Contents

1 Outline and Setup

15/02/2023	finalise details of Granger scoring function
01/03/2023	finalising specifics for experiment
20/04/2023	finalising code for experiment
01/06/2023	finalising description and analysis
20/06/2023	draft version thesis
30/06/2023	final version thesis

2 Introduction

#TODO: update when experimental section has been written

Causal discovery is the task of learning the causal structure underlying a set of variables from observational data. In this setting, causal structure is interpreted as a set of relations between variables or, equivalently, as a graphical model over variables [12, 35, 18]. In the literature, two types of methods for extracting these structures dominate: constraint-based methods and score-based methods [18, 2, 35]. Constraint-based methods first exploit edge constraints to derive an undirected graph over the variables and afterwards apply orientation rules to return an equivalence class of graphs [2, p. 780]. Score-based methods, on the other hand, start with a scoring function and a set of candidate causal graphs interpreted as Bayesian Networks. After assigning a quality score to each candidate based on the candidate's fit on the data and parameter complexity, the network that optimises the scoring function is returned [2, p. 794].

In the setting of Bayesian Networks, a usual assumption on scoring functions is *score equivalence*: whenever two graphs belong to the same Markov equivalence class, equivalent scores must be assigned to these graphs [32, 3, 29]. Clearly, this assumption is inaccurate in the causal setting: since arc directions are given a causal interpretation, distinct graphs within the same equivalence class represent distinct causal models [23, p. 372]. In this thesis, we develop a causal scoring function that rejects the score equivalence assumption by exploiting bivariate Granger causality as a proxy measure of causality. Since the aim of the scoring function is to distinguish between otherwise score equivalent graphs, we assume that the scoring function is always applied within a Markov equivalence class. Explicitly stated, the research question of this thesis is as follows:

> Can bivariate Granger causality provide a causal scoring function that reliably determines causal structure from otherwise score equivalent graphs?

To answer the research question, we posit the following subquestions:

(A) How can bivariate Granger causality be leveraged to construct a causal scoring function?

(B) How well does the proposed causal scoring function recover true causal structure within the class of linear causal models?

(C) How well does the proposed causal scoring function recover true causal structure within the class of non-linear causal models?

Subquestion (A) builds on a literature survey, resulting in the conclusion that bivariate Granger causality can be exploited to counteract score equivalence. More specifically, p-values from bivariate Granger causality tests are leveraged to construct causal scores over candidate causal graphs. Subquestion (B) and (C) are evaluated by, first, defining structural causal models and generating observational data from those models. Subsequently, the proposed causal scoring function is applied within the equivalence class of the model's graph to retrieve the best scoring graph or, if non-unique, a set of best scoring graphs. In turn, retrieved graphs are evaluated on their Structural Hamming Distance with respect to the ground truth graph.

The thesis is structured as follows. Section 3 reviews principles of causal discovery, which mainly considers the type of causal discovery methods as well as the assumptions typically imposed to identify causal structure from observational data. In Section 4, we discuss structural causal models, their interpretation in score-based methods as well as Markov equivalence classes. Section 5, in turn, introduces and defines time series, which are essential to Granger causality. The topic of Granger causality itself is discussed in Section 6, alongside with practical and theoretical issues associated with interpreting Granger causality as a causality measure. Section 8 sets out the general form of scoring functions and proposes a scoring function based on this form as well as the requirements of the causal discovery task. Lastly, Section 9 starts with an experimental setup, reports the results from the experiments and closes off with an analysis.

3 Principles of Causal Discovery

3.1 An Overview of Causal Discovery

Causal discovery, as Pearl describes it, is "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 [39, p. 43]. Collection of experimental data, however, can be unethical, expensive or simply technically impossible [43, 35, 25, 36]. Cast in general terms, *automated causal discovery* is the inference task of automatically detecting the causal structure underlying a set of variables from observational data. Causal structure, in turn, is encoded in a *graphical model* in which each vertex represents a variable and each connection represents some causal relationship [12, 35, 18, 21]. Clearly, causal discovery is a highly non-trivial task: an observational distribution does not by itself disclose causal structure. Although correlations between variables are available in the observational distribution, such correlations can be spurious [2, 11]. Furthermore, causal relations are thought to be asymmetrical whilst correla-

tions are symmetrical, making it impossible to directly infer causal relations from observed correlations [52, 11]. In order to detect causal structure, then, causal structure must be identifiable from the observational distribution. The assumptions under which causal structure is identifiable are discussed in Section 3.5.

3.2 Causal Structures

Causal discovery interprets causal structure as a graphical model over variables, representing the data-generative process responsible for the observational data. Although different classes of graphical models are available, we assume the model class of *directed acyclic graphs* (DAGs). Hence, the following assumption is imposed on the graphical model:

Definition 3.2.1 (Acyclicity) *A directed graph $G = (V, A)$ is called acyclic if there exists no directed path π such that π both begins and ends with V_i for some $V_i \in V$.*

Acyclicity is unwarranted if there exists a *feedback loop* in the data: one or more variables that causally affect themselves. Standardly, feedback loops are modelled in two ways: (i) with a cyclic graphical model or (ii) by modelling causal structure over time steps while maintaining acyclicity. At inference time, option (i) runs into the problem that inference in cyclic models tends to be more involved than in acyclic models. Since our focus is, moreover, on causal structure over time and since it is plausible that self-directed causal interactions only occur across time, we assume DAGs as model class [12, p. 84]. Under the acyclicity assumption, a causal structure is defined as follows:

Given the acyclicity assumption, causal structure over a set of variables is defined as follows:

Definition 3.2.2 (Causal Structure) *A causal structure over a set of variables $X = \{X_1, ..., X_n\}$ is a directed graph $G = (V, A)$ such that $V = X$ and each arc $(V_i, V_j) \in A$ represents a direct functional relation between represented elements of X.*

Given a causal structure $G = (V, A)$ and two variables $V_i, V_j \in V$, V_i is called a *direct cause* of V_j and V_j a *direct effect* of V_i just in case there exists an arc from V_i to V_j and V_i is called an *indirect cause* of V_j and V_j an *indirect effect* of V_i if there exists a directed path of length at least two starting at V_i and ending at V_j.

3.3 Constraint-based and Score-based Methods

In the literature, the standard classes of causal discovery methods are *constraint-based* and *score-based methods* [18, 2, 35]. Constraint-based methods exploit constraints to derive causal structure in two steps. In the *skeleton phase*, conditional independence constraints are applied to retrieve an undirected graph over the variables [30, p. 444]. In the *orientation phase*, a set of orientation rules is applied to transform edges into arcs, encoding causal dependencies [35, p. 7]. Since it is in general not possible to decide all orientations from rules, constraint-based methods do not guarantee a uniquely identified causal graph and usually return a partial causal structure or a class of candidate graphs instead [18, 12, 52]. Score-based methods, on the other hand, consist of a scoring function and a search space of candidate graphs, each interpreted as a *Bayesian Network* (BN). Given a scoring function that measures the network's fit on the data as well as the network's complexity, the aim is to find the network that optimises the scoring function [2, p. 794].

