Modeling and Estimation of High-dimensional Vector Autoregressions

by

Sumanta Basu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2014

Doctoral Committee:

Professor George Michailidis, Chair
Professor Susan Murphy
Assistant Professor Ali Shojaie
Professor Roman Vershynin
Assistant Professor Shuheng Zhou

in memory of Kalyan Kumar Bhattacharyya

ACKNOWLEDGEMENTS

I would like to thank Professor George Michailidis for his guidance, encouragement and continued support during the preparation of this thesis. I would also like to thank the members of my doctoral committee for their helpful comments and suggestions. Finally, I would like to thank the Department of Statistics at University of Michigan for giving me the opportunity to pursue my doctoral degree in such a wonderful academic environment.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Modeling and Estimation of High-dimensional Vector Autoregressions

by

Sumanta Basu

Chair: George Michailidis

Vector Autoregression (VAR) represents a popular class of time series models in applied macroeconomics and finance, widely used for structural analysis and simultaneous forecasting of a number of temporally observed variables. Over the years it has gained popularity in the fields of control theory, statistics, economics, finance, genetics and neuroscience. In addition to the "curse of dimensionality" introduced by a quadratically growing dimension of the parameter space, VAR estimation poses considerable challenges due to the temporal and cross-sectional dependence in the data.

In the first part of this thesis, we discuss modeling and estimation of high-dimensional VAR from short panels of time series, with applications to reconstruction of gene regulatory network from time course gene expression data. We investigate adaptively thresholded lasso regularized estimation of VAR models and propose a thesholded group lasso regularization framework to incorporate *a priori* available pathway information in the model. The properties of the proposed methods are assessed both theoretically and via numerical experiments. The study is illustrated on

two motivating examples coming from functional genomics and financial econometrics.

The second part of this thesis focuses on modeling and estimation of high-dimensional VAR in the traditional time series setting, where one observes a single replicate of a long, stationary time series. We investigate the theoretical properties of ℓ_1-regularized and thresholded estimators in high-dimensional VAR, stochastic regression and covariance estimation problems in a non-asymptotic framework. We establish consistency of the estimators under high-dimensional scaling and propose a measure of stability that provides insight into the effect of temporal and cross-sectional dependence on the accuracy of the regularized estimates. We also propose a low-rank plus sparse modeling strategy of high-dimensional VAR in the presence of latent variables. We study the theoretical properties of the proposed estimator in a non-asymptotic framework, establish its estimation consistency under high-dimensional scaling and compare its performance with existing methods via extensive simulation studies.

CHAPTER I

Introduction

Recent advances in information technology have made high-dimensional time series datasets increasingly common in many biomedical and economic applications. Examples include structural analysis and forecasting of many macroeconomic variables (*Stock and Watson*, 2006; *Bańbura et al.*, 2010), large volatility matrix estimation in asset pricing (*Fan et al.*, 2011), reconstruction of regulatory network from time course gene expression data (*Michailidis and d'Alché Buc*, 2013) and discovering functional and effective connectivity amongst brain regions from fMRI data (*Smith*, 2012).

Despite inherent difference in the focus of these problems together with unique statistical and computational challenges from a modeling perspective, a central underlying theme is to understand the interactions among the components of a large dynamic system from temporal datasets.

This thesis focuses on developing rigorous and computationally efficient modeling strategies for vector autoregressive models (VAR) from high-dimensional datasets. Vector autoregression refers to a popular class of models in economics and control theory, commonly used for studying complex interrelationships among the components of a multivariate time series. In the next few sections, we provide a brief description of VAR models in the statistics and economics literature and their recent applica-

bility in the fields of genomics and neuroscience, outline key technical challenges in high-dimensional settings and summarize our contributions to the existing literature. We conclude this chapter with an outline of subsequent chapters.

1.1 A Short Overview of Vector Autoregressions (VAR)

Autoregressive modeling of multivariate stationary processes originated in control theory, where vector-valued autoregressive moving average (VARMA) and state-space representations were used as canonical tools for identification of linear dynamic systems (*Kumar and Varaiya*, 1986; *Hannan and Deistler*, 2012). Some authors advocated the use of higher-order VAR over more general VARMA models (*Lütkepohl*, 2005) due to numerous identification issues of the latter model class. A strong theoretical justification of such a modeling strategy comes from the famous Wold decomposition theorem, which ensures that a large class of stationary processes can be represented as potentially infinite order VAR processes (*Fournier et al.*, 2006).

VAR models gained popularity in the economics literature following the seminal works of Granger and Sims. *Granger* (1969b) proposed the notion of Granger causality, a statistical framework for determining whether a time series X^t is useful in forecasting another one Y^t. Sims proposed VAR models as a theory-free method for estimating economic relationships (*Sims*, 1980). Since then VAR models have been widely used for testing Granger-causal relationships among macroeconomic variables, including government spending and taxes on economic output (*Blanchard and Perotti*, 2002), stock price and volume (*Hiemstra and Jones*, 1994).

Formally, for a p-dimensional stochastic process $X^t = (X_1^t, \ldots, X_p^t)$, a finite-order VAR model of order d, often denoted as VAR(d), takes the form

$$X^t = A_1 X^{t-1} + A_2 X^{t-2} + \ldots + A_d X^{t-d} + \epsilon^t,\ \mathbb{E}(\epsilon^t) = 0, \mathrm{Var}(\epsilon^t) = \Sigma_\epsilon \tag{1.1}$$

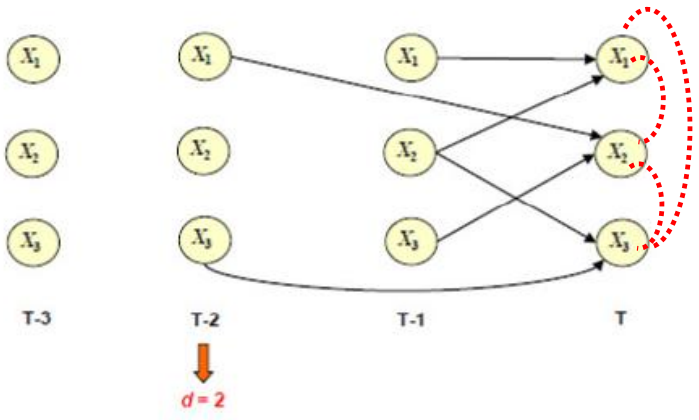


Figure 1.1: Graphical representation of the VAR model (4.3): directed edges (solid) correspond to the entries of the transition matrices, undirected edges (dashed) correspond to the entries of Σ_ϵ^{-1}

where $A_1, \ldots, A_d$ are $p \times p$ matrices and $\{\epsilon^t\}$ is white noise process. The matrices $A_1, \ldots, A_d$, commonly referred to as *transition* matrices, capture temporal relationship among the individual system components, while the error covariance matrix Σ_ϵ (or the precision matrix Σ_ϵ^{-1}) captures additional conemporaneous dependence among them. For structure learning and forecasting problems, one is primarily interested in estimating the transition matrices, although incorporating information about the contemporaneous dependence often results in improved estimation and prediction accuracy.

VAR models provide a natural interpretation as a directed network of interactions among the individual time series, as illustrated in Figure 1.1. The network of directed edges, where the edge weights are represented by the entries of $A_1, \ldots, A_d$, is often referred to as a Granger-causal network and the transition matrices are referred to as the adjaceny matrices.

The macroeconomic applications described above involve learning Granger causal networks in a classical time series setup, where the data consist of a single, long, stationary snapshot of the vector-valued process $\{X^1, \ldots, X^T;\ T \text{ large}\}$. Another important line of research in microeconomics considers learning Granger causal networks among several economic variables from temporal panel data, where one observes

a panel of subjects (individuals, firms, households etc.) over a short period of time (*Cao and Sun*, 2011; *Binder et al.*, 2005).

In the last ten years, the Granger causal framework and VAR modeling have also found diverse applications in biological sciences. An important example is the reconstruction of regulatory networks from time course gene expression data, a canonical problem in functional genomics. The Granger-causal network of interactions among multiple genes are obtained via VAR modeling of short panels, since the time course data generally consist of short time series (typically $5 - 20$ time points), but one has access to replicates from different biological samples or patients (*Michailidis and d'Alché Buc*, 2013).

Another motivating example comes from neuroscience, where the main interest is in finding dynamic connectivity measures among different regions of human brain from time-course fMRI data. Despite increasing use of Granger causality and VAR modeling in the neuroscience literature, their sensitivity to latency difference in Haemodynamic Response Function (HRF) and platform specific issues like downsampling remain unclear (*Seth et al.*, 2013). In this thesis, we do not consider applications in neuroscience and mention them as interesting future research directions.

1.2 High-dimensional VAR: Challenges and Current Work

The problem of high-dimensionality occurs when the ambient dimension of the model parameter space exceeds the available sample size. VAR models are intrinsically high-dimensional due to a quadratically growing parameter space. For instance, fitting a VAR(4) model for $p = 10$ time series requires estimating $dp^2 = 400$ parameters (excluding estimation of the error covariance Σ_ϵ). However, such large number of stationary observations are seldom available in practice. In classical time series settting (T large, $n = 1$), dependence among the observations further reduces the effective sample size. Moreover, recent applications in macoecnomics and genetics

require analysis of hundreds of time series or genes. As a result, consistent estimation and prediction is not possible without making some low-dimensional structural assumption on the underlying model.

In Bayesian econometrics, researchers used several sparsity and structure inducing priors to deal with this curse of dimensionality. Examples include gaussian, double-exponential and Minnesota priors (*Litterman*, 1986; *De Mol et al.*, 2008). The Gaussian and double-exponential priors are related to ridge and lasso penalized regression in the frequentist framework. More recently, the problem of high-dimensional VAR estimation in the time series context has been addressed in the statistics literature by several authors. For instance, *Song and Bickel* (2011) and *Negahban and Wainwright* (2011) proposed lasso, group lasso and nuclear norm penalized estimation procedures to encourage sparsity and structural pattern in the underlying network models. *Davis et al.* (2012) proposed a regularized log-likelihood and two-stage estimation procedure for encouraging sparsity in the model. In the genetic applications of short panel VAR, *Lozano et al.* (2009a); *Fujita et al.* (2007a) and *Shojaie and Michailidis* (2010b) proposed several modeling strategies for estimating sparse Granger causal networks from time course gene expression data.

Despite its long history and wide applicability in many important problems, considerable challenges and interesting questions remain in the statistical analysis of high-dimensional VAR. First, due to the large dimensionality of time series datasets in modern applications, dimension reduction via the mere assumption of sparsity is often not adequate. In many applications, external information available to practitioners can help reduce dimensionality in a meaningful way. In macroeconomic applications, failure to correct for hidden factors can result in a non-sparse network of interaction among the variables. Second, theoretical analysis of regularized estimates commonly used to fit high-dimensional VAR models is yet incomplete. As we describe in Chapter IV, the results of *Song and Bickel* (2011) and *Negahban and*

Wainwright (2011) rely on stringent assumptions which do not hold beyond a small class of stationary VAR(1) models. Lastly, with growing dimension of the datasets, the computational complexity of many of these methods increase dramatically. It is important to come up with scaleable algorithms, often achieved via distributed and parallel implementation, for analyzing large VAR.

1.3 Contribution of this work

This thesis makes several contributions to the existing literature of high-dimensional VAR models.

On the modeling front, we adopt the framework of regularized estimation with convex penalties *inducing low-dimensional structure* on the model space. By and large, all the methods proposed in this thesis can be viewed as variants of an M-estimator of the following form

$$\underset{A_1,\ldots,A_p}{\operatorname{argmin}} \mathcal{L}(X^t, \sum_{i=1}^{d} A_i X^{t-i}) + \mathcal{P}(A_1, \ldots, A_p) \tag{1.2}$$

where $\mathcal{L}(.,.)$ is a loss function (least squares or negative log-likelihood) and $\mathcal{P}(.)$ is a convex penalty encouraging structured sparsity in the solution (lasso, group lasso, nuclear norm or some combination of these). Regularized regression with convex penalties is popular in the statistics and machine learning community for providing a flexible and computationally efficient framework to incorporate in the model external information *a priori* available to practitioners. In the context of structure learning and prediction, we show that efficient use of appropriate penalties can achieve similar goals and reduce effective dimensionality of the problem, leading to more accurate estimation and forecasting strategies. In Chapter III, we show that a thresholded variant of group lasso regularization can incorporate pathway membership of individual genes towards accurate identification of gene regulatory networks. A salient

feature of our proposed method is that it can handle moderate misspecification in the a priori available knowledge. In Chapter V, we show that a low-rank + sparse modeling of VAR(1) model can effectively correct for hidden latent factors in the reconstruction of Granger causal networks from time course data.

On the theoretical front, we present a rigorous, non-asymptotic analysis of high-dimensional VAR estimation problems under the above regularization methods, both in the context of classical time series and short panels. In the past decade, a significant amount of research has been conducted on the theoretical properties of regularized estimation in the regression context. However, most of these analyses crucially rely on the availability of independent and identically distributed samples and hence do not directly apply in the time series context. A key challenge in the theoretical analysis of high-dimensional time series is to capture the effect of temporal and cross-sectional dependence present in the data. To this end, in Chapter IV we develop a novel measure of stability for stationary processes, based on their spectral representation. We derive non-asymptotic upper bounds on the estimation errors of the proposed regularized estimates in the time series context, ensuring consistent estimation under high-dimensional scaling. We show that the proposed measure of stability provides insight into how the dependence present in the data affects the accuracy of these estimates. Our proposed measure of stability is fundamental to the nature of multivariate stationary processes and provides meaningful results in the context of other important problems in high-dimensional time series including stochastic regression and covariance estimation.

For all the methods proposed in this thesis, we discuss computationally efficient implementation strategies. The thresholded group lasso presented in Chapter III relies on hard-thresholding and is computationally efficient than the competing methods for bi-level selection. In Chapter IV we develop a block coordinate descent algorithm for analyzing regularized likelihood based VAR estimates, that is amenable to par-

allel implementation and scales easily for large datasets than the original, sequential version suggested in *Davis et al.* (2012). We also demonstrate the advantages of the proposed methods over competing methods via numerical experiments and applications on real data.

1.4 Organization of Thesis

This thesis consists of two main parts, each including two chapters. In the first part, we discuss modeling and estimation of high-dimensional VAR from short panels of time series, with applications to reconstruction of gene regulatory network from time course gene expression data. In Chapter II, we propose an adaptively thresholded estimation of Granger causal effects obtained from the lasso penalization method. We establish the asymptotic properties of the proposed technique, and discuss the advantages it offers over competing methods, such as the truncating lasso. Its performance and that of its competitors is assessed on a number of simulated settings and it is applied on a data set that captures the activation of T-cells. In Chapter III, we extend the above method to incorporate *a priori* available grouping structure on the individual time series. To that end, we introduce a group lasso regression regularization framework, and also examine a thresholded variant to address the issue of group misspecification. Further, the norm consistency and variable selection consistency of the estimates are established, the latter under the novel concept of direction consistency. The performance of the proposed methodology is assessed through an extensive set of simulation studies and comparisons with existing techniques. The study is illustrated on two motivating examples coming from functional genomics and financial econometrics.

The second part, consisting of Chapters IV and V, focuses on VAR models in the traditional time series settings, where we observe a single, long, stationary realization of the multiple time series. In Chapter IV, we investigate the theoretical properties

of ℓ_1-regularized VAR estimates along with two other important statistical problems in the context of high-dimensional time series - stochastic regression with serially correlated errors and sparse covariance matrix estimation from temporal data. For all three problems, we derive non-asymptotic upper bounds on the estimation errors, thus establishing that consistent estimation is possible via ℓ_1-regularization and thresholding for a large class of stationary time series under sparsity constraints. In Chapter V, we consider the problem of estimating high-dimensional VAR models in the presence of unobserved latent factors. We propose a low-rank plus sparse modeling of the transition matrix of a VAR(1) model and show that a regularized estimator based on an infimal convolution of nuclear norm and ℓ_1 norm can approximate the low-rank and the sparse components with high accuracy. Using the techniques developed in Chapter IV, we establish non-asymptotic upper bound on the approximation error and show that consistent estimation is possible under high-dimensional scaling.

CHAPTER II

Adaptive Thresholding for Reconstructing Regulatory Networks from Time Course Gene Expression Data

2.1 Introduction

Reconstructing gene regulatory networks is a critical problem in systems biology. Gene regulation is carried out by binding of protein products of transcription factors (TF) to cis-regulatory elements of genes, which results in change of expression levels of the regulated genes. Such relationships are often represented in the form of directed graphs with transcription factors (TF) regulating target genes. This interpretation of effects of transcription factors on regulated genes, as a physical intervention mechanism therefore implies that regulatory interactions among genes are by definition causal.

In the theory of graphical models, causal relationships among random variables are modeled using directed (acyclic) graphs, where an edge among two random variables indicates a direct causal effect. Statistical methods based on observational data can only determine associations among random variables and causal discovery requires additional assumptions and/or information about the underlying system. This implies that, reconstructing gene regulatory networks may be only feasible through

carefully designed perturbation experiments. Such experiments are often expensive and only possible in case of model organisms and cell lines. However, regulatory mechanisms become evident if the expression level of gene Y is affected by changes in expression levels of gene X. Time course gene expression data provide a dynamic view of expression levels of all the genes under study, and therefore, can provide cues to the causal relationships among genes, which can be used to reconstruct the gene regulatory network.

Two of the most popular approaches for inferring gene regulatory networks using time course gene expression data are dynamic Bayesian Networks, *Murphy* (2002) and Granger causality, *Granger* (1969a). Dynamic Bayesian Networks (DBNs), generalize the notion of Bayesian networks to allow for cycles in the graph, through expanding the state space of the model by replicating the variables in the network over time points. Cyclic networks are then transformed to directed acyclic graphs (DAGs) by breaking down cycles into interactions between variables at two different time points. *Ong et al.* (2002) and *Perrin et al.* (2003) discuss applications of DBNs for inferring regulatory networks from time course gene expression data.

On the other hand, Granger causality is motivated by a practical interpretation of predictability among random variables. In particular, given two random variables X and Y, if the autoregressive model of Y based on past values of both variables significantly outperforms the model based on Y alone, X is said to be Granger-causal for Y. In the context of gene expression analysis, this definition implies that changes in expression levels of Y could be explained by expression levels of X from previous time points. Exploring Granger causal relationships is closely related to analysis of vector autoregressive (VAR) models. Therefore, while applying DBNs to high-dimensional applications may be computationally prohibitive, statistical methods can be used to derive Granger causal relationships among genes from time-course gene expression data using standard techniques for analysis of VAR models (see *Yamaguchi et al.*

(2007); *Opgen-Rhein and Strimmer* (2007) for examples of such approaches).

Unlike the original application area of Granger causality in econometrics, in gene regulatory network applications, the number of available samples is often small compared to the number of genes in the study. As a result, sparse VAR models have been explored by a number of researchers, including *Fujita et al.* (2007a) and *Mukhopadhyay and Chatterjee* (2007), to obtain reliable estimates of gene regulatory networks when the number of genes, p is large compared to the sample size, n.

Penalized estimation methods provide sparse estimates of high dimensional statistical models. *Arnold et al.* (2007) use the lasso (or ℓ_1) penalty to discover the structure of graphical models based on the concept of Granger causality in a financial setting. More recently, a similar framework, using the group lasso penalty was used by *Lozano et al.* (2009a) to group the effect of observations of each variable over past time points.

A main challenge in applying both DBN and Granger causality models to discover gene regulatory networks is that as the number of time points increases, the number of variables used in the replicated representation of the network also increases. As a result, many available methodologies simply ignore possible effects of genes on each other from time points far in the past, resulting in possible loss of information. To overcome this challenge, *Shojaie and Michailidis* (2010a) proposed to simultaneously estimate the order of the vector auto-regressive model, as well as the interactions among variables using a non-convex penalty, called the truncating lasso penalty, and showed that when the effects of variables on each other decay over time, the proposed penalty consistently estimates the order of the time series, as well as the structure of the regulatory network in high dimensional sparse settings.

The decay condition in *Shojaie and Michailidis* (2010a) (referred to as S-M henceforth) is a natural assumption in many time series models. However, when this condition is not satisfied, the truncating lasso penalty may fail to correctly estimate

the order of the time series. In this study, we discuss examples where the decay assumption of S-M may fail to hold, and propose a new estimator, based on adaptive thresholding of lasso estimates, which can be used to simultaneously estimates the order of the VAR model and the structure of the network. The new estimator is based on the assumption that if the true VAR model includes non-ignorable effects at any given time point, the number of edges in the network should exceed a certain threshold. We formally state this assumption in Section 2.3, where we also investigate the effect of violations of this assumption on false positive and false negative errors.

The remainder of the chapter is organized as follows. In Section 2.2, we review some background material and present the new methodology and discuss its asymptotic properties. Section 2.4 includes a comparative analysis of the performance of the proposed estimator over a set of simulation studies, whereas applications to time-course gene expression data from T-cell activation are presented in Section 2.5. Section 2.6 discusses some final remarks on the choice of appropriate penalty, and methods for evaluating the validity of underlying structural assumption.

2.2 Estimation of Regulatory Networks from Time Course Gene Expression Data

We start this section by a brief introduction of two classes of statistical models for analysis of genetic networks using time series observations, namely dynamic Bayesian Networks (DBN) and graphical Granger causality. We then discuss penalized methods for estimation of gene regulatory networks and introduce our new estimator based on an adaptively thresholded lasso penalty. Computational issues and asymptotic properties of the proposed estimator are discussed at the end of the section.

2.2.1 Dynamic Bayesian Nework and Network Granger Causality

Bayesian networks models (BN) correspond to probability distributions over a directed acyclic graph (DAG). More specifically, let $\mathcal{G} = (V, E)$, denote a DAG with the node set V and the edge set $E \subset V \times V$. Denote the random variables on the nodes of the graph by $X_1, \ldots, X_p$, where $p = |V|$ is the cardinality of the set V. For a DAG $\mathcal{G}$, it is clear that if $(i, j) \in E \Rightarrow (j, i) \notin E$. We represent E through the adjacency matrix A of the graph, a $p \times p$ matrix whose $(j, i)-$th entry indicates whether there is an edge (and its weight) from node j to node i. We represent an edge from j to i by $j \rightarrow i$, and denote by $\mathrm{pa_i}$ the set of parents of node i.

A probability distribution $\mathcal{P}$ is said to be (Markov) compatible with $\mathcal{G}$ if it admits the following decomposition based on the set of parents of each node in the graph (*Pearl* (2000a)):

$$\mathcal{P}(X_1, \ldots, X_p) = \Pi_{i \in V} \mathcal{P}(X_i | \mathrm{pa_i}). \tag{2.1}$$

Pearl (2000a) shows that if $\mathcal{P}$ is strictly positive, the Bayesian network $\mathcal{G}$ associated with $\mathcal{P}$ is unique and $\mathcal{P}$ and $\mathcal{G}$ are compatible. This implies that the joint Gaussian distributions defined according to (2.1) on nodes of $\mathcal{G}$ are uniquely defined and Markov compatible with $\mathcal{G}$. Markov compatible probability distributions on DAGs can be defined using structural equation models, where each variable is modeled as a (nonlinear) function of its parents. Given latent variables $Z_i, i = 1, \ldots, p$ for each node i, the general form of these models is given by:

$$X_i = f_i(\mathrm{pa_i}, \mathrm{Z_i}), \quad \mathrm{i} = 1, \ldots, \mathrm{p} \tag{2.2}$$

In (2.2), the latent variables represent the unexplained variation in each node, which is independent of the effect of its parents. For Gaussian random variables, the function f_i is linear, in the sense that it corresponds to the linear regression of X_i on the set of

its parents pa_i. In other words, for Gaussian random variables (2.2) takes the form:

$$X_i = \sum_{j \in \text{pa}_\text{i}} \rho_{ij} X_j + Z_i, \quad i = 1, \ldots, p \tag{2.3}$$

where ρ_{ij} represent the effect of gene j on i for $j \in \text{pa}_\text{i}$ and ρ_{ij} are the coefficients of the linear regression model of X_i on $X_j, j \in \text{pa}_\text{i}$. Note that in this case $\rho_{ij} = 0$ whenever $j \notin \text{pa}_\text{i}$.

The main limitation of Bayesian networks is the requirement that the underlying graph needs to be a DAG. However, gene regulatory networks often include cycles (e.g. the cell cycle) or feedback loops that control the expression levels of genes. Thus, a more general class of probability distributions on graphs is needed that allows for the presence of directed cycles. To overcome this shortcoming, *Murphy* (2002) introduced a generalization of Bayesian networks for analysis of time series data, called dynamic Bayesian networks (DBN). In DBNs, random variables in the study are replicated over time, and directed edges are only allowed from variables in each time point to those in the future time points. In its simplest form, edges in DBN are limited to those from variables in t to variables in $t+1$. Such a model corresponds to a Markov model. More generally, for variables $X_1, \ldots X_p$ observed over time points $t = 1, \ldots, T$, edges are allowed from any time point t to future time points $t' > t$.

