Causal Discovery from Episodic Data

Osman Mian[°], Sarah Mameche[°], Jilles Vreeken

CISPA Helmholtz Center for Information Security
osman.mian@cispa.de, sarah.mameche@cispa.de, jv@cispa.de

Abstract

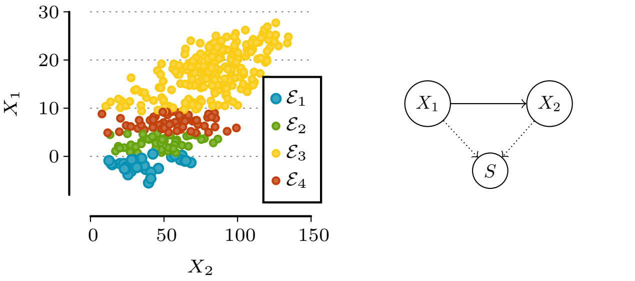
In numerous real-world applications ranging from long-term medical studies to environmental monitoring, data does not arrive in a single batch but accumulates over time, so that learning has to occur in sequential episodes. As each episode preferentially includes samples from a specific time domain, it may be subject to selection bias, posing significant challenges for modeling and adapting the causal structure underlying the observed variables. We address this problem by proposing an approach for learning a fully directed causal graph over a set of observed variables progressively from episodical data. Crucially, we maintain a set of valid candidate models at each given time point, towards converging to the true model when we have access to enough episodes to be representative of the overall distribution. Central to this approach is an information-theoretic perspective of causality which allows us to assess the quality of a causal model for describing a given dataset and hence can effectively guide when and how to update our candidate models as new data arrives. Initial experiments on synthetic data showcase that our approach works well in practice.

Introduction

Determining causality given observational data is a fundamental question throughout the sciences (Pearl 2009). The traditional paradigm in causal learning assumes a single, homogeneous dataset sampled from a single, stationary distribution. In various domains, however, we obtain data over time in multiple episodes.

Consider an application in environmental monitoring, where we take measurements of two markers X_1 and X_2 over the course of a year, with a new batch of measurements arriving each season. For example, in Fig. 1 (left) we show data over X_1: *Temperature* and X_2: *Ozone Level* (Mooij et al. 2014). We obtain observations in episodes $\{\mathcal{E}_1, ...\mathcal{E}_4\}$ at subsequent timepoints $\{t_1, ...t_4\}$ corresponding to winter, spring, summer, and autumn seasons, respectively, and color samples according to the season.

In our example taken from the Tübingen cause-effect pairs (Mooij et al. 2014), there exists a causal function or mechanism that generates X_2 from X_1, and the overall data suggest a roughly linear trend (Fig. 1). However, consider that we first access data from the winter season (blue). From $\mathcal{E}_1$ alone, it appears that the variables are uncorrelated. The same is true for $\mathcal{E}_2$ and $\mathcal{E}_4$, and only when we add $\mathcal{E}_3$ to the picture can we see a linear increasing trend. That is, each episode can be thought of as a snapshot representing only a subregion of the data, and the corresponding data distribution may be subject to selection bias. Combining all snapshots by obtaining sufficiently many samples from each season is essential to obtain a complete picture.



(a) Data over X_1, X_2 arriving in episodes $\mathcal{E}_1$–$\mathcal{E}_4$ over time.

(b) Causal model over X_1, X_2 with selection variable S.

Figure 1: We consider the variables X_1: *Temperature* and X_2: *Ozone Level* from the Tübingen cause-effect pairs (Mooij et al. 2014). When we obtain the data in sequential episodes corresponding to each season (left), both variables appear uncorrelated (e.g. $\mathcal{E}_1$), and only when we have a complete picture ($\mathcal{E}_1$–$\mathcal{E}_4$) can we see the causal trend. To model that each episode is a subsample from a region of the data, we introduce a selection variable S (right).

Classical methods for causal discovery (Spirtes et al. 2000; Chickering 2002; Pearl 2009) commonly build on the assumption that all data is identically distributed. As a consequence, they would need access to samples over all episodes to return a consistent result, as well as need to relearn a causal model from the pooled data whenever a new episode arrives. Not only is this computationally impractical, but it is also difficult to verify whether we observed all relevant episodes, as selection bias is not straightforward to identify unless we make additional assumptions (Kaltenpoth

[°] These authors contributed equally.

and Vreeken 2023). Moreover, a domain expert may want to gain preliminary insights into the causal relationship between X_1 and X_2 already based on data from the earlier episodes, as well as be ready to update these insights as new data becomes available.

Motivated by this, we consider the problem of discovering a causal model progressively from episodical data, addressing both its theoretical and algorithmic aspects. On the modeling front, we propose a causal model with episodical selection bias, which we depict in Fig. 1 (right). Novel to our work is the addition of a categorical latent selection variable S, which in our example models the seasonality effect and takes a different value $S = s_k$ for each particular season s_i. Conditioning on s_i has the effect of selecting samples from a region of the data, i.e. considering a biased distribution $P(X \mid S = s_k)$. We obtain an identically distributed dataset $P(X)$ in the limit, once all seasons have been observed.

