Graphical Causal Models with Discretized Data and Background Information

Nalan Baştürk[1], Chumasha Rajapakshe[1], and Rui Jorge Almeida[1,2]

[1] Department of Quantitative Economics, School of Business and Economics, Maastricht University, P.O. Box 61, 6200 MD Maastricht, The Netherlands
c.rajapakshe@alumni.maastrichtuniversity.nl, n.basturk@maastrichtuniversity.nl

[2] Department of Data Analytics and Digitilisation, School of Business and Economics, Maastricht University, P.O. Box 61, 6200 MD Maastricht, The Netherlands rj.almeida@maastrichtuniversity.nl

Abstract. In several application areas, discretized variables represent an underlying continuous variable. For example, the level of certain medical measures can be 'low', 'medium' or 'high', while the underlying measure is a continuous variable. The estimation of graphical causal models for data with discretized variables leads to biased estimates and underestimated causal relations. In this work, we study the effect of incorporating background information on causal relations when estimating causal models with discretized variables. We show that incorporating background information on the relations between variables improves graphical causal model estimates in case of discretized variables. We find particularly large gains in reducing omitted causal relations and in estimating causal relations correctly. We relate these improvements to the hyperparameter choice in graphical causal models and properties of the variables in the model.

Keywords: Causal discovery · Discretized data · Graphical causal models · Mixed data.

1 Introduction

Graphical causal models have been developed to estimate causal relations within the data without making explicit assumptions about the direction of causality [21, 27]. These models have been applied successfully in economics, psychology and genetics, among other areas [6, 27, 23, 4] and extended to incorporate mixed (discrete and continuous) variables [25, 10, 29]. In many applications with mixed variables, discrete variables represent an underlying continuous variable. For instance, the level of a disease in an individual's medical records can be categorized as 'mild', 'moderate' or 'strong' while the disease intensity prior to categorization is a continuous variable. For linear and non-linear regression, discretization of variables increases statistical uncertainty, causes biased and inconsistent causal effect estimates, and spurious relations between variables [17, 1, 28]. The effects

of discretization on the estimation of graphical causal models have been empirically studied, showing that causal relations tend to be underestimated and that causal effect estimates tend to be biased [11]. Data discretization has also been studied in Bayesian networks which model probabilistic relationships between discretized and remaining variables [8, 12]. Data discretization in Bayesian networks rely on distributional assumptions, and a discretization policy [8, 2].

We study the effects of incorporating background information on the estimation of causal relations in a dataset with discretized and continuous variables. The uncertain causal relations are estimated using graphical causal models, which do not require distributional assumptions or a discretization policy. Incorporating background information, represented as probability distributions, is shown to improve parameter estimates in other models such as latent growth models and multilevel regressions [18, 9]. Methods have been developed to incorporate background information in graphical causal models [13, 7, 22]. In these methods, background information, added as restrictions on causal links, is shown to improve estimates of causal relations in graphical causal models [19, 24]. To the best of our knowledge, the effects of including background knowledge on estimates of graphical causal models with discretized data have not been studied in the literature.

The effects of incorporating prior knowledge on estimation results are relatively straightforward in conventional models such as linear regression. We first illustrate the implications of incorporating background information in a linear regression model with discretized data, where background information is expressed as probability distributions for model parameters. For more complex models such as graphical causal models, derivation of these effects is more involved. Therefore, we empirically analyze the implications of incorporating prior beliefs in graphical causal models. In this case, background information is represented as causal links between part of the variables. We use the results of the illustration to set up the empiricial analysis and discuss the findings. We find that the inclusion of these priors increases the number of correctly estimated causal relations. This result holds particularly for correlated variables in the model. However, the specified background information does not lead to improvements in the bias of the estimated causal effects.

2 Graphical Causal Models

Graphical causal models aim to capture the causal structure of multivariate data using a graph structure G, defined as the ordered pair $< V, E >$ for a set of vertices V and a set of edges, E, representing the variables [26, 27]. An undirected graph indicates which pairs of vertices in E are correlated and a directed acyclic graph (DAG) represents the causal relations between variables [21]. DAGs are often estimated using the PC algorithm [15], which estimates a completed partially directed acyclic graph (CPDAG). The estimated CPDAG includes directed and undirected edges representing the conditional independence of the variables.

