© 2017 Seyedjalal Etesami

CAUSAL STRUCTURE OF NETWORKS OF STOCHASTIC PROCESSES

BY

SEYEDJALAL ETESAMI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Industrial Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Associate Professor Negar Kiyavash, Chair
Adjunct Associate Professor Todd P. Coleman, University of California San Diego
Professor Rayadurgam Srikant
Assistant Professor Ali Belabbas
Assistant Professor Niao He

ABSTRACT

We propose different approaches to infer causal influences between agents in a network using only observed time series. This includes graphical models to depict causal relationships in the network, algorithms to identify the graphs in different scenarios and when only a subset of agents are observed. We demonstrate the utility of the methods by identifying causal influences between markets and causal flow of information between media sites.

We study the statistical and functional dependencies in network of processes. Statistical dependencies can be encoded by directed information graphs (DIGs) and functional relationships using functional dependency graphs (FDGs), both of which are graphical models where nodes represent random processes. DIGs are based on directed information that is an information theoretic quantity. To capture the functional dependencies in a dynamical system, we introduce a new measure in this work and show that the FDGs are a generalization of DIGs. We also establish sufficient conditions under which the FDG defined by our measure is equivalent to the DIG. As an example, we study the relationship between DIGs and linear dynamical graphs (LDGs), that are also a type of graphical models to encode functional dependencies in linear dynamical systems. In this case, we show that any causal LDGs can be reconstructed through learning the corresponding DIGs.

Another contribution is to propose an approach for learning causal interaction network of mutually exciting linear Hawkes processes. In such processes, a natural notion of functional causality exists between processes that is encoded in their corresponding excitation matrices. We show that such causal interaction network is equivalent to the DIG of the processes. Furthermore, We present an algorithm for learning the support of excitation matrix (or equivalently the DIG). The performance of the algorithm is evaluated for a synthesized multivariate Hawkes network as well as real world dataset.

We also study the problem of causal discovery in presence of latent variables, in which only a subset of processes can be observed. We propose an approach for learning latent directed polytrees as long as there exists an appropriately defined discrepancy measure between the observed nodes. Specifically, we use our approach for learning directed information polytrees. We prove that the approach is consistent for learning minimal latent directed trees. Furthermore, we study the problem of structural learning in vector autoregressive (VAR) models with latent variables. In this case, we extend the identifiability to a broader class of structures. In particular, we show that most of the causal structure of a VAR model can be recovered successfully when only the causal network among the latent variables is a directed tree.

Last but not least, we introduce a new statistical metric inspired by Dobrushin's coefficient [1] to measure the dependency or causal direction between variables from observational or interventional data. Our metric has been developed based on the paradigm that the conditional distribution of the variable of interest given all the direct causes will not change by intervening on other variables in the system. We show the advantageous of our measure over other dependency measures in the literature.

We demonstrate the effectiveness of the proposed algorithms through simulations and data analysis.

ACKNOWLEDGMENTS

Many people have contributed to making my time at the University of Illinois enjoyable and fruitful. It would not have been possible without the guidance and support of my advisor, Prof. Negar Kiyavash and my collaborators Prof. Todd Coleman, Prof. Kun Zhang, and Prof. Niao He. Prof. Kiyavash has always been a source of encouragement and inspiration for new paths to pursue as well as helping hands as I learned to walk down those paths. I am also grateful for my committee – Prof. Srikant, Prof Coleman, Prof. Belabbas, Prof. Olshevsky, and Prof. Niao for their time and insight.

Many friends at UIUC have made my experience amazing and colorful – Chris, Anh, Farzad, Yingxiang, Amiremad, Saber to name a few, as well as Amin, and Daniel. Most important of all has been the constant love and dedication of my parents, siblings, and my dear Christiane.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

We begin by describing the main focus of our work, causal inference, and its applications in different disciplines. We continue by summarizing our contributions and the related works. Finally, we conclude by introducing the notations used in the rest of this report.

1.1 Problem Overview and Significance

Research in many disciplines, including biology, economics, social sciences, computer science, and physics, involves studying large networks of interacting agents. The goal of this dissertation is to establish a framework to infer the causal structure in a network of interacting random processes/variables and to succinctly represent it using graphical models. Such a framework is comprised of three components: metrics to measure causal inference, well-defined graphical models that meaningfully represent causal influences between the variables, algorithms that identify such graphs when all or a subset of the processes are observed.

In systems where a notion of time (past/future) exists, the causal influences between the variables maybe categorized into strictly causal and simultaneous. In strictly causal systems, the direction of influences is only from past to present. Such influences govern phenomena in the real word, while the simultaneous effects are usually due to the following two artifacts: i) lack of a natural time axis or loss of it due to measurement effects (e.g, low resolution measurements); ii) existence of confounders that were not factored into the model. Yet both of aforementioned factors commonly occur in practice. As such a framework to capture both causal and simultaneous influences are essential. In this dissertation, we study two types of networks: those with a notion of time referred to as network of random processes or time series and those with only simultaneous influences referred to as network of random variables.

A simple example of network of time series arises in quantitative finance. A market analyst observes the value of different market indices or the price of different stocks for a period of time and his goal is to learn the causal influences between the financial institutions during the observation time. Such knowledge may be subsequently used to design investment strategies or regulatory actions. An example of network with only simultaneous influences is in biological gene perturbation dataset. In this experiment, the activities of different genes are observed or manipulated in order to identify the causal structure among them.

1.2 Our Contribution

Our first contribution is to study the connection between statistical and functional dependencies in a dynamical system. Most of the existing approaches to discover functional dependencies are based on intervention [2–4]. Yet it is often impossible to perform such interventional manipulations. In this dissertation, we

define a statistical measure that is able to capture the functional dependency among processes (variables) in dynamical systems. Subsequently, using this measure, we define a new type of graphical model, functional dependency graph (FDGs) that encodes such dependencies. While our metric is defined using basically an interventional paradigm, we establish a relationship between our measure and the directed information measure for capturing interdependencies in dynamical systems, which is calculated via mere observation [5]. More precisely, we show that the statistical dependency structure of a system (captured by DIG) does not necessary reveal all the functional dependencies of that system (captured by FDG) in general. We also introduce sufficient conditions under which the two graphical models (FDGs and DIGs) are equivalent. In other words, learning the statistical relationships is enough to identify the functional dependencies without any need for intervention.

Our second contribution is to propose an approach for learning causal interaction network of mutually exciting Hawkes processes. In multivariate or mutually exciting point processes, occurrence of an event (arrival) in one process affects the conditional probability of new occurrences that is captured by the excitation matrix of the network. We prove that for linear multivariate Hawkes processes, the causal relationships implied by the excitation matrix is equivalent to a specific factorization of the joint distribution of the system called *minimal generative model*. Minimal generative models encode causal dependencies based on a generalized notion of Granger causality, measured by causally conditioned directed information [6]. One significance of this result is that it provides a surrogate to directed information measure for capturing causal influences for Hawkes processes. We also provide an estimation method for learning the support of excitation matrices with exponential form using second-order statistics of the Hawkes processes [7].

Our third contribution is to develop an approach for structure learning of directed graphical model with polytree structure, when only a subset of random *processes* are observed. Specifically, we will consider the scenario of latent directed information polytrees, where the directed information graph representing observed and unobserved processes is a tree with multiple roots. Learning such graphs requires both finding the number of hidden processes as well as recovering the connections among all hidden and observed nodes. To perform the learning task, we define a discrepancy measure between nodes of a directed polytree and introduce an algorithm that identifies the structure given the discrepancies between only a subset of nodes (observed nodes). Furthermore, we study the problem of structural learning in vector autoregressive (VAR) models with latent variables. We show that the entire causal structure can be identified successfully when the topology of the VAR model is a directed tree and every latent node has at least two children and two parents. Extending this result, we propose a set of sufficient conditions under which the causal influences from latent to observed nodes, between observed nodes, and also between latent nodes can be recovered when only the causal structure between the latent nodes is a directed tree [8]. We also propose an algorithm that finds all possible minimal latent networks (networks with minimum number of latent nodes) if there exists at most one directed path of each length between any two observed nodes through the latent part.

Our last but not least contribution is the introduction of a statistical metric inspired by Dobrushin's coefficient [1] to measure the dependency or causal direction between variables from observational or interventional data. Our metric has been developed based on the paradigm that the conditional distribution of the variable of interest given all the direct causes will not change by intervening on other variables in the system. Despite other dependency measures in the literature, this measure does not have shortcomings in detecting direct influences and it has the ability for group selection in order to have effective interventions. We show the applicability of the proposed algorithms through simulating both synthetic and real-world dataset.

1.3 Literature Review

Causality Granger causality and the principle of intervention are two of most commonly used frameworks to identify causal interactions in a network. The principle of intervention or the Pearl's notion of causality [9] infers the causal relationships by fixing certain variables and allowing others to change, to see how these changes influence the statistics or the values of the other variables. This method was developed mainly based on structural equation modeling (SEM).

The idea of Granger causality is that a random process $\mathbf{X}$ is causing $\mathbf{Y}$, if incorporating the past of $\mathbf{X}$ improves the prediction of the future of $\mathbf{Y}$. Granger [10, 11] proposed a framework to capture such influences in an auto-regressive (linear) setting. The work in [12] extended previous works on linear setting such as [13–15] to more general settings using conditional independence tests known as "strong Granger causality" [16, 17].

Sims [18] proposed an alternative test for causality of autoregressive time series, equivalent to Granger's. He proposed that $\mathbf{X}$ influences $\mathbf{Y}$ if X_t is correlated with the whole future Y_{t+1}^n given the past. The works in [16, 17] developed general forms of Granger and Sims causality using conditional independencies. These works only discussed testing the presence of statistical relationships, not measuring the strengths of such relationships.

Graphical Models In order to visualize the causal structure in a network of random variables or time series, several graphical models have been developed. Bayesian networks [9] and ancestral graphs [19] are the two main graphical models for network of random variables. We will briefly explain the Bayesian network in Chapter 2 but refer the interested reader to [4] and [20] for more details. Dynamic Bayesian network (DBN) [21] is another class of graphical models that extends Bayesian networks to model probability distributions over semi-infinite collection of random variables. For example, Hidden Markov Models (HMMs) can be represented as DBNs. Since the size of DBNs depends on the time-homogeneity and the Markov order of the time series, in general, the graphs can grow with time. Thus, they are not well suited for providing succinct visualization of relationships between the past and the future of time series. As an example, the DBN graph of a vector autoregressive (VAR) process $\underline{\mathbf{X}}(t) \in \mathbb{R}^m$ of order L requires mL nodes [13]. Directed information graphs (DIGs), the alternative graphical model that we study, represent each random process as a node in the graph. Therefore, their size neither depends on the Markov order of processes nor the time (for the VAR example above, the size is m).

As part of the effort to generalize Granger's causality to more general settings, another class of graphical models called the Granger causality graph was developed [13–15, 22]. This class of graphs consists of a mix of both directed and undirected edges for multivariate autoregressive time series and the nodes in the graph represent the time series.

Causality Measures Along side developing different paradigms to define the causal influences, several measures have also been developed to capture such influences.

Average causal effect between X and Y that is defined based on do-operation [23] and it is given by [24], $ACE(X \to Y) := P(Y|do(X) = 0) - P(Y|do(X) = 1)$. Here, it is assumed that X is binary. Since this measure focuses on pairwise influences, it is not suitable for capturing influences in a network. Other measures are conditional mutual information [25] and information flow [26] that are defined analogous to each other. The former compares two conditional probability measures without do-operation and the latter compares them after do-operation. Recently, the authors in [27] developed a new measure based on four postulates

to quantify the causal influence. Their measure is similar to the information flow as defined in [28]. By studying the limitations of these measures, we will propose a new dependency measure in Section 7.2.

Influence measures that have been developed to quantify causal influence between time series are directed information and transfer entropy.

Directed information (DI) is an information-theoretic quantity that generalizes Granger causality beyond linear models [29,30]. DI was first introduces by Marko [31] and then later formalized by Massey [32]. Marko assumed there is no instantaneous influence between time series, and showed the mutual information between the input $\mathbf{X}$ and the output $\mathbf{Y}$ decomposes to a sum of directed information terms from $\mathbf{X}$ to $\mathbf{Y}$ and from $\mathbf{Y}$ to $\mathbf{X}$. Since then, DI has been used in many applications to infer causal relationships. For example, it has been used for analyzing neuroscience data [33–37], gene regulatory data [38], and video recordings [39]. Directed information graphs (DIGs) define a graphical model that captures the generalization of Granger causality using the DI metric among stochastic processes [40]. DIGs subsume Granger causal graphs. It was shown in [41] that in order to guarantee uniqueness of directed information graphs, the joint dynamics of the system must be strictly causal.

Transfer entropy, introduced by [42], is another measure of causality in the literature [43, 44]. The relationship of Granger causality and transfer entropy is discussed in detail in [40, 45]. Transfer entropy is only defined for processes that satisfy Markov property, in which case the DI can be written as a sum of a sequence of transfer entropies.

Causal Learning Learning causal influences of a network of random variables may be done via passive learning techniques that use mere observation of the network's autonomous behavior [46–48]. To mention some, [46,49] proposed an algorithm called LiNGAM that relies on a statistical method known as independent component analysis (ICA). LiNGAM can discover the complete causal structure of continuous-valued data, under the assumptions that the data generating process is linear, there are no unobserved confounders, and disturbance variables have non-Gaussian distributions of non-zero variances.

On the other hand, active learning approaches allow for experimental manipulations (interventions). That is, the learner may actively intervene and control some variable in the system and observe the effects on other variables [2,50]. The difference between two aforementioned approaches has been compared to learning from watching and learning by doing [23, 51, 52].

The authors in [41] proposed various algorithms to learn the causal structure of a stochastic systems using directed information quantity. They also developed several efficient algorithms that recover the DIG when upper bounds on the in-degrees are known.

Most of the learning methods in the literature for causal discovery of time series rely on finding a surrogate to DI or transfer entropy. For instance, [53] proposed linear dynamical graphs as a graphical model to describe the causal interactions in linear dynamical systems which depend only on the coefficient matrices of the model. It was shown in [54] that such graphical model are equivalent to DIGs. Moreover, [53] developed an algorithm based on Wiener filtering to learn the causal structure of such systems when the underlying network is a tree. Later, [55] extended that result to a more general setting in which the causal structure does not have cycles. Independently, [56] investigated learning tree structured networks of linear dynamical systems.

Another parametric dynamical systems in which recovering the corresponding causal structure can be done by learning a set of parameters in the model, excitation matrix, are multivariate Hawkes processes (MHP). We will study such processes in chapter 4 but for more details, we refer the readers to [57]. Several

approaches have been developed in the literature for learning the excitation matrix of an MHP. Most of the existing works assume that the entries of the excitation matrix belong to a set of predefined parametric functions, e.g., the exponential functions in [58–62] and the power-law functions in [63]. For instance, [64,65] considered learning the excitation matrix of symmetric Hawkes processes. In a symmetric Hawkes process, it is assumed that the excitation functions are exponential, the Laplace transform of the excitation matrix can be factored into product of a diagonal matrix and a constant unitary matrix, and the expected values of all intensities are the same.

The authors in [66] proposed the first non-parametric model of one dimensional Hawkes process based on ordinary differential equation. This later has been extended to multi-dimensional case in [60, 67, 68]. For example, in [60], the authors assumed that the excitation functions can be written as a linear combination of a set of basis, then the coefficients as well as the basis functions were being iteratively updated such that the likelihood function of the MHP is maximized. A similar approach were being used in [69] for learning sparse MHPs. One potential drawback of such adaptive approaches is that they require a set of i.i.d. samples for their training phases, which can be hard or costly to acquire in some scenarios.

The authors in [70] proposed a non-parametric estimator that solves a set of Wiener-Hopf equations. Another non-parametric strategy is the contrast function-based estimation in [71] that estimates the excitation functions by linear combinations of a fixed dictionary. To force sparsity in this method, they used an ℓ_1-penalized least squares criterion to learn the coefficients. The work in [72] proposed discretizing the point process by considering the increments over equidistant time points and then fitting a vector autoregressive model by least squares method. This discretization causes approximation error. To avoid that, [73] decomposed the excitation functions into basis functions using polynomial approximation. Finally, [61] proposed an online learning algorithm that simultaneously learns the excitation matrix and tracks the dynamic (intensity functions) of an MHP. However, they assumed that the triggering function are exponential with known exponent parameters.

Tree Causal Structures Polytree models have applications in real world. For instance, polytrees were implemented to enhance caching strategies in distributed databases [74]. Dependency polytrees were also applied to develop an inference framework that optimizes hardware components according the performance and price of architectures [75]. In [76], the authors applied polytree structure graphical model for ozone prediction in Mexico City, where ozone level is used as global indicator for the air quality. Moreover, Protein signaling pathways might be modeled by causal polytrees. For instance, NFkB protein signaling pathway, which activates mammalian immune system cells to produce antibodies against inflammation [77]. In [78], authors characterize dependency graphical models that are isomorphic to polytree graphs.

Even if the underlying structure is not a tree, there are efficient algorithms that approximate the underlying causal structure by the best directed tree such as [79–82]. In [79], authors introduce an algorithm similar to Chow-Liu algorithm [83] to construct a polytree-shaped network to approximates the probability distribution of the network.

Since in a directed polytree, a natural notion of hierarchy (depth) exists, polytree approximation can be used to infer the influence hierarchy among the processes. Such an inference could be helpful in, for instance, determining root causes of events or where to intervene for regulatory action such that it could effectively trickle down.

Latent Graphical Models In practice it is usually difficult and even impossible to collect all relevant time series when doing causal analysis on given ones. Herein, we review some of the previous relevant latent learning algorithms. We categorize the learning approaches to graphs that represent conditional independence relationships among (I) random processes such as DI graphs and (II) random variables such as Bayesian networks or ancestral graphs. Note that some of the learning methods proposed for the latter can be extended to the former, but the methods such as the one presented in this work that requires the notion of time among processes are only applicable to the first type of graphical models. Timing not only aids with causal inference, it also has been proven useful for general other signaling, inference, and control purposes complex network [84–86].

One approach for learning latent graphical models is to fix the number of latent vertices and the structural relationships between latent and observed variables and subsequently use the expectation maximization (EM) algorithm to estimate the model parameters. Given that often the optimization is over a non-convex function, the performance depends on initialization, and the algorithm may get trapped in sub-optimal local minima [87].

The work in [88] considered learning a VAR model with hidden components. The model is identifiable under the assumptions that connections between observed variables are sparse and each latent variable interacts with many observed variables. Two other papers in [89] and [90] applied a method based on EM algorithm to infer properties of partially observed Markov processes. The work in [89] relaxed the finite-state condition required by [90] and provided sufficient conditions under which the partially observed Markov process is identifiable. Essentially, they showed that when the noise is independent and non-Gaussian or the observed variables do not influence the hidden variables, the model is identifiable.

Authors in [91] showed that if the exogenous noises of a VAR model are independent non-Gaussian and additional so-called genericity assumptions hold, then the links between the observed processes as well as the links from latent to observed processes are uniquely identifiable. They presented a result in which they allowed Gaussian noises in their VAR model and obtained a set of conditions under which they can recover the links among the observed processes up to several candidate. Their learning approach is also based on EM and approximately maximizes the likelihood of a parametric VAR model with a mixture of Gaussians as noise distribution. Somewhat similar results for linear models but with random variables were presented in [92].

In [93], the authors considered learning latent graphical models in the setting in which the latent and observed variables are jointly Gaussian, the conditional statistics of the observed variables given the latent variables is a sparse graph, and the number of latent nodes is small relative to the number of observed variables. They proposed a tractable convex program based on regularized maximum-likelihood for latent-variable graphical model selection. Note that the proposed approach in this work does not specify any model for the joint distribution between the observed and the latent variables. Furthermore, it may have a relatively large number of latent variables.

An alternative method was proposed in [94] that is based on a greedy, combinatorial heuristic. It assigns latent variables to groups of observed variables via clustering of the observed ones. This approach has no consistency guarantees. In contrast, our approach guarantees consistency under mild assumptions.

Recently, the quartet[1]-based approaches were applied to learn linear multivariate tree models when only the leaves are observed [95]. In such trees, nodes are multivariate random vectors. In [95], it is further assumed that the conditional expected value of each node given the parent is a linear function of its parents. Recursive

[1] A quartet is an un-rooted binary tree on a set of four observed nodes.

grouping (RG) and Chow-Liu recursive grouping (CLRG) proposed in [96] are two other distance-based learning algorithms that can recover latent Markov graphical models, in which some of the observed nodes are internal. Both RG and CLRG can only recover latent models on a set of hidden and observed random variables that are jointly Gaussian or have a symmetric discrete joint distribution. No such restrictions on the joint are required in the proposed approach in this thesis.

A provably sound[2] algorithm known as FCI was developed for leaning maximal ancestral graphs (MAG) [19, 97]. A MAG is a mixed graph consisting of both directed and undirected edges on the set of observable variables that probabilistically represents the conditional independence among both latent and observable variables in an accompanying DAG. More precisely, consider any DAG (e.g. G over $V = O \cup L \cup S$) that encodes a set of conditional independence relations among nodes in V, where O and L denotes the set of observed and latent variables, respectively, and S denotes a set of unobserved selection variables to be conditioned upon. Suppose there exists a MAG, $M(G)$, over O such that for any three disjoint sets of variables $A, B, C \subseteq O$, A and B are conditionally independent given $C \cup S$ in G if and only if A and B are conditionally independent given C in $M(G)$. In this case, $M(G)$ is said to probabilistically represent G. FCI algorithm does not recover the latent nodes and the relations between latent and observed nodes, but rather the MAG on the set of observed nodes. Our algorithm, on the other hand, recovers the graph on both observed and latent nodes.

Classical approaches to learning latent graphical models, in which nodes represent random variables are of the following flavors; latent cluster models (LCMs) learn a tree structured Bayesian network, in which only one single hidden variable exists [98]. Hierarchical latent class (HLC) models generalize the previous model by allowing multiple hidden variables but they confine the observed variables to the leaves of the tree [99]. Since in HLC models, root walking[3] leads to a marginally equivalent model (two models are marginally equivalent if they share the same conditional distribution between the observable variables given the latent variables), it is impossible to learn edge orientation from the data. Furthermore, learning algorithms for such models has a greedy structure, which is both computationally expensive and not guaranteed to be consistent.

Other popular learning methods for latent Markov graphical models use quartet-based distances [100, 101] to discover the structure. Quartet-based methods first construct a set of quartets for all subsets of four observable nodes and then combine them to form a latent tree. It is known that the problem of determining a latent tree that agrees with the maximum number of quartets is NP-hard [102]. As a result many heuristics have been developed [103], [104]. Authors in [105] propose a quartet based approach which uses rank characterization of the tensor associated with the marginal distribution of a quartet. This characterization allows them to design a nuclear norm based test for resolving quartet relations. Additionally, in practice, quartet-based methods are often much less accurate than neighbor-joining (NJ) method [106]. NJ [107] is another distance-based algorithm that proceeds by repeatedly pairing the two *closest* nodes from the list by adding a new latent node as their parent and replacing the pair with the newly added node. Both NJ and the quartet-based methods rely on the existence of a notion of distance between nodes of a tree, which may not exist in many practical scenarios. In this work, we propose a new method based on a discrepancy measure between the observed nodes, which is not required to be a distance measure.

[2]Soundness is defined as follows: given a perfect oracle of conditional independence, the algorithm outputs the Markov equivalence class of the true causal maximal ancestral graph.

[3]Root walking is an operation on a directed tree that reverses an arrow which goes from the root to one of its neighbors.

1.4 Notation and Definitions

- For a sequence $a_1, a_2, \ldots$, denote $(a_i, \ldots, a_j)$ as a_i^j. Denote $[m] := \{1, \ldots, m\}$, $-\{j\} := [m] \setminus \{j\}$, and the power set of $[m]$ by $2^{[m]}$. We will consider m random processes where the ith (with $i \in \{1, \ldots, m\}$) random process at time t takes values in a Borel space X. Denote the ith random variable at time t by $X_{i,t} : \Omega \to \mathsf{X}$ and the whole collection of m random processes by $\underline{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)^\top$.
- For any Borel space Z, denote its Borel sets by $\mathcal{B}(\mathsf{Z})$ and the space of probability measures on $(\mathsf{Z}, \mathcal{B}(\mathsf{Z}))$ as $\mathcal{P}(\mathsf{Z})$.
- Consider two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{P}(\mathsf{Z})$. $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ ($\mathbb{P} \ll \mathbb{Q}$) if $\mathbb{Q}(A) = 0$ implies that $\mathbb{P}(A) = 0$ for all $A \in \mathcal{B}(\mathsf{Z})$. If $\mathbb{P} \ll \mathbb{Q}$, denote the Radon-Nikodym derivative as the random variable $\frac{d\mathbb{P}}{d\mathbb{Q}} : \mathsf{Z} \to \mathbb{R}$ that satisfies

$$\mathbb{P}(A) = \int_{z \in A} \frac{d\mathbb{P}}{d\mathbb{Q}}(z) \mathbb{Q}(dz), \quad A \in \mathcal{B}(\mathsf{Z}).$$

  For example, for almost all $\mathbf{x}$, we have $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \ll P_{\mathbf{Y}}$.
- The *Kullback-Leibler divergence* between $\mathbb{P} \in \mathcal{P}(\mathsf{Z})$ and $\mathbb{Q} \in \mathcal{P}(\mathsf{Z})$ is defined as $D(\mathbb{P}\|\mathbb{Q}) := \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right]$ if $\mathbb{P} \ll \mathbb{Q}$ and ∞ otherwise. Moreover, the conditional KL divergence is given by

$$D\big(P_{\mathbf{Y}|\underline{\mathbf{W}}} \| Q_{\mathbf{Y}|\underline{\mathbf{W}}} | P_{\underline{\mathbf{W}}}\big) = \int_{\underline{\mathsf{W}}} D\big(P_{\mathbf{Y}|\underline{\mathbf{W}}=\underline{\mathbf{w}}} \| Q_{\mathbf{Y}|\underline{\mathbf{W}}=\underline{\mathbf{w}}}\big) \, P_{\underline{\mathbf{W}}}(d\underline{\mathbf{w}}). \tag{1.1}$$

  Note that $D\big(P_{\mathbf{Y}|\underline{\mathbf{W}}} \| Q_{\mathbf{Y}|\underline{\mathbf{W}}} | P_{\underline{\mathbf{W}}}\big) = 0$ if and only if $P_{\mathbf{Y}|\underline{\mathbf{W}}=\underline{\mathbf{w}}}(d\mathbf{y}) = Q_{\mathbf{Y}|\underline{\mathbf{W}}=\underline{\mathbf{w}}}(d\mathbf{y})$ with $P_{\underline{\mathbf{W}}}$ probability one.
- With slight abuse of notation, we denote *causally conditioned* distribution [108] of $\mathbf{Y}$ given $\mathbf{X}$ as

$$P_{\mathbf{Y}\|\mathbf{X}}(d\mathbf{y}\|\mathbf{x}) := P_{\mathbf{Y}\|\mathbf{X}=\mathbf{x}} := \prod_{t=1}^{n} P_{Y_t|Y^{t-1},X^{t-1}}\big(dy_t | y^{t-1}, x^{t-1}\big). \tag{1.2}$$

  Note that in (1.2) the future (x_t^n) is not conditioned on. Through this dissertation, for simplicity, we will drop the term $(dy_t|y^{t-1}, x^{t-1})$ from the probabilities.

$$P_{\mathbf{X}_j|| \, \underline{\mathbf{X}}_{-\{j\}}} := \prod_{t=1}^{n} P_{X_{j,t}|X_{1,1}^{t-1},\ldots,X_{j,1}^{t-1},\ldots,X_{m,1}^{t-1}}. \tag{1.3}$$

- In equation (1.3) the random process $\mathbf{X}_j$ depends on the set of random processes $\underline{\mathbf{X}}_{-\{j\}}$ by one time delay. This notation may be generalized to d-step delay ($d \in \mathbb{N}$). We denote the causal conditioned distribution with d-step delay as follows

$$P_{\mathbf{X}_j||_d \, \underline{\mathbf{X}}_{\mathcal{K}}} := \prod_{t=1}^{n} P_{X_{j,t}|X_{j,1}^{t-1},\underline{X}_{\mathcal{K},1}^{t-d}}, \tag{1.4}$$

  where $\underline{X}_{\mathcal{K},1}^{t-d}$ stands for $(X_{k_1,1}^{t-d}, \ldots, X_{k_s,1}^{t-d})$. Figure 1.1 illustrates the time dependencies between two processes for $d = 1$ and $d = 3$.
  It is easy to see that for $d = 1$, equation (1.4) becomes Kramer's causal conditioned distribution (1.3). For simplicity, we will write $P_{\mathbf{X}||\mathbf{Y}}$ instead of $P_{\mathbf{X}||_1\mathbf{Y}}$.
- Let $\underline{\mathbf{W}} = \underline{\mathbf{X}}_{\mathcal{A}}$ for some $\mathcal{A} \subseteq -\{i, k\}$. The mutual information, *directed information* [31], and causally

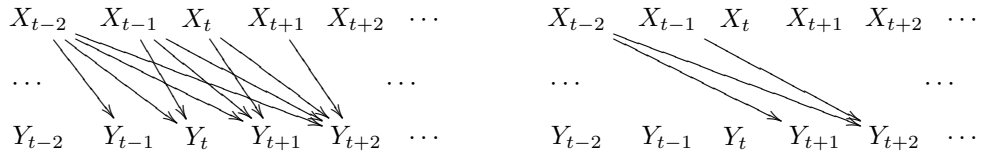


Figure 1.1. Time dependencies between random processes $\mathbf{X}$ and $\mathbf{Y}$ for a unit delay and 3-step delay. Directed edges show the causal conditioned dependencies between variables in process $\mathbf{Y}$ and the corresponding variables in process $\mathbf{X}$.

conditioned directed information [108] are given by

$$\mathrm{I}(\mathbf{X};\mathbf{Y}) := D\big(P_{\mathbf{Y}|\mathbf{X}}\|P_{\mathbf{Y}}|P_{\mathbf{X}}\big) = \textstyle\sum_{t=1}^{n} \mathrm{I}(X^n;Y_t|Y^{t-1}), \tag{1.5}$$

$$\mathrm{I}(\mathbf{X}\to\mathbf{Y}) := D\big(P_{\mathbf{Y}\|\mathbf{X}}\|P_{\mathbf{Y}}|P_{\mathbf{X}}\big) = \textstyle\sum_{t=1}^{n} \mathrm{I}(X^{t-1};Y_t|Y^{t-1}), \tag{1.6}$$

$$\mathrm{I}(\mathbf{X}\to\mathbf{Y}\|\underline{\mathbf{W}}) := D\big(P_{\mathbf{Y}\|\mathbf{X},\underline{\mathbf{W}}}\|P_{\mathbf{Y}\|\underline{\mathbf{W}}}|P_{\mathbf{X},\underline{\mathbf{W}}}\big) = \textstyle\sum_{t=1}^{n} \mathrm{I}(X^{t-1};Y_t|Y^{t-1},\underline{W}^{t-1}). \tag{1.7}$$

Conceptually, mutual information and directed information are related. However, while mutual information quantifies statistical correlation, directed information quantifies statistical *causation*. Note that $\mathrm{I}(\mathbf{X};\mathbf{Y}) = \mathrm{I}(\mathbf{Y};\mathbf{X})$, but $\mathrm{I}(\mathbf{X}\to\mathbf{Y}) \neq \mathrm{I}(\mathbf{Y}\to\mathbf{X})$ in general.

- Consequently, the directed information rate and the conditional directed information rate are defined, respectively, as

$$I_\infty(\mathbf{X}\to\mathbf{Y}) := \lim_{t\to-\infty} \frac{1}{n-t+1} I(\mathbf{X}_t^n \to \mathbf{Y}_t^n),$$

$$I_\infty(\mathbf{X}\to\mathbf{Y}||\underline{\mathbf{W}}) := \lim_{t\to-\infty} \frac{1}{n-t+1} I(\mathbf{X}_t^n \to \mathbf{Y}_t^n||\underline{\mathbf{W}}_t^n).$$

  Since in this work, the length of processes are assumed to be finite, $n<\infty$, the directed information and conditional directed information are finite. Thus, it suffices to work with (1.6) and (1.7). If $n\to\infty$, the same proof ideas hold by replacing (1.6) and (1.7) with the aforementioned information rates instead.

- A *path* between two nodes in an undirected graph is a sequence of distinct vertices such that every vertex in the sequence is adjacent to its predecessor and its successor, all nodes except the end-nodes on a path are called internal nodes. Two paths are called *disjoint* if they do not have any internal vertex in common. A path of the form $v\to\cdots\to u$, on which every edge is an arrow with the arrowheads pointing toward u is a *directed path* form v to u. The set of *parents* and *children* of a node v in $\overrightarrow{T}$ are defined, respectively, by

$$\mathcal{PA}(v) := \{u\in V : (u,v)\in \overrightarrow{E}\}, \quad \mathcal{CH}(v) := \{u\in V : (v,u)\in \overrightarrow{E}\}. \tag{1.8}$$

  Node w is called an *ancestor* of node v in $\overrightarrow{T}$ if there exists a directed path from w to v. In this case, v is called a *descendant* of w. A node v is a non-descendant of w, if there is no directed path from w to v.

CHAPTER 2

GRAPHICAL MODELS

In this chapter, we review the most commonly used graphical models for succinctly representing the causal structure of networks of stochastic processes; Bayesian networks, minimum generative model graphs and directed information graphs.

2.1 Bayesian Networks

A Bayesian network is a graphical model that represents the conditional independencies among a set of random variables via a directed acyclic graph (DAG) [47]. A set of random variables $\underline{X}$ is Bayesian with respect to a DAG $\overrightarrow{G}$, if

$$P(\underline{\mathbf{X}}) = \prod_{i=1}^{m} P(\mathbf{X}_i | \underline{\mathbf{X}}_{\mathcal{P}_{\mathcal{A}_i}}). \tag{2.1}$$

Up to some technical conditions [109], this factorization is equivalent to the *causal Markov* condition. Causal Markov condition states that a DAG is only acceptable as a possible causal hypothesis if every node is conditionally independent of its non-descendant given its parents.

Corresponding DAG of a joint distribution possesses *Global Markov* condition if for any disjoint set of nodes $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ for which $\mathcal{A}$ and $\mathcal{B}$ are *d-separated* by $\mathcal{C}$, then $\underline{\mathbf{X}}_{\mathcal{A}} \perp\!\!\!\perp \underline{\mathbf{X}}_{\mathcal{B}} | \underline{\mathbf{X}}_{\mathcal{C}}$, i.e.,

$$I(\underline{\mathbf{X}}_{\mathcal{A}}; \underline{\mathbf{X}}_{\mathcal{B}} | \underline{\mathbf{X}}_{\mathcal{C}}) = 0.$$

Before defining d-separation in DAGs, let us introduce concept of a collider. In a DAG, a non-endpoint vertex c on a path is said to be a *collider* if both edges are directed toward c on this path. For example, X in Figure 2.2(a) is a collider on the path $Y \to X \leftarrow Z$.

Definition 1. *Let $\overrightarrow{G} = (V, \overrightarrow{E})$ be a DAG and U, W, and Z be three disjoint subsets of V. Z d-separates U from W, if for every path (not necessarily directed) from a node in U to a node in W, there exists a node c such that either*

1. *c is not a collider and it belongs to Z or*
2. *c is a collider and neither c nor any of c's descendants are in Z.*

Remark 1. *It is possible that two DAGs $\overrightarrow{G}_1$ and $\overrightarrow{G}_2$ with the same vertex set capture the same independence relations, i.e., for all disjoint sets U, W, and Z, where U and W are non-empty, Z d-separates U from W in $\overrightarrow{G}_1$ if and only if Z d-separates U from W in $\overrightarrow{G}_2$. In this case, it is said that $\overrightarrow{G}_1$ and $\overrightarrow{G}_2$ are Markov*

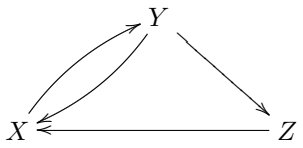


Figure 2.1. A graphical model of the causal influences in the stochastic dynamical system of Example 1.

equivalent. For example, two DAGs in Figure 2.2 are Markov equivalent. [110] gives simple conditions for determining whether two DAGs are Markov equivalent.

It is shown in [109] that causal Markov condition and Global Markov condition are equivalent.

Moreover, a joint distribution is called *faithful* with respect to a DAG if all the conditional independence (CI) relationships implied by the distribution can also be found from its corresponding DAG using d-separation and vice versa[1] [4].

2.2 Minimal Generative Model Graphs

Dynamical systems have a natural representation that is the coupled differential equations which characterize the dynamics of the system over time. Such a representation explicitly describes the inter-dependencies among the processes.

Example 1. *Let x_t, y_t, and z_t be three processes comprising a deterministic dynamical system. Suppose that the differential equations*

$$dx = g_1(x, y, z)dt, \quad dy = g_2(x, y)dt, \quad dz = g_3(y, z)dt,$$

are known. For small Δ, the system becomes

$$x_{t+\Delta} = x_t + \Delta g_1(x_t, y_t, z_t), \quad y_{t+\Delta} = y_t + \Delta g_2(x_t, y_t), \quad z_{t+\Delta} = z_t + \Delta g_3(y_t, z_t). \tag{2.2}$$

A natural graphical representation simply depicts the remaining dependencies. See Figure 2.1. Note that for sufficiently small Δ, (2.2) is strictly causal (e.g. $x_{t+\Delta}$ depends on y_t but not $y_{t+\Delta}$).

Consider a stochastic dynamical system $\underline{\mathbf{X}}$ of m processes with joint distribution $P_{\underline{\mathbf{X}}}$. The dynamics of the system are fully described by $P_{\underline{\mathbf{X}}}$. First, we can factorize $P_{\underline{\mathbf{X}}}$ over time as follows

$$P_{\underline{\mathbf{X}}}(d\underline{\mathbf{x}}) \;=\; \prod_{t=1}^{n} P_{\underline{\mathbf{X}}_t|\underline{\mathbf{X}}^{t-1}}.$$

If $P_{\underline{\mathbf{X}}}$ is strictly causal, then similar to difference equations (2.2) in Example 1, it can be factorized over the indices of the processes,

$$P_{\underline{\mathbf{X}}}(d\underline{\mathbf{x}}) = \prod_{t=1}^{n}\prod_{i=1}^{m} P_{\underline{\mathbf{X}}_{i,t}|\underline{\mathbf{X}}^{t-1}} = \prod_{i=1}^{m} P_{\mathbf{X}_i\|\underline{\mathbf{X}}_{-\{i\}}}. \tag{2.3}$$

[1] The set of distributions that do not satisfy this assumption has measure zero [111].

Notice that each $\mathbf{X}_i$ is still conditioned on the full past of every other process. We will assume that $P_{\underline{\mathbf{X}}}$ is both non-degenerate and strictly causal.

Assumption 1. *For the remainder of this dissertation, we only consider joint distributions that are strictly causal and non-degenerate, i.e., there exists a measure ϕ such that $P_{\underline{\mathbf{X}}}$ is absolutely continuous with respect to ϕ ($P_{\underline{\mathbf{X}}} \ll \phi$) and $\frac{dP_{\underline{\mathbf{X}}}}{d\phi}(\underline{\mathbf{x}}) > 0$ for all $\underline{\mathbf{x}}$ in the support of $P_{\underline{\mathbf{X}}}$.*

Remark 2. *Assumption 1 is to avoid degenerate cases that arise with deterministic relationships. Moreover, this assumption holds for any continuous-time generative model described by coupled stochastic differential equations such as the one presented in Example 1.*

Next, we remove unnecessary dependencies between processes in (2.3). For each process $\mathbf{X}_i$, let $A(i) \subseteq -\{i\}$ denote a subset of processes that does not contain i-th process and define the corresponding induced probability measure P_A,

$$P_A(d\underline{\mathbf{x}}) := \prod_{i=1}^{m} P_{\mathbf{X}_i \| \underline{\mathbf{X}}_{A(i)}}.$$

We want to pick the sets $\{A(i)\}_{i=1}^m$ so that their cardinalities are small, while still capturing the full dynamics[2] of $P_{\underline{\mathbf{X}}}$,

$$D\big(P_{\underline{\mathbf{X}}} \| P_A\big) = 0. \tag{2.4}$$

In Example 1, we have $A(X) = \{Y, Z\}$, $A(Y) = \{X\}$, and $A(Z) = \{Y\}$.

Definition 2. *Under Assumption 1, for a joint distribution $P_{\underline{\mathbf{X}}}$, a* minimal generative model *is a function $A : [m] \to 2^{[m]}$ where the cardinalities of the sets $\{|A(i)|\}_{i=1}^m$ are minimal and (2.4) holds.*

Minimal generative models represent reduced factorizations of the joint distribution of the system. They encode causal relationships by only selecting those subsets of processes that are necessary and sufficient to describe the full dynamics. This model was motivated by reducing coupled differential equations for deterministic systems.

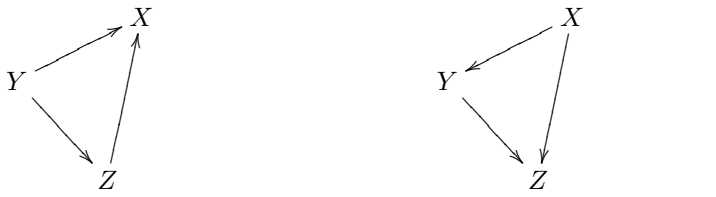
Definition 3. *A* minimal generative model graph *is a directed graph for a minimal generative model, where each process is represented by a node, and there is a directed edge from $\mathbf{X}_k$ to $\mathbf{X}_i$ for $i, k \in [m]$ if and only if $k \in A(i)$.*

Note that unlike Bayesian networks, minimal generative model graphs can have directed loops, as is the case in Figure 2.1.

2.3 Directed Information Graphs

In 1969, motivated by earlier work by Wiener [112], Nobel laureate Clive Granger proposed a framework for identifying when one process statistically "causes" another [11]: "We say that $\mathbf{X}$ is causing $\mathbf{Y}$ if we are better able to predict [the future of] $\mathbf{Y}$ using all available information than if the information apart from [the past

[2]The $A(i)$'s are defined over the whole time horizon. The $A(i)$'s could be defined over sliding windows of time, but that is outside the scope of this dissertation.



$(a) P_{X,Y,Z} = P_Y P_{Z|Y} P_{X|Y,Z}$. $(b) P_{X,Y,Z} = P_X P_{Y|X} P_{Z|X,Y}$.

Figure 2.2. Two possible Bayesian networks for one joint distribution.

of] $\mathbf{X}$ had been used." While this definition is general, its first formulations have mostly been restricted to specific classes of models, such as autoregressive linear models.

It is shown in [113] that for any class of distributions, the directed information explicitly quantifies Granger's statement in the setting of sequential prediction with causal side information. We now define a graphical model using directed information.

Definition 4. *For a set of random processes $\underline{\mathbf{X}}$, the* directed information graph *is a directed graph where each node represents a process and there is a directed edge from process $\mathbf{X}_k$ to process $\mathbf{X}_i$ (for $i, k \in [m]$) if and only if*

$$\mathrm{I}(\mathbf{X}_k \to \mathbf{X}_i \parallel \underline{\mathbf{X}}_{-\{i,k\}}) > 0.$$

Since edges are found separately, directed information graphs are unique. Also, directed cycles are possible. Minimal generative model graphs and directed information graphs are alternative graphical models to characterize the relationships in stochastic dynamical systems. Next result shows their relationship.

