On Learning Time Series DAGs: A Frequency Domain Approach

Aramayis Dallakyan StataCorp, College Station, TX 77845, USA

August 30, 2023

Abstract

The fields of time series and graphical models emerged and advanced separately. Previous work on the structure learning of continuous and real-valued time series utilizes the time domain, with a focus on either structural autoregressive models or linear (non-)Gaussian Bayesian Networks. In contrast, we propose a novel frequency domain approach to identify a topological ordering and learn the structure of multivariate time series. In particular, we define a class of complex-valued Structural Causal Models (cSCM) at each frequency of the Fourier transform of the time series. Assuming that the time series is generated from the transfer function model, we show that the topological ordering and the corresponding summary directed acyclic graph can be uniquely identified from cSCM. The performance of our algorithm is investigated using simulation experiments and real datasets. Code implementing the proposed algorithm is available at Supplementary Materials.

Keywords: Directed Acyclic Graphs; Time Series Analysis; complex-valued SCM

1 Introduction

Structure learning in time series is used in many applications such as machine learning (Peters et al., 2017), economics (Bessler and Yang, 2003; Demiralp and Hoover, 2003), climate research (Runge et al., 2019), and earth science (Runge et al., 2019). There are two general approaches depending on the time-resolution of the data (Breitung and Swanson, 2002; Rajaguru and Abeysinghe, 2008; Hyvärinen et al., 2010). First, if the time-resolution of the measurements is higher than the time scale of the causal influence, then the structure can be learned from the autoregressive model with time-lagged variables. Conversely, if the measurements have a lower time resolution than the causal influence, a model can be used in which the causal influences are contemporaneous or instantaneous (White and Lu, 2010). For details on structure learning from undersampled time series, see Danks and Plis (2013); Gong et al. (2015); Plis et al. (2015).

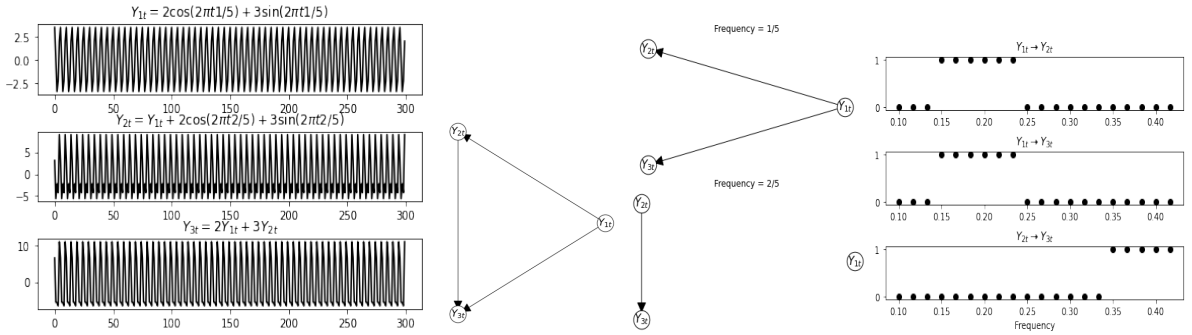
In multivariate time series literature, Structural vector autoregressive (SVAR) models are powerful tools for learning the structure of time series. SVAR allows causal influences to occur contemporaneously and with time lags. Swanson and Granger (1997); Demiralp and Hoover (2003); Moneta and Spirtes (2006); Runge et al. (2019) exploit constraint-based methods, such as PC (Peter-Clark) (Spirtes and Glymour, 1991) algorithm for the SVAR estimation. Such methods rely on Gaussianity and/or faithfulness assumption (see Section 2 for definitions). Hyvärinen et al. (2010); Moneta et al. (2013); Dallakyan (2020) propose methods for non-Gaussian data. Entner and Hoyer (2010); Malinsky and Spirtes (2018) exploit the FCI algorithm to allow for the unmeasured confounding effects. Chu and Glymour (2008) introduced additive non-linear time series models (ANLTSM) with linear contemporaneous effects for performing relaxed conditional independence tests. Peters et al. (2013) generalize ANLTSM and allow for the non-linear contemporaneous effects in their time series models with independent noise (TiMINo) approach. Recently, Pamfil et al.

(2020) propose a fully continuous optimization approach for learning the structure of time series by exploiting a novel characterization of acyclicity constraint introduced in Zheng et al. (2018).

In this work, we depart squarely from the time domain and propose a novel approach to learn structure of time series in the frequency domain. Exploring dependence in the frequency domain is especially of interest in the analysis of EEG data, or brain signals (for a recent overview, see Ombao and Pinto (2022)), as well as in the analysis of macro-economic datasets, such as for identifying business cycles (Croux et al., 2001). In sharp contrast to existing algorithms that learn DAG in the time domain, which by design are unable to identify dependencies in the frequencies, our approach defines a complex-valued SCM in each frequency and recovers the underlying DAG for each frequency. Our approach consists of a two step procedure. In the first step, we recover a topological ordering in each frequency and then use a penalized log likelihood to identify edges. We refer to our procedure as **Fre**quency **Dom**ain structure learning (FreDom). Additionally, we extend our approach to jointly estimate the frequencies under the assumption that the DAG structure is shared across specified frequencies.

We illustrate our method with a small toy example consisting of three time series generated according to Figure 1(a). As shown in Figure 1(b), the time domain algorithm (VARLINGAM) recovers a complete Directed Acyclic Graph (DAG) but fails to capture the different DAG structures present in frequencies 1/5 and 2/5. However, when we apply FreDom to these frequencies, the correct DAGs are recovered (see Figure 1(c)). Figure 1(d) shows the estimation of three edges across all frequency points. For example, in the top figure of Figure 1(d), 1 indicates that the algorithm estimated the edge $Y_{1t} \to Y_{2t}$ and 0 indicates the absence of the edge at the given frequency. As expected, this edge exists in frequencies closer to 1/5 but is missing in other frequencies.

To the best of our knowledge, the only frequency domain approaches for learning the



(a) (b) (c) (d)

Figure 1: Comparison of the VARLiNGAM algorithm with FreDom in a toy example with $p = 3$ variables and $n = 300$ observations. **(a)**: True time series. **(b)**: DAG estimated by the VARLiNGAM algorithm. **(c)**: DAG estimated by FreDom at frequencies 1/5 and 2/5, respectively. **(d)**: Edge estimation by FreDom across all frequencies. For a given frequency, 1 indicates that the particular edge has been estimated, while 0 indicates that the edge is missing.

structure of time series are proposed in Shajarisales et al. (2015); Besserve et al. (2021) and partial directed coherence metric in Baccala and Sameshima (2001); Baccala et al. (2013). The former is limited only to cases when the number of series is equal to two, while the latter heavily relies on VAR estimation. It is important to note that VAR mismodeling can significantly affect this metric (Ombao and Pinto, 2022, Section 5.2).

Compared to the existing methods, another feature of FreDom is that it allows to work with a complex-valued time series or sequence data. The latter is naturally used in telecommunications, robotics, bioinformatics, image processing, radar, and speech recognition (Peter and Scharf, 2010; Wolter and Yao, 2018a,b; Yang et al., 2020; Lee et al., 2022).