Both methods come with their own advantages and disadvantages. In general, constraint-based methods tend to be more efficient but more prone to error propagation than score-based methods. Moreover, constraint-based methods are not applicable for determining causal direction in the bivariate case, given the absence of a conditional independence relation [18, p. 5]. Simultaneously, score-based methods can be computationally expensive if the full set of candidate graphs is large. The problem of finding a globally optimal network is, furthermore, known to be NP-hard. To resolve these problems, one can either restrict the space of candidate networks to a suitable subclass or apply a greedy heuristic search. The latter method, it should be mentioned, is susceptible to local optima [2, 35]. Despite these disadvantages, a notable advantage of score-based methods is their ability to impose an order on the set of candidate graphs. Such an order provides more fine-grained information about the candidates: it says how well the best candidate improves on the other candidates [35, p. 7]. Countering relative disadvantages and preserving advantages of constraint-based and score-based methods is done with *hybrid methods*, which effectively combine elements from both methods with the aim of delivering a robust and reliable causal discovery method [30, p. 444]

3.4 A Graphical Criterion for Conditional Independence

In the next section, we discuss standard identifiability assumptions for the causal discovery task. Before doing that, however, we first should review the d-separation criterion. Pearl [40] proposed d-separation as an efficient method for deriving the set of conditional independencies that a directed acyclic graph over a set of variables imposes on a probability distribution over those variables. Formally:

Definition 3.4.1 (d-separation on paths) *Let $G = (V, A)$ be a DAG and $V_i, V_j \in V$ be variables and Z a set of variables. Then, a path π from V_i to V_j in G is d-separated by a set of nodes 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 3.4.2 (d-separation on sets) *A set X is d-separated from a set Y by a set Z iff Z blocks every path π from a node $V_i \in X$ to a node $V_j \in Y$. Otherwise, X is d-connected with respect to Y.*

It may be useful to spell out the intuitions behind d-separation in information-theoretic terms. Condition (i), firstly, states that the information flow between V_i and V_k becomes blocked if that information has to pass through X. Condition (ii), on the other hand, states that the information between the information between the common causes V_i and V_k is blocked unless information about the common effect is available. Intuitively, information about the effect is capable of making information of one cause available when information about the other cause is available [39, pp. 16–17].

3.5 Identifiability Assumptions for Causal Discovery

A causal structure over a set of variables is called *identifiable* just in case the structure can be uniquely determined from the joint distribution over those variables. In the general case, causal structure is *not* identifiable: the joint distribution does not disclose the underlying causal structure [41, p. 44]. Under certain conditions, however, causal structure becomes identifiable. Common identifiability assumptions relevant to the DAG setting are the causal Markov condition, faithfulness and causal sufficiency [46].

Now, an observational distribution over variables does not itself disclose underlying causal structure. Consequently, learning causal structure from observational distributions requires further background assumptions about the relationship between causal structure and observational distributions. Although no consensus or theoretical proof exists concerning the correct set of assumptions, two common simplifying assumptions are the causal Markov condition and the faithfulness condition [2, 23, 35, 52, 24, 12].

Definition 3.5.1 (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 3.5.1 imposes that non-causation involves non-association or, contrapositively, that association involves causation. This is an oversimplification, as the absence of causation between two variables is consistent with association as in well-known cases of spurious correlation [52, pp. 6–7].

Definition 3.5.2 (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 3.5.2 states the reverse of Definition 3.5.1: non-association involves non-causation or, contrapositively, causation involves association. This, in turn, is an oversimplification since two variables can be causally related without being associated, as in the case where effects of the causal variable are cancelled out due to the effect of other variables [11, p. 7].

Under the causal Markov and the faithfulness condition, there exists a one-to-one correspondence between d-separations in the graph and conditional independence in the distribution [23, p. 5]. As Peters, Janzing, and Schölkopf point out, finding the causal structure only requires the causal Markov condition and a weaker condition called the minimality condition [41, p. 197]. Still, assuming both the causal Markov and the faithfulness condition is desirable for the following reason: it allows identification of the Markov Equivalence Class from the distribution, as further described in Section 4.4.

A further common assumption is causal sufficiency, which assumes that all causal variables are included in the observational data:

Definition 3.5.3 (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.*

Under the causal sufficiency assumption, the causal structure learned over observed variables necessarily includes all causal variables [35, p. 5]. It can occur, however, that not all relevant causal variables are measured. If a latent variable is excluded from the model, spurious relations can emerge between the included effect variables. If a latent variable U is a common cause of observed variables X and Y, for instance, then excluding U can result in the wrong conclusion that X and Y are causally related [2, 11, 43, 35]. Usually, this is resolved by modelling latent variables in a mixed acyclic graph (MAG), which models the presence of latent variables with bi-directed arcs between spuriously correlated variables [16, pp. 94–96]. Although we believe that causal sufficiency is a theoretically undesirable assumption, we restrict our focus to DAG models and leave an expansion to MAG models to future work.

4 Pearl Causality

4.1 Overview of Pearl Causality

#TODO: give short conceptual overview

4.2 Causal Models

Within Pearl's causality framework, the basic unit is what is called a *structural causal model* (SCM). Intuitively speaking, a SCM over a set of variables dictates which causal dependencies hold between those variables as well as the nature of those dependencies. More specifically, a SCM is a model that specifies the causal relations between variables in terms of a qualitative and a quantitative part: (i) *causal structure* and (ii) *causal influence*. Causal structure is, as we saw, represented as a graphical structure over variables. Causal influence, in turn, is defined in terms of *structural equations* that define the value of given variables in terms of other variables interpreted as causes joined with mutually independent noise terms interpreted as influences external to the model [39, p. 27]. More formally, SCMs are defined as follows:

Definition 4.2.1 (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* $U_i \in U$ *and independent of all other* $U_j \in U$

(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. Endogenous variables represent observed causes and effects which are internal to the model. Importantly, each endogenous variable is defined in terms of some exogenous variable. In particular, if the value of each exogenous variable is observed, then the value of each endogenous variable can be inferred using the structural equations [38, pp. 47–48].

4.3 Bayesian Network Interpretation

As intended in Definition 4.2.1, *compatibility* of a causal structure $G = (V, A)$ and set of parameters $\Theta = (U, V, F)$ requires that G and 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 3.2.2 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 the est of structural equations F defined by Θ.

Alternatively, the compatibility requirement can be defined in terms of the BN interpretation of a causal graph. Since this interpretation is central to score-based methods, it is worth discussing. Given a causal structure $G = (V, A)$, the BN representation $\mathcal{M} = (G, \Theta')$ is retrieved by applying parameter estimation to obtain the joint probability distribution $\Pr^{\Theta'}(v_1, ..., v_n) = \prod_{i=1}^{n} \Pr^{\Theta'}(v_i|pa_i)$, under the constraint that PA_i are the parents of V_i in G for each term $\Pr^{\Theta'}(v_i|pa_i)$. In effect, compatibility of a graph G with a distribution Pr is the requirement that Pr respect the decomposition of G or, equivalently, that the graph is in principle capable of generating data from Pr given the appropriate set of parameters:

Definition 4.3.1 (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.

Effectively, this is the causal Markov condition defined before. Although the BN representation preserves graphical structure, it does not preserve causal information: dependencies between X_i and $\{PA_i, U_i\}$ are maintained, but the nature of the causal dependencies defined by f_i are ignored [39, pp. 44–45].

