Ordinal Causal Discovery

Yang Ni

Bani Mallick

1Department of Statistics, Texas A&M University, College Station, Texas, USA

Abstract

Causal discovery for purely observational, categorical data is a long-standing challenging problem. Unlike continuous data, the vast majority of existing methods for categorical data focus on inferring the Markov equivalence class only, which leaves the direction of some causal relationships undetermined. This paper proposes an identifiable ordinal causal discovery method that exploits the ordinal information contained in many real-world applications to uniquely identify the causal structure. The proposed method is applicable beyond ordinal data via data discretization. Through real-world and synthetic experiments, we demonstrate that the proposed ordinal causal discovery method combined with simple score-and-search algorithms has favorable and robust performance compared to state-of-the-art alternative methods in both ordinal categorical and non-categorical data. An accompanied R package OCD is freely available at the first author’s website.

1 INTRODUCTION

Causal discovery [Spirtes et al., 2000; Pearl, 2009] is becoming increasingly more popular in machine learning and finds numerous applications, e.g., biology [Sachs et al., 2005], psychology [Steyvers et al., 2003], and neuroscience [Shen et al., 2020], of which the prevailing goal is to discover causal relationships of variables of interest. The discovered causal relationships are useful for predicting a system’s response to external interventions [Pearl, 2009], a key step towards understanding and engineering that system. While the gold standard for causal discovery remains the controlled experimentation, it can be too expensive, unethical, or even impossible in many cases, particularly on human beings. Therefore, inferring the unknown causal structures of complex systems from purely observational data is often desirable and, sometimes, the only option.

This paper considers causal discovery for ordinal categorical data. Categorical data are common across multiple disciplines. For example, psychologists often use questionnaires to measure latent traits such as personality and depression. The responses to those questionnaires are often categorical, say, with five levels (5-point Likert scale): "strongly disagree", "disagree", "neutral", "agree", and "strongly agree". In genetics, single-nucleotide polymorphisms are categorical variables with three levels (mutation on neither, one, or both alleles). Categorical data also arise as a result of discretization of non-categorical (e.g., continuous and count) data. For instance, in biology, gene expression data are often trichotomized to “underexpression”, “normal expression”, and “overexpression” [Parmigiani et al., 2002; Pe’er, 2005; Sachs et al., 2005] in order to reduce sequencing technical noise while retaining biological interpretability.

While causal discovery for purely observational categorical data have been extensively studied, the vast majority of existing methods [Heckerman et al., 1995; Chickering, 2002] have exclusively focused on Bayesian networks (BNs) with nominal (unordered) categorical variables. It has been well established that a nominal/multinomial BN is generally only identifiable up to Markov equivalence class in which all BNs encode the same Markov properties. For example, $X \rightarrow Y$ and $Y \rightarrow X$ are Markov equivalent and also distribution equivalent [Spirtes and Zhang, 2016] with a multinomial likelihood; therefore, they are non-identifiable with purely observational data.

In many real-world applications, categorical data (including the aforementioned Likert scale, single-nucleotide polymorphisms, and discretized gene expression data) contain ordinal information. In this paper, we show that this often-overlooked ordinal information is crucial in causal discovery for categorical data. We propose an ordinal causal discovery (OCD) method via an ordinal BN. Assuming causal Markov and causal sufficiency, we prove OCD to be identifiable.
in general for ordinal categorical data. Score-and-search BN structure learning algorithms are developed – exhaustive search for small networks (e.g., bivariate data) and greedy search for moderate-sized networks. Through extensive experiments with real-world and synthetic datasets, we demonstrate that the proposed OCD is identifiable, robust, applicable to both categorical and non-categorical data, and competitive against a range of state-of-the-art causal discovery methods. To the best of our knowledge, we are the first to exploit the ordinal information for causal discovery in categorical data. Our major contributions are four-fold.

