Evaluation of Causal Structure Learning Algorithms via Risk Estimation

Marco F. Eigenmann
Seminar für Statistik
ETH Zurich
Zurich, Switzerland

Sach Mukherjee
German Center for Neurodegenerative Diseases (DZNE)
Bonn, Germany

Marloes H. Maathuis
Seminar für Statistik
ETH Zurich
Zurich, Switzerland

Abstract

Recent years have seen many advances in methods for causal structure learning from data. The empirical assessment of such methods, however, is much less developed. Motivated by this gap, we pose the following question: how can one assess, in a given problem setting, the practical efficacy of one or more causal structure learning methods? We formalize the problem in a decision-theoretic framework, via a notion of expected loss or risk for the causal setting. We introduce a theoretical notion of causal risk as well as sample quantities that can be computed from data, and study the relationship between the two, both theoretically and through an extensive simulation study. Our results provide an assumptions-light framework for assessing causal structure learning methods that can be applied in a range of practical use-cases.

1 INTRODUCTION

Causal structure learning has seen many recent developments and the literature is growing rapidly. A range of algorithms have been developed under different assumptions. These include, among others, PC (Spirtes, Glymour, & Scheines, 2000), FCI (Spirtes et al., 2000), GES (Chickering, 2002b), LiNGAM (Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006), MMHC (Tsamardinos, Brown, & Aliferis, 2006), GIES (Hauser & Bühlmann, 2012), RFCI (Colombo, Maathuis, Kalisch, & Richardson, 2012), FCI+ (Claassen, Mooij, & Heskes, 2013), order-independent PC (Colombo & Maathuis, 2014), rank PC (Harris & Drton, 2013), CAM (Bühlmann, Peters, & Ernest, 2014), ICP (Peters, Bühlmann, & Meinshausen, 2016), AGES (Eigenmann, Nandy, & Maathuis, 2017), ARGES (Nandy, Hauser, & Maathuis, 2018), LGES (Frot, Nandy, & Maathuis, 2019), and MRCL (Hill, Oates, Blythe, & Mukherjee, 2019).

The majority of these papers contain theoretical guarantees for the developed algorithms as well as simulation studies showing their empirical performance, including comparisons with competing algorithms. In simulation studies estimated graphs can be compared to the ground truth, e.g. using the Structural Hamming Distance (SHD) (Acid & de Campos, 2003; Tsamardinos et al., 2006) or the more causal oriented Structural Intervention Distance (SID) (Peters & Bühlmann, 2015). Alternatively, one can consider particular features related to a graph like the total causal effect between two nodes (Maathuis, Kalisch, & Bhlmann, 2009).

However, by design and scope, simulation studies have some key limitations. In particular, good performance in a simulation does not imply good performance on a given real-world problem, since a real data-generating system may violate model assumptions in such a way as to strongly affect the relevant output. While model assumptions may be tested in principle using various statistical tools, causal assumptions in particular can be difficult if not impossible to test directly. Thus, in practice, given a data set obtained from a specific system, it remains challenging to choose among algorithms, or to assess a given algorithm. Some work has been done to fill this theoretical-empirical gap (Hill, Heiser, Cokelaer, & et al., 2016; Mooij & Heskes, 2013; Sachs, Perez, Pe'er, Lauffenburger, & Nolan, 2005). This usually involves very interesting and challenging interdisciplinary collaborations which allow to infer a ground truth to which causal methods can be compared.

In this paper, we address the question of evaluating causal structure learning in a more general sense. Our approach is rooted in a decision-theoretic view of causal structure learning and leads to procedures that could be applied generally, wherever suitable data is available. Thus, our goal is not to propose a new approach to esti-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

mate causal graphs, but a new approach to assess existing methods in a problem-specific manner.

The remainder of the paper is organized as follows. We begin with a problem statement, clarifying precisely the question we seek to address. We then propose a notion of *causal risk* as well as corresponding sample quantities that could be used to assess causal risk in practice, and study their relationship. We then show results from a large simulation study, covering more than 40,000 data-generating regimes, aimed at investigating the practical performance of the criteria we propose.

2 PROBLEM STATEMENT AND SUMMARY OF CONTRIBUTIONS

Problem statement. We aim to evaluate the performance of causal structure learning algorithms on a given data set containing some observational and some interventional data. Ideally, we would wish to be able to select the best performing algorithm (among those considered) for the specific problem setting. This problem statement acknowledges that different methods may perform better or worse in specific problem settings (this will becomes precise via the decision-theoretic framework we introduce below). We want to construct an assumptions-light framework and therefore will only assume that the data come from a structural equation model (SEM; see Definition 3.1), without imposing many restrictions on the SEM. In particular, we will not assume joint independence or a particular distribution for the noise terms, nor will we assume acyclicity.

Why assessment of causal learning is hard. It is useful to consider at a high-level why empirical assessment of causal structure learning methods is nontrivial and different from familiar non-causal tasks in machine learning and statistics. In typical non-causal tasks (such as classification/regression or probabilistic modelling, e.g. via non-causal graphical models) performance measures rooted in classical sampling theory make sense, because the core assumption is that all data – current and future – share the same probability model. In contrast, a causal model encodes a *collection of distributions* (e.g. arising from different interventions on the system; see Def. 3.1 and 3.2 below) and this limits the scope of familiar sampling theory-based approaches to assessment.

It is instructive to consider this difference with an example. In a regression problem, assuming that a fixed and unique distribution underlies the data permits (i) the use of residuals as proxies for the statistical noise (that can be used to check assumptions about the noise, via e.g., Tukey-Anscombe or QQ-plots) and (ii) the use of various cross-validation-type methods to test prediction accuracy. For variable selection, candidate procedures can be evaluated using likelihood methods applied to selected variables, and similar strategies can be used for non-causal model selection in general. In contrast, for causal problems, the fact that one is dealing with a collection of potentially very different distributions does not allow the use of sampling techniques like cross-validation in a straightforward way. Moreover, in causal systems a good model needs to go beyond out-of-sample performance and cope with (potentially strongly) out-of-distribution scenarios from which no data may be available.

Summary of contributions. Our main contributions are as follows. (i) We show how causal structural learning can be viewed through a decision-theoretic lens, and propose a notion of causal loss that allows assessment via expected loss or risk. (ii) We study the question of estimating causal risk from data and propose assumptions-light risk estimation procedures that can be used in practice using interventional data. (iii) We study the behaviour of our procedures in theory and in an extensive simulation study spanning more than 40,000 data-generating regimes.