4.4 Markov Equivalence Classes

A central notion within causal discovery is that of a Markov Equivalence Class (MEC). As proposed by Verma and Pearl [54], a MEC defines a set of DAGs that (i) share the same skeleton or undirected graph and (ii) encode the same conditional independence statements. Effectively, the first condition can be verified by considering whether each graph $G = (V, A)$ induces the same undirected graph $G' = (V', E)$ where $V' = V$ and $E' = \{\{V_i, V_j\} : (V_i, V_j) \in A \vee (V_j, V_i) \in A\}$. The second condition, on the other hand, is more involved: it demands a method for determining which conditional independence statements are imposed by arbitrary graphs [39, p. 16]. As discussed in Section 3, 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 set of v-structures is the same: triples $V_i \rightarrow V_j \leftarrow V_k$ where V_i and V_k are non-adjacent. Formally, these observations result in the following definition:

Theorem 4.1 (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.*

Intuitively, Markov equivalence amounts to encoding the same dependency structure in the undirected graph as well as encoding the same conditional independence statements in terms of separations in the directed graph. A MEC of a graph, in turn, is simply the set of graphs that are Markov equivalent to that graph. Formally, the MEC $\mathcal{G}_G$ of a graph G is the set $[G]_\sim = \mathcal{G}_G = \{G' \in \mathscr{G} : G \sim G'\}$ where $\mathscr{G}$ is the entire space of directed graphs and $\sim$ is the Markov equivalence relation. Visually, $\mathcal{G}_G$ of G can be represented by a completed partially directed acyclic graph (CPDAG) $G' = (V, A)$ where V is the set of vertices shared by all members of $\mathcal{G}_G$ and the arc (V_i, V_j) is included in A if all members of $\mathcal{G}_G$ contain the arc (V_i, V_j) and the edge $\{V_i, V_j\}$ is included in A if some members of $\mathcal{G}$ contain the arc (V_i, V_j) and others contain the arc (V_j, V_i) [39, pp. 18–19]. As noted before, assuming the causal Markov condition and faithfulness ensures that the MEC is identifiable from the distribution. Formally, 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) $G \notin \mathcal{G}$ if there exists no parameter set $\Theta = (U, V, F)$ over G that induces Pr and (ii) $G \in \mathcal{G}$ if G is Markov and faithful to Pr. The reader is referred to Peters, Janzing, and Schölkopf [41] for the formal proof [41, pp. 44, 135–136].

5 Time Series

5.1 Time Series Analysis

Cross-sectional data consists of observations at a single time point or interval. Such data is *static*: temporal distinctions between observations are ignored, thus treating all observations as if belonging to the same temporal state. Contrastingly, time series is *dynamic*: observations are distributed over a sequence of time points. As a field, time series analysis leverages time series data with the aim

> to 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 [9, p. 1].

Notable disciplines concerned with time series analysis of stochastic processes are climate science, economics, epidemiology, neuroscience and physics. In what follows, we assume that a time series is defined as a collection of time-indexed random variables. Furthermore, observed values of a given time series are called a *realisation* of that time series [4, 48].

5.2 Time Series Notation

In the standard cross-sectional setting, data of a population is defined as a set of random variables $\mathbf{X} = \{X_1, ..., X_n\}$ for some fixed $n \in \mathbb{N}$. Contrastingly, time series data is defined by a set of time series $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$ that each encodes the state of a random variable throughout time. Formally, each time series $\mathbf{X}_i$ is a set $\{X_i^1, ..., X_i^T\}$ of the states of the random variable X_i up and until T for some fixed $T \in \mathbb{N}$. Hence, the j'th element of the i'th time series is denoted X_i^j and represents the state of X_i at time j. In what follows, it is assumed that each single time-indexed random variable X_i^t is a random variable itself and, thus, has multiple value realisations. For ease of notation, $\mathbf{X}_i^{t-k:t}$ represents the states of the random variable 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, $\mathbf{T}^{:t}$ selects the set $\bigcup_{i=1}^n \mathbf{X}_i^{:t}$ consisting of all time-indexed variables up and until time t from the time series elements $\mathbf{X}_i$.

5.3 Causal Models for Time Series

In the cross-sectional setting, causal models define causal influence over a set of variables at a fixed point in time. In the time series case, however, causal influence is defined over variables at distinct points in time. Most notably, this involves that variables can influence both other variables as well as themselves across time. More formally, it can occur that a variable X_i at time t is a function of itself at time $t - \tau$ for some time lag $0 < \tau$. Fortunately, time series data can still be defined as an SCM. A difference is that causal influence is not defined over time series elements but, instead, over the full set of time-indexed random variables retrieved from the union $\bigcup_{i=1}^n \mathbf{X}_i$ of time series included in $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$ [41, p. 199]. Defined more explicitly:

Definition 5.3.1 (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* $U_i \in U$ *and independent of all other* $U_j \in U$

(ii) $V = \bigcup_{i=1}^{n} \mathbf{X}_i$ *is a set of time-indexed endogenous variables from a time series set* $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$ *where each* $\mathbf{X}_i = \{X_i^1, ..., X_i^T\}$ *for a fixed* $T \in \mathbb{N}$

(iii) F *is a set of structural equations* f_i^t *for* $1 \leq t \leq T$ *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* $0 < \tau$

The only differences with a regular SCM consist in clauses (ii) and (iii). Clause (ii) defines the set of vertices, simply, as the set of time-indexed random variables included in the elements of the time series set. Clause (iii), in turn, defines the structural equation of each time-indexed random variable as a dependence with parent sets at previous points in time up and until some time lag. Since contemporaneous causation is not discussed, we assume that contemporaneous parent sets are excluded from each structural equation.

5.4 Causal Structures for Time Series

An important observation is that clause (iii) from Definition 5.3.1 implicitly defines the causal structure G of a DSCM as a graph over the full set of time-indexed variables $\bigcup_{i=1}^{n} \mathbf{X}_i$. Formally, this representation is defined as a *full time causal graph*:

Definition 5.4.1 (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* $V = \bigcup_{i=1}^{n} \mathbf{X}_i$ *and* $(V_i^{t-\tau}, V_j^t) \in A$ *for* $0 < \tau$ *iff* $V_i^{t-\tau}$ *occurs in the structural equation* f_j^t *of* V_j^t.

The full time representation, however, has two important issues. Firstly, as time series get longer, the graph grows and "can become unwieldy and difficult to interpret" [15, 1]. Furthermore, the model selection process is hampered given that the possible graphs on a set of vertices rapidly grows with the size of that set [10, 13]. Alternatively, causal structure can be modelled with a *window causal graph*. This representation restricts the set of vertices to a window of length γ, effectively representing the largest time gap between causes and effects in the DSCM:

Definition 5.4.2 (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* $V = \bigcup_{i=1}^{n} \mathbf{X}_i^{t-\gamma:t}$ *for some time* $0 < t$ *and time lag* $0 < \gamma$ *and* $(V_i^{t-\tau}, V_j^t) \in A$ *for some time lag* $0 < \tau \leq \gamma$ *iff* $V_i^{t-\tau}$ *occurs in the structural equation* f_j^t *of* V_j^t.