1. We advocate the usefulness of ordinal information of categorical data in causal discovery, which has been overlooked in the literature.
2. We propose the first causal discovery method, OCD, for ordinal categorical data.
3. We prove that OCD is generally identifiable for bivariate data, in contrast to the non-identifiability of multinomial BNs.
4. We demonstrate the strong utility of OCD by comparison with state-of-the-art alternatives using real-world and synthetic datasets.

1.1 RELATED WORK

For brevity, we review causal discovery methods that are fully identifiable with observational data.

**Non-Categorical Data.** Model-based BNs for continuous data are often represented as additive noise models. Under such representation, BNs are generally identifiable if the noises are non-Gaussian [Shimizu et al. 2006], if the functional form of the additive noise model is nonlinear [Hoyer et al. 2009], or if the noise variances are equal [Peters and Bühlmann 2014]. Also see much of the recent literature that focuses on bivariate causal discovery [Mooij et al. 2010, Janzing et al. 2012, Chen et al. 2014, Sgouritsa et al. 2015, Hernandez-Lobato et al. 2016, Marx and Vreeken 2017, Blobaum et al. 2018, Marx and Vreeken 2019, Tagasovska et al. 2020]. For count data, Park and Raskutti 2015 proposed a Poisson BN and showed that it is identifiable based on the overdispersion property of Poisson BNs. By replacing overdispersion property with constant moments ratio property, Park and Park 2019 extended Poisson BNs to the generalized hypergeometric family which contains many count distributions such as binomial, Poisson, and negative binomial. Recently, Choi et al. 2020 developed a zero-inflated Poisson BN for zero-inflated count data.

**Categorical Data.** For nominal categorical data, causal identification is possible under certain assumptions [Peters et al. 2010, Suzuki et al. 2014, Liu and Chan 2016, Cai et al. 2018, Compton et al. 2020, Qiao et al. 2021], e.g., when the categories admit hidden compact representations or when data follow a discrete additive noise model. However, to the best of our knowledge, causal discovery for ordinal data, which are very common in practice, has not been studied. Whether a categorical variable is ordinal or not is, in our opinion, easier to comprehend than the aforementioned assumptions of categorical data (e.g., discrete additive noise). We remark that a recent paper [Luo et al. 2021] also considered ordinal data. However, their work is substantially different from ours. The most prominent difference is that the causal graph of [Luo et al. 2021] is only identifiable up to Markov equivalence classes whereas the proposed method is uniquely identifiable, which is proved for the bivariate case.

**Mixed Data.** There are recent developments for mixed data causal discovery [Cui et al. 2018, Tsagris et al. 2018, Sedgewick et al. 2019], some of which include categorical data. However, the ordinal nature of the categorical data is not exploited for causal identification; therefore, these algorithms output Markov equivalent BNs instead of individual BNs. The latent variable approach by [Wei et al. 2018] could in principle be extended to ordinal data. However, the causal Markov assumption of latent variables cannot translate to the observed variables and the inferred causality does not have direct causal interpretation on the observed variables.

2 BIVARIATE ORDINAL CAUSAL DISCOVERY

We first introduce the proposed OCD method for bivariate data, which will be extended to multivariate data in Section 3. Let \((X, Y) \in \{1, \ldots, S\} \times \{1, \ldots, L\}\) denote a pair of ordinal variables with \(S\) and \(L\) levels, of which the possible causal relationships, \(X \rightarrow Y\) or \(Y \rightarrow X\), are under investigation. Throughout the paper, we make the causal Markov and causal sufficient assumptions, which are frequently adopted in the causal discovery literature [Pearl 2009]. The former allows us to interpret the proposed model causally (beyond conditional independence) whereas the latter asserts that there are no unmeasured confounders.