Given the estimated CPDAGs, causal effects between variables can be obtained using the intervention when the DAG is absent (IDA) method [16, 7].

2.1 Graphical Causal Models for Mixed Data

The PC algorithm first estimates the causal relations through the correlation matrix of the variables with the sequence of independence tests. Let $X_{k,n}$ denote the variables included in the graphical causal model for $k = 1, \ldots, K$ variables and $n = 1, \ldots, N$ observations. In case of mixed data, estimating $\rho_{k,l}$, the correlation between variables k and l, is not straightforward. Therefore a kernel based correlation estimation is proposed [5, 10]. We consider, $\mathbf{K}()$, the kernel density estimator, in the form of a radial basis function for continuous variables:

$$\mathbf{K}_k(X_{k,n_1}, X_{k,n_2}) = \exp\left(-(X_{k,n_1} - X_{k,n_2})^2 / \left(2\sigma^2\right)\right), n_1, n_2 = 1, \ldots, N. \quad (1)$$

For categorical variables, the following kernel function is used:

$$\mathbf{K}_k(X_{k,n_1}, X_{k,n_2}) = h_\theta(P(X_{k,n_1})) \times I(X_{k,n_1} = X_{k,n_2}), \quad (2)$$

where $P(x)$ is the probability that a categorical variable X takes the value x, $h_\theta(x) = (1 - x^\theta)^{\frac{1}{\theta}}$ and $I()$ is the indicator function that takes the value 1 if its argument is true and 0 otherwise. The correlation between variables are then calculated using the kernel alignment method [10, 14]:

$$\hat{\rho}_{k,l} = \frac{\sum_{n_1,n_2=1}^{N} \mathbf{K}_k(X_{k,n_1}, X_{k,n_2}) \mathbf{K}_l(X_{k,n_1}, X_{l,n_2})}{\sqrt{\left(\sum_{n_1=1}^{N} \mathbf{K}_k(X_{k,n_1}, X_{k,n_2})^2\right)\left(\sum_{n_2=1}^{N} \mathbf{K}_l(X_{l,n_1}, X_{l,n_2})^2\right)}}. \quad (3)$$

The causal relations are estimated through a sequence of conditional independence test for each variable k, l, conditioning on the remaining variable set S:

$$H_0(k, l|S) : \rho_{k,l|S} = 0, \quad H_A(k, l|S) : \rho_{k,l|S} \neq 0, \quad (4)$$

where the rejection of H_0 indicates a causal relation between variables:

$$|Z_{k,l|S}|\sqrt{N - |S| - 3} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right); \text{ for } Z_{k,l|S} = \frac{1}{2}\log\left(\frac{1 + \hat{p}_{k,l|S}}{1 - \hat{p}_{k,l|S}}\right), \quad (5)$$

for a given significance level α. The conditioning set S is updated by eliminating links between edges iteratively according to (4) and (5).

The next step in the PC algorithm is the DAG or CPDAG estimation which aims at identifying causal relations. In this step, CPDAGs represent all possible DAGs that satisfy the dependence estimates in (4). For all CPDAGs, the directions between edges are estimated as follows: The algorithm considers v-structures, defined as all triplets (k, l, m), where k and l are adjacent, l and m are adjacent, but k and m are not adjacent. For all such triplets, both edges are directed towards l if and only if m was not part of the conditioning set that made

the edge between k and l drop out. In addition, the remaining directions are found by iteratively applying the PC algorithm [14]. Estimated causal relations can be represented by a $K \times K$ adjacency matrix $\hat{A}$:

$$\hat{A}_{k,l} = I(\hat{\rho}(X_k, X_l) \neq 0). \tag{6}$$

Note that the obtained DAGs are not unique, as several causal relations can lead to the same conditional independence relationships in (6).

2.2 Graphical Causal Models with Background Information

A modification of the IDA method has been proposed to incorporate background knowledge [7, 22], which allows the manual inclusion and exclusion of edges based on prior knowledge. This extension uses a set of constraints denoted as B. The subset B_n for $n = 1, \ldots, N$ restricts the relationship between variables, preventing one variable from being an ancestor of another, while the subset B_d for $d = 1, \ldots, N$ focuses on direct causal information. By incorporating these constraints, the causal inference process aligns better with known background knowledge, guiding the exploration of causal relationships [7].