A closely related model for analysis of time series, which we adapt in this work, was developed in the econometrics literature based on the work of *Granger* (1969a). In this framework, called Granger causality, interactions among variables are defined if past observations of one variable result in improved prediction of other variable. More specifically, let $X^{1:T} \equiv \{X\}_{t=1}^{T}$ and $Y^{1:T} \equiv \{Y\}_{t=1}^{T}$, be trajectories of two stochastic processes X and Y up to time T. Then, X is said to be Granger-causal for Y if the joint prediction model in (2.4) significantly outperforms the model in (2.5).

$$Y^T = AY^{1:T-1} + BX^{1:T-1} + \varepsilon^T \tag{2.4}$$

$$Y^T = AY^{1:T-1} + \varepsilon^T \tag{2.5}$$

Network Granger causal models (NGC) extend the notion of Granger causality among two variables to p variables. In general, define a vector time series $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top$ and consider the corresponding vector auto-regressive (VAR) model (*Lütkepohl* (2005), Chapter 2):

$$\mathbf{X}^T = A^1\mathbf{X}^{T-1} + \ldots A^d\mathbf{X}^{T-d} + \varepsilon^T. \tag{2.6}$$

Here, d denotes the order of the time series and $A^t, t = 1, \ldots, d$ are $p \times p$ matrices whose coefficients represent the magnitude of interaction effects among variables at different time points.

In this model, X_j^{T-t} is considered Granger-causal for X_i^T if the corresponding coefficient, $A_{i,j}^t$ is statistically significant. It is then easy to see that, the NGC corresponds to a DAG with $p \times (d+1)$ variables, in which the ordering of the set of p-variate vectors $\mathbf{X}^{T-d}, \ldots, \mathbf{X}^T$ is determined by the temporal index and the ordering among the elements of each vector is arbitrary. As with DBNs, the interactions in NGCs are only allowed to be forward in time, i.e. of the form $X_j^{T-t} \to X_i^T, t = 1, \ldots, d$.

2.2.2 Penalized Likelihood Estimation Methods for Gene Regulatory Networks

In the analysis of gene regulatory networks, the number of genes often exceeds the available samples of the gene expression data. As a result, an estimate of the gene regulatory network based on graphical Granger causality may include spurious edges that do not correspond to interactions among the genes. In such situations, penalized estimation methods can improve the accuracy of the model, especially for reconstructing the true regulatory network. *Shojaie and Michailidis* (2010b) show that for Gaussian random variables, when the variables inherit a natural ordering,

the likelihood function can be written as a function of the adjacency matrix of the corresponding DAG. They also show that the penalized estimate of the adjacency matrix can be obtained by solving $p-1$ penalized regression problems. Using this connection, general weighted lasso estimates of gene regulatory networks can be found by solving the following p distinct ℓ_1-regularized least squares problems for $i = 1, \ldots, p$:

$$\underset{\theta^t \in \mathbb{R}^p}{\operatorname{argmin}} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t}\theta^t\|_2^2 + \lambda \sum_{t=1}^{d}\sum_{j=1}^{p} |\theta_j^t| w_j^t \tag{2.7}$$

where $\mathcal{X}^t$ denotes the $n \times p$ matrix of observations at time t, and $\mathcal{X}_i^t$ denotes the i^{th} column of $\mathcal{X}^t$. In this formulation, $w_j^t = 1$ corresponds to lasso estimates, and adaptive lasso estimates are obtained by setting $w_j^t = |\hat{A}_{ij}^t|^{-\gamma}$, where $\hat{A}_{ij}^t$ is a consistent estimate of A_{ij}^t. *Shojaie and Michailidis* (2010b) consider a modification of the adaptive lasso, which they call 2-stage lasso in which $w_j^t = 1 \vee |\hat{A}_{ij}^t|^{-\gamma}$, and $\hat{A}_{ij}^t$ is obtained using an initial lasso estimate and $\gamma = 1$.

As pointed out in S-M, the order of the VAR model d is often unknown. Therefore, to estimate the NGC, one either has to include all the previous time points by setting $d = T - 1$, or set d to an arbitrary value. While the latter choice may result in ignoring some of the edges from the true network, the former results in a model with $p(T-1)$ covariates, which in turn exhibits inferior performance. To overcome this shortcoming, the authors propose to estimate the NGC using the truncating lasso penalty, which is given as the solution of the following non-convex optimization problem, for $i = 1, \ldots, p$:

$$\underset{\theta^t \in \mathbb{R}^p}{\operatorname{argmin}} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t}\theta^t\|_2^2 + \lambda \sum_{t=1}^{d} \Psi^t \sum_{j=1}^{p} |\theta_j^t| w_j^t \tag{2.8}$$

$$\Psi^1 = 1, \quad \Psi^t = M^{I\{\|A^{(t-1)}\|_0 < p^2\beta/(T-1)\}},\ t \geq 2$$

where M is a large constant, and β is the allowed false negative rate. S-M propose

an efficient algorithm for solving the optimization problem in (2.8), and show that the proposed penalty gives a consistent estimate of the order of the underlying VAR model, as well as the structure of the network if the model satisfies a decay assumption.

2.3 Adaptively Thresholded Lasso Estimate

The decay assumption for the truncating lasso estimate considered in S-M is a natural assumption in many applications. However, there are examples of VAR models that do not satisfy this assumption. As an example, consider the VAR model whose adjacency matrix is depicted in the top panel of Figure 2.2. In this case, observations at time T are affected by those in time $T-1$ and $T-3$, whereas no significant effects exists from observations in time $T-2$. In Section 2.5, we show that the time series model of T-cell regulation shows a similar pattern of influence. In such cases when the decay assumption fails to hold, the truncating lasso penalty of S-M may not give a correct estimate of the order of the time series, which results in an incorrect estimate of the regulatory network. Examples of such cases are given in Sections 2.4.2 and 2.5.

To address this shortcoming, here we propose to consider the use of adaptive thresholding to provide a consistent estimate of the regulatory networks from time course gene expression data. The main idea for the proposed penalty (which replaces the decay assumption of S-M) is that a given time point includes true effects in the VAR model only if the number of edges in the network should exceed a certain threshold (we formalize this assumption in the following discussion).

Thresholding of lasso estimates has been also considered as a tool to improve the accuracy of lasso estimates in *Wasserman and Roeder* (2009); *Meinshausen and Yu* (2009). More recently, *Zhou* (2010) considered iterative thresholding of both lasso and Dantzig selector estimates for estimation of high dimensional sparse regression models with random design matrix. The author studied asymptotic properties of the thresholded estimator and shows that it results in accurate model selection, as well

as nearly optimal ℓ_2 loss.

To obtain consistent estimates of the order d, as well as edges of the regulatory network, we modify the thresholding framework of *Zhou* (2010) so that only adjacency matrices with significant number of edges are included in the estimate of the regulatory network. Consider, as before, random variables $\mathbf{X}^1, \dots \mathbf{X}^T$ from a VAR model of order d with Gaussian noise, i.e.

$$\mathbf{X}^T = A^1\mathbf{X}^{T-1} + \dots A^d\mathbf{X}^{T-d} + \varepsilon^T, \quad \varepsilon^T \sim N(0, \sigma^2 I_p) \tag{2.9}$$

where I_p denotes the $p \times p$ identity matrix. The adaptively thresholded lasso estimate of NGC is found through the following three-step procedure:

(i) Obtain the regular lasso estimate of the adjacency matrices of NGC $\tilde{A}^t_{\lambda_n}$ by solving (2.7) with tuning parameter $\lambda = \lambda_n$

(ii) Define $\Psi^t = \exp\left(M\mathbf{1}_{\left\{\|\tilde{A}^t\|_0 < p^2\beta/(T-1)\right\}}\right), t = 1, \dots T$, and find the thresholded estimator by setting:

$$\hat{A}^t_{ij} = \tilde{A}^t_{ij}\mathbf{1}_{\left\{|\tilde{A}^t_{ij}| \geq \tau\Psi^t\right\}} \tag{2.10}$$

Here M is a large constant and τ is the tuning parameter for the thresholding step.

(iii) Estimate the order of the time series by setting

$$\hat{d} = \max_t \left\{t : \|\hat{A}^t\|_0 \geq p^2\beta/(T-1)\right\}$$

Before discussing the asymptotic properties of the proposed adaptively thresholded lasso estimator, we compare some features of the new estimator with the truncating lasso estimator of S-M, and discuss the appropriate choice of tuning parameters λ_n and τ.

The proposed adaptively thresholded estimate is found by first obtaining an estimate of the adjacency matrices using regular lasso. Then, in the thresholding step, simultaneous sparsity and order selection in VAR models is achieved by setting small values of the estimated adjacency matrix to zero, while controlling for the total number of nonzero elements of the adjacency matrix. Finally, the index of the last time point in which a significant number of nonzero elements exist in the estimated adjacency matrix is defined as the estimate of the order of VAR model.

As pointed out earlier, the thresholded estimator requires less stringent assumptions about the structure of the time series model, and as shown in Theorem II.1, the consistency of the estimates of the adjacency matrix and the order of the time series are achieved under the usual sparsity and restricted eigenvalue (RE) assumptions. In addition, since the thresholded estimator is found by adaptive thresholding of the regular lasso estimates, the resulting optimization problem is convex. In contrast, although the algorithm for finding the truncating lasso estimate of S-M is shown to be convergent, the resulting estimate may correspond to a local optimum. On the other hand, the thresholded estimator requires appropriate values of two tuning parameters λ_n and τ, and hence the truncating lasso estimate may be obtained more directly. In particular, S-M propose the following error-based choice of tuning parameter, which controls a version of false positive probability:

$$\lambda_e = 2n^{-1/2} Z^*_{\frac{\alpha}{2(T-1)p^2}} \tag{2.11}$$

where α is the probability of false positive determined by the user, and Z^*_q denotes the upper qth quantile of the standard normal distribution. This alleviates the need for searching over the parameter space for appropriate values of λ and provides an intuitive connection to the original definition of Granger causality between two time series given earlier.

Based on the asymptotic properties of the thresholded lasso estimator, and given $\lambda_0 = \sqrt{2\log((T-1)p)/n}$, *Zhou* (2010) suggests the following choices for tuning parameters λ_n and τ:

$$\lambda_n = c_1\sigma\lambda_0$$

$$\tau = c_2\sigma\lambda_0$$

for positive constants c_1 and c_2. Considering the fact that, the choice of the thresholding parameter β is determined by the acceptable degree of false negative error, for $\lambda_0 = \sqrt{2\log((T-1)p)/n}$, and an estimate σ, tuning parameters for the proposed adaptively thresholded estimator amount to appropriate choices of constants c_1 and c_2. A common strategy is to use cross validation (C.V.) over a grid of possible values of c_1 and c_2. We refer the interested reader to *Zhou* (2010) for additional details on connections between c_1 and c_2 and constants that are defined based on the conditions of the problem. For selection consistency of the estimate, we require $c_1 \geq 2\sqrt{1+\theta}$ for some constant $\theta > 0$ and $c_2 = 4c_1$. The quantity θ controls the rate at which the estimator performs consistent variable selection as reflected in Theorem 1. In Sections 2.4 and 2.5, we provide additional guidelines on practical choices of tuning parameters for the data examples considered.

We begin the discussion of asymptotic properties by providing additional notations and statements of the main assumptions.

Denote by $\mathcal{X} = [\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^{T-1}]$ the $n \times p(T-1)$ matrix of "past" observations, and define:

$$\Lambda_{\min}(m) := \min_{\nu \neq 0, \|\nu\|_0 \leq m} \frac{\|\mathcal{X}\nu\|_2^2}{n\|\nu\|_2^2} > 0$$

Denote by $E^t = \{(i,j) : A_{ij}^t \neq 0\}$ the edge set of the adjacency matrix at time lag $t = 1, \ldots, d$ and let $E = \{(i,j) : \exists 1 \leq t \leq d : A_{ij}^t \neq 0\}$ be the set of all edges in the NGC model.

Let $s = \max_i |\mathrm{pa_i}|$ be the maximum number of parents of each node in the NGC

model, and define

$$a_0 = \min_{1\le t\le d} \min_{1\le i,j\le p, A_{ij}\neq 0} |A^t_{ij}|$$

The asymptotic analysis for the thresholded lasso in *Zhou* (2010) incorporates the framework of *Bickel et al.* (2009), based on the restricted eigenvalue condition $RE(\mathcal{X})$, which states that for some integer $1 \le s \le (T-1)p$ and a number k, and for all $\nu \neq 0$ we have

$$\frac{1}{K(s,k)} := \min_{J\subset V, |J|\le s} \min_{\|\nu_{J^c}\|_1 \le k\|\nu_J\|_1} \frac{\|\mathcal{X}\nu\|_2}{n^{1/2}\|\nu_J\|_2} > 0$$

In this case, we say that $RE(\mathcal{X})$ holds with $K(s,k)$. Based on these assumptions, we have the following result on the consistency of network estimation and order selection.

Theorem II.1 (Consistency of Adaptively Thresholded Lasso)**.** *In VAR(d) model of* (2.6) *with independent Gaussian noise with variance* σ^2*, suppose* $RE(\mathcal{X})$ *holds with* $K(s,3)$*, and that* $\lambda_n \ge 2\sigma\sqrt{1+\theta}\lambda_0$ *for some* $\theta > 0$*. Also, assume* $a_0 > c\lambda_n\sqrt{s}$*, for some constant* c *depending on* $\Lambda_{\min}(2s)$ *and* $K(s,3)$*. Finally, assume* $|E| = \zeta\, p^2(T-1)$[1] *for some* $0 < \zeta < 1$*.*

Then for $b = 3K^2(s,3)/4$ *and for any* $\beta > \frac{(T-1)\,b\,s}{p}$*, with probability at least* $1 - p(\sqrt{\pi\, log(T-1)p}[(T-1)p]^{\theta})^{-1}$*, the following hold for the adaptively thresholded lasso estimator with thresholding parameter* β*:*

(i) Control of Type-I error: $FPR \le \frac{b\,s}{(T-1)\,p\,(1-\zeta)}$

(ii) Control of Type-II error: if there exists $\delta > 0$ *such that* $\min_{A^t\neq 0} \|A^t\|_0 > \gamma p^2$ *and* β *is chosen such that* $\beta < \delta/(T-1)$*, then* $FNR = 0$*, otherwise,* $FNR \le \frac{\beta}{(T-1)\,\zeta}$

(iii) Order selection consistency: under the condition in (ii), $\hat{d} = d$

[1]This assumption is made for simplicity of representation. The proof can be written in terms of $|E|$, without making any explicit assumptions on the number of true edges.

Proof. The proof here builds on the results in *Zhou* (2010) (in particular Theorems 1.1 and 3.1), with modifications to account for adaptive thresholding, control of FPR and FNR, and the time series structure. For simplicity, denote by FP and FN, the total number of false positives and false negatives. Also, let $P \equiv |E| = \zeta\,(T-1)p^2$ be total number of positives (i.e. total number of edges) and $N \equiv (T-1)p^2 - |E| = (T-1)p^2\,(1-\zeta)$ denote the number of zeros in the true adjacency matrix.

First, note that from the decomposition of likelihood in *Shojaie and Michailidis* (2010b) it follows that the adaptively thresholded estimator is found by solving p regular lasso regression problems according to (2.7), followed by the thresholding step according to (2.10).

Next note that, by definition of s and the RE condition, each of the p regressions satisfies the $RE(\mathcal{X})$ holds with $K(s,3)$. Therefore, for $\beta = 0$ results of *Zhou* (2010) apply to each individual regression.

Following *Zhou* (2010) we consider, for each $\theta \geq 0$, the set

$$\mathcal{T}_{\theta,i} = \left\{ \epsilon_i^T : \left\| \frac{1}{n} \mathcal{X}^T \epsilon_i^T \right\|_\infty \leq \lambda_{\sigma,\theta,p}, \text{ where } \lambda_{\sigma,\theta,p} = \sigma\sqrt{1+\theta}\lambda_0 \right\}$$

for which $\mathbb{P}(\mathcal{T}_{\theta,i}) \geq 1 - (\sqrt{\pi log(T-1)p}((T-1)p)^\theta)^{-1}$. It then follows from Theorem 1.1 of *Zhou* (2010) that for $\beta = 0$, on the set $\mathcal{T}_\theta = \prod_{i=1}^{p} \mathcal{T}_{\theta,i}$, we have, for all $i = 1, \ldots p$, $\mathrm{pa_i} \subseteq \hat{\mathrm{pa}}_\mathrm{i}$. This implies that for all $t = 1, \ldots, d$, on the set $\mathcal{T}_\theta$, we have

$$E^t \subseteq \hat{E}^t$$

To obtain the upper bound on FPR, we follow the proof of theorem 3.1 in *Zhou* (2010) for each of the p regressions separately. First note that from the results of *Bickel et al.* (2009) it follows that on the set $\mathcal{T}_{\theta,i}$, for $\tilde{v}_i = vec(\tilde{A}_{i:}^{1:T} - A_{i:}^{1:T})$,

$$\|\tilde{v}_{i,\mathrm{pa_i}}\|_2 \leq B_0 \lambda_n \sqrt{s} \text{ and } \|\tilde{v}_{i,\mathrm{pa_i^c}}\|_1 \leq B_1 \lambda_n\, s \tag{2.12}$$

where $B_0 = 4K^2(s,3)$ and $B_1 = 3K^2(s,3)$. If we threshold the lasso estimate by $4\lambda_n$, then it readily follows from (2.12) that (see *Zhou* (2010) for more details) on $\mathcal{T}_{\theta,i}$

$$|\hat{\mathrm{pa}}_{\mathrm{i}} \backslash \mathrm{pa}_{\mathrm{i}}| \leq \frac{\|\tilde{\mathrm{v}}_{\mathrm{i},\mathrm{pa}_{\mathrm{i}}^{\mathrm{c}}}\|_1}{4\lambda_{\mathrm{n}}} \leq \frac{\mathrm{B}_1\,\mathrm{s}}{4} \tag{2.13}$$

Hence $|\hat{\mathrm{pa}}_{\mathrm{i}} \backslash \mathrm{pa}_{\mathrm{i}}| \leq \mathrm{B}_1\,\mathrm{s}/4$, for all $i = 1, \ldots p$, on $\mathcal{T}_\theta$. This implies $FP \leq pbs$ where $b = 3K^2(s,3)/4$.It then follows that on $\mathcal{T}_\theta$ for $\beta = 0$, we have $FNR = 0$ and

$$FPR = FP/N \leq \frac{b\,s}{(T-1)\,p\,(1-\zeta)}.$$

To complete the proof, it suffices to show that for $\beta > \frac{(T-1)\,b\,s}{p}$, FPR does not increase (or is improved) and $FNR \leq \beta/(T-1)\,\zeta$. The fact that adaptive thresholding does not increase FPR follows immediately from the definition of the estimator, as the thresholding coefficient for the adaptively thresholded procedure is at least as large as the procedure of *Zhou* (2010).

Now suppose $A^t \neq 0$ for some $1 \leq t \leq T-1$. It follows from $E \subset \hat{E}$ that $\|\hat{A}^t\|_0 \geq \|A^t\|_0$ and hence, if $\|\hat{A}^t\|_0 < \frac{\beta p^2}{T-1}$, A^t must satisfy the same inequality. Now, if there exists $\delta > 0$ such that $\min_{A^t \neq 0} \|A^t\|_0 > \gamma p^2$ and β is chosen such that $\beta < \delta/(T-1)$, then $\|A^t\|_0 < \delta\,p^2$, which implies that $A^t \equiv 0$, and hence $FNR = 0$. On the other hand, if the condition in (ii) is not satisfied, FN could be at most βp^2, which implies that

$$FNR \leq (\beta p^2)/|E| = \frac{\beta}{(T-1)\zeta}.$$

Finally, to show that $\hat{d} = d$, note that when $A^t \neq 0$, the condition in (ii) guarantees that $\hat{A}^t \neq 0$. On the other hand, if $A^t = 0$, $\|\hat{A}^t\|_0/p^2 \leq \frac{b\,s}{p}$ and hence when $\beta \geq \frac{(T-1)\,b\,s}{p}$, $\hat{A}^t \equiv 0$, which completes the proof. □ □

Before investigating the small sample performance of the proposed estimator in Section 2.4, we offer some remarks regarding asymptotic properties of the estimator.

1. Consider the asymptotic regime with $n \to \infty$, $p = O(n^a)$, for some $a > 0$, and $s = o(p)$. Assume the constant $K(s, 3)$ is uniformly bounded above (see the remark below on the validity of this assumption). Then theorem 1 says that with probability tending to 1, $FPR \to 0$ as long as ζ stays away from 1, i.e., the network is truly sparse. On the other hand, even if no constant δ exists to satisfy the condition in part (ii) of the Theorem, the lower bound on β, given by $\frac{(T-1)\,b\,s}{p}$, converges to zero, indicating that we can make FNR arbitrarily small as long as ζ stays away from zero, i.e., the network is not extremely sparse. The conditions on β are set to achieve a tradeoff between FPR and FNR.

2. The false positive rate in the above theorem can be improved by considering a multi-step thresholding procedure where at the second step the estimate of d is used to restrict the number of time points considered in the estimation. It can be shown that the numerator of the upper bound of FPR can be improved from $b\,s$ to $b\,\sqrt{s}$ (refer to *Zhou* (2010) for more details on the multi-step thresholding). However, this requires an additional assumption on the number of parents of each node in the graph, and is hence not pursued here.

3. The RE condition has been shown to hold for many non-trivial classes of Gaussian design matrices (see for example *van de Geer and Bühlmann* (2009a), *Raskutti et al.* (2010)). In particular *Raskutti et al.* (2010) shows that $RE(\mathcal{X})$ holds with high probability if the sample size n is sufficiently large ($\sim O(klogp)$) and $RE(\Sigma^{1/2})$ holds, where the rows of $\mathcal{X} \sim N(0, \Sigma)$. Hence it is sufficient to ensure that $\lambda_{min}(\Sigma)$ is bounded away from zero as $n, p \to \infty$, which is not very restrictive since every node of the NGC network is a noisy observation with i.i.d innovation of variance σ^2. For the special case of stationary vector autoregressive processes, in Chapter III we use spectral density representation of time series to show a stationary VAR(d) process satisfies this condition if the

spectral matrix operator has continuous eigenvalues and eigenvectors and the adjacency matrices for $t = 1, \ldots T$ are bounded above in spectral norm.

4. The results in Theorem 1 are non-asymptotic and are derived in the regime $n, p \to \infty$ and $p \gg n$, without any restrictions on the length of the time series T. However, it can be seen that if $T \to \infty$, then FPR and FNR converge to 0. In addition, the increase in T also improves the probability of the events under study.

2.4 Numerical Studies

In this section, we evaluate the performance of the proposed thresholded lasso penalty in reconstructing temporal Granger causal effects, and compare it with the performances of (adaptive) lasso and truncating (adaptive) lasso penalties. To this end, we first present the estimated adjacency matrices of two small networks with $p = 20$ and different sparsity patterns to better understand the properties of the thresholded lasso penalty. We then evaluate the phase transition behavior of the competing estimators as the sample size n and the signal to noise ratio (SNR) is varied. To compare the performances of different estimators, we consider three different criteria: (1) the False Positive Rate (FPR), (2) the True Positive Rate (TPR) and (3) the F_1 measure. The F_1 measure is the harmonic mean of $precision(P)$ and $recall(R)$ (i.e. $F_1 = 2PR/(P + R)$) for the estimated graphs. The value of this summary measure ranges between 0 and 1, with higher values corresponding to better estimates.

2.4.1 Illustrative Examples

To illustrate the effect of the proposed estimator, we begin with a simple VAR model that satisfies the decay assumption of S-M. Here $T = 20$, $d = 2$, $p = 20$ and

$s \simeq min\{0.025p^2, n\}$, and every edge has an effect of $\rho = \pm 0.6$. We simulate $n = 30$ independent and identically distributed observations according to the VAR(d) model in (2.6), with $\sigma = 0.3$. The values of α and β are set to 0.1 each.

To obtain comparable results, we set the tuning parameter λ for all estimators to $\lambda = 0.6\lambda_e$, where λ_e is defined in (2.11). The thresholding parameter τ in the second stage of the thresholded lasso penalty is chosen to be $0.7\lambda\sigma$. The results over 50 replications of the above simulation and estimation procedure are presented in Figure 2.1 and Table 2.1[2].

As expected, the truncating lasso estimator outperforms the lasso and thresholded lasso estimators, and provides a consistent estimate of the order d. On the other hand, the thresholded lasso estimator offers additional improvements over its non-thresholded counterpart.

Next, we consider a more complicated structure, where the decay assumption is not satisfied. In particular, we construct a network with the same parameters as before except with $d = 3$ in such a way that there is no edge in the adjacency matrix from lag 2 (i.e., $A^2 = 0$). True and estimated adjacency matrices for this simulation setting are shown in Figure 2.2. The performances of the estimators in terms of TPR, FPR, and F_1 are given in Table 2.2.

It can be seen that the truncating lasso penalty incorrectly estimates the order of VAR as $\hat{d} = 1$, resulting in increased false positive and false negative errors. On the

[2]Here we present the results of simulation for adaptive versions of lasso and truncating lasso estimators; the behavior of the regular versions of these estimators were similar and were excluded to save space

	Alasso	TAlasso	Thlasso
TPR	0.3341 (0.0311)	0.4083 (0.0375)	0.3485 (0.0339)
FPR (×1000)	0.9843 (0.494)	0.8155 (0.4068)	0.4593 (0.2712)
F_1	0.4725 (0.0405)	0.5534 (0.0433)	0.5024 (0.0405)

Table 2.1: F_1, FPR and TPR for (adaptive) lasso, truncating (adaptive) lasso and thresholded lasso. Numbers in the table show mean and standard deviations (in parentheses) over 50 replication.