Essentially, the window graph aims to contract the full time representation into a more manageable representation. Under the assumption of *causal stationarity* or *consistency throughout time* further discussed in Section 6.3, the full time causal graph is reducible to the window causal graph [1, pp. 1–2]. In the worst case, however, the largest time gap γ spans the entire time series, inducing the full time representation itself. A *summary causal graph* forces a condensed representation by modelling the time series instead of the underlying time-indexed variables. Within this representation, an arc between vertices corresponds to some causal influence existing between time-indexed variables of the represented time series. Formally:

Definition 5.4.3 (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 $0 < t$ and lag $0 < \tau$ such that $V_i^{t-\tau}$ occurs in the structural equation f_j^t of V_j^t.*

Although summary graphs impose condense representations, this sort of condensity comes at the expense of temporal information: the existence of an arc entails nothing about the times at which causal influence occurred nor about the number of times it occurred. A second important observation is that the window graph is acyclic whenever the full time graph is acyclic, given that the former is simply an induced subgraph of the latter. Contrastingly, the summary graph is not ensured to be acyclic. A variable at a given time t can, for instance, depend on itself at a previous time $t - \tau$, thus producing a self-loop in the summary graph [1, 41]. Examples of a full time, window and summary graph on time series $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are shown in Figure 1-3.

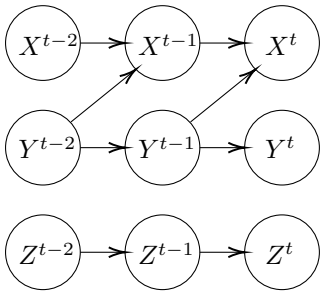


Figure 1: Full Time Causal Graph

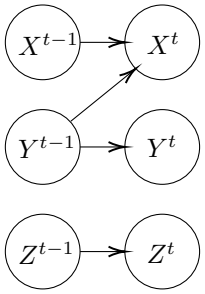


Figure 2: Window Causal Graph

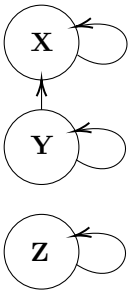


Figure 3: Summary Causal Graph

6 Granger Causality

6.1 The Concept of Granger Causality

Granger causality or G-causality is a well-established method or detecting causality [1, 20, 42]. Importantly, G-causality is not a theoretical framework aimed at defining what counts as true causation. Instead, it is a method that uses forecasting ability to define what are called G-causes: variables that contain unique information which aids in the prediction of other variables [20, p. 430]. Consequently, G-causality differs from what is called *structural causality*: the existence of a G-cause between variables is neither necessary nor sufficient for the existence of a true causal mechanism [42, p. 10]. Given these considerations, G-causes ought to be interpreted as *potential causes* [14, 55]. Nevertheless, G-causality is characterised by two principles believed to be essential to causation: *temporal precedence* and *uniqueness*. First of all, temporal precedence holds that causes must precede their effects in time. Secondly, uniqueness demands that causal time series include unique information about caused time series, which would be unavailable in the absence of those causal series [19, 14].

It is worth discussing some advantages and disadvantages of G-causality. 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 observed data [13, 14, 17]. A disadvantage of using a model-free approach is that the quality of G-causal conclusions is subject to statistical conditions such as appropriate sampling, non-instantaneous causation and stationarity [33, pp. 87, 98]. From a theoretical perspective, G-causality has an appealing interpretation. Firstly, the principle of uniqueness coincides with the observation that in causal relations, changes in the effect are due to the purported cause and not to something different in the system. Secondly, introducing a temporal ordering respects the asymmetric nature of the arrow of causation, which has been met with increased attention and application in recent years [2, 22, 39, 56]. At the same time, it should be noted that imposing a temporal ordering over variables is not sufficient for deriving a *correct* causal ordering: spurious correlations still occur across time [31, 56].

6.2 Formal Definitions

Broadly viewed, we can distinguish between two interpretations of what G-causes are. In the definitions that Granger [20] originally proposed, G-causes are interpreted in terms of conditional independence of *time series* across time. Under the lag-specific definition from Runge [43], contrastingly, this logic is applied to *time-indexed variables* instead. Whilst Granger's interpretation locates causation as a relation between the time expansions of random variables, then, Runge's approach defines causation on specific time instances of those random variables. Consequently, the latter approach makes it transparent which specific time-indexed variables in the time series induce conditional independence.

In Granger's original definition, G-causality is defined against the background of the full domain of time series in the universe. Theoretically speaking, this is to ensure that the information from a given G-cause is unique to that G-cause and does not, instead, derive from elsewhere. Formally:

Definition 6.2.1 (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* $Y^{t+1} \perp\!\!\!\perp \mathbf{X}^{:t} | \mathbf{\Omega}^{:t} \setminus \mathbf{X}^{:t}$ *for all* $t \in \mathbb{N}$*. Otherwise,* $\mathbf{X}$ *is said to Granger cause* $\mathbf{Y}$*.*

In natural terms, a G-cause holds just in case there exists some point in time where the causal and effect variables are dependent given the information from all other previous time series. Clearly, Definition 6.2.1 is unrealistic: real-world data at most provides access to a minute subset of $\mathbf{\Omega}$. Given this consideration, G-causality is adapted to a finite set of time series:

Definition 6.2.2 (1-step Granger Causality) *Let* $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$ *be a finite time series. If* $\mathbf{X}, \mathbf{Y} \in \mathbf{T}$*, then* $\mathbf{X}$ *does not Granger cause* $\mathbf{Y}$ *if* $Y^{t+1} \perp\!\!\!\perp \mathbf{X}^{:t} | \mathbf{T}^{:t} \setminus \mathbf{X}^{:t}$ *for all* $t \in \mathbb{N}$*. Otherwise,* $\mathbf{X}$ *is said to Granger cause* $\mathbf{Y}$*.*

In the above definition, the time window between G-causes and G-effects are fixed at time lag $\tau = 1$. Since G-causes are definable over longer time windows and since setting $\tau = 1$ is a strong assumption, the final adaptation is to generalise to G-causes with time windows of $1 \leq \tau$:

Definition 6.2.3 (h-step Granger Causality) *Let* $\mathbf{T} = \{\mathbf{X}_1, ..., \mathbf{X}_n\}$ *be a finite time series. If* $\mathbf{X}, \mathbf{Y} \in \mathbf{T}$*, then* $\mathbf{X}$ *does not Granger cause* $\mathbf{Y}$ *if* $Y^{t+1} \perp\!\!\!\perp \mathbf{X}^{:t-\tau} | \mathbf{T}^{:t} \setminus \mathbf{X}^{:t-\tau}$ *for all* $t \in \mathbb{N}$ *and* $0 \leq \tau$*. Otherwise,* $\mathbf{X}$ *is said to Granger cause* $\mathbf{Y}$*.*

A final remark is that in the practical setting, the domain of time indices will span some finite interval $\mathcal{T} = \{1, ..., T\}$ for some $T \in \mathbb{N}$ [20, 19, 13, 14].

