Learning Causal Networks from Episodic Data

Osman Mian°
osman.mian@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Sarah Mameche°
sarah.mameche@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Jilles Vreeken
vreeken@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

ABSTRACT

In numerous real-world domains, spanning from environmental monitoring to long-term medical studies, observations do not arrive in a single batch but rather over time in episodes. This challenges the traditional assumption in causal discovery of a single, observational dataset, not only because each episode may be a *biased* sample of the population but also because multiple episodes could *differ* in the causal interactions underlying the observed variables. We address these issues using notions of context switches and episodic selection bias, and introduce a framework for causal modeling of episodic data. We show under which conditions we can apply information-theoretic scoring criteria for causal discovery while preserving consistency. To in practice discover the causal model progressively over time, we propose the CONTINENT algorithm which, taking inspiration from continual learning, discovers the causal model in an online fashion without having to re-learn the model upon arrival of each new episode. Our experiments over a variety of settings including selection bias, unknown interventions, and network changes showcase that CONTINENT works well in practice and outperforms the baselines by a clear margin.

CCS CONCEPTS

• **Mathematics of computing** → **Causal networks**; **Information theory**; • **Theory of computation** → *Online algorithms.*

KEYWORDS

Causal Discovery, Continual Learning, Selection Bias

ACM Reference Format:
Osman Mian, Sarah Mameche, and Jilles Vreeken. 2024. Learning Causal Networks from Episodic Data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671999

1 INTRODUCTION

Determining causality is of fundamental interest throughout the sciences [30]. As controlled experiments are often not feasible, the question of how to do so given observational data alone is gaining increased attention. Classical algorithms for discovering causal networks assume as their starting point a single, homogeneous dataset sampled from a single, stationary distribution [6, 30, 39].

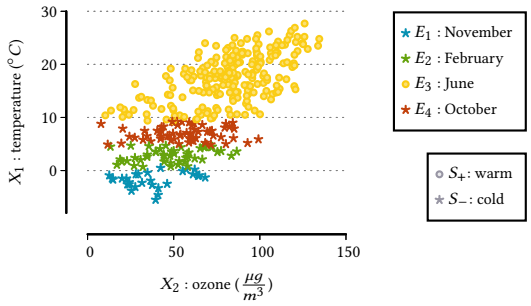

Figure 1: Cause X_1 and effect X_2 [29] measured in episodes over time (E_1-E_4). Each episode comes from an underlying season (S_+, S_-), and an unknown context, here Switzerland.

However, a more realistic setting is one where we obtain observations in batches over time. Not only does this mean that we need to learn and update our causal hypothesis over time, but each batch likely contains samples from a specific time period or subpopulation, resulting in a biased distribution. Even the collective data distribution over *all* such episodes is often not identically distributed since the causal interactions could differ across domains.

To motivate the episodic setting and illustrate its challenges, consider an example in environmental monitoring where we measure two markers X_1: *temperature* and X_2: *ozone concentration* at different times of the year. Suppose we obtain monthly measurements, resulting in episodes $\{E_1, ...E_4\}$ at timepoints $\{t_1, ...t_4\}$ as shown in Fig. 1. In our example taken from the Tübingen cause-effect pairs [29], X_1 is considered the cause of X_2 and the overall data suggest a roughly linear trend of the causal mechanism relating them. Considering a winter month such as E_1 (blue) on its own would however suggest that both variables are uncorrelated. Only when including the summer month E_3 (yellow) do we obtain a complete picture. In this example, there is a high-temperature season S_+ (circle) as well as a low-temperature season S_- (star), and episodes coming from only one such season offer a biased picture.

This simplistic example suggests that combining all episodes is a good practice to remove seasonal bias. This however can lead to its own set of issues. Consider a dataset that stems from a different geographical region or context, where due to local measuring devices noise levels are different, or even the underlying causal relationship changes. For instance, a phenomenon known as ozone suppression [42] creates a situation where ozone levels are no longer

°Both authors contributed equally to this research.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671999

positively correlated with temperature. As ozone suppression only occurs above a certain temperature threshold, it is not visible in the data obtained in Switzerland shown in Fig. 1 but could affect a dataset E_5 from a region with exceptionally high temperatures. Overall, whereas episodes $E_1 - E_4$ should be combined to remove seasonal bias, combining samples from different contexts $E_1 - E_5$ obscures context-specific causal relationships [44].

While recent work in causal discovery considers different contexts [28, 40, 44], it neither addresses episodes nor allows for structural changes in the causal model across contexts. In contrast, we propose a causal modeling framework for episodes with selection bias where an unknown number of causal networks underlie the data-generating process. We show that in principle, we can use a consistent scoring criterion for causal discovery in this setting so long as we observe sufficiently many episodes.

From a practical perspective, existing algorithms for causal discovery [6, 27, 30] start from a single batch of data and hence would need to relearn the causal model whenever a new episode arrives. This is clearly computationally impractical; rather, it would be desirable for a domain expert to gain preliminary insights into the causal relationships based on some earlier episodes and perenially update these as new data becomes available.

Given these limitations, we develop the Continent algorithm for discovering causal models over episodic data, more specifically multiple fully directed causal networks over a set of contexts. Taking inspiration from continual learning, we hereby avoid fully re-learning the causal model upon the arrival of each episode but learn it in an online fashion. We propose a strategy to update the causal hypothesis as new episodes arrive, using distribution matching and an information-theoretic perspective of causality, and show that our updating strategy is consistent. We show in experiments that Continent discovers causal networks reliably from data with episodic selection bias, under interventions, as well as with structural changes in causal networks. Not only does it compare favorably to its competitors, but only Continent is able to learn the causal model adaptively over time. It can also address an experimental setting where we assign a new, unseen episode to one of the causal networks inferred from previous episodes.

Contributions. To summarize our main contributions, we

- introduce a causal modeling framework for episodic data,
- show under which conditions we can use an information-theoretic consistent scoring criterion to identify a set of causal networks underlying such data,
- develop the practical approach Continent to learn such causal networks in a continual fashion,
- confirm in experiments that Continent works in practice.

We structure our exposition according to the above, first introducing notation and preliminaries, then introducing our causal model and practical algorithm, and concluding with an experimental evaluation and discussion.

2 PRELIMINARIES

First, we outline our problem setting and review causal modeling techniques for independent and identically distributed (i.i.d.) data.

2.1 Notation and Problem Setting

Throughout our work, we consider a batch setting where we obtain observations as a sequence of datasets $\{E_0, \ldots, E_N\}$ at timepoints $\{t_0, \ldots, t_N\}$, and refer to dataset E_i at time t_i as an *episode*. We denote the dataset that combines all episodes up to time t_i as $D_N = \cup_{i=1}^{N} E_i$. In each episode, we observe a fixed set of continuous random variables $X = \{X_1, \ldots, X_M\}$ with distribution $P(X)$.

Episodes can belong to different domains or environments, which we call *contexts* denoted by $\{C_0, \ldots, C_R\}$. Each episode E_i is a member of a unique context C_r, which we write as $C(E_i)$, and we write X^r, P^r to refer to variables, resp. distributions, in the rth context. Novel to our work is that we neither know how many contexts R exist nor which context $C(E_i)$ each episode comes from.

In addition to coming from different contexts, episodes are not necessarily i.i.d. but rather could preferentially include samples from a certain *subpopulation* S. To illustrate, consider the warm season S_+ in Fig. 1. Episode E_3 exhibits selection bias in that it only includes i.i.d. samples from this season. We can represent S_+ through a binary variable S with values $S = \circ$, $S = *$, with the interpretation that samples $S = \circ$ are observed, $S = *$ are missing from E_3, so that it follows a biased distribution. In general, we consider a categorical variable S with values $\{s_1, \ldots, s_K\}$ modeling subpopulations of $P(X)$, so that each episode results from selecting an unknown population s_k and sampling from $P(X \mid S = s_k)$. As we could obtain multiple episodes from the same subpopulation, for example repeated monthly episodes over multiple years, we do not assume the number $K \leq N$ to be known a priori.