The core idea of our approaches is to exploit information given by the interventional data to generalize the performance of the algorithms to other unseen interventions. To this end, we use simple statistical tests that do not require the same types of assumptions as causal structure learning methods, and whose output allows us to estimate a useful notion of causal risk. Causal relationships have been estimated in a risk minimization framework (Arjovsky, Bottou, Gulrajani, & Lopez-Paz, 2019), and held-out interventional data has been used in applications (Hill et al., 2016), but to the best of our knowledge, the present work is the first formal risk estimation framework for causal structure learning.

3 CAUSAL RISK

3.1 PRELIMINARIES AND NOTATION

We associate vertices in a directed graph G with variables and say that variable X_j is a parent of X_i if the directed edge $X_j \to X_i$ is included in G. We say that X_j is a descendant of X_i in G if there is a directed path $X_i \to \ldots \to X_j$ in G. We denote the set of parents and descendants of X_i in G by $\mathrm{Pa}(G, i)$ and $\mathrm{De}(G, i)$, respectively. We let $[p] := \{1, \ldots, p\}$.

We assume that the data come from a structural equation model (SEM).

Definition 3.1. *A structural equation model is a system of equations $\mathcal{S} = \{S_1, \ldots, S_p\}$ on a set of variables*

$\{X_1, \ldots, X_p\}$:

$$S_i : X_i \leftarrow f_i(X_{\mathrm{Pa}(G,i)}, \varepsilon_i), \qquad i \in [p], \qquad (1)$$

where G denotes the directed graph associated with the SEM, and the noise terms $\varepsilon_1, \ldots, \varepsilon_p$ have mean 0, and finite variance.

The assignment arrow in Equation (1) emphasizes the causal relationship between its left and right hand sides. In other words, $\mathcal{S}_i$ is understood as the generating mechanism of X_i. A SEM can be represented by a directed graph G, where for any pair (X_i, X_j), there is a directed edge $X_j \to X_i$ if X_j is involved in structural equation $\mathcal{S}_i$, that is, if $j \in \mathrm{Pa}(G, i)$. Thus, a direct edge represents a direct effect and $X_{\mathrm{Pa}(G,i)}$ are the direct causes of X_i.

Some causal modelling frameworks require acyclicity of the graph G and independence of the noise terms, but we do not assume this here.

Definition 3.2. *An intervention on a set of nodes $\{X_i : i \in I\}$ is modelled by replacing the respective structural equations by*

$$\widetilde{S}_i : X_i \leftarrow \widetilde{f}_i(X_{\mathrm{Pa}(\widetilde{G},i)}, \widetilde{\varepsilon}_i), \qquad i \in I,$$

where $\widetilde{f}_i$, $\widetilde{G}$ and $\widetilde{\varepsilon}_i$ are respectively the functional form, the directed graph, and the noise variable under the intervention. The structural equations for $i \notin I$ remain unchanged.

We assume that we have n_0 i.i.d. observations from an unknown SEM (see Def. 3.1), as well as some i.i.d. observations from different interventions (see Def. 3.2). For ease of exposition, we assume that we only have single interventions, meaning that an intervention affects exactly one node or structural equation. We denote by $\boldsymbol{\iota} \subseteq [p]$ the collection of nodes on which we have data from single interventions, and we let n_i, $i \in \boldsymbol{\iota}$ be the corresponding sample sizes. It will turn out to be convenient to use an augmented set $\bar{\boldsymbol{\iota}}$ that includes the observational data, denoted by a 0, i.e. $\bar{\boldsymbol{\iota}} = \{0\} \cup \boldsymbol{\iota}$. The corresponding sample sizes are denoted by $\mathbf{n} = (n_i : i \in \bar{\boldsymbol{\iota}})$. The total sample size is $N = \sum_{i \in \bar{\boldsymbol{\iota}}} n_i$.

We denote by $\boldsymbol{\Theta}$ all parameters necessary to fully represent the SEM and its interventions. For instance, the necessary parameters for a linear Gaussian SEM would be the edge weights, the means and variances of the noise terms in the original regime, and the new weights, means, and variances under the interventions in $\boldsymbol{\iota}$. To emphasize that the true underlying graph representing our SEM is unknown, we will from now on denote it by $G^\star$. Further, we denote by $G^\star_{\boldsymbol{\Theta}}$ the multivariate distributions that arise with the parameters in $\boldsymbol{\Theta}$.

Data coming from the SEM and its interventions are denoted by $\mathbf{X}_{\mathbf{n},\bar{\boldsymbol{\iota}},\boldsymbol{\Theta}} \sim G^\star_{\boldsymbol{\Theta}}$, where $\mathbf{X}_{\mathbf{n},\bar{\boldsymbol{\iota}},\boldsymbol{\Theta}} \in \mathbb{R}^{N \times p}$. We emphasize that this represents a sample from a *collection* of $|\,\boldsymbol{\iota}\,| + 1$ multivariate distributions arising from the underlying causal system. To simplify notation, we will in the sequel suppress the dependence on $\mathbf{n}$ and $\boldsymbol{\Theta}$, and indicate only the set of interventions $\bar{\boldsymbol{\iota}}$. That is, we write $\mathbf{X}_{\bar{\boldsymbol{\iota}}}$ instead of $\mathbf{X}_{\mathbf{n},\bar{\boldsymbol{\iota}},\boldsymbol{\Theta}}$. We also consider leaving out data on certain interventions, considering only a subset $\bar{\boldsymbol{\eta}} \subset \bar{\boldsymbol{\iota}}$. In that case, we write $\mathbf{X}_{\bar{\boldsymbol{\eta}}}$. We consider either all or none of the samples under a certain intervention, so that the sample sizes corresponding to $\bar{\boldsymbol{\eta}}$ equal the corresponding entries of $\mathbf{n}$.

3.2 THE ORACLE RISK FUNCTION

We first define a theoretical notion of risk that involves the true graph G^*. We emphasize that this theoretical quantity cannot be computed in practice. We will consider practically applicable estimates of the theoretical risk in Section 3.3.