Throughout the paper, we use the following notation: scalars are denoted by lowercase letters, except when they indicate the length of time series or frequency. To distinguish a (random) vector from a matrix, we highlight the former in bold. The dependence of vector

or matrix from the time (frequency) index is represented by $\mathbf{X}(t)$ and $B(t)$, respectively, and the ith element of the vector $\mathbf{X}(t)$ is denoted by $X_i(t)$. The conjugate, and the conjugate transpose of the complex-valued matrix is denoted by $\boldsymbol{B}^*$ and $\boldsymbol{B}^H$, respectively. In addition, we place all appendices in Supplementary Materials.

2 Methods

We start by reviewing the existing literature on structure learning for *iid* data. The goal of structure learning is to recover the underlying structure of variables X_i, $i \in V$, given the samples from the distribution $\mathbf{P}$. We let $\mathcal{G}(V, E)$ be a directed acyclic graph (DAG) on E that describes the relationship between variables. Independence-based (also called constraint-based) methods (Spirtes and Glymour, 1991; Pearl, 2009), score-based methods (Heckerman et al., 1995; Chickering, 2002; Teyssier and Koller, 2005; Loh and Bühlmann, 2014), and functional-based methods (Shimizu et al., 2006; Peters et al., 2014; Zhang et al., 2015; Chen et al., 2019) are three popular approaches to learning the structure of the underlying DAG.

Independence-based methods, such as the inductive causation (IC) (Pearl, 2009) and PC (Peter-Clark) (Spirtes and Glymour, 1991) algorithm, utilize conditional independence tests to detect the existence of edges between each pair of variables. The method assumes that the distribution is Markovian and faithful for the underlying DAG, where $\mathbf{P}$ is faithful to the DAG $\mathcal{G}$ if all conditional independencies in $\mathbf{P}$ are entailed in $\mathcal{G}$, and Markovian if the factorization property $\mathbf{P}(X_1, \ldots, X_p) = \prod_{j=1}^{p} \mathbf{P}(X_j | \Pi_j^{\mathcal{G}})$ is satisfied. Here $\Pi_j^{\mathcal{G}}$ is the set of all parents of a node j. In contrast to constraint-based methods, the score-based approach treats structure learning as a combinatorial optimization problem. In particular, in the DAG space, they search and test various graph structures by assigning a score to each graph and selecting the one that best fits the data. Finally, the functional-based methods

restrict the functional class and the error term distributions so as to achieve identification.

2.1 Bayesian Networks and SCM

The SCM for a random vector $\mathbf{X} = \{X_i | i \in E\}$ is a 4-tuple $(\mathbf{X}, \boldsymbol{\varepsilon}, \mathcal{F}, P(\boldsymbol{\varepsilon}))$, where $\boldsymbol{\varepsilon}$ is a set of background (exogenous) variables, $\mathcal{F}$ is a set of functions $\{f_1, f_2, \ldots, f_p\}$ where each f_i maps $\varepsilon_i \cup \Pi_i^{\mathcal{G}}$ to X_i, and $P(\boldsymbol{\varepsilon})$ is a probability function defined over the domain of $\boldsymbol{\varepsilon}$. SCM posits casual relations, such that for all $i \in E$, $X_i := f_i(\Pi_i^{\mathcal{G}}, \varepsilon_i)$, where $\varepsilon_i, i \in E$ are jointly independent and the causal structure is encoded in a DAG $\mathcal{G}$ (Pearl, 2009; Bareinboim et al., 2020). For the recent overview of SCM in the context of econometrics, see Hünermund and Bareinboim (2023).

For example, if f_i's are linear and have additive noise, SCMs can be written as

$$X_j := \sum_{k \in \Pi_j^{\mathcal{G}}} \beta_{jk} X_k + \varepsilon_j, \ j = 1, \ldots, p, \tag{1}$$

Denoting the weighted adjacency matrix $B = (\beta_{jk})$ with zeros along the diagonal, the vector representation of (1)

$$\mathbf{X} := B\mathbf{X} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_p)'$ and $\mathbf{X} := (X_1, \ldots, X_p)'$. A DAG admits a topological ordering $\varrho(\cdot)$ with which a $p \times p$ permutation matrix P_ϱ can be associated such that $P_\varrho \boldsymbol{x} = (x_{\varrho(1)}, \ldots, x_{\varrho(p)})$, for $\boldsymbol{x} \in R^p$. The existence of a topological order leads to the permutation-similarity of B to a strictly lower triangular matrix $B_\varrho = P_\varrho B P_\varrho'$ by permuting rows and columns of B, respectively (Bollen, 1989).

2.2 Complex-Valued Bayesian Networks and cSCM

We define $\boldsymbol{Y} \in C^p$, be iid complex-valued, proper random vectors. The complex-valued SCM and corresponding DAG $\mathcal{G}$ can be defined analogously to real-valued SCM by

$$\boldsymbol{Y} := f(\boldsymbol{Y}, \boldsymbol{\varepsilon}_c) \tag{3}$$

For example, for linear Gaussian BN $\boldsymbol{Y} \sim N_c(0, \Sigma_c)$, then $E[\boldsymbol{Y}\boldsymbol{Y}^H] = \Sigma_c \in C^{p\times p} = \sigma^2(I-B)^{-1}\{(I-B)^H\}^{-1}$, and the weighted adjacency matrix $B \in C^{p\times p}$ is potentially complex-valued where the subscript c indicates that the distribution is complex-valued and A^H denotes the conjugate transpose $(A^*)'$. For details on complex-valued Gaussian distribution, see Chapter 2 in Andersen et al. (1995).

3 Complex-Valued Bayesian Networks For Time Series

We now return to structure learning for time series, given $\mathbf{X}(t) \in R^p$ or C^p for $t = 1, \ldots, T$ such that the autocovariance function satisfies $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, i.e. the spectral density matrix exists (Brockwell and Davis, 1986). Recall that the discrete Fourier transform (DFT) for the time series $\mathbf{X}(t)$ is

$$\boldsymbol{d}(\omega_k) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{X}(t) \exp(-2\pi i \omega_k t), \tag{4}$$

$\boldsymbol{d}^*(\omega_k) = \boldsymbol{d}(-\omega_k) = \boldsymbol{d}(1-\omega_k)$ and from (Brillinger, 1981, Theorem 4.4.1) as $T \to \infty$, $\boldsymbol{d}(\omega_k),\ k = 2, 3, \ldots, (T/2) - 1$ are independent complex Gaussian $N_c(0, S(\omega_k))$ random vectors and for $k = \{1, T/2, T\}$, $\boldsymbol{d}(\omega_k)$ are independent real Gaussian $N_r(0, S(\omega_k))$, where $S(\omega_k)$ is the spectral density matrix at the Fourier frequency ω_k. We assume that the DFT $\boldsymbol{d}(\omega_k)$ satisfies the cSCM with the additive error at each Fourier frequency $\omega_k,\ k = 1, \ldots, T/2$:

$$\boldsymbol{d}(\omega_k) = f(\boldsymbol{d}(\omega_k)) + \boldsymbol{\varepsilon}(k). \tag{5}$$

We denote the adjacency matrix of the graph G by W, where $W_{ij} = 1$ if $d(\omega_k)_j \to d(\omega_k)_i$. Note that if f is linear then the coefficient matrix B has the same non-zero pattern as W.