We consider background information in the form of restrictions on the adjacency matrix $A_{k,l}$ in (6). If node indicators k, l are included in the subset in the restricted set, the connection between the edge $k - l$ is never considered for removal. Hence $\hat{A}_{k,l} = 1$ in (6). These restrictions affect the conditioning set S in the sequential conditional correlation tests in (4) and reduce the number of viable DAGs derived from the CPDAG G^*. Consistency between the constraints B and the CPDAG G^* is achieved when at least one DAG within the equivalence class of G^* adheres to the specified constraints. Once a possible CPDAG with background information is obtained, all possible effects between variables are estimated using the IDA method summarized in Algorithm 1 [16, 7]. IDA enumerates all possible causal effects of variable $X \in \{X_1, \ldots, X_K\} \setminus Y$ on variable $Y \in \{X_1, \ldots, X_K\}$ by listing all possible parental sets of X, $\text{pa}(X, G^*)$, and siblings of X, $\text{sib}(X, G^*)$, defined as variables with directed and undirected edges with X, respectively.

Algorithm 1: Modified IDA algorithm [7, 22]

Require: A CPDAG G^*, a target variable Y
Ensure : $\{\Theta_X\}_{X \in V}$, where Θ_X stores all possible causal effects of X on Y
for *each variable* $X \in V$ **do**
  Set $\Theta_X = \emptyset$;
  for *each* $S \subseteq sib(X, G^*)$ *such that orienting* $S \to X$ *and* $X \to sib(X, G^*) \setminus S$ *does not introduce any v-structure collided on* X **do**
    Estimate the causal effect of X on Y by adjusting for $S \cup \text{pa}(X, G^*)$, and add the causal effect to Θ_X;
return $\{\Theta_X\}_{X \in V}$

3 Illustration of Incorporating Background Information in Models with Discretized Data

Studying the effects of discretized variables and background knowledge on graphical causal models theoretically is not straightforward. We therefore illustrate the effects of discretization and background information with a linear regression model. The background information expressed as prior beliefs for the model coefficients. The data generation process (DGP) is as follows:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{7}$$

where Y is the $N \times 1$ vector of dependent variable, X_1, X_2 are $N \times 1$ vectors of independent variables generated from standard normal distributions, and $\varepsilon \sim N(0, I_N)$ is the $N \times 1$ vector of residuals where I is the $N \times N$ identity matrix.

Next, we define a discretized dummy variable X_3 that represents the underlying continuous X_2. We set the elements of X_3 as $X_{n,3} = 1$ for observation n if $X_{n,2} > 0$ and $X_{n,3} = 0$ otherwise. The regression in (7) with the discretized variable is:

$$Y = X\beta + \eta, \tag{8}$$

where $X = (X_1, X_3)$, $\beta = (\beta_1, \beta_2)'$, and $\eta = (\beta_2 X_2 - \beta_3 X_3 + \varepsilon)$ is the error term of the linear regression with the discretized variable.

Consider the following normal prior distribution on the parameters: $\beta \sim N\left(\underline{\beta}, \tau^{-2} I_2\right)$. The posterior of the model in (8) is:

$$p(\beta|Y, X) \propto \exp\left(-\frac{1}{2}\left((X'X + \tau^{-2}I)^{-\frac{1}{2}}(Y'X + \tau^2\underline{\beta})\right)^2\right) \tag{9}$$

which is equivalent to $p(\beta|Y, X) = N(\bar{\beta}, \bar{V}_\beta)$. The posterior mean is a weighted average of the OLS estimate and the prior:

$$\bar{\beta} = (X'X + \tau^{-2}I)^{-\frac{1}{2}}(Y'X + \tau^2\underline{\beta}) = (X'X + \tau^2 I)^{-1}(X'X\hat{\beta}_{\text{OLS}} + \tau^2\underline{\beta}). \tag{10}$$