Theorem 1. *[6] For any joint distribution $P_{\underline{\mathbf{X}}}$ satisfying Assumption 1, the corresponding minimal generative model graph and directed information graph are equivalent.*

In the remainder of this dissertation, we will refer to generative model graphs and directed information graphs interchangeably.

2.3.1 Bayesian Networks and Directed Information Graphs

As we mentioned, Bayesian networks represent conditional dependencies in a reduced factorization of the joint distribution. Hence, Bayesian networks depend on the order variables. Figure 2.2 shows two possible Bayesian network pertaining to $P_{X,Y,Z}$.

Notice that the Bayesian networks are acyclic, since a variable can only have incoming arrow from the preceding variables. Therefore, in general, DIGs are not in the family of Bayesian networks. However, DIGs and the Bayesian networks share some similar properties, which we review next.

Analogous to Bayesian networks, the causal independences in a DIG can be determined through a graphical separation criterion which we call *c-separation*.

Definition 5. *Let $\overrightarrow{G} = (V, \overrightarrow{E})$ be a DIG and U and Z be two disjoint subsets of V, and $w \in V \setminus (U \cup Z)$. Z c-separates U from w if for every path between a node in U and w there exists a node on that path in $Z \cup w$ with an outgoing arrow or a collider in $V \setminus (Z \cup w)$.*

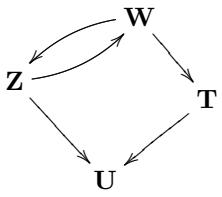


Figure 2.3. An example of DIG with 4 random processes.

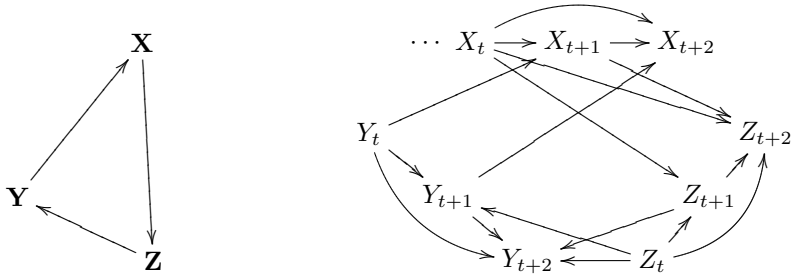


Figure 2.4. The DIG and its underlying variable dependences.

For example, in Figure 2.3, $\mathbf{Z}$ c-separates $\mathbf{U}$ from $\mathbf{W}$. Notice that c-separation, unlike d-separation, is not symmetric, i.e., if Z c-separates U from W, it is not necessary that Z c-separates W from U. A directed graph is said to satisfy *global causal Markov property*, if each c-separation corresponds to a causal independences. In other words, if there exists three disjoint subsets U, $\{w\}$, and Z such that Z c-separates U from w, the corresponding process sets $\underline{\mathbf{X}}_U$ and $\mathbf{X}_w$ are causally independent given the processes in $\underline{\mathbf{X}}_Z$, i.e.,

$$I(\underline{\mathbf{X}}_U \to \mathbf{X}_w || \ \underline{\mathbf{X}}_Z) = 0.$$

Theorem 2. *For any joint distribution $P_{\underline{\mathbf{X}}}$ that satisfies Assumption 1, the DIG is a minimal directed graph with global causal Markov property.*

Proof. See Appendix A.1.1. □

Next, we study the relationship between the DIG of a set of random processes and the independence map among the underlying random variables. Let V be a network of dependent variables, and let σ be an ordering $\{v_1, ..., v_m\}$ of the elements in V. The *boundary strata* of this network relative to σ is an ordered set of subsets of V, $(B_1, B_2..)$, such that each B_i is a Markov boundary of v_i with respect to the set $V_i := \{v_1, ..., v_{i-1}\}$, i.e., B_i is a minimal set satisfying $B_i \subseteq V_i$ and v_i is independent of $V_i \setminus B_i$ given B_i. The DAG created by designating each B_i as parents of vertex v_i is called a boundary DAG of this network relative to σ. By [114], boundary DAGs are Bayesian networks (minimal independence maps under d-separation).

A simple observation is that due to the nature of random processes; there already exists an ordering among the underlying variables, which is time. Hence, if $\underline{\mathbf{X}}$ is a set of random processes that satisfies Assumption 1 with the corresponding minimal generative model graph $\overrightarrow{G}$, then one can define a unique boundary DAG for the underlying variables relative to time ordering. Notice that the boundary DAG relative to time ordering is unique since there are no simultaneous influences between variable, and therefore any causal ordering results in the same DAG. Now, by the definition of minimal generative model graph, the Markov boundary of the t-th variable in process $\mathbf{X}_i$ contains $X_{j,t'}, t' < t$ if and only if $\mathbf{X}_j$ is a parent of $\mathbf{X}_i$ in $\overrightarrow{G}$. For example, in Figure 2.4, Y_t is in the Markov boundary of X_{t+1}, hence $\mathbf{Y}$ must be a parent of $\mathbf{X}$ in the corresponding DIG.

CHAPTER 3

MEASURING FUNCTIONAL DEPENDENCIES

In dynamical systems with specified functional dependencies among the variables, a natural notion of causation exists. That is, a process (or a variable) $\mathbf{X}$ is influenced by another process (or a variable) $\mathbf{Y}$, if $\mathbf{X}$ is a function of $\mathbf{Y}$. Given that the goal of introduction of various graphical models in statistical learning theory is to understand the causal influence structure among the processes, the following natural question arises. *Does a statistical measure of influence that can capture functional relationships exist?* In this chapter, we give an affirmative answer to this question. We define a statistical measure that is able to capture the functional dependency among processes (variables) in dynamical systems. Subsequently, using this measure, we define a new type of graphical model, functional dependency graph (FDGs) that encodes such dependencies. Moreover, we study the relationship between FDG and DIG of a dynamical system.

3.1 Functional Dependencies

Most of the existing approaches to discover functional dependencies are based on intervention [2, 3]. Discovering causal structure by intervention measures the influence of a variable (potential cause) on another variable (effect) in a network through the following processes. The behavior of the effect variable is observed when different values are assigned to the potential cause, while other variables' effects are removed [4]. We use similar paradigm to define our functional dependency measure.

Let $(\mathcal{E}, d)$ be a complete, metric, and separable space equipped with the Borel field $\mathcal{B}$. Consider a causal discrete-time dynamical system with output processes[1] $\underline{\mathbf{X}} = \{\mathbf{X}_1, ..., \mathbf{X}_m\}$ such that $X_{i,t} \in \mathcal{E}$ and is given by

$$X_{i,t} = F_i(\underline{X}^{t-1}, W_{i,t}, t), \quad i = 1, ..., m, \tag{3.1}$$

where F_is are arbitrary functions, and $\mathbf{W}_i$s are exogenous independent random processes such that $W_{i,t}$ is independent of $\underline{X}^{t-1}$ for any i and any t. In this setting, a natural notion of causation among the processes. Namely, $\mathbf{X}_j$ causes $\mathbf{X}_i$, if F_i is a function of $\mathbf{X}_j$.

Remark 3. *Notice that in (3.1), it is assumed that there are no simultaneous influences among the processes. However, our results in this chapter can be extended to the setting in which simultaneous influences are also allowed. For more details see [5].*

To visualize the causal structure in (3.1), we introduce a graphical representation of the causal dependency among the processes. In this graph, nodes represent random processes and there is an arrow from node j to

[1] We use the terminology output that is taken from the context of system identification in control theory. Input processes are $\underline{\mathbf{W}}$.

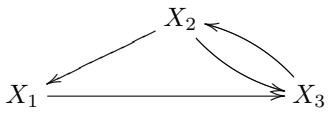


Figure 3.1. Functional dependency graph of Example 2.

nodes i, if $\mathbf{X}_i$ functionally depends on $\mathbf{X}_j$. The following example demonstrates a simple causal system and its corresponding graphical model.

Example 2. *Consider a causal system with 3 processes such that their dynamic is given by*

$$\begin{aligned} X_{1,t} &= e^{-t|X_{1,t-1}|}/3 - X_{2,t-1}/2 + W_{1,t}, \\ X_{2,t} &= X_{2,t-3}/3 + \sin(t\pi + X_{3,t-1})W_{2,t}, \\ X_{3,t} &= \sqrt{|X_{1,t-2}X_{2,t-1}|} + W_{3,t}, \end{aligned}$$

where $\boldsymbol{W}_i$s are independent exogenous noises. Figure 3.1 depicts the functional dependency graph of this system.

We say a random process $\mathbf{X}_i$ functionally depends on process $\mathbf{X}_j$ over the time horizon n, if there exists a time $1 \leq t' \leq n$ such that changing the value of $X_{j,t'}$ while keeping all the other variables fixed results in a change in $\mathbf{X}_i$ at some time $1 \leq t \leq n$ ($t' < t$). Next, we define our measure to capture such functional dependencies in systems whose joint dynamics is described by (3.1).

Definition 6. *We define functional dependency of $X_{i,t}$ on $X_{j,t'}$ in a causal dynamical system, for $t' < t$ and $i \neq j$ as follows:*

$$\alpha_{i,j}(t,t') := \sup_{\substack{\underline{x}=\underline{y} \\ \text{off } X_{j,t'}}} \left[\mathbb{E}_{\boldsymbol{W}_i} \frac{d^2\Big(F_i(\underline{x}, W_{i,t}, t), F_i(\underline{y}, W_{i,t}, t)\Big)}{d^2(x,y)} \right]^{1/2}, \tag{3.2}$$

where d denotes the metric. In this equation, $\underline{x}$ and $\underline{y}$ are two realizations of $\underline{X}^{t-1}$ that are the same everywhere except at $X_{j,t'}$. Further, assume $\underline{x}$ at position $X_{j,t'}$ equals x and $\underline{y}$ equals y ($y \neq x$) at this position.

Equation (3.2) measures whether varying the value of $X_{j,t'}$ while keeping the other variables fixed, changes the value of $X_{i,t}$. Clearly, $\alpha_{i,j}(t,t')$ is always non-negative and if it is positive, it implies the functional dependency of $X_{i,t}$ on $X_{j,t'}$.

Remark 4. *For real-valued random variables, i.e., $\mathcal{E} = \mathbb{R}$, one possible choice for the metric d in (3.2) is Euclidean metric given by*

$$d(x,y) = |x - y|. \tag{3.3}$$

Figure 3.2 summarizes the above definitions. In this figure, columns represent the index of processes and rows represent time. To observe the dependency of $X_{i,t}$ on $X_{j,t'}$, we change the value of (t', j)-th entry that is symbolized by a hammer and observe the value of (t, i)-th entry that is symbolized by an eye while fixing all the other entries.

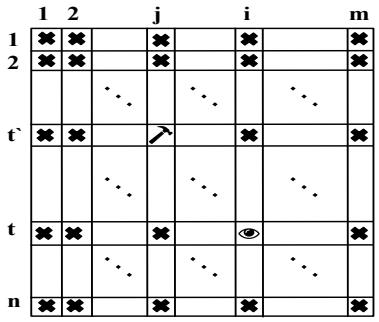


Figure 3.2. A representation of m processes each of length n. Rows represent time and columns represent index of processes. To observe the dependency of $X_i(t)$ on $X_j(t')$, we change the value of $X_j(t')$ (symbolized by a hammer), fix all other variables except $X_i(t)$ (symbolized by crosses), and observe the value of $X_i(t)$ (symbolized by an eye).

Equation (3.2) captures the causal dependency of $X_{i,t}$ on $X_{j,t'}$. To capture the overall causal functional dependency of process $\mathbf{X}_i$ on process $\mathbf{X}_j$, we aggregate the dependencies over the time and define the *functional dependency* graph (FDG) of a causal system as follows:

Definition 7. *Consider a set of random processes $\underline{\boldsymbol{X}} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_m\}$ whose joint dynamics is given by (3.1). The corresponding functional dependency graph of this system $\overrightarrow{G}_{FD} = (V, \overrightarrow{E}_{FD})$ is defined as follows: $V = \{1, ..., m\}$ and for $i \neq j$, $(j, i) \in \overrightarrow{E}_{FD}$ if and only if*

$$\alpha_{i,j} := \frac{1}{n} \sum_{t=1}^{n} \sum_{t'=1}^{t-1} \alpha_{i,j}(t, t') > 0, \tag{3.4}$$

where $\alpha_{i,j}(t, t')$ is given by (3.2).

Consequently, $(j, i) \notin \overrightarrow{E}_{FD}$ iff $\alpha_{i,j} = 0$.

It is important to emphasize that in general, the FDGs are not necessary unique.

Example 3. *Consider the following dynamical system, $X_{1,t} = W_{1,t} = W_1$, $X_{2,t} = X_{1,t-1}$. In this system, we have*

$$F_1(\underline{X}^{t-1}, W_{1,t}, t) = W_{1,t},$$
$$F_2(\underline{X}^{t-1}, W_{2,t}, t) = X_{1,t-1}.$$

Following the Definition 7, the corresponding FDG of this system is $\boldsymbol{X}_1 \to \boldsymbol{X}_2$. However, the above equations can be written as

$$X_{1,t} = X_{2,t-1}, \quad X_{2,t} = W'_{2,t},$$

where $W'_{2,t} = W_1$. In this new setup, the corresponding FDG is $\boldsymbol{X}_2 \to \boldsymbol{X}_1$. Such situations occur because of fully deterministic relationship between processes, i.e., degeneracy.

This phenomena arises because of degenerate relationships between processes. For instance, in Example 3, $P_{\mathbf{X}_1,\mathbf{X}_2}$ is not positive since $P_{\mathbf{X}_1|\mathbf{X}_2}$ is a point mass.

Theorem 3. *Consider a system with positive joint distribution (satisfies Assumption 1) whose dynamic is described by (3.1). Then the corresponding FDG of this system is unique.*

Proof. See Appendix A.2.1. □

Another observation is that the functional dependency measure of $X_{i,t}$ on $X_{j,t'}$, $\alpha_{i,j}(t,t')$, is not necessary bounded. However, if F_is are Lipschitz functions, i.e.,

$$d(F_i(\underline{x},w,t),F_i(\underline{y},w,t)) \leq L\ d(\underline{x},\underline{y}),$$

for some constant L, then it is straightforward to show that $\alpha_{i,j}(t,t')$ is bounded.

In the special case where the functions F_i in (3.1) are real-valued and differentiable, i.e., $\mathcal{E} \subseteq \mathbb{R}$, and $\partial F_i/\partial X_{j,t'}$ exists for all j and t', then we can easily verify whether process $\mathbf{X}_j$ influences process $\mathbf{X}_i$ (i.e., $\alpha_{i,j} > 0$) by calculating partial derivatives of function F_i with respect to $\mathbf{X}_j$. More precisely, suppose there exists a realization of $\underline{X}^{t-1}$, say $\underline{x}$, such that

$$\mathbb{E}_{\mathbf{W}_i}\left[|\frac{\partial F_i}{\partial X_{j,t'}}|\Big|\underline{X}^{t-1}=\underline{x}\right] \neq 0.$$

This implies that there exist two realizations of $\underline{X}^{t-1}$, $\underline{x}$ and $\underline{x}+\underline{w}$, such that

$$\mathbb{E}_{\mathbf{W}_i}\left[\Big|\frac{F_i(\underline{x},W_{i,t},t)-F_i(\underline{x}+\underline{w},W_{i,t},t)}{\underline{w}}\Big|\right] \neq 0,$$

and consequently, $\alpha_{i,j}(t,t') \neq 0$.

In general, learning the corresponding FDG of a causal system by evaluating (3.4) is complicated. However, if some side information about the system dynamic exists, learning its FDG can be significantly simplified. In Section 3.1.2, we discuss a special scenario in which the side information on the dynamics of the system, i.e., knowing the dynamics are linear, allows us to learn its causal structure efficiently.

3.1.1 FDGs for Random Variables

Equation (3.4) in Definition 7 determines whether process $\mathbf{X}_j$ influences process $\mathbf{X}_j$ by identifying a time index (or indices) in $\mathbf{X}_j$ which influence $\mathbf{X}_j$ at some point. Thus, FDGs essentially define influences among random variables. This sets them apart from directed information graphs [41] and linear dynamical graphs [55] which are only defined for random processes. Below, we establish the definition of FDG for a set of random variables.

Suppose a system of random variables $\underline{\zeta} := \{\zeta_1,...,\zeta_m\}$ such that their dependency is captured by $\zeta_i = G_i(\underline{\zeta}_{-i},\omega_i),\ i=1,...,m$, where $\underline{\zeta}_{-i} = \underline{\zeta}\setminus\{i\}$, G_is are arbitrary functions, and ω_is are exogenous independent random variables independent of $\underline{\zeta}_{-i}$. In this case, similar to the random processes scenario, we can define the corresponding FDG of the system as a directed graph whose nodes represent random variables. There is an arrow from node j to i, i.e., ζ_i functionally depends on ζ_j, if and only if

$$\alpha_{i,j} := \sup_{\substack{\underline{\zeta}=\underline{\zeta}'\\ \text{off } \zeta_j}} \left[\mathbb{E}_{\omega_i}\frac{d^2\Big(G_i(\underline{\zeta},\omega_i),G_i(\underline{\zeta}',\omega_i)\Big)}{d^2(\zeta,\zeta')}\right]^{1/2} > 0,$$

where $\underline{\zeta}$ and $\underline{\zeta}'$ are two realizations of $\underline{\zeta}_{-i}$ that are the same everywhere except at ζ_j. Further, assume ζ_j equals ζ in $\underline{\zeta}$ and it equals ζ' in $\underline{\zeta}'$.

3.1.2 Linear Dynamical systems

Perhaps the most studied class of functional dependencies are linear systems which come with their own graphical model, the so-called linear dynamical graphs (LDGs) [55]. Linear dynamical systems are a major subclass of dynamical systems that have been studied extensively in literature and are used in different fields such as economy, finance [115], climatology [116], and biology [117]. In linear systems, the causal influence structure is easy to assess by looking at an appropriate set of coefficients [54]. Furthermore, different approaches for discovering causal structure of such systems given observation of the output processes exist in the literature [118].

Specifically, in [55], the authors study the causal structure in a subclass of causal linear time-invariant systems and introduce a type of graphical model called *linear dynamical* graphs to capture causal structure in this subclass of linear systems. Similar to the FDGs, linear dynamical graphs are also defined based on the functional dependencies. Next, we formally define the linear dynamical graphs and establish their connection with FDGs.

Consider a set of m real-valued random processes $\{\mathbf{X}_1, ..., \mathbf{X}_m\}$ such that their joint dynamics is given by

$$X_{i,t} = \sum_{j=1}^{m} \sum_{s>0} g_{i,j}(s) X_{j,t-s} + W_{i,t}, \quad i = 1, ..., m, \tag{3.5}$$

where $\mathbf{W}_i$ are exogenous independent random processes and $g_{i,i}(s) = 0$ for every i.

Let $\widetilde{\mathbf{X}}_i$ denotes the $\mathcal{Z}$-transform of $\mathbf{X}_i$. Then the set of equations in (3.5) can be represented in the $\mathcal{Z}$-domain, by taking $\mathcal{Z}$-transform of both sides of the equations:

$$\underline{\widetilde{\mathbf{X}}}(z) = \mathbf{G}(z)\underline{\widetilde{\mathbf{X}}}(z) + \underline{\widetilde{\mathbf{W}}}(z), \tag{3.6}$$

where $\mathbf{G}(z)$ is an $m \times m$ matrix whose (i, j)th entry is

$$G_{i,j}(z) := \sum_{s>0} g_{i,j}(s) z^{-s},$$

and $G_{i,i}(z) = 0$. We denote such a system by $(\mathbf{G}(z), \underline{\widetilde{\mathbf{W}}})$.

Definition 8. *[55] The associated linear dynamical graph of a system described in (3.6) is a directed graph, where random processes are represented by nodes and there is an arrow from j to i if and only if $G_{i,j}(z) \neq 0$.*

Next example demonstrates a simple linear time-invariant system and its corresponding linear dynamical graph (LDG).

Example 4. *Consider the following linear systems*

$$\underline{\widetilde{\boldsymbol{X}}}(z) = \begin{bmatrix} 0 & G_{1,2}(z) & 0 & G_{1,4}(z) \\ G_{2,1}(z) & 0 & 0 & G_{2,4}(z) \\ 0 & G_{3,2}(z) & 0 & G_{3,4}(z) \\ 0 & 0 & 0 & 0 \end{bmatrix} \underline{\widetilde{\boldsymbol{X}}}(z) + \begin{bmatrix} \widetilde{\boldsymbol{W}}_1 \\ \widetilde{\boldsymbol{W}}_2 \\ \widetilde{\boldsymbol{W}}_3 \\ \widetilde{\boldsymbol{W}}_4 \end{bmatrix}.$$

Figure 3.3 depicts its corresponding linear dynamical graph (LDG).

Proposition 1. *Consider a causal linear time-invariant system $(\boldsymbol{G}(z), \underline{\widetilde{\boldsymbol{W}}})$. Then, the corresponding linear dynamical graph and the FDG of this system are equivalent.*

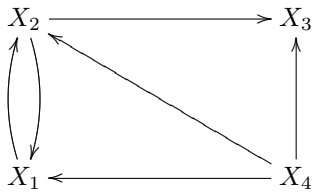


Figure 3.3. Linear dynamical graph of Example 4.

Proof. See Appendix A.2.2. □

Next, we generalize the linear dynamical graphs using FDGs to encompass causal linear systems with time varying coefficient. Let

$$X_{i,t} = \sum_{j=1}^{m} \sum_{s>0} f_{i,j}(t,s) X_{j,t-s} + W_{i,t}, \quad i = 1, ..., m. \tag{3.7}$$

If the coefficients in (3.7) are time invariant, i.e., $f_{i,j}(t,s) = g_{i,j}(s)$, and $f_{i,i}(t,s) = 0$, the system reduces to a causal linear time-invariant system described by (3.5). Next result characterizes the corresponding FDG of the system given by (3.7).

Proposition 2. *In a linear causal system of (3.7), using Euclidean metric (3.3), we have* $\alpha_{i,j}(t, t-s) = |f_{i,j}(t,s)|$, *and consequently,*

$$\alpha_{i,j} = \frac{1}{n} \sum_{t=1}^{n} \sum_{s>0} |f_{i,j}(t,s)|. \tag{3.8}$$

Proof. From equations (3.1) and (3.7), we have

$$d\Big(F_i(\underline{x}, W_{i,t}, t), F_i(\underline{y}, W_{i,t}, t)\Big) = |f_{i,j}(t,s)(x-y)|,$$

where $\underline{x}$ and $\underline{y}$ are two realizations of $\underline{X}^{t-1}$ that are the same everywhere except at $X_{j,t-s}$. Further, $\underline{x}$ at position $X_{j,t'}$ equals x and $\underline{y}$ equals y ($y \neq x$) at this position. Substituting this result into Equation (3.2) implies the results. □

The following example shows a linear causal system and its corresponding FDGs. Note that linear dynamical graph is not able to capture the causal structure of such system.

Example 5. *Consider a linear causal system with 3 processes and the following dynamic*

$$\begin{aligned}
X_{1,t} &= X_{1,t-1}/3 - e^{-t} X_{3,t-1} + W_{1,t}, \\
X_{2,t} &= X_{2,t-2}/3 + e^{-t} X_{1,t} + W_{2,t}, \\
X_{3,t} &= \tan(t\pi/2) X_{1,t-1} + X_{2,t-1}/6 + W_{3,t},
\end{aligned}$$

where $\boldsymbol{W}_i$*s are independent exogenous noises. For example, to asses whether there exists an edge from* $\boldsymbol{X}_3$ *to* $\boldsymbol{X}_1$*, we have to check* $\alpha_{1,3}$ *form (3.4). By Proposition 2, we have*

$$\alpha_{1,3} = \frac{1}{n} \sum_{t=1}^{n} e^{-t} > 0$$

Figure 3.4 demonstrates the corresponding FDG of this system.

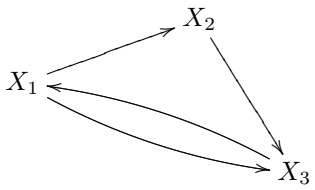


Figure 3.4. FDG of the network in Example 5.

3.2 Measuring Functional Causal Dependency via Directed Information Graphs

In this section, we explore the relationship between statistical dependencies captured by directed information graphs and functional dependencies captured by FDGs in dynamical systems. We further study different conditions under which the two graphical models are equivalent.

Theorem 4. *Consider a set of m random processes with joint dynamics captured by (3.1). Assume their joint distribution is positive. Then $\overrightarrow{E}_{DI} \subseteq \overrightarrow{E}_{FD}$.*

Proof. See Appendix A.2.3. □

It is important to emphasize that in general, the converse of Theorem 4 does not hold. Thus, the FDG of a dynamical system might contain an arrow between two processes while the corresponding DIG does not recover such a relationship. In another words, in a causal dynamical system, existence of statistical dependency implies also functional dependency among processes but not the other way around. We illustrate this using an example.

Example 6. *Consider, a system of two random processes $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ such that $X_{1,t} \sim U[0,1]$ (U stands for uniform distribution), and*

$$X_{2,t} = X_{1,t-1} + W_t \mod 1,$$

where $W_t \sim U[0,1]$ and it is independent of $X_{1,t-1}$. Note that the joint distribution of this system is positive. In this case, $X_{2,t}$ will also have uniform distribution over $[0,1]$ and it is independent of $X_{1,t-1}$. This implies that $I(\boldsymbol{X}_1 \to \boldsymbol{X}_2) = 0$. However, by the definition in (3.2), we obtain $\alpha_{2,1}(t, t-1) > 0$. The corresponding DIG and FDG of this systems are, $(\boldsymbol{X}_1 \quad \boldsymbol{X}_2)$ and $(\boldsymbol{X}_1 \longrightarrow \boldsymbol{X}_2)$, respectively.

As we mentioned earlier, the corresponding DIG of a system is recoverable via mere observation by estimating the directed information quantities in (1.7) [40, 119] or a surrogate when side information is available [120]. Hence, in general, by learning the corresponding DIG of a system using observational data, we can identify some functional dependencies as well.

3.2.1 Special Case: Equivalence between DIGs and FDGs

Previously, we showed that the statistical dependencies recovered by directed information measure as captured by DIGs imply functional dependencies captured by the FDGs in a dynamical system. In this section, we study special dynamical systems and introduce conditions under which the functional dependencies in these systems also imply the statistical dependencies. In other words, their corresponding DIGs and FDGs are equivalent.

Nonlinear Systems with Additive Exogenous Noise:

Let $\mathcal{E} \subseteq \mathbb{R}$ and $d(\cdot,\cdot)$ be the Euclidean metric. Further, consider a special subclass of dynamical system mode of (3.1),

$$F_i = f_i(\underline{X}^{t-1}, t) + g_i(\underline{X}^{t-1}, t)W_{i,t}, \quad i = 1, ..., m, \tag{3.9}$$

where f_i and g_i are arbitrary real-valued functions.

Next result introduces a set of sufficient conditions under which the FDG and the DIG in such systems are equivalent. Hence, a possible approach for learning FDGs of dynamical systems with additive exogenous noise is via estimating the directed information quantities in (1.7). Before stating our result, we need the following definition.

Definition 9. *A random variable W_t is called symmetric if $W_t - \mathbb{E}[W_t]$ and $-W_t + \mathbb{E}[W_t]$ have the same distribution. W is called asymmetric otherwise.*

For instance, standard normal variable, $\mathcal{N}(0,1)$, is a symmetric random variable.

Theorem 5. *Consider a dynamical system with positive joint distribution described by (3.9) with corresponding DIG, $\overrightarrow{G}_{DI}$, and FDG, $\overrightarrow{G}_{FD}$. Further, suppose that for any given t, either $W_{i,t}$ is asymmetric or for all $t' < t$*

$$\sup_{\substack{\underline{x}=\underline{y} \\ off\ X_{j,t'}}} |g_i(\underline{x}, t) - g_i(\underline{y}, t)| = 0. \tag{3.10}$$

Then, $\overrightarrow{E}_{DI} = \overrightarrow{E}_{FD}$.

Proof. See Appendix A.2.4. □

Next example shows a simple system with additive exogenous noise that does not satisfy the conditions in Theorem 5. In this example, while clearly a functional dependency exists between the processes, no statistical dependency is identified by the DI measure among them.

Example 7. *Consider the following dynamical system with two output processes,*

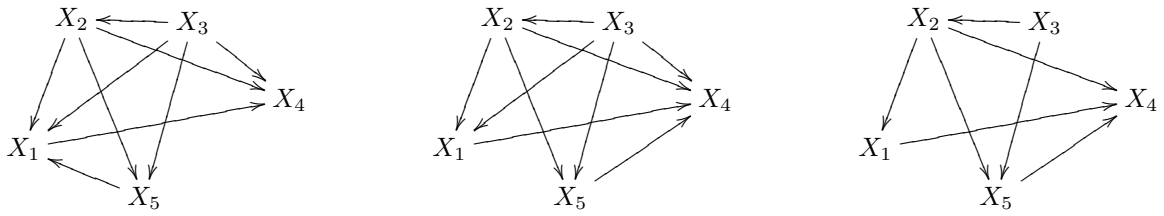
$$X_{1,t} = W_{1,t}, \quad X_{2,t} = (-1)^{\lfloor X_{1,t-1} \rfloor} W_{2,t},$$

where $W_{1,t}$ and $W_{2,t}$ are distributed i.i.d. according to normal distribution with mean zero and variance 1 and $\lfloor x \rfloor$ denotes the floor of x. This system has a positive joint distribution. Furthermore, it is easy to check that $I(\boldsymbol{X}_1 \to \boldsymbol{X}_2) = 0$. However, there is functional dependency between the two processes i.e., $\alpha_{2,1} \neq 0$.

Linear systems:

The linear systems described in Section 3.1.2 are clearly a subclass of dynamical systems with additive exogenous noises, when for all t and $1 \leq i \leq m$,

$$\begin{cases} g_i(\underline{X}^{t-1}, t) = 1, \\ f_i(\underline{X}^{t-1}, t) = \sum_{j=1}^{m} \sum_{s>0} f_{i,j}(t,s) X_{j,t-s}. \end{cases}$$



(a) $N = 50$ (b) $N = 60$ (c) $N = 80$

Figure 3.5. Recovered CFDG of $(\mathbf{G}(z), \widetilde{\underline{W}})$ for sample sizes $\{50, 60, 80\}$ are depicted. The graph (c) is the true FDG.

In this case, the condition in (3.10) holds. Thus, the corresponding FDG and DIG of a causal linear system are equivalent. Combining this result and Proposition 4 imply that the corresponding DIG and the linear dynamical graphs of a causal linear time-invariant system are also equivalent. This result was previously proven tediously in [54] using information-theoretical tools and under more restrictive assumptions.

Different approaches have been developed in literature to learn the coefficients of a linear time-invariant system [118]. These approaches depend on different parameters of the system such as their underlying causal structure. For instance, in [55], the authors propose a learning method for self-kin linear networks. In such systems, there is at least one arrow between any two nodes which share a common child. In [121], the authors study linear systems in which their underlying causal structure is a directed acyclic graph (DAG) by observing all the output processes.

3.3 Experimental Results

Herein, we present two simulation results for both linear and nonlinear systems. Note that in both systems there is no control variables that allows the learner to intervene the system. Thus, discovering the causal structure of these systems via intervention is not straightforward. However, because both systems satisfy conditions of Theorem 5 their DFGs and DIGs are equivalent. Thus, it is possible to learn the causal dependencies via mere observation by estimating the directed information quantities in (1.7).

Linear System:

In this section, we consider a causal linear time-invariant system and reconstruct its corresponding FDG by observing all the output processes. The dynamic is given by $(\mathbf{G}(z), \widetilde{\underline{\mathbf{W}}})$, where $\widetilde{\underline{\mathbf{W}}} \sim \mathcal{N}(0, \Sigma_1)$, $\Sigma_1 = diag\{.2, .5, .3, .2, .5\}$, and

$$\mathbf{G}_1(z) = \frac{1}{6}\begin{bmatrix} 0 & z^{-2} & 0 & 0 & 0 \\ 0 & 0 & 2z^{-2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 3z^{-1} & 1.2z^{-1} & 0 & 0 & z^{-3} \\ 0 & z^{-3} & \sqrt{2}z^{-4} & 0 & 0 \end{bmatrix}.$$

We learned the FDG by learning the corresponding DIG based on Theorem 5. To do so, we sampled each of the output processes, N times over a time horizon of length $n = 10$. Because this system is jointly Gaussian,

the directed information between each pair of the output processes is given by [54],

$$I(\mathbf{X} \to \mathbf{Y}||\underline{\mathbf{Z}}) = \frac{1}{2}\sum_{t=1}^{n} \log \frac{|\Sigma_{Y_1^t \underline{Z}_1^{t-1}}||\Sigma_{X_1^{t-1}Y_1^{t-1}\underline{Z}_1^{t-1}}|}{|\Sigma_{Y_1^{t-1}\underline{Z}_1^{t-1}}||\Sigma_{X_1^{t-1}Y_1^t\underline{Z}_1^{t-1}}|}, \tag{3.11}$$

where $\Sigma_{Y_1^t \underline{Z}_1^{t-1}}$ is the covariance matrix of $(Y_1, ..., Y_t, \underline{Z}_1, ..., \underline{Z}_{t-1})$. We estimated the directed information using the above equation with sample covariance matrix. Using the concentration result for empirical mutual information of Gaussian distribution [122], we decided on whether the estimated DI were positive with confidence $1 - \delta$ by comparing them against the following threshold

$$\tau = \min_{i,j} I(\mathbf{X}_j \to \mathbf{X}_i||\underline{\mathbf{X}}_{-\{i,j\}}) - \mathcal{O}(\sqrt{\log(M/\delta)/N}),$$

where $0 < \delta < 1$, $M = o(nmp)$, and p denotes the Markov-order of the system. In this example $p = 4$ and $\tau = 0.53$. Figure 3.5 depicts the recovered DIG (equivalently FDG) for different sample sizes $N \in \{50, 60, 80\}$.

Note that the above system is a self-kin network. Therefore, as we discussed in Section 3.2.1, the corresponding FDG of this system is also identifiable by learning the corresponding linear dynamical graph using the approach of [55].

Nonlinear System with Additive Exogenous Noise:

We simulated a network of $m = 6$ processes with the following joint dynamics

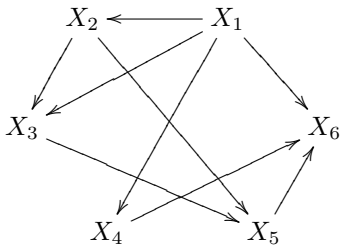
$$\begin{aligned}
X_{1,t} &= 0.2X_{1,t-3} + 0.1X_{1,t-2}^2 + W_{1,t},\\
X_{2,t} &= X_{1,t-1}^2/\sqrt{2} - 0.1|W_{2,t}|,\\
X_{3,t} &= 0.1X_{2,t-1} - 0.5\sqrt{|X_{1,t-1}|} + W_{3,t},\\
X_{4,t} &= -0.2X_{2,t-1} + 0.3\sqrt{|X_{2,t-3}|^3} + W_{4,t},\\
X_{5,t} &= 0.2X_{3,t-2} - 0.2X_{2,t-1} + \sqrt{|W_{5,t}|},\\
X_{6,t} &= 0.3X_{5,t-2} - 0.5X_{4,t-2} + |W_{6,t}|,
\end{aligned} \tag{3.12}$$

where $\mathbf{W}_i$s were generated i.i.d. Gaussian with mean zero and variance one. The output processes $\{\mathbf{X}_1, ..., \mathbf{X}_6\}$ were each of length $n = 20$ and $N \in \{5 \times 10^3, 10^4\}$ number of samples from each of them was collected. In order to estimate the directed information measures, i.e., Equation (1.7), we used the fact that the directed information can be written as a sum of different mutual information [41] and then estimated them using K-nearest neighbor method of [123]. The recovered networks are depicted in Figure 5.7.

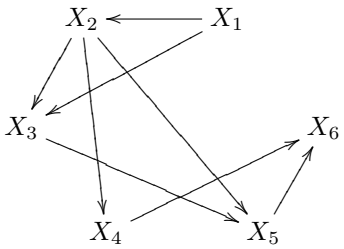
FDGs and DIGs Are Not the Same in General:

We simulated a network of $m = 3$ processes with the following dynamics

$$\begin{aligned}
X_{1,t} &= 0.45X_1(t-1) + W_{1,t},\\
X_{2,t} &= \begin{cases} 0.2W_{2,t}, & \text{if } X_{1,t-1} \in \{0,1\} \\ 0.3X_{2,t-1} + W_{2,t}, & \text{otherwise} \end{cases}\\
X_{3,t} &= 0.4X_{2,t-1} + W_{3,t},
\end{aligned} \tag{3.13}$$



(a) $N = 5000$



(b) $N = 10000$

Figure 3.6. Recovered causal structure of the non-linear model in (3.13) are depicted. The graph (b) is the true FDG.

Name	code
Dell Inc.	DEL
Hewlett-Packard	HP
Intel	INT
Texas Instruments	TXN
International Business Machines	IBM
Cisco Systems	CSCO
Apple Inc.	APPL
Oracle	ORC
Xerox	XRX
Google Inc.	GOG
Microsoft	MSFT
EMC Corporation	EMC

Table 3.1. List of Companies in the analysis

where $W_{i,t} \sim \mathcal{N}(0,1)$ for $i = 1, 2, 3$. The output processes $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ are each of length $n = 10$. $N = 70$ samples from each process is observed. Directed information cannot discover the relationship between $\mathbf{X}_1$ and $\mathbf{X}_2$, because they are statistically independent with probability one. Thus, the DIG of this system is $(\mathbf{X}_1 \quad \mathbf{X}_2 \to \mathbf{X}_3)$.

In order to recover the FDG of this network, we intervened in $\mathbf{X}_1$ and set it to fixed values in $\{-1, 0, 1\}$ and observed the values of the other two processes over the time horizon of length $n = 10$. The recovered FDG is $(\mathbf{X}_1 \to \mathbf{X}_2 \to \mathbf{X}_3)$ that is the correct functional dependencies.

Stock Price Analysis

In this section, we analyse the causal relationship between stock prices of 12 technology companies (Table I) of the New York Stock Exchange sourced from Google Finance. These prices were sampled every 2 minutes for twenty market days (03/03/2008 - 03/28/2008). We assumed the underlying joint dynamics was jointly Gaussian. Therefore, directed information values were estimated using Equation 3.11. The resulting DIG is shown in Figure 3.7.

Fig. 3.7 illustrates interesting interactions between these companies during 2008. For instance the DIG suggests that one of the most influential companies in that period of time was HP. Looking into the global PC market share during 2008, we can find that the Hewlett-Packard company had place one among others.[2]

[2]Gartner, `http://www.gartner.com/newsroom/id/856712`

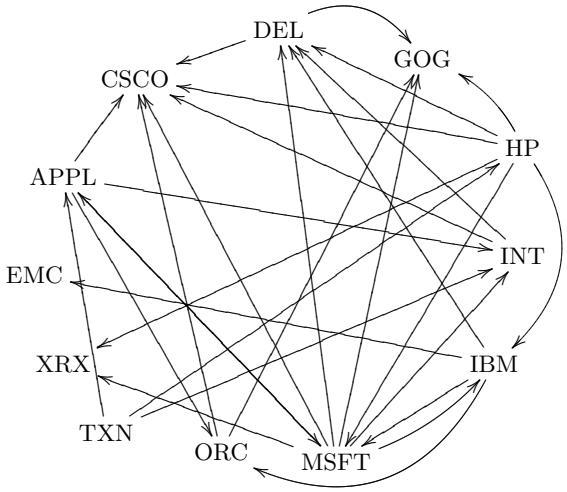


Figure 3.7. The DIG obtained for the stock market using estimating the directed information.

Another example is that Apple has been using Intel processors in its products since 2006. However, in 2008 Apple released MacBook Air and upgraded the processors of MacBook and MacBook Pro to Intel core 2 Duo Penryn.[3] This was a kind of revolution in laptop's market. Hence, we see an influence from Apple on Intel during that period of time. We have also applied similar method to learn the interconnections between the financial institutions by analysing the monthly returns of different banks, brokers, and insurance companies [124].

[3]Apple Press Info, `http://www.apple.com/pr/`

CHAPTER 4

CAUSAL STRUCTURE OF MULTIVARIATE HAWKES PROCESSES

In this chapter, we study the causal structure of a specific type of time series, multivariate linear Hawkes process [120]. Hawkes processes were originally motivated by the quest for statistical models for earthquake occurrences. Since then, they have been successfully applied to seismology [125], biology [126], criminology [127], computational finance [57, 128, 129], etc.

In multivariate or mutually exciting point processes, occurrence of an event (arrival) in one process affects the conditional probability of new occurrences, i.e., the *intensity* function of other processes in the network. Such interdependencies between the intensity functions of a linear Hawkes process are modeled as follows: the intensity function of processes j is assumed to be a linear combination of different terms, such that each term captures only the effects of one other process (See Section 4.1).

This dependency is captured by the support of the excitation matrix of the network. As a result, estimation of the excitation (kernel) matrix of multivariate processes is crucial both for learning the structure of their causal network and for other inference tasks and has been the focus of research.

4.1 Multivariate Hawkes Processes

Fix a complete probability space $(\Omega, \mathcal{F}, P)$. Let $\mathbf{N}_t$ denotes the counting process representing the cumulative number of events up to time t and let $\{\mathcal{F}^t\}_{t\geq 0}$ be a set of increasing σ-algebras such that $\mathcal{F}^t = \sigma\{N^t\}$. The non-negative, $\mathcal{F}^t$-measurable process $\lambda(t)$ is called the intensity of $\mathbf{N}_t$ if

$$P(\mathbf{N}_{t+dt} - \mathbf{N}_t = 1|\mathcal{F}^t) = \lambda(t)dt + o(dt).$$

A classical example of mutually exciting processes, a multivariate Hawkes process [120], is a multidimensional process $\underline{\mathbf{N}} = \{\mathbf{N}_1, ..., \mathbf{N}_m\}$ such that for each $i \in [m]$

$$\begin{aligned} &P\left(dN_{i,t} = 1|\underline{\mathcal{F}}^t\right) = \lambda_i(t)dt + o(dt), \\ &P(dN_{i,t} > 1|\underline{\mathcal{F}}^t) = o(dt), \end{aligned} \tag{4.1}$$

where $\underline{\mathcal{F}}^t = \sigma\{\underline{N}^t\}$. The above equations imply that $\mathbb{E}[dN_{i,t}/dt|\underline{\mathcal{F}}^t] = \lambda_i(t)$. Furthermore, the intensities are all positive and are given by

$$\lambda_i(t) = v_i + \sum_{k=1}^{m} \int_0^t \gamma_{i,k}(t - t')dN_k(t'). \tag{4.2}$$

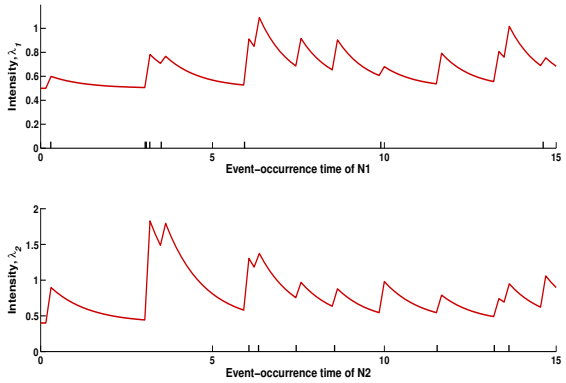


Figure 4.1. Intensities of the multivariate Hawkes process.

