Discovering Interactions in Multivariate Time Series

Alex Tank

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Emily B. Fox, Chair

Ali Shojaie

Don Percival

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Discovering Interactions in Multivariate Time Series

Alex Tank

Chair of the Supervisory Committee:
Professor Emily B. Fox
Department of Statistics

In large collections of multivariate time series it is of interest to determine interactions between each pair of time series. Classically, interactions between time series have been studied using linear vector autoregressive models. However, new methodology must be developed to determine time series interactions in settings that depart from the classical stationary linear model. For example, many time series interactions may be non-linear or non-stationary. Some time series datasets also undergo subsampling or mixed-frequency sampling, so that classical methods cannot be directly applied. Furthermore, many collections of time series are not real valued, but may consist of categorical or event time data. In this thesis we develop methodology for inferring time series interactions in five domains that demand methodology beyond the classical linear model for real valued, fully observed time series.

First, we explore a Bayesian framework for inferring graphical models of time series. The goal is to determine conditional independence relations between entire time series, which for stationary series, are encoded by zeros in the inverse spectral density matrix. We place priors on (i) the graph structure and (ii) spectral matrices given the graph. We leverage a Whittle likelihood approximation and define a conjugate prior—the hyper complex inverse Wishart—on the complex-valued and graph-constrained spectral matrices. Due to conjugacy, we analytically marginalize the spectral matrices and obtain a closed-form marginal likelihood of the time series given a graph.

Second, we take a regularized likelihood approach and formulate a convex estimation proce-

dure for the multiple transition distribution (MTD) model of multivariate categorical time series. Traditionally, the MTD model is plagued by a nonconvex objective, non-identifiability, and presence of many local optima. Our new convex formulation facilitates the application of MTD to high-dimensional multivariate time series using convex penalties. Our formulation also allows identifiability conditions to be stated and imposed. We further derive a novel projected gradient algorithm for optimization.

Third, we study identifiability and estimation of the structural vector autoregressive model under both subsampled and mixed frequency scenarios. We find that when the errors are non-Gaussian and independent, both the lagged linear effects and instantaneous causal effects are identifiable. This implies that the full directed acyclic graph structure of the dynamic causal model is identifiable under arbitrary subsampling and mixed frequencies. An expectation-maximization algorithm is developed for inference.

Fourth, we develop two penalized neural network models based on a multilayer perceptron (MLP) network and a recurrent long-short term memory (LSTM) network able to detect nonlinear Granger causality. In both cases, we add group or hierarchical group lasso penalties to the outgoing weights of an input, shrinking all weights of an input time series to zero when there is no Granger causality between two series. We find that both MLP and LSTM models give state-of-the-art performance for detecting Granger causal connections in the genomics DREAM challenge.

Finally, we develop an efficient linear time alternating direction method of multipliers algorithm to segment locally stationary multivariate time series. The efficiency of our algorithm relies on recasting the global problem of the algorithm in a state space form allowing the use of a fast Kalman filter-smoother algorithm for optimization.

Taken together, these projects provide new methodology for inferring interactions in multivariate time series across data types, sampling regimes, and model classes.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank many people for their support and guidance during my graduate studies. First, and foremost, I would like to thank my parents, David and Karen, and brother Spencer. Your support has been constant and unwavering, and I certainly would not have been able to complete my studies without your help.

I would also like to thank my advisor, Emily Fox, and collaborator, Ali Shojaie, for jointly guiding me through my graduate research. You have both supported me endlessly, provided countless feedback, and always helped keep me focused and on track. In particular, I appreciate how you both have gradually let me learn how to perform research and ask questions independently.

The members of the DYNAMODE lab have also been a constant source of help, feedback, and friendship. I have watched over the years as the lab has grown into its mature form, and am happy to have been there at the beginning. In particular, I'm grateful to have worked with Ian Covert, Nick Foti, Chris Xie, Alec Greaves-Tunnel, Yian Ma, Rahul Nadkarni, Chris Aicher, and Chris Glynn.

I would also like to thank other members of the Statistics Department that I have become friends with and grown with over the years: Rebecca Ferrell, Sean Jewell, Nilanjana Laha, Samson Koelle, and Anya Mikh.

Finally, I'd like to thank Jenny Choi, Yunqi Bu, Jason Xu, Brennan Vincent and Connie Fu for being my family in Seattle and my `s p i c y b o y z` for support and lifelong friendships.

DEDICATION

To my family.

Chapter 1
INTRODUCTION

Multiple time series analysis is a vast area of statistical research [130], spanning many applied domains such as econometrics, neuroscience, genetics, and computer vision. One central scientific question in any data set with multiple recorded time series is how the series in question *interact*. That is, if and to what extent do certain series influence the behavior of other series.

1.1 Types of Interactions

In the multivariate time series literature, modeling strategies for exploring interactions may be categorized into four broad groups: 1) Granger causality, 2) instantaneous correlation and causality, 3) frequency domain interactions, and 4) latent variable methods. Extensions of the first three of these concepts are considered in this thesis, but we discuss all four concepts for completeness.

1.1.1 Granger causality

By far the most popular type of interaction in multivariate time series, Granger causality [71] quantifies how one series influences the future evolution of another series. It was first developed for the case of bivariate time series, but has since seen application for much higher dimensional series [175, 140]. Granger causality is not a true statistical causality measure [91], but instead it is intimately linked to prediction; if the past of one series x_t is helpful in forecasting series y_t then we say x_t *Granger causes* y_t. It is an extremely popular approach with a vast literature and has been used extensively in neuroscience [102, 46], econometrics [130], and genetics [175].

The earliest approaches to inferring Granger causality leveraged vector autoregressive models (VAR), where the conditional mean of a time series given the history of all other series is given by a *linear model*. In these models, Granger causality between two series occurs if the coefficient

in the linear model for predicting the conditional mean of a series is nonzero [130]. A frequentist approach to inference formally proceeds by performing significance analysis on each linear coefficient for being nonzero. Typically, parameters are estimated using either maximum likelihood approaches or method of moments approaches, commonly referred to as the Yule-Walker equations [208, 130]. Asymptotic confidence intervals for the VAR model used in the significance analysis may be derived using martingale central limit theorems [130]. A Bayesian approach to Granger causality in linear models proceeds by placing priors on the linear interaction coefficients. These priors may be continuous Gaussian priors centered at zero [106], or discrete 'spike and slab' priors that place significant prior probability that each interaction coefficient is exactly zero [132, 107]. If the posterior probability places most of the probability mass around zero then no Granger causal interaction is inferred.

In higher dimensional series, L_1 penalized model selection approaches have become popular [43, 175]. These approaches add a penalty, or prior, onto the log-likelihood that bias many inferred interaction parameters of the maximum likelihood solution to be zero, indicating no Granger causality. L_1 penalized methods can be thought of as convex relaxations to the full combinatorial problem of searching over all possible interaction graphs [81]. A related approach using a sparse Danzig selector combined with the Yule-Walker equations has also been explored [77]. An important research focus in these high dimensional settings is bounding the convergence rates between the parameter estimates and the true parameter values. In the time series context, this problem is more difficult since the rates must take into account the auto-correlation, or *memory*, in the observations that are typically absent or ignored when the data are independent and identically distributed. Recently this has been approached by either using notions of α and β-mixing in stochastic processes [35] or by using concepts from the spectral analysis of time series [11].

An important question in all Granger causality analyses is the *time lags* at which an interaction between two series occur. Specifically, how far back in time must one go before the value of x_t no longer impacts the future of y_t. Correct lag specification is essential for determining Granger causality interactions, since lower lag models may incorrectly imply Granger causal interactions when none in fact exist [130]. When the specified lag is too short, then *confounding* may occur

since there are unobserved variables (higher lags) that when excluded from the analysis lead to erroneous causality judgments. When the specified lag is too large, estimation variance increases due to more parameters being estimated. Classical lag determination methods utilize analysis of variance (ANOVA), Bayesian information criteria (BIC) or Akaike information criteria (AIC) model selection approaches and all are used extensively in practice [130]. More sophisticated modern methods have employed convex penalties [140, 175]. Recently, hierarchical group lasso penalties have been employed that specifically select the lag of each Granger causal interaction in a high dimensional time series [140]. Finally, we note that Granger causality in VAR models may be additionally confounded by *latent* unobserved series. Recent work addresses this problem by fitting a sparse plus low rank penalty on the transition matrix to separate out the sparse Granger causal interaction matrix and the spurious interaction matrix due to the latent components [121].

Non-linear models may also be used to assess the extent of Granger causality in time series. In these models, the conditional mean of the time series becomes a *non-linear* function of observations at past lags [188, 97]. Some popular model based approaches are additive models [119, 188] and multivariate threshold autoregressive (TAR) models [191]. Additive models decompose the conditional mean of the series into a sum of univariate non-linear functions of past lags. TAR models are piecewise linear models where the particular linear dynamics are specified by the value of a threshold variable. TAR models were initially introduced as a parsimonious approach for modeling complex dynamics in discrete time dynamical systems [191], but have since been extended to the multivariate case for assessing changes in multivariate dynamics in differing regimes [192]. While TAR models discrete jumps in dynamics, in functional coefficient models the parameters in a VAR model change smoothly with respect to some other variable. When that variable is *time*, these models reduce to time varying VAR models and may be used to assess how Granger causality relationships vary over time [188]. These type of methods may be used to asses Granger causality in non-stationary settings. We develop a linear time algorithm for detecting these changing Granger causality structures in Chapter 8 and a neural network approach to identifying nonlinear Granger causality in Chapter 7.

While the above approaches model Granger causality in the conditional mean of a series, many

frameworks exist for modeling other moments and aspects of the autoregressive conditional distribution. The conditional variance of a series may be modeled using a generalized autoregressive conditional heteroscedasticity (GARCH) model [188, 130]. A multivariate extension, the multivariate GARCH may assess Granger causality in second order moments [130]. Granger causality with respect to the complete conditional distribution function itself has also been studied using autoregressive quantile regression [200]. These methods may be helpful for determining how the tail behavior of series interact. Copula based models have also been introduced for assessing interactions in non-Gaussian time series models [9].

1.1.2 Instantaneous correlation and causality

While Granger causality investigates how series influence each other from the past to the future, instantaneous correlation and causality quantify how series influence each other *contemporaneously*. Time series models that specify instantaneous causal interactions are referred to as *structural* time series models and are popular in econometric modeling [130]. The most common variety is the structural VAR model (SVAR), that augments the linear autoregression with a structural linear model [194, 94]. The structural linear model is a tool in causality analysis that models a cascade of linear effects between variables. In particular, the cascade follows a *directed acyclic graph* (DAG), which specifies a causal ordering to the causal effects [146]. In structural VAR models, it is common for the causal ordering to be specified before hand. In econometrics, for example, prior knowledge or econometric theory commonly suggests a particular ordering to the instantaneous effects [130]. The structural effects are given by a lower triangular matrix and may be learned simultaneously with the lagged autoregressive effects at inference time.

There has also been work on simultaneously learning the structural matrix and the causal ordering of the instantaneous effects. This line of investigation leverages related efforts in the non-time series context [174] and models the instantaneous innovations as non-Gaussian and independent. The non-Gaussianity leads to identifiability of both the autoregressive effects, the causal ordering of the structural effects, and the structural effects themselves [90]. In this thesis we study learning of the instantaneous causal effects under different missing data scenarios in Chapter 5.

Finally, finding latent structure in the instantaneous covariance matrix may also be informative about unobserved latent interactions. For example, recently [159] has proposed a time series clustering model that clusters time series according to their instantaneous correlation and with it analyzes latent structure in a housing price data set. These latent clusters provide information about which housing neighborhoods react similarly to endogenous effects to the housing market.

1.1.3 Frequency domain notions of interaction

In many applications, interactions in time series are more naturally expressed in the *frequency*, or spectral, domain [24]. The spectral approach decomposes interactions across the frequency range. This approach has become popular in neuroscience, where different brain regions interact and communicate across different frequency bands [23, 37]. The *spectral density matrix* is the frequency domain analog of the covariance matrix. Each non-diagonal entry is given by the *cross spectra* between two series, a metric defined at each frequency that provides the covariance in magnitude of a frequency component between two series. Normalizing these cross-spectra by the diagonal elements gives the *spectral coherence* between two series. Just as the inverse of the covariance matrix provides information about interactions and conditional independence in the i.i.d. non-time series case, the inverse of the spectral density matrix also provides information about interactions in the frequency domain. In particular, normalized entries in the inverse are known as the *partial spectral coherence* and have been used extensively for exploring brain networks [23]. Furthermore, they may be used to define and learn graphical models of time series. In particular, zero entries across all frequencies of the inverse spectral density matrix indicates conditional independence between two series. Approaches to time series graphical models have utilized hypothesis testing [40] and model search [7]. In this thesis we explore a Bayesian solution to this problem in Chapter 4.

The classical frequency domain notions of interaction explore interactions between the amplitudes of different frequency bands. However, we note that recently many researchers have begun to study different types of interactions in the frequency domain. For example, phase-amplitude interactions play an increasingly important role in understanding interactions between brain regions where the phase of a slowly oscillating brain region drives the changing amplitudes of particular

frequencies in another region. A parametric approach to studying this phenomenon has recently been explored [50].

There are strong links between frequency domain notions of interaction and the time lagged Granger causality notions. In fact, the original papers on Granger causality [70] defined many of the Granger causality metrics in terms of the spectral densities of the multiple time series. We explore these links further in detail in Chapter 2. In practice, the type of interaction, frequency domain or Granger causality, a practitioner utilizes will depend on the specifics of the application.

1.1.4 Latent Variable Models

Latent variable models of multivariate time series explain the correlations in a time series data set using a reduced set of latent, unobserved variables. They do not explicitly provide a map of interactions in a data set, but instead provide a smaller set of components that mediate potential interactions. The general modeling assumption is that most of the variability in the data set may be explained by a reduced set of factors that evolve over time. Discrete latent factors are modeled using hidden Markov models [154], while continuous multivariate latent components are typically handled using state-space models [78]. Linear state-space models may then be used to determine interactions in this latent space; this approach has recently been used for inferring brain connectivity from magnetoencephalography (MEG) data [205]. However, in general, these latent time series models represent a different class of modeling strategies than models that try to find specific interactions between series.

1.2 From Interactions to Networks

As the number of time series in a data set grows, the question of time series influence can be recast as learning a *network* of such interactions. This network of interactions gives a tight representation of the wiring diagram of interactions and influences, showing how information may be propagated through a complex system. Indeed, this network approach to interactions in time series has entered many fields simultaneously. In neuroscience it has been used to estimate task dependent connectiv-

ity between brain regions [182]. This analysis gives a picture of how information percolates across brain regions, and can give insight into which information pathways in the brain may be disrupted under mental disease [13]. In genetics, applying a network time series analysis to time course gene expression data gives a network of gene regulation, displaying which genes up or down regulate other genes [122, 128]. A statistical network analysis for i.i.d data has also been used to find *hub* genes that influence and regulate many other gene expression pathways [185]. In finance, a time series interaction analysis will find certain companies or sectors of the economy that influence other companies or sectors. This analysis may give insight into how stable the economy is to shocks to particular sectors of the economy.

The network approach to learning time series interactions is intimately linked to notions of *sparsity*, discussed briefly in Section 1.1.1. In the network time series context, the sparsity assumption means that many series do not influence or are not influenced by each other. Leveraging the sparsity assumption, when the true network is sparse, leads to improved efficiency in estimating the network time series structure. Learning sparse networks has been extensively studied in the graphical modeling literature, where either an undirected or directed graph may be learned from i.i.d. data. Indeed, there has been some work on framing the time series network learning problem in the language of classical graphical model structure learning [52, 53]. Similarly, in the machine learning community, the problem of learning a time series network is referred to as dynamic Bayesian structure learning [65].

1.3 Types of Multiple Time Series

Many of the original methodologies and notions developed for the interactions discussed in Section 1.1 were for real valued multiple time series all observed at the same sampling rate. However, as time series methodology has spread into new domains, many of these classical methodologies have been adapted to deal with both 1) different data types and 2) types of missing data. We discuss both directions.

1.3.1 Data Types

The original papers on Granger causality were developed for real valued data observed at discrete time intervals. However, in many applications the time series in question take on different data types such as binary, count, or even event times. For example, incidences of disease in different regions are naturally modeled as counts, while neural spike data may be either thought of as event time or binary data . To deal with count and binary data data types, Poisson or Binomial generalized autoregressive models may be used [75, 9]. These model the evolution of the natural parameter of the generalized linear model as a linear model of the past values of the time series. Since these models are convex, L_1 penalties may be utilized to select for Granger causality, just as in the typical VAR case. In Chapter 6 we develop methods for inferring interpretable Granger causality networks from multivariate categorical time series.

For event time data, the multivariate Hawkes model provides a powerful approach for inferring Granger causality networks in high dimensions [214]. Examples of multivariate event time data are the times each neuron fires in a neural population [160, 124] or the exact times that a person in a social network tweets [214]. Both Bayesian methods using spike-and-slab priors [124] and frequentist approaches using L_1 penalties [214] have been used successfully for inferring Granger causality networks in this setting.

1.3.2 Missing Data and Different Data Sampling Schemes

Many time series data sets also contain *missing data*. In the literature there are broadly three types of missing data in time series:

Discrete Time Random: Entries in each series are missing at random, but the joint observed and missing data occurs at discrete intervals.

Discrete Time Subsampled: All series are *subsampled* at the same rate. For example, some economic indicators might evolve at a weekly rate but we only observe them on the monthly scale.

Discrete Time Mixed Frequency: Entries in each series are *subsampled*, but at different rates. In the classical mixed frequency case, only a subset of series are subsampled at the same rate, and the other contain no missingness.

- **Continuous Time Irregular Sampling** Sampling times occur at random and may take on any real number.

Granger causality has been studied in the discrete, random sampling case [33] where Bayesian or EM algorithms are used to simultaneously fill in missing data and infer the parameters of the VAR models. VAR estimation under irregular sampling has also been approached with kernel methods [8] or using continuous time dynamical systems [31, 82]. Estimation of spectral properties of time series under irregular sampling has also been extensively studied using the Lomb-Scargle periodogram [125, 168].

In the classical mixed frequency case, Granger causality and structural causality have been explored using VAR models [58, 4]. Recently, however, it has been shown that VAR models are only generically identifiable under mixed frequencies from the first two observable moments [4]. In particular, non-identifiable models arise when there is very weak interactions between the series under consideration, precisely the case one wishes to test when doing a Granger causality analysis. Two common approaches to estimation are likelihood based methods [59] and the extended Yule Walker equations [30]. Likelihood methods result in non-convex objectives and typically require multiple optimization restarts while the extended Yule Walker equations are method of moments approaches with a unique solution; the EM methods are generally more efficient, however, since the extended Yule-Walker equations only utilize some of the observable moments. To our knowledge, spectral estimation in the mixed frequency setting has not been studied, perhaps due to the effect of aliasing [147].

Classically, subsampled data has been studied in econometrics using dissagregation methodologies [177]. Standard VAR models suffer more severely from identifiability problems in the subsampled case, implying it is in fact extremely difficult to estimate both Granger causality and instantaneous causality in this setting. Recently, identification of VAR model with no instanta-

neous correlation has been established when the series innovations are *non-Gaussian* and independent [69]. However, to our knowledge there has been no attempt at estimation of Granger causality or instantaneous causality in the subsampled case under correlated innovations. In Chapter 5 we develop methods for inferring instantaneous and Granger causality in both mixed frequency and subsampled time series.

1.4 Contributions and Thesis Outline

The work in this thesis builds on three of the time series interaction frameworks: Granger causality, instantaneous causality, and spectral approaches. We outline the proceeding chapters.

Time Series and Optimization Background In Chapter 2 we provide the technical background on modeling interactions in multivariate time series. In particular, we introduce the technical aspects of Granger causality, insantaneous causality, and the spectral approach to time series modeling. In Chapter 3 we introduce statistical and optimization techniques that are leveraged in later chapters to fit our time series models of interactions to data. We introduce sparsity promoting penalized estimation, proximal gradient algorithms, and the Kalman filter.

Bayesian structure learning In Chapter 4 we develop a Bayesian modeling framework for analyzing conditional independence relationships between time series. The methodology harnesses tools from the spectral approach to multivariate time series combined with Bayesian approaches to modeling graphical models.

Structural VAR models for subsampled and mixed-frequency data In Chapter 5 we develop an approach for assessing both Granger causality and instantaneous causality in subsampled and mixed-frequency time series. This work harnesses non-Gaussian errors to develop novel identifiability results and estimation approaches for VAR models in the subsampled and mixed frequency settings.

Granger causality for categorical time series In Chapter 6 we develop a novel approach to inferring Granger causality in multivariate categorical time series. This work harnesses a convex substitution trick of the mixture transition distribution [157] to scale inference to many series.

Neural Granger causality for nonlinear time series In Chapter 7 an approach for inferring Granger causality in non-linear series based on sparse neural network models is developed. We develop novel architectures for learning time series interactions based on multilayer perceptrons and recurrent networks.

Change point detection for interactions In Chapter 8 we introduce a linear time algorithm for detecting changing Granger causality structures in VAR models. This work uses a alternating direction method of multiplies algorithm where one step is solved using a Kalman smoother.

The research in this thesis was performed collaboratively with many people. The work in Chapter 4 was done jointly with Nick Foti and Emily Fox. The work in Chapter 7 is joint work with both Ian Covert, Nick Foti, Ali Shojaie, and Emily Fox. The work in Chapters 5, 6, and 8 were developed jointly with Ali Shojaie and Emily Fox.

Chapter 2

BACKGROUND: TIME SERIES MODELING

2.1 Introduction

In this section we introduce crucial concepts in multivariate time series that are essential for modeling and studying interactions. In particular, we introduce and compare the following notions:

1. Lagged covariance (Section 2.2)
2. Granger causality (Section 2.3)
3. Instantaneous correlation and causality (Sections 2.3 and 2.4)
4. Spectral coherence and the spectral density matrix (Section 2.5.1)
5. Partial spectral coherence and inverse spectral density matrix (Section 2.5.4).

We also provide background on statistical inference and optimization concepts crucial to our approaches in Chapter 3. Here we provide general background and discussion of existing concepts; specific background details are additionally provided for each main contribution chapter.

2.2 Stationary Time Series

Let $X = \{\ldots, x_{t-1}, x_t, x_{t+1}, \ldots\}$ denote a p-dimensional real valued time series such that $x_t \in \mathbb{R}^p$ $\forall t$. A time series is second order *stationary* if the first and second moments of the process are time invariant. Specifically, the mean of the process is constant over time,

$$E[x_t] = \mu, \tag{2.1}$$

for some mean $\mu \in \mathbb{R}^p$. The expectation here is over random draws of the entire process, X. Furthermore, the covariance between all x_t and x_{t-h} for all integers $h \in \mathbb{Z}$ only depends on h:

$$E[(x_t - \mu)(x_{t-h} - \mu)^T] = \Sigma_h. \tag{2.2}$$

h is referred to the *lag* and $\Sigma_h \in \mathbb{R}^{p \times p}$ is referred to as the lag-h covariance matrix. Equation (2.2) states that in second order stationary series the lag-h covariance matrix does *not* depend on t, but instead is constant across t. All time series considered in this thesis, except for those in Chapter 8, will be stationary.

The ijth entry of the lag covariance matrix, $\Sigma_{h,ij}$, gives the covariance between time series i and j at lag h. The lag-h covariance between two series provides the most basic metric of interaction, as it gives the amount of linear dependence between two series. From the definition given in Equation 2.2, $\Sigma_{-h,ji} = \Sigma_{h,ij}$, so that $\Sigma_{-h} = \Sigma_h^T$, implying it is sufficient to only consider the lag h covariances at positive lags. For independent series i and j, the covariance will be zero across all lags; a high covariance, relative to the variance of each series, suggests a strong level of interaction between series. In many natural processes, the covariance tends to decay with longer lags but may also show periodic patterns or local peaks at nonzero lags. These plots are useful because they may suggest not only if there is any interaction between two series, but also the extent of such interaction and at which lags it may occur.

The covariance matrices may be estimated directly from an observed series. Let $(x_1, \ldots, x_T)$ be an observed realization of a stationary time series. The covariance Σ_h may be estimated using the empirical covariances

$$\hat{\Sigma}_h = \frac{1}{T} \sum_{t=h+1}^{T} (x_t - \hat{\mu})(x_{t-h} - \hat{\mu})^T. \tag{2.3}$$

where $\hat{\mu}$ is the empirical mean of the process. Note that for a given lag h there are only $T - h$ observations that may be used to compute the covariance, thus covariance at higher order lags becomes more difficult and noisy to estimate due to less available data.

2.3 Granger Causality and VAR models

Granger causality quantifies the extent to which one time series is able to predict the future evolution of another time series. Granger causality for real valued stationary time series is classically studied using using vector autoregressive models (VAR) [130]. A kth order vector autoregressive model (VAR(k)) is a time series that obeys the following linear recursion

$$x_t = \sum_{h=1}^{k} A^h x_{t-h} + \epsilon_t, \tag{2.4}$$

where $A^h \in \mathbb{R}^{p \times p}$ is the lag-h *transition matrix* and ϵ_t is a white noise process such that $E[\epsilon_t] = 0$, $E[\epsilon_t \epsilon_{t+1}^T] = \Gamma$, and ϵ_t and ϵ_{t+h} are independent for all t and $h \neq 0$. The VAR model decomposes interactions into two components - the transition matrices A^h and the instantaneous covariance Γ.

We say series i *Granger causes* series j if $A^h_{ij} \neq 0$ at all time lags h. The ijth entry of the transition matrix at lag h, A^h_{ij}, denotes the linear effect of time series i on time series j at lag h conditional on the remaining series. Intuitively, if the linear interaction coefficient is zero across all lags, then series i is not useful in predicting series j.

Γ provides the *instantaneous* covariance. Its entries Γ_{ij} denote the instantaneous covariance between two series i and j, *conditional* on the past of all series. Instantaneous correlation and causality in VAR models are treated in more detail in Section 2.4.

2.3.1 Comparing Lagged Correlation and Transition Matrices in VAR models

Both the lagged covariances and transition matrices provide information about interactions between two series. However, the transition matrix provides information about direct interactions between series, while the covariance is a much more global measure. For example, there may be large covariance between two series i and j even when there is no Granger causality between them, i.e. $A^h_{ij} = A^h_{ji} = 0$. This may occur, for example, if there is a third series that Granger causes both series i and j.

A formal relation between the transition matrices, instantaneous covariance, and lagged co-

variance matrices is given by the *Yule-Walker* equations. To derive these equations we rewrite the VAR(k) model as a VAR(1) model, $\tilde{x}_t = (x_t^T, x_{t-1}^T, \ldots, x_{t-k}^T)^T$. The evolution of $\tilde{x}_t$ is given by

$$\tilde{x}_t = \mathbf{A}\tilde{x}_{t-1} + \tilde{\epsilon}_t \tag{2.5}$$

where $\tilde{\epsilon}_t = (\epsilon_t^T, \mathbf{0}_{pk}^T)^T$ and $\mathbf{0}_{pk}$ is a length pk vector of 0s, and

$$\mathbf{A} = \begin{pmatrix} A^1 & A^2 & \ldots & A^k \\ I_p & 0_p & \ldots & 0_p \\ 0_p & I_p & \ldots & 0_p \\ \vdots & \ldots & \ddots & \vdots \\ 0_p & 0_p & \ldots & I_p \end{pmatrix}, \tag{2.6}$$

where I_p is a p-dimensional identity matrix. Post-multiplying both sides of Equation (2.5) by $\tilde{x}_{t-1}^T$ and taking expectations gives

$$E[\tilde{x}_t \tilde{x}_{t-1}^T] = \mathbf{A}E[\tilde{x}_{t-1}\tilde{x}_{t-1}^T], \tag{2.7}$$

where we have further assumed that $E(\tilde{\epsilon}_t \tilde{x}_{t-1}^T) = 0$, i.e. there is no correlation between the innovations, ϵ_t, and the values of the time series before time t, x_{t-h}. Assuming that $E[\tilde{x}_{t-1}\tilde{x}_{t-1}^T]$ is full rank, the solution is given by

$$\mathbf{A} = E[\tilde{x}_t \tilde{x}_{t-1}^T]E[\tilde{x}_{t-1}\tilde{x}_{t-1}^T]^{-1}. \tag{2.8}$$

$E[\tilde{x}_{t-1}\tilde{x}_{t-1}^T]^{-1}$. and $E[\tilde{x}_t \tilde{x}_{t-1}^T]$ are block matrices with entries given by the lagged covariances of

the VAR(p) process,

$$E[\tilde{x}_{t-1}\tilde{x}_{t-1}^T] = \begin{pmatrix} \Sigma_0 & \Sigma_1 & \dots & \Sigma_{p-1} \\ \Sigma_{-1} & \Sigma_0 & \dots & \Sigma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{-p+1} & \Sigma_{-p+2} & \dots & \Sigma_0 \end{pmatrix} \tag{2.9}$$

and

$$E[\tilde{x}_t\tilde{x}_{t-1}^T] = \begin{pmatrix} \Sigma_1 & \Sigma_2 & \dots & \Sigma_{p-2} \\ \Sigma_0 & \Sigma_1 & \dots & \Sigma_{p-3} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{-p+2} & \Sigma_{-p+3} & \dots & \Sigma_1 \end{pmatrix}. \tag{2.10}$$

Finally, the instantaneous covariance matrix for a VAR(p) model is given by:

$$\text{vec}(\Gamma) = (I_{p^2} - \mathbf{A}) \otimes \text{vec}(E[\tilde{x}_{t-1}\tilde{x}_{t-1}^T]) \tag{2.11}$$

where I_{p^2} is an identity matrix of size p^2, and the vec operation stacks a matrices columns to form a vector. Equation (2.8) and (2.11) may be rearanged to solve for the lagged covariance matrices given $\mathbf{A}$ and Γ implying there is a one–to–one mapping between the lagged covariance matrices and parameters of a VAR model.

2.3.2 Estimation of Granger Causality in VAR models

Classical estimation in VAR models typically proceeds using either least squares estimation or the Yule-Walker equations. The least squares approach estimates each row of A^h in Equation (2.4) independently using standard linear regression of the conditional mean model. Let a_i be the ith row of the concatenated lag matrices $(A^1, \dots, A^h)$. Letting $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_\mathbf{p}, \dots, \tilde{\mathbf{x}}_{\mathbf{T-1}})$ be the design

matrix and $\mathbf{x}_i = (x_{p+1}, \ldots, x_T)^T$ the response, the least squares estimate of a_i is given by

$$\hat{a}_i = (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)^{-1}\tilde{\mathbf{X}}^T\mathbf{x}_i. \tag{2.12}$$

The instantaneous covariance is then estimated using the empirical covariance of the *residuals* of the process

$$\hat{\Gamma} = \frac{1}{T}\sum_{t=p+1}^{T} \hat{e}_t \hat{e}_t^T, \tag{2.13}$$

where $\hat{e}_t = x_t - \sum_{h=1}^{k} \hat{A}^h x_{t-h}$. The Yule-Walker estimate on the other hand is given by (2.8) to obtain an estimate $\hat{\mathbf{A}}$

$$\hat{\mathbf{A}} = \begin{pmatrix} \hat{\Sigma}_1 & \hat{\Sigma}_2 & \ldots & \hat{\Sigma}_{p-2} \\ \hat{\Sigma}_0 & \hat{\Sigma}_1 & \ldots & \hat{\Sigma}_{p-3} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}_{-p+2} & \hat{\Sigma}_{-p+3} & \ldots & \hat{\Sigma}_1 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_0 & \hat{\Sigma}_1 & \ldots & \hat{\Sigma}_{p-1} \\ \hat{\Sigma}_{-1} & \hat{\Sigma}_0 & \ldots & \hat{\Sigma}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}_{-p+1} & \hat{\Sigma}_{-p+2} & \ldots & \hat{\Sigma}_0 \end{pmatrix}^{-1}, \tag{2.14}$$

where the lagged covariances $\hat{\Sigma}_h$ are estimated using Equation (2.3). The instantaneous covariance is then estimated by plugging in the $\hat{\mathbf{A}}$ estimate and similar covariance estimates into Equation (2.11). The main difference between the least squares and Yule-Walker estimates is in differences in computing the lagged covariances – Yule-Walker estimates use all the available data to calculate the lagged covariances in Equation (2.3) while the least squares estimates only use the $x_{p:T}$ data set. Finally, to assess the significance of a Granger causality effect between two series, asymptotic confidence intervals may then be used to determine significance of each entry of $\hat{A}^h$ [130].

2.4 Structural VAR models

Structural VARs (SVAR) model causal interactions within the instantaneous covariance matrix, Γ [101]. In particular, SVAR methods assume a instantaneous linear effect between series at each

time step. Formally, an SVAR method posits the following regressive structure

$$x_t = \mathbf{B}x_t + \sum_{h=1}^{k} \mathbf{D}^h x_{t-h} + e_t \tag{2.15}$$

where $\mathbf{B}$ is the matrix of instantaneous causal effects and e_t is a white noise process where all components of e_t are also assumed independent, i.e. e_{ti} is independent of e_{tj} for all $i \neq j$ and $E(e_t e_t^T) = \Lambda$ where Λ is a diagonal matrix. The matrix $\mathbf{B}$ models the instantaneous effects of one series on the other. A zero entry in element $\mathbf{B}_{ij}$ indicates no instantaneous causal effect between series i and series j. The SVAR model may be rewritten in a VAR form as

$$x_t = \sum_{h=1}^{k} (I - \mathbf{B})^{-1} \mathbf{D}^h x_{t-h} + (I - \mathbf{B})^{-1} e_t. \tag{2.16}$$

Equation (2.16) suggests an interpretation of the parameters in a SVAR model in terms of the parameters of a VAR model. In particular, the VAR transition matrices are given by $\mathbf{A}^h = (I - \mathbf{B})^{-1}\mathbf{D}^h$. Furthermore, letting $\epsilon_t = (I - \mathbf{B})^{-1} e_t$ the instantaneous covariance is given by $\Gamma = (I - \mathbf{B})^{-1}\Lambda(I - \mathbf{B})^{-T}$.

For identifiability purposes, it is typically assumed that $\mathbf{B}$ is a lower triangular matrix with 0s on the diagonal, which implies that $(I - \mathbf{B})^{-1}$ is also lower triangular. The lower diagonal assumption implicitly assumes a *causal ordering* of the instantaneous effects. Since $\mathbf{B}$ is lower diagonal, the ordering of the variables determines which variables are allowed to have instantaneous causal effects on other variables. For example, if a variable comes after another variable in the ordering, then it is constrained to have zero instantaneous causal effect. In econometric applications where sVAR models are predominantly used, the causal ordering is informed by both prior knowledge and economic theory [101]. Further discussion is contained in Chapter 5, in the context of identifiability and estimation of the causal ordering and the parameters of a SVAR model in subsampled and mixed frequency time series.

2.5 Concepts from the Spectral Approach to Time Series

Many interactions in time series are more naturally expressed and analyzed in the *frequency*, or spectral, domain. Here, interactions are spread across frequency ranges and may be more relevant to particular scientific hypothesis than interactions typically inferred from a Granger causality analysis using VAR models.

2.5.1 The Spectral Density Matrix and Spectral Coherence

If we assume that the lagged covariance sequence for a stationary process is absolutely summable,

$$\sum_{h=-\infty}^{\infty} |\Sigma_h| < \infty, \tag{2.17}$$

then we may define the *spectral density matrix* $S(\lambda) \in \mathbb{C}^{p \times p}$, as the discrete Fourier transform of the set of lagged auto-covariance matrices:

$$S(\lambda) = \sum_{h=-\infty}^{\infty} \Sigma_h e^{-ih\lambda}. \tag{2.18}$$

for $\lambda \in [-\frac{1}{2}, \frac{1}{2}]$. The spectral density matrix is the central object in the spectral approach to time series analysis.

2.5.2 Spectral Representation Theorem

For interpretation purposes below it will be useful to also consider the spectral representation of stationary time series. Briefly, this theorem states that under certain mild conditions any draw from a stationary process, like the ones we have defined above, may be represented as:

$$x_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i\lambda} dZ(\lambda) \tag{2.19}$$

where $dZ(\lambda)$ is an independent increments processes whereby $E(dZ(\lambda)) = 0 \; \forall \lambda \in [-\frac{1}{2}, \frac{1}{2}]$, and the covariance of the process is given by the spectral density matrix, $E\left(dZ(\lambda)dZ(\lambda)^T\right) = S(\lambda)$. The cross covariance at distinct frequencies is zero: $E\left(dZ(\lambda)dZ(\lambda')^T\right) = 0 \; \forall \lambda \neq \lambda'$. Intuitively, this theorem states that we can think of our stationary process as being the sum of many sine and cosine functions where the coefficients for each term are independent across frequency.

Importantly, within a frequency, λ ,the covariance between different time series in the independent increments process need not be zero, i.e. $E\left(dZ(\lambda)_i dZ(\lambda)_j\right) \in \mathbb{C}$. This naturally leads to the definition of the *cross spectrum* between time series i and time series j at frequency λ as the ijth entry of the spectral density matrix, $S(\lambda)_{ij}$. Due to the spectral representation theorem we can interpret this quantity as the covariance in the independent increments process, providing a measure of association in magnitude between time series i and j at frequency λ. A normalized version of the cross spectrum is the *coherence*:

$$\gamma_{ij}(\lambda) = \frac{S(\lambda)_{ij}}{\sqrt{S(\lambda)_{ii}}\sqrt{S(\lambda)_{jj}}}, \tag{2.20}$$

with $|\gamma_{ij}| < 1$, a frequency domain analog of the correlation coefficient.

2.5.3 Estimation of the Spectral Density Matrix

Given a finite observed time series, $(x_1, \ldots, x_T)$, the most basic estimator of the spectral density matrix at frequency λ is given by the periodogram

$$P(\lambda) = \frac{1}{T} d(\lambda) d(\lambda)^*, \tag{2.21}$$

where $d(\lambda)$ is the discrete Fourier transform (DFT) of the observed time series at frequency λ

$$d(\lambda) = \sum_{t=0}^{T-1} x_{t+1} e^{-i\lambda t}. \tag{2.22}$$

In many applications the periodogram matrix is typically computed at the *Fourier frequencies* $\lambda_k = \frac{2\pi k}{T}$ for $k \in (0, \frac{T}{2})$. In this case we use the notation

$$P_k = P(\lambda_k) \tag{2.23}$$

$$d_k = d(\lambda_k) \tag{2.24}$$

to denote the periodogram and Fourier transform at the kth Fourier frequency.

While the periodogram estimator, $P(\lambda)$, is asymptotically unbiased, it is not consistent since its variance does not go to zero. The periodogram estimates at different frequencies, $P(\lambda)$ and $P(\lambda')$, are also asymptotically uncorrelated as $T \to \infty$.

To obtain a consistent estimator of the spectral density matrix a common frequentist method is to *smooth* the periodogram:

$$\hat{S}(\lambda_k) = \sum_{|j|<m} W_T(j) P_{k+j} \tag{2.25}$$

where P_k is the periodogram at frequency λ_k as introduced in the main text and $W_T(j) \geq 0$, $\sum_{|j|<m} W_T(j) = 1$ are a smoothing window for a length T series and m is the smoothing window parameter that controls the width of smoothing. This approach was used in the frequentist graph estimation frameworks in [95, 7, 40]. To ensure consistency as $T \to \infty$ we must have $m \to \infty$, $\frac{m}{T} \to 0$, and $\sum_{|j|<m} W_T(j)^2 \to 0$ [25]. The asymptotic variance of $\hat{S}_k$ scales as $\sum_{|j|\leq m} W_T^2(j)$, implying that the asymptotic effective sample size for a smoothed estimate of this form is $(\sum_{|j|\leq m} W_T^2(j))^{-1}$ [25]. The Daniell smoother corresponds to taking $W_T(j) = \frac{1}{2m+1}$ and has an intuitive (effective) sample size of $2m+1$, the size of the smoothing window. Intuitively, this holds asymptotically since as $T \to \infty$ the sample periodograms become independent at different frequencies implying a sample size of $2m+1$, the number of (asymptotically) independent samples.

2.5.4 The Inverse Spectral Density Matrix and Partial Coherence

The cross spectrum between time series x_{ti} and x_{tj} above is defined as the Fourier transformation of the lagged auto-covariance between i and j, $\Sigma_{h,ij}$, the ijth entry of Σ_h. Note that this measure is completely independent of the other time series under consideration, $X_{t,Z}$, where $Z = \{k; k \neq i, j\}$ since the covariance only depends on the two series under consideration. Instead, to get a context driven measure that quantifies association between time series conditional on the other series we must analyze the time series after removing the effects of the other time series, $X_{t,Z}$. With this goal in mind, we define the vector valued linear filter $d_i^*(k) \in \mathbb{R}^{p-2}, k \in \mathbb{Z}$ as that which minimizes the expected mean square loss between the time series x_i and a linear combination of X_Z:

$$d_i^* = \text{argmin}_{d_i} E\left(x_{ti} - \sum_{k=-\infty}^{\infty} d_i(t-k)^T X_{k,Z}\right)^2. \tag{2.26}$$

Note that this optimal predictor is with respect to the entire sequence, $X_{(-\infty,\infty),Z}$, not just the part of the sequence before time t. We now define the stationary processes ϵ_{ti} as the residual between series x_{ti} and the optimal linear estimate given X_Z:

$$\epsilon_{ti} = x_{ti} - \sum_{k=-\infty}^{\infty} d_i^*(t-k)^T X_{kZ} \tag{2.27}$$

and the associated bivariate auto-covariance sequence as

$$\gamma_{k,\epsilon_i\epsilon_j} = E\left(\epsilon_{ti}\epsilon_{(t+k)j)}\right). \tag{2.28}$$

The partial cross spectrum is then defined analogously to the cross spectrum as:

$$r_{ij}(\lambda) = \sum_{k=-\infty}^{\infty} \gamma_{k,\epsilon_i\epsilon_j} e^{-i\lambda k}. \tag{2.29}$$

Normalization gives the *partial coherence*,

$$\rho_{ij}(\lambda) = \frac{r_{ij}(\lambda)}{\sqrt{r_{ii}(\lambda)}\sqrt{r_{jj}(\lambda)}}. \tag{2.30}$$

Comparing Eqs. (2.28) and (2.30) to Eqs. (2.19) and (2.20) we see that the partial coherence between x_i and x_j is simply the coherence at frequency λ of their respective residual series, ϵ_i and ϵ_j, where the linear effects of the remaining series $X_{t,Z}$ across all time have been removed.

While intuitive, the above construction of the partial coherence provides little practical guidance for estimating the partial coherence from data. Fortunately, the spectral domain provides a cleaner approach via the entries of the *inverse* of the spectral density matrix $S(\lambda)$. In particular, we have the following result [40],

$$\rho_{ij}(\lambda) = \frac{(S(\lambda)^{-1})_{ij}}{\sqrt{(S(\lambda)^{-1})_{ii}}\sqrt{(S(\lambda)^{-1})_{jj}}}, \tag{2.31}$$

where $(S(\lambda)^{-1})_{ij}$ is the ijth entry of the inverse of the spectral density matrix. This result says that we can compute the partial coherence for all pairs of time series under consideration by normalizing the off diagonal terms of the inverse spectral density matrix, $S(\lambda)^{-1}$.

2.6 Comparison of Coherence, Partial Coherence, and Granger Causality in VAR models

To gain intuition about the relationship between all the interaction notions developed so far, in this section we compare them in the context of a VAR model. The spectral density matrix for the VAR(k) model defined in Section 2.3 is given by:

$$S(\lambda) = \frac{1}{2\pi}\left(I - A(\lambda)\right)^{-1}\Gamma\left(\left(I - A(\lambda)\right)^{-1}\right)^{*} \tag{2.32}$$

where $A(\lambda)$ is the Fourier transform of the transition matrix sequence, $A(\lambda) = \sum_{h=1}^{k} A^h e^{-i\lambda h}$, and $*$ indicates conjugate transpose. The partial coherence for this processes can then be determined

by forming the inverse spectral density matrix:

$$S(\lambda)^{-1} = 2\pi \left(I - A(\lambda)\right)^* \Gamma^{-1} \left(I - A(\lambda)\right). \tag{2.33}$$

To gain intuition, consider the simple case when $\Gamma = I$ and $k = 1$. The inverse spectral density is then given by

$$S(\lambda)^{-1} = 2\pi(I - Ae^{-i\lambda} - A^T e^{i\lambda} + A^T A), \tag{2.34}$$

where we can see that the ijth entry, $S(\lambda)^{-1}_{ij}$, depends on both A_{ij}, A_{ji} and $(A^T A)_{ij} = \sum_k A_{ki} A_{kj}$. Intuitively, this means that the partial spectral coherence depends on both the directional effects between series i and series j and the degree to which both series have similar directed effects on the remaining processes in the series. If the two series have lagged effects on a disjoint set of other time series, the term will be small, whereas if they effect the set of same time series with the same sign, the term will be large. This last cross product term formalizes the fact that the partial coherence is dependent on what other variables are under consideration: if more time series are considered, $\sum_k^p A_{ki} A_{kj}$ may change considerably. Also note that this last term is *not* frequency dependent but instead acts as a baseline across all frequencies considered. In summary, we conclude that in the VAR(1) model, the partial coherence quantifies both the degree to which two time series influence each other and the degree to which they similarly influence the remaining series.

To contrast partial coherence and coherence for the VAR model we can expand the matrix inverse in Equation (2.32):

$$S(\lambda) = \frac{1}{2\pi}(I + A(\lambda) + A(\lambda)^2 + A(\lambda)^3 \ldots)\Gamma(I + A(\lambda) + A(\lambda)^2 + A(\lambda)^3 \ldots)^* \tag{2.35}$$

which for the simple case of $k = 1$ and $\Gamma = I$ becomes

$$S(\lambda) = \frac{1}{2\pi}\left(I + \sum_{h=1}^{\infty}(A)^h e^{-ih\lambda} + \sum_{h=1}^{\infty}(A)^{h^T} e^{ih\lambda} + \sum_{h=1}^{\infty}\sum_{h'=1}^{\infty}(A)^h (A)^{h'^T} e^{-i\lambda(h-h')}\right). \tag{2.36}$$

We first focus on the interpretation of the second term, $\sum_{h=1}^{\infty}(A)^h_{ij} e^{-ih\lambda}$. For $k = 1$, we obtain the

negative of the first term in the partial coherence, quantifying the single lag effect from x_i to x_j between series. For $h = 2$, $(A)^2_{ij}$ quantifies the expected linear effect from x_j to x_i at two time steps in the future. To see this note that $(A)^2_{ij} = \sum_{m=1}^{p} A_{im}A_{mj}$, the sum over all weights in the path from j to i. Similarly, $(A)^3_{ij} = \sum_{m=1}^{p}\sum_{m'=1}^{p} A_{i'm}A_{mm'}A_{m'j}$, the expected linear effects of x_j onto x_i three time steps into the future. Continuing on in this manner, $(A)^h$ is the expected linear effect from series j to series i h steps into the future. Taken together, we then see the terms $\sum_{h=1}^{\infty} A^h e^{-ih\lambda} + \sum_{h=1}^{\infty} {A^h}^T e^{ih\lambda}$ as the Fourier transforms of the expected linear effects for $i \to j$ and $j \to i$, respectively.

The terms in the last sum can be compared to the A^TA term in the partial coherence. While $(A^TA)_{ij}$ quantifies the similarity in *future* effects for x_j and x_i, $(AA^T)_{ij} = \sum_{m=1}^{p} A_{mj}A_{mi}$ quantifies the similarity of how x_i and x_j are influenced by the state of the full time series at the *past* time step. Similarly, entries in $(A)^h{(A)^{h'}}^T$ quantify the similarity between how series i and j are affected by the states of the time series at time lags h and h', respectively. Taken together, we see the terms in the partial coherence look only at 1) the direct interactions between x_i and x_j and 2) the similarity of their direct effects while the coherence takes into account 1) the marginal interactions between series i and j mediated via other time series and 2) the similarity of how pairs of series i and j are influenced across time lags.

2.6.1 Example

As an example, we plot the magnitude of the coherence and partial coherence in Figures 2.1 and 2.2 for two different $p = 4$ VAR(1) processes with transition matrices:

$$A = \begin{pmatrix} .2 & .2 & 0 & 0 \\ 0 & .2 & 0 & .2 \\ 0 & 0 & .2 & .2 \\ 0 & 0 & 0 & .2 \end{pmatrix} \qquad \tilde{A} = \begin{pmatrix} .2 & 0 & 0 & 0 \\ .2 & .2 & 0 & 0 \\ 0 & 0 & .2 & 0 \\ 0 & .2 & .2 & .2 \end{pmatrix} = A^T. \tag{2.37}$$

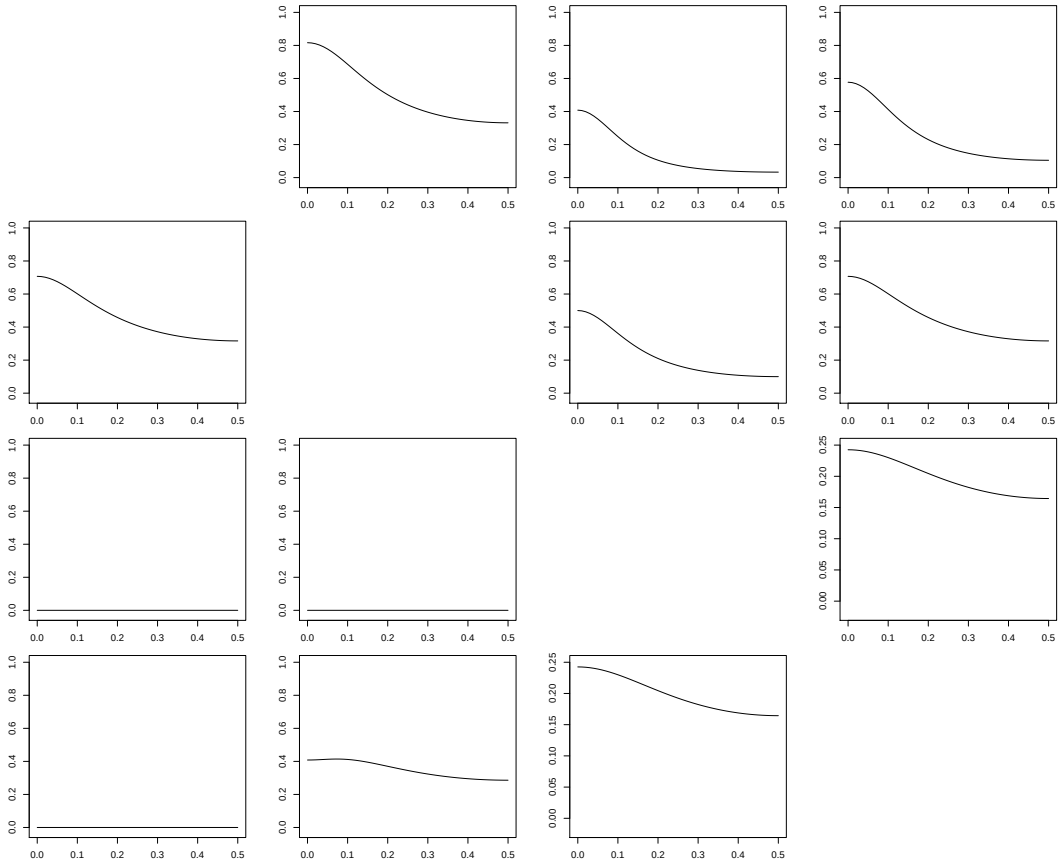


Figure 2.1: The magnitude of the partial coherence (*below diagonal*) and coherence (*above diagonal*) for the example VAR(1) time series in Equation (2.37). In all plots the x axis indicates frequency and y axis indicates the complex magnitude. The ijth plots is the magnitude of the coherence/partial coherence between x_i and x_j.