In this episodic setting, we want to discover how many and which causal models there are.

Problem Statement (Informal). *Given datasets $\{E_0, \ldots, E_N\}$ where each episode E_i is generated from the causal model in an unknown context C_r and by conditioning on an unknown value s_k of S, we want to discover the set of causal models over X.*

Before we address this problem, we take a step back to address causal discovery in an i.i.d. setting and introduce the concepts and assumptions that we build on.

2.2 Causal Discovery

For now, consider the case of a single context without selection bias. We can specify a causal model over the variables X by a directed acyclic graph (DAG) $G = (X, E)$ with node set X and edges $(i, j) \in E$ whenever the variable X_i is a cause of X_j [30]. To denote the set of direct causes of X_j we write pa_j where we leave G implicit. Together with the network structure in G, we assume a structural causal model over the variables, where each effect is generated from its causes through a causal function or mechanism f_j,

$$X_j = f_j(pa_j, N_j)$$

where N_j is a noise variable implicit in G with $N_j \perp\!\!\!\perp X_j$.

A causal model is *identifiable* when we can determine it uniquely from an observational distribution [30]. In general, identifiability of the causal DAG 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 *causal Markov* and *faithfulness* conditions, which together imply

that edge separations in the graphical model 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 G [10].

Identification of causal directions beyond the MEC is possible using additional information about how the system reacts to interventions [11, 22, 44]. In the absence of such information, we need to make additional assumptions, such as restricting the functional dependencies f to nonlinear functions with additive noise [5, 12, 25]. As an example of this approach, a family of methods build on the algorithmic framework of causation [14] and derive consistent scoring criteria that can be used for causal discovery within a given class of functional models. This is the approach we will follow here.

2.3 Information-theoretic Causal Discovery

The algorithmic model of causation [14] 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. The Kolmogorov complexity $K(P)$ over a distribution P defines the length of the shortest program p^* that approximates P up to precision q on a universal Turing machine $\mathcal{U}$ given input $\langle x, q\rangle$ [18],

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

Using Kolmogorov complexity, we can state the centerpiece of the algorithmic view of causal networks, namely the Algorithmic Markov Condition (AMC) [14].

Algorithmic Markov Condition. The AMC postulates that causal mechanisms correspond to *programs* that encode the observed distributions most concisely in terms of Kolmogorov complexity. More precisely, it assumes that each causal mechanism f_j for a given X_j can be described by a program p_j that independently generates the distribution $P(X_j \mid pa_j)$. The AMC posits that the complexity of the overall distribution $P(X)$ corresponds to the summed complexities over these independent programs,

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

which holds up to a constant, i.e. the complexities can differ by that of a program with constant length.

Causal Discovery using the AMC. Kolmogorov complexity cannot be computed for arbitrary programs [18], but can be approximated from above via Minimum Description Length (MDL) [9] for a fixed model class. Eq. (1) is therefore commonly stated for a flexible class of functions, such as non-parametric regression models.

In detail, MDL defines a description length L of X together with its optimal causal model G^*, given by

$$L(X; G^*) = L(G^*) + \sum_{X_j \in G^*} (X_j \mid pa_j, G^*) \ . \tag{2}$$

The score $L(X; G)$ is given by the length, in bits, of first encoding the model itself and then encoding the data under the model. Specifically, $L(G)$ encodes the network structure of G and the functional relationships f_j using a model class of choice with MDL score $L(f_j)$. The remaining term poses according to the causal factorization and describes each variable X_j from its causal parents pa_j. Using L, Eq. (1) suggests estimating the causal model as the one minimizing the overall description length $L(X; G)$.

Various instantiations of L exist, addressing, for example, the bivariate [24] and multivariate case [27], latent confounding [16], and interventional data [21, 22, 26]. Throughout this work, we assume a given score L that decomposes as in Eq. 2 and is consistent in the sense that it allows estimating a DAG $G \sim G^*$ that is Markov equivalent to G^* in the limit, $\lim_{n\to\infty} P(\hat{G} \sim G^*) = 1$ for i.i.d. data with sample size n. We refer to Mian et al. [27] for definitions of L in a multivariate setting and a consistent algorithm for discovering G in from an i.i.d. data distribution. As consistency results and practical algorithms have only been explored in the i.i.d. case [27] or interventional data [22, 26], we turn to episodic data here.

3 THEORY

In this section, we introduce our causal model for episodic data.

3.1 Causal Model

Our causal model comprises a set of causal DAGs $\mathbf{G} = \{G_1, \ldots, G_R\}$ over a common set of variables $X \cup \{S\}$, where X are measured, continuous random variables of interest, and S is an unmeasured categorical variable with values $S = \{s_1, \ldots, s_K\}$. Each DAG G_r is a causal model over X^r, i.e., it describes the causal relationships in all episodes from a given context C_r. The additional variable S models that certain observations may be missing in each episode.

To do so, we extend upon a missingness framework commonly used to handle selection bias [2, 35]. To explain, consider the nth observation, where we represent S using a one-hot encoding,

$$\big(X_1^{(n)}, \ldots, X_M^{(n)}, s_1^{(n)}, \ldots, s_K^{(n)}\big)$$

where we suppress the dependency on the context to avoid clutter. Above, $X^{(n)}$ is associated to indicators s_k where $s_k = 1$ if $X^{(n)}$ is *observed*, else $s_k = 0$ if it is *missing* in a distribution k. We obtain K biased distributions $P(X \mid S = s_k)$ of which episodes are subsamples. Exactly which samples are observed could depend on X; in Fig. 1, for instance, $S_+ = \circ$ holds for the temperature range $X_1 \geq 10$. In general, we assume that any unknown mechanism assigns S,

$$S = g(X, N_s), \qquad N_s \perp\!\!\!\perp S \ ,$$

where g maps each sample to an assignment of S using input X, which is noisy through N_s. We therefore include S in the causal model together with edges $X_j \to S$ for all X_j, and assume that S is a sink node. We include a node S^r in G_r in each C_r with the same K for simplicity, although our framework can be extended to include a dependency K^r. We assume causal sufficiency over $X^r \cup \{S\}$.

To summarize, our causal model is the following.

Assumption 3.1 (Causal model with contexts and selection)**.** Our causal model is given by a set of DAGs $\mathbf{G} = \{G_1, \ldots, G_R\}$ over $X \cup \{S\}$ from a finite number of contexts R such that in context C_r, each observed variable X_j is generated as

$$X_j^r = f_j^r(pa_j^r, N_j^r), \qquad N_j^r \perp\!\!\!\perp X_j^r \ ,$$

where pa_j^r denote the causal parents of X_j^r in G_r and N_j^r is an independent noise term. The latent variable S is generated as

$$S^r = g^r(X_j^r, N_s^r), \qquad N_s^r \perp\!\!\!\perp S^r \ .$$

The above describes an unbiased generating process where each variable X_j is a function of its causal parents pa_j and noise N_j.

In addition, the mechanism g with noise N_s generates S. This generating process happens independently in each context.

We assume that episodes result from conditioning on a specific value of the unobserved selection variable.

Assumption 3.2 (Episodic data)**.** Under the causal model in Assumption 3.1, after generating an unbiased distribution $P^r(X, S)$ from the DAG G_r in each context C_r, all episodes E coming from context $C(E) = C_r$ have distribution $P^r(X \mid S = s_k)$ for some specific $s_k \in \{s_1, \ldots, s_K\}$.

With no assumption on the selection mechanism g, number of contexts R, or number of selection regions K, our model can encompass general cases of episodic data. This invariably also makes it more challenging to discover the causal model from data. To do so, we first state the algorithmic Markov condition for our model.