The exciting functions $\gamma_{i,k}(\cdot)$s are in ℓ_1 such that $\lambda_i(t) \geq 0$ for all $t > 0$. Equivalently, in matrix representation:

$$\Lambda(t) = \mathbf{v} + \int_0^t \Gamma(t - t')d\underline{N}(t'), \tag{4.3}$$

where $\Gamma(\cdot)$ denotes an $m \times m$ matrix with entries $\gamma_{i,j}(\cdot)$; $d\underline{N}, \Lambda(\cdot)$, and $\mathbf{v}$ are $m \times 1$ arrays with entries $dN_i, \lambda_i(\cdot)$, and v_i, respectively. Matrix $\Gamma(\cdot)$ is called the excitation (kernel) matrix. Figure 4.1 illustrates the intensities of a multivariate Hawkes process comprised of two processes ($m = 2$) with the following parameters

$$\mathbf{v} = \begin{pmatrix} 0.5 \\ 0.4 \end{pmatrix} \quad , \quad \Gamma(t) = \begin{pmatrix} 0.1e^{-t} & 0.3e^{-1.1t} \\ 0.5e^{-0.9t} & 0.3e^{-t} \end{pmatrix} u(t),$$

where $u(t)$ is the unit step function.

4.2 Two Equivalence Notations of Causality for Hawkes Processes

Next, we establish the relationship between the excitation matrix of multivariate Hawkes processes and their generative model graph. First notice that the corresponding minimal generative model graph and the DIG of a causal dynamical system are equivalent [6]. Thus, to characterize the minimal generative model graphs of a multivariate Hawkes system, we study the properties of its corresponding DIG.

Recall that the directed information as it is defined in 1.7 is for discrete time dynamical systems. However, multivariate Hawkes processes are continuous processes. Hence, the first step would be to generalized the directed information to continuous time dynamical systems.

Notice that in a DIG, to determine whether $\mathbf{X}_j$ causes $\mathbf{X}_i$ over a time horizon $[0, T]$ in a network of m random processes, two conditional probabilities are compared in KL-divergence sense: one is the conditional probability of $X_{i,t+dt}$ given full past, i.e., $\underline{\mathcal{F}}^t := \sigma\{\underline{X}^t\}$ and the other one is the conditional probability of $X_{i,t+dt}$ given full past except the past of $\mathbf{X}_j$, i.e., $\underline{\mathcal{F}}^t_{-\{j\}} := \sigma\{\underline{X}^t_{-\{j\}}\}$. It is declared that there is no influence from $\mathbf{X}_j$ on $\mathbf{X}_i$, if the two conditional probabilities are the same. More precisely, there is an influence from $\mathbf{X}_j$ on $\mathbf{X}_i$ if and only if the following directed information measure is positive [41],

$$I_T(\mathbf{X}_j \to \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,j\}}) := \inf_{\mathbf{t} \in \mathcal{T}(0,T)} \tilde{I}_{\mathbf{t}}(\mathbf{X}_j \to \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,j\}}), \tag{4.4}$$

where $\mathcal{T}$ denotes the set of all finite partitions of the time interval $[0, T]$ [130], and

$$\tilde{I}_{\mathbf{t}}(\mathbf{X}_j \to \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,j\}}) := \sum_{k=0}^{n} I\left(X_{i,t_{k-1}}^{t_k}; X_{j,0}^{t_k} | \mathcal{F}_{-\{j\}}^{t_{k-1}}\right),$$

where $\mathbf{t} := (0 = t_0, t_1, ..., t_n = T)$.

Proposition 3. *Consider a set of mutually exciting processes $\underline{\boldsymbol{N}}$ with excitation matrix $\Gamma(t)$. Under Assumption 1, $I_T(\boldsymbol{N}_j \to \boldsymbol{N}_i || \underline{\boldsymbol{N}}_{-\{i,j\}}) = 0$ if and only if $\gamma_{i,j} \equiv 0$ over time interval $[0, T]$.*

Proof. See Appendix A.3.1. □

Proposition 3 signifies that the support of the excitation matrix $\Gamma(\cdot)$ determines the adjacency matrix of the DIG and vice versa. Therefore, learning DIG of a mutually exciting Hawkes processes satisfying Assumption 1 is equivalent to learning the excitation matrix given samples from each of the processes. In other word, in the presence of side information that the processes are Hawkes, it is more efficient to learn the causal structure through learning the excitation matrix rather than the directed information needed for learning the DIG in general.

4.3 Learning the Excitation Matrix

Herein, we present an approach for learning the causal structure of a stationary Hawkes network with exponential exciting functions through learning the excitation matrix. This method is based on second order statistic of the Hawkes processes and it is suitable for the case when no i.i.d. samples are available. Note that when i.i.d. samples are available, non-parametric methods for learning the excitation matrix such as MMEL algorithm [131] exist. As mentioned earlier, we focus on learning the excitation matrix of multivariate Hawkes processes with exponential exciting functions. This class of Hawkes processes has been widely applied in many areas such as seismology, criminology, and finance [125–128].

Definition 10. *The excitation matrix of a multivariate Hawkes processes with exponential exciting functions is defined as follows*

$$\mathcal{E}xp(m) := \left\{ \sum_{d=1}^{D} A_d e^{-\beta_d t} u(t) : \ A_d \in \mathbb{R}^{m \times m}, \ (\sum_{d=1}^{D} A_d e^{-\beta_d t})_{i,j} \geq 0, \ \rho(\sum_{d=1}^{D} \frac{A_d}{\beta_d}) < 1, D \in \mathbb{N} \right\},$$

where $\{\beta_d\} > 0$ is called the set of exciting modes.

Example 8. *Consider a set of $m = 5$ mutually exciting processes with the following exponential excitation matrix*

$$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & 0 \\ 0 & 1.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \frac{e^{-t}}{20} + \begin{pmatrix} 0 & 0 & .5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2.5 & 0 \\ .1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \frac{e^{-1.4t}}{20} + \begin{pmatrix} 1 & 1.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \frac{e^{-2t}}{20} \tag{4.5}$$

In this example $D = 3$ and the exciting modes are $\{1, 1.4, 2\}$. By Proposition 3, the adjacency matrix of the corresponding DIG of this network is given by the support of its excitation matrix. Figure 4.2 depicts the corresponding DIG.

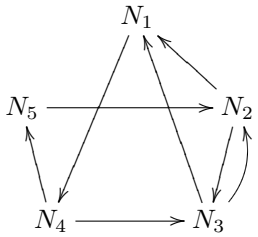


Figure 4.2. Corresponding DIG of the network in Example 8 with the excitation matrix given by (4.5)

Before describing our algorithm, we need to derive some useful properties of moments of the process. A multivariate Hawkes process with the excitation matrix Γ has stationary increments, i.e., the intensity processes is stationary, if and only if the following assumption holds [120, 132]:

Assumption 2. *The spectral radius (the supremum of the absolute values of the eigenvalues) of the matrix $\overline{\Gamma}$, where $[\overline{\Gamma}]_{i,j} = ||\gamma_{i,j}||_1$ is strictly less than one, i.e., $\rho(\overline{\Gamma}) < 1$.*

In this case, from (4.3) and Equation (4.1), we obtain

$$\Lambda = \mathbb{E}[\Lambda(t)] = \mathbf{v} + \int_0^t \Gamma(t-t')\mathbb{E}[d\underline{N}(t')] = \mathbf{v} + \int_0^t \Gamma(t-t')\Lambda dt' = \mathbf{v} + \overline{\Gamma}\Lambda. \tag{4.6}$$

By Assumption 2, $\sum_{i\geq 0} \overline{\Gamma}^i$ converges to $(I-\overline{\Gamma})^{-1}$, thus $\Lambda = (I-\overline{\Gamma})^{-1}\mathbf{v}$. The normalized covariance matrix of a stationary multivariate Hawkes process with lag τ and window size $z > 0$ is defined by

$$\Sigma_z(\tau) := \frac{1}{z}\mathbb{E}\left[\int_t^{t+z} d\underline{N}(x) \int_{t+\tau}^{t+\tau+z} (d\underline{N}(y))^T\right] - \Lambda\Lambda^T z, \tag{4.7}$$

where $\int_t^{t+t'} d\underline{N}(x)$ denotes the number of events in time interval $(t, t+t']$.

Theorem 6. *[64] The Fourier transform of the normalized covariance matrix of a stationary multivariate Hawkes process with lag τ and window size $z > 0$ is given by*

$$\mathcal{F}[\Sigma_z](-\omega) = 4\frac{\sin^2 z\omega/2}{\omega^2 z}\left(I - \mathcal{F}[\Gamma](\omega)\right)^{-1} diag(\Lambda)\left(I - \mathcal{F}[\Gamma](\omega)\right)^{-\dagger}, \tag{4.8}$$

where $A^\dagger$ denotes the Hermitian conjugate of matrix A, and $diag(\Lambda)$ is a diagonal matrix with vector Λ as the main diagonal.

In order to learn the excitation matrix with exponential exciting functions, we need to learn the exciting modes $\{\beta_d\}$, the number of components D, and coefficient matrices $\{A_d\}$. Next results establishes the relationship between the exciting modes and the number of components D with the normalized covariance matrix of the process.

Corollary 1. *Consider a network of a stationary multivariate Hawkes processes with excitation matrix $\Gamma(t)$ belonging to $\mathcal{E}xp(m)$. Then the exciting modes of $\Gamma(t)$ are the absolute values of the zeros of $1/\operatorname{Tr}\mathcal{F}[\Sigma_z]^{-1}(\omega)$.*

Proof. See Section A.3.2. □

Next, we need to find the coefficient matrices $\{A_d\}$. To do so, we use the covariance density of the

processes. The covariance density of a stationary multivariate Hawkes process for $\tau > 0$ is defined as [120]

$$\Omega(\tau) := \mathbb{E}\left[\left(\frac{d\underline{N}(t+\tau)}{dt} - \Lambda\right)\left(\frac{d\underline{N}(t)}{dt} - \Lambda\right)^T\right]. \tag{4.9}$$

Since the processes have stationary increments, we have $\Omega(-\tau) = \Omega^T(\tau)$.

Lemma 1. *[120] We have*

$$\Omega(\tau) = \Gamma(\tau)diag(\Lambda) + \Gamma * \Omega(\tau), \tau > 0. \tag{4.10}$$

It has been shown in [70] that the above equation admit a unique solution for $\Gamma(\tau)$. Next proposition provides a system of linear equations that allows us to learn the coefficient matrices.

Proposition 4. *Consider a network of a stationary multivariate Hawkes processes with excitation matrix $\Gamma(t) \in \mathcal{E}xp(m)$, and exciting modes $\{\beta_1, ..., \beta_D\}$. Then $\{A_d\}$ are a solution of the linear system of equations: $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{H}$, where $\boldsymbol{H}_{m^2 \times m^2}$ is a block matrix with (i,j)th block given by*

$$\boldsymbol{H}_{i,j} = \frac{diag(\Lambda) + \mathcal{L}[\Omega](\beta_j) + \mathcal{L}[\Omega]^T(\beta_i)}{\beta_j + \beta_i},$$

and $\boldsymbol{A} = [A_1, ..., A_D]$ and $\boldsymbol{S} = [\mathcal{L}[\Omega](\beta_1), ..., \mathcal{L}[\Omega](\beta_D)]$.

Proof. See Section A.3.3. □

Combining the results of Corollary 1 and Proposition 4 allows us to learn the excitation matrix of exponential multivariate Hawkes processes from the second order moments. Consequently applying Proposition 3, the causal structure of the network can be learned by drawing an arrow from node i to j, when $\sum_{d=1}^{D} |(A_d)_{j,i}| > 0$.

4.3.1 Estimation and Algorithm

This section discusses estimators for the second order moments, namely the normalized covariance matrix and the covariance density of a stationary multivariate Hawkes processes from data. Once such estimators are available, the approach of previous section maybe used to learn the network. The most intuitive estimator for Λ defined by Equation (4.6) is $\underline{\mathbf{N}}_T/T$. It turns out that this estimator converges almost surely to Λ as T goes to infinity [133]. Furthermore, [133] proposes an empirical estimator for the normalized covariance matrix as follows

$$\widehat{\Sigma}_{z,T}(\tau) := \frac{1}{T}\sum_{i=1}^{\lfloor T/z \rfloor} (\mathbf{X}_{iz} - \mathbf{X}_{(i-1)z})(\mathbf{X}_{iz+\tau} - \mathbf{X}_{(i-1)z+\tau})^T, \tag{4.11}$$

where $\mathbf{X}_t := \underline{\mathbf{N}}_t - \Lambda t$. In the same paper, it has been shown that under Assumption 2, the above estimator converges in ℓ_2 to the normalized covariance matrix (4.7), i.e., $\widehat{\Sigma}_{z,T}(\tau) \underset{T\to\infty}{\longrightarrow} \Sigma_z(\tau)$. Notice that the normalized covariance matrix and the covariance density are related by $\Sigma_{dt}(\tau)/dt = \Omega^T(\tau)$. Therefore, we can estimate the covariance density matrix using Equation (4.11) by choosing small enough window size $z = \Delta$. Namely, $\widehat{\Omega}^T_\Delta(\tau) = \widehat{\Sigma}_\Delta(\tau)/\Delta$.

Algorithm 1

1: *Input* : $\underline{\mathbf{N}}^T$.
2: *Output* : DIG.
3: $\widehat{\Lambda} \leftarrow \underline{\mathbf{N}}_T/T$
4: Choose $\sigma > 0$, $z > 0$, and small $\Delta > 0$.
5: Compute $\widehat{\Sigma}_{z,T}(\tau)$ and $\widehat{\Omega}_{\Delta}(\tau)$ using (4.11).
6: $\{\widehat{\beta}_d\}_{d=1}^{\widehat{D}} \leftarrow$ Zeros of $1/\operatorname{Tr}\mathcal{F}[\Sigma_z]^{-1}(\omega)$.
7: Compute $\mathcal{L}[\widehat{\Omega}_{\Delta}](\widehat{\beta}_d)$ for $d = 1, ..., \widehat{D}$.
8: Solve the set of equations arises from (A.21) for $\widehat{A}_d$.
9: Draw (j, i) if $\sum_{d=1}^{\widehat{D}} |(\widehat{A}_d)_{i,j}| \geq \sigma$.

Algorithm 1 summarizes the steps of our proposed approach for learning the excitation matrix and consequently the causal structure of an exponential multivariate Hawkes process.

4.4 Experimental Results

In this section, we present our experimental results for both synthetic and real data.

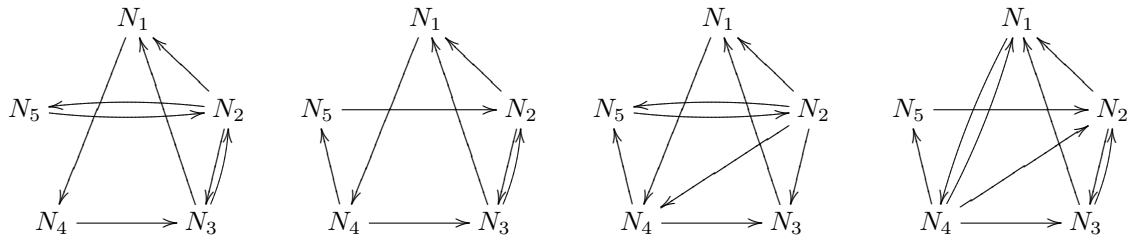
Synthetic Data:

We applied the proposed algorithm to learn the causal structure of the multivariate Hawkes network in Example 8 with $\mathbf{v} = (0.5, 0.4, 0.5, 1, 0.3)^T$. This network satisfies Assumption 2, since $\rho(\overline{\Gamma}) \approx 0.16$. The exciting modes are $\{1, 1.4, 2\}$. We observed the arrivals of all processes during a time period T. Figure 4.3 depicts the outputs of algorithms 1 for $\Delta = 0.2$, $z = 2$, and observation lengths $T \in \{1000, 2100\}$. As illustrated in Figure 4.3, by increasing the length of observation T, the output graph converges the true DIG shown in Figure 4.2. As a comparison, we applied the MMEL algorithm proposed in [131] to learn the excitation matrix for this example and the numerical method based on Nystrom method proposed in [70] with $T = 2100$ and the number of quadrature $Q = 70$. Since MMEL requires i.i.d. samples, we generate 35 i.i.d. samples each of length 60 to obtain Figure 4.3(MMEL). Our proposed algorithm outperforms both MMEL and the numerical method of [70].

Furthermore, we conducted another experiment for a network of 15 processes with 102 edges illustrated in Figure 4.4. For a sample of length $T = 2500$, our algorithm was able to recover 70 edges correctly but identified 34 false arrows. MMEL could only recover 58 arrows correctly while detecting another 41 false arrows. The input for MMEL was 25 sequences each of length 100.

Stock Market Data:

As an example of how our approach may discover causal structure in real-world data, we analyzed the causal relationship between stock prices of 12 technology companies of the New York Stock Exchange sourced from Google Finance. The prices were sampled every 2 minutes for twenty market days (03/03/2008 - 03/28/2008). Every time a stock price changed by $\pm 1\%$ of its current price an event was logged on the stock's process. In order to prevent the substantial changes in stock's prices due to the opening and closing of the market, we



(a) $T = 1000$ (b) $T = 2100$ (c) Numerical method (d) MMEL

Figure 4.3. Recovered DIG of the network in Example 8 with the excitation matrix given by (4.5), (a), (b) Algorithm 1 with $\Delta = 0.2$, $z = 2$, and $T \in \{1000, 2100\}$, (c) the numerical method of [70] with $Q = 70$ and $T = 2100$, and (d) MMEL with 35 i.i.d. samples each of length 60. Our approach learns the graph with $T = 2100$, while other approaches fail at the same sample size.

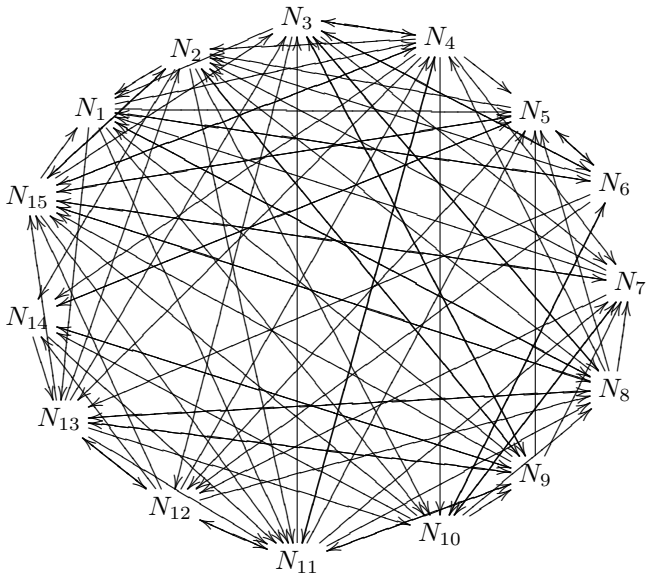


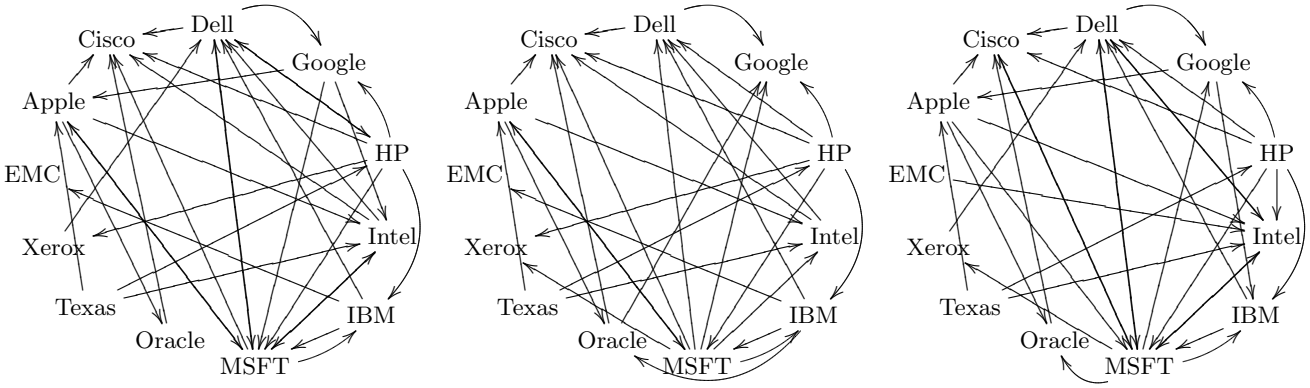
Figure 4.4. True causal structure of the synthesized example.

ignored the samples at the beginning and at the end of each working day. For this part, we have assumed that the jumps occurring in stock's prices are correlated through a multivariate Hawkes process. This model class was advocated in [133, 134]. Figure 5.8(a) illustrate the causal graph resulting from Algorithm 1, with $z = 30$ and $\Delta = 2$ minutes.

To compare our learning approach with other approaches, we applied the MMEL algorithm to learn the corresponding causal graph. For this scenario, we assumed that the data collected from each day is generated i.i.d. Hence, a total of 20 i.i.d. samples were used. Figure 5.8(c) illustrates the resulting graph. As one can see, Figures 5.8(a) and 5.8(c) convey pretty much a similar causal interactions in the dataset. For instance both of these graphs suggest that one of the most influential companies in that period of time was Hewlett-Packard (HP). Looking into the global PC market share during 2008, we find that this was indeed the case.[1]

To use another modality, we derive the corresponding DIG of this network applying Equation (4.4). For this part, we used the market based on the Black-Scholes model [135] in which the stock's prices are modeled

[1]Gartner, `http://www.gartner.com/newsroom/id/856712`



(a) Algorithm 1. (b) DIG. (c) MMEL.

Figure 4.5. Causal structures for the S&P (a) using Algorithm 1, (b) by estimating the directed information DIG, and (c) using MMEL algorithm.

	Alg. 1	DIG	MMEL
Alg. 1	33	25	26
DIG	25	30	24
MMEL	26	24	34

Table 4.1. Number of edges that the approaches jointly recover.

via a set of coupled stochastic PDEs. We assumed that the logarithm of the stock's prices are jointly Gaussian and therefore the corresponding DIs were estimated using Equation (3.11). The resulting DIG is shown in Figure 5.8(b). Note that this DIG is derived from the logarithm of prices and not the jump processes we used earlier. Still it shares a lot of similarities with the two other graphs. For instance, it also identifies HP as one of the most influential companies and Microsoft as one the most influenced companies in that time period. Table 4.1 shows the number of edges that each of the above approaches recovers and the number of edges that they jointly recover. This demonstrates the power of exponential kernels even when data does not come from such a model class.

MemeTracker Data:

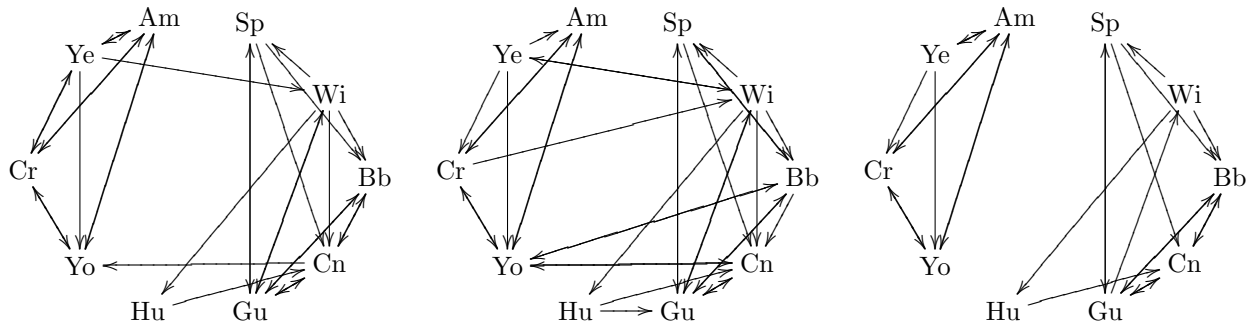
We also studied causal influences in a blogosphere. The causal flow of information between media sites may be captured by studying hyperlinks provided in one media site to others. Specifically, the time of such linking can be modeled using a linear multivariate Hawkes processes with exponential exciting functions [131, 136]. This model is also intuitive in the sense that after emerging a new hot topic, in the first several days, the blogs or websites are more likely feature that topics and it is also more likely that the topic would trigger further discussions and create more hyperlinks. Thus, exponential exciting functions are well suited to capture such phenomenon as the exiting functions should have relatively large values at first and decay fast as time elapses.

For this experiment, we used the MemeTracker[2] dataset. The data contains time-stamped phrase and hyperlink information for news media articles and blog posts from over a million different websites. We extracted the times that hyperlinks to 10 well-known websites listed in Table 4.2 are created during August

[2] http://memetracker.org/data/links.html

Cr	craigslist.org
Ye	yelp.com
Am	amazon.com
Sp	spiegel.de
Wi	wikipedia.org
Yo	youtube.com
Cn	cnn.com
Gu	guardian.co.uk
Hu	humanevents.com
Bb	bbc.co.uk

Table 4.2. List of websites studied in MemeTracker experiment.



(a) Alg. 1 (30). (b) MMEL. (c) Both Alg. 1 & MMEL (110).

Figure 4.6. Recovered causal structure of the MemeTracker dataset using (a) Algorithm 1, (b) MMEL for 30 different phrases, and (c) both Algorithm 1 and MMEL for 110 different phrases.

2008 to April 2009. When a hyperlink to a website is created at a certain time, an arrival events is recorded at that time. More precisely, in this experiment, we picked 30 different phrases that appeared on different websites at different times. If a website that published one of the phrases at time t also contained a hyperlink to one of the 10 listed websites, an arrival event was recorded at time t for that website in our list.

Figure 4.6(a) illustrates the resulting causal structure learned by Algorithm 1 for $z = 12$ hours and $\Delta = 1$ hour. In this graph, an arrow from a node to another, say node Ye to Yo, means creating a hyperlink to `yelp.com` triggers creation of further hyperlinks to `youtube.com`.

We also applied the MMEL algorithm with one exponential kernel function to learn the excitation matrix. For this experiment, the data corresponding to each phrase was treated as an i.i.d. realization of the system. The resulting causal structure is depicted in Figure 4.6(b).

As Figure 4.6(a) illustrates, the nodes can be clustered into two main groups: {Cr, Ye, Am, Yo} and {Bb, Cn, Gu, Hu, Sp, Wi}. The first group consists of mainly merchandise and reviewing websites and the second group contains the broadcasting websites. However, this is not as clear in Figure 4.6(b). This is because MMEL requires more i.i.d. samples (phrases) to be able to identify the correct arrows. Note that as we increase the number of phrases (110), Figure 4.6(c), both graphs become similar with two clearly visible main clusters.

CHAPTER 5

LEARNING MINIMAL LATENT POLYTREES

In practice it is often difficult and even impossible to collect all the relevant time series when performing causal analysis on a dataset. The causal structure recovery literature currently is of two flavors when it comes to dealing with latent variables: one assumes that the underlying network has a specific causal structure, that is the flavor of this chapter. The other assumes a model that describes the dynamic among the latent and observed processes, which is the flavor of Chapter 6.

This chapter studies the problem of learning the causal structure of dynamics, where only a subset of random processes are observed. More specifically, we develop an approach for recovering directed graphs whose underlying structure is a polytree and introduced an algorithm that can learn the entire casual structure (observed and latent nodes) using a so-called discrepancy measure.

5.1 Minimal Latent Polytree

Consider a set of random processes $\underline{\mathbf{X}}$ whose directed information graph is a polytree $\overrightarrow{T} = (V, \overrightarrow{E})$, abbreviated as DIT. Denote $\underline{\mathbf{O}} = \{\mathbf{X}_1, ..., \mathbf{X}_m\}$ as the set of observable processes and their corresponding nodes in the DIT is denoted by O. Likewise, denote $\underline{\mathbf{L}} = \{\mathbf{Y}_1, ..., \mathbf{Y}_k\}$ as the set of latent processes and their corresponding nodes are denoted by L. Briefly, $\underline{\mathbf{X}} = \underline{\mathbf{O}} \cup \underline{\mathbf{L}}$ is the set of random processes and $V = O \cup L$ is their corresponding nodes in the DIT.

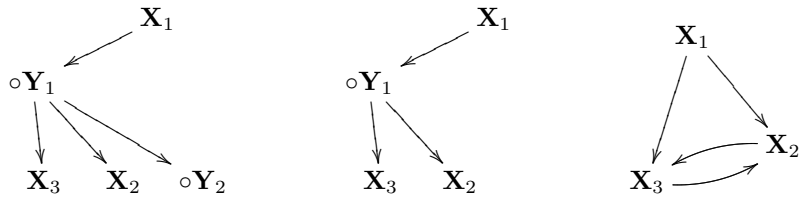
A probability distribution $P_{\underline{\mathbf{O}}}$ is called *polytree-decomposable* if there exists a joint distribution of the form $P_{\underline{\mathbf{O}} \cup \underline{\mathbf{L}}}$ that satisfies Assumption 1 and its corresponding DIG is a polytree. In this case, $P_{\underline{\mathbf{O}} \cup \underline{\mathbf{L}}}$ is called a polytree-extension of $P_{\underline{\mathbf{O}}}$.

Example 9. *Consider an array of five random processes* $\underline{\boldsymbol{X}} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{Y}_1, \boldsymbol{Y}_2)$ *with the joint dynamics:*

$$\underline{\boldsymbol{X}}_t = \underline{\boldsymbol{X}}_{t-1}\boldsymbol{A} + \underline{\boldsymbol{X}}_{t-2}\boldsymbol{B} + \underline{\boldsymbol{W}}_t,$$

where $\underline{\boldsymbol{X}}_t$ *is the row vector* $(X_{1,t}, X_{2,t}, X_{3,t}, Y_{1,t}, Y_{2,t})$*, and* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are* 5×5 *real matrices such that their non-zero entries are* $\boldsymbol{A}(4,2), \boldsymbol{A}(1,4), \boldsymbol{A}(4,5)$*, and* $\boldsymbol{B}(4,3)$ *and they are all equal to 0.5.* $\underline{\boldsymbol{W}}$ *is a set of 5 jointly independent random processes. Figure 5.1(a) illustrates the corresponding DIG of the whole system. Figure (b) and (c) are obtained by marginalizing over* $\boldsymbol{Y}_2$ *and* $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2\}$*, respectively. Since there exists at least one joint distribution such that its corresponding DIG has polytree structure,* $P_{\underline{\boldsymbol{O}}}$ *is polytree-decomposable, where* $O = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$.

A latent node $h \in L$ is called redundant if the directed information graph corresponding to the joint distribution of observed and latent nodes excluding $\mathbf{Y}_h$, $(P_{\underline{\mathbf{O}}, \underline{\mathbf{L}} \setminus \{\mathbf{Y}_h\}})$ remains a forest, i.e., a collection of



(a) (b) (c)

Figure 5.1. The DIGs of Example 3. (a) shows the DIT corresponding to $P_{\underline{\mathbf{X}}}$. (b) is the DIT corresponding to $P_{\underline{\mathbf{X}}\setminus\{\mathbf{Y}_2\}}$. (c) is the DIG corresponding to $P_{\underline{\mathbf{O}}}$. Latent nodes are indicated by circles.

polytrees. For instance in Example 9, Y_2 is a redundant hidden node. A latent directed information polytree (LDIT) is called *minimal* if it has no redundant hidden nodes[1]. The polytree in Figure 9(b) is minimal.

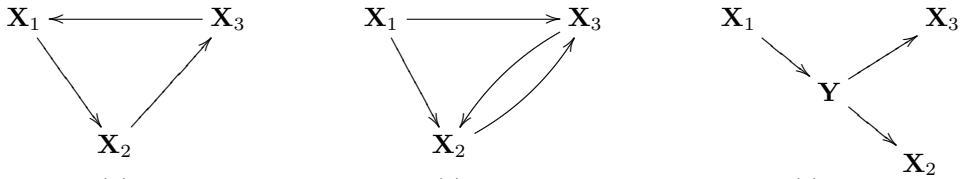
Assumption 3. *We assume that the joint distribution of the set of observed processes is polytree-decomposable.*

The next example demonstrates cases in which one is polytree-decomposable and the other is not.

Example 10. *Consider a set of 3 observable processes $\underline{\boldsymbol{X}}$ comprising a physical, dynamical system, such that the evolution of the processes over time satisfies the following stochastic equations:*

$$\begin{aligned} X_{1,t} &= X_{3,t-1}/3 + V_{1,t}, \\ X_{2,t} &= X_{1,t-1}/2 + V_{2,t}, \\ X_{3,t} &= X_{2,t-1}/2 + V_{3,t}, \end{aligned} \tag{5.1}$$

where (V_1, V_2, V_3) are three exogenous, independent processes. Figure 5.2(a) demonstrates the corresponding



(a) (b) (c)

Figure 5.2. Directed information graphs of Example 10.

DIG. For this small example, by checking all possible sets of auxiliary variables, we can conclude that there is no set of auxiliary variables $\underline{\boldsymbol{L}}$ such that $P_{\underline{\boldsymbol{X}}\cup\underline{\boldsymbol{L}}}$ both satisfies Assumption 1 and its corresponding DIG is a polytree. Now, consider the following discrete-time dynamical system with the corresponding DIG shown in Figure 5.2(b):

$$\begin{aligned} X_{1,t} &= V_{1,t}, \\ X_{2,t} &= X_{1,t-2}/2 + V_{4,t-1}/2 + V_{2,t}, \\ X_{3,t} &= X_{1,t-2}/3 + V_{4,t-1}/3 + V_{3,t}, \end{aligned} \tag{5.2}$$

[1]A redundant hidden node in [4] is defined as a hidden node that the joint distribution without it remains a tree instead of a forest.

where (V_1, V_2, V_3, V_4) are exogenous, independent processes. By defining $Y_t := X_{1,t-1} + V_{4,t}$, we can obtain a DIT as shown in Figure 5.2(c).

5.1.1 Some Properties of a Minimal LDIT

This section presents some properties of the DIT and the minimal LDIT, which will be used in Section 5.2 for structure learning.

Lemma 2. *Let $\overrightarrow{T} = (V, \overrightarrow{E})$ be the DIT corresponding to the joint distribution of a collection of random processes $\underline{\boldsymbol{X}}$. Let $\boldsymbol{X} \in \underline{\boldsymbol{X}}$ and $\mathcal{A}_1$ and $\mathcal{A}_2$ be two disjoint subsets of the parents of $\boldsymbol{X}$, i.e., $\mathcal{PA}(\boldsymbol{X})$. Then $\underline{\boldsymbol{X}}_{\mathcal{A}_1}$ and $\underline{\boldsymbol{X}}_{\mathcal{A}_2}$ are independent.*

Proof. See Appendix A.4.1. □

Lemma 3. *In a minimal LDIT, all hidden nodes have at least two children.*

Proof. See Appendix A.4.2. □

Lemma 4. *Consider a collection of random processes $\underline{\boldsymbol{X}}$ with a DIT $T = (V, \overrightarrow{E})$. If there is a directed path from j to i of length d, i.e., there is a sequence of nodes $(i_1, ..., i_{d-1})$ where j is the parent of i_1, i_k is the parent of i_{k+1} for $(1 \leq k \leq d-2)$, and i_{d-1} is the parent of i then*

$$D(P_{\boldsymbol{X}_i|\boldsymbol{X}_j} || \; P_{\boldsymbol{X}_i||_d \boldsymbol{X}_j}) = 0 \; . \tag{5.3}$$

Proof. See Appendix A.4.3. □

Lemma 4 implies that by walking along the path between two random process $\mathbf{X}_i$ and $\mathbf{X}_j$, each time we pass a node, the time dependency between $\mathbf{X}_i$ and $\mathbf{X}_j$ is shifted by at least one unit. In the next sections we will see that these time delays will help us recover the structure of a minimal LDIT. Time delays have also been used for infernce tasks in network forensic applications such as traffic analysis [137–140].

Lemma 5. *Suppose there exist two disjoint directed paths from $\boldsymbol{W}$ to $\boldsymbol{X}$ and $\boldsymbol{Y}$ in a minimal LDIT. Then*

$$D\left(P_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{W}} || P_{\boldsymbol{W}} P_{\boldsymbol{X}||\,\boldsymbol{W}} P_{\boldsymbol{Y}||\,\boldsymbol{W}}\right) = 0. \tag{5.4}$$

Proof. See Appendix A.4.4. □

Lemma 6. *In a minimal LDIT, if the root ancestors [2] of two nodes are disjoint, they are independent.*

Proof. See Appendix A.4.5. □

Another property which plays an essential role in learning the latent structure is what we call *sibling resemblance*.

Definition 11. *A collection of random processes $\underline{\boldsymbol{X}}$ with a corresponding minimal LDIT, $\overrightarrow{T} = (V, \overrightarrow{E})$, satisfies sibling resemblance property, if for every pair $(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $(i \neq j)$, of sibling with common parent $\boldsymbol{X}_k$ the following property holds: If there exists a time s such that $I(X_{i,1}^s; \boldsymbol{X}_k) > 0$, then $I(X_{i,s}; \boldsymbol{X}_j | X_{i,1}^{s-1}) > 0$*

[2] The set of roots that are ancestors of a given node in a directed tree is called root ancestors of that node.

This property simply states that in a minimal LDIT, the information inherited from a node to its children is not independent. Many dynamical systems such as autoregressive models satisfy this property. Next example illustrates the importance of this property for learning latent polytrees.

Example 11. *Consider a minimum LDIT with two observable and one latent random processes denoted by* $\underline{\boldsymbol{X}} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}_1\}$. *Let* $X_{1,t+1} = 0.2Y_{1,2t-1} + \epsilon_{1,t+1}$ *and* $X_{2,t+1} = -0.9Y_{1,2t} + \epsilon_{2,t+1}$, *where* $\epsilon_{1,t}$, $\epsilon_{2,t}$, *and* $\boldsymbol{Y}_1$ *are jointly independent. The corresponding DIG of this system is* $\boldsymbol{X}_1 \leftarrow \boldsymbol{Y}_1 \rightarrow \boldsymbol{X}_2$. *Suppose that* $\{Y_{1,2t}\}$ *and* $\{Y_{1,2t-1}\}$, *i.e., the even and odd sub processes of* $\boldsymbol{Y}_1$ *are independent. In this case* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *are independent and detecting the hidden confounder between them is impossible. This system does not satisfy the sibling resemblance property since* $\boldsymbol{X}_1$ *and* $\boldsymbol{X}_2$ *are sibling with* $\boldsymbol{Y}_1$ *as their common parent and* $I(X_1^2; \boldsymbol{Y}_1) > 0$, *(*$s = 2$*), but* $I(X_{1,2}; \boldsymbol{X}_2 | X_{1,1}) = 0$.

5.1.2 Presence of Simultaneous Influences

Excluding simultaneous influences helps us write equation (2.3) which consequently leads to the definition of generative model graphs in Section 2.2. Now the question is, what if there were in fact simultaneous influences?

In this section, we show that if there are simultaneous influences between processes, the corresponding DIG is not a polytree and hence it cannot be recovered by our proposed method. To make the statement rigorous, we need to modify the definition of the directed information graph by using the original Kramer's causal conditioning that allows for simultaneous influences. For $\mathcal{K} \subseteq -\{j\}$ define

$$\widetilde{P}_{\mathbf{X}_j || \, \underline{\mathbf{X}}_{\mathcal{K}}} := \prod_{t=1}^{n} P_{X_{j,t} | X_{j,1}^{t-1}, \underline{X}_{\mathcal{K},1}^{t}}, \tag{5.5}$$

and the modified conditional directed information as

$$\widetilde{I}\left(\mathbf{X}_j \rightarrow \mathbf{X}_i || \underline{\mathbf{X}}_{\mathcal{K}}\right) := \mathbb{E}_{P_{\underline{\mathbf{X}}_{\mathcal{K} \cup \{i,j\}}}} \left[\log \frac{d\widetilde{P}_{\mathbf{X}_i || \mathbf{X}_j, \underline{\mathbf{X}}_{\mathcal{K}}}}{d\widetilde{P}_{\mathbf{X}_i || \underline{\mathbf{X}}_{\mathcal{K}}}} \right]$$

Using the above measure, we are able to define the *modified directed information graph* (MDIG) that captures the simultaneous effects as such: there is an arrow from node j to node i for $i, j \in \{1, ..., m\}$ in the MDIG if and only if

$$\widetilde{I}\left(\mathbf{X}_j \rightarrow \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,j\}}\right) > 0.$$

Theorem 7. *Let* $\overrightarrow{T}$ *to be a MDIG over a set of random processes* $\underline{\boldsymbol{X}}$ *which is a polytree and let* $\mathcal{PA}(\boldsymbol{X})$ *to be the parent set of* $\boldsymbol{X}$ *in* $\overrightarrow{T}$, *then*

$$D\left(\widetilde{P}_{\boldsymbol{X} || \ \underline{\boldsymbol{X}}_{\mathcal{PA}(\boldsymbol{X})}} || P_{\boldsymbol{X} || \ \underline{\boldsymbol{X}}_{\mathcal{PA}(\boldsymbol{X})}}\right) = 0.$$

Proof. See Appendix A.4.6. □

A consequence of the above result is that the corresponding DIG of a system with simultaneous influences is not a polytree. This is because, when the corresponding MDIG of a dynamical system is a polytree, based on the above result, all the simultaneous influences can be dropped.

5.2 Recovery of Latent Polytrees

A simple observation about a directed ploytree is that each pair of nodes that are the descendants of the same root has a unique common ancestor. In this section, we define a notion of distance on a polytree in order to determine the distance of each pair of nodes to their common ancestor, if it exists. Moreover, we will show that given these distances for a subset of nodes, the graph is uniquely recoverable.

Definition 12. *Given a polytree $\overrightarrow{T} = (V, \overrightarrow{E})$ with the root set $\mathcal{R}$, every function $\gamma : V \times V \to \mathbb{R}$ that satisfies the following criterion is called a discrepancy on $\overrightarrow{T}$. γ assigns a real number to the path from v_1 to the common ancestor of v_1 and v_2, such that*

1. *$\gamma(v_1, v_2) = 0$ if and only if either v_1 is the ancestor of v_2 or $v_1 = v_2$.*
2. *If the common ancestor of v_1 and v_2 is the same as the common ancestor of v_1 and v_3, then*

$$\gamma(v_1, v_2) = \gamma(v_1, v_3).$$

3. *If the common ancestor of v_1 and v_2 is on the path from the common ancestor of v_1 and v_3 to v_1, then*

$$\gamma(v_1, v_2) < \gamma(v_1, v_3).$$

4. *$\gamma(v_1, v_2) < 0$ if and only if v_1 and v_2 have no common ancestor.*

The image of such these functions can be presented by the discrepancy matrix:

$$\Gamma_V := [\gamma_r(v_i, v_j)]\,,\ v_i, v_j \in V.$$

Note that for a given polytree, the discrepancy matrix is not unique. Any function that satisfies the conditions in Definition 12 is a valid discrepancy measure.