The usefulness of the lag-specific interpretation emerges from a loss of information in the time series case. Observe that in the time series case, a G-cause $\mathbf{X} \to \mathbf{Y}$ holds just in case there exists some time t and time lag τ for which the following conditional dependence statement holds [41, p. 207]:

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

A major problem, however, is that the conditional statement leaves implicit which variables render the dependence. Since this information is important when evaluating the impact of states of time series on states of other time series, a sensible choice is to evaluate the specific lags $0 \leq \tau$ at which $X^{t-\tau}$ generates dependence. Runge [43] defines lag-specific G-causation as follows: instead of evaluating conditional dependence of an entire history of $\mathbf{X}$ with Y^{t+1}, independence is evaluated between Y^{t+1} and a single lagged instance $X^{t-\tau}$. Formally:

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

Since the conditional dependence imposed in Equation 2 regards $X^{t-\tau}$ and Y^{t+1}, only $X^{t-\tau}$ becomes removed from the conditional set [43, 46, 44]

6.3 Further Assumptions for Causal Discovery

In addition to the identifiability assumptions outlined in Section 3.5, we outline a number of assumptions relevant to structural causal models as well as G-causality.

A first assumptions concerns the *dependency type*: whether the structural equations in models are linear or non-linear [43, p. 8]. Secondly, there is the assumption of *causal stationarity*: that causal relationships "remain constant in direction throughout time" or, equivalently, that causal mechanisms are invariant with respect to changes in time. More formally, this involves that each structural equation f_j^t encodes exactly the same dependency for all points in time [43, 46, 1, 2, 5]. In the CI-based setting, this results in stationarity on the conditional independence relation Runge [43, p. 8]. A last assumption is the assumption of *non-instantaneous effect*: that causal relationships occur only across time points and, consequently, cannot occur within the span of a single time point. Essentially, this assumption accords with the Granger's intuitive time precedence principle, which states that causes always occur earlier than effects [43, p. 8]. Despite its theoretical appeal, the non-instantaneous effects assumption involves a practical problem: data recordings may not be sufficiently fine-grained to ensure that causal relations never occur within the span of a single time point. A first case is the problem of *subsampling*: when the rate of sampling is slower than the causal process generating the time series data. Secondly, *temporal aggregation* of time series data for size reduction is prone to merge data of causes and effects even if the original data separated causes and effects [2, 18, 41].

7 Background

7.1 Related Work

In this section, we discuss work relevant to this thesis. In the subsequent section, we describe the main contributions of this thesis. As we described in Section 3, causal discovery methods are classified into constraint-based, score-based and hybrid methods. An influential constraint-based method is the PC algorithm from [50], which effectively aims to learn the CPDAG representation of the Markov Equivalence Class. In the first step, PC initialises the complete undirected graph $G = (V, A)$ and constructs the skeleton by evaluating for all adjacent nodes V_i, V_j and $k = 0, 1, 2, ...$ the conditional independence $X_i \perp\!\!\!\perp X_j | Z$ where $|Z| = k$ and $Z = \{X_k : V_k \in adj(V_i)\}$ or $Z = \{X_k : V_k \in adj(V_j)\}$. Whenever $X_i \perp\!\!\!\perp X_j | Z$ holds, the edge between V_i and V_j is removed and Z is stored. In the second step, v-structures in the graph are identified using stored d-separation sets and orientation rules. In the final step, further orientation rules are applied to orient remaining edges as much as possible [50, 18, 23, 51]. Runge et al. [45] developed PCMCI, which combines Granger causality with the PC algorithm to learn a causal graph from time series. Similar to PC, the first step of PCMCI iteratively tests the hypothesis $H_0 : X_i^{t-\tau} \perp\!\!\!\perp X_j^t | Z$ with $|Z| = k$ from $k = 0, 1, 2, ...$ onwards. If

the p-value for H_0 is below a fixed threshold α, $X_i^{t-\tau} \to X_j^t$ is removed. In the second stage, a momentary conditional independence test is applied on each pair of variables $X_i^{t-\tau}, X_j^t$ to control for false positives that occur in highly interdependent time series [45, 46].

A central score-based method is Greedy Equivalence Search (GES), introduced by Meek [34] and further developed by Chickering [7]. GES starts with the empty graph and consists of two further phases in which candidate graphs are scored using the Bayesian Information Criterion (BIC). In the first phase, arcs are added to the graph until a local maximum is reached. In the second phase, arcs are removed until a local maximum is reached, which marks the termination of the algorithm, after which GES outputs the graph that locally maximises BIC [7, 34, 41]. Pamfil et al. [37] proposed a score-based method called DYNOTEARS for learning a window graph over time-indexed variables representing both contemporaneous and lagged relationship. Effectively, the method learns graphical structure represented as adjacency matrices by minimising a least squares loss objective subject to ℓ_1-penalisation as well as an acyclicity constraint [37]. It has been shown, however, that methods such as DYNOTEARS are oriented towards finding parsimonious graphs that best explain the data, making them unsuitable for the causal discovery task [1, 27]. A hybrid method for learning non-causal Bayesian Networks called the Max-Min Hill-Climbing (MMHC) algorithm has been presented by Tsamardinos, Brown, and Aliferis [53]. Combining techniques from local learning, constraint-based and score-based methods, the MMHC algorithm consists of two phases. In the first phase, a skeleton graph is learned using a local discovery algorithm called the Max-Min Parents and Children algorithm. The second phase is a greedy hill-climbing search starting from the empty graph. During this phase, edge addition, deletion and reversal are applied with the constraints that added edges were present in the skeleton found with the MMPC algorithm in the first phase [53, p. 33].

7.2 Main Contribution

#TODO: describe main contribution of thesis as well as properties of the method

8 Scoring Functions

8.1 Generalised 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, find a graph that optimises 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, this aim is formalised by selecting $\hat{G}$ in the following equation:

$$\hat{G} = \text{argmax}_{G \in \mathcal{G}} \ \phi(\mathcal{D}, G) \tag{3}$$

An important advantage of score-based methods is their ability to induce an order on the set of candidate graphs, which allows for a qualitative comparison of graphs. Formally, this is because ϕ assigns real-valued scores: since each $G \in \mathcal{G}$ is assigned a real 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 $\phi(\mathcal{D}, G), \phi(\mathcal{D}, G') \in \mathbb{R}$.

Given that the qualitative order on $\mathcal{G}$ is defined by ϕ, an important task of score-based methods consists in defining an appropriate scoring function. Minimally, a scoring function should define a graph's quality in terms of its fit on the data and its complexity [2, 3]. On its own, a graph defines only qualitative relations between variables. In order to define a graph's fit on the data, the graph's BN interpretation is adopted [2]. Commonly, a parametric model is assumed, which can be estimated using properties of the graph. After estimating the set of parameters, a BN $\mathcal{M} = (G, \Theta)$ results. Since $\mathcal{M}$ defines a joint probability distribution over the variables of interest, it has a fit or likelihood score with respect to the data $\mathcal{L}(\mathcal{D}, \mathcal{M})$. Furthermore, $\mathcal{M}$ has a complexity score $dim(G)$ proportionate to the number of parameters in Θ [41, pp. 148–149]. Given a function f that determines the trade-off between fit and complexity, the following scheme results:

$$\phi(\mathcal{D}, G) = f(\mathcal{L}(\mathcal{D}, \mathcal{M}), dim(G)), \tag{4}$$

Next to fit and complexity, it is common to impose two further restrictions on scoring functions. Firstly, *score decomposability*: the score over the whole graph must be decomposable as a sum of independent local scores defined over single variables joined with their parents [26, 32]. Formally:

$$\phi(\mathcal{D}, G) = \sum_{V_i \in V} f(\mathcal{D}, V_i \cup PA_i), \tag{5}$$

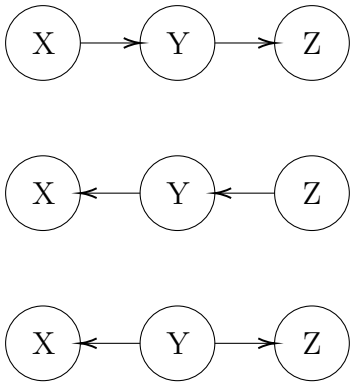
where f defines local scores. Decomposability is desirable for at least two reasons. First of all, the existence of local scores allows for subdividing the complex task of computing the total score of the graph into locally and efficiently computable subtasks. Secondly, the fact that total scores are sums of independent scores implies that each score can be stored for reuse, increasing the efficiency of greedy heuristics such as hill climbing search [8, pp. 50–52].