In practice, however, the values of S need not exactly correspond to our episodes $\mathcal{E}$, such as when we obtain episodes $\mathcal{E}_1 - \mathcal{E}_{12}$ in monthly time intervals in our example. That is, we may not know how many selectors s_k exist and which subregion s_k of the data in a given episode $\mathcal{E}_i$ represents. Furthermore, episodes may arrive in an unknown order from different subregions, for example, when we obtain data in batches from different hospitals $\mathcal{E}_i$ that represent different subpopulations s_k.

To address these challenges, we develop a practical algorithm to discover the causal model in an online fashion. The main idea is that, instead of enforcing a single causal model that may be biased and needs to be re-learned upon arrival of each episode, we allow *multiple* plausible causal models. The reasoning is that these candidates provide insights into the system at a given point in time and can eventually be merged into a single causal hypothesis. This warrants a strategy to decide, upon arrival of each new episode, whether the causal model over this episode is sufficiently similar to an existing candidate or should be included as a new hypothesis. To do this, we turn to an information-theoretic view of causal networks which allows us to assess whether a given model accurately and concisely describes a given dataset. We develop this approach in the remainder of this work.

Contributions To summarize our main contributions, we

- introduce a framework for causal modeling of data from time-dependent episodes under selection bias and show how the algorithmic model of causation can be used to decide whether the same or a different causal model applies across episodes,
- develop a practical approach for episodical learning of fully directed causal networks,
- conduct initial experiments to test our method in practice.

We structure our exposition according to the above, first introducing notation and preliminaries that we build on, then introducing our causal model, information-theoretic perspective, and practical algorithm, and concluding with an experimental evaluation and discussion.

Preliminaries

We begin by introducing notation, describing our problem setting, and outlining causal modeling techniques for independent and identically distributed (i.i.d.) data that we build on and extend to episodical data.

Problem Setting Throughout our work, we consider a sequence $\{\mathcal{E}_0, ...\mathcal{E}_n\}$ of datasets associated with timepoints $\{t_0, ...t_n\}$. We consider a batch setting where we obtain data in subsequent episodes, where $\mathcal{E}_i$ arrives at timepoint t_i.

In all episodes, we measure a fixed set of continuous random variables $X = \{X_1, ...X_m\}$ with distribution $P^{(i)}(X)$ in episode $\mathcal{E}_i$, having overall distribution $P(X)$.

Novel to our work is that we do *not* assume that $P^{(i)}(X)$ is i.i.d. with respect to $P(X)$. That is, there could be *selection bias* where the episode $\mathcal{E}_i$ is not representative of the population X. As selection bias results from preferential inclusion of samples based on a downstream causal variable (Kaltenpoth and Vreeken 2023), we will address it using a causal model over X and $\mathcal{E}$. Before we state our model, we introduce the relevant concepts and assumptions in causal discovery that we build on.

Causal Discovery

As is common, we assume that the causal model over the observed variables X is given as a directed acyclic graph (DAG) $\mathcal{G} = (X, E)$ with node set X and edges $(i, j) \in E$ whenever the variable X_i is a cause of X_j (Pearl 2009). We write the set of direct causes of X_j as pa_j. In particular, we assume that there is a single, fixed DAG over all episodes, with the reasoning that the underlying causal connections between observed variables remain the same.

We call a causal model *identifiable* when we can uniquely determine it from a purely observational distribution (Pearl 2009). Identifiability of the causal DAG $\mathcal{G}$ is only possible under additional assumptions. Hence, we assume causal *sufficiency*, which states that no latent variable jointly causes any of the observed variables, as well as the *Markov* and *faithfulness* conditions, which together imply that separations in the graphical model $\mathcal{G}$ correspond to independence constraints in the observed distribution P. Under these assumptions, it is well known that identifiability holds up to the Markov Equivalence Class (MEC) of $\mathcal{G}$ (Hauser and Bühlmann 2013).

Identification of causal directions beyond the MEC is only possible when we obtain additional information about how the system reacts to interventions (Hauser and Bühlmann 2014; Zhang et al. 2017; Mameche, Kaltenpoth, and Vreeken 2023) or make additional assumptions, such as restricting the functional dependencies to nonlinear functions with additive noise (Bühlmann et al. 2014; Hoyer et al. 2009; Marx and Vreeken 2021). The latter is the approach we will follow in this work. In particular, we take an information-theoretic view that we describe next.

Algorithmic Framework of Causality

The algorithmic model of causation (Janzing and Schölkopf 2010) reasons about the complexity of causal mechanisms in describing the observed data. To this end, it uses the concept of Kolmogorov complexity. Kolmogorov complexity defines, for binary strings $x \in \{0,1\}^*$, the length $K(x)$ of the shortest binary program x^* that outputs x and halts. Similarly, for a distribution P, the Kolmogorov complexity $K(P)$ corresponds to the length of the shortest program p^* approximating P up to a given precision q on a universal Turing machine $\mathcal{U}$ on input $\langle x, q\rangle$ (Li and Vitányi 2009),

$$K(P) = \min_{p^* \in \{0,1\}^*} \{|p^*| : \mathcal{U}\langle x, q\rangle - P(x)| \leq \tfrac{1}{q}\} \; .$$

Using Kolmogorov complexity, we can formalize the idea that truly causal mechanisms provide concise explanations of the data distributions.