Postulate 3.3 (Algorithmic Markov Condition)**.** *Under Assumptions 3.1 and 3.2, a set of causal DAGs $\mathbf{G} = \{G_1, \ldots, G_R\}$ is only admissible as the causal hypothesis over X and S if*

$$K\big(P(X \cup \{S\})\big) \stackrel{+}{=} \sum_{r=1}^{R} \sum_{j=1}^{M} K\big(P^r(X_j \mid pa_j)\big) + K(P^r(S \mid X))$$
$$\stackrel{+}{=} K\big(P(X)\big) + K\big(P(S \mid X)\big)$$

where $\stackrel{+}{=}$ holds up to an additive constant.

As S is not included in any parent set, we can in principle consider the complexity of, and hence causal structure over, X *independently* of the complexity of S. This motivates the idea of using a consistent scoring criterion to find the causal structure over X in each context.

As a complication, we hereby need to discover the number of contexts. Suppose we obtained data D_n accumulated over n episodes. There could be any number R of different causal models, with $1 \leq R \leq n$. Thus, we need to consider any partition of our samples into R disjoint sets, which we write as $\Pi(D_n) = \{X^1, \ldots X^R\}$. In each set, we propose discovering the causal DAG using the consistent score $L(X^r; G)$, and overall find the partition minimizing this score.

To summarize, our objective is as follows.

Problem Statement. *Given variables X and data D_n over n episodes, we aim to discover the partition $\Pi(D_n)$ of D_n into contexts and the causal model $\hat{G}r$ in each context minimizing*

$$\min_{\Pi(D)} \sum_{r=1}^{|\Pi(D)|} \min_{G_r} L(X^r; G_r) \ . \tag{3}$$

where we write X^r for the data in the r-th set of $\Pi(D)$.

This leaves us with two questions; first, ensuring that the above is a consistent way of identifying the causal model, and second, how to efficiently minimize it in practice.

3.2 Asymptotic Guarantees

We first want to establish conditions under which L can be used in a consistent way to discover the causal DAGs in all contexts.

This revolves around whether the *biased* distributions in each episode eventually allow us to estimate the relevant distributions in Postulate 3.3 in an *unbiased* way so that we can apply Eq. (3). That is, estimation of each causal mechanism should not depend on the selection variable. We hence make the following assumption.

Assumption 3.4 (Ignorability)**.** Under the causal model in Assumption 3.1 and given D^N over N episodes, in each context C_r, we assume the following *ignorability* of selection bias,

$$X_j^r \perp\!\!\!\perp S^r \mid \mathbf{Z}^r$$

for each X_j^r and conditioning set $\mathbf{Z}^r \subseteq X^r \setminus \{X_j^r, S^r\}$.

Examples of when the above holds are cases known as Missing At Random (MAR) or Missing Completely At Random (MCAR) [2, 3, 35], for example, when a biased $P(X \mid S = s_k)$ is a uniform sample from the population $P(X)$. A more realistic case is the one in Fig. 1 where the selection mechanism depends on temperature X_1. We can see that episodes from the cold season $P(X \mid S = *)$ indeed do not allow an unbiased view of the causal mechanism, however once we obtain enough episodes from both S_-, S_+ then ignorability holds. More generally, we ensure via Assumption 3.4 that we eventually obtain enough samples from the support of X.

With this, we can show that an MDL-based score L can be used for causal discovery with unknown contexts.

Theorem 3.5 (Consistency of L in the episodic setting)**.** *For the causal model in Assumption 3.1 and given data D_n over n episodes as in Assumption 3.2, under Assumption 3.4, a consistent scoring criterion L that decomposes as in Eq. 2 remains consistent,*

$$\lim_{|D_n| \to \infty} P(\hat{G}_r \sim G_r^*) = 1 \qquad \text{for all } r \in \{1, \ldots, R\} \ .$$

However, this does not make it obvious how to apply L in practice. First, note that the result relies on enough episodes being observed so that selection is ignorable, that is, we did not yet address how to deal with non-ignorable selection at each time point when we only observed a subset of episodes. Second, even when observing enough episodes, searching over the space of DAGs to minimize L as in Eq. 3 is prohibitive even for a single causal model due to the super-exponential search space over DAGs [6]. While there exist greedy algorithms to do so, such as the MDL-based Globe, applying such methods to any partition of the data with an unknown number of contexts is not favorable as it could violate the i.i.d. assumption required for these methods. To address these issues, we propose an algorithm for causal discovery over episodic data in the following.

4 ALGORITHM

In this section, we introduce our algorithm Continent.

4.1 Overview

To motivate our algorithm setup, let us revisit our motivating example in Fig. 1 showing episodes obtained in winter E_1, spring E_2, summer E_3, and autumn E_4. We consider a fixed number of seasons, here S_+, S_-. All episodes E_1-E_4 shown come from a context C_1 but any number of future episodes could arrive from a different C_2.

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

input : episodes E arriving over time, residual test $\mathcal{T}$, causal discovery algorithm $\mathcal{A}$ with score L
output: causal model $\mathbf{G} = \{G_1, ...G_R\}$

1 $\mathbf{G} \leftarrow \{\}$
2 $\tau \leftarrow 0$
3 **while** *a new episode* E_i *arrives* **do**
4 $\quad \mathbf{G} \leftarrow$ UPDATE $(\mathbf{G}, E_i, \mathcal{A}, \mathcal{T})$
5 $\quad \tau \leftarrow \tau + 1$
6 $\quad$ **if** $\tau \geq \tau_{max}$ **then**
7 $\quad\quad \mathbf{G} \leftarrow$ MERGE $(\mathbf{G}, \mathcal{A}, \mathcal{T})$
8 $\quad\quad \tau \leftarrow 0$
9 $\quad$ **end**
10 **end**
11 $\mathbf{G} \leftarrow$ MERGE $(\mathbf{G}, \mathcal{A})$
12 **return** $\mathbf{G}$

Given a learner $\mathcal{A}$ for greedy DAG search with a consistent scoring criterion L, we aim to discover the underlying causal DAG G_1 over E_1-E_4, and possibly add a causal model G_2 if future episodes from a different C_2 arrive. Applying $\mathcal{A}$ to all episodes at each time point is not only impractical, but may also not be consistent given that selection bias is not ignorable until all episodes arrive. Instead, we propose an algorithm CONTINENT that maintains plausible causal models $\mathbf{G} = \{G_1, \ldots, G_R\}$ at each time t_i and uses a strategy for *updating* $\mathbf{G}$ when a new episode E_{i+1} arrives.

Model Updating. In our example, say that we obtained episodes E_1-E_3 and the current causal model is $\mathbf{G} = \{G_1\}$. As we already observed episodes from both seasons S_+ resp. S_- we likely already learned an unbiased model G_1. As the autumn episode E_4 arrives, we want to assign it to G_1 without re-learning the causal model from scratch. To this end, we propose using a two-sample testing procedure $\mathcal{T}$ to decide whether a given episode matches an existing causal model. Here, after checking with $\mathcal{T}$ that E_1-E_4 can be stacked we combine the data E_1-E_4 and keep the model G_1 as is.

On the other hand, say episode E_5 from a different context C_2 arrives[1] and $\mathcal{T}$ decides that it does not match any current causal model. Then we apply the learner $\mathcal{A}$ to learn a new model G_2 over E_5 and add it to our set of models, $\mathbf{G} = \{G_1, G_2\}$.

Note that the above assumes that we already learned an unbiased causal model over the available episodes. We also need to consider the case where a causal model is biased, such that we need to update it after *merging* data from multiple episodes.

Model Merging. Say that we observed episodes E_1-E_2 to learn a causal model G_0. From the winter seasons S_- alone, it appears that X_1, X_2 are uncorrelated, hence G_0 is biased. When E_3 from summer season S_+ arrives, we need to *merge* the data to the previous episodes and learn a new model G_1.

To do this, we attempt merging data over multiple episodes at regular time intervals. We again apply $\mathcal{T}$ to check whether a merge is possible, and if so, check whether merging any two causal models

[1]This could be e.g. readings obtained from a different geographical region where causal mechanism between X_1 and X_2 is different/non-existent.