3.1 Linear Case

In this section, we assume f is linear in (5)

$$\boldsymbol{d}(\omega_k) := B(\omega_k)\boldsymbol{d}(\omega_k) + \boldsymbol{\varepsilon}(k). \tag{6}$$

where $B(\omega_k) \in C^{p\times p}$ entails the underlying structure of the DAG at frequency ω_k, $k = 1, \ldots, T/2$. Consequently, from the inverse Fourier transform and (6), **the time series is generated** from the transfer function model

$$\mathbf{X}(t) = \sum_{k=1}^{T} (I_p - B(\omega_k))^{-1} \exp(2\pi i \omega_k t)\boldsymbol{\varepsilon}(k), \tag{7}$$

where $i = \sqrt{-1}$, $\omega_k = k/T,\ k = 1, \ldots, T$ and $\boldsymbol{\varepsilon}(k)$ are independent $N_c(0, (1/T)I_p)$, $\boldsymbol{\varepsilon}(k) = \boldsymbol{\varepsilon}^*(T-k)$ for $\omega_t \neq \{0, 0.5, 1\}$, and real Gaussian $N_r(0, (1/T)I_p)$ otherwise. Moreover, from (5), the spectral density matrix can be estimated by periodogram, defined as:

$$I(\omega_k) = \frac{1}{T}(I_p - B(\omega_k))^{-1}\{(I_p - B(\omega_k))^{-1}\}^H. \tag{8}$$

In general, $I(\omega_k)$ is an unbiased estimator of the spectral density $S(\omega_k)$, but not consistent even under classical fixed-p asymptotics (Brillinger, 1981, Theorem 5.2.4). To insure consistency, it is common to use a smoothed periodogram estimator

$$\hat{S}(\omega_k) = \frac{1}{2m+1} \sum_{|k|\leq m} I(\omega_{j+k}) \tag{9}$$

where m is chosen as $m = o(\sqrt{T})$.

A point of departure for our algorithm is an important result for the real-valued SCM, which state that the graph $\mathcal{G}$ and the parameters B can be identified from the covariance matrix under equal variance and causal sufficiency assumptions (Peters and Bühlmann, 2013). Ghoshal and Honorio (2018); Chen et al. (2019) observe that the ordering of certain conditional variances implies the identifiability of parameters. Consequently, by ordering the estimates of those variables, the authors establish a fast method to learn the topological ordering of the variables. Next lemma, which is the extension of Chen et al. (2019,

Lemmas 1) to cSCM defined in (6), is used to recover such topological ordering for cSCM. The proof is provided in Appendix A.1, for completeness.

Lemma 1 *Let $\boldsymbol{Y} \in C^p$ is generated as in (6). If the parent set $\Pi_j^G = \emptyset$ then var$(\boldsymbol{Y}_j) = 1/T$, otherwise var$(\boldsymbol{Y}_j) \geq 1/T * (1 + \eta) > 1/T$, where $\eta = \min_{(k,j)\in E} \beta_{jk}\beta_{jk}^*$.*

The findings in Lemma 1 enable the modification of Chen et al. (2019, Algorithm 1) to accommodate the complex-valued case. In this modified algorithm, the topological ordering of the Fourier transforms $\boldsymbol{d}(\omega_k)$ is estimated at each Fourier frequency by iteratively selecting a source node. This selection is based on comparing variances conditional on the previously chosen variables. Notably, the findings imply that the conditional variance of var$(\boldsymbol{Y}_j|\boldsymbol{Y}_C)$, where C is a set, equals $1/T$ if the parents of j form a subset of C, and is greater than or equal to $1/T(1 + \eta)$ otherwise.

Remark 1 *We note that the assumption of equal error variance is commonly used in application with variables from a similar domain, spatial or time series data (Rajaratnam and Salzman, 2013; Park, 2020).*

The main distinction between FreDom and Chen et al. (2019) is that, at each frequency point, the conditional variances are derived from the (inverse) spectral density matrix, rather than the covariance matrix. Algorithm 1 provides a summary of the main steps.

To select a source node by comparing conditional variances, in Algorithm 1, we minimize the frequency domain analog of Chen et al. (2019) criterion

$$\begin{aligned} f(S(\omega_k), \Theta[k,(i-1)], j) &= [\hat{S}(\omega_k)]_{j,j} - [\hat{S}(\omega_k)]_{j,\Theta}[\hat{S}(\omega_k)]^{-1}_{\Theta,\Theta}[\hat{S}(\omega_k)]_{\Theta,j} \\ &= \frac{1}{([\hat{S}(\omega_k)]^{-1}_{\Theta\cup\{j\},\Theta\cup\{j\}})_{j,j}}. \end{aligned} \tag{10}$$

The next theorem shows that, under certain assumptions on $N = 2m + 1$, at each frequency ω_k, the Algorithm 1 recovers a topological ordering of the underlying true graph with probability at least $1 - \varepsilon$. The proof is provided in Appendix A.2.

Algorithm 1 Stage 1 of FreDom Algorithm

Input:

$\hat{S}(\omega_k) \leftarrow$ ***spectral density matrix at frequency*** ω_k

$\Theta \in R^p \leftarrow \emptyset$

for $i = 1$ **to** p **do**

$\theta \leftarrow \arg\min_{j \in V/\Theta[i-1]} f(\hat{S}(\omega_k), \Theta[(i-1)], j)$

$\Theta[i] = \theta$

end for

Output:Θ

Theorem 1 *Let* $\boldsymbol{d}(\omega_k)$ *satisfies (6) and time series is generates as in (7). Suppose that the spectral density matrix* $S(\omega_k)$ *has minimum eigenvalues* $\lambda_{\min} > 0$*. If*

$$N > p^2 \log\Big(\frac{16p^2}{\varepsilon}\Big) 12800 \max_i([S(\omega_k)]^2_{ii})\Big(\frac{\eta\lambda_{\min} + (2/T)(1+\eta)}{\eta\lambda^2_{\min}}\Big),$$

then Algorithm 1 using criterion (10) recovers a topological ordering with probability at least $1 - \varepsilon$*.*

3.2 Recovering DAG from topological ordering

In the first stage of FreDom, Algorithm 1 returns the topological ordering of a DAG at each frequency. In Stage 2 of FreDom, we recover the DAG given the topological ordering. As discussed in Section 2.1, given a topological ordering ϱ, B_ϱ is lower triangular. Similarly, from (5) and (8), given the ordering, $B_\varrho(\omega_k)$ is lower triangular, and $L_\varrho(\omega_k) = \sqrt{T}(I - B_\varrho(\omega_k))$ is the Cholesky factor of the inverse spectral density matrix $\Omega_\varrho(\omega_k) = S^{-1}_\varrho(\omega_k) = L^H_\varrho(\omega_k)L_\varrho(\omega_k)$. From now on, whenever there is no confusion, we drop the subscript ϱ. From (4), in frequency ω_k, the pdf for $\boldsymbol{d}(\omega_k)$ is

$$g(\boldsymbol{d}(\omega_k)) = \frac{\exp(-\boldsymbol{d}^H(\omega_k)L^H(\omega_k)L(\omega_k)\boldsymbol{d}(\omega_k))}{\pi^p \det(S(\omega_k))} \tag{11}$$