Algorithmic Markov Condition We now turn to the centerpiece of the algorithmic view of causal networks, namely the Algorithmic Markov Condition (AMC) (Janzing and Schölkopf 2010). It postulates that causal mechanisms correspond to *programs* that encode the observed distributions most concisely in terms of Kolmogorov complexity. More precisely, the AMC posits that each causal mechanism f_j for X_j can be described as a program p_j that generates the distribution $P(X_j \mid pa_j)$ independently from all other variables, such that the complexity of the overall distribution $K(P(X))$ corresponds to the summed complexities over all causal conditionals $K\big(P(X_j \mid pa_j)\big)$ up to a program of constant length,

$$K(P(X)) = \sum_{j=1}^{m} K\big(P(X_j \mid pa_j)\big) \tag{1}$$

where equality holds up to an additive constant. Simply stated, AMC postulates that the true underlying causal network compresses data the most. Several recent approaches have used this formulation to infer the most likely causal hypothesis from observational data, and we give a brief outline of the common approach to doing so.

Causal Discovery using AMC As Kolmogorov complexity cannot be computed for arbitrary programs (Li and Vitányi 2009), but can be approximated from above via Minimum Description Length (Grünwald 2007) for a known model class, a common approach is to instantiate Eq. (1) for a fixed model class and using it to discover the causal model. There are different approaches to doing so, addressing, for example, the bivariate case (Marx and Vreeken 2019), latent confounding (Kaltenpoth and Vreeken 2019), and multi-context data (Mameche, Kaltenpoth, and Vreeken 2022; Mian, Kamp, and Vreeken 2023). While our approach can in principle work together with any of these instantiations, we will build on the GLOBE framework (Mian, Marx, and Vreeken 2021) as it provides a general framework and efficient algorithm to discovering fully directed causal networks, although only from i.i.d. data.

GLOBE models causal functions through non-parametric multivariate regression. It considers the model class consisting of causal DAGs where each edge is modeled through a nonlinear function with additive noise describing an effect from its cause. The complexity of the data under a given model $\mathcal{M}$ is measured through its MDL score $L(X, \mathcal{M})$.

In accordance with Eq. (1), this leads to the objective

$$\begin{aligned} \mathcal{M}^* &= \arg\min_{\mathcal{M}} L(X, \mathcal{M}) \\ &= \arg\min_{\mathcal{M}} \Big(L(\mathcal{M}) + \sum_j L(X_j \mid pa_j, \mathcal{M}) \Big) \; . \end{aligned}$$

The idea is that the MDL score of a dataset X together with its causal model $\mathcal{M}$ is given by the length, in bits, of encoding the model itself, $L(\mathcal{M})$, as well as the cost of the data given this model. The latter term decomposes according to the causal factorization, that is, we describe each variable X_j from its causal parents pa_j in $\mathcal{M}$. This amounts to encoding the non-parametric regression function associated with the causal relationship. We refer to Mian, Marx, and Vreeken (2021) for the formal definitions of L. Given that the above objective involves a super-exponential search space over all DAG models for a given set of variables, Mian, Marx, and Vreeken (2021) develop the greedy algorithm GLOBE to discover fully directed causal DAGs in practice.

While GLOBE is consistent for an i.i.d. dataset (Mian, Marx, and Vreeken 2021), these guarantees do not apply in the episodical case that we study. We hence adapt the causal model and algorithm to episodical data in the following.

Theory

In this section, we introduce our causal model for episodical data with selection bias.

Causal Model for Episodical Data

We assume that each episode corresponds to a subpopulation or subregion of the data, such as seasons or age groups. To model this, we define episodical selection bias as conditioning on an unobserved variable S. Hereby S is a categorical variable taking values $S = \{s_1, ...s_K\}$ such that in each episode $\mathcal{E}_i$, the value $S = s_k$ remains fixed. The variable S can be thought of as grouping data into subregions corresponding to biased subsamples, where each episode will be drawn from one of the subregions.

We assume that there are no unobserved variables except the selection variable S. We state this assumption as follows.

Assumption 1 (Episodical selection bias) *We assume that an unobserved categorical variable S with only ingoing edges, $X_j \rightarrow S$ for some $X_j \in X$, so that*

1. *In each episode $\mathcal{E}_j$, $P^{(j)}(X \mid S = s_k)$ is identically distributed for some specific $s_k \in \{s_1, ...s_K\}$,*
2. *The data over all episodes $P(X) = \bigcup_k P(X \mid S = s_k)$ is identically distributed.*

The first part of the assumption asserts that the bias within each episode $\mathcal{E}_j$ is fixed. The second part asserts that we eventually observe enough episodes to remove the selection effect, e.g. $P(X)$ is representative of the population. We treat S as hidden, that is, we know neither how many subpopulations and thus values $\{s_1, ...s_K\}$ of the variable S exist, nor in which episode we observe which subpopulation, that is, for which $\mathcal{E}_i$ and which s_k condition 1.1 holds, nor how many episodes we need to observe to obtain an unbiased distribution where assumption 1.2 holds. Armed with this, we are ready to state our causal model.

