An Information Theoretic Framework for Continual Learning of Causal Networks.

Osman Mian OSMAN.MIAN@CISPA.DE

Sarah Mameche SARAH.MAMECHE@CISPA.DE

CISPA Helmholtz Center for Information Security
Stuhlsatzenhaus 5,
66123 Saarbrücken

Editors:

Abstract

Discovering causal networks, especially from observational data alone, is a fundamental yet challenging task. Existing causal discovery algorithms not only rely on strict assumptions such as having i.i.d data, but are also limited to working with static, fully-specified datasets, rendering them incapable of learning causal networks in a continual fashion. In this short paper, we propose an information-theoretic approach that can learn causal networks in a continual fashion, does not require the i.i.d assumption on continually arriving data, and converges to the true underlying causal network as samples within the accumulated batches of data converge to the underlying data generating distribution. Our proposed approach, ConCausD, leverages the Algorithmic Markov Condition, a postulate by Janzing and Schölkopf (2010b), to discover causal networks in an online fashion. ConCausD is not only capable of continual learning, it also provides multiple plausible causal graphs at the end of each iteration, while the existing approaches can only predict a single causal network.

1. Introduction

Discovering causal dependencies from observational data is one of the most fundamental problems in science (Pearl, 2009). While there exist a plethora of approaches for discovering causal networks designed for single (Spirtes et al., 2000a; Chickering, 2002; Shimizu et al., 2006; Peters et al., 2014; Huang et al., 2018) or multiple but fully specified Shimizu (2012); Mooij et al. (2016); Zhang et al. (2017); Mian et al. (2023) datasets, causal discovery over continually arriving data still remains an open problem. It is easy to see that in practical applications such as healthcare, stock-markets, and weather-forecasting, we continuously receive new data over time. In such cases we have data arriving in batches, perennially.

The traditional paradigm in causal structure discovery assumes a single, homogeneous dataset sampled from a single, stationary distribution. In various domains where we obtain data over time in multiple batches, there is no guarantee that combining all the batches would result in a single homogeneous dataset. This makes learning causal networks over such data challenging. Existing methods that work with tabular data require that each time we receive a new batch of data, we include the new data together with already existing data and rediscover the causal network from scratch. This already results in two major problems.

First, as the number of samples grows, so does the learning time. This can be a profound problem, especially for Kernel-based approaches (Huang et al., 2018) where the runtime complexity is cubic in the number of samples. Second, given the independent and identically distributed (i.i.d) assumption that lies at the heart of many causal discovery algorithms, distribution shifts across different batches mean that blindly grouping all data together from multiple batches will result in biased or incorrect estimation of causal networks (Lee and Tsui, 1982; Tillman, 2009). Even if we do not stack all the data together and instead learn individual models over each incoming batch, it is non-trivial to combine those models into a single model.

To address these problems, we propose an information-theoretic framework to learn causal networks for data arriving over multiple batches over time. We build this framework on the algorithmic model of causality and use the Algorithmic Markov Condition (AMC) (Janzing and Schölkopf, 2010b), a postulate stating that the true causal factorization of the joint distribution has the lowest Kolmogorov complexity. This allows us to uniquely identify a fully directed overall causal network in a continual learning fashion. While the Kolmogorov complexity is not computable itself, it can be instantiated in a statistically well-founded manner using the Minimum Description Length (MDL) principle (Marx and Vreeken, 2021).

We propose the ConCausD framework, which can be implemented using any AMC-based causal discovery approach, can leverage already learned information about causal structure, and does not require re-learning of causal networks over *all* of the data unless strictly necessary. Moreover, ConCausD provides more flexibility compared to the existing causal discovery approaches as it keeps track of multiple potential causal structures at each time step instead of forcing the prediction of a single causal model. We postulate that as a joint distribution over all the received batches gets closer to the true underlying distribution, ConCausD will converge to the true causal model.

This paper is organized as follows: In Section 2, we describe the problem setup and the preliminaries required to formalize our approach. We describe our proposed approach, ConCausD, in Section 3, and discuss its implications as well as ongoing work in Section 4 before providing concluding remarks in Section 5.