The bivariate OCD considers the following probability distribution for causal model \(X \rightarrow Y\),

\[
p_{X \rightarrow Y}(X, Y) = p(X)p(Y|X),
\]

where \(p(X)\) is a multinomial/categorical distribution with probabilities \(\pi = (\pi_1, \ldots, \pi_S)\) with \(\sum_{s=1}^{S} \pi_s = 1\), and \(p(Y|X)\) is defined by an ordinal regression model [Agresti 2003].

\[
Pr(Y \leq \ell|X) = F(\gamma_\ell - \beta_X), \quad \ell = 1, \ldots, L,
\]

where \(\beta_X\) is a generic notation of \(\beta_1, \ldots, \beta_S\) for \(X = 1, \ldots, S\). Typical choices of the link function \(F\) are probit and inverse logit, which are empirically quite similar;
hereafter we always use the probit link except for the identifiability theory, which is valid for both link functions. We fix $\gamma_1 = 0$ for ordinal regression parameter identifiability [Agresti 2003]. Equation (2) implies the conditional probability distribution $P(Y = l | X = s) = F(\gamma_l - \beta_s)$ for $l = 1, \ldots, L$ and $s = 1, \ldots, S$ where $\gamma_0 = -\infty$ and $\gamma_L = \infty$. Let $\beta = (\beta_1, \ldots, \beta_S)$ and $\gamma = (\gamma_2, \ldots, \gamma_{L-1})$. We denote the model $p_{X \rightarrow Y}$ by $p_{X \rightarrow Y}(X, Y | \pi, \beta, \gamma)$. Similarly, we define the probability model $p_{Y \rightarrow X}$ as $p_{Y \rightarrow X}(Y, X | \rho, \alpha, \eta)$. If the maximum likelihood estimate $\hat{p}_{X \rightarrow Y}$ given observations of $(X, Y)$ is strictly larger than $\hat{p}_{Y \rightarrow X}$, then $X \rightarrow Y$ is deemed a more likely data generating causal model.

3 IDENTIFIABILITY

We will show that the proposed OCD is generally identifiable.

Definition 1 (Distribution Equivalence) $p_{X \rightarrow Y}(X, Y | \pi, \beta, \gamma)$ and $p_{Y \rightarrow X}(Y, X | \rho, \alpha, \eta)$ are distribution equivalent if for any values of $(\pi, \beta, \gamma)$ there exist values of $(\rho, \alpha, \eta)$ such that $p_{X \rightarrow Y}(X, Y | \pi, \beta, \gamma) = p_{Y \rightarrow X}(Y, X | \rho, \alpha, \eta)$ for any $X, Y$, and vice versa.

Distribution equivalent causal models are clearly not distinguishable from each other by examining their observational distributions. The well-known multinomial BNs are distribution equivalent as illustrated in the following example.

Example 1 (Multinomial BN) Consider a bivariate multinomial BN of $X \rightarrow Y$ whose conditional $p(Y | X)$ and marginal $p(X)$ probability distributions are given in Figure 1(a), and the joint distribution $p(X, Y)$ is given in Figure 1(b). Because of the multinomial assumption, we can find a set of parameters, i.e., the conditional $p(X | Y)$ and marginal $p(Y)$ probabilities (Figure 1(c)) of the reverse causal model $Y \rightarrow X$, which leads to the same joint distribution. Therefore, the probability distribution does not provide information for causal identification.

Incorporating the underappreciated ordinal information, we will show that $p_{X \rightarrow Y}(X, Y | \pi, \beta, \gamma)$ and $p_{Y \rightarrow X}(Y, X | \rho, \alpha, \eta)$ are generally not distribution equivalent and are, therefore, identifiable.

Theorem 1 (Identifiability of OCD) Let $X \in \{1, \ldots, S \}$ and $Y \in \{1, \ldots, L \}$ where $S, L > 2$. Suppose $X \rightarrow Y$ is the data generating causal model and the observational probability distribution of $(X, Y)$ is given by

$$p(X, Y) = p_{X \rightarrow Y}(X, Y | \pi, \beta, \gamma).$$

For almost all $(\pi, \beta, \gamma)$ with respect to the Lebesgue measure, the distribution cannot be equivalently represented by the reverse causal model, i.e., there does not exist $(\rho, \alpha, \eta)$ such that

$$p(X, Y) = p_{Y \rightarrow X}(Y, X | \rho, \alpha, \eta), \forall X, Y.$$
We develop simple score-and-search learning algorithms to estimate the structure of causal graphs, which already show strong empirical performance (see Section 6), although more sophisticated learning methods such as Bayesian inference could be adopted to further improve the performance.