As we discussed in Section 3.1, to achieve a consistent estimator of spectral density (9), we smooth it along the $N = 2m+1$ frequencies, where $m \approx o(\sqrt{T})$ is the half-window size. Therefore, the pdf for $\boldsymbol{d}(\omega_k)$ can be written as

$$g(\boldsymbol{d}(\omega_k)) = \frac{\exp\{-N\mathrm{tr}(\hat{S}(\omega_k)L^H(\omega_k)L(\omega_k))\}}{\pi^{Np}\det(L^H(\omega_k)L(\omega_k))^{-N}}, \tag{12}$$

where $\hat{S}(\omega_k)$ is given in (9) and the convex penalized log-likelihood function is

$$W_{\mathrm{FreDom}}[L(\omega_k)] = \log\det(L^H(\omega_k)L(\omega_k)) - \mathrm{tr}(\hat{S}(\omega_k)L^H(\omega_k)L(\omega_k)) + \lambda_N \sum_{i>j} |L(\omega_k)_{ij}|, \tag{13}$$

with $\lambda_N = \lambda/N$. Let $\boldsymbol{x}^i = \{L_{ij}(\omega_k)\}_{j=1}^i$ denote the vector of lower triangular and diagonal entries in the ith row of $L(\omega_k)$ and $\hat{S}_i(\omega_k)$ is the $i \times i$ submatrix of $\hat{S}(\omega_k)$ for $1 \le i \le p$. After some algebra, it follows from (13)

$$\begin{aligned} W_{\mathrm{FreDom}}[L(\omega_k)] &= \sum_{i=1}^{p} (\boldsymbol{x}^i)^H \hat{S}_i(\omega_k)\boldsymbol{x}^i - 2\log \boldsymbol{x}_i^i + \lambda \sum_{j=1}^{i-1} \sqrt{|x_j^i|} \\ &= \sum_{i=1}^{p} W_{\mathrm{FreDom,i}}(\boldsymbol{x}^i) \end{aligned} \tag{14}$$

where for $2 \le j \le p$

$$W_{\mathrm{FreDom,i}}(\boldsymbol{x}^i) = (\boldsymbol{x}^i)^H \hat{S}_i(\omega_k)\boldsymbol{x}^i - 2\log x_i^i + \lambda \sum_{j=1}^{i-1} \sqrt{|x_j^i|}$$

and

$$W_{\mathrm{FreDom,1}}(\boldsymbol{x}^i) = [L(\omega_k)]_{11}^2 [S(\omega_k)]_{11} - 2\log[L(\omega_k)]_{11}$$

Equation (14) demonstrates that the optimization of $W_{\mathrm{FreDom}}[L(\omega_k)]$ decomposes into an optimization of p parallel functions, and the functions depend on disjoint sets of parameters. Similar to Khare et al. (2019, Lemmas 2.2 and 2.3), it is easy to show that any

global minimum of $W_{\text{FreDom}}[L(\omega_k)]$ over the open set $\mathcal{L}_p$ (a set of complex-valued lower triangular matrices with positive diagonals) lies in $\mathcal{L}_p$.

We now provide an algorithm to minimize $W_{\text{FreDom}}[L(\omega_k)]$. Since $\{\boldsymbol{x}^i\}_{i=1}^p$ separates the non-zero parameters in $L(\omega_k)$, it follows that optimizing $W_{\text{FreDom}}[L(\omega_k)]$ is equivalent to a parallel optimization of $W_{\text{FreDom},1}(\boldsymbol{x}^i)$, $1 \leq i \leq p$. It is important and timely to note that $\boldsymbol{x}^i$ is complex-valued. Thus, we resort to Wirtinger calculus (Wirtinger, 1927; Brandwood, 1983; Dallakyan et al., 2022), together with the definition of Wirtinger subgradients (Bouboulis et al., 2012) for the optimization. The next lemma show that a minimizer of $W_{\text{FreDom},1}(\boldsymbol{x}^i)$ can be computed in a closed form. The proof is provided in Appendix A.3.

Lemma 2 *A minimizer of $W_{FreDom,i}(\boldsymbol{x}^i)$ can be computed in a closed form.*

$$\boldsymbol{x}_j^i = -\Big(1 - \frac{\lambda}{2|\sum_{l\neq j}(\hat{S}_i(\omega_k))_{lj}x_l|}\Big)_+ \frac{\sum_{l\neq j}(\hat{S}_i(\omega_k))_{lj}x_l}{(\hat{S}_i(\omega_k))_{jj}} \tag{15}$$

for $1 \leq j \leq i-1$, and

$$x_i^i = \frac{-Re(\sum_{l\neq k}(\hat{S}_i(\omega_k))_{li}x_l) + \sqrt{Re(\sum_{l\neq i}(\hat{S}_i(\omega_k))_{li}x_l)^2 + 4(\hat{S}_i(\omega_k))_{ii}}}{2(\hat{S}_i)_{ii}} \tag{16}$$

Using Lemma 2, we develop a cyclic coordinatewise algorithm, where the elements of $\boldsymbol{x}^i$ are iteratively updated until convergence. Algorithm 2 summarizes Stage 2 of the FreDom algorithm.

Note that given a topological ordering, Algorithm 2 solves p modified complex-lasso problems. Under similar assumptions and techniques, as in Tugnait (2022, Theorem 1) and Deb and Basu (2023), it can be shown that our proposed estimator converges to true $L(\omega_k)$ with high probability.

Algorithm 2 Stage 2 of FreDom: minimization algorithm for W_{FreDom}

Input:

$\hat{L}(\omega_k)^{(0)}, \leftarrow$ ***initial estimate***

for $i = 1$ **to** p **do**

Set $\hat{\boldsymbol{x}}^{\boldsymbol{i}}$ **to be minimizer of** $W_{\mathbf{FreDom,i}}(\boldsymbol{x}^{\boldsymbol{i}})$ **by using coordinatewise algorithm**

Construct $\hat{L}(\omega_k)$ **by setting its** i**th row as** $\hat{\boldsymbol{x}}^{\boldsymbol{i}}$

end for

4 Application to Stock Volatility Data

We utilize the FreDom method to analyze data on stock return volatility. Section 6 presents additional simulation results. The data used in this study is sourced from Demirer et al. (2018), where authors estimate the global bank network connectedness. The original dataset comprises 96 banks from 29 developed and emerging economies (countries), spanning the period from September 12, 2003, to February 7, 2014. To facilitate clarity, we specifically focus on economies with more than four banks, resulting in a selection of 54 banks (for further details, please refer to Demirer et al. (2018)). Figure 2 visualizes the estimated adjacency matrix generated by the FreDom algorithm at various frequencies. The rows and columns of the matrix are sorted by country. The tuning parameter for the FreDom algorithm is determined using the extended Bayesian information criterion (BIC) (Foygel and Drton, 2010). We define a search space grid $[\lambda_{min}, \lambda_{max}]$, where the selection of λ_{min} and λ_{max} is made to avoid excessively dense or sparse models. To initiate the search, we find the value of λ^* that yields a graph without edges and set $\lambda_{max} = \lambda/2$ to prevent highly sparse models. Additionally, we employ a "warm" starting strategy across the grid to expedite convergence.