A second restriction on scoring functions is what is called *score equivalence*: graphs that belong to the same equivalence class are assigned the same score. More formally, a scoring function ϕ is score equivalent just in case the following

entailment holds: for all MECs $\mathcal{G}$ and graphs $G, G' \in \mathcal{G}$, $\phi(\mathcal{D}, G) = \phi(\mathcal{D}, G')$ [32, 6, 3, 29]. The motivation behind the score equivalence assumption is what is called the *independence interpretation* of graphical structure. Under the independence interpretation, graphical structures are interpreted as sets of independence constraints on a probability distribution. Since structures within the same equivalence class by definition impose the same independence constraints, graphs in the same equivalence class should be assigned the same score [7, p. 448].

8.2 Causal Scoring Functions

Since the independence interpretation defines graphical structures as the set of imposed independencies, it seems reasonable to assume score equivalence. In the causal setting, however, the independence interpretation imposes an incorrect interpretation of graphical structure: graphical structures are not defined by imposed independencies but, instead, are given a causal interpretation. In particular, distinct graphs within the same Markov equivalence class encode distinct causal structures [7, p. 448]. As a simple illustration, consider the following three structures:



Since the three structures above belong to the same MEC, the score equivalence assumption implies that each should be assigned the same score. In the causal setting, this is a non-trivial decision: $X \to Y$, for instance, differs from $Y \to X$.

If a scoring function is to account for fine-grained distinctions in causal direction, then it should, firstly, not be score equivalent. Hence, the scoring function must be able to assign different scores to graphs belonging to the same equivalence class. An important further desideratum is that differences in scores reflect the graph's ability to capture *causal* information: otherwise, the scoring function would not be very useful in discovering causal structure. Although the first desideratum seems relatively easy to satisfy, the second desideratum is highly non-trivial

given the very premise of the causal discovery task: the true causal structure is unavailable and must be retrieved from the observational distribution [39, p. 43]. Simultaneously, the assumption that the true causal structure is unknown is consistent with the assumption that *some* causal information is available. Suppose that we are given some non-symmetric bivariate causality measure $\kappa : \mathscr{D} \times \mathscr{A} \to \mathbb{R}$ that scores how well a given arc captures directional causal information, derived in some way from the data. Given the bivariate scores from κ, a causal score $\varkappa(\mathcal{D}, G)$ could 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:

$$\phi(\mathcal{D}, G) = g(\mathcal{L}(\mathcal{D}, \mathcal{M}), dim(G), \varkappa(\mathcal{D}, G)) \tag{6}$$

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.

8.3 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 $\kappa : \mathscr{A} \to [0, 1]$. 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 matrix M in which each element m_{ij} is defined as follows:

$$m_{ij} = \begin{cases} 1 & \text{if } (V_i, V_j) \in A \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Since M indicates which vertices are adjacent in G, M 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 of a digraph G is a square $n \times n$ matrix that can be translated back into G by defining $G = (V, A)$ with $V = \{V_1, ..., V_n\}$ and $(V_i, V_j) \in A$ if and only if $m_{ij} = 1$ [49, p. 6].

Under the matrix representation, assigning causal scores to arcs effectively reduces to an element-wise matrix multiplication of the adjacency matrix with a causal scoring matrix. Given a non-symmetric function $\kappa : \mathscr{A} \to [0, 1]$ that

measures causal information of single arcs, we can construct a non-symmetric $n \times n$ score matrix M^* where each elements is defined as $m^*_{ij} = \kappa((V_i, V_j))$. Since $m^*_{ij} \in [0, 1]$ for each $(m^*_{ij})_{1 \leq i,j \leq n}$ and since $m_{ij} \in \{0, 1\}$ for each $(m_{ij})_{1 \leq i,j \leq n}$, the product $m_{ij} \cdot m^*_{ij}$ returns a score in the interval $[0, 1]$. Now, if $m_{ij} \cdot m^*_{ij} = 0$, then either $m_{ij} = 0$ or $m^*_{ij} = 0$. More intuitively, zero scores indicate either the absence of an arc or a poor grasp of causal information of a present arc. Conversely, non-zero scores indicate the degree to which a present arcs captures causal information. Using M and M^*, we can apply the Frobenius inner product $\langle M, M^* \rangle_F = \sum_{i,j}^{n} m_{ij} \cdot m^*_{ij}$ to construct a total score. Given the noted properties, higher values of the resulting score indicate a better grasp of causal information whilst lower values indicate a poor grasp of causal information.

Defining $\varkappa(\mathcal{D}, G)$ as $\langle M, M^* \rangle_F = \sum_{i,j}^{n} m_{ij} \cdot m^*_{ij}$ satisfies both score decomposability and score non-equivalence. Since the whole score $\varkappa(\mathcal{D}, G)$ is by definition a decomposable sum over individual arcs, the score decomposability property is satisfied. Consider a graph $G = (V, A)$ and an arbitrary variable $V_k \in V$. By Equation 7, the parents of V_k correspond to the k'th column of the adjacency matrix M of G and the associated causal scores correspond to the k'th column of M^* or, more formally, the column vectors $(m_{ik})_{1 \leq i \leq n}$ and $(m^*_{ik})_{1 \leq i \leq n}$. Since the full score $\varkappa(\mathcal{D}, G)$ is simply the sum over scores $\sum_{i=1}^{n} m_{ik} \cdot m^*_{ik}$ at each k'th column, $\varkappa(\mathcal{D}, G)$ can be decomposed into local scores $\sum_{i=1}^{n} m_{ik} \cdot m^*_{ik}$ defined at each variable V_k joined with its parents. To see that causal scores are non-equivalent, observe that κ is a non-symmetric causality measure: generally, $\kappa((V_i, V_j)) \neq \kappa((V_j, V_i))$ for arbitrary vertices V_i, V_j. From this, it follows that whenever $A \neq A'$ for two graphs $G = (V, A), G' = (V', A') \in \mathcal{G}$ in the same Markov equivalence class $\mathcal{G}$, $\varkappa(\mathcal{D}, G) \neq \varkappa(\mathcal{D}, G')$ is not excluded.

8.4 Granger Scoring Function

In this section, we propose a scoring function that uses p-values from G-causality tests as causal scores. The motivation behind using p-values is twofold. First of all, a p-value on a null hypothesis that $X^{t-\tau}$ is not a G-cause of Y^{t+1} provides a natural interpretation of the arc's quality: lower values of p indicate that the arc $X^{t-\tau} \to Y^{t+1}$ captures G-causal information better whilst higher values of p indicate that $X^{t-\tau} \to Y^{t+1}$ has a poor grasp of G-causal information. Secondly, well-performing constraint-based methods such as PCMCI from Runge et al. [45] already leverage these p-values to maintain or remove arcs given a significance level α. Since α is effectively used as a decision threshold on the p-values of given arcs, using these scores on arcs directly does not seem far-fetched.

