test (ZCI) [9]: designed for gaussian data, tests for zero partial correlation via Fisher's z-transformation.
- Kernel Conditional Independence test (KCI) [6,21]: based on normalized cross-covariance operators on reproducing kernel Hilbert spaces.
- Stochastic complexity-based Conditional Independence criterium (SCI) [14]: conditional mutual information is used as a measure for conditional independence and approximated using stochastic complexity.
- Mutual information conditional independence test (MCI) [5]: an information-theoretic distance measure designed for discrete data.

The experiments were carried out in the R language and made publicly available[4].

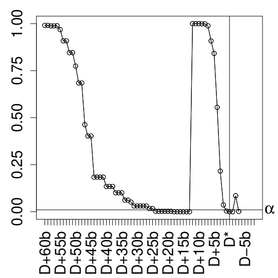
[4] `https://github.com/lucieK-J/CausalDiscretization.git`. The experiments were carried out on a MacBookPro18.4 with a total of 10 cores and 64 GB of memory.



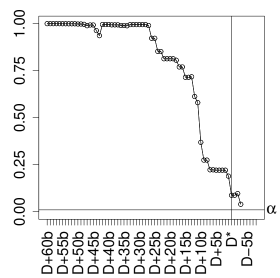
(a) Curves for *cri*, TP and TN are median values and colored regions represent their 0.90 confidence intervals over 100 distributions.



(b) Tracking of $p''_{V_1-V_2}$, case of truly present edge.



(c) Tracking of $p''_{V_1-V_3}$, case of truly present edge.



(d) Tracking of $p''_{V_2-V_3}$, case of truly absent edge.

Fig. 2: For the V-structure and the MCI test: degradation of the optimal discretization ($D*$) by removing i borders ($D - ib$) or adding i borders ($D + ib$).

5.1 Study of the *cri* behaviour

Let us now conduct an experimental study to verify the behavior of our index *cri* with respect to the quality of the discretization in terms of causal relevance. For this, we choose to degrade the optimal partition denoted D^* in a bilateral way, on the one hand by successively and randomly removing borders, and on the other hand by successively adding borders randomly to a randomly chosen node variable in $\mathcal{V}$.

Setting $n = 100$, and considering the same conditions as in our previous example (Table 2), we chose to present one such trajectory of degradation in Figure 2. Since the obtained *cri*, true positive, and true negative depend on the data distribution, we perform a sensitivity analysis in sub-figure 2a. For illustrating our index construction from the edge-wise level, we include the edge-wise p-value estimate (defined in (1)) in the sub-figures 2b-2d for one data distribution.

Note that we stop calculating the index if a variable becomes constant due to the suppression of too many borders. On a global view, we observe that our index reflects the true positive mean trend and that the optimal discretization obtains the highest index value. We can also comment that *cri* correctly captures the uncertainty arising from the data distribution, as the sensitivity analysis shows a matching trend between true positive and *cri* confidence intervals. On the edge-by-edge level, we can see that the optimal partition D^* corresponds to low values of the edge-wise p-value estimate regardless of the status of the edge (present or absent). Note that the estimated edge-wise p-values are below α for D^* if the edges are truly present and below if truly absent.

5.2 Results

We present the results of causal graph estimations via the PC algorithm ($\alpha = 0.01$) over $n = 100$ observations for the five structures in terms of average true positive and true negative in Table 3. Results are averaged over 500 simulations for each structure. We have carried out the tests considering three situations for the discretization: baselines, our proposed approach (causal discretization), and perfect discretization given.

Table 3: Average true positive and true negative rates with $n = 100$ on 500 simulations. Negative true rates are not defined for the mediator due to no absent edge in this structure.

			chain		fork		V-structure		mediator		diamond	
			TP	*TN*	TP	*TN*	TP	*TN*	TP	*TN*	TP	*TN*
Baselines	no discretisation	ZCI	0.21	*0.96*	0.21	*0.98*	0.19	*0.99*	0.20	*NA*	0.20	*0.981*
		KCI	0.47	*0.99*	0.48	*0.99*	0.45	*0.99*	0.44	*NA*	0.44	*0.99*
	Equal frequency discretization	SCI	0.36	*1*	0.37	*0.99*	0.30	*1*	0.30	*NA*	0.31	*1*
		MCI	0.45	*0.99*	0.47	*0.99*	0.40	*1*	0.42	*NA*	0.40	*1*
Causal discretization		SCI	0.48	*0.99*	0.49	*0.98*	0.45	*1*	0.42	*NA*	0.35	*1*
		MCI	0.56	*0.95*	0.62	*0.93*	0.51	*0.92*	0.46	*NA*	0.43	*0.99*
Perfect discretization given		SCI	0.83	*0.99*	0.87	*0.97*	0.84	*0.98*	0.73	*NA*	0.66	*1*
		MCI	0.74	*1*	0.79	*1*	0.79	*1*	0.39	*NA*	0.39	*1*

For baselines, two configurations are considered: estimating the graph from non-discretized data using ZCI and KCI tests, and applying equal frequency discretization with specified bin numbers. The causal graphs are estimated using SCI and MCI for both the equal frequency discretization and the proposed method. The proposed method uses SCI and MCI in both discretization and causal graph estimation, running 30 random initializations with 10 initial bins for each simulation. For perfect discretization, the causal graph is estimated using SCI and MCI tests.