Let $\hat{H}$ be a causal structure learning algorithm that returns a graph. Let $\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}})$ denote the graph that is returned when the algorithm is applied to data set $\mathbf{X}_{\bar{\boldsymbol{\iota}}}$. The general form of the risk function we propose is the following:

$$R_{\bar{\boldsymbol{\iota}}}(\hat{H}) = \mathbb{E}_{\mathbf{X}_{\bar{\boldsymbol{\iota}}} \sim G^\star_{\boldsymbol{\Theta}}} \left[L(G^\star, \hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}})) \right], \qquad (2)$$

where L is a loss function acting on a pair of directed graphs.

Note that this notion of risk is problem-specific in the sense that it quantifies the finite sample efficacy of method $\hat{H}$ in the specific context defined by the system $G^\star_{\boldsymbol{\Theta}}$. This allows for the possibility that a given method may do well in some settings but not in others.

There are several choices to be made in order to define a concrete risk function to study. In particular, we must define a loss function. As we will motivate in Section 3.2.1 below, we will consider a node-wise loss function that compares the *descendants* of nodes of the graphs $G^\star$ and $\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}})$. In particular, we consider the following special case of Equation (2)

$$R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}) = \mathbb{E}_{\mathbf{X}_{\bar{\boldsymbol{\iota}}} \sim G^\star_{\boldsymbol{\Theta}}} \frac{1}{p} \sum_{i=1}^{p} \left(J(\mathrm{De}(G^\star, i), \mathrm{De}(\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}), i) \right), \qquad (3)$$

where $J(A, B)$ denotes the Jaccard distance between two sets A and B, defined as $J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$ if $A \cup B \neq \emptyset$ and 0 otherwise.

We note that due to identifiability issues, some causal structure learning algorithms do not return a directed

graph, but for example a partially directed graph. In that case, $\mathrm{De}(\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}), i)$ should be adapted, depending on the interpretation of the output graph. For example, if the output is a completed partially directed acyclic graph (CPDAG; Andersson, Madigan, & Perlman, 1997; Chickering, 2002a), the descendants of X_i can be replaced by the set of possible descendants of X_i, i.e., the set of nodes X_j for which there is a partially directed path from X_i to X_j. We will make this concrete in Section 4. To set things up, however, we will first omit these issues and think in terms of directed graphs.

3.2.1 Remarks on the Oracle Risk Function

A key challenge in working with Equation (2) is the presence of $G^\star$. In particular, if $\boldsymbol{\iota} \subsetneq [p]$ there is at least one intervention that is of interest but under which we have no data. This limits the utility of standard likelihood-based and cross-validation-type approaches since we cannot sample from the unobserved intervention. Therefore, we need an approach that does not rely on explicit knowledge of $G^\star$.

When looking at general interventions as defined in Definition (3.2) we can quickly see that the observational and interventional data distinguish themselves on the descendants of the intervened node. Indeed, the distribution of all descendants of X_i is potentially different. It is therefore quite natural to use the descendants as a feature that captures information that is causally relevant, estimable and that may generalize to unobserved interventions.

3.3 RISK ESTIMATORS

3.3.1 Descendant Estimation

Since the true underlying graph G^* is unknown, we must construct an estimate for $\mathrm{De}(G^\star, i)$ in Equation (3). This is virtually impossible for $i \notin \boldsymbol{\iota}$. For $i \in \boldsymbol{\iota}$, however, it is feasible by comparing the observational data, $\mathbf{X}_{\{0\}} \in \mathbb{R}^{n_0 \times p}$, to the data under the intervention on node i, $\mathbf{X}_{\{i\}} \in \mathbb{R}^{n_i \times p}$. In particular, we can compare the jth column of $\mathbf{X}_{\{0\}}$ to the jth column of $\mathbf{X}_{\{i\}}$. If these are significantly different, we conclude that the intervention on X_i has affected X_j, and hence that X_j is a descendant of X_i. This approach only works if the aggregated effect of directed paths from X_i to X_j in G^* does not cancel, if there is at least one such path. The latter is related to the common faithfulness assumption (Spirtes et al., 2000).

Concretely, for each intervention node $i \in \boldsymbol{\iota}$, we conduct $p-1$ two-sample tests, comparing the jth column of $\mathbf{X}_{\{0\}} \in \mathbb{R}^{n_0 \times p}$ to the jth column of $\mathbf{X}_{\{i\}} \in \mathbb{R}^{n_i \times p}$, for $j \in [p] \setminus \{i\}$. If a significant difference is found, X_j is declared to be an estimated descendant of X_i. With a slight abuse of notation we denote the set of estimated descendants of node X_i by $\widehat{\mathrm{De}}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}, i)$. This will substitute $\mathrm{De}(G^\star, i)$ in the risk estimators.

We note that our framework is general, in the sense that we do not specify the type of interventions (e.g., do-interventions or shift interventions) nor the statistical tests that should be used. We will require, however, that one has some knowledge of the type of interventions, so that one can conduct appropriate statistical tests to detect the descendants. We refer to Section 4.2 for the concrete example we used in the simulation study.

3.3.2 The Main Risk Estimator

We consider three risk estimators: the first two (Equation (5) and (6)) serve as auxiliary risk estimators and are needed to construct the third one (Equation (7)), which is the main risk estimator we propose.

Ideally, we would like to compute

$$\hat{R}^J_{\bar{\boldsymbol{\iota}},\text{oracle}}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) = \frac{1}{p}\sum_{i=1}^{p} J(\mathrm{De}(G^\star, i), \mathrm{De}(\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}), i)), \tag{4}$$

where $\mathrm{De}(G^\star, i)$ should be replaced by some estimate of it. As discussed in Section 3.3.1, however, we can only estimate $\mathrm{De}(G^\star, i)$ by $\widehat{\mathrm{De}}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}, i)$ for $i \in \boldsymbol{\iota}$.[1] Restricting the node-wise sum to these terms and scaling appropriately, we obtain our first risk estimator:

$$\hat{R}^J_{\bar{\boldsymbol{\iota}},\text{naive}}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) = \frac{1}{|\boldsymbol{\iota}|}\sum_{i \in \boldsymbol{\iota}} J(\widehat{\mathrm{De}}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}, i), \mathrm{De}(\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}), i)). \tag{5}$$

Equation (5) is a natural first step since it contains everything we can estimate. Indeed, for $i \notin \boldsymbol{\iota}$, $\widehat{\mathrm{De}}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}, i)$ is not defined and cannot be defined in a reasonable and natural way. Hence, in order to sum over all nodes as done in Equation (4), we would have to make outside assumptions that cannot be supported by data. This should be avoided and Section 3 of the supplementary material illustrates failure cases that can arise when doing this.