2. Preliminaries

Problem Setup We consider a setting where we have data arriving perennially in batches over time $\mathcal{D} = \{d_1, d_2, ...\}$, where d_t represents the batch of data arriving at time-step t and $\mathcal{D}_t$ represents the subset $\{d_1, ..., d_t\}$. Each dataset $d_i \in \mathcal{D}$ is defined over the identical set of m variables $\mathcal{X} = \{X_1, X_2, ..., X_m\}$. For the framework described in this work, this set can either consist of continuous-valued data or discrete valued data, or a mixture of both, as long as we can define a lossless compressor to approximate the Kolmogorov Complexity (described later in this section) of the incoming data under a proposed model. Each $d_i \in \mathcal{D}$ does not need to have i.i.d samples. We do assume, however, that $\lim_{t\to\infty} \mathcal{D}_t$ is i.i.d. Simply stated, we assume that with a large enough number of batches, we will have i.i.d data.

We work under the setting where variables in $\mathcal{X}$ are causally related to each other, and a Structural Causal Model (SCM) $\mathcal{S}$ (Pearl, 2009) over $\mathcal{X}$ models a joint distribution P over $\mathcal{X}$ corresponding to the observation distribution of the system. A causal Directed

Acyclic Graph (DAG) G over $\mathcal{X}$ is a graph where the nodes represent random variables $\{X_1, X_2, ..., X_m\}$ and edges show the causal relationship between those variables as entailed by $\mathcal{S}$. A directed edge $X_i \to X_j$ implies that X_i is the *causal parent* or a *direct cause* of X_j. We define pa_j to be the set of all causal parents of X_j. We assume that the true underlying causal DAG that captures the structure of the physical process between the variables remains the same throughout. Note that this setup does not rule out the presence of interventional datasets in $\mathcal{D}$ as we discuss in Sec. 4.

When working with causal DAGs, we make the common assumptions, namely the 1) causal Markov condition (Spirtes et al., 2000b), 2) faithfulness condition (Spirtes et al., 2000b), and 3) causal sufficiency (Pearl, 2009). The combination of these assumptions implies that each separation present in the true graph G is an independence in the (true) joint distribution P over the random variables $\mathcal{X}$ and vice versa. This allows us to discover causal networks from observational data up to the Markov equivalence class (Glymour et al., 2019). With additional assumptions over the data generating mechanisms (Peters et al., 2017), such as 4) non-linearity of the causal relation alongside independent additive Gaussian noise term (Hoyer et al., 2009), or 5) the low-noise assumption (Blöbaum et al., 2018; Marx and Vreeken, 2019) it is possible to go beyond the Markov equivalence class and discover a fully oriented causal network (Shimizu et al., 2006; Peters et al., 2014; Mian et al., 2021). For our proposed framework, we will require assumptions 1-3, and either assumption 4. or 5.

Information Theoretic Causal Discovery Information theoretic causal discovery builds on top of Kolmogorov Complexity (Kolmogorov, 1965). The Kolmogorov complexity of a finite binary string x is defined as the ***length*** of the ***shortest*** binary program, p^*, for a Universal Turing machine $\mathcal{U}$ that produces x as its output and then ***halts***. One could think of p^* as the ultimate lossless compressor for x, and its length to be the best lossless compression of x. This idea is similarly extendable to a probability distribution P. The Kolmogorov complexity $K(P)$ of a probability distribution P is the length of the shortest program that outputs $P(x)$ up to the specified precision q (Li and Vitányi, 2009). Formally stated,

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

Using Kolmogorov complexity, Janzing and Schölkopf (2010b) postulate the Algorithmic Markov Condition (AMC).

Postulate 1 (Janzing and Schölkopf (2010b)) *A causal DAG G over random variables $\mathcal{X}$ with joint density P is only acceptable if the shortest description of P factorizes as*

$$K(P(X_1, \ldots, X_m)) = \sum_{j=1}^{m} K(P(X_j \mid pa_j)) \ . \tag{1}$$

which holds up to an additive constant.

Postulate 1 states that under the assumption that the underlying causal graph over variables can be modeled by a DAG, the true causal DAG will be the minimizer of Eq. (1).

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

input : MDL-based causal discovery algorithm $\mathcal{A}$, episodes $\mathcal{E}$ arriving over time
output: candidate causal models $\mathcal{M}$

$\mathcal{M} \leftarrow \{\}$
$\tau \leftarrow 0$
$\tau_{\max} \leftarrow k$
while a new episode $\mathcal{E}_i$ arrives **do**
  $\mathcal{M} \leftarrow$ Update $(\mathcal{E}_i, \mathcal{M}, \mathcal{A})$
  $\tau \leftarrow \tau + 1$
  if $\tau == \tau_{\max}$ **then**
    $\mathcal{M} \leftarrow$ Merge $(\mathcal{M}, \mathcal{A})$
    $\tau \leftarrow 0$
  end
  yield $\mathcal{M}$
end

Hence, in an ideal world, if one could compute Kolmogorov complexity of the data under a proposed DAG, the true causal DAG will result in the best *compression* of the data.

