Granger Significance Scoring

A Granger Causality-Based Scoring Function for the Time Series Causal Discovery Task

Author: Thierry Orth (6176178)

1st Supervisor: Dr. M. van Ommen
2nd Supervisor: Dr. A.J. Feelders

A thesis presented for the degree of
Artificial Intelligence MSc

Artificial Intelligence
Department of Information and Computing Sciences
Utrecht University

June, 2023

Abstract

A usual assumption in score-based methods for deriving graphical models from data is score equivalence. In brief terms, score equivalence requires that scores are the same between Markov equivalent graphs, that is, structures that encode the same conditional independencies. In the causal discovery task, however, this assumption is inaccurate: since Markov equivalent graphs can differ on arcs sets, their causal meaning is different. In this thesis, we propose what we call the Granger scoring function (GSF): a scoring function designed to learn causal summary graphs from time series data. More specifically, this function combines significance values from lag-specific Granger causality tests on time series data to infer a score on candidate graphs. In line with standard constraint-based methods, this score is interpreted as the amount of evidence in favor of graphs and is applied, subsequently, to select a graph or set of graphs that maximises the amount of evidence. In order to evaluate if the GSF reliably retrieves causal structure, we perform two experiments: first, an experiment to evaluate how well the GSF recovers the true graph from equivalent graphs and, second, an experiment to assess how well the GSF's performance fares in settings where the true graph is not necessarily accessible. From the experiments, it is concluded that the GSF significantly improves on random chance as well as on PCMCI$^+$, a similar structure discovery method. Furthermore, it is shown that the GSF's performance is maximised in the case of non-linear dependencies and higher proportion of lagged causes.

Keywords— *time series causal discovery, Granger causality, score non-equivalence, scoring functions, score-based methods*

Contents

1 Introduction

Causal discovery is the problem of learning causal relations from non-experimental, observational data. In the automated causal discovery task, this problem is interpreted as the task of learning a graphical model describing causal relationships between observed variables [21, 57, 30]. In the literature, two prominent procedures for extracting causal structure from data are constraint-based and score-based methods [30, 3, 57]. Constraint-based methods first exploit conditional independence constraints to derive an undirected graph and afterwards apply orientation rules to return an equivalence class of graphs [3, p. 780]. In constrast, score-based methods apply a scoring function on a set of candidate graphs and return the graphs that optimise the selected scoring function [3, p. 794].

A standard assumption on scoring functions is *score equivalence*: graphs that encode the same set of conditional independencies are assigned the same score [38, 9, 47]. In the causal setting, however, this assumption is inaccurate: graphs within an equivalence class can differ on arcs sets, effectively encoding different causal structures [39, p. 372]. In this thesis, we propose what we call the Granger scoring function (GSF): a causal scoring function that rejects score equivalence. In particular, the GSF combines p-values from lag-specific Granger causality tests on time series data to retrieve the strength of the evidence in favor candidate graphs, which is subsequently used to select an optimal candidate from the candidate space. Stated in explicit terms, the research question of this thesis is as follows:

Can the GSF reliably determine causal structure from otherwise Markov equivalent graphs? (**RQ**)

In evaluating the GSF, we perform two experiments. In the first experiment, we assume access to the true MEC to evaluate how well the scoring function recovers the true graph from equivalent graphs. The second experiment assumes access to an estimate of the MEC and aims to assess the scoring function's performance under a more realistic setting. In both experiments, a wide range of randomised causal model parametrisations are evaluated to ensure representative performance. In addition, performance scores are compared with those of Runge et al. [70] to benchmark performance. Since the subquestions of **(RQ)** depend on further details of a time series causal models, we postpone them until §8.2.

At this point, it is relevant to mention a number of invoked assumptions to make the scope of this work explicit. A first assumption is that causal structure comes in the form of a *directed acyclic graph* (DAG): cycles, edges and bidirectional arcs are excluded. Secondly, we assume the *causal Markov condition* and *faithfulness* to ensure identifiability of the MEC from the joint distribution.

In addition, the availability of a MEC further necessitates restriction to an acyclic segment of a subclass of summary graphs, as further outlined in §7.3.3. A fourth assumption is *causally sufficiency*, which states that all causal variables are included, allowing us to use the DAG representation. Last but not least, *causal stationarity* is assumed, which assumes that whenever a causal relation occurs at a given time point, that causal relation occurs at all time points. This assumptions allows the use of correlation tests to decide conditional independence. In future work, a subset of these assumptions may be weakened.

In broad lines, the thesis is structured as follows. First, §2-§5 discuss the preliminaries required for motivating and outlining the GSF in §6 and §7, respectively. In turn, §8 and §9 are dedicated to the evaluation of the GSF. In conclusion, §10 discusses the results, the limitations of this work as well as possibilities for future work. Appendix A includes a notation table for reference; Appendix B sets out the procedures relevant to our method; additional proofs are stated in Appendices D-E; results are included in Appendices F-G.

2 Principles of Causal Discovery

2.1 An Overview of Causal Discovery

Causal discovery is, as Pearl describes it, "an induction game that scientists play against Nature": scientists conduct experiments, collect data and apply inductive inference to infer the causal structure underlying the data-generating process [61, p. 43]. In some cases, collecting experimental data can be unethical, expensive or simply technically impossible [68, 57, 41, 58]. Cast in general terms, *automated causal discovery* is the inference task of automatically detecting causal structure from observational data, without performing experiments. Causal structure, in turn, is encoded as a causally interpreted *graphical model* consisting of vertices that represent variables in the data and connections representing causal relationships between those variables [21, 57, 30, 34]. Next to providing a graphical depiction of causal relations, functional mechanisms can be estimated using the causal graph, which allow for the identification and estimation of causal effects without the need for performing experiments [61, 90].

Unfortunately, causal discovery is a highly non-trivial inference task and is, in fact, known to be NP-hard in the general case. A first problem concerns a combinatorial explosion in the search space as the number of variables increases [48]. A second issue is the fact that the observational distribution does not disclose causal relationships. On the one hand, correlations in the data are prone to be *spurious* due to, for instance, latent common causes [3, 20]. On the other hand, correlation cannot justify causation as correlations are symmetric relations whilst

causation is an asymmetric relation [80, 20]. In order to make causal structure identifiable from the observational distribution, further assumptions are required, as further discussed in Section 2.6.

2.2 Relevance for Artificial Intelligence

In this section, we briefly discuss the relevance of automated causal discovery to the field of artificial intelligence. A first point relates to a central aspiration of artificial intelligence: the automation of tasks believed to be specific to human intelligence at human-level performance [72]. The discovery of causal relations falls under this scope: identifying causal relations is, as Pearl for instance puts it, a "hallmark of human cognition" [63]. A second point of relevance concerns the intuitive and interpretable nature of causal graphs as qualitative descriptions of systems. Performing inference tasks with the aid of causal graphs, it may be argued, is preferable to performing inference in more opaque models [90, p. 101].

Causal discovery is, furthermore, relevant for addressing a multitude of problems within the paradigm of *predictive artificial intelligence*. A first notable problem relates to the desideratum of robustness to dataset shift. Since learning from spurious correlations reduces robustness, using causal structure as constraints on the learning process of machine learning algorithms can reduce overfitting and increase generalisation of models. A second problem concerns the lack of interpretability. Integrating causal structure into the learning process makes it clearer which aspects the learning algorithm used in constructing the model [18, 48, 7, 92, 73, 56]. An alternative approach is to learn causal models from input-output pairs of an opaque model, which can subsequently be used to generate *contrastive explanations* or *counterfactual explanations* for the model's predictions [84, 54].

2.3 Causal Structure

Within the causal discovery task, the standard interpretation of causal structure is a graph over variables. From an intuitive point of view, this graph represents the causal dependencies of the generative process underlying the observational data. A graph $G = (V, E)$ is a tuple consisting of vertices V and a set of edges $E \subseteq V \times V$. In what follows, an edge $V_i - V_j$ is called undirected if $(V_i, V_j) \in E$ and $(V_j, V_i) \in E$. A directed edge $V_i \to V_j$ is called an *arc* and corresponds to $(V_i, V_j) \in E$ and $(V_j, V_i) \notin E$. Given these notions, an *undirected graph* $G = (V, E)$ is a graph consisting exclusively of undirected edges; a *directed graph* $G = (V, A)$ consists exclusively of arcs. In addition, the *adjacencies* of a vertex V_i in a graph $G = (V, E)$ is defined as the set of vertices $adj(V_i) = \{V_j : (V_j, V_i) \in E \text{ or } (V_i, V_j) \in E\}$ connected with V_i. In turn, a path $\pi = V_1 ... V_n$ is a sequence of distinct vertices

of length $2 \leq n$ such that each vertex V_i is adjacent to V_{i+1} for $1 \leq i < n$. A directed path, in turn, is a sequence of arcs pointing in the same direction: $\pi = V_1...V_n$ such that $V_i \rightarrow V_{i+1}$ for $1 \leq i < n$ [87, p. 2]. A causal structure is simply a directed graph over a set of variables representing functional relationships between variables:

Definition 2.3.1 (Causal Structure) *A causal structure over a set of variables* $\mathbf{X} = \{X_1, ..., X_n\}$ *is a directed graph* $G = (V, A)$ *such that* V *represents* $\mathbf{X}$ *and each arc* $(V_i, V_j) \in A$ *represents that* V_j *is a direct function of* V_i*.*

Within a causal structure $G = (V, A)$, a variable V_i is called a *direct cause* of a variable V_j if the arc $V_i \rightarrow V_j$ is included in G. In similar fashion, V_i is an *indirect cause* of V_j just in case there exists a directed path $V_i...V_j$ in G [61, p. 44].

In this work, we assume *directed acyclic graphs* (DAGs) as the model class for causal structure. In formal terms, this assumption corresponds to the acyclicity condition on causal structure:

Definition 2.3.2 (Acyclicity) *A directed graph* $G = (V, A)$ *is called acyclic if there exists no* $V_i \in V$ *for which there exists a directed path* $\pi = V_i...V_i$*.*

The acyclicity assumption is unwarranted in cases of *feedback loops*: self-directed causal influence of variables. Standardly, feedback loops are modelled in either of two ways: (i) with a cyclic graphical model or (ii) as an acyclic causal structure over time steps. As further explained in §4.4 and §7.3, our attention is restricted to a subgraph of a subclass of causal summary graphs guaranteed to be acyclic. If one were to adapt the search space to include cyclic models, the acyclicity assumption may be relaxed [21, p. 84].

2.4 Constraint-Based and Score-Based Methods

In the literature, the standard classes of causal discovery methods are *constraint-based* and *score-based methods* [30, 3, 57]. Constraint-based methods exploit constraints to derive causal structure in two steps, usually called the *skeleton phase* and the *orientation phase*. In the skeleton phase, conditional independence constraints are applied to derive an undirected graph over variables [50, p. 444]. In the orientation phase, a set of orientation rules is applied to transform edges into arcs as far as warranted [57, p. 7]. Since it is generally not possible to orient all edges, identification of a unique causal graph is not guaranteed. Instead, constraint-based methods generally return a partial causal structure encoding an equivalence class of graphs [30, 21, 80]. Score-based methods, in turn, consist of a scoring function and a search space of graphs interpreted as *Bayesian Networks* (BNs).

The first component of score-based methods is the *scoring function*: a function that measures fit on the data as well as complexity of candidate networks. The second component is the search procedure: an algorithmic method that efficiently traverses the selected search space. Given these two components, the aim of score-based methods is to extract from the search space a network or set of networks that optimise the scoring function [3, p. 794].

At this moment, it is relevant to describe relative advantages and disadvantages of both method classes. In general, constraint-based methods tend to be more efficient than score-based methods as long as the number of conditional independence tests is restricted. Contrastingly, score-based methods are less prone to error propagation, due to localised computation of scores discussed in §7.2.1. A further notable advantage of score-based methods is their ability to impose a total order on the set of candidate graphs, allowing for more fine-grained assessment and comparison of candidates [57, p. 7]. A last advantage is that, in the general case, score-based methods can determine can determine causal direction in the bivariate case whilst constraint-based methods cannot do so as conditional independence testing requires variable triples [30, p. 5].

At the same time, score-based methods are computationally expensive if the full set of candidate graphs is large. Moreover, the computational tasks of finding or approximating a globally optimal network are each known to be NP-hard. In order to resolve the computational problem, two tactics are commonplace: (i) a restriction of the space of candidate networks to a small and suitable subclass or (ii) an efficient search procedure that restricts the candidates to be evaluated. Since the latter method typically amounts to greedy heuristic search, a problem of local optima is involved. On the other hand, it should be clear that restriction to "a suitable subclass" can similarly involve local optima in cases where the globally optimal graph is excluded due to an inappropriate choice of the subclass [3, 57]. Next to constraint-based and score-based methods, *hybrid methods* effectively combine elements from both methods with the aim of countering their relative disadvantages and delivering a more robust and more reliable causal discovery procedure [50, p. 444].

2.5 A Graphical Criterion for Conditional Independence

In order to discuss the standard identifiability assumptions in causal discovery, we should first review the graphical notion of d-separation. Pearl [62] proposed d-separation as an efficient method for deriving the set of conditional independencies that a DAG imposes on the observational distribution. Formally, d-separation is defined as follows:

Definition 2.5.1 (d-separation on paths) *Let $G = (V, A)$ be a DAG and consider $V_i \in V$, $V_j \in V$ and $Z \subseteq V$. Then, a path $\pi = V_i ... V_j$ is d-separated by Z iff either of the following conditions holds:*

(i) π contains a chain $V_i \to X \to V_j$ or fork $V_i \leftarrow X \to V_j$ such that $X \in Z$;

(ii) π contains a collider $V_i \to X \leftarrow V_j$ such that $\sigma^*(X) \cap Z = \varnothing$.

Otherwise, π is called d-connected.

Definition 2.5.2 (d-separation on sets) *A set X is d-separated from a set Y by a set Z iff for all variables $V_i \in X$ and $V_j \in Y$ and paths $\pi = V_i ... V_j$, π is d-separated by Z. Otherwise, X and Y are d-connected.*

It may be helpful to spell out the intuitions behind d-separation in terms of conditioning on variables. Condition (i) states, firstly, that information about an intermediary cause between indirect causes removes dependence and, secondly, that information on a common cause removes dependence between direct effects. Condition (ii), on the other hand, states that if two variables share a common effect, then information about that effect or one of its effects can show dependence between the initial two variables [61, pp. 16–17].

2.6 Identifiability Assumptions for Causal Discovery

A causal structure is called *identifiable* just in case it can be uniquely determined from the observational distribution [66, p. 44]. In the general case, however, the observational distribution does not disclose causal structure. Under a number of assumptions, however, causal structure becomes identifiable. Although different sets of assumptions suffice for identifiability, we adhere to the following assumptions: the *causal Markov condition*, *faithfulness* and *causal sufficiency*. Conceptually, the first two assumptions jointly entail that d-separations in the causal structure correspond to conditional independencies in the distribution. On the other hand, causal sufficiency asserts that all causal variables are modelled in the graph [71, 3, 39, 57, 80, 40, 21].

Definition 2.6.1 (Causal Markov Condition) *A causal DAG $G = (V, A)$ and distribution* Pr *satisfy the causal Markov condition if for any $X, Y, Z \subseteq V$, if X and Y are d-separated by Z in G, then X and Y are independent given Z in* Pr.

Definition 2.6.1 imposes that d-separations are present in the graph only if there are corresponding conditional independencies in the observational distribution. Contrapositively stated, conditional dependencies are present in the distribution only if there are no corresponding d-separation statements.

Definition 2.6.2 (Faithfulness) *A causal DAG $G = (V, A)$ and distribution* Pr *satisfy the faithfulness condition if for any $X, Y, Z \subseteq V$, if X and Y are independent given Z in* Pr*, then X and Y are d-separated by Z in G.*

Definition 2.6.2 states the reverse of Definition 2.6.1: conditional independencies in the distribution are present only if there are corresponding d-separations in the graph or, contrapositively, the absence of d-separations implies corresponding conditional dependencies in the graph.

On accepting both the causal Markov and faithfulness conditions, it is clear that a one-to-one correspondence follows between d-separation properties in the graph and conditional independencies in the observational distribution. Combined with acyclicity, these assumptions make the Markov equivalence class identifiable from the observational distribution [39, 40]. Note that this set of assumptions has been contested. The causal Markov condition is, for example, violated in indeterministic contexts where d-separated variables show conditional dependence due to randomly coordinated variation [36, 8]. Faithfulness, in turn, fails if d-connected variables are rendered conditionally independent due to cancelling effects on mutual causal influence [20, p. 7]. In this work, we assume these two conditions to facilitate identification of the Markov equivalence class from the observational distribution. If the search space is shifted, these assumptions can be weakened [66, p. 197].

A further assumption is causal sufficiency, which assumes that all causal variables are observed:

Definition 2.6.3 (Causal Sufficiency) *A set of variables V is called causally sufficient if every variable Z that is a common cause of variables $X \in V$ and $Y \in V$ is a member of V.*

A problem with the causal sufficiency assumption is that it can occur that not all causal variables are measured. If a variable is excluded from the model, spurious correlations can emerge between its effects. Given a latent variable U and a causal structure consisting of its effect variables X and Y, excluding U results in a dependence between X and Y that cannot be traced back to U within the model [3, 20, 68, 57]. A usual model choice is to represent latent variables in a *mixed acyclic graph* (MAG), which models the presence of a latent variable with a bi-directed arc between its effects [27, pp. 94–96]. Although causal sufficiency is a theoretically unsatisfactory assumption, we restrict our focus to DAGs and leave an expansion to MAGs for future work.

3 Causal Modelling

3.1 Causal Models

Within Pearl's causal modelling framework, causal relationships are modelled in terms of a *structural causal model* (SCM) over variables. In basic terms, a causal model consists of a qualitative and a quantitative part: a graph defining *causal relations* between variables and, in addition, a set of structural equations defining *causal influence* between variables. As Pearl puts it, the causal structure is "a blueprint for forming a 'causal model' – a precise specification of how each variable is influenced by its parents in the DAG". In colloquial terms, causal relations in the graph translate to causal influence in structural equations [61, pp. 44–45]. In formal terms, a SCM consists of a DAG and a set of functions that define the value of variables in terms of other variables joined with mutually independent noise terms:

Definition 3.1.1 (Structural Causal Model) *A structural causal model (SCM) is a tuple* $\mathcal{M} = (G, \Theta)$ *such that* G *is a causal structure and* $\Theta = (U, V, F)$ *a set of parameters compatible with* G *consisting of*

(i) U *is a set of exogenous variables distributed according to* $\Pr(u_i)$ *for each* $U_i \in U$ *with* $U_i \perp\!\!\!\perp U_j$ *for* $U_i \neq U_j$

(ii) V *is a set of endogenous variables*

(iii) F *is a set of structural equations* f_i *that defines* $v_i = f_i(pa_i, u_i)$ *for each* $V_i \in V$ *where* PA_i *are the parents of* V_i *in* G

Here, exogenous variables are the noise terms that represent influences external to the model. One the other hand, endogenous variables represent observed causes and effects that are internal to the model. Condition (iii) shows that every endogenous variable is defined in terms of some exogenous variable as well as a possibly empty subset $PA_i \subseteq V$. An important remark is that a complete valuation of the exogenous variables is sufficient to infer all values of the endogenous variables: if the value u_i of each exogenous variable $U_i \in U$ is observed, then recursive application of the structural equations F provides the value v_j of each endogenous variable $V_j \in V$ [60, 63].

3.2 Markov Equivalence Class

A central notion for causal discovery is that of a Markov equivalence class (MEC). Verma and Pearl [83] defined a MEC as a set of DAGs that (i) share the same

undirected graph or "skeleton" and (ii) encode the same conditional independence statements. Intuitively, Markov equivalence tracks if graphs encode the same dependency structure in the undirected graph as well as if the same conditional independencies are encoded in their directed graph. The first condition can be easily verified by considering whether graphs $G = (V, A)$ induces the same undirected graph, which can be retrieved by defining $G' = (V', E)$ where $V' = V$ and $E = \{(V_i, V_j) : (V_i, V_j) \in E \text{ or } (V_j, V_i) \in E\}$. On the other hand, the second condition is more involved: it demands an effective method to decide for arbitrary graphs which conditional independence statements are imposed [61, p. 16]. As discussed in §2, Verma and Pearl proposed *d*-separation as an effective criterion for determining the conditional independence statements imposed by arbitrary graphs. As it turns out, two graphs share the same set of *d*-separation just in case their *v*-structures are the same: triples $V_i \rightarrow V_j \leftarrow V_k$ where V_i and V_k are non-adjacent, so-called *unshielded triples*. Hence, Markov equivalence between graphs is fully definable in terms of sameness of skeleton and *v*-structures:

Theorem 3.1 (Markov Equivalence) *Two DAGs G, G' are called Markov equivalent iff (i) their skeletons are the same and (ii) their set of v-structures is the same.*

A MEC of a graph is, in turn, simply the set of graphs containing that graph and closed under the Markov equivalence relation. In formal terms, the MEC $\mathcal{G}_G$ of G is the set $\mathcal{G}_G = [G]_\sim = \{G' \in \mathscr{G} : G' \sim G\}$ where $\mathscr{G}$ is the entire space of directed graphs and $\sim$ is the Markov equivalence relation. A more convenient representation of $\mathcal{G}_G$ is a completed partially directed graph (CPDAG). A CPDAG $C_\mathcal{G}$ representing a MEC $\mathcal{G}$ is, simply, the union graph over all members of $\mathcal{G} = \{G^1 = (V^1, A^1), ..., G^m = (V^m, A^m)\}$ defined as $C_\mathcal{G} = (V, E)$ with vertices $V = \bigcup_{i=1}^m V^i = V^i$ and the union of all arcs $E = \bigcup_{i=1}^m A^i$ of the members in $\mathcal{G}$. Hence, an arc is included in $C_\mathcal{G}$ just in case that arc is included in all members of $\mathcal{G}$; an edge is added whenever members disagree about arc direction [61, pp. 18–19].

As noted before, assuming the causal Markov condition and faithfulness ensures that the MEC over a set of variables is identifiable from the joint distribution over that set of variables. Formally, this involves that the distribution Pr suffices for assessing if $G \in \mathcal{G}$ for an arbitrary graph G and arbitrary MEC $\mathcal{G}$ induced by the set of conditional independencies from Pr. Briefly, this is established by the following two claims: (i) $G \in \mathcal{G}$ if G and Pr satisfy the Markov and faithfulness condition and (ii) $G \notin \mathcal{G}$ if there exists no parameter set $\Theta = (U, V, F)$ over G that induces Pr. The reader is referred to Peters, Janzing, and Schölkopf [66] for further details as well as the formal proof [66, pp. 44, 135–136].

3.3 The Bayesian Network Interpretation of Causal Models

As intended in Definition 3.1.1, *compatibility* of a causal structure $G = (V, A)$ and set of parameters $\Theta = (U, V, F)$ demands that the causal structure G and the causal influence defined in Θ do not conflict: a causal arc (V_i, V_j) is included in the graph just in case V_i occurs as a causal variable in the structural equation f_j of V_j. Definition 2.3.1 implicitly defines this requirement by stating that in a causal structure $G = (V, A)$, each $(V_i, V_j) \in A$ represents a direct functional relation between V_i and V_j. This functional relation is, in turn, given by f_j from the set of structural equations F in the parameter set Θ.

An alternative formulation of the compatibility requirement relevant to score-based methods is the Bayesian network (BN) interpretation of causal graphs. Within this formulation, the important part is that the graph decomposition corresponds to independence properties in the distribution:

Definition 3.3.1 (Compatibility) *Given a DAG $G = (V, A)$ and sets of variables X, Y, Z, if* Pr *is a distribution compatible with G, then d-separation of X and Y given Z implies $X \perp\!\!\!\perp Y|Z$ with respect to* Pr.

Clearly, this is the causal Markov condition defined before. In brief terms, the underlying thought is that G's decomposition allows estimation of a BN $\mathcal{M} = (G, \Theta')$ with joint probability distribution $\Pr^{\Theta'}(v_1, ..., v_n) = \prod_{i=1}^{n} \Pr^{\Theta'}(v_i|pa_i)$ that approximates $\Pr(v_1, ..., v_n)$. In intuitive terms, the graph is, in principle, capable of generating data from Pr given the appropriate set of parameters. Although the BN representation preserves the structure of the graph, it clearly does not preserve causal information: causal dependencies between X_i and $\{PA_i, U_i\}$ are maintained, but the functional dependency defined by f_i is ignored [61, pp. 44–45].

4 Time Series Causal Discovery

4.1 Time Series Analysis

Time series analysis is a field of data analysis that analyses variables with a time sequential ordering. In this way, it differs from usual cross-sectional, which ignores temporal distinctions and treats variables as if belonging to the same temporal state. The aims of time series analysis are as follows:

> [T]o understand or model the stochastic mechanism that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series and, possibly, other related series or factors [17, p. 1].

Major disciplines concerns with time series analysis are climate science, economics, epidemiology, neuroscience and physics [10, 76].