An important observation derived from this exercise is that Fourier frequencies in close proximity exhibit a tendency to possess a similar causal structure and topological order-

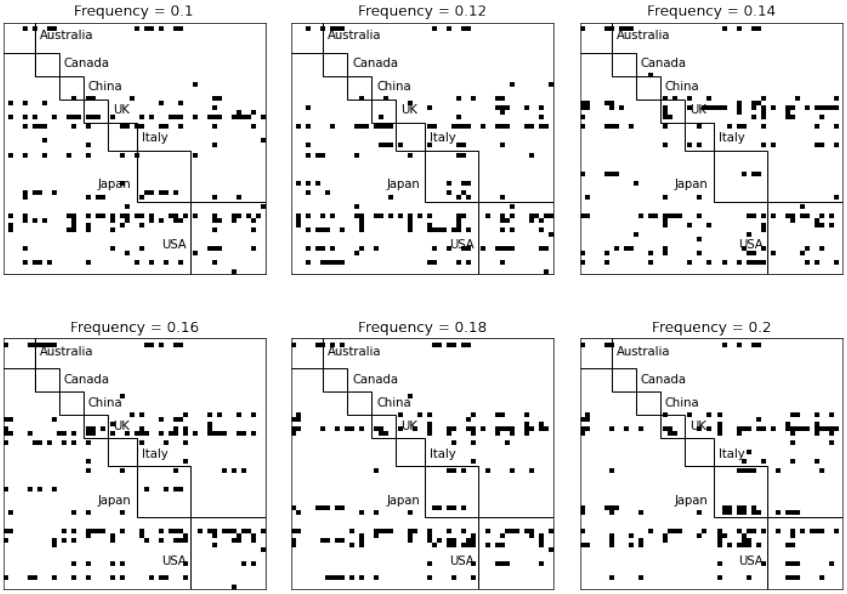


Figure 2: Estimated adjacency matrix for different frequencies with rows and columns sorted by country.

ing. This claim is supported by the block diagonal structure depicted in Figure 3, which showcases the Spearman correlation of the topological orderings across frequencies.

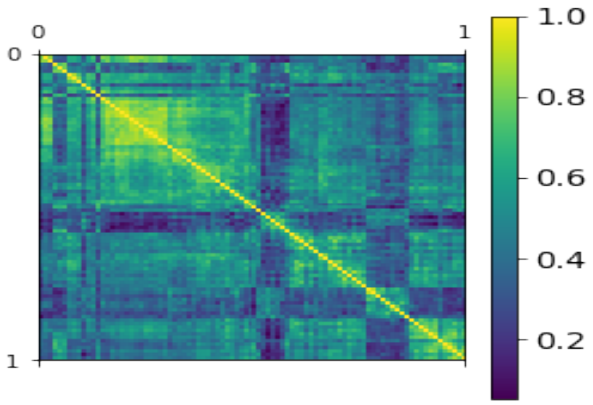


Figure 3: Spearman correlation of topological orderings over frequencies.

Hence, it is justifiable to assume that the structure remains invariant for frequencies that are in close proximity to each other. The joint estimation of graphical models has shown to be more accurate in practice than separate estimation (Danaher et al., 2014).

Further details on this assumption are provided in the subsequent section.

5 Joint estimation

Let $J = \{\omega_1, \ldots, \omega_M\}$ be the set of selected (possibly in close proximity) frequencies with cardinality $M = |J|$. The discussion in the previous section motivates the following assumption:

Assumption 1 *(Structure Invariance) The structure of time series* $\mathbf{X}(t)$ *and DFT* $\boldsymbol{d}(\omega_k)$, $(k, t = 1, \ldots, T)$ *remains unchanged across the time and* M *frequency points.*

Next, we define a **summary** DAG for the frequency domain.

Definition 1 *A summary DAG* $\mathcal{G}$ *for the time series* $\mathbf{X}(t)$ *and specified frequency points* $J = \{\omega_1, \ldots, \omega_M\}$ *is a DAG which has an arrow from* $\mathbf{X}(t)_i$ *to* $\mathbf{X}(t)_j$, $i \neq j$, *if the corresponding adjacency matrix* $W_{ji}(\omega_k) \neq 0$ *for some* $k = 1, \ldots, M$.

To accommodate joint estimation over selected frequencies, Algorithms 1 and 2 need a slight modification.

Algorithm 1 modification: After estimating topological ordering over $J = \{\omega_1, \ldots, \omega_M\}$ frequencies, we select the most commonly occurring ordering.

Algorithm 2 modification: Given the topological ordering, the joint log-likelihood function over M frequencies is

$$W(L[\cdot]) = \sum_{k=1}^{M} N[\log\det(L^H(\omega_k)L(\omega_k)) - \mathrm{tr}(\tilde{S}(\omega_k)L^H(\omega_k)L(\omega_k))].$$

From Assumption 1, $B(\omega_k)$, and therefore $L(\omega_k)$, have the same structure over $k = 1, \ldots, M$ frequency point. Thus, to impose a structure similarity assumption on the Fourier frequency points, we define the following constrained optimization problem

$$\begin{aligned} \min_{L[\cdot],Z} \quad & -W(L[\cdot]) + P(Z,\lambda), \\ \text{s.t.} \quad & L(\omega_k) = Z,\ k = 1,\ldots,M, \end{aligned} \tag{17}$$

where

$$\begin{aligned} P(Z,\lambda) &= \lambda \sum_{ij} |Z_{ij}| \\ L[\cdot] &= \{L(\omega_1),\ldots,L(\omega_M)\}. \end{aligned} \tag{18}$$

The constraints $L(\omega_k) = Z,\ k = 1,\ldots,M$ is used to ensure that Assumption 1 is satisfied, i.e., in each Fourier frequency the summary DAG structures are the same and the penalty $P(Z,\lambda)$ introduces sparsity. The minimization problem (17) is convex, and the existence of a minimizer is guaranteed for any choice of $\lambda \geq 0$ (Rockafellar, 1970, Theorem 27.2). We appeal to the ADMM (alternating direction method of multipliers) algorithm for minimizing (17) (Boyd et al., 2011; Dallakyan et al., 2022; Ng and Zhang, 2022). The ADMM minimizes the scaled augmented Lagrangian

$$\begin{aligned} \mathcal{L}_\rho(\Theta[\cdot], Z, U[\cdot]) = & \sum_{n=1}^{M} N[-\log\det(L^H(n)L(n)) + \mathrm{tr}(\tilde{S}(n)L^H(n)L(n))] \\ & + \rho \sum_{n=1}^{M} (\|L(n) - Z + U(n)\|_F^2 - \|U(n)\|_F^2) + P(Z,\lambda), \end{aligned} \tag{19}$$

where $\rho > 0$ is the penalty coefficient, $U(n)$, $n = 1,\ldots,n$ are the Lagrangian multipliers, and $\|X(n)\|_F^2 = \sum_{ij} |X_{ij}(n)|^2$. Given $(L^{(k)}[\cdot], Z^{(k)}, U^{(k)}[\cdot])$ matrices in the kth iteration, the ADMM algorithm implements the following three updates for the next $(k+1)$ iteration:

(a) $L^{(k+1)}[\cdot] \leftarrow \arg\min_{L[\cdot]} \mathcal{L}_\rho(\Theta[\cdot], Z^{(k)}, U^{(k)}[\cdot])$

(b) $Z^{(k+1)} \leftarrow \arg\min_{Z]} \mathcal{L}_\rho(L^{(k+1)}[\cdot], Z, U^{(k)}[\cdot])$

(c) $U^{(k+1)}[\cdot] \leftarrow U^{(k)}[\cdot] + (L^{(k+1)}[\cdot] - Z^{(k+1)})$

Interestingly, as we show in Appendix B, each of the updates (a)-(b) has closed-form solutions. Moreover, in contrast to real-valued ADMM formulation, where L is real-valued,

in (17), $L[\cdot]$ is complex-valued. As before, to solve complex-valued optimization, we resort to Wirtinger calculus (Wirtinger, 1927; Brandwood, 1983).