Algorithm 2: TESTRESIDUALEQ $(G_r, E_i, D, \mathcal{T})$

input : causal model G, episode E_i, data D, residual test $\mathcal{T}$
output: test result

1 **foreach** X_j *with parent set* $\mathbf{Z}$ *in* G **do**
2 $\quad p_j \leftarrow \mathcal{T}$.TEST$(H_0 : P^D(X_j \mid \mathbf{Z}) \equiv P^i(X_j \mid \mathbf{Z}); \alpha)$
3 **end**
4 $\mathbf{p} \leftarrow \mathcal{T}$.CORRECT$(\{p_1, ...p_M\})$
5 **if** $\mathcal{T}$.SIGNIFICANT$(\mathbf{p})$ **return** TRUE **else return** FALSE

results in an improved model, judging by our score L. As stacking may be sufficient when we already gained sufficient evidence for a candidate model, in practice, we attempt merging at regular time intervals using a pre-specified tolerance parameter τ_{max}.

Combining the model updating and model merging described above, we have our proposed approach, CONTINENT .

CONTINENT. We show the pseudocode of CONTINENT in Alg. 1. We maintain a set of models $\mathbf{G}$ throughout, where we associate each $G \in \mathbf{G}$ to a dataset D of episodes, initially empty (Line 1).

As new episodes arrive, we update $\mathbf{G}$ at each time step using the UPDATE function (Line 4). In short, it checks using hypothesis testing whether a new episode E_i matches the data D under an existing model, in which case we stack the datasets E_i and D; else we apply $\mathcal{A}$ to E_i to discover a new model G_i which we add to $\mathbf{G}$. We show our hypothesis test in Alg. 2, and UPDATE in the appendix.

After a pre-specified number of episodes, we attempt merging existing models (Line 7), with a tolerance parameter τ keeping track of the time since a merge last happened (Line 8). In essence, MERGE performs pairwise comparison of models G, G'. If appropriate, it learns a new model $G_\cup$ after pooling the resp. datasets D, D' of the pair. During the algorithm, we only allow such a merge if $\mathcal{T}$ marks the residual distributions of D, D' as compatible, for which we again apply our hypothesis test in Alg. 2. We include the pseudocode for MERGE in Appendix B.

Our alternation of updating and merging continues as long as new episodes arrive. We conclude with a final merge (Line 11). Compared to merge steps throughout our algorithm which we protect by $\mathcal{T}$, we consider all remaining possible merges of model pairs G, G' in this step given that no more episodes arrive (Line 11).

4.2 Consistency

Naturally, we want to make sure that our adaptive strategy is consistent. At any time point t_i, however, we only have access to a subset of the episodes so that ignorability in Assumption 3.4 unlikely holds, and hence any causal model inferred using $\mathcal{A}$ may be incorrect. Nevertheless, we need to avoid merging episodes with different underlying models. We now show that we can do so without knowing the true models. To do so, we assume a hypothesis test $\mathcal{T}$ testing

$$H_0 : P^1(X_j \mid \mathbf{Z}) \equiv P^2(X_j \mid \mathbf{Z})$$

for a given variable X_j, conditioning set $\mathbf{Z}$ and two datasets P^1, P^2. Given any causal DAG, we test H_0 for each variable given its estimated parent set and include a multiple testing correction, as

shown in Alg. 2. We can show that our updating strategy protected by this test is consistent under the following condition.

Assumption 4.1 (Detectable selection)**.** We assume that selection *detectable* for a variable X_j and pair of contexts C_r, C'_r meaning

$$P^r(X_j \mid pa_j) \neq P^{r'}(X_j \mid pa_j)$$
$$\Rightarrow P^r(X_j \mid pa_j, S = s_k) \neq P^{r'}(X_j \mid pa_j, S = s_k)$$

holds for each value s_k of S.

Unlike ignorability in Assumption 3.4 which requires full independence of the causal mechanism and selection mechanism, that is, ensures that we can estimate the causal mechanism for each variable in a fully unbiased way, Assumption 4.1 only requires that distribution differences of $P(X)$ hold also in the biased distribution $P(X \mid S = s_k)$. Given that the latter are subsamples of the overall distribution, this is reasonable in practice. With this, we can show that our updating strategy is consistent.

Theorem 4.2 (Consistency of updating using $\mathcal{T}$)**.** *With discrepancy test $\mathcal{T}$ we will never merge a new episode E_{i+1} with a set $\hat{X}^r$ from an incorrect context where $C(E_{i+1}) \neq C(E)$ for some $E \in \hat{X}^r$.*

This shows that our updating step is safe in the sense that we always discover subsets of the correct contexts. When we observed all episodes, we can also recover the exact sets of contexts if ignorability holds, based on Thm. 3.5.

Corollary 4.3 (Consistency of Continent)**.** *Given a consistent DAG search algorithm $\mathcal{A}$ and score L, under assumption 3.4 our algorithm is consistent, so that*

$$\lim_{|D_n| \to \infty} P(\hat{G}_r \sim G_{r*}) = 1 \qquad \text{for all } r \in \{1 \dots, R\}$$

holds after we obtain n episodes D_n and perform the merge step.

As the final step in this section, we address practical considerations around our algorithm.

4.3 Instantiation

We conclude this section by giving details on the components of Continent.

Causal Discovery Algorithm $\mathcal{A}$. We assume a score-based causal discovery algorithm $\mathcal{A}$ that allows discovering a causal DAG G from an i.i.d. dataset D. While in principle, this could be any score-based method with a consistent scoring criterion L decomposing according to Eq. (2), we use an MDL-based approach in our practical instantiation as it allows for a principled way for model comparison. We instantiate $\mathcal{A}$ with Globe [27] which is an efficient algorithm for discovering causal networks. It models causal functions through non-parametric multivariate regression with additive noise.

Residual Test $\mathcal{T}$. Our method can also work together with any hypothesis test $\mathcal{T}$ for differences in conditional distributions under a causal model. As Globe models causal functions through non-parametric spline regression, a natural choice is testing residual distributions under a given model for equality. As we apply a test per each variable, we perform Bonferroni correction to obtain a p-value from the test results $\{p_1, \dots, p_M\}$. Unless otherwise stated, we apply the non-parametric Kolmogorov-Smirnov [1, 38] test in our evaluations.

5 RELATED WORK

Discovering causal models that faithfully describe the interactions between variables of interest given observational data alone is an actively studied problem and finds applications in almost all areas of science. Approaches to do so typically fall into the categorizations of constraint-based methods, such as PC [30], or score-based methods, such as GES [6, 34]. As these approaches discover a Markov Equivalence Class (MEC) of the causal DAG [10], 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 [5, 33], such as LiNGAM [37] which assumes linear non-Gaussian models. Another branch of work builds on the algorithmic model of causality [14], such as Globe [27]. However, the examples given up to this point assume an i.i.d. data distribution where a single causal network can capture the causal interactions, and where neither selection bias nor contexts exist.

Selection Bias. Missingness is a well-studied problem in statistical inference and in particular, many approaches exist for *correcting* for missingness and selection bias [4, 8, 43]; see Little and Rubin [19] for an overview. Only very recent work studies assumptions for *identifying* whether selection bias holds in a given dataset [17]. Our perspective is different as we are interested in *adapting* causal discovery to the presence of missingness. An important line of work studies *recoverability* [3, 31] from selection bias in causal discovery, modeled through unobserved sink node S in the causal graph. We also adopt this model here using multiple missingness regions, and in addition consider the presence of multiple contexts in the form of varying causal mechanisms.