The difference between the posterior mean and true coefficients is:

$$\bar{\beta} - \beta^\star = (X'X + \tau^2 I)\left(X'X(\hat{\beta}_{\text{OLS}} - \beta^\star) + \tau^2(\underline{\beta} - \beta^\star)\right) \tag{11}$$

where $\beta^\star = (\beta_1, \beta_2)'$ and the bias of the OLS estimator is [11]:

$$\hat{\beta}_{\text{OLS}} - \beta^\star = (X'X)^{-1}\begin{pmatrix} X_3' X_3 X_1'\eta - X_1' X_3 X_3'\eta \\ X_1' X_1 X_3'\eta - X_1' X_3 X_1'\eta \end{pmatrix}, \tag{12}$$

i.e. correlation of variables X_1, X_3 indicate that both the discretized and continuous variables are biased due to discretization. From (12) and the weighted average property in (11), it is straightforward that bias is reduced when the prior satisfies the inequality $\underline{\beta} - \beta^\star < \hat{\beta}_{\text{OLS}} - \beta^\star$, but the weighted effect of this reduction depends on the correlation of data, represented by $X'X$ in (11). In this illustration, a possible way to include background information on causal relations, instead of causal effects is a limiting case:

$$\lim_{\tau \to \infty} \bar{\beta} = \lim_{\tau \to \infty} (X'X + \tau^2 I)^{-1}\tau^2\underline{\beta} = \underline{\beta}. \tag{13}$$

4 Effects of Background Knowledge on Graphical Causal Models with Discretized Data

Graphical causal models have been applied to mixed data successfully in the literature [25, 10, 29]. However, in case of mixed data, discretization is shown to be problematic in graphical causal models. In this case, graphical causal models lead to wrong causal estimates and a higher bias [11]. In graphical causal models, the effects of incorporating prior beliefs, or similar background information, on the obtained results are more complex than the illustration in Section 3. This complexity is due to the graphical causal model estimation, as well as the property that the causal estimate is not necessarily unique [21, 16, 15]. We study the effects of incorporating background information empirically, where background information is represented as potential links between variables.

The graphical causal model in Section 2 is estimated in two steps. First, the causal relations represented in the adjacency matrix in (6) are estimated. Second, the causal effects between variables are estimated based on the estimated adjacency matrix. Thus, including background information can affect the estimated the causal relations, i.e. the graph structure, or the estimated causal effects at the second stage. To differentiate these two effects, we first report the effects of background information in the estimated adjacency matrices. We next report the effects of adding background information on the obtained causal effects.

4.1 Simulation Setup

For the simulation study, we consider a linear regression as the underlying DGP:

$$Y_n = \beta_0 + \sum_{k=1}^{K} \beta_k X_n^k + \varepsilon_n, \tag{14}$$

where $\{\beta_0, \ldots, \beta_K\}$ are the model parameters, Y_n for $n = 1, \ldots, N$ are the observed output variables, X_n^k for $k = 1, \ldots, K$ are the input variables and the error terms have the following distribution $\varepsilon_n \sim NID(0, \sigma^2)$. The graphical causal model in Section 2 is defined over $K + 1$ variables $X^1, \ldots, X^K, Y$, where we use the notation Y to clarify the endogenous variable in the simulations. The adjacency matrix for this DGP is given in Figure 1.

We simulate data with $n = 150$ observations in a sample, and $K = 9$ input variables, and all coefficients are fixed as $\beta_0 = \ldots = \beta_9 = 5$. Thus the network has $p = 10$ nodes. We generate X^k variables with different properties: X_n^k for $k = 1, 2, 3$ are discrete data, while the remaining variables are continuous. An important distinction in the model is the difference between discrete and discretized data. The discrete data are defined by

$$X_n^k \sim \text{Bernoulli}(0.5), \text{ for } k = 1, 2, 3, \forall n. \tag{15}$$

Generating the outcome variable in (14) with these discrete variables in (15) indicate that the correct model is in the set of models considered in the graphical

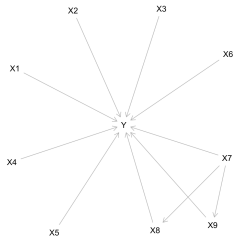

Fig. 1. True underlying DAG for the simulated data according to (14).