Score. We score causal graphs by the Bayesian information criterion (BIC). We choose BIC over AIC because it favors a more parsimonious causal graph due to the heavier penalty on model complexity and generally has a better empirical performance. Let $x = (x_1, \ldots, x_n)$ denote $n$ realizations of $X$. The score of $G$ (smaller is better) is given by

$$\text{BIC}(G|x) = -2 \sum_{i=1}^{n} \log \hat{p}(x_i|G) + K \log(n),$$

where $K$ is the number of model parameters and $\hat{p}(x_i|G)$ is the joint distribution evaluated at $x_i$ given the MLE of model parameters.

**Exhaustive Search.** For small networks (say $q = 2$ or 3), we compute the scores for all networks $G$, and identify $G = \text{arg min}_{G \in \mathcal{G}} \text{BIC}(G|x)$. While this approach is exact and useful for bivariate OCD, it becomes computationally infeasible for moderate-sized networks as the number of networks $|\mathcal{G}|$ grows super-exponentially in $q$.

**Greedy Search.** We use a simple iterative greedy search algorithm [Chickering, 2002, Scutari et al., 2019] for moderate-sized networks. At each iteration, we score all the graphs that can be reached from the current graph by an edge addition, removal, or reversal. We replace the current graph by the graph with the largest improvement (largest decrease in BIC) and stop the algorithm when the score can no longer be improved. The greedy search algorithm is summarized in Algorithm 1, which is guaranteed to find a local optimal graph. The algorithm can be improved by tabu search and random non-local moves [Scutari et al., 2019] but we do not pursue this direction as the simple greedy algorithm already yields favorable results against state-of-the-art alternative methods. The worst per iteration cost is $O(q f(n, m, L))$ for $q$ nodes, $n$ observations, $m$ maximum number of parents, and $L$ maximum levels, where $f(n, m, L)$ is the computational complexity of an ordinal regression with $m$ regressors. This is because at most $2q$ score evaluations are required at each iteration [Scutari]
6 EXPERIMENTS

We evaluate the proposed and state-of-the-art alternative causal discovery methods with synthetic as well as three sets of real data. The real data are not categorical and therefore allow us to extend our comparison to causal models designed for continuous data.

6.1 SYNTHETIC ORDINAL DATA

We simulate low-dimensional, higher-dimensional, and bi-variate (with confounders) synthetic ordinal data.

6.1.1 Low-Dimensional Multivariate Ordinal Data

We consider synthetic ordinal data \( (n = 500, q = 10) \). To mimic survey data with 5-point Likert-scale questionnaires, we simulate data from the proposed OCD model with \( L_j = L = 5, \forall j \). The true BN is generated randomly (Figure 2(a)), which has one \( \beta \)-structure (i.e., subgraph \( j \rightarrow k \leftarrow i \)). Its Markov equivalence class, represented by a completed partially directed acyclic graph (CPDAG), can be obtained by removing the directionality of the red dashed edges in Figure 2(a). We consider 6 scenarios with different levels of signal strength by generating simulations are shown in Figure 2(b)-(c) as functions of signal strength \( \sigma \). Since multinomial BNs with BDe and BIC, and OSEM only estimate CPDAGs, we report the lower bounds of their SID. There are several conclusions that can be drawn. First, OCD is empirically identifiable because both SHD and SID quickly approach 0 as signal becomes stronger. Second, OCD uniformly outperforms the alternative methods in both SHD and SID across all signal levels, which suggests that exploiting the ordinal nature of ordinal categorical data is crucial for causal discovery. Third, BIC+ is better than BIC and BDe in SHD but not necessarily in SID, suggesting the estimated causal ordering from BIC+ is biased. Fourth, although OSEM also accounts for ordinal data, it is not identifiable and may be sensitive to the tuning parameter, which is hard to be objectively tuned. Therefore, we drop OSEM in the subsequent simulations.