In figure 2.1 for A, we see nonzero coherence between x_3 and (x_1, x_2). This occurs because x_3 and (x_1, x_2) are driven by x_4. In Figure 2.2 for $\tilde{A}$ the coherence between x_3 and (x_1, x_2) are zero. This occurs because in $\tilde{A}$, x_3 and (x_1, x_2) neither Granger cause each other, nor have the same causal ancestors.

When we look at partial coherence we also see a large difference in the x_3 and x_2 relationship. The partial coherence is zero for A since x_3 and x_2 both do not directly influence each other nor influence the same other series, while for $\tilde{A}$ it is nonzero since they both influence x_4.

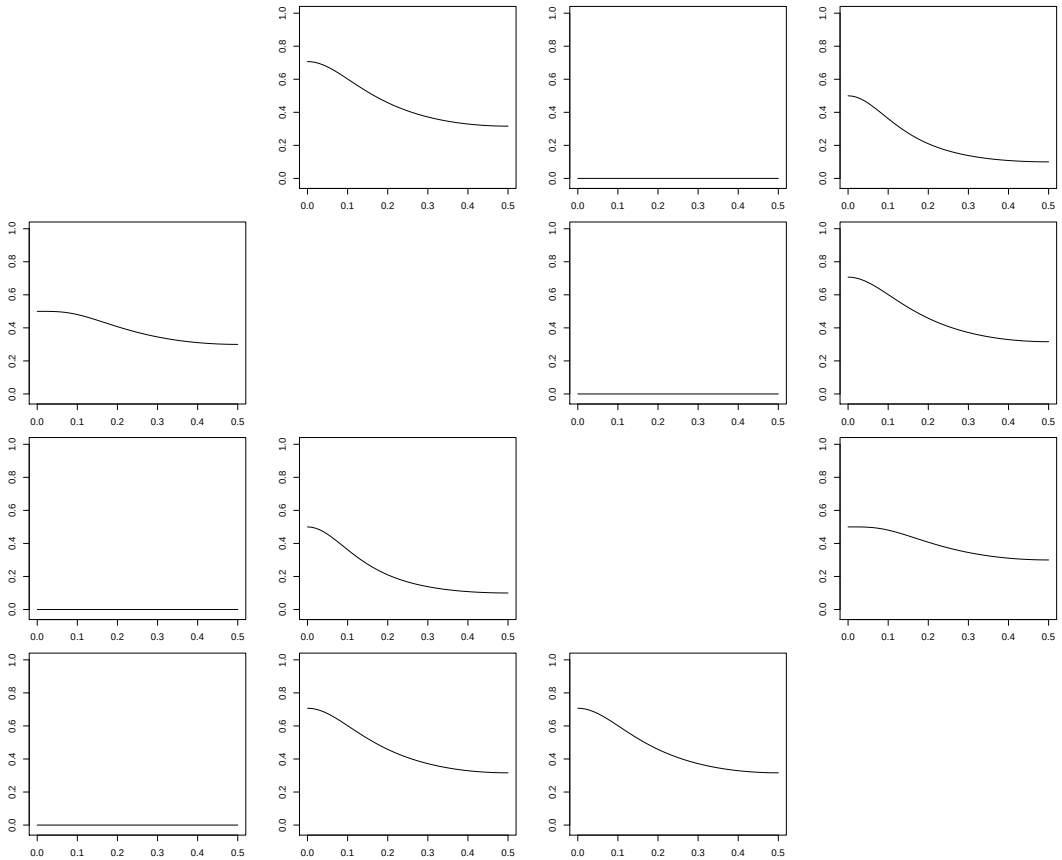


Figure 2.2: The magnitude of the partial coherence (*below diagonal*) and coherence (*above diagonal*) for the example VAR(1) time series in Equation (2.37). In all plots the x axis indicates frequency and y axis indicates the complex magnitude. The ijth plots is the magnitude of the coherence/partial coherence between x_i and x_j.

Chapter 3

BACKGROUND: STATISTICS AND OPTIMIZATION CONCEPTS

3.1 Introduction

The methodology in this thesis makes heavy use of optimization techniques. Numerical methods are required since estimators for interactions, like Granger causality, are solutions to optimization problems, typically with no closed form solutions. Numerical methods follow an algorithm that compute a path of solutions that converge to a local, and sometimes global, solution to the optimization problem. In this section we jointly discuss optimization and its use in time series modeling of interactions.

3.2 Maximum Likelihood

In this thesis many of the statistical estimators for time series interactions are given by the solution to a *maximum likelihood* problem. In this setting, one assumes a data generating model for the observed time series. Typically, the model will depend on some parameters θ contained in some set Θ, $\theta \in \Theta$. In our applications, θ will be parameters that specify which and to what extent series interact with each other. For example, in the VAR case $\theta = (A^1, \ldots, A^k)$ is the lagged transition matrices. Let $x_{1:T}$ denote the observed time series. The maximum likelihood estimator of θ is given by

$$\underset{\theta \in \Theta}{\text{minimize}} \quad -\log f(x_{1:T}|\theta) \tag{3.1}$$

where $f_\theta(x_{1:T}) = p(x_{1:T}|\theta)$ is the probability density function of the time series data generating model. In most of the chapters of this thesis we consider *autoregressive* time series models, the VAR model introduced in Chapter 2 being one example. For a lag K autoregressive process, the probability density function of the entire series is specified through the conditional distributions of

x_t given the past:

$$p(x_{1:T}|\theta) = p(x_{1:K}|\theta) \prod_{t=K+1}^{T} p(x_t|x_{t-K:t-1}, \theta). \tag{3.2}$$

In many cases, the term $p(x_{1:K}|\theta)$ is given by the stationary distribution of x_t under the parameters θ. For many models this term is more complicated, so it is common to approximate the full likelihood, $f_\theta(x_{1:T})$, with an approximate likelihood that simply conditions on the first K lags:

$$\underset{\theta \in \Theta}{\text{minimize}} \quad -\log g(x_{K+1:T}|\theta, x_{1:K}) \tag{3.3}$$

where $g(x_{K+1:T}|\theta, x_{1:K}) = \prod_{t=K+1}^{T} p(x_t|x_{t-K:t-1}, \theta)$.

In some cases the optimization of these problems may be given in closed form; however, in more complicated models numerical methods are required to compute an approximate solution. The maximum likelihood estimator in many time series models has a number of desirable properties, such as consistency and asymptotic normality [130].

3.3 Regularized Maximum Likelihood

In many settings one has *prior knowledge* about properties of the underlying parameter θ. In this case one may bias the solution of the maximum likelihood problem to parts of the parameter space that represent the prior knowledge through a regularization function $\Omega(\theta)$. Low values of $\Omega(\theta)$ encode more probable states of θ under prior information and high values less probable states. The final estimator for θ is given by a *regularized* maximum likelihood problem

$$\underset{\theta \in \Theta}{\text{minimize}} \quad -\log f(x_{1:T}|\theta) + \Omega(\theta), \tag{3.4}$$

and a regularized approximate liklelihood problem when the first K observations are only conditioned on

$$\underset{\theta \in \Theta}{\text{minimize}} \quad -\log g(x_{K+1:T}|\theta, x_{1:K}) + \Omega(\theta). \tag{3.5}$$

A common prior assumption in time series interaction modeling is that interactions encoded in the transition matrices A are sparse, i.e. that only a small number of series interact with each other. In this case $\Omega(\theta)$ typically is given by the L_1 norm of the lagged set of transition matrices, $\Omega(\theta) = ||\theta||_1$. This norm encourages solutions to Problem 3.4 to be sparse. For a more involved discussion of sparsity assumption in regularized maximum likelihood problems and other sparsity inducing regularizers see [26].

We provide a classic example of an approximate regularized maximum likelihood for inferring Granger causality in VAR(k) models. Here, the L_1 norm penalty is used to encourage many entries in the lagged transition matrices, A^h, to be zero. The regularized approximate maximum likelihood problem corresponding to Problem (3.5) for selecting Granger causality in a VAR(k) model with independent Gaussian errors with fixed identity covariance matrix, $\Gamma = I$, is

$$\underset{A^1,\ldots,A^h}{\text{minimize}} \quad \sum_{t=k+1}^{T} \left\| x_t - \sum_{h=1}^{k} A^h x_{t-h} \right\|_2^2 + \lambda \sum_{h=1}^{k} \|A^h\|_1 \tag{3.6}$$

where $||A||_1 = \sum_{i,j} |A_{ij}|$ is the L_1 matrix norm and λ is a regularization parameter that controls the level of sparsity in the final solution; higher λ leads to a sparser solution. Sparsity in the entries of A^h means that only a few pairs of series Granger cause one another. If $A^h_{ij} = 0$ for all h then time series i does not Granger cause series j. Note that this problem may be solved independently for each row of $(A^1, \ldots, A^k)$. In particular, each independent problem is given by a classical LASSO regression [189], for which there are numerous fast and scalable algorithms. We describe and provide a proximal gradient solver for the sparse Granger causality problem in Section 3.4. In Chapter 6 we develop a similar penalized maximum likelihood formulation for inferring Granger causality networks in *categorical* time series.

Another common penalty for Granger causality selection in VAR models is the *group lasso* penalty [207], which shrinks the interaction coefficients between two series jointly across *all lags*. The group lasso penalized optimization problem for the approximate VAR likelihood with fixed

identity covariance matrix, $\Gamma = I$, is given by [128]

$$\underset{A^1,\ldots,A^h}{\text{minimize}} \sum_{t=1}^{T} \left\| x_t - \sum_{h=1}^{k} A^h x_{t-h} \right\|_2^2 + \lambda \sum_{i,j} \|(A_{ij}^1, \ldots, A_{ij}^k)\|_2 \tag{3.7}$$

where the group penalty is given by $||(A_{ij}^1, \ldots, A_{ij}^k)||_2$, the L_2 norm of the coefficients between series i and j. Solutions to Problem (3.7) will generally lead to sparser inferred Granger causality networks than those of Problem (3.6). This is because solutions to Problem (3.7) will generally have $A_{ij}^h = 0$ many (i, j) pairs across all lags h. We harness the power of group lasso penalties in time series modeling of interactions in Chapters 6, 7, and 8.

3.3.1 Convexity and Nonconvexity

In many cases the log-likelihood and regularization function are *convex* functions of the parameters θ. A convex function g is one that may be written as

$$g\left(\gamma\theta_1 + (1-\gamma)\theta_2\right) \leq \gamma g(\theta_1) + (1-\gamma) g(\theta_2) \tag{3.8}$$

for all $\gamma \in (0, 1)$ and $\theta_1, \theta_2 \in \Theta$. The regularized VAR optimization in Problem (3.6) is an example of a convex regularized maximum likelihood problem. For more information on properties of convex functions see [20].

Convexity confers many benefits for statistical modeling and inference. For one, locally optimal solutions to the regularized maximum likelihood problem coincide with globally optimal solutions. This implies that one may trust that the final solution truly is the maximum likelihood solution. In the non-convex setting, multiple random restarts are typically required to attain a good locally optimal solution. However, still one may not obtain the global optima; extensive simulations are required to show that in many simulation settings random restarts are sufficient at finding the global solution and this is done in Chapters 4 and 5. The problem with locally optimal solutions is that they may not represent the true data generating parameter, and in many cases do not have any of the statistical guarantees of the maximum likelihood solution. In an applied setting, a local

optima may provide very misleading information about what interactions are occurring within a multivariate time series.

Second, many powerful algorithms exist for optimizing convex functions with convergence guarantees that do not exist when the function is non-convex. Below we discuss one type of optimization algorithm that we use in this monograph.

3.4 Proximal Gradient Descent

In Chapter 6 we utilize two instances of *proximal gradient* algorithms. Proximal gradient algorithms are applicable to optimizing functions of the following form

$$\underset{\theta}{\text{minimize}} \;\; h(\theta) + \lambda g(\theta), \tag{3.9}$$

where $h(\theta)$ is a differentiable convex function and $g(\theta)$ is a convex, but not necessarily differentiable function. Given an initial parameter estimate $\theta^{(0)}$, proximal algorithms first take a step in the direction of the gradient $\nabla h(\theta^{(0)})$, then perform a *proximal mapping* with respect to the function g. Specifically, the proximal gradient algorithm iterates the following recursion starting with $l = 0$:

$$\theta^{(l+1)} = \text{Prox}_{\lambda g}\left(\theta^{(l)} - \gamma^{(l)} \nabla h(\theta^{(l)})\right) \tag{3.10}$$

where $\nabla h(\theta^{(l)})$ is the gradient with respect to the smooth function h at $\theta^{(l)}$, $\gamma^{(l)}$ is a stepsize at iteration l. $\text{Prox}_{\lambda g}(\tilde{\theta})$ is the proximal map for g at $\tilde{\theta}$ given by

$$\underset{\theta}{\text{minimize}} \;\; \lambda g(\theta) + \left|\left|\tilde{\theta} - \theta\right|\right|_2^2. \tag{3.11}$$

Proximal algorithms are popular for sparse regularized maximum likelihood problems since typically the log–likelihood is a smooth function of the parameters, while the sparse regularizer is not smooth with a closed form proximal mapping. As an example, the L_1 norm regularizer used to promote sparse Granger causality estimates in the VAR transition matrix has a proximal mapping over each individual entry of the transition matrices and is given by *soft thresholding* on each element

of θ

$$\text{Prox}_{\lambda||.||_1}(A_{ij}^h) = \text{soft}(A_{ij}^h, \lambda) \tag{3.12}$$

$$= (\theta - \lambda \text{sign}(\theta))_+ \tag{3.13}$$

where λ is the L_1 regularization parameter and $(.)_+$ is the thresholding function

$$(x)_+ = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases} \tag{3.14}$$

The proximal mapping of the group lasso penalty used in Problem 3.7 decomposes independently over each vector θ in a group and is given by a group soft thresholding on θ:

$$\text{Prox}_{\lambda||.||_2}(\theta) = \text{gsoft}(\theta, \lambda). \tag{3.15}$$

$$= \left(1 - \frac{\lambda}{||\theta||_2}\right)_+ \theta \tag{3.16}$$

A full proximal gradient algorithm for solving the sparse VAR model for Granger causality estimation is given in Algorithm 1 and a similar one for the group lasso is given in Algorithm 2.

Algorithm 1 Proximal gradient algorithm for Granger causality selection for VAR(k) with lasso penalty (Problem 3.6).

Data: $x_{1:T}$
Result: $A^{1*}, \ldots, A^{k*}$
Initialize $A^{1,(0)}, \ldots, A^{1,(0)}$ **for** $i \in \{1, \ldots, p\}$ **do**
  $l = 0$ **while** $A_{i:}^{1,(l)}, \ldots, A_{i:}^{k,(l)}$ *not converged* **do**
    $\nabla L(A^{(l)})_{i:}^h = \sum_{t=0}^{T-1} 2(x_{ti} - \sum_{h=1}^k A_{i:}^{h,(l)} x_{t-h}) x_{t-h}^T$, $\forall h$ determine γ^l by line search [144]
    $\tilde{A}_{i:}^h = A_{i:}^h - \gamma^{(l)} \nabla L(A^{(l)})_{i:}^h \; \forall h$ **for** $j \in \{1, \ldots, p\}$ *and* $h \in \{1, \ldots, k\}$ **do**
      $A_{ij}^{h(l+1)} = \text{soft}(\tilde{A}_{ij}^h, \gamma^l \lambda)$
    end
    $l = l + 1$
  end
end

Algorithm 2 Proximal gradient algorithm for Granger causality selection for VAR(k) with group lasso penalty (Problem 3.7) .

Data: $x_{1:T}$
Result: $A^{1*}, \ldots, A^{k*}$
Initialize $A^{1,(0)}, \ldots, A^{1,(0)}$ **for** $i \in \{1, \ldots, p\}$ **do**
  $l = 0$ **while** $A_{i:}^{1,(l)}, \ldots, A_{i:}^{k,(l)}$ *not converged* **do**
    $\nabla L(A^{(l)})_{i:}^{h} = \sum_{t=0}^{T-1} 2(x_{ti} - \sum_{h=1}^{k} A_{i:}^{h,(l)} x_{t-h}) x_{t-h}^{T} \quad \forall h$ determine $\gamma^{(l)}$ by line search [144]
    $\tilde{A}_{i:}^{h} = A_{i:}^{h} - \gamma^{(l)} \nabla L(A^{(l)})_{i:}^{h} \; \forall h$ **for** $j \in \{1, \ldots, p\}$ **do**
      $(A_{ij}^{1}, \ldots, A_{ij}^{h})^{(l+1)} = \text{gsoft}\left((\tilde{A}_{ij}^{1}, \ldots, \tilde{A}_{ij}^{k}), \gamma^{(l)}\lambda\right)$
    end
    $l = l + 1$
  end
end

When g is an indicator function for a convex set C, then the proximal gradient algorithm reduces to the *projected gradient* algorithm and the proximal operator is simply the *projection* onto the set C given by

$$\underset{\theta \in C}{\text{minimize}} \left|\left|\tilde{\theta} - \theta\right|\right|_2^2. \tag{3.17}$$

When the set C admits a fast projection method, projected gradient methods can be very efficient solvers. We use a projected gradient algorithm in Chapter 6.

For convex problems, and under certain conditions on the step size choice γ_l, both projected and proximal gradient algorithms converge to the globally optimal solution. Details on step size selection can be found in [144].

3.5 State-Space Models and the Kalman Filter-Smoother

Many optimization and inference routines for time series models that are used in this thesis (Chapters 5 and 8) leverage the powerful modeling framework of Gaussian state-space models. State-space models introduce a m-dimensional latent time series, $z_t \in \mathbb{R}^m$, that explains the observed series y_t. State-space models may be used to model complex, higher dimensional dynamics, smooth time series trends, low dimensional structure, and missing data [78, 51]. In this thesis, the state-space framework is leveraged for computational benefits, namely linear time posterior inference of

the latent z_t sequence given the observed x_t sequence.

Formally, a Gaussian state-space model is given by

$$z_t = A_t z_{t-1} + \mu_t + \eta_t \tag{3.18}$$

$$y_t = C_t z_t + e_t, \tag{3.19}$$

where $A_t \in \mathbb{R}^{m \times m}$, is the state-space transition matrix, $C_t \in \mathbb{R}^{m \times p}$ is the output matrix, $\mu_t \in \mathbb{R}^m$ is a fixed input to the system. $\eta_t \sim N(0, Q_t)$ and $e_t \sim N(0, R_t)$ are the m and p-dimensional errors introduced to the latent series and observed series, respectively, where $Q_t \in \mathbb{R}^{m \times m}$ and $R_t \in \mathbb{R}^{p \times p}$ are the associated covariance matrices. In the most general form, the parameters $\theta = (A_t, C_t, \mu_t, Q_t, R_t)_{t=1}^T$ may all vary with time t.

Typically, we are interested in inferring the posterior distribution over the latent sequence $(z_1, \ldots, z_T)$ given an observed sequence $(y_1, \ldots, y_T)$ and the state-space parameters, θ, in a computational efficient manner. Since y_t and z_t are jointly Gaussian distributed, the conditional distribution of the z_t sequence given the x_t sequence is Gaussian. Thus, inference requires determining both the mean and covariance of the z_t sequence. Computing the first , $E(z_t|y_{1:T})$, and covariance, $\text{Cov}(z_t|y_{1:T})$ may be performed using the Kalman filter-smoother in time linear in the length of the time series, T.

In the Kalman filter-smoother algorithm we first compute the *forward pass* followed by a *backward pass*. The forward pass updates are given recursively by the predict equations

$$\hat{z}_{t|t-1} = A_t \hat{z}_{t-1|t-1} + \mu_t \tag{3.20}$$

$$P_{t|t-1} = A_t P_{t-1|t-1} A_t^T + Q_t \tag{3.21}$$

$$K_t = P_{t|t-1} C_t^T (R_t + C_t P_{t|t-1} C_t^T)^{-1} \tag{3.22}$$

$$\hat{z}_{t|t} = \hat{z}_{t|t-1} + K_t (y_t - C_t z_{t|t-1}) \tag{3.23}$$

$$P_{t|t} = (I - K_t C_t) P_{t|t-1}, \tag{3.24}$$

where $\hat{z}_{t|t-1} = E(z_t|y_{1:(t-1)}, \theta)$, $\hat{z}_{t|t} = E(z_t|y_{1:t})$, $P_{t|t-1} = \text{Cov}(z_t|y_{1:(t-1)})$, and $P_{t|t} = \text{Cov}(z_t|y_{1:t})$.

The backward pass recursively computes the final posterior expectations and covariances

$$\hat{z}_{t|T} = \hat{z}_{t|t} + W_t(\hat{z}_{t+1|T} - \hat{z}_{t+1|t}) \tag{3.25}$$

$$P_{t|T} = P_{t|t} + W_t(P_{t+1|T} - P_{t+1|t})W_t^T, \tag{3.26}$$

where $W_t = P_{t|t}A_t^T P_{t+1|t}^{-1}$ and $\hat{z}_{t|t}$, $\hat{z}_{t+1|t}$, $P_{t|t}$, and $P_{t+1|t}$ were previously computed in the forward pass.The final posterior expectations and covariances are given by $E(z_t|y_{1:T}) = \hat{z}_{t|T}$ and $P_{t|T} = \text{Cov}(z_t|y_{1:T})$. A complete derivation for this algorithm may be found in [78]. This exact Kalman filter-smoother algorithm will be used in both Chapters 5 and 8.

Chapter 4

BAYESIAN STRUCTURE LEARNING OF STATIONARY TIME SERIES

4.1 Introduction

Probabilistic graphical models (PGMs)—which compactly encode a set of conditional independence statements—have become a defacto tool for defining probabilistic models over large sets of random variables. When faced with time series, dynamic Bayesian networks (DBNs) are commonly deployed and specify sparse between- and within-time dependencies, often encoded by a *template model* replicated across time to straightforwardly model the growing set of random variables [105]. Learning template models requires specifying the set of dependency lags to be considered [217, 180]. In many applications, one instead aims to infer conditional independence between entire data streams accounting for interactions at all possible lags, represented by a *time series graphical model* (TGM). For example, imagine recording brain activity from multiple regions of the brain over time. Inference of a TGM in this setting would provide insight into the functional connectivity of different brain regions, an object of substantial scientific interest [182, 136]. TGMs have also been applied to intensive care monitoring [64] and financial time series [181].

The pioneering work of Dahlhaus [40] introduced the concept of undirected graphical models for stationary time series. The key insight was to transform the series to the *frequency domain* and express the graph relationships in the resulting spectral representation. For jointly Gaussian stationary time series, Dahlhaus [40] showed that conditional independencies between time series are encoded by zeros in the inverse spectral density matrices. This result is the frequency-domain analog to Gaussian graphical modeling in the i.i.d. (non-time-series) setting, where zeros in the inverse covariance matrix, or *precision matrix*, encode the conditional independencies between observed dimensions [115]. Dahlhaus' insight was first exploited to perform independent hypothesis tests of conditional independence between each pair of time series [40], with more recent work

correcting for multiple comparisons [134, 203].

A likelihood-based approach leveraging the *Whittle approximation* [201] has also been introduced [7]. The Whittle approximation casts the likelihood in the frequency domain with terms depending on the spectral density matrices critical to TGM structure learning, and independently so across frequencies. One approach scores graphs using AIC [7]. A recent penalized likelihood variant [95] places a joint graphical lasso [41] across frequencies to enforce a common zero pattern in the spectral density matrices. A penalized likelihood approach restricted to finite vector autoregressive processes has also been considered [181].

We instead consider a Bayesian approach to TGM structure learning, with all the benefits garnered from the Bayesian paradigm, including modeling within a generative framework where information from multiple sources can integrated and combined with available prior knowledge. For example, neural data are notoriously noisy, and robust inferences often rely on integrating time series across multiple trials and individuals or recording platforms (e.g., EEG/MEG). Our approach also leverages the Whittle likelihood. We then introduce a novel hyper Markov law [44], the *hyper complex inverse Wishart* distribution, that serves as a conjugate prior for the spectral density matrices whose inverses have a zero pattern specified by a graph. For decomposable graphs, this formulation leads to a closed-form expression for the marginal likelihood of a multivariate time series given a graph. By placing a prior on graph structures, we achieve a fully Bayesian approach to TGM structure learning for stationary time series. For our graph prior, we consider a multiplicity correcting prior [27]. Our analytic expression for the marginal likelihood is critical to the practicality of our approach since we can avoid inference of the large set of high-dimensional, complex spectral density matrices. In particular, for a length T series of dimension p, there are T spectral matrices to consider, each of size $p \times p$. In the i.i.d. setting, inference of just a single $p \times p$ graph-constrained covariance matrix is challenging; in this setting, inference of the T such matrices is prohibitive.

Hyper Markov laws based on the hyper inverse Wishart are a popular tool for Bayesian graphical model selection in the i.i.d. setting [67, 93]. Indeed, many powerful Bayesian structure learning algorithms based on this framework have been developed, both for decomposable [171, 73] and

non-decomposable [6, 138] graphs. By framing TGM structure learning in this common framework, we are able to apply existing state-of-the-art inference machinery for standard structure learning to the time series case. In this Chapter we use the feature-inclusion stochastic search (FINCS) procedure [171] for inference in decomposable models; however, many other MCMC and search schemes may be used. Importantly, future computational advances in Bayesian inference for i.i.d. graphical models may be easily extended using our framework to the time series case.

We test our methods on data simulated from vector autoregressive models with randomly generated TGMs. Our approach reaches almost perfect TGM recovery as the length of the time series or number of independent replicates increases. We then demonstrate the utility of our methods on a global stock indices dataset and MEG neuroimaging data of auditory attention switching tasks. In both cases we find meaningful, intuitive structure in the data.

The Chapter is organized as follows. We provide background on graphical models and stationary time series in Sec. 4.2. Our proposed TGM method is in Sec. 4.3, first introduced in the context of multiple independent realizations and then adapted to perform efficient inference of the TGM from only a single realization. In Sec. 4.5, we discuss how existing Bayesian structure learning methods may be modified to fit our formulation. Simulated results are in Sec. 4.6, with our stock and MEG analyses in Secs. 4.7 and 4.8, respectively.

4.2 Background

4.2.1 Graphs

Let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \ldots, p\}$ and edge set E, where $E \subset \{(i, j) \in V \times V : i \neq j\}$. Nodes i and j are adjacent, or *neighbors*, if $(i, j) \in E$. A *complete graph* is one having $(i, j) \in E$ for every $i, j \in V$ and complete subgraphs $C \subset V$ are termed *cliques*. A triple of subgraphs (A, S, B) where $V = A \cup B$ and $S = A \cap B$ with S complete is called a *decomposition* if every path from a node in A to a node in B must pass through S, the *separator*. Recursively decomposing A and B in this fashion results in the *prime*

components of a graph. If the prime components are complete then the graph is *decomposable*. We let the sets $\mathcal{C} = \{C_1, \ldots, C_K\}$ and $\mathcal{S} = \{S_2, \ldots, S_K\}$ each denote the prime components and their separators, respectively, generated by the decomposition. For simplicity, we restrict our attention to decomposable graphs but stress that our formulation is extensible to the non-decomposable case (see Sec. 4.9).

4.2.2 Hyper Markov distributions

For a given set of of random variables X, with realization $x \in \mathcal{X}$, dimensionality p, and joint density $p(x)$, an undirected graphical model G can be constructed by stating that an edge $(i, j) \notin E$ if X_i and X_j are conditionally independent given the remaining variables, i.e. $X_j \perp\!\!\!\perp X_i | X_{Z_{ij}}$ where $Z_{ij} = V \setminus \{i, j\}$. If the graph is decomposable, the joint density decomposes over cliques and separators:

$$p(x) = \frac{\prod_{C \in \mathcal{C}} p(x_C)}{\prod_{S \in \mathcal{S}} p(x_S)} \tag{4.1}$$

where $p(x_A)$ for $A \subset V$ denotes the marginal distribution of the set of variables x_A.

A hyper Markov law [44] is a distribution over probability measures that is concentrated on distributions that obey the Markov properties specified by G. Examples include the hyper Wishart and hyper Dirichlet distribution [44, 67]. Such distributions have proven pivotal in Bayesian graphical modeling by serving as conjugate priors for the graph parameters conditioned on the graph structure G. For example, in Gaussian graphical models (GGMs), the hyper inverse Wishart distribution provides a conjugate prior for covariance matrices that obey a zero pattern in the precision, as specified by G. By integrating over the hyper Markov distribution, one can obtain the *marginal likelihood* of the data conditioned on the structure G alone.

4.2.3 Stationary time series

Let $X_t = (X_{t1}, ..., X_{Tp})^T \in \mathbb{R}^p$ for $t \in \mathbb{Z}$ be a multivariate Gaussian stationary time series such that:

$$E(X_t) = \mu \quad \forall t \in \mathbb{Z} \tag{4.2}$$

$$\text{Cov}(X_t, X_{t+h})) = \Gamma(h) \quad \forall t, h \in \mathbb{Z}. \tag{4.3}$$

A time series probabilistic graphical model (TGM), $G = (V, E)$, may be constructed by letting $(i, j) \notin E$ denote that the entire time series $X_i(:)$ and $X_j(:)$ are conditionally independent given the remaining collection of time series $X_{Z_{ij}}$ where $Z_{ij} = V \setminus \{i, j\}$. For the Gaussian stationary series we consider, one can show that conditional independence holds between time series iff [40]

$$\text{Cov}(X_{it}, X_{j(t+h)}|X_{Z_{ij}}) = 0 \ \ \forall h \in \mathbb{Z}. \tag{4.4}$$

The *spectral density matrix* of a stationary time series is defined as the Fourier transform of the lagged covariance matrices, $\Gamma(h) = \text{Cov}(X_t, X_{t+h})$:

$$S(\lambda) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-i\lambda h} \tag{4.5}$$

for $\lambda \in [0, 2\pi]$ and $S(\lambda) \in \mathbb{C}^{p \times p}$ and Hermitian positive semidefinite. The marginal dependencies between time series are captured by $S(\lambda)$, and from Eq. (4.5), $S(\lambda)_{ij} = 0$ for all $\lambda \in [0, 2\pi]$ iff $\Gamma(h)_{ij} = 0$ for all $h \in \mathbb{Z}$. Furthermore, conditional independence between Gaussian stationary time series holds iff

$$S(\lambda)^{-1}_{ij} = 0 \quad \forall \lambda \in [0, 2\pi], \tag{4.6}$$

implying that inferring zeros in the inverse spectral density matrices across frequencies equates with inferring the TGM structure [40].

4.2.4 The Complex Normal and Complex Inverse Wishart Distributions

The complex normal distribution is a generalization of the multivariate normal distribution to the complex domain. Let $Z \in \mathbb{C}^p$ be a complex random variable. Z is distributed as a complex normal distribution, $\mathcal{N}_c(0, \Sigma)$, with zero mean and complex Hermitian positive definite covariance matrix

$\Sigma \in \mathbb{C}^{p\times p}$ if it has density given by

$$p(z) = \frac{1}{\pi^p|\Sigma|}e^{-z^*\Sigma^{-1}z}, \tag{4.7}$$

where $z^* = \bar{z}^T$ denotes the conjugate transpose of z. If $Z \sim \mathcal{N}_c(0,\Sigma)$ then the distribution over Z can be represented equivalently as a joint distribution over the real and imaginary elements of $Z = X + iY$, $X, Y \in \mathbb{R}^p$

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N(0, \begin{bmatrix} \text{Re}\Sigma & -\text{Im}\Sigma \\ \text{Im}\Sigma & \text{Re}\Sigma \end{bmatrix}, \tag{4.8}$$

where $\text{Re}\Sigma$ and $\text{Im}\Sigma$ indicate the real and imaginary components of Σ, respectively. Thus we see that the real and imaginary components are independent iff $\text{Im}\Sigma = 0$. As in the non-complex case, the marginal likelihood of X_A for some subset of nodes $A \subseteq \{1,\ldots,p\}$, is given by $X_A \sim \mathcal{N}_c(0,\Sigma_A)$, where Σ_A is the matrix formed by selecting the rows and columns of Σ in A.

The conjugate prior distribution for Σ is given by the complex inverse Wishart, $\Sigma \sim IW_c(\delta, W)$, with degrees of freedom parameter $\delta > 0$ and centering matrix $W \in \mathbb{C}^{p\times p}$, Hermitian positive definite. Its density is given by

$$p(\Sigma|W,\delta) = B(W,\delta)|\Sigma|^{-(\delta+2p)}e^{-\text{tr}W\Sigma^{-1}} \tag{4.9}$$

with normalization constant

$$B(W,\delta) = \frac{|W|^{\delta+p}}{\pi^{\frac{p(p-1)}{2}}\prod_{j=1}^{p}(\delta+p-j)!}.$$

Note that we have used an alternative parameterization of the inverse Wishart distribution commonly used in the graphical modeling literature [67]. The marginal distribution of Σ_A where $A \subseteq \{1,\ldots,p\}$ is given by $\Sigma_A \sim IW_c(\delta, W_A)$.

4.3 A Bayesian Approach

There are two standard approaches to Bayesian inference in graphical models: (1) placing a prior that jointly specifies the graph structure and associated parameters or (2) placing a prior on graph structures and then a prior on parameters given a graph; both rely on specifying a likelihood model. We opt for the second approach and describe the various components in this section. At a high level, our methods combine existing Whittle likelihood based methods [7, 95] with the hyper Markov framework to Bayesian graphical modeling [93, 67]. In the context of our TGMs, we introduce a conjugate *hyper complex inverse Wishart* prior on graph-constrained spectral density matrices. By integrating out the spectral density matrices, we obtain a marginal likelihood of the time series given the graph structure, G, allowing us to straightforwardly leverage state-of-the-art computational methods for i.i.d. Bayesian structure learning.

4.3.1 Whittle likelihood

Let $\mathbf{X} = [X_1, \ldots, X_T]$, with $x_t \in \mathbb{R}^p$ a realization of a p-dimensional stationary Gaussian time series observed at T time points, and $\mathbf{X}_{1:N} = \{\mathbf{X}^1, \ldots, \mathbf{X}^N\}$ be the collection of N independent realizations. We move to the frequency domain by transforming each $\mathbf{X}^n$ using a discrete Fourier transform. Let $d_{nk} \in \mathbb{C}^p$ denote the discrete Fourier coefficient associated with the nth time series at frequency $\lambda_k = \frac{2\pi k}{T}$:

$$d_{nk} = \sum_{t=0}^{T-1} x_{n(t+1)} e^{-i\lambda_k t}. \tag{4.10}$$

The Whittle approximation [201] assumes the Fourier coefficients are independent *complex normal random variables* with mean zero and covariance given by the corresponding spectral density matrix $S_k = S(\lambda_k)$:

$$d_{nk} \sim \mathcal{N}_c(0, S_k) \quad k = 0, \ldots, T-1, \tag{4.11}$$

such that the likelihood of $\mathbf{X}_{1:N}$ is approximated as

$$p(\mathbf{X}_{1:N}|S_{0:T-1}) \approx \prod_{n=1}^{N}\prod_{k=0}^{T-1}\frac{1}{\pi^p|S_k|}e^{-d_{nk}^* S_k^{-1} d_{nk}}, \tag{4.12}$$

where $\frac{1}{\pi^p|S|}e^{-z^* S^{-1} z}$ is the density of a complex normal distribution, $\mathcal{N}_c(0, S)$, with $S \in \mathbb{C}^{p\times p}$ and Hermitian positive definite. The Whittle approximation holds asymptotically with large T [25, 24, 201]. This approximation has been used in the Bayesian context in [161, 109]

Recall that conditional independencies are encoded in the off diagonal elements of S_k^{-1}. If time series $X_i(t)$ and $X_j(t)$ are conditionally independent, then the Whittle approximation says that as T gets large the ith and jth elements of the Fourier coefficients d_{nk} are conditional independent across all frequencies. Thus, if G is decomposable, Eq. (4.12) can be rewritten as

$$p(\mathbf{X}_{1:N}|G, S_{0:(T-1)}) \approx \prod_{k=0}^{T-1}\frac{\prod_{C\in\mathcal{C}}\frac{1}{\pi^{N|C|}|S_{kC}|^N}e^{-\mathrm{tr}\tilde{P}_{kC}S_{kC}^{-1}}}{\prod_{S\in\mathcal{S}}\frac{1}{\pi^{N|S|}|S_{kS}|^N}e^{-\mathrm{tr}\tilde{P}_{kS}S_{kS}^{-1}}} \tag{4.13}$$

where

$$\tilde{P}_k = T\sum_{n=1}^{N} P_{nk} = T\sum_{n=1}^{N}\frac{1}{T}d_{nk}d_{nk}^* \tag{4.14}$$

is the scaled aggregate *periodogram* over the N time series at frequency $\frac{2\pi k}{T}$. For $A \subset V$, S_{kA} and $\tilde{P}_{kA}$ are the restriction of both matrices to the elements in A and $|A|$ denotes the cardinality of the set A.

4.3.2 Hyper complex inverse Wishart prior on graph-constrained spectral density matrices

We seek a prior for the spectral density matrices, S_k, whose inverses each have zeros dictated by a graph G. Recall that these S_k matrices are complex-valued and restricted to be Hermitian positive definite. As discussed in Sec. 4.2.2, the hyper inverse Wishart distribution serves as a prior for real-valued, positive-definite matrices with pre-specified zeros in the inverse, and is a conjugate

prior for the covariance of a zero-mean GGM. Motivated by the connection between GGMs and our TGMs, and the analogous structure of our TGM-based Whittle likelihood of Eq. (4.13) to that of a GGM with N i.i.d. observations, we propose a novel *hyper complex inverse Wishart* prior with density function

$$p(\Sigma|\delta, W, G) \propto \mathbf{1}_{\Sigma \in M^+(G)} |\Sigma|^{-(\delta+2p)} e^{-\mathrm{tr} W\Sigma^{-1}} \tag{4.15}$$

for *degrees of freedom* $\delta > 0$, *scale matrix* $W \in \mathbb{C}^{p\times p}$ positive definite and Hermitian, and graph G. We have used an analogous parameterization to that of the hyper inverse Wishart [44]. Here, $\Sigma \in M^+(G)$ denotes that Σ is in the set of all Hermitian positive-definite matrices with $(\Sigma^{-1})_{ij} = 0$ for all $(i, j) \notin E$. When G is decomposable, the normalization constant is available and the density decomposes over cliques and separators:

$$p(\Sigma|\delta, W, G) = \frac{\prod_{C\in\mathcal{C}} \mathrm{IW}_c(\Sigma_C|\delta, W_C)}{\prod_{S\in\mathcal{S}} \mathrm{IW}_c(\Sigma_C|\delta, W_C)} \tag{4.16}$$

$$= \frac{\prod_{C\in\mathcal{C}} B(W_C, \delta)|\Sigma_C|^{-(\delta+2|C|)} e^{-\mathrm{tr} W_C\Sigma_C^{-1}}}{\prod_{S\in\mathcal{S}} B(W_S, \delta)|\Sigma_S|^{-(\delta+2|S|)} e^{-\mathrm{tr} W_S\Sigma_S^{-1}}}, \tag{4.17}$$

where IW_c denotes the complex inverse Wishart [24] detailed Section 4.3.4 with normalizer

$$B(W, \delta) = \frac{|W|^{\delta+p}}{\pi^{\frac{p(p-1)}{2}} \prod_{j=1}^{p}(\delta + p - j)!}. \tag{4.18}$$

We denote our proposed prior as $HIW_c(\delta, W, G)$ and specify

$$S_k \mid G \sim HIW_c(\delta_k, W_k, G) \quad k = 0, \ldots, T-1. \tag{4.19}$$

In Section 4.3.4, we show that this prior specification is *conjugate* to the TGM-based Whittle likelihood of Eq. (4.13). Also note that the graph, G, is shared across all frequencies.

4.3.3 *Marginal likelihood*

Due to conjugacy of our proposed hyper complex inverse Wishart prior, the marginal likelihood of the time series $\mathbf{X}_{1:N}$ given a decomposable graph G, integrating out the spectral density matrices $S_{0:T-1}$, has a closed form which is derived in Section 4.3.4 and given by

$$p(\mathbf{X}_{1:N}|G) \approx \pi^{-NTp} \prod_{k=0}^{T-1} \frac{h(W_k, \delta_k, G)}{h(W_k^*, \delta_k^*, G)}. \tag{4.20}$$

Here, $\delta_k^* = \delta_k + N$, $W_k^* = W_k + P_k$, and

$$h(W, \delta, G) = \frac{\prod_{C \in \mathcal{C}} B(W_C, \delta)}{\prod_{S \in \mathcal{S}} B(W_S, \delta)}. \tag{4.21}$$

From the definition of δ_k^*, we see that N, the number of time series, acts as the effective number of observations in this case. For the i.i.d. GGM, N represents the number of independent vector-valued observations; in our TGM, N plays the same role, but represents the number of independent *time series* observations. Likewise, as in standard inverse Wishart based modeling of covariances for i.i.d. Gaussian data, based on a set of N i.i.d. complex normal observations of Fourier coefficients d_{nk} with covariance S_k (see Eq. (4.12)), we update the prior scale matrix W_k with the outer product $P_k = \sum_{n=1}^{N} d_{nk} d_{nk}^*$, which is the aggregate *periodogram* (see Eq. (4.14)).

Having an analytic marginal likelihood of the time series given a PGM allows us to perform inference directly over graphs, sidestepping any thorny issues with inference directly on the T spectral matrices, each of size $p \times p$, themselves. This is a critical feature of the practicality of our approach.

4.3.4 *Complete details on the complex hyper inverse Wishart distribution and full derivation of the marginal Whittle likelihood*

Derivation of the Marginal Likelihood for the complex hyper inverse Wishart We have defined the hyper-complex inverse Wishart distribution for a graph $G = \{V, E\}$ as the restriction of the complex inverse Wishart distribution to $\Sigma \in \mathbb{C}^{p \times p}$ with a zero pattern in Σ^{-1} specified by G.

Its density is given by:

$$p(\Sigma|\delta, W, G) = \mathbf{1}_{\Sigma \in M^+(G)} h(W, \delta, G) |\Sigma|^{-(\delta+2p)} e^{-\mathrm{tr} W\Sigma^{-1}} \tag{4.22}$$

where $h(W, \delta, G)$ is a normalization constant and $M^+(G)$ is the set of positive definite matrices with zeros in their inverse that obey the conditional independence properties of G.

Due to the fact that the complex inverse Wishart distribution is conjugate to the complex normal distribution for an unrestricted Σ, by Proposition 5.1 in [44] it follows that the hyper complex inverse Wishart distribution is a strong hyper-Markov distribution. It follows that for decomposable G the complex hyper inverse Wishart density can be written in terms of the cliques, $\mathcal{C}$, and separators, $\mathcal{S}$, of G:

$$p(\Sigma|\delta, W, G) = \mathbf{1}_{\Sigma \in M^+(G)} \frac{\prod_{C \in \mathcal{C}} p(\Sigma_C|W_C, \delta)}{\prod_{S \in \mathcal{S}} p(\Sigma_S|W_S, \delta)} \tag{4.23}$$

where $p(\Sigma_C|W_C, \delta)$ is the unrestricted complex inverse Wishart density for Σ_C. This decomposition implies that the normalization constant for Equation (4.22) is also given by the ratio of complex inverse Wishart normalization constants for cliques and separators

$$h(W, \delta, G) = \frac{\prod_{C \in \mathcal{C}} B(W_C, \delta)}{\prod_{S \in \mathcal{S}} B(W_S, \delta)}. \tag{4.24}$$

If $Z_1, ...Z_N \overset{i.i.d.}{\sim} \mathcal{N}_c(0, \Sigma)$, then the joint distribution of $Z_1, \ldots, Z_N$, and Σ can be written as:

$$p(z_1, \ldots, z_N, \Sigma|G, W, \delta) \propto \mathbf{1}_{\Sigma \in M^+(G)} \frac{h(W, \delta, G)}{\pi^{Np}} |\Sigma|^{-(\delta+N+2p)} \quad e^{-\mathrm{tr}(W + \sum_{i=1}^N z_i z_i^*)\Sigma^{-1}}. \tag{4.25}$$

Since the part dependent on Σ is the kernel for a $HIW_c(W + \sum_{i=1}^N z_i z_i^*, \delta + N, G)$ distribution, it follows that the marginal distribution of $Z_1, \ldots, Z_n|G, W, \delta$ is given by the ratio of prior and posterior normalization constants of the complex hyper inverse Wishart distribution times a likelihood constant:

$$p(z_1, \ldots, z_n|G, W, \delta) = \frac{h(W, \delta, G)}{\pi^{Np} h(W + \sum_{i=1}^N z_i z_i^*, \delta + N, G)}. \tag{4.26}$$

Marginal Likelihood for the full Whittle Likelihood The model in the main text places independent $HIW_c(W_k, \delta_k, G)$ priors on each spectral density matrix in the Whittle likelihood, $S_k \sim HIW_c(W_k, \delta_k, G)\ \forall k \in [T-1]$. Applying the above marginal likelihood result to each frequency component in the Whittle approximation shows that the marginal likelihood of the data given a graph, a set of centering matrices ,$W_0, \ldots, W_{T-1}$, and degrees of freedom, $\delta_0, \ldots \delta_{T-1}$, for each frequency can be approximated by a product of the normalization constants across frequencies:

$$p(\mathbf{X}_{1:N}|G) \approx \pi^{-NTp} \prod_{k=0}^{T-1} \frac{h(W_k, \delta_k, G)}{h(W_k^*, \delta_k^*, G)}. \tag{4.27}$$

where $W_k^* = W_k + P_k$ and $\delta_k^* = \delta_k + N$. Indeed, this derivation shows that our prior specification for spectral density matrices is conjugate to the entire Whittle likelihood.

4.3.5 Fractional priors for model selection

Marginal likelihoods used for model comparison [99] are notoriously sensitive to the choice of prior parameters, or *hyperparameters*. In our case, the marginal likelihood in Eq. (4.27) depends strongly on the hyper complex inverse Wishart scale matrix, W_k. Since the scale and shape of the spectral density matrices are not known a priori, and vary dramatically across frequencies, we employ *fractional priors* [142] over each S_k. Fractional priors effectively hold out some fraction of the data, and utilize that fraction to determine an adequate hyperparameter setting for each model. The rest of the data are then used for model comparison. Fractional priors have been deployed for graphical model selection in i.i.d. graphs and have a number of desirable properties such as information consistency and demonstrated robustness [171]. In our case, under a fractional prior with parameter $g \in (0, 1)$, the fractional marginal likelihood is

$$p(\mathbf{X}_{1:N}|g, G) = \pi^{-NTp} \prod_{k=0}^{T-1} \frac{h(gP_k, gN, G)}{h(P_k, N, G)}. \tag{4.28}$$

Here, we see that g controls the fraction of data used for prior formulation versus model comparison. Importantly, we now have just a single, scalar, and interpretable parameter g to tune. Default

settings are suggested in [142, 171].

4.3.6 Graph prior

There are two common approaches in the literature to specifying a prior distribution on graphs. The first approach places a uniform distribution on the space of all possible graphs [67, 45, 162]. As noted in [5], this prior puts high weight on graphs with a medium number of edges and significantly less weight on graphs with small or many edges. In response to this problem, it has been proposed to place a prior directly on the size of the graph and then consider a conditionally uniform prior on all graphs of the same size [5, 47, 93]. We follow this later approach and place a binomial distribution on the number of edges, k:

$$p(G) \propto r^k (1-r)^{m-k}, \tag{4.29}$$

where r is the prior probability that each of $m = \frac{p(p-1)}{2}$ possible undirected edges $(i,j) \in V \times V$ is included. Since r is unknown, we further place a Beta(a,b) prior over r. This acts as a prior on our prior. Integrating out r gives the marginal prior over graphs

$$p(G) \propto \frac{\beta(a+k, b+m-k)}{\beta(a,b)} \tag{4.30}$$

where $\beta(.,.)$ is the beta function. As explored in [171], this is a multiplicity correcting prior [170] over graphs with the desirable property of diminishing false positive edge discoveries as extra unconnected nodes are added to the graph.

4.4 Methods for Single Time Series

In some applications of interest one observes only a single multivariate time series, $N = 1$, from which the graph must be inferred. Two challenges arise in this setting: (1) the effective number of observations informing Eq. (4.27) is just one and (2) the periodogram used in computing W_k^* is noisy regardless of the length of the series, T. The periodogram is a notoriously poor estima-

tor of the spectral density, and when the spectral density itself is of primary interest, a common frequentist method is to smooth the periodogram to obtain a consistent spectral density estimator [95, 7, 40]. One could imagine using the smoothed periodogram as a plug-in estimator in Eq. (4.27), scaled by the effective degrees of freedom, and we outline this approach at the end of this section for completeness. An alternative variance-reduction technique is the Bartlett method [10], that divides the length T series into M shorter series of length $\frac{T}{M}$ and averages the resulting M periodograms, but at the cost of reduced resolution (i.e., number of considered frequencies). This approach mimics the implicit smoothing that occurs when we compute the periodogram based on N truly independent series each of length T, as in Eq. (4.14).

In contrast to a plug-in estimator, a natural Bayesian approach enforces smoothing across frequencies via a prior distribution over the set of spectral densities [161]. Previous approaches have coupled elements of a Cholesky decomposition of each spectral density matrix across frequencies; however, this approach is unsuitable to our case since 1) it does not enforce sparsity in the inverse spectral density and 2) a prior of this form will remove the simple marginal likelihood structure in Eq. (4.27) that we harness for efficient inference. Motivated by our aims to both share information across frequencies and maintain the form of the marginal, we utilize a piecewise constant prior over spectral densities given a graph, G. We partition the interval $[0, \pi]$ into M intervals $w_1 = \left[0, \frac{\pi}{M}\right), \ldots, w_j = \left[\frac{\pi(j-1)}{M}, \frac{\pi j}{M}\right), \ldots, w_M = \left[\frac{\pi(M-1)}{M}, \pi\right]$ and then draw a separate positive definite Hermitian matrix from a HIW_c distribution for each interval:

$$\tilde{S}_j \sim HIW_c(\delta, W_j, G) \quad j = 1, \ldots, M. \tag{4.31}$$

Our resulting spectral density is simply

$$S(\lambda) = \sum_{j=1}^{M} \mathbf{1}_{\lambda \in w_j} \tilde{S}_j \quad \forall \lambda \in [0, \pi]. \tag{4.32}$$

Under this prior, the marginal likelihood for the single ($N = 1$) time series becomes

$$p(\mathbf{X}|G) \approx \pi^{-Mp} \prod_{j=1}^{M} \frac{h(W_j, \delta_j, G)}{h(W_j^*, \delta_j^*, G)} \tag{4.33}$$

where $\delta_j^* = \delta_j + \sum_{k=0}^{T-1} \mathbf{1}_{\lambda_k \in w_j}$ and $W_j^* = W_j + \sum_{k=0}^{T-1} \mathbf{1}_{\lambda_k \in w_j} P_k$. By setting $M = \lfloor \sqrt{T} \rfloor$, we obtain an asymptotically approximate nonparametric prior distribution over continuous spectral density matrices: for T large enough the prior puts positive support on spectral density matrices arbitrarily close to any continuous spectral density over $[0, 2\pi]$. Furthermore, under this setting as $T \to \infty$, the number of Fourier frequencies, and thus number of samples $\sum_{k=0}^{T-1} \mathbf{1}_{\lambda_k \in w_j}$, within each interval grows as $\sqrt{T}$.