We next define a risk estimator that focuses more explicitly on the out-of-distribution aspect. This risk estimator uses a leave-one-out cross-validation-like approach on the intervention nodes and is defined as

$$\hat{R}^J_{\bar{\boldsymbol{\iota}},CV}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) = \frac{1}{|\boldsymbol{\iota}|}\sum_{i \in \boldsymbol{\iota}} J(\widehat{\mathrm{De}}(\mathbf{X}_{\{0,i\}}, i), \mathrm{De}(\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}}\setminus\{i\}}), i)). \tag{6}$$

This expression uses a *distributional splitting* scheme in which the interventional data is split into two disjoint

[1] The "hat" on $R^J_{\bar{\boldsymbol{\iota}},\text{oracle}}$ in equation (4) is used to indicate that this is a random variable that depends on the data $\mathbf{X}_{\bar{\boldsymbol{\iota}}}$.

groups.[2] This is fundamentally different from randomly splitting the sample as in classical cross-validation.

In words, Equation (6) does the following. It applies the algorithm under investigation, $\hat{H}$, $|\,\boldsymbol{\iota}\,|$ times. For each $i \in \boldsymbol{\iota}$, we pass to the algorithm all observational data and the interventional data corresponding to interventions in $\boldsymbol{\iota} \setminus \{i\}$, $\mathbf{X}_{\bar{\boldsymbol{\iota}} \setminus \{i\}}$, and determine the descendants of X_i in the resulting graph, $\mathrm{De}(\hat{H}(\mathbf{X}_{\bar{\boldsymbol{\iota}} \setminus \{i\}}), i)$. At the same time, we use the observational data and interventional data corresponding to the intervention on i, $\mathbf{X}_{\{0,i\}}$, to estimate the descendants of X_i using some two-sample tests, yielding $\widehat{\mathrm{De}}(\mathbf{X}_{\{0,i\}}, i)$. By comparing these two estimated sets of descendants, we emulate the evaluation of the performance of $\hat{H}$ on unseen interventions. Finally, this is averaged over $i \in \boldsymbol{\iota}$.

Our main risk estimator $\hat{R}^J_{\bar{\boldsymbol{\iota}},w}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}})$ combines Equation (5) and (6) as a weighted sum, where the weights, $\frac{|\,\boldsymbol{\iota}\,|}{p}$ and $\frac{p-|\,\boldsymbol{\iota}\,|}{p}$, correspond to the proportion of nodes with and without interventions, respectively:

$$\hat{R}^J_{\bar{\boldsymbol{\iota}},w}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) = \frac{|\,\boldsymbol{\iota}\,|}{p} \hat{R}^J_{\bar{\boldsymbol{\iota}},\text{naive}}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) + \frac{p-|\,\boldsymbol{\iota}\,|}{p} \hat{R}^J_{\bar{\boldsymbol{\iota}},CV}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}}). \tag{7}$$

This estimator balances both aspects, the in-distribution performance on seen interventions through $\hat{R}^J_{\bar{\boldsymbol{\iota}},\text{naive}}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}})$, and the out-of-distribution performance on unseen interventions through $\hat{R}^J_{\bar{\boldsymbol{\iota}},CV}(\hat{H}, \mathbf{X}_{\bar{\boldsymbol{\iota}}})$.

3.4 PROPERTIES OF THE MAIN RISK ESTIMATOR

We now investigate under which circumstances we can, in expectation, rank two algorithms correctly. Concretely, for a given setting characterized by $G^*_{\boldsymbol{\Theta}}$ and $\boldsymbol{\iota}$ and two causal structure learning methods $\hat{H}_1$ and $\hat{H}_2$, we investigate when the difference $R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_1) - R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_2)$ in the oracle risks (defined in Equation (3)) and the corresponding expected difference with respect to our proposed risk estimator $\hat{R}^J_{\bar{\boldsymbol{\iota}},w}$, i.e.,

$$\mathbb{E}_{\mathbf{X}_{\bar{\boldsymbol{\iota}}} \sim G^\star_{\boldsymbol{\Theta}}} \left[\hat{R}^J_{\bar{\boldsymbol{\iota}},w}(\hat{H}_1, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) - \hat{R}^J_{\bar{\boldsymbol{\iota}},w}(\hat{H}_2, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) \right],$$

have the same sign. Of course, this task should be easier if the difference in oracle risks is larger. Our results will therefore depend on

$$\delta := \left| R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_1) - R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_2) \right|.$$

Since our focus is on the performance of the risk estimator, we will assume for simplicity that the descendant estimation has oracle performance.

[2]Note that $\widehat{\mathrm{De}}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}, i) = \widehat{\mathrm{De}}(\mathbf{X}_{\{0,i\}}, i)$.

Assumption 1. *(Oracle performance of the descendant estimation)*

We assume that the descendant estimation via $\widehat{\mathrm{De}}(\boldsymbol{X}_{\bar{\boldsymbol{\iota}}}, i)$ *achieves oracle performance with respect to* $G^\star_{\boldsymbol{\Theta}}$ *and* $\boldsymbol{\iota}$*:* $\widehat{\mathrm{De}}(\boldsymbol{X}_{\bar{\boldsymbol{\iota}}}, i) = \mathrm{De}(G^\star, i)$ *for all* $\boldsymbol{X}_{\bar{\boldsymbol{\iota}}} \sim G^*_{\boldsymbol{\Theta}}$ *and* $i \in \boldsymbol{\iota}$*.*

Assumption 1 is essentially one of correctness of the statistical decisions in a classical testing sense. It allows us to write $\mathrm{De}(G^\star, i)$ instead of $\widehat{\mathrm{De}}(\mathbf{X}_{\bar{\boldsymbol{\iota}}}, i)$ for $i \in \boldsymbol{\iota}$ in the risk estimators.