6.1.2 Higher-Dimensional Multivariate Ordinal Data

We fix the sample size \( n = 500 \) and the number of categories \( L = 5 \) but vary the number of nodes \( q = 10, 20, \ldots, 100 \) and the signal strength \( \sigma = 0.25, 0.5, 0.75, 1, 1.25, 1.5 \). The graphs are kept at the same sparsity as in Section 6.1.1 across \( q \).

## Algorithm 1 Greedy Search

**Input:** data \( x \), initial empty graph \( G \)

1. Compute \( \text{BIC}(G|x) \) and set \( \text{BIC} = \text{BIC}(G|x) \).
2. **repeat**
   1. Initialize \( \text{Improvement} = \text{false} \).
   2. **for** all graphs \( G' \) reachable from \( G \) **do**
      1. Compute \( \text{BIC}(G'|x) \).
      2. **if** \( \text{BIC}(G'|x) < \text{BIC} \), **then**
         1. Set \( G = G' \) and \( \text{BIC} = \text{BIC}(G'|x) \)
         2. Set \( \text{Improvement} = \text{true} \).
   **end if**
3. **end for**
4. **until** \( \text{Improvement} = \text{false} \)

**Output:** graph \( G \)

We use \texttt{polr} function in the R package \texttt{MASS} for ordinal regression which appears to scale linearly in \( n, m \), and \( L \), empirically.
(denser graphs will be considered later). The SHD is shown in Figure[5] whereas the SID is provided in the Supplementary Materials. The proposed OCD uniformly outperforms the competing methods BDe, BIC, and BIC+ across q and σ. In general, OCD is quite stable as q increases when the signal strength is moderate to moderately large σ ≥ 0.5 whereas the competing methods quickly deteriorate with q regardless of the signal strength.

**Scalability.** We investigate the scalability of the proposed OCD with respect to n, L, and q. We vary n = 500, 750, . . . , 2750 (keeping q = 10 and L = 5), L = 5, . . . , 14 (keeping n = 500 and q = 10), and q = 10, 20, . . . , 100 (keeping n = 500 and L = 5). The total CPU times in seconds on a 2.9 GHz 6-Core Intel Core i9 laptop are provided in the Supplementary Materials. The greedy search appears to scale linearly in n and L, and quadratically in q, which agrees with the complexity analysis in Section[5] It is moderately scalable: e.g., for q = 100, the search completes in about 3 hours.

**Denser Graphs.** In the Supplementary Materials, we present additional simulation scenarios with denser graphs for q = 50 nodes and more v-structures, which lead to similar conclusions, i.e., OCD significantly outperforms the competing methods in SHD and SID.

**6.1.3 Bivariate Ordinal Data with Unmeasured Confounders**

While our identifiability theory assumes no unmeasured confounders, we now empirically test the sensitivity of OCD to unmeasured confounders for bivariate ordinal data. We generate trivariate ordinal data (X₁, X₂, X₃) with L = 5 from the following true causal graph,

![True DAG](image)

We hide X₃ as a confounder and apply OCD to (X₁, X₂). In the simulation truth, we assume βᵢjkℓ, for each ℓ = 1, . . . , L, to be the same for all j ≠ k, i.e., the confounding effect is the same as the causal effect, which is simulated from N(0, σ²). We consider different levels of signal strength σ = 0.25, 0.5, 0.75, 1, 1.25, 1.5 and different sample sizes n = 100, 200, . . . , 1000. Under each combination of (σ, n), we repeat the experiment 100 times, and report the average accuracy (ACC) for forced decisions. The forced decision forces methods to choose between X₁ → X₂ and X₂ → X₁. The same metric has been used in similar bivariate causal discovery problems [Mooij et al., 2016, Tagasovska et al., 2020]. OCD is relatively robust to confounders (Figure[4]a): it is able to correctly identify the causal direction given a large enough sample size or when the signal is sufficiently strong. For comparison, we apply a recent causal discovery method for bivariate nominal categorical data, HCR [Cai et al., 2018]. Its average ACC is shown in Figure[4]b). We find the ACC of HCR is uniformly lower than that of OCD although we note that HCR is not specifically designed for this task.