Different Contexts. A wealth of recent literature studies causal discovery from different environments, experimental regimes, or contexts [13, 20, 40, 44]; prominent examples include the constraint-based JCI framework [28], additive noise model based multi-group Lingam [36], and score-based approaches [7, 21, 22, 26] for discovering causal DAGs from multi-context data. While studies of latent confounding in such data exist [23], latent selection remains underexplored. In particular, existing work assumes that each context is an identically distributed (i.i.d.) sample with fixed causal model. We make this setting more general in that we obtain biased samples from each context, which need to be combined to result in i.i.d. data. To our knowledge, we are the first to allow a different causal model with episodical bias in different contexts, and also address the algorithmic challenges associated with discovering causal networks in an online fashion.

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

6 EVALUATION

Since to the best of our knowledge, there is no specific algorithm designed for causal discovery from continually arriving episodic data, we look at the nearest possible modifications of existing algorithms for comparison. As baseline we compare to Globe [27], Resit [33] and Ges [6, 34]. We modify these algorithms as follows — we first learn a causal network over each individual incoming

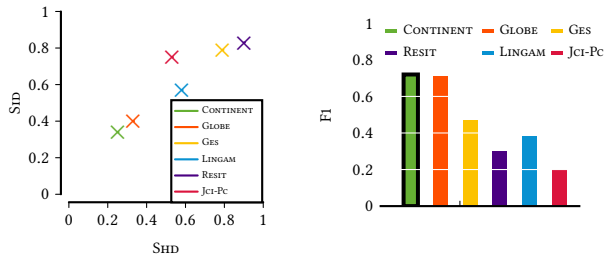

Figure 2: Normalized SHD and SID [Left, Closer to origin is better] and Orientation F1 [Right, Higher is better] for networks learned over episodic data with selection bias.

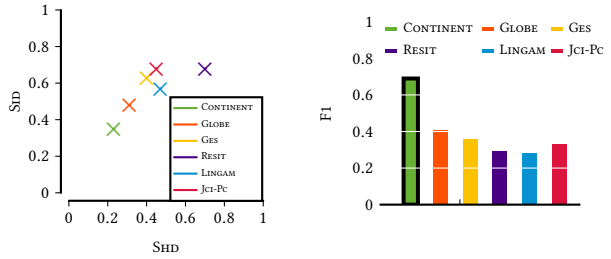

Figure 3: Normalized SHD and SID [Left, Closer to origin is better] and Orientation F1 [Right, Higher is better] for networks learned over episodic data with unknown interventions.

episode of data, and then take a union over the edges. This is correct, under the assumption underlying each of these approaches, that each episode comes from the *same* causal network [26]. We also compare to multi-environment causal discovery approaches such as the JCI-framework [28] using the PC algorithm [39], as well as Multi-Group Lingam (LINGAM) [36]. The latter two approaches, however, require that all episodes are available to learn a causal network. Hence, we provide all episodes in one go to these approaches. This constitutes an advantage as they can learn from complete data.

To measure the quality of the predicted causal structures we use the Structural Hamming Distance (SHD) [15] , the Structural Intervention Distance (SID) [32], as well as the Orientation-F1 score over learned networks. SHD counts the number of edges where the predicted causal network differs from the true causal network, SID counts pairs of variables for which intervention estimation differs across predicted resp. true causal network and the F1 score allows us to see how accurately edges are oriented in the learned network. Next, we discuss results over both synthetic and real-world data.

6.1 Synthetic Data

For each of the proposed experiment setups, we generate random graphs using Erdős-Rényi model for network sizes $d = \{5, 10, 15\}$, and generate data for effects using functions of the following form, $X_i = \sum_{x \in pa_i} f(x) + \mathcal{N}_i$, 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 $\mathcal{N}_i$ is either Gaussian or Uniform. For each graph/function combination, we generate a total of 10,000

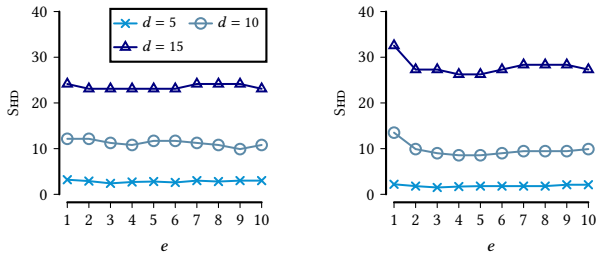

Figure 4: [Lower is better] Change in SHD over increasing number of episodes e for data with selection bias (left) and unknown interventions (right) for graph sizes $d = \{5, 10, 15\}$.

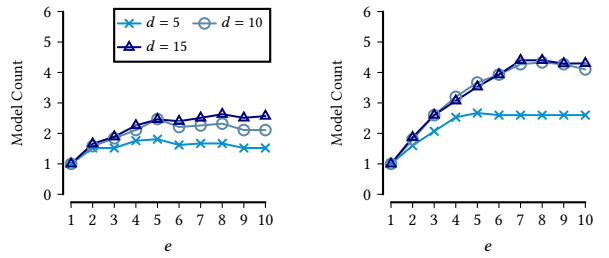

Figure 5: Model Count over increasing episodes e for data with selection bias (left) and data with (unknown) interventions (right) for graphs of size $d = \{5, 10, 15\}$. There are 1 resp. 3 true underlying models for bias resp. intervention cases.

samples and then split them into 10 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 LINGAM are provided all episodes together, we only measure performance over the final network.

Primarily, we investigate the following questions.

- **Q1** Can CONTINENT reliably discover causal networks when the incoming episodes come from the same underlying causal network?
- **Q2** How well does CONTINENT perform when episodes contain unknown interventions?
- **Q3** Can CONTINENT identify causal networks from episodic data containing *different* causal mechanisms?
- **Q4** How does CONTINENT's performance change over time as episodes arrive?

As CONTINENT is designed without the assumption that each data comes from the same underlying causal network, it therefore maintains a list of candidate networks for groups of episodes. For comparability to other approaches, we *force* CONTINENT to predict a single causal network for cases $Q1$ and $Q2$ by taking a union over the edges in candidate models [26]. We further provide an analysis of the individually learned causal networks for evaluation in $Q3$. We release all our code and data for research purposes[2]. Next, we show results for each of the four questions.

[2] https://eda.rg.cispa.io/prj/continent/

Experiment	Nodes	SHD	SID	F1
Interventions	5	0.23	0.15	0.68
	10	0.25	0.34	0.54
	15	0.29	0.50	0.43
Mechanism Changes	5	0.21	0.15	0.74
	10	0.26	0.38	0.64
	15	0.36	0.65	0.41

Table 1: Normalized SHD [Lower is better], normalized SID [Lower is better] and Orientation F1 [Higher is better] for networks predicted by CONTINENT for held-out episodes for interventional data as well as mechanism changes.

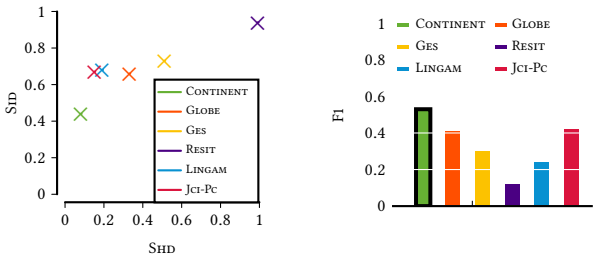

Figure 6: Normalized SHD and SID [Left, Closer to origin is better] and Orientation F1 [Right, Higher is better] for networks learned REGED Lung cancer gene expression dataset.

Q1. Identical Networks. We first test all methods on the cases where each incoming episode comes from the same underlying causal network, both for i.i.d. as well as selection bias. Interesting for us is the latter where episodes can contain selection bias. We generate this case by choosing a variable at random from our dataset and sorting the entire data over that variable before splitting the data into episodes and transmitting it. We show the results for this in Fig 2 where we see that CONTINENT shows superior performance to the competition. CONTINENT not only discovers causal network structurally closer to the ground truth, but also clearly performs well when orienting the edges as can be seen by the F1 score in Fig. 2. This shows that CONTINENT performs well under selection bias. We provide the results for i.i.d. data in the appendix.