Our approach is also amenable to general smoothing estimators for spectral densities as described in Section 2.5.3 of Chapter 2, although the Bayesian interpretation breaks down in this setting. This approach is inspired by the use of more sophisticated smoothing techniques in previous TGM procedures [95, 7, 40] and we develop a similar procedure tailored to our objective function in Eq. (4.27). We may plug in a smoothed estimate of the spectral density matrix, scaled by the asymptotic effective degrees of freedom, for the priodogram, P_k, in Eq. (4.27). Specifically, we set $W_k^* = W_k + (\sum_{|j| \leq m} W_T^2(j))^{-1} \hat{S}_k$. The degrees of freedom parameter δ_k^* is similarly updated by adding the effective sample size of the smoother to the prior degrees of freedom: $\delta_k^* = \delta_k + (\sum_{|j| \leq m} W_T^2(j))^{-1}$. If we use the Daniell smoother outlined above the updates become $W_k^* = W_k + \sum_{|j| \leq m} P_{k+j}$ and $\delta^* = \delta_k + 2m + 1$. In practice we may set $m = \lfloor \frac{\sqrt{T}}{2} \rfloor$ to ensure that the conditions for consistency of $\hat{S}_k$ are met. We note that the Whittle approximation does not necessarily hold anymore when we use this type of smoothing plug in estimator; however, our simulations demonstrate that this method still performs well when only a single series is observed.

4.5 Inference

Bayesian structural learning algorithms for decomposable graphs come in two flavors: MCMC samplers and stochastic search procedures [171, 6]. By placing decomposable graphical inference

for time series in the same framework as for the i.i.d. case via our analytic $p(\mathbf{X}_{1:N} \mid G)$, we can easily modify both types of existing methods for the time series case.

Classical MCMC samplers for decomposable graphs sample from the posterior over graphs via Metropolis-Hastings (MH) by proposing single edge addition and deletion moves that keep the graph decomposable [67, 5]. While it is possible to obtain any decomposable graph from any other decomposable graph via a sequence of edge additions and deletions, the path may be hard to reach leading to prohibitive converge times for even a moderate number of vertices p. More recent graph samplers add more global moves by either randomly generating new decomposable graphs [216] or by generating from a Markov chain over a junction tree representation of the graph [73]. To compute the MH acceptance ratio, these samplers rely on computing ratios of present and proposed marginal likelihoods. For simple edge additions and deletions, this ratio simplifies into a function of only the cliques and separators that change between moves. For our case, the ratio expands into a product over frequencies of the same affected cliques and separators, allowing simple modifications to the existing implementations of these samplers to handle TGMs.

All current MCMC samplers struggle to scale to even moderate numbers of nodes. In contexts where point estimates suffice, we can instead consider stochastic search procedures. We utilize a modification of the efficient feature-inclusion stochastic search (FINCS) [171] for inference in our TGMs. FINCS interleaves three moves: 1) single edge addition and deletion moves for local changes to the graph, 2) global sampling moves where edges are added independently to an empty graph and the final graph is triangulated to maintain decomposability, and 3) resampling at step t a full graph from a list of past visited models, $\{G_1, G_2, \ldots, G_{t-1}\}$, in proportion to their posterior probabilities. In steps 1) and 2), to enforce exploration of high probability regions, edge additions that tend to continually improve the model probability are preferentially selected in proportion to a current heuristic estimate of the posterior edge probability

$$\hat{q}_{ij}(t) = \frac{\sum_{k=1}^{t} 1_{\{i,j\} \in E_t} p(X_{1:N}|G_t)p(G_t)}{\sum_{k=1}^{t} p(X_{1:N}|G_t)}, \tag{4.34}$$

where E_t is the current edge set. Edge deletions are performed proportional to $\hat{q}_{ij}^{-1}(t)$. As in

MCMC samplers [67, 5], the junction tree representation of the graph can be efficiently updated after each local move since the two graphs only differ by a single clique and its corresponding separators, allowing a quick computation of the marginal likelihood of a proposed graph in Eq. (4.27). Importantly, the FINCS algorithm depends on the data only through the marginal likelihoods of the cliques C—used to compute the full graph marginal likelihood—which in our TGM case is a product over T frequencies:

$$\prod_{k=0}^{T-1} \frac{B(W_{k,C}, \delta)}{B(W^*_{k,C}, \delta^*)}. \tag{4.35}$$

That is, our implementation simply modifies the original FINCS definition of the clique marginal likelihood.

4.6 Simulations

To test our TGM methods, we consider simulated setups for both $N > 1$ and $N = 1$ time series each generated from an order-1 vector autoregressive process, denoted VAR(1), for $p = 20$ dimensions. Specifically, we simulated data from the model

$$x_t = Ax_{t-1} + \epsilon(t), \tag{4.36}$$

where $x(t) \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$, and $\epsilon(t) \sim N(0, I_{p \times p})$. x_0 is initialized randomly from $N(0, I_{p \times p})$. The inverse spectral density of a VAR(1) process is given by [181]

$$S(\lambda)^{-1} = I + A^T A + e^{-i\lambda} A + e^{i\lambda} A^T. \tag{4.37}$$

Random sparse TGMs were generated by first restricting A to be upper triangular. Following the simulated setup in [181], we set the diagonal elements to a constant $A_{ii} = .5$ and sample the upper diagonal elements as $a_{ij} \sim .5\delta_{ij}$, where $\delta_{ij} \sim$ Binomial(ρ) with $\rho = .2$ for all simulations. The graph G was then determined by identifying the zeros in $S(\lambda)^{-1}$ using Eq. (4.37). Proposed A matrices were accepted when both the absolute value of all eigenvalues of A were less than one, making the series stationary, and the graph G determined by A was decomposable.

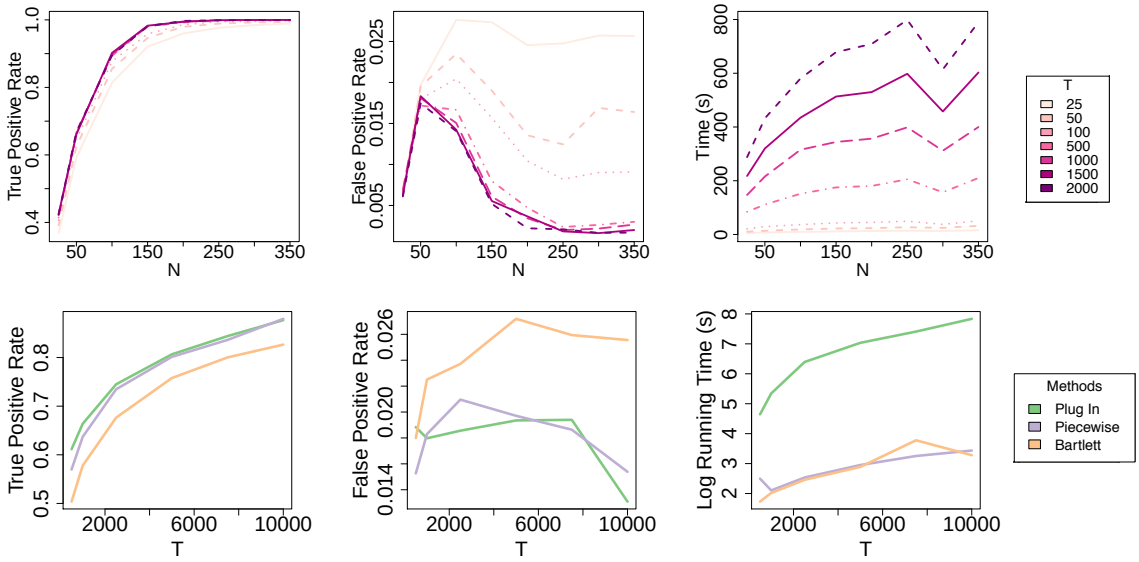


Figure 4.1: **Top:** As a function of the number of time series N, and plotted for various values of their length T, (*left*) mean true positive rate, (*middle*) median false positive rate, and (*right*) mean running time computed across the 200 replicates. Standard error bars are small relative to the scale of the plots and are omitted for clarity. **Bottom:** Same plots as a function of T for a single time series ($N = 1$), and plotted for various periodogram smoothing techniques.

We note that since our formulation reduces to a standard structure learning problem, our emphasis is less on assessing performance with respect to p, which should follow from whichever structure learning algorithm is selected; instead, our focus is on N and T, which are specific to the time series and spectral analysis. For example, in the FINCS algorithm [171], it is quoted that the method can handle graphs with up to roughly $p = 100$ nodes.

4.6.1 Multiple time series

To analyze how our TGM structure learning performance varies with the number of time series replicates, N, we simulated data for $N \in \{20, 50, 100, 150, 200, 250, 300, 350\}$ and $T \in \{25, 50, 100, 500, 1000, 1500, 2000\}$. This process was repeated 200 times for each combination of N and T. Each time series is first decomposed into its discrete Fourier components. We then ran 10,000 iterations of the FINCS algorithm using the fractional marginal likelihood in Eq. (4.28)

with $g = \frac{4}{N}$, a default setting [142, 171]. Our graph prior followed the multiplicity correcting form in Eq. (4.30) with $a = b = 1$. The graph visited with highest posterior probability was then selected and true and false positive rates were computed. Results are displayed in Fig. 4.1. Across T, the true positive rate increases quickly with the number of series, N, achieving an almost perfect true positive rate by about $N = 150$. We also see that the rate of increase in the true positive rate increases with the length of the series T, which relates to the number of considered Fourier frequencies. It is interesting to note that for all T under consideration, the false positive rate tends to start very low ($\approx .005$) for $N = 20$ replicates then spike at $N \in \{50, 100\}$ before declining again. This occurs due to the fact that at low N, very few edges are introduced at all, perhaps due to an Occam's razor type effect of marginal likelihoods penalizing model complexity. As N starts to increase, more edges are introduced, both correct and incorrect, and as N further increases, the false edges are pruned and true edges are retained, leading to a decline in the false positive rate. Note that the false positive spike tends to be more pronounced for time series of smaller length, $T \in \{25, 50\}$. One would expect to see significant improvements, especially for small N, by leveraging the piecewise constant prior of Sec. 4.4 and explored in Sec. 4.6.2 where we show that we are able to learn graphs from just $N = 1$ time series. However, we chose not to include this prior in this analysis so as not to confound its effect with our performance. Here, the noisy periodogram is smoothed implicitly by averaging over N.

Finally, in Fig. 4.1 we see that runtime increases as a function of T due to the dependence on T in the marginal likelihood computation of Eq. (4.27), though significant cost reductions can be achieved through parallelizations leveraging the product form.

4.6.2 Single time series: comparison of methods

To assess the performance of our single-time-series methods outlined in Sec. 4.4, we simulated a time series with $T \in \{500, 1000, 2500, 5000, 7500, 10000\}$. For the piecewise constant prior method, we use $M = \lfloor \sqrt{T} \rfloor$ pieces. We compare against the Bartlett time-series-splitting approach with the number of splits set to $\lfloor \sqrt{T} \rfloor$. We also examine a smoothed plug-in estimator of the spectral density using a Daniell smoother with $m = \lfloor \frac{\sqrt{T}}{2} \rfloor$ for a total window size of $2\lfloor \frac{\sqrt{T}}{2} \rfloor +$

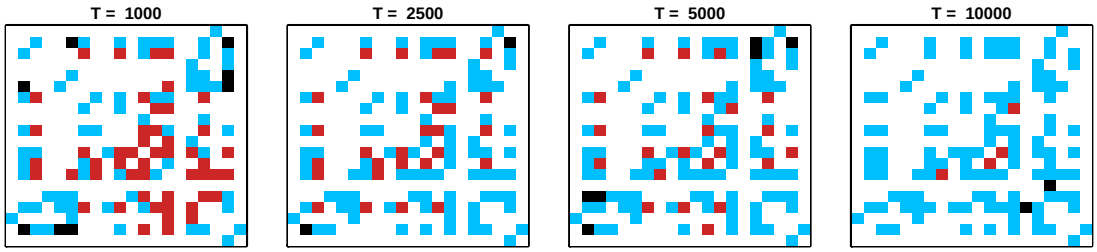


Figure 4.2: Example evolution of error types for the piecewise prior method as a function of series length, $T \in \{1000, 2500, 5000, 10000\}$ and $N = 1$, for a selected graph. Blue, red, **black**, and white entries indicate true positives, false negatives, false positives, and true negatives, respectively. The graph was selected by choosing the graph out of 200 replications with median true positive rate at $T = 2500$. The plots display the connectivity graphs where pixel (i, j) denotes the existence of absence of an edge between series i and j.

$1 \approx \lfloor \sqrt{T} \rfloor$. For each method, the FINCS algorithm was run for 10,000 iterations and the highest scoring graph was selected and used to compute true and false positive rates. This process was repeated 200 times with results displayed in Fig. 4.1 with a replicate representative of our median performance shown in Fig. 4.2. The true positive rate increases for all three methods as a function of T, achieving a final value of about .9 for both the plug-in and piecewise constant prior methods and .79 for the Bartlett method at $T = 10000$. All methods maintain a low false positive rate around .02. Overall, the Bartlett method performs uniformly worse in terms of both true and false positive performance, while the piecewise prior method performs on par with the plug-in method, but at a fraction of the computational cost.

4.7 Global Stock Indices

We explore the utility of our method in discovering conditional independencies between countries inherent in the global financial system. A similar experiment was conducted in [181] using a penalized-likelihood approach to learn TGMs, but restricted to finite-order VAR models with pre-specified order. (Recall that our method only assumes Gaussian stationarity, which includes the class of possibly infinite order VAR processes.) Using `www.globalfinancialdata.com`, we acquired the daily closing prices of 17 stock indices in US dollars for various countries around the world from June 3, 1997 to June 30, 1999. Missing prices were backfilled and only days

where all exchanges traded were considered which resulted in time series of length 542. Following standard practice when analyzing stock prices, we converted the closing prices, p_t, on day t to log-returns according to

$$r_t = 100 \log(p_t/p_{t-1}).$$

We compare the graphical models inferred under two settings: (i) treating the log-returns as independent (as in [171]) and (ii) using our methods to learn a TGM treating the log-returns as a time series. The best graphical models learned in each scenario are depicted in Fig. 4.3.

For our TGM algorithm, we computed the periodogram for the 17-dimensional time series, resulting in 542 complex-valued matrices of dimension 17×17. Since we only have one realization of the time series, we smoothed the periodogram using the techniques and settings discussed in Sec. 4.6.2. We then ran the FINCS algorithm for 100,000 iterations. We compare the resulting highest-probability graph (see Fig. 4.3) to that learned treating the time series as independent based on the model in [171], again using 100,000 iterations of the FINCS algorithm, but in its originally proposed form for non-temporal data.

In Figure 4.3, we see that in both cases we recover some geographical relationships between countries. However, the independent model returns a significantly denser graph than that learned by our TGM approach. Since the independent model is not taking the temporal nature of the data into account, some edges are likely spurious due to random correlations. The TGM, on the other hand, provides an interpretable and intuitive structure with strong geographic connections. For example, there is a distinct United Kingdom / eurozone cluster of Germany 'DE', Finland 'FI', Netherlands 'NL', Belgium 'BE', Switzerland 'CH', Austria 'AT', Spain 'ES', Italy 'IT', Portugal 'PT', and the United Kingdom 'UK'. Another distinct cluster includes the United States 'US', Canada 'CA', Hong Kong 'HK' (whose currency is linked to the USD), and Australia 'AU' (whose currency is correlated with the US S&P), with Japan 'JP' hanging off this cluster. One perhaps strange missing link is between Ireland 'IE' and the UK, though the US and Ireland have a long history of economic connections possibly explaining why Ireland is included in the separator between these two distinct clusters.

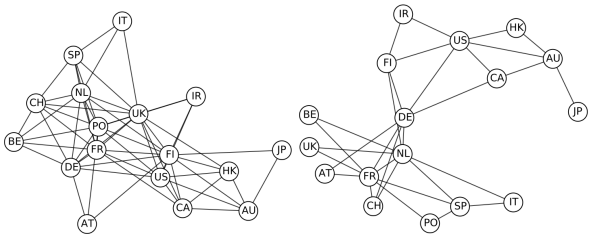


Figure 4.3: Graphical models with the highest posterior probability for the stock index data. **Left:** Treating the log-returns as independent. **Right:** Using our TGM algorithm. In both cases, we see regional connections, but our TGM algorithm results in a sparser and more interpretable graph.

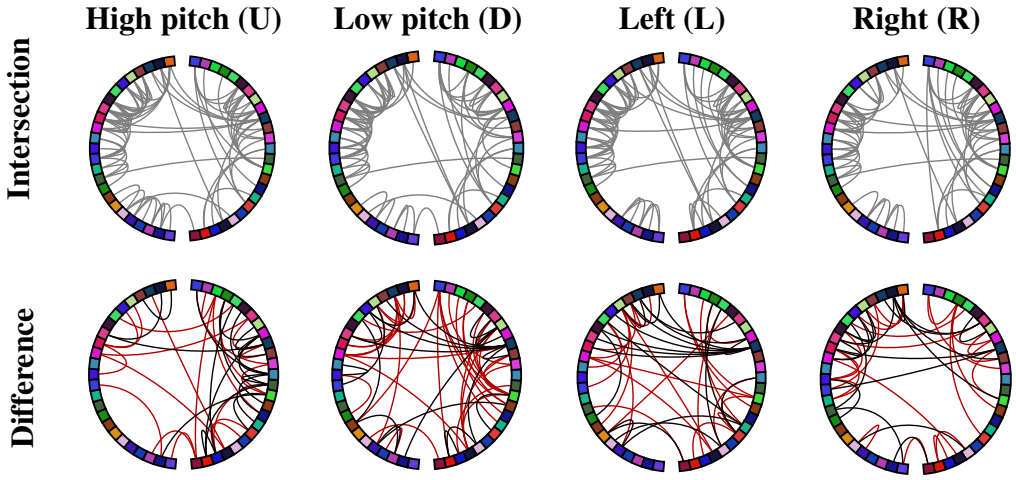


Figure 4.4: Learned TGMs for different MEG conditions. Each node on the periphery represents a brain region with location indicating anatomical location. **Top:** Intersection of learned edges between switching and non-switching conditions. **Bottom:** *Black* edges indicating those in the non-switching condition but not in the switching and *red* vice versa.

4.8 Magnetoencephalography Data

Next we learn TGMs to capture the structure of underlying cortical dynamics from magnetoencephalography (MEG) data collected from ten subjects who were asked to perform a task while maintaining focus on an audio stream and then again while switching focus [114]. Our goal is to discover differences in the underlying TGMs between the non-switching and switching attention conditions. Such differences provide further understanding into the neural underpinnings of auditory selective selection, an important constituent to communication.

The data were collected for each subject performing the experiment in the *switching* (S) and *non-switching* (N) attention conditions. For both S and N conditions, each subject performed the task under an auditory condition of *high* (U) and *low* (D) pitch, and spatial conditions of *left* ear (L) and *right* ear (R) attending. For each of the eight possible conditions, MEG recordings were collected resulting in a 150-dimensional time series of length 992 where each dimension corresponds to a localized region of the brain. We have between 17 and 30 trials for each subject, resulting in about 200 replicate time series per condition.

Often with MEG data, many of the dimensions are dominated by noise due to limited brain activity in that region. We reduced the number of brain regions we studied from 150 to 50 by only considering those with largest variance. In particular, for each trial we mean-centered all of the time-series and computed the variance and retained the top 50 most volatile regions.

We computed the periodogram for each trial and averaged across trials within each condition, resulting in eight periodograms. We ran our spectral TGS version of the FINCS algorithm on these periodograms for 100,000 iterations with fractional prior parameter $4/N_c$, where N_c is the number of trials for condition $c \in \{\text{S, N}\} \times \{\text{U, D, L, R}\}$. We also ran the algorithm for 1.7 million iterations and saw no difference in the resulting graphs.

In Figure 4.4, we depict the intersections and differences between the learned graphs for each experimental condition. We see in the top row that there are a lot of shared connections between the switching and non-switching conditions for each auditory condition. In the bottom row, the differences between the switching and non-switching conditions are depicted where red edges are those in the switching condition but not the non-switching, and black edges are the reverse. The difference plots show that there seems to be substantial "rewiring" for many of the conditions with many edges connecting frontal to back regions. Interestingly, we again see consistencies in these rewirings across conditions. Additionally, we reliably uncover local connections between adjacent brain regions across experimental conditions. Such observations provide guidance for developing experiments and methods to discern the underlying mechanisms that give rise to these different structures.

4.9 Discussion

We introduced a Bayesian approach to graphical model structure learning for time series. In particular, we propose a prior—the *hyper complex inverse Wishart* distribution—for the spectral density matrices in a Whittle likelihood approximation. For decomposable graphs, this prior is conjugate and leads to a closed-form expression of the marginal likelihood of the time series given the graph, marginalizing the spectral density matrices across frequencies. Being able to integrate out this large collection of complex matrices—one for each time point—is critical to developing a practical and scalable inference algorithm. For this, exploiting the fact that our marginal likelihood is analogous to that for i.i.d. Gaussian graphical models [93] but with a product over the number of Fourier frequencies, allows us to deploy straightforward modifications to existing MCMC and stochastic search algorithms. Our simulations show that when many time series are observed, our method recovers the correct graph. When a single time series is observed, we proposed a method to increase robustness of our graph estimation using a piecewise constant prior. Our results on the stock and MEG datasets demonstrated our ability to discover intuitive and interpretable structure in these datasets, importantly leveraging the temporal dependencies.

Extensions to non-decomposable graphs are possible using the i.i.d. graph approaches in both [162] and [6]. A Laplace approximation to the marginal likelihood for non-decomposable graphs is proposed in [6], which we could similarly utilize to approximate the frequency-specific marginal at each term in Equation (4.27). Parallelizing the Laplace approximation computation across frequencies would lead to a scalable method for inference in non-decomposable time series graphs.

Chapter 5

IDENTIFIABILITY AND ESTIMATION IN SUBSAMPLED AND MIXED FREQUENCY STRUCTURAL VECTOR AUTOREGRESSIVE MODELS

5.1 Introduction

Classical approaches to multivariate time series and Granger causality assume that all time series are sampled at the same rate. However, due to data integration across heterogeneous sources, many data sets in econometrics, health care, environment monitoring, and neuroscience comprise multiple series sampled at different rates, referred to as mixed−frequency time series. Furthermore, due to the cost or technological challenge of data collection, many series may be sampled at a rate lower than the true causal scale of the underlying physical process. For example, many econometric indicators, such as GDP and housing price data, are recorded at quarterly and monthly scales [137]. However, there may be important interactions between these indicators at the weekly or bi−weekly scales [137, 183, 18]. In neuroscience, imaging technologies with high spatial resolution, like functional magnetic resonance imaging or fluorescent calcium imaging, have relatively low temporal resolutions. However, it is well−known that many important neuronal processes and interactions happen at finer time scales [213]. A causal analysis rooted at a slower time scale than the true causal time scale may miss true interactions and add spurious ones [213, 177, 18, 22]. A comprehensive approach to Granger causality in multivariate time series should be able to simultaneously accommodate both mixed−frequency and subsampled data.

Recently, causal discovery in subsampled time series has been studied with methods in causal structure learning using graphical models [42, 150, 87]. These methods are model−free, and automatically infer a sampling rate for causal relations most consistent with the data. We maintain a similar goal, but take a model−based approach and examine the identifiability of structural vector autoregressive models under both subsampling and mixed−frequency settings. Structural models

are an important tool in time series analysis [130, 79] and are a mainstay in econometrics and macro−economic policy analysis. These models combine classical linear autoregressive models with structural equation modeling [194] to allow analysis of both instantaneous and lagged causal effects between time series. However, structural models are commonly applied to regularly sampled data, where each series is observed at the same regular intervals. Moreover, the time scale of such an analysis is typically restricted to this shared sampling scale.

[69] recently explored identifiability and estimation of vector autoregressive models under subsampling with independent innovations, i.e., no instantaneous causal effects or error correlations. They showed that with non−Gaussian errors, the transition matrix is identifiable under subsampling, implying that Granger causality estimation is possible. Unfortunately, their results do not cover the case of correlated errors, a common and important aspect of many real−world time series [130]. Interestingly, non−Gaussian errors have also been shown to aide model identifiability in structural autoregressive models with standard sampling assumptions [212, 112, 90, 89, 148]. This line of work applies techniques developed for structural equation modeling with non−Gaussian errors and independent component analysis [88] to the structural time series context. Non−Gaussian errors allow identification of the structural model without other identifying restrictions [112], and also allow identification of the causal ordering of the instantaneous effects if these are known to follow a directed acyclic graph [90]. These models have been successfully applied to many non−Gaussian time series in econometrics [112, 110, 111, 83].

Our approach to subsampling unifies existing approaches to identifiability along two complementary directions. First, our work connects the non−Gaussian subsampled autoregressive model with the independent innovations method [69] with the non−Gaussian structural autoregressive framework [212, 112, 90, 89, 148] by proving identifiability of a structural vector autoregressive model of order one under arbitrary subsampling. As a result, we find that not only can one identify the causal structure of lagged effects from subsampled data with correlated errors, but also the directed acyclic graph of the instantaneous effects without prior knowledge of the causal ordering.

Second, we generalize our results to the mixed−frequency setting with arbitrary subsampling, where the subsampling level may be different for each time series. In doing so, we provide a

unified theoretical approach and estimation methodology for subsampled and mixed−frequency cases. Deriving identifiability conditions on the model parameters in the mixed−frequency case is difficult [4] and has only been studied based on the first two moments of the mixed−frequency process. Our work takes a complementary direction by leveraging higher order moments and provides the first set of specific model conditions for mixed−frequency structural models needed for identifiability. Furthermore, previous mixed−frequency approaches have assumed a causal ordering, while our results indicate this may be estimated by leveraging non−Gaussianity. Finally, our approach to identifiability allows us to move beyond the classical mixed−frequency setting, where the time scale is fixed at the most finely sampled series [4]; we instead consider identifiability and estimation in more general mixed−frequency cases. The four sampling types covered by our approach are shown Figure 5.1. To simplify the presentation, we first introduce our theoretical results for subsampled series of case (a) in Section 5.3. We then generalize the results to the mixed−frequency cases (b), (c), and (d) in Section 5.4.

We introduce an exact expectation−maximization algorithm for inference for both subsampled and mixed−frequency cases. [69] also utilize such an algorithm, but because they formulate inference directly on the subsampled process by marginalizing the missing data, their approach requires an extra approximation. Our approach instead casts inference as a missing data problem and uses a Kalman filter to exactly compute the E−step for both subsampled and mixed−frequency cases. We validate our estimation and identifiability results via extensive simulations and apply our method to evaluate causal relations in a subsampled climate data set and a mixed−frequency econometric data set. Taken together, we present a unified theoretical analysis and estimation methodology for subsampled and mixed−frequency cases, which have previously been studied separately. A summary of our contributions is presented in Table 5.1.

5.2 Background

Let $x_t \in \mathbb{R}^p$ $(t \in \{1, \ldots, T\})$ be a p−dimensional multivariate time series generated at a fixed sampling rate. We collect all x_ts into the matrix $X = (x_1, \ldots, x_T)$. We assume that the dynamics

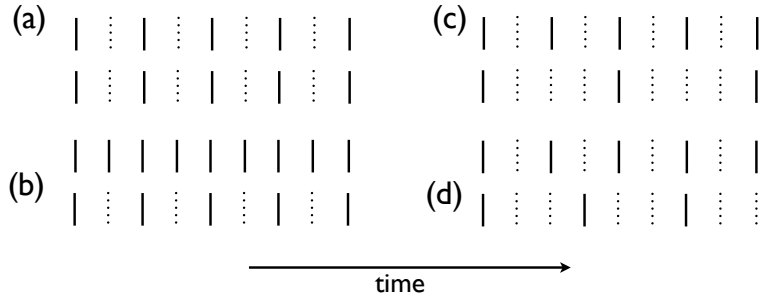


Figure 5.1: Four types of structured sampling. Black lines indicate observed data and dotted lines indicate missing data. (a) Both series are subsampled. (b) The standard mixed–frequency case, where only the second series is subsampled. (c) A subsampled version of (b) where each series is subsampled at different rates. (d) A subsampled mixed–frequency series with no common factor across sampling rates and is thus not a subsampled version of (b).

of x_t follow a combination of instantaneous effects, autoregressive effects, and independent noise,

$$x_t = Bx_t + Dx_{t-1} + e_t, \tag{5.1}$$

where $B \in \mathbb{R}^{p\times p}$ is the structural matrix that determines the instantaneous time linear effects, $D \in \mathbb{R}^{p\times p}$ is an autoregressive matrix that specifies the lag one effects conditional on the instantaneous effects, and $e_t \in \mathbb{R}^p$ is a white noise process such that $E(e_t) = 0$ for all t, and e_{ti} is independent of $e_{t'j}$ for all i, j, t, t' such that $(i, t) \neq (j, t)$. We assume e_{tj} is distributed as $e_{tj} \sim p_{e_j}$. Solving (5.1) in terms of x_t gives the following lag one structural vector autoregressive process for the evolution of x_t,

$$x_t = (I - B)^{-1}Dx_{t-1} + (I - B)^{-1}e_t = Ax_{t-1} + Ce_t. \tag{5.2}$$

Under the representation in (5.2), each element A_{ji} denotes the lag one linear effect of series i on series j and $C \in \mathbb{R}^{p\times p}$ is the structural matrix. The error e_{tj} is referred to as the shock to series j at time t and the element C_{ji} is the linear instantaneous effect of e_{tj} on x_{ti}.

Conditions on C, or equivalently B, for model identifiability and estimation have been explored [79, 101]. The most typical condition is that C is a lower triangular matrix with ones on the

Table 5.1: Summary of contributions of our work to identifiability and estimation in mixed–frequency sampling for structural autoregressive models. The subsampling types are as in Figure 5.1. Citations indicate previous work and check marks indicate our contributions. The notation (ce) indicates computationally expensive; see the discussion at the end of Section 5.7. Hyv08 represents [89], Gong15 represents [69] and Lut05 represents [130]. $C = I$ indicates no instantaneous correlation in the errors where as C free does; we introduce this notation below.

sampling type		none	A	B	C	D
$C = I$	identifiability	Lut05	Gong15	✓	✓	✓
	estimation	Lut05	Gong15 (approx), ✓	✓	✓(ce)	✓(ce)
C free	identifiability	Hyv08	✓	✓	✓	✓
	estimation	Hyv08	✓	✓	✓(ce)	✓(ce)

diagonal, implying a known causal ordering of the instantaneous effects. In this case, one may interpret the instantaneous effects as a directed acyclic graph [115], i.e., a graph, $G = (V, E)$, with vertices $V = \{1, \ldots, p\}$ and directed edge set E, with no directed cycles. A causal ordering is an ordering of the vertices into a sequence, π, such that if j comes before i in π then E does not contain a path of edges from i to j; see, e.g., [176] for details. In the structural context, for $i \neq j$ there exists a directed edge $i \to j$ from x_i to x_j in E, if and only if C_{ji} is nonzero. Classical estimation for structural models with known causal ordering typically proceeds by simultaneously fitting A and C with the identifiability constraint that C is lower triangular. When there are no unobserved confounders, as we assume throughout this work, we may refer to the entries in C as instantaneous causal effects.

A recent line of work [212, 112, 90] focuses on estimating A and C when π is unknown. When the errors, e_t, are non−Gaussian, both the causal ordering and instantaneous effects C may be inferred directly from the data using techniques from independent component analysis [90]. Alternatively, one may dispense with orderings and lower triangular restrictions and directly estimate C under non−Gaussian errors [112]. Our analysis continues these directions, leveraging non−Gaussianity of the structural model with subsampling and or mixed−frequencies.

5.3 Subsampled Structural Vector Autoregressive Models

5.3.1 The subsampled process

Subsampling occurs when, due to low temporal resolution, we only observe x_t every k time steps, as displayed graphically in case A in Figure 5.1. In this case, we only have access to observations $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{\tilde{T}}) \equiv \left(x_1, x_{1+k}, \ldots, x_{1+(\tilde{T}-1)k}\right)$, where $\tilde{T}$ is the number of subsampled observations. By marginalizing out the unobserved x_t, we obtain the evolution equations for $\tilde{x}_t$,

$$\begin{aligned}\tilde{x}_t = x_{1+tk} &= Ax_{1+tk-1} + Ce_{1+tk} = A\left(Ax_{1+tk-2} + Ce_{1+tk-1}\right) + Ce_{1+tk} \\ &= A^k\tilde{x}_{t-1} + \sum_{l=0}^{k-1} A^l Ce_{1+tk-l} \qquad (5.3) \\ &= A^k\tilde{x}_{t-1} + L\tilde{e}_t, \qquad (5.4)\end{aligned}$$

where $\tilde{e}_t = \left(e^T_{1+tk}, \ldots e^T_{2+(t-1)k}\right)^T$ is the stacked vector of errors for time $1+tk$ and the unobserved time points between times $1+tk$ and $1+(t-1)k$ and $L = \left(C, AC, \ldots, A^{k-1}C\right)$. Equation (5.3) states that the subsampled process is a linear transformation of the past subsampled observations with transition matrix A^{k-1} and a weighted sum of the shocks across all unobserved time points. Each shock is weighted by A raised to the power of the time lag. We provide an example of Equation (5.3) in Section 5.9.

Equation (5.4) appears to take a similar form to the structural process in (5.1); however, now the vector of shocks, $\tilde{e}_t$, is of dimension kp, with special structure on both the structural matrix L and the distributions of the elements in $\tilde{e}_t$. Unfortunately, this representation does not have the interpretation of instantaneous causal effects described in Section 5.2, as there are now multiple shocks per individual time series. We will refer to the full parametrization of the subsampled structural model in (5.4) as $(A, C, p_e; k)$. Identifiability of this model means that there is a unique pair of matrices A and C consistent with the joint distribution of $\tilde{X}$ at subsampling rate k.

5.3.2 Lagged and Instantaneous Causality Confounds of Subsampling

A classical analysis based on $\tilde{x}_t$ that does not account for subsampling would incorrectly estimate lagged Granger causal effects in A^k, because $A_{ij} = 0$ does not imply that $(A^k)_{ij} = 0$, and vice versa [69]. Similarly, estimation of structural interactions may also be biased if subsampling is ignored. Classical structural estimation methods that assume a known causal ordering of the instantaneous shocks simply estimate the covariance of the error process, $\Sigma = E(Ce_t e_t^T C^T) = C\Lambda C^T$, and let the estimated structural matrix be the Cholesky decomposition of Σ. Under subsampling, the covariance of the error process is

$$E(L\tilde{e}_t\tilde{e}_t^T L^T) = L\left(I_k \otimes \Lambda\right) L^T, \tag{5.5}$$

where $\otimes$ is the Kronecker product and I_k is the identity matrix of size k. The causal structure given by zeros in the Cholesky decomposition of (5.5) need not be the same as those implied by C.

Example 1. *Consider the process [69]*

$$A = \begin{pmatrix} 0.8 & 0.5 \\ 0 & -0.8 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

so that $C\Lambda C^T = I_p$. Then, for a subsampling of $k = 2$,

$$A^k = \begin{pmatrix} 0.64 & 0 \\ 0 & 0.64 \end{pmatrix}, \quad L\left(I_k \otimes \Lambda\right) L^T = \begin{pmatrix} 1.89 & -0.4 \\ -0.4 & 1.64 \end{pmatrix};$$

this implies no lagged causal effect between x_1 and x_2, but a relatively large instantaneous interaction, contrary to the true data generating model; see Figure 5.2.

5.3.3 Identifiability of L under subsampling

While both lagged Granger causality and instantaneous structural interactions are confounded by subsampling, we show here that, by accounting for subsampling, we may, under some conditions,

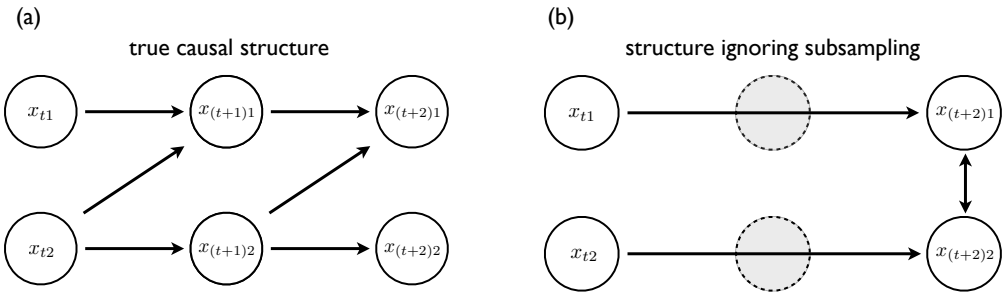


Figure 5.2: Graphical depiction of how subsampling confounds both causal analysis of lagged and instantaneous effects. (a) The true causal diagram for the regularly sampled data. (b) The estimated causal structure of the subsampled process when the effects of subsampling are ignored.

still estimate the A and C matrices of the underlying process directly from the subsampled data. As a first step to proving the identifiability of A and C, we show that the matrix $L = \left(C, \ldots, A^{k-1}C\right)$ in (5.4) is identifiable up to permutation and scaling of columns when the p_{e_j}, the distribution of e_{tj}, are all non−Gaussian.

Proposition 1. *Suppose that all p_{e_j} are non−Gaussian. Given a known subsampling factor k and subsampled data $\tilde{X}$ generated according to equation (*5.4*), L may be determined up to permutation and scaling of columns.*

The proof closely follows that of Proposition 1 in [69] and depends on the following fundamental result in independent component analysis [55].

Lemma 2. *Let $\hat{e} = Jr$ and $\hat{e} = Ms$ be two representations of the n−dimensional random vector $\hat{e}$, where J and M are constant matrices of orders $n \times l$ and $n \times m$, respectively, and $r = (r_1, \ldots, r_l)^T$ and $s = (s_1, \ldots, s_m)^T$ are random vectors with independent components.*

If the ith column of J is not proportional to any column of M, then r_i is Gaussian. Moreover, if the ith column of J is proportional to the jth column of M, then the logarithms of the characteristic functions of r_i and s_j differ by a polynomial in a neighborhood of the origin.

This result states that if r is non−Gaussian with independent elements, and if $Jr = Ms$, then M and J must be equal up to permutation and scaling of columns. This implies that one may estimate

M from only observations of $\hat{e}$ and that the estimate of M should be equal up to permutations and scalings of the true M.

To prove Proposition 1 using Lemma 2, note that A^k is identifiable by linear regression. Thus, the error component $\hat{e} = \tilde{x}_t - A^k \tilde{x}_{t-1} = L\tilde{e}_t$ satisfies the conditions of Lemma 2 and L is identifiable up to permutations and scalings since $\tilde{e}_t$ are non−Gaussian.

5.3.4 Complete identifiability of the structural autoregressive model when $C = I$

Using the identifiability result for L in Proposition 1 we can derive identifiability statements and conditions for C and A in the subsampled case. We require a few mild assumptions:

Assumption 1. *x_t is stationary so that all singular values of A have modulus less than one;*

Assumption 2. *The distributions p_{e_j} are distinct for each j after rescaling e_j by any non−zero scale factor, their characteristic functions are all analytic, or they are all non−vanishing, and none of them has an exponent factor with polynomial of degree at least two,*

Assumption 3. *All p_{e_j} are asymmetric.*

Assumption 1 is standard in time series modeling [130] and Assumption 2 is common in independent component analysis. While many of our identifiability results for C only require that the p_{e_j} distributions are non-Gaussian, our identifiability results for A in part b) of Theorems 1 and 2 and part c) of Theorems 3 and 4, further require Assumption 3, that p_{e_j} is asymmetric. In practice, assuming fully Gaussian errors may be unrealistic, where aspects of non-Gaussianity, like asymmetry [197, 113, 80], heavy tails [39, 155], or stochastic volatility [96] are often observed. Not only are non-Gaussian errors empirically appealing, but furthermore, theoretical and modeling approaches that harness the higher order moments of non-Gaussian distributions aid in identifying model parameters that are unidentifiable from the first two moments alone.

[69] give identifiability results under Assumptions 1 and 2 for the subsampled autoregression with no error correlations, $C = I$. We restate their result in our framework.

Theorem 1. *[69] Suppose that e_{tj} is non−Gaussian for all t, j, and that the data $\tilde{x}_t$ are generated by equation (5.2) with $C = I_p$. Assume that the process admits another representation $(A', I_p, p'_e; k)$. If Assumptions 1 and 2 hold, then*

(a) A' can be represented as $A = AD_1$, where D_1 is a diagonal matrix with 1 or -1 on the diagonal. If we constrain the self influences to be positive, represented by the diagonal entries, then $A' = A$.

(b) if Assumption 3 also holds, then $A' = A$.

5.3.5 Complete identifiability of general structural autoregressive model

For identifiability of the full model under subsampling, we require two more assumptions:

Assumption 4. *The variance of each p_{e_j} is equal to one, i.e., $\Lambda = I_p$;*

Assumption 5. *The matrix C is full rank.*

Assumption 4 is common in structural models and removes the non-identifiability between scaling the e_{tj}s and scaling the columns of C. Assumption 5 is mild, and covers the more restrictive assumption in non−Gaussian structural models that C may be row, and column, permuted to a lower triangular matrix [174]. Under these assumptions, we have the following identifiability result for general subsampled structural models.

Theorem 2. *Suppose that e_{tj} are all non−Gaussian and independent, and that the data $\tilde{x}_t$ are generated by Equation (5.2) with representation $(A, C, p_e; k)$. Assume that the process also admits another subsampling representation $(A', C', p'_e; k)$. If Assumptions 1, 2 and 4 hold, then*

a) C is equal to C' up to permutation of columns and scaling of columns by 1 or -1, that is $C' = CP$ where P is a scaled permutation matrix with 1 or -1 elements. This implies $\Sigma = CC^T = C'C'^T = \Sigma^T$;

b) if Assumptions 3 and 5 also hold, then $A = A'$.

The requirement that C is full rank is due to the structure of L. Since one may identify C as the first p columns of L, to obtain A we must premultiply the second set of p columns of L by C^{-1}.

The asymmetry assumption is needed since the scaling of the columns of C and AC by 1s or -1s is ambiguous if the distributions are symmetric; the asymmetry assumption ensures that the unit scalings are identifiable. See Section 5.9 for a full proof.

If the instantaneous causal effects follow a directed acyclic graph, we may identify the structure without any prior information about causal ordering of the variables.

Corollary 3. *If Assumptions 1, 2, and 4 hold and the true structural process corresponds to a directed acyclic graph G, that is, it has a lower triangular structural matrix C with positive diagonals, and it admits another representation with structural matrix C', then $C = C'$. Hence the structure of G is identifiable without prior specification of the causal ordering of G.*

This result follows because C may be identified up to a column permutation. Based on the identifiability results of [174], if C corresponds to an acyclic graph, it may be row and column permuted to a unique lower triangular matrix. The row permutations identify the causal ordering, and the nonzero elements below the diagonal identify the edges in G. See [174] for more details on identifiability and estimation of the graph from C.

Taken together, the results of Theorem 2 and Corollary 3 imply that when the shocks, e_t, are independent and asymmetric, a complete causal diagram of the lagged effects and the instantaneous effects are fully identifiable from the subsampled time series, $\tilde{X}$.

5.4 Mixed–Frequency Structural Autoregressive Models

5.4.1 Background and Motivation

Estimation and forecasting of mixed−frequency time series are commonly approached using autoregressive models [59, 169]. Typically, the model is fit at the same scale as the fastest sampled time series, which is depicted in Figure 5.1 (c). Due to costly data collection, especially for macroeconomic indicators like GDP, this scale is generally arbitrary and may not reflect the true causal dynamics, leading to confounded Granger and instantaneous causality judgments [213, 22]. If the true causal scale, or one of interest to an analyst, is at a lower rate, as in case (d) in Figure 5.1, then

an analysis at the observed rate will run into the same problems as those for the single frequency subsampling case discussed in Section 5.3.2. We provide an example at the end of Section 5.4.2.

Identifiability conditions for mixed−frequency autoregressive models with no subsampling at the fastest scale (Figure 5.1b) was an open problem for many years [30]. [4] recently showed that the mixed−frequency non−structural autoregressive model of Figure 5.1 (b) is generically identifiable from the first two observed moments, so unidentifiable models make up a set of measure zero of the parameter space. Explicit identifiability conditions for the lag 1, bivariate case from the first two moments have also been established [4]. However, no explicit identifiability conditions for structural models or models in higher dimensions have been explored.

In this section, we generalize our identifiability results from Section 5.3 to the mixed−frequency case with arbitrary levels of subsampling for each time series. Our analysis indicates that Granger and instantaneous causal effects can be accurately estimated from mixed−frequency time series. Specifically, we use the results from Section 5.3 to provide explicit identifiability conditions for mixed−frequency structural models under arbitrary subsampling, cases (b), (c), and (d) in Figure 5.1, with non−Gaussian error assumptions. Together, our framework provides a unified way of deriving explicit identifiability conditions for both subsampling and mixed−frequency cases. While case (c) in Figure 5.1 is a subsampled version of the standard mixed−frequency case, our results also cover mixed−frequency subsampling like case (d). To our knowledge, these results are the first identifiability results for subsampled mixed−frequency cases like (c) and (d).

5.4.2 Mixed–Frequency Structural Autoregressive Models

We assume each time series in $x_t \in \mathbb{R}^p$ is sampled at one of two sampling rates, a slow subsampling ratesk_s and a fast subsampling rate k_f. We write $x_t = (x_t^s, x_t^f)$, where x_t^s are those series subsampled at k_s and x_t^f are those subsampled at k_f. Let $k \in \{k_s, k_f\}^p$ be the list of subsampling rates for each time series. In Figure 5.1(b), $k_f = 1$ and $k_s = 2$, whereas in Figure 5.1(c), $k_f = 2$ and $k_s = 4$. Analogous to the subsampled case, we refer to a parameterization of a mixed−frequency structural model as $(A, C, p_e; k)$, where k is now a $p-$vector. Let k^* be the smallest multiple of both k_s and k_f; for example, in Figure 5.1(c), $k^* = 4$ and in Figure 5.1(d) $k^* = 6$.

We may derive a similar representation to (5.4) for mixed−frequency series. Fix a time point t such that all series are observed. Let $I^{(q)}$ be a modified $p \times p$ identity matrix where all rows i such that x_{ti} is not observed at time $t-q$ are set to zero. Further, let $I^{(\bar{q})} = I - I^{(q)}$, $A^{(q)} = I^{(q)}A$, and $A^{(\bar{q})} = I^{(\bar{q})}A$. Then

$$\begin{aligned} x_t &= Ax_{t-1} + Ce_t = AI^{(1)}x_{t-1} + AI^{(\bar{1})}x_{t-1} + Ce_t \\ &= AI^{(1)}x_{t-1} + A(A^{(\bar{1})}x_{t-2} + C^{(\bar{1})}e_{t-1}) + Ce_t \\ &= F\tilde{x}_{t-1} + L\tilde{e}_t, \end{aligned} \tag{5.6}$$

where

$F = (A, AA^{(\bar{1})}, \ldots, AA^{(\bar{1})}\ldots A^{(\overline{k^*-1})})$, $L = (C, AC^{(\bar{1})}, AA^{(\bar{1})}C^{(\bar{2})}, \ldots, AA^{(\bar{1})}\ldots A^{(\overline{k^*-2})}C^{(\overline{k^*-1})})$, $\tilde{x}_{t-1} = (I^{(1)}x_{t-1}, \ldots, I^{(k)}x_{t-k^*})$, and $\tilde{e}_t = (e_t, e_{t-1}, \ldots, e_{t-k^*+1})$.

Equation (5.6) has the same form as (5.4), suggesting that similar identifiability results will hold. We provide an explicit example of (5.6) for a mixed−frequency series in Section 5.9.

In a subsampled mixed−frequency setting where the fastest rate is greater than unity, Figure 5.1 (c), not accounting for subsampling not only can lead to the kind of mistaken inferences discussed in Section 5.3.2, but also to further mistakes unique to the mixed−frequency case.

Example 2. *Consider a subsampled mixed−frequency structural process generated by (5.6) with the (A, C) parameters given by Example 1. Suppose subsampling is not taken into account and $\tilde{X}$ is analyzed instead as a classical mixed−frequency series, case (b), based on the first two moments [4]. We consider two cases.*

Case 1: The sampling rate is $k = (2, 4)$. In this case, if $\tilde{X}$ is analyzed at the rate $(1, 2)$ using the first two moments, then A and Σ are not identifiable at this rate since both off diagonal elements of A are zero [4]. Thus, no inference of both the instantaneous correlations and lagged effects are even possible.

Case 2: The sampling rate is $k = (2, 6)$. In this case, if $\tilde{X}$ is analyzed at the rate $(1, 3)$ using

the first two moments, the estimated A and covariance Σ will be the same as that in Example 1 [4], leading to an incorrect inference that there is an instantaneous effect but not any directed lagged effect.

5.4.3 Identifiability of Mixed–Frequency Structural Autoregressive Models

We provide generalizations of Theorems 1 and 2 to the mixed−frequency case.

Theorem 3. *Suppose the e_{tj} are non−Gaussian and independent for all t and j, and that the data $\tilde{x}_t$ are generated by equation (5.2) with $C = I_p$. Assume that the process also admits another mixed−frequency representation $(A', I_p, p'_e; k)$. If Assumptions 1 and 2 hold, then*

(a) A' can be represented as $A' = AD_1$, where D_1 is a diagonal matrix with 1 or -1 on the diagonal.

(b) If any multiple of k_i is 1 smaller than some multiple of k_j, then $A_{ij} = A'_{ij}$. If $A_{ij} \neq 0$ this implies that $(D_1)_{jj} = 1$, i.e. the jth columns of A and A' are equal: $A_{:j} = A'_{:j}$.

(c) If Assumption 3 also holds, then $A' = A$.

Proof. Parts (a) and (c) follow since we may further subsample all series in x_t to a subsampling rate of k^*. This gives a subsampled $\tilde{X}$ with representation $(A, I, p(e); k^*)$. Applying Theorem 1 gives the result. Furthermore, if some multiple of k_i is equal to some multiple of k_j minus one, then there exists a set of ts for (5.6) where series i is observed at time $t-1$ and series j is observed at time t. By identifiability of linear regression, $A'_{ij} = A_{ij}$. This resolves the sign ambiguity of the columns in 3, so that $A_{:j} = A'_{:j}$. □

Theorem 4. *Suppose the e_{tj} are non−Gaussian and independent for all t and j, and that the data $\tilde{x}_t$ are generated by equation (5.2) with representation $(A, C, p_e; k)$. Assume that the process also admits another mixed−frequency subsampling representation $(A', C', p'_e; k)$. If assumptions 1, 2, and 4 hold, then*

(a) C is equal to C' up to permutation of columns and scaling of columns by 1 or -1, i.e. $C' = CP$ where P is a scaled permutation matrix with 1 or -1 elements. This implies that $\Sigma = CC^T = C'C'^T = \Sigma'$.