Standard causal discovery estimates causal structure from cross-sectional data. In contrast, time series causal discovery estimates causal structure from time series data [3]. This branch of causal discovery has enjoyed increasing interest due to rapid growth of available time series data [55]. In addition, time series causal discovery seems a more suitable approach than standard causal discovery in the case of variables with an inherent time-ordering, such as carbon dioxide emissions, energy use and economic production [90, p. 124]. In such cases, time-ordering can provide valuable information and, moreover, can be important to reflect in the inferred causal structure.

4.2 Time Series Notation

In the standard cross-sectional setting, data of a population is defined as a set of random variables $\mathbf{X} = \{X_1, ..., X_d\}$ for some fixed dimensionality $d \in \mathbb{N}$. Contrastingly, time series data is defined by a set of time series $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$ where each $\mathbf{X}_i$ encodes a set of time-ordered states of a random variable throughout time. A time series, then, is interpreted as a collection of time-indexed instances of some random variable: $\mathbf{X}_i = \{X_i^1, ..., X_i^T\}$ where $T \in \mathbb{N}$ is the final time point. Hence, X_i^t represents the state of X_i at time t.

In what follows, it is assumed that each time-indexed random variable X_i^t has multiple possible value realisations. The *realisation* of a time series $\mathbf{X}_i$, in turn, is a complete valuation $x_i^1, ..., x_i^T$ for all time-indexed variables X_i^t for $1 \leq t \leq T$. In practical settings, it is usual that we are given a single rather than multiple value realisations [68, p. 3]. For ease of notation, $\mathbf{X}_i^{t-k:t}$ represents the states of the time series $\mathbf{X}_i$ from time $t-k$ up and until time t defined by the set $\{X_i^s \in \mathbf{X}_i : k \leq s \leq t\}$. Similarly, $\mathbf{X}_i^{:t}$ defines the states of X_i from the initial time up and until time t given by $\{X_i^s \in \mathbf{X}_i : s \leq t\}$. With a slight abuse of notation, $\mathbf{T}^{t-k:t}$ selects the set $\bigcup_{i=1}^d \mathbf{X}_i^{t-k:t}$ consisting of all time-indexed variables up and until time t from the time series elements $\mathbf{X}_i$ and $\mathbf{T}^{:t}$ is defined analogously.

4.3 Causal Models for Time Series

Since time series data differs from cross-sectional data in relevant respects, we should be explicit and adapt the SCM representation in the case of time series data. Instead of defining a SCM directly on time series, the key is to, instead, use the full set of time-indexed variable defined as the union $\bigcup_{i=1}^d \mathbf{X}_i$ of time series included in $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$ [66, p. 199]. In explicit terms:

Definition 4.3.1 (Dynamic Structural Causal Model) *A dynamic structural causal model (DSCM) is a tuple* $\mathcal{M} = (G, \Theta)$ *where* G *is a causal structure and* $\Theta = (U, V, F)$ *is a set of parameters compatible with* G *such that*

(i) U *is a set of exogenous variables distributed according to* $\Pr(u_i)$ *for each* $U_i \in U$ *with* $U_i \perp\!\!\!\perp U_j$ *for* $U_i \neq U_j$

(ii) $V = \bigcup_{i=1}^{d} \mathbf{X}_i$ *is a set of time-indexed endogenous variables from a time series set* $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$

(iii) F *is a set of structural equations* f_i^t *for* $1 \leq i \leq d$, $1 \leq t \leq T$ *that defines* $v_i^t = f_i^t(pa_i^{t-\tau}, ..., pa_i^{t-1}, u_i^t)$ *for each* $V_i^t \in V$ *where* $PA_i^{t-\tau}, ..., PA_i^t$ *are the parents of* V_i^t *in* G *starting from a fixed time lag* $0 \leq \tau$

Although clauses (ii) and (iii) in Definition 4.3.1 differ from those stipulated in Definition 3.1.1, these specifications can be verified to formally satisfy clauses (ii) and (iii) in Definition 3.1.1 by treating each time-indexed variable as a random variable.

4.4 Causal Structures for Time Series

In the setting of time series data, three standard graphs for encoding causal structure are the full time causal graph, the window causal graph and the summary causal graph. Although we adopt the summary causal graph as model class, a short discussion of the full time causal graph and the window causal graph is useful for situating our choice. For reference, examples for all graphs are shown in Figures 1-3. First of all, the full time causal graph defines causal relations over the full set of time-indexed variables:

Definition 4.4.1 (Full Causal Time Graph) *A full time causal graph on a parameter set* $\Theta = (U, V, F)$ *over a time series* $\mathbf{T}$ *is a directed graph* $G = (V, A)$ *such that* $V = \bigcup_{i=1}^{d} \mathbf{X}_i$ *and* $(V_i^{t-\tau}, V_j^t) \in A$ *iff* $V_i^{t-\tau}$ *occurs in the structural equation* f_j^t *of* V_j^t *with* $0 < \tau$ *for* $i = j$ *and* $0 \leq \tau$ *for* $i \neq j$.

Although the full time representation is a fine-grained representation, it involves three problems. First of all, a small increase in the number of time series or the number of time indices results in a growth in the graph that quickly results in an "unwieldy and difficult to interpret" representation [24, 2]. Secondly, inferring the full time causal graph from real-world data is usually impossible as there is typically only a single observation x_i^t available for each time-indexed variable X_i^t [3]. Thirdly, the model selection process becomes hampered due to combinatorial

explosion: the number of possible graphs on a set of vertices rapidly grows as the number of vertices increases [19, 22]. An alternative representation is the *window causal graph*, which restricts the set of vertices up to a maximal lag $\tau_{\max}$ that represents the largest time gap between causes and effects in the DSCM:

Definition 4.4.2 (Window Causal Graph) *A window causal graph on a parameter set $\Theta = (U, V, F)$ over a time series $\mathbf{T}$ is a directed graph $G = (V, A)$ such that $V = \bigcup_{i=1}^{d} \mathbf{X}_i^{t-\tau_{\max}:t}$ for time t and maximal lag $\tau_{\max}$ and $(V_i^{t-\tau}, V_j^t) \in A$ iff $V_i^{t-\tau}$ occurs in the structural equation f_j^t of V_j^t with $0 < \tau \leq \tau_{\max}$ for $i = j$ and $0 \leq \tau \leq \tau_{\max}$ for $i \neq j$.*

In essence, the window graph aims to contract the full time representation into a more manageable representation. Under the causal stationarity assumption discussed in §4.5, the window graph fully captures the causal structure of the full time graph [2, pp. 1–2]. In the worst case, however, the largest time gap $\tau_{\max}$ spans the entire length of the time series, which simply results in the full time representation. A *summary causal graph* resolves this problem by forcing a condensed representation. Instead of modelling the time-indexed variables, this representation effectively models the time series themselves with arcs corresponding to causal influence at some point in time:

Definition 4.4.3 (Summary Causal Graph) *A summary causal graph on a parameter set $\Theta = (U, V, F)$ over a time series $\mathbf{T}$ is a directed graph $G = (V, A)$ such that $V = \mathbf{T}$ and $(V_i, V_j) \in A$ iff there exists a time t and lag τ such that $V_i^{t-\tau}$ occurs in the structural equation f_j^t of V_j^t with $0 < \tau$ for $i = j$ and $0 \leq \tau$ for $i \neq j$.*

Note, first, that this condensed representation comes at the expense of temporal information: the existence of an arc entails nothing about the time at which causal influence occurred nor about the frequency of causal influence. As a second point, the window graph is ensured to be acyclic whenever the full time graph is acyclic, since the former is simply an induced subgraph of the latter. Contrastingly, summary graphs are not ensured to be acyclic [2, 66]. A first way in which cyclicity can occur is due to autogenerative dependence of variables throughout time $X^{t-\tau} \to X^t$, so-called "self-loops". A second way in which cycles can occur is in the form of a feedback loop between distinct time series or, more precisely, a directed cycle between distinct time series. A simple example is a time window causal graph with causes $X^{t-1} \to Y^t$ and $Y^{t-1} \to X^t$, which results in a summary causal graph with causes $\mathbf{X} \to \mathbf{Y}$ and $\mathbf{Y} \to \mathbf{X}$.

Since the assumed search space in this work is a MEC, we restrict attention to summary graphs in which all cycles are due to self-loops. This subclass of summary graphs, effectively, consist of two subgraphs: (i) a cyclic segment of self-loops and

(ii) an acyclic segment of causal relations between distinct variables. Since the second subgraph is a DAG, a MEC becomes definable. Further details about this model choice are included in §7.3.3. In future work, a less restrictive search space can be adopted; the point of this work is to evaluate if the scoring function reliably determines true causal structures from Markov equivalent structures.

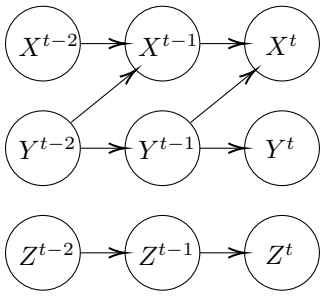


Figure 1: Full Time Causal Graph

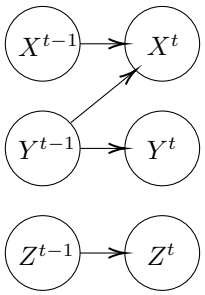


Figure 2: Window Causal Graph

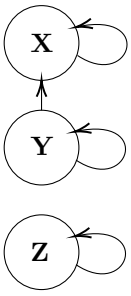


Figure 3: Summary Causal Graph

4.5 Further Assumptions for Time Series Causal Discovery

In addition to the identifiability assumptions outlined in Section §2.6, we here outline a number of modelling choices and assumptions relevant to the G-causality framework. An important modelling choice concerns the *dependency type* of variables in structural equations. This concerns, for example, linear versus non-linear transformations [68, p. 8]. An important assumption is the *causal stationarity assumption*: causal relations "remain constant in direction throughout time" or, identically, that causal mechanisms are invariant with respect to changes in time. In addition, assuming causal stationarity avoids requiring multiple realisations of time series and allows for regression-based conditional independence tests, as outlined in §5.4. Within the context of DSCMs, causal stationarity involves that structural equations f_j^t are the same for each time t [68, 71, 2, 3, 11].

In what follows, we will not adopt the assumption of *non-instantaneous effects*: that causal relations occur only across time and, consequently, not within the span of a single time point. Although this violates Granger's time precedence principle, the non-instantaneous effects assumption has a practical limitation: data recordings may be too coarse-grained to ensure that causal influence only occurs across time points. A first reason is *subsampling*: if the sampling rate is slower than the causal process generating the data, then causal influence can occur within the span of a single time point. Secondly, the practice of *temporal aggregation* is an entrenched size reduction method. This, however, is prone to merge causal and effect variables, even if the non-aggregated data succeeded in a full separation of causes and effects. If one wants to ensure time precedence, then the granularity of the time series data has to fit the causal domain of interest, i.e., the data recordings must ensure that relevant causal influence only occurs across and not within timepoints [3, 30, 66, 68].

5 Granger Causality

5.1 The Concept of Granger Causality

In these sections, we discuss the concept and definitions of Granger causality as well as associated assumptions for the causal discovery task and further relevant details. Granger causality, G-causality for short, is a well-established method for detecting causal relations from time series data [2, 31, 67]. Simultaneously, it must be emphasised that G-causes do not, in general, represent true causal mechanisms. Rather, G-causes are understood in terms of *forecasting ability*: a variable G-causes another variable just in case it contains unique information that aids in prediction [31, p. 430]. Since this clearly does not track causal mechanisms, the existence of a G-cause is neither necessary nor sufficient for the existence of a true underlying

causal mechanism: G-causes are best understood as *potential causes* [67, 52, 61, 23, 85]. Still, G-causality is preferable over standard forecasting due to two principles: (i) *temporal precedence*, which demands that causes precede their effects in time and (ii) *uniqueness*, which requires that causal time series include information about caused time series unavailable in the absence of those causal series [33, 23].

At this point, a number of advantages and disadvantages of G-causality are worth noting. An important advantage is that G-causality is a model-free approach: it does not assume a causal model and can, thus, be directly applied to data [22, 23, 28]. A disadvantage of using such a model-free approach is that in the absence of sufficient background knowledge, the quality of G-causal conclusions becomes subject to statistical conditions such as appropriate sampling, non-instantaneous causation and stationarity [52, pp. 87, 98]. Another important advantage is the theoretical appeal of G-causality. First of all, the principle of uniqueness coincides with the notion that a cause's change in the effect are due to properties of that cause alone. Secondly, temporal precedence respects the notion that the arrow of causation is asymmetric, which has been met with increased application in recent years [3, 35, 61, 86]. As a caveat, a temporal ordering over variables remains insufficient for strict derivation of a true causal ordering: spurious correlations are known to occur across time [51, 86].

5.2 Formal Definitions of Granger Causality

Broadly viewed, two interpretations of G-causes are current in the literature. According to the lag-general interpretation that Granger [31] originally proposed, G-causal relations are defined on *time series*, encoding the presence of causal influence at an unspecified point in time. The lag-specific interpretation exemplified in work from Assaad, Devijver, and Gaussier [3] and Runge [68], on the other hand, defines G-causal relations on *time-indexed variables* instead. In this framework, the time at which G-causal influence occurs is specified. At the same time, specifying the time of causal interaction demands availability of multiple realisations of time series or assuming causal stationarity, as further discussed in §4.5 [68].

Under Granger's interpretation, G-causality is defined as conditional dependence of a causal and a caused time series conditional on the full domain of time series in the universe excluding the causal series. From a theoretical point of view, this is to ensure that the information uniquely derives from the purported G-cause. Formally:

Definition 5.2.1 (General Granger Causality) *Let $\mathbf{\Omega}$ be the set of all time series in the universe. If $\mathbf{X}$, $\mathbf{Y} \in \mathbf{\Omega}$, then $\mathbf{X}$ does not Granger cause $\mathbf{Y}$ if $Y^t \perp\!\!\!\perp \mathbf{X}^{:t-1} | \mathbf{\Omega}^{:t-1} \setminus \mathbf{X}^{:t-1}$ for all $t \in \mathbb{N}$. Otherwise, $\mathbf{X}$ is said to Granger cause $\mathbf{Y}$.*

In natural terms, $\mathbf{X} \rightarrow_G \mathbf{Y}$ just in case there exists a point in time where the value of $\mathbf{Y}$ is conditionally dependent on the past values of $\mathbf{X}$ given the past information of all other time series. A clear advantage of G-causality is that it tackles the problem of correlational symmetricity discussed in §2.1: $\mathbf{X} \rightarrow_G \mathbf{Y}$ and $\mathbf{Y} \rightarrow_G \mathbf{X}$ encode distinct dependencies. Now, it should be clear that Definition 5.2.1 is unrealistic: real-world data at most provides access to a minute subset of $\mathbf{T} \subset \mathbf{\Omega}$. Given this consideration, Granger [31] adapted G-causality as follows:

Definition 5.2.2 (Granger Causality) *Let* $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$ *be a finite time series. If* $\mathbf{X}, \mathbf{Y} \in \mathbf{T}$*, then* $\mathbf{X}$ *does not Granger cause* $\mathbf{Y}$ *if* $Y^t \perp\!\!\!\perp \mathbf{X}^{:t-1} | \mathbf{T}^{:t-1} \setminus \mathbf{X}^{:t-1}$ *for all* $t \in \mathbb{N}$*. Otherwise,* $\mathbf{X}$ *is said to Granger cause* $\mathbf{Y}$*.*

In the practical case, the domain of time indices will span a finite interval $\mathcal{T} = \{1, ..., T\}$ for some fixed $T \in \mathbb{N}$ [31, 33, 22, 23]. As previously mentioned, the lag-general interpretation involves a loss of information: a G-cause $\mathbf{X} \rightarrow_G \mathbf{Y}$ holds just in case the following conditional dependence statement holds for some $1 < t \leq T$:

$$Y^t \not\perp\!\!\!\perp \{X^1, ..., X^{t-1}\} | \mathbf{T}^{:t-1} \setminus \{X^1, ..., X^{t-1}\} \tag{1}$$

From $\mathbf{X} \rightarrow_G \mathbf{Y}$ alone, however, it is left implicit which subset of $\{X^1, ..., X^{t-1}\}$ renders conditional dependence. Since such fine-grained information is important when evaluating the time-relative impact, a sensible choice is to evaluate the specific time lags $0 \leq \tau$ at which $X^{t-\tau}$ generates dependence. Runge [68] defines lag-specific G-causes as follows: instead of evaluating conditional dependence of an entire history of $\mathbf{X}^{:t-1}$ with Y^t, independence is evaluated between a time-indexed instances $X^{t-\tau}$ and Y^t. In turn, lag-specific Granger causality correspond to the following conditional dependence:

$$Y^t \not\perp\!\!\!\perp X^{t-\tau} | \mathbf{T}^{:t-1} \setminus \{X^{t-\tau}\} \tag{2}$$

Since the conditional dependence imposed in (2) concerns the relation $X^{t-\tau} \rightarrow_G Y^t$, only $X^{t-\tau}$ is removed from the conditional set. In what follows, a G-causal link is called instantaneous if $\tau = 0$ and lagged for $0 < \tau$ [68, 71, 69, 66].

5.3 Vector Autoregression Models

A common mathematical model for capturing relationships between time series variables as well as for generating time series realisations is the Vector Autoregression (VAR) model [67, 68, 55, 57]. In addition, such models are used for establishing G-causes by evaluating if including past values of the purported G-cause results in a significant change in prediction of a variable of interest [26, 52, 67, 68, 55, 57, 65, 81, 31, 32, 1, 55]. In §8, VAR models are used to define the

structural equations of DSCMs. Before defining VAR models, it is convenient to first consider the functional form of linear structural equations within Pearl's framework. Given a set of random variables $\mathbf{X} = \{X_1, ..., X_d\}$ and a set of linear equations $\{f_1, ..., f_d\}$, the value of each X_i is defined as a weighted sum of the other variables plus additive noise:

$$X_i = f_i(\mathbf{X}) = \sum_{k=1, k\neq i}^{n} \beta_{i,k} X_k + \epsilon_i, \tag{3}$$

where each coefficient $\beta_{i,k}$ weights the influence of X_k on X_i and ϵ_i is an additive noise term on X_i, playing the role of an exogenous variable [39, 66].

A VAR model over a time series is, in turn, a linear model defined by a set of linear equations f_i^t that each define the value of X_i^t in terms of *past values* of itself and other variables. More formally, a VAR model over a causally stationary time series $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$ each of length T consists of d linear equations f_i for $1 \leq i \leq d$ that each define the value of variable X_i^t in terms of past values for variables included in $\mathbf{T}^{:t-1}$. A τ'th order VAR model, denoted, denoted VAR(τ), has a maximal lag τ that defines the time window of the model. Effectively, this restricts the past information to $\mathbf{T}^{t-\tau:t-1}$ [81, 49]. A single equation on X_i^t is then given as follows:

$$X_i^t = f_i(\mathbf{T}^{t-\tau:t-1}) = \sum_{k=1}^{d} \sum_{\gamma=1}^{\tau} \beta_{i,k}^{t-\gamma} X_k^{t-\gamma} + \epsilon_i, \tag{4}$$

where $\beta_{i,k}^{t-\gamma}$ weights the influence of $X_k^{t-\gamma}$ at time $t-\gamma$ and the noise term ϵ_i is defined in the same way for every time t. Since a time series can be ordered index-wise, we can switch to a more convenient time series vector representation $\mathbf{T} = (\mathbf{X}_1, ..., \mathbf{X}_d)$ and define the values of the full time series vector $\mathbf{T}$ at time t as follows:

$$\mathbf{T}^t = \sum_{\gamma=1}^{\tau} \Phi^\gamma \mathbf{T}^{t-\gamma} + \vec{\epsilon}, \tag{5}$$

where Φ^γ is a $n \times n$ coefficient matrix at time lag γ such that each $\Phi_{i,k}^\gamma$ weights the influence of $X_k^{t-\gamma}$ on X_i^t and $\vec{\epsilon}$ is a vector of noise terms $(\epsilon_1, ..., \epsilon_d)^\top$ coupled with each X_i for $1 \leq i \leq d$ [68, p. 2]. Although VAR coefficients and past values are combined in a linear way, one can generalise these models to a *non-linear additive noise model*:

$$\mathbf{T}^t = g\left(\sum_{\gamma=1}^{\tau} \Phi^\gamma \mathbf{T}^{t-\gamma}\right) + \vec{\epsilon}, \tag{6}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a non-linear transformation applied on the linear sum from before [41, 58]

5.4 Regression-Based Conditional Independence Tests

The formal exposition in §5.2 defined G-causes in terms of conditional independencies. Unfortunately, practical evaluation of conditional independencies is subject to a number of limitations. First of all, it is unlikely that exact conditional independence can be found under finite sampling conditions [66]. Secondly, conditional independence testing is subject to the *curse of dimensionality*: as the size of the conditional set Z grows, the amount of data for evaluating the null hypothesis quickly increases whilst the amount of available data becomes increasingly sparse [92]. In addition, the time series case involves the problem that standard conditional independence tests require multiple realisations of time series [68, p. 3]

Instead, the aim is to establish conditional independence by evaluating a null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ against an alternative hypothesis $H_1 : X \not\perp\!\!\!\perp Y|Z$ given observed samples $\{x_i, y_i, z_i\}_{i=1}^n$. In model-free approaches, conditional independence is evaluated directly without assuming functional dependencies between variables. Contrastingly, regression-based methods impose dependencies $X = f_X(Z) + \epsilon_X$ and $Y = f_Y(Z) + \epsilon_Y$. Zhang et al. [92] has shown that for identifiable additive noise models $Y = f(X) + \epsilon$, independence between residuals $\hat{r}_X = X - \hat{f}_X(Z)$ and $\hat{r}_Y = Y - \hat{f}_Y(Z)$ is a sufficient condition for $X \perp\!\!\!\perp Y|Z$ [92, pp. 1250–1251]. In evaluating $X \perp\!\!\!\perp Y|Z$, regression-based methods evaluate the independence of residuals from regressing X on Z and regressing Y on Z. Here, the dependencies $X = f_X(Z) + \epsilon_X$ and $Y = f_Y(Z) + \epsilon_Y$ assume that X and Y are centered and that ϵ_X and ϵ_Y are independent and identically distributed. In the first step, models $\hat{f}_X$ and $\hat{f}_Y$ are estimated using a sample $\{x_i, y_i, z_i\}_{i=1}^n$. In the next step, the residuals $\hat{r}_X = X - \hat{f}_X(Z)$ and $\hat{r}_Y = Y - \hat{f}_Y(Z)$ are computed. In the final step, independence between the residuals $\hat{r}_X$ and $\hat{r}_Y$ is evaluated with a statistical test. Partial correlation tests assume that f_X and f_Y are linear and, furthermore, evaluate independence of residuals with a regular t-test. In the case of non-parametric regression such as Gaussian Process regression, independence of residuals is evaluated with a non-parametric test such as a distance correlation test. Under the assumption of causal stationarity, $X^{t-\tau} \to_G Y^t$ is evaluated using the sample $\{x^{i-\tau}, y^i, t^{:i-1} \setminus x^{i-\tau}\}_{i=\tau+1}^T$, thus avoiding the need for multiple realisations [68, 92, 41, 94, 46].

6 Background

6.1 Related Work

In this section, we discuss work relevant to this thesis and embed our own contribution within the current literature. In §6.1.1-§6.1.3, we survey a number of relevant constraint-based, score-based and hybrid methods. In turn, §6.2 discusses the proposed scoring function, primarily building on work from Runge et al. [70]. In the same section, we further motivate the use of MECs as search spaces.

6.1.1 Constraint-Based Methods

A central constraint-based method in the literature is Spirtes, Glymour, and Scheines [79]'s PC algorithm. Under the assumptions of acyclicity, faithfulness and sufficiency, PC infers a CPDAG from the observational distribution in three consecutive steps: (i) derivation of the skeleton, (ii) determination of v-structures and (iii) orientation of remaining edges using a set of orientation rules. Step (i) effectively starts with the complete undirected graph $G = (V, E)$ and performs a level-wise search to remove edges. In algorithmic terms, PC starts with $k = 0$ and tests for each adjacent pair (V_i, V_j) and conditioning set $Z \subseteq \mathcal{A}(V_i) \setminus \{V_j\}$ of size k the conditional independence $V_i \perp\!\!\!\perp V_j | Z$ using a conditional independence tester $\mathcal{I} : \mathcal{X} \times \mathcal{X} \times \wp(\mathcal{X}) \to [0, 1]$ and significance level $\alpha_{\text{PC}} \in [0, 1]$. Given $H_0 : V_i \perp\!\!\!\perp V_j | Z$ as null hypothesis and $H_1 : V_i \not\perp\!\!\!\perp V_j | Z$ as alternative hypothesis, the edge $V_i - V_j$ is removed if $\mathcal{I}(V_i, V_j, Z) \not< \alpha_{\text{PC}}$ and Z is stored in the separation set of V_i and V_j; otherwise $V_i - V_j$ is kept in the graph. After each iteration, the update $k \leftarrow k + 1$ is performed until no conditioning set Z of size k can be found. Step (ii) generates v-structures $V_i \to V_j \leftarrow V_k$ whenever pairs (V_i, X) and (X, V_j) are adjacent (V_i, V_j) is not adjacent and X is not included in the separation set of (V_i, V_j). In step (iii), a set of orientation rules that exploit acyclicity and v-structure constraints are applied to orient remaining edges as far as warranted [78, 30, 39, 79].