causal model estimation. On the other hand, the discretized data, X_n^k for $k = 1, 2, 3$ are generated from a continuous variable, $\tilde{X}_n^k$:

$$\tilde{X}_n^k \sim NID(0,1),\ X_n^k = I(\tilde{X}_n^k > 0),\ \text{for } k = 1,2,3, \forall n, \tag{16}$$

where the DGP in (14) is adjusted with $\tilde{X}_n^k$:

$$Y_n = \beta_0 + \sum_{k=1}^{3} \beta_k \tilde{X}_n^k + \sum_{k=4}^{9} \beta_k X_n^k + \varepsilon_n, \forall n, \tag{17}$$

where $\tilde{X}_n^k$ is not an input to the graphical causal model. Hence, the DGP in (17) is not in the consideration set of the graphical causal model. The remaining variables X^k for $k = 4, 5, 6$ are data from independent normal distributions, and X^k for $k = 7, 8, 9$ are generated from a dependent normal distribution:

$$X_n^k \sim NID(0,1),\ \text{ for } k = 4, \ldots, 7, \forall n \tag{18}$$

$$X_n^k = X_n^7 + \varepsilon_n^k,\ \varepsilon_n^k \sim NID(0,1),\ \text{ for } k = 8, 9, \forall n. \tag{19}$$

These different specifications for X variables aim to study the estimation of graphical causal models in discrete data, discretized data, continuous data and continuous data with correlations. Furthermore, we replicate each simulation 100 times to reduce the effect of simulation noise.

4.2 Effects of Background Knowledge on Adjacency Matrices

We present the effects of including background information on the estimated causal relations, namely the estimation of the adjacency matrix, between different types of variables, with and without discretization and two types of background information. For the graphical causal model estimation, we consider three significance levels $\alpha = \{0.2, 0.5, 0.8\}$. Intuitively, the optimized graph is expected to have more connections with a relatively low α. Results with and without background information are obtained using the PC algorithm [27, 3] and fast causal inference (FCI) algorithm [20], respectively.

We analyze the effects of discretization and including background information through five scenarios: (1) 'no bg' corresponds to discretized data with no

Table 1. Mean TP, FP, FN, TN for causal effects between all, discrete (dis), mixed, continuous independent (cts-indep), dependent (cts-indep) variables without background knowledge (left), with priors (middle panels) and without discretization (right).

	no bg				inacc bg				acc bg				no dis. acc bg			
	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN	TP
$\alpha = 0.2$																
all	69.3	9.7	9.6	1.4	77.7	1.3	5.0	6.0	79.0	0.0	5.3	5.7	79.0	0.0	5.7	5.3
dis	5.7	0.3	0.0	0.0	5.9	0.1	0.0	0.0	6.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0
mixed	44.3	6.7	3.0	0.0	50.5	0.5	0.8	2.2	51.0	0.0	0.7	2.3	51.0	0.0	1.1	1.9
cts.indep	5.8	0.2	0.0	0.0	6.0	0.1	0.0	0.0	6.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0
cts.dep	13.5	2.5	6.6	1.4	15.4	0.6	4.2	3.8	16.0	0.0	4.6	3.4	16.0	0.0	4.5	3.5
$\alpha = 0.5$																
all	69.1	9.9	10.1	0.9	76.1	2.9	4.9	6.1	79.0	0.0	5.1	5.9	79.0	0.0	5.3	5.7
dis	5.6	0.4	0.0	0.0	5.9	0.1	0.0	0.0	6.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0
mixed	44.5	6.5	3.0	0.0	49.2	1.8	0.6	2.4	51.0	0.0	0.6	2.4	51.0	0.0	0.8	2.1
cts.indep	5.4	0.6	0.0	0.0	5.8	0.2	0.0	0.0	6.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0
cts.dep	13.6	2.4	7.1	0.9	15.2	0.8	4.3	3.7	16.0	0.0	4.4	3.6	16.0	0.0	4.4	3.6
$\alpha = 0.8$																
all	71.6	7.4	10.5	0.5	74.2	4.8	4.7	6.3	79.0	0.0	4.7	6.3	79.0	0.0	5.0	6.0
dis	5.7	0.3	0.0	0.0	5.6	0.4	0.0	0.0	6.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0
mixed	46.2	4.8	2.9	0.1	48.1	2.9	0.7	2.3	51.0	0.0	0.6	2.4	51.0	0.0	0.7	2.3
cts.indep	5.6	0.4	0.0	0.0	5.5	0.6	0.0	0.0	6.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0
cts.dep	14.1	1.9	7.6	0.4	15.0	1.0	4.0	4.0	16.0	0.0	4.1	3.9	16.0	0.0	4.3	3.7