Empirically, the zero type II error hypothesis is generally respected, resulting in high true negative rates. ZCI struggles with detecting causal links in continuous data baselines due to its assumption of Gaussian data, while KCI performs better, highlighting the drawbacks of arbitrary discretization. Knowing the ideal discretization generally yields the best results, emphasizing the importance of intelligent discretization.

Our causal discretization with the MCI test often outperforms the ideal discretization in true positive mean, particularly in mediator and diamond structures. This approach highlights causal links even with limited examples in intervals, especially for structures with more edges. A separate study confirms the relevance of ideal discretization for larger sample sizes.

Our method outperforms equal-frequency discretization in discrete data tests (MCI or SCI) in true positive mean across all structures. The best results are seen with perfect discretization, indicating potential for further improvement. MCI detects more causal links than SCI but with a lower true negative rate. However, when perfect discretization is known, SCI surpasses MCI in identifying causal links, which may be due to SCI challenges in calibrating p-values affecting our index calculation.

6 Conclusion

In causal discovery, data discretization can be a major issue, as it has a considerable impact on the relevance of the results. However, this discretization is indispensable, for instance, when using tests that require the data to be used in bins as the proxy variable approach or when the application requires a discretization. To assess the impact of the discretization on the causal discovery, we proposed a causal relevance index approaching the true positive mean of causal links estimated via the PC algorithm. We performed a sensitivity analysis of our proposed index, which showed that it correctly captures the data uncertainty to reflect the true positive rate. Moreover, we proposed an approach that seeks for the more causally relevant multivariate partition based on this index.

We have tested on simulations several approaches for causal discovery with and without discretization. Our results show that the proposed index reflects the actual true positive mean. We also show that our optimization procedure tends to approach the "upper bound" given by the perfect partitioning of the data, which is usually unknown in real-world problems. However, there is still room for improvement. The results also indicate the impact of p-values calibration on every approach. Also, the false discovery rate leans on the Type II error rate of the independence tests. We recommend using a well-calibrated conditional independence test to search for the partition. Once the partition is identified, one could use another and more efficient test for the causal discovery phase.

One limitation of our work is that each iteration requires the estimation of a causal graph with the PC algorithm. In future work, we plan to reduce the deriving complexity by considering an alternative search procedure using stochastic optimization. Another perspective to improve our approach is to investigate the possibilities to optimize simultaneously both the true positive and the true negative rate and free ourselves from the zero type II errors hypothesis, which is rarely respected in practice. To do this, we could benefit from Strob's work, which proposes a modification of the PC Algorithm to control the false positive rate despite non-zero Type II error rates using the bounded p-values and the Benjamini-Yekutieli procedure.

References

1. Bay, S.D.: Multivariate discretization for set mining. Knowledge and information systems **3**(4), 491–512 (2001)
2. Bay, S.D., Pazzani, M.J.: Detecting change in categorical data: Mining contrast sets. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 302–306 (1999)
3. Bonissone, P., Henrion, M., Kanal, L., Lemmer, J.: Equivalence and synthesis of causal models. In: Uncertainty in artificial intelligence. vol. 6, p. 255. Elsevier Science & Technology (1991)
4. Colombo, D., Maathuis, M.H., et al.: Order-independent constraint-based causal structure learning. J. Mach. Learn. Res. **15**(1), 3741–3782 (2014)
5. Edwards, D.: Introduction to graphical modelling. Springer Science & Business Media (2012)
6. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. Advances in neural information processing systems **20** (2007)
7. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: Problems and methods. ACM Computing Surveys (CSUR) **53**(4), 1–37 (2020)
8. Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., Schölkopf, B.: Causal discovery from heterogeneous/nonstationary data. The Journal of Machine Learning Research **21**(1), 3482–3534 (2020)
9. Kalisch, M., Bühlman, P.: Estimating high-dimensional directed acyclic graphs with the pc-algorithm. Journal of Machine Learning Research **8**(3) (2007)
10. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical review E **69**(6), 066138 (2004)
11. Liu, M., Sun, X., Qiao, Y., Wang, Y.: Causal discovery with unobserved variables: A proxy variable approach. arXiv preprint arXiv:2305.05281 (2023)
12. Malinsky, D., Danks, D.: Causal discovery algorithms: A practical guide. Philosophy Compass **13**(1), e12470 (2018)
13. Margaritis, D.: Distribution-free learning of bayesian network structure in continuous domains. In: AAAI. vol. 5, pp. 825–830 (2005)
14. Marx, A., Vreeken, J.: Testing conditional independence on discrete data using stochastic complexity. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 496–505. PMLR (2019)
15. Nogueira, A.R., Pugnana, A., Ruggieri, S., Pedreschi, D., Gama, J.: Methods and tools for causal discovery and causal inference. Wiley interdisciplinary reviews: data mining and knowledge discovery **12**(2), e1449 (2022)
16. Pearl, J.: Causality. Cambridge university press (2009)
17. Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology **78**(5), 947–1012 (2016)
18. Ruiz, A., Villa, D.: Causal inference for data science. O'Reilly media (2023)
19. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
20. Strobl, E.V., Spirtes, P.L., Visweswaran, S.: Estimating and controlling the false discovery rate of the pc algorithm using edge-specific p-values. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(5), 1–37 (2019)
21. Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based conditional independence test and application in causal discovery. arXiv preprint arXiv:1202.3775 (2012)