In order to evaluate the null hypothesis $Y^{t+1} \perp\!\!\!\perp X^{t-\tau} | \mathbf{T}^{:t} \setminus \{X^{t-\tau}\}$ for arbitrary inputs, a conditional independence oracle is required. Suppose we are given an

oracle $\mathcal{I}$ that takes as input observations of sets of random variables $(x_i, y_i, z_i)_{i=1}^n$ and returns a p-value for the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$: formally, $p = \Pr(T_n \geq \hat{T}_n|H_0)$ where $\hat{T}_n$ is the observed value of the test statistic. Given $\mathcal{I}$, we can then define the following causal measure $\kappa_{\mathcal{I}} : \mathbf{T} \times \mathbf{T} \times \mathbf{T} \to [0, 1]$:

$$\kappa_{\mathcal{I}}(X^{t-\tau}, Y^{t+1}, \mathbf{T}^{:t} \setminus X^{t-\tau}) = 1 - \mathcal{I}(\{X^{t-\tau}\}, \mathbf{T}^{:t} \setminus X^{t-\tau}, \{Y^{t+1}\}) = 1 - p \quad (8)$$

Intuitively viewed, $\kappa_{\mathcal{I}}$ assigns a score that indicates the likelihood of the alternative hypothesis that $X^{t-\tau}$ is a G-cause of Y^{t+1} or, alternatively, how well $X^{t-\tau} \to Y^{t+1}$ captures causal information.

Under the causal stationarity assumption, it suffices to estimate the time window causal graph [1, p. 1]. Under the assumption of access to the Markov equivalence class $\mathcal{G}$ of the true window graph, the causal scoring matrix is of size $n \times n$ where n is the number of vertices of each member $G' \in \mathcal{G}$.[1] In advance, we can set $m^*_{ij} = 0$ whenever (i) the represented edge is absent in the skeleton underlying $\mathcal{G}$ or (ii) the time indices of the represented arc do not satisfy the time precedence condition. In the remaining non-zero cells m^*_{ij} of M^*, $\kappa_{\mathcal{I}}$ can be applied to derive a causal scoring matrix as follows: define $m^*_{ij} = \kappa_{\mathcal{I}}(X^{t-\tau}, Y^{t+1}, \mathbf{T}^{:t} \setminus X^{t-\tau})$, where it is assumed that m^*_{ij} represents $X^{t-\tau} \to Y^{t+1}$. Given a candidate graph $G = (V, A)$ over the n vertices together and its $n \times n$ adjacency matrix, the full causal score is computed as follows:

$$\varkappa(\mathcal{D}, G) = \langle M, M^* \rangle_F = \sum_{i,j}^{n} m_{ij} \cdot m^*_{ij} \quad (9)$$

Since we assumed that $\varkappa$ establishes score non-equivalence amongst members of a MEC, 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 $\mathcal{G}$ and, thus, redundant. By construction, then, differences between $G, G' \in \mathcal{G}$ results wholly 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 Equation 10 reduces to the following objective in the setting where the set of candidates $\mathcal{G}$ is a MEC:

$$\hat{G} = \text{argmax}_{G \in \mathcal{G}} \ \phi(\mathcal{D}, G) = \text{argmax}_{G \in \mathcal{G}} \ \varkappa(\mathcal{D}, G) \quad (10)$$

[1] Note: since these vertices correspond to time-indexed variables, m^*_{ij} does not designate the arc (V_i, V_j).

8.5 Scoring Procedure

At this point, we are in a position to set out our method in explicit terms. Algorithm 1 describes the retrieval of the causal scoring matrix. Scoring a Markov equivalence class as well as retrieving the best scoring candidate are outlined in Algorithm 2.

Algorithm 1 Significance Testing Phase

function GET_SCORING_MATRIX($\mathcal{D}, \mathcal{G}, d, \mathcal{I}$)
  $V \leftarrow$ VERTICES($\mathcal{G}$)
  $\gamma \leftarrow$ MAX_LAG(V)
  $d \leftarrow |V|$
  $S \leftarrow [d, d]$

  for $c \in \{1, ..., d\}$ **do**
    for $e \in \{1, ..., d\}$ **do**
      $X_i^t \leftarrow V[c]$
      $X_j^{t'} \leftarrow V[e]$
      if $0 < t - t' \leq \gamma$ **then**
        $S[c, e] \leftarrow \mathcal{I}(X_i^t, X_j^{t'}, V \setminus \{X_i^t\})$
      else if $t - t' = 0$ or $\gamma < t - t'$ **then**
        $S[c, e] \leftarrow 0$
      end if
    end for
  end for
  return S
end function

9 Experiments

9.1 Evaluating Causal Methods on Synthetic Data

Before outlining the experiment, it is important to emphasise 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, causal discovery becomes infeasible [41, pp. 50, 138]. Shimizu [47] and Hoyer et al. [25] discuss positive identifiability results for linear additive models with non-Gaussian noise and non-linear additive

Algorithm 2 Scoring Phase

function SCORE_EQUIVALENCE_CLASS($\mathcal{M}, T, \mathcal{G}, d, \mathcal{I}$)
  $\mathcal{D} \leftarrow$ GENERATE_DATA($\mathcal{M}, T, d$)
  $M^* \leftarrow$ GET_SCORING_MATRIX($\mathcal{D}, \mathcal{G}, d, \mathcal{I}$)
  $d \leftarrow |\mathcal{G}|$
  $S \leftarrow [d]$
  for $k \in \{1, ..., d\}$ **do**
    $M \leftarrow$ GET_ADJACENCY_MATRIX($\mathcal{G}[k]$)
    $S[k] \leftarrow \sum_{i,j}^{n} m_{ij} \cdot m^*_{ij}$
  end for
  return S
end function

function GET_BEST_CANDIDATE($\mathcal{M}, T, \mathcal{G}, d, \mathcal{I}$)
  $S \leftarrow$ SCORE_EQUIVALENCE_CLASS($\mathcal{M}, T, \mathcal{G}, d, \mathcal{I}$)
  $G \leftarrow \text{argmax}_{i \in \{1,...,|\mathcal{G}|\}} S[i]$
  if $|G| = 1$ **then**
    $G^* \leftarrow G$
  else if $1 < |G|$ **then**
    $G^* \leftarrow \text{argmin}_{i \in \{1,...,|G|\}} S[i]$
  end if
  return G^*
end function

models, respectively. In addition, Runge et al. [45] 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 should be evaluated on a large number of distinct structures so as to reduce biased conclusions. In this context, one may consider the number of variables in the model, the density of causal structure, the dependency type of equations and the coefficients defining those equations. A third and similarly related desideratum is *model dimensionality*: since high-dimensionality is likely to affect the method"s performance, we should evaluate the method's performance when the number of variables increases. A last desideratum concerns *sample size*: it should be evaluated how the method fares against varying amounts of available data. In our setting, this concerns the number of times series observations [43, pp. 13–14]