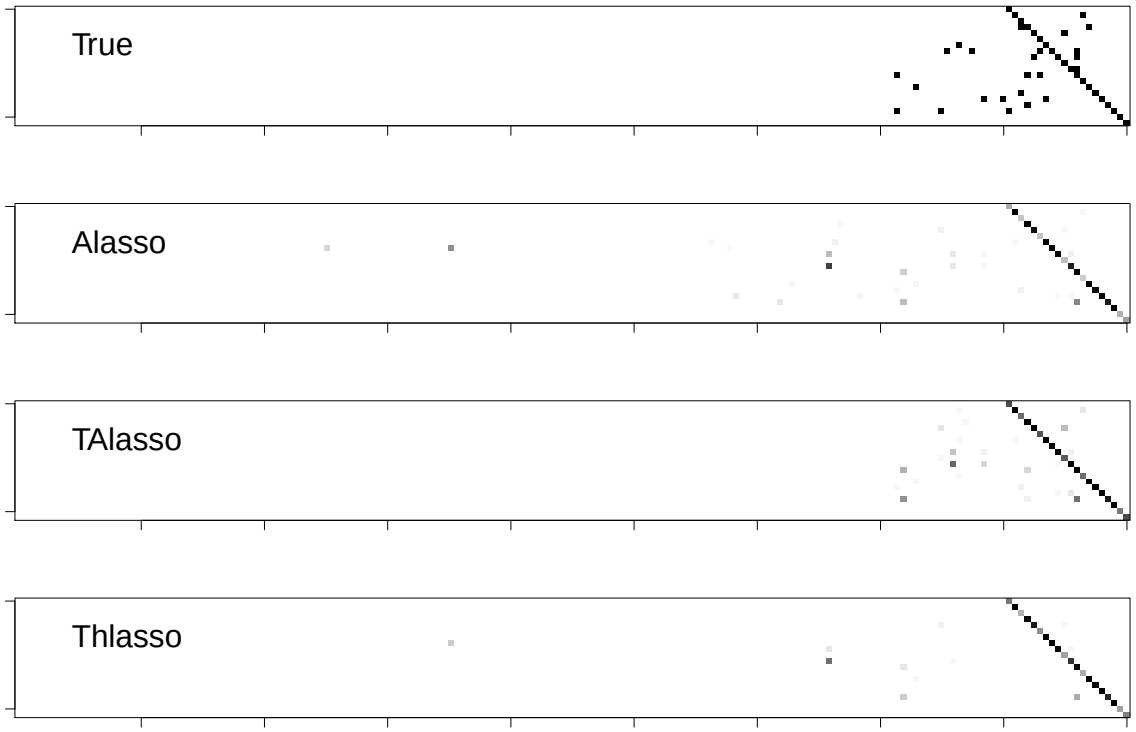


Figure 2.1: True and estimated adjacency matrices of graphical Granger model (a) with T=10, d=2, p=20, n=30, SNR=2.4, the gray-scale images of the estimates represent the percentage of times an edge has been detected in the 50 iterations.

other hand, the (adaptive) lasso estimate includes many edges in later time lags, while failing to include some of the edges in the first time lag. This simulation illustrates the logic and advantages of the proposed thresholded lasso estimator.

2.4.2 Study of Phase Transition Behavior

In this section, we study the phase transition of three performance metrics as the values of (a) sample size (n) and (b) signal-to-noise ratio ($SNR = \rho/\sigma$) is varied for

	Alasso	TAlasso	Thlasso
TPR	0.3462 (0.0529)	0.3077 (0.0558)	0.6288 (0.0698)
FPR (×1000)	0.8254 (0.3454)	0.7694 (0.3729)	0.7415 (0.2611)
F_1	0.4729 (0.0591)	0.4338 (0.0654)	0.7251 (0.0581)

Table 2.2: F_1, FPR and TPR for (adaptive) lasso, truncating (adaptive) lasso and thresholded lasso. Numbers in the table show mean and standard deviations (in parentheses) over 50 replication.

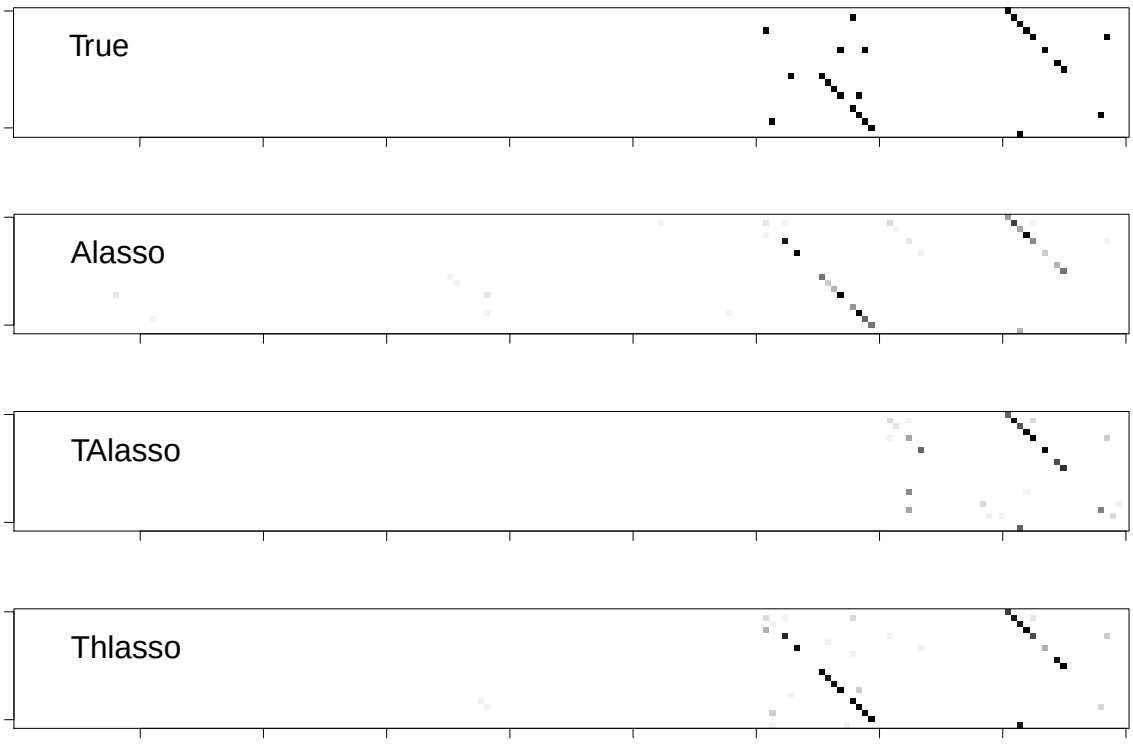


Figure 2.2: True and estimated adjacency matrices of graphical Granger model (b) with T=10, d=3, p=20, n=30, SNR=2.4, the gray-scale images of the estimates represent the percentage of times an edge has been detected in the 50 iterations.

different combinations of $n,\ p,\ \rho$ and σ. The results showing phase transitions for sample size are based on $p = 100,\ \rho = 0.9,\ \sigma = 0.3$, while those for phase transitions for SNR use $p = 150,\ n = 120, \sigma = 0.3$. Similar results were obtained for other choices of these parameters.

Figure 2.3 summarizes the phase transition results for sample size n. It can be seen that the phase transition occurs at a much smaller sample size for thresholded lasso compared to (adaptive) lasso and truncating (adaptive) lasso. However, the performances of thresholded lasso and regular lasso are almost similar when n is almost as large as p. For smaller sample sizes, thresholded lasso slightly affects the number of false positives, but greatly improves on the false negatives, resulting in a better F_1 than regular lasso.

Results of phase transition for SNR presented in Figure 2.4 also indicate that

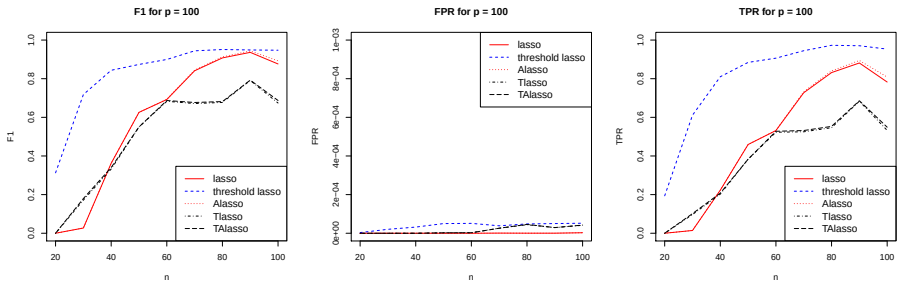


Figure 2.3: Phase transition of F_1, FPR and TPR with increase in sample size

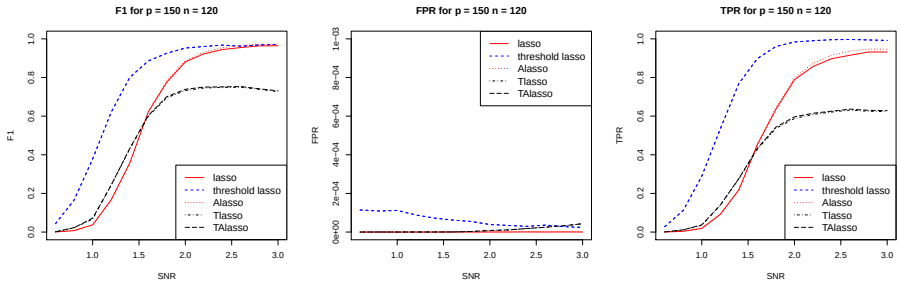


Figure 2.4: Phase transition of F_1, FPR and TPR with increase in SNR

phase transition occurs at a smaller SNR for thresholded lasso compared to (adaptive) lasso and truncating (adaptive) lasso. As in the previous case, the performance of thresholded lasso and regular lasso become more similar as SNR increases. Also, it can be seen that for smaller SNR, thresholded lasso slightly affects the number of false positives while greatly improves the false negatives, which results in significant gain in the overall performance of the proposed estimator in terms of the F_1 measure.

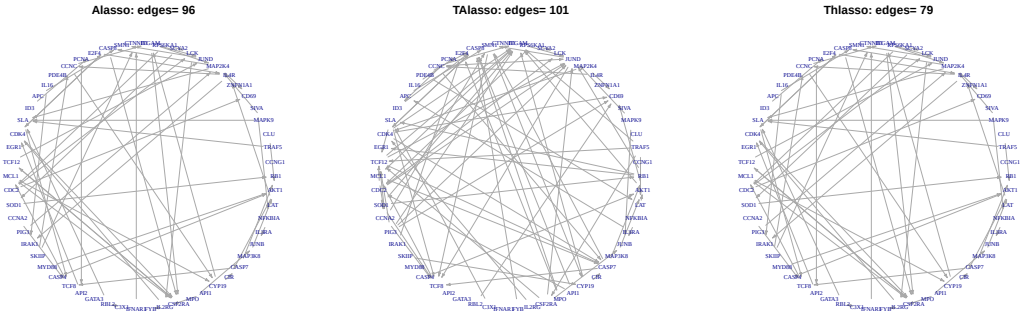
Comparison of phase transition behaviors of lasso, truncating lasso and the adaptively thresholded lasso procedures indicates that the proposed estimator provides a better estimate of Granger causal effects over the range of values of n and SNR. In addition, this advantage becomes more significant in problems with smaller sample size and/or signal to noise ratio.

2.5 Analysis of T-Cell Activation

We illustrate the application of NGC models in reconstructing gene regulatory networks using the time course gene expression data of *Rangel et al.* (2004) on T-cell activation. Activated T-cells are involved in regulation of effector cells (e.g. B-cells) and play a central role in mediating immune response. The data set comprises of $n = 44$ gene expression samples of $p = 58$ genes involved in activation of T-cells, measured over 10 time points. In this study, the activity levels of genes are measured at $t = 0, 2, 4, 6, 8, 18, 24, 32, 48, 72$ hours after stimulation of cells using a T-cell receptor independent activation mechanism. Since changes in regulations often occur at early stages of activation, and to simplify the analysis from the unbalanced experiments, we consider only the earliest 5 time points.

Estimated networks of T-cell activation using the adaptive lasso, the truncating adaptive lasso and the thresholded lasso estimators are shown in Figure 2.5. The tuning parameters for different estimators are determined as in Section 2.4, where the value of σ is estimated using the standard pooled estimate. Lasso and truncating lasso estimates provided similar estimates to their adaptive counterparts and considering the advantages of the adaptive estimators over the regular estimators are not presented. The networks in Figure 2.5 are obtained by drawing an edge between gene i and gene j whenever there is an nonzero element in one of the adjacency matrices $\hat{A}^t_{ij}, T - \hat{d} \leq t \leq T - 1$. Comparison of the estimated networks reveals a significant overlap between the adaptive lasso and thresholded lasso estimates, whereas the truncating adaptive lasso estimate seems to give a different estimate. This is highlighted by the summary measures in Table 2.3, where the total number of edges in each network, along with the structural Hamming distance (SHD) between pairs of two networks, defined as the number of edges different between each two networks, are given.

The striking difference between the estimated regulatory networks using the trun-



(a) Adaptive Lasso (b) Truncating Lasso (c) Thresholded Lasso

Figure 2.5: Estimated Gene Regulatory Networks of B-cell activation. Edges indicate nonzero entries in the estimated adjacency matrix in at least one time lag.

cating lasso estimate raises the question of whether the decay condition necessary for the performance of the truncating lasso estimator is satisfied. Although the true regulatory mechanism in this biological system is unknown, the gray-scale images of the estimated adjacency matrices in Figure 2.6 suggest that in this case the decay condition may be indeed violated. This example underscores the advantage of our newly proposed estimator in cases where the conditions required for the truncating lasso estimate of S-M are not met.

2.6 Discussion

Time course gene expression data provide a valuable source of information for the study of biological systems. Simultaneous analysis of changes in expressions of thousands of genes over time reveals important cues to the dynamic behavior of

	Alasso	TAlasso	Thlasso
Alasso	(96)	–	–
TAlasso	99	(101)	–
Thlasso	35	102	(79)

Table 2.3: Structural Hamming Distance between different estimates of the T-cell regulatory network. Diagonal numbers in parentheses show the total number of edges in each network.

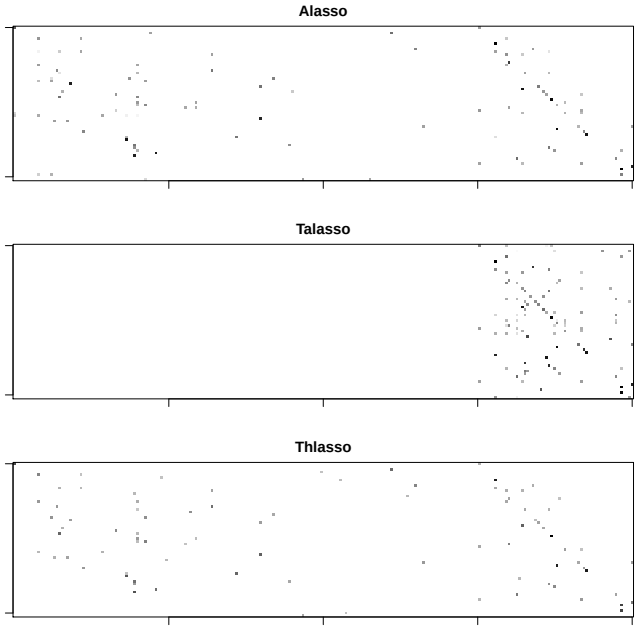


Figure 2.6: Adjacency Matrices of Estimated B-Cells Networks.

the organism and provides a unique window for discovering regulatory interactions among genes. A main challenge in applying statistical models for inferring regulatory networks from time course gene expression data stems from the unknown order of the time series. Simplified methods that ignore effects of genes from time points farther in the past may suffer from loss of information, and could fail to include significant regulatory interactions that are manifested after a long time lag. In contrast, methods that incorporate all of the past information may suffer from an unnecessary curse of dimensionality, and could result in inferior inference especially when the sample size is small.

To overcome this challenge, we proposed a new penalized estimation method for inferring gene regulatory networks from time series observations, based on adaptive thresholding of lasso estimates. The proposed estimator builds upon the previously proposed truncating lasso estimator *Shojaie and Michailidis* (2010a). Both of these estimators attempt to simultaneously estimate the order of the VAR model and the structure of the network, under two different structural assumptions. While the

truncating lasso estimate is based on the assumption that the effects of genes on each other decay over time, the newly proposed adaptively thresholded lasso estimator relies on a less stringent structural assumption that sets a lower bound on the number of edges in the adjacency matrix of the NGC at each time point (see Section 2.3 for a formal statement of this assumption). The relaxation of the decay assumption allows the new estimator to correctly estimate the order of the time series in a broader class of models. However, while the truncating lasso penalty may fail in situations where the decay assumption is violated, it offers advantages in favorable settings.

A natural question therefore arises on the choice of the appropriate penalty for simultaneous estimation of the order of the time series and the structure of the NGC model. The truncating lasso penalty can be advantageous if its underlying assumption is satisfied, but its performance degrades markedly if it does not hold. In absence a formal methodology for determining which of the two assumptions may be more appropriate, the regular (adaptive) lasso estimate can guide the user: if the estimate from the (adaptive) lasso clearly supports the decay assumption, then one could apply the truncating lasso penalty, otherwise, the thresholded lasso penalty provides a more reliable estimate of the NGC.

CHAPTER III

Network Granger Causality with Inherent Grouping Structure

3.1 Introduction

We consider the problem of learning a directed network of interactions among a number of entities from time course data. A natural framework to analyze this problem uses the notion of Granger causality (*Granger*, 1969b). Originally proposed by C.W. Granger this notion provides a statistical framework for determining whether a time series X is useful in forecasting another one Y, through a series of statistical tests. It has found wide applicability in economics, including testing relationships between money and income (*Sims*, 1972), government spending and taxes on economic output (*Blanchard and Perotti*, 2002), stock price and volume (*Hiemstra and Jones*, 1994), etc. More recently the Granger causal framework has found diverse applications in biological sciences including functional genomics, systems biology and neurosciences to understand the structure of gene regulation, protein-protein interactions and brain circuitry, respectively.

It should be noted that the concept of Granger causality is based on associations between time series, and only under very stringent conditions, true causal relationships can be inferred (*Pearl*, 2000b). Nonetheless, this framework provides a powerful

tool for understanding the interactions among random variables based on time course data.

Network Granger causality (NGC) extends the notion of Granger causality among two variables to a wider class of p variables. Such extensions involving multiple time series are handled through the analysis of vector autoregressive processes (VAR) (*Lütkepohl*, 2005). Specifically, for p stationary time series $X_1^t, \ldots, X_p^t$, with $X^t = (X_1^t, \ldots, X_p^t)'$, one considers the class of models

$$X^t = A^1 X^{t-1} + \ldots + A^d X^{t-d} + \epsilon^t, \tag{3.1}$$

where $A^1, A^2, \ldots, A^d$ are $p \times p$ real-valued matrices, d is the *unknown* order of the VAR model and the innovation process satisfies $\epsilon^t \sim N(0, \sigma^2 I)$. In this model, the time series $\{X_j^t\}$ is said to be Granger causal for the time series $\{X_i^t\}$ if $A_{i,j}^h \neq 0$ for some $h = 1, \ldots, d$. Equivalently we can say that there exists an edge $X_j^{t-h} \to X_i^t$ in the underlying network model comprising of $(d+1) \times p$ nodes (see Figure 3.1). We call $A^1, \ldots, A^d$ the adjacency matrices from lags $1, \ldots, d$. Note that the entries A_{ij}^h of the adjacency matrices are not binary indicators of presence/absence of edges between two nodes X_i^t and X_j^{t-h}. Rather, they represent the direction and strength of influence from one node to the other.

The temporal structure induces a natural partial order among the nodes of this network, which in turn simplifies significantly the corresponding estimation problem (*Shojaie and Michailidis*, 2010a) of a directed acyclic graph. Nevertheless, one still has to deal with estimating a high-dimensional network (e.g. hundreds of genes) from a limited number of samples.

The traditional asymptotic framework of estimating VAR models requires observing a long, stationary realization $\{X^1, \ldots, X^T, T \to \infty\, p, d \text{ fixed}\}$ of the p-dimensional time series. This is not appropriate in many biological applications

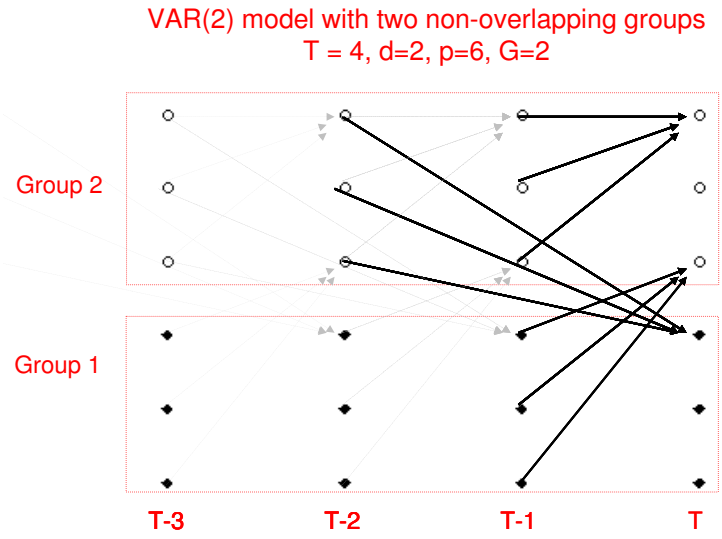


Figure 3.1: An example of a Network Granger causal model with two non-overlapping groups observed over T = 4 time points

for the following reasons. First, long stationary time series are rarely observed in these contexts. Second, the number of time series (p) being large compared to T, the task of consistent order (d) selection using standard criteria (e.g., AIC or BIC) becomes challenging. Similar issues arise in many econometric applications where empirical evidence suggests lack of stationarity over a long time horizon, although the multivariate time series exhibits locally stable distributional properties.

A more suitable framework comes from the study of panel data, where one observes several replicates of the time series, with possibly short T, across a panel of n subjects. In biological applications replicates are obtained from test subjects. In the analysis of macroeconomic variables, households or firms typically serve as replicates. After removing panel specific fixed effects one treats the replicates as independent samples, performs regression analysis under the assumption of common slope structure and studies the asymptotic properties under the regime $n \to \infty$. Recent works of *Cao and Sun* (2011) and *Binder et al.* (2005) analyze theoretical properties of short panel VARs in the low-dimensional setting ($n \to \infty$, T, p fixed).

The focus of this work is on estimating a *high-dimensional NGC model* in the panel data context (p, n large, T small to moderate). This work is motivated by two application domains, functional genomics and financial econometrics. In the first application (presented in Section 3.6) one is interested in reconstructing a gene regulatory network structure from time course data, a canonical problem in functional genomics (*Michailidis*, 2012). The second motivating example examines the composition of balance sheets of the $n = 50$ largest US banks by size, over $T = 9$ quarterly periods, which provides insight into their risk profile.

The nature of high-dimensionality in these two examples comes from both estimation of p^2 coefficients for each of the adjacency matrices $A^1, \ldots, A^d$, but also from the fact that the order of the time series d is often unknown. Thus, in practice, one must either "guess" the order of the time series (often times, it is assumed that the data is generated from a VAR(1) model, which can result in significant loss of information), or include all of the past time points, resulting in significant increase in the number of variables in cases where $d \ll T$. Thus, efficient estimation of the order of the time series becomes crucial.

Latent variable based dimension reduction techniques like principal component analysis or factor models are not very useful in this context since our goal is to reconstruct a network among the observed variables. To achieve dimension reduction we impose a group sparsity assumption on the structure of the adjacency matrices $A_1, \ldots, A_d$. In many applications, structural grouping information about the variables exists. For example, genes can be naturally grouped according to their function or chromosomal location, stocks according to their industry sectors, assets/liabilities according to their class, etc. This information can be incorporated to the Granger causality framework through a group lasso penalty. If the group specification is correct it enables estimation of denser networks with limited sample sizes (*Bach*, 2008; *Huang and Zhang*, 2010; *Lounici et al.*, 2011). However, the group lasso penalty can achieve

model selection consistency only at a group level. In other words, if the groups are misspecified, this procedure can not perform within group variable selection (*Huang et al.*, 2009), an important feature in many applications.

Over the past few years, several authors have adopted the framework of network Granger causality to analyze multivariate temporal data. For example, *Fujita et al.* (2007b) and *Lozano et al.* (2009b) employed NGC models coupled with penalized ℓ_1 regression methods to learn gene regulatory mechanisms from time course microarray data. Specifically, *Lozano et al.* (2009b) proposed to group all the past observations, using a variant of group lasso penalty, in order to construct a relatively simple Granger network model. This penalty takes into account the average effect of the covariates over different time lags and connects Granger causality to this average effect being significant. However, it suffers from significant loss of information and makes the consistent estimation of the signs of the edges difficult (due to averaging). *Shojaie and Michailidis* (2010b) proposed a truncating lasso approach by introducing a truncation factor in the penalty term, which strongly penalizes the edges from a particular time lag, if it corresponds to a highly sparse adjacency matrix.