Next, we need to link the expected estimated difference in performance based on the cross-validation risk estimator

$$\mathbb{E}_{\mathbf{X}_{\bar{\boldsymbol{\iota}}} \sim G^\star_{\boldsymbol{\Theta}}} \left[\hat{R}^J_{\bar{\boldsymbol{\iota}},CV}(\hat{H}_1, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) - \hat{R}^J_{\bar{\boldsymbol{\iota}},CV}(\hat{H}_2, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) \right], \tag{8}$$

using the seen interventions in $\boldsymbol{\iota}$, to the true difference in performance of $\hat{H}_1$ and $\hat{H}_2$ on unseen interventions on nodes $i \notin \boldsymbol{\iota}$,

$$\mathbb{E}_{\mathbf{X}_{\bar{\boldsymbol{\iota}}} \sim G^\star_{\boldsymbol{\Theta}}} \left[\frac{1}{p - |\,\boldsymbol{\iota}\,|} \sum_{i \notin \boldsymbol{\iota}} \Big(J(\mathrm{De}(G^\star, i), \mathrm{De}(\hat{H}_1(\mathbf{X}_{\bar{\boldsymbol{\iota}}}), i)) - J(\mathrm{De}(G^\star, i), \mathrm{De}(\hat{H}_2(\mathbf{X}_{\bar{\boldsymbol{\iota}}}), i)) \Big) \right]. \tag{9}$$

This link must only be made if there actually are unseen interventions, i.e., $|\,\boldsymbol{\iota}\,| < p$.

Now consider two algorithms $\hat{H}_1$ and $\hat{H}_2$ and a setting defined by $G^\star_{\boldsymbol{\Theta}}$ and $\boldsymbol{\iota}$ with $|\,\boldsymbol{\iota}\,| < p$. We say that $\hat{H}_1$ and $\hat{H}_2$ satisfy expected relative δ-performance on unseen interventions with respect to $G^\star_{\boldsymbol{\Theta}}$ and $\boldsymbol{\iota}$ if $\big|(8) - (9)\big| < \frac{p}{p-|\,\boldsymbol{\iota}\,|}\delta$. This will serve as our second assumption.

Assumption 2. *(Expected relative* δ*-performance on new interventions)*

We assume that algorithms $\hat{H}_1$ *and* $\hat{H}_2$ *satisfy expected relative* δ*-performance on new interventions with respect to* $G^\star_{\boldsymbol{\Theta}}$ *and* $\boldsymbol{\iota}$ *(with* $|\,\boldsymbol{\iota}\,| < p$*).*

The expected relative δ-performance on unseen interventions incorporates the following two components: (i) The performance of the algorithms using all data $\mathbf{X}_{\bar{\boldsymbol{\iota}}}$ must be similar to the performance of the algorithms when the data on one intervention is omitted, i.e., using $\mathbf{X}_{\bar{\boldsymbol{\iota}} \setminus \{i\}}$ for $i \in \boldsymbol{\iota}$, as is done in the cross-validation risk estimator. (ii) The performance of the algorithms on the seen interventions in $\boldsymbol{\iota}$ must be representative of the algorithms' performance on unseen interventions.

The latter point is most important. We note that it is not testable; it is in essence an extrapolation type assumption

that allows us to generalize the performance of the cross-validation risk estimator from seen to unseen interventions, and it becomes increasingly strong as the number of interventions decreases.

We now obtain the following theorem. Its proof can be found in Section 1 of the supplementary material.

Theorem 3.3. *Consider a setting defined by $G^*_{\boldsymbol{\Theta}}$ and $\boldsymbol{\iota}$ ($|\boldsymbol{\iota}| > 1$) and two algorithms $\hat{H}_1$ and $\hat{H}_2$ with oracle risk difference $\delta = \left|R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_1) - R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_2)\right|$.*

If $|\boldsymbol{\iota}| = p$ and $\hat{H}_1$ and $\hat{H}_2$ satisfy Assumption 1 with respect to $G^\star_{\boldsymbol{\Theta}}$ and $\boldsymbol{\iota}$, or if $|\boldsymbol{\iota}| < p$ and and $\hat{H}_1$ and $\hat{H}_2$ satisfy Assumptions 1 and 2 with respect to $G^\star_{\boldsymbol{\Theta}}$ and $\boldsymbol{\iota}$, then

$$R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_1) - R^J_{\bar{\boldsymbol{\iota}}}(\hat{H}_2)$$

and

$$\mathbb{E}_{\boldsymbol{X}_{\bar{\boldsymbol{\iota}}} \sim G^\star_{\boldsymbol{\Theta}}}\left[\hat{R}^J_{\bar{\boldsymbol{\iota}},w}(\hat{H}_1, \boldsymbol{X}_{\bar{\boldsymbol{\iota}}}) - \hat{R}^J_{\bar{\boldsymbol{\iota}},w}(\hat{H}_2, \boldsymbol{X}_{\bar{\boldsymbol{\iota}}})\right]$$

have the same sign.

This result says that, under the given assumptions, the expected estimated difference in risk of the two algorithms has the correct sign. This is reassuring, as this is a property that a sensible risk estimator should have. Theorem 3.3 does not guarantee, however, that the algorithms are correctly ranked for a particular realization of the data. In the simulations in Section 4 we will assess the practical performance of our proposed risk estimator.

We note that the assumption of expected relative δ-performance acts as expected in the following ways: (i) If the true difference in oracle risks δ is larger, then the condition is weaker. This makes sense, since we have more "room for error" in the estimation before we flip the sign. (ii) If we have interventional data on a large proportion of the nodes, then the factor $\frac{p}{p-|\boldsymbol{\iota}|}$ is large and the condition also becomes weaker. This can be explained by the fact that in this case the weighted risk estimator gives a large weight to the naive risk estimator and only a small weight to the cross-validation risk estimator. In other words, there is less out-of-distribution assessment to be done.

4 SIMULATION STUDY

In this Section, we empirically investigate the behaviour of the proposed risk estimation procedure via a simulation study. The basic strategy is as follows: we simulate data from many different known SEMs, that is, from many distributions $G^\star_{\boldsymbol{\Theta}}$. In each such regime, since we know the true graph $G^\star$, we can compute the oracle risk (defined in Section 3.2) and thereby empirically assess agreement with our proposed risk estimator. The goal is to investigate behaviour in a range of finite-sample settings, where all estimation is done using available data, as would be the case in practical applications.

In line with the theoretical framework, we want to understand whether it is possible to distinguish, in an entirely data-driven manner, whether a certain method is more effective than another. This question is most urgent when two methods differ greatly in oracle risk, since then an incorrect choice means high cost or regret. Hence, we require a set of approaches for learning structure that would be collectively expected to span a range of performance levels. To this end we included both principled causal methods and simple non-causal estimators (that were expected to perform poorly). We note that we are not surveying all potentially useful methods for any particular setting, and acknowledge that many valid algorithms for the simulated settings have not been considered. We emphasize that the goal of the simulation is not to offer guidance on specific methods that might work well in specific settings, but to study risk estimation *per se*.