Example 12. *Consider the polytree depicted in Fig.5.3 with roots $\{v_5, v_6\}$ and the following discrepancy matrix:*

$$\Gamma_V = \begin{pmatrix} 0 & 2 & 3 & 1 & 3 & 4 \\ 0 & 0 & -2 & 0 & -1 & 1 \\ 1 & -3 & 0 & 1 & 1 & -3 \\ 0 & 1 & 2 & 0 & 2 & 3 \\ 0 & -1 & 0 & 0 & 0 & -2 \\ 0 & 0 & -1 & 0 & -1 & 0 \end{pmatrix}.$$

For instance, looking at the third row, this particular discrepancy function assigns 1 to the path from v_3 to its common ancestor with v_1, i.e., v_5. Since v_2 and v_3 have no common ancestor, $\Gamma_V(3, 2) < 0$.

We prove that the discrepancy matrix suffices to uniquely learn the topology of a polytree $\overrightarrow{T} = (V, \overrightarrow{E})$. We also present an algorithm that learns the structure of a polytree given the discrepancies between all the pairs of observed nodes.

Definition 13. *In a polytree $\overrightarrow{T} = (V, \overrightarrow{E})$, we call a subset $L \subset V$ learnable, if every node $v \in L$ has at least two outgoing arrows. We call $O := V \setminus L$ the set of observed nodes.*

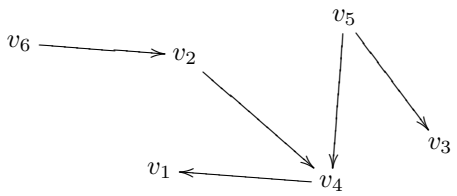


Figure 5.3. The directed tree of Example 12.

For example, $\{v_5\}$ is a learnable subset of the polytree shown in Figure 5.3. From Definition 13, if L is a learnable subset of a polytree, then all the leaves belong to $O = V \setminus L$.

Theorem 8. *Let $\overrightarrow{T} = (V, \overrightarrow{E})$ be a polytree with the root set $\mathcal{R}$ and let $L \subseteq V$ be a learnable subset. Then existence of a discrepancy matrix Γ_O for $O = V \setminus L$ suffices for learning $\overrightarrow{T}$.*

Proof. See Appendix A.4.7. □

Next, inspired by the steps in the proof of Theorem 8, we present an algorithm for structure learning of polytrees.

5.2.1 Structure Recovery Algorithm

The rational of the proposed algorithm in this section follows the three main steps of proof of Theorem 8: the first step is to discover the number of roots $|\mathcal{R}|$ of the underlying polytree and all their descendants in the set of observed nodes (O) given the discrepancy matrix Γ_O. This can be done by fixing a node $v \in O$ and finding a maximal subset of O containing v in which every pairs of nodes have positive discrepancy (Algorithm 2).
Next step is to recover the underlying tree for every root $r \in \mathcal{R}$ given its discovered descendants in the set O. This can be done using the recursive approach summarized in Algorithm 3.
The last step is to merge the recovered trees from the previous step to recover the underlying polytree. This too is possible, since if two recovered trees are connected, their common subgraph is also a tree; thus, it can be learned using Algorithm 3. Algorithm 4 describes the required steps.

Next, we present our algorithm that learns a polytree given a discrepancy matrix on its observed nodes using the aforementioned three main steps.. A simple example that illustrates the algorithm is also provided. First, we need the following definition.

Definition 14. *A tree merger is an operator that takes two directed trees $\overrightarrow{T}_1, \overrightarrow{T}_2$ and a given sub-tree of both of them, say $\overrightarrow{T}_3$ and merges them at $\overrightarrow{T}_3$. We denote this operation by $\overrightarrow{T}_1 \circ \overrightarrow{T}_2|_{\overrightarrow{T}_3}$.*

Figure 5.4 depicts one such tree merger.

Polytree(Γ_O) presents an algorithm for learning the polytree $\overrightarrow{T}(V, \overrightarrow{E})$ with the root set $\mathcal{R}$ given the discrepancy matrix Γ_O on its observed nodes O. First, it calls the subroutine **Separation**(Γ_O) which finds subsets O_is, where $O = \cup_i O_i$ such that each subset corresponds to observed nodes in a directed tree with a single root. Each of these single rooted sub-trees can be learned by Algorithm **Tree**(O). To complete the task, Algorithm **Polytree**(Γ_O) must connect these sub-trees to recover the original polytree. This is done by using the fact that if a polytree $\overrightarrow{T}$ and a directed tree $\overrightarrow{T}_i$ have an intersection, then their intersection will be a directed tree. Thus it also could be learned by Algorithm **Tree**(O). After learning the intersection part,

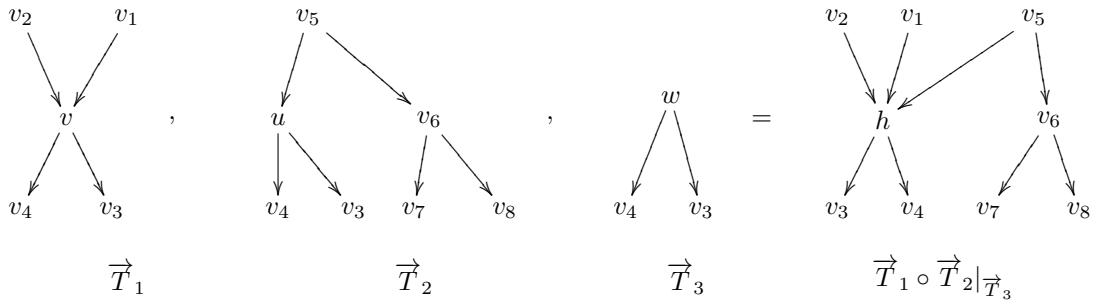


Figure 5.4. An example that illustrates the merger operator between two directed trees.

Algorithm 2 : **Separation**(Γ_O)

Input : Γ_O
Output : $O_1, ..., O_{|\mathcal{R}|}$
$M \leftarrow \emptyset$, $i \leftarrow 1$
while $O \setminus M \neq \emptyset$ **do**
  Choose v in $O \setminus M$
  Find all $\mathcal{C} \subseteq O$ such that $v \in \mathcal{C}$ and
  for all $(u, w) \in \mathcal{C} \times \mathcal{C}, \gamma(u, w) \geq 0$.
  $O_i \leftarrow$maximal $\mathcal{C}$
  Return O_i
  $M \leftarrow M \cup O_i$
  $i \leftarrow i + 1$
end while

Algorithm **Polytree**(Γ_O) uses the tree merger operator defined in Definition 14 to connect these together. In Algorithm **Tree**(O), $\overrightarrow{T}_1 \oplus \overrightarrow{T}_2(h)$ is an operator that connects a directed tree $\overrightarrow{T}_1 = (V_1, \overrightarrow{E}_1)$ with root r_1 to a polytree $\overrightarrow{T}_2 = (V_2, \overrightarrow{E}_2)$ given a leaf of $\overrightarrow{T}_2$, h, by simply substituting h in $\overrightarrow{T}_2$ by $\overrightarrow{T}_1$. More precisely,

$$\overrightarrow{T}_1 \oplus \overrightarrow{T}_2(h) := \left(V_1 \cup V_2 \setminus \{h\}, \overrightarrow{E}\right),$$

where

$$\overrightarrow{E} = \overrightarrow{E}_1 \cup \{(\mathcal{PA}_2(h), r_1)\} \cup \overrightarrow{E}_2 \setminus \{(\mathcal{PA}_2(h), h)\},$$

and $\mathcal{PA}_2(h)$ is given by (1.8) and it represents the set of parents of h in $\overrightarrow{T}_2$. Figure 5.5(b) depicts an example.

Example 13. *Consider the polytree in Example 12. Assume* $O = \{v_1, v_2, v_3, v_4, v_6\}$. *Then, by the definition* $V \setminus O = \{v_5\}$ *is a learnable subset. Given the discrepancy matrix*

$$\Gamma_O = \begin{pmatrix} 0 & 2 & 3 & 1 & 4 \\ 0 & 0 & -2 & 0 & 1 \\ 1 & -3 & 0 & 1 & -3 \\ 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & -1 & 0 & 0 \end{pmatrix},$$

Algorithm 4 calls ***Separation*** *to find all sub-trees with single roots, which are* $O_1 = \{v_1, v_2, v_4, v_6\}$ *and* $O_2 = \{v_1, v_3, v_4\}$. *As one can see in Figure 5.3, the sub-trees induced by* O_1 *and* O_2 *each have a single root.*

Algorithm 3 : **Tree**(O)

1: *Input* : Γ_O
2: *Output* : $\overrightarrow{T} = (V, \overrightarrow{E})$
3: For all $v \in O$
4: $B_v \leftarrow \arg\min_{u \in O \setminus \{v\}} \gamma(v, u)$
5: **if** $B_v = O \setminus \{v\}\ \forall v \in O$ **then**
6: **if** $\exists w \in O : \min_{u \in O \setminus \{w\}} \gamma(w, u) = 0$ **then**
7: $\overrightarrow{T}$ is a star graph with w as the root in the center.
8: **else**
9: $\overrightarrow{T}$ is a star graph with a hidden node as the root in the center.
10: **end if**
11: **else**
12: Choose w such that $B_w \neq O \setminus \{w\}$
13: $\overrightarrow{T}' \leftarrow$**Tree**$(B_w \cup \{w\})$
14: $\overrightarrow{T}'' \leftarrow$**Tree**$(O \setminus B_w)$
15: Substitute w in $\overrightarrow{T}''$ by another node, say h.
16: $\overrightarrow{T} \leftarrow \overrightarrow{T}' \oplus \overrightarrow{T}''(h)$
17: **end if**

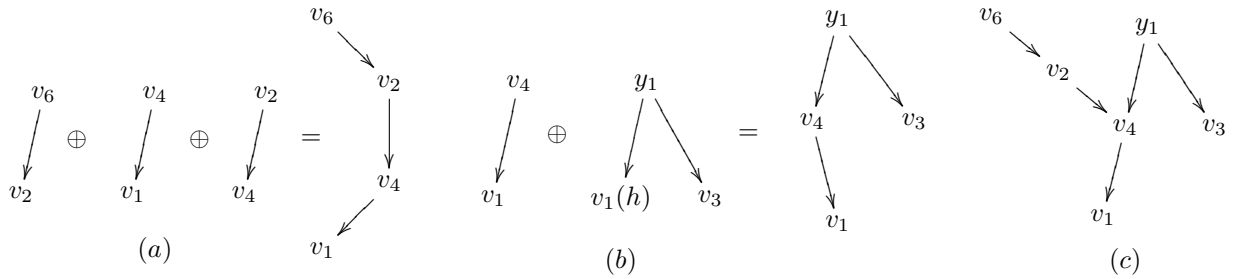


Figure 5.5. (a) Illustrate the steps and outputs of **Tree**$(\{v_1, v_2, v_4, v_6\})$. (b) Illustrate the steps and outputs of **Tree**$(\{v_1, v_3, v_4\})$. (c) Illustrates merging the first two directed trees by sharing their common sub-tree which is obtained by **Tree**$(\{v_1, v_4\})$.

Subsequently, Algorithm 4 calls ***Tree*** *to build the sub-trees. Figures 5.5(a) and 5.5(b) illustrate these sub-trees. For instance, the subtree in Figure 5.5(a) is obtained as follows: Algorithm 3 computes B_{v_i}s for $i \in \{1, 2, 4, 6\}$ at step 4. Since $B_{v_2} = \{v_1, v_4\} \neq O_1 \setminus \{v_2\}$, the condition in step 5 is not satisfied and Algorithm 3 will jump to step 12 and chooses w to be v_2. In step 13 and 14, the algorithm recursively calls itself but this time given $\{v_1, v_2, v_4\}$ and $\{v_2, v_6\}$, respectively. Since the sub-tree induced by $\{v_2, v_6\}$ is a star, it will be constructed in steps 5 to 10. On the other hand, the sub-tree induced by $\{v_1, v_2, v_4\}$ is not a star. It is learned by breaking it into two stars as shown in Fig. 5.5(a).*

Finally, Algorithm 4 must reconnect the sub-trees depicted in Fig. 5.5(a) and 5.5(b). To do so, it finds the common sub-tree between them at steps 8 and 9, and it merges the trees in Fig. 5.5(a) and 5.5(b) together at step 11. The final result is shown in Figure 5.5(c).

Algorithm 4 : **Polytree**(Γ_O)

Input : Γ_O
Output : $\overrightarrow{T} = (V, \overrightarrow{E})$
Separation(Γ_O).
$\overrightarrow{T} \leftarrow$**Tree**($O_1$)
$\mathcal{S} \leftarrow O_1, \mathcal{I} \leftarrow \{1\}$
while $\mathcal{I} \neq \{1, 2, ..., |\mathcal{R}|\}$ **do**
  Find $i \in \{1, 2, ..., |\mathcal{R}|\} \setminus \mathcal{I}$ such that $O_i \cap \mathcal{S} \neq \emptyset$
  $\overrightarrow{T}_{sub} \leftarrow$**Tree**($\mathcal{S} \cap O_i$)
  $\overrightarrow{T}_i \leftarrow$**Tree**($O_i$)
  $\overrightarrow{T} \leftarrow \overrightarrow{T} \circ \overrightarrow{T}_i|_{\overrightarrow{T}_{sub}}$
  $\mathcal{S} \leftarrow \mathcal{S} \cup O_i$
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{i\}$
end while

5.3 Discrepancy Measure for Latent Directed Information Polytrees

In this section, we establish a discrepancy measure for learning minimal directed information polytrees. Recall that Lemma 4 states that the lag between random processes grows by walking along the directed paths in a minimal DIT. This allows us to have the following definition in such graphs.

Definition 15. *For any pair of random processes $(\boldsymbol{X}_j, \boldsymbol{X}_k) \in \underline{\boldsymbol{O}} \times \underline{\boldsymbol{O}}$, we define the directed measure from $\boldsymbol{X}_j$ to $\boldsymbol{X}_k$ denoted by $\gamma(\boldsymbol{X}_j, \boldsymbol{X}_k)$ as follows: If $I(\boldsymbol{X}_k; \boldsymbol{X}_j) = 0$, then $\gamma(\boldsymbol{X}_j, \boldsymbol{X}_k) = -1$, and*

$$\gamma(\boldsymbol{X}_j, \boldsymbol{X}_k) := \begin{cases} \max_{d \geq 0} \left\{ d : \ I\left(X_{j,1}^d; \boldsymbol{X}_k\right) = 0 \right\} & j \neq k \\ 0 & j = k. \end{cases} \tag{5.6}$$

Note that $I\left(X_{j,1}^0; \boldsymbol{X}_k\right) = 0$.

Theorem 9. *Let $\underline{\boldsymbol{X}} = \underline{\boldsymbol{O}} \cup \underline{\boldsymbol{L}}$ be a collection of random processes which form a minimal LDIT, $\overrightarrow{T} = (V, \overrightarrow{E})$, where $V = O \cup L$. If $\underline{\boldsymbol{X}}$ satisfies Assumptions 1, 3, and the sibling resemblance property, then the directed measure defined above is an admissible discrepancy and L is a learnable subset.*

Proof. See Appendix A.4.8. □

5.4 Sample Complexity for Empirical Estimator

This section studies the complexity of the proposed algorithm to recover the minimal LDIT given N i.i.d. samples of the observed random processes, $\{\underline{\mathbf{O}}^{(1)}, ..., \underline{\mathbf{O}}^{(N)}\}$, where $\underline{\mathbf{O}}^{(q)} = \{\mathbf{X}_1^{(q)}, ..., \mathbf{X}_m^{(q)}\}$ denotes the q-th sample from all the m processes. $\mathbf{X}_i^{(q)} \in \mathcal{X}^n$ for each i. Consider the case that the alphabet set $\mathcal{X}$ is finite. In order to learn the minimal LDIT we need to estimate the directed measures introduced in the previous section between all pairs of observed processes. To do so, first we estimate the joint distributions for each pair $(\mathbf{X}_i, \mathbf{X}_j)$ using the empirical estimator defined as

$$\widehat{P}_{\mathbf{X}_i, \mathbf{X}_j}(\mathbf{x}_i, \mathbf{x}_j) := \frac{1}{N} \sum_{q=1}^{N} \mathbb{I}_{\left\{(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{X}_i^{(q)}, \mathbf{X}_j^{(q)})\right\}}, \tag{5.7}$$

where $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{X}^n \times \mathcal{X}^n$ and $\mathbb{I}$ is the indicator function. Using the empirical distribution of (5.7), we can compute the empirical entropies and consequently, the empirical mutual information.

Lemma 7. *Given N i.i.d. samples of two random processes, $\boldsymbol{X}_1 \in |\mathcal{X}|^{d_1}$ and $\boldsymbol{X}_2 \in |\mathcal{X}|^{d_2}$, $d_1, d_2 \leq n$, we have*

$$\mathbb{P}\left(|I(\boldsymbol{X}_1; \boldsymbol{X}_2) - \widehat{I}(\boldsymbol{X}_1; \boldsymbol{X}_2)| \geq \epsilon\right) \leq 6|\chi|^{2n}\ e^{-N\xi_n(\epsilon)},$$

where $\xi_n(\epsilon) > 0$ and it is given by

$$\xi_n(\epsilon) = 2\exp\left(\frac{2\log\frac{\epsilon}{3|\chi|^{2n}}}{\log\frac{\epsilon}{3|\chi|^{2n}} - 1}\log\frac{\epsilon}{3|\chi|^{2n}\log\frac{3|\chi|^{2n}}{\epsilon}}\right). \tag{5.8}$$

Proof. See Appendix A.4.9. □

As long as there exists an estimator for the mutual information $\widehat{I}(\cdot\ ;\ \cdot)$, such as the empirical estimator in (5.7), we can estimate the directed measure (5.6) from $\mathbf{X}_i$ to $\mathbf{X}_j$ by estimating $\widehat{I}(\mathbf{X}_j; X_{i,1}^d)$ for $d = 1, ..., n$. After choosing an appropriate threshold $\rho > 0$, our estimate of directed measure will be the smallest d for which $\widehat{I}(\mathbf{X}_j; X_{i,1}^d) > \rho$:

$$\widehat{\gamma}(\mathbf{X}_i, \mathbf{X}_j) := \min\{d : \widehat{I}(\mathbf{X}_j; X_{i,1}^d) > \rho\}. \tag{5.9}$$

Theoretically, the best possible threshold is

$$\rho^* := \min_{i \neq j}\left\{I(X_{i,1}^{\gamma(\mathbf{X}_i, \mathbf{X}_j)+1}; \mathbf{X}_j)\right\}. \tag{5.10}$$

The next theorem presents a concentration bound for our estimate.

Theorem 10. *Given N i.i.d. samples of two random processes $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ each of length n, and threshold $0 < \rho \leq \rho^*$ in (5.10), we have*

$$\mathbb{P}\left(\gamma(\boldsymbol{X}_1, \boldsymbol{X}_2) \neq \widehat{\gamma}(\boldsymbol{X}_1, \boldsymbol{X}_2)\right) \leq 6n|\chi|^{2n}\ e^{-N\xi_n(\rho)},$$

where $\xi_n(\cdot)$ is given in (5.8).

Proof. Using definition (5.10), one can show $\{\gamma(\mathbf{X}_1, \mathbf{X}_2) \neq \widehat{\gamma}(\mathbf{X}_1, \mathbf{X}_2)\} \subseteq \bigcup_{k=1}^n \{|I_k - \widehat{I}_k| \geq \rho\}$, where

$$I_k := I(X_{1,1}^k; \mathbf{X}_2),\ \widehat{I}_k := \widehat{I}(X_{1,1}^k; \mathbf{X}_2).$$

Applying the union bound and Lemma 7 concludes the proof. □

Most of the practical dynamical systems have finite memory, i.e., they have finite Markov order. In such scenarios, the sample complexity reduces extensively. More precisely, consider a dynamical system with finite Markov order p, then in order to estimate $I(X_{i,1}^d; \mathbf{X}_j)$, it suffices to estimate the estimating mutual information between two random processes each of length at most $p + 1$. This is true because for a process $\mathbf{X}_j$ of length n and finite Markov order p, we have

$$H(\mathbf{X}_j) = \sum_{t=1}^n H(X_{j,t}|X_{j,t-p}^{t-1}) = \sum_{t=1}^n H(X_{j,t-p}^t) - H(X_{j,t-p}^{t-1}). \tag{5.11}$$

Using the result of Lemma 7, Theorem 10, and Equation (5.11), we obtain the following sample complexity for a network with finite Markov order.

Corollary 2. *Given N i.i.d. samples of two random processes $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ each of length n with finite Markov order p, and threshold $0 < \rho \leq \rho^*$ in (5.10), we have*

$$\mathbb{P}\left(\gamma(\boldsymbol{X}_1, \boldsymbol{X}_2) \neq \widehat{\gamma}(\boldsymbol{X}_1, \boldsymbol{X}_2)\right) \leq 6n^2|\chi|^{2p+2}\ e^{-N\xi_{p+1}(\rho/n)}.$$

Let $\widehat{\overrightarrow{T}}_N = (\widehat{V}_N, \widehat{\overrightarrow{E}}_N)$ denote the reconstructed polytree using the empirical directed measures (5.9) given N i.i.d. samples from the observable processes and assume that the true minimal LDIT was $\overrightarrow{T} = (V, \overrightarrow{E})$. Define the error event as

$$\{\overrightarrow{T} \not\equiv \widehat{\overrightarrow{T}}_N\} := \{V \neq \widehat{V}_N\} \cup \{\overrightarrow{E} \neq \widehat{\overrightarrow{E}}_N\}.$$

That is, an error occurs in the reconstruction algorithm, if the set of constructed nodes and edges are not precisely those of the true polytree $\overrightarrow{T}$.

Corollary 3. *Consider a minimal LDIT $\underline{\boldsymbol{X}} = \underline{\boldsymbol{O}} \cup \underline{\boldsymbol{L}}$ consisting of m observable nodes. Given N i.i.d. samples from each of the observable processes,*

$$\mathbb{P}\left(\overrightarrow{T} \not\equiv \widehat{\overrightarrow{T}}_N\right) \leq 12\binom{m}{2} n|\chi|^{2n}\ e^{-N\xi_n(\rho)},$$

where $0 < \rho \leq \rho^*$ and $\xi_n(\cdot)$ is given in (5.8).

Proof. Theorem 8 states that given the discrepancies between all pair of observed nodes, $\overrightarrow{T}$ is recoverable. Since there are m such nodes, $2\binom{m}{2}$ directed measures need to be estimated. Theorem 10 and union bound establish the result. □

5.5 Experimental Results

In this section, we present our experimental results for both synthetic linear system and non-linear system, and a real dataset.

Autoregressive Model:

We simulated a network of 14 processes corresponding to a polytree with 3 roots in which 4 processes were latent. We observed $N \in \{2000, 4000\}$ i.i.d. samples from every observed process each of length $n = 20$. They were modeled as zero-mean multivariate normal autoregressive time-series such that $\underline{\mathbf{Z}}_t = \sum_{i=1}^{3} \mathbf{A}_i \underline{\mathbf{Z}}_{t-i} + \underline{\mathbf{W}}_t$, where $\underline{\mathbf{Z}}_t, \underline{\mathbf{W}}_t \in \mathbb{R}^{14}$ and $\mathbf{A}_i \in \mathbb{R}^{14\times 14}$. $\mathbf{W}_i$s were generated i.i.d. Gaussian with mean zero and variance one. The non-zero entries of $\mathbf{A}_i$s are given in Table 5.1. The first four processes of $\underline{\mathbf{Z}}$ denoted by $(\mathbf{Y}_1, ..., \mathbf{Y}_4)$ were the latent ones.

Mutual information between two jointly Gaussian random processes $\mathbf{X}$ and $\mathbf{Y}$ is given by [25] $I(\mathbf{X}; \mathbf{Y}) = -0.5 \log \frac{|\Sigma_{\mathbf{X},\mathbf{Y}}|}{|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|}$, where $\Sigma_{\mathbf{X}}$ is the covariance matrix of process $\mathbf{X}$, and $\Sigma_{\mathbf{X},\mathbf{Y}}$ is the covariance matrix of $(\mathbf{X}, \mathbf{Y})$. Hence, we were able to estimate the discrepancies (5.9) by estimating the covariance matrices between the observed processes. Figure 5.6(a) and 5.6(b) illustrate the recovered structure for $N = 2000$ and $N = 4000$, respectively.

To compute each directed measure pair γ_{jk}, we estimated quantities $f_{j,k}(d) := \widehat{I}\left(X_{j,1}^d; \mathbf{X}_k\right)$, for $1 \leq d \leq 20$ using the above expression for the mutual information. If for all $1 \leq d \leq 20$, $f_{j,k}(d)$ is less than τ, a sufficiently

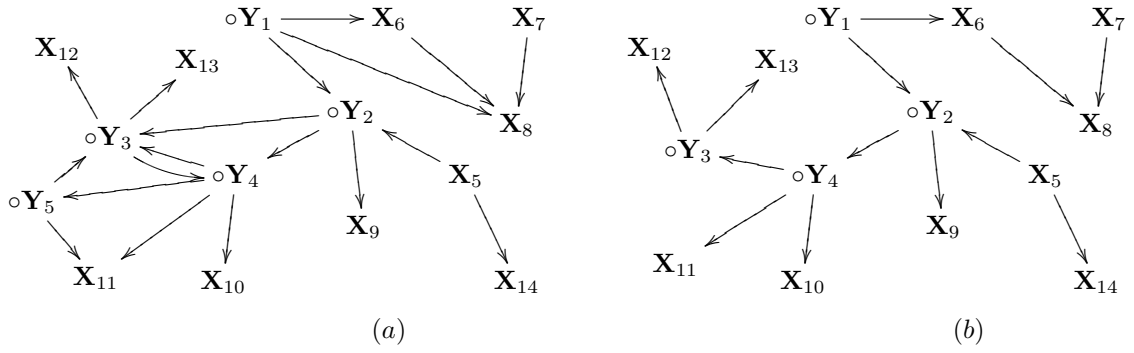


Figure 5.6. Recovered polytree of the AR model. Latent nodes are indicated by circles.

small threshold (in this example $\tau = 0.05$), we set the directed measure from j to k, $\gamma_{j,k}$ to -1. Otherwise, it is set to equal a value d^*, where d^* is the first value at which $f_{j,k}(d)$ makes a significant jump. That is, $f_{j,k}(d^*)$ is greater than its preceding values $\{f_{j,k}(i), i < d^*\}$. This means ρ in Section 5.4 was set to equal $f_{j,k}(d^* - 1)$.

The reason we see cycles for small number of sample is because of estimation errors. When the number of samples are not sufficient to estimate the entries of the discrepancy matrix correctly, the resulting discrepancy matrix will violate some constraints in Definition 12, particularly constraint (2), which will enforce the algorithm to add cycles in order to be consistent with the estimated discrepancy matrix.

$\mathbf{A}_1$	$A_1(1,1) = 1,\ A_1(2,1) = 1,\ A_1(2,2) = 0.5,\ A_1(2,5) = \sqrt{2}/2,\ A_1(3,4) = 1,\ A_1(5,5) = 1,\ A_1(6,1) = -2,\ A_1(8,8) = 1,\ A_1(8,7) = 0.1,\ A_1(10,10) = 0.3,\ A_1(12,12) = \sqrt{2},\ A_1(13,13) = -0.2,\ A_1(13,3) = -1,\ A_1(14,5) = 0.2.$
$\mathbf{A}_2$	$A_2(3,3) = -1,\ A_2(5,5) = 0.2,\ A_2(7,7) = \sqrt{2},\ A_2(8,8) = 1,\ A_2(8,7) = 0.2,\ A_2(9,9) = 3,\ A_2(9,2) = 2.5,\ A_2(10,4) = -1,\ A_2(11,11) = 1,\ A_2(12,3) = -\sqrt{2}.$
$\mathbf{A}_3$	$A_3(4,2) = \sqrt{3},\ A_3(6,6) = 1,\ A_3(8,6) = 0.6,\ A_3(11,4) = -2.$

Table 5.1. Non-zero coefficients of the AR model.

A Non-linear Model:

We simulated a network of 7 processes, which formed a polytree with 2 roots in which 2 processes were latent. Denoting the latent processes with Y and the observed ones with X, the model is expressed as

$$
\begin{aligned}
Y_{1,t} &= Y_{1,t-3} + 0.1Y_{1,t-2}^2 + \zeta_{1,t},\\
X_{1,t} &= X_{1,t-1}^2/\sqrt{2} - 0.1|\zeta_{2,t}|,\\
Y_{2,t} &= Y_{2,t-1} - X_{1,t-1} + 1.5\sqrt{|Y_{1,t-1}|} + \zeta_{3,t},\\
X_{2,t} &= -2Y_{2,t-1} + 0.3\sqrt{|X_{2,t-3}|^3} + \zeta_{4,t},\\
X_{3,t} &= 2X_{3,t-2} - 0.2Y_{2,t-1} + \zeta_{5,t},\\
X_{4,t} &= X_{4,t-1} + \sqrt{|2X_{4,t-2}|} - Y_{1,t-1} + 2Y_{1,t-2} + 0.7\log|Y_{1,t-3}| + \zeta_{6,t},\\
X_{5,t} &= 3X_{5,t-2} + 2.5X_{4,t-2} + \zeta_{7,t},
\end{aligned}
$$

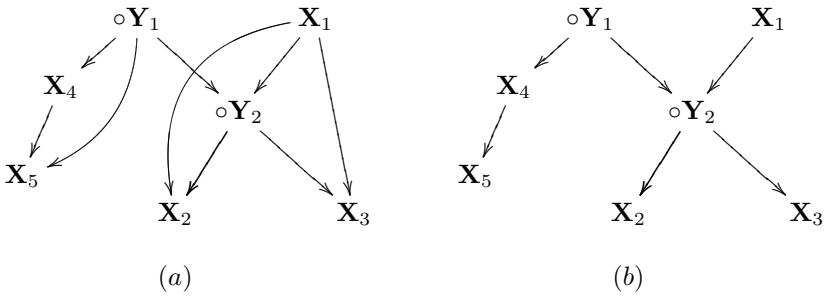


Figure 5.7. Recovered polytree of the non-linear model. Latent nodes are indicated by circles.

where ζ_is were generated i.i.d. Gaussian with mean zero and variance one. The observed variables $\{X_1, ..., X_5\}$ were each of length $n = 20$ and $N \in \{10^3, 10^4\}$ number of samples from each of them was collected. The directed measures were estimated using Equation (5.9) and the mutual information was estimated using 1-nearest neighbour method in [141]. The same thresholding procedure of Section 5.5 was used to decide whether the estimated mutual information are zero or positive. The recovered networks are depicted in Figure 5.7.

Market Analysis:

As an example of how our approach may discover causal structure in real-world data, we analyzed the causal relationship between stock prices of 10 technology companies of the New York Stock Exchange sourced from Google Finance for twenty market days (03/03/2008 -03/28/2008). In this simulation, we assumed that the underlying causal structure did not change during the sampling period. Furthermore, we assumed that influences took a business day to propagate among the stocks. Hence, the difference between, t and $t+1$, is one business day. To obtain i.i.d. samples, the price of each stock was sampled every two minutes during a business day. This amounted to $N = 200$ number of i.i.d. samples for each stock and $n = 20$.

For this experiment, we used the Black-Scholes model [135] for the market in which, the stock prices are modeled via a set of coupled stochastic partial differential equations. This model allows to model the logarithm of the stocks prices as an autoregressive model [142]. Thus, the directed measure were estimated similar to Section 5.5 from the logarithm of the stock's prices.

Since the underlying true DIG of these 10 companies is not necessarily a polytree, we first approximated the DIG graph of the network by the best directed tree, where best is in the sense of minimizing the Kullback-Leibler (KL) divergence between the true joint and the one resulting from the directed tree approximation. It was shown in [80] that the optimal approximate directed tree maximizes the sum of pair-wise directed information terms. Thus, to obtain the best tree approximation, we estimated the pair-wise directed information and found the maximum spanning tree. As depicted in Figure 5.8(a), the approximation identified two disjoint trees. In order to obtain a polytree, we connected the two sub-trees by the arrow with maximum directed information weight between the nodes of the two sub-trees. This edge was (HP,EMC) as shown in Figure 5.8(b).

HP and IBM are the roots in polytree depicted in Figure 5.8(b). This suggests that they had significant influences on the other companies' stock prices during 2008. In fact, Gartner, Inc. had ranked IBM as the worldwide share leader in the enterprise portal software market based on total software revenue[3]. Further-

[3]IBM, https://www-03.ibm.com/press/us/en/pressrelease/24507.wss

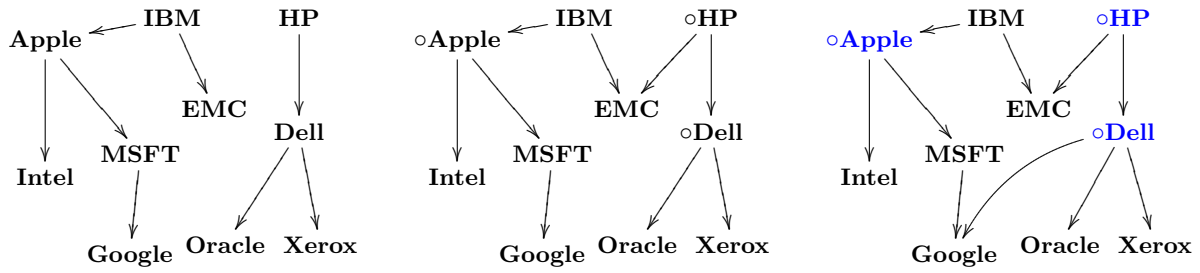


Figure 5.8. The polytree of the market data. In (b) latent nodes are indicated by circles. Recovered polytree of the market is in (c).

more, HP was the global PC market share leader during the same period followed by Dell Inc[4]. Another observation is the detected influence of Apple on Intel and Microsoft. Although Apple had begun using Intel processors in its products since 2006, it was only in 2008 that it released MacBook Air and upgraded the processors of MacBook and MacBook Pro to Intel core 2 Duo Penryn. Thus, it causes Intels stock price to increase. The arrow from Apple to Microsoft might be a result of the following phenomenon, during 2007-2008, Apples Mac OS X posted its biggest gain, while Windows OS market share dived below 90% for the first time[5].

To test out latent learning algorithm, we removed the data for the following three companies: Apple, HP, and Dell in the polytree of Figure 5.8(b) and ran our algorithm with the data from the remaining 7 companies. We used the same thresholding procedure of Section 5.5 to obtain the directed measures. The estimated discrepancy matrix is given in (23) and the recovered polytree is shown in Figure 5.8(c). The algorithm successfully recovered the hidden nodes, but it added one spurious edge. As a result the recovered structure is not a polytree. This could be predicted by investigating the estimated discrepancy matrix in (5.12); since entries {(In,Or),(Go,Or),(Ms,Or),(Go,Xr),(Or,Go),(Or,Ib),(Xr,Go),(Xr,In),(Xr,Ms)} are positive when they should have been -1 due to the fact that these pairs have no common ancestor in Figure 5.8(b). The reason for this is maybe due to estimation error resulting from insufficiency of the number of samples or the fact that the true underlying graph is not a polytree.

$$\Gamma_V = \begin{array}{c} \\ Em \\ Go \\ In \\ Ms \\ Ib \\ Or \\ Xr \end{array} \begin{array}{c} \begin{array}{ccccccc} Em & Go & In & Ms & Ib & Or & Xr \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 0 & 2 & 1 & 3 & 1 & 1 \\ 2 & 1 & 0 & 1 & 2 & 1 & -1 \\ 2 & 0 & 1 & 0 & 2 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 \\ 2 & 1 & -1 & -1 & 1 & 0 & 1 \\ 2 & 1 & 1 & 1 & -1 & 1 & 0 \end{pmatrix} \end{array}. \tag{5.12}$$

[4]Gartner, http://www.gartner.com/newsroom/id/856712

[5]http://www.computerworld.com/article/2529379/microsoft-windows/windows-market-share-dives-below-90–for-first-time.html

CHAPTER 6

LATENT RECOVERY IN VAR MODELS

This chapter studies the dependency graph of vector Auto Regressive (VAR) models from samples when a subset of the variables are latent. More precisely, we assume that the available measurements are a set of random processes $\underline{X}_t \in \mathbb{R}^n$ which, together with another set of latent random processes $\underline{Z}_t \in \mathbb{R}^m$, where $m \leq n$ form a first order VAR model as follows:

$$\begin{bmatrix} \underline{X}_{t+1} \\ \underline{Z}_{t+1} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \underline{X}_t \\ \underline{Z}_t \end{bmatrix} + \begin{bmatrix} \underline{\omega}_{X,t+1} \\ \underline{\omega}_{Z,t+1} \end{bmatrix}. \tag{6.1}$$

As we showed in Section 3.2, in VAR models, the support of the coefficient matrix encodes the causal structure in a VAR model. We propose a learning approach that recovers the observed sub-network (support of A_{11}) from linear regression on the observed variables $\underline{\mathbf{X}}$ as long as the *latent sub-network* (support of A_{22}) is a DAG. We also derive a set of sufficient conditions under which we can uniquely recover the causal influences from latent to observed processes, (support of A_{12}) and also the causal influences among the latent variables, (support of A_{22}). Additionally, we propose a sufficient condition under which the complete causal structure can be recovered uniquely.

6.1 Problem Setting

Consider the VAR model in (6.1). Let $\underline{\omega}_{Z,t} \in \mathbb{R}^m$ be i.i.d random vectors with mean zero. For simplicity, we denote the matrix $[A_{11}, A_{12}; A_{21}, A_{22}]$ by A. Our goal is to recover $Supp(A)$ from observed data, i.e., $\{\underline{X}_t\}$. Rewrite 6.1 as follows

$$\underline{X}(t+1) = \sum_{k=0}^{t} A_k^* \underline{X}_{t-k} + A_{12} A_{22}^t \underline{Z}_0 + \sum_{k=0}^{t-1} \tilde{A}_k \underline{\omega}_{Z,t-k} + \underline{\omega}_{X,t+1},$$

where $A_0^* := A_{11}$, $A_k^* := A_{12} A_{22}^{k-1} A_{21}$ for $k \geq 1$, and $\tilde{A}_k := A_{12} A_{22}^k$. In the remainder, we will assume that the A_{22} is acyclic, i.e., $\exists\ 0 < l \leq m$, such that $A_{22}^l = 0$. Thus, for $t \geq l$, the above equation becomes

$$\underline{X}_{t+1} = \sum_{k=0}^{l} A_k^* \underline{X}_{t-k} + \sum_{k=0}^{l-1} \tilde{A}_k \underline{\omega}_{Z,t-k} + \underline{\omega}_{X,t+1}. \tag{6.2}$$

Note that the limits of summations in (6.2) are changed.

We are interested in recovering the set $\{Supp(A_k^*)\}_{k=0}^l$ because it captures important information about the structure of the VAR model. Specifically, $Supp(A_0^*) = Supp(A_{11})$; so it represents the direct causal influences between the observed variables and $Supp(A_k^*)$ for $k \geq 1$ determines whether at least one directed

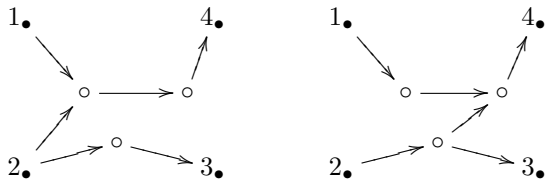


Figure 6.1. Two unobserved networks with the same linear measurements. Observed and latent nodes are depicted by black and white circles, respectively.

path of length $k+1$ exists between any two observed nodes which goes through the latent sub-network[1]. We will make use of this information in our recovery algorithm. We call the set of matrices $\{Supp(A_k^*)\}_{k\geq 0}$, *linear measurements*. In Section 6.3, we present a set of sufficient conditions under which given the linear measurements, we can recover the entire or most parts of the unobserved network uniquely.

Note that in general, the linear measurements cannot uniquely specify the unobserved network. For example, Figure 6.1 illustrates two different unobserved networks that both share the same set of linear measurements,

$$A_1^* = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_2^* = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix},$$

and $A_k^* = 0$ for $k > 2$.

6.2 Identifiability of the Linear Measurements

As we need the linear measurements for our structure learning, in this section, we study the conditions required for recovering the linear measurements from the observed processes $\{\underline{X}_t\}$. To do so, we start off by rewriting Equation (6.2) as follows

$$\underline{X}_{t+1} = \mathcal{A}\underline{\mathcal{X}}_{t-l:t} + \sum_{k=0}^{l-1} \tilde{A}_k \underline{\omega}_{Z,t-k} + \underline{\omega}_{X,t+1}, \tag{6.3}$$

where $\mathcal{A} := [A_0^*, ..., A_l^*]_{n\times n(l+1)}$, and $\underline{\mathcal{X}}_{t-l:t} := [\underline{X}_t; \cdots; \underline{X}_{t-l}]_{n(l+1)\times 1}$.

By projecting $\tilde{A}_k\underline{\omega}_Z(t-k)$ onto the vector space spanned by the observed processes, i.e., $\{\underline{X}(t), ..., \underline{X}(t-l)\}$, we obtain

$$\tilde{A}_k\underline{\omega}_{Z,t-k} = \sum_{r=0}^{l} C_r^s \underline{X}_{t-r} + \underline{N}_{Z,t-k}, \quad 0 \leq k \leq l-1, \tag{6.4}$$

where $\{\underline{N}_{Z,t-k}\}$ denote the residual terms and $\{C_r^s\}$ are the corresponding coefficient matrices. Substituting (6.4) into (6.3) implies

$$\underline{X}_{t+1} = \mathcal{B}\underline{\mathcal{X}}_{t-l:t} + \underline{\theta}_{t+1}, \tag{6.5}$$

[1]Herein, we exclude degenerate cases where there is a direct path from an observed node to another one with length k but the corresponding entry in matrix $Supp(A_k^*)$ is zero. In fact, such special cases can be resolved by small perturbation of nonzero entries in matrix A.

where $\mathcal{B} := [B_0^*, ..., B_l^*]$, and

$$B_k^* := A_k^* + \sum_{s=0}^{l-1} C_k^s, \quad \underline{\theta}_{t+1} := \underline{\omega}_{X,t+1} + \sum_{k=0}^{l-1} \underline{N}_{Z,t-k}.$$

Note that by this representation, $\underline{\theta}_{t+1}$ is orthogonal to $\underline{\mathcal{X}}_{t-l:t}$, i.e., $\mathbb{E}[\underline{\theta}_{t+1}^T \underline{X}_{t-k}] = 0$, for $0 \leq k \leq l$. Hence, Equation (6.5) shows that the minimum mean square error (MMSE) estimator can learn the coeffiecient matrix $\mathcal{B}$ given the observed processes. More precisely, we have

$$\mathcal{B} = [\gamma_X(1), ..., \gamma_X(l+1)] \times \Gamma_X(l)^{-1}, \tag{6.6}$$

where $\Gamma_X(l) := \mathbb{E}\{\underline{\mathcal{X}}_{t-l:t} \ \underline{\mathcal{X}}_{t-l:t}^T\}$. Let us denote the Fourier transform of g by $\mathcal{F}[g]$, that is given by $\sum_{h=-\infty}^{\infty} g(h) e^{-h\Omega j}$.