Assumption 2 (Causal model for episodical data) *Our causal model over X and $\mathcal{E}$ is given by a DAG $\mathcal{G}$ and categorical variables S such that for each X_j,*

$$X_j = f(pa_j, N_j), \qquad N_j \perp\!\!\!\perp X_j \ ,$$

where each X_j is an unbiased sample with distribution $P(X) = \cup_k P^{(k)}(X) = \cup P(X \mid S = s_k)$.

The above states that each variable X_j is generated as a function of its causal parents pa_j and independent noise N_j. We can state the algorithmic Markov condition for this model as follows.

Postulate 1 (Algorithmic Markov Condition) *A causal DAG $\mathcal{G}$ over X and $\mathcal{E}$ is only acceptable if*

$$\begin{aligned} K(P(X)) &= \sum_{j=1}^{m} K\big(P(X_j \mid pa_j)\big) \\ &= \sum_{j=1}^{m} K\left(\bigcup_k P(X_j \mid pa_j, S = s_k)\right) \end{aligned}$$

where = holds up to an additive constant.

While the overall distribution P factorizes according to the causal DAG, our selection variable splits P into potentially biased subsets $P^{(k)}$ that our episodes subsample from. This motivates our problem of discovering the model at a given point in time where possibly only a subset of the relevant episodes were observed. We present our approach to solving the following problem in the next section.

Problem Statement *Given continuous variables X and episodes $\mathcal{E}$, where we observed a subset $\mathcal{E} = \{\mathcal{E}_0, ...\mathcal{E}_i\}$ at a given timepoint t_i, we aim to discover a set of candidate causal DAGs $\mathcal{G}$ over X that are valid at timepoint t_i.*

We describe our approach to this problem next.

Method

In this section, we introduce our algorithm ECD for **E**pisodic **C**ausal **D**iscovery.

To motivate our algorithm setup, let us revisit our motivating example where we obtain data from episodes arriving throughout a year with a hidden seasonality effect. Let us think of our data as split into two distributions S_- and S_+ representing low and high temperatures, respectively, where episodes $\mathcal{E}_1, \mathcal{E}_2$, and $\mathcal{E}_4$ are sampled from S_- and the episode $\mathcal{E}_3$ from S_+, each providing a snapshot of the true generating process. As S_- is biased and our theoretical guarantees for DAG learning do not apply, the causal model M_- learned from episodes of S_- may be biased. Once we obtained enough episodes from *both* S_-, S_+, however, we access an *unbiased* distribution, so that a model M^* learned over the pooled data will be guaranteed to be accurate in the limit of samples. This motivates a strategy where we learn an initial model from the first episode E_0 from S_-, test whether this model can extrapolate to the subsequent episodes from S_-, and update it accordingly. When we obtain an episode from S_+, this may initially lead to a different candidate model M_+, resulting in two distinct models being maintained, but once we observe enough episodes, we conjecture to merge both models to obtain the underlying M^*.

Algorithm 1: ECD $(\mathcal{A}, \mathcal{E})$

```
   input : MDL-based causal discovery algorithm $\mathcal{A}$,
           episodes $\mathcal{E}$ arriving over time
   output: candidate causal models $\mathcal{M}$
 1 $\mathcal{M} \leftarrow \{\}$
 2 $\tau \leftarrow 0$
 3 $\tau_{\text{max}} \leftarrow k$
 4 while a new episode $\mathcal{E}_i$ arrives do
 5     $\mathcal{M} \leftarrow$ UPDATE $(\mathcal{E}_i, \mathcal{M}, \mathcal{A})$
 6     $\tau \leftarrow \tau + 1$
 7     if $\tau == \tau_{max}$ then
 8         $\mathcal{M} \leftarrow$ MERGE $(\mathcal{M}, \mathcal{A})$
 9         $\tau \leftarrow 0$
10     end
11     yield $\mathcal{M}$
12 end
```

Following this idea, we devise an approach where we first learn potentially incomplete models from the currently available episodes and, given that we do not know when we arrived at an unbiased distribution nor in which order episodes arrive from which selector, allow a set of candidate models at any time. To progressively merge models together when appropriate, we consider the description length of an episode under different models. We merge a new model for an incoming episode $\mathcal{E}_i$ with an existing model whenever the new model does not significantly improve upon the description length of a previous model.

To do so, we assume an information-theoretic learner $\mathcal{A}$ that allows discovering a fully directed DAG model M from a given dataset $\mathcal{E}_i$, as well as outputs its description length $L(M)$. While in principle, this could be any MDL-based method, we instantiate $\mathcal{A}$ with GLBOE (Mian, Marx, and Vreeken 2021) unless otherwise stated.