Unfortunately, the PC algorithm gives poor performance on deriving time series graphs for the following reason: samples for performing regression-based conditional independence tests on lag-specific links $X_i^{t-\tau} \to X_i^t$ are often shared with those for $X_i^{t-\tau'} \to X_i^t$ for $\tau' \neq \tau$. Due to this interdependence of conditional independence tests, false positive and false negative rates increase [68, pp. 12–14]. Runge et al. developed the PCMCI algorithm, which assumes the causal Markov condition, faithfulness and causal sufficiency to infer a time series graph whilst counteracting false positives and false negatives via additional conditioning sets. PCMCI is equivalent to PC on steps (ii) and (iii), but step (i) is different. Instead of performing a level-wise search, PCMCI estimates a conditioning set $\hat{\mathcal{P}}(X_j^t)$ for every X_j^t via ranking of test statistic values. Here, the point is to mitigate the lack

of detection power that occurs when the conditioning set grows whilst provably maintaining the removal of incorrect links. In addition, an estimate $\hat{\mathcal{P}}(X_i^{t-\tau})$ of the parents of $X_i^{t-\tau}$ is added to control for false positives due to autocorrelation of the causal variable. Jointly, these constitute the *momentary conditional independence* (MCI) test for edge removal:

$$X_i^{t-\tau} \perp\!\!\!\perp X_j^t | \hat{\mathcal{P}}(X_j^t) \setminus \{X_i^{t-\tau}\}, \hat{\mathcal{P}}(X_i^{t-\tau}) \tag{7}$$

Next to the usual hyperparameter α_{PC}, PCMCI and further variants depend on $\tau_{\max}$. Effectively, $\tau_{\max}$ defines the maximal time delay for evaluating lag-specific conditional independence. Since a higher choice of $\tau_{\max}$ harms performance but not estimation quality, Runge et al. advises to choose a large value in the absence of further background knowledge about the relevant causal system [68, 70].

PCMCI was developed for time series graphs without instantaneous effects. Runge [69] developed an extension called the PCMCI$^+$ algorithm to account for lagged and instantaneous effects. First, PCMCI$^+$ estimates adjacency sets $\hat{\mathcal{B}}(X_i^{t-\tau})$ and $\hat{\mathcal{B}}(X_j^t)$ to account for autocorrelation effects. In the next step, the time series graph G is initialised with all instantaneous adjacencies plus the lagged adjacencies from $\hat{\mathcal{B}}(X_j^t)$. In turn, PCMCI$^+$ tests all adjacent pairs $(X_i^{t-\tau}, X_j^t)$ and iterates through instantaneous conditioning sets $Z \subseteq adj(X_j^t)$ and performs the following MCI test to decide edge removal:

$$X_i^{t-\tau} \perp\!\!\!\perp X_j^t | Z, \hat{\mathcal{B}}(X_j^t) \setminus \{X_i^{t-\tau}\}, \hat{\mathcal{B}}(X_i^{t-\tau}) \tag{8}$$

Note that the adjacency sets are lagged parent sets for $0 < \tau$ as before. In the case of $\tau = 0$, on the other hand, these sets consists of instantaneous adjacencies [69].

Assaad, Devijver, and Gaussier [2] developed a method for discovering an *extended causal summary graph*: a graph consisting of a *past slice* of time series $\mathbf{X}_i^{:t-1}$ as well as a *present slice* consisting of present variables X_i^t. Causal relations are defined between past and present slices as well as within the present slice. A past-to-present link in the graph corresponds to the existence of *some* causal links $X^{t-\tau} \to X_j^t$ for $0 < \tau$; a present-to-presence link encodes the existence of a causal link at $\tau = 0$. The extended summary graph, Assaad, Devijver, and Gaussier [2] argue, is preferable to regular summary graphs: these graph clearly distinguish past from present causation and uphold acyclicity. Using *greedy causation entropy* to decide conditional independence of past-to-present variables and mutual information of present-to-present variables, Assaad, Devijver, and Gaussier [2] infer a graph using the order-independent PC-stable algorithm as well as the FCI algorithm [2].[1]

[1] Note: "order-independence" involves that the algorithm's output does not depend on

6.1.2 Score-Based Methods

Meek [53] and Chickering [14] developed Greedy Equivalence Search (GES), a central score-based method in the literature. Starting from an empty graph and with the Bayesian Information Criterion as scoring function, GES consists of (i) an *arc insertion phase* and (ii) an *arc removal phase*. In step (i), arcs insertions are iteratively performed on the graph until a local maximum is reached. In step (ii), GES iteratively deletes arcs from the graph until a local maximum is reached. After reaching termination, the equivalence class is given as output [14, 53, 66].

Pamfil et al. [59] proposed a score-based method called DYNOTEARS for learning a window graph over time-indexed variables that represents both contemporaneous and lagged relationships. Effectively, the window graph is represented as a contemporaneous and a lagged adjacency matrix and learned by minimising a least squares loss objective subject to ℓ_1-penalisation as well as an acyclicity constraint [59]. A noted problem of method such as DYNOTEARS is their orientation towards finding parsimonious graphs that best explain the data, which is an unsuitable learning task from a causal point of view [2, 44].

6.1.3 Hybrid Methods

A relevant hybrid method for learning non-causal Bayesian Networks is the Max-Min Hill-Climbing (MMHC) algorithm, developed by Tsamardinos, Brown, and Aliferis [82]. Combining techniques from local learning, constraint-based and score-based methods, the MMHC algorithm consists of (i) a *skeleton phase* and (ii) a *greedy phase*. Phase (i) learns a skeleton graph using a local discovery algorithm called the Max-Min Parents and Children algorithm. In phase (ii), greedy hill-climbing search is applied to the empty graph with arcs additions, deletions and reversals subject to edge constraints that derive from the skeleton found in phase (i) [82, p. 33].

6.2 Main Contribution

6.2.1 Granger Scoring Function

The contribution of this work is what we call the Granger scoring function (GSF). From an intuitive point of view, the GSF constructs scores reflecting the amount of evidence in favor of candidate graphs. In the general lag-specific case, this scoring function establishes score non-equivalence in two consecutive steps: (i) extraction of a p-value p_{ij}^{τ} for every lag-specific link $X_i^{t-\tau} \to_G X_j^t$ and (ii) construction of a

the order in which the variables are given as input, which fails for the regular PC algorithm [50].

score p_G on a candidate graph G as a function of p_{ij}^{τ} for all present arcs $X_i^{t-\tau} \to_G X_j^t$ in G. Since our model class is the summary graph, the exposition in what follows replaces step (ii) by, first, combining p_{ij}^{τ} for all τ into a p-value p_{ij} for the lag-general arc $\mathbf{X}_i \to_G \mathbf{X}_j$ and, second, fusion all p_{ij} into a final score p_G.

While the extraction of p-values uses Runge [69] PCMCI$^+$'s MCI test, the consecutive process departs from PCMCI$^+$: instead of removing edges $X_i^{t-\tau} -_G X_j^t$ on the basis of a significance threshold α_{PC}, p-values are used as scores on $X_i^{t-\tau} \to_G X_j^t$ and, subsequently, combined using a p-value combining method. This, we argue, has three advantages over PCMCI$^+$. First of all, the resulting causal structure does not depend on a general hyperparameter α_{PC} whilst PCMCI$^+$ does. Secondly, the method avoids the *multiple testing problem* internal to PCMCI$^+$: in this case, the problem that the error rate of both acceptation and rejection of hypotheses grows as the number of evaluated edges grows [47]. Last of all, our method exploits the information from inferred p-values up and until the final inference step whilst PCMCI$^+$ does not, which incurs a potentially detrimental information loss in the inference process.

6.2.2 Comments on Search Space

Given that the focus in this work lies on the scoring function's capacity for distinguishing otherwise statistically indistinguishable Markov equivalent graphs, we assume the Markov equivalence class of the true graph as the search space. Given a CPDAG representation of a MEC, an efficient algorithm for generating all DAGs within the MEC is given in Wienöbst et al. [87]. A first problem of assuming the MEC of the true graph as search space is that the MEC is, in general, not given. Furthermore, the MEC is in the worst case superexponentially large: given a set of m vertices $V = \{V_1, ..., V_m\}$, the complete graph over $G = (V, A)$ over V has a MEC $\mathcal{G}_G$ is of size $m!$ [29, p. 172]. In the realistic setting, the MEC is unavailable and an estimate $\hat{\mathcal{G}}_G$ has to be estimated from data. An alternative approach would be to estimate the CPDAG, generate a DAG belonging to the represented MEC and perform a greedy search under equivalent constraints similar to GES. Since our focus here is on scoring functions, however, we leave such endeavours to future work. In any case, it should be emphasised that the search space is not essential to our main contribution, namely, the GSF.

7 Scoring Functions

7.1 Overview

In this chapter, we discuss the literature on scoring functions and introduce the Granger scoring function on the basis of considerations in the literature. In §7.2.1, the form of general scoring functions is discussed, together with two usual constraints on such functions: score decomposability and score equivalence. Given this overview, §7.2.2 outlines what is required in the special case of causal scoring functions. Given that Granger causality is bivariate, we assume that the causal scoring function's overall score is a function of bivariate scores on individual arcs. In turn, §7.2.3 we switch from the graph-theoretic representation to a matrix representation to make subsequent mathematical notation more convenient. At the heart of the chapter, §7.3 develops the GSF, which effectively combines p-values from lag-specific Granger causality tests to retrieve a final score on candidate graphs. In turn, §7.3.2 discusses some relevant properties of the GSF, §7.3.3 outlines the search space and §7.3.4 describes the scoring procedures involved.

7.2 Scoring Functions

7.2.1 General Scoring Functions

The central task of score-based methods can be framed as follows: given a set of candidate graphs and an appropriate scoring function, infer a graph or a set of graphs that optimise the value of that function. Given a set of candidate graphs $\mathcal{G}$ and a scoring function $\phi : \mathscr{D} \times \mathscr{G} \to \mathbb{R}$ from data domains and directed acyclic graphs to real-valued scores, the score-based task is formalised as follows:

$$\hat{G} = \text{argmax}_{G \in \mathcal{G}} \ \phi(\mathcal{D}, G) \tag{9}$$

An important advantage of score-based methods over constraint-based methods is the ability to induce an order on the set of candidate graphs, thus allowing a qualitative comparison of graphs. More specifically, this is because ϕ assigns real-valued scores: since each $G \in \mathcal{G}$ is assigned a real-valued number, ϕ induces a non-strict total order $\preceq$ over $\mathcal{G}$ where $G \preceq G'$ for $G, G' \in \mathcal{G}$ if and only if $\phi(\mathcal{D}, G) \leq \phi(\mathcal{D}, G')$ for a fixed domain $\mathcal{D}$. Given $\preceq$, it can be decided for arbitrary $G, G' \in \mathcal{G}$ (i) if G is better than G' and (ii) what the relative improvement of G over G' amounts to.

Since the qualitative order on $\mathcal{G}$ is defined in terms of ϕ, a central task of score-based methods consists in defining an appropriate scoring function. A minimal requirement in the literature is that the scoring function is a function of the graph's fit on the data and the graph's complexity [3, 9]. The graph's fit on the data, in

turn, is computed using the graph's Bayesian network (BN) interpretation [3]. Usually, a parametric model is estimated, resulting in a BN $\mathcal{M} = (G, \Theta)$. Since $\mathcal{M}$ defines a joint probability distribution $\Pr^{\Theta}(v_1, ..., v_n) = \prod_{i=1}^{n} \Pr^{\Theta}(v_i|pa_i)$ over the variables $V = \{V_1, ..., V_n\}$, a fit or likelihood score $\mathcal{L}(\mathcal{D}, \mathcal{M})$ can be defined, which captures how well the model explains the observational data. In addition, a complexity score $dim(G)$ can be defined over $\mathcal{M}$, where $dim(G)$ is proportionate to the number of parameters in Θ [66, pp. 148–149]. Given a function f that determines the trade-off between the model's fit and complexity, the general form of scoring functions is captured in the following scheme:

$$\phi(\mathcal{D}, G) = f(\mathcal{L}(\mathcal{D}, \mathcal{M}), dim(G)), \tag{10}$$

As an example, a classical pair of scoring functions are *Akaike's Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC), respectively defined as follows with N as the number of evaluated datapoints:

$$\phi_{\text{AIC}}(\mathcal{D}, G) = \mathcal{L}(\mathcal{D}, \mathcal{M}) - dim(G) \tag{11}$$

$$\phi_{\text{BIC}}(\mathcal{D}, G) = \mathcal{L}(\mathcal{D}, \mathcal{M}) - \frac{\log(N)}{2} dim(G) \tag{12}$$

Two common restrictions on scoring functions are score decomposability and score equivalence. First of all, *score decomposability* requires that the score over the whole graph is decomposable as a sum of independent local scores defined over single variables joined with their parents [9, 43, 38]. Formally:

$$\phi(\mathcal{D}, G) = \sum_{V_i \in V} g(\mathcal{D}, V_i \cup PA_i), \tag{13}$$

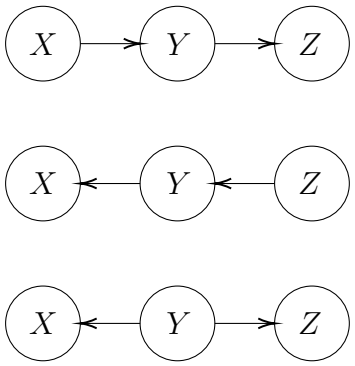
where g defines the local scores. Decomposability is desirable for at least two reasons. First of all, local scores allow for subdivision of the task of computing the graph's total score into locally and efficiently computable subtasks. Secondly, locally computed scores can be stored for reuse, increasing the efficiency of greedy heuristics such as hill climbing search [16, pp. 50–52].

The second restriction on scoring functions is *score equivalence*: graphs that belong to the same equivalence class are assigned the same score. More formally, score equivalence of a scoring function ϕ demands that for all MECs $\mathcal{G}$ and graphs $G, G' \in \mathcal{G}$, $\phi(\mathcal{D}, G) = \phi(\mathcal{D}, G')$ [38, 13, 9, 47]. Score equivalence stems from the *independence interpretation* of graphical structure. Under this interpretation, a graphical structure over variables is equated with the set of independencies it imposes on probability distributions over those variables. Since structures within the same equivalence class impose the same independence constraints, it naturally

follows that scores should be the same for graphs within the same equivalence class [14, p. 448]. In the terminology of Spirtes, Glymour, and Scheines [79], Markov equivalent structures are *statistically indistinguishable*: each graph satisfies the same relevant statistical properties and cannot be distinguished on the basis of those properties [79, p. 59].

7.2.2 Causal Scoring Functions

Under the independence interpretation, imposing score equivalence is a natural choice: structures within the same equivalence class are equivalent up to independencies. In the causal setting, however, score equivalence is a misguided assumption: since arcs are given a causal interpretation, structures within the same equivalence class are, in general, causally non-equivalent [14, p. 448]. Consider the following causal structures:



It is clear that their causal meaning is different: $X \to Y$ and $X \leftarrow Y$, for example, describe different causal relations. Although the structures are statistically indistinguishable, their different causal interpretations makes them *causally distinguishable*.

Given these considerations, a causal scoring function should drop the score equivalence assumption: it should, in principle, be possible that structures within the same equivalence class are assigned different scores. A further desideratum of a causal scoring function is, naturally, that differences in scores reflect differences in causal information in the graphs scores. Although it is quite trivial to construct a scoring function that satisfies the first desideratum, the second desideratum is clearly non-trivial: the very premise of the causal discovery task is, after all, that the true causal structure is unavailable and must be inferred from non-causal, statistical properties of the observational distribution [61, p. 43].

At the same time, the unavailability of the true causal structure is consistent with the assumption that some causal information is available. Suppose, for the

moment, that we are given a non-symmetric bivariate measure $\kappa : \mathscr{D} \times \mathscr{A} \to [0, 1]$ that scores how well a given arc captures directional causal information, derived from statistical properties of the observational distribution. Given the relevant bivariate scores from κ, a causal score $\varkappa(\mathcal{D}, G)$ can be defined on each candidate $G \in \mathcal{G}$. Given a function g that determines the trade-off between fit, complexity and causal score, the scheme for a causal scoring function is then given as follows:

$$\phi(\mathcal{D}, G) = g(\mathcal{L}(\mathcal{D}, \mathcal{M}), dim(G), \varkappa(\mathcal{D}, G)) \tag{14}$$

Although the model selection problem could be reduced to comparing graphs on the scores from $\varkappa$, integrating the scores from $\varkappa$ into the scoring function ϕ is beneficial for the following reason: it allows for weighting the contribution of the causal information, which is not possible if attention is restricted to that causal information.

7.2.3 Matrix Representations

In this section, we switch from a graph-theoretic to a matrix representation of causal structure and specify causal scoring matrices whilst remaining agnostic with respect to the causal measure $\kappa : \mathscr{D} \times \mathscr{A} \to [0, 1]$. In order to switch from the graph-theoretic to the matrix representation, it must be observed that any DAG $G = (V, A)$ stands in a one-to-one correspondence with some square matrix M^G in which each element m_{ij}^G is defined as follows:

$$m_{ij}^G = \begin{cases} 1 & \text{if } (V_i, V_j) \in A \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Since M^G defines which vertices are adjacent in G, M^G is called the *adjacency matrix* of G. Observe, now, that the matrix representation involves no loss of structural information over the graph-theoretic representation: every adjacency matrix M^G of a digraph G is a $d \times d$ matrix that can be translated back into G by defining $G = (V, A)$ with $V = \{V_1, ..., V_d\}$ and $(V_i, V_j) \in A$ if and only if $m_{ij}^G = 1$ [77, p. 6].

Under the matrix representation, assigning causal scores to present arcs effectively reduces to an element-wise matrix multiplication of the adjacency matrix with a causal scoring matrix. Given a non-symmetric function $\kappa : \mathscr{D} \times \mathscr{A} \to [0, 1]$, we first construct a non-symmetric $d \times d$ causal score matrix M^* in which each element is defined as $m_{ij}^* = \kappa(\mathcal{D}, (V_i, V_j))$. Since the adjacency matrix of a graph $G = (V, A)$ is a matrix M^G where $m_{ij}^G = 1$ just in case $(V_i, V_j) \in A$, taking the element-wise matrix product $(M^G \circ M^*)_{ij} = m_{ij}^G \cdot m_{ij}^*$ for $i = 1, ..., d$, $j = 1, ..., d$

defines a matrix M_G^* in which each non-zero element defines how well the arc corresponding to that element captures causal information.

7.3 Granger Scoring Function

7.3.1 Proposal

In this section, we motivate the use of p-values from lag-specific Granger causality tests as bivariate scores and, moreover, outline how these scores are combined into final scores on candidate graphs. First, we discuss the computation of the relevant p-values. Suppose that we want to evaluate the null hypothesis $X^{t-\tau} \perp\!\!\!\perp Y^t | \mathbf{T}^{:t} \setminus \{X^{t-\tau}\}$ for a given arc $X^{t-\tau} \to_G Y^t$. In that case, a necessary requirement is access to a conditional independence tester $\mathcal{I} : \mathcal{X} \times \mathcal{X} \times \wp(\mathcal{X}) \to [0, 1]$ that takes in a sample of observations $\{x^{i-\tau}, y^i, t^{:i-1} \setminus x^{i-\tau}\}_{i=\tau+1}^{T}$ and returns a p-value p^τ for the null hypothesis $H_0 : X^{t-\tau} \perp\!\!\!\perp Y^t | \mathbf{T}^{:t} \setminus \{X^{t-\tau}\}$. Here, we assume that the tester returns a two-tailed p-value $p = 2 \cdot \min\{\Pr(T \geq \hat{T}|H_0), \Pr(T \leq \hat{T}|H_0)\}$ for a fixed test statistic $\hat{T}$. From an intuitive point of view, p^τ is interpreted as the likelihood of observing a result at least as extreme as the observed result, given that the arc is absent. An alternative interpretation of p^τ is as the amount of evidence in favor of the null hypothesis. It is worth emphasising that, as Hubbard and Lindsay [42] discuss, p-values are not a strict measure of evidence in this sense. This interpretation is, instead, a *pragmatic interpretation* that follows the treatment of p-values in constraint-based methods as further explained in Appendix C. Given this interpretation of p-values, we can define a measure $\kappa_{\text{GC}} : \mathscr{D} \times \mathscr{A} \to [0, 1]$ that encodes the amount of evidence that disfavors the hypothesis that a lag-specific arc is absent:

$$\kappa_{\text{GC}}(\mathcal{D}, (X^{t-\tau}, Y^t)) = 1 - p^\tau \tag{16}$$

where it is assumed that κ_{GC} has the required access to the samples from $\mathcal{D}$ as well as $\mathcal{I}$. This score, then, is interpreted as a causal score on $X^{t-\tau} \to Y^t$. Corresponding to this notion, arcs with high scores correspond to arcs with strong evidence whilst arcs with weak evidence get low scores.

Given a p-value p_τ for the null hypothesis $H_0^\tau : X^{t-\tau} \perp\!\!\!\perp Y^t | \mathbf{T}^{:t} \setminus \{X^{t-\tau}\}$ for a number of considered time lags $\tau_{\max}$ as well as a method h for combining p-values, it is moreover possible to define a causal measure $\kappa'_{\text{GC}} : \mathscr{D} \times \mathscr{A} \times \mathbb{N} \to [0, 1]$ on lag-general arcs $\mathbf{X} \to_G \mathbf{Y}$, effectively tracking the amount of evidence disfavoring the null hypothesis $H_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{T} \setminus \mathbf{X}$:

$$\kappa'_{\text{GC}}(\mathcal{D}, (\mathbf{X}, \mathbf{Y}), \tau_{\max}) = 1 - h(p^0, ..., p^{\tau_{\max}}) \tag{17}$$

where it is again assumed that κ'_{GC} has access to the samples from $\mathcal{D}$ as well as the independence tester $\mathcal{I}$. Note, here, that the null hypothesis $H_0 := H_0^0 \wedge ... \wedge H_0^{\tau_{\max}}$ indicates the absence of causal interaction at any point in time whilst $H_1 := \neg(H_0^0 \wedge ... \wedge H_0^{\tau_{\max}})$ encodes the presence of causal interaction at some point in time. Hence, the relevant score is defined as $1 - h(p^0, ..., p_{\tau^{\max}})$, which as before encodes the amount of evidence in favor of the input arc.

Using the matrix representation, we can construct a $d \times d$ causal scoring matrix M^* over a time series $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$ in which each element m^*_{ij} is defined as a causal score $m^*_{ij} = \kappa'_{\mathrm{GC}}(\mathcal{D}, (\mathbf{X}_i, \mathbf{X}_j), \tau_{\max})$. Let M^G be the adjacency matrix of a candidate graph G. Under the suggested interpretation, the score entry m^*_{ij} encodes an evidential weight on the arc $\mathbf{X}_i \to_G \mathbf{X}_j$. As a consequence, the product $m^G_{ij} \cdot m^*_{ij}$ for $m^G_{ij} = 1$ defines an evidential weight for the arc $\mathbf{X}_i \to_G \mathbf{X}_j$ in the candidate G whilst zero entries correspond to absent arcs. Hence, the causal score matrix M^*_G on G defined as the element-wise product $(M^G \circ M^*)_{ij}$ includes the evidential weight on every present arc in the candidate graph and has zero entries elsewhere.

A last step has to be taken to retrieve $\varkappa(\mathcal{D}, G)$ from M^*_G: the scores from the matrix must be combined into a real-valued score. A first option is to define the final score as the Frobenius product of M^G and M^*, which is effectively a simple sum over individual arc scores:

$$\varkappa(\mathcal{D}, G) = \langle M^G, M^* \rangle_F = \sum_{i,j}^{d} m^G_{ij} \cdot m^*_{ij} \tag{18}$$

Although a simple method, two disadvantages are worth mentioning. A first disadvantage of the Frobenius score is that it sums over probabilities, which effectively assumes that the probabilities in question are mutually exclusive events. A second disadvantage is that it is difficult to interpret the resulting score: for d arcs, $\varkappa(\mathcal{D}, G)$ can fall anywhere in the interval $[0, d]$. A method that avoids both problems is to use a p-value combiner h to construct a p-value, indicating the amount of the evidence in favor of the conjunction of present arcs:

$$\varkappa(\mathcal{D}, G) = h(\{m^G_{ij} \cdot m^*_{ij} : m^G_{ij} = 1\}) \tag{19}$$

Recall that each individual product $m^G_{ij} \cdot m^*_{ij}$ for $m^G_{ij} = 1$ defines the amount of evidence in favor of the arc $\mathbf{X}_i \to_G \mathbf{X}_j$. Combining the scores for all present arcs into an overall p-value, then, corresponds to a score on the the amount of evidence in favor of the full graph. The resulting score, then, is interpreted as the amount of evidence favoring the graph as a whole.