(b) If C is lower triangular with positive diagonals, i.e. the instantaneous interactions follow a directed acyclic graph, and if for all i there exists a j such that any multiple of k_i is 1 smaller than some multiple of k_j with $A_{j:}C_{:i} \neq 0$, then $A = A'$.

(c) If Assumptions 3 and 5 also hold, then $A = A'$.

The proofs of parts (a) and (c) follow the same subsampling logic as the proof given for Theorem 3. The proof of part (b) is given in Section 5.9.

Theorems 3 and 4 demonstrate that identifiability of structural models still holds for mixed−frequency series with subsampling under non−Gaussian errors. Properties 1 and 3 in both Theorem 3 and Theorem 4 are the same as their subsampled counterparts; Property 2 in both theorems shows how the mixed−frequency setting provides additional information to resolve parameter ambiguities in the non−Gaussian setting. Specifically, when there is one time step difference between when series x_j and x_i is sampled, then A_{ij} is identifiable. We can then use this information to resolve sign ambiguities in columns of A, which leads to Property 2 in both Theorems 3 and 4. This result applies directly to the standard mixed−frequency setting [4, 169] where one series is observed at every time step in Figure 5.1 (b). It also applies to case (d), since there exists certain time steps where one series is observed directly before a latter series.

5.5 Estimation

5.5.1 Modeling non–Gaussian errors

We model the non−Gaussian errors as a mixture of Gaussians with m components. This approach has been used widely in econometrics and other fields as a flexible and tractable way of modeling non−Gaussianity in innovations [112, 69]. Formally, we assume that e_{tj} is drawn from the mixture distribution

$$z_{tj} \sim \text{Categorical}(\pi_j), \quad e_{tj} \sim \mathcal{N}(\mu_{jz_{tj}}, \sigma^2_{jz_{tj}}),$$

where μ_j, σ_j^2 and π_j are length m vectors specifying the mean, variance, and mixing weight of each mixture component. The z_{tj} component indicators are auxiliary variables introduced to facilitate tractable inference. Furthermore, the mixture model for the errors implies that conditional on the assignment indicators, z_{tj}, the mean and variance of the error distribution for each series j are time−dependent. This mixture of Gaussians model is able to capture the types of non-Guassianity required for identifiability and also those observed in real world time series. Asymmetric errors may be formed when the mixture centers are nonzero and the variances or mixture weights are different. A non-Gaussian and heavy tailed, though symmetric, distribution may be formed by setting the mixture centers to zero but allowing the mixture variances to have different values. The full set of parameters for the structural model is $\Theta = (A, C, \mu, \sigma^2, \pi)$ where μ, σ^2, π concatenate the mixture parameters of the errors across series. For example, μ_{ji} is the mean of the ith mixture component for the jth error distribution, and likewise for σ^2 and π.

5.5.2 Expectation–maximization algorithm

We develop an expectation−maximization algorithm for joint maximum likelihood estimation of the full set of parameters Θ based only on the observed subsampled/mixed−frequency data $\tilde{X}$. Our method is the same for both subsampled and mixed−frequency data, unlike that of [69], which is tailored to the subsampled case. Furthermore, the non−structural specific, i.e., $C = I$, algorithm of [69] introduces auxiliary noise terms to facilitate inference, rendering their resulting algorithm non−exact; in contrast, our algorithm introduces no such approximations. Since the log−likelihood is non−convex, we employ multiple random restarts to avoid poor local optima. For the subsampled case, the local optimum problem is particularly severe due to the nonidentifiability under the first two moments, so many (A, C) parameter values can give a good fit to the data. The basic algorithm also suffers from slow convergence due to the large amount of missing data. To speed up the algorithm we deploy the adaptive over−relaxed method [165].

Let $W = C^{-1}$, and let $z_{tji} = 1$ if error e_{tj} was generated by mixture component i and $z_{tji} = 0$

otherwise. The complete log−likelihood, $\log p(X_{1:T}, z_{1:T} \mid \Theta)$, of our structural model is

$$T \log |W| + \sum_{t=1}^{T} \sum_{j=1}^{p} \sum_{i=1}^{m} z_{tji} \left\{ \log \pi_{ji} - \frac{1}{2} \log 2\pi\sigma_{ji}^2 - \frac{(W_j x_t - W_j A x_{t-1} + \mu_{ji})^2}{2\sigma_{ji}^2} \right\}, \quad (5.7)$$

where W_j is the jth row vector of W. The algorithm alternates between the E−step, where we compute the conditional expectation $E\left\{\log p(X_{1:T}, z_{1:T} \mid \Theta) \mid \tilde{X}\right\}$, and the M−step, where that expectation is maximized with respect to the parameters Θ. We first provide the M−step updates, and then explain how the conditional expectations are computed using a Kalman filter.

5.5.3 M–step

In the M−step, we maximize the expected complete log−likelihood conditional on the observed data, $E\left\{\log p(X_{1:T}, z_{1:t} \mid \Theta) \mid \tilde{X}\right\}$, with respect to Θ via coordinate ascent, cycling through A, W, and (μ, σ^2, π) until convergence. The specific updates are given below.

A update: Each row of A, A_j, may be updated independently,

$$\hat{A}_j = \left\{ \sum_{t=1}^{T} \sum_{i=1}^{m} \frac{E(z_{tji} x_{t-1} x_{t-1}^T \mid \tilde{X})}{\sigma_{ji}^2} \right\}^{-1} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{m} \frac{-\mu_{ji} E(z_{tji} x_{t-1} \mid \tilde{X}) + E(z_{tji} x_{t-1} x_t^T \mid \tilde{X}) W_j^T}{\sigma_{ji}^2} \right\}. \quad (5.8)$$

μ, σ^2, and π update: These may be optimized jointly in one step using

$$\hat{\mu}_{ji} = \left(\sum_{t=1}^{T} E(z_{tji} \mid \tilde{X}) \right)^{-1} \left(\sum_{t=1}^{T} E(z_{tji} x_t \mid \tilde{X}) - W_j A E(z_{tji} x_{t-1} \mid \tilde{X}) \right), \quad \hat{\pi}_{ji} = T^{-1} \sum_{t=1}^{T} E(z_{tji} \mid \tilde{X})$$

$$\hat{\sigma}_{ji}^2 = \frac{1}{\sum_{t=1}^{T} E(z_{tji} \mid \tilde{X})} \Bigg\{ \sum_{t=1}^{T} W_j E(z_{tji} x_t x_t^T \mid \tilde{X}) W_j^T + W_j^T A E(z_{tji} x_{t-1} x_{t-1}^T \mid \tilde{X}) A^T W_j^T + \hat{\mu}_{ji}^2 E(z_{tji} \mid \tilde{X}) - 2\mu_{ji} W_j E(z_{tji} x_t \mid \tilde{X}) - 2 W_j E(z_{tji} x_t x_{t-1}^T) A^T W_j^T + 2\mu_{ji} W_j A E(z_{tji} x_{t-1}) \Bigg\}$$

W update: Since the maximization is not given in closed form, we utilize the Newton−Raphson method. Let $w = \text{vec}(W)$ be the W vectorization. At each step, the next w iterate is

$$w^{l+1} = w^l - H(w^l)^{-1}\nabla l(w^l), \tag{5.9}$$

where $l(w) = E\left\{\log p(X_{1:T}, z_{1:t} \mid \Theta) \mid \tilde{X}\right\}$ and $H(w)$ is the Hessian of $l(w)$ with respect to w. Expressions for the gradient and Hessian are in Section 5.9.

5.5.4 E–step

All conditional expectations in the $M-$step above are computed using the Kalman filtering−smoothing algorithm. For simplicity, consider one block of data, so that $X = x_{1:t}$, where x_1 and x_t are fully observed but $x_{t'}$, $1 < t' < t$, have some missing data, and hence are not included in $\tilde{X}$. Any subsampled/mixed−frequency series can be broken into blocks of this type. The conditional expectation $E(z_{tji}x_t x_{t-1}^T \mid \tilde{X})$ can be computed by noticing that

$$E(z_{tji}x_t x_{t-1}^T \mid \tilde{X}) = E_{z_{1:t}}\left\{z_{tji}E_x\left(x_t x_{t-1}^T \mid \tilde{X}, z_{1:t}\right)\right\}. \tag{5.10}$$

For a fixed $z_{1:t}$, $E_x\left(x_t x_{t-1}^T \mid \tilde{X}, z_{1:t}\right)$ is computed using the Kalman filtering−smoothing algorithm, since for fixed $z_{1:t}$, $\tilde{x}_t$ is a linear Gaussian state−space model with latent observations x_t. We compute $E_x\left(x_t x_{t-1}^T \mid \tilde{X}, z_{1:t}\right)$ for each $z_{1:t}$ combination, then sum them together weighted by $p(z_{1:t} \mid \tilde{X})z_{tji}$. The probability $p(z_{1:t} \mid \tilde{X})$ may be computed by

$$p(z_{1:t} \mid \tilde{X}) \propto p(\tilde{X} \mid z_{1:t})p(z_{1:t}), \tag{5.11}$$

where $p(z_{1:t})$ is the prior mixture component weights, π, and $p(\tilde{X} \mid z_{1:t})$ is the likelihood of the observed data, which may also be computed by one Kalman pass. This processes is repeated for all expectations in the $E-$step. The computational complexity of this exact algorithm scales as $2^{(k+1)p}$, since the Kalman filter must be run for all combinations of $z_{1:t}$ for each block. The approximate algorithm of [69] has the same complexity. Similar to [69], we have explored approximate

inference methods based on variational and Markov chain Monte Carlo methods but found their performance to be poor; we discuss this in Section 5.8.

5.6 Simulations

5.6.1 Estimation Dependence on Subsampling Factor and Number of Observations

We first investigate the performance of the expectation-maximization algorithm under subsampling. We simulate data with $p = 2$ time series and $m = 2$ mixture components. The asymmetric error distributions are given by: $\pi_1 = (0{\cdot}7, 0{\cdot}3)$, $\sigma_1 = (0{\cdot}2, 1)$, $\mu_1 = (0{\cdot}36, -0{\cdot}84)$ for e_{t1} and $\pi_2 = (0{\cdot}7, 0{\cdot}3)$, $\sigma_2 = (0{\cdot}2, 1)$, $\mu_2 = (-0{\cdot}36, 0{\cdot}84)$ for e_{t2}. We consider two cases for A and C:

$$A^{(1)} = \begin{pmatrix} 0.98 & 0 \\ 0.2 & 0.98 \end{pmatrix} A^{(2)} = \begin{pmatrix} 0.98 & 0.31 \\ -0.31 & 0.98 \end{pmatrix}, \quad C^{(1)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} C^{(2)} = \begin{pmatrix} 1 & 0 \\ -0.2 & 1 \end{pmatrix}.$$

Simulations are performed for two subsampling factors, $k \in \{2, 3\}$, and three sample sizes, $T \in \{205, 403, 805\}$. Due to subsampling, the actual sample sizes are reduced. Data from each parameter configuration are generated 10 times and the estimation algorithm is run on each realization using 1000 random restarts. Box plots of the error estimates for two of the scenarios are shown in Figures 5.3 and 5.4; see the Section 5.9 for plots of the other two settings.

We also perform a similar experiment for $p = 3$. We simulate data with parameters

$$A = \begin{pmatrix} 0.57 & 0 & -0.2 \\ 0.2 & 0.57 & 0 \\ 0 & 0.25 & 0.57 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0.2 & 1 & 0 \\ 0.25 & -0.2 & 1 \end{pmatrix}. \tag{5.12}$$

The mixture of normal error distributions for e_{t1} and e_{t2} are the same as that for the $p = 2$ case. The parameters for e_3 are $\mu_3 = (-0.625, 1.875)$, $\sigma_3 = c(0.2, 3)$, and $\pi_3 = (0.75, 0.25)$. The average error rates are displayed in Table 5.2, and indicate increasingly accurate estimation in trivariate structural systems as sample size increases.

k	2			3		
T	203	403	805	203	403	805
A	-2.4	-7.0	-7.5	-0.9	-1.6	-6.8
C	-3.6	-4.8	-5.8	-1.8	-1.8	-3.9

Table 5.2: Average $\log$ MSE of A and C in a $p = 3$ structural system over ten random samples for $k \in (2, 3)$ and three sample sizes.

k	2					3				
γ	1.8	1.2	0.8	0.4	0	1.8	1.2	0.8	0.4	0
$A^{(1)}$ $C^{(1)}$	-9.0	-7.7	-7.3	-7.0	-0.018	-8.1	-7.0	-7.1	-7.0	-7.4
$A^{(1)}$ $C^{(2)}$	-9.0	-7.9	-7.7	-7.4	0.16	-7.9	-7.2	-7.4	-7.2	-7.5
$A^{(2)}$ $C^{(1)}$	-9.1	-7.9	-0.94	-0.26	1.2	-8.0	-0.33	0.71	1.6	1.6
$A^{(2)}$ $C^{(2)}$	-9.1	-8.0	-0.94	0.15	1.3	-8.0	-0.32	1.0	1.4	1.2

Table 5.3: Average $\log$ MSE of A over ten random samplings for both A estimates across multiple settings of the parameters, number of observations, and subsampling factors.

5.6.2 Estimation Dependence on the Asymmetry of Errors

We analyze estimation performance as a function of the skewness of the error distribution, γ, a measure of asymmetry. We simulate data from the same (A, C) parameter configurations as in Subsection 5.6.1 for $k \in (2, 3)$ and $T = 403$. While keeping the variance fixed, we vary the error distributions across a range of γ, $\gamma \in (1.8, 1.2, 0.8, 0.4, 0)$, so that e_{t1} and e_{t2} have the same magnitude skewness but opposite sign. The skewness values are obtained by gradually modifying the μ, σ^2, and π values in a bivariate mixture of normals. See Section 5.9 for the exact parameter values and plots of the simulated error distributions.

The results for A estimation are shown in Table 5.3. First, estimation remains accurate across all skewness settings for $A^{(1)}$ for $k = 3$, while for $k = 2$ the error remains low for $\gamma > 0$ but spikes for $\gamma = 0$. For $A^{(2)}$, estimation is stable until $\gamma = 1.2$ for $k = 2$, but for $k = 3$, estimation is only stable at $k = 1.8$. Taken together, these results suggest that the strength of identifiability depends on a combination of factors, both A, C, k, and the level of asymmetry of the error distributions. Similar results for C are given in Section 5.9.

5.6.3 Estimation Dependence on the Signal to Noise Ratio

We next investigate estimation performance in subsampling and mixed−frequency sampling as a function of the signal to noise ratio. In these experiments we use $A^{(1)}$ and $C^{(2)}$. We scale A by a factor to set its maximum eigenvalue to the desired level. We perform these experiments for both full subsampling of $k = 2$ and 3 and mixed−frequency subsampling where one series is observed at every time point and the other is subsampled. Data from each parameter configuration are generated 40 times. In Figure 5.5 we plot the average absolute error for estimating the A and C matrices as a function of the maximum eigenvalue of A. Estimation under subsampling is stable until the maximum eigenvalue falls to about 0·5, and estimation becomes dramatically worse, indicating unstable estimation in this regime. The increasing error in the estimation of A as a function of signal to noise ratio is also observed in the mixed−frequency case. However, the estimation error increases less dramatically than in the subsampled case, partly due to the presence of fewer local optima in the mixed−frequency case. In the mixed−frequency case, the error in C estimation appears to be constant across the maximum eigenvalue range we considered.

Unstable estimation arises from a combination of two factors. First, under subsampling, the transition matrix of the subsampled process is A^k, indicating that the signal strength between observations scales exponentially as a function of k. Furthermore, the likelihood surface is multimodal, where multiple high probability modes all have approximately the same A^k value. As the signal to noise ratio falls, A^k estimation becomes more difficult due to subsampling, and thus the multimodal estimation becomes more severe, and modes far from the true A occasionally have higher likelihood. Overall, these simulations indicate that, in the subsampling case, there appears to be a threshold on the maximum eigenvalue, below which inference becomes unstable.

The simulations cover cases (a) and (b) in Figure 5.1. Unfortunately, the complexity of the E-step forbids performing simulations in a reasonable time on cases (c) and (d). Future work will explore tractable inference in these cases; see the discussion at the end of Section 5.7.

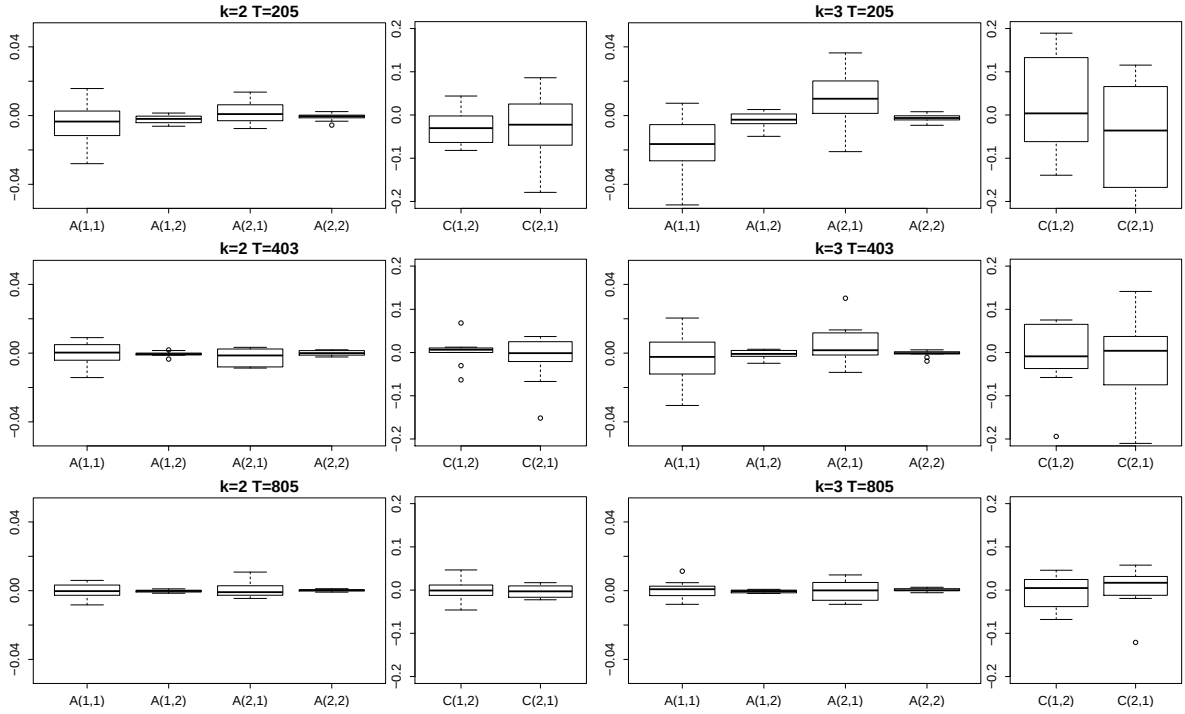


Figure 5.3: Boxplots of errors in $A^{(1)}$ and $C^{(1)}$ parameter estimates over 10 random data samplings. The original series is of length 203 (top), 403 (middle) or 805 (bottom) and then subsampled at $k = 2$ (left) and $k = 3$ (right).

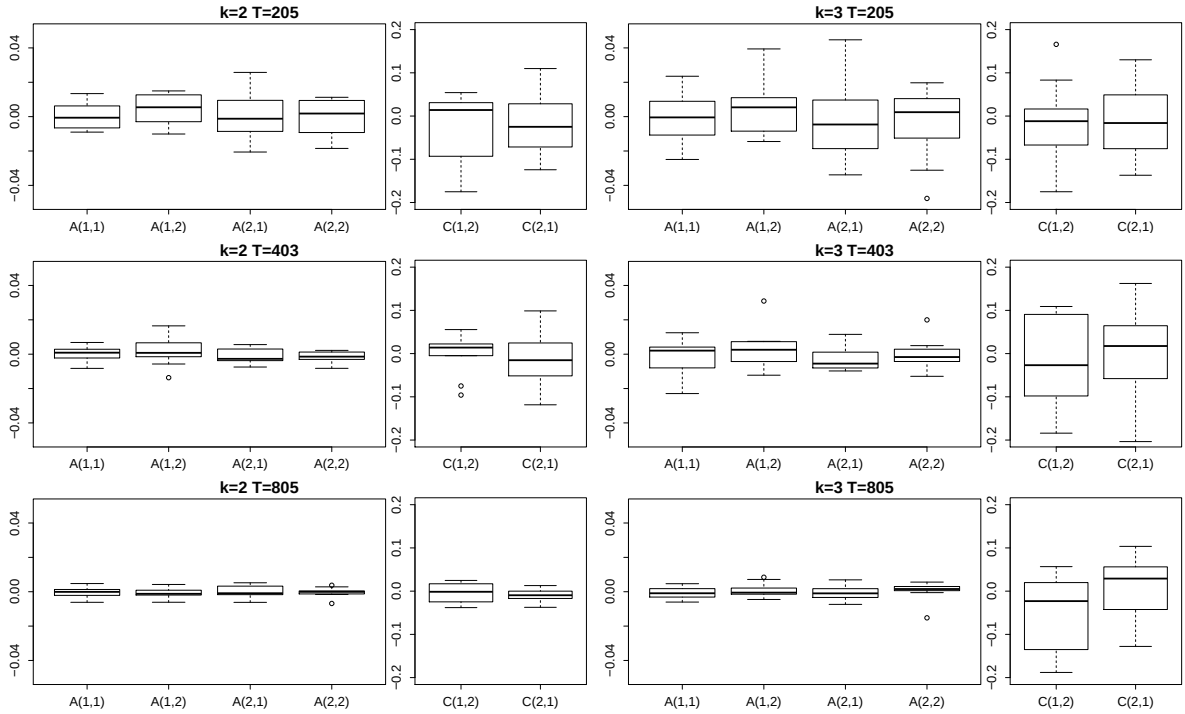


Figure 5.4: As in Fig. 5.3 for $A^{(2)}$ and $C^{(2)}$.

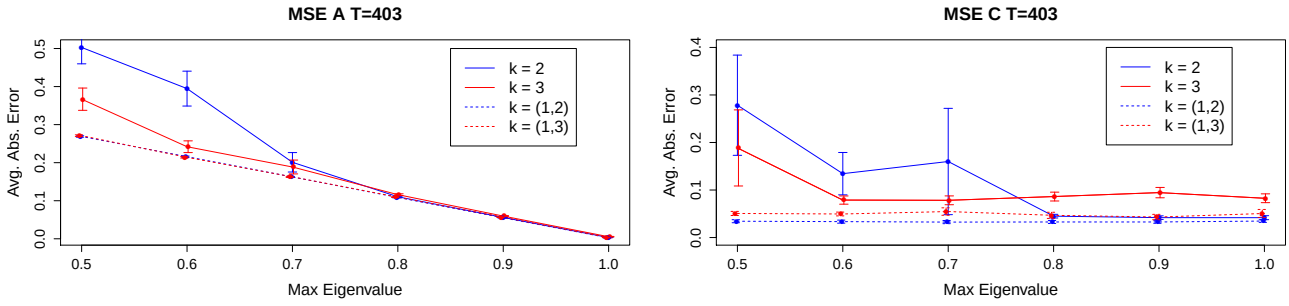


Figure 5.5: Average mean squared error in estimation of A (left) and C (right) as a function of maximum eigenvalue of A. Error bars indicate one standard error from 40 simulation runs.

5.7 Real Data

5.7.1 Subsampled Ozone Data

We use the subsampled structural model to analyze the causal scale and pathways in an ozone and temperature data set. The temperature ozone data are the 50th causal effect pair from the website `https://webdav.tuebingen.mpg.de/cause-effect/`, and were also considered in [69]. The dataset consists of temperature and ozone concentration, sampled daily, for $T = 365$ days. First we standardize each time series to mean zero and unit variance. We fit the subsampled structural model to the preprocessed series for $k = (1, 2, 3, 4)$ subsampling regimes under both independent errors, $C = I$, and structural covariance in the instantaneous errors, C free. To ensure that good optima are found we perform 30,000 restarts and run the adaptive over−relaxed algorithm until the relative change in log−likelihood is less than 10^{-6}.

The estimated $\hat{A}$ for $k = 1$ is

$$\hat{A} = \begin{pmatrix} 0.669 & 0.175 \\ -0.050 & 0.992 \end{pmatrix},$$

with maximum eigenvalue of 0·962, suggesting that accurate estimation of subsampled parameters is possible. The Bayesian information criterion score for all models is displayed in Table 5.4. Across all subsampling rates, the structural model, C free, has lower score, indicating that the two extra parameters of the structural model, the off diagonal elements of C, provide necessary flexibility. Furthermore, the best performing model is the structural model with subsampling rate $k = 2$. The estimated transition matrix at $k = 2$ is

$$\hat{A} = \begin{pmatrix} 0.849 & 0.058 \\ -0.027 & 0.981 \end{pmatrix},$$

similar to that given by [69] for $C = I$. After normalizing columns, we obtain

$$\hat{C} = \begin{pmatrix} 1.0 & 0.2 \\ 0.29 & 1.0 \end{pmatrix}, \quad \hat{\Sigma} = \hat{C}\hat{\Lambda}(e_t)\hat{C}^T = \begin{pmatrix} 0.199 & 0.054 \\ 0.053 & 0.054 \end{pmatrix}.$$

Model / k	1	2	3	4
$C = I$	901·96	791·02	839·56	797·00
C free	784·53	777·78*	790·46	791·23

Table 5.4: Bayesian information criterion scores of subsampling and covariance types on the atmospheric dataset. An asterisk denotes the lowest value.

These results indicate weak lagged effects at the subsampled scale, but stronger instantaneous effects between temperature and ozone. Furthermore, the temperature series obtains most of its power from a strong error variance, while the ozone series is driven relatively more by the autoregressive component. See Section 5.9 for quantile-quantile plots of the inferred mixture of error distributions.

5.7.2 Mixed–Frequency: GDP and Treasury Bonds

We perform a structural autoregressive analysis on the mixed−frequency data set of quarterly gross domestic product (GDP) and monthly price of treasury bonds. The data set has been previously compiled and analyzed in the mixed−frequency setting by [169] and is available on the author's website. We follow [169] and compute the log of both series. Furthermore, as is common in mixed−frequency analysis [30, 209], we compute first differences to remove first order non−stationarities.

There are multiple approaches to modeling mixed−frequency GDP in the literature. Some authors treat GDP as a flow variable and use state-space models to directly model the aggregation over months in a quarter [169, 66]. Flow variables are ones that are aggregates over a period of time, rather than stock variables which are snapshots of a value at a particular time. Other authors ignore the generative subsampling structure and instead jointly model the high and low frequency variables in a quarter using mixed data sampling methods [66]. We follow another line of work [30, 54, 4, 172, 210] and treat GDP as a purely subsampled series and apply our mixed−frequency structural autoregressive model at the monthly rate. Indeed, recent theoretical work on mixed−frequency autoregressive models for GDP also focuses on the purely subsampled,

rather than aggregated, case [4]. Extensions of our framework to handle aggregated variables is an important line of future work.

In the traditional approaches to mixed−frequency analyses, A and the instantaneous covariance Σ are generically identifiable from the first two moments [4]. What sets our non−Gaussian approach apart in this mixed−frequency domain is the ability to uniquely identify the ordering of the instantaneous causal effects in the structural matrix C. To highlight this ability, we perform model selection on the zero entries in C to determine the causal ordering of the instantaneous effects. Specifically, we calculate the Bayesian information criterion for the models $M : C_{1,1} = C_{2,1} = 0$, $M_{GDP \to TB} : C_{1,2} = 0$, $M_{TB \to GDP} : C_{2,1} = 0$, and $M_{GDP \to TB, TB \to GDP}$. Models M, $M_{GDP \to TB}$ and $M_{TB \to GDP}$ represent acyclic structures on the instantaneous effects while the unrestricted model $M_{GDP \to TB, TB \to GDP}$ does not. The scores for all models in Table 5.5 indicate that $M_{GDP \to TB}$ performs best. The estimated matrices of

$$\hat{A} = \begin{pmatrix} 0.297 & -0.068 \\ 0.012 & 0.658 \end{pmatrix} \qquad \hat{C} = \begin{pmatrix} 0.950 & 0.0 \\ 0.280 & 0.695 \end{pmatrix},$$

suggest a slight negative lagged interaction from GDP to treasury bonds and an instantaneous interaction at the monthly scale from treasury bonds to GDP. See Section 5.9 for quantile-quantile plots of the inferred mixture of error distributions.

The above analysis fits a structural model at the time scale of months, the same sampling rate as the treasury bond time series. The results from Section 5.4 indicate that we could uniquely identify models at bi−monthly, or even more granular, time scales. However, even at the bi−monthly rate, the computational complexity of the E−step of our algorithm becomes prohibitive due to the large number of combinations of error mixture components in a data block, as discussed in Section 5.5.4. Since the E−step requires running the forward−backward algorithm many times, a considerable computational speed up could be achieved from a parallel implementation.

Model	M	$M_{TB\to GDP}$	$M_{GDP\to TB}$	$M_{GDP\to TB,TB\to GDP}$
	1984·00	1983·41	1981·08*	1987·55

Table 5.5: Bayesian information criterion scores of different instantaneous causality structures on the GDP data set. An asterisk denotes the lowest value.

5.8 Discussion

Our results provide sufficient conditions for identifiability of structural autoregressive models for both subsampled and mixed frequency series. The causal diagram of both lagged and instantaneous effects is identifiable under arbitrary subsampling and non−Gaussian errors.

We developed an exact expectation−maximization algorithm for estimation and analyzed its performance via simulations. Our algorithm has two drawbacks: high complexity due to a Kalman filter evaluation for all mixture error assignments within a time block; and many local optima due to weak identifiability. Our simulations show that the latter is more severe under even subsampling factors and low signal to noise regimes.

An ongoing line of work is to develop approximate inference for these models using Markov chain Monte Carlo or variational methods. Unfortunately, we have found that the local optima makes sampling difficult. A Gibbs sampler we have explored gets stuck in one local mode and requires the same number of random restarts as our algorithm to find a good solution. Perhaps incorporating recent advances in sampling [131] may prove beneficial. We have also attempted a variational algorithm but found that performance was poor. [69] also reported significantly worse results for a variational approach than for their approximate expectation−maximization algorithm. By breaking the dependence between the unobserved, subsampled x_t and the auxiliary z_ts, the variational approach avoids the combinatorial evaluation of a Kalman filter; however, this dependence is critical for correctly evaluating the probable trajectories of the latent x_t, without which inference of A suffers.

As a future direction, it would also be interesting to explore method of moments estimation for this problem. Such an approach may side step both the local optima problem and the combinatorial

complexity of our algorithm.

5.9 Appendix

5.9.1 Example of Subsampled and Mixed Frequency Structural Autoregressive Process

Subsampled Process

To provide intuition about the forms of the subsampled and mixed frequency structural autoregressive processes given in Equation (5.3) and Equation (5.6), respectively, of the main Chapter and how these equations explicitly relate to the A and C parameters, we consider the following bivariate example. Let the parameters of the process be given by

$$A = \begin{pmatrix} 0.8 & 0.1 \\ -0.2 & 0.8 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0.2 \\ 0 & 1 \end{pmatrix}.$$

For a subsampled process with a subsampling factor of $k = 3$ we have that

$$A^3 = \begin{pmatrix} 0.464 & 0.190 \\ -0.380 & 0.464 \end{pmatrix}, \quad AC = \begin{pmatrix} 0.8 & 0.26 \\ -0.2 & 0.76 \end{pmatrix}, \quad A^2C = \begin{pmatrix} 0.62 & 0.284 \\ -0.32 & 0.556 \end{pmatrix},$$

which gives the evolution equation for the subsampled process as:

$$\begin{aligned} \tilde{x}_t &= A^3\tilde{x}_{t-1} + Ce_{1+tk} + ACe_{1+tk-1} + A^2Ce_{1+tk-2} \\ &= \begin{pmatrix} 0.464 & 0.190 \\ -0.380 & 0.464 \end{pmatrix} \tilde{x}_{t-1} \\ &\quad + \begin{pmatrix} 1 & 0.2 \\ 0 & 1 \end{pmatrix} e_{1+tk} + \begin{pmatrix} 0.8 & 0.26 \\ -.2 & 0.76 \end{pmatrix} e_{1+tk-1} + \begin{pmatrix} 0.62 & 0.284 \\ -0.32 & 0.556 \end{pmatrix} e_{1+tk-2}. \end{aligned}$$

Mixed Frequency Process

For the mixed frequency case of $k = (1, 3)$ we have that

$$A^{(\bar{1})} = A^{(\bar{2})} = \begin{pmatrix} 0 & 0 \\ -0.2 & 0.8 \end{pmatrix}, \quad C^{(\bar{1})} = C^{(\bar{2})} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

so that

$$AA^{(\bar{1})} = \begin{pmatrix} -0.02 & 0.08 \\ -0.16 & 0.64 \end{pmatrix}, \quad AA^{(\bar{1})}A^{(\bar{2})} = \begin{pmatrix} -0.016 & 0.064 \\ -0.128 & 0.512 \end{pmatrix}$$

and

$$AC^{(\bar{1})} = \begin{pmatrix} 0 & 0.1 \\ 0 & 0.8 \end{pmatrix}, \quad AA^{(\bar{1})}C^{(\bar{2})} = \begin{pmatrix} 0 & 0.08 \\ 0 & 0.64 \end{pmatrix}.$$

The entire mixed frequency structural autoregressive process may then be written as

$$\begin{aligned} x_t &= F\tilde{x}_t + L\tilde{e}_t \qquad (5.13) \\ &= AI^{(1)}x_{t-1} + AA^{(\bar{1})}I^{(2)}x_{t-2} + AA^{(\bar{1})}A^{(\bar{2})}I^{(3)}x_{t-3} + Ce_t + AC^{(\bar{1})}e_{t-1} + AA^{(\bar{1})}C^{(\bar{2})}e_{t-2} \\ &= \begin{pmatrix} 0.8 \\ -0.2 \end{pmatrix} x_{(t-1)1} + \begin{pmatrix} -0.02 \\ -0.16 \end{pmatrix} x_{(t-2)1} + \begin{pmatrix} -0.016 & 0.064 \\ -0.128 & 0.512 \end{pmatrix} \begin{pmatrix} x_{(t-3)1} \\ x_{(t-3)2} \end{pmatrix} \\ &\quad + \begin{pmatrix} 1 & 0.2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e_{t1} \\ e_{t2} \end{pmatrix} + \begin{pmatrix} 0 & 0.1 \\ 0 & 0.8 \end{pmatrix} \begin{pmatrix} e_{(t-1)1} \\ e_{(t-1)2} \end{pmatrix} + \begin{pmatrix} 0 & 0.08 \\ 0 & 0.64 \end{pmatrix} \begin{pmatrix} e_{(t-2)1} \\ e_{(t-2)2} \end{pmatrix}. \end{aligned}$$

5.9.2 Proof of Theorem 2 (subsampled case only)

We prove it for the subsampled case. The structural vector autoregressive model can be decomposed as:

$$\begin{aligned}\tilde{x}_t &= A^k \tilde{x}_{t-k} + L\tilde{e}_t \\ &= A^k \tilde{x}_{t-1} + \vec{e}_t,\end{aligned}$$

where $L = (C, AC, \ldots, A^{k-1}C)$ and $\vec{e}_t = L\tilde{e}_t$. We may determine A^k uniquely by linear regression and thus determine the distribution of $\vec{e}_t$. Proposition 1 states that each column of L' is a scaled version of a column of L. Denote by L_{lp+i}, $l = 0, \ldots, k-1$, $i = 1, \ldots, p$ the $(lp+i)$th column of L, and similarly for L'_{lp+i}. From the Uniqueness Theorem in Erikson and Koivunen 2004 [55], we know that under Assumption 2, for each i, there exists one and only j such that the distribution of $e_{(t-l)i}, l = 1, \ldots, k-1$ is the same as the distribution of $e'_{(t-l)j}$, $l = 1, \ldots, k-1$ up to changes in location and scale. This implies that each column in L_{lp+i}, $l = 0, \ldots, k-1$, is proportional to at least one of the nonzero columns in L_{lp+j}, $l = 1, \ldots, k-1$, and vice versa. The proportionality must be either 1 or -1 since we have standardized the p_e to have unit variance. Furthermore, it must be the case that each L_{lp+i} is proportional to some column L'_{lp+j} for all $j, i \in (1, \ldots, p)$ since the columns are ordered in magnitude in both L and L', ie $\|L_{lp+i}\|_2 > \|L_{(l+1)p+i}\|_2$,

$$\begin{aligned}\|L_{(l+1)p+i}\|_2 &= \|AA^l C_{:i}\|_2 \\ &\leq \|A\|_2 \|A^l C_{:i}\|_2 \\ &< \|A^l C_{:i}\|_2 \\ &= \|L_{lp+i}\|_2.\end{aligned}$$

This implies that L' may be written as:

$$\begin{aligned} L' &= LP \\ &= \left(CP_0, ACP_1, \ldots, A^{k-1}CP_{k-1}\right), \end{aligned}$$

where P_i is a scaled permutation matrix with either 1s or -1 scaling factors, and P_i and P_j have the same permutation pattern but potentially different scaling factors. This proves the first assertion, i.e. $C' = CP_0$ and $\Sigma' = C'C'T = CP_0P_0^TC^T = CC^T = \Sigma$. Now, if the p_e are restricted to be nonsymmetric then the scaling factors must all be 1 so that all the P_i are equal.

$$\begin{aligned} A'C' &= A'CP \\ &= ACP \end{aligned}$$

and since C is full rank, CP is full rank so that $A' = A$, as desired.

5.9.3 Proof of Theorem 4 part 2

If C is lower triangular then $C = C'$. Now, $AC = A'C'P_1 = A'CD$ where D is diagonal with either 1 or -1 on the diagonal. This implies that $L'_{p+1:2p} = ACD$. We proceed by induction. Since the last column of C, $C_{:p}$, is zeros everywhere except the last element, we must have that $C_{pp}A_{:p}D_{pp} = L'_{2p} = C_{pp}A'_{:p}$, so that $A_{:p}D_{pp} = A'_{:p}$. Following the same logic as the proof to part (b) of Theorem 3, if there exists some j such that a multiple of k_p is one less than a multiple of k_j and $A_{pj} \neq 0$, then we can identify A_{pj}, and hence its sign, implying $A_{:p} = A'_{:p}$.

Assume that $A_{:i} = A'_{:i}$ for $i > j$. Since C is lower diagonal we must have that

$$\begin{aligned} L'_{p+j} &= \left(C_{jj}A'_{:j} + \sum_{i>j} C_{ij}A_{:i}\right) \\ &= D_{jj}\left(C_{jj}A_{:j} + \sum_{i>j} C_{ij}A_{:i}\right). \end{aligned}$$

Since $A_{lj} = A'_{lj}$ with $C_{jj}A_{lj} + \sum_{i>j} C_{ij}A_{li} \neq 0$ for some l, this implies $D_{jj} = 1$, so that $A_{:j} = A'_{:j}$. Taken together, $A = A'$.

5.9.4 Expectation-maximization algorithm details

The gradient of the expected joint log probability given in the main text with respect to $W = C^{-1}$ is

$$\begin{aligned}\nabla l(W) = TW^{-T} + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i=1}^{m}\frac{1}{\sigma_{ji}^2}\Big[&- E(z_{tji}x_t x_t^T \tilde{X})W_j^T - AE(z_{tij}x_{t-1}x_{t-1}^T \mid \tilde{X})A^T W_j^T \\ &+ \left\{E(z_{tji}x_t x_{t-1}^T \mid \tilde{X})A^T + AE(z_{tjj}x_{t-1}x_t \mid \tilde{X})\right\}W_i^T \\ &+ E(z_{tji}x_t \mid \tilde{X})\mu_{ji} - AE(z_{tji}x_{t-1}|\tilde{X})\mu_{ji}\Big].\end{aligned}$$

The Hessian with respect to $w = \text{vec}(W)$ is

$$H(w) = -T\Omega(W^{-T} \otimes W^{-1}) + \sum_{t=1}^{T}\sum_{j=1}^{p}\sum_{i=1}^{m}\Gamma_{tji} \otimes D^{(j)},$$

where

$$\begin{aligned}\Gamma_{tji} = &\frac{1}{\sigma_{ji}^2}\Big\{ - E(z_{tji}x_t x_t^T \mid \tilde{X}) - AE(z_{tji}x_{t-1}x_{t-1}^T)A^T + E(z_{tji}x_t x_{t-1}^T \mid \tilde{X})A^T \\ &+ AE(z_{tji}x_{t-1}x_t^T \mid \tilde{X})\Big\}\end{aligned} \tag{5.14}$$

and $D^{(j)}$ is a $p \times p$ matrix with $D_{jj}^{(j)} = 1$ and all other entries zero. Ω is a permutation matrix with all zero entries except with $\Omega_{nm} = 1$ for all $n \in (1 \ldots p^2)$ and $m = (n-1)\text{mod}(p) + \lfloor (n-1)/p \rfloor + 1$. Finally, note there is a non-identifiability between the scale of the errors, e_t, and the magnitude of C. For algorithmic stability we fix the first mixture component for each e_t to have variance set to one, $\sigma_{j1}^2 = 1$ for all j.

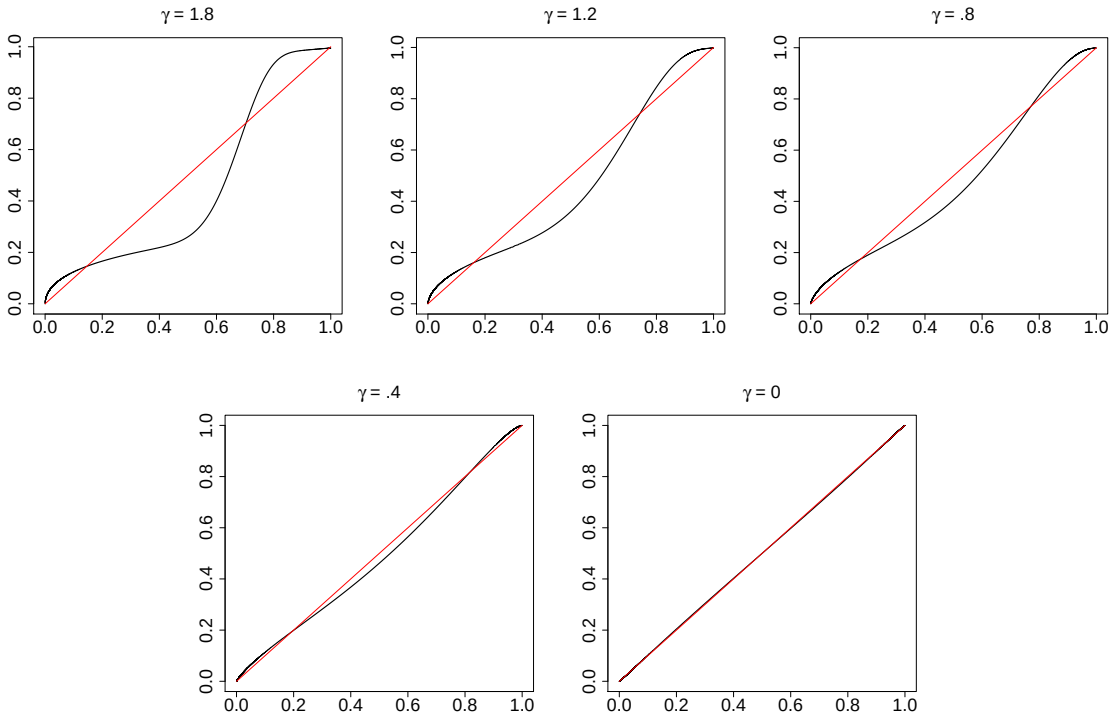


Figure 5.6: Quantile-quantile plots of the five mixture of normal distributions with varying skewness, γ, values used in the asymmetry simulation experiments.

5.9.5 Non-Gaussian Errors and Skewness Simulations

Here we provide additional details and analysis for the skewness simulation experiment in the main Chapter. The μ,σ^2, and π values used for the skewness values are shown in Table 5.6 and the quantile-quantile plots comparing each distribution to that of a Gaussian with the same variance are shown in Figure 5.6. For comparison with the real data experiments, we also show similar quantile-quantile plots for the estimated error distributions in our two real data experiments on the Ozone and mixed frequency GDP data in Figure 5.7.

The average MSE values for estimating the structural matrix C are displayed in Table 5.7. The simulation patterns for C estimation are similar to that of A with a few differences. First, while for A there was stable estimation for $A^{(1)}$ at $k = 2$ and $\gamma = 0$, it seems that estimation is unstable for C at $\gamma = 0$ across all conditions. However, overall C estimation appears more robust in general to

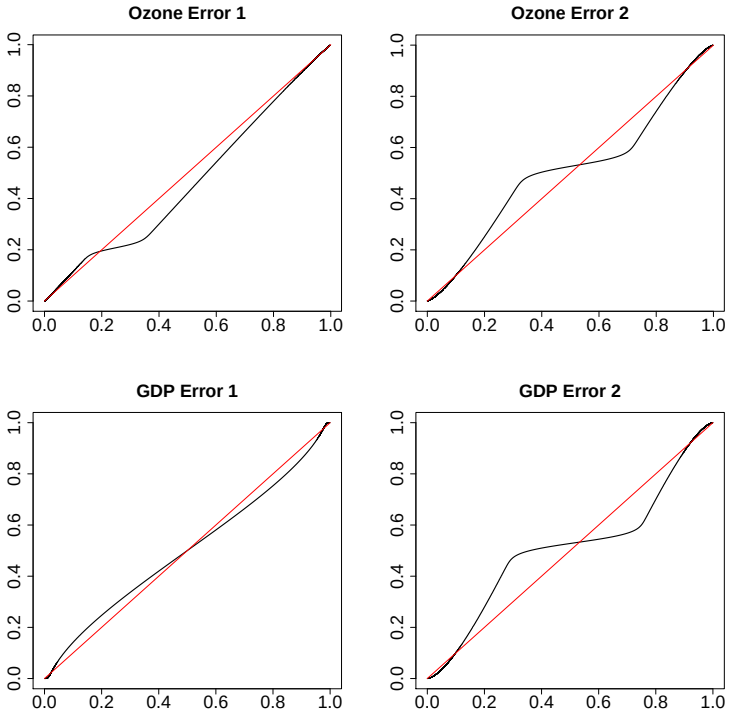


Figure 5.7: Quantile-quantile plots of the estimated mixture of Gaussians error distributions from the (top) Ozone data fit at a rate of $k = 2$ and (bottom) mixed frequency GDP data.

γ	1·8	1·2	0·8	0·4	0
μ_{11}	0·36	0·380	0·403	0·409	0·408
μ_{12}	-0·84	-0·705	-0·604	-0·499	-0.408
σ_{11}	0·2	0·344	0·419	0·553	0·681
σ_{12}	1·0	0·904	0·839	0·757	0·681
π_{11}	0·7	0·65	0·6	0·55	0·5
π_{12}	0·3	0·35	0·4	0·45	0·5

Table 5.6: A table displaying the bivariate mixture model parameter values for e_1 in the different error distributions used in the skewness simulation experiments. The e_2 distributions were the same except for $\mu_{21} = -\mu_{11}$ and $\mu_{22} = -\mu_{12}$, resulting in a negative skewness of the same magnitude for e_2.

k			2					3		
γ	1.8	1.2	0.8	0.4	0	1.8	1.2	0.8	0.4	0
$A^{(1)}\, C^{(1)}$	-6.0	-4.2	-3.1	-1.4	-0.49	-4.6	-2.4	-2.2	-1.1	-0.33
$A^{(1)}\, C^{(2)}$	-6.0	-4.2	-3.2	-1.2	-0.66	-4.6	-2.3	-2.1	-1.1	-0.4
$A^{(2)}\, C^{(1)}$	-5.3	-3.8	-2.5	-1.2	-0.38	-3.3	-2.1	-1.3	-0.5	-0.84
$A^{(2)}\, C^{(2)}$	-5.2	-3.9	-2.5	-0.9	-0.35	-3.5	-2.0	-0.94	-0.48	-0.55

Table 5.7: A table displaying the $\log$ of the average C MSE over ten random samplings for both A and C estimates across multiple settings of the parameters, skewness of the error distributions, and subsampling factors.

smaller γ values as estimation seems to become unstable around $\gamma = .4$ for $k = 4$ and $\gamma = .8$ or $\gamma = .4$ for $k = 3$.

5.9.6 Additional Simulation Plots

Here we provide additional histogram plots from simulations in the main text. Figures 5.8 and 5.9 provide estimates for the remaining $(A^{(1)}, C^{(2)})$ and $(A^{(2)}, C^{(1)})$ simulation parameter configurations.

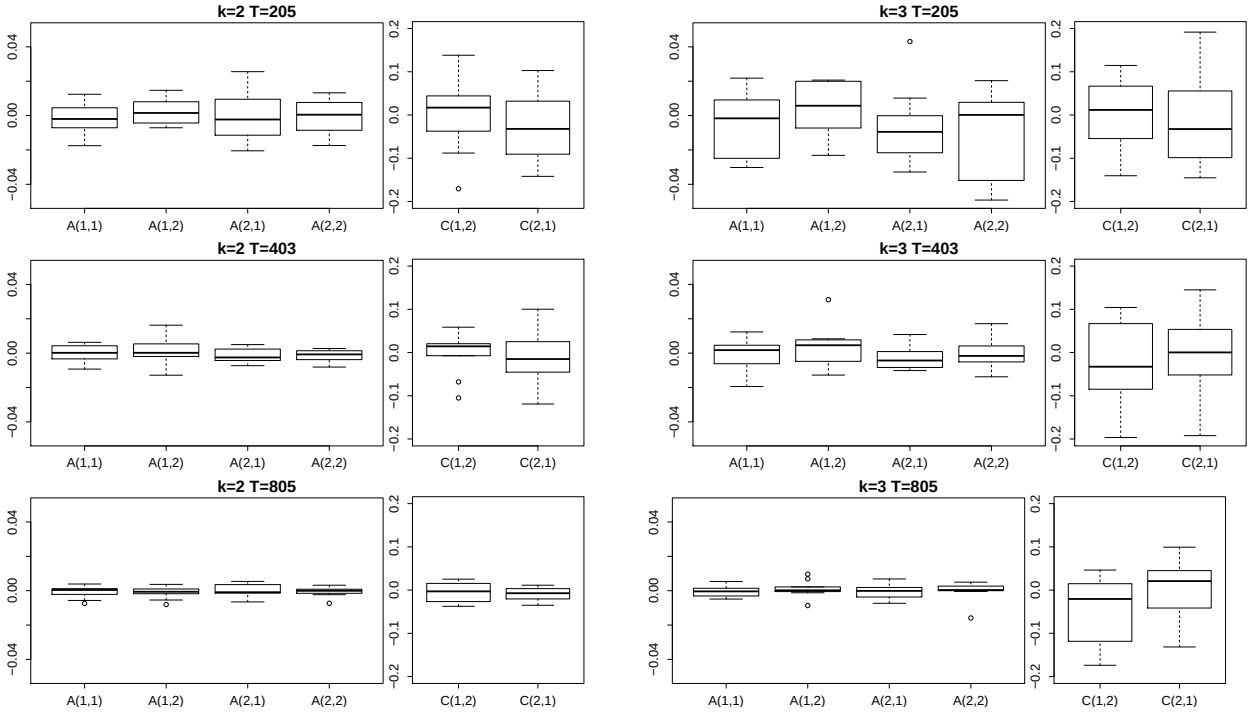


Figure 5.8: Boxplots of $A^{(2)}$ and $C^{(1)}$ parameter estimates over 10 random data samplings. The original series is either of length 203 (top), 403 (middle) or 805 (bottom) and then subsampled at (left) $k = 2$ and (right) $k = 3$.