Proposition 5. *For the stationary VAR model in (6.1) in which the latent sub-network is a DAG, i.e., $A_{22}^l = 0$, we have*

$$\max_{0 \leq k \leq l} ||B_k^* - A_k^*||_1 \leq \sqrt{nl \frac{M}{L}} ||A_{12}||_2,$$

where $L := \inf_{\Omega \in [0,2\pi]} \lambda_{min}(\mathcal{F}[\gamma_X])$ and $M :=\sup_{\Omega \in [0,2\pi]} \lambda_{max}(\mathcal{F}[\gamma_{\omega_Z}])$.

Proof. See Appendix A.5.1. □

This result implies that we can asymptotically recover the support of $\{A_k^*\}_{k=0}^l$ as long as the absolute values of non-zero entries of $\{A_k^*\}_{k=0}^l$ are bounded away from zero by $2\sqrt{nl\frac{M}{L}}||A_{12}||_2$. Note that the direct causal influences among the observed nodes (support of A_{11}) can be recovered from A_0^*. We will make use of $\{Supp(A_k^*)\}_{k>0}$ to recover the unobserved network in the next section.

Proposition 6. *Let Σ_X and Σ_Z be the autocovariance matrices of $\underline{\omega}_{X,t}$ and $\underline{\omega}_{Z,t}$, respectively. Then, the ratio M/L strictly increases by decreasing σ_X^2/σ_Z^2 where $\Sigma_X = \sigma_X^2 I_{n\times n}$ and $\Sigma_Z = \sigma_Z^2 I_{m\times m}$.*

Proof. See Appendix A.5.2. □

When only a finite number of samples from the observed processes are available, say $\{\underline{X}_t\}_{t=1}^T$, we can estimate the coefficient matrix $\mathcal{B}$, using an empirical estimator for $\Gamma_X(l)$, $\{\gamma_X(h)\}$, and then applying (6.6). Denote the result of this estimation by $\mathcal{B}_T$. It can be shown that [143]

$$\sqrt{T}\text{vec}(\mathcal{B}_T - \mathcal{B}) \xrightarrow[T\to\infty]{d} \mathcal{N}(0, \Gamma_X^{-1}(l) \otimes \Sigma).$$

where $\xrightarrow{d}$ denotes convergence in distribution. Matrix Σ is given by $(\Sigma_X + \sum_{k=0}^{l-1}(\tilde{A}_k \Sigma_Z \tilde{A}_k^T))$. The vec$(\cdot)$ operator transforms a matrix to a vector by stacking its columns and $\otimes$ is the Kronecker product.

6.3 Learning the Unobserved Network

Recall that we refer to $Supp([0, A_{12}; A_{21}, A_{22}])$ as the unobserved network and $Supp(A_{22})$ as the latent sub-network. We present three algorithms that take linear measurements $\{Supp(A_k^*)\}_{k\geq 0}$ as their input. First algorithm recovers the entire unobserved network uniquely as log as it is a directed tree and each latent node has at least two parents and two children. The output of the second algorithm is $Supp([0, A_{12}; \widehat{A}_{21}, A_{22}])$,

where $Supp(A_{21}) \subseteq Supp(\widehat{A}_{21})$. This means that $[A_{21}]_{ij} = 0$ whenever $[\widehat{A}_{21}]_{ij} = 0$. This output is guaranteed whenever the latent sub-network is a directed tree and some extra conditions are satisfied on how the latent and observed variables are connected (see Assumption 4 in Section 6.3.2). Third algorithm finds the set of all possible networks that are consistent with the measurements and have the minimum number of latent nodes. This algorithm is able to do so when there exists at most one directed "latent path" of any arbitrarily length between two observed nodes (see Assumption 5 in Section 6.3.3). A directed path is called latent if all the intermediate variables on that path are latent.

6.3.1 Unobserved Network is a Directed Tree

The work in [144] introduced a necessary and sufficient condition and also an algorithm to recover a weighted directed tree uniquely[2] from a valid distance matrix D defined on the observed nodes. The condition is as follows: every latent node must have at least two parents and two children. A matrix D, in [144], is a valid distance matrix over a weighted directed tree, when $[D]_{ij}$ equals the sum of all the weights of those edges that belong to the directed path from i to j, and $[D]_{ij} = 0$, if there is no directed path from i to j.

The algorithm in [144] has two phases. In the first phase, it creates a directed graph among the observed nodes with the adjacency matrix $Supp(D)$. In the second phase, it recursively finds and removes the circuits[3] by introducing latent nodes for each circuit. For more details see [144].

In order to adopt [144]'s algorithm for learning the unobserved network, we introduce a valid distance matrix using our linear measurements as follows,

$$[D]_{ij} = \begin{cases} k+1 & [Supp(A_k^*)]_{ji} \neq 0, \\ 0 & \text{Otherwise.} \end{cases}$$

Recall that $[Supp(A_k^*)]_{ji}$ indicates whether there exists a directed latent path from i to j of length $k+1$ in the unobserved network. From theorem 8 in [144], it is easy to show that the unobserved network can be recovered uniquely from above distance matrix if its topology is a directed tree.

6.3.2 Latent Sub-network is a Directed Tree

We need the following definition to present our results.

Definition 16. *We denote the subset of observed nodes that are parents of a latent node h by $\mathcal{P}_h^O$ and denote the subset of observed nodes that h is their parent, by $\mathcal{C}_h^O$. We further denote the set of all leaves in the latent sub-network by $\mathcal{L}$.*

We consider learning an unobserved network G that satisfies the following assumptions.

Assumption 4. *Assume that the latent sub-network of G is a directed tree. Furthermore, for any latent node h in G; (i) $\mathcal{P}_h^O \not\subseteq \cup_{h \neq j} \mathcal{P}_j^O$ and if h is a leaf of the latent sub-network, then (ii) $\mathcal{C}_h^O \not\subseteq \cup_{i \in \mathcal{L}, i \neq h} \mathcal{C}_i^O$.*

This assumption states that the latent sub-network of G must be a directed tree such that each latent node in G has at least one unique parent in the set of observed nodes. That is, a parent who is not shared with any other latent node. Furthermore, each latent leaf has at least one unique child among the observed

[2]The skeleton of the recovered tree is the same as the original one but not necessary the weights.
[3]In a directed graph, a circuit is a cycle after removing all the directions.

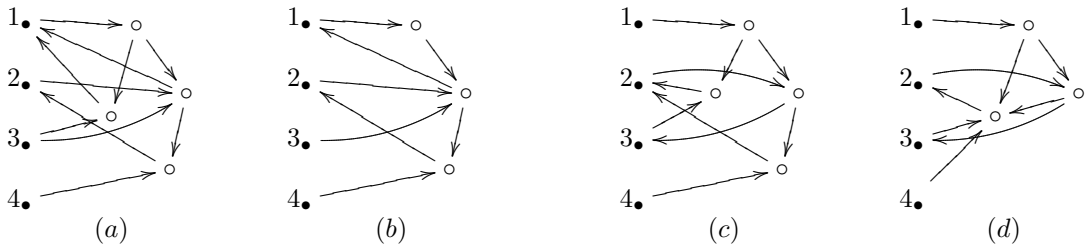


Figure 6.2. Observed and latent nodes are indicated by black and white circles, respectively. Graph (a) satisfies (ii) but not (i) and it can be reduced to (b). Graph (c) satisfies (i) but not (ii) and it can be reduced to (d).

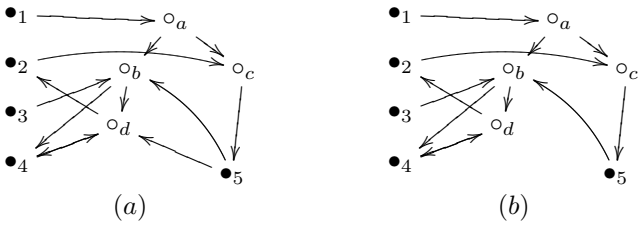


Figure 6.3. Both graphs satisfying Assumption 4 and have the same induced linear measurements but $Supp(A_{21})_{(b)} \subset Supp(A_{21})_{(a)}$.

nodes. For instance, when $Supp(A_{22})$ represents a directed tree and both $Supp(A_{12})$ and $Supp(A_{21})$ contain identity matrices, Assumption 4 holds.

Figure 6.3(a) illustrates a simple network that satisfies Assumption 4 in which the unique parents of latent nodes a, b, c, and d are $\{1\}$, $\{3\}$, $\{2\}$, and $\{4\}$, respectively. The unique children of latent leaves c and d are $\{5\}$ and $\{2, 4\}$, respectively.

Theorem 11. *Among all unobserved networks that are consistent with the linear measurements induced from (6.1), graph G that satisfies Assumption 4 has the minimum number of latent nodes.*

Proof. See Appendix A.5.3. □

Note that if Assumption 4 is violated, one can find many unobserved networks that are consistent with the linear measurements but are not minimum (in terms of the number of latent nodes). For example, the network in Figure 6.2(a) satisfies Assumption 4 (ii) but not (i). Figure 6.2(b) depicts an alternate network with the same linear measurements as the network in Figure 6.2(a) but it has fewer number of latent nodes. Similarly, the graph in Figure 6.2(c) satisfies Assumption 4 (i) but not (ii). Figure 6.2(d) shows an alternate graph with one less latent node.

Theorem 12. *Consider an unobserved network G with adjacency matrix $Supp([0, A_{12}; A_{21}, A_{22}])$. If G satisfies Assumption 4, then its corresponding linear measurements uniquely identify G upto $Supp([0, A_{12}; \widehat{A}_{21}, A_{22}])$, where $Supp(A_{21}) \subseteq Supp(\widehat{A}_{21})$.*

Proof. See Appendix A.5.4. □

Figure 6.3(a) gives an example of a network satisfying Assumption 4 and an alternate network, Figure 6.3(b), with the same linear measurements which departs from the Figure 6.3(a) in A_{21} component.

Next, we propose the directed tree recovery (DTR) algorithm that takes the linear measurements of an unobserved network G satisfying Assumption 4 and recovers G upto the limitation in Theorem 12. This

Algorithm 5 The DTR Algorithm

1: **Input:** $\{Supp(A_k^*)\}_{k\geq 1}$
2: Find $\{l_i\}$ using (6.7) and set $U := \emptyset$.
3: **for** $i = 1, ..., n$ **do**
4: Find R_i, M_i from (6.8) and (6.9)
5: $Y_i := \{j : j \neq i \wedge l_j = l_i\}$
6: **if** $\forall j \in Y_i$, $(R_j \not\subseteq R_i) \vee (R_j = R_i \ \wedge M_i \subseteq M_j)$ **then**
7: Create node h_i and set $\mathcal{P}_{h_i} = \{i\}$, $U \leftarrow \{i\} \cup U$
8: **end if**
9: **end for**
10: **for** every latent node h_s **do**
11: **if** $\exists h_k, (l_k = l_s + 1) \wedge (R_s \subseteq R_k)$ **then**
12: $\mathcal{P}_{h_s} \leftarrow \{h_k\} \cup \mathcal{P}_{h_s}$
13: **end if**
14: $\mathcal{C}_{h_s} \leftarrow \{j : [A_1^*]_{js} \neq 0\}$
15: **end for**
16: **for** $i = 1, ..., n$ **do**
17: **if** $\exists\ j \in U$, s.t. $M_j \subseteq M_i$ **then**
18: $\mathcal{P}_{h_j} \leftarrow \{i\} \cup \mathcal{P}_{h_j}$
19: **end if**
20: **end for**

algorithm consists of three main loops. Recall that Assumption 4 implies that each latent node has at least one unique observed parent. The first loop finds all the unique observed parents for each latent node (lines: 3-9). The second loop reconstructs $Supp(A_{22})$ and $Supp(A_{12})$ (lines: 10-15). And finally, the third loop constructs $Supp(\widehat{A}_{21})$ such that $Supp(A_{21}) \subseteq Supp(\widehat{A}_{21})$ (lines: 16-20).

The following lemma shows that the first loop of Algorithm 5 can find all the unique observed parents from each latent node. To present the lemma, we need the following definitions.

Definition 17. *For a given observed node i, we define*

$$l_i := \max\{k : [A_{k-1}^*]_{si} \neq 0, \ \textit{for some } s\}, \tag{6.7}$$

$$R_i := \{j : [A_{l_i-1}^*]_{ji} \neq 0\}, \tag{6.8}$$

$$M_i := \{(j, r) : [A_{r-1}^*]_{ji} \neq 0\}. \tag{6.9}$$

In the above equations, l_i denotes the length of longest directed latent path that connects node i to any other observed node. R_i is the set of all observed nodes that can be reached by i with a directed latent path of length l_i and set M_i consists of all pairs (j, r) such that there exists a directed latent path from i to j with length r.

Lemma 8. *Under Assumption 4, an observed node i is the unique parent of a latent node if and only if for any other observed node j s.t. $l_i = l_j$, we have*

$$(R_j \not\subseteq R_i) \vee (R_j = R_i \ \wedge M_i \subseteq M_j).$$

Proof. See Appendix A.5.5. □

The second loop recovers $Supp(A_{22})$ based on the following observation. If a latent node h_k is the parent of latent node h_s, then h_k can reach all the observed nodes in R_s, i.e, $R_s \subseteq R_k$ and $l_k = l_s + 1$ (line: 11).

Furthermore, $Supp(A_{12})$ can be recovered using the fact that an observed node j is a children of a latent node h_s, if a unique parent of h_s, e.g., s can reach j by a directed latent path of length 2 (line: 14). Finally, the third loop reconstructs $Supp(\widehat{A}_{21})$ by adding an observed node i to the parent set of latent node h_j, if i can reach all the observed nodes that a unique parent of h_j, e.g., j reaches (lines: 17-18).

Proposition 7. *Suppose network G satisfies Assumption 4. Then given its corresponding linear measurements, Algorithm 5 recovers G upto the limitation in Theorem 12.*

Proof. See Appendix A.5.6. □

6.3.3 Learning More General Unobserved Networks with Minimum Number of Latent Nodes

In general, there may not be a unique minimal unobserved network consistent with the linear measurements (see Fig. 6.1). Hence, we try to find an efficient approach for recovering all possible minimal unobserved networks under some conditions. In fact, without any extra conditions, finding a minimal unobserved network is NP-hard.

Theorem 13. *Finding an unobserved network that is both consistent with a given linear measurements and has minimum number of latent nodes is NP-hard.*

Proof. See Appendix A.5.7. □

In the remainder of this section, after some definitions, we propose the Node-Merging (NM) algorithm. This algorithm returns all possible unobserved networks with minimum number of latent nodes that are consistent with the linear measurements if we consider the following assumption.

Assumption 5. *Assume that there exists at most one directed latent path of each length between any two observed nodes.*

For example, the graph in Figure 6.3-right satisfies this assumption but not the one in Figure 6.3-left. This is because there are two directed latent paths of length 2 from node 5 to node 4.

Definition 18. *(Merging) We define merging two nodes i' and j' in graph G as follows: remove node j' and the edges between i' and j', then give all the parents and children of $j'$to i'. We denote the resulting graph after merging i' and j' by Merge(G, i', j'). We say that two nodes i' and j' are mergeable if Merge(G, i', j') is consistent with the linear measurements of G.*

Definition 19. *(Contentedness) Consider an undirected graph $\bar{G}$ over the observed nodes which is constructed as follows: there is an edge between two nodes i and j in $\bar{G}$, if there exists $k \geq 1$ s.t. $Supp([A_k^*]_{ij}) = 1$ or $Supp([A_k^*]_{ji}) = 1$; We say that two observed nodes i and j are "connected" if there exist a path between them in $\bar{G}$.*

It can be seen that if pairs i, j and j, k are connected then node i and k are also connected. Thus, we can define a *connected class*. That is, a subset of observed nodes in which any two nodes are connected.

The Node-Merging algorithm has two phases: initialization and merger.

Initialization: We first find the set of all connected classes, say $S_1, S_2, ..., S_C$. For each class S_c, we create a directed graph $G_{0,c}$ that is consistent with the linear measurements. To do so, for any two observed

Algorithm 6 The Node-Merging (NM) Algorithm

Initialization: Construct graph G_0.
$\mathcal{G}_0 := G_0$, $\mathcal{G}_s := \emptyset, \forall s > 0$
$k := 0$
while $\mathcal{G}_k \neq \emptyset$ **do**
  for $G \in \mathcal{G}_k$ **do**
    for $i', j' \in G$ **do**
      if Check(G, i', j') **then**
        $\mathcal{G}_{k+1} := \mathcal{G}_{k+1} \cup \text{Merge}(G, i', j')$.
      end if
    end for
  end for
  $k := k + 1$
end while
Output: $\mathcal{G}_{out} := \mathcal{G}_{k-1}$

nodes $i, j \in S_c$, if $[A_r^*]_{ji} \neq 0$, we construct a directed path with length $r+1$ from node i to node j by adding r new latent nodes to $G_{0,c}$.

Merger: In this phase, for any $G_{0,c}$ from the initialization phase, we merge its latent nodes iteratively until no further latent pairs can be merged. Since order of mergers leads to different networks with minimum number of latent nodes, the output of this phase will be the set of all such networks. Algorithm 6 summarizes the steps of the NM algorithm. In this algorithm, subroutine Check(G, i', j') checks whether two nodes i' and j' are mergeable.

Theorem 14. *Under Assumption 5, the NM algorithm returns the set of all networks that are consistent with the linear measurements and have minimum number of latent nodes.*

Proof. See Appendix A.5.8. □

6.4 Experimental Results

Synthetic Data:

We considered a directed random graph denoted by DRG(p, q), such that there exists a directed link from an observed node to a latent node and vice versa independently, with probability p. Furthermore, there is a directed link from a latent node to any other latent node with probability q. If there is a link between two nodes, we set the weight of that link uniformly from $\{-a, a\}$.

In order to evaluate how well we can estimate the linear measurements, we generated 1000 instances of DRG$(0.4, 0.4)$ with $n+m = 100$, $\mathbb{E}\{[\omega_X(t)]_i^2\} = \mathbb{E}\{[\omega_Z(t)]_i^2\} = 0.1$, and $a = 0.1$. The length of time series was set to $T = 1000$. We considered two cases for estimating A_{11} using linear regression in (6.5) with lag length $l = 1$ and $l = 3$. Let $\hat{A}_{11}$ be the output of linear regression. We computed $Supp(\hat{A}_{11})$ by setting entry (i, j) to one if $|[\hat{A}_{11}]_{ij}| > a/2$. In Figure 6.4-left, the expected estimation error, i.e. $||Supp(\hat{A}_{11}) - Supp(A_{11})||_F^2/n^2$, is computed where $||.||_F$ is the Frobenius norm. As it can be seen, the estimation error decrease as we increase the lag length.

We also studied the effect of observed to latent noise power ratio (OLNR), $\mathbb{E}\{[\omega_X(t)]_i^2\}/\mathbb{E}\{[\omega_Z(t)]_i^2\}$, in estimating the linear measurements. We generated 1000 instances of DRG$(0.1, 0.1)$ with $n = 10$, $m = 5$,

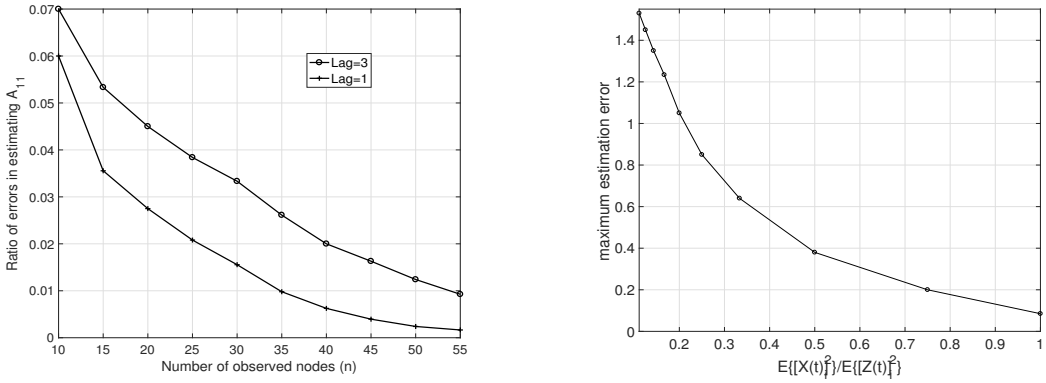


Figure 6.4. Average error in computing linear measurements. Left: The average normalized error versus number of observed nodes. Right: The average of maximum estimation error versus OLNR.

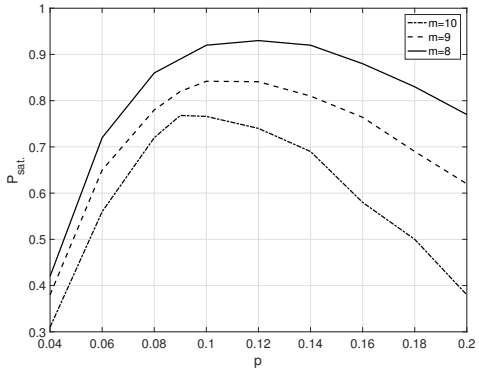


Figure 6.5. The probability $P_{sat.}$ versus the parameter p.

and $a = 0.5$. Figure 6.4-right illustrates $\max_k ||Supp(\hat{A}_k^*) - Supp(A_k^*)||_F^2$, as a function of OLNR. As it can be seen, the average of maximum estimation error decreases as OLNR increases which is expected from Proposition 6.

We investigated what percentage of instances of random graphs satisfy Assumption 4. We generated 1000 instances of DRG$(p, 1/n)$ with $n = 100$, and $p \in [0.04, 0.2]$. In Figure 6.5, the probability of satisfying Assumption 4, $P_{sat.}$, is depicted versus p for different number of latent variables in the VAR model. As it can be seen, for large value of m, the probability $P_{sat.}$ decreases. This is because it becomes less likely to see a unique observed parent for each latent node. For a fixed number of latent nodes, the same event will occur if we increase p. Furthermore, for small p, there might exist some latent nodes that have no observed parent or no observed children.

We also evaluated the performance of the NM algorithm in random graphs. We generated 1000 instances of DRG$(1/2n, 1/2n)$ with $n = 10, 20, ..., 100$, $m = n/2$, and computed the linear measurements. If for a class of connected nodes, the number of latent nodes generated in the initial phase exceeds 40, we assumed that the corresponding instance cannot be recovered efficiently in time and did not proceed to the merging phase. In Figure 6.6-left, we depicted the percentage of instances in which the algorithm can recover all possible minimal unobserved networks. As it is shown, large portion of instances (at least 96.9%) can be recovered

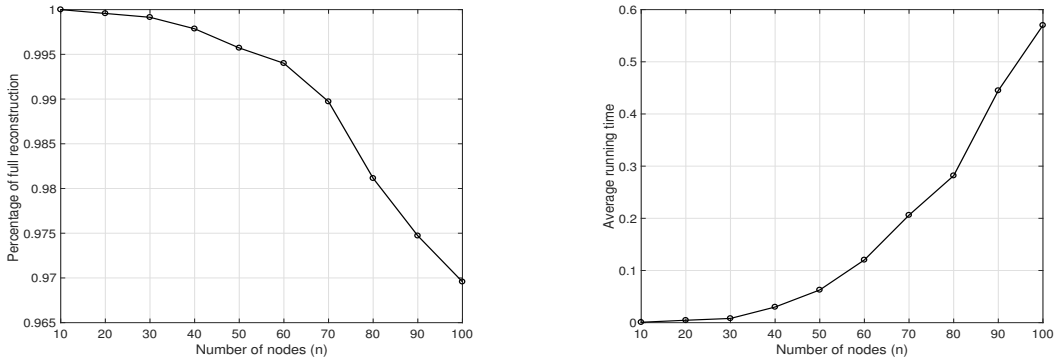


Figure 6.6. Recovering the minimal unobserved network: Results are averaged over 1000 instances of DRG(p, q) where $n = 10, 20, \cdots, 100$, $m = n/2$, and $p = q = 1/(2n)$. Left: The percentage of instances that can be reconstructed efficiently in time. Right: Average run time of the algorithm.

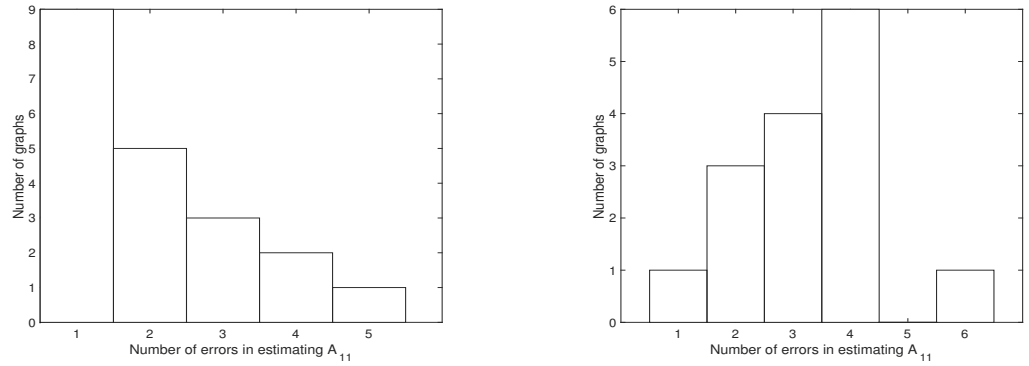


Figure 6.7. The histogram of $||Supp(\hat{A}_{11}) - Supp(A_{11})||_F^2$ for high power and low power conditions.

even for the case $n = 100$. In Figure 6.6-right, the average run time of the algorithm is depicted[4]. This plot shows that we can recover all possible minimal unobserved networks for a large portion of instances efficiently even in relatively large networks. This observation is not surprising since we know that the size of each connected class nodes is of order $\log(n)$ in sparse random graphs [145].

US Macroeconomic Data:

We considered the following set of time series from the quarterly US macroeconomic data for the period from 31-Mar-1947 to 31-Mar-2009 collected from the St. Louis Federal Reserve Economic Database (FRED) (`http://research.stlouisfed.org/fred2/`): gross domestic product (GDP), gross domestic product price deflator (GDPDEF), paid compensation of employees (COE), non-farm business sector index of hours worked (HOANBS), three-month treasury bill yield (TB3MS), personal consumption expenditures (PCEC), and gross private domestic investment (GPDI).

We selected any four times series as observed processes and computed $Supp(\hat{A}_{11})$ with lag length $l = 3$. We divided the $\binom{7}{4} = 35$ possible selections into two classes: 1) High power: $\text{tr}(\mathbb{E}\{\omega_X(t)\omega_X(t)^T\}) > \tau$ for a fixed threshold τ. 2) Low power: $\text{tr}(\mathbb{E}\{\omega_X(t)\omega_X(t)^T\}) < \tau$. In this experiment, we set $\tau = 0.02$. In Figure

[4]This experiment was performed on a on a Mac Pro with 2×2.4 GHz 6-Core Intel Xeon processor and 32 GB of RAM.

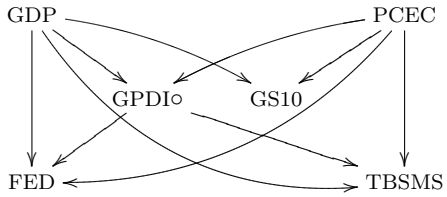


Figure 6.8. The causal structure in US macroeconomic data.

6.7, we plotted the histograms of $||Supp(\hat{A}_{11}) - Supp(A_{11})||_F^2$ for these two classes. As it can be seen, in the high power regime, most of the possible selections have small estimation error.

We also considered the following six time series of US macroeconomic data during 1-Jun-2009 to 31-Dec-2016 from the same database: GDP, GPDI, PCEC, TBSMS, effective federal funds rate (FEDFUND), and ten-year treasury bond yield (GS10). We obtained the causal structure among these six time series using a linear regression with lag length $l = 1$ and considered the result as our ground truth (see Figure 6.8). Then, we removed GPDI from the dataset and considered the remaining five time series as observe processes. We performed a linear regression with lag length $l = 2$ to obtain the linear measurements and detected non-zero entries of linear measurements by considering a threshold of 2.2. Algorithm 5 recovered the ground truth in Figure 6.8 correctly.

Dairy Prices and West German Macroeconomic Data:

A collection of three US dairy prices has been observed monthly from January 1986 to December 2016 (`http://future.aae.wisc.edu/tab/prices.html`): milk price, butter price, and cheese price. We performed a linear regression with lag length $l = 1$ on the whole time series and considered the resulting graph as our ground truth (see Figure 6.9-left). We used 0.25 as the threshold to detect the non-zero entries of the coefficient matrix. Next, we omitted the butter prices from the dataset and considered the milk price and cheese prices as observed processes. We performed the linear regression with lag length $l = 2$ and detected the nonzero entries with a threshold of 0.15. The linear measurements were: $Supp(A_0^*) = Supp(A_{11}) = [1, 1; 1, 0]$ and $Supp(A_1^*) = [0, 0; 1, 0]$. Algorithm 5 recovered correctly the true causal graph using this linear measurements.

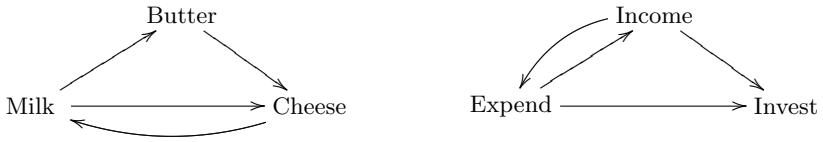


Figure 6.9. The true causal structure. Left: US Dairy prices. Right: West German macroeconomic data.

We also considered the quarterly West German consumption expenditures X_1, fixed investment X_2, and disposable income X_3 during 1960-1982 (`http://www.jmulti.de/data_imtsa.html`). Similar to the previous experiment with dairy prices, we found entire causal structure among $\{X_1, X_2, X_3\}$ using a threshold of 0.2. Figure 6.9 depicts the resulting graph. Next, we considered X_3 to be latent and used $\{X_1, X_2\}$ to estimate the linear measurements $Supp(A_0^*) = Supp(A_{11}) = [0, 0; 1, 1]$ and $Supp(A_1^*) = [1, 0; 1, 0]$, where the threshold for detecting nonzero entries was set to 0.1. Using this linear measurements, Algorithm 5 recovered correctly the true network in Figure 6.9-right.

CHAPTER 7

A DEPENDENCY MEASURE BASED ON WASSERSTEIN DISTANCE

By studying the limitations of the existing dependencies measures such as their shortcomings in detecting direct influences or their lack of ability for group selection in order to have effective interventions, we introduce a new dependency measure to overcome them. More precisely, we define a new measure that is capable of capturing dependencies that occur rarely or even over a zero measure set. On contrary, this is not possible via other measures such as mutual information that are limited to those realizations with positive probability.

Despite other measures such as conditional mutual information, our measure can encode the direct influence between two variables in a network independent of the other indirect influences between them. As a result, the direct influence between two variables can still be detected using this measure even when some variables in the indirect causal path depend on the cause almost deterministically.

This new measure has computational advantageous over other similar measures such as mutual information and information flow. Furthermore, it allows identifying the range of covariates in which the causal influence is obvious, or to find the group of subjects on which the treatment is most effective. In other words, we can determine the range for a common cause of two variables in which the influence between these two variables is maximized or minimized.

7.1 Defination

Pearl in [23] proposes that the influence of a variable (potential cause) on another variable (effect) in a network is assessed by assigning different values to the potential cause, while other variables' effects are removed, and observing the behavior of the effect variable. This can be done by intervention or "do-operation". This proposal defines a paradigm that can be used to identify the dependency or influence between the variables of a network. That is the conditional distribution of a variable given all its direct causes will not change by assigning different values to other variables in the system. Herein, we use this paradigm to define a new dependency measure.

Consider $\underline{X}$ a collection of m "random variables". In order to identify the dependency of X_i on X_j, we select a set of indices $\mathcal{K}$, where $\mathcal{K} \subseteq -\{i,j\}$ and consider the following two probability measures:

$$\begin{aligned} \mu_i(\underline{x}_{\mathcal{K}\cup\{j\}}) &:= P\Big(X_i \Big| \underline{X}_{\mathcal{K}\cup\{j\}} = \underline{x}_{\mathcal{K}\cup\{j\}}\Big), \\ \mu_i(\underline{y}_{\mathcal{K}\cup\{j\}}) &:= P\Big(X_i \Big| \underline{X}_{\mathcal{K}\cup\{j\}} = \underline{y}_{\mathcal{K}\cup\{j\}}\Big), \end{aligned} \tag{7.1}$$

where $\underline{x}_{\mathcal{K}\cup\{j\}}$ and $\underline{y}_{\mathcal{K}\cup\{j\}} \in E^{|\mathcal{K}|+1}$ are two realizations for $\underline{X}_{\mathcal{K}\cup\{j\}}$ that are the same every where except at X_j. Further, assume $\underline{x}_{\mathcal{K}\cup\{j\}}$ at position X_j equals x and $\underline{y}_{\mathcal{K}\cup\{j\}}$ equals y $(y \neq x)$ at this position. If there exists a subset $\mathcal{K} \subseteq -\{i,j\}$ such that for all such realizations $\mu_i(\underline{x}_{\mathcal{K}\cup\{j\}})$ and $\mu_i(\underline{y}_{\mathcal{K}\cup\{j\}})$ are the same, then

we say X_i has zero dependency on X_j [146]. This is analogous to the conditional independence that states if X_j and X_i are independent given some $\underline{X}_{\mathcal{K}}$, then there is no causal influence between them. Note that using mere observational data, comparing the two conditional probabilities in (7.1) reveals the dependency between X_i and X_j. However, when interventional data is available, we can identify whether X_j causes X_i, i.e., the direction of influence.

In order to compare the two probability measure in (7.1), a metric on the space of probability measures is required. There are several metrics that can be used such as KL-divergence, total variation, etc [147]. For instance, using the KL-divergence will lead to develop CI test-based approaches [148]. In this work, we use Wasserstein distance. We will discuss the advantage of using such metric later in Sections 7.3 and 7.3.1.

Definition 20. *Let (E, d) be a metrical complete and separable space equipped with the Borel field $\mathcal{B}$, and let $\mathcal{M}$ be the space of all probability measures on $(E, \mathcal{B})$. Given $\nu_1, \nu_2 \in \mathcal{M}$, the Wasserstein metric between ν_1, ν_2 is given by*

$$W_d(\nu_1, \nu_2) := \inf_{\pi} \left(\mathbb{E}_{\pi}[d(x, y)]\right), \tag{7.2}$$

where the infimum is taken over all probability measures π on $E \times E$ such that its marginal distributions are ν_1 and ν_2, respectively.

Using the above distance, we define the dependency of X_i on X_j given $\mathcal{K} \subseteq -\{i, j\}$ as follows:

$$c_{i,j}^{\mathcal{K}} := \sup_{\substack{\underline{x}_{\mathcal{K}\cup\{j\}} = \underline{y}_{\mathcal{K}\cup\{j\}} \\ \text{off } j}} \frac{W_d\Big(\mu_i(\underline{x}_{\mathcal{K}\cup\{j\}}), \mu_i(\underline{y}_{\mathcal{K}\cup\{j\}})\Big)}{d(x, y)}. \tag{7.3}$$

The suprimum is over all realizations $\underline{x}_{\mathcal{K}\cup\{j\}}$ and $\underline{y}_{\mathcal{K}\cup\{j\}}$ that only differ at the jth variable. Moreover, we assume $\underline{x}_{\mathcal{K}\cup\{j\}}$ at jth position equals x and $\underline{y}_{\mathcal{K}\cup\{j\}}$ equals y $(y \neq x)$ at this position. When $\mathcal{K} = -\{i, j\}$, $c_{i,j}^{\mathcal{K}}$ is called Dobrushin's coefficient [1]. Similarly, we define the dependency of a set of nodes $\mathcal{B}$ on a disjoint set $\mathcal{A}$ given $\mathcal{K}$, where $\mathcal{K} \cap (\mathcal{A} \cup \mathcal{B}) = \emptyset$, as follows,

$$c_{\mathcal{B},\mathcal{A}}^{\mathcal{K}} := \sup_{\substack{\underline{x}_{\mathcal{K}\cup\mathcal{A}} = \underline{y}_{\mathcal{K}\cup\mathcal{A}} \\ \text{off } \mathcal{A}}} \frac{W_d\Big(\mu_{\mathcal{B}}(\underline{x}_{\mathcal{K}\cup\mathcal{A}}), \mu_{\mathcal{B}}(\underline{y}_{\mathcal{K}\cup\mathcal{A}})\Big)}{d(\underline{x}_{\mathcal{A}}, \underline{y}_{\mathcal{A}})}. \tag{7.4}$$

Remark 5. *An alternative way of interpreting the above measure is via an equivalent network in which all the nodes in the set $\mathcal{K} \cup \{j\}$ are injected with independent inputs that have distributions equal to their marginals, i.e., node k is injected with an independent random variable that has distribution $P(X_k)$. In this equivalent network, the dependency of i on j given $\mathcal{K}$ can be expressed by*

$$\int_E \prod_{k\in\mathcal{K}} P(X_k = x_k) P(X_j = y) P(X_j = x) \frac{W_d\Big(\mu_i(\underline{x}_{\mathcal{K}\cup\{j\}}), \mu_i(\underline{y}_{\mathcal{K}\cup\{j\}})\Big)}{d(x, y)} dx_k dx dy.$$

Clearly, this expression is bounded above by (7.3).

7.1.1 Maximum Mean Discrepancy

Using a special case of the duality theorem of Kantorovich and Rubinstein [149], we obtain an alternative approach for computing the Wasserstein metric in (7.2) as follows:

$$W_d(\nu_1, \nu_2) = \sup_{f \in \mathcal{F}_L} \left| \int_E f d\nu_1 - \int_E f d\nu_2 \right|, \tag{7.5}$$

where $\mathcal{F}_L$ is the set of all continuous functions satisfying the Lipschitz condition:

$$||f||_{\text{Lip}} := \sup_{x \neq y} |f(x) - f(y)|/d(x, y) \leq 1.$$

This representation of the Wasserstein metric is a special form of integral probability metric (IPM) [150] that has been studied extensively in probability theory [151] with applications in empirical process theory [152], transportation problem [149], etc. IPM is defined similar to (7.5) but instead of $\mathcal{F}_L$, the suprimum is taken over a class of real-valued bounded measurable functions on E.

One particular instance of IPM is maximum mean discrepancy (MMD) in which the suprimum is taken over $\mathcal{F}_{\mathcal{H}} := \{f : ||f||_{\mathcal{H}} \leq 1\}$. More precisely, MMD is defined as

$$\text{MMD}(\nu_1, \nu_2) := \sup_{f \in \mathcal{F}_{\mathcal{H}}} \left| \int_E f d\nu_1 - \int_E f d\nu_2 \right|, \tag{7.6}$$

Here, $\mathcal{H}$ represents a reproducing kernel Hilbert space (RKHS) [153] with reproducing kernel $k(\cdot, \cdot)$. MMD has been used in statistical applications such as independence testing and testing for conditional independence [154–156].

It is shown in [157] that when $\mathcal{H}$ is a universal RKHS [158], defined on the compact metric space E, then $\text{MMD}(\nu_1, \nu_2) = 0$ if and only if $\nu_1 = \nu_2$. In this case, MMD can also be used to compare the two conditional distributions in (7.1). This is because, $\text{MMD}(\mu_i(\underline{x}_{\mathcal{K} \cup \{j\}}), \mu_i(\underline{y}_{\mathcal{K} \cup \{j\}})) = 0$ implies that the two conditional distributions are the same. This allows us to define a new dependency measure which we denoted it by $\tilde{c}^{\mathcal{K}}_{i,j}$ similar to (7.3) that uses MMD instead of Wasserstein distance, i.e.,

$$\tilde{c}^{\mathcal{K}}_{i,j} := \sup_{\substack{\underline{x}_{\mathcal{K} \cup \{j\}} = \underline{y}_{\mathcal{K} \cup \{j\}} \\ \text{off } j}} \frac{\text{MMD}\Big(\mu_i(\underline{x}_{\mathcal{K} \cup \{j\}}), \mu_i(\underline{y}_{\mathcal{K} \cup \{j\}})\Big)}{d(x, y)}. \tag{7.7}$$

It is straight forward to show that this measure has similar properties as the one in (7.3). The main difference between these two measures is their estimation method that we discuss in Section 7.3.1.

7.2 Comparison With Other Dependency Measures

In this section, we study the relationship between our measure in (7.3) and other measures in the literature that are introduced to encode the dependencies between variables of a network.

7.2.1 Mutual Information

Conditional mutual information is an information theoretic measure that has been used in the literature to identify the independence structure of a network. This measure compares two probability measures $P(X_i|X_j, \underline{X}_{\mathcal{K}})$ and $P(X_i|\underline{X}_{\mathcal{K}})$ as follows,

$$I(X_i; X_j|\underline{X}_{\mathcal{K}}) := \sum_{x_i, x_j, \underline{x}_{\mathcal{K}}} P(x_i, x_j, \underline{x}_{\mathcal{K}}) \log \frac{P(x_i|x_j, \underline{x}_{\mathcal{K}})}{P(x_i|\underline{x}_{\mathcal{K}})}. \tag{7.8}$$

This measure is symmetric and hence it cannot capture the direction of influence. Moreover, it only compares the probability measures over all pairs (X_i, X_j) that have positive probability.

Example 14. *Consider a network of two variables X and Y, in which $X \sim \mathcal{N}(0,1)$ is a zero mean Gaussian variable and Y is $\mathcal{N}(0,1)$ whenever X is a rational number and $\mathcal{N}(1,2)$ otherwise. In this network, X has influence on Y but it cannot be captured using CI. This is because $I(X;Y) = 0$. On the other hand, we have $c_{y,x} > 0$ and $c_{x,y} = 0$.*

Note that any other measures in the literature that is based on conditional independence test such as the kernel-based methods in [156, 159] have the similar limitation.

7.2.2 A Better Measure for Direct Influences

Consider a network comprises of three random variables $\{X, Y, Z\}$, in which $Y = f(X, W_1)$ and $Z = g(X, Y, W_2)$, where W_1 and W_2 are independent exogenous noises. Functions f and g belong to appropriately constrained functional class that the transformations from (X, W_1) to (X, Y) and from (X, Y, W_1) to (X, Y, Z) are invertible. In other words, there exist functions ϕ and φ such that $W_1 = \phi(X, Y)$ and $W_2 = \varphi(X, Y, Z)$. Furthermore, f is an injective function in its first argument, i.e., if $f(x_1, w) = f(x_2, w)$ for some w, then $x_1 = x_2$.

In order to measure the direct influence from X to Z, one may compute the conditional mutual information between X and Z given Y, i.e., $I(X; Z|Y)$. However, this is not a good measure because as the dependency of Y on X grows, i.e., $H(Y|X) \to 0$, then $I(X; Z|Y) \to 0$. This can be seen by the definition of the conditional mutual information,

$$I(X; Z|Y) = H(Y|X) + \mathbb{E}\left[\log \frac{\sum_{x'} P_{Y|X}(y|x')P_X(x')}{P_X(x)}\right] + \mathbb{E}\left[\log \frac{P_{Y|X}(y|x)P_X(x)P_{Z|X,Y}(z|x,y)}{\sum_{x'} P_{Y|X}(y|x')P_X(x')P_{Z|X,Y}(z|x',y)}\right]. \tag{7.9}$$

As $H(Y|X)$ goes to zero, in other words, as P_{W_1} tends to a Dirac measure, i.e., $\delta_{w_0}(W_1)$ for some fixed value w_0, then by specifying the value of X, the ambiguity about the value of Y will go to zero. In this case, given $X = x$, we imply that Y will take $f(x, w_0)$ with high probability. Thus, using the injective property of f, it is straight forward to see that the right hand side of (7.9) tends to zero.