Figure 4 illustrates joint estimation of the stock volatility data for the frequencies considered in Section 4.

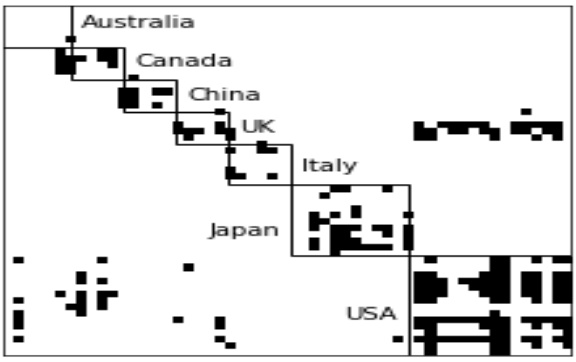


Figure 4: Estimated adjacency matrix using joint FreDom. Rows and columns are sorted by country.

As can be seen, compared to findings in Figure 2, banks from the same country tend to compose tighter groups, meanwhile being connected to banks from other countries. The latter result has been confirmed in many macro-economic studies (Demirer et al., 2018). The other interesting finding that needs more investigation is the causal relationship between UK and US banks.

6 Numerical Experiments

In this section, we present the performance of FreDom, Joint FreDom on various simulated time series data. To facilitate comparison, we introduce an extended version of FreDom, denoted as exFreDom, in Appendix C. exFreDom is an extended version of NOTEARS (Zheng et al., 2018) adapted to the frequency domain. It is utilized in Experiment 3 and the Air Pollution data analysis in Section 6.1.

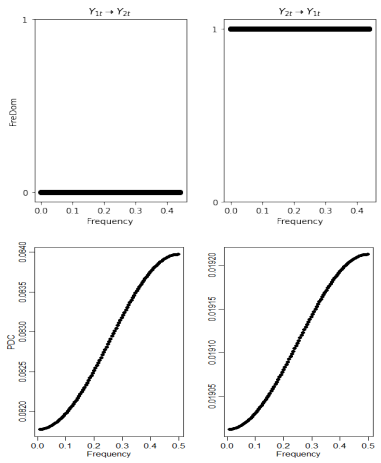


Figure 5: Comparing FreDom (top row) with PDC (bottom row) over frequencies. Columns correspond to causal directions.

One challenge in adopting the joint frequency domain approach is the need for parameter estimation at each Fourier frequency. For instance, in the case of FreDom with a p-dimensional time series and M Fourier frequencies, it estimates $Mp(p+1)/2$ parameters. Based on simulations, satisfactory results are obtained by choosing $M = (5, 10)$.

For all experiments, the time series length is set to $T = 1000$, the half-window size is $m = \sqrt{T}/2$, and each simulation is repeated 50 times. In our simulations, we assume oracle sparsity for all methods. This means that each method employs tuning parameters that yield the true number of edges.

Experiment 1: Two dimensional Time Series. In this experiment, we compare FreDom with partial directed coherence (PDC) introduced in Baccala and Sameshima (2001). We generate two-dimensional series following the procedure described in the subsequent experiment, ensuring that $Y_{2t} \to Y_{1t}$ in each Fourier frequency.

Figure 5 summarizes the results. Each column represents the causal direction $Y_{1t} \to Y_{2t}$ and $Y_{2t} \to Y_{1t}$ respectively, while the rows correspond to the outcomes from FreDom and PDC. In the case of FreDom, a value of 0 indicates the absence of an edge, while 1 signifies

the presence of an edge. For PDC, absolute values closer to 1 indicate causal dependence, while values closer to 0 suggest no relationship. As observed, FreDom successfully captures the true causal direction, while PDC yields small values in both directions, incorrectly indicating no relationship.

Experiment 2: Time Series from (7)

Experiment 2: Time Series from (7). We utilize (Dai and Guo, 2004, Theorem 1) (for details, see Appendix E), which states that (7) can be used to generate a time series whose topological order and spectrum are identical to the given order and spectrum at Fourier frequencies. We simulate complex-valued time series from the provided random summary DAG for $p = 5, 10, 15$.

For each Fourier frequency ω_k, we construct the Cholesky factor of the inverse spectral density by following these steps: (1) Fix the order and fill the adjacency matrix with zeros, (2) Replace every matrix entry in the lower triangle (below the diagonal) by independent realizations of Bernoulli(s) random variables with success probability s, $0 < s < 1$, where s reflects the sparseness of the model. We select $s = 0.4$ for this experiment. (3) Finally, in the adjacency matrix replace each entry with a 1 by the independent realizations of a $c_1 \cos(4\pi\omega_k) + 1.2ic_2 \sin(2\pi\omega_k)$, where c_1, c_2 are randomly selected from the $U[-0.1, -1] \cup [0.1, 1]$ distribution. The above procedure ensures that the DAG structure of the generated time series is the same for all frequencies, and $B(\omega_k)$ in (5) is only a function of ω_k. In Figure 6 we compare the performance of Joint FreDom and VARLINGAM (Hyvärinen et al., 2010) using the structural hamming distance (SHD). We can see that FreDom outperforms VARLINGAM for all K.

Experiment 3: Data from the non-linear SVAR Model. Similar to Peters et al. (2013), we simulate dataset from $X_1(t) = b_{11}X_2(t)^2 + b_{12}X_1(t-1) + b_{13}X_2(t-1)^2 +$

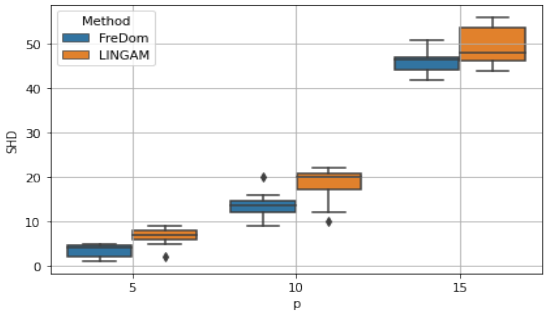


Figure 6: SHD metrics for Experiment 2.