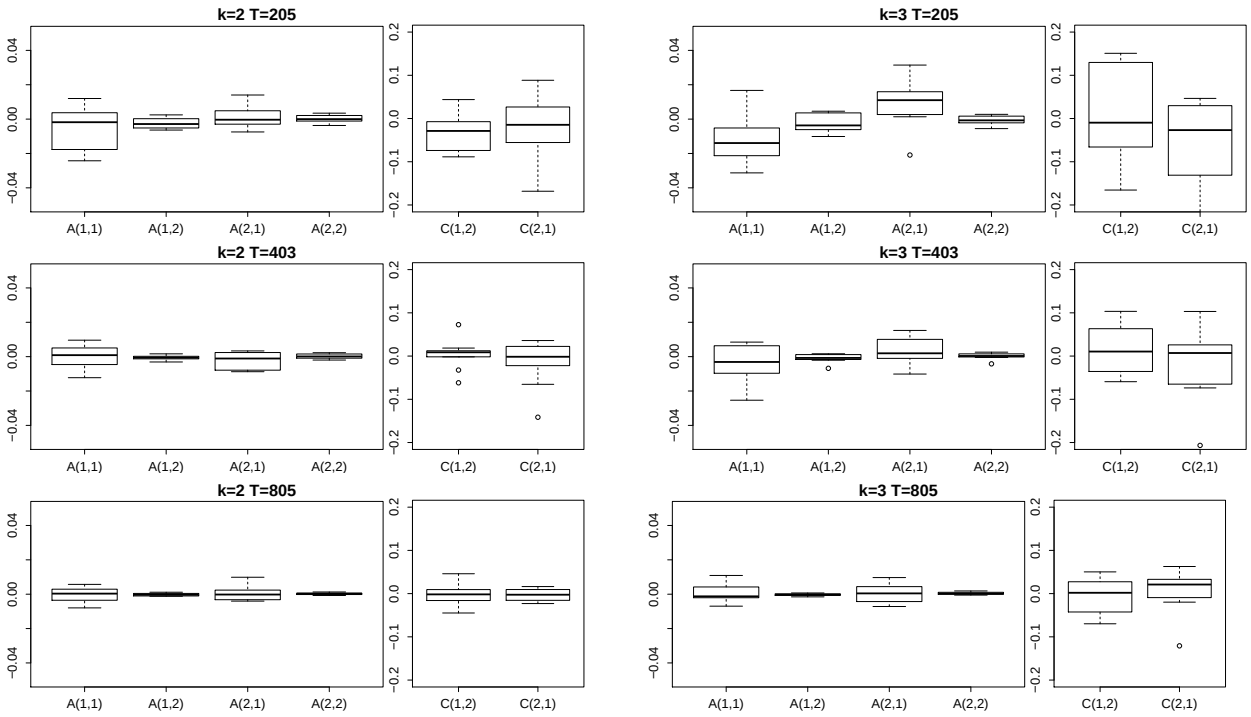


Figure 5.9: Boxplots of $A^{(1)}$ and $C^{(2)}$ parameter estimates as in Figure 5.8.

Chapter 6

GRANGER CAUSALITY FOR MULTIVARIATE CATEGORICAL TIME SERIES

6.1 Introduction

Granger causality [71] is a popular framework for assessing the relationships between time series, and has been widely applied in econometrics, neuroscience, and genomics, amongst other fields. Given two time series x and y, the idea is to use the temporal structure of the data to assess whether the past values of one, say x, are predictive of future values of the other, y, beyond what the past of y can predict alone; if so, x is said to *Granger cause* y. Recently, the focus has shifted to inferring Granger causality networks from multivariate time series data, with the goal of uncovering a sparse set of Granger causal relationships amongst the individual univariate time series. Building on the typical autoregressive framework for assessing Granger causality, a majority of approaches for inferring Granger causal networks have focused on real-valued Gaussian time series using the vector autoregressive model (VAR) with sparsity inducing penalties [76, 175]. More recently, this approach has been extended to non-Gaussian data such as multivariate point processes using sparse Hawkes processes [214], count data using autoregressive Poisson generalized linear models [74], or even time series with heavy tails using VAR models with elliptical errors [153]. In contrast, inferring networks for multivariate *categorical* time series has not been studied under this paradigm.

Multivariate categorical time series arise naturally in many domains. For example, we might have health states from various indicators for a patient over time, voting records for a set of politicians, action labels for players on a team, social behaviors for kids in a school, or musical notes in an orchestrated piece. There are also many datasets that can be viewed as binary multivariate time series based on the presence or absence of an action for some set of entities. Furthermore, in some applications, collections of continuous-valued time series are each quantized into a small

set of discrete values, like the weather data from multiple stations analyzed in [48], wind data in [157], stock returns in [141], or sales volume for a collection of products in [32].

The *mixture transition distribution* (MTD) model [15, 157], originally proposed for parsimonious modelling of higher order Markov chains, can provide an approach to modeling multivariate categorical time series [32, 141, 215]. The MTD model reduces each categorical interaction to a standard single dimensional Markov transition probability table. While alluring due to its elegant construction and intuitive interpretation, widespread use of the MTD model has been limited by a non-convex objective with many local optima, a large number of parameter constraints, and unknown identifiability conditions [141, 215, 14]. For this reason, most applications of the MTD model to multivariate time series have looked at a maximum of three or four time series. To bypass the limitations of MTD, autoregressive generalized linear models have been advocated for categorical time series. In particular, autoregressive generalized linear binomial models are often used for the special case of two categories per series [74, 9]. However, their multinomial-output extension to a larger number of states per series has not been widely adopted. See [100] for an application to the univariate time series case.

We refer to the autoregressive multinomial GLM as the mixture logistic transition distribution (mLTD). The mLTD model uses a logistic function to bypass parameter constraints, results in a convex objective, and has well-known identifiability conditions. However, these advantages of mLTD come at the cost of reduced interpretability, mainly because the transition distribution in mLTD depends nonlinearly on the model parameters. [141] has recently proposed a constrained autoregressive probit model that improves interpretability. However, the probit model is both highly non-convex and inference is computationally intensive, limiting applications to higher dimensional series. As such, one is still torn between a computational and interpretability tradeoff. We address this issue by going back to the interpredability of the MTD framework and showing how one can dramatically improve its computationational drawbacks.

In particular, we recast inference in the MTD model as a convex problem through a novel reparameterization. We further develop a regularized estimation framework for identifying Granger causality for multivariate categorical time series. We also establish for the first time conditions for

identifiability in the MTD model and compare the identifiability conditions for MTD and mLTD models. We find that while the identifiability conditions for the MTD model are given by a non-convex set, we may easily enforce the constraints using our convex re-parameterization trick by augmenting the likelihood with appropriate convex penalties. We then develop an efficient projected gradient algorithm for optimizing the penalized convex MTD objective. Our efficient algorithm depends on a Dykstra splitting method for projection onto the constraint sets of the MTD model. This computational approach for MTD provides enormous gains over past methods, enabling this model to be applied to large, modern datasets for the first time. Importantly, the computational insights provided in this Chapter carry over to the suite of other applications of MTD models, such as higher order Markov chains, beyond the multivariate categorical time series which are the focus herein.

As a comparison benchmark we also develop a penalized mLTD model for Granger causality in multivariate Markov chains. While straightforward, the application of the penalized mLTD framework to multivariate categorical time series with more than two categories is new. We compare MTD and mLTD methods under multiple simulation conditions and use the MTD method to uncover Granger causality structure in a music data set. Studying the potential theoretical benefits of one framework over the other is left as future work.

6.2 Categorical Time Series and Granger Causality

6.2.1 Granger Causality

Let $x_t = (x_{1t}, \dots x_{dt}), \in \mathcal{X}$ denote a d-dimensional categorical random variable indexed by time where $\mathcal{X} = (\mathcal{X}_1 \times \mathcal{X}_2 \dots \times \mathcal{X}_d)$, with $\mathcal{X}_i$ denoting the set of possible values of x_{it}. Let $m_i = |\mathcal{X}_i|$ be the cardinality of set $\mathcal{X}_i$, i.e. the number of categories series i may take. A length T multivariate categorical time series is the sequence $X = \{x_1, \dots, x_t, \dots, x_T\}$. An order k multivariate Markov chain models the transition probability between the categories at lagged times $t-1, \dots, t-k$ and

those at time t using a transition probability distribution:

$$p(x_t|x_{t-1},\ldots) = p(x_t|x_{t-1},\ldots,x_{t-k}). \tag{6.1}$$

Due to the complexity of fully parameterizing this transition distribution, it is common to simplify the model and assume that the categories at time t are conditionally independent of one another given the past realizations:

$$p(x_t|x_{t-1},\ldots,x_{t-k}) = \prod_{i=1}^{d} p(x_{it}|x_{t-1},\ldots,x_{t-k}). \tag{6.2}$$

For simplicity, we assume $k = 1$, but stress that all models and results equally apply to higher orders of k. Based on the decomposition assumption, Equation (6.2), the problem of estimation and inference decomposes into independent subproblems over each series i. Using this decomposition, we define Granger non-causality for two categorical time series x_{it} and x_{jt} as follows.

Definition 1. *Time series x_j is not Granger causal for time series x_i iff*

$$p(x_{it}|x_{1(t-1)},\ldots,x_{j(t-1)},\ldots x_{d(t-1)}) = p(x_{it}|x_{1(t-1)},\ldots,x_{(j-1)(t-1)},x_{(j+1)(t-1)},\ldots,x_{d(t-1)}).$$

Definition 1 states that x_{jt} is not Granger causal for time series x_{it} if the probability that x_{it} is in a given state at time t is conditionally independent of the value of $x_{j(t-1)}$ at time $t-1$ given the values of all other series $x_{k(t-1)}$, $k \neq i, j$, at time lag $t-1$. Definition 1 is natural since it implies that if x_{it} does not Granger cause x_{jt}, then knowing $x_{i(t-1)}$ does not help predicting the future state of series j, x_{jt}. For real-valued data, classical definitions of Granger non-causality generally state that the conditional mean, in homoskedastic models, or conditional variance, in heteroskedastic models, of x_{jt} do not depend on the past values x_{it}. Thus, Definition 1 is a generalization of the classical case to multivariate categorical data, where notions like conditional mean and variance are less applicable. While this definition of Granger causality is intuitive and similar to other definitions for real-valued data, it has not been explicitly stated for multivariate categorical time

series and represents a contribution of our work.

6.2.2 Tensor Representation for Categorical Time Series

Each individual conditional distribution in Equation (6.2) can be represented as a conditional probability tensor $\tilde{\mathbf{P}}^i$ with $d+1$ modes of dimension $m_i \times m_1 \times \ldots \times m_d$. Each entry of the tensor is given by

$$\tilde{\mathbf{P}}^i_{x_{it},x_{1(t-1)},\ldots,x_{d(t-1)}} = p(x_{it}|x_{1(t-1)},\ldots,x_{d(t-1)}). \tag{6.3}$$

Definition 1 may be stated equivalently using the language of tensors: x_j does not Granger cause x_i if all subtensors along the mode associated with x_j are equal. Specifically,

$$\tilde{\mathbf{P}}^i_{1:m_i,1:m_1,\ldots,x_{j(t-1)}=1,\ldots,1:m_d} = \tilde{\mathbf{P}}^i_{1:m_i,1:m_1,\ldots,x_{j(t-1)}=2,\ldots,1:m_d} = \ldots = \tilde{\mathbf{P}}^i_{1:m_i,1:m_1,\ldots,x_{j(t-1)}=m_j,\ldots,1:m_d}. \tag{6.4}$$

This subtensor view of Granger non-causality in categorical time series is displayed graphically in Figure 6.1.

The tensor interpretation suggests a naive penalized likelihood method to select for Granger non-causality in categorical time series: perform penalized maximum likelihood estimation of the conditional probability tensor with a penalty that enforces equality among the subtensors of each mode. While we have explored the above approach in low dimensions, $d \leq 5$, memory, and in turn, computational requirements for storing the complete probability tensor becomes infeasible for even moderate dimensions since $\tilde{\mathbf{P}}^i$ has $m_i \times m_1 \times \ldots m_d$ entries. Other authors have modeled the conditional probability distribution of Markov chains using a Bayesian nonparametric higher order singular value decomposition [167] that adaptively shrinks the number of parameters needed to represent the high dimensional tensor. We take an alternative approach and instead, in Sections 6.2.3 and 6.2.4, present tensor parameterizations where the number of parameters needed to represent the full conditional probability tensor grows linearly with d. We establish Granger non-causality conditions and associated penalized likelihood methods for estimation under these

parameterizations in Sections 6.3 and 6.4, respectively.

In specifying our models, and throughout the remainder of the chapter, we focus in on a single conditional of x_{it} given x_{t-1} in Equation (6.2). For notational simplicity, we drop the i index.

6.2.3 The MTD model

The MTD model [157] provides an elegant and intuitive parameterization of the multivariate Markov transition distribution as a convex combination of pairwise transition probabilities. Specifically, the MTD model is given by:

$$p(x_{it}|x_{1(t-1)},\ldots,x_{d(t-1)}) = \gamma_0 p_0(x_{it}) + \sum_{j=1}^{d} \gamma_j p_j(x_{it}|x_{j(t-1)}), \tag{6.5}$$

where p_0 is a probability vector, $p_j(.|.)$ is a pairwise transition probability table between $x_{j(t-1)}$ and x_{it} and $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_d)$ is a $d+1$ dimensional probability distribution, i.e. that $\mathbf{1}^T\gamma = 1$ with $\gamma_j \geq 0$, $j = 0, \ldots, d$. We let the matrix $\mathbf{P}^j \in \mathbb{R}^{m_i \times m_j}$ denote the pairwise transitions $\mathbf{P}^j_{x_{it},x_{j(t-1)}} = p_j(x_{it}|x_{j(t-1)})$. Thus, $\mathbf{1}^T\mathbf{P}^j = \mathbf{1}^T$, $\mathbf{P}^j_{lk} \geq 0$, $l = 1, \ldots, m_i$, $k = 1, \ldots, m_j$. We also let $\mathbf{p}^0 \in \mathbb{R}^{m_i}$ denote the intercept, where $\mathbf{p}^0_{x_{it}} = p_0(x_{it})$. While past formulations of the MTD model neglect the p_0 intercept term, we show below that the intercept is crucial for model identifiability and, consequently, Granger causality inference. Finally, we note that the MTD model may be extended by adding in interaction terms for pairwise effects [15], such as $p_{jk}(x_{it}|x_{j(t-1)}, x_{k(t-1)})$, though we focus our presentation on the simple case above.

6.2.4 The mLTD model

The multinomial logistic transition distribution (mLTD) model is given by:

$$p(x_{it}|x_{1(t-1)},\ldots,x_{d(t-1)}) = \frac{\exp\left(\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{d} \mathbf{Z}^j_{x_{it},x_{j(t-1)}}\right)}{\sum_{x' \in \mathcal{X}_i} \exp\left(\mathbf{z}^0_{x'} + \sum_{j=1}^{d} \mathbf{Z}^j_{x',x_{j(t-1)}}\right)} \tag{6.6}$$

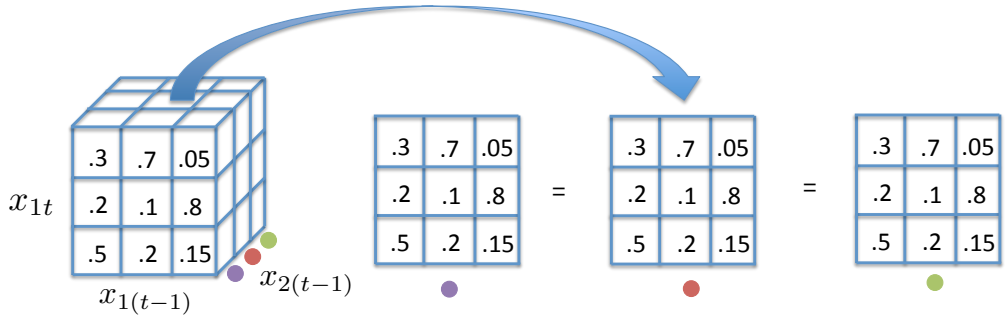


Figure 6.1: Illustration of Granger non-causality in an example with $d = 2$ and $m_1 = m_2 = 3$. Since the tensor represents conditional probabilities, the columns of the front face of the tensor, the vertical x_{1t} axis, must sum to one. Here, x_2 is not Granger causal for x_1 since each slice of the conditional probability tensor along the x_2 mode is equal.

where $\mathbf{Z}^j \in \mathbb{R}^{m_i \times m_j}$ and $\mathbf{z}^0 \in \mathbb{R}^{m_i}$. While not used before to model multivariate categorical time series with $m > 2$ categories, its close cousin, the probit model, has been utilized for this purpose [141]. The model in [141] is not a natural fit for inferring Granger causality networks both due to the non-convexity of the probit model and the non-convex constraints imposed on the $\mathbf{Z}^j$ matrices. Note that, like the MTD model, the mLTD model naturally allows adding interaction terms, though we focus again our presentation on the simple case above.

6.2.5 Comparing MTD and mLTD models

Both MTD and mLTD models represent the full conditional probability tensor using individual matrices for each x_j series, $\mathbf{P}^j$ for MTD and $\mathbf{Z}^j$ for mLTD. However, how these matrices are composed and restrictions on their domains differ substantially between the two models. The MTD model is a convex combination of pairwise probability tables whereas mLTD is a nonlinear function of the unresricted $\mathbf{Z}^j$s. MTD may thus be thought of as a linear tensor factorization method for conditional probability tensors, where the tensor is created by summing probability table slices along each dimension. This interpretation of MTD is displayed graphically in Figure 6.2.

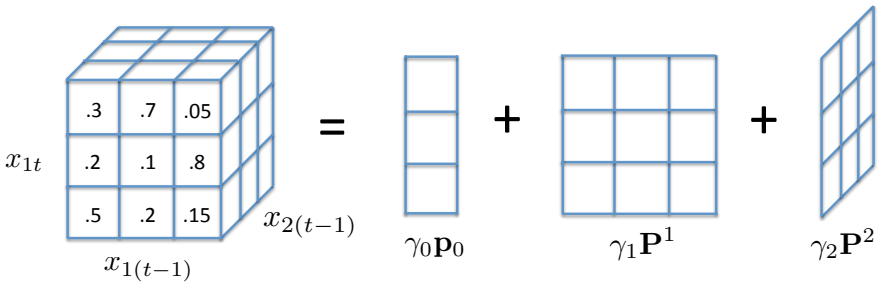


Figure 6.2: Schematic of the MTD factorization of the conditional probability tensor $p(x_{t1}|x_{(t-1)1}, x_{(t-1)2})$ for $d = 2$ time series and $m = 3$ categories.

6.3 Convexity, Identifiability and Granger Causality

In this section, we first introduce a novel reparamaterization of the MTD model that renders the log-likelihood of the MTD model *convex*. The convex formulation alone opens up an array of possibilities for the MTD framework beyond our multivariate categorical time series focus, eliminating the primary barrier to adoption of this method, i.e. non-convexity and associated computationally demanding inference procedures. The proposed change-of-variables also allows us to derive both novel identifiability conditions for the MTD model and Granger causality restrictions that hold for both MTD and mLTD models. The non-identifiability of the MTD model was first pointed out by [117], but no explicit conditions or general framework for identifiability were given. We show that while the identifiability conditions for MTD are non-convex, they may be enforced implicitly by adding an appropriate convex penalty to the convex log-likelihood objective. The proofs of all results are given in Section 6.9.

6.3.1 Convex MTD

Maximum likelihood for the MTD model under the $(\gamma, \mathbf{P})$ parameterization is given by the non-convex optimization problem:

$$\begin{aligned} &\underset{\mathbf{P},\gamma}{\text{minimize}} - \sum_{t=1}^{T} \log \left(\gamma_0 \mathbf{p}^0_{x_{it}} + \sum_{j=1}^{d} \gamma_j \mathbf{P}^j_{x_{it}\ x_{j(t-1)}} \right) \\ &\text{subject to} \quad \mathbf{1}^T \mathbf{P}^j = \mathbf{1}^T, \ \ \mathbf{P}^j \geq 0, \ \forall j \qquad \mathbf{1}^T \gamma = 1 \ , \gamma \geq 0. \end{aligned}$$

The log-likelihood surface is highly non-convex, following from the multiplication of the γ_j and $\mathbf{P}^j$ terms in the log term. It also contains many local optima due to the general non-identifiability. Indeed, the set of equivalent models forms a non-convex region in the $(\gamma, \mathbf{P})$ parameterization (i.e., the convex combination of equivalent models is not necessarily another equivalent model), leading to many non-convex shaped ridges and sets of equal probability.

Fortunately, optimization may be recast into a convex program using the re-parameterization $\mathbf{Z}^j = \gamma_j \mathbf{P}^j$ and $\mathbf{z}^0 = \gamma_0 \mathbf{p}^0$. Using this reparameterization we can rewrite the factorization of the conditional probability tensor for MTD in Equation (6.5) as

$$p(x_{it} | x_{1(t-1)}, \ldots, x_{p(t-1)}) = \mathbf{z}^0_{x_{it}} + \sum_{j=1}^{d} \mathbf{Z}^j_{x_{it}, x_{j(t-1)}}. \tag{6.7}$$

The full optimization problem for maximum log-likelihood including constraints then becomes:

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}} - \sum_{t=1}^{T} \log \left(\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{d} \mathbf{Z}^j_{x_{it}\ x_{j(t-1)}} \right) \\ &\text{subject to} \quad \mathbf{1}^T \mathbf{Z}^j = \gamma_j \mathbf{1}^T, \ \ \mathbf{Z}^j \geq 0, \ \forall j \qquad \mathbf{1}^T \gamma = 1 \ , \gamma \geq 0. \end{aligned} \tag{6.8}$$

Problem (6.8) is convex since the objective function is a linear function composed with a log function and only involves linear equality and inequality constraints [20].

The $\mathbf{Z}^j$ reparameterization in Equation (6.7) also provides clear intuition for why the MTD model may not be identifiable. Since the probability function is a linear sum of $\mathbf{Z}^j$s, one may move

probability mass around, taking some from some $\mathbf{Z}^j$ and moving to some $\mathbf{Z}_i$, $i \neq j$, while keeping the conditional probability tensor constant. These sets of equivalent MTD parameterizations have the following appealing property:

Proposition 2. *The set of MTD parameters, $\mathbf{Z}$, that yield the same factorized conditional distribution $p(x_{it}|x_{(t-1)})$ forms a convex set.*

Taken together, the convex reparameterization and Proposition 2 imply that the convex function given in Equation (6.8) has no local optima, and that the globaly optimal solution to Problem (6.8) is given by a convex set of equivalent MTD models.

6.3.2 Identifiability

Identifiability for the MTD model

The re-parameterization of the MTD model in terms of $\mathbf{Z}^j$ instead of γ_j and $\mathbf{P}^j$, combined with the introduction of an intercept term, allows us to explicitly characterize identifiability conditions for this model.

Theorem 3. *Every MTD distribution has a unique parameterization where the minimal element in each row of $\mathbf{P}^j$ (and thus $\mathbf{Z}^j$) is zero for all j.*

The intuition for this result is simple — any excess probability mass on a row of each $\mathbf{Z}^j$ may be pushed onto the same row of the intercept term $\mathbf{z}^0$ without changing the full conditional probability. This operation may be done until the smallest element in each row is zero, but no more without violating the positivity constraints of the pairwise transitions. The identifiability condition in Theorem 3 also offers an interpretation of the parameters in the MTD model. Specifically, the element $\mathbf{Z}^j_{mn}$ denotes the additive increase in probability that x_i is in state m given that x_j is in state n. Furthermore, the γ_j parameters now represent the total amount of probability mass in the full conditional distribution explained by categorical variable x_j, providing an interpretable notion of dependence in categorical time series. The mLTD model, however, does not readily suggest a

simple and interpretable notion of dependence from the $\mathbf{Z}^j$ matrix due to the non-linearity of the link function. The identifiability conditions are displayed pictorially in Figure 6.3.

Unfortunately, the necessary constraint set for identifiability specified in Theorem 3 is a non-convex set since the locations of the zero elements in each row of $\mathbf{Z}^j$ are unknown. Naively, one could search over all possible locations for the zero element in each row of each $\mathbf{Z}^j$; however, this quickly becomes infeasible as both m and d grow.

Instead, we add a penalty term $\Omega(\mathbf{Z})$, or prior, that biases the solution towards the uniqueness constraints. This regularization also aids convergence of optimization since the maximum likelihood solution without identifiability constraints is not unique. Letting

$$L_{\text{MTD}}(\mathbf{Z}) = -\sum_{t=1}^{T} \log\left(\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{p} \mathbf{Z}^j_{x_{it}\ x_{j(t-1)}}\right), \tag{6.9}$$

the regularized estimation problem is given by

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}} \;\; L_{\text{MTD}}(\mathbf{Z}) + \lambda\Omega(\mathbf{Z}) \\ &\text{subject to} \quad \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T, \;\; \mathbf{Z}^j \geq 0 \;\forall j, \;\; \mathbf{1}^T\gamma = 1 \;, \gamma \geq 0. \end{aligned} \tag{6.10}$$

Theorem 4. *For any $\lambda > 0$ and $\Omega(\mathbf{Z})$ that does not depend on $\mathbf{z}^0$ and is increasing with respect to the absolute value of entries in $\mathbf{Z}^j$, the solution to the problem in Equation (6.32) is contained in the set of identifiable MTD models described in Theorem 3.*

Intuitively, by penalizing the entries of the $\mathbf{Z}^j$ matrices, but not the intercept term, solutions will be biased to having the intercept contain the excess probability mass, rather than the $\mathbf{Z}^j$ matrices. Thus, even with a very small penalty, we constrain the solution space to the set of identifiable models. Theorem 4 characterizes an entire *class* of regularizers that enforce the identifiability constraints for MTD. As we explain in Section 6.4.1, a convenient choice for $\Omega(\mathbf{Z})$ for our case coincides with a regularizer for selecting for Granger causality.

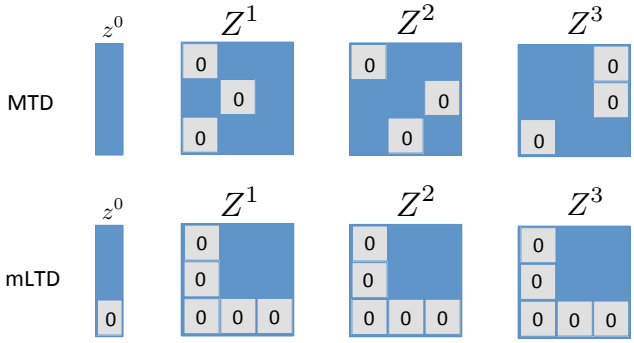


Figure 6.3: Schematic displaying the identifiability conditions for the MTD model (*top*) and the mLTD model (*bottom*) for a $d = 3$ and $m_1 = m_2 = m_3 = 3$ example. Identifiability for MTD requires a zero entry in each row of $\mathbf{Z}^j$, while for mLTD the first column and last row must all be zero. In MTD the columns of each $\mathbf{Z}^j$ must also sum to the same value, and must sum to one across all $\mathbf{Z}^j$.

Identifiability for the mLTD model

The non-identifiability of multinomial logistic models is also well-known, as is the non-identifiability of generalized linear models with categorical covariates. Combining the standard identifiability restrictions for both settings gives [1]:

Proposition 5. *([1]) Every mLTD has a unique parameterization such that first column and last row of* $\mathbf{Z}^j$ *are zero for all* j *and the last element of* $\mathbf{z}^0$ *is zero.*

These conditions are displayed pictorially in Figure 6.3. Under the identifiability constraints for both MTD and mLTD models, at least one element in each row must be zero. For MTD this zero may be in any column, while for mLTD the zero may be placed in the first column of each row without loss of generality. For mLTD the last row of $\mathbf{Z}^j$ must also be zero due to the logistic output (one category serves as the 'baseline'); in MTD, instead, each column of $\mathbf{P}^j$ must sum to one.

6.3.3 Granger Causality in MTD and mLTD

Under the $\mathbf{Z}^j$ MTD parameterization and the mLTD specification of Equation (6.6), we have the following simple result for Granger non-causality conditions:

Proposition 6. *In both the MTD model of Equation (6.7) and the mLTD model of Equation (6.6), time series x_j is Granger non-causal for time series x_i iff the columns of $\mathbf{Z}^j$ are all equal.*

Intuitively, if all columns of $\mathbf{Z}^j$ are equal, the transition distribution for x_{it} does not depend on $x_{j(t-1)}$. This result for mLTD and MTD models is analogous to the general Granger non-causality result for the slices of the conditional probability tensor being constant along the $x_{j(t-1)}$ mode being equal. Based on Proposition 6, we might select for Granger non-causality by penalizing the columns of $\mathbf{Z}^j$ to be the same. While this approach is potentially interesting, a more direct, stable method takes into account the conditions required for identifiability of the $\mathbf{Z}^j$ under both models.

Under the identifiability constraints for both MTD and mLTD given in Theorems 3 and Proposition 5, respectively, then x_j is Granger non-causal for x_i iff $\mathbf{Z}^j = 0$ (a special case of all columns being equal). For both MTD and mLTD models this fact follows from each row having at least one zero element; for all the columns to be equal as stated in Proposistion 6, all elements in each row must also be equal to zero. Taken together, if we enforce the identifiability constraints, we may uniquely select for Granger non-causality by encouraging some $\mathbf{Z}^j$ to be zero.

6.4 Granger Causality Selection

We now turn to procedures for inferring Granger non-causality statements from observed multivariate categorical time series. In Section 6.3, we derived that if $\mathbf{Z}^j = 0$, then x_j is Granger non-causal for x_i in both MTD and mLTD models. To perform model selection, we take a penalized likelihood approach and present a set of penalty terms that encourage $\mathbf{Z}^j = 0$ while maintaining convexity of the overall objective. The final parameter estimates automatically satisfy the identifiability constraints for MTD. We also develop analogous penalized criterion for selecting Granger causality in the mLTD model.

6.4.1 Model selection in MTD

We now explore penalties that encourage the $\mathbf{Z}^j$ matrices to be zero. Under the $\mathbf{P}^j, \gamma_j$ parameterization this is equivalent to encouraging the γ_j to be zero. We first introduce an L_0 penalized problem in terms of the original γ_j parameterization, and then show how convex relaxations of the L_0 norm on γ_j lead to natural convex penalties on $\mathbf{Z}^j$. Ideally, we would solve the penalized L_0 problem:

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}} \;\; L_{\text{MTD}}(\mathbf{Z}) + \lambda||\gamma_{1:d}||_0 \\ &\text{subject to} \quad \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T, \;\; \mathbf{Z}^j \geq 0 \;\forall j, \;\; \mathbf{1}^T\gamma = 1 \;, \gamma \geq 0 \end{aligned} \tag{6.11}$$

where $\lambda \geq 0$ is a regularization parameter and $||\gamma_{1:d}||_0$ is the L_0 norm over the γ weights and the intercept weight γ_0 is not regularized. The L_0 penalty simply counts the number of non-zero γ_j, which is equivalent to the number of non-zero $\mathbf{Z}^j$. This results in a non-convex objective. Instead, we develop alternative convex penalties suited to model selection in MTD. Importantly, we require that any such penalty $\Omega(\mathbf{Z})$ fall in the intersection of two penalty classes: 1) $\Omega(\mathbf{Z})$ must be a convex relaxation to the L_0 norm in Problem (6.11) to promote sparse solutions and 2) $\Omega(\mathbf{Z})$ must satisfy the conditions of Theorem 4 to ensure the final parameter estimates satisfy the MTD identifiability constraints. We propose and compare two penalties that satisfy these criteria.

Our first proposal is the standard L_1 relaxation, as in lasso regression, which simply sums the absolute values of γ_j. This penalty encourages *soft-thresholding*, where some estimated γ_j are set exactly to zero while others are shrunk relative to the estimates from the unpenalized objective. Note that due to the greater than zero constraint, the L_1 norm on $\gamma_{1:d}$ is simply given by the sum $\sum_{j=1}^{d} \gamma_j$. If γ_0 were included in the L_0 regularization, the L_1 relaxation would fail due to the γ simplex constraints $1^T\gamma = 1, \gamma \geq 0$ so the L_1 norm would always be equal to one over the feasible set [149]. Our addition of an unpenalized intercept to the MTD model allows us to sidestep this issue and leverage the sparsity promoting properties of the L_1 penalty for model selection in MTD.

The L_1 regularized MTD problem is thus given by

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}} \;\; L_{\text{MTD}}(\mathbf{Z}) + \lambda \sum_{j=1}^{d} \gamma_j \\ &\text{subject to} \;\; \mathbf{1}^T \mathbf{Z}^j = \gamma_j \mathbf{1}^T, \;\; \mathbf{Z}^j \geq 0 \; \forall j, \;\; \mathbf{1}^T \gamma = 1 \; , \gamma \geq 0, \end{aligned} \tag{6.12}$$

Equation (6.12) may be rewritten solely in terms of the $\mathbf{Z}^j$ terms by noting that $\gamma_j = \frac{1}{m_j} 1^T \mathbf{Z}^j 1$. Defining $\tilde{z}^T = (\text{vec}(\mathbf{Z}^1)^T, \ldots, \text{vec}(\mathbf{Z}^d)^T)$, and assuming $|\mathcal{X}_i| = m \; \forall i$ for simplicity of presentation, we can rewrite the MTD constraints as

$$(I_d \otimes A)\tilde{z} = 0, \;\; \mathbf{1}^T \tilde{z} = m, \;\; \tilde{z} \geq 0,$$

where

$$A = \begin{pmatrix} \mathbf{1}_m^T & -\mathbf{1}_m^T & 0 & 0 & \ldots \\ 0 & \mathbf{1}_m^T & -\mathbf{1}_m^T & 0 & \ldots \\ \ldots & \ldots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{1}_m^T & -\mathbf{1}_m^T \end{pmatrix} \tag{6.13}$$

I_d is a d-dimensional identity matrix. This gives the final penalized optimization problem only in terms of $\mathbf{Z}^j$ as

$$\begin{aligned} &\underset{\mathbf{Z}}{\text{minimize}} \;\; L_{\text{MTD}}(\mathbf{Z}) + \lambda \sum_{i=1}^{d} \frac{1}{m} \mathbf{1}^T \mathbf{Z}^j \mathbf{1} \\ &\text{subject to} \;\; (I_d \otimes A)\tilde{z} = 0, \;\; \mathbf{1}^T \tilde{z} = m, \;\; \tilde{z} \geq 0 \end{aligned} \tag{6.14}$$

Writing the L_1 penalized problem in this form shows that the L_1 penalty increases with the absolute value of the entries in $\mathbf{Z}^j$ and does not penalize the intercept, thus satisfying the conditions of Theorem 4. As a result, the solution to the problem given in Equation (6.14) automatically satisfies the MTD identifiability constraints. Furthermore, the solution will lead to Granger causality estimates since many of the $\mathbf{Z}^j$ will be zero due to the L_1 penalty.

Another natural convex relaxation of the objective in Equation (6.11) is given by a group lasso penalty on each $\mathbf{Z}^j$. The relaxation is derived by writing the L_0 norm as a rank constraint in terms of $\mathbf{Z}^j$, which then is relaxed to a group lasso. Specifically, assume all time series have the same number of categories, $m_j = m \ \ \forall j$. Due to the equality and greater than zero constraints

$$\begin{aligned} ||\gamma_{1:d}||_0 &= || \left(\mathbf{1}^T \text{vec}(\mathbf{Z}^1), \ldots, \mathbf{1}^T \text{vec}(\mathbf{Z}^d) \right) ||_0 \\ &= \text{rank}(\mathbf{Q}^T \mathbf{Q}) \\ &= \text{rank}(\mathbf{Q}) \end{aligned}$$

where

$$\mathbf{Q} = \begin{pmatrix} \text{vec}(\mathbf{Z}^1) & 0 & \ldots & 0 \\ 0 & \text{vec}(\mathbf{Z}^2) & \ldots & 0 \\ 0 & \ldots & \ddots & \vdots \\ 0 & \ldots & \ldots & \text{vec}(\mathbf{Z}^d) \end{pmatrix}.$$

Thus we can use the nuclear norm on $\mathbf{Q}$ as a convex relaxation to $||\gamma_{1:d}||_0$. Furthermore, the nuclear norm of $\mathbf{Q}$ is given by the sum of $\mathbf{Z}^j$ Froebenius norms,

$$||\mathbf{Q}||_* = \sum_{j=1}^{d} ||\mathbf{Z}^j||_F,$$

where $||.||_*$ is the nuclear norm and $||.||_F$ is the Froebenius norm. This group penalty gives the final problem

$$\begin{aligned} &\underset{\mathbf{Z}}{\text{minimize}} \ \ L_{\text{MTD}}(\mathbf{Z}) + \lambda \sum_{j=1}^{d} ||\mathbf{Z}^j||_F \\ &\text{subject to} \quad (I_d \otimes A)\tilde{z} = 0, \ \ \mathbf{1}^T \tilde{z} = m, \ \ \tilde{z} \geq 0. \end{aligned} \tag{6.15}$$

Here, we penalize $\mathbf{Z}^j$ directly, rather than indirectly via γ_j. The group lasso penalty drives all elements of $\mathbf{Z}^j$ to zero together, such that the optimal solution automatically selects some $\mathbf{Z}^j$ to

be all zero and others not. This effect naturally coincides with our conditions of Granger non-causality that *all* elements of $\mathbf{Z}^j = 0$. The group lasso penalty also satisfies the conditions of Theorem 4 since the L_2 norm is increasing with respect to each element in $\mathbf{Z}^j$ and the intercept is not penalized. Thus, solutions to Problem (6.15) automatically enforce the MTD identifiability constraints.

6.4.2 Model selection in mLTD

To select for Granger causality in the mLTD model, we add a group lasso penalty to each of the $\mathbf{Z}^j$ matrices, analogously to Equation (6.15), leading to the following optimization problem:

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \sum_{t=1}^{T} \mathbf{z}^0_{x_{it}} + \sum_{j=1}^{d} \mathbf{Z}^j_{x_{it},x_{j(t-1)}} + \log\left(\sum_{x' \in \mathcal{X}_i} \exp\left(\mathbf{z}^0_{x'} + \sum_{j=1}^{d} \mathbf{Z}^j_{x',x_{j(t-1)}}\right)\right) + \lambda \sum_{j=1}^{d} ||\mathbf{Z}^j||_F \\ \text{subject to} \quad & \mathbf{Z}^j_{1:m_i,1} = 0, \mathbf{Z}^j_{m_i,1:m_j} = 0 \ \ \forall j. \end{aligned} \tag{6.16}$$

For two categories, $m_i = 2 \ \ \forall i$, this problem reduces to sparse logistic regression for binary time series, which was recently studied theoretically [74]. As in the MTD case, the group lasso penalty shrinks some $\mathbf{Z}^j$ entirely to zero.

6.5 Optimization

For both penalized MTD and mLTD models we use proximal gradient based methods for optimization. For the mLTD model we perform gradient steps with respect to the mLTD likelihood followed by a proximal step with respect to the group lasso penalty. This leads to a gradient step of the smooth likelihood followed by separate soft group thresholding [144] on each $\mathbf{Z}^j$.

For the MTD model, our proximal algorithm reduces to a projected gradient algorithm [144]. Projected gradient algorithms take steps along the gradient of the objective, and then project the result onto the feasible region defined by the constraints. In comparison to other MTD optimization methods, our projected gradient algorithm under the $\mathbf{Z}^j$ parameterization is guaranteed to reach the global optima of the MTD log-likelihood. The gradient of the regularized MTD model with respect

to entries in $\mathbf{Z}^j$ over the feasible set is given by

$$\frac{dL}{d\mathbf{Z}^j_{x',x''}} = \sum_{t=1}^{T} 1_{\{x_{it}=x',x_{j(t-1)}=x''\}} \frac{1}{\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{p} \mathbf{Z}^j_{x_{it},x_{j(t-1)}}} + \lambda \frac{d\Omega}{d\mathbf{Z}^j_{x',x''}}. \tag{6.17}$$

For the L_1 norm, $\Omega(\mathbf{Z})$ is not differentiable when an element in any $\mathbf{Z}^j$ is zero. For the L_2 group norm, $\Omega(\mathbf{Z})$ is not differentiable when *every* element in at least one $\mathbf{Z}^j$ is zero. However, the MTD constraints enforce that $\mathbf{Z}^j \geq 0$. Since the point of non-differentiability for the L_2 norm in our case occurs when elements are identically zero, we modify the constraints so that $\mathbf{Z}^j \geq \epsilon$ for some small ϵ when using the group penalty. This allows us to ignore non-differentiability issues, and instead take gradient steps directly along the penalized MTD objective.

Following the notation from the end of Section 6.4.1, let the set $C = \{\tilde{z} | \tilde{z} \geq \epsilon, (I_d \otimes A)\tilde{z} = 0, 1^T\tilde{z} = m\}$ denote the modified MTD constraints with respect to the $\mathbf{Z}^j$ parameterization. We perform projected gradient descent by taking a step along the regularized MTD gradient of Equation (6.17) and then projecting the result onto C. Specifically, the algorithm iterates the following recursion starting at iteration $k = 0$

$$\tilde{z}^{k+1} = \mathcal{P}_C\left(\tilde{z}^k - \delta_k \frac{dL}{d\tilde{z}}\right), \tag{6.18}$$

where δ_k is a step size chosen by line search [144]. We have written the projected gradient steps in terms of the vectorized variables $\tilde{z}$, rather than the $\mathbf{Z}^j$ matrices, for ease of presentation. The $\mathcal{P}_C(x)$ operation is the projection of a vector x onto the modified MTD constraint set C:

$$\begin{aligned} &\underset{z}{\text{minimize}} \quad ||z - x||_2^2 \\ &\text{subject to} \quad z \geq \epsilon, \quad (I_d \otimes A)z = 0, \quad \mathbf{1}^T z = m. \end{aligned}$$

where $\epsilon = 0$ for the L_1 penalty and $\epsilon > 0$ but small for the group lasso penalty. This is a quadratic program and we use the dual method [68] as implemented in the R quadratic programming package *quadprog* [193]. However, we have found that this standard R solver scales poorly as the number of

time series d gets large. Instead, we have developed a fast projection algorithm based on Dykstra's splitting algorithm [21] that harnesses the particular structure of the constraint set for much faster projection, as described in Section 6.5.1. The full projected gradient algorithm for MTD is given in Algorithm 3.

6.5.1 Dykstra's Splitting Algorithm for Projection onto the MTD Constraints

The set C may be written as the intersection of two simpler sets: $C = S \cap B$, where S is the simplex constraint set of the first column of each $\mathbf{Z}^j$ matrix and the greater than zero constraint for all entries of $\mathbf{Z}^j$. Specifically,

$$S = \left\{ \{\mathbf{Z}^j \in \mathbb{R}^{m \times m}\}_{j=0}^{d} \middle| \sum_{j=0}^{d} \sum_{i=1}^{m} \mathbf{Z}_{1i}^j = 1, \mathbf{Z}^j \geq 0 \forall j \right\}. \tag{6.19}$$

On the other hand, $B = \cup_{j=1}^{d} B_j$, where B_j is the constraint set that all columns in $\mathbf{Z}^j$ sum to the same value:

$$B_j = \left\{ \mathbf{Z}^j \in \mathbb{R}^{m \times m} \middle| A \text{ vec}(\mathbf{Z}^j) = \mathbf{0} \right\}, \tag{6.20}$$

where the matrix A is given in Equation (6.13). Dykstra's algorithm alternates between projecting onto the simplex constraints S and the equal column sums B by iterating the following steps. Let $w^0 = x, u^0 = v^0 = 0$ and repeatedly update starting with iteration number $l = 0$:

$$y^l = \mathcal{P}_S(w^l + u^l)$$

$$u^{l+1} = w^l + u^l - y^l$$

$$w^l = \mathcal{P}_B(y^l + v^l)$$

$$v^{l+1} = y^l + v^l - w^l$$

where $\mathcal{P}_S$ is the projection onto the set S and $\mathcal{P}_B$ is the linear projection onto the set B. The $\mathcal{P}_S$ projection may be split into a simplex projection for the constraint $\sum_{j=0}^{d} \sum_{i=1}^{m} \mathbf{Z}_{1i}^j = 1, \mathbf{Z}_{1i}^j \geq$

$0\ \forall i, j$ and a greater than zero constraint $\mathbf{Z}^j_{ni} \geq 0\ \forall i, j$ and $n > 1$. We perform the simplex projection in $(dm)\log(dm)$ time using the algorithm of [49] and the greater than zero projection is simply thresholding elements at zero and is performed in linear time. The $\mathcal{P}_B$ linear projection is performed separately for each $\mathbf{Z}^j$:

$$\mathcal{P}_{B_j}(x) = \left(I - \left(A\left(AA^T\right)^{-1}A^T\right)\right)x \tag{6.21}$$

where $\left(I - \left(A\left(AA^T\right)^{-1}A^T\right)\right)$ may be precomputed so the per-iteration complexity for the full B projection is dm^4 since A is a $(m-1) \times m^2$ matrix. Importantly, this projection scheme harnesses the structure of the constraint set by splitting the projections into components that admit fast and simple low-dimensional projections. The full projection algorithm is given in Algorithm 4.

We compare projection times of the Dykstra algorithm to the active set method of [68] implemented in the R package *quadprog* [193]. The Dykstra projection for the MTD constraints was implemented in C++. Elements of $\mathbf{Z}^j$ were drawn independently from a normal distribution with standard deviation .7 and then projected onto C. Average run times across 10 random realizations for $d \in (10, 20, 30, 40, 50, 60, 70)$ series and $m = 5$ categories are displayed in Figure 6.4. The Dykstra algorithm was run until iterates changed by less than 10^{-10}. For each run, the elementwise maximum difference between the Dykstra projection the *quadprog* projection was always on the scale of 10^{-10}. Across this range of d the *quadprog* runtime appears to scale quadratically in d, with a total run time on the scale of seconds for $d \geq 20$. The Dykstra projection method, however, appears to scale near linearly in this range with run times on the order of milliseconds. We also performed experiments with differing standard deviations for the independent draws of $\mathbf{Z}^j$ and the results were all very similar.

6.5.2 Comparing model selection and optimization in MTD and mLTD

Approaches to model selection in MTD and mLTD models are conceptually similar; both add regularizing penalties to enforce elements in $\mathbf{Z}^j$ to zero. However, these two approaches differ in practice. We explore the differences in selecting for Granger causality between these two ap-

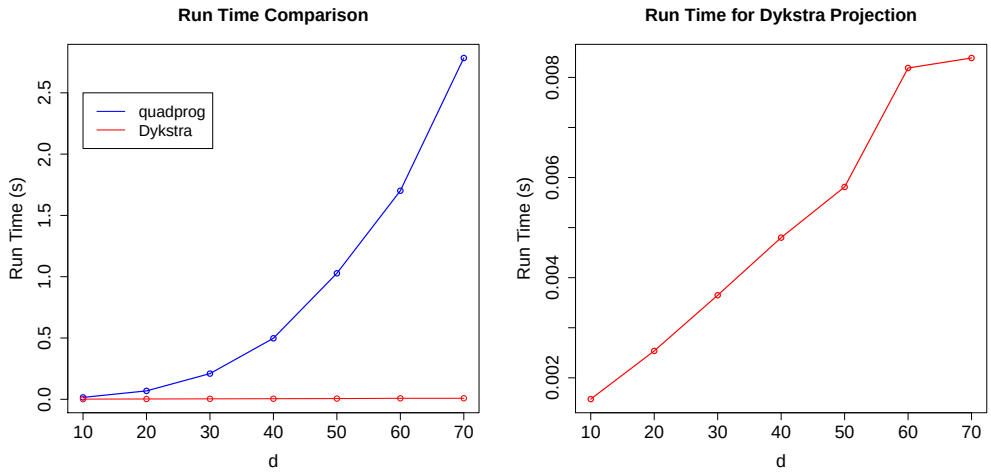


Figure 6.4: (*left*) A runtime comparison of the quadprog projection method and the Dykstra projection method on a range of time series dimensions. (*right*) A zoom in of only the compute time of the Dykstra method.

proaches via extensive simulations in Section 6.6.

Both MTD and mLTD models take gradient steps followed by a proximal operation. In the mLTD model this proximal operation is given by soft thresholding on the elements of $\mathbf{Z}^j$. In the MTD optimization the proximal operation reduces to a projection onto the MTD constraint set. Importantly, due to the restricted domain of the MTD parameter set, the normally non-smooth penalty terms become smooth over the constraint set and we thus include them in the gradient step. In mLTD, the soft threshold proximal operation is performed in linear time while in MTD the projection is performed by iteratively using the Dykstra algorithm, where each step of the Dykstra algorithm is performed in log-linear time.

Algorithm 3 Projected gradient algorithm for MTD using Dykstra projections.

Data: X
Result: $\hat{\mathbf{Z}}$
Initialize $\mathbf{Z}^0 \quad \forall j \quad k = 0$ **while** $\mathbf{Z}^k$ *not converged* **do**
  compute $\nabla L(\mathbf{Z}^k)$ via Equation (6.17) determine γ^k by line search [144] $\mathbf{Z}^{k+1} = DykstraMTD\left(\mathbf{Z}^k + \gamma^k \nabla L(\mathbf{Z}^k)\right) \quad k = k + 1$
end

Algorithm 4 *DykstraMTD*: Zykstra algorithm for projection onto the MTD constraints.

Data: $\mathbf{Z}$

Result: $P_C(\mathbf{Z})$

$z = \left((\mathbf{z}^0)^T, vec(\mathbf{Z}^1)^T, \ldots, vec(\mathbf{Z}^p)^T\right)^T$ Let S be the ordered indices of z whose elements belong in the first column of some $\mathbf{Z}^j,\ j > 0$ or in $\mathbf{z}^0$ Let (j) refer to ordered indices of z whose elements belong to $\mathbf{Z}^j\ \forall j$. $w_0 = z\ \ u_0 = v_0 = 0\ \ l = 0$ **while** w^l *not converged* **do**

  $y^l_S = SimplexProjection(w^l_S + p^l_S)$ via [49] $y^l_{\setminus S} = PositiveThreshold\left(w^l_{\setminus S} + u^l_{\setminus S}\right)$

  $u^{l+1} = w_l + u_l - y_l\ \ w^k_{(0)} = y^l_{(0)} + v^l_{(0)}$ **for** *j = 1:p* **do**

    $w^l_{(j)} = P_{B_j}\left(y^l_{(j)} + v^l_{(j)}\right)$ via Equation (6.21)

  end

  $v^{(l+1)} = y^l + q^l - w^l\ \ l = l + 1$

end

6.6 Experiments

6.6.1 Simulation Set Up

We perform a set of simulation experiments to compare the MTD and mLTD model selection methods. Specifically, we compare the MTD group lasso, L_1-MTD, and mLTD group lasso methods on simulated categorical time series generated first from a sparse MTD model. We find that the group lasso MTD outperforms the MTD L_1 and thus only compare MTD group lasso and mLTD group lasso on two further simulated scenarios: a sparse mLTD model and a sparse latent vector autoregressive model (VAR) with quantized outputs. For all experiments we consider time series of length $T \in (200, 400)$, dimension $d \in (15, 25)$, and number of categories $m \in (2, 3, 4, 5, 6)$. We first explain the details of each simulation condition and then discuss the results.