Since we assumed that $\varkappa$ is the function that establishes score non-equivalence, it is safe to assume that $\mathcal{L}(\mathcal{D}, \mathcal{M})$ and $dim(G)$ in the expression $\phi(\mathcal{D}, G) =$

$g(\mathcal{L}(\mathcal{D}, \mathcal{M}), dim(G), \varkappa(\mathcal{D}, G))$ are equivalent across all members of an arbitrary MEC $\mathcal{G}$. Consequently, differences between $G, G' \in \mathcal{G}$ results from differences in the terms $\varkappa(\mathcal{D}, G)$ and $\varkappa(\mathcal{D}, G')$. Since higher scores scores for $\varkappa$ are assumed to reflect a better grasp of causal information, the objective posed in (14) reduces to the following objective in the setting, where the set of candidates $\mathcal{G}$ is a MEC:

$$\hat{G} = \text{argmax}_{G \in \mathcal{G}} \ \phi(\mathcal{D}, G) = \text{argmax}_{G \in \mathcal{G}} \ \varkappa(\mathcal{D}, G) \tag{20}$$

Within the context of MECs, the Granger scoring function is therefore defined as follows:

$$\phi_{\text{GSF}}(\mathcal{D}, G) = \varkappa(\mathcal{D}, G) \tag{21}$$

Although it is possible to formulate a generalised version following the scheme of (14), our attention in this work is restricted to the case of equivalent graphs, making such a formulation unnecessary.

7.3.2 Properties of the GSF

At this point, it is useful to discuss whether the GSF meets the restrictions of standard scoring functions. Concerning score decomposability, it is clear that each score is computed locally: each score $\sum_{i=1}^{d} m_{ik}^{G} \cdot m_{ik}^{*}$ for column k is computed independently from $\sum_{i=1}^{n} m_{ik'}^{G} \cdot m_{ik'}^{*}$ for $k' \neq k$ and each k'th column represents all parents of the k'th variable. Hence, our method has the usual advantages that score decomposability entails: computation of scores is subdivided into manageable subcomputations and scores can be stored for later reuse. Simultaneously, the final score derived from h is not, in the general case, a decomposable sum. Since computations are nevertheless subdivided and stored, this alone does not obstruct the use of greedy heuristics, which is the usual motivation for score decomposability as discussed in §7.2.1.

In order to see that causal scores from $\varkappa$ are non-equivalent, it suffices to observe that $\kappa_{\mathcal{I}}$ is a non-symmetric causality measure: generally, $\kappa_{\mathcal{I}}((V_i, V_j)) \neq \kappa_{\mathcal{I}}((V_j, V_i))$ for arbitrary vertices V_i and V_j since $\kappa_{\mathcal{I}}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{T} \setminus \mathbf{X}_i, \tau_{\max})$ evaluates $X_i^{t-\tau} \not\to_G X_j^t$ for different values of τ whilst $\kappa_{\mathcal{I}}(\mathbf{X}_j, \mathbf{X}_i, \mathbf{T} \setminus \mathbf{X}_j, \tau_{\max})$ evaluates $X_j^{t-\tau} \not\to_G X_i^t$ for different values of τ. In the general case, the underlying data samples for evaluation will be distinct: for $\tau = 1$, for example, the first statement is evaluated on the pairs $(X_i^1, X_j^2), ..., (X_i^{T-1}, X_j^T)$ whilst the second statement is evaluated with $(X_j^1, X_i^2), ..., (X_j^{T-1}, X_i^T)$. From this, it follows that whenever $A \neq A'$ for two graphs $G = (V, A)$ and $G' = (V', A')$ in the same Markov equivalence class $\mathcal{G}$, it can occur that $\varkappa(\mathcal{D}, G) \neq \varkappa(\mathcal{D}, G')$.

7.3.3 Search Space

A last step before outlining the experimental setup is defining the search space to which the scoring function $\varkappa$ is applied. Since MECs are defined over DAGs and since summary causal graphs can be cyclic, we restrict our attention to acyclic subgraphs of a specific class of summary graphs: namely, graphs where all cycles are self-loops. Consider a summary graph $G = (V, A)$ over a time series set $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$ such that a cycle occurs if and only if that cycle is a self-loop from a variable to itself or, equivalently, if that cycle is of length one. In that case, we can define G as the union graph $G = G^- \cup G^+$ of two graphs G^- and G^+ defined as follows: $G^- = (V^-, A^-)$ with $V^- = V$ and $A^- = \{(V_i^-, V_i^-) : V_i^- \in V^-\}$ and $G^+ = (V^+, A^+)$ with $V^+ = V$ and $A^+ = A \setminus A^-$. By construction, it follows that G^+ is guaranteed to be acyclic: all cycles in G are self-loops and, thus, moved to G^-. Intuitively, this subgraph encodes causal interaction between distinct time series under an acyclicity constraint Since G^+ is a DAG, a Markov equivalence class $\mathcal{G}_{G^+}$ becomes definable. Given $\mathcal{G}_{G^+}$, we can construct a $d \times d$ scoring matrix M^* as defined in §7.3.1. In order to reduce computation, we can set $m^*_{ij} = 0$ for arcs (V_i, V_j) not included in $C_{\mathcal{G}_{G^+}}$ as these are, by construction, not present in any candidate graph and therefore redundant. Given the matrix M^*, the scoring function is applied to the adjacency matrix M' of each candidate $G' \in \mathcal{G}_{G^+}$ to derive a causal score on G'. After scoring each G', the highest scoring graph is selected as the optimal candidate.

7.3.4 Scoring Procedure

At this point, it is helpful to outline the full scoring procedure in precise algorithmic terms. Algorithm 5 describes the process of generating a data sample $\mathcal{D}$ from an input DSCM $\mathcal{M}$, as required as input for constructing a score matrix. Given the data sample, Algorithm 7 constructs a matrix of causal scores. Algorithm 8, in turn, takes the resulting score matrix as input and returns a score on each member in the input Markov equivalence class using a p-value combiner h. In the last step, the vector of scores is used to select an optimal candidate $G^* \in \mathcal{G}$ in Algorithm 6, whilst accounting for the possibility of score ties.[2]

[2] Note: as long as score ties are not guaranteed between all members $G, G' \in \mathcal{G}$ for all equivalence classes $\mathcal{G}$, the rejection of the score equivalence assumption is consistent with occurrence of score ties.

Algorithm 1 Data Generation Phase

function `generate_data`$(\mathcal{M}, t)$
  $U, V, F \leftarrow \mathcal{M}$
  $n \leftarrow |V|$
  $\mathcal{D} \leftarrow [t, n]$

  for i in $1, ..., t$ **do**
    for j in $1, ..., n$ **do**
      $\mathcal{D}[i, j] \leftarrow f_j(\mathcal{D}^{1:i-1})$
    end for
  end for

  return $\mathcal{D}$
end function

Algorithm 2 Graph Retrieval Phase

function `get_best_candidate`$(\mathcal{G}, scores)$
  $m \leftarrow |\mathcal{G}|$
  $G \leftarrow \text{argmax}_{i \in \{1,...,m\}}\ scores[i]$

  if $|G| = 1$ **then**
    $G^* \leftarrow G$
  else if $1 < |G|$ **then**
    $G^* \leftarrow G[k]$ `for random index` $1 \leq k \leq |G|$
  `end if`

  `return` G^*
`end function`

Algorithm 3 Score Matrix Phase

function `get_scoring_matrix`$(\mathcal{D}, \mathcal{G}, \kappa_{\mathcal{I}}, \tau_{\max})$
  $d \leftarrow |V_{\mathcal{G}}|$
  $M^* \leftarrow [d, d]$

  for $i \in \{1, ..., d\}$ **do**
    for $j \in \{1, ..., d\}$ **do**
      $\mathbf{X}_i \leftarrow \mathcal{D}[:, i]$
      $\mathbf{X}_j \leftarrow \mathcal{D}[:, j]$
      $M^*[i, j] \leftarrow \kappa_{\mathcal{I}}(\mathbf{X}_i, \mathbf{X}_j, \mathcal{D} \setminus \{\mathbf{X}_i\}, \tau_{\max})$
    end for
  end for

  return M^*
end function

Algorithm 4 Scoring Phase

function `score_equivalence_class`$(\mathcal{G}, M^*, h)$
  $m \leftarrow |\mathcal{G}|$
  $scores \leftarrow [m]$

  for $k \in \{1, ..., m\}$ **do**
    $M \leftarrow M^{\mathcal{G}[k]}$
    $scores[k] \leftarrow h(M, M^*)$
  end for

  return *scores*
end function

8 Experimental Setup

8.1 Desiderata for Evaluating Causal Methods on Synthetic Data

Before outlining the experimental setup, we consider a number of desiderata involved in evaluating causal methods on *synthetic data*. Firstly, the causal discovery task requires that the structural equations of the employed data-generating models

allow for *identifiability*: otherwise, the causal discovery task becomes infeasible [66, pp. 50, 138]. Shimizu [74] and Hoyer et al. [41] discuss positive theoretical identifiability results for linear additive models with non-Gaussian noise and non-linear additive models, respectively. In addition, Runge et al. [70] enumerates a number of specific linear and non-linear dependencies for the time series causal discovery task. A second desideratum is *model realism*: properties of synthetic data should be close to real-world data. Salient properties are non-linearity, autocorrelation and noise. Relatedly, *model diversity* is desirable: the method ought to be evaluated on a large number of distinct causal models to reduce biased conclusion and augment external validity of the results. In this context, relevant properties are the number of variables in the model, the density of causal structure, the dependency type of structural equations as well as the value of coefficients defining those equations. A third and similarly related desideratum is *model dimensionality*: in all likelihood, larger dimensions affect the method's performance. A last desideratum concerns *sample size*: if the available time series data is sparse, for example, we expect that a method's performance to decrease [68, pp. 13–14].

8.2 Experimental Setup

The aim of the experiments is to evaluate how well the GSF retrieves causal structures close to the true causal structure, measured using a set of metrics specified in §8.6. Given the similarities with PCMCI$^+$ noted in §6.2, a comparison with PCMCI$^+$'s performance scores provides a meaningful point of reference. Combining this observation with the modelling desiderata from §8.1, the following subquestions are central to answering our research question:

> **(SQ.1)** How does the Granger scoring function perform compared with PCMCI$^+$?
>
> **(SQ.2)** How does dependency type affect performance?
>
> **(SQ.3)** How does model dimensionality affect performance scores?
>
> **(SQ.4)** How does sample size affect performance scores?
>
> **(SQ.5)** How does lagged versus instantaneous link proportion affect performance scores?

In order to evaluate the subquestions, two experiments are performed.[3] In the first experiment, we evaluate how well the GSF recovers the true graph from

[3] Note: the code for the experiments was written in Python and can be accessed at https://github.com/ThierryOrth/Granger-Scoring-Function/.

equivalent graphs. Hence, we assume access to the true MEC that the underlying causal model induces. Given the aim of this experiment, singleton equivalence classes are excluded. In the second experiment, we assess the GSF's performance under realistic conditions: instead of assuming access to the true MEC, we assume access to an estimate of the MEC, which is obtained using PCMCI$^+$. Since this experiment concerns the GSF's performance and not the ability to distinguish between equivalent graphs, singleton equivalence classes are not excluded. Since an estimate of the MEC does not guarantee access to the true graph, we show the scores of the optimal graph, which is the graph within the estimated MEC with the highest relative accuracy score, as defined in §8.6. The procedure for generating causal models is the same for both experiments. On each dimensionality parameter d, we run 100 iterations. Within the span of an iteration, we randomly generate a d-dimensional parameter configuration, from which a linear and non-linear causal model is constructed, as further detailed in Section §8.4. From each model, we draw observational data samples of five different sizes. On each of the five drawn samples, PCMCI$^+$ and the GSF are each applied to infer an optimal causal structure, which is subsequently evaluated using the structure evaluation metrics, as outlined in §8.6.

8.3 Synthetic Data

A causally stationary DCSM is, at its core, defined as a set of time series $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_d\}$ and structural equations $\{f_{\mathbf{X}_1}, ..., f_{\mathbf{X}_d}\}$ defining the value of those time series on each time instant. In generating observational data for a time series $\mathbf{X}_j = (X_j^1, ..., X_j^T)$, we assume the following *additive noise model* (ANM) employed by Runge et al. [70]:

$$X_j^t = f_{\mathbf{X}_j}(\mathbf{T}^{:t-1}) = \beta_j X_j^{t-1} + g\left(\sum_{i=1, j\neq i}^{d} \beta_i X_i^{t-\tau_i}\right) + \epsilon_j^t, \tag{22}$$

where g is some linear or non-linear function and $\epsilon_j^t \sim \mathcal{N}(0, 1)$ is assumed to be standard Gaussian noise. ANMs are special instances of structural equations as described in §4.3, with the constraint of weight parametrisation and additive noise. A relevant point is that ANMs allow for identifiability of causal direction underlying the data-generating process, making them suitable for generating data for the causal discovery task [66, 41].

In addition to identifiability of causal structure, the other desiderata outlined in 8.1 are met as follows. First, model realism is satisfied using both linearities and non-linearities in the place of g, by inducing autocorrelation via the constraint $\beta_i \neq 0$ whenever $i = j$ in (22) and through the use of noise terms. Secondly,

model diversity is accomplished by considering different model dimensionalities, randomising included causal links, varying the density of both lagged and instantaneous links and through arbitrary selection of coefficients and time lags. Last of all, we evaluate the effect of sample size by evaluating the method on different sample sizes.

8.4 Hyperparameters

In generating causal models, we largely follow the experimental setup from Runge et al. [70]. Observe that a DSCM is defined in terms of a dimensionality, causal links between variables, coefficients and time lags. Let $\{d_1, ..., d_n\}$ be a set of dimensionalities with $d_1 = 2$, $\{\tau_1, ..., \tau_m\}$ a set of time lag parameters and $\{\beta_1, ..., \beta_k\}$ a set of coefficients. Given a fixed d_i, a d_i-dimensional DSCM is generated as follows. First, the upper bound on the number of lagged causes as well as that on instantaneous causes is defined as the parameter $L = d_i$ if $2 < d_i$ and $L = 1$ for $d_i = 2$, in line with Runge et al. [70]. Given L, a random number of instantaneous causes as well as a random number of lagged causes are selected, both from the set $\{0, ..., L\}$. In turn, all causes are coupled with random coefficients β_j plus, in the cases of lagged causes, with random time lags τ_k. Given the resulting parameter configuration, a linear model is a model in which the function g from (22) is the identity function whilst a non-linear model uses a non-linear function in the place of g. Given a causal model and a sample size n, an observational data sample of size n are generated using (22).

In the experiment, we consider dimensionality parameters from the set $\{2, ..., 25\}$. In turn, links are generated according to the parameter L as defined before, with the exception that we always assume autogenerative dependencies. Hence, the number of links for a fixed d has a lower bound of d and an upper bound of $(2 \cdot L) + d$. Coefficients β determining causal influence are drawn from the interval $[-0.9, 0.9]$ with stepsize 0.2 and excluding 0; time lags determining the time of causal influence are drawn from the set $\{1, 2\}$. The dependency g in (22) is defined as $g(x) = x$ in the linear case and $g(x) = (1 + 5xe^{-x^2/20})x$ in the non-linear case. Last but not least, we use the set $\{20, 40, 60, 80, 100\}$ as sample size parameters for generating observational data samples.

Since non-stationarity prevents identifiability, attention is restricted to time stationary models [75, pp. 312–313]. In addition, models violating the restrictions in §7.3.3 are excluded from the experiments. Last but not least, PCMCI$^+$ has two further hyperparameters that are part of the experimental setup. First, the threshold value α_{PC} used in the edge removal step. Here, we use Runge et al.'s model selection procedure for optimising the value of α_{PC}, as further described in Runge et al. [70, p. 13]. Secondly, the parameter $\tau_{\max}$ that decides that maximal

time lag at which to evaluate causal relationships. In line with the experimental setup of Runge, we assume $\tau_{\max} = 5$ throughout the experiments [69].

8.5 Computing and Combining p-values

In addition to the hyperparameters outlined in Section 8.4, a conditional independence tester $\mathcal{I}$ and method h for combining p-values are to be defined. In choosing $\mathcal{I}$, we follow Runge et al. [70] in performing a t-test on the partial correlation coefficient in the linear case and a Gaussian Process regression combined with a distance correlation test (GPDC) in the non-linear case. Both fall under the regression-based approach of conditional independence testing and are described by Runge [68] and Runge et al. [70].

From a conceptual point of view, a p-value combiner is interpreted as a likelihood-ratio test of the null hypothesis against variations of the alternative hypothesis: it compares the evidence for the null hypothesis with that for the alternative hypothesis. Given a vector of p-values $\vec{p} = (p_1, ..., p_k)$ and a vector of weights $\vec{w} = (w_1, ..., w_k)$ defining the contribution for each p_i, a first candidate for a p-value combiner $h : V \times V \to [0, 1]$ is Fisher's method:

$$h_F(\vec{p}, \vec{w}) = -2 \sum_{i=1}^{k} w_i \log(p_i) \tag{23}$$

A second method is Stouffer's method, which transforms each p-value using the inverse of the standard normal cumulative distribution function:

$$h_S(\vec{p}, \vec{w}) = \sum_{i=1}^{k} w_i \Phi^{-1}(p_i) \tag{24}$$

A problem with both Fisher's and Stouffer's method is, however, that each p_i is assumed to be independent [37, 88]. A method that drops the independence assumption on p-values is Wilson's *harmonic mean p-value*:

$$h_H(\vec{p}, \vec{w}) = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} w_i / p_i} \tag{25}$$

A number of relevant properties of h_H are (i) robustness to positive dependency between p-values, (ii) insensitivity to the number of tests k, (iii) robustness to the distribution of the weights $\vec{w}$ and (iv) high influence by small p-values. Concerning point (iii), weights are informed by prior knowledge about the relative importance of each p_i, for example, due to differences in sample size [88]. Since such prior knowledge is not available in our experiments, we assume a uniform

weight distribution $w_i = 1$ for all $1 \leq i \leq k$. Intuitively, point (iv) says that in assessing the evidence for a conjunctive hypothesis, low amounts of evidence on any of the conjunct hypothesis has a bigger effect than high amounts of evidence. In this sense, h_H is conservative with respect to the conjunctive null hypothesis. In our case, this involves that time-lagged arcs $X_i^{t-\tau} \rightarrow_G X_j^t$ with low evidence exert bigger influence on reducing the evidence for summary arcs $\mathbf{X}_i \rightarrow_G \mathbf{X}_j$ and that, similarly, that summary arcs $\mathbf{X}_i \rightarrow_G \mathbf{X}_j$ with low amounts of evidence play a bigger role in reducing the amount of evidence for the full summary graph.

8.6 Evaluation Metrics

In the literature, two classes of evaluation metrics for causal structure learning are distinguished: (1) *graph distance-based measures* and (2) *classification-based measures*. In the experiments, we will use measures from both classes. A first distance-based measure is Structural Hamming Distance (SHD), which is defined as the minimal number of arc insertions, removals or reversals required for turning an estimated graph $\hat{G}$ into the ground-truth graph G. In formal terms, SHD is a function $\text{SHD} : \mathscr{G} \times \mathscr{G} \rightarrow \mathbb{N}$ defined as follows:

$$|\{(V_i, V_j) : (V_i, V_j) \in E_G \,\Delta\, E_{G'} \text{ or } (V_i, V_j) \in E_{G'} \,\Delta\, E_G\}|, \tag{26}$$

where $A \,\Delta\, B := (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference of A and B. A second distance measure is the Frobenius norm, defined as the square root of the sum of squared values of a matrix:

$$||A||_F = \sqrt{\text{trace}\,(A^T A)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2} \tag{27}$$

Hence, $||M^\delta||_F$ is interpreted as the square root of the sum of squared differences between M^G and $M^{\hat{G}}$. In our case, the relevant Frobenius norm is $||M^\delta||_F$ with $M^\delta = M^G - M^{\hat{G}}$: the difference matrix of the adjacency matrices of G and $\hat{G}$.

Classification measures perform graph evaluation as in a categorical classification problem: entries m_{ij}^G are interpreted as classes $c \in \{0, 1\}$ that encode absent and present arcs [12, 64]. In what follows, we restrict our attention to *accuracy scores*.[4] The standard definition of accuracy is as follows:

$$\frac{|m_{ij}^* = m_{ij}|}{|m_{ij}^* = m_{ij}| + |m_{ij}^* \neq m_{ij}|}, \tag{28}$$

[4] Note: as my supervisor pointed out, precision and recall are uninformative measures in this case, as these are identical in the setting of a Markov equivalence class; the proof is included in Appendix D.

where m^*_{ij} are entries in the adjacency matrix of the estimated graph G^* and m_{ij} are entries in the adjacency matrix of the true graph G. An additional measure of interest in our case is accuracy relativised to the points of estimation, which we refer to as *relative accuracy* (RA):

$$\frac{|\{m^*_{ij} = m_{ij} : c_{ij} = c_{ji} = 1\}|}{|\{m^*_{ij} = m_{ij} : c_{ij} = c_{ji} = 1\}| + |\{m^*_{ij} \neq m_{ij} : c_{ij} = c_{ji} = 1\}|} \tag{29}$$

where c_{ij} are entries in the CPDAG, which are effectively the points of estimation in the MEC setting. The reason for adopting this measure is as follows. Since the MEC already fixes a number of arcs, standard accuracy scores or *absolute accuracy* (AA) does not allow us to distinguish between the MEC's arcs and the GSF's performance on unresolved edges. Hence, we in addition adopt RA to specifically evaluate the GSF's contribution to the final scores.

Last but not least, we use two statistical tests for comparing performance across dependencies as well as across methods. The first test is the two-sample Kolmogorov-Smirnov (KS) test. In brief terms, the two-sample KS test computes as test statistic the greatest absolute distance between two empirical distribution functions $F_{1,n}$ and $F_{2,m}$:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{30}$$

In turn, $D_{n,m}$ is used to derive the likelihood of observing these samples given that their underlying probability distribution is the same [5, 15]. Secondly, the Spearman rank-order correlation coefficient computes dependence between variables using value-ranked lists of the variables. Given ranked variables, the pointwise difference for every i'th entry contributes to the Spearman correlation coefficient r_s as follows:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{31}$$

As usual, the coefficient's range is the interval $[-1, 1]$ [15, 91]. In further sections, we assume for both statistical significance tests a significance level of $\alpha = 0.05$.

9 Results

9.1 Experiment I

In this section, we included the part of the results most pertinent to the analysis; the entirety of results are included in Appendix F. The reader is referred to Subappendix F.1 for the relevant plots, whilst descriptive statistics are to be found

in Subappendix F.2. A description of the included plots and tables are in order. Figure 14 shows the performance scores of the GSF and PCMCI$^+$ in the linear and the non-linear setting against the number of variables in the model, with performance scores averaged over all iterations and over all sample size parameters. Figure 15, in turn, plots a number of relevant CPDAG properties against dimensionality: the number of DAGs in a MEC defined as $m = |\mathcal{G}|$, the number of edges k in the corresponding CPDAG $C_{\mathcal{G}}$ as well as the fraction of edges p in $C_{\mathcal{G}}$, which is computed as the number of edges divided by the number of edges and arcs.

Figure 16 depicts the performance scores against model dimensionality, averaged over all iterations and with each row corresponding to a fixed sample size parameter. Figure 17 and 18 show the relationship between performance scores and the proportion of instantaneous causes and lagged causes, respectively. Note that although the trends of these proportions are simply mirroring trends, we nevertheless included both plots for ease of reference. In the plots, every row corresponds to a specific structure discovery method. Proportion values were rounded to the closest number divisible by 0.025; metric scores for identical rounded proportion values were averaged. The number of evaluated models per datapoint are indicated with a color gradient that darkens the color for larger numbers. The darkest datapoint at $p = 0.0$ and $p = 1.0$ for the instantaneous and lagged link plots respectively corresponds to a total number of 358 evaluated models; the count of the rest of the datapoints are averaged at $\mu = 102$ with standard deviation $\sigma = 42$.

Table 1 records values of the Spearman correlation coefficient between CPDAG properties and performance metrics. In turn, Table 2 and 3 contain the sample-specific and average mean and standard deviation of line heights, respectively; the average mean and standard deviation of the distances across methods as well as Kolmogorov-Smirnov test results are included in Table 4. Table 5 records the correlation between sample size and performance scores; Table 6 and 7 describe the correlation between performance scores and the proportion of instantaneous and lagged causes, respectively.

9.1.1 Method Comparison

Figure 4 shows that the GSF improves on PCMCI$^+$ on all metrics. Table 3 and 4 confirm the GSF's improvement over PCMCI$^+$: the mean performance of the GSF consistently improves over that of PCMCI$^+$. In the case of SHD and FN, a significant difference in performance is witnessed between the GSF and PCMCI$^+$ across dependencies. A similar observation holds for both proportion scores, with an interesting case of convergence of the GSF on AA around $d = 15$. More salient is the GSF's performance on RA, i.e., the proportion of correctly oriented arcs in the CPDAG. It is clear that the GSF's performance exceeds that of PCMCI$^+$, which

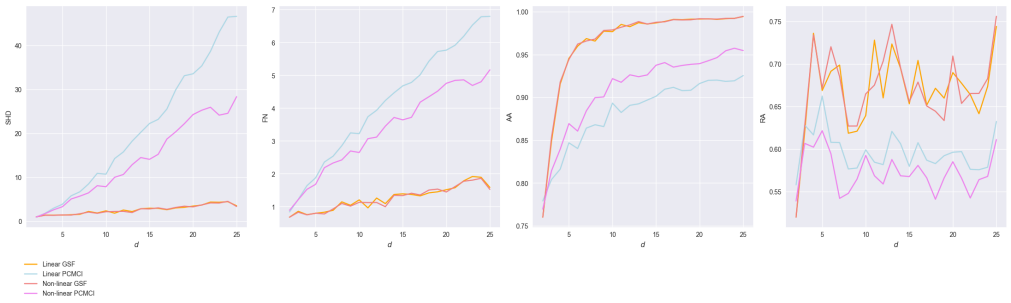


Figure 4: Plots of dimensionality values d against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores.