**6.2 SACHS’S SINGLE-CELL FLOW CYTOMETRY DATA**

We evaluate the proposed OCD on the well-known single-cell flow cytometry dataset [Sachs et al., 2005], which contains measurements of q = 11 phosphorylated proteins under different experimental conditions. Sachs et al. [2005] provided a consensus causal network of these proteins, which could be used to gauge the performance of causal discovery algorithms. As in Tagasovska et al. [2020] we consider the cd3cd28 dataset with 853 cells subject to the same experimental condition.

**Implementations.** Since the raw measurements are highly skewed and heavy-tailed, Sachs et al. [2005] discretized the data into L = 3 levels (“low”, “average”, and “high”) and fit a multinomial BN based on the BDe score. As we will
(b) SHD in \(q (\sigma = 0.25)\)

(a) Average ACC of OCD

Figure 3: SHD (lower is better) for OCD, BDe, BIC, and BIC+ as functions of \(q\) in the synthetic ordinal data with the sample size fixed at \(n = 500\) and different signal strength \(\sigma \in \{0.25, 0.5, 0.75, 1\}\).

see, this approach throws away the ordinal information inherent in the raw measurements and hence significantly underperforms OCD (with greedy search). For comparison, we also apply ANM [Hoyer et al., 2009], LiNGAM [Shimizu et al., 2006], RESIT with the Gaussian process implementation [Peters et al., 2014], bivariate causal discovery methods (HCR, bQCD [Tagasovska et al., 2020], GR-AN [Hernandez-Lobato et al., 2016], IGCI with uniform measure [Janzing et al., 2012], SLOPE [Marx and Vreeken, 2017]), and methods inferring Markov equivalence classes (PC [Spirtes et al., 2000], CPC [Ramsey et al., 2012], GES [Chickering, 2002], IAMB [Tsamardinos et al., 2003], multinomial BNs with BIC and BDe), and the mixed data approach MXM [Tsagris et al., 2018] to the raw continuous data. For bivariate causal discovery methods, we follow a similar ad hoc procedure in Tagasovska et al. [2020] first run CAM [Bühlmann et al., 2014] and then orient the estimated edges by the bivariate methods. HCR is the closest competitor as it is also designed for categorical data although with a very different scope (only applicable to bivariate nominal categorical data and assuming the existence of hidden compact representations). We still compare the proposed OCD with OSEM. To address the tuning parameter issue of OSEM, we tune it in an oracle way on an evenly-spaced 12-grid from 0.5 to 6.0.

**Metrics.** We use the same SHD and SID metrics as in Section 6.1. For methods that output CPDAGs instead of BNs, we report the lower and upper bounds of SID.

**Results.** In Table 1 we summarize the SHD and SID. OCD shows very strong performance comparing to state-of-the-art alternatives. It has the lowest SHD and the second lowest SID, which shows benefit of discretization for highly noisy data. The substantial improvement of OCD from multinomial BN with BDe (SHD 14 vs 21) highlights the importance of exploiting the ordinal information of discrete data for causal discovery. While there is strong motivation (e.g., biological interpretation) to use \(L = 3\) for this dataset, we test OCD with \(L\) up to 10. OCD stays very competitive within this range: the SID remains 62 for all \(L\) whereas the SHD slightly increases as \(L\) increases possibly due to relatively small sample size, e.g., SHD = 16 for \(L = 10\), which is still quite competitive (second to SHD = 15 for bQCD and IGCI). The smallest SHD that OSEM achieves
over the range of tuning parameter is 18.