Sparse MTD For the MTD model, we randomly generate parameters by $\gamma_{ij} \sim \frac{z_{ij}\phi_{ij}}{\sum_{l=1}^{p} z_{il}\phi_{il}}$ where $\phi_i \sim$ Dirichlet(α) and $z_{ij} \sim$ Binomial(δ). We let $\delta = .15, \alpha = 5$. Columns of $\mathbf{Z}^{ij}$ are generated according to $\mathbf{Z}^{ij}_{:l} \sim$ Dirichlet(γ) with $\gamma = .7$. (Note that here we have added a superscript i to $\mathbf{Z}$ to specifically indicate the j to i interaction, whereas previously we dropped the i index for notational simplicity by assuming we were just looking at the series i term.) To ensure that the columns are not close to identical in $\mathbf{Z}^{ij}$ (which would imply Granger non-causality), $\mathbf{Z}^{ij}$ is sampled until the

average total variation norm between the columns is greater than some tolerance ρ. This ensures that non-causality occurs only due to which $\mathbf{Z}^j$ are zero, and not due to equal columns in the simulation. For our simulations, we set $\rho = .3$. A lower value of ρ makes it more difficult to learn the Granger causality graph since some true interactions might be extremely weak.

Sparse mLTD For the mTLD model, the nonzero $\mathbf{Z}^{ij}$ parameters are generated by $\mathbf{Z}^{ij}_{lk} \sim z_{ij} N(0, \sigma_Z^2)$ where $z_{ij} \sim$ Binomial(δ) with $\delta = .15$.

Sparse Latent VAR To examine data generated from neither of the models considered, we simulate data from a continuous time series $y_t \in \mathbb{R}^p$ according to a sparse VAR(1):

$$y_t = A y_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2 I_p)$. The sparse matrix A is generated by first sampling entries $B_{ij} \sim N(0, \sigma_A^2)$ and then setting $A_{ij} = B_{ij} z_{ij}$, where $z_{ij} \sim$ Binomial(δ) with $\delta = .15$. We then quantize each dimension, y_{ti}, into m categories to create a categorical time series x_{ti}. For example, when $m = 3$, $x_{ti} = 1$ if y_{ti} is in the $(0, .33)$ quantile of $\{y_{1i}, \ldots y_{Ti}\}$, and so forth.

6.6.2 Simulation Results

For all methods - MTD L_1, MTD group lasso, and mLTD group lasso - we compute the area under the ROC curve between the true Granger causality graph and the sparse graph that results when varying λ across a range of values.

The results are displayed as histograms across all simulation runs in Figures 6.5, 6.6, and 6.7 for the categorical time series generated by MTD, mLTD, and latent VAR, respectively. We note that the mLTD group lasso model performs best when the data are generated from a mLTD, and likewise the MTD group lasso performs best when the data are generated from a MTD. Furthermore, the MTD L_1 estimator tends to outperform the MTD group lasso across most settings. Interestingly, for data generated from mLTD we see improved performance as a function of the number of categories

m for all n and d settings, while for MTD performance starts high, dips and goes back up with increasing m. This is probably due to the simulation conditions, as in both MTD and mLTD models Granger causality can be quantified as the difference between the columns of $\mathbf{Z}^j$. When there are more categories, there is higher probability under our simulation conditions that there will be some columns with large deviation from other columns in $\mathbf{Z}^j$. This leads to improved Granger causality detection when it exists.

In the latent VAR simulation, MTD group and mLTD group perform similarly in the $T = 200$ simulation condition, but mLTD consistently outperforms MTD in the $T = 400$ case. Taken together, though, both methods perform comparably. There is also evidence of improved performance for both MTD and mLTD methods as the quantization of the latent VAR processes becomes finer, and the number of categories increases. For the MTD model the average AUC increases roughly monotonically with quantization level, though for the mLTD average performance appears to peak at $m = 4$ categories and then levels off or slightly declines. When the quantization is too coarse, say for $m = 2$ or $m = 3$, some Granger causality interactions may become hard to detect since there is much less information about the underlying VAR processes contained in the quantized series with fewer categories.

As expected, across all simulation conditions and estimation methods increasing the sample size T leads to improved performance while increasing the dimension d worsens performance.

6.7 Music Data Analysis

We analyze Granger causality connections in the 'Bach Choral Harmony' data set available at the UCI machine learning repository [120] (`https://archive.ics.uci.edu/ml/datasets/Bach+Chorales`). This data set has been used previously in [156, 56]. The data set consists of 60 chorales for a total of 5665 time steps. At each time step 15 unique discrete events are recorded. There are 12 harmony notes, {C, C#, D, F#, D#, E, F, G, G#, A , A#, B}, that take values either 'on' (played) or 'off' (not played), i.e. $x_{tj} \in \{0, 1\}$ for $j \in \{1, \ldots, 12\}$. There is a 'meter' category taking values in $\{1, \ldots, 5\}$, where lower numbers indicate less accented events and higher numbers higher accented events. There is also the 'pitch class of the base note', taking 12 different

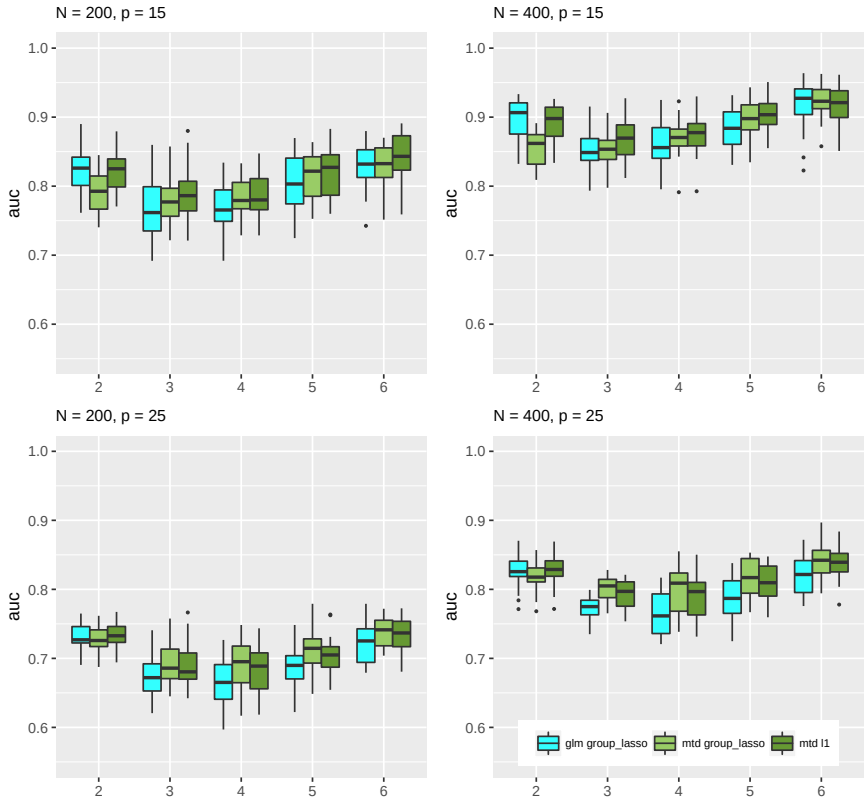


Figure 6.5: AUC for data generated by a sparse MTD process. Boxplots over 20 simulation runs.

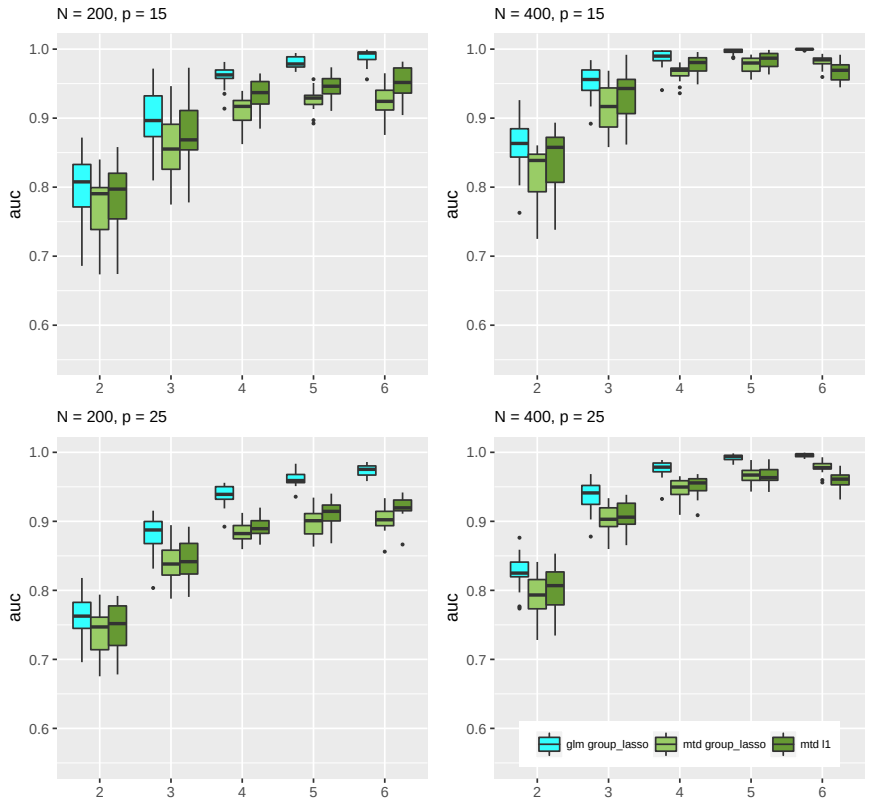


Figure 6.6: AUC for data generated by a sparse latent mLTD process. Boxplots over 20 simulation runs.

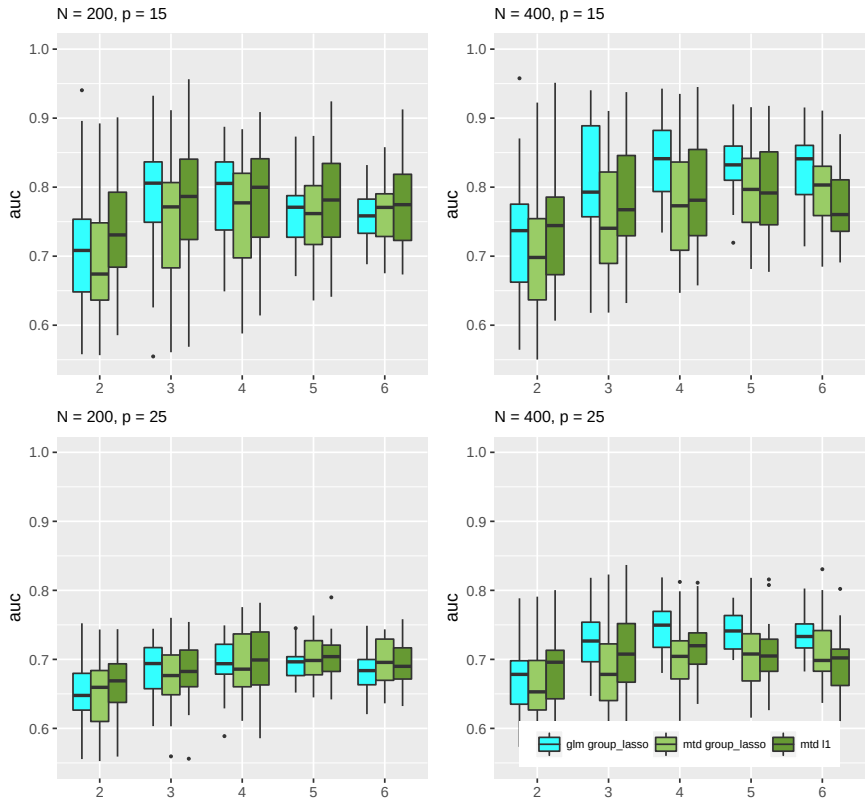


Figure 6.7: AUC for data generated by a sparse latent VAR process. Boxplots over 20 simulation runs.

values and a 'chord' category. We group all chords that occur less than 200 times into one group, giving a total of 12 chord categories.

We apply the sparse MTD model for Granger causality selection and choose the tuning parameter λ by a five-fold cross validation over a grid of λ values. We threshold the γ weights at .01 and plot the estimated resulting Granger causality graph in Figure 6.8. For further interpretability we bold all edges with γ weight magnitudes greater than .06. As mentioned in Section 6.3.2, the MTD model is much more appropriate than the mLTD model for this type of exploratory Granger causality analysis: The γ weights intuitively describe the amount of probability mass that is accounted for in the conditional probability table, giving an intuitive notion of dependence between categorical variables. In the mLTD model, however, it is not clear how to define strength of interaction and dependence given a set of estimated $\mathbf{Z}^j$ parameters due to the non-linearity of the softmax function.

The harmony notes in the graph are displayed in a circle corresponding to the circle of fifths. The circle of fifths is a sequence of pitches where the next pitch in the circle is found seven semitones higher or lower, and it is a common way of displaying and understanding relationships between pitches in western classical music. Plotting the graph in this way shows substantially higher connections with respect to sequences on this circle. For example, moving both clockwise and counter-clockwise around the circle of fifths we see strong connections between adjacent pitches, and in some cases strong connections between pitches that are two hops away on the circle of fifths. Strong connections to pitches far away on the circle of fifths are much rarer. Together, this indicates that in these chorales there is strong dependence in time between pitches moving in both the clockwise and counter-clockwise direction on the circle of fifths.

We also note that the 'chord' category has very strong outgoing connections implying it has strong Granger causality selection with all harmony pitches. This result is intuitive, as it implies that there is strong dependence between what chord is played at time step t and what harmony notes are played at time step $t + 1$. The bass pitch is also influenced by 'chord' and tends to both influence and be influenced by most harmony pitches. Finally, we note that the 'meter' category has much fewer and weaker incoming and outgoing connections, capturing the intuitive notion that

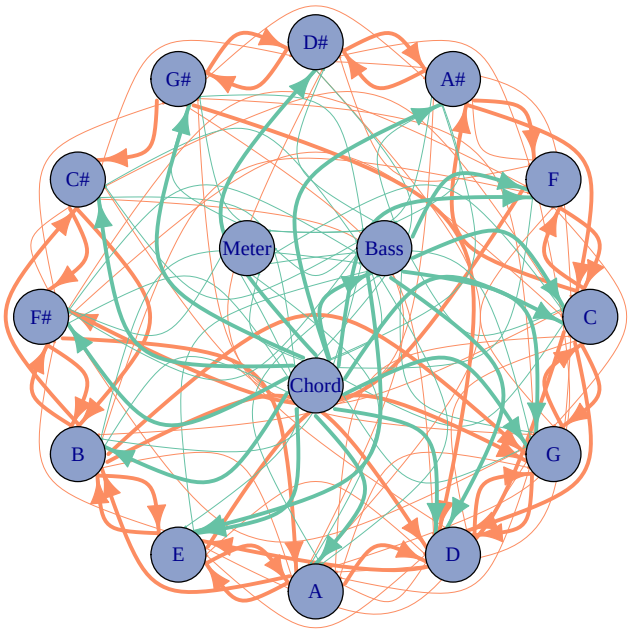


Figure 6.8: The Granger causality graph for the 'Bach Choral Harmony' data set using the penalized MTD method. The harmony notes are displayed around the edge in a circle corresponding to the circle of fifths. Orange links display directed interactions between the harmony notes while green links display interactions to and from the 'bass', 'chord', and 'meter' variables.

the level of accentuation of certain notes does not really relate to what notes are played.

We also performed a connectivity analysis using the penalized mLTD model. However, the mLTD model presents some extra difficulties. Importantly, due to the non-linearity of the softmax function there is not as an intuitive interpretation of 'link strength' between two categorical variables in mLTD as there is in the MTD model. For this reason, it is not clear how to define the strength of interaction and dependence given a set of estimated $\mathbf{Z}^j$ parameters. We chose to use the normalized L_2 norm of each $\mathbf{Z}^j$ matrix, $\frac{||Z_j^i||}{\sqrt{m_i}\sqrt{m_j}}$, as a measure of connection strength in the mLTD model. However, this metric does not have a direct interpretation with respect to the conditional probability tensor. Due to these interpetational difficulties we present the results of the mLTD Bach analysis in the Appendix. We note here that the final graph shows some of the structure of the MTD analysis, strong connections between chord and the harmony notes and some strong connections between notes on the circle of fifths. However, in general, the resulting graph is much less sparse

and interpretable than the MTD graph.

6.8 Discussion

We have proposed a novel convex framework for the MTD model as well as two penalized estimation strategies that simultaneously promotes sparsity in Granger causality estimation and constrain the solution to an identifiable space. We have also introduced the mLTD model as a baseline for multivariate categorical time series that although straightforward, has not been explored in the literature. Novel identifiability conditions for the MTD have been derived and compared to those for the mLTD model. For optimization, we have developed a novel projected gradient algorithm for the MTD model that harnesses the new convex formulation. We also develop a novel Dykstra projection method to quickly project onto the MTD constraint set, allowing the MTD model to scale to much higher dimensions. Our experiments demonstrate the utility of both the MTD and mLTD model for inference of Granger causality networks from categorical time series, even under model misspecification.

There are a number of potential directions for future work. Since we have formulated both MTD and mLTD models as convex problems, the general theory for high dimensional estimators based on convex losses [139] may be leveraged to prove consistency of both models. Recently, [75] established consistency of high dimensional autoregressive GLMs with univariate natural parameters for each series. An interesting direction would be to combine these general techniques for dealing with dependent observations with those of [139] to derive rates for both the MTD and mLTD models.

Further theoretical comparison between mLTD and MTD is also important. For example, to what extent may a mLTD distribution be represented by an MTD one, and vice versa; or, to what extent are both models consistent for Granger causality estimation under model misspecification. Our simulations results suggest that both methods perform well under model misspecification but more general theoretical results are certainly needed.

It would also be interesting to explore other regularized MTD objectives, such as the nuclear norm on $\mathbf{Z}^j$ when the number of categories per time series is large. This penalty would both select

for sparse dependencies while simultaneously sharing information about transitions within each $\mathbf{Z}^j$. Another possibility includes the hierarchical group lasso over lags for higher order Markov chains, as in [140] for VARs, to automatically obtain the order of the Markov chain. Overall, the methods presented herein open up many new opportunities for analyzing multivariate categorical time series both in practice and theoretically.

6.9 Appendix

6.9.1 mLTD Bach Analysis

For the mLTD Bach analysis we performed a 5-fold cross validation to select the λ tuning parameter then thresholded the final connection weights, given by the standardised L_2 norm of $\mathbf{Z}^j$, at .01, as in the MTD case. First, we note that the final mLTD model is much less sparse than the MTD case with only 5 total zero weights. We display the final graph in Figure 6.9.1, where, for interpretability, we bold edges with total weight greater than .45. In this graph there are strong connections in the counter clockwise direction between G#, C#, F#, and B. However, the other connections on the circle of fifths are relatively weaker, and there are many more connections between notes far away on the circle of fifths. The mLTD graph also shows that the chord note both affects and is affected by many harmony notes. Furthermore, we see that the bass category is effected by most harmony notes as well. Overall, however, this graph is much less interpretable than the mTD graph and fails to find the full circle of fifths structure.

6.9.2 Proofs

Proof of Proposition 6 If the columns of $\mathbf{Z}^j$ are all equal then for all fixed values of $x_{\backslash j(t-1)}$ the conditional distribution is the same for all values of $x_{j(t-1)}$. If one column is different then the conditional distribution for all values of $x_{\backslash j(t-1)}$ will depend on $x_{j(t-1)}$.

Proof of Theorem 3 Let $\mathbf{Z}$ be the parameter set for an MTD model. For each $\mathbf{Z}^j$ let the vector α^j be the minimal element in each row, $\alpha^j_k = \min \mathbf{Z^j}_{k:}$. Let $\tilde{\mathbf{Z}}^j = \mathbf{Z}^j - \alpha_j$ and $\tilde{\mathbf{z}}^0 = \mathbf{z}^0 + \sum_{j=1}^d \alpha_j$.

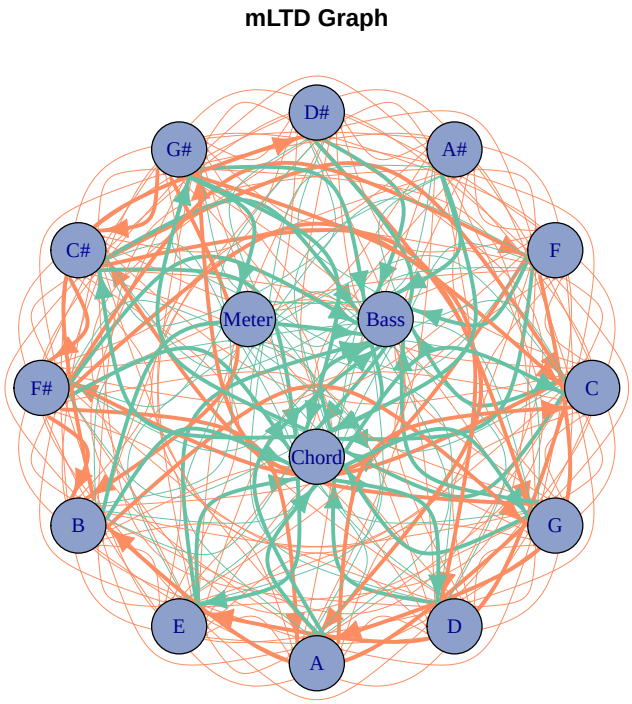


Figure 6.9: The Granger causality graph for the 'Bach Choral Harmony' data set using the mLTD method. The harmony notes are displayed around the edge in a circle corresponding to the circle of fifths. Orange links display directed interactions between the harmony notes while green links display interactions to and from the 'bass', 'chord', and 'meter' variables.

This $\tilde{\mathbf{Z}}$ gives the same MTD distribution as $\mathbf{Z}$. Furthermore, this $\tilde{\mathbf{Z}}$ has a zero element in each row of each $\tilde{\mathbf{Z}}^j$ by construction.

Suppose two parameter sets $\mathbf{X}$ and $\mathbf{Y}$ provide the same MTD distribution. Let $\tilde{\mathbf{X}}$ be as above for $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ of $\mathbf{Y}$.

We use a proof by contradiction. Suppose that $\tilde{\mathbf{Y}} \neq \tilde{\mathbf{X}}$. There must exist some j and some row k such that $\tilde{\mathbf{X}}^j_{k:} \neq \tilde{\mathbf{Y}}^j_{k:}$. Let l_X be the index of the zero element for $\mathbf{X}^j$, i.e. such that $\tilde{\mathbf{X}}^j_{kl} = 0$, and likewise for l_Y. If there are more than one zero element, pick any. Furthermore, if $\tilde{\mathbf{X}}^j_{k:}$ and $\tilde{\mathbf{Y}}^j_{k:}$ share a zero in the same location (if there one or more zero elements in each), then let l_X and l_Y be that index so that $l_X = l_Y$.

If $l_X = l_Y$, let l' be an index such that $\tilde{\mathbf{X}}^j_{kl'} \neq \tilde{\mathbf{Y}}^j_{kl'}$. This index must exist by construction. Let the categories of other series (not for series j), $x_{\setminus j(t-1)}$, be fixed arbitrarily. The difference between the conditional distributions for $\mathbf{X}$ are

$$\begin{aligned}
\tilde{\mathbf{X}}^j_{kl'} &= \tilde{\mathbf{X}}^j_{kl'} - \tilde{\mathbf{X}}^j_{kl_X} && (6.22)\\
&= (\tilde{\mathbf{X}}^j_{kl'} + \alpha_{jk}) - (\tilde{\mathbf{X}}^j_{kl_X} + \alpha_{jk}) && (6.23)\\
&= \mathbf{X}^j_{kl'} - \mathbf{X}^j_{kl_X} && (6.24)\\
&= (\mathbf{x}^0_k + \sum_{i \in \setminus j} \mathbf{X}^i_{kx_{i(t-1)}} + \mathbf{X}^j_{kl'}) - (\mathbf{x}^0_k + \sum_{i \in \setminus j} \mathbf{X}^i_{kx_{i(t-1)}} + \mathbf{X}^j_{kl_X}) && (6.25)\\
&= p_X(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l') - p_X(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X), && (6.26)
\end{aligned}$$

and a similar calculation for $\mathbf{Y}$ shows that

$$\tilde{\mathbf{Y}}^j_{kl'} = p_Y(x_t = k | x_{\setminus j(t-1)}, x_{(t-1)j} = l') - p_Y(x_t = k | x_{\setminus j(t-1)}, x_{(t-1)j} = l_Y). \quad (6.27)$$

However, $\tilde{\mathbf{Y}}^j_{kl'} \neq \tilde{\mathbf{X}}^j_{kl'}$, thus showing that

$$p_Y(x_t = k|x_{\setminus j(t-1)}, x_{(t-1)j} = l') - p_Y(x_t = k|x_{\setminus j(t-1)}, x_{(t-1)j} = l_Y) \neq \\ p_X(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l') - p_X(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l_X). \tag{6.28}$$

This inequality contradicts our assumption that the MTD distributions parametrized by $\mathbf{X}$ and $\mathbf{Y}$ are the same since $l_X = l_Y$.

If $l_X \neq l_Y$, then

$$p_X(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y) - p_X(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l_X) = \tilde{\mathbf{X}}^j_{kl_Y} \tag{6.29}$$

and

$$p_Y(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y) - p_Y(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l_X) = -\tilde{\mathbf{Y}}^j_{kl_X}. \tag{6.30}$$

However, $-\tilde{\mathbf{Y}}^j_{kl_X} \neq \tilde{\mathbf{X}}^j_{kl_Y}$ since at least one of $\tilde{\mathbf{Y}}^j_{kl_X}$ and $\tilde{\mathbf{X}}^j_{kl_Y}$ are nonzero and both are nonnegative. Again, this shows that

$$p_Y(x_t = k|x_{\setminus j(t-1)}, x_{(t-1)j} = l') - p_Y(x_t = k|x_{\setminus j(t-1)}, x_{(t-1)j} = l_Y) \neq \\ p_X(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l') - p_X(x_t = k|x_{\setminus j(t-1)}, x_{j(t-1)} = l_X), \tag{6.31}$$

which contradicts our assumption that the MTD distributions parametrized by $\mathbf{X}$ and $\mathbf{Y}$ are the same.

The same argument shows that the reduction is unique.

Proof of Proposition 2 For any two MTD factorizations $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ that have the same conditional distribution $p(x_{kt}|x_{t-1})$ for all x_{kt} and $x_{(t-1)}$, then for any $0 < \alpha < 1$, the probability tensor of the

MTD model for the parameter set $\alpha\mathbf{Z} + (1-\alpha)\tilde{\mathbf{Z}}$ is given by

$$\begin{aligned}
&\alpha\mathbf{z}^0_{x_{kt}} + (1-\alpha)\tilde{\mathbf{z}}^0_{x_{kt}} + \sum_{j=1}^{d}\left(\alpha\mathbf{Z}^j_{x_{kt}x_{j(t-1)}} + (1-\alpha)\tilde{\mathbf{Z}}^j{}_{x_{kt}x_{j(t-1)}}\right) \\
&= \alpha\left(\mathbf{z}^0_{x_{kt}} + \sum_{j=1}^{d}\mathbf{Z}^j_{x_{kt}x_{j(t-1)}}\right) + (1-\alpha)\left(\tilde{\mathbf{z}}^0_{x_{kt}} + \sum_{i=1}^{d}\tilde{\mathbf{Z}}^j{}_{x_{kt}x_{j(t-1)}}\right) \\
&= \alpha p(x_{kt}|x_{(t-1)}) + (1-\alpha)p(x_{kt}|x_{(t-1)}) \\
&= p(x_{kt}|x_{(t-1)}).
\end{aligned}$$

which shows that $\alpha\mathbf{Z} + (1-\alpha)\tilde{\mathbf{Z}}$ has the same distribution as both $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, so that the set of parameters with the same distribution is a convex set.

Proof of Theorem 4 First, we note that a solution always exists since the log likelihood $L(\mathbf{Z}) = -\sum_{t=1}^{T}\log\left(\mathbf{z}^0_{x_{jt}} + \sum_{i=1}^{d}\mathbf{Z}^j_{x_{jt}\ x_{i(t-1)}}\right)$ and penalty are both bounded below by zero and the feasible set is closed and bounded. Suppose an optimal solution is $\mathbf{Z}$ such that there exists some i such that one row, call it k, of $\mathbf{Z}^j$ does not have a zero element. Let $\alpha = \min(\mathbf{Z}^j_{k:})$ be the minimum value in row k and let $\tilde{\mathbf{Z}}^j$ be equal to $\mathbf{Z}^j\ \forall i$ except that $\tilde{\mathbf{Z}}^j{}_{k:} = \mathbf{Z}^j_{k:} - \alpha$ and $\tilde{z}^j_k = z^j_k + \alpha$. Due to the nonidentifiability of the MTD model $L(\tilde{\mathbf{Z}}) = L(\mathbf{Z})$, while we have that $\Omega(\tilde{\mathbf{Z}}^j) < \Omega(\mathbf{Z}^j)$, implying for $\lambda > 0$

$$L(\tilde{\mathbf{Z}}) + \lambda\Omega(\tilde{\mathbf{Z}}) < L(\mathbf{Z}) + \lambda\Omega(\mathbf{Z}),$$

showing that $\mathbf{Z}$ cannot be an optima.

6.9.3 Optimization Algorithms

In the main text we presented a projected gradient algorithm for optimization. Here we present some alternative methods for optimization of the penalized MTD objective and discuss in what contexts they might be applicable.

6.9.4 Majorization-Minimization

Here we use the convex formulation of MTD in the main text to derive a majorization-minization (MM) algorithm [86]. The closed form updates are only given when there is no penalty function $\Omega(\mathbf{Z})$, so that this algorithm is not as generally applicable as the projected gradient algorithm. Interestingly, we find that the MM updates of the convex formulation correspond exactly to the MTD EM algorithm of [117] for the non-convex parameterization. This proves that the EM algorithm for MTD converges to a global optima even though the log-likelihood is non-convex.

We derive the MM algorithm for the convex MTD formulation with no penalty term (and no intercept):

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}} \;\; L_{\text{MTD}}(\mathbf{Z}) \\ &\text{subject to} \quad \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T, \;\; \mathbf{Z}^j \geq 0 \;\forall j, \;\; \mathbf{1}^T\gamma = 1 \;, \gamma \geq 0. \end{aligned} \tag{6.32}$$

To derive the MM algorithm we first form the surrogate function

$$Q(\mathbf{Z}, \mathbf{Z}^{(n)}) = \sum_{t=1}^{T}\sum_{j=1}^{d} p_{jt} \log \frac{Z^j_{x_{it}x_{(jt-1)}}}{p_{jt}} + \lambda \sum_j \sum_{ik} \log Z^j_{ik} \tag{6.33}$$

where $p_{jt} = \frac{Z^{j(n)}_{x_{it}x_{j(t-1)}}}{\sum_{l=1}^{d} Z^{l(n)}_{x_{it}x_{l(t-1)}}}$. $Q(\mathbf{Z}, \mathbf{Z}^{(n)})$ satisfies the MM algorithm conditions that $Q(\mathbf{Z}, \mathbf{Z}^{(n)}) \geq L_{\text{MTD}}(\mathbf{Z}) + \lambda\Omega(\mathbf{Z})$ and $Q(\mathbf{Z}, \mathbf{Z}) = L_{\text{MTD}}(\mathbf{Z})$. This implies we may iteratively minimize $Q(\mathbf{Z}, \mathbf{Z}^{(n)})$:

$$\mathbf{Z}^{(n+1)} = \underset{Z,\gamma}{\text{argmin}} \; Q(\mathbf{Z}, \mathbf{Z}^{(n)}). \tag{6.34}$$

The optimization problem for the MM update is given by

$$\underset{\mathbf{Z},\gamma}{\text{minimize}} - \sum_{t=1}^{T}\sum_{j=1}^{d} p_{jt} \log \frac{Z^j_{x_{it}x_{(jt-1)}}}{p_{jt}} \tag{6.35}$$

$$\text{subject to} \quad \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T \;\forall j, \;\; \mathbf{1}^T\gamma = 1. \tag{6.36}$$

where we have removed the greater than zero constraints because these are automatically enforced in the $\log$ terms of the $Q(Z, Z^{(n)})$ objective. We may first rewrite the objective equivalently as

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}} - \sum_{j=1}^{d}\sum_{l=1}^{m}\sum_{k=1}^{m} \tilde{p}_{lk}^{j} \log Z_{lk}^{j} \\ &\text{subject to} \quad \mathbf{1}^T\mathbf{Z}^j = \gamma_j \mathbf{1}^T \ \forall j, \quad \mathbf{1}^T\gamma = 1. \end{aligned} \tag{6.37}$$

where $\tilde{p}_{lk}^{j} = \sum_{t=1} p_{jt} 1_{(x_{it}=l, x_{j(t-1)}=k)}$.

We derive the solution by solving the KKT conditions. The Lagrangian of Problem (6.37) is given by

$$\sum_{j=1}^{d}\sum_{l=1}^{m}\sum_{k=1}^{m} \tilde{p}_{lk}^{j} \log Z_{lk}^{j} + \sum_{j}\sum_{k} \lambda_k^j \left(\left(\sum_{l} Z_{lk}^{j} \right) - \gamma_j \right) + \nu(1^T\gamma - 1) \tag{6.38}$$

where λ_j^k and ν are Lagrange multipliers. The solution must satisfy the KKT conditions [20]

$$Z_{lk}^{j} = \frac{\tilde{p}_{lk}^{j}}{\lambda_k^j} \ \ \forall j, l, k, \tag{6.39}$$

$$\nu = \sum_{k} \lambda_k^j \ \ \forall j, \tag{6.40}$$

$$\mathbf{1}^T\mathbf{Z}^j = \gamma_j \mathbf{1}^T \ \forall j, \quad \mathbf{1}^T\gamma = 1. \tag{6.41}$$

Summing over Equation (6.39) for all rows l gives

$$\gamma_j = \frac{\sum_l \tilde{p}_{lk}^{j}}{\lambda_k^j}. \tag{6.42}$$

Re-arranging and summing over k gives

$$\frac{\sum_{lk} \tilde{p}_{lk}^{j}}{\gamma_j} = \sum_{k} \lambda_k^j = \nu, \tag{6.43}$$

and finally re-aranging once more and summing over j gives

$$\frac{\sum_j \sum_{lk} \tilde{p}_{lk}^j}{\nu} = \sum_j \gamma_j = 1. \tag{6.44}$$

Plugging these results back into those above implies that $\nu = \sum_j \sum_{lk} \tilde{p}_{lk}^j$, $\gamma_j = \frac{\sum_{lk} \tilde{p}_{lk}^j}{\sum_j \sum_{lk} \tilde{p}_{lk}^j}$, $\lambda_k^j = \frac{(\sum_l \tilde{p}_{lk}^j)(\sum_j \sum_{lk} \tilde{p}_{lk}^j)}{\sum_{lk} \tilde{p}_{lk}^j}$. Plugging into Equation (6.39) gives the final update for $\mathbf{Z}^{(n+1)}$ as

$$Z_{lk}^{j(n+1)} = \left(\frac{\tilde{p}_{lk}^j}{\sum_l \tilde{p}_{lk}^j} \right) \left(\frac{\sum_{lk} \tilde{p}_{lk}^j}{\sum_j \sum_{lk} \tilde{p}_{lk}^j} \right). \tag{6.45}$$

This update is identical to the updates for the EM algorithm in the original $(\mathbf{P}, \gamma)$ parameterization [117]. Since the MM algorithm on a convex problem converges to a global optima, it follows that the EM algorithm for the original non-convex MTD parameterization also converges to a global optima.

Chapter 7

NEURAL GRANGER CAUSALITY FOR NONLINEAR TIME SERIES

7.1 Introduction

In many scientific applications of multivariate time series it is important to go beyond prediction and forecastability and instead interpret the structure within the time series itself. Typically, this structure provides information about the contemporaneous and lagged relationships within and between individual series and how these series interact. For example, in neuroscience it is important to determine how brain activation spreads through brain regions [182, 196]; in finance it is important to determine groups of stocks with high covariance to design low risk portfolios [173]; and, in biology to infer gene regulatory networks from time series of gene expression levels [63, 128]. However, for a given statistical model or methodology, there is often a tradeoff between the *interpretability* of these structural relationships and *expressivity* of the model dynamics.

Among the many choices for understanding relationships between series, Granger causality [70, 130] is a commonly used framework for time series structure discovery that quantifies the extent to which the past of one time series aids in predicting the future evolution of another time series. When an entire system of time series is studied, networks of Granger causal interactions may be uncovered [12]. This is in contrast to other types of structure discovery, like coherence [166] or lagged correlation [166], that analyze strictly *bivariate* covariance relationships. That is, Granger causality metrics depend on the activity of the entire *system* of time series under study, making them more appropriate for understanding high-dimensional complex data streams. Methodology for estimating Granger causality may be separated into two camps, model-based and model-free.

Most classical model-based methods assume linear time series dynamics and utilize the popular vector autoregressive (VAR) model [130, 128]. In this case, the time lags of a series have a linear effect on the future of each other series, and the magnitude of the linear coefficients quantifies the

Granger causal effect. Sparsity inducing regularizers, like the Lasso [189] or group lasso [207], help scale linear Granger causality estimation in VAR models to the high-dimensional setting [128, 11]. In classical linear VAR methods, one must explicitly specify the maximum time lag to consider when assessing Granger causality. If the specified lag is too short, Granger causal connections occurring at longer time lags between series will be missed while overfitting may occur if the lag is too large. Lag selection penalties, like the hierarchical lasso [140] and truncating penalties [175], have all been utilized to automatically select the relevant lags while protecting against overfitting. Furthermore, these penalties lead to a sparse network of Granger causal interactions, where only a few Granger causal connections exist for each series — a crucial property for scaling Granger causal estimation to the high-dimensional setting, where the number of time series and number of potentially relevant time lags all scale with the number of observations [26].

Model-based methods may fail in real world cases when the relationships between the past of one series and future of another falls out of the model class [188, 190, 129]. This typically occurs when there are *nonlinear* dependencies between the past of one series and the future. Model-free methods, like transfer entropy [196] or directed information [3], are able to detect these nonlinear dependencies between past and future with minimal assumptions about the predictive relationships. However, these estimators have high variance and thus require lots of data for reliable estimation. These approaches also suffer from a curse of dimensionality [163] when the number of series grows, making them inappropriate in the high-dimensional setting.

Neural networks are capable of representing complicated, nonlinear, and non-additive interactions between inputs and outputs. Indeed, their time series variants, such as autoregressive multilayer perceptrons (MLPs) [158, 104, 17] and recurrent neural networks (RNNs) like long-short term memory networks (LSTM) [72], have shown impressive performance in forecasting multivariate time series given their past [206, 211, 118]. While these methods have shown impressive predictive performance, they are essentially black box methods and provide little interpretability of the multivariate structural relationships in the series. A second drawback is that jointly modeling a large number of series leads to many network parameters. As a result, these methods require much more data to fit reliably and tend to perform poorly in high-dimensional settings.

We present a framework for structure learning in MLPs and RNNs that leads to interpretable nonlinear Granger causality discovery. The proposed framework harnesses the impressive flexibility and representational power of neural networks. It also sidesteps the black-box nature of many network architectures by introducing *component-wise* architectures that disentangle the effects of lagged inputs on individual output series. For interpretability and an ability to handle limited data in the high-dimensional setting, we place sparsity-inducing penalties on particular groupings of the weights that relate the histories of individual series to the output series of interest. We term these *sparse component-wise* models, e.g. cMLP and cLSTM, when applied to the MLP and LSTM, respectively. In particular, we select for Granger causality by adding *group sparsity* penalties [207] on the outgoing weights of the inputs.

As in linear methods, appropriate lag selection is crucial for Granger causality selection in nonlinear approaches — especially in highly parametrized models like neural networks. For the MLP, we introduce two more structured group penalties [140, 85] that automatically detect both nonlinear Granger causality and also the lags of each inferred interaction. Our proposed cLSTM model, on the other hand, sidesteps the lag selection problem entirely since the recurrent architecture efficiently models long range dependencies [72]. When the true network of nonlinear interactions is sparse, both cMLP and cLSTM approaches will select a subset of the time series that Granger-cause the output series, no matter the lag of interaction. To our knowledge, these approaches represent the first set of nonlinear Granger causality methods applicable in high dimensions without requiring precise lag specification.

We first validate our approach and associated penalties via simulations on both linear VAR and nonlinear Lorenz-96 data [98], showing that our nonparametric approach accurately selects the Granger causality graph in both linear and nonlinear settings. Second, we compare our cMLP and cLSTM models with existing Granger causality approaches [123, 116] on the difficult DREAM3 gene regulatory network recovery benchmark datasets [152] and find that our methods outperform a wide set of competitors across all five datasets. Finally, we use our cLSTM method to explore Granger causal interactions between body parts during natural motion with a highly nonlinear and complex dataset of human motion capture [36, 62].

Traditionally, the success stories of neural networks have been on prediction tasks in large datasets. In contrast, here our performance metrics relate to our ability to produce interpretable structures of interaction amongst the observed time series. Furthermore, these successes are achieved in limited data scenarios. Our ability to produce interpretable structures and train neural network models with limited data can be attributed to our use of structured sparsity-inducing penalties and the regularization such penalties provide, respectively. We note that sparsity inducing penalties have been used for architecture selection in neural networks [2, 126]. However, the focus of the architecture selection was on improving predictive performance rather than on returning interpretable structures of interaction amongst observed quantities. More generally, our proposed formulation shows how structured penalties common in regression [103, 85] may be generalized for structured sparsity and regularization in neural networks. This opens up new opportunities to use these tools in other neural network context, especially as applied to structure learning problems. In concurrent work, a similar notion of sparse-input neural networks were developed for high-dimensional regression and classification tasks for independent data [57].

7.2 Linear Granger Causality

Let $\mathbf{x_t} \in \mathbb{R}^{\mathbf{p}}$ be a p-dimensional stationary time series and assume we have observed the process at T time points, $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$. Using a model-based approach, as is our focus, Granger causality in time series analysis is typically studied using the vector autoregressive model (VAR) [130]. In this model, the time series at time t, x_t, is assumed to be a combination of the past K lags of the series

$$\mathbf{x}_t = \sum_{k=1}^{K} A^{(k)} \mathbf{x}_{t-k} + e_t, \tag{7.1}$$

where $A^{(k)}$ is a $p \times p$ matrix that specifies how lag k affects the future evolution of the series and e_t is zero mean noise. In this model, time series j does not Granger-cause time series i if and only if for all $k, A^{(k)}_{ij} = 0$. A Granger causal analysis in a VAR model thus reduces to determining which values in $A^{(k)}$ are zero over all lags. In higher dimensional settings, this may be determined by

solving a group lasso regression problem [127]

$$\min_{A^{(1)},\ldots,A^{(K)}} \sum_{t=K+1}^{T} \|\mathbf{x}_t - \sum_{k=1}^{K} A^{(k)}\mathbf{x}_{t-k}\|_2^2 + \lambda \sum_{ij} \|(A_{ij}^{(1)},\ldots,A_{ij}^{(K)})\|_2, \tag{7.2}$$

where $||.||_2$ denotes the L_2 norm which acts as a group penalty jointly shrinking all values of $(A_{ij}^{(1)},\ldots,A_{ij}^{(K)})$ to zero [207] and $\lambda > 0$ is a hyperparameter that controls the level of group sparsity.

The group penalty in Equation 7.2 may be replaced with a structured hierarchical penalty [92, 85] that automatically selects the lag of each Granger causal interaction [140]. Specifically, the hierarchical lag selection problem is given by

$$\min_{A^{(1)},\ldots,A^{(K)}} \sum_{t=K+1}^{T} \|\mathbf{x}_t - \sum_{k=1}^{K} A^{(k)}\mathbf{x}_{t-k}\|_2^2 + \lambda \sum_{ij} \sum_{k=1}^{K} \|(A_{ij}^{(k)},\ldots,A_{ij}^{(K)})\|_2, \tag{7.3}$$

where $\lambda > 0$ now controls the lag order selected for each interaction. Specifically, at higher values of λ there exists a k for each (i,j) pair such that the entire contiguous set of lags $(A_{ij}^{(k)},\ldots,A_{ij}^{(K)})$ is shrunk to zero. If $k = 1$ for a particular (i,j) pair, then all lags are equal to zero and series i does not Granger-cause series j; thus, this penalty simultaneously selects for Granger non-causality and the lag of each Granger causal pair.

7.3 Models for Neural Granger Causality

7.3.1 Adapting Neural Networks for Granger causality

A *nonlinear* autoregressive model (NAR) allows x_t to evolve according to more general nonlinear dynamics [17]

$$\mathbf{x}_t = g(x_{<t1},\ldots,x_{<tp}) + e_t, \tag{7.4}$$

where $x_{<ti} = \big(...,x_{(t-2)i},x_{(t-1)i}\big)$ denotes the past of series i and we have assumed an additive zero mean noise e_t.

In a forecasting setting, it is common to jointly model the full nonlinear functions g using neural networks. Neural networks have a long history in NAR forecasting, using both traditional architectures [34, 16, 17] and more recent deep learning techniques [118, 206, 187]. These approaches either utilize an MLP where the inputs are $x_{<t} = x_{(t-1):(t-K)}$, for some lag K, or a recurrent network, like an LSTM.

There are two problems with applying the standard neural network NAR model in the context of inferring Granger causality. The first is that these models act as black boxes that are difficult to interpret. Due to sharing of hidden layers, it is difficult to specify sufficient conditions on the weights that simultaneously allows series j to Granger cause series i but not Granger cause series i' for $i \neq i'$. Second, a joint network over all x_{ti} $\forall i$ assumes that each time series depends on the same past lags of the other series. However, in practice, each x_{ti} may depend on different past lags of the other series.

To tackle these challenges, we propose a structured neural network approach to modeling and estimation. First, instead of modeling g jointly across all outputs x_t, as is standard in multivariate forecasting, we instead focus on each output *component*:

$$x_{ti} = g_i(x_{<t1}, \ldots, x_{<tp}) + e_{ti}.$$

Here, g_i is a function that specifies how the past lags are mapped to series i. In this context, Granger non-causality between two series j and i means that the function g_i does not depend on $x_{<tj}$, the past lags of series j. More formally,

Definition 7. *Time series j is* Granger non-causal *for time series i if for all $(x_{<t1}, \ldots, x_{<tp})$ and all $x'_{<tj} \neq x_{<tj}$,*

$$g_i(x_{<t1}, \ldots, x_{<tj}, \ldots, x_{<tp}) = g_i(x_{<t1}, \ldots, x'_{<tj}, \ldots x_{<tp});$$

that is, g_i is invariant to $x_{<tj}$.

In Section 7.3.2 and 7.3.3 we consider these component models in the context of MLPs and

LSTMs. We examine a set of sparsity inducing penalties as in Equations (7.2) and (7.3) that allow us to infer the invariances of Definition 7 that lead us to identify Granger non-causal statements.

7.3.2 Sparse Input MLPs

One first approach is to model each output component g_i with a separate MLP, so that we can easily disentangle the effects from inputs to outputs. We refer to this approach as a *componentwise* MLP (cMLP). Let g_i take the form of an MLP with $L-1$ layers and let the vector $h_t^\ell \in \mathcal{R}^m$ denote the values of the m-dimensional ℓth hidden layer at time t. The parameters of the neural network are given by weights $\mathbf{W}$ and biases $\mathbf{b}$ at each layer, $\mathbf{W} = \{W^1, \ldots, W^L\}$ and $\mathbf{b} = \{b^1, \ldots, b^L\}$. To draw an analogy with the time series VAR model, we further decompose the weights at the first layer across time lags, $W^1 = \{W^{11}, \ldots, W^{1K}\}$. The dimensions of the parameters are given by $W^1 \in \mathbb{R}^{m \times pK}$, $W^\ell \in \mathbb{R}^{m \times m}$ for $1 < \ell < L$, $W^L \in \mathbb{R}^m$, $b^\ell \in \mathbb{R}^m$ for $\ell < L$ and $b^L \in \mathbb{R}$. Using this notation, the vector of first layer hidden values at time t is given by

$$h_t^1 = \sigma\left(\sum_{k=1}^{K} W^{1k}\mathbf{x}_{t-k} + b^1\right), \tag{7.5}$$

where σ is an activation function. Typical activation functions are either `logistic` or `tanh` functions. The vector of hidden units in subsequent layers is given by a similar form, each with a σ activation function:

$$h_t^\ell = \sigma\left(W^\ell h_t^{\ell-1} + b^\ell\right). \tag{7.6}$$

After passing through the $L-1$ hidden layers, the time series output, x_{ti}, is given by a linear combination of the final hidden layer

$$x_{ti} = g_i(x_{<t}) + e_{ti} = W_L h_t^{L-1} + b^L + e_{ti} \tag{7.7}$$

where w_L is the linear output decoder and h_t^L is the final hidden output from the final $L-1$th layer. The error term, e_{ti}, is modeled as mean zero Gaussian noise with identity covariance. We chose this linear output decoder since our primary motivation involves real-valued multivariate time series. However, other decoders like a `logistic`, `softmax`, or poisson likelihood with exponential link function [135], could be used to model nonlinear Granger causality in multivariate binary [74], categorical [186], or positive count time series [74].

Penalized Selection of Granger Causality in the cMLP

In Equation (7.5), if the jth column of the first layer weight matrix, $W_{:j}^{1k}$, contains zeros for all k, then series j does not Granger-cause series i. That is, $x_{(t-k)j}$ for all k does not influence the hidden unit h_t^1 and thus the output x_{ti}. Per Definition 7, we see g_i is invariant to $x_{<tj}$. Thus, analogously to the VAR case, one may select for Granger causality by applying a group penalty to the columns of the W^{1k} matrices for each g_i,

$$\min_{\mathbf{W}} \sum_{t=K+1}^{T} \left(x_{it} - g_i(x_{(t-1):(t-K)}\right)^2 + \lambda \sum_{j=1}^{p} \Omega(W_{:j}^1) \tag{7.8}$$

where Ω is a penalty that shrinks the entire set of first layer weights for input series j, i.e., $W_{:j}^1 = (W_{:j}^{11}, \ldots, W_{:j}^{1K})$, to zero. We consider three different penalties, which, together, show how we recast structured regression penalties to the neural network case.