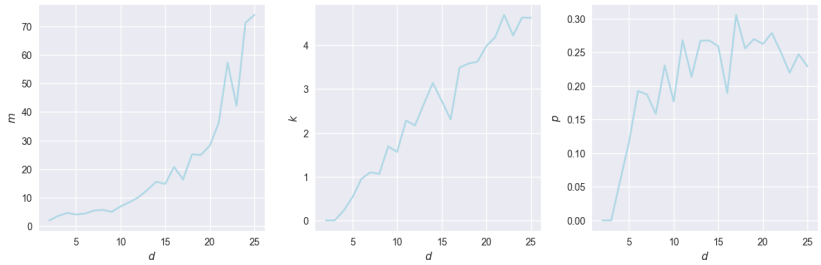


Figure 5: Plots of dimensionality values d against the number of DAGs in the MEC (m), the number of edges (k) and the proportion of edges over edges and arcs (p).

can be confirmed in the line distance statistics recorded in Table 4. Simultaneously, Table 3 shows a higher amount of variation for the GSF. Still, Figure 4 indicates that, on average, the additional variation does not result in a performance lower than PCMCI$^+$ for any value of d. From Table 4, it can be witnessed that the KS test scores are significant between the two methods across both dependencies, suggesting that their respective scores are not drawn from the same distribution.

9.1.2 Model Dimensionality

Figure 4 shows that dimensionality increases coincide with increases in SHD and FN for both discovery methods and dependencies. The magnitude of increase, however, is much higher for PCMCI$^+$ than for the GSF on both dependencies. Table 1 verifies this results numerically: the positive correlation with d is much higher for PCMCI$^+$. An interesting observation is that SHD and FN scores correlate more with the equivalence class size m than with d in the case of the GSF.

From Figure 4, it can in addition be seen that AA increases with model dimensionality whilst highly erratic behavior is witnessed for RA. Interestingly, the GSF shows converging behavior for AA around $d = 20$. Table 1 indicates a no-

table difference across the discovery methods: whilst PCMCI$^+$'s AA scores mostly correlate with d and m, those of the GSF mostly correlate with d, k and p. In the case of RA, correlations with d vanish across both discovery methods, although a small correlation with m is seen for the linear case of the GSF and the non-linear case for both the GSF as well as PCMCI$^+$.

9.1.3 Dependency Type

Figure 4 shows stark differences in performance across dependencies for PCMCI$^+$, but minimal differences in the case of the GSF. This observation is supported by Table 4, which shows that, indeed, the mean line distance between the linear and non-linear case tends to be much bigger for PCMCI$^+$, with the exception of RA scores. Furthermore, Table 4 shows that the KS test returns a significant value for PCMCI$^+$ on all metrics, whilst none of the values for the GSF are significant. This suggests that the samples for PCMCI$^+$ are drawn from the same distribution, whilst the opposite is the case for the GSF.

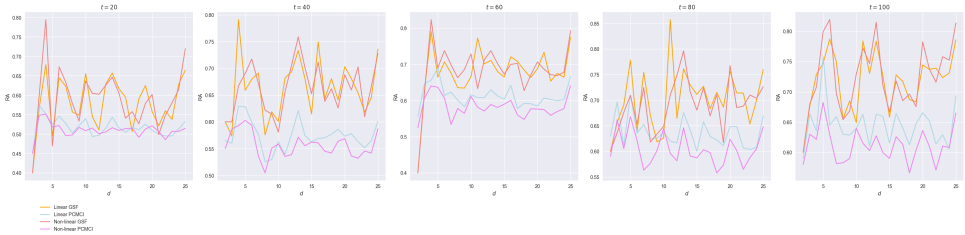


Figure 6: Plots of dimensionality values d against relative accuracy (RA) scores, with columns corresponding to specific sample size parameters.

9.1.4 Sample Size

Figure 6 indicates that SHD and FN scores generally decrease for the GSF. On the other hand, no clear trend is discernible for PCMCI$^+$. Table 2 confirms this observation: SHD and FN scores always decrease for the GSF, but not unilaterally so for PCMCI$^+$. Table 5 shows that, in addition, a weak negative correlation occurs for the GSF whilst the correlation for PCMCI$^+$ is almost zero. A second relevant observation is that for both discovery methods, AA and RA scores generally increase as sample size increases. This fact can be independently verified in Table 2. Table 5 indicates a weak positive correlation for the GSF with AA and RA; a stronger correlation is seen for PCMCI$^+$ in the case of RA, yet a vanishing correlation is observed for AA scores.

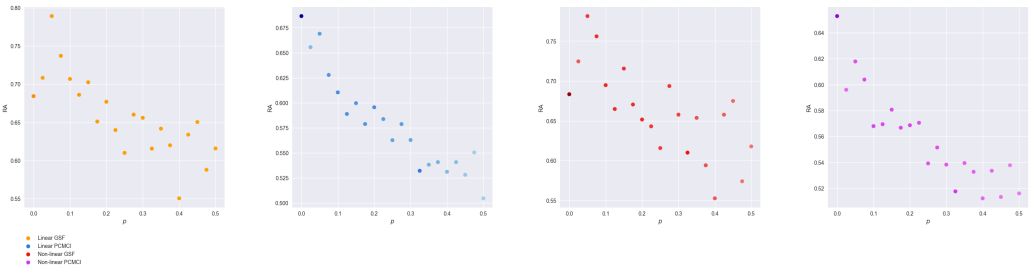


Figure 7: Plots of instantaneous link proportion values against relative accuracy (RA) scores, with columns corresponding to causal discovery methods.

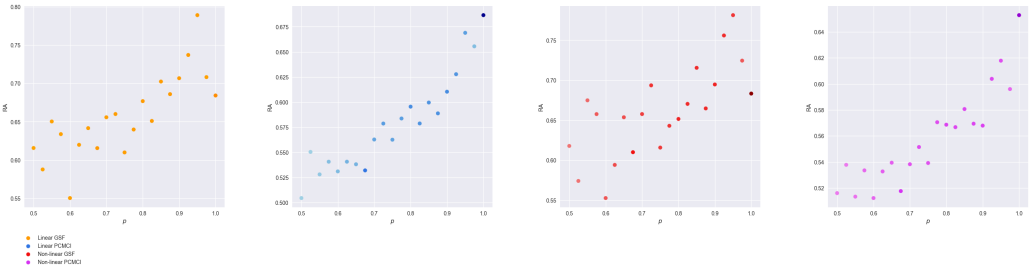


Figure 8: Plots of lagged link proportion values against relative accuracy (RA) scores, with columns corresponding to causal discovery methods.

9.1.5 Instantaneous and Lagged Causes

Figure 7 and 8 indicate a weak positive correlation with SHD and FN in the instantaneous case and a weak negative correlation for AA and RA. Since any increase in the proportion of instantaneous causes involve a commensurate decrease in the proportion of lagged causes, this trend is reversed in the lagged case: a weak negative correlation occurs for SHD and FN whilst a weak positive correlation holds for AA and RA. Table 6 and 7 can be consulted to independently verify these observations. A salient finding in Figure 8 is that a proportion of lagged causes close to one gives notably high RA scores. In the linear case, RA scores are seen to center around the interval $[0.70, 0.75]$. In the non-linear case, RA scores cluster around the interval $[0.70, 0.80]$.

9.2 Experiment II

The results of the second experiment are included in Appendix G. Plots and descriptive statistics are analogous to those from the first experiment, with the exception that we included the performance scores of the optimal graph from the MEC in Figure 9 and 11 as well as Table 10 and 11 as a meaningful point of

reference. The optimal graph, to repeat, is the graph within the estimated MEC with the highest RA score. In this experiment, the number of evaluated models at the darkest point at $p = 0.0$ and $p = 1.0$ for the instantaneous and lagged link plots equals 354; the count of the other datapoints are averaged at $\mu = 102$ with standard deviation $\sigma = 43$.

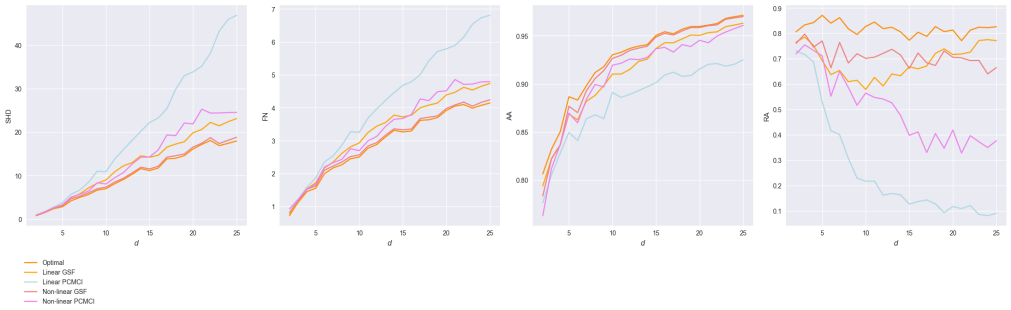


Figure 9: Plots of dimensionality values d against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores.

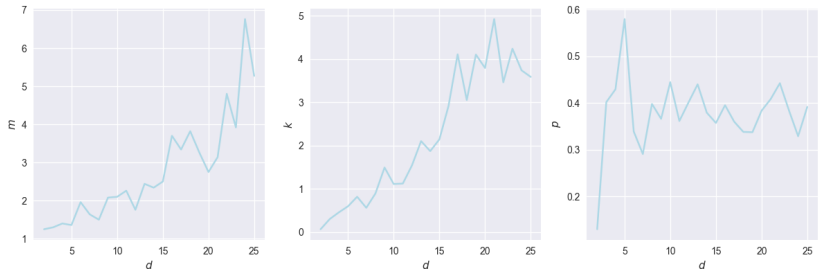


Figure 10: Plots of dimensionality values d against the number of DAGs in the MEC (m), the number of edges (k) and the proportion of edges over edges and arcs (p).

9.2.1 Method Comparison

Figure 9 shows that the GSF outperforms PCMCI$^+$. Table 10 and 11 confirm this result: the scores of the GSF consistently improve on those of PCMCI$^+$. In particular, the GSF is significantly closer to optimal performance than PCMCI$^+$. In the case of SHD and FN, the distance between the GSF's scores and the optimal score is notably smaller than those of PCMCI$^+$, especially in the non-linear case. For AA, a similar observation holds. In the case of the RA scores, however, the GSF's performance falls short of optimal performance. Simultaneously, it is clear that the GSF significantly improves on both PCMCI$^+$ as well as random chance.

As in the previous experiment, the amount of variation for both discovery methods is relatively high. Table 11 shows that across both dependencies and all metrics, the null is rejected for the point-wise method comparison, suggesting that the scores are not drawn from the same distribution. Interestingly, the point-wise comparisons with the optimal scores show that the null is rejected everywhere except for the GSF in the non-linear case for SHD, FN and AA, suggesting that the scores for the GSF in these cases are drawn from the same distribution as the optimal scores.

9.2.2 Model Dimensionality

Figure 9 indicates that all scores increase with model dimensionality. In the case of SHD and FN, all discovery methods show a highest positive correlation with d, as well as a strong correlation with m and k and a weak correlation with p. For AA, a strong correlation again occurs for d, as well as a weak to modest correlation with m, k and p in the case of the GSF; a small positive correlation for p is witnessed for PCMCI$^+$ across dependencies, as well as a weak to modest correlation with m and k in the linear case. Concerning RA scores, diverging trends are observed. In the case of the GSF, a weak correlation is witnessed for d, p and m, whilst a strong negative correlation with m occurs in the non-linear case, as well as a weak to modest correlation with d, k and p. PCMCI$^+$, on the other hand, shows an extremely strong negative correlation with d and a moderately strong negative correlation with k in the linear case as well as a weaker negative correlation for m and p. In the non-linear case, however, the negative correlation with d significantly shrinks, whilst the negative correlation with k and especially that with m grows.

9.2.3 Dependency Type

Figure 9 shows significant differences in performance across dependencies for PCMCI$^+$ but minimal differences in the case of the GSF. This observation is supported by Table 11, which shows that, indeed, the mean line distance between the linear and non-linear case tends to be much bigger for PCMCI$^+$. In the point-wise dependency comparison of both discovery methods, Table 11 shows that the null hypothesis should be rejected for all metrics.

9.2.4 Sample Size

Figure 11 suggests that both SHD and FN scores decrease for the GSF whilst no clear trend is witnessed for PCMCI$^+$. Table 12 indicates a unilateral performance improvement for the GSF. With minor exceptions, a general improvement of AA

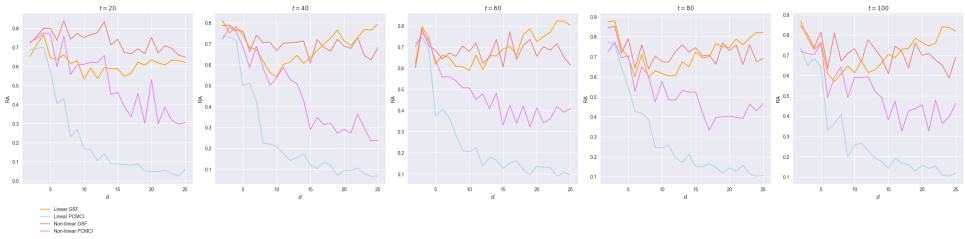


Figure 11: Plots of dimensionality values d against relative accuracy (RA) scores, with columns corresponding to specific sample size parameters.

and RA scores can be witnessed for both discovery methods. Interestingly, Table 12 reports the presence of a weak correlation for the GSF and almost no correlation for PCMCI$^+$ across all scores except RA. In the case of RA, the correlation seems to vanish for the non-linear case in particular, whilst the correlation for the linear case seems almost identical for both discovery methods.

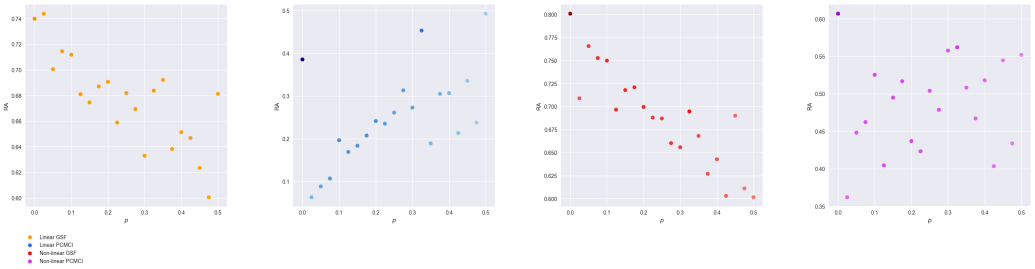


Figure 12: Plots of instantaneous link proportion values against relative accuracy (RA) scores, with columns corresponding to causal discovery methods.

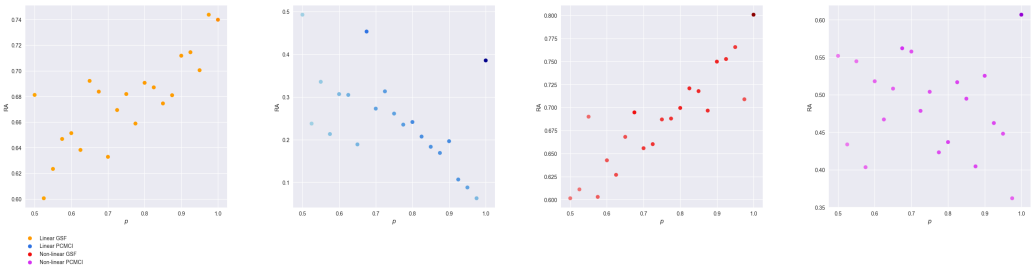


Figure 13: Plots of lagged link proportion values against relative accuracy (RA) scores, with columns corresponding to causal discovery methods.

9.2.5 Instantaneous and Lagged Causes

Figure 12 and 13 indicate a similar pattern as in the first experiment: an increase in the proportion of instantaneous causes impairs scores, whilst an increase in the proportion of lagged causes improves scores. Table 13 and 14 confirm this observation. A striking result is that in the linear case of PCMCI$^+$, the correlation between the proportion of instantaneous causes and RA scores is positive. Although more instantaneous causes seem to be beneficial in this specific case, it is clear that this does not help the AA scores, where the correlation is negative. Hence, the initial observation is sound. We observe that, as before, a high proportion of lagged causes for the GSF results in high RA scores, especially in the non-linear case. As before, scores in the linear case fall in the interval $[0.7, 0.75]$ whilst scores in the non-linear case cluster around $[0.75, 0.8]$.

10 Discussion

10.1 Analysis

10.1.1 Method Comparison

The results of the first experiment indicate that the GSF significantly improves on PCMCI$^+$ on the task of deciding the correct orientation of edges in the CPDAG. In particular, we can corroborate this observation on the basis of the significantly higher performance of the GSF on RA scores, which directly document the proportion of correctly oriented arcs in the CPDAG. A notable observation is the converging trend for the GSF's AA scores, which seems wholly absent for PCMCI$^+$ within the range of tested dimensionality parameters. An interesting result from the second experiment is that in the absence of the true Markov equivalence class, the GSF's performance on RA scores stays similar to in the first experiment, indicating that the GSF's performance on RA scores can be extended to the practical case where edges in the estimated MEC may not correspond to those in the true MEC. Furthermore, it was seen that the GSF performs quite well for higher sample sizes, larger proportions of lagged causes and non-linear dependencies. Simultaneously, performance was shown to fall short of optimal performance on average, suggesting that the PCMCI$^+$'s CPDAG output may be preferable depending on the size of the model, sample size, time granularity and expected dependency.

10.1.2 Dependency Type

Both experiments show that the GSF's performance between the linear and non-linear case are, in general, negligible. An interesting exception concerns the RA

scores. In the first experiment, a small difference in performance was witnessed, but the KS test suggested that the distributions underlying the linear and non-linear scores were the same. In the second experiment, on the other hand, the KS test warranted rejection of the null hypothesis, in line with the observed differences across dependencies as observed in Figure 9. A plausible explanation for this difference relates to the number of evaluated edges, which is bigger in absolute and relative terms as seen in Figure 5 and 10. Plausibly, the increase in edges makes the difference in identifiability more salient across the linear and non-linear case. This is not surprising given the additive Gaussian noise in the ANM: it is more difficult to fit a line in the linear regression step of the partial correlation test. Since the GPDC test is capable of fitting a broader range of functions than the partial correlation test, it may be that higher identifiability in the non-linear case is partially or wholly due to the GPDC test rather than the underlying non-linear dependency. In any case, the non-parametric nature of GPDC may be preferred in that it is capable of capturing a broader range of functions than partial correlation tests, thus increasing identifiability of causal dependencies.

10.1.3 Model Dimensionality

In both experiments, it was seen that increases in model dimensionality correspond to worse performance on SHD and FN metrics whilst augmenting performance on AA and RA scores. On observing the correlation values, it must be concluded that model dimensionality on its own cannot fully explain the metric scores. In the first experiment, SHD and FN scores were seen to correlate significantly more with m than with d. Plausibly, this is because a larger candidate space includes more graphs more distant from the true graph on SHD and FN scores, which results in an increasing likelihood of making more errors on SHD and FN scores. Whilst AA scores show a high correlation with d, k and p, no strong correlation is found between RA scores and any of the CPDAG properties. In the second experiment, it was observed that all CPDAG properties and specifically d and k correlate well with SHD, FN and AA. The shrinkage of the correlation with m can be explained as follows: since the true graph is not necessarily available, increases in m do not necessarily correspond to more errors on SHD and FN scores. Although an increase has occurred in the correlation of the CPDAG properties with RP scores, the correlation coefficients are still marginal. Hence, RA scores cannot be explained in terms of just CPDAG properties. Rather, it is plausible that RA scores depend on a bigger set of factors, which may include the number of lagged causes, sample size as well as further factors.

10.1.4 Sample Size

Both experiments show that, in general, a higher sample size is conducive to performance on all metrics. To be precise, a negative correlation is witnessed for SHD and FN scores whilst a positive correlation is seen for both AA and RA scores. For the GSF, correlations stay almost constant across experiment. The second experiment shows higher absolute mean scores for SHD, FN and AA scores as witnessed when comparing Table 3 and Table 10. Two observations are in order. First, the mean line scores end at a similar endpoint for $t = 100$, indicating that the GSF's performance on high sample sizes is robust on distinguishing between equivalent graphs between the two experiments. Secondly, a difference in trends is witnessed across experiments. Whilst line distances proportionately grow with sample size in the first experiment, this is not the case in the second experiment. The highest values in the non-linear case, for example, are attained at $t = 20$ and $t = 80$. The counterintuitive conclusion seems to be that higher sample size does not increase identifiability. Since this may change for higher sample sizes, we refrain from drawing the general conclusion here.

10.1.5 Link Proportion

Across the experiments, we witnessed the following trend: a higher proportion of instantaneous causes harms the GSF's performance on all metrics. In addition, the correlations between link proportion and metrics performance were generally similar between the first and second experiment. A clear point of difference concerns the performance in the linear case, which is seen to decrease in the second experiment. Since the performance in the non-linear case does not decrease, we suggest that this, again, results from a performance differences across dependencies as described in §10.1.2.

A salient result from Figure 18 and 23 concerns the beneficial effect of lagged link proportion: if the proportion of lagged links is near one, then RA scores get a score around the $[0.70, 0.75]$ in the linear case and $[0.70, 0.80]$ in the non-linear case. Hence, it is beneficial to ensure higher granularity of time series data, as discussed in §4.5. Plausibly, lower identifiability for more instantaneous causes can be explained in terms of a problem of symmetry. Since evaluating $X^{t-\tau} \to_G Y^t$ and $Y^{t-\tau} \to_G X^t$ for instantaneous causes is measured at $\tau = 0$, their samples are the same and so their final p-value becomes the same. For lagged causes, p-values will in general be asymmetric, as discussed in §7.3.2. As a consequence, it becomes more difficult to infer causal direction as the number of instantaneous causes increases, analogous to the problem of correlational symmetry discussed in §2.

10.2 Limitations

10.2.1 Model Class

A first issue relating to the model class is the focus on summary causal graphs. Although temporal information is used at inference time, the final graph representation wholly ignores this information: arcs encode the presence of causal interaction yet the time of interaction is not disclosed. As a consequence, summary graphs become difficult to interpret in scenarios in which time of causal interaction is relevant. A second limitation of the model class concerns the ability to indicate latent confounders. In particular, the DAG representation has no resources for expressing the presence of a cause external to the variables internal to the model: all connections between variables, after all, represent direct causal relationships. In the literature, the usual model class for latent confounders is a mixed acyclic graph, which models latent causes using bi-directed arcs between effects [27, pp. 94–96].

10.2.2 Search Space

A limitation concerning the search space is the use of a Markov equivalence class as candidate class. First of all, it is not clear if the MEC is the optimal candidate class choice for the GSF. In practice, the equivalence class has to be inferred using a prior procedure, meaning that the GSF's performance will depend on the prior procedure's degree of success in estimating the equivalence class of the true graph. In addition, scoring the full equivalence class is computationally inefficient. In the worst case, the equivalence class is superexponentially large in the number of variables: given a DAG over n vertices with the complete undirected graph as skeleton, for example, the resulting equivalence class has a total of $n!$ members [29].

A last problem of using an equivalence class is that our attention was restricted to a special subclass of summary graphs: those in which all cycles are due to autogenerative dependencies. Since evaluating causal interaction in our setting amounts to evaluating one-directional causal influence, it is unlikely that performance will significantly change for summary graphs with cycles. Still, this conclusion cannot be drawn at this point and, moreover, scoring cyclic summary graphs naturally requires a different search procedure than a simple exhaustive search within the equivalence class.

10.2.3 Scoring Procedure

A notable challenge of the proposed scoring function is that the choice of an adequate scoring function is shifted to the choice of an adequate p-value combiner. In the literature, a number of different methods are available [89, 4, 6, 88, 45]. Although Wilson [88]'s method has the advantage of accounting for dependence as well as prior information, it is not unique in these properties [93]. In order to optimise the GSF's performance, it is instructive to evaluate performance under different combination methods. In addition, such an evaluation is bound to provide better insight into the overall effectiveness of the GSF. An additional limitation relates to the computational costs of constructing the causal score matrix. If we are given a maximal time lag $\tau_{\max}$ and a CPDAG with n edges, then the total number of conditional independence tests equals $\tau_{\max} \cdot (2 \cdot n)$. If the conditional independence tester itself is already computationally expensive, such as with GPDC, this is clearly undesirable [70, p. 13].