Table 1: Sachs’s data. Methods (marked by *) that are only applicable to bivariate data are combined with CAM, PC, CPC, GES, IAMB, BIC, BDe, and MXM only learn CPDAGs; we provide the lower and upper bounds of SID. Lower SHD and SID are better.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SHD</th>
<th>SID</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCD</td>
<td>14</td>
<td>62</td>
</tr>
<tr>
<td>bQCD*</td>
<td>15</td>
<td>69</td>
</tr>
<tr>
<td>IGCI*</td>
<td>15</td>
<td>82</td>
</tr>
<tr>
<td>GR-AN*</td>
<td>16</td>
<td>80</td>
</tr>
<tr>
<td>HCR*</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>SLOPE*</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>ANM</td>
<td>17</td>
<td>86</td>
</tr>
<tr>
<td>LiNGAM</td>
<td>16</td>
<td>86</td>
</tr>
<tr>
<td>PC</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>CPC</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>GES</td>
<td>50-83</td>
<td>50-80</td>
</tr>
<tr>
<td>IAMB</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>BIC</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>BDe</td>
<td>49-104</td>
<td>49-104</td>
</tr>
<tr>
<td>MXM</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>RESIT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: CEP data. Metrics of OCD and HCR are averaged over different values of $L = 10, \ldots, 20$ with standard errors given within the parentheses. Higher ACC and AUC are better.

<table>
<thead>
<tr>
<th>Methods</th>
<th>ACC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCD</td>
<td>0.73 (0.01)</td>
<td>0.51 (0.00)</td>
</tr>
<tr>
<td>HCR</td>
<td>0.76 (0.00)</td>
<td>0.76 (0.02)</td>
</tr>
<tr>
<td>bQCD*</td>
<td>0.70</td>
<td>0.56</td>
</tr>
<tr>
<td>CAM</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>CPU</td>
<td>1s</td>
<td>1s</td>
</tr>
<tr>
<td>SHD</td>
<td>50-83</td>
<td>50-80</td>
</tr>
<tr>
<td>SLOPE*</td>
<td>50-80</td>
<td>50-70</td>
</tr>
<tr>
<td>ANM</td>
<td>50-80</td>
<td>79-70</td>
</tr>
<tr>
<td>LiNGAM</td>
<td>49-104</td>
<td>49-104</td>
</tr>
<tr>
<td>RESIT</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>

6.3 CAUSEEFFECTPAIRS (CEP) BENCHMARK DATA

We consider the CauseEffectPairs (CEP) benchmark data [Mooij et al., 2016] (version: 12/20/2017), which contain 108 datasets from 37 domains (e.g., biology, economy, engineering, and meteorology). Each dataset contains a pair of variables ($X, Y$) for which the causal relationship is clear from the context, e.g., older "age" causes higher "glucose". We retain the same 99 pairs as in Tagasovska et al., 2020 that have univariate non-binary cause and effect variables.

Implementations. We compare OCD with HCR, bQCD, IGCI, CAM, SLOPE, LiNGAM, and RESIT. To apply OCD and HCR, we discretize each variable at $L - 1$ quantiles for $L \in \{10, \ldots, 20\}$. All other methods are applied to the (standardized) continuous data without discretization.

Metrics. We compute the ACC for forced decisions as in Section 6.3 and, additionally, the area under the receiver operating curve (AUC) for ranked decision. The ranked decision ranks the confidence of the causal direction [Mooij et al., 2016] Tagasovska et al., 2020. The simple heuristic confidence [Mooij et al., 2016] is adopted here. For instance, for the proposed OCD, we define the confidence of $X \rightarrow Y$ to be $C_{X \rightarrow Y} = BIC(Y \rightarrow X|x) - BIC(Y \rightarrow Y|x)$.