We first consider a *group lasso* penalty over the entire set of outgoing weights across all lags for time series j, $W_{:j}^1$,

$$\Omega(W_{:j}^1) = \|W_{:j}^1\|_F, \tag{7.9}$$

where $\|.\|_F$ is the Froebenius matrix norm. This penalty shrinks all weights associated with lags for input series j equally. For large enough λ, the solutions to Equation (7.8) with the group penalty in Equation (7.9) will lead to many zero columns in each W^{1k} matrix, implying only a small number of estimated Granger causal connections. This group penalty is the neural network analogue of the

group penalty across lags in Equation 7.2 for the VAR case.

To detect the lags where Granger causal effects exists, we propose a new penalty called a *group sparse group lasso* penalty. This penalty assumes that only a few lags of a series j are predictive of series i, and provides both sparsity across groups (a sparse set of Granger causal time series) and sparsity within groups (a subset of relevant lags)

$$\Omega(W^1_{:j}) = \alpha \|W^1_{:j}\|_F + (1-\alpha) \sum_{k=1}^{K} \|W^{1k}_{:j}\|_2. \tag{7.10}$$

where $\alpha \in (0, 1)$ controls the tradeoff in sparsity across and within groups. This penalty is a related to, and is a generalization of, the sparse group lasso [178].

Finally, we may simultaneously select for both Granger causality and the lag order of the interaction by replacing the group lasso penalty in Equation (7.8) with a *hierarchical* group lasso penalty [140] in the MLP optimization problem,

$$\Omega(W^1_{:j}) = \sum_{k=1}^{K} \|(W^{1k}_{:j}, \ldots, W^{1K}_{:j})\|_F. \tag{7.11}$$

The hierarchical penalty leads to solutions such that for each j there exists a lag k such that all $W^{1k'}_{:j} = 0$ for $k' > k$ and all $W^{1k'}_{:j} \neq 0$ for $k' \leq k$. Thus, this penalty effectively selects the lag of each interaction. The hierarchical penalty also sets many columns of W^{1k} to be zero across all k, effectively selecting for Granger causality. In practice, the hierarchical penalty allows us to fix K to a large value, ensuring that no Granger causal connections at higher lags are missed.

While the primary motivation of our penalties is for efficient Granger causality selection, the lag selection penalties in Equations (7.10) and (7.11) are also of independent interest to nonlinear forecasting with neural networks. In this case, overspecifying the lag of a NAR model leads to poor generalization and overfitting [17]. One proposed technique in the literature is to first select the appropriate lags using forward orthogonal least squares [17]; our approach instead combines model fitting and lag selection into one procedure.

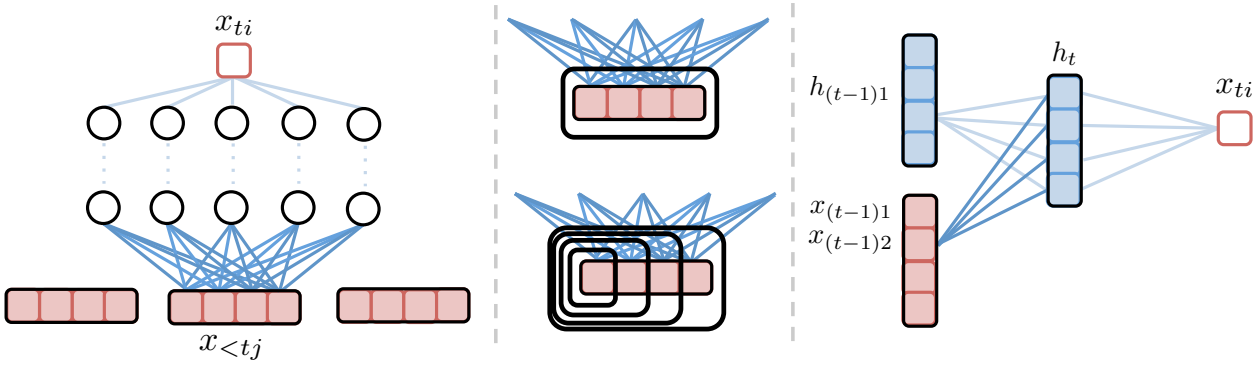


Figure 7.1: (left) Schematic for modeling Granger causality using cMLPs. If the outgoing weights for series j, shown in dark blue, are penalized to zero, then series j does not Granger-cause series i. (middle) The group lasso penalty jointly penalizes the full set of outgoing weights while the hierarchical version penalizes the nested set of outgoing weights, penalizing higher lags more. (right) Schematic for modeling Granger causality using a cLSTM. If the dark blue outgoing weights to the hidden units from an input $x_{(t-1)j}$ are zero, then series j does not Granger-cause series i.

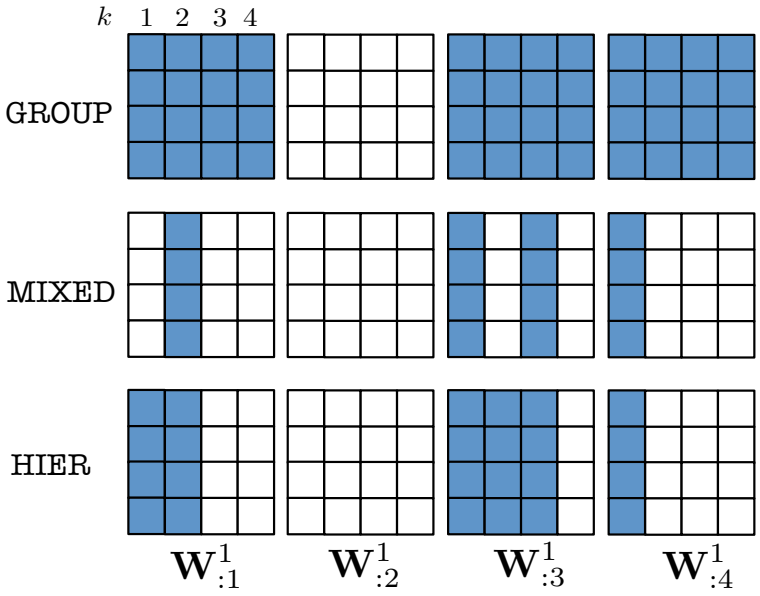


Figure 7.2: Example of group sparsity patterns of the first layer weights of a cMLP with four first layer hidden units and four input series with maximum lag $k = 4$. Differing sparsity patterns are shown for the three different structured penalties of group lasso (GROUP) from Equation (7.9), group sparse group lasso (MIXED) from Equation (7.10) and hierarchical lasso (HIER) from Equation (7.11).

7.3.3 Sparse Input cRNNs

Recurrent neural networks (RNNs) are particularly well suited for modeling time series, as they compress the past of a time series into a hidden state, aiming to capture complicated nonlinear dependencies at longer time lags than traditional time series models. As with MLPs, time series forecasting with RNNs typically proceeds by jointly modeling the entire evolution of the multivariate series using a single recurrent network.

As in the MLP case, it is difficult to disentangle how each series affects the evolution of another series when using an RNN. This problem is even more severe in complicated recurrent networks like LSTMs. To model Granger causality with RNNs, we follow the same strategy as with MLPs and model each g_i function using a separate RNN. For simplicity, we assume a single-layer RNN, but our formulation may be easily generalized to accommodate more layers.

Let $h_t \in \mathbb{R}^m$ represent the m-dimensional hidden state at time t, representing the historical context of the time series for predicting a component x_{ti}. The hidden state at time $t+1$ is updated recursively

$$h_t = f(\mathbf{x}_t, h_{t-1}), \tag{7.12}$$

where f is some nonlinear function that depends on the particular recurrent architecture.

Due to their effectiveness at modeling complex time dependencies, we choose to model the recurrent function f using an LSTM [72]. The LSTM model introduces a second hidden state variable c_t, referred to as the cell state, giving the full set of hidden parameters as (c_t, h_t). The

LSTM model updates its hidden states recursively as

$$\begin{aligned} f_t &= \sigma\left(W^f \mathbf{x}_t + U^f h_{(t-1)}\right) \\ i_t &= \sigma\left(W^{in} \mathbf{x}_t + U^{in} h_{(t-1)}\right) \\ o_t &= \sigma\left(W^o \mathbf{x}_t + U^o h_{(t-1)i}\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \sigma\left(W^c \mathbf{x}_t + U^c h_{(t-1)}\right) \\ h_t &= o_t \odot \sigma(c_t), \end{aligned} \tag{7.13}$$

where $\odot$ denotes componentwise multiplication and i_t, f_t, and o_t represent input, forget and output gates, respectively, that control how each component of the state cell, c_t, is updated and then transferred to the hidden state used for prediction, h_t. In particular, the forget gate, f_t, controls the amount that the past cell state influences the future cell state whereas the input gate i_t controls the amount that the current observation influences the new cell state. The additive form of the cell state update in the LSTM allows it to encode long-range dependencies, since cell states from far in the past may still influence the cell state at time t if the forget gates remain close to one. In the context of Granger causality, this flexible architecture can represent long-range, nonlinear dependencies between time series.

As in the cMLP, the output for series i at time t is given by a linear decoding of the hidden state

$$x_{ti} = g_i(x_{<t}) + e_{ti} = W^2 h_t + e_{ti}, \tag{7.14}$$

where W^2 are the output weights and we let $\mathbf{W} = (W^1, W^2)$ be the full set of parameters where $W^1 = \left((W^f)^T, (W^{in})^T, (W^o)^T, (W^c)^T\right)^T$ are the full set of first layer weights. As in the MLP case, other decoding schemes could be used in the case of categorical or count data.

Granger Causality Selection in LSTMs

In Equation (7.13) the set of input matrices W^1 controls how the past time series affects the forget gates, input gates, output gates, and cell updates, and, consequently, the update of the hidden

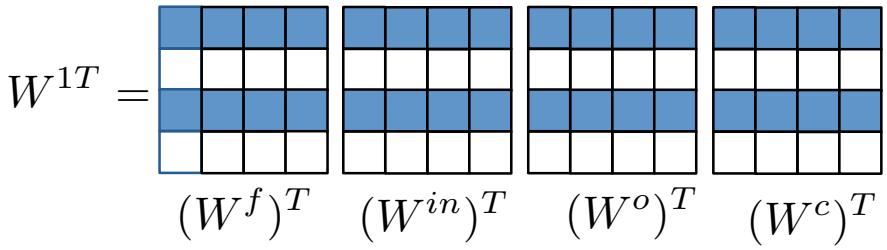


Figure 7.3: Example of the group sparsity patterns in a sparse cLSTM model with a four dimensional hidden state and four input series. Due to the group lasso penalty on the columns of W, the W^f, W^{in}, W^o, and W^c matrices will share the same column sparsity pattern.

representation. Like in the MLP case, for this componentwise LSTM model (cLSTM) a sufficient condition for Granger non-causality of an input series j on an output i is that all elements of the jth column of W are zero, $W_{:j} = 0$. Thus, we may select series that Granger-cause series i using a group lasso penalty across columns of W by

$$\min_{W,U,w^O} \sum_{t=2}^{T} \left(x_{it} - g_i(x_{<t})\right)^2 + \lambda \sum_{j=1}^{p} \|W^1_{:j}\|_2. \tag{7.15}$$

For a large enough λ, many columns of W will be zero, leading to a sparse set of Granger causal connections.

7.4 Optimizing the Penalized Objectives

7.4.1 Optimizing the Penalized cMLP Objective

We optimize the nonconvex objectives of Equation (7.8) using proximal gradient descent with line search [144]. Line search is preferred over gradient descent alone because it aids in convergence to a local optimum. Proximal optimization is important in our context because it leads to *exact zeros* in the columns of the input matrices, a critical requirement for interpreting Granger non-causality in our framework. The algorithm updates the network weights $\mathbf{W}$ iteratively starting with $\mathbf{W}^{(0)}$ by

$$\mathbf{W}^{(l+1)} = \text{prox}_{\gamma^{(l)}\lambda\Omega}\left(\mathbf{W}^{(l)} + \gamma^{(l)}\nabla\mathcal{L}(\mathbf{W}^{(l)})\right), \tag{7.16}$$

where $\mathcal{L}(\mathbf{W}) = \sum_{t=K+1}^{T}(x_{ti} - g_i(x_{<t}))^2$ is the neural network reconstruction loss and $\text{prox}_{\lambda\Omega}$ is the proximal operator with respect to the sparsity inducing penalty function Ω. The entries in $\mathbf{W}^{(0)}$ are initialized randomly from a standard normal distribution. The scalar $\gamma^{(l)}$ is the step size determined by line search [144]. While the objectives in Equation (7.8) are nonconvex, we find that no random restarts are required to accurately detect Granger causality connections.

Since the sparsity promoting group penalties are only on the input weights, the proximal step for weights at the higher levels is simply the identity function. The proximal step for the group lasso penalty on the input weights is given by a group soft-thresholding operation on the input weights [144]:

$$\text{prox}(\gamma^{(l)}\lambda W^1_{:k}) = \text{soft}(W^1_{:k}, \gamma^{(l)}\lambda) \tag{7.17}$$

$$= \left(1 - \frac{\lambda\gamma^{(l)}}{||W^1_{:j}||_F}\right)_+ W^1_{:k}, \tag{7.18}$$

where $(x)_+ = \max(0, x)$. For the group sparse group lasso, the proximal step on the input weights is given by group-soft thresholding on the lag specific weights, followed by group-soft thresholding on the entire resulting input weights for each series. See Algorithm 6. The proximal step on the input weights for the hierarchical penalty is given by iteratively applying the group soft-thresholding operation on each nested group in the penalty, from the smallest group to the largest group [92], and is shown in Algorithm 7.

Since all datasets we study are relatively small, the gradients are with respect to the full data objective; for larger datasets one could use proximal stochastic gradient [204].

7.4.2 Optimizing the Penalized cLSTM Objective

Similar to the cMLP, we optimize Equation (7.15) using proximal gradient descent with line search. When the data consists of many replicates of short time series, like in the DREAM3 data in Section 7.7, we perform a full backpropagation through time (BPTT) to compute the gradients. However, for longer series we truncate the backpropagation through time by unlinking the hidden sequences.

Algorithm 5 Proximal gradient descent with line search algorithm for solving Equation (7.8). Proximal steps given in Equation (7.17) for the group lasso penalty and in Equation (7) for the hierarchical penalty.

Require: $\lambda > 0$
  $m = 0$, initialize $\mathbf{W}^{(0)}$
  while not converged **do**
    $m = m + 1$
    determine γ by line search.
    for $j = 1$ to p **do**
      $W_{:j}^{1(m+1)} = \text{prox}_{\gamma\lambda\Omega}\left(W_{:j}^{1(m)} + \gamma\nabla_{W_{:j}^{1}}\mathcal{L}(\mathbf{W}^{(m)})\right)$
    end for
    for $l = 2$ to L **do**
      $W^{l(m+1)} = \mathbf{W}^{l(m)} + \gamma\nabla_{W^{l}}\mathcal{L}(\mathbf{W}^{(m)})$
    end for
  end while
  return$(\mathbf{W}^{(m)})$

Algorithm 6 One pass algorithm to compute the proximal map for the group sparse group lasso penalty for relevant lag selection in the cMLP model.

Require: $\lambda > 0$, $\gamma > 0$, $(W_{:j}^{11}, \ldots, W_{:j}^{1K})$
  for $k = K$ to 1 **do**
    $W_{:j}^{1k} = \text{soft}\left(W_{:j}^{1k}, \gamma\lambda\right)$
  end for
  $(W_{:j}^{11}, \ldots, W_{:j}^{1K}) = \text{soft}\left(\left(W_{:j}^{11}, \ldots, W_{:j}^{1K}\right), \gamma\lambda\right)$
  return$(W_{:j}^{11}, \ldots, W_{:j}^{1K})$

Algorithm 7 One pass algorithm to compute the proximal map for the hierarchical group lasso penalty for automatic lag selection in the cMLP model.

Require: $\lambda > 0$, $\gamma > 0$, $(W_{:j}^{11}, \ldots, W_{:j}^{1K})$
  for $k = K$ to 1 **do**
    $(W_{:j}^{1k}, \ldots, W_{:j}^{1K}) = \text{soft}\left(\left(W_{:j}^{1k}, \ldots, W_{:j}^{1K}\right), \gamma\lambda\right)$
  end for
  return$(W_{:j}^{11}, \ldots, W_{:j}^{1K})$

In practice, we do this by splitting the dataset up into equal sized batches, and treating each batch as an independent realization. Under this approach, the gradients used to optimize Equation (7.15) are only approximations of the gradients of the full component-wise LSTM model. This is very common practice in the training of of RNNs [202, 184, 199]. The full optimization algorithm for training is shown in Algorithm 8.

Algorithm 8 Proximal gradient descent with line search algorithm for solving Equation (7.8). Proximal steps given in Equation (7.17) for the group lasso penalty and in Equation (7) for the hierarchical penalty.

Require: $\lambda > 0$
  $m = 0$, initialize $\mathbf{W}^{(0)}$
  while not converged **do**
    $m = m + 1$
    compute $\nabla \mathcal{L}(\mathbf{W}^{(m)})$ by BPTT (truncated for large T)
    determine γ by line search.
    for $j = 1$ to p **do**
      $W_{:j}^{1(m+1)} = \text{soft}\left(W_{:j}^{1(m)} + \gamma \nabla_{W_{:j}^1} \mathcal{L}(\mathbf{W}^{(m)}), \gamma\lambda\right)$
    end for
    $W^{2(m+1)} = W^{2(m)} + \gamma \nabla_{W^2} \mathcal{L}(\mathbf{W}^{(m)})$
  end while
  return$(\mathbf{W}^{(m)})$

7.5 Comparing cMLP and cLSTM Models for Granger Causality

Both cMLP and cLSTM frameworks model each component function g_i using independent networks for each i. For the cMLP model, one needs to specify a maximum possible model lag K. However, our lag selection strategy (Equation 7.11) allows one to set that to a large value and the weights for higher lags are automatically removed from the model. On the other hand, the cLSTM model requires no maximum lag specification, and instead automatically learns the memory of each interaction. As a consequence, the cMLP and cLSTM differ in the amount of data used for training, as noted by a comparison of the t index in Equation (7.15) and Equation (7.11). For a length T series, the cMLP and cLSTM models use $T - K$ and $T - 1$ data points, respectively. While insignificant for large T, when the data consist of independent replicates of short series, as in

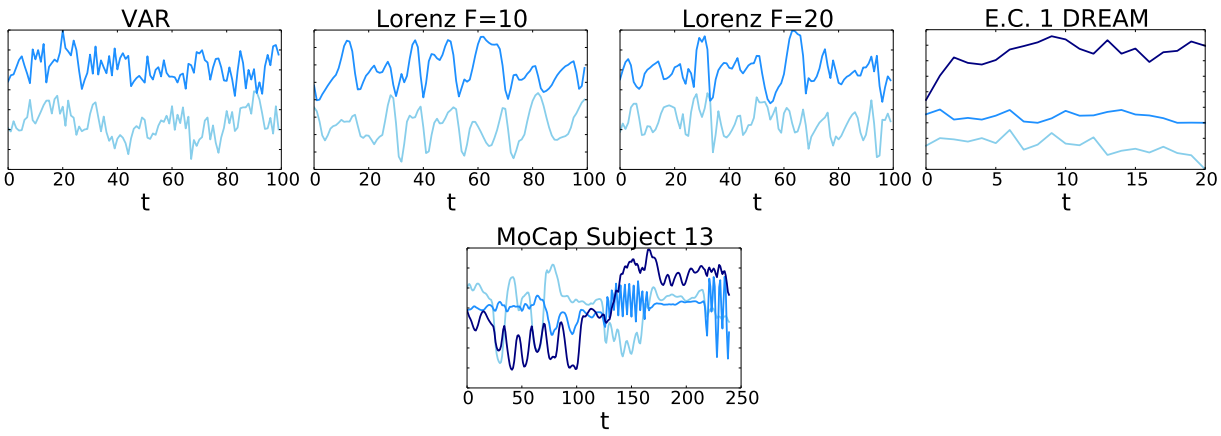


Figure 7.4: Example multivariate linear (VAR) and nonlinear (Lorenz, DREAM, and MoCap) series that we analyze using both cMLP and cLSTM models. Note as the forcing constant, F, in the Lorenz model increases, the data become more chaotic.

the DREAM3 data in Section 7.7, the difference may be important. This ability to simultaneously model longer range dependencies while harnessing the full training set may explain the impressive performance of the cLSTM in the DREAM3 data in Section 7.7.

Finally, the zero outgoing weights in both the cMLP and cLSTM are a sufficient but not necessary condition to represent Granger non-causality. Indeed, series i could be Granger non-causal of series j through a complex configuration of weights that exactly cancel each other. However, because we wish to *interpret* the outgoing weights of the inputs as a measure of dependence, it is important that these weights reflect the true relationship between inputs and outputs. Our penalization schemes in both cMLP and cLSTM acts as a prior that biases the network to represent Granger non-causal relationships with zeros in the outgoing weights of the inputs, rather than through other configurations. Our simulation results in Section 7.6 validate this intuition.

7.6 Simulation Experiments

7.6.1 cMLP and cLSTM Simulation Comparison

To compare and analyze the performance of our two approaches, cMLP and cLSTM, we apply both methods to detecting Granger causality networks in simulated linear VAR data and simulated

Lorenz-96 data [98], a nonlinear model of climate dynamics. Overall, our results show that our methods can accurately reconstruct the underlying Granger causality graph in both linear and nonlinear settings. We first describe the results from the Lorenz experiment in detail and present the VAR results subsequently.

Lorenz-96 Model

The continuous dynamics in a p-dimensional Lorenz model are

$$\frac{dx_{ti}}{dt} = \left(x_{t(i+1)} - x_{t(i-2)}\right) x_{t(i-1)} - x_{ti} + F, \tag{7.19}$$

where $x_{t(-1)} = x_{t(p-1)}$, $x_{t0} = x_{tp}$, $x_{t(p+1)} = x_{t1}$ and F is a forcing constant which determines the level of nonlinearity and chaos in the series. Example series for two settings of F are displayed in Figure 7.4. We numerically simulate a $p = 20$ Lorenz-96 model with a sampling rate of $\Delta_t = 0.05$, which results in a multivariate, nonlinear time series with sparse Granger causal connections. In particular each series i has Granger connections from series $(i-2)$, $(i-1)$, and $(i+1)$.

Average values of area under the ROC curve (AUROC) across five simulation seeds are listed in Table 7.1 for the cMLP and cLSTM models under two different data set lengths, $T \in (500, 1000)$, and forcing constants, $F \in (10, 40)$. We use $m = 10$ hidden units for both methods. While more layers may prove beneficial, for all experiments we fix the number of hidden layers, L, to one and leave the effects of additional hidden layers to future work. For the cMLP, we use the hierarchical penalty with model lag of $K = 5$; see Section 7.6.2 for a performance comparison of several possible penalties across model input lags.

The ROC values are computed by sweeping λ across a range of values; discarded edges (inferred Granger non-causality) for a particular λ setting are those whose associated L_2 norm of the input weights of the neural network is exactly zero. Note that our proximal gradient algorithm sets many of these groups to be exactly zero.

As expected, the results in Table 7.1 indicate that the cMLP and cLSTM performance improves as the data set size T increases. Furthermore, the cMLP outperforms the cLSTM in the less chaotic

F	10	10	40	40
T	500	1000	500	1000
cMLP	**95.7**	**99.1**	86.5	92.0
cLSTM	76.1	90.0	**87.5**	**96.25**

Table 7.1: Mean AUROC comparisons between cMLP and cLSTM Granger causality selection across five simulated Lorenz datasets, as a function of the forcing constant F, and the length of the time series T.

regime of $F = 10$, but underperforms in the more chaotic regime when $F = 40$.

VAR model

To analyze the performance of our methods when the true underlying dynamics are linear, we simulate data from $p = 20$ VAR(1) and VAR(2) models with randomly generated sparse transition matrices. To generate sparse dependencies for each time series i, we create self dependencies and randomly select three more dependencies among the other $p-1$ time series. Where series i depends on series j, we set $A^k_{ij} = 0.1$ for $k = 1, 2$. All other entries of A are set to zero. In the VAR(1) model, all dependencies occur at the first time lag, but in the VAR(2) model all dependencies occur at a time lag of 2. This is chosen to see how well our methods detect Granger causality at longer time lags, even though no time lag is explicitly specified in our methods. Our results are averages over five randomly generated sparse dependency graphs.

The average AUROC results are displayed in Table 7.2 for both cLSTM and cMLP models for $T \in (500, 1000, 5000)$. As expected, the performance of both models improves at larger T. Interestingly, the cMLP outperforms the cLSTM in both VAR(1) and VAR(2) cases. The cLSTM appears to have more difficulty in the VAR(2) case, but eventually performs well at larger T; we attribute this to the difficulty of our shallow cLSTM in memorizing information from previous lags, which is required to represent a VAR(2) model with dependencies only at the second lag. By contrast, the cMLP can directly use information from previous lags.

Model	VAR(1)			VAR(2)		
T	500	1000	5000	500	1000	5000
cMLP	97.9	99.6	100.0	92.6	95.9	99.9
cLSTM	80.1	93.8	99.9	67.6	88.6	97.7

Table 7.2: AUROC comparisons between cMLP and cLSTM Granger causality selection averaged over five simulated VAR datasets, as a function of the length of the time series T and VAR lag order.

7.6.2 Quantitative Analysis of the Hierarchical Penalty

K	5	10	20
GROUP	89.2	85.6	84.2
MIXED	90.2	88.5	87.4
HIER	**94.5**	**93.5**	**94.3**

Table 7.3: AUROC comparisons between different cMLP Granger causality selection penalties on simulated Lorenz data as a function of the input model lag, K.

We next quantitatively compare three possible structured penalties for Granger causality selection in the cMLP model. In Section 7.3.2 we introduced the full group lasso (GROUP) penalty over all lags (Equation 7.8), the group sparse group lasso (MIXED) (Equation 7.10) and the hierarchical (HIER) lag selection penalty (Equation 7.11). We compare these approaches across various choices of the cMLP model's maximum lag, $K \in (5, 10, 20)$. We use $m = 10$ hidden units for data simulated from the nonlinear Lorenz model with $F = 20$, $p = 20$, and $T = 750$. As in Section 7.6.1, we compute the mean AUROC over five simulation runs and display the results in Table 7.3. Importantly, the hierarchical penalty outperforms both group and mixed penalties across all model input lags K. Furthermore, performance significantly declines as K increases in both group and mixed settings while the performance of the hierarchical penalty stays roughly *constant* as K

increases. This result suggests that performance of the hierarchical penalty for nonlinear Granger causality selection is robust to the input lag, implying that precise lag specification is unnecessary. In practice, this allows one to set the model lag to a large value without worrying that nonlinear Granger causality detection will be compromised.

7.6.3 Qualitative Analysis of the Hierarchical Penalty

To qualitatively validate the performance of the hierarchical group lasso penalty for automatic lag selection, we apply our penalized cMLP framework to data generated from a sparse VAR model with longer interactions. Specifically, we generate data from a $p = 10$, VAR(3) model as in Section 7.1. To generate sparse dependencies for each time series i, we create self dependencies and randomly select two more dependencies among the other $p - 1$ time series. Where series i depends on series j, we set $A^k_{ij} = .1$ for $k = 1, 2, 3$. All other entries of A are set to zero. This implies that the Granger causal connections that do exist are all of lag order 3. We run the cMLP with the hierarchical group lasso penalty and a maximal lag order of $K = 5$.

We visually display the selection results for one cMLP (i.e., one output series) across a variety of λ settings in Figure 7.5. For the lower $\lambda = 0.01$ setting, the cMLP both (i) overestimates the lag order for a few input series and (ii) allows some false positive Granger causal connections. For the higher $\lambda = 0.037$, lag selection performs almost perfectly, in addition to correct estimation of the Granger causality graph. Higher λ values lead to larger penalization on longer lags, resulting in weaker long-lag connections. While we show results for multipple λ for visualization, in practice one may use cross validation to select the appropriate λ, and in turn, the per interaction lag.

7.7 DREAM Challenge

We next apply our methods to determine Granger causality networks from a realistically simulated time course gene expression data set. The data are from the DREAM3 challenge [152] and provide a difficult, nonlinear data set for rigorously comparing methods for Granger causality detection [123, 116]. The data is simulated using continuous gene expression and regulation dynamics, with

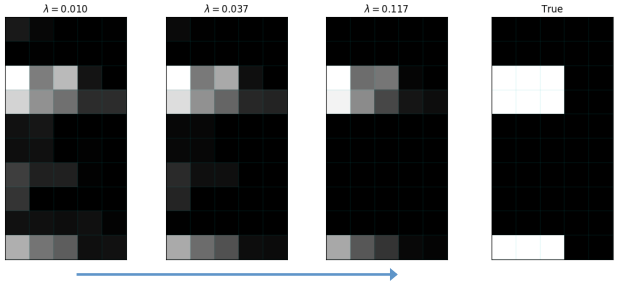


Figure 7.5: Qualitative results of the cMLP automatic lag selection using a hierarchical group lasso penalty and maximal lag of $K = 5$. The true data are from a VAR(3) model. The images display results for a single cMLP (one output series) using various penalty strengths λ. The rows of each image correspond to different input series while the columns correspond to the lag, with $k = 1$ at the left and $k = 5$ at the right. The magnitude of each entry is the L_2 norm of the associated input weights of the neural network after training. The true lag interactions are shown in the rightmost image. White represents positive magnitudes and black zero.

multiple hidden factors that are not observed. The challenge contains five different simulated data sets, each with different ground truth Granger causality graphs: two E. Coli (E.C.) data sets and three Yeast (Y.) data sets. Each data set contains $p = 100$ different time series, each with 46 replicates sampled at 21 time points for a total of 966 time points. This represents a very limited data scenario relative to the dimensionality of the networks and complexity of the underlying dynamics of interaction. Three time series components from a single replicate of the E. Coli 1 data set are shown in Figure 7.4.

We apply both the cMLP and cLSTM to all five data sets. Due to the short length of the series replicates, we choose the maximum lag in the cMLP to be $K = 2$ and use 5 and 10 hidden units for the cMLP and LSTM, respectively. For our performance metric, we consider the DREAM3 challenge metrics of area under the ROC curve (AUROC) and area under the precision recall curve (AUPR). Both curves are computed by sweeping λ over a range of values, as described in Section 7.6.

In Figure 7.6, we compare the AUROC and AUPR of our cMLP and cLSTM to previously published AUROC and AUPR results on the DREAM3 data [123]. These comparisons include both linear and nonlinear approaches: (i) a linear VAR model with a lasso penalty (LASSO) [128],

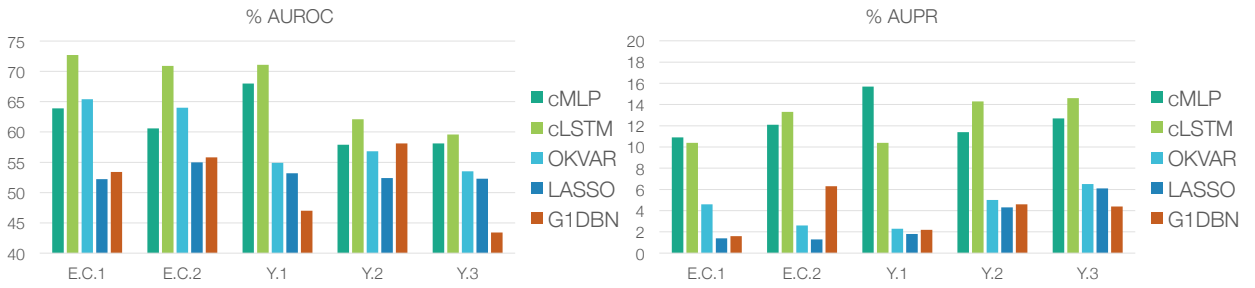


Figure 7.6: (Top) AUROC and (bottom) AUPR (given in %) results for our proposed regularized cMLP and cLSTM models and the set of methods—OKVAR, LASSO, and G1DBN—presented in [123]. These results are for the DREAM3 size-100 networks using the original DREAM3 data sets.

(ii) a dynamic Bayesian network using first-order conditional dependencies (G1DBN) [116], and (iii) a state-of-the-art multi-output kernel regression method (OKVAR) [123]. The latter is the most mature of a sequence of nonlinear kernel Granger causality detection methods [179, 133]. In terms of AUROC, our cLSTM outperforms all methods across all five datasets. Furthermore, the cMLP method outperforms previous methods on two datasets, Y.1 and Y.3, ties G1DBN on Y.2, and slightly under performs OKVAR in E.C.1 and E.C.2. In terms of AUPR, both cLSTM and cMLP methods do much better than all previous approaches, with the cLSTM outperforming the cMLP in three datasets. The raw ROC curves for cMLP and cLSTM are displayed in Figure 7.7.

These results clearly demonstrate the importance of taking a nonlinear approach to Granger causality detection in a (simulated) real-world scenario. Among the nonlinear approaches, the neural network methods are extremely powerful. Furthermore, the cLSTM's ability to efficiently capture longer range dependencies (without relying on long-lag specifications) appears to be particularly useful. This result validates many findings in the literature where LSTMs outperform MLPs. An interesting facet of these results, however, is that the impressive performance gains are achieved in a limited data scenario and on a task where the goal is recovery of interpretable structure. This is in contrast to the standard story of prediction on large datasets. To achieve these results the regularization and induced sparsity of our penalties is critical.

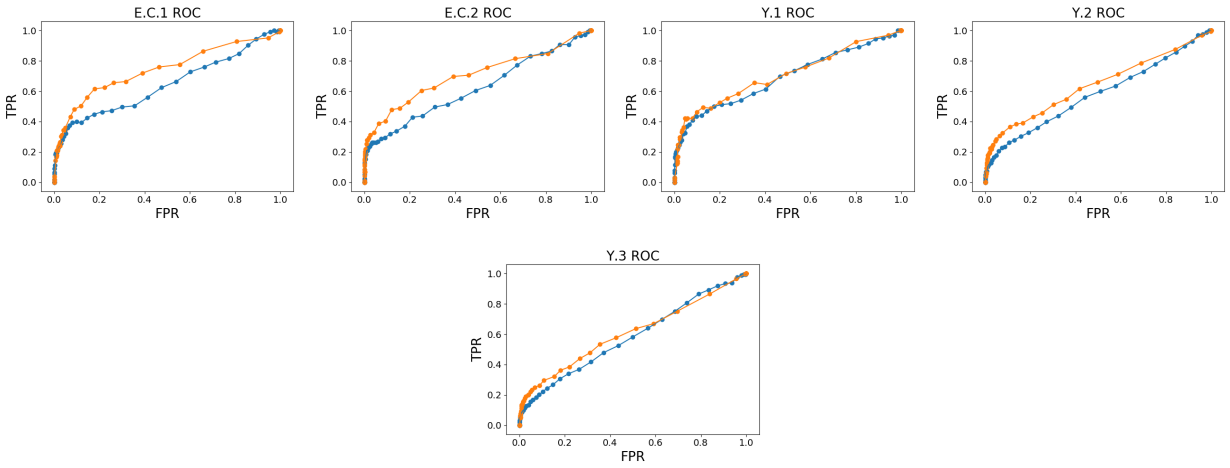


Figure 7.7: ROC curves for the cMLP and cLSTM models on the five DREAM datasets.

7.8 Dependencies in Human Motion Capture Data

We next apply our methodology to detect complex, nonlinear dependencies in human motion capture (MoCap) recordings. In contrast to the DREAM3 challenge results, this analysis allows us to more easily visualize and interpret the learned network. Human motion has been previously modeled using both linear dynamical systems [84], switching linear dynamical systems [145, 62] and also nonlinear dynamical models using Gaussian processes [198]. While the focus of previous work has been on motion classification [84] and segmentation [62], our analysis delves into the potentially long-range, nonlinear dependencies between different regions of the body during natural motion behavior. We consider a data set from the CMU MoCap database [36] previously studied in [62]. The data set consists of $p = 54$ joint angle and body position recordings across two different subjects for a total of $T = 2024$ time points. In total, there are recordings from 24 unique regions because some regions, like the thorax, contain multiple angles of motion corresponding to the degrees of freedom of that part of the body.

We apply the cLSTM model with $m = 8$ hidden units to this data set. For computational speed ups, we break the original series into length 20 segments and fit the regularized cLSTM model from Equation (7.15) over a range of λ values. To develop a weighted graph for visualization, we let

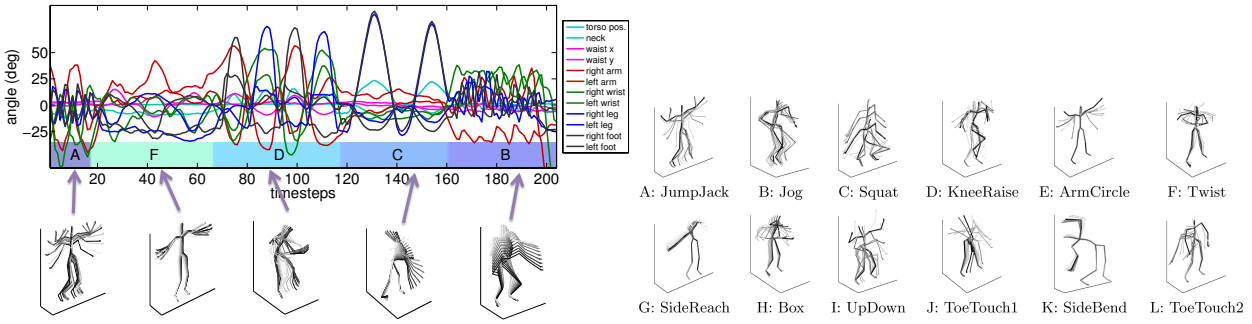
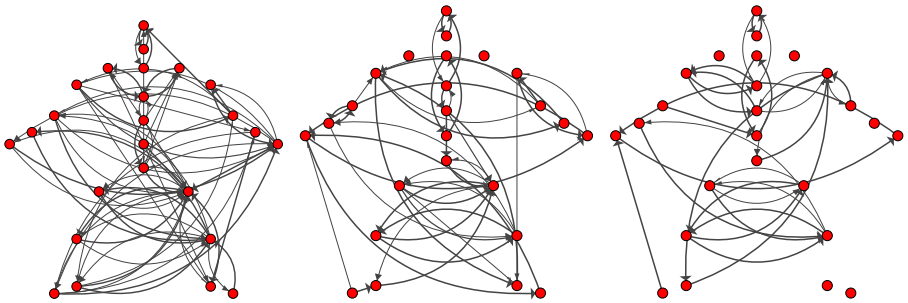


Figure 7.8: (top) Example time series from the MoCap data set paired with their particular motion behaviors. (bottom) Skeleton visualizations of 12 possible exercise behavior types observed across all sequences analyzed in the main text.

the edge weight w_{ij} between components be the norm of the outgoing cLSTM weights from input series j to output component series i, standardized by the maximum such edge weight associated with the cLSTM for series i. Edges associated with more than one degrees of freedom (angle directions for the same body part) are averaged together. Finally, to aid visualization, we further threshold edge weights $.01$ and below.

The resulting estimated graphs are displayed in Figure 7.9 for multiple values of the regularization parameter, λ. While we present the results for multiple λ, one may use cross validation to select λ if one graph is required. To interpret the presented skeleton plots, it is useful to understand the full set of motion behaviors exhibited in this data set. These behaviors are depicted in Figure 7.8, and include instances of *jumping jacks*, *side twists*, *arm circles*, *knee raises*, *squats*, *punching*, various forms of *toe touches*, and *running in place*. Due to the extremely limited data for any individual behavior, we chose to learn interactions from data aggregated over the entire collection of behaviors. In Figure 7.9, we see many intuitive learned interactions. For example, even in the more sparse graph (largest λ) we learn a directed edge from right knee to left knee and a separate edge from left knee to right. This makes sense as most human motion, including the motions in this dataset involving lower body movement, entail the right knee leading the left and then vice versa. We also see directed interactions leading down each arm, and between the hands and toes for toe touches.



(a) $\lambda = .18$ (b) $\lambda = .36$ (c) $\lambda = .40$

Figure 7.9: Nonlinear Granger causality graphs inferred from the human MoCap data set using the regularized cLSTM model. Results are displayed for a range of λ values. Each node corresponds to one location on the body.

7.9 Discussion

We have presented a framework for nonlinear Granger causality selection using regularized neural network models of time series. To disentangle the effects of the past of an input series on the future of an output, we model each output series using a separate neural network. We then apply both the component multilayer perceptron (cMLP) and component long-short term memory (cLSTM) architectures, with associated sparsity promoting penalties on incoming weights to the network, and select for Granger causality. Overall, our results show that these methods outperform existing Granger causality approaches on the challenging DREAM3 data set and furthermore discover interpretable and insightful structure on a MoCap data set.

Our work opens the door to multiple exciting avenues for future work. While we are the first to use a hierarchical lasso penalty in a neural network, it would be interesting to also explore other types of structured penalties, such as tree structured penalties [103].

Furthermore, although we have presented two relatively simple approaches, based off MLPs and LSTMs, our general framework of penalized input weights easily accommodates more powerful architectures. Exploring the effects of multiple hidden layers, powerful recurrent and convolutional architectures, like clockwork RNNs [108] and dilated causal convolutions [195], opens up a wide swath of research directions and the potential to detect long-range and complex dependencies.

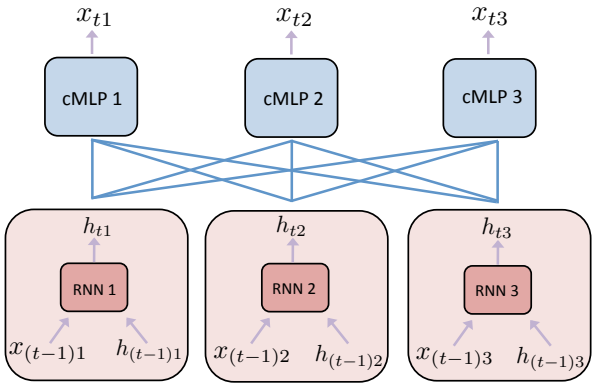


Figure 7.10: Proposed architecture for detecting nonlinear Granger causality that combines aspects of both the cLSTM and cMLP models. A separate hidden representation, h_{tj}, is learned for each series j using an RNN. At each time point, the hidden states are each fed into a sparse cMLP to predict the individual output for each series x_{ti}. Joint learning of the whole network with a group penalty on the input weights of the individual cMLPs would allow the network to share infomration about hidden features in each h_{tj} while also allowing interpretable structure learning between the hidden states of each series and each output.

Finally, while we consider sparse input models, a different *sparse output* architecture would use a network, like an RNN, to learn hidden representations of each individual input series, and then model each output component as a sparse nonlinear combination across the hidden states of all time series, allowing a shared hidden representation across component tasks. A schematic of the proposed architecutre that combines ideas from our cMLP and cLSTM models is shown in Figure 7.10.

Chapter 8

AN EFFICIENT ADMM ALGORITHM FOR STRUCTURAL BREAK DETECTION IN MULTIVARIATE TIME SERIES

8.1 Introduction

In many applied fields, such as neuroscience and economics, it is necessary to segment a non-stationary and multivariate signal into stationary regimes. Many methods have been proposed to accomplish this important challenge. Bayesian approaches [61] typically define a generative model, like a vector autoregressive model (VAR), for each stationary regime and a switching Markov process to model switches between regimes. Nonparametric methods [143, 151] directly analyze jumps in the spectral density of the process over time.

Recently, many authors have explored segmentation procedures based on convex optimization [28, 29, 164]. Similar to Bayesian methods, convex approaches model each regime using an autoregressive model. Instead of using a Markov switching process, fused group lasso penalties enforce the constraint that the autoregressive parameters of the process tend to stay constant over time, and only rarely switch to new parameter values. In practice, segmentation methodologies using fused group lasso penalties have relied on an approximate group least angle regression [207] solver for optimization [28, 29]. While an intuitive and widely used algorithm, group least angle regression does not provide any guarantees for returning the optimal solution for the convex segmentation problem.

Instead, we develop an efficient alternating direction method of multipliers (ADMM) algorithm [19] that directly solves the convex segmentation problem with group fused lasso penalties. Our ADMM approach splits the convex problem into a global quadratic program that may be solved in time linear with the series length and a simple group lasso proximal update. The work in this chapter is only partially complete; more work is required to study the theoretical properties of our

estimator, similar to [28, 164], and apply it to real data.

Both code for the ADMM algorithm and code to reproduce our experiments may be found at `bitbucket.org/atank/convex_tar`.

8.2 Background

Let $x_t \in \mathbb{R}^p$ be a p-dimensional multivariate time series. We assume that x_t follows a locally stationary vector autoregressive model with break points. Specifically, let L be the number of break points occurring at times $(t_1, \ldots, t_L)$. For each $t \in (t_i, t_{i+1}]$, x_t follows a stationary VAR of lag order K

$$x_t = \sum_{k=1}^{K} A^{ik} x_{t-k} + e_t, \tag{8.1}$$

where $\left(A^{ik}, \ldots, A^{iK}\right)$ are the K $p \times p$ matrices of the ith VAR process and $e_t \in \mathbb{R}^p$ is mean zero noise, $E(e_t) = 0$, with covariance $E(e_t e_t^T) = \Gamma^k$.

Given an observed time series at N time points, $(x_1, \ldots, x_N)$, the goal of estimation is to segment the series into $\hat{L} + 1$ stationary blocks, where $\hat{L}$ is the estimated number of change points. To do this, estimates of the break points, $\left(\hat{t}_1, \ldots, \hat{t}_{\hat{L}}\right)$, and estimates of the autoregressive VAR parameters, $\hat{A}^{ik}$ for $k \in (1, \ldots, K)$ and $i \in (1, \ldots, \hat{L})$, in each stationary segment must be determined.

While our approach to segmentation does not take the instantaneous covariance, Γ^k, within each segment into account, we do use a least square approach which should, in practice, be robust to differing instantaneous correlation structures. We leave the investigation of our approach under differing instantaneous error correlations as future work.

8.3 Estimation

We follow previous work and formulate structural break estimation in autoregressive models via a convex optimization problem with fused group lasso penalties [164, 29, 28]. First, we introduce local autoregressive parameters $A^t = \left(A^{t1}, \ldots, A^{tK}\right)$ that specify the autoregressive dynamics at

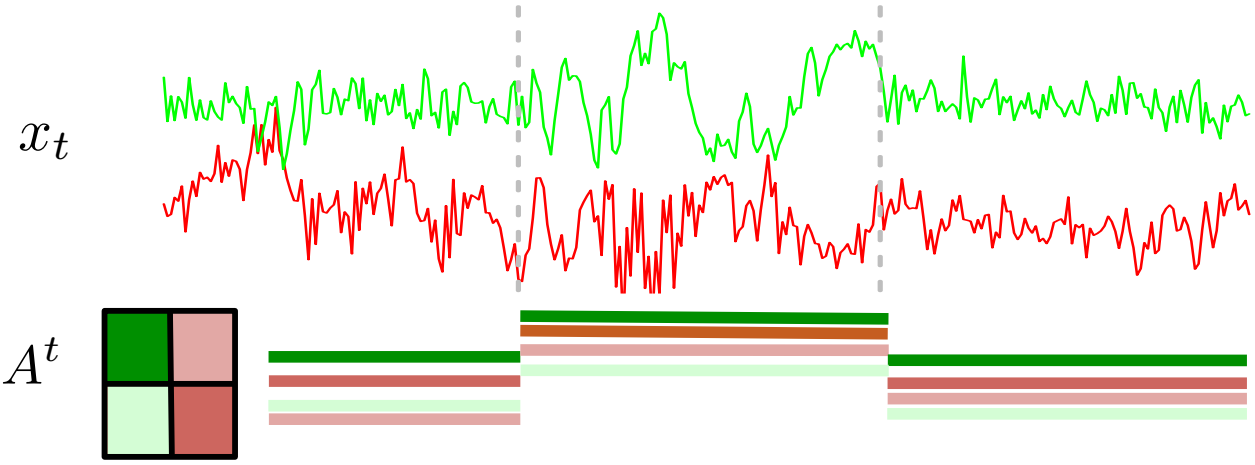


Figure 8.1: Schematic showing a bivariate time series with break point locations in dotted grey. The parameters of the VAR model are constant within a stationary segment but change value after a change point.

each time point. We then solve the following penalized least squares optimization problem

$$\min_{A^1,\ldots,A^N} \sum_{t=1}^{N} \|x_t - A^t \tilde{x}_t\|_2^2 + \lambda \sum_{t=2}^{N} \|A^t - A^{t+1}\|_F, \tag{8.2}$$

where $\tilde{x_t} = (x_{t-1}^T, \ldots, x_{t-K}^T)^T$, $\|.\|_F$ is the Frobenius norm that acts as a group lasso penalty and $\lambda > 0$ is a tuning parameter that controls the number of estimated break points. In this setting, the fused group lasso penalty shrinks the A^t and A^{t+1} parameter estimates to be identical. Let $(\hat{A}^1, \ldots, \hat{A}^N)$ be the solution to Problem (8.2). Then the change point estimates $(t_1, \ldots t_{\hat{L}})$ are times t when $\hat{A}^t \neq \hat{A}^{t+1}$ and $\hat{L}$ is the number of such time points.

8.4 ADMM Algorithm

Problem (8.2) poses multiple challenges for efficient optimization. First, the number of parameters to be estimated in this model, the full sequence of $(A^1, \ldots, A^N)$, grows linearly with the length of the sequence. Thus we seek an optimization method that efficiently deals with the large number of parameters. Second, the group lasso penalty needed to select for break points, $||.||_F$, is *non-*

smooth. This means that many standard smooth optimization methods, like gradient descent or Newton-Raphson, are not applicable.

To address the above challenges, we develop an efficient ADMM algorithm that simultaneously scales linearly with the length of the time series, N, and also efficiently deals with the $N-1$ number of non-smooth group lasso penalty terms. Furthermore, the resulting algorithm solves Problem (8.2) exactly.

First, we introduce a change of variables parameterization $\theta^1 = A^1$ and $\theta^t = A^t - A^{t+1}$ for $t > 1$. The reparameterization lets us rewrite Problem (8.2) as

$$\min_{\theta^1,\ldots,\theta^N} \|Y - X\theta\|_F + \lambda \sum_{t=2}^{N} \|\theta^t\|_F, \tag{8.3}$$

where $\theta = (\theta^1, \ldots, \theta^N)^T$, $Y = (x_1, \ldots, x_N)^T$ and

$$X = \begin{pmatrix} \tilde{x}_1^T & 0 & 0 & \ldots & 0 \\ \tilde{x}_2^T & \tilde{x}_2^T & 0 & \ldots & 0 \\ \tilde{x}_3^T & \tilde{x}_3^T & \tilde{x}_3^T & \ldots 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tilde{x}_N^T & \tilde{x}_N^T & \tilde{x}_N^T & \ldots & \tilde{x}_N^T \end{pmatrix}. \tag{8.4}$$

Since Problem (8.3) takes the form of a group lasso regression problem, approximate solvers like group least angle regression may be used [29, 28]. However, we instead develop an efficient ADMM algorithm to solve Problem (8.3) exactly.