Kolmogorov complexity, however, is not computable due to the halting problem. We can nevertheless approximate it from above in a statistically well-founded way through lossless compression (Li and Vitányi, 2009), using the Minimum Description Length (MDL) principle (Rissanen, 1978; Grünwald, 2007). Marx and Vreeken (2021) prove a formal connection between AMC and MDL by showing that the MDL formulation gives (on expectation) the same inference result as the original postulate. Therefore, in the limit where the number of samples $n \to \infty$, finding the true DAG can be achieved by finding the minimizer of a suitable lossless MDL score.

MDL, for a given model class $\mathcal{M}$, chooses the best model $M \in \mathcal{M}$ for data D as the one that minimizes

$$L(D, M) = L(M) + L(D \mid M),$$

where $L(M)$ is the length in bits of the description of M, and $L(D \mid M)$ is the length in bits of the description of data D given M. Stated with reference to Eq. (1), we consider our model class M to consist of all tuples of the form (G, S_G), where G can be any graph from the space of DAGS, and S_G is the SCM, or the generating mechanism for variables, under their parents specified in G. Simply put, we seek to find that structure, and corresponding set of functional mappings, that results in the best compression for a given data D.

Armed with the knowledge of AMC, we now describe how we can leverage this idea of AMC-based causal discovery to propose a framework for continual learning of causal networks.

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 $\mathcal{M}$

$\hat{M} \leftarrow \mathcal{A}.\textsc{Learn}(\mathcal{E}_i)$
$L^\star \leftarrow \textsc{Sum\_Costs}(\mathcal{M}) + \textsc{Cost}(\hat{M})$
$\mathcal{M}^\star \leftarrow \mathcal{M} \oplus \hat{M}$
foreach $M_j \in \mathcal{M}$ **do**
  $\tilde{\mathcal{M}} \leftarrow \textsc{Copy}(\mathcal{M})$
  $\tilde{\mathcal{M}}[j].\textsc{Add\_Data}(\mathcal{E}_i)$
  $\tilde{\mathcal{M}}[j] \leftarrow \mathcal{A}.\textsc{Resume}(\tilde{\mathcal{M}}[j])$
  $L_j \leftarrow \textsc{Sum\_Costs}(\tilde{\mathcal{M}})$
  if $L_j < L^\star$ **then**
    $(L^\star, \mathcal{M}^\star) \leftarrow (L_j, \tilde{\mathcal{M}})$
  end
end
return $\mathcal{M}^\star$

3. Continual Causal Discovery

In this section we propose a framework to continually discover causal networks over data that arrives in batches over time. We first provide a motivating example to explain our idea, and then use this motivating example to describe the framework.

As a motivating example, consider the scenario where we already have the knowledge of the best model $M^* = (G^*, S^*)$ given to us a priori. Then, for each data batch d_t that arrives, we can be guaranteed that M^* is going to compress this given batch the best, by the virtue of Eq. (1). Next, consider that we also know apriori that each batch d_t is not i.i.d. but in fact has selection bias such that each d_t only has samples either from S^*_+ or S^*_-, where S^*_+ resp. S^*_- are two disjoint subsets of the domain of S^*, such that their union again gives the full SCM. In this new scenario, we can have two separate best models, namely M^*_+ and M^*_-, depending on which kind of d_t we receive. Nevertheless, by merging these two models into one, we can arrive at the true model once more. It is straightforward to see that this argument can be extended to more than two splits of S^* e.g. three different regimes shown in Fig. 1. The key point is that as long as we can merge the models, we can arrive at the true model eventually.

The example above describes our scenario exactly, but in reverse order. In practice, each d_i we receive can be thought of as a local *snapshot* of the overall joint distribution over $\mathcal{X}$. If we can stitch enough snapshots together, we could arrive at the full picture. This is what we propose.