We show the pseudocode of ECD in Alg. 1. We use M_0 to initialize a set of candidate models $\mathcal{M} = \{M_{\mathcal{S}}\}$ that we maintain throughout, where we write each candidate as $M_{\mathcal{S}}$, subscripted by a set $\mathcal{S} = \{E_{s_1}, ...E_{s_n}\}$ of episodes that it represents. Starting from the first episode E_0, as new episodes arrive, we test and update the current list of models at each timestep using the UPDATE function (Line 5). The tolerance parameter τ keeps track of timesteps since the last model merge (Line 6). After a prespecified number of timesteps, we try and merge existing models (Line 8). This continues perennially.

The main component of ECD is the UPDATE function shown in Alg. 2, which updates the list of models each time a new episode comes in. It does so as follows: we first learn a new model M using the learner $\mathcal{A}$ and record its compression score (Lines 1-3). We then evaluate whether there already is a competing model $M_{\mathcal{S}_j}$ that compresses the new episode E_i as effectively as M (Lines 5-8). We can straightforwardly derive a hypothesis test for this based on no-hypercompression (Grünwald 2007). We use this no-

Algorithm 2: UPDATE $(\mathcal{A}, \mathcal{E}_i, \mathcal{M})$

```
   input : MDL-based causal discovery algorithm $\mathcal{A}$,
           episode $\mathcal{E}_i$ at timepoint $t_i$,
           current set of candidate models $\mathcal{M}$
   output: Updated candidate models $\mathcal{M}$
1  $M \leftarrow \mathcal{A}.\text{LEARN}(\mathcal{E}_i)$
2  $M^\star \leftarrow M$
3  $L^\star \leftarrow M^\star.\text{SCORE}(\mathcal{E}_i)$
4  foreach $M_\mathcal{S} \in \mathcal{M}$ corresponding to episode set $\mathcal{S}$
    do
5      $L \leftarrow M_\mathcal{S}.\text{SCORE}(\mathcal{E}_i)$
6      if $L < L^\star$ then
7          $L^\star \leftarrow L$
8          $M^\star \leftarrow M_\mathcal{S}$
9      end
10 end
11 if $M^\star$ is $M$ then
12     $M_{\{\mathcal{E}_i\}} \leftarrow M.\text{ADDDATA}(\mathcal{E}_i)$
13     $\mathcal{M} \leftarrow \mathcal{M} \cup M_{\{\mathcal{E}_i\}}$
14 else
15     $M^\star.\text{ADDDATA}(\mathcal{E}_i)$
16 end
```

hypercompression test to decide if M has a statistically different compression than the existing best-case $M_\mathcal{S}$. If so, we merge the episodes together to obtain a pooled dataset $\mathcal{S}' = \mathcal{S} \cup \{E_i\}$ (Line 15) and obtain $M_{\mathcal{S}'}$ and discard the model M. Otherwise, we include M as a new candidate model (Lines 12-13). The MERGE procedure (Alg. 1, Line 8) works analogously to our UPDATE, except that we repeatedly test models learned between pairs of existing models, and keep the joint model if it improves overall compression.

Overall, we conjecture that as the joint distribution over all received episodes gets closer to the true distribution, our approach will converge to a single causal hypothesis while maintaining useful models throughout the process.

Related Work

While many approaches address correcting for selection bias in a train and test setting with covariate shift (Gretton et al. 2008; Sugiyama, Krauledat, and Müller 2007), we are not aware of previous work that addresses episodical selection bias while accounting for the causal structure among the observed variables. As our main focus is causal discovery, we give an overview of the existing literature here.

Discovering causal relationships from observational data is an actively studied problem with applications in almost all areas of science. Classical approaches typically fall into the categorizations of constraint-based methods, such as PC (Pearl 2009), or score-based methods, such as GES (Chickering 2002; Ramsey et al. 2017). As these approaches discover a Markov Equivalence Class (MEC) of the causal model (Hauser and Bühlmann 2013), recent approaches study under which assumptions we can determine causal directions beyond the MEC. One line of work does so by constraining the functional model (Peters et al. 2014; Bühlmann et al. 2014), such as LiNGAM (Shimizu et al. 2006) which assumes linear non-Gaussian models. Another branch of work builds on the algorithmic model of causality (Janzing and Schölkopf 2010). However, all aforementioned methods assume that no selection effects are present.

To address selection bias, recent work shows how to identify *whether* selection bias holds in a single dataset (Kaltenpoth and Vreeken 2023), but we aim to address *how* to adapt causal discovery methods in the presence of it. Besides this, we study a different form of selection bias where episodes are biased datasets representing a fixed selection effect and the overall distribution is unbiased, which warrants a different problem setup than the general selection bias in Kaltenpoth and Vreeken (2023) or the outcome-specific selection bias studied by Zhang et al. (2016).