This analysis shows that $I(X; Z|Y)$ fails to capture the direct influence between X and Z when the dependence can be explained by Y, which depends on X almost in a deterministic manner. However, looking at $c^y_{z,x}$, we have

$$c^y_{z,x} = \sup_{y,x,x'} \frac{W_d\left(P_{x,y}(Z), P_{x',y}(Z)\right)}{d(x, x')},$$

where

$$P_{x,y}(Z) := P_{W_2}(\varphi(x,y,Z))|\frac{\partial g}{\partial W_2}(x,y,\varphi(x,y,Z))|^{-1}.$$

This distribution depends only on realizations of (X,Y) and it is independent of $P_{X,Y}$. Hence, changing the dependency between X and Y will not affect $c^y_{z,x}$, which makes it a better candidate to measure the direct influences between variables of a network. As an illustration, we present the following simple example.

Example 15. *Consider a network of three variables $\{X,Y,Z\}$ in which $Y = aX+W_1$ and $Z = bX+cY+W_2$ for some non-zero coefficients $\{a,b,c\}$ and exogenous noises W_1 and W_2. In this example, it is straight forward to see that*

$$I(X;Z|Y) = H(bX+W_2|aX+W_1) - H(W_2). \tag{7.10}$$

As we mentioned earlier, by reducing the variance of W_1, the first term in (7.10) tends to $H(bX+W_2|X) = H(W_2)$. Hence, the conditional mutual information goes to zero. But, using the result of Theorem 15, we have $c^y_{z,x} = |b|$, which is independent of the variance of W_1.

Theorem 15. *Consider a linear system $\overline{X} = \boldsymbol{A}\overline{X}+\overline{W}$, where $\boldsymbol{A}$ has zero diagonals and its support represents a DAG. $\overline{W}$ is a vector of m independent random variables with mean zero. Then, $c^{Pa_i\setminus\{j\}}_{i,j} = |A_{i,j}|$.*

Proof. See Appendix A.6.1. □

7.2.3 Information Flow

Another quantity that has been introduced in the literature to capture the strength of the impact of interventions is information flow [26]. This quantity is defined using Pearls do-calculus [23]. Intuitively, the intervention on X_i removes the dependencies of X_i on its parents, and thus replaces $P(X_i|\underline{X}_{Pa_i})$ with the delta function.

Below, we introduce the formal definition of information flow. Consider three disjoint subsets A, B, and $\mathcal{K}$ of V. The information flow from $\underline{X}_A$ to $\underline{X}_B$ imposing $\underline{X}_{\mathcal{K}}$ is defined by

$$I(\underline{X}_A \to \underline{X}_B|do(\underline{X}_{\mathcal{K}})) := \sum_{\underline{x}_{A\cup B\cup\mathcal{K}}} P(\underline{x}_{\mathcal{K}})P(\underline{x}_A|do(\underline{x}_{\mathcal{K}}))P(\underline{x}_B|do(\underline{x}_{A\cup\mathcal{K}}))\log\frac{P(\underline{x}_B|do(\underline{x}_{A\cup\mathcal{K}}))}{\sum_{\underline{x}'_A}P(\underline{x}'_A|do(\underline{x}_{\mathcal{K}}))P(\underline{x}_B|do(\underline{x}'_A,\underline{x}_{\mathcal{K}}))}. \tag{7.11}$$

This is defined analogous to the conditional mutual information in (7.8). But unlike the conditional mutual information, the information flow is defined for all pairs $(\underline{x}_A;\underline{x}_C)$ rather than being limited to those with positive probability. Similar measures are introduced in [27, 28] which are also based on do-calculation.

Our measure in (7.3) is more similar to the aforementioned measures than the mutual information, in the sense that it is defined for all pairs rather than being limited to those with positive probability.

However, since Wasserstein metric can be estimated using a linear programming (see Section 7.3.1), our measure has computational advantageous over the information flow or other similar causal measures that uses KL-divergence. Another advantage of (7.3) over the information flow is that it requires less number of interventions. More precisely, calculating (7.11) requires at least two do-operations that are $P(\underline{x}_B|do(\underline{x}_{A\cup\mathcal{K}}))$ and $P(\underline{x}_A|do(\underline{x}_{\mathcal{K}}))$ but (7.3) requires only one such intervention. There are also some technical differences between our measure and information flow that we show one such differences through a simple example.

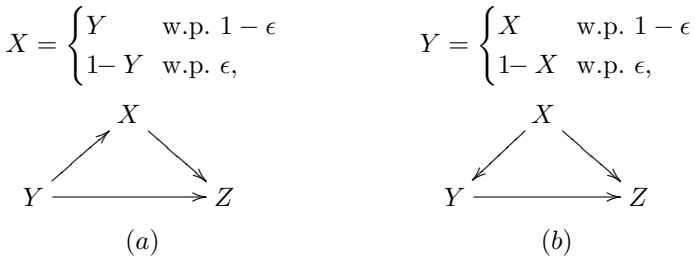


Figure 7.1. DAGs for which information flow fails to capture the influence.

Example 16. *Consider a network of three binary random variables $\{X, Y, Z\}$ with $Z = X \oplus Y$ an XOR. Suppose the underlying DAG of this network is given by Figure 7.1(b), in which X takes zero with probability b. In this case, $I(X \to Z|do(Y)) = H(B(b))$, where H denotes the entropy function and $B(b)$ denotes Bernoulli distribution with parameter b. This is because for this DAG, we have $P(X|do(Y)) = P(X)$.*

However, if the underlying DAG is given by Figure 7.1(a), we have $I(X \to Z|do(Y)) = H(B(\epsilon))$, because $P(X|do(y)) = P(X|y)$. Now, consider a scenario in which ϵ tends to zero. In this scenario, both DAGs describe a system in which $X = Y$ and $Z = X \oplus Y$. However, in the first DAG, we have $I(X \to Z|do(Y)) = H(B(b)) > 0$ while in the second DAG, we have $I(X \to Z|do(Y)) \to 0$. Hence, the information flow depends on the underlying DAG. But $c^y_{z,x}$ in both DAGs is independent of ϵ and it is positive.

7.2.4 Group Selection for Effective Intervention

Consider the network shown in Figure 7.2 in which C is a common cause for two variables X and Y. In this network, to measure the influence of X on Y, one may consider $P(Y|do(X))$ that is given by $\sum_c P(Y|X,c)P(c) = \mathbb{E}_c[P(Y|X,c)]$. See, e.g., the back-door criterion in [23]. This conditional distribution is an average over all possible realizations of the common cause C.

Consider an experiment that is been conducted on a group of people with different ages C in which the goal is to identify the effect of a treatment X on a special disease Y. Suppose that this treatment has clearer effect on that disease for elderly people and less obvious effect for younger ones. In this case, averaging the effect of the treatment on the disease for all people with different ages, i.e., $P(Y|do(X))$ might not reveal the true effect of the treatment. Hence, it is important to identify a regime (in this example age range) of C in which the influence of X on Y is maximized. As a consequence, we can identify the group of subjects on which the intervention is effective.

Note that this problem cannot be formalized using do-operation or other measures that take average over all possible realizations of C. However, using the measure in (7.3), we can formulate this problem as follows: given $X = x$ and two different realizations for C, say c and c', we obtain two conditional probabilities $P(Y|x,c)$ and $P(Y|x,c')$. Then, we say in group $C = c$, the causal influence between X and Y is more obvious compare to the group $C = c'$, if given $C = c$, changing the assignments of X leads to larger variation of the conditional probabilities compared to changing the assignment of X given $C = c'$. More precisely, if $c^{C=c}_{y,x} \geq c^{C=c'}_{y,x}$, where

$$c^{C=c}_{y,x} := \sup_{x \neq x'} \frac{W_d\Big(P(Y|x,c), P(Y|x',c)\Big)}{d(x,x')}. \tag{7.12}$$

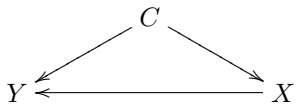


Figure 7.2. C is a common cause for X and Y.

Note that $c^c_{y,x} = \sup_c c^{C=c}_{y,x}$, where $c^c_{y,x}$ is given in (7.3). Using this new formulation, we define the range of C in which the influence from X to Y is maximized as $\arg\max_c c^{C=c}_{y,x}$.

Example 17. *Suppose that $Y = CX + W_2$ and $X = W_1/C$, where C takes value from $\{1, ..., M\}$ w.p. $\{p_1, ..., p_M\}$ and $W_1, W_2 \sim \mathcal{N}(0, 1)$. In this case, we have $c^{C=c}_{y,x} = |c|$. Thus, $C = M$ will show the influence of X on Y more clearer. On the hand, such property cannot be detected using other measures. For instance, considering the information flow (that is the same as mutual information in this example), we obtain*

$$I(X \to Y|do(C) = c) = I(X; Y|C = c) = 0.5 \log(2).$$

This is because, $(Y|X = x, C = c) \sim \mathcal{N}(cx, 1)$, $(X|C = c) \sim \mathcal{N}(0, 1/c^2)$, and $(Y|C = c) \sim \mathcal{N}(0, 2)$.

7.3 Properties of the Measure

Herein, we study the properties of our measure.

Lemma 9. *The measure defined in (7.3) possesses the following properties:*

- *Asymmetry: In general $c^{\mathcal{K}}_{i,j} \neq c^{\mathcal{K}}_{j,i}$. $c^{\mathcal{K}}_{i,j} \geq 0$ and when it is zero, we have $X_i \perp\!\!\!\perp X_j|\underline{X}_{\mathcal{K}}$.*
- *Decomposition: $c^{\mathcal{K}}_{i,\{j,k\}} = 0$ implies $c^{\mathcal{K}}_{i,j} = c^{\mathcal{K}}_{i,k} = 0$.*
- *Weak union: If $c^{\mathcal{K}}_{i,\{j,k\}} = 0$, then $c^{\mathcal{K}\cup\{k\}}_{i,j} = c^{\mathcal{K}\cup\{j\}}_{i,k} = 0$.*
- *Contraction: If $c^{\mathcal{K}}_{i,j} = c_{i,\mathcal{K}} = 0$, then $c_{i,\mathcal{K}\cup\{j\}} = 0$.*
- *Intersection: If $c^{\mathcal{K}\cup\{k\}}_{i,j} = c^{\mathcal{K}\cup\{j\}}_{i,k} = 0$, then $c^{\mathcal{K}}_{i,\{j,k\}} = 0$.*

Proof. See Appendix A.6.2. □

Note that unlike the intersection property of the conditional independence, which does not always hold, the intersection property of the dependency measure in (7.3) always holds. This is due to the fact that (7.3) is defined for all realizations $(x_j, \underline{x}_{\mathcal{K}})$ not only those with positive measure. See Example 14 for the asymmetric property of $c^{\mathcal{K}}_{i,j}$.

We say a DAG possesses global Markov property with respect to our measure if for any node i and disjoint sets $\mathcal{B}$, and $\mathcal{C}$ for which i is d-separated from $\mathcal{B}$ by $\mathcal{C}$, we have $c^{\mathcal{C}}_{i,\mathcal{B}} = c^{\mathcal{C}}_{\mathcal{B},i} = 0$.

Theorem 16. *Consider a faithful network of m random variables whose causal structure that is captured by the measure in (7.3) can be represented by a DAG. The corresponding joint distribution of this network can be factorized as in (2.1). Furthermore, its corresponding DAG possesses the global Markov property.*

Proof. See Appendix A.6.3. □

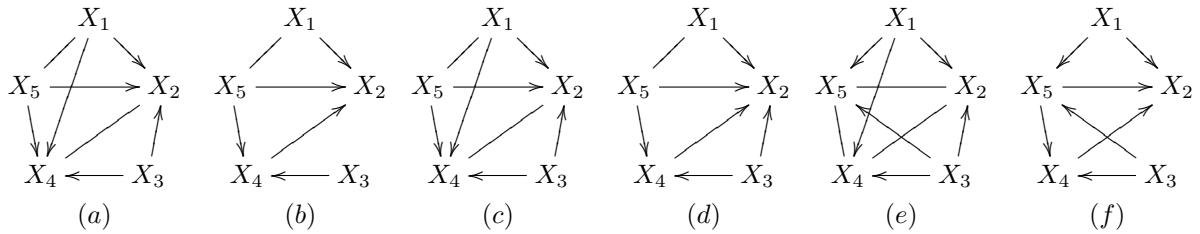


Figure 7.3. Recovered DAGs of the system given in (7.16) for different sample sizes. (a)-(b) use the measure in (7.3) and pure observation. (c)-(d) use kernel-based method and pure observation. (e)-(f) use the measure in (7.3) and interventional data.

Similar to the Bayesian networks, the global Markov property can be used to develop a reconstruction algorithm for the causal structure of a network defined using the measure in (7.3). The output of this algorithm will be a mixed graph that belongs to the Markov equivalence class of the true influence structure graph.

7.3.1 Estimation

The measure introduced in (7.3) can be computed explicitly for special probability measures. For instance, if the joint distribution of $\underline{X}$ is Gaussian with mean $\vec{\mu}$ and covariance matrix Σ, then using the results of [160] about the Wasserstein distance between two Gaussian distributions and Equation (7.5), we obtain

$$c_{i,j}^{\mathcal{K}} = |\Sigma_{i,\{j,\mathcal{K}\}}\big(\Sigma_{\{j,\mathcal{K}\},\{j,\mathcal{K}\}}\big)^{-1}\,\mathbf{e}_1|,$$

where $\Sigma_{i,\{j,\mathcal{K}\}}$ denotes the sub-matrix of Σ comprising row i and columns $\{j,\mathcal{K}\}$. In this equation, we have $\mathbf{e}_1 = (1,0,...,0)^T$. Hence, in such systems, one can estimate the dependency measure by estimating the covariance matrix. However, this is not the case in general. Therefore, we introduce a non-parametric method for estimating our dependency measure using kernel method.

Given $\{x^{(1)},...,x^{(N_1)}\}$ and $\{x^{(N_1+1)},...,x^{(N_1+N_2)}\}$ that are i.i.d. samples drawn randomly from ν_1 and ν_2, respectively, the estimator of (7.5) is given by [161],

$$\widehat{W}_d(\hat{\nu}_1,\hat{\nu}_2) := \max_{\{\alpha_i\}} \frac{1}{N_1}\sum_{i=1}^{N_1}\alpha_i - \frac{1}{N_2}\sum_{j=1}^{N_2}\alpha_{j+N_1}, \tag{7.13}$$

such that $|\alpha_i - \alpha_j| \le d(x^{(i)},x^{(j)}),\ \forall i,j$. In this equation, $\hat{\nu}_1$ and $\hat{\nu}_2$ are empirical estimator of ν_1 and ν_2, respectively.

The estimator of MMD is given by [161]

$$(\widehat{\mathrm{MMD}}(\hat{\nu}_1,\hat{\nu}_2))^2 := \sum_{i,j=1}^{N_1+N_2} y_i y_j k(x^{(i)},x^{(j)}), \tag{7.14}$$

where $y_i := 1/N_1$ for $i \le N_1$ and $y_i := -1/N_2$, elsewhere. $k(\cdot,\cdot)$ in the above equation represents the reproducing kernels of $\mathcal{H}$.

It is shown in [161] that (7.13) converges to (7.5) as $N_1, N_2 \to \infty$ almost surely as long as the underlying metric space is totally bounded. It is important to mention that the estimator in (7.13) depends on $\{x^{(j)}\}$s only through the metric $d(\cdot,\cdot)$, and thus its complexity is independent of the dimension of $x_{(i)}$, unlike the

KL-divergence estimator [162]. The estimator in (7.14) also converges to (7.7) almost surely with the rate of order $\mathcal{O}(1/\sqrt{N_1} + 1/\sqrt{N_2})$, when $k(\cdot,\cdot)$ is measurable and $\sup_{x\in E} k(x,x)$ is bounded.

Consider a network of m random variables $\underline{X}$. Given N i.i.d. realizations of $\underline{X}$, $\{\underline{z}^{(1)}, ..., \underline{z}^{(N)}\}$, where $\underline{z}^{(l)} \in E^m$, we use (7.13) and define

$$\widehat{c}_{i,j}^{\mathcal{K}} := \max_{1\leq l,k\leq N} \frac{\widehat{W}_d\Big(\hat{\mu}_i\left(\underline{z}^{(l)}_{\mathcal{K}\cup\{j\}}\right), \hat{\mu}_i\left(\underline{z}^{(k)}_{\mathcal{K}\cup\{j\}}\right)\Big)}{d(z_j^{(l)}, z_j^{(k)})}, \tag{7.15}$$

such that $\underline{z}^{(l)}_{\mathcal{K}\cup\{j\}} = \underline{z}^{(k)}_{\mathcal{K}\cup\{j\}}$ off j. Similarly, one can introduce an estimator for $\tilde{c}_{i,j}^{\mathcal{K}}$ using (7.14). By applying the result of Corollary 5 in [163], we obtain the following result.

Corollary 4. *Let (E,d) be a totally bounded metric space and a network of random variables with positive probabilities, then $\widehat{c}_{i,j}^{\mathcal{K}}$ converges to $c_{i,j}^{\mathcal{K}}$ almost surely as N goes to infinity.*

Proof. This is a direct consequence of Corollary 5 in [163] and the fact that all the influences occur with positive probability. □

7.4 Experimental Results

We simulated the following synthesized non-linear system and learned its corresponding causal structure form samples of observational and interventional data, respectively.

$$\begin{aligned} &X_1 = W_1, \quad X_2 = X_1^2 + 2X_4 - |X_5| + W_2, \\ &X_3 = W_3, \quad X_4 = X_3 - X_5 + W_4, \\ &X_5 = W_5, \text{ if } X_3 \text{ is natural}, \quad X_5 = 2\sqrt{|X_1|} + W_5, \text{ o.t}, \end{aligned} \tag{7.16}$$

where $W_i \sim U[-1,1]$.

Learning from Observational Data:

We used the estimator of MMD given in (7.14) with Gaussian kernels and estimated the dependency measures. We obtained the corresponding DAG of this network given a set of observation of size $N \in \{900, 2500\}$. Using the results on the convergence rate of the MMD estimator, we used a threshold of order $\mathcal{O}(1/\sqrt{N})$ to distinguish positive and zero measure. Figure 7.3 depicts the resulting DAGs. We also compared the performance of our measure with the kernel-based method proposed in [159]. Note that in this particular example, since the influence of X_3 on X_5 is not detectable by mere observation, the best we can learn from mere observation is the DAG presented in Figure 7.3(b). In this DAG, the direction of edge between X_5 and X_1 is not identifiable using the Meek rule.

Learning via Intervention:

We intervened at node X_3 and fixed its value to be natural number and irrational, separately and observed the outcome of the other nodes for different sample sizes. Figure 7.3 depicts the outcome of the learning algorithm that uses our measure. In this case, $X_3 \to X_5$ was identified and then the Meek rules helped to detect all the directions even the direction of $X_1 - X_5$ as it is shown in Figure 7.3(f).

CHAPTER 8

CONCLUSION AND FUTURE DIRECTIONS

8.1 Conclusion

In this dissertation, we studied the causal influences between variables in a network. We used graphical models to depict causal influences between variables in a well-defined manner. More specifically, we studied the functional and statistical dependencies in a dynamical systems and established their connection. To do so, we defined a statistical measure that is able to capture the functional dependency among processes of dynamical systems. Subsequently, using this measure, we defined a new type of graphical model, functional dependency graph that can encode functional dependencies. We showed that the statistical dependency structure of a system (captured by DIG) does not necessary reveal all the functional dependencies of that system (captured by FDG) in general.

We proposed an approach for learning causal interaction network of a specific network of point processes, mutually exciting linear Hawkes processes. We proved that for such point processes, the causal relationships implied by the excitation matrix is equivalent to a specific factorization of the joint distribution of the system called *minimal generative model*. One significance of this result is that it provides a surrogate to directed information measure for capturing causal influences for Hawkes processes. Furthermore, we provided an estimation method for learning the support of excitation matrices with exponential kernels using second-order statistics of the Hawkes processes.

We then developed an approach for structure learning of directed graphical model when only a subset of processes are observed. Specifically, we studied the scenario in which the directed information graph representing observed and unobserved processes is a directed tree with multiple roots. Learning such graphs requires both finding the number of hidden processes as well as recovering the connections among all hidden and observed nodes. We defined a discrepancy measure between nodes of a directed tree and introduced an algorithm that identifies the structure given the discrepancies between only the observed nodes. Moreover, we studied the problem of learning the dependency graph between variables of a vector autoregressive model with latent variables and showed that the entire or most of the causal structure can be identified successfully under some sufficient topological constraints.

At last, we introduced a new statistical measure to capture the dependency or causal direction between variables of a network from observational or interventional data. We discussed the advantageous of this dependency measures over other related measures in the literature.

We then showed how useful this framework can be in practice by finding the causal structure between different technology companies by analyzing their stock prices as well as influences between media sites by studying hyperlinks provided in one media site to others.

8.2 Future Directions

This thesis studied the causal influences between variables of a network in different scenarios such as linear dynamical systems, multivariate Hawkes processes, and VAR models. There are a number of avenues for extending this dissertation. In particular, latent processes, sparse networks, and Bayesian methods are important lines of future work.

In this dissertation, we developed an algorithm to recover the causal network of systems that have polytree structure. Also, the algorithm do not require any parametric model, the important step will be to extend these results beyond polytrees. However, due to the challenge of the general problem, extensions might only be feasible for specific classes of distributions or parametric models.

Suppose the causal network of a system is sparse. The proposed algorithms in this work do not incorporate such knowledge. There is a large body of work on sparse model selection, such as with L_1 regularization. For linear regression, lasso is an example of a sparsity-inducing fitting procedure using L_1 regularization [164]. For Markov networks of jointly Gaussian variables, [165, 166] and references therein use the lasso to identify sparse graphical models. An important avenue of future research will be to identify when similar methods could be adopted to identify sparse directed information graphs or sparse approximations for more general classes of distributions.

Another direction of future research is to extend the proposed algorithm for learning the excitation matrix of a multivariate Hawkes process with exponential kernels to a broader class of functions. More specifically, there are plenty of works that applied online learning methods in reproducing kernel Hilbert space to identify a set of parameters (e.g. exciting functions) by minimizing a certain loss function [153, 167, 168]. Developing similar online learning algorithms for learning the causal structure in Hawkes processes will be another direction for future research.

Notice that through this dissertation, there was an important assumption that the underlying causal structure does not change over the time of analysis. Although, this is a valid assumption for many real application, there are several situations in which the causal structure might vary by joining new processes to the dynamic, vanishing some of the processes, or changing the direction of influences. There are not many works that address this problem in the literature. New line of research will be studying such problem and developing algorithms that not only can identify the time that causal network changes but also learn the structure of the network as it varies.

CHAPTER 9

REFERENCES

[1] R. L. Dobrushin, "Prescribing a system of random variables by conditional distributions," *Theory of Probability & Its Applications*, vol. 15, no. 3, pp. 458–486, 1970.

[2] F. Eberhardt, "Causation and intervention," *Unpublished doctoral dissertation, Carnegie Mellon University*, 2007.

[3] K. Shanmugam, M. Kocaoglu, A. G. Dimakis, and S. Vishwanath, "Learning causal graphs with small interventions," in *Advances in Neural Information Processing Systems*, 2015, pp. 3195–3203.

[4] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[5] J. Etesami and N. Kiyavash, "Measuring causal relationships in dynamical systems through recovery of functional dependencies," *IEEE Transactions on Signal and Information Processing over Networks*, 2016.

[6] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Equivalence between minimal generative model graphs and directed information graphs," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 293–297.

[7] J. Etesami, K. Zhang, N. Kiyavash, and K. Singhal, "Learning network of multivariate hawkes processes: A time series approach," in *Proceedings of the 32rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

[8] S. Salehkaleybar, J. Etesami, and N. Kiyavash, "Learning latent networks in vector auto regressive models," *arXiv preprint arXiv:1702.08575*, 2017.

[9] J. Pearl, *Causality: models, reasoning and inference*. Cambridge university press, 2009.

[10] C. W. J. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series*. Princeton University Press., 1964.

[11] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[12] M. Eichler, "Graphical modelling of multivariate time series," *Probability Theory and Related Fields*, pp. 1–36, 2012.

[13] R. Dahlhaus and M. Eichler, "Causality and graphical models in time series analysis," *Oxford Statistical Science Series*, pp. 115–137, 2003.

[14] R. Dahlhaus, "Graphical interaction models for multivariate time series 1," *Metrika*, vol. 51, no. 2, pp. 157–172, 2000.

[15] M. Eichler, "Granger causality and path diagrams for multivariate time series," *Journal of Econometrics*, vol. 137, no. 2, pp. 334–353, 2007.

[16] J. Florens and M. Mouchart, "A note on noncausality," *Econometrica: Journal of the Econometric Society*, pp. 583–591, 1982.

[17] G. Chamberlain, "The general equivalence of Granger and Sims causality," *Econometrica: Journal of the Econometric Society*, pp. 569–581, 1982.

[18] C. Sims, "Money, income, and causality," *The American Economic Review*, vol. 62, no. 4, pp. 540–552, 1972.

[19] J. Zhang, "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias," *Artificial Intelligence*, vol. 172, no. 16, pp. 1873–1896, 2008.

[20] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[21] K. P. Murphy, "Dynamic bayesian networks: representation, inference and learning," Ph.D. dissertation, University of California, 2002.

[22] J. Runge, "Detecting and quantifying causality from time series of complex systems," Ph.D. dissertation, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2014.

[23] J. Pearl, "Causality: models, reasoning, and inference," *Econometric Theory*, vol. 19, pp. 675–685, 2003.

[24] P. W. Holland, "Causal inference, path analysis and recursive structural equations models," *ETS Research Report Series*, vol. 1988, no. 1, 1988.

[25] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[26] N. Ay and D. Polani, "Information flows in causal networks," *Advances in complex systems*, vol. 11, no. 01, pp. 17–41, 2008.

[27] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, B. Schölkopf et al., "Quantifying causal influences," *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.

[28] N. Ay and D. C. Krakauer, "Geometric robustness theory and biological networks," *Theory in biosciences*, vol. 125, no. 2, pp. 93–121, 2007.

[29] J. Rissanen and M. Wax, "Measures of mutual and causal dependence between two time series (corresp.)," *Information Theory, IEEE Transactions on*, vol. 33, no. 4, pp. 598–601, 1987.

[30] S. Salehkaleybar, J. Etesami, and N. Kiyavash, "Identifying nonlinear 1-step causal influences in presence of latent variables," *arXiv preprint arXiv:1701.06605*, 2017.

[31] H. Marko, "The bidirectional communication theory—a generalization of information theory," *IEEE Transactions on Communications*, vol. 21, no. 12, pp. 1345–1351, Dec 1973.

[32] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*. Citeseer, 1990, pp. 303–305.

[33] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using granger causality and fmri," *Neuroimage*, vol. 25, no. 1, pp. 230–242, 2005.

[34] M. Besserve, B. Schölkopf, N. K. Logothetis, and S. Panzeri, "Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis," *Journal of computational neuroscience*, vol. 29, no. 3, pp. 547–566, 2010.

[35] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A granger causality measure for point process models of ensemble neural spiking activity," *PLoS computational biology*, vol. 7, no. 3, p. e1001110, 2011.

[36] Y. Liu and S. Aviyente, "Information theoretic approach to quantify causal neural interactions from eeg," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010, pp. 1380–1384.

[37] S. Kim, C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Dynamic and succinct statistical analysis of neuroscience data," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 683–698, 2014.

[38] A. Rao, A. O. Hero III, J. D. Engel et al., "Motif discovery in tissue-specific regulatory sequences using directed information," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, p. 3, 2007.

[39] X. Chen, A. O. Hero, and S. Savarese, "Multimodal video indexing and retrieval using directed information," *Multimedia, IEEE Transactions on*, vol. 14, no. 1, pp. 3–16, 2012.

[40] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[41] C. Quinn, N. Kiyavash, and T. P. Coleman, "Directed information graphs," *Transactions on Information Theory*, vol. 61, no. 12, pp. 6887–6909, 2015.

[42] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[43] M. Chávez, J. Martinerie, and M. Le Van Quyen, "Statistical assessment of nonlinear causality: application to epileptic eeg signals," *Journal of neuroscience methods*, vol. 124, no. 2, pp. 113–128, 2003.

[44] B. Gourévitch and J. J. Eggermont, "Evaluating information transfer between auditory cortical neurons," *Journal of Neurophysiology*, vol. 97, no. 3, pp. 2533–2543, 2007.

[45] K. Hlavácková-Schindler, "Equivalence of granger causality and transfer entropy: a generalization," *Applied Mathematical Sciences*, vol. 5, no. 73, pp. 3637–3648, 2011.

[46] S. Shimizu, A. Hyvarinen, Y. Kano, and P. O. Hoyer, "Discovery of non-gaussian linear causal models using ica," in *Proceedings of the twenty-first conference on Uncertainty in artificial intelligence*, 2005.

[47] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[48] A. Hyvärinen and S. M. Smith, "Pairwise likelihood ratios for estimation of non-gaussian structural equation models," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 111–152, 2013.

[49] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *The Journal of Machine Learning Research*, vol. 7, pp. 2003–2030, 2006.

[50] Y.-B. He and Z. Geng, "Active learning of causal networks with intervention experiments and optimal designs," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[51] S. Tong and D. Koller, "Active learning for structure in Bayesian networks," in *International Joint Conference on Artificial Intelligence*, vol. 17, no. 1, 2001, pp. 863–869.

[52] Y. Hagmayer, S. A. Sloman, D. A. Lagnado, and M. R. Waldmann, "Causal reasoning through intervention," *Causal learning: Psychology, philosophy, and computation*, pp. 86–100, 2007.

[53] D. Materassi and G. Innocenti, "Topological identification in networks of dynamical systems," *Automatic Control, IEEE Transactions on*, vol. 55, no. 8, pp. 1860–1871, 2010.

[54] J. Etesami and N. Kiyavash, "Directed information graphs: A generalization of linear dynamical graphs," in *American Control Conference (ACC), 2014*. IEEE, 2014, pp. 2563–2568.

[55] D. Materassi and M. V. Salapaka, "On the problem of reconstructing an unknown topology via locality properties of the Wiener filter," *IEEE Transactions on Automatic Control*, vol. 57, no. 7, pp. 1765–1777, 2012.

[56] V. Tan and A. Willsky, "Sample complexity for topology estimation in networks of LTI systems," in *2011 IEEE Conference on Decision and Control (CDC)*. IEEE, 2011.

[57] T. J. Liniger, "Multivariate hawkes processes," Ph.D. dissertation, Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403, 2009, 2009.

[58] M. Farajtabar, N. Du, M. G. Rodriguez, I. Valera, H. Zha, and L. Song, "Shaping social activity by incentivizing users," in *Advances in neural information processing systems*, 2014, pp. 2474–2482.

[59] J. G. Rasmussen, "Bayesian inference for hawkes processes," *Methodology and Computing in Applied Probability*, vol. 15, no. 3, pp. 623–642, 2013.

[60] K. Zhou, H. Zha, and L. Song, "Learning triggering kernels for multi-dimensional hawkes processes," in *Proceedings of The 30th International Conference on Machine Learning*, 2013, pp. 1301–1309.

[61] E. C. Hall and R. M. Willett, "Tracking dynamic point processes on networks," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 4327–4346, 2016.

[62] J. Etesami, N. Kiyavash, and T. Coleman, "Learning minimal latent directed information polytrees," *Neural Computation*, 2016.

[63] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "Seismic: A self-exciting point process model for predicting tweet popularity," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1513–1522.

[64] E. Bacry, K. Dayri, and J.-F. Muzy, "Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data," *The European Physical Journal B*, vol. 85, no. 5, pp. 1–12, 2012.

[65] S.-H. Yang and H. Zha, "Mixture of mutually exciting processes for viral diffusion," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1–9.

[66] E. Lewis and G. Mohler, "A nonparametric em algorithm for multiscale hawkes processes," *Journal of Nonparametric Statistics*, pp. 1–16, 2011.

[67] D. Luo, H. Xu, Y. Zhen, X. Ning, H. Zha, X. Yang, and W. Zhang, "Multi-task multi-dimensional hawkes processes for modeling event sequences," 2015.

[68] H. Xu, M. Farajtabar, and H. Zha, "Learning granger causality for hawkes processes," 2016.

[69] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes." in *AISTATS*, vol. 31, 2013, pp. 641–649.

[70] E. Bacry and J.-F. Muzy, "Second order statistics characterization of hawkes processes and non-parametric estimation," *preprint arXiv:1401.0903*, 2014.

[71] N. R. Hansen, P. Reynaud-Bouret, V. Rivoirard et al., "Lasso and probabilistic inequalities for multivariate point processes," *Bernoulli*, vol. 21, no. 1, pp. 83–143, 2015.

[72] M. Eichler, R. Dahlhaus, and J. Dueck, "Graphical modeling for multivariate hawkes processes with nonparametric link functions," *Journal of Time Series Analysis*, 2016.

[73] R. Lemonnier and N. Vayatis, "Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate hawkes processes," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 161–176.

[74] O. Messaouda, J. B. Oommen, and S. Matwin, "Enhancing caching in distributed databases using intelligent polytree representations," in *Advances in Artificial Intelligence*. Springer, 2003, pp. 498–504.

[75] M. S. Zaveri and D. Hammerstrom, "Cmol/cmos implementations of bayesian polytree inference: Digital and mixed-signal architectures and performance/price," *Nanotechnology, IEEE Transactions on*, vol. 9, no. 2, pp. 194–211, 2010.

[76] L. E. Sucar, J. Pérez-Brito, J. C. Ruiz-Suárez, and E. Morales, "Learning structure from data and its application to ozone prediction," *Applied Intelligence*, vol. 7, no. 4, pp. 327–338, 1997.

[77] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell et al., *Molecular cell biology*. WH Freeman New York, 2000, vol. 4.

[78] L. M. de Campos, "Independency relationships in singly connected networks," Citeseer, Tech. Rep., 1994.

[79] G. Rebane and J. Pearl, "The recovery of causal poly-trees from statistical data," *Third Conference on Uncertainty in Artificial Intelligence*, 1987.

[80] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Efficient methods to compute optimal tree approximations of directed information graphs," *IEEE Transactions on Signal Processing*, vol. 61, no. 12, pp. 3173–3182, 2013.

[81] C. Quinn, J. Etesami, N. Kiyavash, and T. Coleman, "Robust Directed Tree Approximations for Networks of Stochastic Processes," *IEEE International Symposium on Information Theory (ISIT)*, 2013, submitted.

[82] C. J. Quinn, A. Pinar, and N. Kiyavash, "Bounded degree approximations of stochastic networks," *arXiv preprint arXiv:1506.04767*, 2015.

[83] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *Information Theory, IEEE Transactions on*, vol. 14, no. 3, pp. 462–467, 1968.

[84] S. Kadloor, X. Gong, N. Kiyavash, T. Tezcan, and N. Borisov, "Low-cost side channel remote traffic analysis attack in packet networks," in *Communications (ICC), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1–5.

[85] A. Houmansadr, N. Kiyavash, and N. Borisov, "Multi-flow attack resistant watermarks for network flows," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1497–1500.

[86] S. Kadloor and N. Kiyavash, "Delay optimal policies offer very little privacy," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 2454–2462.

[87] G. Elidan, N. Friedman, and D. M. Chickering, "Learning hidden variable networks: The information bottleneck approach." *Journal of Machine Learning Research*, vol. 6, no. 1, 2005.

[88] A. Jalali and S. Sanghavi, "Learning the dependence graph of time series with latent factors," *arXiv preprint arXiv:1106.1887*, 2011.

[89] P. Geiger, K. Zhang, B. Schoelkopf, M. Gong, and D. Janzing, "Causal inference by identification of vector autoregressive processes with hidden components," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 1917–1925.

[90] X. Boyen, N. Friedman, and D. Koller, "Discovering the hidden structure of complex dynamic systems," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 91–100.

[91] P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf, "Causal inference by identification of vector autoregressive processes with hidden components," in *Proceedings of 32th International Conference on Machine Learning (ICML 2015)*, 2015.

[92] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen, "Estimation of causal effects using linear non-gaussian causal models with hidden variables," *International Journal of Approximate Reasoning*, vol. 49, no. 2, pp. 362–378, 2008.

[93] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 1610–1613.

[94] G. Elidan, I. Nachman, and N. Friedman, "" ideal parent" structure learning for continuous variable bayesian networks." *Journal of Machine Learning Research*, vol. 8, no. 8, 2007.

[95] A. Anandkumar, K. Chaudhuri, D. Hsu, S. M. Kakade, L. Song, and T. Zhang, "Spectral methods for learning multivariate latent tree structure." in *NIPS*, 2011, pp. 2025–2033.

[96] M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning latent tree graphical models," *The Journal of Machine Learning Research*, vol. 12, pp. 1771–1812, 2011.

[97] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 499–506.

[98] P. F. Lazarsfeld and N. W. Henry, *Latent structure analysis*. Houghton, Mifflin, 1968.

[99] N. L. Zhang and T. Kocka, "Efficient learning of hierarchical latent class models," in *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*. IEEE, 2004, pp. 585–593.

[100] T. Jiang, P. Kearney, and M. Li, "A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application," *SIAM Journal on Computing*, vol. 30, no. 6, pp. 1942–1961, 2001.

[101] P. L. Erdos, M. A. Steel, L. Székely, and T. J. Warnow, "A few logs suffice to build (almost) all trees: Part ii," *Theoretical Computer Science*, vol. 221, no. 1, pp. 77–118, 1999.

[102] M. Steel, "The complexity of reconstructing trees from qualitative characters and subtrees," *Journal of classification*, vol. 9, no. 1, pp. 91–116, 1992.

[103] J. S. Farris, "Estimating phylogenetic trees from distance matrices," *American Naturalist*, pp. 645–668, 1972.

[104] S. Sattath and A. Tversky, "Additive similarity trees," *Psychometrika*, vol. 42, no. 3, pp. 319–345, 1977.

[105] M. Ishteva, H. Park, and L. Song, "Unfolding latent tree structures using 4th order tensors," *International Conference on Machine Learning*, 2013.

[106] K. S. John, T. Warnow, B. M. Moret, and L. Vawter, "Performance study of phylogenetic methods:(unweighted) quartet methods and neighbor-joining," *Journal of Algorithms*, vol. 48, no. 1, pp. 173–193, 2003.

[107] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[108] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, 1998.

[109] S. L. Lauritzen, *Graphical models*. Oxford University Press, 1996.

[110] R. A. Ali, T. S. Richardson, and P. Spirtes, "Markov equivalence for ancestral graphs," *The Annals of Statistics*, pp. 2808–2837, 2009.

[111] C. Meek, "Strong completeness and faithfulness in bayesian networks," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 411–418.

[112] N. Wiener, "The theory of prediction," in *Modern Mathematics for Engineers*, 1st ed., E. F. Beckenback, Ed. McGraw-Hill, 1956, pp. 165–190.

[113] C. J. Quinn, T. P. Coleman, and N. Kiyavash, "A generalized prediction framework for granger causality," in *IEEE Conf. on Computer Communications Workshops (INFOCOM WKSHPS),*, 2011, pp. 906–911.

[114] T. S. Verma and J. Pearl, "Causal networks: Semantics and expressiveness," *Fourth Conference on Uncertainty in Artificial Intelligence*, 1988.

[115] J. Geweke, "Inference and causality in economic time series models," *Handbook of econometrics*, vol. 2, pp. 1101–1144, 1984.

[116] U. Triacca, "Is granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature?" *Theoretical and applied climatology*, vol. 81, no. 3-4, pp. 133–135, 2005.

[117] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[118] L. Ljung, "System identification," in *Signal Analysis and Prediction*. Springer, 1998, pp. 163–173.

[119] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[120] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[121] D. Materassi, "Reconstruction of topologies for acyclic networks of dynamical systems," in *Proceedings of the IEEE American control conference*, 2011.

[122] A. Anandkumar, V. Y. Tan, F. Huang, and A. S. Willsky, "High-dimensional gaussian graphical model selection: Walk summability and local separation criterion," *Journal of Machine Learning Research*, vol. 13, no. Aug, pp. 2293–2337, 2012.

[123] K. Sricharan, R. Raich, and A. O. Hero III, "Empirical estimation of entropy functionals with confidence," *arXiv preprint arXiv:1012.4188*, 2010.

[124] J. Etesami, A. Habibnia, and N. Kiyavash, "Econometric modeling of systemic risk: Going beyond pairwise comparison and allowing for nonlinearity," *Systemic Risk Centre, The London School of Economics and Political Science*, 2017.

[125] Y. Ogata, "Seismicity analysis through point-process modeling: A review," *Pure and Applied Geophysics*, vol. 155, no. 2-4, pp. 471–507, 1999.

[126] P. Reynaud-Bouret, S. Schbath et al., "Adaptive estimation for hawkes processes; application to genome analysis," *The Annals of Statistics*, vol. 38, no. 5, pp. 2781–2822, 2010.

[127] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, 2011.

[128] C. G. Bowsher, "Modelling security market events in continuous time: Intensity based, multivariate point process models," *Journal of Econometrics*, vol. 141, no. 2, pp. 876–912, 2007.

[129] I. Muni Toke and F. Pomponio, "Modelling trades-through in a limited order book using hawkes processes," *Economics discussion paper*, no. 2011-32, 2011.

[130] T. Weissman, Y.-H. Kim, and H. H. Permuter, "Directed information, causal estimation, and communication in continuous time," *Information Theory, IEEE Transactions on*, vol. 59, no. 3, pp. 1271–1287, 2013.

[131] K. Zhou, H. Zha, and L. Song, "Learning triggering kernels for multi-dimensional hawkes processes," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1301–1309.

[132] P. Brémaud and L. Massoulié, "Stability of nonlinear hawkes processes," *The Annals of Probability*, pp. 1563–1588, 1996.

[133] E. Bacry, S. Delattre, M. Hoffmann, and J.-F. Muzy, "Some limit theorems for hawkes processes and application to financial statistics," *Stochastic Processes and their Applications*, vol. 123, no. 7, pp. 2475–2499, 2013.

[134] S. W. Linderman and R. P. Adams, "Discovering latent network structure in point process data," *preprint arXiv:1402.0914*, 2014.

[135] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *The journal of political economy*, pp. 637–654, 1973.

[136] J. C. L. Pinto, T. Chahed, and E. Altman, "Trend detection in social networks using hawkes processes," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 1441–1448.

[137] V. Shmatikov and M.-H. Wang, "Timing analysis in low-latency mix networks: Attacks and defenses," in *Computer Security–ESORICS 2006*. Springer, 2006, pp. 18–33.

[138] N. Kiyavash and T. Coleman, "Covert timing channels codes for communication over interactive traffic," in *Acoustics, Speech and Signal processing, 2009. ICASSP 2009. IEEE international conference on*. IEEE, 2009, pp. 1485–1488.