Table 2. MSE and variance of the estimated versus true effects for, discrete (dis), mixed, continuous independent (cts-indep), dependent (cts-indep) variables without background information (left), with background information (middle panels) and without discretization (right).

	no bg		inacc bg		acc bg		no dis. acc bg	
	mean	var	mean	var	mean	var	mean	var
$\alpha = 0.2$								
all	0.62	0.64	1.20	0.91	1.18	0.93	0.27	0.34
dis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
mixed	0.82	0.81	0.91	0.73	0.87	0.81	0.26	0.36
cts.indep	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
cts.dep	0.65	0.72	2.56	1.85	2.54	1.74	0.48	0.54
$\alpha = 0.5$								
all.1	0.71	0.82	1.26	1.06	1.22	1.03	0.39	0.55
dis.1	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
mixed.1	0.89	0.90	0.95	0.85	0.97	0.91	0.29	0.46
cts.indep.1	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
cts.dep.1	0.81	1.19	2.65	2.17	2.51	1.92	0.83	1.10
$\alpha = 0.8$								
all.2	0.78	0.91	1.19	1.07	1.12	0.92	0.43	0.63
dis.2	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
mixed.2	0.88	0.91	0.93	0.86	0.84	0.67	0.31	0.45
cts.indep.2	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
cts.dep.2	1.09	1.50	2.46	2.16	2.39	1.98	0.94	1.40

background information, (2) 'inacc bg' corresponds to discretized data with inaccurate background information, (3) 'acc bg' corresponds to discretized data estimated with relatively accurate background information, (4) 'no dis. acc bg' corresponds data without discretization and with accurate background information. Data generations for the four cases are given in (14)–(19).

The accuracy of the background information is defined according to the probability that the background information indicates a correct link between variables and the probability that it indicates an incorrect link between variables. The background information puts restrictions on the adjacency matrix:

$$\begin{aligned} A_{i,10} \text{ for } i = 1, \ldots, 9 &= \begin{pmatrix} 1 & \text{with probability } p_{TP} \\ \text{NA} & \text{with probability } 1 - p_{TP} \end{pmatrix} \\ A_{i,k} \text{ for } i = 1, \ldots, 8,\ k > i &= \begin{pmatrix} 1 & \text{with probability } p_{FP} \\ \text{NA} & \text{with probability } 1 - p_{FP} \end{pmatrix} \end{aligned} \tag{20}$$

i.e. the background information provides the true links as in the DGP in (14) with probability p_{TP}, and it indicates wrong links, which are not part of the DGP in (14) with probability p_{FP}. When including background information, only the NA (missing) elements of the adjacency matrix are estimated using the graphical causal model. For the case of inaccurate background information, we set $p_{TP} = 0.8$ and $p_{FP} = 0.2$, where the background information is still stronger for indicating true links, but it also has a probability to indicate non-existing links. For the case of accurate background information, we set $p_{TP} = 0.8$ and $p_{FP} = 0$, hence the background information never indicates non-existing links.

Table 1 summarizes the obtained causal relations over $M = 100$ simulation replications. The entries indicate correctly and incorrectly estimated causal links calculated according to the difference between the true adjacency relations $A_{k,l}$ in Figure 1 and the estimated adjacency relations $\hat{A}_{k,l}$ in (6):

$$\begin{aligned} \text{TP} &= \sum_{k=1}^{K} \sum_{l \neq k} I(\hat{A}_{k,l} = 1, A_{k,l} = 1), \quad \text{FP} = \sum_{k=1}^{K} \sum_{l \neq k} I(\hat{A}_{k,l} = 1, A_{k,l} = 0), \\ \text{TN} &= \sum_{k=1}^{K} \sum_{l \neq k} I(\hat{A}_{k,l} = 0, A_{k,l} = 0), \quad \text{FN} = \sum_{k=1}^{K} \sum_{l \neq k} I(\hat{A}_{k,l} = 0, A_{k,l} = 1). \end{aligned} \tag{21}$$

The comparison of the first and last four columns in Table 1 shows the effects of discretization compared to the no discretization case. True positives and true negatives are smaller or equal under discretization for almost all variables and α values. The inclusion of background information, in the form of accurate or inaccurate information, improve correctly identified relations (TP) as well as the correctly excluded relations (TN) compared to the discretized case without prior information. This result holds particularly for mixed variable connections, i.e. connections between discretized and continuous variables including the outcome variable Y. Hence the correct DGP in (14) is more likely to be attained even under relatively inaccurate background information.