Despite recent use of NGC in applications involving high dimensional data, theoretical properties of the resulting estimators have not been fully investigated. For example, *Lozano et al.* (2009b) and *Shojaie and Michailidis* (2010b) discuss asymptotic properties of the resulting estimators, but neither address in depth norm consistency properties, nor do they examine under what vector autoregressive structures the obtained results hold.

In this chapter, we develop a general framework that accommodates different variants of group lasso penalties for NGC models. It allows for the simultaneous estimation of the order of the times series and the Granger causal effects; further, it allows for variable selection even when the groups are misspecified. In summary, the key contributions of this work are: (i) investigate in depth *sufficient conditions* that

explicitly take into consideration the structure of the VAR(d) model to establish norm consistency, (ii) introduce the novel notion of *direction consistency*, which generalizes the concept of sign consistency and provides insight into the properties of group lasso estimates within a group, and (iii) use the latter notion to introduce an easy to compute thresholded variant of group lasso, that performs within group variable selection in addition to group sparsity pattern selection even when the group structure is misspecified.

All the obtained results are non-asymptotic in nature, which help provide insight into the properties of the estimates under different asymptotic regimes arising from varying growth rates of T, p, n, group sizes and the number of groups.

3.2 Model and Framework

Notation. Consider a VAR model

$$\underbrace{X^t}_{p\times 1} = \underbrace{A^1}_{p\times p} X^{t-1} + \ldots + A^d X^{t-d} + \epsilon^t,\ \epsilon^t \sim N(0_{p\times 1}, \sigma^2 I_{p\times p}) \tag{3.2}$$

observed over T time points $t = 1, \ldots, T$, across n panels. The index set of the variables $\mathbb{N}_p = \{1, 2, \ldots, p\}$ can be partitioned into G non-overlapping groups $\mathcal{G}_g$, i.e., $\mathbb{N}_p = \cup_{g=1}^{G} \mathcal{G}_g$ and $\mathcal{G}_g \cap \mathcal{G}_{g'} = \phi$ if $g \neq g'$. Also $k_g = |\mathcal{G}_g|$ denotes the size of the g^{th} group with $k_{max} = \max\limits_{1\leq g\leq G} k_g$. In general, we use $\lambda_{\min}$ and $\lambda_{\max}$ to denote the minimum and maximum of a finite collection of numbers $\lambda_1, \ldots, \lambda_m$.

For any matrix A, we denote the i^{th} row by $A_{i:}$, j^{th} column by $A_{:j}$ and the collection of rows (columns) corresponding to the g^{th} group by $A_{[g]:}$ ($A_{:[g]}$). The transpose of a matrix A is denoted by A' and its Frobenius norm by $||A||_F$. For a symmetric/Hermitian matrix Σ, its maximum and minimum eigenvalues are denoted by $\Lambda_{\min}(\Sigma)$ and $\Lambda_{\max}(\Sigma)$, respectively. The symbol $A^{1:h}$ is used to denote the concatenated matrix $\left[A^1 : \cdots : A^h\right]$, for any $h > 0$. For any matrix or vector D, $\|D\|_0$ denotes

the number of non-zero coordinates in D. For notational convenience, we reserve the symbol $\|.\|$ to denote the ℓ_2 norm of a vector and/or the spectral norm of a matrix. For a pre-defined set of non-overlapping groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$ on $\{1, \ldots, p\}$, the mixed norms of vectors $v \in \mathbb{R}^p$ are defined as $\|v\|_{2,1} = \sum_{g=1}^{G} \|v_{[g]}\|$ and $\|v\|_{2,\infty} = \max_{1 \leq g \leq G} \|v_{[g]}\|$. Also for any vector β, we use β_j to denote its j^{th} coordinate and $\beta_{[g]}$ to denote the coordinates corresponding to the g^{th} group. We also use $supp(v)$ to denote the support of v, i.e., $supp(v) = \{j \in \{1, \ldots, p\} | v_j \neq 0\}$.

Network Granger causal (NGC) estimates with group sparsity. Consider n replicates from the NGC model (3.2), and denote the $n \times p$ observation matrix at time t by $\mathcal{X}^t$. In econometric applications the data on p economic variables across n panels (firms, households etc.) can be observed over T time points. For time course microarray data one typically observes the expression levels of p genes across n subjects over T time points. After removing the panel specific fixed effects one assumes the common slope structure and independence across the panels. The data are high-dimensional if either T or p is large compared to n. In such a scenario, we assume the existence of an underlying group sparse structure, i.e., for every $i = 1, \ldots, p$, the support of the i^{th} row of $A^{1:T-1} = \left[A^1 : \cdots : A^{T-1}\right]$ in the model (3.2) can be covered by a small number of groups s_i, where $s_i \ll (T-1)G$. Note that the groups can be misspecified in the sense that the coordinates of a group covering the support need not be all non-zero. Hence, for a properly specified group structure we shall expect $s_i \ll \|A_{i:}^{1:T}\|_0$. On the contrary, with many misspecified groups, s_i can be of the same order, or even larger than $\|A_{i:}^{1:T}\|_0$.

Learning the true network of Granger causal effects $\{(i,j) \in \{1, \ldots, p\} : A_{ij}^t \neq 0 \text{ for some } t\}$ is equivalent to recovering the correct sparsity pattern in $A^{1:(T-1)}$ and consistently estimating the non-zero effects A_{ij}^t. In the high-dimensional regression problems this is achieved by simultaneous regularization and selection operators like

lasso and group lasso. The group Granger causal estimates of the adjacency matrices $A^1, \ldots, A^{T-1}$ are obtained by solving the following optimization problem

$$\hat{A}^{1:T-1} = \underset{A^1,\cdots,A^{T-1}}{\operatorname{argmin}} \frac{1}{2n} \left\| \mathcal{X}^T - \sum_{t=1}^{T-1} \mathcal{X}^{T-t} \left(A^t\right)' \right\|_F^2 + \lambda \sum_{t=1}^{T-1} \sum_{i=1}^{p} \sum_{g=1}^{G} w_{i,g}^t \|A_{i:[g]}^t\| \tag{3.3}$$

where $\mathcal{X}^t$ is the $n \times p$ observation matrix at time t, constructed by stacking n replicates from the model (3.2), w^t is a $p \times G$ matrix of suitably chosen weights and λ is a common regularization parameter. The optimization problem can be separated into the following p penalized regression problems:

$$\hat{A}_{i:}^{1:T-1} = \underset{\theta^1,\cdots,\theta^{T-1}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|\mathcal{X}_{:i}^T - \sum_{t=1}^{T-1} \mathcal{X}^{T-t}\theta^t\|^2 + \lambda \sum_{t=1}^{T-1} \sum_{g=1}^{G} w_{i,g}^t \|A_{i:[g]}^t\|, \quad i = 1, \cdots, p \tag{3.4}$$

The order d of the VAR model is estimated as $\hat{d} = \underset{1\le t\le T-1}{\max} \{t : \hat{A}^t \neq \mathbf{0}\}$.

Different choices of weights $w_{i:g}^t$ lead to different variants of NGC estimates. The regular NGC estimates correspond to the choices $w_{i,g}^t = 1$ or $\sqrt{k_g}$, while for adaptive group NGC estimates the weights are chosen as $w_{i,g}^t = \left\|\hat{A}_{i:[g]}^t\right\|^{-1}$, where $\hat{A}^t$ are obtained from a regular NGC estimation. For $\hat{A}_{i:[g]}^t = \mathbf{0}$, the weight $w_{i,g}^t$ is infinite, which is interpreted as discarding the variables in group g from the optimization problem.

Thresholded NGC estimates are calculated by a two-stage procedure. The first stage involves a regular NGC estimation procedure. The second stage uses a bi-level thresholding strategy on the estimates $\hat{A}^t$. First, the estimated groups with ℓ_2 norm less than a threshold ($\delta_{grp} = c\lambda$, $c > 0$) are set to zero. The second level of thresholding (within group) is applied if the *a priori* available grouping information is not entirely reliable. $\hat{A}_{ij}^t$within an estimated group $\hat{A}_{i:[g]}^t$ is thresholded to zero if $\left|\hat{A}_{ij}^t\right| / \left\|\hat{A}_{i:[g]}^t\right\|$ is less than a threshold $\delta_{misspec} \in (0, 1)$. So, for every $t = 1, \ldots, T-1$,

if $j \in \mathcal{G}_g$, the thresholded NGC estimates are

$$\tilde{A}_{ij}^t = \hat{A}_{ij}^t I\left\{\left|\hat{A}_{ij}^t\right| \geq \delta_{misspec}\left\|\hat{A}_{i:[g]}^t\right\|\right\} I\left\{\left\|\hat{A}_{i:[g]}^t\right\| \geq \delta_{grp}\right\}$$

The tuning parameters λ_{grp} and $\delta_{misspec}$ are chosen via cross-validation. The rationale behind this thresholding strategy is discussed in Section 3.4.

3.3 Estimation Consistency of NGC estimates

In this section we establish the norm consistency of regular group NGC estimates. The regular NGC estimates in (3.3) are obtained by solving p separate group lasso programs with a *common* design matrix $\mathbf{X}_{n\times p(T-1)} = [\mathcal{X}^1 : \cdots : \mathcal{X}^{T-1}]$. This design matrix has $\bar{p} = (T-1)p$ columns which can be partitioned into $\bar{G} = (T-1)G$ groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_{\bar{G}}\}$. We denote the sample Gram matrix by $C = \mathcal{X}'\mathcal{X}/n$. For the i^{th} optimization problem, these $\bar{G} = (T-1)G$ groups are penalized by $\lambda_{(t-1)G+g} := \lambda\, w_{i,g}^t$, $1 \leq t \leq T-1$, $1 \leq g \leq G$, with the choice of weights $w_{i,g}^t$ described in Section 3.2. Following *Lounici et al.* (2011) one can establish a non-asymptotic upper bound on the ℓ_2 estimation error of the NGC estimates $\hat{A}^t$ under certain restricted eigenvalue (RE) assumptions. These assumptions are common in the literature of high-dimensional regression (*Lounici et al.*, 2011; *Bickel et al.*, 2009; *van de Geer and Bühlmann*, 2009b) and are known to be sufficient to guarantee consistent estimation of the regression coefficients even when the design matrix is singular. Of main interest, however, is to investigate the validity of these assumptions in the context of NGC models. This issue is addressed in Proposition III.2.

For $L > 0$, we say that a **Restricted Eigenvalue** (RE) assumption RE(s, L) is

satisfied if there exists a positive number $\phi_{RE} = \phi_{RE}(s) > 0$ such that

$$\min_{\substack{J \subset \mathbb{N}_{\bar{G}}, |J| \leq s \\ \Delta \in \mathbb{R}^{\bar{p}} \setminus \{\mathbf{0}\}}} \left\{ \frac{\|\mathbf{X}\Delta\|}{\sqrt{n}\|\Delta_{[J]}\|} : \sum_{g \in J^c} \lambda_g \|\Delta_{[g]}\| \leq L \sum_{g \in J} \lambda_g \|\Delta_{[g]}\| \right\} \geq \phi_{RE} \tag{3.5}$$

The following proposition provides a non-asymptotic upper bound on the ℓ_2-estimation error of the group NGC estimates under RE assumptions. The proof follows along the lines of *Lounici et al.* (2011) and is delegated to Appendix 3.8.3.

Proposition III.1. *Consider a regular NGC estimation problem* (3.4) *with* $s_{\max} = \max_{1 \leq i \leq p} s_i$ *and* $s = \sum_{i=1}^{p} s_i$. *Suppose* λ *in* (3.3) *is chosen large enough so that for some* $\alpha > 0$,

$$\lambda_g \geq \frac{2\sigma}{\sqrt{n}} \sqrt{\left\|C_{[g][g]}\right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log \bar{G}} \right) \text{ for every } g \in \mathbb{N}_{\bar{G}}, \tag{3.6}$$

Also assume that the common design matrix $\mathbf{X} = [\mathcal{X}^1 : \cdots : \mathcal{X}^{T-1}]$ *in the* p *regression problems* (3.4) *satisfy* $RE(2s_{\max}, 3)$. *Then, with probability at least* $1 - 2p\bar{G}^{1-\alpha}$,

$$\left\| \hat{A}^{1:T-1} - A^{1:T-1} \right\|_F \leq \frac{4\sqrt{10}}{\phi^2_{RE}(2s_{\max})} \frac{\lambda^2_{\max}}{\lambda_{\min}} \sqrt{s} \tag{3.7}$$

Remark. Consider a high-dimensional asymptotic regime where $\bar{G} \asymp n^B$ for some $B > 0$, $k_{\max}/k_{\min} = O(1)$, $s = O(n^{a_1})$ and $k_{\max} = O(n^{a_2})$ with $0 < a_1,\ a_2 < a_1 + a_2 < 1$ so that the total number of non-zero effects is $o(n)$. If $\{\|C_{[g][g]}\|,\ g \in \mathbb{N}_{\bar{G}}\}$ are bounded above (often accomplished by standardizing the data) and $\phi^2_{RE}(2s_{\max})$ is bounded away from zero (see Proposition III.2 for more details), then the NGC estimates are norm consistent for any choice of $\alpha > 2 + a_2/B$.

Note that group lasso achieves faster convergence rate (in terms of estimation and prediction error) than lasso if the groups are appropriately specified. For example, if all the groups are of equal size k and $\lambda_g = \lambda$ for all g, then group lasso can achieve an ℓ_2 estimation error of order $O\left(\sqrt{s}(\sqrt{k} + \sqrt{\log \bar{G}})/\sqrt{n}\right)$. In contrast, lasso's error is

known to be of the order $O\left(\sqrt{\|A^{1:d}\|_0 \log \bar{p}/n}\right)$, which establishes that group lasso has a lower error bound if $s \ll \|A_{i:}^{1:d}\|_0$. On the other hand, lasso will have a lower error bound if $s \asymp \|A_{i:}^{1:d}\|_0$, i.e., if the groups are highly misspecified.

Validity of RE assumption in Group NGC problems. In view of Theorem III.1, it is important to understand how stringent the RE condition is in the context of NGC problems. It is also important to find a lower bound on the RE coefficient ϕ_{RE}, as it affects the convergence rate of the NGC estimates. For the panel-VAR setting, we can rigorously establish that the RE condition holds with overwhelming probability, as long as $n,\ p$ grow at the same rate required for ℓ_2-consistency.

The following proposition achieves this objective in two steps. Note that each row of the design matrix $\mathbf{X}$ (common across the p regressions) is independently distributed as $N(\mathbf{0}, \Sigma)$ where Σ is the variance-covariance matrix of the $(T-1)p$-dimensional random variable $\left((X^1)', \ldots, (X^{T-1})'\right)'$. First, we exploit the spectral representation of the stationary VAR process to provide a lower bound on the minimum eigenvalue of Σ. In the next step, we establish a suitable deviation bound on $\mathbf{X} - \Sigma$ to prove that $\mathbf{X}$ satisfies RE condition with high probability for sufficiently large n.

Proposition III.2. *(a) Suppose the VAR(d) model of* (3.2) *is stable, stationary. Let* Σ *be the variance-covariance matrix of the* $(T-1)p$*-dimensional random variable* $\left((X^1)', \ldots, (X^{T-1})'\right)'$. *Then the minimum eigenvalue of* Σ *satisfies*

$$\Lambda_{min}(\Sigma) \geq \sigma^2 \left[\max_{\theta \in [-\pi, \pi]} \|\mathcal{A}(e^{-i\theta})\|\right]^{-2} \geq \sigma^2 \left[1 + \sum_{t=1}^{d} \|A^t\|\right]^{-2} \geq \sigma^2 \left[1 + \frac{1}{2}(\mathbf{v}_{in} + \mathbf{v}_{out})\right]^{-2}$$

where $\mathcal{A}(z) := I - A^1 z - A^2 z^2 - \ldots - A^d z^d$ *is the reverse characteristic polynomial of the VAR(d) process, and* $\mathbf{v}_{in}$, $\mathbf{v}_{out}$ *are the maximum incoming and outgoing effects*

at a node, cumulated across different lags

$$\mathbf{v}_{in} = \sum_{t=1}^{d} \max_{1\le i\le p} \sum_{j=1}^{p} |A_{ij}^t|, \ \mathbf{v}_{out} = \sum_{t=1}^{d} \max_{1\le j\le p} \sum_{i=1}^{p} |A_{ij}^t|$$

(b) In addition, suppose the replicates from different panels are i.i.d. Then, for any $s > 0$*, there exist universal positive constants* c_i *such that if the sample size* n *satisfies*

$$n \ > \frac{\Lambda_{\max}^2(\Sigma)}{\Lambda_{\min}^2(\Sigma)}(2 + L\lambda_{\max}/\lambda_{\min})^4 \, c_0 s(k_{\max} + c_1 \log(e\bar{G}/2s))$$

then $\mathbf{X}$ *satisfies* $RE(s, L)$ *with* $\phi_{RE}^2 \ge \Lambda_{\min}(\Sigma)/2$ *with probability at least* $1-c_2\exp(-c_3\, n)$.

Remark. Proposition III.2 has two interesting consequences. First, it provides a lower bound on the RE constant ϕ_{RE} which is independent of T. So if the high dimensionality in the Granger causal network arises only from the time domain and not the cross-section ($T \to \infty,\ p,\ G$ fixed), the stationarity of the VAR process guarantees that the rate of convergence depends only on the true order (d), and not T. Second, this result shows that the NGC estmates are consistent even if the node capacities $\mathbf{v}_{in}$ and $\mathbf{v}_{out}$ grow with $n,\ p$ at an appropriate rate.

3.4 Variable Selection Consistency of NGC estimates

In view of (3.4), to study the variable selection properties of NGC estimates it suffices to analyze the variable selection properties of p generic group lasso estimates with a common design matrix.

The problem of group sparsity selection has been thoroughly investigated in the literature (*Wei and Huang*, 2010; *Lounici et al.*, 2011). The issue of selection and sign consistency within a group, however, is still unclear. Since group lasso does not impose sparsity within a group, all the group members are selected together (*Huang et al.*, 2009) and it is not clear which ones are recovered with correct signs. This

also leads to inconsistent variable selection if a group is misspecified, i.e., not all the members within a group has non-zero effect. Several alternate penalized regression procedures have been proposed to overcome this shortcoming (*Breheny and Huang*, 2009; *Huang et al.*, 2009). The main idea behind these procedures is to combine ℓ_2 and ℓ_1 norms in the penalty to encourage sparsity at both group and variable level. These estimators involve nonconvex optimization problems and are computationally expensive. Also their theoretical properties in a high dimensional regime are not well studied.

We take a different approach to deal with the issue of group misspecification. Although group lasso penalty does not perform exact variable selection within groups, it performs regularization and shrinks the individual coefficients. We utilize this regularization to detect misspecification within a group. To this end, we formulate a generalized notion of sign consistency, henceforth referred as "direction consistency", that provides insight into the properties of group lasso estimates within a single group. Subsequently, these properties are used to develop a simple, easy to compute, thresholded variant of group lasso which, in addition to group selection, achieves variable selection and sign consistency within groups.

We consider a generic group lasso regression problem of the linear model $y = X\beta^0 + \epsilon$ with p variables partitioned into G non-overlapping groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$ of size k_g, $g = 1, \ldots, G$. Without loss of generality, we assume $\beta^0_{[g]} \neq \mathbf{0}$ for $g \in S = \{1, 2, \ldots, s\}$ and $\beta^0_{[g]} = \mathbf{0}$ for all $g \notin S$ and consider the following group lasso estimate

of β^0:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{g=1}^{G} \lambda_g \|\beta_{[g]}\| \tag{3.8}$$

$$\underbrace{\beta^0}_{p\times 1} = [\underbrace{\beta^0_{[1]}, \ldots, \beta^0_{[s]}}_{k_1+\ldots+k_s=q}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{p-q}] = [\beta^0_{(1)} : \beta^0_{(2)}] \tag{3.9}$$

$$\underbrace{\mathbf{X}}_{n\times p} = [\underbrace{\mathbf{X}_{(1)}}_{n\times q} : \underbrace{\mathbf{X}_{(2)}}_{n\times(p-q)}] \; C = \frac{1}{n}\mathbf{X}'\mathbf{X} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \tag{3.10}$$

Direction Consistency. For an m-dimensional vector $\tau \in \mathbb{R}^m \backslash \{\mathbf{0}\}$ define its direction vector $D(\tau) = \tau / \|\tau\|$, $D(\mathbf{0}) = \mathbf{0}$. In the context of a generic group lasso regression (3.10), for a group $g \in S$ of size k_g, $D(\beta^0_{[g]})$ indicates the direction of influence of $\beta^0_{[g]}$ at a group level in the sense that it reflects the relative importance of the influential members within the group. Note that for $k_g = 1$ the function $D(\cdot)$ simplifies to the usual $sgn(\cdot)$ function.

Definition. An estimate $\hat{\beta}$ of a generic group lasso problem (3.8) is ***direction consistent*** at a rate δ_n, if there exists a sequence of positive real numbers $\delta_n \to 0$ such that

$$\mathbb{P}\left(\|D(\hat{\beta}_{[g]}) - D(\beta^0_{[g]})\| < \delta_n, \; \forall g \in S, \; \hat{\beta}_{[g]} = \mathbf{0}, \; \forall g \notin S\right) \to 1 \text{ as } n, p \to \infty. \tag{3.11}$$

Now suppose $\hat{\beta}$ is a direction consistent estimator. Consider the set $\tilde{S}^n_g := \{j \in \mathcal{G}_g \; : \; |{\beta^0}_j| \,/\, \|{\beta^0}_{[g]}\| > \delta_n\}$. $\tilde{S}^n_g$ can be viewed as a collection of influential group members within a group $\mathcal{G}_g$, which are "detectable" with a sample of size n. Then, it readily follows from the definition that

$$\mathbb{P}(sgn(\hat{\beta}_j) = sgn(\beta_j), \; \forall j \in \tilde{S}^n_g, \forall g \in \{1, \ldots, s\}) \to 1 \text{ as } n, p \to \infty. \tag{3.12}$$

The latter observation connects the precision of group lasso estimates to the accuracy of *a priori* available grouping information. In particular, if the pre-specified grouping structure is correct, i.e., all the members within a group have non-zero effect, then for a sufficiently large sample size we have $\tilde{S}_g^n = \mathcal{G}_g$ for all $g \in S$. Hence, if the group lasso estimate is direction consistent, it will correctly estimate the sign of all the variables in the support. On the other hand, in case of a misspecified *a priori* grouping structure (numerous zero coordinates in β_g for $g \in S$), group lasso will correctly estimate only the signs of the influential group members.

Example. We demonstrate the property of direction consistency using a small example. Consider a linear model with 8 predictors

$$y = 0.5x_1 - 3x_2 + 3x_3 + x_4 - 2x_5 + 3x_8 + e,\ e \sim N(0,1)$$

The coefficient vector β^0 is partitioned into four groups of size 2, viz., $(0.5, -3)$, $(3,1)$, $(-2,0)$ and $(0,3)$. The last two groups are misspecified. We generated $n = 25$ samples from this model and ran group lasso regression with the above group structure. Figure 3.2 shows the true coefficient vectors (solid) and their estimates (dashed) from five iterations of the above exercise. Note that even though the ℓ_2 errors between $\beta^0_{[g]}$ and $\hat{\beta}_{[g]}$ vary largely across the four groups, the distance between their projections on the unit circle, $\left\| D(\beta^0_{[g]}) - D(\hat{\beta}_{[g]}) \right\|$, are comparatively stable across groups. In fact, Theorem 3.4.1 shows that under certain irrepresentable conditions (IC) on the design matrix, it is possible to find a uniform (over all $g \in S$) upper bound δ_n on the ℓ_2 gap of these direction vectors. This motivates a natural thresholding strategy to correct for the misspecification in groups (cf. Proposition 3.4.2). Even though a group $\beta^0_{[g]}$ is misspecified (i.e., lies on a coordinate axis), direction consistency ensures, with high probability, that the corresponding coordinate in $D(\hat{\beta}_{[g]})$ will be smaller than a threshold δ_n which is common across all groups in the support.