Results. In Table 1 we summarize the ACC, AUC, and CPU times. For OCD and HCR, the average metrics over $L = 10, \ldots, 20$ as well as their standard errors are reported. The proposed OCD is highly competitive in all metrics. OCD has the second highest ACC and AUC, and is fast; it completes the analysis of 99 datasets in 36 seconds. Only IGCI, CAM, and LiNGAM are faster but they have worse ACC and AUC than OCD. SLOPE has slightly higher ACC and AUC than OCD. However, SLOPE is about 1 or 2 orders of magnitude slower than OCD and relatively sensitive to small added noise (see the additional experiments that investigate the "Sensitivity to Small Added Noise" in the Supplementary Materials). Finally, the small standard errors of the performance metrics of OCD indicate its relative robustness with respect to the number $L$ of levels of discretization for the considered datasets and range.

6.4 SINGLE-CELL RNA-SEQUENCING DATA

We further validate the proposed OCD with a publicly available single-cell RNA-sequencing (scRNA-seq) dataset of 2,717 murine embryonic stem cells [Klein et al., 2015]. We obtain a list of literature-curated pairs of transcription factor ($X$) and its target ($Y$) from the TRRUST database [Han et al., 2018], which provides biological ground truth of the causal relationships, namely $X \rightarrow Y$. We then extract the corresponding genes from the scRNA-seq dataset. Removing genes with more than 90% zeros (these genes have very low statistical variability), we retain 6701 pairs for causal validation, which still have 62% zeros. The zeros in scRNA-seq data are either (a) true biological zero counts or (b) small counts that are too low to detect. In either case, they can be regarded as "low expression". We compare OCD with the best performing methods in Section 6.3 bQCD and SLOPE, as well as the closest competitor HCR. We are not able to generate results (runtime errors) from CAM, LiNGAM, and RESIT possibly because of the large percentages of zeros. To apply OCD and HCR, we dichotomize the data at 0 and the median of the non-zero expression (i.e., "low", "average", and "high" expression). ACC and CPU time are reported in Table 2. OCD is the best and is the only method that is better than random guess (p-value = 10^{-75}, binomial test with $H_0 : p = 0.5$ vs $H_a : p > 0.5$) for this dataset possibly because of its highly non-standard distribution due to zero-inflation. Therefore, although discretizing continuous or count data may lose information, it often improves the robustness by not having to impose a particular distributional assumption on the raw data.
Table 3: Single-cell RNA-seq data. Higher ACC is better.

<table>
<thead>
<tr>
<th></th>
<th>OCD</th>
<th>HCR</th>
<th>bQCD</th>
<th>SLOPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>0.61</td>
<td>0.36</td>
<td>0.45</td>
<td>0.50</td>
</tr>
<tr>
<td>CPU</td>
<td>19m</td>
<td>22m</td>
<td>3.4h</td>
<td>2h</td>
</tr>
</tbody>
</table>

7 CONCLUSION

There are several limitations of the current work, which we plan to address in our future work. First, the current score-and-search algorithm outputs a point estimate of the causal graph with no uncertainty quantification and no global convergence guarantee. We plan to develop a fully Bayesian approach by assigning sparse priors (i.e., spike-and-slab priors on $\beta$’s) and carrying out posterior inference via the Markov chain Monte Carlo. Second, we have empirically assessed the identifiability of the proposed OCD for multivariate data and for bivariate data with unmeasured confounders. The identifiability theory for multivariate categorical data or bivariate categorical data with unmeasured confounders is in general lacking in the causal discovery literature. Third, we have not explicitly addressed the problem of choosing the number $L$ of categories in data discretization. We picked $L = 3$ for genomic data by convention and assessed its robustness up to $L = 10$. For non-genomic data, there is no obvious/universal choice of $L$. Instead of picking a specific $L$, we have tested the proposed OCD in a range of values. In the future, we plan to propose data-driven ways (e.g., via BIC) to objectively choose $L$.

Acknowledgements

Ni’s research was partially supported by National Science Foundation (DMS-2112943 and DMS-1918851). Mallick’s research was partially supported by TRIPods National Science Foundation (CCF-1934904) and National Cancer Institute of the National Institutes of Health (R01CA194391).

References