Next, we report the results of incorporating background information based on variable types. Table 1 shows that the confusion matrix elements improve under background information particularly for mixed data (connections between discrete and continuous data) and for connections between continuous and dependent variables (X^7, X^8, X^9, Y). CM elements for discretized variables are close to each other in all considered cases. This indicates that even without background information, the links between discretized variables are accurately estimated. In this simulation, discretized variables do not correlate with other variables. We conjecture that this lack of correlation is the reason for the discretized variables' CMs to be better than those of the continuous correlated variables.

An important observation from Table 1 is that the true positives and false negatives are exactly zero in all cases for discrete (dis) and continuous independent (cts.indep) variables. These results arise from our simulation setup. The DGP has no causal links between discrete variables and between continuous dependent variables. Therefore, the estimation results cannot indicate a true positive or a false negative. Finally, the effect of the hyperparameter α on the obtained results is minimal. With priors, a smaller α, $\alpha = 0.2$ leads to slight improvements in the confusion matrix elements, but there is no general link between the value of α and improvements in TP or TN.

4.3 Effects of Background Knowledge on Causal Effects

The second step in estimating graphical causal models is to obtain causal effects. Causal effects are estimated based on the links between variables according to the estimated adjacency matrices in Section 4.2. Note that estimated causal estimates are not unique in graphical causal models. The conventional method is to consider the minimum and maximum values of causal relations based on the graph estimates, and report the lower and upper bounds of these estimates [16, 15]. We report the effects of incorporating prior beliefs on the mean squared differences between estimated and true causal effects for all model parameters.

We report four cases to compare the effects of discretization as well as the inclusion of background information: (1) 'no bg' corresponds to discretized data with no background information, (2) 'inacc bg' corresponds to discretized data with inaccurate background information, (3) 'acc bg' corresponds to discretized data estimated with relatively accurate background information, (4) 'no dis. no bg' corresponds to data without discretization and without background information. Data generations for the four cases are given in (14)–(19). The accuracy of the background information is defined as in (20).

Table 2 presents average MSEs and its variance across 100 simulation replications. MSE values in the first and last panels (no bg vs no dis. acc bg) are substantially different with much higher values for discretized data without background information (no bg). This indicates the bias arising from data discretization, which is in line with the literature [11]. For discrete and continuous independent variables, the estimation bias is close to zero in all scenarios.

5 Conclusion

In this work, we study the effect of incorporating background information in graphical causal models with discretized variables. We show that incorporating prior beliefs on the relations between variables improves graphical causal model estimates with a particular reduction in omitted causal relations in estimates and an increase in correctly identified causal relations. The inclusion of background information, in inaccurate or accurate form, increase the overall bias in all variables, and specifically in continuous dependent variables. This indicates that adding background information improves the estimation of links between variables, but has a deteriorating effect on the estimation bias in graphical causal models. Furthermore, the bias in continuous dependent variables increases with background information since our priors do not indicate relations between continuous and correlated variables. A potential solution to this is to include background information in the form of priors on model parameters. Finally, uncertainty in background information can be explicitly considered in combining background information and data information.

Acknowledgements. N. Baştürk is partially supported by an NWO grant number 195.187.