Finally, a wealth of recent literature has studied causal discovery from different environments, experimental regimes, or contexts; examples include the JCI framework (Mooij, Magliacane, and Claassen 2016) and multi-group Lingam (Shimizu 2012) for discovering causal DAGs from multi-context data. While the episodes we study are reminiscent of environments, they are different for two reasons. First, existing work assumes that environments are obtained in a single batch, whereas we specifically address the case where we obtain episodes in turn which comes with new algorithmic challenges. More importantly, the modeling assumptions of both cases are different. Each environment is typically assumed to be identically distributed (i.i.d.) while there can be distribution shifts between environments. In contrast, each episode can represent a biased sample, but episodes follow an identical distribution when combined. This difference becomes apparent in the causal model, where environments can be modeled by a categorical variable with edges *towards* observed variables (Mooij, Magliacane, and Claassen 2016), whereas we model episodes using an *outgoing* edge. That is, an *exogenous* categorical variable can represent regime changes with distribution shifts, whereas *conditioning* on a categorical variable as we do corresponds to taking snapshots and considering a biased subsample of the distribution.

To demonstrate how classical and environment-based causal discovery approaches fare with episodical selection bias in practice, we next compare them against ECD.

Evaluation

In this section we discuss results of initial experiments that we have conducted using ECD.

Experimental Setup Since to the best of our knowledge, there is no specific algorithm designed for causal discovery from episodic data, we look at the nearest possible modifications of existing algorithms for comparison. As baseline we compare to LINGAM (Shimizu et al. 2006), RESIT (Peters et al. 2014) and GES (Chickering 2002; Ramsey et al. 2017). We modify these algorithms as follows — we first learn a causal network over each individual incoming episode of data. As taking an intersection (resp. union) over networks for each episode might be too conservative (resp. prone to false positives), we compute the predicted causal

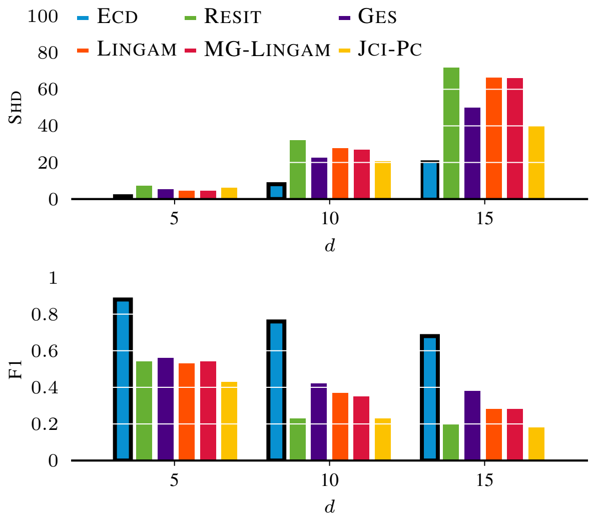


Figure 2: SHD [Top, lower is better] and Orientation F1 score [Bottom, higher is better] for networks learned over episodical i.i.d. data for networks sizes $d \in \{5, 10, 15\}$. ECD discovers causal networks better than the baselines.

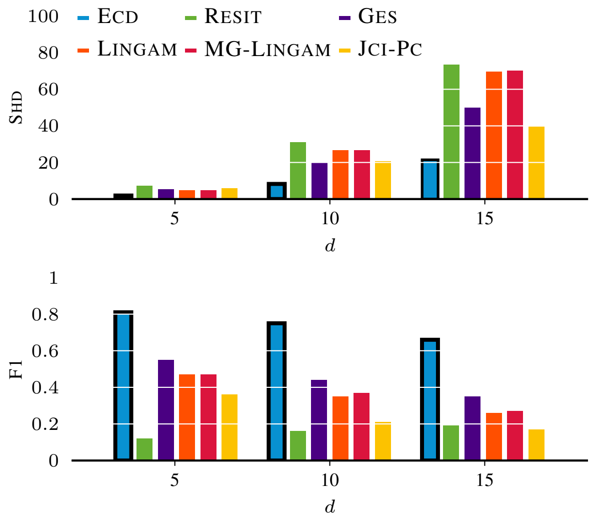


Figure 3: SHD [Top, lower is better] and Orientation F1 score [Bottom, higher is better] for networks learned over episodical data with selection bias for networks sizes $d \in \{5, 10, 15\}$. ECD has a slightly degraded performance compared to the i.i.d. case, but still outperforms the competition.

network using a majority vote over edges learned across each episode. In addition to baselines, we compare to multi-environment causal discovery approaches such as the JCI-framework (Mooij, Magliacane, and Claassen 2016) using the PC algorithm (Spirtes et al. 2000), as well as Multi-Group Lingam (MG-LINGAM) (Shimizu 2012). The latter two approaches, however, require that all episodes are available to learn the causal network. Hence, we provide all episodes in one go to these approaches. This constitutes an advantage as they can learn from complete data from the start. To measure the quality of the predicted causal structures we use the Structural Hamming Distance (SHD) (Kalisch and Bühlmann 2007) as well the F1 score over edge orientations. SHD counts the number of edges where the predicted causal network differs from the true causal network, whereas the F1 score gives us a sense of how accurately we determine the correct edge orientations in terms of precision and recall.