10.3 Future Work

A number of research directions naturally follow from the limitations noted in Section 10.2. A first line of research is to generalise the GSF to graph representations such as extended summary or window graphs. An interesting direction would be to follow Pamfil et al. [59]'s division of the discovery process into a lagged and instantaneous step. In the former step, time-ordering constraints can be applied whilst the latter step could use an equivalence search. As an advantage, it is easy to infer summary graphs from these graphs if so desired [2]. An additional line of research is to use MAGs as model class, so as to account for latent causes. In both lines of research, an advantage of the GSF is the following: since the GSF just evaluates bivariate causal connections, the scoring function is model agnostic and allows for different graph representations.

A third extension of the current research is to use an efficient search procedure to query the Markov equivalence class, instead of generating the full equivalence class. A promising procedure would be an equivalence search in the spirit of Meek [53] and Heckerman, Meek, and Cooper [38]. In addition, the construction of the causal score matrix could be integrated into the scoring procedure. In this way, scores can be computed and stored for arcs present in the candidates, which may reduce the total number of required computations. A last line of research is to perform an empirical evaluation on different p-value combiners to assess relative differences in their effect on the GSF's performance.

11 Conclusion

In this thesis, we proposed a causally informed scoring function that we called the Granger scoring function. In this way, we sought to meet two important desiderata of causal scoring function: first, that it breaks the score equivalence assumption and, second, that it is causally informed. The resulting scoring function combines significance values from lag-specific Granger causality to construct a score on candidate graphs, which is interpreted as the amount of evidence in favor of those graphs. The experiments were constructed to evaluate if the GSF meets the two desiderata, that is, whether it reliably recovers causal structure from otherwise Markov equivalent graphs. First of all, we performed an experiment to evaluate how well the GSF recovers the true graph from equivalent graphs, given that the true graph is accessible within the candidate space. Secondly, an experiment was performed to evaluate how well the GSF's performance fares in settings where an estimation of the MEC is given, which effectively involves that the true graph is not necessarily accessible within the candidate space.

The experiments indicated a relatively high performance on all metrics. In the first experiment, it was clear that the GSF improved on PCMCI$^+$. Across experiments, relative accuracy scores of the GSF stayed around the same level, indicating that performance remains robust in the more realistic case where the true graph is not necessarily accessible. On the one hand, it was clear that the effects of model dimensionality and sample size on relative accuracy scores are not unilateral across parameter settings. On the other hand, non-linear dependencies and larger proportions of lagged causes resulted in improved performance, which can be explained in terms of better identifiability in the former case and the presence of asymmetry in p-values in the latter case. In the former case, it seems advantageous to apply a flexible conditional independence test such as GPDC, which induces identifiability through the ability to fit a wider range of functions. Concerning the latter, we conclude that granularity of time series data is important for augmenting the GSF's performance. Hence, it is key that the sampling process underlying the data is in tune with the minimal time period of causal interaction within the domain of interest. Instead of deciding which causal interactions take place, the required domain knowledge thus reduces to knowledge of the minimal time frame separating cause and effect.

A Notation Table

Notation	*Interpretation*
$\wp(\cdot)$	The power set of the input set.
$\mathbf{T}$	A set of time series $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$ with $\mathbf{X}_i = \{X_i^1, ..., X_i^T\}$.
τ	A discrete time lag.
$\mathscr{G}$	The space of directed acyclic graphs.
$\mathscr{D}$	The space of data samples.
$\mathcal{G}_G$, $[G]_\sim$	The Markov equivalence class of digraph G.
ϵ	A noise or error term.
$\|\| \cdot \|\|_F$	The Frobenius norm of the input matrix.
$\langle \cdot, \cdot \rangle_F$	The Frobenius inner product of the input matrices.
PA_i	The parent set of V_i in a given graph $G = (V, A)$ defined as $\{V_j : (V_j, V_i) \in A\}$.
pa_i	A value configuration of the parent set PA_i.
π	A path in a graph defined as a sequence of adjacent vertices $V_1 ... V_n$.
$\sigma^*(\cdot)$	Descendant set, includes the input vertex and all vertices reachable with a directed path.
$\mathcal{A}(\cdot)$	The set of adjacencies of the input variable.
$X \perp\!\!\!\perp Y \| Z$	Shorthand for "X is conditionally independent of Y given Z".
$X \to Y$	Shorthand for "X is a cause of Y".
$X \to_G Y$	Shorthand for "X is a G-cause of Y".
$\mathcal{X}$	An arbitrary domain of time series or standard random variables.
$\mathcal{I} : \wp(\mathcal{X})^3 \to [0, 1]$	A conditional independence tester returning a probability value.
$\phi : \mathscr{D} \times \mathscr{G} \to \mathbb{R}$	A general scoring function.
$\kappa : \mathscr{D} \times \mathscr{A} \to \mathbb{R}$	A bivariate, non-symmetric causal measure.
$\varkappa : \mathscr{D} \times \mathscr{G} \to \mathbb{R}$	A causal scoring function.
$h : V \times V \to [0, 1]$	A p-value combiner from p-value vectors and weight vectors to a p-value.
M_G	The adjacency matrix of graph G.
M^*	A causal scoring matrix.
M_G^*	The matrix of scores retrieved from element-wise multiplication of M^* with M_G.

B Procedures

Algorithm 5 Data Generation Phase

function `generate_data`($\mathcal{M}, t$)
  $U, V, F \leftarrow \mathcal{M}$
  $n \leftarrow |V|$
  $\mathcal{D} \leftarrow [t, n]$

  for i in $1, ..., t$ **do**
    for j in $1, ..., n$ **do**
      $\mathcal{D}[i, j] \leftarrow f_j(\mathcal{D}^{1:i-1})$
    end for
  end for

  return $\mathcal{D}$
end function

Algorithm 6 Graph Retrieval Phase

function `get_best_candidate`($\mathcal{G}, scores$)
  $m \leftarrow |\mathcal{G}|$
  $G \leftarrow \text{argmax}_{i \in \{1,...,m\}}\ scores[i]$

  if $|G| = 1$ **then**
    $G^* \leftarrow G$
  else if $1 < |G|$ **then**
    $G^* \leftarrow G[k]$ `for random index` $1 \leq k \leq |G|$
  `end if`

  `return` G^*
`end function`

Algorithm 7 Score Matrix Phase

function `get_scoring_matrix`$(\mathcal{D}, \mathcal{G}, \kappa_{\mathcal{I}}, \tau_{\max})$
  $d \leftarrow |V_{\mathcal{G}}|$
  $M^* \leftarrow [d, d]$

  for $i \in \{1, ..., d\}$ **do**
    for $j \in \{1, ..., d\}$ **do**
      $\mathbf{X}_i \leftarrow \mathcal{D}[:, i]$
      $\mathbf{X}_j \leftarrow \mathcal{D}[:, j]$
      $M^*[i, j] \leftarrow \kappa_{\mathcal{I}}(\mathbf{X}_i, \mathbf{X}_j, \mathcal{D} \setminus \{\mathbf{X}_i\}, \tau_{\max})$
    end for
  end for

  return M^*
end function

Algorithm 8 Scoring Phase

function `score_equivalence_class`$(\mathcal{G}, M^*, h)$
  $m \leftarrow |\mathcal{G}|$
  $scores \leftarrow [m]$

  for $k \in \{1, ..., m\}$ **do**
    $M \leftarrow M^{\mathcal{G}[k]}$
    $scores[k] \leftarrow h(M, M^*)$
  end for

  return $scores$
end function

C Explanation: Interpretation of p-values

Within PC and PCMCI and their variants, conditional independence in the skeleton phase is decided with a conditional independence tester $\mathcal{I} : \mathcal{X} \times \mathcal{X} \times \wp(\mathcal{X}) \to [0, 1]$ and significance level $\alpha \in [0, 1]$ as follows:

$$\alpha < \mathcal{I}(X, Y, Z) \Rightarrow X \perp\!\!\!\perp Y | Z \tag{32}$$

$$\mathcal{I}(X, Y, Z) \leq \alpha \Rightarrow X \not\perp\!\!\!\perp Y | Z \tag{33}$$

In other words, the p-value $p = \mathcal{I}(X, Y, Z)$ is used for accepting the null hypothesis $H_0 : X \perp\!\!\!\perp Y | Z$ in (32) and accepting the alternative hypothesis $H_1 : X \not\perp\!\!\!\perp Y | Z$ in (33). Since the skeleton phase starts with complete undirected graph $G = (V, E)$ and given faithfulness and the contrapositive of the causal Markov condition, this results in removal and acceptance of the edge $X - Y$, respectively. In this way, p is treated as the probability that the edge is included given the data, with α acting as a threshold. Under this interpretation, higher values of p correspond to a higher amount of evidence in favor of the null hypothesis [90, pp. 109–110].

D Proof: Precision-Recall Collapse

We prove that precision and recall of a candidate graph are equal whenever that graph belongs to the Markov equivalence class of the true graph. Given the definitions below, it suffices to show that FP = FN for any candidate graph whenever that graph belongs to the Markov equivalence class of the true graph.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Pick an arbitrary graph $G = (V, A)$ and consider the MEC $\mathcal{G}_G$ as well as a candidate graph $G' \in \mathcal{G}_G$. On expanding FP and FN, we see that $\text{FP} = |\{(V_i, V_j) : (V_i, V_j) \in A' \text{ and } (V_i, V_j) \notin A\}|$ and $\text{FN} = |\{(V_i, V_j) : (V_i, V_j) \notin A' \text{ and } (V_i, V_j) \in A\}|$. Consider two arbitrary vertices $V'_i, V'_{j'}$. Since G and G' are DAGs and since $(V'_i, V'_j) \in A'$ if and only if $(V'_i, V'_j) \in A$ or $(V'_j, V'_i) \in A$, the following equivalence holds:

$$
\begin{gathered}
(V_i', V_j') \in \{(V_i, V_j) : (V_i, V_j) \in A' \text{ and } (V_i, V_j) \notin A\} \\
\Leftrightarrow \\
(V_i', V_j') \in A' \text{ and } (V_i', V_j') \notin A \\
\Leftrightarrow \\
(V_j', V_i') \notin A' \text{ and } (V_j', V_i') \in A \\
\Leftrightarrow \\
(V_j', V_i') \in \{(V_i, V_j) : (V_i, V_j) \notin A' \text{ and } (V_i, V_j) \in A\}
\end{gathered}
$$

Since V_i' and V_j' were arbitrary, it follows that $|\{(V_i, V_j) : (V_i, V_j) \in A' \text{ and } (V_i, V_j) \notin A\}| = |\{(V_i, V_j) : (V_i, V_j) \notin A' \text{ and } (V_i, V_j) \in A\}|$ and so FP = FN. Since G, $\mathcal{G}$ and G' were moreover arbitrary, precision and recall of a candidate graph are equal whenever that graph belongs to the Markov equivalence class of another graph.

E Proof: CPDAG Score

We prove that if a candidate DAG belongs to a MEC, then the number of incorrect arcs in the CPDAG will always be at least the number of incorrect arcs in the DAG.

Consider a MEC estimate $\mathcal{G}^*$, the corresponding CPDAG $C_{\mathcal{G}^*} = (V_{\mathcal{G}^*}, A_{\mathcal{G}^*})$ and an arbitrary candidate graph $G = (V, A) \in \mathcal{G}^*$. Observe that since $G \in \mathcal{G}^*$, the set of arcs in G and $C_{\mathcal{G}^*}$ are the same. Hence, any difference in scores must result from the edges in $C_{\mathcal{G}^*}$. Suppose that the number of edges equals m and that the number of errors on arcs equals k. Next, take an arbitrary edge $\{V_i, V_j\}$. Since $\{V_i, V_j\}$ if and only if $(V_i, V_j) \in A_{\mathcal{G}^*}$ and $(V_j, V_i) \in A_{\mathcal{G}^*}$, $C_{\mathcal{G}^*}$ gets at least one error. Since G is a DAG included in $\mathcal{G}^*$, it follows that either $(V_i, V_j) \in A$ or $(V_j, V_i) \in A$. In either case, G receives at most one error. Since this argument applies to all edges in $C_{\mathcal{G}^*}$, the number of errors for G equals $k + i$ for some $i \leq m$ whilst the number of errors for $C_{\mathcal{G}^*}$ equals $k + j$ for some $m \leq j$. Since therefore $k + i \leq k + j$, the desired conclusion follows.

F Results Experiment I

F.1 Plots

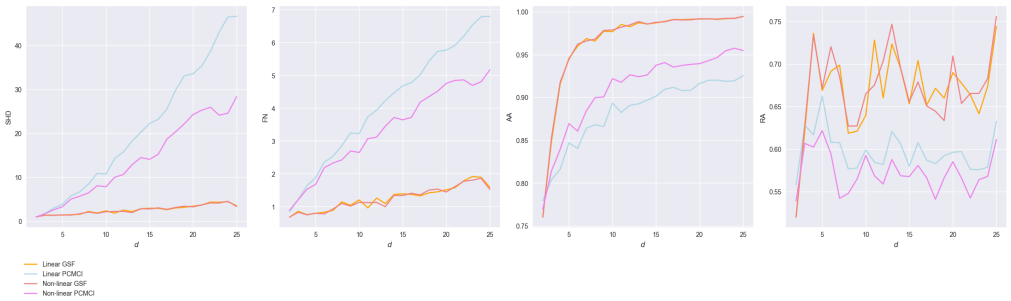


Figure 14: Plots of dimensionality values d against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores.

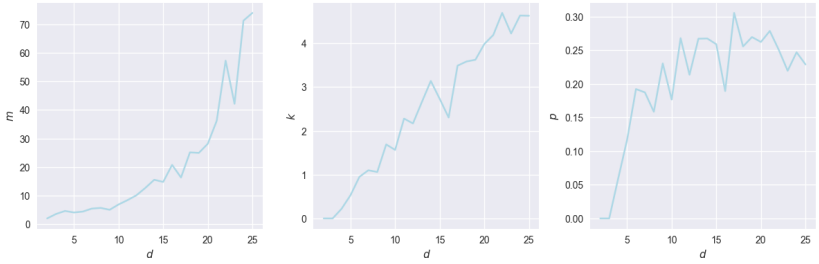


Figure 15: Plots of dimensionality values d against the number of DAGs in the MEC (m), the number of edges (k) and the proportion of edges over edges and arcs (p).

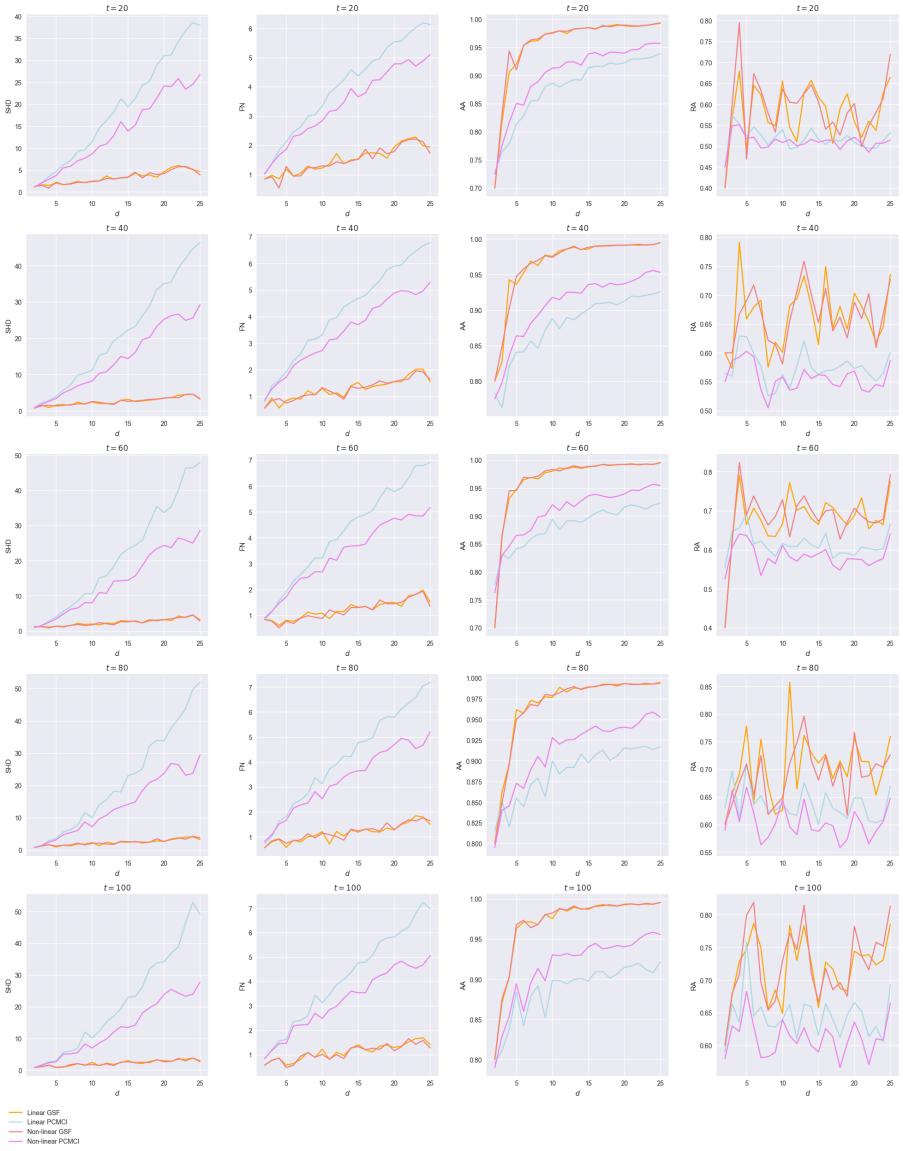


Figure 16: Plots of dimensionality values d against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores, with rows corresponding to specific sample size parameters.

Figure 17: Plots of instantaneous link proportion values against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores, with rows corresponding to causal discovery methods.

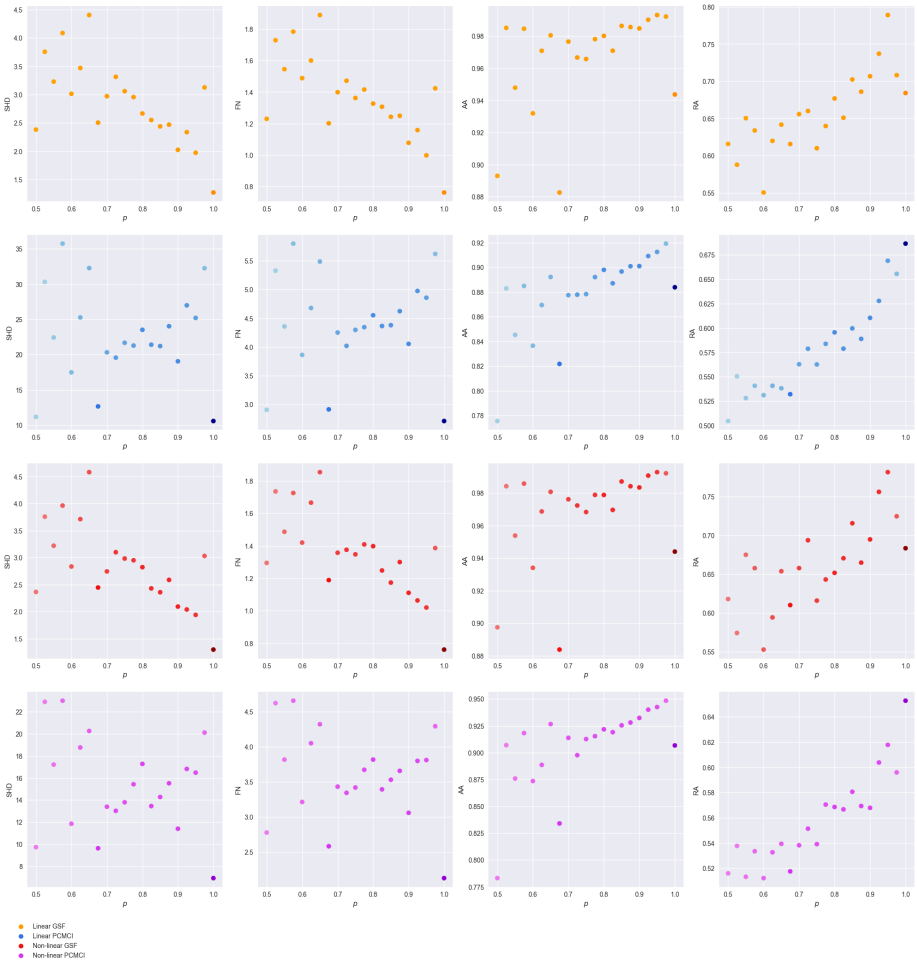


Figure 18: Plots of lagged link proportion values against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores, with rows corresponding to causal discovery methods.

F.2 Descriptive Statistics

			SHD	FN	AP	RP
GSF	Linear	d	$r_s = 0.476, p = 0.0$	$r_s = 0.456, p = 0.0$	$r_s = 0.58, p = 0.0$	$r_s = 0.069, p = 0.017$
		m	$r_s = 0.72, p = 0.0$	$r_s = 0.675, p = 0.0$	$r_s = 0.1, p = 0.001$	$r_s = 0.106, p = 0.0$
		k	$r_s = 0.215, p = 0.0$	$r_s = 0.2, p = 0.0$	$r_s = 0.397, p = 0.0$	$r_s = 0.037, p = 0.204$
		p	$r_s = -0.013, p = 0.647$	$r_s = -0.015, p = 0.613$	$r_s = 0.413, p = 0.0$	$r_s = 0.044, p = 0.132$
	Non-linear	d	$r_s = 0.467, p = 0.0$	$r_s = 0.445, p = 0.0$	$r_s = 0.578, p = 0.0$	$r_s = 0.073, p = 0.012$
		m	$r_s = 0.718, p = 0.0$	$r_s = 0.667, p = 0.0$	$r_s = 0.094, p = 0.001$	$r_s = 0.111, p = 0.0$
		k	$r_s = 0.214, p = 0.0$	$r_s = 0.199, p = 0.0$	$r_s = 0.403, p = 0.0$	$r_s = 0.045, p = 0.122$
		p	$r_s = -0.013, p = 0.651$	$r_s = -0.01, p = 0.723$	$r_s = 0.425, p = 0.0$	$r_s = 0.05, p = 0.081$
PCMCI$^+$	Linear	d	$r_s = 0.977, p = 0.0$	$r_s = 0.976, p = 0.0$	$r_s = 0.708, p = 0.0$	$r_s = 0.007, p = 0.812$
		m	$r_s = 0.666, p = 0.0$	$r_s = 0.666, p = 0.0$	$r_s = 0.443, p = 0.0$	$r_s = 0.04, p = 0.163$
		k	$r_s = 0.64, p = 0.0$	$r_s = 0.642, p = 0.0$	$r_s = 0.098, p = 0.001$	$r_s = -0.008, p = 0.78$
		p	$r_s = 0.446, p = 0.0$	$r_s = 0.447, p = 0.0$	$r_s = -0.037, p = 0.197$	$r_s = -0.015, p = 0.595$
	Non-linear	d	$r_s = 0.81, p = 0.0$	$r_s = 0.807, p = 0.0$	$r_s = 0.627, p = 0.0$	$r_s = 0.057, p = 0.047$
		m	$r_s = 0.574, p = 0.0$	$r_s = 0.574, p = 0.0$	$r_s = 0.392, p = 0.0$	$r_s = 0.083, p = 0.004$
		k	$r_s = 0.694, p = 0.0$	$r_s = 0.697, p = 0.0$	$r_s = 0.095, p = 0.001$	$r_s = -0.008, p = 0.77$
		p	$r_s = 0.536, p = 0.0$	$r_s = 0.538, p = 0.0$	$r_s = -0.02, p = 0.483$	$r_s = -0.024, p = 0.409$

Table 1: Table of Spearman correlation with dimensionality (d), the number of DAGs in the MEC (m), the number of edges (k) and the proportion of edges over edges and arcs (p).