As is standard in ADMM, we introduce auxilliary parameters $W = (W^1, \ldots, W^N)^T$ and the constraint $W = \theta$ to break apart the least squares term and the group lasso penalty in Problem (8.3):

$$\min_{W,\theta} \|Y - X\theta\|_F + \lambda \sum_{t=2}^{N} \|W^t\|_F \quad \text{such that } W = \theta. \tag{8.5}$$

The augmented Lagrangian for Problem (8.5) is given by

$$\min_{W,\theta} \|Y - X\theta\|_F + \lambda \sum_{t=1}^{N} \|W^t\|_F + \frac{\rho}{2}\|\theta - W\|_F^2 + \text{trace}(\Omega^T(\theta - W)), \tag{8.6}$$

where $\Omega \in \mathbb{R}^{p \times pKN}$ are Lagrange multipliers and $\rho > 0$. The scaled ADMM update steps for iteratively solving the augmented Lagrangian of Problem (8.6) are given by [19]

$$\theta^{(l+1)} = \min_{\theta} \|Y - X\theta\|_F + \frac{\rho}{2}\|\theta - W^{(l)} + \Omega^{(l)}\|_F^2 \tag{8.7}$$

$$W^{(l+1)} = \min_{W} \lambda \sum_{i=2}^{N} \|W^i\|_F + \frac{\rho}{2}\|\theta^{(l+1)} - W + \Omega^{(l)}\|_F^2 \tag{8.8}$$

$$\Omega^{(l+1)} = \Omega^{(l)} + \theta^{(l+1)} - W^{(l+1)}. \tag{8.9}$$

We refer to Problem (8.7) as the global problem and Problem (8.8) as the local problem. The left hand side of the objectives in Problems (8.7) and (8.8) correspond to the reconstruction error and group lasso terms of the primary objective in Problem (8.5). The right hand side of each objective may be thought of as energy terms that shrink W and θ towards eachother.

The θ and W variables are referred to as the *primal* variables whereas Ω denotes the set of *dual* variables that enforce the constraint that $\theta = W$. The algorithm converges when both primal and dual variables converge. Furthermore, at the primal and dual solution, $W = \theta$. It is easy to see that the dual updates converge when $W = \theta$ since the update in Problem (8.9) reduces to $\Omega^{(l+1)} = \Omega^{(l)}$. In practice, θ and W variables may start far apart but converge to the same value. Furthermore, the step size hyper parameter ρ controls how fast θ and W are brought together. For more information about the derivation of the scaled ADMM steps and explanations of ADMM convergence see [19].

We present the specific form for solving Problems (8.7) and (8.8) in Sections 8.4.1 and 8.4.4 below.

8.4.1 Global θ update

Although the global ADMM subproblem given in Equation (8.7) is a quadratic program with p^2KN variables, we develop an efficient linear time algorithm with complexity $O(N)$. First, we reintroduce the $\theta^t = A^{t+1} - A^t$ parameterization, which gives

$$\min_{A^1,\ldots,A^N} \sum_{t=1}^{N} \|x_t - A^t \tilde{x}_t\|_2^2 + \frac{\rho}{2} \sum_{t=1}^{N} \|A^{t+1} - A^t - W^{t(l)} + \Omega^{t(l)}\|_F^2. \tag{8.10}$$

Problem (8.10) may be decomposed into p independent problems which may be solved in parallel for each row of $A^t = \left(a_1^t, \ldots, a_p^t\right)^T$. The problem for each $(a_j^1, \ldots, a_j^N)$ is given by

$$\min_{a_j^1,\ldots,a_j^N} \sum_{t=1}^{N} \left(x_{tj} - \tilde{x}_t^T a_j^t\right)^2 + \frac{\rho}{2} \sum_{t=1}^{N} \|a_j^{t+1} - a_j^t - w_j^{t(l)} + \phi_j^{t(l)}\|_2^2. \tag{8.11}$$

where $w_j^{t(l)}$ is the jth row of $W^{t(l)}$ and $\phi_j^{t(l)}$ is the jth row of $\Omega^{t(l)}$. Problem (8.11) may be solved efficiently by noting that it takes the same form as a canonical smoothing problem for the $\left(a_j^1, \ldots, a_j^N\right)$ in a state space model [38]. This follows from the fact that Problem (8.11) is the negative log-likelihood of a Gaussian state-space model.

8.4.2 Global Problem as Inference in a Gaussian State-Space Model

To elucidate the connection between Problem (8.11) and smoothing in a state-space model note that Gaussian state-space models [38] take the following canonical generative form:

$$z_t = B_t z_{t-1} + \mu_t + \eta_t \tag{8.12}$$

$$y_t = C_t z_t + e_t \tag{8.13}$$

for a d-dimensional state variable sequence $(z_t, \ldots, z_N)$, where by convention $z^0 = 0$, and m-dimensional observation sequence $(y_1, \ldots, y_N)$. Equation (8.12) specifies the transition distribu-

tion between past and future states of z_t. $B_t \in \mathbb{R}^{d\times d}$ is the transition matrix, $\mu_t \in \mathbb{R}^d$ is the bias at time t, and $\eta_t \sim N(0, Q_t)$ is the Gaussian transition noise with $p \times p$ covariance matrix Q_t. Equation (8.13) specifies the emission distribution for the observation sequence $y_t \in \mathbb{R}^m$ as a linear transformation with observation matrix $C_t \in \mathbb{R}^{d\times m}$ of the state sequence z_t plus Gaussian noise, $e_t \sim N(0, W_t)$, where W_t is the $m \times m$ observation noise covariance.

The joint negative log-likelihood of a state sequence and observation sequence under the Gaussian state-space model of Equations (8.12) and (8.13) is given by

$$\sum_{t=1}^{N} \frac{1}{2}(y_t - C_t z_t)Q_t^{-1}(y_t - C_t z_t) + \sum_{t=1}^{N} \frac{1}{2}(z_t - B_t z_{t-1} - \mu_t)R_t^{-1}(z_t - B_t z_{t-1} - \mu_t). \qquad (8.14)$$

Comparison of the general log-likelihood in (8.14) with the global ADMM problem in (8.11) identifies the objective in Problem (8.11) as a Gaussian state-space model with

1. $(a_j^1, \ldots, a_j^N) = (z_1, \ldots, z_t)$ as the state variables of dimension $d = pK$,

2. $(x_{1j}, \ldots, x_{Nj}) = (y_1, \ldots, y_N)$ as the observation sequence with dimension $m = 1$,

3. $B = I_{pK\times p_k}$ as the state transition matrix,

4. $\mu_t = w_j^{t(l)} - \phi_j^{t(l)}$ as the transition bias,

5. $Q_t = \frac{1}{\rho} I_{pK\times pK}$ as the state evolution noise covariance matrix and

6. $R_t = \frac{1}{2}$ as the observation noise variance.

The connection between the global ADMM optimization problem and a Gaussian state-space model is displayed pictorially in Figure 8.2 and the parameters of the state-space model for the global step are shown in Table 8.1.

state-space param.	z_t	y_t	μ_t	C_t	B_t	R_t	Q_t
global update param.	a_t^j	x_{jt}	$w_j^{t(l)} - \phi_j^{t(l)}$	$\tilde{x}_t$	$I_{pK \times pK}$	$\frac{1}{\rho} I_{pK \times pK}$	$\frac{1}{2}$

Table 8.1: Correspondence between the parameters of a Gaussian state-space model (Equations 8.12 and 8.13) and the terms in the global update of the ADMM algorithm (Problem 8.11). The correspondence is used to develop a Kalman filter-smoother algorithm for inferring the optimal a_t^j sequence of the global problem in linear time.

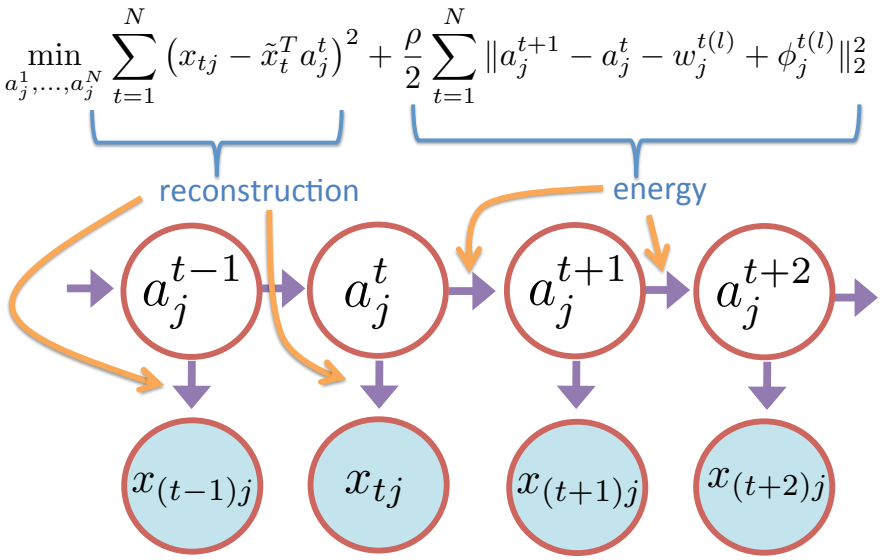


Figure 8.2: Graphical model schematic showing the connection between the global ADMM optimization problem (Problem 8.11) and a Gaussian state-space model (Equations 8.12 and 8.13). In particular, the reconstruction error between the future and the linear prediction of the past, $(x_{tj} - \tilde{x}^T a_j^t)$, corresponds to the Gaussian emissions of the state-space model. The energy term, that biases the first differences $a_j^{t+1} - a_j^t$ to be close to $w_j^{t(l)} - \phi_j^{t(l)}$, corresponds to the state transitions.

8.4.3 Global Update with Kalman Filter-Smoothing

Inference for the maximum likelihood state sequence $\left(a_j^1, \ldots, a_j^N\right)$ in this model may be solved using a *Kalman filtering-smoothing* algorithm. Kalman smoothers compute the expected value of the latent sequence given the observations. Due to Gaussianity, this expected value is the same as the maximum of the log-likelihood. Many smoothing algorithms exist but here we employ the classical Rauch-Tung-Streibel smoother [38]. In our case, this smoothing algorithm reduces to first computing the following forward filtering steps initialized with $\hat{a}_j^{1|1} = \mu_1$ and $\Sigma^{1|1} = \rho^{-1} I_{pK \times pK}$, and then for $t > 1$ recursively computing :

$$\begin{aligned}
\hat{a}_j^{t|(t-1)} &= \hat{a}_j^{(t-1)|(t-1)} + \mu_t \\
\Sigma^{t|(t-1)} &= \Sigma^{(t-1)|(t-1)} + \rho^{-1} I_{pK \times pK} \\
K^t &= \frac{\Sigma^{t|(t-1)} \tilde{x}_t}{\frac{1}{2} + \tilde{x}_t^T \Sigma^{t|(t-1)} \tilde{x}_t} \\
\hat{a}_j^{t|t} &= \hat{a}_j^{t|(t-1)} + K^t \left(x_{tj} - \tilde{x}_t^T \hat{a}_j^{t|(t-1)}\right) \\
\Sigma^{t|t} &= (I - K^t \tilde{x}_t^T) \Sigma^{t|(t-1)}.
\end{aligned}$$

The full forward pass is completed once $\hat{a}_j^{t|t}$ and $\Sigma^{t|t}$ have been computed for $t = 1, \ldots, N$. Since computing $\hat{a}_j^t|t$ and $\Sigma^{t|t}$ depends only on $\hat{a}_j^{(t-1)|(t-1)}$ and $\Sigma^{(t-1)|(t-1)}$, respectively, the forward pass may be computed in $O(N)$ operations.

Once the forward pass has been completed, the optimal state sequence $(\hat{a}_j^1, \ldots \hat{a}_j^N)$ solution to Problem 8.7 may be computed with a recursive *backwards* smoothing pass that starts at $\hat{a}_j^N$ and works backwards. Specifically, we initialize the backward pass with $\hat{a}^N = \hat{a}^{N|N}$ and $\Sigma^N = \Sigma^{N|N}$. We next recursively compute the rest of the backward messages, starting with $t = N - 1$ and ending at $t = 1$:

$$\hat{a}_j^t = \hat{a}_j^{t|t} + C^t \left(\hat{a}^{t+1} - \hat{a}^{t+1|t}\right) \tag{8.15}$$

$$\Sigma^t = \Sigma^{t|t} + C^t \left(\Sigma^{t+1} - \Sigma^{t+1|t}\right) C^{tT}, \tag{8.16}$$

where $C^t = \Sigma^{t|t}\left(\Sigma^{(t+1)|t}\right)^{-1}$. The result of the backward pass, $(\hat{a}_j^1, \ldots, \hat{a}_j^N)$, is the optimal solution to Problem (8.11). The updates in Equations 8.15 and 8.16 depend on the future values of $\hat{a}_j^{t+1}$ and Σ^{t+1} from the backward pass and also on the $\hat{a}_j^{t|t}$, $\hat{a}_j^{t+1|t}$, $\Sigma^{t|t}$, and $\Sigma^{t+1|t}$ from the forward pass. Since each update in the backward pass only depends on the previous backward pass update, the full $(\hat{a}_j^1, \ldots, \hat{a}_j^N)$ is computed in $O(N)$.

The forward and backward update steps presented here is a simplification of the full forward-backward algorithm for the general state-space model since in the global ADMM problem the transition matrix $B_t = I_{pK\times pK}$ and transition covariance $Q_t = \frac{1}{\rho} I_{pK\times pK}$ are diagonal and constant across time. Furthermore, the observation noise $R_t = \frac{1}{2}$ is one dimensional and also constant across time. See Section 3.5 for the filtering-smoothing algorithm in the general case. Since both forward and backward passes require $O(N)$ operations, the full smoothing computation to solve Problem (8.11) is $O(N)$. The forward messages and backward messages are displayed pictorially in Figure 8.3.

8.4.4 W update

The W update in Problem (8.8) is given separately for each W^t. Specifically, it is given by the proximal operator for the group lasso penalty:

$$W^{t(l+1)} = \text{prox}_{\frac{\lambda}{\rho}||.||_F}(\theta^{t(l+1)} + \Omega^{t(l)}). \tag{8.17}$$

where the prox operator for the group lasso penalty is given by separate soft group thresholding for each time t

$$W^{t(l+1)} = \begin{cases} 0 & \text{if } \|\theta^{t(l+1)} + \Omega^{t(l)}\|_F < \frac{\lambda}{\rho} \\ \left(1 - \frac{\lambda}{\rho\|\theta^{t(l+1)} + \Omega^{t(l)}\|_F}\right)\left(\theta^{t(l+1)} + \Omega^{t(l)}\right) & \text{otherwise.} \end{cases} \tag{8.18}$$

Intuitively, this proximal operator shrinks certain parameter increments to zero when the norm of $\theta^{t(l+1)} + \Omega^{t(l)}$ is less than $\frac{\lambda}{\rho}$. Recall that change points are detected when $W^t = A^{t+1} - A^t = 0$.

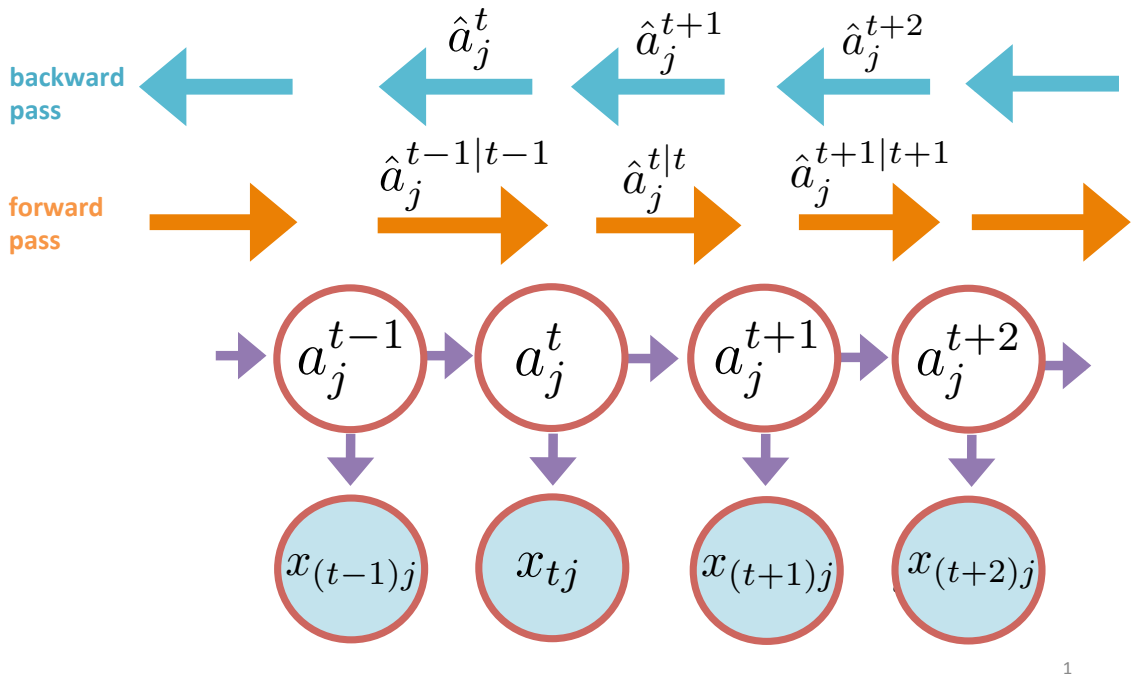


Figure 8.3: Schematic displaying the forward and backward messages in the Kalman Filtering-Smoothing algorithm used to compute the optimal state sequence $(\hat{a}_j^1, \ldots, \hat{a}_j^N)$ from the global ADMM problem in Equation (8.11). The forward pass consists of recursively computing messages $\hat{a}_j^{t|t}$ in linear time. The backward pass consists of starting at time N and recursively computing the backward messages, which in this case are the optimal state sequences, $\hat{a}_j^t$, also in linear time.

Thus, this group lasso proximal step effectively determines the break points.

Taken together, the global step of the ADMM algorithm is computed in $O(N)$ time and the proximal update for the group lasso for each W^t is computed in $O(N)$. Thus the full algorithm has a total per iteration complexity of $O(N)$.

8.4.5 Automatic Step Size Tuning

Convergence speed of the proposed ADMM algorithm depends on the step size parameter, ρ. Unfortunately, the optimal setting for ρ depends on the dataset so it is difficult to specify default settings. Instead, we utilize an adaptive step size routine [19] that increases or decreases ρ by a constant factor to ensure that primal and dual convergence is on same scale. See [19] for more information on automatic step size tuning in ADMM.

8.5 Simulation

To test our algorithm, we detect break points on a $p = 10$ series with $N = 300$ time points. We randomly generate a series with two structural break points at times $t \in (100, 200)$ for a total of three stationary regions each generated by a different VAR(1) process. The noise covariance for each stationary region is isotropic, i.e. $\Gamma = \sigma^2 I_{p\times p}$. We run our algorithm for three lambda values, $\lambda \in (1, 3, 5)$, and declare a break point is detected if $||\theta^t||_F > .005$. The estimated break points are shown in Figure 8.5. Overall, the $\lambda = 5$ case accurately detects the break points, while smaller λ values tend to overestimate the number of break points.

8.6 Discussion and Future Work

8.6.1 Scaling to higher dimensions

The global step of our ADMM algorithm solves p independent smoothing problems. While the Kalman filter we utilize has runtime linear in N, each recursive step during the backward smoothing phase requires calculating the inverse of a $pK \times pK$ matrix. While this computation is viable

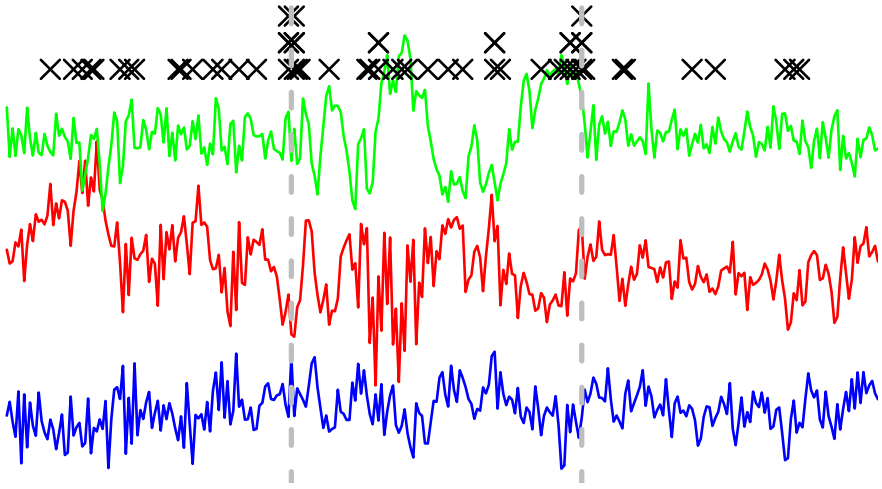


Figure 8.4: Three of ten series series from a $p = 10$ $N = 300$ series with estimated change points. Each row of $\times$s indicate detected change points for a different $\lambda \in (1, 3, 5)$. True change point times shown in dotted grey.

for moderate sized p, future work aims to explore alternate Gaussian state-space smoothers that scale better with p for high dimensional applications.

8.6.2 Quadratic time pruning of change points for consistent estimation

One can show that estimation of break points using solutions to Problem (8.2) is not consistent for the true break points [28, 29, 164]. In fact, change point detection with fused group lasso penalties tends to overestimate the number of change points. However, it can be shown that with high probability, as the length of each stationary segment increases, the true change points are contained in the set of estimated change points. Thus, multiple authors have proposed a further pruning stage that consistently picks the true set of change points from the candidate change points estimated via the fused group lasso estimation stage [28, 29, 164].

Unfortunately, the pruning stage requires identifying the subset of candidate change points that gives the model with highest Akaike information criteria (AIC) score [28, 29, 164]. To identify this optimal subset, previous authors have had to exhaustively compute the AIC for all change point subsets of the candidate set and then choose the one with smallest value. However, we have

noticed that it is possible to determine the optimal change points using a dynamic programming algorithm on the candidate change points that only takes time quadratic in the number of change points. Taken together, our full algorithm for structural change point detection utilizes an ADMM algorithm with linear per iteration complexity to identify candidate change points followed by a dynamic programming pruning stage that takes time quadratic with the number of candidate change points.

Chapter 9

THESIS CONTRIBUTIONS AND FUTURE DIRECTIONS

We first review the technical contributions presented in this thesis and then discuss future directions

9.1 Contributions

Bayesian structure learning The work in Chapter 4 introduces a novel Bayesian methodology for inferring conditional independence graphs from time series. We introduce independent complex hyper inverse Wishart prior distributions over spectral densities at each frequency. Combined with a prior over graphs and the Whittle likelihood, this presents a full Bayesian model for time series with a sparse inverse spectral density matrix. For estimation we marginalize out the infinite dimensional spectral matrices and use graph search to find the best fitting graph.

Structural VAR models for subsampled and mixed-frequency data Chapter 5 introduces novel theory and estimation methodology for inferring lagged Granger and instantaneous causality from subsampled and mixed-frequency data. The theory provides conditions on the parameters of a SVAR model that lead to model identifiability. A novel EM algorithm is presented for inference and validated through numerous simulation studies and real world data applications.

Granger causality for categorical time series Chapter 6 introduces novel methodology for inferring Granger causality networks for categorical time series. We develop a novel convex parameterization of the mixture transition distribution and novel identifiability restrictions on the parameters. We further explore sparsity promoting penalties for model selection in the resulting convex model. A projected gradient algorithm is also developed for estimation that scales to high dimensions.

Neural Granger causality for nonlinear time series In Chapter 7 we introduce novel neural network architectures for estimation of nonlinear Granger causality. In particular, we apply group lasso penalties to the outgoing weights of the input layer of a neural network. We develop the framework for both static multilayer perceptrons and recurrent networks. In the multilayer perceptron case we also introduce structured penalties for automatic lag selection of nonlinear Granger causality.

Change point detection for interactions In Chapter 8 we develop an ADMM algorithm for segmenting multivariate time series into stationary segments. Importantly, our algorithm is the first one that exactly solves the group penalized dynamics change point problem. Furthermore, one step of the algorithm may be computed in linear time by utilizing a Kalman smoother.

9.2 Future Directions

Bayesian Spectral Modeling The work in Chapter 4 utilizes a Bayesian model where the Whittle approximation to multivariate Gaussian time series is used as a likelihood. This approach of using a multivariate complex normal likelihood to represent the Whittle approximation in a Bayesian model is under utilized and invites multiple extensions. For example, the work of [60] develops a Bayesian time varying low rank covariance matrix to model heteroskedastic Gaussian data. A similar approach, with similar sampling algorithms, may be used to develop a low rank, nonparametric, model of the spectral density matrix that smoothly changes over time. Specifically, a $2-$dimensional complex Gaussian process prior may be placed on entries of a low rank approximation to the spectral density matrix at each frequency time point pair. Combined with the Whittle likelihood at each frequency and distinct time windows, the approach would automatically smooth the spectral density estimates across frequency and time.

Granger causality for subsampled and mixed frequency time series For the work in Chapter 5 on Granger and instantaneous causality in sub-sampled and mixed frequency time series the biggest bottleneck is computational. New methods need to be developed that may scale inference

in these non-Gaussian models to higher dimensions. One avenue may be in utilizing Bayesian sampling approaches that are able to more easily hop between modes of the posterior [131]. The work in this thesis also only developed theory and estimation methodology for VAR models of one lag. Future work should develop similar theory and methodology for higher order models as well.

In higher dimensions, Granger causality estimation in mixed frequency time series is difficult due to the presence of missing data. Standard approaches in non-missing time series utilize maximum likelihood approaches combined with L_1 penalties to select for which coefficients are zero. Unfortunately, L_1 penalized estimation of the VAR transition matrix in subsampled series is a non-convex optimization problem, so may be difficult in practice to actually select the correct Granger causality network pattern. An alternative approach would utilize a sparse solution to the extended Yule-Walker equations [30]. The extended Yule-Walker equations are a method of moments estimator of the VAR transition matrix for subsampled time series. [76] develop a sparse estimator of the transition matrix from the standard Yule-Walker equations for normally sampled data; applying the method of [76] to the extended Yule-Walker equations would provide a fast, scalable, and convex way of assessing Granger causality in mixed frequency VAR models.

Granger causality for categorical time series In Chapter 6 methodology was developed for inferring Granger causality networks in categorical time series. Future work in this area should aim at developing high dimensional consistency theory for both the penalized MTD and penalized mLTD model. It would be very interesting to see how these estimation rates compare to those of the MTD model. It would also be interesting to study convex MTD models with interactions, and also ones with longer lags. Lag selection penalties could be used in this context. At the end of Chapter 6 we present a method for robust and convex estimation of matrix factorization models, including non-negative matrix factorization and topic modeling, based on the convex substitution trick used for the convex MTD framework. Developing both estimation theory and efficient optimization routines for these problems is another important direction for future work.

Neural Granger causality for nonlinear time series There are also multiple potential extensions of the work in Chapter 7. A first avenue to explore would be a method that combines both MLP and LSTM approaches by adding an MLP on top of the LSTM representation and applying the group lasso penalties at the layer between the LSTM and MLP for Granger causality selection. A second approach would be to use Jacobian penalties to directly penalize the effect of an input series on an output series. This approach would allow joint modeling of the full multivariate dynamics while simultaneously selecting for Granger causality between some inputs and some outputs.

Change point detection for interaction In Chapter 8, we propose an efficient ADMM algorithm to fit a structural change point model of stationary dynamics. Unfortunately, the proposed estimator will tend to overestimate the number of change points and a further pruning stage is necessary. Previous work maximizes an information criterion to prune the estimated change points [28]. This maximization may actually be solved more efficiently by dynamic programming. Thus, together, our efficient ADMM algorithm and a dynamic programming pruning stage will efficiently solve the full change point problem.

BIBLIOGRAPHY

[1] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[2] Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems*, 2016.

[3] Pierre-Olivier Amblard and Olivier JJ Michel. On directed information theory and Granger causality graphs. *Journal of computational neuroscience*, 30(1):7–16, 2011.

[4] Brian DO Anderson, Manfred Deistler, Elisabeth Felsenstein, Bernd Funovits, Lukas Koelbl, and Mohsen Zamani. Multivariate AR systems and mixed frequency data: G-identifiability and estimation. *Econometric Theory*, pages 1–34, 2015.

[5] H. Armstrong, C. Carter, K. Wong, and R. Kohn. Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing*, 19(3):303–316, 2009.

[6] M. Baback, E. Khan, K. M. Murphy, and B. M. Marlin. Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models. In *NIPS*, pages 1285–1293, 2009.

[7] F. R. Bach and M. I. Jordan. Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing*, 52(8):2189–2199, 2004.

[8] Mohammad Taha Bahadori and Yan Liu. Granger causality analysis in irregular time series. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 660–671. SIAM, 2012.

[9] Mohammad Taha Bahadori and Yan Liu. An examination of practical Granger causality inference. In *Proceedings of the 2013 SIAM International Conference on data Mining*, pages 467–475. SIAM, 2013.

[10] M. S. Bartlett. Smoothing periodograms from time series with continuous spectra. *Nature*, 161(4096):686–687, 1948.

[11] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

[12] Sumanta Basu, Ali Shojaie, and George Michailidis. Network Granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 2015.

[13] Eugene Belilovsky, Gaël Varoquaux, and Matthew B Blaschko. Testing for differences in Gaussian graphical models: applications to brain connectivity. In *Advances in Neural Information Processing Systems*, pages 595–603, 2016.

[14] Andre Berchtold. Estimation in the mixture transition distribution model. *Journal of Time Series Analysis*, 22(4):379–397, 2001.

[15] André Berchtold and Adrian E Raftery. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, pages 328–356, 2002.

[16] SA Billings and S Chen. The determination of multivariable nonlinear models for dynamic systems using neural networks. 1996.

[17] Stephen A Billings. *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons, 2013.

[18] John CG Boot, Walter Feibes, and Johannes Hubertus Cornelius Lisman. Further methods of derivation of quarterly figures from annual data. *Applied Statistics*, pages 65–75, 1967.

[19] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[20] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[21] James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer, 1986.

[22] Jörg Breitung and Norman R Swanson. Temporal aggregation and spurious instantaneous causality in multiple time series models. *Journal of Time Series Analysis*, 23(6):651–665, 2002.

[23] Steven L Bressler. Spectral methods in neural data analysis: Overview. In *Encyclopedia of Computational Neuroscience*, pages 89–92. Springer, 2015.

[24] D.R. Brillinger. *Time Series: Data Analysis an Theory*. Holden-Day, 2001.

[25] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, NY, 1991.

[26] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[27] C. M. Carvalho and J. G. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512, 2009.

[28] Ngai Hang Chan, Chun Yip Yau, and Rong-Mao Zhang. Group lasso for structural break time series. *Journal of the American Statistical Association*, 109(506):590–599, 2014.

[29] Ngai Hang Chan, Chun Yip Yau, and Rong-Mao Zhang. Lasso estimation of threshold autoregressive models. *Journal of Econometrics*, 189(2):285–296, 2015.

[30] Baoline Chen and Peter A Zadrozny. An extended Yule-Walker method for estimating a vector autoregressive model with mixed-frequency data. *Advances in Econometrics*, 13:47–74, 1998.

[31] Shizhe Chen, Ali Shojaie, and Daniela M Witten. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, pages 1–11, 2017.

[32] Wai-Ki Ching, Eric S Fung, and Michael K Ng. A multivariate Markov chain model for categorical data sequences and its applications in demand predictions. *IMA Journal of Management Mathematics*, 13(3):187–199, 2002.

[33] Ching Wai Jeremy Chiu, Bjørn Eraker, Andrew T Foerster, Tae Bong Kim, Hernán D Seoane, et al. Estimating vars sampled at mixed or irregular spaced frequencies: A Bayesian approach. Technical report, 2011.

[34] S Reynold Chu, Rahmat Shoureshi, and Manoel Tenorio. Neural networks for system identification. *IEEE Control systems magazine*, 10(3):31–35, 1990.

[35] K. Chung Wong, Z. Li, and A. Tewari. Lasso Guarantees for Time Series Estimation Under Subgaussian Tails and β-Mixing. *ArXiv e-prints*, February 2016.

[36] CMU. Carnegie Mellon University graphics lab motion capture database. Available at http://mocap.cs.cmu.edu/. 2009.

[37] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT Press, 2014.

[38] Jacques JF Commandeur and Siem Jan Koopman. *An introduction to state space time series analysis*. Oxford University Press, 2007.

[39] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[40] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172, 2000.

[41] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *JRSS(B)*, 76(2):373–397, 2014.

[42] David Danks and Sergey Plis. Learning causal structure from undersampled time series. 2013.

[43] Richard A Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096, 2016.

[44] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3):1272–1317, 1993.

[45] P. Dellaportas, P. Giudici, and G. Roberts. Bayesian inference for nondecomposable graphical Gaussian models. *Sankhy: The Indian Journal of Statistics*, 65(1):43–55, 2003.

[46] Mingzhou Ding, Yonghong Chen, and Steven L. Bressler. *Granger Causality: Basic Theory and Application to Neuroscience*, pages 437–460. Wiley, 2006.

[47] A. Dobra, C. Hans, B. Jones, J.R. Nevins, Joseph R., G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

[48] Finale Doshi, David Wingate, Josh Tenenbaum, and Nicholas Roy. Infinite dynamic Bayesian networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 913–920, 2011.

[49] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

[50] Grenier Y. Dupre la Tour, T. and A. Gramfort. Parametric estimation of spectrum driven by an exogenous signal. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, USA, Feb 2017.

[51] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*, volume 38. OUP Oxford, 2012.

[52] Michael Eichler. Fitting graphical interaction models to multivariate time series. *arXiv preprint arXiv:1206.6839*, 2012.

[53] Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. *arXiv preprint arXiv:1206.5246*, 2012.

[54] Bjørn Eraker, Ching Wai Chiu, Andrew T Foerster, Tae Bong Kim, and Hernán D Seoane. Bayesian mixed−frequency VARs. *Journal of Financial Econometrics*, 13(3):698–721, 2014.

[55] Jan Eriksson and Visa Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *Signal Processing Letters, IEEE*, 11(7):601–604, 2004.

[56] Roberto Esposito and Daniele P Radicioni. Carpediem: Optimizing the viterbi algorithm and applications to supervised sequential learning. *Journal of Machine Learning Research*, 10(Aug):1851–1880, 2009.

[57] J. Feng and N. Simon. Sparse-Input Neural Networks for High-dimensional Nonparametric Regression and Classification. *ArXiv e-prints*, 2017.

[58] Claudia Foroni and Massimiliano Marcellino. Mixed frequency structural vector autoregressive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):403–425, 2016.

[59] Claudia Foroni and Massimiliano Giuseppe Marcellino. A survey of econometric methods for mixed-frequency data. *Available at SSRN 2268912*, 2013.

[60] Emily Fox and David Dunson. Bayesian nonparametric covariance regression. *Arxiv preprint*, 2011.

[61] Emily Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 457–464, 2009.

[62] Emily B Fox, Michael C Hughes, Erik B Sudderth, and Michael I Jordan. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics*, 2014.

[63] André Fujita, Patricia Severino, João Ricardo Sato, and Satoru Miyano. Granger causality in systems biology: modeling gene networks in time series microarray data using vector autoregressive models. In *Brazilian Symposium on Bioinformatics*, pages 13–24. Springer, 2010.

[64] U. Gather, M. Imhoff, and R. Fried. Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, 21(18), 2002.

[65] Zoubin Ghahramani. Learning dynamic Bayesian networks. *Adaptive processing of sequences and data structures*, pages 168–197, 1998.

[66] Eric Ghysels. Macroeconomics and the reality of mixed−frequency data. *Journal of Econometrics*, 193(2):294–314, 2016.

[67] P. Giudici and P. J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

[68] Donald Goldfarb and Ashok Idnani. Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis*, pages 226–239. Springer, 1982.

[69] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1898–1906, 2015.

[70] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[71] Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

[72] Alex Graves. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012.

[73] P. J. Green and A. Thomas. Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91–110, 2013.

[74] E. C. Hall, G. Raskutti, and R. Willett. Inference of high-dimensional autoregressive generalized linear models. *ArXiv e-prints*, May 2016.

[75] Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

[76] Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *arXiv preprint arXiv:1307.0293*, 2013.

[77] Fang Han, Huasnran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150, 2015.

[78] Jeff Harrison and Mike West. *Bayesian forecasting & dynamic models*, volume 1030. Springer New York City, 1999.

[79] Andrew C Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, 1990.

[80] Campbell R Harvey and Akhtar Siddique. Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3):1263–1295, 2000.

[81] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*. CRC press, 2015.

[82] James Henderson and George Michailidis. Network reconstruction using nonparametric additive ode models. *PloS one*, 9(4):e94003, 2014.

[83] Helmut Herwartz and Martin Plödt. The macroeconomic effects of oil price shocks: Evidence from a statistical identification approach. *Journal of International Money and Finance*, 61:30–44, 2016.

[84] Eugene Hsu, Kari Pulli, and Jovan Popović. Style translation for human motion. In *ACM Transactions on Graphics (TOG)*. ACM, 2005.

[85] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 2011.

[86] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[87] Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. Causal discovery from subsampled time series data by constraint optimization. *arXiv preprint arXiv:1602.07970*, 2016.

[88] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[89] Aapo Hyvärinen, Shohei Shimizu, and Patrik O Hoyer. Causal modelling combining instantaneous and lagged effects: An identifiable model based on non-Gaussianity. In *Proceedings of the 25th international conference on Machine learning*, pages 424–431. ACM, 2008.

[90] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *The Journal of Machine Learning Research*, 11:1709–1731, 2010.

[91] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[92] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 2011.

[93] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400, 2005.

[94] Karl G Jöreskog. Testing structural equation models. *Sage focus editions*, 154:294–294, 1993.

[95] A. Jung, G. Hannak, and N. Görtz. Graphical LASSO based model selection for time series. *ArXiv e-prints*, 2014.

[96] Alejandro Justiniano and Giorgio E Primiceri. The time−varying volatility of macroeconomic fluctuations. *American Economic Review*, 98(3):604–41, 2008.

[97] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge University Press, 2004.

[98] A Karimi and Mark R Paul. Extensive chaos in the Lorenz-96 model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2010.

[99] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[100] Benjamin Kedem and Konstantinos Fokianos. Regression models for categorical time series. *Regression Models for Time Series Analysis*, pages 89–137, 2005.

[101] Lutz Kilian and Helmut Lütkepohl. *Structural vector autoregressive analysis*. Cambridge University Press, 2017.

[102] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput Biol*, 7(3):e1001110, 2011.

[103] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-task regression with structured sparsity. 2010.

[104] Özgür Kişi. River flow modeling using artificial neural networks. *Journal of Hydrologic Engineering*, 9(1):60–63, 2004.

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

[106] Gary Koop, Dimitris Korobilis, et al. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, 3(4):267–358, 2010.

[107] Dimitris Korobilis. Var forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230, 2013.

[108] Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber. A clockwork RNN. In *International Conference on Machine Learning*, 2014.

[109] R. T. Krafty, O. Rosen, D. S. Stoffer, D. J. Buysse, and M. H. Hall. Conditional spectral analysis of replicated multiple time series with application to nocturnal physiology. *ArXiv e-prints*, 2015.

[110] Markku Lanne and Helmut Lütkepohl. Structural vector autoregressions with non-normal residuals. *Journal of Business & Economic Statistics*, 28(1):159–168, 2010.

[111] Markku Lanne, Helmut Lütkepohl, and Katarzyna Maciejowska. Structural vector autoregressions with Markov switching. *Journal of Economic Dynamics and Control*, 34(2):121–131, 2010.

[112] Markku Lanne, Mika Meitz, Pentti Saikkonen, et al. Identification and estimation of non-Gaussian structural vector autoregressions. *CREATES, Arhus Univerisity, Technical report*, 2015.

[113] Markku Lanne and Saikkonen Pentti. Modeling conditional skewness in stock returns. *The European Journal of Finance*, 13(8):691–704, 2007.

[114] E. Larson and A.K.C. Lee. Switching auditory attention using spatial and non-spatial features recruits different cortical networks. *NeuroImage*, 84:681–687, 2014.

[115] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[116] Sophie Lèbre. Inferring dynamic genetic networks with low order independencies. *Statistical applications in genetics and molecular biology*, 2009.

[117] Sophie Lèbre and Pierre-Yves Bourguignon. An EM algorithm for estimation in the mixture transition distribution model. *Journal of Statistical Computation and Simulation*, 78(8):713–729, 2008.

[118] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Graph convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[119] Yehua Li and Marc G Genton. Single-index additive vector autoregressive time series models. *Scandinavian Journal of Statistics*, 36(3):369–388, 2009.

[120] M. Lichman. UCI machine learning repository, 2013.

[121] R. Liégeois, B. Mishra, M. Zorzi, and R. Sepulchre. Sparse plus low-rank autoregressive identification in neuroimaging time series. *ArXiv e-prints*, March 2015.

[122] Néhémy Lim, Cédric Auliac, George Michailidis, et al. Operator-valued kernel-based vector autoregressive models for network inference. *Machine learning*, 99(3):489, 2015.

[123] Néhémy Lim, Florence dAlché Buc, Cédric Auliac, and George Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine learning*, 2015.

[124] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014.

[125] Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–462, 1976.

[126] C. Louizos, K. Ullrich, and M. Welling. Bayesian Compression for Deep Learning. *ArXiv e-prints*, 2017.

[127] Aurélie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 2009.

[128] Aurelie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical Granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[129] Bethany Lusch, Pedro D. Maia, and J. Nathan Kutz. Inferring connectivity in networked dynamical systems: Challenges using Granger causality. *Phys. Rev. E*, 2016.

[130] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[131] Y.-A. Ma, T. Chen, L. Wu, and E. B. Fox. A Unifying Framework for Devising Efficient and Irreversible MCMC Samplers. *ArXiv e-prints*, August 2016.

[132] David Madigan and Adrian E Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

[133] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel-Granger causality and the analysis of dynamical networks. *Physical review E*, 2008.

[134] Y. Matsuda. A test statistic for graphical modelling of multivariate time series. *Biometrika*, 93(2):pp. 399–409, 2006.

[135] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[136] T. Medkour, A. T. Walden, Burgess A. P., and Strelets V. B. Brain connectivity in positive and negative syndrome schizophrenia. *Neuroscience*, 169(4):1779 – 1788, 2010.

[137] Filippo Moauro and Giovanni Savio. Temporal disaggregation using multivariate structural time series models. *The Econometrics Journal*, 8(2):214–234, 2005.

[138] A. Mohammadi and E. C. Wit. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

[139] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[140] W. B. Nicholson, J. Bien, and D. S. Matteson. Hierarchical vector autoregression. *ArXiv e-prints*, December 2014.

[141] João Nicolau. A new model for multivariate Markov chains. *Scandinavian Journal of Statistics*, 41(4):1124–1135, 2014.

[142] A. O'Hagan. Fractional Bayes factors for model comparison. *JRSS(B)*, 57(1):99–138, 1995.

[143] Hernando Ombao, Rainer Von Sachs, and Wensheng Guo. Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531, 2005.

[144] Neal Parikh and Stephen P Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[145] Vladimir Pavlovic, James M Rehg, and John MacCormick. Learning switching linear models of human motion. In *Advances in neural information processing systems*, 2001.

[146] Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

[147] Donald B Percival and Andrew T Walden. *Spectral analysis for physical applications*. Cambridge University Press, 1993.

[148] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.

[149] Mert Pilanci, Laurent E Ghaoui, and Venkat Chandrasekaran. Recovery of sparse probability measures via convex programming. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2012.

[150] Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. Rate-agnostic (causal) structure learning. In *Advances in Neural Information Processing Systems*, pages 3285–3293, 2015.

[151] P. Preuß, R. Puchstein, and H. Dette. Detection of multiple structural breaks in multivariate time series. *ArXiv e-prints*, September 2013.

[152] Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PloS one*, 2010.

[153] Huitong Qiu, Sheng Xu, Fang Han, Han Liu, and Brian Caffo. Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1843–1851, 2015.

[154] Lawrence Rabiner and B Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[155] Svetlozar Todorov Rachev. *Handbook of heavy tailed distributions in finance: Handbooks in finance*, volume 1. Elsevier, 2003.

[156] Daniele P Radicioni and Roberto Esposito. Breve: An hmperceptron-based chord recognition system. In *Advances in Music Information Retrieval*, pages 143–164. Springer, 2010.

[157] Adrian E Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 528–539, 1985.

[158] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Multistep Neural Networks for Data-driven Discovery of Nonlinear Dynamical Systems. *ArXiv e-prints*, January 2018.

[159] You Ren, Emily B Fox, and Andrew Bruce. Achieving a hyperlocal housing price index: Overcoming data sparsity by Bayesian dynamical modeling of multiple data streams. *arXiv preprint arXiv:1505.01164*, 2015.

[160] Patricia Reynaud-Bouret, Vincent Rivoirard, and Christine Tuleau-Malot. Inference of functional connectivity in neurosciences via hawkes processes. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 317–320. IEEE, 2013.

[161] O. Rosen and D. S. Stoffer. Automatic estimation of multivariate spectra via smoothing splines. *Biometrika*, 94(2):335–345, 2007.

[162] A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Stat*, 29(3):391–411, 2002.

[163] Jakob Runge, Jobst Heitzig, Vladimir Petoukhov, and Jürgen Kurths. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Physical review letters*, 108(25):258701, 2012.

[164] A. Safikhani and A. Shojaie. Structural Break Detection in High-Dimensional Non-Stationary VAR models. *ArXiv e-prints*, August 2017.

[165] Ruslan Salakhutdinov and Sam Roweis. Adaptive overrelaxed bound optimization methods. In *ICML*, pages 664–671, 2003.

[166] Koichi Sameshima and Luiz Antonio Baccala. *Methods in brain connectivity inference through multivariate time series analysis*. CRC press, 2016.

[167] A. Sarkar and D. B. Dunson. Bayesian Nonparametric Modeling of Higher Order Markov Chains. *ArXiv e-prints*, June 2015.

[168] Jeffrey D Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

[169] Frank Schorfheide and Dongho Song. Real-time forecasting with a mixed-frequency var. *Journal of Business & Economic Statistics*, 33(3):366–380, 2015.

[170] J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136:2144–2162, 2006.

[171] J. G. Scott and C. M. Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *J. Comp. Graph. Stat.*, 17(4):790–808, 2008.

[172] Byeongchan Seong. Cointegration analysis with mixed−frequency data of quarterly GDP and monthly coincident indicators. *Korean Journal of Applied Statistics*, 25(6):925–932, 2012.

[173] William Sharpe, Gordon J Alexander, and Jeffrey W Bailey. Investments. 1998.

[174] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[175] Ali Shojaie and George Michailidis. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

[176] Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

[177] Andrea Silvestrini and David Veredas. Temporal aggregation of univariate and multivariate time series models: A survey. *Journal of Economic Surveys*, 22(3):458–497, 2008.

[178] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 2013.

[179] Vikas Sindhwani, Ha Quang Minh, and Aurélie C. Lozano. Scalable matrix-valued kernel learning for high-dimensional nonlinear multivariate regression and Granger causality. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

[180] M. R. Siracusa and J. W. Fisher III. Tractable Bayesian inference of time-series dependence structure. In *AISTATS*, 2009.

[181] J. Songsiri and L. Vandenberghe. Topology selection in graphical models of autoregressive processes. *JMLR*, 11:2671–2705, 2010.

[182] Olaf Sporns. *Networks of the Brain*. MIT press, 2010.

[183] Daniel O Stram and William WS Wei. A methodological note on the disaggregation of time series totals. *Journal of Time Series Analysis*, 7(4):293–302, 1986.

[184] Ilya Sutskever. *Training recurrent neural networks*. University of Toronto Toronto, Ontario, Canada, 2013.

[185] Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela M Witten. Learning graphical models with hubs. *Journal of Machine Learning Research*, 15(1):3297–3331, 2014.

[186] A. Tank, E. B. Fox, and A. Shojaie. Granger causality networks for categorical time series. *ArXiv e-prints*, June 2017.

[187] Y. Tao, L. Ma, W. Zhang, J. Liu, W. Liu, and Q. Du. Hierarchical Attention-Based Recurrent Highway Networks for Time Series Prediction. *ArXiv e-prints*, June 2018.

[188] Timo Terasvirta, Dag Tjostheim, Clive WJ Granger, et al. Modelling nonlinear economic time series. *OUP Catalogue*, 2010.

[189] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[190] Howell Tong. Nonlinear time series analysis. In *International Encyclopedia of Statistical Science*. Springer, 2011.

[191] Howell Tong and Keng S Lim. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 245–292, 1980.

[192] Ruey S Tsay. Testing and modeling multivariate threshold models. *Journal of the American Statistical Asssociation*, 93(443):1188–1202, 1998.

[193] BA Turlach and A Weingessel. quadprog R package. available online, 2013.

[194] Jodie B Ullman and Peter M Bentler. *Structural equation modeling*. Wiley Online Library, 2003.

[195] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[196] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropya model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

[197] W David Walls. Modelling heavy tails and skewness in film returns. *Applied Financial Economics*, 15(17):1181–1188, 2005.

[198] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 2008.

[199] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[200] Halbert White, Tae-Hwan Kim, and Simone Manganelli. Var for var: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics*, 187(1):169–188, 2015.

[201] P. Whittle. The analysis of multiple stationary time series. *JRSS(B)*, 15(1):125–139, 1953.

[202] Ronald J Williams and David Zipser. Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation: Theory, architectures, and applications*, 1:433–486, 1995.

[203] R. J. Wolstenholme and A. T. Walden. An efficient approach to graphical modeling of time series. *ArXiv e-prints*, 2015.

[204] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *ArXiv e-prints*, March 2014.

[205] Ying Yang, Elissa Aminoff, Michael Tarr, and Kass E Robert. A state-space model of cross-region dynamic connectivity in meg/eeg. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1234–1242. Curran Associates, Inc., 2016.

[206] R. Yu, S. Zheng, A. Anandkumar, and Y. Yue. Long-term Forecasting using Tensor-Train RNNs. *ArXiv e-prints*, October 2017.

[207] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[208] G Udny Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226:267–298, 1927.

[209] Peter A Zadrozny. *Estimating a multivariate ARMA model with mixed frequency data: an application to forecasting US GNP at monthly intervals*, volume 90. Federal Reserve Bank of Atlanta, 1990.

[210] Peter A Zadrozny. Extended Yule−Walker identification of VARMA models with single or mixed−frequency data. *Journal of Econometrics*, 193(2):438–446, 2016.

[211] G Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, 2003.

[212] Kun Zhang and Aapo Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Machine learning and knowledge discovery in databases*, pages 570–585. Springer, 2009.

[213] Douglas Zhou, Yaoyu Zhang, Yanyang Xiao, and David Cai. Analysis of sampling artifacts on the Granger causality analysis for topology extraction of neuronal dynamics. *Frontiers in computational neuroscience*, 8, 2014.

[214] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 641–649, 2013.

[215] Dong-Mei Zhu and Wai-Ki Ching. A new estimation method for multivariate Markov chain model with application in demand predictions. In *Business Intelligence and Financial Engineering (BIFE), 2010 Third International Conference on*, pages 126–130. IEEE, 2010.

[216] H. Zhu, N. Strawn, and D. B. Dunson. Bayesian graphical models for multivariate functional data. *ArXiv e-prints*, 2014.

[217] B. D. Ziebart, A. K. Dey, and J. A. Bagnell. Learning selectively conditioned forest structures with applications to dbns and classification. *UAI*, 2007.