Q2. Interventions. After our sanity check using i.i.d. data and dominant performance over data with selection bias, we level up the difficulty by introducing episodes that contain interventions. To do so, we generate 3 datasets. The first dataset is observational, whereas for the other two, we select a subset of at most $\log_2(d)$ variables and perform a *do*-intervention [30] on that subset, before generating the data. This gives us data sampled from three different distributions. We further split each of these datasets into episodes before transmitting them. We never provide information about these interventions to any of the methods beforehand.

We show the results of this experiment in Fig. 3, where we see that while GLOBE degrades slightly, CONTINENT's performance does not degrade compared to the setup in $Q1$. CONTINENT , in fact, continues to clearly outperform the competition.

Q3. Changing Mechanisms. As the next challenging step, we introduce episodes containing different causal networks/mechanisms over the same variables. This rules out using any of our competitors as they can not handle such data. To evaluate CONTINENT in this setting, we additionally generate a hold-out set of episodes that we do not learn over. Once CONTINENT has learned over the training episodes, we try to *predict* the causal network for hold-out episodes, without learning it explicitly, using the existing learned models. We do so by simply taking the model that compresses this hold-out episode best (Eq. (1)) and comparing the predicted network to the ground truth. We show the results in Table. 1 where we observe that CONTINENT continues to perform well overall for $d = 5, 10$, and at least structurally for $d = 15$. We see that for this challenging setting with changing mechanisms, CONTINENT can find a reasonable skeleton (lower SHD) but conflicting mechanisms cause it to get edge directions wrong more often (higher SID). Nevertheless, we see that CONTINENT's performance does not degrade massively compared to previous settings, even in this challenging case.

Q4. Performance over time. We measure how the individual models present inside CONTINENT evolve over time. To that end, we show how the SHD (Fig. 4) as well as the model count (Fig. 5) progresses as we receive new episodes. For the case of SHD, we find that CONTINENT always ends up with a lower SHD at the final episode, than the one it starts with, this effect is more profound for networks of size $d = 15$ than $d = 5$ as it might be harder to identify the correct network over a larger number of variables in the beginning. We see that CONTINENT is able to improve as the number of episodes increase. For data with selection bias, we see that CONTINENT keeps on average 2 models throughout the learning as shown in Fig. 5. More interestingly CONTINENT ends up converging to almost 3 models for interventional data as shown in Fig. 5, which is exactly the actual number of different networks present across episodes.

6.2 Lung cancer gene expression data

After measuring the efficacy of our approach using synthetic data, we turn to (pseudo) real-world REGED dataset [41] containing 20,000 samples over 500 variables for lung cancer gene-expressions. We split the samples into ten non-overlapping episodes and consider two non-overlapping networks of sizes $d = 5, 15$ within the ground truth network and run a total of 10 experiments as follows. First, we randomly choose a subset of 5 episodes, merge them and introduce selection bias over the stacked data akin to $Q2$, before splitting it back. We show the results for REGED5 in the appendix and for REGED15 in Fig. 6 where once again CONTINENT comes out on top.

7 DISCUSSION AND CONCLUSION

Our interest in this work is determining causality when data arrives progressively over time in multiple episodes, each representing subsamples of the population or subregions of the data that need to be pooled *together* to avoid bias. At the same time, we address that the causal relationships may not be stationary over time, and treat episodes from different contexts under a *separate* causal model. To address this setting, we propose a causal model over a set of latent contexts leading to a set of different causal networks, as well

as model episodic bias through a hidden selection variable. We show that information-theoretic scoring criteria remain consistent for this model in the limit if we obtain sufficiently many episodes so that selection bias becomes ignorable. To address the more realistic setting where episodes arrive one by one over time with non-ignorable selection bias, we propose the CONTINENT algorithm to learn the causal model adaptively over time. It maintains a set of causal networks over all episodes and incorporates new episodes into the model, using a residual testing strategy to avoid combining episodes from different contexts.

Our experimental results show that our method performs reliably in the presence of selection bias, under unknown interventions, and even when different causal models underlie the data-generating process, which to our knowledge no existing methods can address. Future directions of our work include addressing non-ignorability further by using correction or extrapolation techniques, and addressing practical considerations such as the instantiation choices.

REFERENCES

[1] KOLMOGOROV AN. 1933. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att* 4 (1933), 89–91.

[2] Elias Bareinboim and Judea Pearl. 2012. Controlling Selection Bias in Causal Inference. In *AISTATS*, Vol. 22. PMLR, 100–108.

[3] Elias Bareinboim, Jin Tian, and Judea Pearl. 2014. Recovering from Selection Bias in Causal and Statistical Inference. *AAAI* 28, 1 (2014).

[4] P Boeken, Noud de Kroon, Mathijs de Jong, Joris M. Mooij, and Onno Zoeter. 2023. Correcting for selection bias and missing response in regression using privileged information. In *UAI*. PMLR, 195–205.

[5] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals Stat.* 42, 6 (2014), 2526–2556.

[6] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3 (2002), 507–554.

[7] Daniel Eaton and Kevin Murphy. 2007. Exact Bayesian structure learning from uncertain interventions. In *AISTATS*. PMLR, 107–114.

[8] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. 2008. Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press.

[9] Peter Grünwald. 2007. *The Minimum Description Length Principle*. MIT Press.

[10] Alain Hauser and Peter Bühlmann. 2013. Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Statist. Soc. B* 77 (03 2013).

[11] Alain Hauser and Peter Bühlmann. 2014. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning* 55, 4 (jun 2014), 926–939.

[12] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *NeurIPS*, Vol. 21. Curran.

[13] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. 2020. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems* 33 (2020).

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

[15] Markus Kalisch and Peter Bühlmann. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *JMLR* 8, Mar (2007), 613–636.

[16] David Kaltenpoth and Jilles Vreeken. 2019. We Are Not Your Real Parents: Telling Causal from Confounded using MDL. In *SDM*. SIAM, 199–207.

[17] David Kaltenpoth and Jilles Vreeken. 2023. Identifying Selection Bias from Observational Data. *AAAI* (2023), 8177–8185.

[18] M. Li and P. Vitányi. 2009. *An Introduction to Kolmogorov Complexity and its Applications*. Springer.

[19] Roderick Little and Donald Rubin. 2019. Statistical Analysis with Missing Data, Third Edition. (04 2019). https://doi.org/10.1002/9781119482260

[20] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. 2018. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. In *NIPS*, Vol. 31.

[21] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. 2022. Discovering Invariant and Changing Mechanisms from Data. In *KDD*. ACM, 1242–1252.

[22] Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. 2023. Learning Causal Mechanisms under Independent Changes. In *NeurIPS*.

[23] Sarah Mameche, Jilles Vreeken, and David Kaltenpoth. 2024. Identifying Confounding from Causal Mechanism Shifts. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4897–4905.

[24] Alexander Marx and Jilles Vreeken. 2019. Identifiability of Cause and Effect using Regularized Regression. In *KDD*. ACM.

[25] Alexander Marx and Jilles Vreeken. 2021. Formally Justifying MDL-based Inference of Cause and Effect. *arXiv preprint arXiv:2105.01902* (2021).

[26] Osman Mian, Michael Kamp, and Jilles Vreeken. 2023. Information-theoretic causal discovery and intervention detection over multiple environments. In *AAAI-23*.

[27] Osman Mian, Alexander Marx, and Jilles Vreeken. 2021. Discovering fully oriented causal networks. In *AAAI*.

[28] Joris M Mooij, Sara Magliacane, and Tom Claassen. 2016. Joint causal inference from multiple contexts. *JMLR* 21 (2016).

[29] Joris M. Mooij, J. Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Scholkopf. 2014. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *ArXiv* abs/1412.3773 (2014).

[30] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

[31] Judea Pearl. 2012. A solution to a class of selection bias problems. (2012).