For each of the proposed experiment setups, we generate random graphs using Erdős-Rényi model for network sizes $\{5, 10, 15\}$, and generate data for effects using functions of the following form,

$$X_i = \sum_{x \in pa_i} f(x) + \mathcal{U}(-1, 1),$$

where $f(x)$ is either a polynomial function or a combination of sine and cosine functions defined over each parent $x \in pa_i$ of X_i, and $\mathcal{U}$ denotes uniform noise. For each graph/function combination, we generate a total of $10,000$ samples and then split them into 10 different episodes of size 1000 each. We *transmit* these episodes to each algorithm one at a time. After each episode, we note the updated causal network for each of the methods. As JCI-PC and MG-LINGAM are provided all 10 episodes in one go, we only measure the performance over the final network for them. For our preliminary findings, we ask the following three questions.

1. Can ECD reliably discover causal networks when the incoming episodes are all i.i.d?
2. How well does ECD perform when the incoming episodes have selection bias?
3. How does ECD's performance change over the course of arriving episodes?

We subsequently investigate each of these questions.

With i.i.d. data We begin with evaluating the behavior of ECD when the i.i.d. assumption holds for all incoming data. This means that each of the 10 episodes that are transmitted over time is sampled from the same distribution. We measure the final predicted network at the end of the last episode and report the results in Fig. 2, where we see that overall ECD has the lowest SHD among other approaches and the highest orientation F1 score among the same, resulting in overall best performance. This implies that ECD not only finds the correct causal skeleton, but also orients the edges with high precision. In contrast, we see that while LINGAM for 5 variables has the second best SHD, it performs worse than GES in terms of edge orientation as indicated by a lower F1 score in contrast to GES.

With selection bias After testing that our proposed approach works reliably in the i.i.d. setting, we introduce selection bias within each episode next. To do so, we first sort all data on one randomly chosen variable before splitting it into 5 disjoint datasets, resulting in five different *seasons* $\{s_1, .., s_5\}$. Next, we shuffle the samples within each disjoint dataset and split it into 2 further parts, leaving us again with a total of 10 *episodes*. This induces a distribution shift across

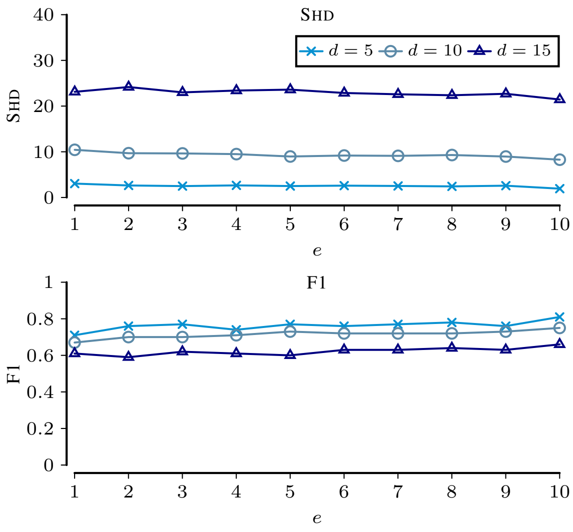


Figure 4: SHD [Top, lower is better] and Orientation F1 score [Bottom, higher is better] over increasing number of episodes, e for data with selection bias. ECD always finishes at a better SHD resp. F1 score from the one it starts from.

different episodes, with a total of five different environment distributions present in the incoming stream and exactly two episodes per environment.

We report the results for this experiment in Fig. 3 where ECD results in a slightly worse performance than the i.i.d. case, but still beats the competition by a clear margin. Other approaches except RESIT, also degrade slightly. RESIT, while surprisingly having a comparable SHD to the i.i.d. case, ends up incorrectly orienting the edges as is reflected by a drop in F1 score.

Performance over multiple episodes As the third step in our preliminary evaluations, we measure how the performance of ECD evolves over multiple arriving episodes. For i.i.d. data, we found that both SHD and F1 remained consistent across 10 episodes, indicating the expected behavior that there is no drastic change in the learned model over time. For data with selection bias we show the results in Fig. 4. We observe that overall there is a slight improvement in performance with more episodes. In each of the cases, we always end up with a lower (resp. higher) SHD (resp. F1 score) compared to the start. The improvement in performance, although slow, suggests that the model is indeed updated along the way. One explanation for the slower improvement trend could be that 1000 samples per episode are already informative enough to learn most of the structure despite distribution shifts and therefore, the room for improvement across episodes is small.

In addition to network accuracy, we also keep a record of the average number of models present inside ECD across episodes and report the results in Fig. 5. For the biased-data case, we see that more models are created initially, but on average we always merge back to around 2 models. For the i.i.d. case, we observed that the number of models on average is between 1 and 2, indicating that new models are not created so often. This is the expected behavior since each incoming data then comes from the same distribution.

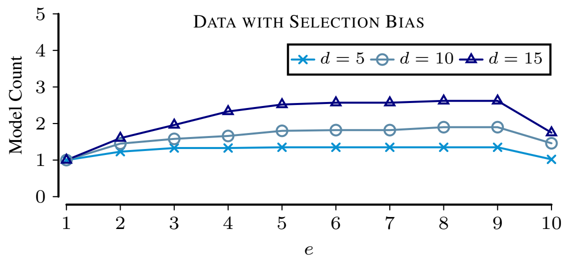


Figure 5: Model Count [lower is better] over increasing number of episodes e for data with selection bias for network sizes $d \in \{5, 10, 15\}$. The model count is never greater than 4 because ECD is able to merge similar models as we receive more episodes.