4.1 CONSIDERED SETTINGS

In order to cover a wide variety of settings, we sampled a large parameter space, in a similar manner to the simulation study in Heinze-Deml, Maathuis, and Meinshausen (2018). The settings are defined below, and all parameters were sampled uniformly from the given ranges. We considered settings in which the statistical tests (for descendant estimation) were appropriate, as well as settings which violated assumptions of the tests.

The causal graph. The causal graph was taken to be a directed acyclic graph, obtained by choosing a causal order on an Erdős-Rényi graph with $p \in \{25, 50, 100, 200\}$ nodes and expected neighborhood size ENS $\in \{1.5, 2.5\}$.

The SEM. We took SEMs of the following form

$$\mathcal{S}_i : X_i \leftarrow \sum_{j=1}^{i-1} f(b_{ji}, X_j) + \varepsilon_i,\ i \in [p],$$

where the variables are assumed to be in a causal order, and $b_{ji} \neq 0$ if and only if $X_j \to X_i$ in $G^\star$. Here the nonzero b_{ji}'s are sampled uniformly from $[-3, -1] \cup [1, 3]$. For a given SEM, the link functions are all of the same type, and are chosen to be either linear, or sigmoidal (expressions appear below).

The noise variables are taken to be jointly independent. For a given SEM, they all have the same type of distribution, which is chosen to be either Gaussian or lognormal, both with mean zero. The noise variance is set to 1 for source nodes. For non-source nodes, the noise variance

and the edge weights were scaled to obtain variables with unit variance and a signal to noise ratio of 5. For details we refer to Section 2 of the supplementary material.

Interventions. We consider two types of interventions: Shift and Do-and-Shift. Both have a mean-shift component which is set to 5 throughout, meaning that the noise distribution of an intervened node is shifted by 5. For Do-and-Shift, we additionally delete all incoming edges and set the noise variance of the node to 1.

For a given SEM, the interventions were either all Shift interventions, or all Do-and-Shift interventions, and each node had a probability $P_{\boldsymbol{\iota}}$ to be intervened upon, independently of each others, where $P_{\boldsymbol{\iota}} \in \{0.1, 0.2, 50.5, 1\}$.

Data. For each SEM, one data set was generated, consisting of both observational and interventional data. The interventional sample sizes n_i, $i \in \boldsymbol{\iota}$ for a SEM were taken to be identical and equal to $n_{int} \in \{10, 100, 1000\}$. The observational sample size n_0 was set to $\max(n_{int}, 100)$.

The parameter space is summarized below. Additional details regarding the simulations can be found in Section 2 of the supplementary material.

1. Causal graph
   - Number of variables $p \in \{25, 50, 100, 200\}$
   - Expected neighborhood size ENS $\in \{1.5, 2.5\}$
2. SEM
   - Non-zero edge weights $b_{ji} \in [-3, -1] \cup [1, 3]$
   - Link functions $f(b_{ji}, X_j)$ in the SEM:
     - linear: $b_{ji}X_j$
     - sigmoidal: $b_{ji}(\frac{10}{(1+\exp(-0.65*X_j))} - 5)$
   - Noise distribution:
     - $\mathcal{N}(0, 1)$
     - log-normal$(0, 1) - e^{0.5}$
   - The noise and edge weights were scaled to obtain variables with unit variance and a signal to noise ratio of 5 for non-source nodes.
3. Interventions
   - Probability for each node to be intervened upon (independently): $P_{\boldsymbol{\iota}} \in \{0.1, 0.2, 0.5, 1\}$
   - Intervention type: Shift or Do-and-Shift, both with a mean shift of 5
4. Data
   - Sample sizes of the interventions: $n_i = n_{int}$ for all $i \in \boldsymbol{\iota}$, with $n_{int} \in \{10, 100, 1000\}$, and observational data size $n_0 = \max(n_{int}, 100)$

For each sampled setting we generated a data set and ran GES and GIES, giving estimates $\widehat{GES}$ and $\widehat{GIES}$, respectively. These are principled causal algorithms that were run with standard settings as implemented in the `R`-package `pcalg` (Kalisch, Mächler, Colombo, Maathuis, & Bühlmann, 2012). GIES is expected to perform better because it is geared towards settings with interventional data. We also considered graphs based on the Pearson correlation (expected to perform worse than the causal methods). In particular, we considered the graph with the same expected neighborhood size as $G^\star$, using in essence an oracle cut-off to the matrix of (absolute) correlations and also, for comparison, the almost empty graph consisting of only one undirected edge between the nodes with the largest correlation coefficient. We denote these two algorithms by $\widehat{ACor}$ and $\widehat{Empty}$, respectively[3].

For all algorithms we interpreted undirected edges as possibly directed edges. Accordingly, we replaced descendant sets in the risk estimator by their corresponding possible descendant sets. In order to have more stable and meaningful results we imposed the following two conditions on the settings used. First, we considered only data sets that contain at least two interventions, so that the cross-validation based estimator can use some interventional data. Second, the intervened nodes were required to have at least three descendants in total with respect to $G^\star$. Note that this is a condition on the number of descendants (not out-degree) and serves to limit the occurrence of situations with zero true positives, in which case the Jaccard loss can only take the values 0 or 1.

4.2 DESCENDANT ESTIMATION

We used two-sample t-tests for a difference in mean between observational and interventional data to obtain $\widehat{\mathrm{De}}(\mathbf{X}_{\boldsymbol{\iota}}, i) = \widehat{\mathrm{De}}(\mathbf{X}_{\{0,i\}}, i)$, which is required in the building blocks (5) and (6) of our main risk estimator (7). For each intervention we test for a difference in mean between the observational and interventional data of every node but the intervened one. The cut-off was computed with a multiplicity correction based on an empty graph, and under the assumption of Gaussian noise terms. Please see Section 2.1 of the supplementary material for details.

We expect the t-tests to be a good choice in the linear Gaussian SEM. For the linear log-normal and sigmoidal case, behavior should still be reasonable for large sam-

[3]We note that $\widehat{ACor}$ uses oracle information (true ENS of $G^\star$); this is intended to provide a simple point of comparison with correct sparsity, but with performance expected to be below an appropriate causal algorithm but better than random guessing.

ple sizes due to the central limit theorem. In the sigmoidal log-normal case, however, we can run into some issues. Indeed, a non-linear transformation applied to a non-symmetric distribution introduces an artificial mean, in the sense that one can obtain a mean shift that is not due to an intervention. This likely yields more false positives in the descendant estimation. This latter scenario is meant to assess the sensitivity of our framework to incorrect descendant estimation. (But we note that in a real-world use-case, one could center every node based on the observational data and avoid this problem.)