[32] Jonas Peters and Peter Bühlmann. 2015. Structural intervention distance for evaluating causal graphs. *Neural computation* 27, 3 (2015), 771–799.

[33] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. Causal Discovery with Continuous Additive Noise Models. *JMLR* 15 (2014).

[34] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 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.* (2017).

[35] Donald B. Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.

[36] Shohei Shimizu. 2012. Joint estimation of linear non-Gaussian acyclic models. *Neurocomputing* 81 (2012).

[37] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *JMLR* 7 (2006).

[38] Nickolay Smirnov. 1948. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics* 19, 2 (1948), 279–281.

[39] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT Press.

[40] Chandler Squires, Yuhao Wang, and Caroline Uhler. 2020. Permutation-based causal structure learning with unknown intervention targets. In *UAI*. PMLR, 1039–1048.

[41] Alexander Statnikov, Sisi Ma, Mikael Henaff, Nikita Lytkin, Efstratios Efstathiadis, Eric R. Peskin, and Constantin F. Aliferis. 2015. Ultra-Scalable and Efficient Methods for Hybrid Observational and Experimental Local Causal Pathway Discovery. *JMLR* 16 (2015), 3219–3267.

[42] Allison L. Steiner, Adam J. Davis, Sanford Sillman, Robert C. Owen, Anna M. Michalak, and Arlene M. Fiore. [n. d.]. Observed suppression of ozone formation at extremely high temperatures due to chemical and biophysical feedbacks. *Proceedings of the National Academy of Sciences* ([n. d.]).

[43] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J. Mach. Learn. Res.* 8 (dec 2007), 985–1005.

[44] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI*.

A THEORY

We provide the technical details of our results in the following.

For ease of exposition, we separate out the case of a single context with one causal model in Thm. 3.5 and show it first.

Lemma A.1 (Consistency of L for a single causal model)**.** *Assume a causal model in Assumption 3.1 with one context C, $R = 1$, and true causal DAG G^* in C. Given data D_n over n episodes from C as in Assumption 3.2, under Assumption 3.4, a consistent scoring criterion L that decomposes as in Eq. 2 remains consistent,*

$$\lim_{n\to\infty} P(\hat{G} \sim G^*) = 1 \ .$$

Proof.

In the underlying causal model in assumption 3.1 with $R = 1$, consider data D_n^* from the true causal DAG G^* over $X \cup \{S\}$ where S is observed. By consistency of L, we know that

$$\lim_{|D_n^*|\to\infty} P(G^* \sim \arg\min_G L(X \cup \{S\}; G)) = 1 \ .$$

Using that L is decomposable as in Eq. 2, we can write

$$\begin{aligned}\min_G L(X \cup \{S\}; G) &= \min_{G(X,S)} \Big(L(G(X,S)) + \sum_{j=1}^{M} L(X_j \mid pa_j(G)) + L(S \mid X)\Big) \\ &= \min_{G(X)} \Big(L(G(X)) + \sum_{j=1}^{M} L(X_j \mid pa_j(G))\Big) + \min_{G(S|X)} \Big(L(G(S \mid X)) + L(S \mid X)\Big) = \min_{G(X)} L(X; G(X)) + \min_{G(S|X)} L(S; G(S \mid X))) \ .\end{aligned}$$

Above, we separated the graph structure G into two subgraphs: $G(X)$ over X, and $G(S \mid X)$ which includes S as well as all edges towards it. We can do so as S is a sink node and L is decomposable. Hence, when S is observed, the subgraph $G(X)$ can be identified with our objective by construction. As S is unobserved, however, we only access data D_n over n episodes inducing a biased distribution $\tilde{X}$. In that case, assume we obtain a different minimiser $\tilde{G} = \min_{G(\tilde{X})} L(\tilde{X}; G)$ with $\tilde{G} \nsim G^*$ and $L(\tilde{X}; \tilde{G}) < L(X; G^*)$. Then for at least one X_j, $pa_j(\tilde{G}) \neq pa_j(G^*)$. Due to ignorabiliy in Assumption 3.4, $\tilde{X}_j \perp\!\!\!\perp S \mid \mathbf{Z}$ holds for the conditioning sets $\mathbf{Z}_1 = \tilde{pa}_j(\tilde{G})$ and $\mathbf{Z}_2 = \tilde{pa}_j(G^*)$, therefore S does not affect the local score for $\tilde{X}_j$ given either conditioning set. That is, $L(\tilde{X}_j \mid \tilde{pa}_j(G^*)) = L(X_j \mid pa_j(G^*)) < L(X_j \mid pa_j(\tilde{G})) = L(\tilde{X}_j \mid \tilde{pa}_j(\tilde{G}))$, which contradicts that $\tilde{G}$ is a minimizer. Therefore, we also have

$$\lim_{|D_n|\to\infty} P(G^* \sim \arg\min_G L(X; G)) = 1 \ .$$

□

With this, we move to our full causal model with multiple contexts. For ease of access, we restate our objective in Eq. (3),

$$\big(\hat{\Pi}(D), \hat{G}\big) = \min_{\Pi(D)} \sum_{r=1}^{|\Pi(D)|} \min_{G_r} L(X^r; G_r) \ .$$

Theorem A.2 (Consistency of L for multiple causal models)**.** *For the causal model in Assumption 3.1 and given data D_n over n episodes as in Assumption 3.2, under Assumption 3.4, a consistent scoring criterion L that decomposes as in Eq. 2 remains consistent,*

$$\lim_{|D_n|\to\infty} P(\hat{G}_r \sim G_r^*) = 1 \qquad \textit{for all } r \in \{1, \ldots, R\} \ .$$

Proof. First, assume the number of contexts R and the context $C(E)$ that each episode E belongs to is known. That is, for the data $D = D_N$ over all episodes, we know the true partitioning $\Pi(D) = \{X^1, \ldots, X^R\}$ into disjoint, non-empty subsets $X^r \subseteq D$ such that $\cup_r X^r = D$ and each X^r is generated from a fixed causal model G^r in the rth context. Due to independent data generation from each G^r, we can apply Lemma A.1 separately in each context and obtain

$$\lim_{|D|\to\infty} P(G_1^*, \ldots G_R^* \sim \min_{G_1,\ldots G_R} \sum_{r=1}^{R} L(X^r; G_r(X))) = \lim_{|D|\to\infty} P(G_1^*, \ldots G_R^* \sim \min_{G_1,\ldots G_R} \sum_{r=1}^{R} L(X^r, S^r; G_r(X))) = 1 \ .$$

Left to show is the case where R and $\Pi(D)$ are unknown. We compare

- the true model $\mathbf{G}^* = \{G_1^*, \ldots, G_{R^*}^*\}$ and subsets $\Pi^*(D) = \{X^{*1}, \ldots, X^{*R^*}\}$, and
- the estimated model $\hat{\mathbf{G}} = \{\hat{G}_1, \ldots, \hat{G}_{\hat{R}}\}$ and subsets $\hat{\Pi}(D) = \{\hat{X}^1, \ldots, \hat{X}^{\hat{R}}\}$ minimizing Eq. 3 with score $L(\hat{\mathbf{G}}) = \sum_{r=1}^{\hat{R}} L(\hat{X}^r; \hat{G}_r(\hat{X}^r))$.