			SHD	FN	AP	RP
GSF	Linear	$t = 20$	$\mu = 3.303, \sigma = 3.107$	$\mu = 1.51, \sigma = 1.012$	$\mu = 0.957, \sigma = 0.096$	$\mu = 0.58, \sigma = 0.334$
		$t = 40$	$\mu = 2.582, \sigma = 2.623$	$\mu = 1.268, \sigma = 0.987$	$\mu = 0.966, \sigma = 0.084$	$\mu = 0.664, \sigma = 0.322$
		$t = 60$	$\mu = 2.427, \sigma = 2.509$	$\mu = 1.222, \sigma = 0.967$	$\mu = 0.965, \sigma = 0.089$	$\mu = 0.679, \sigma = 0.318$
		$t = 80$	$\mu = 2.319, \sigma = 2.55$	$\mu = 1.157, \sigma = 0.99$	$\mu = 0.968, \sigma = 0.082$	$\mu = 0.703, \sigma = 0.313$
		$t = 100$	$\mu = 2.227, \sigma = 2.517$	$\mu = 1.117, \sigma = 0.989$	$\mu = 0.97, \sigma = 0.08$	$\mu = 0.718, \sigma = 0.305$
	Non-linear	$t = 20$	$\mu = 3.222, \sigma = 3.049$	$\mu = 1.478, \sigma = 1.019$	$\mu = 0.959, \sigma = 0.093$	$\mu = 0.59, \sigma = 0.335$
		$t = 40$	$\mu = 2.569, \sigma = 2.602$	$\mu = 1.266, \sigma = 0.983$	$\mu = 0.966, \sigma = 0.082$	$\mu = 0.664, \sigma = 0.324$
		$t = 60$	$\mu = 2.371, \sigma = 2.536$	$\mu = 1.193, \sigma = 0.974$	$\mu = 0.966, \sigma = 0.09$	$\mu = 0.684, \sigma = 0.322$
		$t = 80$	$\mu = 2.391, \sigma = 2.6$	$\mu = 1.187, \sigma = 0.991$	$\mu = 0.967, \sigma = 0.083$	$\mu = 0.69, \sigma = 0.317$
		$t = 100$	$\mu = 2.148, \sigma = 2.523$	$\mu = 1.077, \sigma = 0.994$	$\mu = 0.97, \sigma = 0.079$	$\mu = 0.725, \sigma = 0.308$
PCMCI$^+$	Linear	$t = 20$	$\mu = 18.914, \sigma = 12.899$	$\mu = 4.031, \sigma = 1.633$	$\mu = 0.88, \sigma = 0.077$	$\mu = 0.517, \sigma = 0.121$
		$t = 40$	$\mu = 20.924, \sigma = 15.066$	$\mu = 4.191, \sigma = 1.832$	$\mu = 0.88, \sigma = 0.068$	$\mu = 0.574, \sigma = 0.152$
		$t = 60$	$\mu = 21.335, \sigma = 15.724$	$\mu = 4.211, \sigma = 1.898$	$\mu = 0.883, \sigma = 0.062$	$\mu = 0.613, \sigma = 0.156$
		$t = 80$	$\mu = 21.305, \sigma = 16.322$	$\mu = 4.182, \sigma = 1.953$	$\mu = 0.887, \sigma = 0.057$	$\mu = 0.637, \sigma = 0.161$
		$t = 100$	$\mu = 21.203, \sigma = 16.53$	$\mu = 4.166, \sigma = 1.962$	$\mu = 0.888, \sigma = 0.056$	$\mu = 0.645, \sigma = 0.165$
	Non-linear	$t = 20$	$\mu = 13.809, \sigma = 10.534$	$\mu = 3.416, \sigma = 1.463$	$\mu = 0.903, \sigma = 0.076$	$\mu = 0.51, \sigma = 0.086$
		$t = 40$	$\mu = 14.269, \sigma = 11.924$	$\mu = 3.419, \sigma = 1.606$	$\mu = 0.908, \sigma = 0.071$	$\mu = 0.557, \sigma = 0.127$
		$t = 60$	$\mu = 14.153, \sigma = 12.331$	$\mu = 3.385, \sigma = 1.642$	$\mu = 0.909, \sigma = 0.069$	$\mu = 0.583, \sigma = 0.137$
		$t = 80$	$\mu = 13.75, \sigma = 12.586$	$\mu = 3.311, \sigma = 1.67$	$\mu = 0.913, \sigma = 0.065$	$\mu = 0.604, \sigma = 0.149$
		$t = 100$	$\mu = 13.21, \sigma = 12.356$	$\mu = 3.231, \sigma = 1.664$	$\mu = 0.916, \sigma = 0.065$	$\mu = 0.612, \sigma = 0.148$

Table 2: Table of line height statistics per metric score, with rows corresponding to sample sizes

	SHD	FN	AP	RP
Linear GSF	$\mu = 2.572, \sigma = 2.078$	$\mu = 1.255, \sigma = 0.744$	$\mu = 0.965, \sigma = 0.067$	$\mu = 0.669, \sigma = 0.194$
Linear PCMCI$^+$	$\mu = 20.736, \sigma = 14.879$	$\mu = 4.156, \sigma = 1.815$	$\mu = 0.883, \sigma = 0.054$	$\mu = 0.597, \sigma = 0.11$
Non-linear GSF	$\mu = 2.54, \sigma = 2.074$	$\mu = 1.24, \sigma = 0.74$	$\mu = 0.965, \sigma = 0.066$	$\mu = 0.671, \sigma = 0.197$
Non-linear PCMCI$^+$	$\mu = 13.838, \sigma = 11.57$	$\mu = 3.352, \sigma = 1.569$	$\mu = 0.91, \sigma = 0.064$	$\mu = 0.573, \sigma = 0.1$

Table 3: Table of line height statistics per metric score, averaged over sample size parameters.

	SHD	FN	AP	RP
Linear GSF/Linear PCMCI$^+$	$\mu = 18.208, \sigma = 13.928$	$\mu = 2.921, \sigma = 1.564$	$\mu = 0.09, \sigma = 0.044$	$\mu = 0.149, \sigma = 0.124$
	$\delta_{KS} = 0.725, p = 0.0$	$\delta_{KS} = 0.728, p = 0.0$	$\delta_{KS} = 0.851, p = 0.0$	$\delta_{KS} = 0.315, p = 0.0$
Linear GSF/Non-linear GSF	$\mu = 0.621, \sigma = 0.978$	$\mu = 0.241, \sigma = 0.352$	$\mu = 0.006, \sigma = 0.017$	$\mu = 0.074, \sigma = 0.114$
	$\delta_{KS} = 0.017, p = 0.893$	$\delta_{KS} = 0.023, p = 0.531$	$\delta_{KS} = 0.015, p = 0.961$	$\delta_{KS} = 0.018, p = 0.815$
Non-linear GSF/Non-linear PCMCI$^+$	$\mu = 11.362, \sigma = 10.932$	$\mu = 2.138, \sigma = 1.406$	$\mu = 0.064, \sigma = 0.052$	$\mu = 0.169, \sigma = 0.129$
	$\delta_{KS} = 0.634, p = 0.0$	$\delta_{KS} = 0.647, p = 0.0$	$\delta_{KS} = 0.644, p = 0.0$	$\delta_{KS} = 0.41, p = 0.0$
Linear PCMCI$^+$/Non-linear PCMCI$^+$	$\mu = 6.957, \sigma = 9.422$	$\mu = 0.824, \sigma = 0.983$	$\mu = 0.03, \sigma = 0.03$	$\mu = 0.045, \sigma = 0.067$
	$\delta_{KS} = 0.227, p = 0.0$	$\delta_{KS} = 0.226, p = 0.0$	$\delta_{KS} = 0.366, p = 0.0$	$\delta_{KS} = 0.16, p = 0.0$

Table 4: Table of line distance and KS test statistics, averaged over sample size parameters.

	SHD	FN	AP	RP
Linear GSF	$r_s = -0.124, p = 0.0$	$r_s = -0.124, p = 0.0$	$r_s = 0.115, p = 0.0$	$r_s = 0.141, p = 0.0$
Linear PCMCI$^+$	$r_s = 0.021, p = 0.019$	$r_s = 0.021, p = 0.019$	$r_s = -0.005, p = 0.594$	$r_s = 0.308, p = 0.0$
Non-linear GSF	$r_s = -0.123, p = 0.0$	$r_s = -0.123, p = 0.0$	$r_s = 0.111, p = 0.0$	$r_s = 0.133, p = 0.0$
Non-linear PCMCI$^+$	$r_s = -0.053, p = 0.0$	$r_s = -0.053, p = 0.0$	$r_s = 0.058, p = 0.0$	$r_s = 0.297, p = 0.0$

Table 5: Table of Pearson correlation of metric scores with sample size parameters.

	SHD	FN	AP	RP
Linear GSF	$r_s = 0.31, p = 0.0$	$r_s = 0.313, p = 0.0$	$r_s = -0.224, p = 0.0$	$r_s = -0.222, p = 0.0$
Linear PCMCI$^+$	$r_s = 0.12, p = 0.0$	$r_s = 0.122, p = 0.0$	$r_s = -0.301, p = 0.0$	$r_s = -0.474, p = 0.0$
Non-linear GSF	$r_s = 0.3, p = 0.0$	$r_s = 0.302, p = 0.0$	$r_s = -0.214, p = 0.0$	$r_s = -0.209, p = 0.0$
Non-linear PCMCI$^+$	$r_s = 0.182, p = 0.0$	$r_s = 0.188, p = 0.0$	$r_s = -0.24, p = 0.0$	$r_s = -0.398, p = 0.0$

Table 6: Table of Spearman correlation of metric scores with the proportion of instantaneous causes.

	SHD	FN	AP	RP
Linear GSF	$r_s = -0.31, p = 0.0$	$r_s = -0.313, p = 0.0$	$r_s = 0.224, p = 0.0$	$r_s = 0.222, p = 0.0$
Linear PCMCI$^+$	$r_s = -0.12, p = 0.0$	$r_s = -0.122, p = 0.0$	$r_s = 0.301, p = 0.0$	$r_s = 0.474, p = 0.0$
Non-linear GSF	$r_s = -0.3, p = 0.0$	$r_s = -0.302, p = 0.0$	$r_s = 0.214, p = 0.0$	$r_s = 0.209, p = 0.0$
Non-linear PCMCI$^+$	$r_s = -0.182, p = 0.0$	$r_s = -0.188, p = 0.0$	$r_s = 0.24, p = 0.0$	$r_s = 0.398, p = 0.0$

Table 7: Table of Spearman correlation of metric scores with the proportion of lagged causes.

G Results Experiment II

G.1 Plots

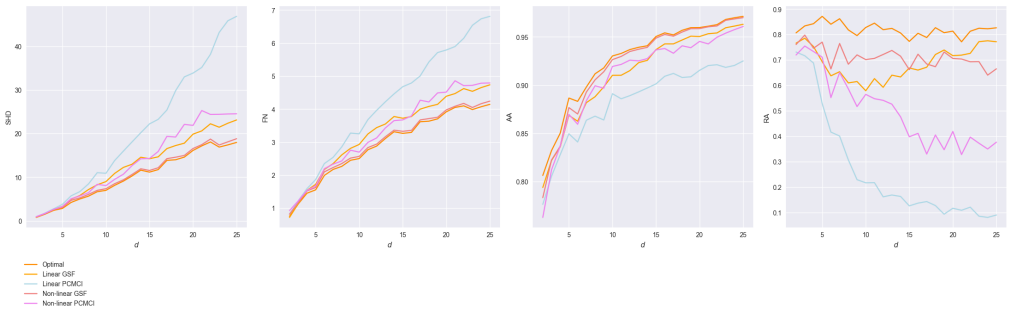


Figure 19: Plots of dimensionality values d against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores.

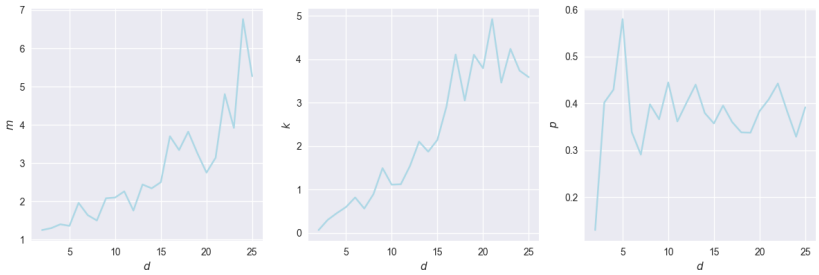


Figure 20: Plots of dimensionality values d against the number of DAGs in the MEC (m), the number of edges (k) and the proportion of edges over edges and arcs (p).

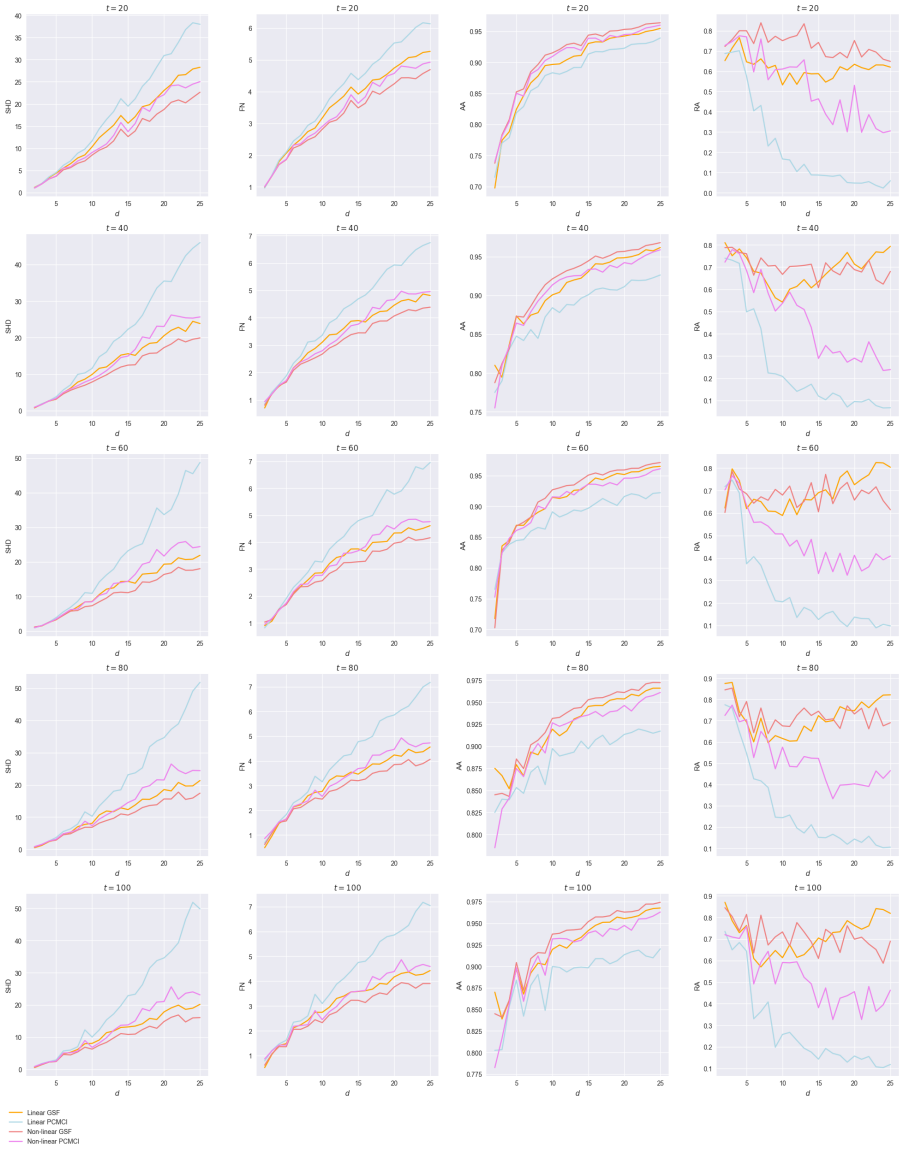


Figure 21: Plots of dimensionality values d against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores, with rows corresponding to specific sample size parameters.

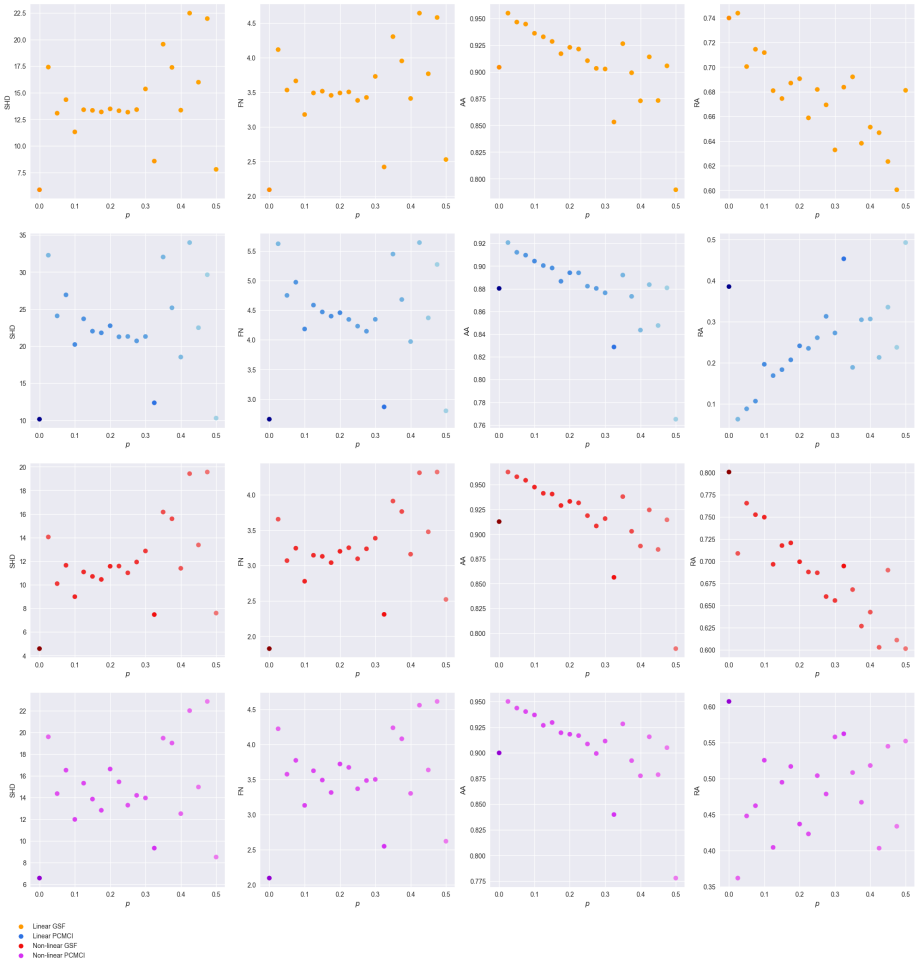


Figure 22: Plots of instantaneous link proportion values against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores, with rows corresponding to causal discovery methods.

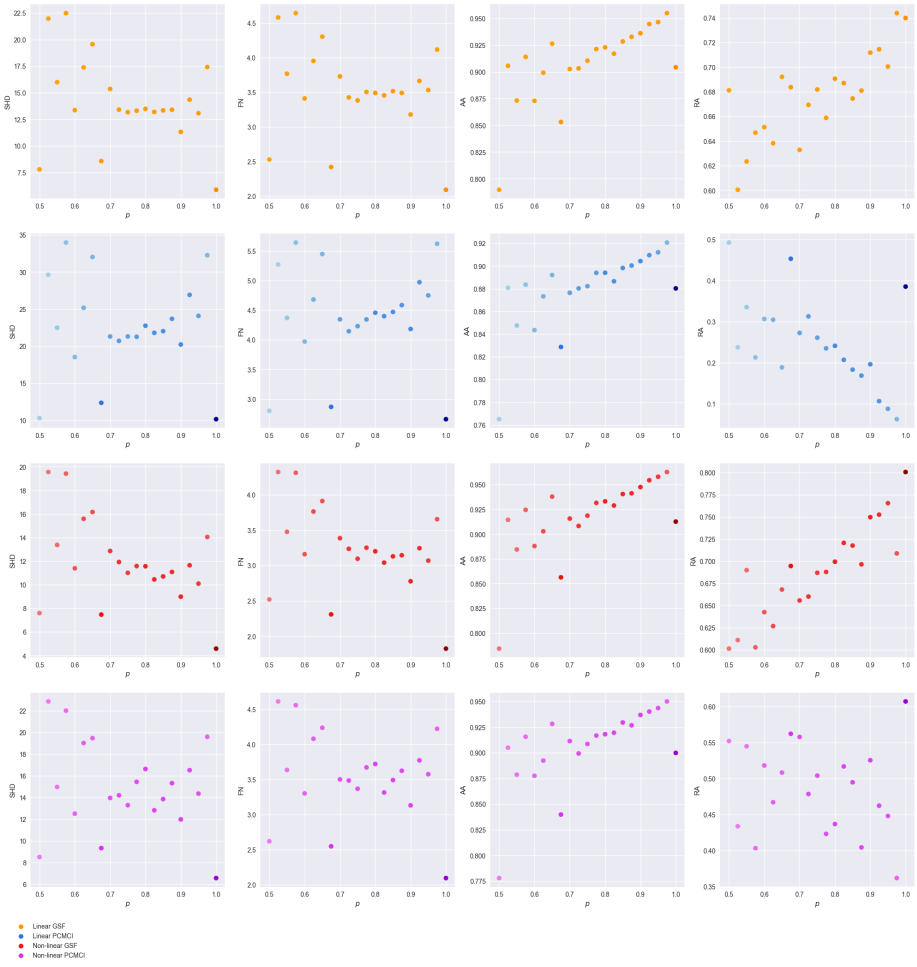


Figure 23: Plots of lagged link proportion values against Structural Hamming Distance (SHD), Frobenius norm (FN), absolute accuracy (AA) and relative accuracy (RA) scores, with rows corresponding to causal discovery methods.

G.2 Descriptive Statistics

			SHD	FN	AP	RP
GSF	Linear	d	$r_s = 0.882, p = 0.0$	$r_s = 0.879, p = 0.0$	$r_s = 0.818, p = 0.0$	$r_s = 0.167, p = 0.0$
		m	$r_s = 0.381, p = 0.0$	$r_s = 0.381, p = 0.0$	$r_s = 0.201, p = 0.0$	$r_s = -0.091, p = 0.0$
		k	$r_s = 0.451, p = 0.0$	$r_s = 0.45, p = 0.0$	$r_s = 0.39, p = 0.0$	$r_s = 0.083, p = 0.0$
		p	$r_s = 0.105, p = 0.0$	$r_s = 0.105, p = 0.0$	$r_s = 0.171, p = 0.0$	$r_s = 0.114, p = 0.0$
	Non-linear	d	$r_s = 0.798, p = 0.0$	$r_s = 0.794, p = 0.0$	$r_s = 0.734, p = 0.0$	$r_s = -0.154, p = 0.0$
		m	$r_s = 0.417, p = 0.0$	$r_s = 0.416, p = 0.0$	$r_s = 0.114, p = 0.0$	$r_s = -0.466, p = 0.0$
		k	$r_s = 0.563, p = 0.0$	$r_s = 0.557, p = 0.0$	$r_s = 0.204, p = 0.0$	$r_s = -0.159, p = 0.0$
		p	$r_s = 0.113, p = 0.0$	$r_s = 0.108, p = 0.0$	$r_s = 0.142, p = 0.0$	$r_s = 0.189, p = 0.0$
PCMCI$^+$	Linear	d	$r_s = 0.977, p = 0.0$	$r_s = 0.976, p = 0.0$	$r_s = 0.721, p = 0.0$	$r_s = -0.719, p = 0.0$
		m	$r_s = 0.37, p = 0.0$	$r_s = 0.371, p = 0.0$	$r_s = 0.166, p = 0.0$	$r_s = -0.115, p = 0.0$
		k	$r_s = 0.477, p = 0.0$	$r_s = 0.476, p = 0.0$	$r_s = 0.346, p = 0.0$	$r_s = -0.411, p = 0.0$
		p	$r_s = 0.125, p = 0.0$	$r_s = 0.124, p = 0.0$	$r_s = 0.184, p = 0.0$	$r_s = -0.237, p = 0.0$
	Non-linear	d	$r_s = 0.797, p = 0.0$	$r_s = 0.793, p = 0.0$	$r_s = 0.649, p = 0.0$	$r_s = -0.451, p = 0.0$
		m	$r_s = 0.422, p = 0.0$	$r_s = 0.423, p = 0.0$	$r_s = 0.064, p = 0.002$	$r_s = -0.457, p = 0.0$
		k	$r_s = 0.685, p = 0.0$	$r_s = 0.681, p = 0.0$	$r_s = -0.005, p = 0.794$	$r_s = -0.664, p = 0.0$
		p	$r_s = 0.136, p = 0.0$	$r_s = 0.13, p = 0.0$	$r_s = 0.114, p = 0.0$	$r_s = -0.026, p = 0.198$

Table 8: Table of Spearman correlation with dimensionality (d), the number of DAGs in the MEC (m), the number of edges (k) and the proportion of edges over edges and arcs (p).