9.2 Regression-based Conditional Independence Tests

The theoretical exposition of G-causality in Section 6 defined G-causes in terms of conditional independence relations. Strictly speaking, practical estimation of G-causality between X and Y then amounts to testing the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ against the alternative hypothesis $H_1 : X \not\perp\!\!\!\perp Y|Z$ given observed samples $\{x_i, y_i, z_i\}_{i=1}^n$. Unfortunately, such conditional tests are subject to the insistent *curse of dimensionality*: when the size of the conditional set Z grows, the amount of data required for evaluating the null hypothesis increases whilst the available data becomes increasingly sparse. In the literature, the dimensionality problem for conditional independence testing is countered with various approaches [43, 57, 25, 58, 28]. A general distinction within these approaches is that of model-free and regression-based approaches. Model-free approaches evaluate conditional independence directly without an assumption on the functional dependencies between variables. Contrastingly, regression-based methods impose dependencies $X = f_X(Z) + \epsilon_X$ and $Y = f_Y(Z) + \epsilon_Y$ to evaluate independence of X and Y given Z. Since the former are known to be computationally demanding and given the limited amount of computational resources at our disposal, we follow the regression-based approach.

Within the regression-based approach, $X \perp\!\!\!\perp Y|Z$ is assessed by evaluating independence of residuals from regressing X on Z and regressing Y on Z. Zhang et al. [57] has shown that in the case of identifiable additive noise models of the form $Y = f(X) + \epsilon$, 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 $X \perp\!\!\!\perp Y|Z$ [57, pp. 1250–1251]. In the assumed dependencies $X = f_X(Z) + \epsilon_X$ and $Y = f_Y(Z) + \epsilon_Y$, X and Y are assumed to be centered and ϵ_X and ϵ_Y are assumed to be independent and identically distributed. In the first step, models $\hat{f}_X$ and $\hat{f}_Y$ are estimated given 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 [43, 57]. 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 the distance correlation coefficient [43, p. 8].

9.3 Experimental Setup

9.3.1 Hyperparameters

#TODO: describe hyperparameters: T, d, $D = |E|/|V|$, c_j, dependency type

9.3.2 Synthetic Data

#TODO: describe generation of synthetic data i.r.t. desiderata from 9.1

A Notation Table

Notation	Interpretation
$\mathbf{X}$	random variable set $\{X_1, ..., X_n\}$ or time series $\{X^1, ..., X^m\}$
$\mathbf{T}$	time series set $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$ where $\mathbf{X}_i = \{X_i^1, ..., X_i^T\}$
τ	discrete time lag or window
γ	largest time gap between variables
$\mathscr{G}$	space of directed acyclic graphs
$\mathscr{D}$	space of data samples
$\mathcal{G}$	Markov equivalence class or set of candidate graphs, depending on context
$[G]_\sim$	Markov equivalence class of G
$\mathcal{D}$	data sample
ϵ	noise or error term
$\langle \cdot, \cdot \rangle_F$	Frobenius inner product of matrices
PA_i	parent set of V_i in $G = (V, A)$ defined as $\{V_j : (V_j, V_i) \in A\}$
$\sigma^*(V_i)$	descendant set defined as $\{V_j : \exists \pi = V_1 ... V_n : \pi_1 = V_i \text{ and } \pi_n = V_j\}$
pa_i	concrete value configuration of PA_i
$X \perp\!\!\!\perp Y \mid Z$	conditional independence of X and Y given Z
$X \to Y$	X is a cause of Y
$\mathcal{I} : \mathcal{X} \times \mathcal{X} \times \mathcal{X} \to [0, 1]$	conditional independence oracle that returns a probability
$\phi : \mathscr{D} \times \mathscr{G} \to \mathbb{R}$	general scoring function
$\kappa : \mathscr{D} \times \mathscr{A} \to \mathbb{R}$	non-symmetric bivariate causal measure
$\varkappa : \mathscr{D} \times \mathscr{G} \to \mathbb{R}$	causal scoring function

References

[1] 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.

[2] 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.

[3] Alexandra M. Carvalho and Tópicos Avançados. "Scoring Functions for Learning Bayesian Networks". In: 2009.

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

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

[6] David Maxwell Chickering. "A Transformational Characterization of Equivalent Bayesian Network Structures". In: *arXiv preprint arXiv:1302.4938* (2013).

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

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

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

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

[11] Marek J Druzdzel. "The Role of Assumptions in Causal Discovery". In: *8th Workshop on Uncertainty Processing (WUPES-09)*. Sept. 2009, pp. 57–68. URL: http://d-scholarship.pitt.edu/6017/.

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

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

[14] 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. DOI: 10.1098/rsta.2011.0613.

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

[16] 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. DOI: https://doi.org/10.1016/j.laa.2014.02.032. URL: https://www.sciencedirect.com/science/article/pii/S0024379514000998.

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

[18] 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.

[19] C.W.J. Granger. "Testing for Causality: A Personal Viewpoint". In: *Journal of Economic Dynamics and Control* 2 (1980), pp. 329–352. ISSN: 0165-1889. DOI: https://doi.org/10.1016/0165-1889(80)90069-X. URL: https://www.sciencedirect.com/science/article/pii/016518898090069X.

[20] 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.

[21] 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.

[22] Abdulnasser Hatemi-j. "Asymmetric Causality Tests with an Application". In: *Empirical Economics* 43.1 (2012), pp. 447–456.

[23] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. *Causal Structure Learning*. 2017. DOI: 10.48550/ARXIV.1706.09141. URL: https://arxiv.org/abs/1706.09141.

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

[25] Patrik Hoyer et al. "Nonlinear Causal Discovery With Additive Noise Models". In: *Advances in neural information processing systems* 21 (2008).

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

[27] 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.

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

[29] 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.

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

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

[32] David Maxwell Chickering, David Heckerman, and Christopher Meek. "A Bayesian Approach to Learning Bayesian Networks with Local Structure". In: *arXiv e-prints* (2013), arXiv–1302.

[33] 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.

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

[35] 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.

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

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

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

[39] Judea Pearl. *Causality. Models, Reasoning, and Inference*. 2nd ed. Cambridge, UK: Cambridge University Press, 2009. ISBN: 978-0-521-89560-6. DOI: 10.1017/CBO9780511803161.

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

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

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

[43] J. 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. DOI: 10.1063/1.5025050. eprint: https://doi.org/10.1063/1.5025050. URL: https://doi.org/10.1063/1.5025050.

[44] 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.

[45] Jakob Runge et al. "Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets". In: *Science Advances* 5.11 (Nov. 2019). DOI: 10.1126/sciadv.aau4996. URL: https://doi.org/10.1126%2Fsciadv.aau4996.

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

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

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

[49] 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. DOI: 10.3390/e21111080. URL: https://doi.org/10.3390%2Fe21111080.

[50] 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.

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

[52] Peter Spirtes and Kun Zhang. "Causal Discovery and Inference: Concepts and Recent Methodological Advances". In: *Applied Informatics* 3 (Dec. 2016). DOI: 10.1186/s40535-016-0018-x.

[53] 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.

[54] 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.

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

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

[57] 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.

[58] Kun Zhang et al. "Kernel-Based Conditional Independence Test and Application in Causal Discovery". In: *arXiv preprint arXiv:1202.3775* (2012).