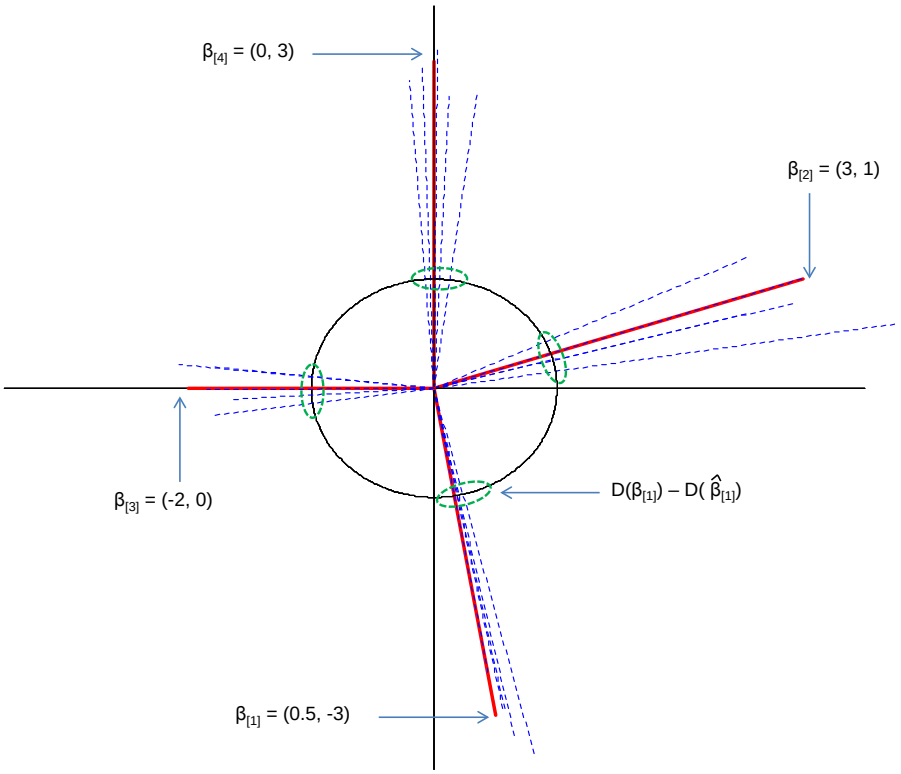


Figure 3.2: Example demonstrating direction consistency

Group Irrepresentable Conditions (IC). Next, we define the IC required for direction consistency of group lasso estimates. Irrepresentable conditions are common in the literature of high-dimensional regression problems (*Zhao and Yu*, 2006; *van de Geer and Bühlmann*, 2009b) and are shown to be sufficient (and essentially necessary) for selection consistency of the lasso estimates. Further these conditions are known to be satisfied with high probability, if the population analogue of the Gram matrix belongs to the Toeplitz family (*Zhao and Yu*, 2006; *Wainwright*, 2009). In NGC estimation the population analogue of the Gram matrix $\Sigma = Var(\mathbf{X}^{1:(T-1)})$ is block Toeplitz, so the irrepresentable assumptions are natural candidates for studying selection consistency of the estimates. Consider the notations of (3.8) and (3.10). Define $K = diag\,(\lambda_1 \mathbf{I}_{k_1}, \lambda_2 \mathbf{I}_{k_2}, \ldots, \lambda_s \mathbf{I}_{k_s})$.

Uniform Irrepresentable Condition (IC) is satisfied if there exists $0 < \eta < 1$ such that for all $\tau \in \mathbb{R}^q$ with $\|\tau\|_{2,\infty} = \max_{1 \le g \le s} \|\tau_{[g]}\|_2 \le 1$

$$\frac{1}{\lambda_g} \left\| \left[C_{21}(C_{11})^{-1} K\tau \right]_{[g]} \right\| < 1 - \eta, \ \forall g \notin S = \{1, \ldots, s\} \tag{3.13}$$

Note that the definition reverts to the usual IC for lasso when all groups correspond are singletons.

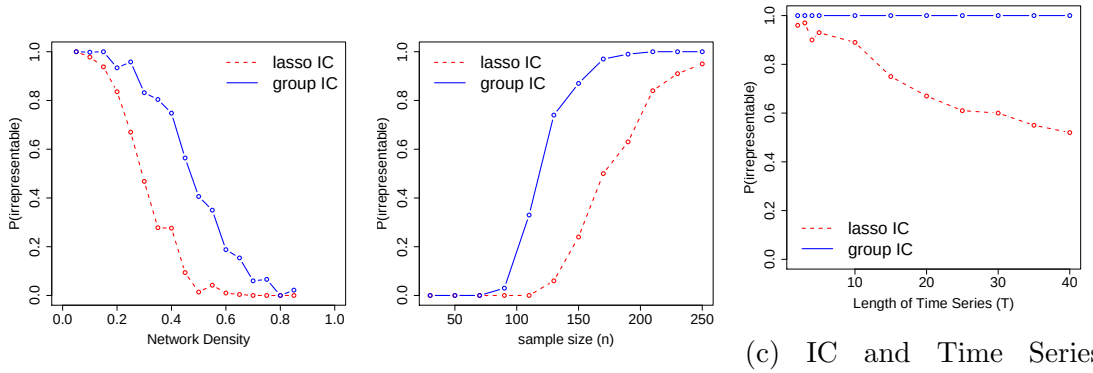
The IC is more stringent than the RE condition and is rarely met if the underlying model is not sparse. It can be shown that a slightly weaker version of this condition is necessary for direction consistency. We refer the readers to Appendix 3.8.4 for further discussion on the different irrepresentable assumptions and their properties. Numerical evidence suggests that the group IC tends to be less stringent than the IC required for the selection consistency of lasso. We illustrate this using three small simulated examples.

Simulation 1. We constructed group sparse NGC models with $T = 5$, $p = 21$, $G = 7, k_g = 3$ and different levels of network densities, where the network edges were selected at random and scaled so that $\|A^1\| = 0.1$. For each of these models, we generated 100 samples of size $n = 150$ and calculated the proportions of times the two types of irrepresentable conditions were met. The results are displayed in Figure 3.3a.

Simulation 2. We selected a VAR(1) model from the above class and generated samples of size $n = 20, 50, \ldots, 250$. Figure 3.3b displays the proportions of times (based on 100 simulations) the two ICs were met.

Simulation 3. We generated $n = 200$ samples from the VAR(1) model of example 2 for $T = 2, 3, 4, 5, 10, \ldots, 40$. Figure 3.3c displays the proportions of times (based on 100 simulations) the two ICs were met.

Selection consistency for generic group lasso estimates. For simplicity, we



(a) IC and Network Density (b) IC and Sample Size (c) IC and Time Series length

Figure 3.3: Comparison of lasso and group irrepresentable conditions in the context of group sparse NGC models. (a) group ICs tend to be met for dense networks where lasso IC fails to meet. (b) For the same network group IC is met with smaller sample size than required by lasso. (c) For longer time series group IC is satisfied more often than lasso IC.

discuss the selection consistency properties of a generic group lasso regression problem with a common tuning parameter across groups, i.e., $\lambda_g = \lambda$ for every $g \in \mathbb{N}_G$. Similar results can be obtained for more general choices of the tuning parameters.

Theorem 3.4.1. *Assume that the group uniform IC holds with* $1-\eta$ *for some* $\eta > 0$. *Then, for any choice of* $\alpha > 0$,

$$\begin{aligned} \lambda &\geq \max_{g \notin S} \frac{1}{\eta} \frac{\sigma}{\sqrt{n}} \sqrt{\left\| (C_{22})_{[g][g]} \right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right) \text{ and} \\ \delta_n &\geq \max_{g \in S} \frac{1}{\left\| \beta^0_{[g]} \right\|} \left(\lambda \sqrt{s} \left\| (C_{11})^{-1} \right\| + \frac{\sigma}{\sqrt{n}} \sqrt{\left\| (C_{11})^{-1}_{[g][g]} \right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right) \right), \end{aligned}$$

with probability greater than $1 - 4G^{1-\alpha}$, *there exists a solution* $\hat{\beta}$ *satisfying*

1. $\hat{\beta}_{[g]} = 0$ *for all* $g \notin S$,
2. $\left\| \hat{\beta}_{[g]} - \beta^0_{[g]} \right\| < \delta_n \left\| \beta^0_{[g]} \right\|$, *and hence* $\left\| D(\hat{\beta}_{[g]}) - D(\beta^0_{[g]}) \right\| < 2\delta_n$, *for all* $g \in S$. *If* $\delta_n < 1$, *then* $\hat{\beta}_{[g]} \neq 0$ *for all* $g \in S$.

Remark. The tuning parameter λ can be chosen of the same order as required for ℓ_2 consistency to achieve selection consistency within groups in the sense of (3.12). Further, with the above choice of λ, δ_n can be chosen of the order of $O(\sqrt{s}(\sqrt{k_{max}} + \sqrt{\log G})/\sqrt{n})$. Thus, group lasso correctly identifies the group sparsity pattern and is direction consistent if $\sqrt{s}(\sqrt{k_{max}} + \sqrt{\log G})/\sqrt{n} \to 0$, the same scaling required for ℓ_2 consistency.

Thresholding in Group NGC estimators. As described in Section 3.2, regular group NGC estimates can be thresholded both at the group and coordinate levels. The first level of thresholding is motivated by the fact that lasso can select too many false positives [cf. *van de Geer et al.* (2011), *Zhou* (2010) and the references therein]. The second level of thresholding employs the direction consistency of regular group NGC estimates to perform within group variable selection with high probability. The following proposition demonstrates the benefit of these two types of thresholding. The second result is an immediate corollary of Theorem 3.4.1. Proof of the first result (thresholding at group level) requires some additional notations and is delegated to Appendix 3.8.5.

Theorem 3.4.2. *Consider a generic group lasso regression problem* (3.8) *with common tuning parameter* $\lambda_g = \lambda$.

(i) Assume the RE(s, 3) condition of (3.5) *holds with a constant* ϕ_{RE} *and define* $\hat{\beta}^{thgrp}_{[g]} = \hat{\beta}_{[g]}\mathbf{1}_{\|\hat{\beta}_{[g]}\|>4\lambda}$. *If* $\hat{S} = \{g \in \mathbb{N}_G : \hat{\beta}^{thgrp}_{[g]} \neq \mathbf{0}\}$, *then* $|\hat{S}\backslash S| \leq \frac{s}{\phi^2_{RE}/12}$, *with probability at least* $1 - 2G^{1-\alpha}$.

(ii) Assume that uniform IC holds with $1 - \eta$ *for some* $\eta > 0$. *Choose* λ *and* δ_n *as in Theorem 3.4.1 and define*

$$\hat{\beta}^{thgrp}_j = \hat{\beta}_j \mathbf{1}\{|\hat{\beta}_j|/\|\hat{\beta}_{[g]}\| > 2\,\delta_n\} \text{ for all } j \in \mathcal{G}_g$$

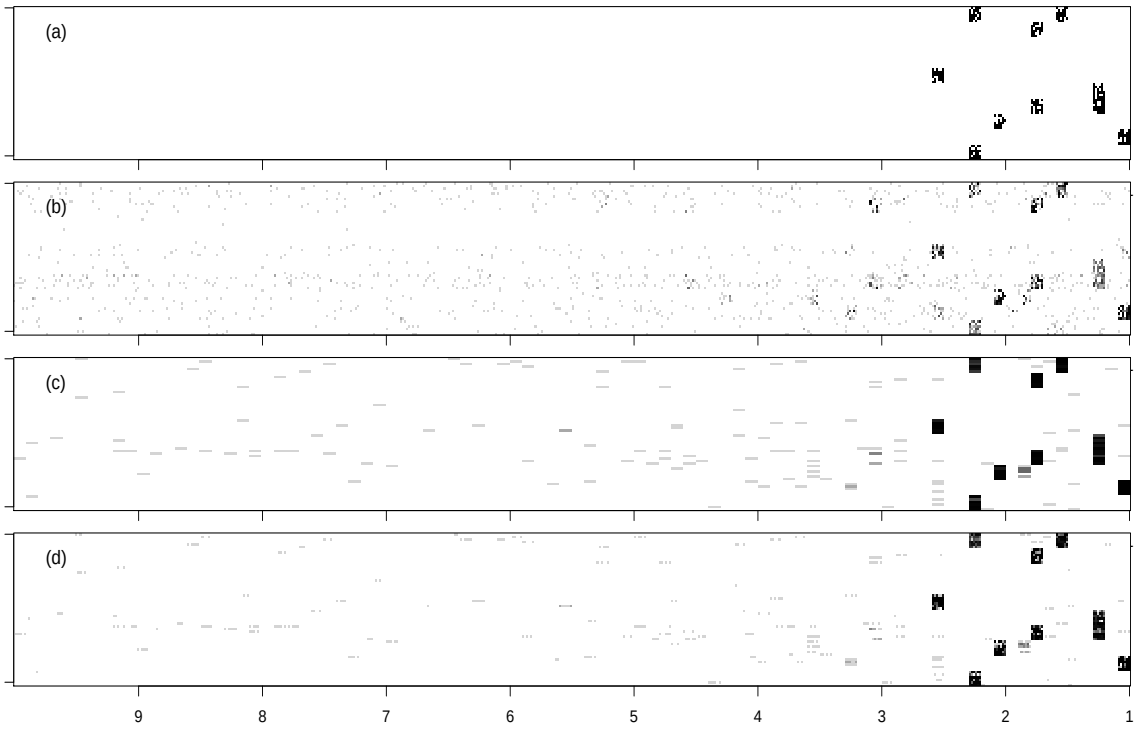


Figure 3.4: Estimated adjacency matrices of a misspecified NGC model with p = 60, T = 10, n = 60: (a) True, (b) Lasso, (c) Group Lasso, (d) Thresholded Group Lasso. The grayscale represents the proportion of times an edge was detected in 100 simulations.

Then $sgn(\beta_j^0) = sgn(\hat{\beta}_j^{thgrp})\ \forall\, j \in \mathbb{N}_p$ *with probability at least* $1-4G^{1-\alpha}$, *if* $\min_{j \in supp(\beta^0)} |\beta_j^0| > 2\delta_n \|\beta_{[g]}^0\|$ *for all* $j \in \mathcal{G}_g$, *i.e., if the effect of every non-zero member in a group is "visible" relative to the total effect from the group.*

3.5 Performance Evaluation

We evaluate the performances of regular, adaptive and thresholded variants of the group NGC estimators through an extensive simulation study, and compare the results to those obtained from lasso estimates. The R package `grpreg` (*Breheny and Huang*, 2009) was used to obtain the group lasso estimates. The settings considered are:

(a) *Balanced groups of equal size*: i.i.d samples of size $n = 60, 110, 160$ are generated from lag-2 ($d = 2$) VAR models on $T = 5$ time points, comprising of $p = 60, 120, 200$

nodes partitioned into groups of equal size in the range 3-5.

(b) *Unbalanced groups:* We retain the same setting as before, however the corresponding node set is partitioned into one larger group of size 10 and many groups of size 5.

(c) *Misspecified balanced groups:* i.i.d samples of size $n = 60, 110, 160$ are generated from lag-2 ($d = 2$) VAR models on $T = 10$ time points, comprising of $p = 60, 120$ nodes partitioned into groups of size 6. Further, for each group there is a 30% misspecification rate, namely that for every parent group of a downstream node, 30% of the group members do not exert any effect on it.

Using a 19 : 1 sample-splitting, the tuning parameter λ is chosen from an interval of the form $[C_1\lambda_e, C_2\lambda_e]$, $C_1, C_2 > 0$, where $\lambda_e = \sqrt{2 \log p/n}$ for lasso and $\sqrt{2 \log G/n}$ for group lasso. The thresholding parameters are selected as $\delta_{grp} = 0.7\lambda\sigma$ at the group level and $\delta_{misspec} = n^{-0.2}$ within groups. These parameters are chosen by conducting a 20-fold cross-validation on independent tuning datasets of same sizes, using intervals of the form $[C_3\lambda, C_4\lambda]$ for δ_{grp} and $\{n^{-\delta}, \delta \in [0, 1]\}$ for $\delta_{misspec}$. Finally, within group thresholding is applied only when the group structure is misspecified.

The following performance metrics were used for comparison purposes: (i) $Precision = TP/(TP + FP)$, (ii) $Recall = TP/(TP + FN)$ and (iii) Matthew's Correlation coefficient (MCC) defined as

$$\frac{(TP \times TN) - (FP \times FN)}{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))^{1/2}}$$

where TP, TN, FP and FN correspond to true positives, true negatives, false positives and false negatives in the estimated network, respectively. The average and standard deviations (over 100 replicates) of the performance metrics are presented for each setup.

The results for the balanced settings are given in Table 3.1. The Recall for $p = 60$ shows that even for a network with $60 \times (5 - 1) = 240$ nodes and $|E| = 351$ true

Table 3.1: Performance of different regularization methods in estimating graphical Granger causality with **balanced** group sizes and no misspecification; $d = 2$, $T = 5$, $SNR = 1.8$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

		$p = 60$, $\|E\| = 351$ Group Size=3			$p = 120$, $\|E\| = 1404$ Group Size=3			$p = 200$, $\|E\| = 3900$ Group Size=5		
	n	160	110	60	160	110	60	160	110	60
P	Lasso	80(2)	75(2)	66(4)	69(1)	62(2)	52(2)	52(1)	47(1)	38(1)
	Grp	95(2)	91(4)	83(7)	91(3)	80(5)	68(7)	78(4)	72(3)	59(6)
	Thgrp	96(1)	92(3)	86(6)	93(3)	83(5)	70(7)	82(4)	76(3)	64(6)
	Agrp	96(2)	92(4)	83(7)	92(3)	82(5)	69(7)	81(3)	74(3)	60(6)
R	Lasso	71(2)	54(2)	31(2)	54(1)	40(1)	22(1)	38(1)	28(1)	15(1)
	Grp	99(1)	93(3)	71(7)	91(2)	81(2)	48(8)	84(1)	70(2)	41(4)
	Thgrp	99(1)	93(3)	71(7)	91(2)	81(2)	48(8)	84(2)	69(2)	41(3)
	Agrp	99(1)	93(3)	71(7)	91(2)	81(2)	47(8)	84(1)	69(2)	40(4)
MCC	Lasso	75(2)	63(2)	45(3)	60(1)	49(1)	33(1)	43(1)	35(1)	23(1)
	Grp	97(1)	92(3)	76(5)	91(1)	80(2)	56(2)	81(2)	70(2)	48(2)
	Thgrp	98(1)	93(2)	78(5)	92(1)	81(2)	57(3)	83(2)	72(2)	50(3)
	Agrp	97(1)	92(3)	76(5)	91(1)	81(2)	56(3)	82(2)	71(2)	48(2)
ERR	Lasso	10.5	11.3	13.9	16.63	17.37	16.69	19.79	20	18.52
LAG	Grp	3.19	6.95	12.76	4.86	10.77	12.65	4.21	5.27	7.8
	Thgrp	2.83	5.87	10.01	3.98	9.03	11.19	3.06	3.91	5.68
	Agrp	3.13	6.89	12.59	4.63	10.37	12.34	3.58	4.87	7.59

edges, the group NGC estimators recover about 71% of the true edges with a sample size as low as $n = 60$, while lasso based NGC estimates recover only 31% of the true edges. The three group NGC estimates have comparable performances in all the cases. However thresholded lasso shows slightly higher precision than the other group NGC variants for smaller sample sizes (e.g., $n = 60, p = 200$). The results for $p = 60, n = 110$ also display that lower precision of lasso is caused partially by its inability to estimate the order of the VAR model correctly, as measured by ERR LAG=Number of falsely connected edges from lags beyond the true order of the VAR model divided by the number of edges in the network ($|E|$). This finding is nicely illustrated in Figure 3.4 and Table 3.1. The group penalty encourages edges from the nodes of the same group to be picked up together. Since the nodes of the same group are also from the same time lag, the group variants have substantially lower ERR LAG. For example, average ERR LAG of lasso for $p = 200,\ n = 160$ is 19.79% while the average ERR LAGs for the group lasso variants are in the range $3.06\% - 4.21\%$.

The results for the unbalanced networks are given in Table 3.2. As in the balanced

Table 3.2: Performance of different regularization methods in estimating graphical Granger causality with **unbalanced** group sizes and no misspecification; $d = 2$, $T = 5$, $SNR = 1.8$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

		$p = 60$, $\|E\| = 450$ Groups=$1 \times 10, 11 \times 5$			$p = 120$, $\|E\| = 1575$ Groups=$1 \times 10, 23 \times 5$			$p = 200$, $\|E\| = 4150$ Groups=$1 \times 10, 39 \times 5$		
	n	160	110	60	160	110	60	160	110	60
P	Lasso	72(2)	69(3)	62(2)	51(1)	48(1)	41(1)	61(1)	53(1)	42(2)
	Grp	84(4)	79(6)	76(9)	55(5)	47(5)	40(6)	86(3)	77(5)	66(7)
	Thgrp	86(4)	82(7)	78(11)	60(6)	50(7)	40(5)	88(2)	79(6)	69(6)
	Agrp	85(3)	81(5)	77(9)	59(5)	51(5)	42(6)	88(2)	78(5)	67(6)
R	Lasso	45(2)	35(2)	22(2)	43(1)	34(1)	22(1)	23(1)	15(0)	7(0)
	Grp	94(3)	87(5)	61(8)	88(2)	75(5)	48(6)	73(3)	49(6)	22(5)
	Thgrp	95(2)	88(4)	62(8)	89(3)	77(4)	50(5)	73(3)	50(6)	21(5)
	Agrp	94(3)	87(5)	61(8)	88(2)	75(5)	48(6)	73(3)	49(6)	22(5)
MCC	Lasso	56(2)	48(2)	35(2)	46(1)	39(1)	29(1)	36(1)	28(1)	17(1)
	Grp	89(3)	82(4)	67(5)	68(3)	58(3)	42(3)	79(1)	61(3)	37(3)
	Thgrp	90(3)	84(4)	68(6)	72(4)	61(4)	43(2)	80(1)	62(3)	37(3)
	Agrp	89(3)	83(4)	67(6)	71(3)	60(3)	43(3)	79(1)	61(3)	37(3)
ERR	Lasso	10.59	10.74	11.76	18.3	18.72	18.76	11.54	10.93	9.29
LAG	Grp	7.04	9.85	13.04	12.53	14.71	13.06	4.8	6.41	6.85
	Thgrp	6.58	8.98	11.1	9.6	11.9	10.9	4.06	5.65	5.7
	Agrp	6.74	9.19	12.96	10.81	12.78	11.79	4.55	6.2	6.81

group setup, in almost all the simulation settings the group NGC variants outperform the lasso estimates with respect to all three performance metrics. However the performances of the different variants of group NGC are comparable and tend to have higher standard deviations than the lasso estimates. Also the average ERR LAGs for the group NGC variants are substantially lower than the average ERR LAG for lasso demonstrating the advantage of group penalty. Although the conclusions regarding the comparisons of lasso and group NGC estimates remain unchanged it is evident that the performances of all the estimators are affected by the presence of one large group, skewing the uniform nature of the network. For example the MCC measures of group NGC estimates in a balanced network with $p = 60$ and $|E| = 351$ vary around $97 - 98\%$ which lowers to $89\% - 90\%$ when the groups are unbalanced.

The results for misspecified groups are given in Table 3.3. Note that for higher sample size n, the MCC of lasso and regular group lasso are comparable. However, the thresholded version of group lasso achieves significantly higher MCC than the rest. This demonstrates the advantage of using the directional consistency of group

Table 3.3: Performance of different regularization methods in estimating graphical Granger causality with **misspecified** groups (30% misspecification); $d = 2$, $T = 10$, $SNR = 2$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

		$p = 60$, $\|E\| = 246$ Group Size=6			$p = 120$, $\|E\| = 968$ Group Size=6		
	n	160	110	60	160	110	60
P	Lasso	88(2)	85(3)	77(5)	59(1)	55(1)	49(2)
	Grp	65(2)	66(2)	66(3)	43(3)	44(4)	38(4)
	Thgrp	87(3)	88(3)	85(3)	56(6)	56(6)	51(7)
	Agrp	65(2)	66(2)	66(3)	45(2)	45(4)	39(4)
R	Lasso	80(3)	63(3)	37(2)	66(1)	54(1)	35(1)
	Grp	100(0)	98(2)	82(6)	87(2)	78(3)	59(4)
	Thgrp	100(0)	98(2)	79(6)	86(2)	79(3)	57(4)
	Agrp	100(0)	98(2)	82(6)	86(2)	78(3)	58(3)
MCC	Lasso	84(2)	73(2)	53(3)	62(1)	54(1)	41(1)
	Grp	81(1)	80(2)	74(4)	61(2)	58(3)	47(2)
	Thgrp	93(2)	93(2)	82(4)	69(4)	66(4)	53(3)
	Agrp	81(1)	80(2)	74(4)	62(2)	59(2)	47(2)
ERR	Lasso	12.63	17.05	22.41	45.09	49.68	53.4
LAG	Grp	9.43	8.78	15.12	18.22	18.43	29.26
	Thgrp	6.45	5.34	8.02	11.81	12.84	15.57
	Agrp	9.11	8.78	14.96	16.32	16.9	27.69

lasso estimators to perform within group variable selection. We would like to mention here that a careful choice of the thresholding parameters δ_{grp} and $\delta_{misspec}$ via cross-validation improves the performance of thresholded group lasso; however, we do not pursue these methods here as they require grid search over many tuning parameters or an efficient estimator of the degree of freedom of group lasso.

In summary, the results clearly show that all variants of group lasso NGC outperform the lasso-based ones, whenever the grouping structure of the variables is known and correctly specified. Further, their performance depends on the composition of group sizes. On the other hand, if the a priori known group structure is moderately misspecified lasso estimates produce comparable results to regular and adaptive group NGC ones, while thresholded group estimates outperform all other methods, as expected.