Discussion and Conclusion

In this paper we proposed an approach for the discovery of causal networks for data arriving over multiple episodes, containing selection bias. To achieve our goal we used an information theoretic approach utilizing the Algorithmic Markov Condition postulate (Janzing and Schölkopf 2010).

Initial results show that our method performs reliably both in the case of i.i.d. as well as in data with selection bias. Encouraged by these results, we are currently exploring more challenging experimental settings such as lower sample sizes and larger number of episodes to see how well our method adapts to changing data. In our current setup, the selection bias is introduced using an existing variable within the datasets. As next steps, we would like to make the setting more complex and conduct experiments where data consist of selection bias induced by an unobserved variable.

We currently assume that the causal network across each episode remains fixed. This model does not allow for interventions in incoming data. This might be a restrictive assumption in practice as there may be episodes that contain either implicit or explicit interventions resulting in data being generated from an interventional distribution. Intuitively, our proposed approach should be modifiable to adapt and converge to two different models for observational resp. interventional distributions. As next steps, we aim to investigate how we can incorporate handling of both observational and interventional data into ECD. Furthermore, we are also investigating how to adapt our approach in case of mechanism shifts, i.e. when the underlying generating mechanism between variables changes, without changing the underlying causal connections.

Finally, despite the initially promising results, we assessed our method's efficacy only in terms of its empirical performance. While ECD clearly outperforms the existing approaches, we still need to theoretically justify its soundness. We are currently working on proofs to provide theoretical guarantees for our approach, in the limit where the number of episodes $|\mathcal{E}| \to \infty$.

References

Bühlmann, P.; Peters, J.; Ernest, J.; et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals Stat.*, 42(6): 2526–2556.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *JMLR*, 3: 507–554.

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2008. Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press. ISBN 9780262255103.

Grünwald, P. 2007. *The Minimum Description Length Principle*. MIT Press.

Hauser, A.; and Bühlmann, P. 2013. Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Statist. Soc. B*, 77.

Hauser, A.; and Bühlmann, P. 2014. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4): 926–939.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *NeurIPS*, volume 21. Curran.

Janzing, D.; and Schölkopf, B. 2010. Causal Inference Using the Algorithmic Markov Condition. *IEEE TIT*, 56(10): 5168–5194.

Kalisch, M.; and Bühlmann, P. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *JMLR*, 8(Mar): 613–636.

Kaltenpoth, D.; and Vreeken, J. 2019. We Are Not Your Real Parents: Telling Causal from Confounded using MDL. In *SDM*, 199–207. SIAM.

Kaltenpoth, D.; and Vreeken, J. 2023. Identifying Selection Bias from Observational Data.

Li, M.; and Vitányi, P. 2009. *An Introduction to Kolmogorov Complexity and its Applications*. Springer.

Mameche, S.; Kaltenpoth, D.; and Vreeken, J. 2022. Discovering Invariant and Changing Mechanisms from Data. In *KDD*, 1242–1252. ACM.

Mameche, S.; Kaltenpoth, D.; and Vreeken, J. 2023. Learning Causal Mechanisms under Independent Changes. In *NeurIPS*.

Marx, A.; and Vreeken, J. 2019. Identifiability of Cause and Effect using Regularized Regression. In *KDD*. ACM.

Marx, A.; and Vreeken, J. 2021. Formally Justifying MDL-based Inference of Cause and Effect. *arXiv preprint arXiv:2105.01902*.

Mian, O.; Kamp, M.; and Vreeken, J. 2023. Information-theoretic causal discovery and intervention detection over multiple environments. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI-23*.

Mian, O.; Marx, A.; and Vreeken, J. 2021. Discovering fully oriented causal networks. In *AAAI*.

Mooij, J. M.; Magliacane, S.; and Claassen, T. 2016. Joint causal inference from multiple contexts. *JMLR*, 21.

Mooij, J. M.; Peters, J.; Janzing, D.; Zscheischler, J.; and Scholkopf, B. 2014. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *ArXiv*, abs/1412.3773.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal Discovery with Continuous Additive Noise Models. *JMLR*, 15.

Ramsey, J.; Glymour, M.; Sanchez-Romero, R.; and Glymour, C. 2017. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *J. Data Sci. Anal.*

Shimizu, S. 2012. Joint estimation of linear non-Gaussian acyclic models. *Neurocomputing*, 81.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *JMLR*, 7.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT Press.

Sugiyama, M.; Krauledat, M.; and Müller, K.-R. 2007. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J. Mach. Learn. Res.*, 8: 985–1005.

Zhang, K.; Huang, B.; Zhang, J.; Glymour, C.; and Schölkopf, B. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI*.

Zhang, K.; Zhang, J.; Huang, B.; Schölkopf, B.; and Glymour, C. 2016. On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection. In *UAI*, 825–834.