References

1. Barnwell-Ménard, J.L., Li, Q., Cohen, A.A.: Effects of categorization method, regression type, and variable distribution on the inflation of type-i error rate when categorizing a confounding variable. Statistics in Medicine **34**(6), 936–949 (2015)
2. Cobb, B.R., Rumí, R., Salmerón, A.: Bayesian network models with discrete and continuous variables. Advances in Probabilistic Graphical Models pp. 81–102 (2007)
3. Colombo, D., Maathuis, M.H., et al.: Order-independent constraint-based causal structure learning. Journal of Machine Learning Research **15**(1), 3741–3782 (2014)
4. Cornelisz, I., Cuijpers, P., Donker, T., van Klaveren, C.: Addressing missing data in randomized clinical trials: A causal inference perspective. PloS One **15**(7), e0234349 (2020)
5. Cristianini, N., Shawe-Taylor, J., Elisseeff, A., Kandola, J.: On kernel-target alignment. Advances in Neural Information Processing Systems **14** (2001)
6. Elwert, F.: Graphical causal models. In: Handbook of Causal Analysis for Social Research, pp. 245–273. Springer (2013)
7. Fang, Z., He, Y.: IDA with background knowledge. In: Peters, J., Sontag, D. (eds.) Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI). Proceedings of Machine Learning Research, vol. 124, pp. 270–279. PMLR (03–06 Aug 2020)
8. Friedman, N., Goldszmidt, M., et al.: Discretizing continuous attributes while learning Bayesian networks. In: ICML. pp. 157–165 (1996)
9. Gao, Y., Kennedy, L., Simpson, D., Gelman, A.: Improving multilevel regression and poststratification with structured priors. Bayesian Analysis **16**(3), 719 (2021)

10. Handhayani, T., Cussens, J.: Kernel-based approach to handle mixed data for inferring causal graphs. arXiv preprint arXiv:1910.03055 (2019)
11. Hanoch, O., Baştürk, N., Almeida, R.J., Habtewold, T.D.: Analysis of graphical causal models with discretized data. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 223–234. Springer (2022)
12. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs, vol. 2. Springer (2007)
13. Johnson, S.R., Tomlinson, G.A., Hawker, G.A., Granton, J.T., Feldman, B.M.: Methods to elicit beliefs for bayesian priors: a systematic review. Journal of Clinical Epidemiology **63**(4), 355–369 (2010)
14. Kalisch, M., Bühlman, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research **8**(3) (2007)
15. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. Journal of Statistical Software **47**(11), 1–26 (2012)
16. Maathuis, M.H., Kalisch, M., Bühlmann, P.: Estimating high-dimensional intervention effects from observational data. The Annals of Statistics **37**(6A), 3133–3164 (2009)
17. Maxwell, S.E., Delaney, H.D.: Bivariate median splits and spurious statistical significance. Psychological Bulletin **113**(1), 181 (1993)
18. McNeish, D.M.: Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using m plus. Journal of Educational and Behavioral Statistics **41**(1), 27–56 (2016)
19. Meek, C.: Causal inference and causal explanation with background knowledge. arXiv preprint arXiv:1302.4972 (2013)
20. Mooij, J.M., Magliacane, S., Claassen, T.: Joint causal inference from multiple contexts. The Journal of Machine Learning Research **21**(1), 3919–4026 (2020)
21. Pearl, J., Verma, T.S.: A statistical semantics for causation. Statistics and Computing **2**(2), 91–95 (1992)
22. Perkovic, E., Kalisch, M., Maathuis, M.H.: Interpreting and using cpdags with background knowledge (2017), arXiv preprint arXiv:1707.02171
23. Rohrer, J.M.: Thinking clearly about correlations and causation: Graphical causal models for observational data. Advances in Methods and Practices in Psychological Science **1**(1), 27–42 (2018)
24. Scheines, R., Spirtes, P., Glymour, C., Meek, C., Richardson, T.: The tetrad project: Constraint based aids to causal model specification. Multivariate Behavioral Research **33**(1), 65–117 (1998)
25. Sokolova, E., Groot, P., Claassen, T., Rhein, D.v., Buitelaar, J., Heskes, T.: Causal discovery from medical data: dealing with missing values and a mixture of discrete and continuous data. In: Conference on Artificial Intelligence in Medicine in Europe. pp. 177–181. Springer (2015)
26. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review **9**(1), 62–72 (1991)
27. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, Prediction, and Search. MIT press (2001)
28. Thoresen, M.: Spurious interaction as a result of categorization. BMC Medical Research Methodology **19**(1), 1–8 (2019)
29. Zhong, W., Dong, L., Poston, T.B., Darville, T., Spracklen, C.N., Wu, D., Mohlke, K.L., Li, Y., Li, Q., Zheng, X.: Inferring regulatory networks from mixed observational data using directed acyclic graphs. Frontiers in Genetics **11**, 8 (2020)