[139] S. Kadloor, N. Kiyavash, and P. Venkitasubramaniam, "Mitigating timing based information leakage in shared schedulers," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 1044–1052.

[140] S. Kadloor, N. Kiyavash, and P. Venkitasubramaniam, "Scheduling with privacy constraints," in *Information Theory Workshop (ITW), 2012 IEEE*. IEEE, 2012, pp. 40–44.

[141] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[142] D. Marček, "Stock price prediction using autoregressive models and signal processing procedures," in *Proceedings of the 16th Conference MME*, vol. 98, 1998, pp. 114–121.

[143] H. Lütkepohl and M. Krätzig, *Applied time series econometrics*. Cambridge university press, 2004.

[144] A. N. Patrinos and S. L. Hakimi, "The distance matrix of a graph and its tree realization," *Quarterly of applied mathematics*, pp. 255–269, 1972.

[145] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[146] J. Etesami and N. Kiyavash, "Interventional dependency graphs: An approach for discovering influence structure," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1158–1162.

[147] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International statistical review*, vol. 70, no. 3, pp. 419–435, 2002.

[148] M. Singh and M. Valtorta, "Construction of bayesian network structures from data: a brief survey and an efficient algorithm," *International journal of approximate reasoning*, vol. 12, no. 2, pp. 111–131, 1995.

[149] C. Villani, "Topics in optimal transportation (graduate studies in mathematics, vol. 58)," 2003.

[150] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, pp. 429–443, 1997.

[151] R. M. Dudley, *Real analysis and probability*. Cambridge University Press, 2002, vol. 74.

[152] A. W. Van Der Vaart and J. A. Wellner, *Weak Convergence*. Springer, 1996.

[153] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.

[154] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, "A kernel statistical test of independence," in *Advances in neural information processing systems*, 2007, pp. 585–592.

[155] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence." in *NIPS*, vol. 20, 2007, pp. 489–496.

[156] X. Sun, D. Janzing, B. Schölkopf, and K. Fukumizu, "A kernel-based causal learning algorithm," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 855–862.

[157] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Advances in neural information processing systems*, 2006, pp. 513–520.

[158] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *Journal of Machine Learning Research*, vol. 7, no. Dec, pp. 2651–2667, 2006.

[159] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery." Corvallis, OR, USA: AUAI Press, July 2011, pp. 804–813.

[160] C. R. Givens, R. M. Shortt et al., "A class of wasserstein metrics for probability distributions," *Michigan Math. J*, vol. 31, no. 2, pp. 231–240, 1984.

[161] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. Lanckriet, "Non-parametric estimation of integral probability metrics," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1428–1432.

[162] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.

[163] P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. Aimale, and F. Wimberly, "Constructing bayesian network models of gene expression networks from microarray data," 2000.

[164] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[165] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[166] A. Bolstad, B. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[167] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[168] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[169] A. Jeffrey, *Complex analysis and applications*. CRC Press, 2005, vol. 10.

[170] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[171] J. Gutiérrez-Gutiérrez, P. M. Crespo et al., *Block Toeplitz matrices: asymptotic results and applications*. Now, 2012.

[172] D. S. Johnson, "The np-completeness column: an ongoing guide," *Journal of Algorithms*, vol. 6, no. 3, pp. 434–451, 1985.

APPENDIX A

PROOFS OF THEOREMS

A.1 Proofs of Chapter 2

A.1.1 Proof of Theorem 2

Suppose Z c-separates U from w in a DIG. Then, we need to show

$$I\left(\underline{\mathbf{X}}_U \rightarrow \ \mathbf{X}_w || \underline{\mathbf{X}}_Z\right) = 0.$$

Let $\mathcal{A} := \mathcal{PA}(\underline{\mathbf{X}}_w) \setminus Z$ be the parent set of w except the ones that are already in Z. By the definition of DIG, we have

$$I\left(\underline{\mathbf{X}}_U \rightarrow \ \mathbf{X}_w || \underline{\mathbf{X}}_\mathcal{A}, \underline{\mathbf{X}}_Z\right) = 0. \tag{A.1}$$

If for any t,

$$D\left(P_{\underline{X}^t_{\mathcal{A},1}|\ \underline{X}^t_{U\cup\{w\}\cup Z,1}} \ || \ P_{\underline{X}^t_{\mathcal{A},1}|\ \underline{X}^t_{\{w\}\cup Z,1}}\right) = 0. \tag{A.2}$$

Then, (A.2) and (A.1) will imply the result. In order to show (A.2), we use the d-separation criterion for the corresponding boundary DAG introduced in Section 2.3.1. Notice that every path from a node in U and a node in $\mathcal{A}$ contains at least a node in $Z \cup \{w\}$ with an outgoing arrow, or contains a collider that is not in $Z \cup \{w\}$, which implies that every path in the corresponding boundary DAG between $\underline{X}^t_{\mathcal{A},1}$ and $\underline{X}^t_{U,1}$ is d-separated by $\underline{X}^t_{\{w\}\cup Z,1}$, consequently, (A.2) holds.

A.2 Proofs of Chapter 3

A.2.1 Proof of Theorem 3

We use proof by contradiction. Suppose there exist two FDGs $\overrightarrow{G}_1$ and $\overrightarrow{G}_2$ associated with a dynamical system given by (3.1) with positive joint distribution. Assume (j,i) belongs to $\overrightarrow{G}_1$ but it does not belong to $\overrightarrow{G}_2$. Corresponding to the FDG $\overrightarrow{G}_2$, there exists a set of exogenous noises $\{\mathbf{W}_i\}$ and a set of functions $\{F_i\}$s such that

$$X_{i,t} = F_i(X_{j,t'}, \mathcal{R}, W_{i,t}, t), \tag{A.3}$$

in which $\mathcal{R}$ denotes $\underline{X}^{t-1} \setminus \{X_{j,t'}\}$ and $W_{i,t}$ is independent of $\{\mathcal{R}, X_{j,t'}\} \cup \{W_k^t : k \neq i\}$. We define a new random process as follows

$$\tilde{X}_{i,t} := F_i(x, \mathcal{R}, W_{i,t}, t), \tag{A.4}$$

where $x \in \mathcal{E}$ is a realization of $X_{j,t'}$. We will show that $d(\tilde{X}_{i,t}, X_{i,t}) = 0$ with probability one. Hence, $X_{i,t}$ can be written as a function of $(\mathcal{R}, W_{i,t}, t)$, i.e., there exists a function Ψ, such that

$$X_{i,t} = \Psi(\mathcal{R}, W_{i,t}, t). \tag{A.5}$$

To show this we use the fact that (j,i) does not belong to $\overrightarrow{G}_2$. Therefore, $\alpha_{i,j}(t,t')$ define in (3.2) equals zero for all t and t'. This implies that for any triple (x,y,R) in which $d(x,y) > 0$, measure of the following set is zero,

$$\mathcal{S}_1 := \{w : d(F_i(x,R,w,t), F_i(y,R,w,t)) > 0\},$$

where R denotes a realization of $\mathcal{R}$. In another words, for every pair (R,y), we have

$$P\left(d(\tilde{X}_{i,t}, X_{i,t}) = 0 | \mathcal{R} = R, X_{j,t'} = y\right) = 1.$$

Using the total probability law and the above equality, we obtain

$$P\left(d(\tilde{X}_{i,t}, X_{i,t}) = 0\right) = \sum_{R,y} P\left(d(\tilde{X}_{i,t}, X_{i,t}) = 0 | \mathcal{R} = R, X_{j,t'} = y\right) \times P\left(\mathcal{R} = R, X_{j,t'} = y\right) = 1.$$

On the other hand, corresponding to the FDG $\overrightarrow{G}_1$, there exists a set of exogenous noises $\{\mathbf{W}'_i\}$ and a set of functions $\{G_i\}$s such that

$$X_{i,t} = G_i(X_{j,t'}, \mathcal{R}, W'_{i,t}, t), \tag{A.6}$$

where $W'_{i,t}$ denotes the exogenous noise and it is independent of $\{\mathcal{R}, X_{j,t'}\} \cup \{W_k'^t : k \neq i\}$. Using (A.5), (A.6), with probability one, we have

$$G_i(X_{j,t'}, \mathcal{R}, W'_{i,t}, t) = \Psi(\mathcal{R}, W_{i,t}, t). \tag{A.7}$$

Recall that $\mathcal{R}$ denotes $\underline{X}^{t-1} \setminus \{X_{j,t'}\}$ and both $W_{i,t}$ and $W'_{i,t}$ are independent of $\mathcal{R} \cup \{X_{j,t'}\}$. Below, we use this independency and the fact that (j,i) belongs to $\overrightarrow{G}_1$ to derive the contradiction.

Because (j,i) belongs to $\overrightarrow{G}_1$ and using the Definition 7, we obtain that there exist t and t' such that $\alpha_{i,j}(t,t') > 0$. Consequently, there exist realizations (x^*, y^*, R^*) in which $d(x^*, y^*) > 0$ and the following set has positive measure

$$\mathcal{S}_2 := \{w' : d(G_i(x^*, R^*, w', t), G_i(y^*, R^*, w', t)) > 0\}.$$

Equivalently,

$$d(G_i(x^*, R^*, W'_{i,t}, t), G_i(y^*, R^*, W'_{i,t}, t)) > 0,$$

with positive probability. We define tow random variables as follows,

$$\begin{aligned} Z_0 &:= \Psi(R^*, W_{i,t}, t), \\ Z_1 &:= G_i(X_{j,t'}, R^*, W'_{i,t}, t). \end{aligned} \tag{A.8}$$

Note that such random variables are well define because of positivity assumption, i.e., $P(X_{j,t'} = x^* | \mathcal{R} = R^*) > 0$. Because $W_{i,t}$ is independent of $X_{j,t'}$, Z_0 is also independent of $X_{j,t'}$, and because Z_0 is not a function of $X_{j,t'}$, varying $X_{j,t'}$ will not change the value of Z_0, i.e., the following set has measure one,

$$\{w : d(Z_0|_{X_{j,t'}=x^*}, Z_0|_{X_{j,t'}=y^*}) = 0\}, \tag{A.9}$$

where $Z_0|_{X_{j,t'}=x^*}$ denotes the value of Z_0 after fixing the value of $X_{j,t'}$ to be x^*. On the other hand, from (A.7) and (A.8), we imply

$$\begin{aligned} Z_0|_{X_{j,t'}=x^*} &= G_i(x^*, R^*, W'_{i,t}, t), \\ Z_0|_{X_{j,t'}=y^*} &= G_i(y^*, R^*, W'_{i,t}, t). \end{aligned} \tag{A.10}$$

Combining (A.9), (A.10), and the fact $\mathcal{S}_2$ has positive measure, the contradiction will follow.

A.2.2 Proof of Proposition 1

We prove it by showing that (j, i) does not belong to the linear dynamical graph if and only if $(j, i) \notin \overrightarrow{E}_{FD}$. Suppose, (j, i) does not belong to the linear dynamical graph, by the Definition 8, $G_{i,j}(z) = 0$, equivalently, $g_{i,j}(s) = 0$ for $s > 0$. This implies that $\alpha_{i,j} = 0$, i.e.,

$$\sum_{s>0} |g_{i,j}(s)| = 0.$$

The converse can be shown similarly.

A.2.3 Proof of Theorem 4

In order to prove the above statement, we show that if $(j, i) \notin \overrightarrow{E}_{FD}$, then $(j, i) \notin \overrightarrow{E}_{DI}$.
Suppose, $(j, i) \notin \overrightarrow{E}_{FD}$, then by the Definition 7, $\alpha_{i,j} = 0$, which implies $\alpha_{i,j}(t, t') = 0$ for all t and $t' \leq t$. Consequently, using Equation (3.2), we obtain that for every t and $(x, y) \in \mathcal{E}^2$, the following set has measure zero,

$$\Big\{ w \in \mathcal{E} : d\Big(F_i(\underline{x}, w, t), F_i(\underline{y}, w, t)\Big) > 0 \Big\}.$$

Recall that $\underline{x}$ and $\underline{y}$ are two realizations of $\underline{X}^{t-1}$. We consider the following conditional probability for an event set $\mathcal{W} \in \mathcal{B}$,

$$P(X_{i,t} \in \mathcal{W} | \underline{X}^{t-1}\} = \underline{x}) = P(F_i(\underline{x}, W_{i,t}, t) \in \mathcal{W} | \underline{X}^{t-1}\} = \underline{x}) = P(F_{i,\underline{x}}^{-1}(\mathcal{W})), \tag{A.11}$$

where $F_{i,\underline{x}}^{-1}(\mathcal{W}) := \{w \in \mathcal{E} : F_i(\underline{x}, w, t) \in \mathcal{W}\}$.

Observe that $F_{i,\underline{y}}^{-1}(\mathcal{W})$ can be written as the union of the following two events:

$$\Big\{w : F_i(\underline{y}, w, t) \in \mathcal{W}, d\Big(F_i(\underline{x}, w, t), F_i(\underline{y}, w, t)\Big) = 0\Big\}$$
$$\cup \Big\{w : F_i(\underline{y}, w, t) \in \mathcal{W}, d\Big(F_i(\underline{x}, w, t), F_i(\underline{y}, w, t)\Big) > 0\Big\}.$$

Note that the second term in the above expression has zero measure and the first term is a subset of $F_{i,\underline{x}}^{-1}(\mathcal{W})$. Hence, $F_{i,\underline{y}}^{-1}(\mathcal{W}) \subseteq F_{i,\underline{x}}^{-1}(\mathcal{W})$. Similarly, one can show $F_{i,\underline{x}}^{-1}(\mathcal{W}) \subseteq F_{i,\underline{y}}^{-1}(\mathcal{W})$ and thus, we have

$$F_{i,\underline{y}}^{-1}(\mathcal{W}) = F_{i,\underline{x}}^{-1}(\mathcal{W}),$$

with probability one. This implies

$$P(X_{i,t} \in \mathcal{W} | \underline{X}^{t-1}\} = \underline{x}) = P(X_{i,t} \in \mathcal{W} | \underline{X}^{t-1}\} = \underline{y}),$$

for all $\mathcal{W} \in \mathcal{B}$ and $\underline{x}$ and $\underline{y}$ that are only different in $X_{j,t'}$. Using this fact and total probability law, we obtain

$$\begin{aligned}
&P(X_{i,t} | \underline{X}^{t-1} \setminus \{X_{j,t'}\} = \underline{x} \setminus \{x\}) = \int_{x \in \mathcal{E}} P\left(X_{i,t}, X_{j,t'} \middle| \underline{X}^{t-1} \setminus \{X_{j,t'}\} = \underline{x} \setminus \{x\}\right) dX_{j,t'} \\
&= \int_{x \in \mathcal{E}} P\left(X_{i,t} | \underline{X}^{t-1} = \underline{x}\right) P\left(X_{j,t'} | \underline{X}^{t-1} \setminus \{X_{j,t'}\} = \underline{x} \setminus \{x\}\right) dX_{j,t'} \\
&= P\left(X_{i,t} | \underline{X}^{t-1} = \underline{x}\right).
\end{aligned}$$

The above equation implies that for any t and $t' < t$, $X_{i,t}$ is independent of $X_{j,t'}$ given $\underline{X}^{t-1} \setminus \{X_{j,t'}\}$. Using Assumption 1 and the above result, we will show that for any t, $X_{i,t}$ is also independent of $\cup_{t' \leq b} X_{j,t'}$ given $\underline{X}^{t-1} \setminus \{\cup_{t' \leq b} X_{j,t'}\}$ for any $b < t$. To do this, we use induction on b.
First case, $b = 2$: from the above results, i.e., for any t and t', $X_{i,t}$ is independent of $X_{j,t'}$ given $\underline{X}^{t-1} \setminus \{X_{j,t'}\}$, we have

$$P(X_{i,t} | \underline{X}^{t-1}) = P(X_{i,t} | \underline{X}^{t-1} \setminus \{X_{j,1}\}) = P(X_{i,t} | \underline{X}^{t-1} \setminus \{X_{j,2}\}).$$

By total probability law and the above equalities, we obtain

$$\begin{aligned}
&P(X_{i,t} | \underline{X}^{t-1} \setminus \{X_{j,1}, X_{j,2}\}) = \int_{X_j(2)} P(X_{i,t}, X_{j,2} | \underline{X}^{t-1} \setminus \{X_{j,1}, X_{j,2}\}) dX_{j,2} \qquad \text{(A.12)} \\
&= \int_{X_{j,2}} P(X_{i,t} | \underline{X}^{t-1} \setminus \{X_{j,1}\}) \times P(X_{j,2} | \underline{X}^{t-1} \setminus \{X_{j,1}, X_{j,2}\}) dX_{j,2} \\
&= \int_{X_{j,2}} P(X_{i,t} | \underline{X}^{t-1} \setminus \{X_{j,2}\}) \times P(X_{j,2} | \underline{X}^{t-1} \setminus \{X_{j,1}, X_{j,2}\}) dX_{j,2} \\
&= P(X_{i,t} | \underline{X}^{t-1} \setminus \{X_{j,2}\}).
\end{aligned}$$

Suppose that the case $b = t - 2$ holds true, then following the same steps as in (A.29), we can show the final case $b = t - 1$, i.e.,

$$P(X_{i,t} | \underline{X}^{t-1}) = P(X_{i,t} | \underline{X}_{-\{j\}}^{t-1}).$$

Thus, $X_{i,t}$ is independent of $\mathbf{X}_j$ given $\underline{X}_{-\{j\}}^{t-1}$ for all t, which means $I(\mathbf{X}_j \to \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,j\}}) = 0$.

A.2.4 Proof of Theorem 5

From Theorem 4, we know that if $(j,i) \notin \overrightarrow{E}_{FD}$, then $(j,i) \notin \overrightarrow{E}_{DI}$. Here, we prove the converse. If

$$I(\mathbf{X}_j \to \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,j\}}) = 0,$$

then for any t, X_j^{t-1} and $X_{i,t}$ are independent given $\underline{X}_{-\{j\}}^{t-1}$. In this case, we obtain

$$\mathbb{E}[X_{i,t}|\underline{X}^{t-1} = \underline{x}] = \mathbb{E}[X_{i,t}|\underline{X}^{t-1} = \underline{y}],$$
$$Var[X_{i,t}|\underline{X}^{t-1} = \underline{x}] = Var[X_{i,t}|\underline{X}^{t-1} = \underline{y}],$$

where $\underline{x}$ and $\underline{y}$ are two realizations of $\underline{X}^{t-1}$ that are only different in $X_{j,t'}$ for some $t' < t$.
The above equalities and (3.9) imply

$$f_i(\underline{x},t) + g_i(\underline{x},t)\mu_{i,t} = f_i(\underline{y},t) + g_i(\underline{y},t)\mu_{i,t}, \tag{A.13}$$
$$g_i^2(\underline{x},t) = g_i^2(\underline{y},t), \tag{A.14}$$

where $\mu_{i,t} := \mathbb{E}[W_{i,t}]$. From equations (3.9), (A.13), and (A.14), we obtain

$$d^2\Big(F_i(\underline{x},W_{i,t},t), F_i(\underline{y},W_{i,t},t)\Big)/d^2(x,y) =$$
$$\big((f_i(\underline{x},t) - f_i(\underline{y},t)) + (g_i(\underline{x},t) - g_i(\underline{y},t))W_{i,t}\big)^2/(x-y)^2.$$

On the other hand, by the definition of $\alpha_{i,j}(t,t')$ and (3.9), one can simplify (7.3) as follows:

$$\sup_{\substack{\underline{x}=\underline{y} \\ \text{off } X_j(t')}} \Bigg[\left(\frac{f_i(\underline{x},t) - f_i(\underline{y},t)}{x-y}\right)^2 + \left(\frac{g_i(\underline{x},t) - g_i(\underline{y},t)}{x-y}\right)^2 \sigma_i^2(t) \tag{A.15}$$
$$+2\frac{(f_i(\underline{x},t) - f_i(\underline{y},t))(g_i(\underline{x},t) - g_i(\underline{y},t))}{(x-y)^2}\mu_{i,t}\Bigg]^{1/2}, \tag{A.16}$$

where $\sigma_i^2(t) := \mathbb{E}[W_{i,t}^2]$.
If g_i satisfies (3.10), then (A.13)-(A.15) imply that (A.15) is zero and consequently $\alpha_{i,j}(t,t') = 0$ for all $t' < t$.
Otherwise, assume W_i is asymmetric. Using Equation (A.14), we have either:
(i) $g_i(\underline{x},t) = g_i(\underline{y},t)$ or
(ii) $g_i(\underline{x},t) + g_i(\underline{y},t) = 0$.
In the first case (i), clearly we have $\alpha_{i,j}(t,t') = 0$. In the second case (ii), i.e., $g_i(\underline{x},t) + g_i(\underline{y},t) = 0$, using Equation (A.13), we have

$$f_i(\underline{x},t) + 2g_i(\underline{x},t)\mu_{i,t} = f_i(\underline{y},t). \tag{A.17}$$

Since $X_{i,t}$ and X_j^t are independent given $\underline{X}_{-\{j\}}^t \setminus \{X_{i,t}\}$, the following two random variables must have the same distributions,

$$X_{i,t}|\underline{X}^{t-1} = \underline{x}, \quad X_{i,t}|\underline{X}^{t-1} = \underline{y},$$

where $\underline{x}$ and $\underline{y}$ are two realizations of $\underline{X}^{t-1}$ that only differ at $X_{j,t'}$. By the definition of $X_{i,t}$, we imply

$$f_i(\underline{x},t) + g_i(\underline{x},t)W_{i,t} \sim f_i(\underline{y},t) + g_i(\underline{y},t)W_{i,t}.$$

Substituting (A.17) into the above expression, the fact that $g_i(\underline{x},t)+g_i(\underline{y},t) = 0$, and adding and subtracting $g_i(\underline{x},t)\mu_{i,t}$, we obtain

$$f_i(\underline{x},t) + g_i(\underline{x},t)\mu_{i,t} + g_i(\underline{x},t)\left(W_{i,t} - \mu_{i,t}\right) \sim f_i(\underline{x},t) + g_i(\underline{x},t)\mu_{i,t} - g_i(\underline{x},t)\left(W_{i,t} - \mu_{i,t}\right).$$

This can only happen if $\mathbf{W}_i$ is symmetric, which contradicts our assumption.

A.3 Proofs of Chapter 4

A.3.1 Proof of Proposition 3

Suppose $\gamma_{i,j} \equiv 0$. (4.2) implies that for every $t \leq T$, $\lambda_i(t)$ is $\underline{\mathcal{F}}^t_{-\{j\}}(= \sigma\{\underline{N}^t_{-\{j\}}\})$-measurable and from (4.1), we have

$$P\left(dN_i(t) = 1|\underline{\mathcal{F}}^t\right) = P(dN_i(t) = 1|\underline{\mathcal{F}}^t_{-\{j\}}).$$

Equivalently, for every $0 \leq t_{k-1} < t_k$,

$$I\left(N^{t_k}_{i,t_{k-1}}; N^{t_k}_{j,0}|\mathcal{F}^{t_{k-1}}_{-\{j\}}\right) = 0, \tag{A.18}$$

and thus, $\tilde{I}_{\mathbf{t}}(\mathbf{N}_j \to \mathbf{N}_i||\underline{\mathbf{N}}_{-\{i,j\}}) = 0$, for any finite partition $\mathbf{t} \in \mathcal{T}(0,T)$.
For the converse we use proof by contradiction. Suppose $I_T(\mathbf{N}_j \to \mathbf{N}_i||\underline{\mathbf{N}}_{-\{i,j\}}) = 0$ and $\gamma_{i,j} \neq 0$. Using the definition in (4.4), it is straightforward to observe that for any $t < T$,

$$I_t(\mathbf{N}_j \to \mathbf{N}_i||\underline{\mathbf{N}}_{-\{i,j\}}) = 0.$$

Similarly, $I_{t+dt}(\mathbf{N}_j \to \mathbf{N}_i||\underline{\mathbf{N}}_{-\{i,j\}}) = 0$. Consequently,

$$0 = I_{t+dt}(\mathbf{N}_j \to \mathbf{N}_i||\underline{\mathbf{N}}_{-\{i,j\}}) - I_t(\mathbf{N}_j \to \mathbf{N}_i||\underline{\mathbf{N}}_{-\{i,j\}}) = I\left(dN_{i,t}; N^t_{j,0}|\mathcal{F}^t_{-\{j\}}\right).$$

This implies $P(dN_{i,t} = 1|\underline{\mathcal{F}}^t_{-\{j\}}) = \lambda_i(t)dt + o(dt)$, or $\lambda_i(t)$ is $\underline{\mathcal{F}}^t_{-\{j\}}$-measurable. Since, we have assumed $\gamma_{i.j} \neq 0$, we obtain $N_{j,t}$ is $\underline{\mathcal{F}}^t_{-\{j\}}$-measurable, for all $t \leq T$. In words, jth process is determined by other processes which contradicts with the Assumption 1 that states there is no deterministic relationships between processes.

A.3.2 Proof of Corollary 1

If the excitation matrix belongs to $\mathcal{E}xp(m)$, from Equation (4.8) we have

$$\left(I - \sum_{d=1}^{D} \frac{A_d^T}{j\omega + \beta_d}\right) diag(\Lambda)^{-1} \left(I - \sum_{d=1}^{D} \frac{A_d}{-j\omega + \beta_d}\right) = \frac{4\sin^2 z\omega/2}{\omega^2 z}\mathcal{F}[\Sigma_z]^{-1}(\omega).$$

By evaluating the trace of the above equation, we obtain

$$\sum_{i=1}^{m} \frac{|1-a_{i,i}|^2}{\lambda_i} + \sum_{i\neq j} \frac{|a_{i,j}|^2}{\lambda_i} = \frac{4\sin^2 z\omega/2}{\omega^2 z} \operatorname{Tr} \mathcal{F}[\Sigma_z]^{-1}(\omega), \tag{A.19}$$

where $a_{i,j} = \sum_{d=1}^{D} \frac{a_{i,j}^{(d)}}{-j\omega+\beta_d}$, and $A_d = [a_{i,j}^{(d)}]$. To learn the entire set $\{\pm j\beta_d\}$, we have to show that there are no pole zero cancellations in (A.19). That is, the nominator and denominator of (A.19) have no common roots. Let

$$g(\omega) := \left(\sum_{i=1}^{m} \frac{|1-a_{i,i}|^2}{\lambda_i} + \sum_{i\neq j} \frac{|a_{i,j}|^2}{\lambda_i} \right) \prod_{d=1}^{D} |-j\omega+\beta_d|^2,$$

which is the nominator of Equation (A.19). It is straightforward to check that for $\omega = -j\beta_k$, the above quantity is non-zero, due to the fact that β_ds are distinct and $A_k \neq \mathbf{0}$. Since $g(\omega)$ is a polynomial with real coefficients, from complex conjugate root theorem [169], we have $g(j\beta_k) \neq 0$. Therefore, the set $\{\pm j\beta_d\}$ contains all the poles of (A.19).

A.3.3 Proof of Proposition 4

From Lemma 1, the Laplace transform of the covariance density can be written as

$$\mathcal{L}[\Omega](s) = \mathcal{L}[\Gamma](s)\left(diag(\Lambda) + \mathcal{L}[\Omega](s)\right) + \int_0^\infty \int_t^\infty \Gamma(t')\Omega^T(t) e^{-s(t'-t)} dt' dt. \tag{A.20}$$

When $\Gamma(t) \in \mathcal{E}xp(m)$, it can be shown that (A.20) becomes

$$\mathcal{L}[\Omega](s) = \sum_{d=1}^{D} \frac{A_d}{s+\beta_d} \left(diag(\Lambda) + \mathcal{L}[\Omega](s) + \mathcal{L}[\Omega]^T(\beta_d) \right). \tag{A.21}$$

If the set of exciting modes are given, we can insert $s = \beta_d$, for $d = 1, \ldots, D$ in the above equation and obtain the system of D equations.

A.4 Proofs of Chapter 5

A.4.1 Proof of Lemma 2

We consider two cases: i) If $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{R}$ (the root set). Using the chain rule we have

$$P_{\underline{\mathbf{X}}} = \prod_{t=1}^{n} P_{\underline{X}_t|\underline{\mathbf{X}}_1^{t-1}} =$$
$$\prod_{t=1}^{n} \prod_{a\in\mathcal{A}_1} \prod_{b\in\mathcal{A}_2} P_{X_{a,t}|\underline{\mathbf{X}}_1^{t-1}, \underline{\mathbf{X}}_{\mathcal{S}(a),t}} P_{X_{b,t}|\underline{\mathbf{X}}_1^{t-1}, \underline{\mathbf{X}}_{\mathcal{S}(b),t}} P_{\underline{X}_{-\mathcal{A}_1\cup\mathcal{A}_2,t}|\underline{\mathbf{X}}_1^{t-1}}, \tag{A.22}$$

where for every x, $\mathcal{S}(x) \subseteq -\{x\}$ such that the above equation holds. Note that, if we consider no simultaneous influences, then $\mathcal{S}(x) = \emptyset$ for every x. By the definition of DI, we also have

$$D(P_{\mathbf{X}_a||\underline{\mathbf{X}}_{-\{a\}}} || \; P_{\mathbf{X}_a}) = 0, \; \forall a \in \mathcal{A}_1 \cup \mathcal{A}_2.$$

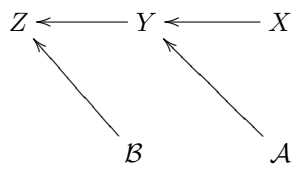


Figure A.1. DIG in Lemma 4. $\mathcal{A} \cup \{\mathbf{X}\}$ is the parent set of $\mathbf{Y}$, and $\mathcal{B} \cup \{\mathbf{Y}\}$ is the parent set of $\mathbf{Z}$.

Combining the above equations implies

$$P_{\underline{\mathbf{X}}} = P_{\underline{\mathbf{X}}_{\mathcal{A}_1}} P_{\underline{\mathbf{X}}_{\mathcal{A}_2}} \prod_{t=1}^{n} P_{\underline{X}_{-\mathcal{A}_1 \cup \mathcal{A}_2, t} | \underline{\mathbf{X}}_1^{t-1}}$$

On the other hand, again using chain rule we have $P_{\underline{\mathbf{X}}} = P_{\underline{\mathbf{X}}_{\mathcal{A}_1, \mathcal{A}_2}} P_{\underline{\mathbf{X}} \setminus (\mathcal{A}_1 \cup \mathcal{A}_2) | \mathcal{A}_1, \mathcal{A}_2}$. The equivalence between the two last equations and the positivity assumption, implies that $\underline{\mathbf{X}}_{\mathcal{A}_1}$ and $\underline{\mathbf{X}}_{\mathcal{A}_2}$ are independent.
ii) Otherwise, let $\mathcal{B}_1$ and $\mathcal{B}_2$ to be the set of all parents of $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. Since the system has a tree structure, then, $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$. Similar to the previous case, one can obtain

$$D\left(P_{\underline{\mathbf{X}}_{\mathcal{A}_1} | \underline{\mathbf{X}}_{\mathcal{A}_2 \cup \mathcal{B}_1 \cup \mathcal{B}_2}} || P_{\underline{\mathbf{X}}_{\mathcal{A}_1} | | \underline{\mathbf{X}}_{\mathcal{B}_1}}\right) = 0.$$

Therefore, $\underline{\mathbf{X}}_{\mathcal{A}_1}$ and $\underline{\mathbf{X}}_{\mathcal{A}_2}$ are independent if $\underline{\mathbf{X}}_{\mathcal{B}_1}$ and $\underline{\mathbf{X}}_{\mathcal{B}_2}$ are independent. By continuing the same procedure, we will end up with two disjoint subsets, $\mathcal{R}_1$ and $\mathcal{R}_2$ of the root set $\mathcal{R}$, such that $\mathcal{R}_i$ is the set of ancestors of $\mathcal{A}_i$. Since $\underline{\mathbf{X}}_{\mathcal{R}_1}$ and $\underline{\mathbf{X}}_{\mathcal{R}_2}$ are independent, $\underline{\mathbf{X}}_{\mathcal{A}_1}$ and $\underline{\mathbf{X}}_{\mathcal{A}_2}$ become independent.

A.4.2 Proof of Lemma 3

Suppose $\mathbf{Y}_h$ is a hidden node in a minimal LDIT with no outgoing edges and let $\{\mathbf{X}_1, ..., \mathbf{X}_s\}$ to be its parents. Since $\mathbf{Y}_h$ has no descendant, by marginalizing over $\mathbf{Y}_h$, we obtain s disjoint subtrees. This is a contradiction with the minimality assumption. Now suppose there exists a latent node, $\mathbf{Y}$, in a minimal LDIT with k parents $\underline{\mathbf{X}}_{\mathcal{K}} := \{\mathbf{X}_1, ..., \mathbf{X}_k\}$ and one child $\mathbf{X}_0$. From the definition of a generative model graph:

$$D(P_{\mathbf{X}_0 | \mathbf{Y}, \underline{\mathbf{X}}_{\mathcal{K}}} || P_{\mathbf{X}_0 | | \mathbf{Y}}) = 0, \quad D(P_{\mathbf{Y} | \underline{\mathbf{X}}_{\mathcal{K}}} || P_{\mathbf{Y}} || \underline{\mathbf{X}}_{\mathcal{K}}) = 0. \tag{A.23}$$

By the chain rule:

$$P_{X_{0,1}^t | \underline{\mathbf{X}}_{\mathcal{K}}} = \sum_{Y_1^{t-1}} P_{X_{0,1}^t | Y_1^{t-1}, \underline{\mathbf{X}}_{\mathcal{K}}} P_{Y_1^{t-1} | \underline{\mathbf{X}}_{\mathcal{K}}} \ . \tag{A.24}$$

From (A.23), (A.24), we have $D(P_{\mathbf{X}_0 | \underline{\mathbf{X}}_{\mathcal{K}}} || P_{\mathbf{X}_0 | | \underline{\mathbf{X}}_{\mathcal{K}}}) = 0$.

A.4.3 Proof of Lemma 4

It suffices to prove the lemma for $d = 2$, as the case for larger d, can be proved by induction. Consider the case where $d = 2$ ($\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$). Let $\mathcal{A} = \underline{\mathbf{X}}_{\mathcal{PA}(\mathbf{Y})} \setminus \{\mathbf{X}\}$ and $\mathcal{B} = \underline{\mathbf{X}}_{\mathcal{PA}(\mathbf{Z})} \setminus \{\mathbf{Y}\}$ to be the set of parents of $\mathbf{Y}$ and $\mathbf{Z}$ excluding $\mathbf{X}$ and $\mathbf{Y}$, respectively, as shown in Fig. A.1. First, we show that

$$D(P_{Z_t | Z_1^{t-1}, \mathbf{X}} || P_{Z_t | Z_1^{t-1}, X_1^{t-2}}) = 0, \ \forall t \leq n \ . \tag{A.25}$$

Note that if (A.25) holds, then by multiplying all terms for $t = 1, ..., n$, we obtain

$$P_{\mathbf{Z}|\mathbf{X}} = \prod_{t=1}^{n} P_{Z_t|Z_1^{t-1},X_1^{t-2}} \ ,$$

which proves our claim. By the chain rule for any t, we have

$$P_{Z_1^t|\mathbf{X}} = \sum_{\mathcal{B}_1^{t-1}Y_1^{t-1}} \ P_{Z_t|Z_1^{t-1},Y_1^{t-1},\mathcal{B}_1^{t-1},\mathbf{X}} \ P_{Z_1^{t-1}|\mathcal{B}_1^{t-1},Y_1^{t-1},\mathbf{X}} \ P_{\mathcal{B}_1^{t-1}|Y_1^{t-1},\mathbf{X}} \ P_{Y_1^{t-1}|\mathbf{X}}. \tag{A.26}$$

Theorem 1, Lemma 2, and the definition of generative model imply the following equalities

$$\begin{aligned} P_{\mathbf{Z}|\mathbf{Y},\mathcal{B},\mathbf{X},\mathcal{A}} = P_{\mathbf{Z}||\mathbf{Y},\mathcal{B}} = P_{\mathbf{Z}||\mathbf{Y},\mathcal{B},\mathbf{X},\mathcal{A}}, \\ P_{\mathcal{B}|\mathbf{Y},\mathbf{X},\mathcal{A}} = P_{\mathcal{B}} = P_{\mathcal{B}||\mathbf{X},\mathcal{A},\mathbf{Z}}, \\ P_{\mathbf{Y}|\mathbf{X},\mathcal{A}} = P_{\mathbf{Y}||\mathbf{X},\mathcal{A}} = P_{\mathbf{Y}||\mathbf{X},\mathcal{A},\mathbf{Z},\mathcal{B}}. \end{aligned} \tag{A.27}$$

The above equalities imply

$$\begin{aligned} P_{Z_t|Z_1^{t-1},Y_1^{t-1},\mathcal{B}_1^{t-1},\mathbf{X}} &= P_{Z_t|Z_1^{t-1},Y_1^{t-1},\mathcal{B}_1^{t-1},X_1^{t-2}}, \\ P_{Z_1^{t-1}|Y_1^{t-1},\mathcal{B}_1^{t-1},\mathbf{X}} &= P_{Z_1^{t-1}|Y_1^{t-1},\mathcal{B}_1^{t-1},X_1^{t-2}}, \\ P_{\mathcal{B}_1^{t-1}|Y_1^{t-1},\mathbf{X}} &= P_{\mathcal{B}_1^{t-1}|Y_1^{t-1},X_1^{t-2}}. \end{aligned} \tag{A.28}$$

Moreover, one can obtain the following equation using chain rule, Lemma 6, and equalities in (A.27)

$$\begin{aligned} P_{Y_1^{t-1}|\mathbf{X}} = \sum_{\mathcal{A}_1^{t-2}} \ P_{Y_1^{t-1}|\mathcal{A}_1^{t-2},\mathbf{X}} \ P_{\mathcal{A}_1^{t-2}|\mathbf{X}} \\ = \sum_{\mathcal{A}_1^{t-2}} \ P_{Y_1^{t-1}|\mathcal{A}_1^{t-2},X_1^{t-2}} \ P_{\mathcal{A}_1^{t-2}|X^{t-2}} = P_{Y_1^{t-1}|X_1^{t-2}}. \end{aligned} \tag{A.29}$$

Substituting (A.28)-(A.29) into the right-hand side of (A.26) proves our claim.

A.4.4 Proof of Lemma 5

It suffices to show

$$D(P_{\mathbf{Y}|\mathbf{W},\mathbf{X}}||P_{\mathbf{Y}||\mathbf{W}}) = 0. \tag{A.30}$$

Suppose the length of the path from $\mathbf{W}$ to $\mathbf{Y}$ is d. We will prove (A.30) by induction on d. For $d = 1$, define $\mathcal{A} := \underline{\mathbf{X}}_{\mathcal{P}_{\mathcal{A}}(\mathbf{Y})} \setminus \{\mathbf{W}\}$. In this case similar to the proof of Lemma 4 the following equalities hold

$$D(P_{\mathbf{Y}|\mathcal{A},\mathbf{W},\mathbf{X}}||P_{\mathbf{Y}||\mathcal{A},\mathbf{W}}) = 0, \quad D(P_{\mathcal{A}|\mathbf{W},\mathbf{X}}||P_{\mathcal{A}}) = 0. \tag{A.31}$$

From chain rule,

$$P_{Y_1^t|\mathbf{W},\mathbf{X}} = \sum_{\mathcal{A}} P_{Y_t|Y_1^{t-1},\mathcal{A},\mathbf{W},\mathbf{X}} P_{Y_1^{t-1}|\mathcal{A},\mathbf{W},\mathbf{X}} P_{\mathcal{A}|\mathbf{W},\mathbf{X}}.$$

Then by applying (A.31) to the above equation, we obtain (A.30).
Assume that equation (A.30) holds for paths of length $d < k$. In order to prove the case $d = k$, let $\mathbf{Z}$ to be the parent of $\mathbf{Y}$ on the path from $\mathbf{W}$ to $\mathbf{Y}$, and $\mathcal{B} := \underline{\mathbf{X}}_{\mathcal{PA}(\mathbf{Y})} \setminus \{\mathbf{Z}\}$. The path from $\mathbf{W}$ to $\mathbf{Z}$ is of length $k-1$ so by induction hypothesis we have

$$D(P_{\mathbf{Z}|\mathbf{W},\mathbf{X}}||P_{\mathbf{Z}||\mathbf{W}}) = 0. \tag{A.32}$$

Moreover, by the definition of generative model graph and Theorem 1:

$$D(P_{\mathbf{Y}|\mathcal{B},\mathbf{Z},\mathbf{W},\mathbf{X}}||P_{\mathbf{Y}||\mathcal{B},\mathbf{Z}}) = 0, \quad D(P_{\mathcal{B}|\mathbf{Z},\mathbf{W},\mathbf{X}}||P_{\mathcal{B}}) = 0. \tag{A.33}$$

Chain rule implies

$$P_{Y^t|\mathbf{W},\mathbf{X}} = \sum_{\mathcal{B},\mathbf{Z}} P_{Y^t|\mathcal{B},\mathbf{Z},\mathbf{W},\mathbf{X}} P_{\mathcal{B}|\mathbf{Z},\mathbf{W},\mathbf{X}} P_{\mathbf{Z}|\mathbf{W},\mathbf{X}}.$$

Applying (A.32) and (A.33) to the above equation proves the claim.

A.4.5 Proof of Lemma 6

Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be two disjoint subsets of the root set $\mathcal{R}$ in a minimal LDIT. Furthermore, assume $\mathcal{R}_1$ and $\mathcal{R}_2$ are root ancestors for nodes $\mathbf{X}$ and $\mathbf{Y}$, respectively. Denote all the nodes on the paths from $\mathcal{R}_1$ to $\mathbf{X}$ by $\mathcal{A}$. It is easy to check that if a node belongs to $\mathcal{A}$, so do all of its parents. Therefore, $\underline{\mathbf{X}}_{\mathcal{PA}(\mathbf{X})} \subseteq \mathcal{A}$, where $\mathcal{PA}(\mathbf{X})$ is the parent set of $\mathbf{X}$. Similarly, we denote all the nodes on the paths from $\mathcal{R}_2$ to $\mathbf{Y}$ by $\mathcal{B}$. By the definition of generative model, we obtain

$$P_{\mathbf{X},\mathbf{Y},\mathcal{R}_1,\mathcal{R}_2,\mathcal{A},\mathcal{B}} = P_{\mathcal{R}_1} P_{\mathcal{R}_2} \Psi_{\mathcal{A},\mathcal{R}_1} \Phi_{\mathcal{B},\mathcal{R}_2} P_{\mathbf{X}||\mathcal{PA}(\mathbf{X})} P_{\mathbf{Y}||\mathcal{PA}(\mathbf{Y})}, \tag{A.34}$$

where Ψ and Φ represent the terms including the causal conditioned distributions of all processes on the paths from $\mathcal{A}_1$ to $\mathbf{X}$, and from $\mathcal{A}_2$ to $\mathbf{Y}$, respectively. On the hand, from chain rule we obtain

$$P_{\mathbf{X},\mathbf{Y},\mathcal{R}_1,\mathcal{R}_2,\mathcal{A},\mathcal{B}} = P_{\mathcal{R}_1,\mathcal{R}_2}\ P_{\mathcal{A}|\mathcal{R}_1,\mathcal{R}_2} P_{\mathcal{B}|\mathcal{A},\mathcal{R}_1,\mathcal{R}_2}\ P_{\mathbf{X}|\mathcal{B},\mathcal{A},\mathcal{R}_1,\mathcal{R}_2}\ P_{\mathbf{Y}|\mathbf{X},\mathcal{B},\mathcal{A},\mathcal{R}_1,\mathcal{R}_2}.$$

The equivalence between (A.34) and (A.35), and the positivity assumption imply that $\mathbf{X}$ and $\mathbf{Y}$ are independent, whenever $\mathcal{PA}(\mathbf{X})$ and $\mathcal{PA}(\mathbf{Y})$ are independent. Continuing the same procedure, we can show $\mathbf{X}$ and $\mathbf{Y}$ are independent, if $\mathcal{R}_1$ and $\mathcal{R}_2$ are independent.