$u_1(t)$, $X_2(t) = b_{22}X_2(t-1) + u_2(t)$, $X_3(t) = b_{31}X_1(t)^3 + b_{32}X_2(t-1)^2 + b_{33}X_3(t-1) + u_3(t)$, $X_4(t) = \exp(b_{41}X_{3t}) + b_{42}X_4(t-1) + u_4(t)$, where $u_i(t) \sim N(0,1)$ and $b_{ij} \sim U[-0.1,-0.4] \cup [0.1,0.4]$. In this experiment, we compare performance of Joint FreDom, ExFreDom, VARLIGNAM and DYNOTEARS (Pamfil et al., 2020) using SHD and Structural Intervention Distance (SID) (Peters and Bühlmann, 2015). The SID quantifies the proximity between two DAGs in terms of their respective causal inference statements. A lower value of SID and SHD indicates a better performance. Figure 7 reports the simulation results.

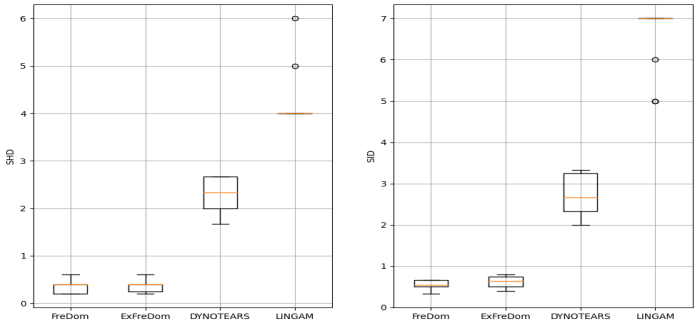


Figure 7: SHD and SID metrics for Experiment 2.

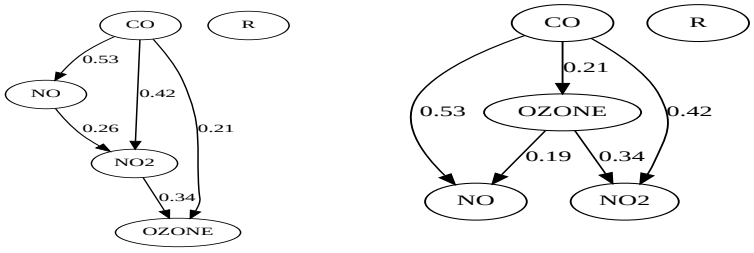
6.1 Air Pollution Data

We use (Ex)FreDom to estimate a summary DAG for 5 time series of air pollutants of length 8370. The series was recorded hourly during the year 2006 at Azusa, California. Data

can be obtained from the Air Quality and Meteorological Information System. Recorded variables include CO and NO (pollutants mainly emitted from cars and industry), NO_2 and O_3 (generated from different reactions in the atmosphere), and the global solar radiation intensity R. The similar datasets were analyzed in Dahlhaus and Eichler (2003) and Davis et al. (2016).

Figure F.2 in Appendix F shows an average daily plot of five variables. Due to early morning traffic, CO and NO increase early, resulting in NO_2 increase. Higher NO_2 levels increase the Ozone (O_3) and the global radiation levels throughout the day.

Following Dahlhaus and Eichler (2003), we apply FreDom to the residual series after subtracting the daily averages, as shown in Figure F.2. The missing values in the original series are filled in by interpolating the residual series using splines. Figure 8(a) and Figure 8(b) report the estimated summary DAGs from FreDom and ExFreDom, respectively. The weights on the edges report the absolute values of the partial spectral coherence, which are frequency domain analogs of partial correlations. Additional results for LINGAM, NOTEARS, and Granger causality are available in Appendix F.



(a) FreDom (b) ExFreDom

Figure 8: The estimated DAG from the air pollution data.

The summary DAG of FreDom correctly capturing the generation of NO_2 from CO and NO and the contemporaneous relation of CO and NO as the latter two pollutants are emitted from cars. However, we cannot validate the direction of the edge from the CO to NO. The edge from NO_2 to O_3 indicates that the latter is created from NO_2. Compared to

FreDom, ExFreDom misses edge from NO to NO_2 and the edge from O_3 to NO_2 is reversed.

7 Conclusion

In this paper, we present a frequency domain approach for recovering the topological ordering of time series. Based on the obtained ordering, we propose a penalized likelihood approach to learn the summary directed acyclic graph (DAG). Additionally, we extend the algorithm to enable joint estimation. The simulation results are highly encouraging and demonstrate the effectiveness of our proposed method.

For future work, we suggest exploring situations in which time series are affected by unobserved confounders or undersampling. These scenarios present interesting challenges that warrant further investigation.

SUPPLEMENTARY MATERIAL

Appendix: "FreDomsupplement.pdf" includes supplementary materials covering proofs and additional simulations for the proposed FreDom algorithm.

Stock and Air Pollution datasets: Datasets used in the illustration of FreDom in Sections 6.1 and 4.

Python code for examples: A Python script for reproducibility. The Stata code is available upon request.

References

Andersen, H., M. Hojbjerre, D. Sorensen, and P. Eriksen (1995). *Linear and Graphical Models for the Multivariate Complex Normal Distribution*. New York: Springer-Verlag.

Baccala, L., C. Brito, D. Takahashi, and K. Sameshima (2013, 07). Unified asymptotic theory for all partial directed coherence forms. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 371*, 20120158.

Baccala, L. and K. Sameshima (2001, 05). Partial directed coherence: A new concept in neural structure determination. *Biological Cybernetics 84*, 463–474.

Bareinboim, E., J. Correa, D. Ibeling, and T. Icard (2020). On Pearl's Hierarchy and the Foundations of Causal Inference. Technical Report R-60, Causal AI Lab, Columbia University.

Besserve, M., N. Shajarisales, D. Janzing, and B. Schölkopf (2021). Cause-effect inference through spectral independence in linear dynamical systems: theoretical foundations.

Bessler, D. A. and J. Yang (2003). The structure of interdependence in international stock markets. *Journal of International Money and Finance 22*(2), 261–287.

Bollen, K. (1989). *Structural Equations with Latent Variables*. New York: John Wiley and Sons.

Bouboulis, P., K. Slavakis, and S. Theodoridis (2012). Adaptive learning in complex reproducing kernel hilbert spaces employing wirtinger's subgradients. *IEEE Transactions on Neural Networks and Learning Systems 23*, 425–438.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn. 3*(1), 1–122.

Brandwood, D. H. (1983). A complex gradient operator and its application in adaptive array theory. *IEE Proceedings F - Communications, Radar and Signal Processing 130*(1), 11–16.

Breitung, J. and N. R. Swanson (2002). Temporal aggregation and spurious instantaneous causality in multiple time series models. *Econometrics eJournal*.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Brockwell, P. J. and R. A. Davis (1986). *Time Series: Theory and Methods*. New York, NY, USA: Springer-Verlag New York, Inc.

Chen, W., M. Drton, and Y. S. Wang (2019, 09). On causal discovery with an equal-variance assumption. *Biometrika 106*(4), 973–980.

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

Chu, T. and C. Glymour (2008). Search for additive nonlinear time series causal models. *Journal of Machine Learning Research 9*(32), 967–991.

Croux, C., M. Forni, and L. Reichlin (2001). A measure of comovement for economic variables: Theory and empirics. *The Review of Economics and Statistics 83*(2), 232–241.