For contradiction, assume that there is no exact correspondence between the true and estimated models, more precisely, that for at least one context r with true model X^{*r} and G_r^* there is no other r' so that $\hat{X}^{r'} = X^{*r}$ and $\hat{G}_{r'} \sim G_r^*$. We can distinguish the following cases,

(1) Case $X^{*r} = \hat{X}^{r'}$ for some $r' \neq r$: then also $\hat{G}_{r'} \sim G^*_r$ by Lemma A.1 as X^{*r} is a dataset from a single context r, which however contradicts the above assumption.
(2) Case $X^{*r} \subset \hat{X}^{r'}$ for some $r' \neq r$: Then the set X^{*r} is wrongly included under the incorrect model $\hat{G}_{r'}$. Then the decomposition of Eq. 3 will contain a suboptimal likelihood term

$$L(X^{*r} \mid \hat{G}_{r'}) = \sum_{j=1}^{M} L(X^{*r}_j \mid pa^r_j(\hat{G}_{r'})) \ .$$

Using that L is decomposable, we can replace the above term in the decomposition of L as follows (keeping all other terms the same),
(a) if $G^*_r \in \hat{\mathbf{G}}$, we can replace $L(X^{*r} \mid \hat{G}_{r'})$ with $L(X^{*r} \mid G^*_r)$.
(b) if $G^*_r \notin \hat{\mathbf{G}}$, we can replace $L(X^{*r} \mid \hat{G}_{r'})$ with the full cost $L(X^{*r}; G^*_r)$ as the likelihood component dominates over $L(G^*_r)$ in the limit of samples, as shown in Mian et al. [27].

In both cases, we can replace $\hat{\mathbf{G}}$ by $\hat{\mathbf{G}} \cup \{G^*_r\}$ and $\hat{\Pi}(D)$ by $\{\hat{X}^1, .., X^{*r}, \hat{X}^{r'}, .., \hat{X}^{\hat{R}}\}$ where we separate X^{*r} and $\hat{X}^{r'}$ and keep all other parts the same, resulting in a favorable model, contradicting that it is the minimizer of Eq. 3.
(3) Case $X \subset \hat{X}^{r'}$ for some $r' \neq r$ and for a set $X \subset X^{*r}, X \neq \emptyset$: This means that a non-empty subset of X^{*r} is included under the incorrect DAG, in which case we can apply the same argument as in case (2).

We can disregard the case $X^{*r} \cap \hat{X}^{r'} = \emptyset$ for all r' as then X^{*r} is not covered by the partition.
Thus, $\hat{R} = R^*$ and each $\hat{X}^r = X^{*r}$ and $G_r \sim \hat{G}_r$ (up to permuting the indices). □

Next, we justify our updating and merging strategy in the presence of selection bias.

Theorem A.3 (Consistency of model updating using $\mathcal{T}$)**.** *Given a set of causal DAGs $\hat{\mathbf{G}} = \{\hat{G}_1, \ldots, \hat{G}_{\hat{R}}\}$ and subsets $\hat{\Pi}(D) = \{\hat{X}^1, \ldots, \hat{X}^{\hat{R}}\}$ over episodes $\cup_r \hat{X}^r = \{E_1, \ldots, E_i\}$. With discrepancy test $\mathcal{T}$ we will never merge a new episode E_{i+1} with a set $\hat{X}^r$ from an incorrect context.*

Proof. We need to show that with a merge protected by $\mathcal{T}$, a merge of E_{i+1} with any set $\hat{X}^r$ can only occur if $C(E_{i'}) = C(E_{i+1})$ for all $i' \leq i$. For induction on the time step i, consider the following cases,

(1) For the base case is $i = 2$, assume $C(E_1) \neq C(E_2)$. We need to show that $\mathcal{T}$ never merges E_1, E_2 from C_1, C_2 From our causal model, we know there is at least one variable in G^*_1, G^*_2 s.t.

$$P(X^1_j \mid pa^1_j) \neq P(X^2_j \mid pa^2_j)$$

From Cor. 4.5 in MSS, this implies that also for any conditioning set $\mathbf{Z}$,

$$P(X^1_j \mid \mathbf{Z}^1) \neq P(X^2_j \mid \mathbf{Z}^2)$$

that is, we have a distribution shift even when $\mathcal{A}$ discovers an incorrect DAG $\hat{G}_1$. Left to show is that it holds also for the biased distributions

$$P(X^1_j \mid \mathbf{Z}^1, S = s_k) \neq P(X^2_j \mid \mathbf{Z}^2, S = s_{k'})$$

which holds under detectable selection. Hence, our test T will detect the difference for X_j given enough data from E_1, E_2 and reject merging.
(2) For the induction step, we can assume that $C(E_{i'}) = C(E_{i''})$ for all i', i'', and apply the above pairwise argument to E_{i+1} and each $E_{i'}$.

Corollary A.4 (Consistency of model merging)**.** *Given a consistent DAG search algorithm $\mathcal{A}$ and score L,* Continent *is consistent under Assumption 3.4, so that*

$$\lim_{|D^n| \to \infty} P(\hat{G}_r \sim G_{r*}) = 1 \qquad \textit{for all } r \in \{1, \ldots, R\} \ .$$

Proof. Consider the estimated model $\hat{\mathbf{G}} = \{\hat{G}_1, \ldots, \hat{G}_{\hat{R}}\}$ and subsets $\hat{\Pi}(D) = \{\hat{X}^1, \ldots, \hat{X}^{\hat{R}}\}$ that we obtain with Continent at time step n. By the previous theorem, we know that episodes from different contexts were not merged incorrectly, $\hat{X}^r \subseteq X^{*r'}$ for some r' for each r where $\hat{R} \leq R$, which we write shorthand as $\hat{\Pi}(D) \subseteq \Pi^*(D)$. In case $\hat{R} < R$, we need to consider any remaining merges among sets in $\hat{X}^r$. If the assumptions of Thm. 3.5 hold, then we can use

$$\min_{\Pi(D), \hat{\Pi}(D) \subseteq \Pi(D)} \sum_{r=1}^{|\Pi(D)|} \min_{G_r} L(X^r; G_r) \ .$$

The above will be minimized for $\Pi^*(D)$ and $\hat{G}_r \sim G_{r*}$ for each r as it considers a subset of the partitions that Thm. 3.5 considers. Hence minimizing L is a consistent way to discover the remaining merges. □

B ALGORITHM

In this section, we include the pseudocode of the components of CONTINENT .

Algorithm 3: UPDATE $(\mathbf{G}, E, \mathcal{A}, \mathcal{T})$

input : episode E,
causal model $\mathbf{G}$,
causal discovery algorithm $\mathcal{A}$ with score L,
residual test $\mathcal{T}$
output: updated causal model $\mathbf{G}$

accepted $\leftarrow$ FALSE
foreach G_r *over data* D *in* $\mathbf{G}$ **do**
  if TESTRESIDUALEQ $(G_r, E, D, \mathcal{T})$ **then**
    accepted $\leftarrow$ TRUE
    $D \leftarrow D$.STACKDATA (E)
  end
end
if *not* accepted **then**
  $G \leftarrow \mathcal{A}$.LEARN(E)
  $\mathbf{G} = \mathbf{G} \cup \{G\}$
end
return $\mathbf{G}$

Algorithm 4: MERGE $(\mathbf{G}, \mathcal{A}, \mathcal{T})$

input : causal model $\mathbf{G}$,
causal discovery algorithm $\mathcal{A}$ with score L,
residual test $\mathcal{T}$
output: updated causal model $\mathbf{G}$

repeat
  foreach G *over data* D *in* $\mathbf{G}$ **do**
    $D^\star \leftarrow D$
    $G^\star \leftarrow G$
    $L^\star \leftarrow G$.SCORE(D)
    foreach G' *over data* D' *in* $\mathbf{G}$ *not seen yet* **do**
      if not TESTRESIDUALEQ $(G', D, D', \mathcal{T})$ **continue**;
      $D^\cup = D \cup D'$
      $G^\cup \leftarrow \mathcal{A}$.LEARN$(D^\cup)$
      $L^\cup \leftarrow G^\cup$.SCORE$(D^\cup)$
      if *TESTSCOREDIFF* $(L^\cup, L^\star)$ **then**
        $D^\star \leftarrow D^\cup$
        $L^\star \leftarrow L^\cup$
        $G^\star \leftarrow G^\cup$
      end
    end
    if $G^\star$ *is not* G **then**
      replace corresponding G, G' with $G^\cup$ in $\mathbf{G}$
    end
  end
until *convergence*;
return $\mathbf{G}$