4.3 RESULTS

Thus, we investigate agreement in ranking under true and estimated risk in a range of data-generating regimes. Figure 1 shows a summary of results (additional results shown in supplementary material). Here a difference in true risk refers to the quantity $\hat{R}^J_{\bar{\boldsymbol{\iota}},\text{oracle}}(\hat{H}_1, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) - \hat{R}^J_{\bar{\boldsymbol{\iota}},\text{oracle}}(\hat{H}_2, \mathbf{X}_{\bar{\boldsymbol{\iota}}})$, where $\hat{H}_1, \hat{H}_2$ are the two causal structure learning methods being compared. A difference in estimated risk refers to the corresponding quantity obtained from the risk estimator: $\hat{R}^J_{\bar{\boldsymbol{\iota}},\text{w}}(\hat{H}_1, \mathbf{X}_{\bar{\boldsymbol{\iota}}}) - \hat{R}^J_{\bar{\boldsymbol{\iota}},\text{w}}(\hat{H}_2, \mathbf{X}_{\bar{\boldsymbol{\iota}}})$. We emphasize that risk estimation uses only the finite sample data generated in the specific example.

How to read the plot. The plot should be read as follows. Each cell in the upper panel corresponds to a specific data-generating regime (defined by combinations of the factors listed above). The specific regime is indicated by the labels shown and the color indicates the difference in true risk. Corresponding cells in the lower panel refer to the same regimes and show how often estimated differences in risks agreed in sign with the true risk difference (for the respective regime). For example, the very top left cell is the regime with a linear SEM with p=200 nodes, Normal noise, Do-and-Shift interventions where the probability for each node to be intervened upon was 0.1, and sample size 10. The colorbar shows that, in this case, the difference in true risk is large. The corresponding cell in the lower panel shows that the estimated difference in risk agrees in sign with the true risk difference.

The choice of the algorithms shown. We chose to show the results on the two algorithms $\widehat{GES}$ and $\widehat{ACor}$, because their comparison shows patterns that appear to hold more generally (see additional plots in Section 3.1 of the supplementary material). We note that some other pairs, such as for example $\widehat{GES}$ and $\widehat{GIES}$ (both principled causal estimators) showed even better results.

Key insights. We can see that the our main risk estimator shows good performance, i.e. agreement with the ranking under the oracle, under many different settings and across a range of true performance differences (from light blue to violet in the upper panel of the Figure). In the more favourable regimes it is often the case that all signs are estimated correctly. These regimes also include examples with a lower percentages of interventions (as shown). Since all risk estimation was done using only the finite sample, regime-specific data (as would be available in a real-world application), these results suggest that our approach, or modifications of it, could be used to assess causal structure learning in practice. Sample size and descendant estimation plays an important role (see also Section 3.3 of the supplementary material). The size of the graph positively impacts risk estimator performance. As expected, more interventions help, as this makes the assumptions of Theorem 3.3 weaker. In the sigmoidal log-normal case we see poor performance in the small graph and small sample setting. This is likely due to model mis-specification with respect to the testing approach. However, we note that this is an issue within the scope of classical statistical testing and could be resolved in practice by appropriate testing methods.

We emphasize that the results in Figure 1 concern risk estimation in specific data-generating regimes. In our framework, the performance of a method is assessed in the context of a data-generating regime. This means that selecting the "best" algorithm should not be interpreted as a general superiority statement. Further, good performance is defined with respect to the specific risk function used here. Our framework could be adapted to different definitions of descendants and different choices of loss functions more tailored to particular research interests (that might lead to different rankings).

Finally, we note that for small differences it can be more instructive to estimate the actual performance in terms of Equation (3), that is, whether both methods have a low or high causal risk. In Section 3.2 of the supplementary material we show some results concerning this task.

5 CONCLUSIONS

We proposed a formal risk estimation framework for the evaluation of causal structure learning algorithms. This new framework represents a practical tool to facilitate the use, assessment and interpretation of causal structure learning methods.

We showed theoretically and empirically that the proposed approach – that involves only quantities that can be computed from available data – is indeed able to agree with a ranking of methods that would be possible given access to a true underlying causal graph. In a simulation study, covering a wide range of data-generating regimes, we found that often a large majority of signs are esti-