Table 3.4: Mean and standard deviation of MSE for different NGC estimates

	Lasso	Grp	Agrp	Thgrp
mean	0.649	0.456	0.457	0.456
stdev	0.340	0.252	0.251	0.252

3.6 Application

Example: T-cell activation. Estimation of gene regulatory networks from expression data is a fundamental problem in functional genomics (*Friedman*, 2004). Time course data coupled with NGC models are informationally rich enough for the task at hand. The data for this application come from *Rangel et al.* (2004), where expression patterns of genes involved in T-cell activation were studied with the goal of discovering regulatory mechanisms that govern them in response to external stimuli. Activated T-cells are involved in regulation of effector cells (e.g. B-cells) and play a central role in mediating immune response. The available data comprising of $n = 44$ samples of $p = 58$ genes, measure the cells response at 10 time points, $t = 0, 2, 4, 6, 8, 18, 24, 32, 48, 72$ hours after their stimulation with a T-cell receptor independent activation mechanism. We concentrate on data from the first 5 time points, that correspond to early response mechanisms in the cells.

Genes are often grouped based on their function and activity patterns into biological pathways. Thus, the knowledge of gene functions and their membership in biological pathways can be used as inherent grouping structures in the proposed group lasso estimates of NGC. Towards this, we used available biological knowledge to define groups of genes based on their biological function. Reliable information for biological functions were found from the literature for 38 genes, which were retained for further analysis. These 38 genes were grouped into 13 groups with the number of genes in different groups ranging from 1 to 5.

Figure 3.5 shows the estimated networks based on lasso and thresholded group lasso estimates, where for ease of representation the nodes of the network correspond

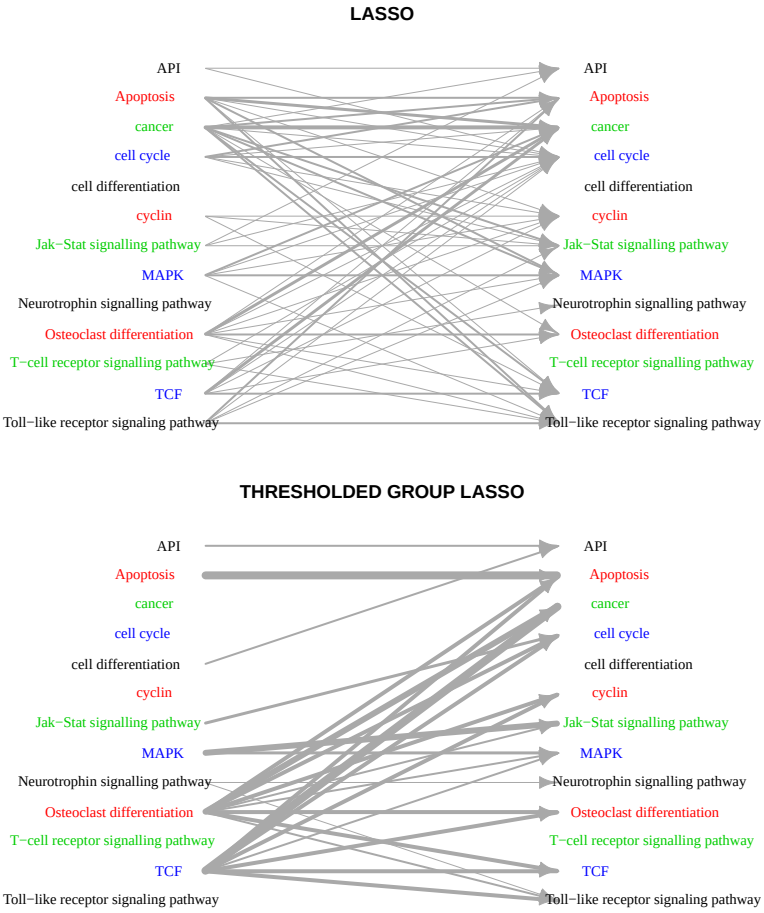


Figure 3.5: Estimated Gene Regulatory Networks of T-cell activation. Width of edges represent the number of effects between two groups, and the network represents the aggregated regulatory network over 3 time points.

to groups of genes. In this case, estimates from variants of group NGC estimator were all similar, and included a number of known regulatory mechanisms in T-cell activation, not present in the regular lasso estimate. For instance, *Waterman et al.* (1990) suggest that TCF plays a significant role in activation of T-cells, which may describe the dominant role of this group of genes in the activation mechanism. On the other hand, *Kim et al.* (2005) suggest that activated T-cells exhibit high levels of osteoclast-associated receptor activity which may attribute the large number of associations between member of osteoclast differentiation and other groups. Finally,

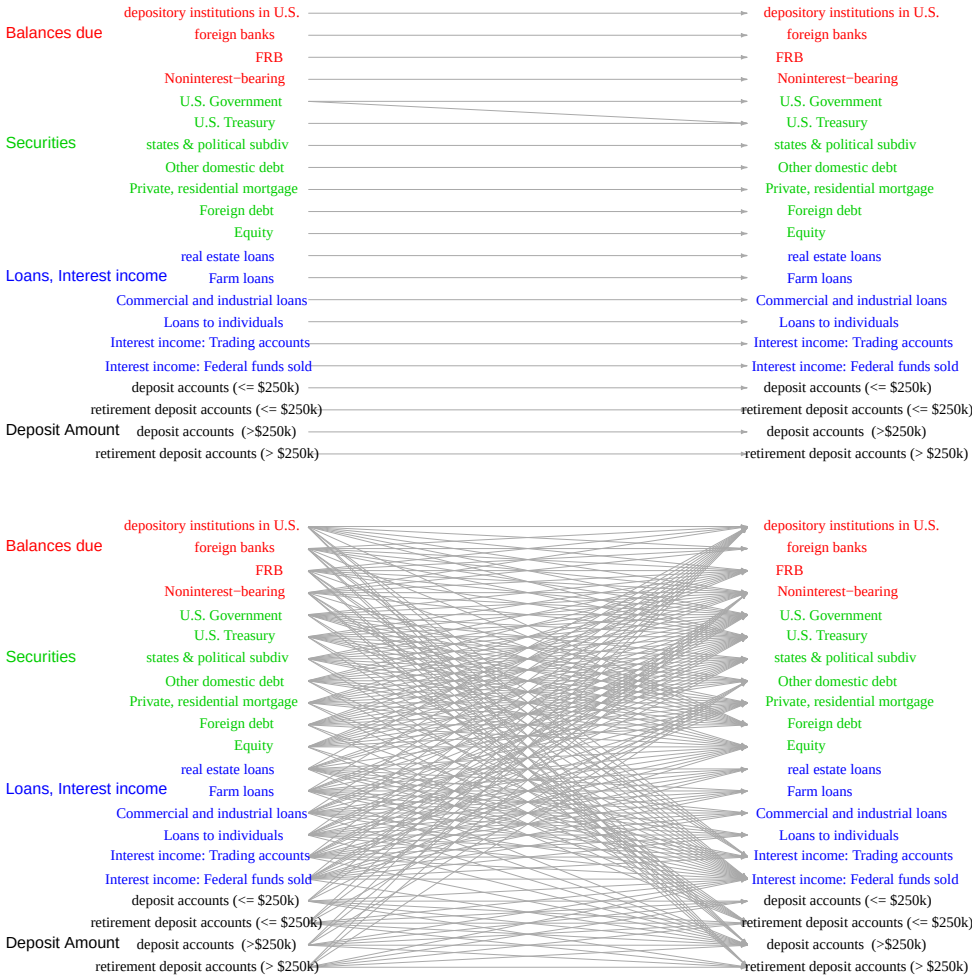


Figure 3.6: Estimated Networks of banking balance sheet variables using (a) lasso and (b) group lasso. The networks represent the aggregated network over 5 time points.

the estimated networks based on variants of group lasso estimator also offer improved estimation accuracy in terms of mean squared error (MSE) despite having having comparable complexities to their regular lasso counterpart (Table 3.4), which further confirms the findings of other numerical studies in that paper.

Example: Banking balance sheets application. In this application, we examine the structure of the balance sheets in terms of assets and liabilities of the $n = 50$ largest (in terms of total balance sheet size) US banking corporations. The data cover 9 quarters (September 2009-September 2011) and were directly obtained from the Fed-

Table 3.5: Mean and standard deviation (in parentheses) of PMSE (MSE in case of Dec 2010) for prediction of banking balance sheet variables.

Quarter	Lasso	Grp	Agrp	Thgrp
Dec 2010	1.59 (0.29)	0.36 (0.05)	0.36 (0.05)	0.37 (0.05)
Mar 2011	1.46 (0.30)	0.47 (0.23)	0.47 (0.23)	0.46 (0.22)
Jun 2011	1.33 (0.26)	0.36 (0.11)	0.36 (0.11)	0.35 (0.11)
Sep 2011	1.72 (0.32)	0.50 (0.18)	0.50 (0.18)	0.47 (0.16)

eral Deposit Insurance Corporation (FDIC) database (available at `www.fdic.gov`). The $p = 21$ variables correspond to different assets (US and foreign government debt securities, equities, loans (commercial, mortgages), leases, etc.) and liabilities (domestic and foreign deposits from households and businesses, deposits from the Federal Reserve Board, deposits of other financial institutions, non-interest bearing liabilities, etc.) We have organized them into four categories: two for the assets (loans and securities) and two for the liabilities (Balances Due and Deposits, based on a \$250K reporting FDIC threshold). Amongst the 50 banks examined, one discerns large integrated ones with significant retail, commercial and investment activities (e.g. Citibank, JP Morgan, Bank of America, Wells Fargo), banks primarily focused on investment business (e.g. Goldman Sachs, Morgan Stanley, American Express, E-Trade, Charles Schwab), regional banks (e.g. Banco Popular de Puerto Rico, Comerica Bank, Bank of the West).

The raw data are reported in thousands of dollars. The few missing values were imputed using a nearest neighbor imputation method with $k = 5$, by clustering them according to their total assets in the most recent quarter (September 2011) and subsequently every missing observation for a particular bank was imputed by the median observation on its five nearest neighbors. The data were log-transformed to reduce non-stationarity issues. The dataset was restructured as a panel with $p = 21$ variables and $n = 50$ replicates observed over $T = 9$ time points. Every column of replicates was scaled to have unit variance.

We applied the proposed variants of NGC estimates on the first $T = 6$ time

points (Sep 2009 - Dec 2010) of the above panel dataset. The parameters λ and δ_{grp} were chosen using a 19 : 1 sample-splitting method and the misspecification threshold $\delta_{misspec}$ was set to zero as the grouping structure was reliable. We calculated the MSE of the fitted model in predicting the outcomes in the four quarters (December 2010 - September 2011). The Predicted MSE (MSE for Dec 2010) are listed in Table 3.5. The estimated network structures are shown in Figure 3.6.

It can be seen that the lasso estimates recover a very simple temporal structure amongst the variables; namely, that past values (in this case lag-1) influence present ones. Given the structure of the balance sheet of large banks, this is an anticipated result, since it can not be radically altered over a short time period due to business relationships and past commitments to customers of the bank. However, the (adaptive) group lasso estimates reveal a richer and more nuanced structure. Examining the fitted values of the adjacency matrices A^t, we notice that the dominant effects remain those discovered by the lasso estimates. However, fairly strong effects are also estimated within each group, but also between the groups of the assets (loans and securities) on the balance sheet. This suggests rebalancing of the balance sheet for risk management purposes between relatively low risk securities and potentially more risky loans. Given the period covered by the data (post financial crisis starting in September 2009) when credit risk management became of paramount importance, the analysis picks up interesting patterns. On the other hand, significant fewer associations are discovered between the liabilities side of the balance sheet. Finally, there exist relationships between deposits and securities such as US Treasuries and other domestic ones (primarily municipal bonds); the latter indicates that an effort on behalf of the banks to manage the credit risk of their balance sheets, namely allocating to low risk assets as opposed to more risky loans.

It is also worth noting that the group lasso model exhibits superior predictive performance over the lasso estimates, even 4 quarters into the future. Finally, in

this case the thresholded estimates did not provide any additional benefits over the regular and adaptive variants, given that the specification of the groups was based on accounting principles and hence correctly structured.

3.7 Discussion

In this chapter, the problem of estimating Network Granger Causal (NGC) models with inherent grouping structure is studied when replicates are available. Norm, and both group level and within group variable selection consistency are established under fairly mild assumptions on the structure of the underlying time series. To achieve the second objective the novel concept of direction consistency is introduced.

The type of NGC models discussed in this study have wide applicability in different areas, including genomics and economics. However, in many contexts the availability of replicates at each time point is not feasible (e.g. in rate of returns for stocks or other macroeconomic variables), while grouping structure is still present (e.g. grouping of stocks according to industry sector). Hence, it is of interest to study the behavior of group lasso estimates in such a setting and address the technical challenges emanating from such a pure time series (dependent) data structure.

3.8 Technical Results

3.8.1 Auxiliary Lemmas

Lemma 3.8.1 (Characterization of the Group lasso estimate)**.** *A vector $\hat{\beta} \in \mathbb{R}^p$ is a solution to the convex optimization problem*

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|Y - X\beta\|^2 + \sum_{g=1}^{G} \lambda_g \|\beta_{[g]}\| \tag{3.14}$$

if and only if $\hat{\beta}$ *satisfies, for some* $\tau \in \mathbb{R}^p$ *with* $\max_{1\leq g\leq G} \left\|\tau_{[g]}\right\| \leq 1$, $\frac{1}{n}\left[X'(Y - X\hat{\beta})\right]_{[g]} = \lambda_g \tau_{[g]} \, \forall g$. *Further,* $\tau_{[g]} = D\left(\hat{\beta}_{[g]}\right)$ *whenever* $\hat{\beta}_{[g]} \neq \mathbf{0}$.

Proof. Follows directly from the KKT conditions for the optimization problem (3.14). □

Lemma 3.8.2 (Concentration bound for multivariate Gaussian)**.** *Let* $Z_{k\times 1} \sim N(0, \Sigma)$. *Then, for any* $t > 0$, *the following inequalities hold:*

$$\mathbb{P}\left(|\|Z\| - \mathbb{E}\|Z\|| > t\right) \leq 2\exp\left(-\frac{2t^2}{\pi^2\|\Sigma\|}\right), \; \mathbb{E}\|Z\| \leq \sqrt{k}\sqrt{\|\Sigma\|}$$

Proof. The first inequality can be found in *Ledoux and Talagrand* (1991) (equation (3.2). To establish the second inequality note that,

$$\mathbb{E}\|Z\| \leq \sqrt{\mathbb{E}\|Z\|^2} = \sqrt{\mathbb{E}\left[\text{trace}\left(ZZ'\right)\right]} = \sqrt{\text{trace}\left(\Sigma\right)} \leq \sqrt{k}\sqrt{\|\Sigma\|}$$

□

Lemma 3.8.3. *Let* $\beta, \; \hat{\beta} \in \mathbb{R}^m \backslash \{\mathbf{0}\}$. *Let* $\hat{u} = \hat{\beta} - \beta$ *and* $r = D(\hat{\beta}) - D(\beta)$. *Then* $\|r\| < 2\delta$ *whenever* $\|\hat{u}\| < \delta\,\|\beta\|$.

Proof. It follows from $\|\hat{u}\| < \delta\,\|\beta\|$ that

$$(1-\delta)\|\beta\| < \|\beta\| - \|\hat{u}\| \leq \|\hat{\beta}\| \leq \|\hat{u}\| + \|\beta\| < (1+\delta)\|\beta\|\,,$$

which implies that $\left|\|\beta\| - \|\hat{\beta}\|\right| < \delta\|\beta\|$. Now,

$$\|\hat{\beta}\|\,\|\beta\|\|r\| = \left\| \; \hat{\beta}\|\beta\| + (\hat{u} - \hat{\beta})\|\hat{\beta}\| \; \right\| \leq \left\|\hat{\beta}\left(\|\beta\| - \|\hat{\beta}\|\right) + \|\hat{\beta}\|\; \hat{u} \; \right\| < \|\hat{\beta}\|\,\|\beta\|(\delta+\delta)$$

since $\left|\|\beta\| - \|\hat{\beta}\|\right| < \delta\|\beta\|$ and $\|\hat{u}\| < \delta\,\|\beta\|$. □

Lemma 3.8.4. *Let $\mathcal{G}_1, \ldots, \mathcal{G}_G$ be any partition of $\{1, \ldots, p\}$ into G non-overlapping groups and $\lambda_1, \ldots, \lambda_G$ be positive real numbers. Define the cone sets $\mathcal{C}(J, L) = \{v \in \mathbb{R}^p : \sum_{g \notin J} \lambda_g \|v_{[g]}\| \leq L \sum_{g \in J} \lambda_g \|v_{[g]}\|\}$ for any subset of groups $J \subseteq \mathbb{N}_G$. Also define the set of group s-sparse vectors $\mathbb{D}(s) := \{v \in \mathbb{R}^p : \|v\| \leq 1,\ supp(v) \subseteq \mathcal{G}_J$ for some $J \subseteq \mathbb{N}_G,\ |J| \leq s\}$. Then*

$$\bigcup_{J \subseteq \mathbb{N}_G, |J| \leq s} \mathcal{C}(J, L) \cap \mathbb{S}^{p-1} \subseteq (2 + L') cl\{conv\{\mathbb{D}(s)\}\} \tag{3.15}$$

where $L' = L\lambda_{\max}/\lambda_{\min}$, $\mathbb{S}^{p-1} = \{v \in \mathbb{R}^p : \|v\| = 1\}$ is the ball of unit norm vectors in $\mathbb{R}^p$ and $cl\{.\}$, $conv\{.\}$ respectively denote the closure and convex hull of a set.

Proof. Note that for any $J \subseteq \mathbb{N}_G$, $|J| \leq s$, and $v \in \mathcal{C}(J, L) \cap \mathbb{S}^{p-1}$, we have

$$\sum_{g \notin J} \|v_{[g]}\| \leq L \frac{\lambda_{\max}}{\lambda_{\min}} \sum_{g \in J} \|v_{[g]}\|$$

which implies

$$\|v\|_{2,1} \leq (L' + 1) \sum_{g \in J} \|v_{[g]}\| \leq (L' + 1)\sqrt{s} \|v_{[J]}\| \leq (L' + 1)\sqrt{s}$$

Hence the union of the cone sets on the left hand side of (3.15) is a subset of $A := \{v \in \mathbb{R}^p : \|v\| \leq 1,\ \|v\|_{2,1} \leq (L' + 1)\sqrt{s}\}$.

We will show that the set A is a subset of $B := (2 + L') cl\{conv\{\mathbb{D}(s)\}\}$, the closed convex hull on the right hand side of (3.15). Since both sets A and B are closed convex, it is enough to show that the support function of A is dominated by the support function of B.

The support fucntion of A is given by $\phi_A(z) = \sup_{\theta \in A} \langle \theta, z \rangle$. For any $z \in \mathbb{R}^p$, let $S \subseteq \{1, \ldots, G\}$ be a subset of top s groups in terms of the ℓ_2 norm of $z_{[g]}$. Thus, $\|z_{[S^c]}\|_{2,\infty} \leq \|z_{[g]}\|$ for all $g \in S$. This implies $\|z_{[S^c]}\|_{2,\infty} \leq (1/s)\|z_{[S]}\|_{2,1} \leq$

$(1/\sqrt{s})\|z_{[S]}\|$. So, we have

$$\begin{aligned}\phi_A(z) = \sup_{\theta\in A}\langle\theta, z\rangle &\leq \sup_{\|\theta_{[S]}\|\leq 1}\langle\theta_{[S]}, z_{[S]}\rangle + \sup_{\|\theta_{[S^c]}\|_{2,1}\leq\sqrt{s}(L'+1)}\langle\theta_{[S^c]}, z_{[S^c]}\rangle \quad (3.16)\\ &\leq \|z_{[S]}\| + (L'+1)\sqrt{s}\|z_{[S^c]}\|_{2,\infty} \leq (L'+2)\|z_{[S]}\| \quad (3.17)\end{aligned}$$

On the other hand, support function of $B := (L'+2)cl\{conv\{\mathbb{D}(s)\}\}$ is given by

$$\phi_B(z) = \sup_{\theta\in B}\langle\theta, z\rangle = (L'+2)\max_{|U|=s,\ U\subseteq\mathbb{N}_G}\sup_{\|\theta_{[U]}\|\leq 1}\langle\theta_{[U]}, z_{[U]}\rangle = (L'+2)\|z_{[S]}\|$$

This concludes the proof. □

Lemma 3.8.5. *Consider a matrix $X_{n\times p}$ with rows independently distributed as $N(0,\Sigma)$, $\Lambda_{\min}(\Sigma) > 0$. Let $\mathcal{G}_1,\ldots,\mathcal{G}_G$ be any partition of $\{1,\ldots,p\}$ into G non-overlapping groups of size $k_1,\ldots,k_g$, respectively. Let $C = X'X/n$ denote the sample Gram matrix and $\mathbb{D}(s)$ denote the set of group s-sparse vectors defined in Lemma 3.8.4. Then, for any integer $s \geq 1$ and any $\eta > 0$, we have*

$$\mathbb{P}\left[\sup_{v\in cl\{conv\{\mathbb{D}(s)\}\}}|v'(C-\Sigma)v| > 6\eta\|\Sigma\|\right] \leq c_0\exp\left[-n\min\{\eta,\eta^2\} + c_1 s(k_{\max} + c_2\log(eG/2s))\right] \quad (3.18)$$

for some universal positive constants c_i.

Proof. We consider a fixed vector $v\in\mathbb{R}^p$ with $\|v\|\leq 1$, the support of which can be covered by a set J of at most s groups, i.e., $supp(v)\subseteq\mathcal{G}_J$, $J\subseteq\mathbb{N}_G$, $|J|\leq s$. Define $Y = Xv$. Then each coordinate of Y is independently distributed as $N(0,\sigma_y^2)$, where $\sigma_y^2 = v'\Sigma v \leq \|\Sigma\|$.

Then, for any $\eta > 0$, Hansen-Wright inequality of *Rudelson and Vershynin* (2013) ensures

$$\mathbb{P}\left[|v'(C-\Sigma)v| > \eta\|\Sigma\|\right] \leq \mathbb{P}\left[\frac{1}{n}|Y'Y - \mathbb{E}Y'Y| > \eta\sigma_y^2\right] \leq 2\exp\left[-cn\min\{\eta,\eta^2\}\right]$$

Next, we extend this deviation bound on all vectors v in the sparse set

$$\mathbb{D}(2s) = \{v \in \mathbb{R}^p : \|v\| \leq 1,\ supp(v) \subseteq \mathcal{G}_J \text{ for some } J \subseteq \mathbb{N}_G,\ |J| \leq 2s\} \tag{3.19}$$

For a given $J \subseteq \mathbb{N}_G$, $|J| = 2s$, we define $\mathbb{D}_J = \{v \in \mathbb{R}^p : \|v\| \leq 1, supp(v) \subseteq \mathcal{G}_J\}$ and note that $\mathbb{D}(2s) = \cup_{|J|=2s}\mathbb{D}_J$. For an $\epsilon > 0$ to be specified later, we construct an ϵ-net $\mathcal{A}$ of $\mathbb{D}_J$. Since $\sum_{g\in J} k_g \leq 2s\, k_{\max}$, it is possible to construct such a net $\mathcal{A}$ with cardinality at most $(1 + 2/\epsilon)^{s\, k_{\max}}$ (*Vershynin*, 2009).