Dahlhaus, R. and M. Eichler (2003). Causality and graphical models in time series analysis. In G. P., H. N, and R. S (Eds.), *Highly Structured Stochastic Systems*, Chapter 4. Oxford: Oxford University Press.

Dai, M. and W. Guo (2004). Multivariate spectral analysis using cholesky decomposition. *Biometrika 91*(3), 629–643.

Dallakyan, A. (2020). Nonparanormal Structural VAR for Non-Gaussian Data. *Computational Economics 0*, 1–21.

Dallakyan, A., R. Kim, and M. Pourahmadi (2022). Time series graphical lasso and sparse var estimation. *Computational Statistics & Data Analysis 176*, 107557.

Danaher, P., P. Wang, and D. M. Witten (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(2), 373–397.

Danks, D. and S. Plis (2013). Learning causal structure from undersampled time series.

Davis, R. A., P. Zang, and T. Zheng (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics 25*(4), 1077–1096.

Deb, N. and S. Basu (2023). Regularized estimation of sparse spectral precision matrices.

Demiralp, S. and K. Hoover (2003). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics 65*, 745–767.

Demirer, M., F. X. Diebold, L. Liu, and K. Yilmaz (2018). Estimating global bank network connectedness. *Journal of Applied Econometrics 33*(1), 1–15.

Entner, D. and P. Hoyer (2010). On causal discovery from time series data using fci.

Foygel, R. and M. Drton (2010). Extended bayesian information criteria for gaussian graphical models. In *Advances in Neural Information Processing Systems 23*, pp. 604–612.

Ghoshal, A. and J. Honorio (2018). Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Volume 84 of *Proceedings of Machine Learning Research*, pp. 1466–1475.

Gong, M., K. Zhang, B. Schoelkopf, D. Tao, and P. Geiger (2015). Discovering temporal causal relations from subsampled data. In *Proceedings of the 32nd International Conference on Machine Learning*, Volume 37 of *Proceedings of Machine Learning Research*, pp. 1898–1906.

Heckerman, D., D. Geiger, and D. M. Chickering (1995, September). Learning bayesian networks: The combination of knowledge and statistical data. *Mach. Learn. 20*(3), 197–243.

Hyvärinen, A., K. Zhang, S. Shimizu, and P. O. Hoyer (2010). Estimation of a structural vector autoregression model using non-gaussianity. *J. Mach. Learn. Res. 11*, 1709–1731.

Hünermund, P. and E. Bareinboim (2023, 03). Causal inference and data fusion in econometrics. *The Econometrics Journal*.

Khare, K., S.-Y. Oh, S. Rahman, and B. Rajaratnam (2019). A scalable sparse cholesky based approach for learning high-dimensional covariance matrices in ordered data. *Machine Learning 108*(12), 2061–2086.

Lee, C., H. Hasegawa, and S. Gao (2022). Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica 9*(8), 1406–1426.

Loh, P.-L. and P. Bühlmann (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *J. Mach. Learn. Res. 15*(1), 3065–3105.

Malinsky, D. and P. Spirtes (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, Volume 92 of *Proceedings of Machine Learning Research*, pp. 23–47.

Moneta, A., D. Entner, P. O. Hoyer, and A. Coad (2013). Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics 75*(5), 705–730.

Moneta, A. and P. Spirtes (2006). Graphical models for the identification of causal structures in multivariate time series models. In *Proceedings of the 9th Joint International Conference on Information Sciences (JCIS-06)*, pp. 613–616. Atlantis Press.

Ng, I. and K. Zhang (2022, 28–30 Mar). Towards federated bayesian network structure learning with continuous optimization. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera (Eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, Volume 151 of *Proceedings of Machine Learning Research*, pp. 8095–8111. PMLR.

Ombao, H. and M. Pinto (2022). Spectral dependence. *Econometrics and Statistics*.

Pamfil, R., N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, P. Beaumont, K. Georgatzis, and B. Aragam (2020). Dynotears: Structure learning from time-series data. *ArXiv abs/2002.00498*.

Park, G. (2020). Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research 21*(75), 1–34.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). USA: Cambridge University Press.

Peter, S. and L. Scharf (2010). *Statistical signal processing of complex-valued data : the theory of improper and noncircular signals*. Cambridge: Cambridge University Press.

Peters, J. and P. Bühlmann (2015). Structural intervention distance for evaluating causal graphs. *Neural Comput. 27*(3), 771–799.

Peters, J. and P. Bühlmann (2013, 11). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika 101*(1), 219–228.

Peters, J., D. Janzing, and B. Schlkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Peters, J., D. Janzing, and B. Schölkopf (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, Volume 26.

Peters, J., J. M. Mooij, D. Janzing, and B. Schölkopf (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research 15*(58), 2009–2053.

Plis, S., D. Danks, C. Freeman, and V. Calhoun (2015). Rate-agnostic (causal) structure learning. *Advances in neural information processing systems*, 3303–3311.

Rajaguru, G. and T. Abeysinghe (2008). Temporal aggregation, cointegration and causality inference. *Economics Letters 101*(3), 223–226.

Rajaratnam, B. and J. Salzman (2013, October). Best permutation analysis. *J. Multivar. Anal. 121*, 193–223.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton Mathematical Series. Princeton University Press.

Runge, J., S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. Mahecha, J. Muñoz-Marí, E. V. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler (2019). Inferring causation from time series in earth system sciences. *Nature Communications 10*.

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances 5*(11).

Shajarisales, N., D. Janzing, B. Schoelkopf, and M. Besserve (2015). Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32nd International Conference on Machine Learning*, Volume 37 of *Proceedings of Machine Learning Research*, pp. 285–294.

Shimizu, S., P. O. Hoyer, A. Hyvärinen, and A. Kerminen (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research 7*(72), 2003–2030.

Spirtes, P. and C. Glymour (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review 9*(1), 62–72.

Swanson, N. R. and C. W. J. Granger (1997). Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association 92*(437), 357–367.

Teyssier, M. and D. Koller (2005). Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, USA, pp. 584–590. AUAI Press.

Tugnait, J. K. (2022). On sparse high-dimensional graphical model learning for dependent time series. *Signal Processing 197*, 108539.

White, H. and X. Lu (2010). Granger causality and dynamic structural systems. *Journal of Financial Econometrics 8*, 193–243.

Wirtinger, W. (1927). Zur formalen theorie der funktionen von mehr komplexen veränderlichen. *Mathematische Annalen 97*, 357–376.

Wolter, M. and A. Yao (2018a). Complex gated recurrent neural networks.

Wolter, M. and A. Yao (2018b). Fourier rnns for sequence prediction. *arXiv: Machine Learning*.

Yang, M., M. Q. Ma, D. Li, Y.-H. H. Tsai, and R. Salakhutdinov (2020). Complex transformer: A framework for modeling complex-valued sequence. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4232–4236.

Zhang, K., Z. Wang, J. Zhang, and B. Schölkopf (2015). On estimation of functional causal models: General results and application to the post-nonlinear causal model. *ACM Trans. Intell. Syst. Technol. 7*(2).

Zheng, X., B. Aragam, P. Ravikumar, and E. P. Xing (2018). Dags with no tears: Continuous optimization for structure learning. In *NeurIPS*.