			SHD	FN	AP	RP
GSF	Linear	$t = 20$	$\mu = 14.947, \sigma = 9.956$	$\mu = 3.605, \sigma = 1.395$	$\mu = 0.894, \sigma = 0.086$	$\mu = 0.618, \sigma = 0.21$
		$t = 40$	$\mu = 13.273, \sigma = 8.99$	$\mu = 3.378, \sigma = 1.365$	$\mu = 0.909, \sigma = 0.07$	$\mu = 0.691, \sigma = 0.223$
		$t = 60$	$\mu = 12.234, \sigma = 8.153$	$\mu = 3.251, \sigma = 1.291$	$\mu = 0.912, \sigma = 0.079$	$\mu = 0.699, \sigma = 0.225$
		$t = 80$	$\mu = 11.617, \sigma = 7.891$	$\mu = 3.15, \sigma = 1.303$	$\mu = 0.923, \sigma = 0.059$	$\mu = 0.719, \sigma = 0.219$
		$t = 100$	$\mu = 11.288, \sigma = 7.731$	$\mu = 3.1, \sigma = 1.295$	$\mu = 0.924, \sigma = 0.06$	$\mu = 0.716, \sigma = 0.223$
	Non-linear	$t = 20$	$\mu = 12.09, \sigma = 8.618$	$\mu = 3.226, \sigma = 1.298$	$\mu = 0.91, \sigma = 0.075$	$\mu = 0.733, \sigma = 0.233$
		$t = 40$	$\mu = 11.098, \sigma = 7.966$	$\mu = 3.073, \sigma = 1.287$	$\mu = 0.919, \sigma = 0.071$	$\mu = 0.702, \sigma = 0.243$
		$t = 60$	$\mu = 10.428, \sigma = 7.478$	$\mu = 2.982, \sigma = 1.238$	$\mu = 0.919, \sigma = 0.079$	$\mu = 0.682, \sigma = 0.313$
		$t = 80$	$\mu = 9.766, \sigma = 7.18$	$\mu = 2.863, \sigma = 1.253$	$\mu = 0.929, \sigma = 0.063$	$\mu = 0.724, \sigma = 0.303$
		$t = 100$	$\mu = 9.402, \sigma = 7.104$	$\mu = 2.793, \sigma = 1.266$	$\mu = 0.932, \sigma = 0.063$	$\mu = 0.713, \sigma = 0.312$
PCMCI$^+$	Linear	$t = 20$	$\mu = 18.89, \sigma = 12.868$	$\mu = 4.028, \sigma = 1.634$	$\mu = 0.88, \sigma = 0.08$	$\mu = 0.222, \sigma = 0.316$
		$t = 40$	$\mu = 20.932, \sigma = 15.124$	$\mu = 4.188, \sigma = 1.842$	$\mu = 0.881, \sigma = 0.064$	$\mu = 0.249, \sigma = 0.245$
		$t = 60$	$\mu = 21.211, \sigma = 15.679$	$\mu = 4.198, \sigma = 1.895$	$\mu = 0.883, \sigma = 0.065$	$\mu = 0.252, \sigma = 0.263$
		$t = 80$	$\mu = 21.184, \sigma = 16.199$	$\mu = 4.172, \sigma = 1.943$	$\mu = 0.888, \sigma = 0.055$	$\mu = 0.242, \sigma = 0.272$
		$t = 100$	$\mu = 21.243, \sigma = 16.497$	$\mu = 4.168, \sigma = 1.967$	$\mu = 0.888, \sigma = 0.056$	$\mu = 0.24, \sigma = 0.264$
	Non-linear	$t = 20$	$\mu = 13.622, \sigma = 10.245$	$\mu = 3.398, \sigma = 1.441$	$\mu = 0.904, \sigma = 0.074$	$\mu = 0.524, \sigma = 0.421$
		$t = 40$	$\mu = 14.049, \sigma = 11.564$	$\mu = 3.401, \sigma = 1.576$	$\mu = 0.907, \sigma = 0.071$	$\mu = 0.464, \sigma = 0.377$
		$t = 60$	$\mu = 13.78, \sigma = 11.813$	$\mu = 3.351, \sigma = 1.597$	$\mu = 0.909, \sigma = 0.07$	$\mu = 0.48, \sigma = 0.355$
		$t = 80$	$\mu = 13.365, \sigma = 12.042$	$\mu = 3.275, \sigma = 1.624$	$\mu = 0.913, \sigma = 0.066$	$\mu = 0.516, \sigma = 0.346$
		$t = 100$	$\mu = 12.862, \sigma = 11.88$	$\mu = 3.196, \sigma = 1.627$	$\mu = 0.916, \sigma = 0.067$	$\mu = 0.518, \sigma = 0.336$

Table 9: Table of line height statistics per metric score, with rows corresponding to sample sizes.

	SHD	FN	AP	RP
Linear GSF	$\mu = 12.672, \sigma = 8.185$	$\mu = 3.297, \sigma = 1.28$	$\mu = 0.912, \sigma = 0.059$	$\mu = 0.688, \sigma = 0.122$
Linear PCMCI$^+$	$\mu = 20.692, \sigma = 14.851$	$\mu = 4.151, \sigma = 1.816$	$\mu = 0.884, \sigma = 0.053$	$\mu = 0.257, \sigma = 0.235$
Non-linear GSF	$\mu = 10.557, \sigma = 7.429$	$\mu = 2.987, \sigma = 1.23$	$\mu = 0.922, \sigma = 0.063$	$\mu = 0.711, \sigma = 0.167$
Non-linear PCMCI$^+$	$\mu = 13.536, \sigma = 11.121$	$\mu = 3.324, \sigma = 1.529$	$\mu = 0.91, \sigma = 0.065$	$\mu = 0.5, \sigma = 0.24$
Optimal	$\mu = 10.166, \sigma = 7.247$	$\mu = 2.921, \sigma = 1.228$	$\mu = 0.926, \sigma = 0.06$	$\mu = 0.819, \sigma = 0.118$

Table 10: Table of line height statistics per metric score, averaged over sample size parameters.

	SHD	FN	AP	RP
Linear GSF/Linear PCMCI$^+$	$\mu = 8.039, \sigma = 8.134$ $\delta_{KS} = 0.3, p = 0.0$	$\mu = 0.859, \sigma = 0.744$ $\delta_{KS} = 0.301, p = 0.0$	$\mu = 0.03, \sigma = 0.017$ $\delta_{KS} = 0.428, p = 0.0$	$\mu = 0.438, \sigma = 0.237$ $\delta_{KS} = 0.765, p = 0.0$
Linear GSF/Non-linear GSF	$\mu = 2.22, \sigma = 2.605$ $\delta_{KS} = 0.138, p = 0.0$	$\mu = 0.348, \sigma = 0.4$ $\delta_{KS} = 0.136, p = 0.0$	$\mu = 0.015, \sigma = 0.022$ $\delta_{KS} = 0.161, p = 0.0$	$\mu = 0.114, \sigma = 0.136$ $\delta_{KS} = 0.155, p = 0.0$
Non-linear GSF/Non-linear PCMCI$^+$	$\mu = 3.051, \sigma = 5.789$ $\delta_{KS} = 0.132, p = 0.0$	$\mu = 0.35, \sigma = 0.58$ $\delta_{KS} = 0.13, p = 0.0$	$\mu = 0.013, \sigma = 0.018$ $\delta_{KS} = 0.156, p = 0.0$	$\mu = 0.239, \sigma = 0.225$ $\delta_{KS} = 0.401, p = 0.0$
Linear PCMCI$^+$/Non-linear PCMCI$^+$	$\mu = 7.225, \sigma = 9.534$ $\delta_{KS} = 0.233, p = 0.0$	$\mu = 0.853, \sigma = 0.996$ $\delta_{KS} = 0.235, p = 0.0$	$\mu = 0.031, \sigma = 0.029$ $\delta_{KS} = 0.378, p = 0.0$	$\mu = 0.25, \sigma = 0.268$ $\delta_{KS} = 0.439, p = 0.0$
Linear GSF/Optimal	$\mu = 2.583, \sigma = 2.751$ $\delta_{KS} = 0.151, p = 0.0$	$\mu = 0.407, \sigma = 0.412$ $\delta_{KS} = 0.152, p = 0.0$	$\mu = 0.019, \sigma = 0.025$ $\delta_{KS} = 0.185, p = 0.0$	$\mu = 0.146, \sigma = 0.141$ $\delta_{KS} = 0.422, p = 0.0$
Linear PCMCI$^+$/Optimal	$\mu = 10.571, \sigma = 9.634$ $\delta_{KS} = 0.373, p = 0.0$	$\mu = 1.251, \sigma = 0.947$ $\delta_{KS} = 0.373, p = 0.0$	$\mu = 0.046, \sigma = 0.028$ $\delta_{KS} = 0.553, p = 0.0$	$\mu = 0.565, \sigma = 0.243$ $\delta_{KS} = 0.856, p = 0.0$
Non-linear GSF/Optimal	$\mu = 0.39, \sigma = 0.599$ $\delta_{KS} = 0.025, p = 0.441$	$\mu = 0.066, \sigma = 0.102$ $\delta_{KS} = 0.026, p = 0.4$	$\mu = 0.005, \sigma = 0.014$ $\delta_{KS} = 0.033, p = 0.148$	$\mu = 0.108, \sigma = 0.167$ $\delta_{KS} = 0.242, p = 0.0$
Non-linear PCMCI$^+$/Optimal	$\mu = 3.375, \sigma = 5.787$ $\delta_{KS} = 0.147, p = 0.0$	$\mu = 0.404, \sigma = 0.578$ $\delta_{KS} = 0.145, p = 0.0$	$\mu = 0.017, \sigma = 0.024$ $\delta_{KS} = 0.184, p = 0.0$	$\mu = 0.32, \sigma = 0.222$ $\delta_{KS} = 0.598, p = 0.0$

Table 11: Table of line distance and KS test statistics, averaged over sample size parameters.

	SHD	FN	AP	RP
Linear GSF	$r_s = -0.124, p = 0.0$	$r_s = -0.124, p = 0.0$	$r_s = 0.172, p = 0.0$	$r_s = 0.168, p = 0.0$
Linear PCMCI$^+$	$r_s = 0.022, p = 0.017$	$r_s = 0.022, p = 0.017$	$r_s = -0.003, p = 0.726$	$r_s = 0.172, p = 0.0$
Non-linear GSF	$r_s = -0.108, p = 0.0$	$r_s = -0.108, p = 0.0$	$r_s = 0.137, p = 0.0$	$r_s = 0.054, p = 0.0$
Non-linear PCMCI$^+$	$r_s = -0.057, p = 0.0$	$r_s = -0.057, p = 0.0$	$r_s = 0.062, p = 0.0$	$r_s = 0.042, p = 0.0$

Table 12: Table of Pearson correlation of metric scores with sample size parameters.

	SHD	FN	AP	RP
Linear GSF	$r_s = 0.235, p = 0.0$	$r_s = 0.238, p = 0.0$	$r_s = -0.276, p = 0.0$	$r_s = -0.241, p = 0.0$
Linear PCMCI$^+$	$r_s = 0.115, p = 0.0$	$r_s = 0.118, p = 0.0$	$r_s = -0.296, p = 0.0$	$r_s = 0.259, p = 0.0$
Non-linear GSF	$r_s = 0.27, p = 0.0$	$r_s = 0.275, p = 0.0$	$r_s = -0.29, p = 0.0$	$r_s = -0.291, p = 0.0$
Non-linear PCMCI$^+$	$r_s = 0.187, p = 0.0$	$r_s = 0.194, p = 0.0$	$r_s = -0.224, p = 0.0$	$r_s = -0.054, p = 0.008$

Table 13: Table of Spearman correlation of metric scores with the proportion of instantaneous causes.

	SHD	FN	AP	RP
Linear GSF	$r_s = -0.235, p = 0.0$	$r_s = -0.238, p = 0.0$	$r_s = 0.276, p = 0.0$	$r_s = 0.241, p = 0.0$
Linear PCMCI$^+$	$r_s = -0.115, p = 0.0$	$r_s = -0.118, p = 0.0$	$r_s = 0.296, p = 0.0$	$r_s = -0.259, p = 0.0$
Non-linear GSF	$r_s = -0.27, p = 0.0$	$r_s = -0.275, p = 0.0$	$r_s = 0.29, p = 0.0$	$r_s = 0.291, p = 0.0$
Non-linear PCMCI$^+$	$r_s = -0.187, p = 0.0$	$r_s = -0.194, p = 0.0$	$r_s = 0.224, p = 0.0$	$r_s = 0.054, p = 0.008$

Table 14: Table of Spearman correlation of metric scores with the proportion of lagged causes.

H References

[1] Murat Akkaya. "Vector Autoregressive Model and Analysis". In: *Handbook of Research on Emerging Theories, Models, and Applications of Financial Econometrics*. Springer, 2021, pp. 197–214.

[2] Charles K. Assaad, Emilie Devijver, and Eric Gaussier. "Discovery of Extended Summary Graphs in Time Series". In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 96–106.

[3] Charles K. Assaad, Emilie Devijver, and Eric Gaussier. "Survey and Evaluation of Causal Discovery Methods for Time Series". In: *Journal of Artificial Intelligence Research* 73 (2022), pp. 767–819.

[4] Allan Birnbaum. "Combining Independent Tests of Significance". In: *Journal of the American Statistical Association* 49.267 (1954), pp. 559–574.

[5] Z. William Birnbaum. "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size". In: *Journal of the American Statistical Association* 47.259 (1952), pp. 425–441.

[6] M.F. Brilhante, D. Pestana, and F. Sequeira. "Combining p-values and Random p-values". In: *Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces*. IEEE. 2010, pp. 515–520.

[7] Philippe Brouillard et al. "Differentiable Causal Discovery from Interventional Data". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 21865–21877.

[8] Nancy Cartwright. "Causal Diversity and the Markov Condition". In: *Synthese* 121.1/2 (1999), pp. 3–27.

[9] Alexandra M. Carvalho. "Scoring Functions for Learning Bayesian Networks". In: *INESC-ID Tec. Rep* 12 (2009), pp. 1–48.

[10] Chris Chatfield. *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC, 2003.

[11] Pu Chen. "A Time Series Causal Model". In: *University Library of Munich, Germany, MPRA Paper* (Jan. 2010).

[12] Lu Cheng et al. "Evaluation Methods and Measures for Causal Learning Algorithms". In: *CoRR* abs/2202.02896 (2022). arXiv: 2202.02896.

[13] David Maxwell Chickering. "A Transformational Characterization of Equivalent Bayesian Network Structures". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 87–98. ISBN: 1558603859.

[14] David Maxwell Chickering. "Learning Equivalence Classes of Bayesian Network Structures". In: *The Journal of Machine Learning Research* 2 (2002), pp. 445–498.

[15] Gregory W. Corder and Dale I. Foreman. *Nonparametric Statistics for Non-Statisticians*. John Wiley & Sons, Inc., 2011.

[16] Madalina Croitoru et al. *Graph Structures for Knowledge Representation and Reasoning*. Springer, 2018.

[17] Jonathan D. Cryer. *Time Series Analysis*. Vol. 286. Springer, 1986.

[18] Peng Cui et al. "Causal Inference Meets Machine Learning". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 3527–3528.

[19] Rainer Dahlhaus and Michael Eichler. "Causality and Graphical Models in Time Series Analysis". In: *Oxford Statistical Science Series* (2003), pp. 115–137.

[20] Marek J. Druzdzel. "The Role of Assumptions in Causal Discovery". In: *8th Workshop on Uncertainty Processing (WUPES-09)*. Sept. 2009, pp. 57–68.

[21] Frederick Eberhardt. "Introduction to the Foundations of Causal Discovery". In: *International Journal of Data Science and Analytics* 3 (2016), pp. 81–91.

[22] Michael Eichler. *Causal Inference in Time Series Analysis*. Wiley Online Library, 2012.

[23] Michael Eichler. "Causal Inference with Multiple Time Series: Principles and Problems". In: *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 371 (July 2013), p. 20110613.

[24] Michael Eichler. "Graphical Modelling of Multivariate Time Series". In: *arXiv preprint math/0610654* (2006).

[25] Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Springer, 1992.

[26] J. P. Florens and M. Mouchart. “A Note on Noncausality”. In: *Econometrica* 50.3 (1982), pp. 583–591. ISSN: 00129682, 14680262.

[27] Christopher J. Fox, Andreas Käufl, and Mathias Drton. “On the Causal Interpretation of Acyclic Mixed Graphs Under Multivariate Normality”. In: *Linear Algebra and its Applications* 473 (2015). Special Issue on Statistics, pp. 93–113. ISSN: 0024-3795.

[28] Karl J. Friston et al. “Granger Causality Revisited”. In: *Neuroimage* 101 (2014), pp. 796–808.

[29] Steven B. Gillispie and Christiane Lemieux. “Enumerating Markov Equivalence Classes of Acyclic Digraph Models”. In: *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. UAI ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 171–177. ISBN: 1558608001.

[30] Clark Glymour, Kun Zhang, and Peter Spirtes. “Review of Causal Discovery Methods Based on Graphical Models”. In: *Frontiers in Genetics* 10 (2019), pp. 1–15.

[31] Clive W.J. Granger. “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods”. In: *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438.

[32] Clive W.J. Granger. “Some Recent Development in a Concept of Causality”. In: *Journal of Econometrics* 39.1-2 (1988), pp. 199–211.

[33] Clive W.J. Granger. “Testing for Causality: A Personal Viewpoint”. In: *Journal of Economic Dynamics and Control* 2 (1980), pp. 329–352. ISSN: 0165-1889.

[34] Ruocheng Guo et al. “A Survey of Learning Causality with Data: Problems and Methods”. In: *ACM Computing Surveys (CSUR)* 53.4 (2020), pp. 1–37.

[35] Abdulnasser Hatemi-J. “Asymmetric Causality Tests with an Application”. In: *Empirical Economics* 43.1 (2012), pp. 447–456.

[36] Daniel Hausman and James Woodward. “Manipulation and the Causal Markov Condition”. In: *Philosophy of Science* 71.5 (2004), pp. 846–856.

[37] Nicholas A. Heard and Patrick Rubin-Delanchy. “Choosing Between Methods of Combining p-values”. In: *Biometrika* 105.1 (2018), pp. 239–246.

[38] David Heckerman, Christopher Meek, and Gregory Cooper. "A Bayesian Approach to Causal Discovery". In: *Innovations in Machine Learning: Theory and Applications*. Ed. by Dawn E. Holmes and Lakhmi C. Jain. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–28. ISBN: 978-3-540-33486-6.

[39] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. "Causal Structure Learning". In: *Annual Review of Statistics and Its Application* 5 (2018), pp. 371–391.

[40] Christopher Hitchcock. "Causal Models". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2022. Metaphysics Research Lab, Stanford University, 2022.

[41] Patrik Hoyer et al. "Nonlinear Causal Discovery With Additive Noise Models". In: *Advances in Neural Information Processing Systems* 21 (2008).

[42] Raymond Hubbard and R. Murray Lindsay. "Why P-Values are not a Useful Measure of Evidence in Statistical Significance Testing". In: *Theory & Psychology* 18.1 (2008), pp. 69–88.

[43] Finn V. Jensen. "Bayesian Networks". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.3 (2009), pp. 307–315.

[44] Marcus Kaiser and Maksim Sipos. "Unsuitability of NOTEARS for Causal Graph Discovery when Dealing with Dimensional Quantities". In: *Neural Processing Letters* 54.3 (2022), pp. 1587–1595.

[45] James T. Kost and Michael P. McDermott. "Combining Dependent P-values". In: *Statistics & Probability Letters* 60.2 (2002), pp. 183–190.

[46] Pascal Lavergne and Valentin Patilea. "Breaking the Curse of Dimensionality in Nonparametric Testing". In: *Journal of Econometrics* 143.1 (2008), pp. 103–122.

[47] Zhifa Liu, Brandon Malone, and Changhe Yuan. "Empirical Evaluation of Scoring Functions for Bayesian Network Model Selection". In: *BMC bioinformatics*. Vol. 13. 15. Springer. 2012, pp. 1–16.

[48] Yunan Luo, Jian Peng, and Jianzhu Ma. "When Causal Inference Meets Deep Learning". In: *Nature Machine Intelligence* 2.8 (2020), pp. 426–427.

[49] Helmut Lütkepohl. "Vector Autoregressive Models". In: *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing, 2013, pp. 139–164.

[50] Marloes Maathuis et al. *Handbook of Graphical Models*. CRC Press, 2018.

[51] David P. MacKinnon. *Introduction to Statistical Mediation Analysis*. Routledge, 2012.

[52] Mariusz Maziarz. "A Review of the Granger-Causality Fallacy". In: *The Journal of Philosophical Economics: Reflections on Economic and Social Issues* 8.2 (2015), pp. 86–105.

[53] Christopher Meek. "Graphical Models: Selecting Causal and Statistical Models". PhD thesis. Carnegie Mellon University, 1997.

[54] Tim Miller. "Contrastive Explanation: A Structural-Model Approach". In: *The Knowledge Engineering Review* 36 (2021), pp. 1–24.

[55] Raha Moraffah et al. "Causal Inference for Time Series Analysis: Problems, Methods and Evaluation". In: *Knowledge and Information Systems* (2021), pp. 1–45.

[56] Raha Moraffah et al. "Causal Interpretability for Machine Learning - Problems, Methods and Evaluation". In: *SIGKDD Explor. Newsl.* 22.1 (May 2020), pp. 18–33. ISSN: 1931-0145.

[57] Ana Nogueira et al. "Methods and Tools for Causal Discovery and Causal Inference". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12 (Mar. 2022). DOI: 10.1002/widm.1449.

[58] Christopher Nowzohour and Peter Bühlmann. "Score-based Causal Learning in Additive Noise Models". In: *Statistics* 50.3 (2016), pp. 471–485.

[59] Roxana Pamfil et al. "Dynotears: Structure Learning from Time-Series Data". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1595–1605.

[60] Judea Pearl. *Causal Inference in Statistics: a Primer*. eng. Chichester, West Sussex: Wiley, 2016 - 2016. ISBN: 9781119186854.

[61] Judea Pearl. *Causality. Models, Reasoning, and Inference*. 2nd ed. Cambridge, UK: Cambridge University Press, 2009. ISBN: 978-0-521-89560-6.

[62] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[63] Judea Pearl. "The Seven Tools of Causal Inference, with Reflections on Machine Learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

[64] Jonas Peters and Peter Bühlmann. "Structural Intervention Distance for Evaluating Causal Graphs". In: *Neural Computation* 27.3 (2015), pp. 771–799.

[65] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. "Causal Inference on Time Series Using Restricted Structural Equation Models". In: *Advances in Neural Information Processing Systems* 26 (2013).

[66] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[67] Florin Popescu and Isabelle Guyon. *Causality in Time Series: Challenges in Machine Learning*. 2013.

[68] Jakob Runge. "Causal Network Reconstruction from Time Series: From Theoretical Assumptions to Practical Estimation". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.7 (2018), p. 075310.

[69] Jakob Runge. "Discovering Contemporaneous and Lagged Causal Relations in Autocorrelated Nonlinear Time Series Datasets". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 1388–1397.

[70] Jakob Runge et al. "Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets". In: *Science Advances* 5.11 (Nov. 2019).

[71] Jakob Runge et al. "Inferring Causation from Time Series in Earth System Sciences". In: *Nature Communications* 10.1 (2019), pp. 1–13.

[72] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Series in Artificial Intelligence. Pearson, 2020. ISBN: 9780134610993.

[73] S. K. Sgaier, V. Huang, and G. Charles. "The Case for Causal AI". In: *Stanford Social Innovation Review* 18.3 (2020), pp. 50–55.

[74] Shohei Shimizu. "LiNGAM: Non-Gaussian Methods for Estimating Causal Structures". In: *Behaviormetrika* 41 (2014), pp. 65–98.

[75] Ali Shojaie and Emily B Fox. "Granger Causality: A Review and Recent Advances". In: *Annual Review of Statistics and Its Application* 9 (2022), pp. 289–319.

[76] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and its Applications*. Vol. 3. Springer, 2000.

[77] Elsa Siggiridou et al. "Evaluation of Granger Causality Measures for Constructing Networks from Multivariate Time Series". In: *Entropy* 21.11 (Nov. 2019), pp. 1–26.

[78] Peter Spirtes and Clark Glymour. "An Algorithm for Fast Recovery of Sparse Causal Graphs". In: *Social Science Computer Review* 9.1 (1991), pp. 62–72.

[79] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. 2000.

[80] Peter Spirtes and Kun Zhang. "Causal Discovery and Inference: Concepts and Recent Methodological Advances". In: *Applied Informatics* 3 (Dec. 2016).

[81] James H. Stock and Mark W. Watson. "Vector Autoregressions". In: *Journal of Economic Perspectives* 15.4 (2001), pp. 101–115.

[82] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. "The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm". In: *Machine Learning* 65 (2006), pp. 31–78.

[83] Thomas Verma and Judea Pearl. "Equivalence and Synthesis of Causal Models". In: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. UAI '90. USA: Elsevier Science Inc., 1990, pp. 255–270. ISBN: 0444892648.

[84] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.

[85] Halbert White and Xun Lu. "Granger Causality and Dynamic Structural Systems". In: *Journal of Financial Econometrics* 8.2 (2010), pp. 193–243.

[86] Wolfgang Wiedermann and Alexander Von Eye. *Statistics and Causality*. Wiley Online Library, 2016.

[87] Marcel Wienöbst et al. “Efficient Enumeration of Markov Equivalent DAGs”. In: *arXiv preprint arXiv:2301.12212* (2023).

[88] Daniel J. Wilson. “The Harmonic Mean p-value for Combining Dependent Tests”. In: *Proceedings of the National Academy of Sciences* 116.4 (2019), pp. 1195–1200.

[89] Sungho Won et al. “Choosing an Optimal Method to Combine p-values”. In: *Statistics in Medicine* 28.11 (2009), pp. 1537–1553.

[90] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. “A Survey on Causal Discovery: Theory and Practice”. In: *International Journal of Approximate Reasoning* 151 (2022), pp. 101–129.

[91] Jerrold H Zar. “Significance Testing of the Spearman Rank Correlation Coefficient”. In: *Journal of the American Statistical Association* 67.339 (1972), pp. 578–580.

[92] Hao Zhang et al. “Causal Discovery Using Regression-based Conditional Independence Tests”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[93] Hong Zhang and Zheyang Wu. “The Generalized Fisher’s Combination and Accurate p-value Calculation Under Dependence”. In: *Biometrics* (2022).

[94] Kun Zhang et al. “Kernel-Based Conditional Independence Test and Application in Causal Discovery”. In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. UAI’11. Barcelona, Spain: AUAI Press, 2011, pp. 804–813. ISBN: 9780974903972.