We want a tail inequality for $M := \sup_{v\in\mathbb{D}_J} |v'\Delta v|$, where $\Delta = C - \Sigma$. Since $\mathcal{A}$ is an ϵ-cover of $\mathbb{D}_J$, for any $v \in \mathbb{D}_J$, there exists $v_0 \in \mathcal{A}$ such that $w = v - v_0$ satisfies $\|w\| \leq \epsilon$. Then

$$|v'\Delta v| = |(w + v_0)'\Delta(w + v_0)| \leq |w'\Delta w| + |v_0'\Delta v_0| + 2|v_0'\Delta w|$$

Taking supremum over all $v \in \mathbb{D}_J$, and noting that $w/\epsilon \in \mathbb{D}_J$, we obtain

$$M \leq \epsilon^2 M + \max_{v_0\in\mathcal{A}} |v_0'\Delta v_0| + \sup_{u,v\in\mathbb{D}_J} 2\epsilon|u'\Delta v| \tag{3.20}$$

To upper bound the third term, note that $(u + v)/2 \in \mathbb{D}_J$, and

$$2|u'\Delta v| \leq |(u + v)'\Delta(u + v)| + |u'\Delta u| + |v'\Delta v|$$

Hence

$$\sup_{u,v\in\mathbb{D}_J} 2\epsilon|u'\Delta v| \leq 4\epsilon M + \epsilon M + \epsilon M = 6\epsilon M$$

From equation (3.20), we now have

$$M \leq (1 - 6\epsilon - \epsilon^2)^{-1} \max_{v_0\in\mathcal{A}} |v_0'\Delta v_0|$$

Choosing $\epsilon > 0$ small enough so that $(1 - 6\epsilon - \epsilon^2) > 1/2$, we obtain

$$\begin{aligned}\mathbb{P}\left[\sup_{v\in\mathbb{D}_J} |v'\Delta v| > 2\eta\|\Sigma\|\right] &\leq \mathbb{P}\left[\max_{v_0\in\mathcal{A}} |v_0'\Delta v_0| > \eta\|\Sigma\|\right] \\ &\leq 2\,(1+2/\epsilon)^{s\,k_{\max}} \exp[-cn\min\{\eta,\eta^2\}]\end{aligned}$$

Taking supremum over $\begin{pmatrix} G \\ 2s \end{pmatrix} \leq (eG/2s)^{2s}$ choices of J, we get

$$\mathbb{P}\left[\sup_{v\in\mathbb{D}(2s)} |v'\Delta v| > 2\eta\|\Sigma\|\right] \leq 2\exp\left[-cn\min\{\eta,\eta^2\} + 2s\log(eG/2s) + 2s\,k_{\max}\log(1+2/\epsilon)\right] \quad (3.21)$$

In order to extend this deviation inequality to $cl\{conv\{\mathbb{D}(s)\}\}$, we note that any v in the convex hull of $\mathbb{D}(s)$ can be expressed as $v = \sum_{i=1}^m \alpha_i v_i$, where $v_1, \ldots, v_m$ are in $\mathbb{D}(s)$ and $0 \leq \alpha_i \leq 1$, $\sum \alpha_i = 1$. Then

$$|v'\Delta v| \leq \sum_{i=1}^m \sum_{j=1}^m \alpha_i\alpha_j |v_i'\Delta v_j|$$

Also, for every $i,\ j$, $(v_i + v_j)/2 \in \mathbb{D}(2s)$, and

$$|v_i'\Delta v_j| \leq \frac{1}{2}\left[|(v_i+v_j)'\Delta(v_i+v_j)| + |v_i'\Delta v_i| + |v_j'\Delta v_j|\right]$$

Hence

$$\sup_{v\in conv\{\mathbb{D}(s)\}} |v'\Delta v| \leq \sum_{i=1}^m \sum_{j=1}^m \alpha_i\alpha_j\ \frac{1}{2}[4+1+1] \sup_{v\in\mathbb{D}(2s)} |v'\Delta v|$$

Together with the continuity of quadratic forms, this implies

$$\sup_{v\in cl\{conv\{\mathbb{D}(s)\}\}} |v'\Delta v| \leq 3 \sup_{v\in\mathbb{D}(2s)} |v'\Delta v|$$

The result then readily follows from equation (3.21). □

3.8.2 Proof of Main Results

Proof of Proposition III.2. (a) Note that Σ is a $p(T-1) \times p(T-1)$ block Toeplitz matrix with $(i,j)^{th}$ block $(\Sigma_{ij})_{1 \le i,j \le (T-1)} := \Gamma(i-j)$, where $\Gamma(\ell)_{p \times p}$ is the auto-covariance function of lag ℓ for the zero-mean VAR(d) process (3.2), defined as $\Gamma(\ell) = \mathbb{E}[\mathbf{X}^t(\mathbf{X}^{\mathbf{t}-\ell})']$.

We consider the cross spectral density of the VAR(d) process (3.2)

$$f(\theta) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\ell\theta}, \ \theta \in [-\pi, \pi] \tag{3.22}$$

From standard results of spectral theory we know that $\Gamma(\ell) = \int_{-\pi}^{\pi} e^{i\ell\theta} f(\theta)\, d\theta$, for every ℓ.

We want to find a lower bound on the minimum eigenvalue of Σ, i.e., $\inf_{\|x\|=1} x'\Sigma x$. Consider an arbitrary $p(T-1)$-variate unit norm vector x, formed by stacking the p-tuples $x^1, \ldots, x^{T-1}$.

For every $\theta \in [-\pi, \pi]$ define $G(\theta) = \sum_{t=1}^{T-1} x^t\, e^{-it\theta}$ and note that

$$\begin{aligned} \int_{-\pi}^{\pi} G^*(\theta) G(\theta)\, d\theta &= \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} (x^t)'(x^\tau) \int_{-\pi}^{\pi} e^{i(t-\tau)\theta}\, d\theta \\ &= \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} (x^t)'(x^\tau)\ (2\pi\, \mathbf{1}_{\{t=\tau\}}) = 2\pi \sum_{t=1}^{T-1} (x^t)'(x^t) = 2\pi\ \|x\|^2 = 2\pi \end{aligned}$$

Also let $\mu(\theta)$ be the minimum eigenvalue of the Hermitian matrix $f(\theta)$. Following

Parter (1961) we have the result

$$\begin{aligned} x'\Sigma x &= \sum_{t=1}^{T-1}\sum_{\tau=1}^{T-1}(x^t)'\Gamma(t-\tau)x^\tau = \sum_{t=1}^{T-1}\sum_{\tau=1}^{T-1}(x^t)'\left(\int_{-\pi}^{\pi} e^{i(t-\tau)\theta}f(\theta)d\theta\right)x^\tau \\ &= \int_{-\pi}^{\pi}\left(\sum_{t=1}^{T-1}(x^t)'e^{it\theta}\right)\, f(\theta)\, \left(\sum_{\tau=1}^{T-1}x^\tau e^{-i\tau\theta}\right)d\theta = \int_{-\pi}^{\pi} G^*(\theta)\, f(\theta)\, G(\theta)\, d\theta \\ &\geq \int_{-\pi}^{\pi}\mu(\theta)\,(G^*(\theta)G(\theta))\;\, d\theta \geq \left(\min_{\theta\in(-\pi,\pi)}\mu(\theta)\right)\int_{-\pi}^{\pi} G^*(\theta)G(\theta)\, d\theta = 2\pi \min_{\theta\in(-\pi,\pi)}\mu(\theta) \end{aligned}$$

So $\Lambda_{min}(\Sigma) \geq 2\pi \min_{\theta\in(-\pi,\pi)}\mu(\theta)$. Since $\mathcal{A}(z) = I - A^1z - A^2z^2 - \ldots - A^dz^d$ is the (matrix-valued) characteristic polynomial of the VAR(d) model (3.2), we have the following representation of the spectral density (see eqn (9.4.23), *Priestley* (1981)):

$$f(\theta) = \frac{1}{2\pi}\sigma^2(\mathcal{A}(e^{-i\theta}))^{-1}(\mathcal{A}^*(e^{-i\theta}))^{-1}$$

Thus, $2\pi\mu(\theta) = 2\pi\Lambda_{min}(f(\theta)) = 2\pi/\Lambda_{max}(f(\theta)^{-1}) \geq \sigma^2/\left\|\mathcal{A}(e^{-i\theta})\right\|^2$. But $\left\|\mathcal{A}(e^{-i\theta})\right\| \leq 1+\sum_{t=1}^{d}\|A^t\|$ for every $\theta \in [-\pi,\pi]$. The result then follows at once from the standard matrix norm inequality (see e.g. *Golub and Van Loan*, 1996, Cor 2.3.2)

$$\|A^t\|_2 \leq \sqrt{\|A^t\|_1\|A^t\|_\infty} \leq \frac{\|A^t\|_1 + \|A^t\|_\infty}{2}\;\, t = 1,\ldots,d$$

where

$$\|A^t\|_1 = \max_{1\leq i\leq p}\sum_{j=1}^{p}|A^t_{ij}|,\;\; \|A^t\|_\infty = \max_{1\leq j\leq p}\sum_{i=1}^{p}|A^t_{ij}|$$

(b) The first part of the proposition ensures that $\Lambda_{min}(\Sigma) \geq \sigma^2\left[1+\frac{1}{2}(\mathbf{v}_{in}+\mathbf{v}_{out})\right]^{-2}$. If the replicates available from different panels are i.i.d, each row of the design matrix is independently and identically distributed according to a $N(\mathbf{0},\Sigma)$ distribution.

To show that RE(s, L) of (3.5) holds with high probability for sufficiently large

n, it is enough to show that

$$\min_{\substack{v \in \mathcal{C}(J,L)\backslash\{0\} \\ J \subset \mathbb{N}_{\bar{G}},\ |J| \leq s}} \frac{1}{n}\frac{\|\mathbf{X}v\|^2}{\|v\|^2} \geq \phi_{RE}^2 \tag{3.23}$$

holds with high probability, where the cone sets $\mathcal{C}(J, L)$ are defined as

$$\mathcal{C}(J, L) := \{v \in \mathbb{R}^{\bar{p}} : \sum_{g \notin J} \lambda_g \|v_{[g]}\| \leq L \sum_{g \in J} \lambda_g \|v_{[g]}\|\} \tag{3.24}$$

for all $J \subset \mathbb{N}_{\bar{G}}$ with $|J| \leq s$. Denote the ball of unit norm vectors in $\mathbb{R}^{\bar{p}}$ by $\mathbb{S}^{\bar{p}-1}$. By scale invariance of $\|\mathbf{X}v\|^2/n\|v\|^2$, it is enough to show that with high probability

$$\min_{\substack{v \in \mathbb{S}^{\bar{p}-1} \cap \mathcal{C}(J,L) \\ J \subset \mathbb{N}_{\bar{G}},\ |J| \leq s}} v'Cv \geq \phi_{RE}^2 \tag{3.25}$$

where $C = \mathbf{X}'\mathbf{X}/n$ is the sample Gram matrix.

By part (a), we already know that $v'\Sigma v \geq \Lambda_{\min}(\Sigma) > 0$ for all $v \in \mathbb{S}^{\bar{p}-1}$. So we only need to show that $|v'(C - \Sigma)v| \leq \Lambda_{\min}(\Sigma)/2$ with high probability, uniformly on the set

$$\bigcup_{J \subseteq \mathbb{N}_{\bar{G}}, |J| \leq s} \mathcal{C}(J, L) \cap \mathbb{S}^{\bar{p}-1} \tag{3.26}$$

The proof relies on two key parts. In the first part, we use an extremal representation to show that the above union of the cone sets sits within the closed convex hull of a suitably defined set of group s-sparse vectors. In particular, it follows from Lemma 3.8.4 that

$$\bigcup_{J \subseteq \mathbb{N}_{\bar{G}},\ |J| \leq s} \mathcal{C}(J, L) \cap \mathbb{S}^{\bar{p}-1} \subseteq (L' + 2)cl\{conv\{\mathbb{D}(s)\}\} \tag{3.27}$$

where $\mathbb{D}(s) = \{v \in \mathbb{R}^{\bar{p}} : \|v\| \leq 1,\ supp(v) \subseteq \mathcal{G}_J \text{ for some } J \subseteq \mathbb{N}_{\bar{G}},\ |J| \leq s\}$,

$L' = L\lambda_{\max}/\lambda_{\min}$ and $cl\{.\}$, $conv\{.\}$ respectively denote the closure and convex hull of a set.

The next part of the proof is an upper bound on the tail probability of $v'(C-\Sigma)v$, uniformly over all $v \in cl\{conv\{\mathbb{D}(s)\}\}$, presented in Lemma 3.8.5. In particular, setting $\eta = \Lambda_{\min}(\Sigma)/12\|\Sigma\|(2+L')^2$ in the above lemma yields

$$\mathbb{P}\left[\sup_{v\in(2+L')cl\{conv\{\mathbb{D}(s)\}\}} |v'(C-\Sigma)v| > \Lambda_{\min}(\Sigma)/2\right] \leq c_0 \exp[-c_1\, n] \tag{3.28}$$

for the proposed choice of n. Together with the lower bound on $\Lambda_{\min}(\Sigma)$ established in part (a), this concludes the proof. □

Proof of Theorem 3.4.1. Consider any solution $\hat{\beta}_R \in \mathbb{R}^q$ of the restricted regression

$$\underset{\beta\in\mathbb{R}^q}{\operatorname{argmin}} \frac{1}{2n}\left\|\mathbf{Y} - X_{(1)}\beta\right\|_2^2 + \lambda\sum_{g=1}^{s}\left\|\beta_{[g]}\right\|_2 \tag{3.29}$$

and set $\hat{\beta} = \left[\hat{\beta}_R' : \mathbf{0}_{1\times(p-q)}\right]'$. We show that such an augmented vector $\hat{\beta}$ satisfies the statements of Theorem 3.4.1 with high probability.

Let $\hat{u} = \hat{\beta}_{(1)} - \beta^0_{(1)} = \hat{\beta}_R - \beta^0_{(1)}$. In view of lemmas 3.8.1 and 3.8.3, it suffices to show that the following events happen with probability at least $1 - 4\,G^{1-\alpha}$:

$$\left\|\hat{u}_{[g]}\right\| < \delta_n \left\|\beta^0_{[g]}\right\|, \text{ for all } g \in S \tag{3.30}$$

$$\frac{1}{n}\left\|\left[X'\left(\epsilon - X_{(1)}\hat{u}\right)\right]_{[g]}\right\| \leq \lambda, \text{ for all } g \notin S \tag{3.31}$$

Note that, in view of Lemma 3.8.1, $\hat{u} = (C_{11})^{-1}\left(\frac{1}{\sqrt{n}}Z_{(1)} - \lambda\tau\right)$ for some $\tau \in \mathbb{R}^q$ with $\left\|\tau_{[g]}\right\| \leq 1$ for all $g \in S$, and $Z = \frac{1}{\sqrt{n}}X'\epsilon = \left[Z'_{(1)} : Z'_{(2)}\right]'$. Thus, for any $g \in S$,

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{u}_{[g]}\right\| > \delta_n\left\|\beta^0_{[g]}\right\|\right) &\leq \mathbb{P}\left(\left\|\left[(C_{11})^{-1}\left(\frac{1}{\sqrt{n}}Z_{(1)} - \lambda\tau\right)\right]_{[g]}\right\| > \delta_n\left\|\beta^0_{[g]}\right\|\right)\\ &\leq \mathbb{P}\left(\left\|\left[(C_{11})^{-1}Z_{(1)}\right]_{[g]}\right\| > \sqrt{n}\left[\delta_n\left\|\beta^0{}_{[g]}\right\| - \lambda\left\|\left[(C_{11})^{-1}\tau\right]_{[g]}\right\|\right]\right)\end{aligned}$$

Note that $V = (C_{11})^{-1} Z_{(1)} \sim N(\mathbf{0}, \sigma^2 (C_{11})^{-1})$. So $V_{[g]} \sim N(\mathbf{0}, \sigma^2 C_{11}^{[g][g]})$, where $\Sigma^{[g][g]} := (\Sigma^{-1})_{[g][g]}$. Also, by the second statement of lemma 3.8.2 we have $\mathbb{E}\left\|V_{[g]}\right\| \leq \sigma\sqrt{k_g}\sqrt{\left\|C_{11}^{[g][g]}\right\|}$. Therefore $\mathbb{P}\left(\left\|\hat{u}_{[g]}\right\| > \delta_n \left\|\beta^0_{[g]}\right\|\right)$ is bounded above by

$$\mathbb{P}\left(\left|\left\|V_{[g]}\right\| - \mathbb{E}\left\|V_{[g]}\right\|\right| > \sqrt{n}\left[\delta_n \left\|\beta^0{}_{[g]}\right\| - \lambda \left\|(C_{11})^{-1}\right\| \sqrt{s}\right] - \sigma\sqrt{k_g \left\|C_{11}^{[g][g]}\right\|}\right)$$
$$\leq 2\exp\left[-\frac{2}{\pi^2\sigma^2\|C_{11}^{[g][g]}\|}\left(\sqrt{n}\delta_n\|\beta^0_{[g]}\| - \sqrt{n}\lambda\|C_{11}^{-1}\|\sqrt{s} - \sigma\sqrt{k_g\|C_{11}^{[g][g]}\|}\right)^2\right]$$

For the proposed choice of δ_n, this expression is bounded above by $2\,G^{-\alpha}$.

Next, for any $g \notin S$, we get

$$\mathbb{P}\left(\frac{1}{n}\left\|\left[X'\left(\epsilon - X_{(1)}\hat{u}\right)\right]_{[g]}\right\| > \lambda\right)$$
$$\leq \mathbb{P}\left(\left\|\left[Z_{(2)} - C_{21}C_{11}^{-1}Z_{(1)}\right]_{[g]}\right\| > \sqrt{n}\lambda\left(1 - \left\|\left[C_{21}C_{11}^{-1}\tau\right]_{[g]}\right\|\right)\right)$$

Defining $W = Z_{(2)} - C_{21}C_{11}^{-1}Z_{(1)} \sim N(\mathbf{0}, \sigma^2(C_{22} - C_{21}C_{11}^{-1}C_{12}))$, the uniform irrepresentable condition implies that the above probability is bounded above by $\mathbb{P}\left(\left\|W_{[g]}\right\| > \sqrt{n}\lambda\eta\right)$.

It can then be seen that $W_{[g]} \sim N(\mathbf{0}, \sigma^2\bar{C}_{[g][g]})$, where $\bar{C} = C_{22} - C_{21}C_{11}^{-1}C_{12}$ denotes the Schur complement of C_{22}. As before, lemma 3.8.2 establishes that

$$\begin{aligned}\mathbb{P}\left(\left\|W_{[g]}\right\| > \sqrt{n}\lambda\eta\right) &\leq \mathbb{P}\left(\left|\left\|W_{[g]}\right\| - \mathbb{E}\left\|W_{[g]}\right\|\right| > \sqrt{n}\lambda\eta - \sigma\sqrt{k_g\|\bar{C}_{[g][g]}\|}\right) \\ &\leq 2\exp\left[-\frac{2}{\pi^2\|\sigma^2\bar{C}_{[g][g]}\|}\left(\sqrt{n}\lambda\eta - \sigma\sqrt{k_g\|\bar{C}_{[g][g]}\|}\right)^2\right],\end{aligned}$$

and the last probability is bounded above by $2G^{-\alpha}$ for the proposed choice of λ. The results in the proposition follow by considering the union bound on the two sets of the probability statements made across all $g \in \mathbb{N}_G$. □

3.8.3 Proof of results on ℓ_2-consistency

We first note that each of the p optimization problems in (3.4) is essentially a generic group lasso regression on n independent samples from a linear model $Y = X\beta^0 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{g=1}^{\bar{G}} \lambda_g \|\beta_{[g]}\| \tag{3.32}$$

where $\mathbf{Y}_{n\times 1} = \mathcal{X}_i^T$, $\mathbf{X}_{n\times \bar{p}} = [\mathcal{X}^1 : \cdots : \mathcal{X}^{T-1}]$, $\beta^0_{\bar{p}\times 1} = vec(A^{1:(T-1)}_{i:})$, $\{1, \ldots, \bar{p}\} = \cup_{g=1}^{\bar{G}} \mathcal{G}_g$, $\bar{p} = (T-1)p$, $\bar{G} = (T-1)G$ and $\lambda_g = \lambda w^t_{i,g}$. In Proposition III.3, we first establish the upper bounds on estimation error in the context of a generic group lasso penalized regression problem. The results for regular group NGC then readily follows by applying the above Proposition on the p separate regressions.

Recall the Restricted Eigenvalue assumption required for the derivation of ℓ_2 estimation and prediction error. Following *van de Geer and Bühlmann* (2009b), we introduce a slightly weaker notion called **Group Compatibility** (GC). For a constant $L > 0$ we say that GC(S, L) condition holds, if there exists a constant $\phi_{compatible} = \phi_{compatible}(S, L) > 0$ such that

$$\min_{\Delta \in \mathbb{R}^p \setminus \{\mathbf{0}\}} \left\{ \frac{\left(\sum_{g\in S} \lambda_g^2\right)^{1/2} \|X\Delta\|}{\sqrt{n} \sum_{g\in S} \lambda_g \|\Delta_{[g]}\|} : \sum_{g \notin S} \lambda_g \|\Delta_{[g]}\| \leq L \sum_{g \in S} \lambda_g \|\Delta_{[g]}\| \right\} \geq \phi_{compatible} \tag{3.33}$$

The fact that GC(S, L) holds whenever RE(s, L) is satisfied (and $\phi_{RE} \leq \phi_{compatible}$) follows at once from Cauchy Schwarz inequality. We shall derive upper bounds on the prediction and $\ell_{2,1}$ estimation error of group lasso estimates involving the compatibility constant. This notion will also be used later to connect the irrepresentable conditions to the consistency results of group lasso estimators.

Proposition III.3. *Suppose the GC condition* (3.33) *holds with* $L = 3$. *Choose* $\alpha > 0$

and denote $\lambda_{min} = \min_{1 \leq g \leq G} \lambda_g$. *If*

$$\lambda_g \geq \frac{2\sigma}{\sqrt{n}} \sqrt{\left\| C_{[g][g]} \right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right)$$

for every $g \in \mathbb{N}_G$, *then, the following statements hold with probability at least* $1 - 2G^{1-\alpha}$,

$$\frac{1}{n} \left\| X \left(\hat{\beta} - \beta^0 \right) \right\|^2 \leq \frac{16}{\phi^2_{compatible}} \sum_{g=1}^{s} \lambda_g^2 \tag{3.34}$$

$$\|\hat{\beta} - \beta^0\|_{2,1} \leq \frac{16}{\phi^2_{compatible}} \frac{\sum_{g=1}^{s} \lambda_g^2}{\lambda_{min}}. \tag{3.35}$$

If, in addition, RE(2s, 3) holds, then, with the same probability we get

$$\|\hat{\beta} - \beta^0\| \leq \frac{4\sqrt{10}}{\phi^2_{RE}(2s)} \frac{\sum_{g=1}^{s} \lambda_g^2}{\lambda_{min} \sqrt{s}}. \tag{3.36}$$

Proof of Proposition (III.3). Since $\hat{\beta}$ is a solution of the optimization problem (3.32), for all $\beta \in \mathbb{R}^p$, we have

$$\frac{1}{n} \|Y - X\hat{\beta}\|^2 + 2 \sum_{g=1}^{G} \lambda_g \|\hat{\beta}_{[g]}\| \leq \frac{1}{n} \|Y - X\beta\|^2 + 2 \sum_{g=1}^{G} \lambda_g \|\beta_{[g]}\|.$$

Plugging in $Y = X\beta^0 + \epsilon$, and simplifying the resulting equation, we get

$$\frac{1}{n} \|X(\hat{\beta} - \beta^0)\|^2 \leq \frac{1}{n} \|X(\beta - \beta^0)\|^2 + \frac{2}{n} \sum_{g=1}^{G} \left\| (X'\epsilon)_{[g]} \right\| \left\| (\hat{\beta} - \beta)_{[g]} \right\|$$
$$+2 \sum_{g=1}^{G} \lambda_g \left(\|\beta_{[g]}\| - \|\hat{\beta}_{[g]}\| \right).$$

Fix $g \in \mathbb{N}_G$ and consider the event $\mathcal{A}_g = \left\{ \epsilon \in \mathbb{R}^n : \frac{2}{n} \left\| (X'\epsilon)_{[g]} \right\| \leq \lambda_g \right\}$. Note that

$Z = \frac{1}{\sqrt{n}} X'\epsilon \sim N(\mathbf{0}, \sigma^2 C)$. So $Z_{[g]} \sim N(\mathbf{0}, \sigma^2 C_{[g][g]})$. Then,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{A}_g^c\right) &= \mathbb{P}\left(\left\|Z_{[g]}\right\| > \frac{1}{2}\lambda_g\sqrt{n}\right) \\
&\leq \mathbb{P}\left(\left|\left\|Z_{[g]}\right\| - \mathbb{E}\left\|Z_{[g]}\right\|\right| > \frac{\lambda_g\sqrt{n}}{2} - \sigma\sqrt{k_g}\sqrt{\left\|C_{[g][g]}\right\|}\right),
\end{aligned}$$

where the last inequality follows from the second statement of Lemma 3.8.2. Now, let $x_g = \frac{\lambda_g\sqrt{n}}{2} - \sigma\sqrt{k_g}\sqrt{\left\|C_{[g][g]}\right\|}$. Then, for $x_g > 0$, if

$$2\exp\left(-\frac{2\,x_g^2}{\pi^2\sigma^2\left\|C_{[g][g]}\right\|}\right) \leq 2\,G^{-\alpha},$$

we get

$$\mathbb{P}\left(\mathcal{A}_g^c\right) \leq 2G^{-\alpha}.$$

But this happens if,

$$\sqrt{2}x_g \geq \sqrt{\alpha\,\log\,G}\pi\sigma\sqrt{\left\|C_{[g][g]}\right\|},$$

which is ensured by the proposed choice of λ_g.

Next, define $\mathcal{A} := \cap_{g=1}^G \mathcal{A}_g$. Then, $\mathbb{P}(\mathcal{A}) \geq 1 - 2G^{1-\alpha}$, and on the event $\mathcal{A}$, we have, for all $\beta \in \mathbb{R}^p$,

$$\begin{aligned}
\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|^2 + \sum_{g=1}^G \lambda_g \left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| &\leq \frac{1}{n}\|X(\beta - \beta^0)\|^2 \\
&\quad + 2\sum_{g=1}^G \lambda_g \left(\left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| + \left\|\beta_{[g]}\right\| - \left\|\hat{\beta}_{[g]}\right\|\right).
\end{aligned}$$

Note that $\left(\left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| + \left\|\beta_{[g]}\right\| - \left\|\hat{\beta}_{[g]}\right\|\right)$ vanishes if $g \notin S$ and is bounded above by $\min\{2\left\|\beta_{[g]}\right\|, 2\left(\left\|\beta_{[g]} - \hat{\beta}_{[g]}\right\|\right)\}$ if $g \in S$.