The main idea behind our proposed approach is to allow for more than one causal model at any time step and progressively merge them together, until eventually we converge to one model. Every time we receive a new batch of data, we evaluate its compression based on existing models as well as learn a model over this data. If the existing models provide good

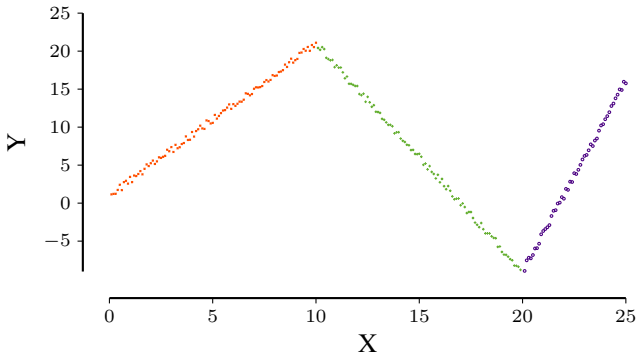

Figure 1: Hypothetical scenario between two variables X and Y with ground truth $X \to Y$. Existing causal discovery approaches will fail if data continually arrives in batches of biased samples from three different regimes i.e $\{[0, 10), [10, 20), [20, 25]\}$ as shown in this figure.

enough compression for this batch, we assign this batch to the best compressing model, else we use the model learned for this data and add it to our list of candidate models. Concretely stated, if evaluating on existing models already results in a better compression for the new data than its own best model, we simply assign the data to the existing model instead of storing a new model just for this data. We periodically check if we can merge the existing models. If merging two models into one results in a better compressing model for data of both models, we perform the merging. We call this proposed framework, Continual Causal Discovery, ConCausD, and describe its outline in Algorithms. 1 and 2. Next we describe the steps in detail.

ConCausD framework can be implemented using any Information-theoretic AMC-based causal discovery method $\mathcal{A}$ including but not limited to the proposals by Mian et al. (2021); Mameche et al. (2022), and Mameche et al. (2023). We start Alg. 1 with an empty set of Models M (L1). Each time a new batch of data d_i arrives (L4), we make an update to the existing set of plausible causal models using the Update function (L5). Note that Update step at L5 would either create an extra model or assign d_i to an existing one, thereby never decreasing the number of models. We, therefore, need a step that can merge causal models if they become sufficiently similar. To do so, we try to merge each of our existing models (L8) after a predefined number of iterations set by the max tolerance threshold, τ_{max} (L7). This intra-model evaluation allows us to merge the models that have similar performance on each others' data. At the end of each batch evaluation, we yield the current set of plausible causal models as output (L10).

The two main components of ConCausD are the Update function described in Alg. 2, and the Merge function. Update works as follows: for each batch of data, $\mathcal{E}_i$ that we receive, we use $\mathcal{A}$ to learn the locally best model, $\hat{M}$, for $\mathcal{E}_i$ (L1), compute the total cost $L^\star$ of storing $\hat{M}$ as a new model alongside already existing models (L2), and create an updated list, which we consider as our best case configuration, containing existing models as well as $\hat{M}$ (L3). Next we iterate over already existing models (L4). For each model, we check

if adding $\mathcal{E}_i$ to this model instead of using $\hat{M}$ costs less than storing $\hat{M}$ explicitly (L5-8). We do this by first adding $\mathcal{E}_i$ to the data in M_j (L6), and then *resuming* causal learner $\mathcal{A}$ to adjust to parameters change induced due to $\mathcal{E}_i$ (L7). Intuitively, if $\mathcal{E}_i$ comes from the same distribution as M_j, the overall cost of additionally storing $\hat{M}$ would be higher due to the overhead of storing (potentially) identical models twice. After evaluating each model, we update the best model configuration, if our evaluation results in a lower number of bits for this configuration (L8-9). Finally, we return the best configuration (L12). The Merge function shown in Alg. 1 line 8 works similarly to Update, except that we repeatedly learn $\hat{M}$ pairwise across existing models, instead of using $\mathcal{E}_i$ from an incoming batch. If a model M_{ij} over combined data costs fewer bits, we use this model instead of models M_i and M_j.

4. Related Work

A growing body of work studies causal discovery from ***observational data***, introducing constraint-based (Pearl, 2009), score-based Spirtes et al. (1999); Huang et al. (2018) and hybrid (Squires et al., 2020) approaches to discover the Markov equivalence class of the causal graph from an i.i.d. data distribution. Other works study assumptions to identify additional causal directions (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2014; Blöbaum et al., 2018), including the line of work of information-theoretic approaches (Marx and Vreeken, 2019; Mian et al., 2021) that we follow, which is inspired by the algorithmic framework of causation going back to Janzing and Schölkopf (2010a).

More recently the interest in causal discovery has turned towards ***interventional data***, and constraint- and score-based (Mooij et al., 2016; Zhang et al., 2017; Squires et al., 2020) as well as information-theoretic approaches (Mian et al., 2023; Mameche et al., 2023) have been proposed for this setting. These methods can discover a shared causal network from data in different contexts with distribution shifts, causal mechanism changes, or interventions, without knowing which variables are affected by such changes. All aforementioned methods however assume datasets that come from multiple contexts that are both known and unbiased.