A.4.6 Proof of theorem 7

Proof consists of two parts: first we show that if $\mathcal{PA}_i$ is the parent set of $\mathbf{X}_i$ in a MDIG, then

$$D\left(P_{X_{i,t}|X_{i,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i\},1}} ||\ P_{X_{i,t}|X_{i,1}^{t-1},\underline{\mathbf{X}}^t_{\mathcal{PA}_i,1}}\right) = 0. \tag{A.35}$$

To do so, we use the definition of MDIG in Section 5.1.2. Let $\mathcal{R} = -\{i\} \setminus \mathcal{PA}_i$ to be the set of all nodes except i and its parents. Since there is no arrow in MDIG from $\mathcal{R}$ to $\mathbf{X}_i$, we have

$$\widetilde{I}(\mathbf{X}_r \to \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,r\}}) = 0, \ \forall r \in \mathcal{R}.$$

The positivity assumption together with the above equalities imply

$$D\left(P_{X_{i,t}|X_{i,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i\},1}} || \ P_{X_{i,t}|X_{i,1}^{t-1},\bigcap_{r\in\mathcal{R}}\underline{\mathbf{X}}^t_{-\{i,r\},1}}\right) = 0.$$

Noticing that $\bigcap_{r\in\mathcal{R}} \underline{\mathbf{X}}^t_{-\{i,r\},1} = \underline{\mathbf{X}}^t_{\mathcal{PA}_i,1}$, one can establish (A.35). Next we will show that if there is an arrow from $\mathbf{X}_j$ to $\mathbf{X}_i$ in a MDIG with polytree structure (e.g., $\widetilde{I}(\mathbf{X}_j \to \mathbf{X}_i || \underline{\mathbf{X}}_{-\{i,j\}}) > 0$), then

$$D\left(P_{X_{i,t}|X_{i,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i\},1}} || \ P_{X_{i,t}|X_{i,1}^{t-1},X_{j,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i,j\},1}}\right) = 0. \tag{A.36}$$

In words, given the past of $\mathbf{X}_j$ is enough for predicting the $X_{i,t}$. To prove (A.36), we use the fact that the graph is a polytree, and thus if there is an arrow from $\mathbf{X}_j$ to $\mathbf{X}_i$, there will be no arrow in the opposite direction, i.e., $\widetilde{I}(\mathbf{X}_i \to \mathbf{X}_j || \underline{\mathbf{X}}_{-\{i,j\}}) = 0$. Consequently,

$$D\left(P_{X_{j,t}|X_{j,1}^{t-1},\underline{\mathbf{X}}^t_{-\{j\},1}} || \ P_{X_{j,t}|X_{j,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i,j\},1}}\right) = 0.$$

On the other hand, the chain rule implies

$$P_{X_{i,t}|X_{i,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i\},1}} = P_{X_{j,t}|X_{j,1}^{t-1},\underline{\mathbf{X}}^t_{-\{j\},1}} \frac{P_{X_{i,t}|X_{i,1}^{t-1},X_{j,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i,j\},1}}}{P_{X_{j,t}|X_{j,1}^{t-1},X_{i,1}^{t-1},\underline{\mathbf{X}}^t_{-\{i,j\},1}}}$$

Combining the last two equations will imply (A.36).

A.4.7 Proof of theorem 8

First we prove that Γ_O suffices to learn $\overrightarrow{T}$ when $\mathcal{R} = \{r\}$. The proof is by induction on $|O|$. The base case, $|O| = 1$ is trivial, since by Definition 13, L must be empty and $\overrightarrow{T}$ is the single node. Suppose, a tree $\overrightarrow{T} = (V, \overrightarrow{E})$ can be recovered, given any learnable subset L such that $|O| \leq k-1$. For the case that $|O| = k$, let $v \in O$ and $B_v := \arg\min_{u\in O\setminus\{v\}} \gamma_r(v,u)$. Note that in a single root tree all the discrepancies must be non-negative. We claim that $\overrightarrow{T}$ is a star with a root in the center if and only if $B_v = O \setminus \{v\}$ for all $v \in O$. If $\overrightarrow{T}$ is a star, then clearly $B_v = O \setminus \{v\}$ for all $v \in O$. The other direction is proved by arguing that if $\overrightarrow{T}$ is not a star then there exists a directed path of length two and because L is learnable, then one can find a node on this path such that $B_v \neq O \setminus \{v\}$.

If there exists $v \in O$ such that $B_v \neq O \setminus \{v\}$, and $\min_{u\in O\setminus\{v\}} \gamma_r(v,u) = 0$, then all the nodes in B_v are the descendants of v. In this case by induction hypothesis, the subtree of $\overrightarrow{T}$ containing v and all its descendant, is recoverable by $B_v \cup \{v\}$ as well as the rest of the tree by $O \setminus B_v$. Similarly for the case $\min_{u\in O\setminus\{v\}} \gamma_r(v,u) > 0$.

We show that if $|\mathcal{R}| > 1$, learning $\overrightarrow{T}$ can be done by learning $|\mathcal{R}|$ single rooted trees, separately.

For $v \in O$, let M_v be a maximal subset of O containing v such that for every $u, w \in M_v$, $\gamma(u,w) \geq 0$. Clearly, if w belongs to M_v, so does all its descendants which are also in O.

Denote the minimal induced polytree of $\overrightarrow{T}$ containing M_v by $\overrightarrow{T}|_{M_v} = (V', \overrightarrow{E}')$. Note that from the maximality of M_v, $O \cap V' \subseteq M_v$. First we show that $V' \setminus M_v$ is a learnable subset in $\overrightarrow{T}|_{M_v}$, i.e., all nodes with out-degree at most one in $\overrightarrow{T}|_{M_v}$ belong to M_v. All leaves in $\overrightarrow{T}|_{M_v}$ belong to M_v otherwise they can be eliminated from $\overrightarrow{T}|_{M_v}$ and it is a contradiction with the minimality assumption on $\overrightarrow{T}|_{M_v}$. Let $u' \in V' \setminus M_v$ be a node with out-degree one in $\overrightarrow{T}|_{M_v}$. Since $O \cap V' \subseteq M_v$, then $u' \in L$. If the out-degree of u' is also one in $\overrightarrow{T}$, then we have a contradiction with the learnability assumption of L. Hence, there exists at least one descendent of u' in O which does not belong to $\overrightarrow{T}|_{M_v}$ in which case, we have a contradiction with the maximality of M_v.

Next, we claim that $\overrightarrow{T}|_{M_v}$ has only one root from the root set $\mathcal{R}$. Suppose $\overrightarrow{T}|_{M_v}$ has more than one root. Since a tree has no cycles, then there must exist at least two nodes with degree one (either a root with degree one or a leaf) with no common ancestor in $\overrightarrow{T}|_{M_v}$, which contradicts the definition of M_v.

The final step is to prove that these single rooted sub-trees can be merged uniquely. This can be done by observing that if two single rooted trees $\overrightarrow{T}_1 = (V_1, \overrightarrow{E}_1)$ and $\overrightarrow{T}_2 = (V_2, \overrightarrow{E}_2)$ have an intersection in $\overrightarrow{T}$, then that intersection is also a single rooted tree that can be learned from $O \cap V_1 \cap V_2$.

A.4.8 Proof of Theorem 9

To show this we prove that the directed measure in (5.6) is a discrepancy measure on T. First it is important to note that by Lemma 3 the set of hidden nodes is a learnable subset in a minimal LDIT. The rest of the proof verifies that directed measure in (5.6) satisfies the properties of a discrepancy measure introduced in Definition 12.

(1) From Definition 15, $\gamma(\mathbf{X}, \mathbf{X}) = 0$. Suppose $\mathbf{X}$ is an ancestor of $\mathbf{Y}$. By the sibling resemblance property, since $\mathbf{X}$ is the common ancestor of $\mathbf{X}$ and $\mathbf{Y}$ and $I(X_1; \mathbf{X}) > 0$, then $I(X_1; \mathbf{Y}) > 0$. In other word $\gamma(\mathbf{X}, \mathbf{Y}) = 0$.

(2) This property is also a consequence of the sibling resemblance property. Let $\mathbf{W}$ to be the common ancestor of $\mathbf{X}$ and $\mathbf{Y}$. If $\gamma(\mathbf{X}, \mathbf{W}) = d$, then by using Lemma 5 we obtain $I(X_1^d; \mathbf{Y}) = 0$. Which implies $\gamma(\mathbf{X}, \mathbf{Y}) \geq d$. On the other hand, since $I(X_1^{d+1}; \mathbf{W}) > 0$ and $I(\mathbf{Y}; \mathbf{W}) > 0$, by sibling resemblance property we obtain $I(X_{d+1}; \mathbf{Y}|X_1^d) > 0$, which implies $\gamma(\mathbf{X}, \mathbf{Y}) = \gamma(\mathbf{X}, \mathbf{W}) = d$.

(3) This is shown by proving that for a given path $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ in a minimal LDIT, if $\gamma(\mathbf{Y}, \mathbf{X}) = l$ and $\gamma(\mathbf{Z}, \mathbf{Y}) = d$ then $\gamma(\mathbf{Z}, \mathbf{X}) > \max\{l, d\}$.
First we prove $\gamma(\mathbf{Z}, \mathbf{X}) > d$. It suffices to show

$$I(Z_{d+1}; \mathbf{X}|Z_1^d) = 0. \tag{A.37}$$

Using the chain rule we obtain

$$P_{Z_{d+1}|Z_1^d,\mathbf{X}} = \sum_{Y_1} P_{Z_{d+1}|Z_1^d,Y_1,\mathbf{X}} P_{Z_1^d|Y_1,\mathbf{X}} \frac{P_{Y_1|\mathbf{X}}}{P_{Z_1^d|\mathbf{X}}}. \tag{A.38}$$

Since $\gamma(\mathbf{Z}, \mathbf{Y}) = d$, $\mathbf{Y}$ is an ancestor of $\mathbf{Z}$, and by using the same argument as in the proof of Lemma 4, we obtain

$$D(P_{Z_1^d|\mathbf{Y},\mathbf{X}} || P_{Z_1^d}) = 0, \quad D(P_{Y_1|\mathbf{X}} || P_{Y_1}) = 0, \tag{A.39}$$

$$D(P_{Z_{d+1}|Z_1^d,Y_1,\mathbf{X}} || P_{Z_{d+1}|Z_1^d,Y_1}) = 0. \tag{A.40}$$

Finally, the claim follows by substituting (A.39) and (A.40) into the right-hand side of (A.38). The statement $\gamma(\mathbf{Z},\mathbf{X}) > l$ may be proven by showing $I(Z_{l+1};\mathbf{X}|Z_1^l) = 0$.

$$P_{Z_{l+1}|Z_1^l,\mathbf{X}} = \sum_{Y_1^l} P_{Z_{l+1}|Z_1^l,Y_1^l,\mathbf{X}} P_{Z_1^l|Y_1^l,\mathbf{X}} \frac{P_{Y_1^l|\mathbf{X}}}{\sum_{Y_1'^{l-1}} P_{Z_1^l|Y_1'^{l-1},\mathbf{X}} P_{Y_1'^{l-1}|\mathbf{X}}},$$

since $\gamma(\mathbf{Y},\mathbf{X}) = l$, and using the same argument as above, one can prove the claim.
(4) This property is a direct consequence of Lemma 6 and Definition 15.

A.4.9 Proof of Lemma 7

First we prove the following Lemma which will be used in the Proof of Lemma 7.

Lemma 10. *Let $1 \leq a/x$ and $x \geq 0$. For any $0 < \lambda < 1$, $x \log \frac{a}{x}$ is bounded from above by $\frac{a^\lambda x^{1-\lambda}}{\lambda}$.*

Proof. Since $1 \leq a/x$, then $\log \left(\frac{a}{x}\right)^\lambda \leq \left(\frac{a}{x}\right)^\lambda$, for any $0 < \lambda < 1$. Hence, $\lambda x \log \frac{a}{x} \leq a^\lambda x^{1-\lambda}$. □

Proof of Lemma: Using the McDiarmid's inequality [170] and the union bound for the empirical estimator (5.7), we obtain

$$\mathbb{P}\left(\max_{(\mathbf{x}_1,\mathbf{x}_2)\in|\mathbf{X}|^{d_1+d_2}} |P_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2) - \widehat{P}_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2)| \geq \delta \right) \;\leq\; 2|\mathcal{X}|^{d_1+d_2} e^{-2N\delta^2} \;\leq\; 2|\mathcal{X}|^{2n} e^{-2N\delta^2}. \quad \text{(A.41)}$$

For simplicity, denote $(\mathbf{X}_1,\mathbf{X}_2)$ by $\mathbf{Z}$. From $||P_{\mathbf{Z}} - \widehat{P}_{\mathbf{Z}}||_1 \;\leq\; |\mathcal{X}|^{2n} \max_{\mathbf{Z}} |P_{\mathbf{Z}}(\mathbf{Z}) - \widehat{P}_{\mathbf{Z}}(\mathbf{Z})|$ and (A.41), we obtain

$$\mathbb{P}\left(||P_{\mathbf{Z}} - \widehat{P}_{\mathbf{Z}}||_1 \geq |\mathcal{X}|^{2n}\delta \right) \;\leq\; 2|\mathcal{X}|^{2n} e^{-2N\delta^2}. \quad \text{(A.42)}$$

Using an ℓ_1-norm bound on entropy [25], if $||P_{\mathbf{Z}} - \widehat{P}_{\mathbf{Z}}||_1 < 0.5$, then

$$|H(\mathbf{Z}) - \widehat{H}(\mathbf{Z})| \;\leq\; ||P_{\mathbf{Z}} - \widehat{P}_{\mathbf{Z}}||_1 \log \frac{|\mathcal{X}|^{d_1+d_2}}{||P_{\mathbf{Z}} - \widehat{P}_{\mathbf{Z}}||_1}.$$

Applying Lemma 10, we have

$$|H(\mathbf{Z}) - \widehat{H}(\mathbf{Z})| \;\leq\; \frac{1}{\lambda} ||P_{\mathbf{Z}} - \widehat{P}_{\mathbf{Z}}||_1^{1-\lambda} |\mathcal{X}|^{\lambda(d_1+d_2)}. \quad \text{(A.43)}$$

Therefore,

$$\mathbb{P}\left(|H(\mathbf{Z}) - \widehat{H}(\mathbf{Z})| \geq \epsilon \right) \;\leq\; \mathbb{P}\left(||P_{\mathbf{Z}} - \widehat{P}_{\mathbf{Z}}||_1^{1-\lambda} \geq \frac{\lambda\epsilon}{|\mathcal{X}|^{\lambda(d_1+d_2)}} \right).$$

From (A.42), we have

$$\mathbb{P}\left(|H(\mathbf{Z}) - \widehat{H}(\mathbf{Z})| \geq \epsilon \right) \;\leq\; 2|\mathcal{X}|^{2n} \exp\left(-2N \left(\frac{\lambda\epsilon}{|\mathcal{X}|^{2n}} \right)^{\frac{2}{1-\lambda}} \right). \quad \text{(A.44)}$$

Using the definition of mutual information, $I(\mathbf{X}_1;\mathbf{X}_2) = H(\mathbf{X}_1) + H(\mathbf{X}_2) - H(\mathbf{X}_1,\mathbf{X}_2)$, we obtain

$$\mathbb{P}\left(|I(\mathbf{X}_1;\mathbf{X}_2) - \widehat{I}(\mathbf{X}_1;\mathbf{X}_2)| \geq \epsilon\right) \;\leq\; \mathbb{P}\left(|H(\mathbf{X}_1) - \widehat{H}(\mathbf{X}_1)| \geq \epsilon/3\right) + \mathbb{P}\left(|H(\mathbf{X}_2) - \widehat{H}(\mathbf{X}_2)| \geq \epsilon/3\right) + \mathbb{P}\left(|H(\mathbf{X}_1,\mathbf{X}_2) - \widehat{H}(\mathbf{X}_1,\mathbf{X}_2)| \geq \epsilon/3\right).$$

Applying the upper bound in (A.44) to the above inequality will conclude the lemma. It only remains to choose λ to minimize the right hand side of (A.44). We choose $\lambda = 1/\log(\frac{3|\mathcal{X}|^{2n}}{\epsilon})$.

A.5 Proofs of Chapter 6

A.5.1 Proof of Proposition 5

The set of equation in (6.4), can be written in a matrix form as follows

$$\widetilde{A}\begin{bmatrix} \underline{\omega}_{Z,t} \\ \vdots \\ \underline{\omega}_{Z,t-l+1} \end{bmatrix} = \mathbf{C}\begin{bmatrix} \underline{X}_t \\ \vdots \\ \underline{X}_{t-l} \end{bmatrix} + \begin{bmatrix} \underline{N}_{Z,t} \\ \vdots \\ \underline{N}_{Z,t-l+1} \end{bmatrix}, \tag{A.45}$$

where $\widetilde{A} = diag(\tilde{A}_0, ..., \tilde{A}_{l-1})$, and $\mathbf{C}$ a block matrix with C_r^s as its (s,r)th block for $s = 0, ..., l-1$ and $k = 0, ..., l$. Since $\underline{\mathbf{N}}_Z$ and $\underline{\mathbf{X}}$ are orthogonal, we imply

$$||\widetilde{A}\Gamma_{\omega_Z}(l-1)\widetilde{A}^T||_2 \;\geq\; ||\mathbf{C}\Gamma_X(l)\mathbf{C}^T||_2. \tag{A.46}$$

Using (A.46) and the relationship between ℓ_2 and ℓ_1 norms of a matrix, we obtain

$$\lambda_{max}\left(\Gamma_{\omega_Z}(l-1)\right)||\widetilde{A}||_2^2 \;\geq\; \lambda_{min}\left(\Gamma_X(l)\right)||\mathbf{C}||_1^2/(nl) \tag{A.47}$$

where $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ denote the minimum and maximum eigenvalues of a given matrix, respectively. Since $\Gamma_X(l)$ and $\Gamma_{\omega_Z}(l-1)$ are block-Toeplitz matrices, their eigenvalues can be bounded as follows [171]:

$$L := \inf_{\Omega\in[0,2\pi]} \lambda_{min}(\mathcal{F}(\gamma_X)) \;\leq\; \lambda_{min}(\Gamma_X(l)), \tag{A.48}$$

$$M := \sup_{\Omega\in[0,2\pi]} \lambda_{max}(\mathcal{F}(\gamma_{\omega_Z})) \;\geq\; \lambda_{max}(\Gamma_{\omega_Z}(l-1)), \tag{A.49}$$

where j denotes $\sqrt{-1}$. Using (A.47)-(A.49) and the fact that $\widetilde{A}$ is diagonal and $||A_{22}||_2 < 1$, we obtain

$$\sqrt{nl\frac{M}{L}}||A_{12}||_2 \;\geq\; \sqrt{nl\frac{M}{L}}||A_{12}||_2 \max_{0\leq k\leq l-1}||A_{22}||_2^k \;\geq\; ||\mathbf{C}||_1. \tag{A.50}$$

From (6.5), we have $B_k^* - A_k^* = \sum_{s=0}^{l-1} C_k^s$, where the right hand side can be obtain by summing up the appropriate columns of matrix $\mathbf{C}$. This implies that $\max_{0\leq k\leq l}||B_k^* - A_k^*||_1 \leq ||\mathbf{C}||_1$. Combining this inequality and the bound in (A.50) concludes the result.

A.5.2 Proof of Proposition 6

The spectral density of matrix $\gamma_X(h)$ can be computed as follows:

$$\mathcal{F}(\gamma_X) = \sigma_X^2 F_X(\Omega) F_X(\Omega)^H + \sigma_Z^2 F_Z(\Omega) F_Z(\Omega)^H \tag{A.51}$$

where H denotes Hermitian of a matrix and

$$F_X(\Omega) := [e^{j\Omega} I_{n\times n} - A_{11} - \sum_{k=0}^{l-1} A_k^* e^{-kj\Omega}]^{-1},$$
$$F_Z(\Omega) := F_X(\Omega) \left(A_{12} \sum_{k=0}^{l-1} A_{22}^k e^{-kj\Omega} \right).$$

We define the function $\psi_{\frac{\sigma_X}{\sigma_Z}}(\Omega, v) := \underline{v}^T \mathcal{F}(\gamma_X)\underline{v}/\sigma_Z^2$ where $\underline{v}$ is a unit vector. Suppose that $(\Omega^*, \underline{v}^*)$ minimizes the function $\psi_{\frac{\sigma_X}{\sigma_Z}}(.)$. By the definition of L and M, the ratio M/L is equal to $1/\psi_{\frac{\sigma_X}{\sigma_Z}}(\Omega^*, \underline{v}^*)$. Now if we decrease $\frac{\sigma_X}{\sigma_Z}$ to $\frac{\sigma'_X}{\sigma'_Z}$, then we have: $\psi_{\frac{\sigma'_X}{\sigma'_Z}}(\Omega^*, \underline{v}^*) < \psi_{\frac{\sigma_X}{\sigma_Z}}(\Omega^*, \underline{v}^*)$. Moreover, for the optimal solution $(\Omega'^*, \underline{v}'^*)$ of $\psi_{\frac{\sigma'_X}{\sigma'_Z}}(.)$, we know that: $\psi_{\frac{\sigma'_X}{\sigma'_Z}}(\Omega'^*, \underline{v}'^*) \leq \psi_{\frac{\sigma'_X}{\sigma'_Z}}(\Omega^*, \underline{v}^*)$. Thus, we can conclude that: $1/\psi_{\frac{\sigma'_X}{\sigma'_Z}}(\Omega'^*, \underline{v}'^*) > 1/\psi_{\frac{\sigma_X}{\sigma_Z}}(\Omega^*, \underline{v}^*)$.

A.5.3 Proof of Theorem 11

First, we show such G has minimum number of latent nodes. We do this by means of contradiction. But first observe that since the latent subnetwork of G is a directed tree, we can assign a non-negative number l_h to latent node h that represents the length of longest directed path from h to its latent descendants. Clearly, all such descendants are leaves which we denote them by $\tilde{L}_h$. For instance, if the latent subnetwork of G is $a \to b \to c$, then $l_a = 2$ and $\tilde{L}_a = \{c\}$.

Suppose that G contains m latent nodes $\{h_1, ..., h_m\}$ and there exists another network G_1 (not necessary with tree-structure induced latent subgraph), with $m_1 < m$ number of latent nodes that it is also consistent with the same linear measurements as G. Due to assumption (i), there is at least m distinct observed nodes that have out-going edges to the latent subnetwork. More precisely, each h_i has at least a unique observed node as its parent. We denote a unique observed parent of node h_i by o_i.

Because $m_1 < m$, there exists at least one observed node in $\bar{O} := \{o_1, ..., o_m\}$ that has shared its latent children with some other latent nodes in G_1. Among all such observed nodes, let o_{i^*} to be the one[1] that its corresponding latent node in G, (h_{i^*}) has maximum $l_{h_{i^*}}$. Furthermore, let $\tilde{I}_{i^*} \subset \{1, ..., m\} \setminus \{i^*\}$ to be the index-set of those observed nodes that o_{i^*} has shared a latent child with them in G_1.

[1] If there are several such observed node, let o_{i^*} to be one of them.

By the choice of o_{i^*}, we know that $l_{h_j} \leq l_{h_{i^*}}$ for all $j \in \tilde{I}_{i^*}$ and if for some $1 \leq k \leq m$, $l_{h_k} > l_{h_{i^*}}$, then o_k has not shared its latent child in G_1 with any other observed nodes in $\bar{O}$. Moreover, there should be at least a latent node h_{j^*} where $j^* \in \tilde{I}_{i^*}$ such that $l_{h_{j^*}} = l_{h_{i^*}}$. Otherwise, G_1 will not be consistent with the linear measurements of G. Let $\tilde{I}_{**} := \{j : l_{h_j} = l_{h_{i^*}}\} \cap \tilde{I}_{i^*}$. Because, o_{i^*} shares its latent children with $\cup_{j \in \tilde{I}_{**}} o_j$ in G_1 and the fact that both G and G_1 consistent with the same linear measurements, then the following holds in graph G,

$$\mathcal{C}^O_{\tilde{L}_{h_{i^*}}}(G) \subseteq \cup_{j \in \tilde{I}_{**}} \mathcal{C}^O_{\tilde{L}_{h_j}}(G),$$

where $\mathcal{C}^O_{\tilde{L}_{h_j}(G)}$ indicates the set of observed children of the set $\tilde{L}_{h_j}$. This indeed contradicts with assumption (ii).

A.5.4 Proof of Theorem 12

First, we require the following definition. For a network G with corresponding latent sub-network that is a tree, we define $U_k(G) := \{h \in G : \ l_h = k\}$. To prove the equivalency, suppose there exists another network G_2 such that its latent sub-network is a tree and has minimum number of latent nodes. Let $\{h_1, ..., h_m\}$ to denote the latent nodes in G. Since G satisfies Assumption (i), for every latent node h_i there exists a unique observed node o_i such that $o_i \in \mathcal{P}^O_{h_i}(G)$ and $o_j \notin \mathcal{P}^O_{h_i}(G)$ for all $j \neq i$.

Since both G and G_2 are consistent with the same linear measurement, it is easy to observe that if $h_i \in U_k(G)$, then o_i must have at least a latent child in G_2, say h'_i, such that $l_{h_i} = l_{h'_i}$. Note that l_{h_i} is computed in G and $l_{h'_i}$ in G_2. Moreover, we must have:

$$\mathcal{C}^O_{\tilde{L}_{h_i}}(G) = \bigcup_{h' \in H'(o_i) \cap U_{l_{h_i}}(G_2)} \mathcal{C}^O_{\tilde{L}_{h'}}(G_2),$$

where $H'(o_i)$ denotes the set of latent nodes in G_2 that have o_i as their observed parent. In other words, observed nodes that can be reached by a directed path of length $l_{h_i} + 2$ from o_i should be the same in both graph G and G_2. This results plus the fact that G satisfies Assumption (ii), imply:
I) For every $h_i \in U_k(G)$, there exists a unique latent node $h'_i \in U_k(G_2)$, such that $o_i \in \mathcal{P}^O_{h'_i}(G_2)$ and $o_j \notin \mathcal{P}^O_{h'_i}(G_2)$ for all $j \neq i$, and

$$\mathcal{C}^O_{\tilde{L}_h}(G) = \mathcal{C}^O_{\tilde{L}_{h'_i}}(G_2).$$

Using I) and knowing that both G and G_2 have the same number of latent nodes, we obtain:
II) $|U_k(G)| = |U_k(G_2)|$, for all k.
Using I) and II), we can define a bijection ϕ between the latent subnetworks of G and G_2 as follows $\phi(h_i) = h'_i$. Using this bijection and Assumption (ii) of G conclude that if $h \in U_k(G)$ is the common parent of $\{h_{j_1}, ..., h_{j_s}\} \subseteq U_{k-1}(G)$, then $\phi(h) \in U_k(G_2)$ should be the common parent of $\{\phi(h_{j_1}), ..., \phi(h_{j_s})\} \subseteq U_{k-1}(G_2)$ and the proof is complete.

A.5.5 Proof of Lemma 8

Suppose that o_i is the unique observed node of a latent node h_i. Then, for any o_j such that $l_i = l_j$, if h_i is not a child of o_j, then from assumption ii we have $R_j \not\subseteq R_i$. If h_i is a child of o_j, since we know that $l_i = l_j$, then $M_i \subseteq M_j$ and $R_i = R_j$.

Now, suppose that the observed node o_i satisfies conditions but it is not unique parent of any latent node. Let h_i and h_i' be children of o_i. At least one of them, say node h_i, can reach an observed node by a path of length $l_i - 1$. If h_i' has the same property, then consider the unique observed parent of h_i', say node o_j. Based on Assumption (ii), we have $R_j \subseteq R_i$, which is in contradiction with the assumption that node o_i satisfies conditions of Lemma. Moreover, if h_i' does not have a path to observed node with a length of $l_i - 1$, then for any observed parent of h_i, one of the conditions in the Lemma is not satisfied. Thus, the proof is complete.

A.5.6 Proof of Proposition 7

Notice that the first loop in Algorithm 5 uses the result of Lemma 8 and finds all the latent nodes and their corresponding unique observed parents. The next loop uses the fact that the latent sub-network is a tree and also it satisfies Assumption 4. Hence, if there exist two latent nodes h and h', one with depth l and the other one with depth $l + 1$, such that $R_h \subseteq R_{h'}$, then h' must be the parent of h in the latent sub-network.

Moreover, since each latent node has a unique observed parent, using A_1^*, Algorithm 5 can identify all the observed children of a latent node. Finally, the last loop in this algorithm locates the rest of observed nodes as the input of the right latent nodes. The algorithm does it by using the fact that if an observed node i shares a latent child with another observed node $j \in U$, then $M_j \subseteq M_i$. Clearly, if the true unobserved network satisfies Assumption 4, the output of this algorithm will have a latent sub-network that is a tree and consistent with the linear measurement. Thus, by the result of Theorem 11, it will be the same as the true unobserved network up to some permutations in $Supp(A_{21})$.

A.5.7 Proof of Theorem 13

Consider the instance of the problem where $A_{22} = 0_{m\times m}$. Without loss of generality, we can assume that entries of A_{12} and A_{21} are just zero or one. Thus, we need to find $[A_{12}]_{n\times k}$ and $[A_{21}]_{k\times n}$ such that $Supp(A_{12}A_{21}) = Supp(A_1^*)$ and k is minimum. We will show that the set basis problem [172] can be reduced to the decision version of finding the minimal unobserved network which we call it the latent recovery problem. But before that, we define the set basis problem:

The Set Basis Problem [172]: given a collection $\mathcal{C}$ of subsets of a finite set $U = \{1, \cdots, n\}$ and an integer k, decide whether or not there is a collection $\mathcal{B} \subseteq 2^U$ of at most k sets such that for every set $C \in \mathcal{C}$, there exists a collection $\mathcal{B}_\mathcal{C} \subseteq \mathcal{B}$ where $\bigcup_{B\in\mathcal{B}_C} B = C$.

Any instance of the basis problem can be reduced to an instance of latent recovery problem. To do so, we encode any set C in collection $\mathcal{C}$ to a row of $A_1^* = A_{12}A_{21}$ where i-th entry is equal to one if $i \in C$, and otherwise zero. It is easy to verify that the rows of matrix A_{21} correspond to sets in collection $\mathcal{B}$ if there exist a solution for the basis problem. Since the basis problem is NP-complete, we can conclude that finding the minimal unobserved network is NP-hard.

A.5.8 Proof of Theorem 14

Consider a minimal unobserved network G_{min}. Pick any latent node i' which its in-degree or out-degree is greater than one. Let $V_{i'}^-$ and $V_{i'}^+$ be the sets of nodes that are going to and incoming from node i', respectively. We omit the node i' and create $|V_{i'}^-| \times |V_{i'}^+|$ latent nodes $\{i'_{j'k'}|j' \in V_{i'}^-, k' \in V_{i'}^+\}$. We also add a direct link from node $j' \in V_{i'}^-$ to $i'_{j'k'}$ and from $i'_{j'k'}$ to $k' \in V_{i'}^+$ in order to be consistent with measurements. We continue this process until there is no latent node with in-degree or out-degree greater than one. Since there exists at most one path with length k from any observed node to another observed node, the resulted graph is exactly equal to graph G_0. Hence we can construct the minimal graph G_{min} just by reversing the process of generating latent nodes from G_{min} to merging latent nodes from G_0. But the NM algorithm consider all the sequence of merging operations. Thus, G_{min} would be in the set $\mathcal{G}_{out}$ and the proof is complete.

A.6 Proofs of Chapter 7

A.6.1 Proof of Theorem 15

In order to complete the proof, we need the following technical lemmas. When $d(\cdot,\cdot)$ is the Euclidean distance, we denote the Wasserstein metric by $W_E(\cdot,\cdot)$.

Lemma 11. *For real-valued random variables, we have*

$$|\mathbb{E}_{\nu_1}[x] - \mathbb{E}_{\nu_2}[y]| \;\leq\; W_E(\nu_1,\nu_2) \;\leq\; \sqrt{\mathbb{E}_{\nu_1}[x^2] + \mathbb{E}_{\nu_2}[y^2] - 2\mathbb{E}_{\pi}[xy]}, \tag{A.52}$$

where π is any joint distribution of x and y such that its marginals are ν_1 and ν_2.

Proof. The lower bound is due to the dual representation of the Wasserstein metric and the fact that $f(x) = x$ is Lipschitz.
For the upper bound, we use the Jensen's inequality, that is

$$W_d(\nu_1,\nu_2) \;\leq\; \inf_{\pi} \left(\mathbb{E}_{\pi}[d^p(x,y)]\right)^{1/p}, \tag{A.53}$$

for $p \geq 1$. For $p = 2$, we use the monotonicity of $\sqrt{x}$, and the fact that the space of probability measures is complete and obtain the result. □

Here, we consider a more general form than a simple linear model. Consider a network of variables in which every variable X_i functionally depends on a subset of other variables $\underline{X}_{Fp_i}$ (the parent set of node i) as follows,

$$X_i = F_i(\underline{X}_{Fp_i}) + G_i(\underline{X}_{Fp_i})W_i, \quad \forall i, \tag{A.54}$$

where F_i, G_i are arbitrary functions such that $G_i \neq 0$. W_is denote exogenous noises with mean zero and variance σ_i^2 and have no influence on each other, i.e., for any $\mathcal{K} \subseteq -\{W_i, W_j\}$, $c^{\mathcal{K}}_{W_i,W_j} = 0$.

Lemma 12. *For a system described by (A.54), the influence of node j on its child i given the rest of i's parents $Fp_i \setminus \{j\}$ under Euclidean metric, is bounded as follows*

$$\sup_{\substack{\overline{x}_{Fp_i}=\overline{y}_{Fp_i} \\ off\ j}} \Big| \frac{F_i(\overline{x}_{Fp_i}) - F_i(\overline{y}_{Fp_i})}{x-y} \Big| \leq \quad c_{i,j}^{Fp_i\setminus\{j\}} \quad \leq$$
$$\sup_{\substack{\overline{x}_{Fp_i}=\overline{y}_{Fp_i} \\ off\ j}} \left[\left(\frac{F_i(\overline{x}_{Fp_i}) - F_i(\overline{y}_{Fp_i})}{x-y} \right)^2 + \left(\frac{G_i(\overline{x}_{Fp_i}) - G_i(\overline{y}_{Fp_i})}{x-y} \sigma_i \right)^2 \right]^{1/2}. \tag{A.55}$$

where the suprimum is taking over all realizations of $\underline{X}_{-\{i\}}$ that are only different at X_j.

Proof. Using the lower bound in Lemma 11 and the fact that W_is have zero mean, we obtain the lower bound in (A.55).
To obtain the upper bound, we again use the result of Lemma 11, with the following joint distribution $\pi(X_i, Y_i)$,

$$\frac{1}{|G_i(\overline{x}_{Fp_i})|} f_{W_i}\left(\Theta_{\overline{x}_{Fp_i}}(X_i)\right) \mathbb{I}_{\{\Theta_{\overline{x}_{Fp_i}}(X_i)=\Theta_{\overline{y}_{Fp_i}}(Y_i)\}},$$

where

$$\Theta_{\overline{x}_{Fp_i}}(X_i) := \frac{X_i - F_i(\overline{x}_{Fp_i})}{G_i(\overline{x}_{Fp_i})},$$

and f_{W_i} denotes the probability density function of W_i and $\mathbb{I}$ denotes the indicator function. Using this joint distribution, we obtain the upper bound in (A.55). □

Applying the above result to a linear system in which $F_i(\overline{y}_{Fp_i}) = (\mathbf{A}\overline{x})_i$ and $G_i(\overline{x}_{Fp_i}) = 1$, we obtain that $c_{i,j}^{Fp_i\setminus\{j\}} = |A_{i,j}|$.

A.6.2 Proof of Lemma 9

• $c_{i,j}^{\mathcal{K}} \geq 0$ since Wasserstein is a metric. If $c_{i,j}^{\mathcal{K}} = 0$, we have $W_d\left(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|y_j, \underline{x}_{\mathcal{K}})\right) = 0$, for all realizations x_j, y_j and $\underline{x}_{\mathcal{K}}$. Using the fact that Wasserstein is a metric on the space of probability measures, the above equality, and total probability law, we obtain

$$P(X_i|\underline{x}_{\mathcal{K}}) = \sum_{x_j} P(X_i|x_j, \underline{x}_{\mathcal{K}}) P(x_j|\underline{x}_{\mathcal{K}}) = P(X_i|y_j, \underline{x}_{\mathcal{K}}) \sum_{x_j} P(x_j|\underline{x}_{\mathcal{K}}) = P(X_i|y_j, \underline{x}_{\mathcal{K}}).$$

The above equality holds for all y_j and $\underline{x}_{\mathcal{K}}$. This implies $X_i \perp\!\!\!\perp X_j|\underline{X}_{\mathcal{K}}$.
• We show this by an example. Let $X = U_{[0,1]}$ to be uniformly distributed between zero and one, and

$$Y = \begin{cases} V_{[0,1]} & \text{if } X \in \mathcal{A}, \\ U_{[0,1]} & \text{otherwise,} \end{cases}$$

where $\mathcal{A} = \{\frac{i}{i+1} : i \in \mathbb{N}\}$, and $V_{[0,1]}$ is a random variable independent of U that is distributed non-uniformly over $[0,1]$. In this case, we have

$$0 \;<\; \frac{W_d(P(Y|X=1/2), P(Y|X=\sqrt{2}))}{d(1/2,\sqrt{2})} \;\leq c_{y,x}.$$

On the other hand, it is easy to see that Y has a uniform distribution over $[0,1]$ almost surely. Furthermore, for two measurable sets C and B in the σ-algebra, we have

$$\begin{aligned}
&P(X \in C|Y \in B) = \frac{P(Y \in B|X \in C)\; P(X \in C)}{P(Y \in B)} = \\
&\frac{P(Y \in B|X \in C \cap \mathcal{A})\; P(X \in C \cap \mathcal{A}) \;+\; P(Y \in B|X \in\; C \setminus \mathcal{A})\; P(X \in C \setminus \mathcal{A})}{P(Y \in B)} \\
&= \frac{P(Y \in B|X \in C \setminus \mathcal{A})\; P(X \in C \setminus \mathcal{A})}{P(Y \in B)} = P(X \in C \setminus \mathcal{A}).
\end{aligned}$$

The last equality uses the fact that $P(Y \in B) = P(Y \in B|X \notin \mathcal{A}) = P(Y \in B|X \in C \setminus \mathcal{A})$. Thus, changing the value of Y will not affect the conditional distribution of X given Y, i.e., $c_{x,y} = 0$.

• If $c^{\mathcal{K}}_{i,\{j,k\}} = 0$, $W_d(P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}), P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}})) = 0$, for all realization $x_j, y_j, x_k, y_k, \underline{x}_{\mathcal{K}}$. By the total probability law, we obtain

$$\begin{aligned}
P(X_i|x_k, \underline{x}_{\mathcal{K}}) &= \sum_{x_j} P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}) P(x_j|x_k, \underline{x}_{\mathcal{K}}) \\
&= P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}}) \sum_{x_j} P(x_j|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}}).
\end{aligned}$$

This implies that $P(X_i|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_k, \underline{x}_{\mathcal{K}})$. Hence, $c^{\mathcal{K}}_{i,k} = 0$. Similarly, we can prove that $c^{\mathcal{K}}_{i,j} = 0$.

• Suppose $c^{\mathcal{K}}_{i,\{j,k\}} = 0$, then from the previous proof, we have $P(X_i|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_k, y_j, \underline{x}_{\mathcal{K}})$, for all realizations $y_j, x_k, y_k, \underline{x}_{\mathcal{K}}$. Thus, $P(X_i|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_k, x_j, \underline{x}_{\mathcal{K}})$ This is equivalent to say $c^{\mathcal{K}\cup\{j\}}_{i,k} = 0$. The other part can be shown similarly.

• If $c^{\mathcal{K}}_{i,j} = c_{i,\mathcal{K}} = 0$, then from $c^{\mathcal{K}}_{i,j} = 0$ and total probability law, we obtain that

$$W_d(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|\underline{x}_{\mathcal{K}})) = 0. \tag{A.56}$$

On the other hand, using the triangle inequality of the Wasserstein metric, we have

$$\begin{aligned}
W_d(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|y_j, \underline{y}_{\mathcal{K}})) \leq &W_d(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|\underline{x}_{\mathcal{K}})) + W_d(P(X_i|\underline{x}_{\mathcal{K}}), P(X_i|\underline{y}_{\mathcal{K}})) \\
&+ W_d(P(X_i|\underline{y}_{\mathcal{K}}), P(X_i|y_j, \underline{y}_{\mathcal{K}})).
\end{aligned}$$

The first and third expressions on the right hand side are zero due to (A.56) and the second expression is zero due to $c_{i,\mathcal{K}} = 0$.

• If $c^{\mathcal{K}\cup\{k\}}_{i,j} = 0$,

$$W_d(P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}), P(X_i|y_j, x_k, \underline{x}_{\mathcal{K}})) = 0.$$

This implies that $P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}) = P(X_i|x_k, \underline{x}_{\mathcal{K}})$ for all realizations x_j, x_k, and $\underline{x}_{\mathcal{K}}$. Similarly, because of $c_{i,k}^{\mathcal{K}\cup\{j\}} = 0$, we have $P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}) = P(X_i|x_j, \underline{x}_{\mathcal{K}})$ for all realizations x_j, x_k, and $\underline{x}_{\mathcal{K}}$. Hence, for all realizations, we have

$$P(X_i|x_j, \underline{x}_{\mathcal{K}}) = P(X_i|x_k, \underline{x}_{\mathcal{K}}).$$

This result and the total probability law will establish the result.

A.6.3 Proof of Theorem 16

Since the influence structure of this network is a DAG, there exists an ordering of the variables such that for every node i, all its parents have indices less that i. Without loss of generality suppose that $\{X_1, ..., X_m\}$ is that ordering. Furthermore, using the chain rule, we have

$$P(\underline{X}) = \prod_{i=1}^{m} P(X_i|\underline{X}_{\{<i\}}), \tag{A.57}$$

where $\underline{X}_{\{<i\}}$ denotes all the variables with indices less than i. Due to the nature of this ordering, all the nodes in $\{< i\}$ that do not belong to Pa_i are non-descendants of node i. Hence, by the definition of ID, they have zero influence on X_i given the parents of i and because of the first property in Lemma 9, they can be dropped from the conditioning in (A.57).
The global Markov property is a direct consequence of Lemma 9 and Theorem 3.27 in [109].