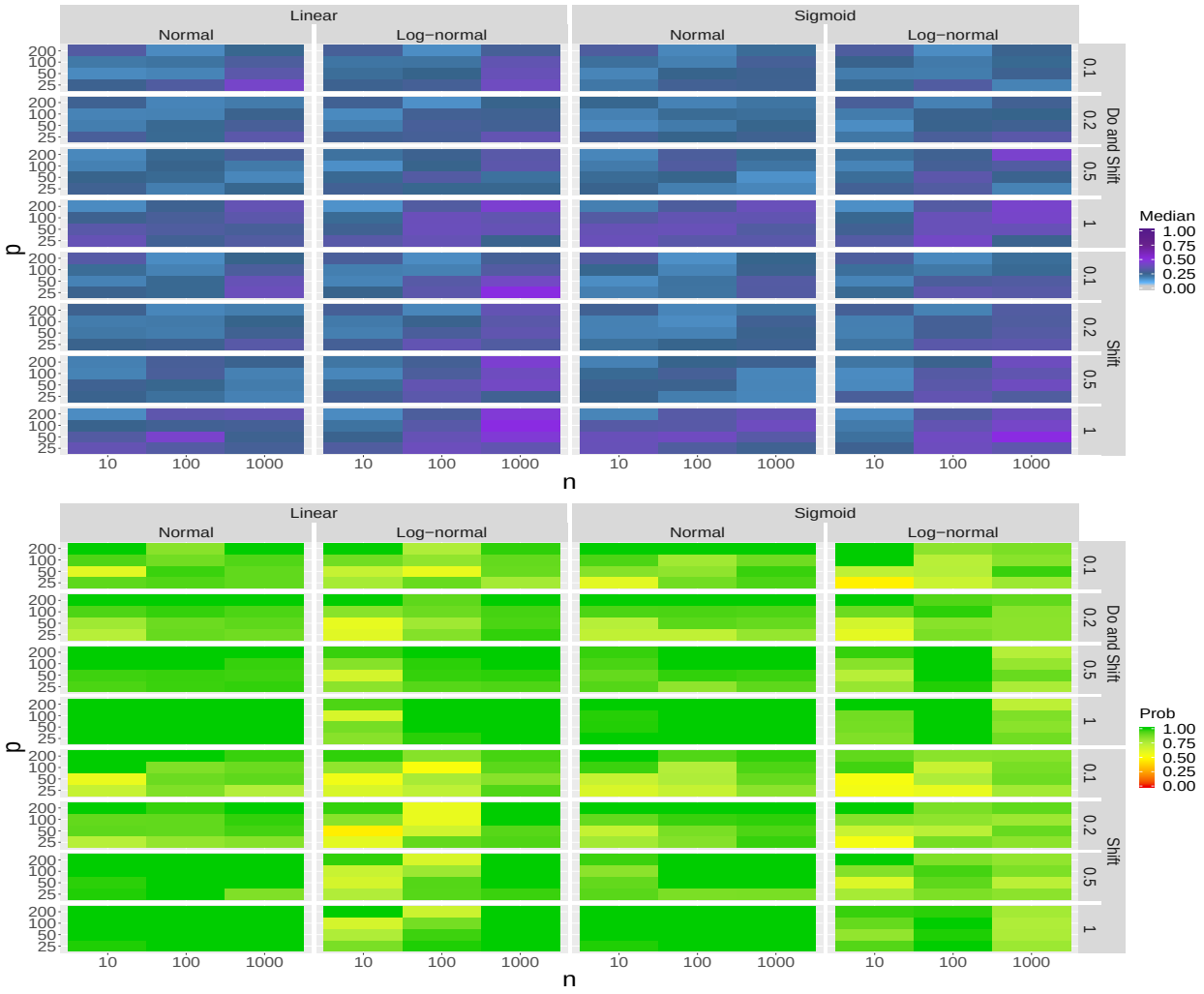

Figure 1: For both the top and bottom panel: Each cell corresponds to a simulation setting, characterized by p (left vertical axis), n (bottom horizontal axis), the link functions and error distribution of them SEM (top horizontal axis) and the type and probability of an intervention (right vertical axis). Upper panel: median difference of true risk between methods *GES* and *ACor* (see text) for different settings. A small value (blue) represents a more difficult situation to evaluate (since the true risks are then similar). Lower panel: empirical probabilities for how often the corresponding difference in estimated risk agrees in sign with the difference of true risk for different settings. A large value (green) means that the risk estimator performed well in this sense. In each cell we consider only settings for which the true risks differ by at least 0.1. A cell is left gray if less than 3 settings are available.

mated correctly. Importantly, these scenarios are not limited to settings with many interventions, large oracle differences, or particular algorithms. This suggests that our approach, or extensions of it, have the potential to allow truly practical assessment of causal structure learning. As the field of causal structure learning continues to advance, we think questions around problem-setting-specific empirical assessment will become ever more important in real-world applications.

Further work will be needed to weaken the assumptions regarding the statistical tests and to make this framework as general as possible. Moreover, different uses of this framework, for instance to tune the parameters of causal algorithms like PC and GES, can expand its scope and lead to additional interesting results and applications.

Acknowledgements

MFE was supported by the Swiss SNF grant 200021_172603. SM is a member of the German Bundesministerium für Bildung und Forschung (BMBF) consortium “MechML”.

References

Acid, S., & de Campos, L. M. (2003). Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. *J. Artif. Intell. Res.*, *18*, 445-490.

Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Stat.*, *25*, 505–541.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). *Invariant risk minimization.* (arXiv:1907.02893)

Bühlmann, P., Peters, J., & Ernest, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann. Stat.*, *42*, 2526-2556.

Chickering, D. M. (2002a). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.*, *2*, 445–498.

Chickering, D. M. (2002b). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, *3*, 507-554.

Claassen, T., Mooij, J. M., & Heskes, T. (2013). Learning sparse causal models is not NP-hard. In A. Nicholson & P. Smyth (Eds.), *Proceedings of UAI 2013* (p. 172181). AUAI Press.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, *15*, 3921-3962.

Colombo, D., Maathuis, M. H., Kalisch, M., & Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.*, *40*, 294-321.

Eigenmann, M. F., Nandy, P., & Maathuis, M. H. (2017). Structure learning of linear Gaussian structural equation models with weak edges. In G. Elidan, K. Kersting, & A. T. Ihler (Eds.), *Proceedings of UAI 2017.* AUAI Press.

Frot, B., Nandy, P., & Maathuis, M. H. (2019). Robust causal structure learning with some hidden variables. *J. Roy. Stat. Soc. B*, *81*, 459487.

Harris, N., & Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, *14*, 3365-3383.

Hauser, A., & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, *13*, 2409–2464.

Heinze-Deml, C., Maathuis, M. H., & Meinshausen, N. (2018). Causal Structure Learning. *Annu. Rev. Stat. Appl.*, *5*, 371-391.

Hill, S. M., Heiser, L., Cokelaer, T., & et al. (2016). Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nat. Methods*, *13*, 310–318.

Hill, S. M., Oates, C. J., Blythe, D. A., & Mukherjee, S. (2019). Causal learning via manifold regularization. *J. Mach. Learn. Res.*, *20*, 1-32.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.*, *47*, 11: 1–26.

Maathuis, M. H., Kalisch, M., & Bhlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Stat.*, *37*, 3133–3164.

Mooij, J. M., & Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. In A. Nicholson & P. Smyth (Eds.), *Proceedings of UAI 2013* (pp. 431–439). AUAI Press.

Nandy, P., Hauser, A., & Maathuis, M. H. (2018). High-dimensional consistency in score-based and hybrid structure learning. *Ann. Stat.*, *46*, 3151-3183.

Peters, J., & Bühlmann, P. (2015). Structural intervention distance for evaluating causal graphs. *Neural Comput.*, *27*, 771-799.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals. *J. Roy. Stat. Soc. B*, *78*, 947–1012.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*, 523–529.

Shimizu, S., Hoyer, P., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, *7*, 2003–2030.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (Second ed.). MIT Press, Cambridge. (With additional material by D. Heckerman, C. Meek, G.F. Cooper and T. Richardson)

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.*, *65*, 31–78.