Finally, fewer works investigate ***selection bias*** where a given dataset is not identically distributed, but rather some part of the relevant domain is observed, and other parts remain unobserved. Pearl (2012) study this problem under a missingness framework (Rubin, 1976; Little and Rubin, 2019) and give conditions under which information about causal mechanisms is recoverable in such cases (Bareinboim et al., 2014). Other approaches exist for correcting selection bias (Boeken et al., 2023) when additional (privileged) information is available. Finally, Kaltenpoth and Vreeken (2023) propose an approach for identifying whether selection bias holds.

However, while multi-context approaches cannot handle selection bias, existing work in selection bias does not consider multiple datasets that need to be matched together to discover a causal model. Neither lines of work address a dynamic setting with data arriving continually over time. This motivates us to propose our framework for ***continual learning*** of causal networks over such data.

5. Discussion and Ongoing Work

We believe that there are a number of advantages to using ConCausD framework. First, unlike existing causal discovery approaches that inherently predict one single network over all data, ConCausD can propose more than one plausible causal network based on the batches of data seen so far. Second. each new incoming data can be assigned to the best compressing model without strictly having the need to learn a new or updated model each time. Moreover, it is straightforward to see that having multiple models allows ConCausD to deal with biased batches of data as shown in Fig. 1, where samples from the orange regime would be assigned to a model different from the one for samples from the indigo regime. Eventually, with enough data from the green regime, merging the data across all models would result in an i.i.d. sample, which can then be used to reliably learn the true underlying causal network using the same causal discovery method $\mathcal{A}$.

While we only provide a framework outline in this short paper, ConCausD can be implemented using any AMC-based causal discovery approach, a number of which have been proposed recently (Mian et al., 2021; Mameche et al., 2022). Furthermore, the application of ConCausD framework is not just limited to observational data. If the incoming batches either come from observational data or from an intervention distribution, one would expect ConCausD to converge to two different models in the limit, one of which models the interventional SCM and the other one which models the observational SCM. Once again, this should be possible to achieve using any AMC-based method that works with interventional data (Mian et al., 2023; Mameche et al., 2023).

While we have the blueprint for ConCausD, there still remain a number of implementation-related questions that need to be answered in practice, the most obvious one being a statistically sound test to decide when two existing models are sufficiently similar enough to be merged. Currently we are working on an implementation of our proposed framework to build a working algorithm, with provable theoretical guarantees.

6. Conclusion

In this short paper we introduced the ConCausD framework for learning causal networks in a continual fashion. We use an information-theoretic approach to continuously build and merge plausible causal models, until we converge to the single underlying causal network. Our proposed framework can be instantiated using any causal discovery algorithm that is based on the Algorithmic Markov Condition. As ongoing work, we are implementing a proof-of-concept for ConCausD, with provable theoretical guarantees.

References

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. 28(1), 2014.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909. PMLR, 2018.

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

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

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 2019.

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

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696, 2009.

B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. In *KDD*. ACM, 2018.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE TIT*, 56(10):5168–5194, 2010a.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE TIT*, 56(10):5168–5194, 2010b.

David Kaltenpoth and Jilles Vreeken. Identifying selection bias from observational data. *AAAI*, pages 8177–8185, 2023.

A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11, 1965.

Sik-Yum Lee and Kwok-Leung Tsui. Covariance structure analysis in several populations. *Psychometrika*, 47, 1982.

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

Roderick Little and Donald Rubin. Statistical analysis with missing data, third edition. 04 2019. doi: 10.1002/9781119482260.

Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Discovering invariant and changing mechanisms from data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1242–1252, 2022.

Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal models under independent changes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Alexander Marx and Jilles Vreeken. Identifiability of cause and effect using regularized regression. In *KDD*. ACM, 2019.

Alexander Marx and Jilles Vreeken. Formally justifying mdl-based inference of cause and effect. *arXiv preprint arXiv:2105.01902*, 2021.

Osman Mian, Alexander Marx, and Jilles Vreeken. Discovering fully oriented causal networks. 2021.

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

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

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

Judea Pearl. A solution to a class of selection bias problems. 2012.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(1):465–471, 1978.

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

Shohei Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81, 2012.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7, 2006.

Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21, 1999.

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

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000b.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

Robert E Tillman. Structure learning with independent non-identically distributed data. In *ICML*, pages 1041–1048, 2009.

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