This leads to the following sparsity oracle inequality, for all $\beta \in \mathbb{R}^p$,

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|^2 + \sum_{g=1}^{G} \lambda_g \left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| \leq \frac{1}{n}\|X(\beta-\beta^0)\|^2 + 4\sum_{g\in S} \lambda_g \min\left\{\left\|\beta_{[g]}\right\|, \left\|\beta_{[g]} - \hat{\beta}_{[g]}\right\|\right\}. \quad (3.37)$$

The sparsity oracle inequality (3.37) with $\beta = \beta^0$, and $\Delta := \hat{\beta} - \beta^0$ leads to the following two useful bounds on the prediction and $\ell_{2,1}$-estimation errors:

$$\frac{1}{n}\|X\Delta\|^2 \leq 4\sum_{g\in S} \lambda_g \left\|\Delta_{[g]}\right\| \quad (3.38)$$

$$\sum_{g\notin S} \lambda_g \left\|\Delta_{[g]}\right\| \leq 3\sum_{g\in S} \lambda_g \left\|\Delta_{[g]}\right\|. \quad (3.39)$$

Now, assume the group compatibility condition 3.33 holds. Then,

$$\frac{1}{n}\|X\Delta\|^2 \leq 4\sum_{g\in S} \lambda_g \left\|\Delta_{[g]}\right\| \leq \sqrt{\sum_{g\in S}\lambda_g^2}\,\frac{\|X\Delta\|}{\sqrt{n}}\,\frac{4}{\phi_{compatible}}, \quad (3.40)$$

which implies the first inequality of proposition III.3. The second inequality follows from

$$\begin{aligned}\lambda_{min}\left\|\hat{\beta}-\beta\right\|_{2,1} &\leq \sum_{g=1}^{G}\lambda_g\left\|\Delta_{[g]}\right\| \leq 4\sum_{g\in S}\lambda_g\left\|\Delta_{[g]}\right\| \\ &\leq 4\sqrt{\sum_{g\in S}\lambda_g^2}\,\frac{\|X\Delta\|}{\sqrt{n}}\,\frac{1}{\phi_{compatible}} \leq \frac{16}{\phi^2_{compatible}}\sum_{g\in S}\lambda_g^2,\end{aligned}$$

where the last step uses (3.40).

The proof of the last inequality of proposition III.3, i.e., the upper bound on ℓ_2 estimation error under $RE(2s)$, is the same as in Theorem 3.1 in *Lounici et al.* (2011) and is omitted. □

Proof of Proposition III.1. Applying the ℓ_2-estimation error of (3.36) on the i^{th} group

lasso regression problem of regular group NGC, we have

$$\|\hat{A}_{i:}^{1:T-1} - A_{i:}^{1:T-1}\| \leq \frac{4\sqrt{10}}{\phi_{RE}^2(2s_i)} \frac{\sum_{g=1}^{s_i} \lambda_g^2}{\lambda_{\min}\sqrt{s_i}} \leq \frac{4\sqrt{10}}{\phi_{RE}^2(2s_{\max})} \frac{\lambda_{\max}}{\lambda_{\min}} \sqrt{s_i}$$

with probability at least $1 - 2\bar{G}^{1-\alpha}$. Combining the bounds for all $i = 1, \ldots, p$ and noting that $s = \sum_{i=1}^{p} s_i$, we have the required result. □

3.8.4 Irrepresentable assumptions and consistency

In this subsection, we discuss two results involving the compatibility and irrepresentable conditions for group lasso. We first show that a stronger version of the uniform irrepresentable assumption implies the group compatibility (3.33), and hence, consistency in $\ell_{2,1}$ norm. Next we argue that a weaker version of the irrepresentable assumption is indeed necessary for the direction consistency of the group lasso estimates. These results generalize analogous properties of lasso (*van de Geer and Bühlmann*, 2009b; *Zhao and Yu*, 2006) to the group penalization framework. The proofs are given under a special choice of tuning parameter $\lambda_g = \lambda\sqrt{k_g}$. Similar results can be derived for the general choice of λ_g, although their presentation is more involved.

Proposition III.4. *Assume uniform irrepresentable condition* (3.13) *holds with* $\eta \in (0, 1)$, *and* $\Lambda_{min}(C_{11}) > 0$. *Then group compatibility(S, L)* (3.33) *condition holds whenever* $L < \frac{1}{1-\eta}$.

Proof. First note that with the above choice of λ_g the Group Compatibility (S, L) condition simplifies to

$$\phi_{compatible} := \min_{\Delta \in \mathbb{R}^p \setminus \{\mathbf{0}\}} \left\{ \frac{\sqrt{q}\|X\Delta\|}{\sqrt{n}\sum_{g \in S} \sqrt{k_g}\|\Delta_{[g]}\|} : \sum_{g \notin S} \sqrt{k_g}\|\Delta_{[g]}\| \leq L \sum_{g \in S} \sqrt{k_g}\|\Delta_{[g]}\| \right\} > 0 \tag{3.41}$$

Also, the uniform irrepresentable condition guarantees that there exists $0 < \eta < 1$ such that $\forall \tau \in \mathbb{R}^q$ with $\|\tau\|_{2,\infty} = \max_{1 \le g \le s} \|\tau_{[g]}\|_2 \le 1$, we have,

$$\frac{1}{\sqrt{k_g}} \left\| \left[C_{21} \left(C_{11} \right)^{-1} K^0 \tau \right]_{[g]} \right\|_2 < 1 - \eta \ \forall g \notin S$$

Here $K^0 = K/\lambda$ is a $q \times q$ block diagonal matrix with s diagonal blocks $\sqrt{k_1}\, \mathbf{I}_{k_1 \times k_1}, \ldots, \sqrt{k_s}\, \mathbf{I}_{k_s \times k_s}$. Define

$$\Delta^0 := \underset{\Delta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{X}\Delta\|_2^2 : \ \sum_{g \in S} \sqrt{k_g} \|\Delta_{[g]}\|_2 = 1, \ \sum_{g \notin S} \sqrt{k_g} \|\Delta_{[g]}\|_2 \le L \right\} \tag{3.42}$$

Note that $\frac{1}{n}\|\mathbf{X}\Delta^0\|_2^2 = \phi^2_{compatible}/q$, and introduce two Lagrange multipliers λ and λ' corresponding to the equality and inequality constraints for solving the optimization problem in (3.42). Also, partition $\Delta^0 = \left[\Delta^0_{(1)} : \Delta^0_{(2)} \right]$ and $\mathbf{X} = \left[\mathbf{X}_{(1)} : \mathbf{X}_{(2)} \right]$ into signal and nonsignal parts as in (3.10). The first q linear equations of the KKT conditions imply that there exists $\tau^0 \in \mathbb{R}^q$ such that

$$C_{11}\Delta^0_{(1)} + C_{12}\Delta^0_{(2)} \ = \ \lambda K^0 \tau^0 \tag{3.43}$$

and, for every $g \in S$,

$$\begin{aligned} \tau^0_{[g]} \ &= \ D(\Delta^0_{[g]}) \text{ if } \Delta^0_{[g]} \neq \mathbf{0} \\ \|\tau^0_{[g]}\|_2 \ &\le \ 1 \text{ if } \Delta^0_{[g]} = \mathbf{0} \end{aligned}$$

It readily follows that $(\tau^0)^T K^0 \Delta^0_{(1)} = \sum_{g \in S} \sqrt{k_g} \|\Delta^0_{[g]}\|_2 = 1$.
Multiplying both sides of (3.43) by $(\Delta^0_{(1)})^T$ we get

$$\left(\Delta^0_{(1)}\right)^T C_{11} \Delta^0_{(1)} + \left(\Delta^0_{(1)}\right)^T C_{12} \Delta^0_{(2)} = \lambda \tag{3.44}$$

Also, (3.43) implies

$$\Delta^0_{(1)} + (C_{11})^{-1} C_{12} \Delta^0_{(2)} = \lambda (C_{11})^{-1} K^0 \tau^0 \tag{3.45}$$

Multiplying both sides of the equation by $(K^0\tau^0)^T = (\tau^0)^T K^0$ we obtain

$$1 = -\left(\tau^0\right)^T K^0 (C_{11})^{-1} C_{12} \Delta^0_{(2)} + \lambda \left(K^0\tau^0\right)^T (C_{11})^{-1} \left(K^0\tau^0\right) \tag{3.46}$$

Note that the absolute value of the first term,

$$\left| \sum_{g \notin S} \left(\Delta^0_{[g]}\right)^T \left[C_{21} (C_{11})^{-1} K^0 \tau^0 \right]_{[g]} \right|, \tag{3.47}$$

is bounded above by

$$(1-\eta) \left(\sum_{g \notin S} \sqrt{k_g} \|\Delta^0_{[g]}\|_2 \right) \leq (1-\eta) L \tag{3.48}$$

by virtue of the uniform irrepresentable condition and the Cauchy-Schwartz inequality.

Assuming the minimum eigenvalue of C_{11}, i.e., $\Lambda_{min}(C_{11})$, is positive and considering $\|K^0\tau^0\|_2 \leq \sqrt{q}$, the second term is at most $\lambda q / \Lambda_{min}(C_{11})$. So (3.46) implies

$$1 \leq (1-\eta) L + \frac{\lambda q}{\Lambda_{min}(C_{11})} \tag{3.49}$$

In particular, $\lambda \geq \Lambda_{min}(C_{11}) (1 - (1-\eta)L) / q$ is positive whenever $L < 1/(1-\eta)$.

Next, multiply both sides of (3.45) by $(\Delta^0_{(2)})^T C_{21}$ to get

$$\left(\Delta^0_{(2)}\right)^T C_{21} \Delta^0_{(1)} + \left(\Delta^0_{(2)}\right)^T C_{21} (C_{11})^{-1} C_{(12)} \Delta^0_{(2)} = \lambda \left(\Delta^0_{(2)}\right)^T C_{21} (C_{11})^{-1} K^0 \tau^0 \tag{3.50}$$

Using the upper bound in (3.48), the right hand side is at least $-\lambda(1-\eta)L$.

Also a simple consequence of the block inversion formula of the non-negative definite matrix C guarantees that the matrix $C_{22} - C_{21}\left(C_{11}\right)^{-1} C_{12}$ is non-negative definite. Hence,

$$\begin{aligned} & \left(\Delta^0_{(2)}\right)^T \left[C_{22} - C_{21}\left(C_{11}\right)^{-1} C_{12}\right] \Delta^0_{(2)} \geq 0 \\ \text{and} \quad & \left(\Delta^0_{(2)}\right)^T C_{22}\Delta^0_{(2)} \geq \left(\Delta^0_{(2)}\right)^T C_{21}\left(C_{11}\right)^{-1} C_{12}\Delta^0_{(2)} \end{aligned}$$

Putting all the pieces together we get

$$\begin{aligned} \phi^2_{compatible}/q &= \frac{1}{n}\|\mathbf{X}\Delta^0\|_2^2 \\ &= {\Delta^0_{(1)}}^T C_{11}\Delta^0_{(1)} + 2{\Delta^0_{(2)}}^T C_{21}\Delta^0_{(1)} + {\Delta^0_{(2)}}^T C_{22}\Delta^0_{(2)} \\ &= \lambda + {\Delta^0_{(2)}}^T C_{21}\Delta^0_{(1)} + {\Delta^0_{(2)}}^T C_{22}\Delta^0_{(2)} \text{ , by (3.44)} \\ &\geq \lambda - \lambda(1-\eta)L \text{ , by (3.50)} \\ &= \lambda(1-(1-\eta)L) \end{aligned}$$

Plugging in the lower bound for λ we obtain the result; namely,

$$\phi^2_{compatible} = \Lambda_{min}(C_{11})\left(1-(1-\eta)L\right)^2 > 0$$

for any $L < \frac{1}{1-\eta}$. □

In this subsection we investigate the necessity of irrepresentable assumptions for direction consistency of group lasso estimates. To this end we first introduce the notion of weak irrepresentability.

For a q-dimensional vector τ define the stacked direction vector $\underbrace{\tilde{D}(\tau)}_{q\times 1} = [\underbrace{D(\tau_{[1]})'}_{k_1\times 1}, \ldots, \underbrace{D(\tau_{[s]})'}_{k_s\times 1}]'$.

Weak Irrepresentable Condition is satisfied if

$$\frac{1}{\lambda_g}\left\|\left[C_{21}(C_{11})^{-1}K\tilde{D}(\beta^0_{(1)})\right]_{[g]}\right\| \leq 1, \ \forall g \notin S = \{1,\ldots,s\} \tag{3.51}$$

We argue the necessity of weak irrepresentable condition for group sparsity selection and direction consistency under two regularity conditions on the design matrix, as $n,\, p \to \infty$:

(A1) The minimum eigenvalue of the signal part of the Gram matrix, viz. $\Lambda_{min}(C_{11})$, is bounded away from zero.

(A2) The matrices C_{21} and C_{22} are bounded above in spectral norm.

As in the last proposition, we set $\lambda_g = \lambda\sqrt{k_g}$ and $K^0 = K/\lambda$. Suppose that the weak irrepresentable condition does not hold, i.e., for some $g \notin S$ and $\xi > 0$, we have,

$$\frac{1}{\sqrt{k_g}} \left\| \left[C_{21}(C_{11})^{-1} K^0 \tilde{D}(\beta^0_{(1)}) \right]_{[g]} \right\| > 1 + \xi$$

for infinitely many n. Also suppose that there exists a sequence of positive reals $\delta_n \to 0$ such that the event

$$E_n := \{ \|D(\hat{\beta}_{[g]}) - D(\beta_{[g]})\|_2 < \delta_n,\ \forall g \in S,\ \text{and}\ \hat{\beta}_{[g]} = \mathbf{0}\, \forall\, g \notin S \}$$

satisfies $\mathbb{P}(E_n) \to 1$ as $p,\, n \to \infty$.

Note that for large enough n so that $\delta_n < \min_g \|D(\beta_{[g]})\|$, we have $\hat{\beta}_{[g]} \neq \mathbf{0},\, \forall\, g \in S$ on the event E_n.

Then, as in the proof of Theorem 3.4.1, we have, on the event E_n,

$$\hat{\mathbf{u}} = (C_{11})^{-1} \left[\frac{1}{\sqrt{n}} Z_{(1)} - \lambda K^0 \tilde{D}(\hat{\beta}_{(1)}) \right] \tag{3.52}$$

$$\text{and} \quad \frac{1}{n} \left\| \left[\mathbf{X}_{(2)}{}^T (\epsilon - \mathbf{X}_{(1)}\hat{u}) \right]_{[g]} \right\| \leq \lambda\sqrt{k_g},\ \forall g \notin S \tag{3.53}$$

Substituting the value of $\hat{u}$ from (3.52) in (3.53), we have, on the event E_n,

$$\frac{1}{\sqrt{n}} \left\| \left[\mathbf{Z}_{(2)} - C_{21}(C_{11})^{-1}\mathbf{Z}_{(1)} + \lambda\sqrt{n} C_{21}(C_{11})^{-1} K^0 \tilde{D}(\hat{\beta}_{(1)}) \right]_{[g]} \right\| \leq \lambda\sqrt{k_g},$$

which implies that

$$\left\| \left[Z_{(2)} - C_{21} \left(C_{11} \right)^{-1} Z_{(1)} \right]_{[g]} \right\| \\ \geq \lambda \sqrt{n} \sqrt{k_g} \left[\frac{1}{\sqrt{k_g}} \left\| \left[C_{21} (C_{11})^{-1} K^0 \tilde{D}(\hat{\beta}_{(1)}) \right]_{[g]} \right\| - 1 \right]. \quad (3.54)$$

Now note that for large enough n, if $\|C_{21}\|$ is bounded above, direction consistency guarantees that the expression on the right is larger than

$$\frac{1}{2} \lambda \sqrt{n} \sqrt{k_g} \left[\frac{1}{\sqrt{k_g}} \left\| \left[C_{21} (C_{11})^{-1} K^0 \tilde{D}(\beta_{(1)}) \right]_{[g]} \right\| - 1 \right]$$

which in turn is larger than $\frac{1}{2} \lambda \sqrt{n} \sqrt{k_g} \, \xi$, in view of the weak irrepresentable condition.

This contradicts $\mathbb{P}(E_n) \to 1$, since the left-hand side of (3.54) corresponds to the norm of a zero mean Gaussian random variable with bounded variance structure $[C_{22} - C_{21}(C_{11})^{-1} C_{12}]_{[g][g]}$ while $\lambda \sqrt{n} \sqrt{k_g}$ diverges with $\sqrt{\log G}$.

3.8.5 Thresholding Group Lasso Estimates.

Proof of Theorem 3.4.2. We use the notations developed in the proof of Proposition III.3. First note that, (ii) follows directly from Theorem 3.4.1. For (i), since the falsely selected groups are present after the initial thresholding, we get $\|\hat{\beta_{[g]}}\| > 4\lambda$ for every such group. Next, we obtain an upper bound for the number of such groups. Specifically, denoting $\Delta = \hat{\beta} - \beta^0$, we get

$$\left| \hat{S} \backslash S \right| \leq \frac{\|\hat{\beta}_{S^c}\|_{2,1}}{4\lambda} = \frac{\sum_{g \notin S} \|\Delta_{[g]}\|}{4\lambda}. \quad (3.55)$$

Next, note that from the sparsity oracle inequality (3.38), the following holds on

the event $\mathcal{A}$,

$$\sum_{g \notin S} \|\Delta_{[g]}\| \leq 3 \sum_{g \in S} \|\Delta_{[g]}\|$$

It readily follows that

$$4 \sum_{g \notin S} \|\Delta_{[g]}\| \leq 3\|\Delta\|_{2,1} \leq \frac{48}{\phi^2}\, s\lambda$$

where the last inequality follows from the $\ell_{2,1}$-error bound of (3.35). Using this inequality together with (3.55) gives the result. □

CHAPTER IV

Regularized Estimation in Sparse High-dimensional Time Series Models

4.1 Introduction

Recent advances in information technology have made high-dimensional time series datasets increasingly common in numerous scientific and socio-economic applications. Examples include structural analysis and forecasting with a large number of macroeconomic variables (*De Mol et al.*, 2008), reconstruction of gene regulatory networks from time course microarray data (*Michailidis and d'Alché Buc*, 2013), portfolio selection and volatility matrix estimation in finance (*Fan et al.*, 2011) and studying coactivation networks in human brains using task based or resting state fMRI data (*Smith*, 2012). These applications require analyzing a large number of temporally observed variables using small to moderate sample sizes (number of time points). Meaningful inference in such situations is often impossible without imposing some lower dimensional structural assumption on the data generating mechanism. The most common structural assumption is that of sparsity on the model parameter space. In high-dimensional regression problems, the notion of sparsity is often incorporated in the estimation procedure by ℓ_1-regularization (*Bickel et al.*, 2009) procedures like lasso, while for covariance matrix estimation problems, sparsity is

enforced via hard thresholding (*Bickel and Levina*, 2008).

The theoretical properties of such regularized estimates under high-dimensional scaling has been the topic of numerous studies over the last few years, under the key assumption that the samples are independent and identically distributed (i.i.d). On the other hand, theoretical analysis of these estimates in a time series context, where the data exhibit temporal and cross-sectional dependence, is rather incomplete. A central challenge in analyzing regularized estimation problems in high-dimensional time series is to quantify the dependence present in the data and its effect on the accuracy of the estimation procedures. In classical asymptotic analysis, this is typically achieved by assuming some mixing condition on the underlying stochastic process. Although suitable for studying limiting behavior of the estimates, mixing conditions are often hard to verify even for standard processes. A more recent approach (*Lam and Souza*, 2013; *Chen et al.*, 2013) is to impose some decay assumption on a functional dependence measure (*Wu*, 2005) of the underlying stationary, causal processes. Despite the intuitive appeal and nice theoretical properties of this functional dependence measure, the decay assumptions often lead to restrictions on the model parameters. Hence, the objective of this study is to examine regularized estimation problems in high-dimensional stationary time series models under sparsity constraints.

Towards this goal, we adopt a novel, non-asymptotic approach to deal with dependence in high-dimensional time series. Our approach is based on ***stability***, a key notion in classical time series analysis and systems theory. For a covariance-stationary process, we introduce a measure of stability using the extreme eigenvalues of its spectral density and show that this measure can be used to capture the effect of dependence on the accuracy of regularized estimates. In particular, we derive non-asymptotic error bounds in three important and widely applicable estimation problems - (a) stochastic regression with serially correlated errors, (b) transition matrix estimation in large vector autoregressive (VAR) models, (c) large covariance matrix

estimation from temporally observed data. In all three problems, we establish that the effect of dependence is minimal as long as the underlying processes are stable. The estimates enjoy nearly the same convergence rates as in the i.i.d. case, with an additional "price" which depends on the stability measure of the process and captures the effect of dependence. Next, we outline the three problems addressed and summarize the contributions of this work.

Stochastic Regression. We start with the problem of stochastic regression with serially correlated errors - a canonical problem in time series analysis (*Hamilton*, 1994). A linear regression model of the form

$$y^t = \langle \beta^*, X^t \rangle + \epsilon^t, \ t = 1, \ldots, n \tag{4.1}$$

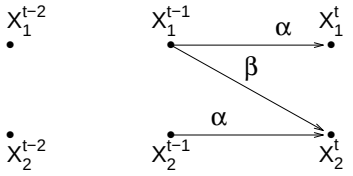
is considered, where the p-dimensional ($n \ll p$) predictors $\{X^t\}$ and the errors $\{\epsilon^t\}$ are generated according to independent, centered, Gaussian stationary processes. Under a sparsity assumption on β^*, we study the properties of the lasso estimate

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \|Y - \mathcal{X}\beta\|^2 + \lambda_n \|\beta\|_1 \tag{4.2}$$

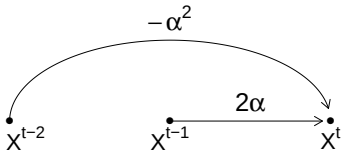
where $Y = [y^n : \ldots : y^1]'$, $\mathcal{X} = [X^n : \ldots : X^1]'$ and $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. Theoretical properties of lasso have been studied for fixed design regression $Y = \mathcal{X}\beta^* + E$, with $E = [e^n : \ldots : e^1]'$, by several authors (*Bickel et al.*, 2009; *Loh and Wainwright*, 2012; *Negahban et al.*, 2012). All the aforementioned papers established consistency of lasso in high-dimensional regime under some form of restricted eigenvalue (RE) or restricted strong convexity (RSC) assumption on $S = \mathcal{X}'\mathcal{X}/n$ and suitable deviation conditions on $\mathcal{X}'E/N$. For a fixed design matrix, verifying RE-type conditions is NP-hard (*Dobriban and Fan*, 2013). For random design regression, these assumptions are known to hold with high probability, as long as the samples are i.i.d. (*Raskutti et al.*, 2010; *Rudelson and Zhou*, 2013). It is not clear, however, whether these conditions

are satisfied with high probability when the observations are dependent. For instance, *Loh and Wainwright* (2012) and *Negahban and Wainwright* (2011) have shown that RE/RSC and deviation conditions are satisfied with high probability if the predictors $\{X^t\}$ are generated according to a Gaussian VAR(1) process $X^t = A_1 X^{t-1} + \xi^t$ with $\|A_1\| < 1$, where $\|.\|$ denotes the operator norm of a matrix. In Figure 4.1 and Lemma 4.8.4, we show that the condition $\|A_1\| < 1$ is **very restrictive** and fails to hold beyond a limited subclass of stable VAR(1) processes. More importantly, this condition is violated by *all* VAR(d) models, whenever $d > 1$, as shown in Figure 4.1. A major contribution of this work is to establish the validity of RE and deviation conditions for a large class of stationary Gaussian processes $\{X^t\}$ and $\{\epsilon^t\}$, even when the errors are serially correlated. The results crucially rely on the proposed measure of stability and use a blend of ideas from spectral theory of multivariate time series, convex geometry and non-asymptotic random matrix theory. An important consequence of these results is to ensure that consistent estimation with lasso is possible in high-dimensional stochastic regression in the presence of serially correlated errors, as long as the underlying processes are stable.

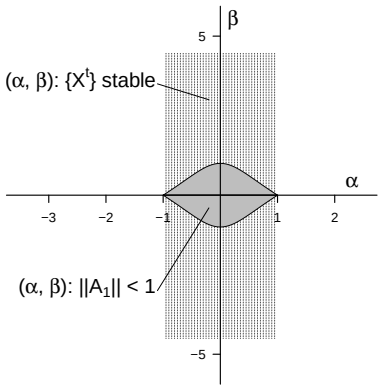
Vector Autoregression. Next, we address the problem of transition matrix estimation in high-dimensional sparse vector autoregressive models (VAR). Vector Autoregression (VAR) represents a popular class of time series models in applied macroeconomics and finance, widely used for structural analysis and simultaneous forecasting of a number of temporally observed variables (*Sims*, 1980; *Bernanke et al.*, 2005a; *Stock and Watson*, 2005). Unlike structural models, VAR provides a broad framework for capturing complex temporal and cross-sectional interrelationship among the time series (*Bańbura et al.*, 2010). In addition to economics, VAR models have been instrumental in linear system identification problems in control theory (*Kumar and Varaiya*, 1986), while more recently, they have become standard tools in functional genomics for reconstruction of regulatory networks (*Lozano et al.*, 2009b; *Shojaie and*



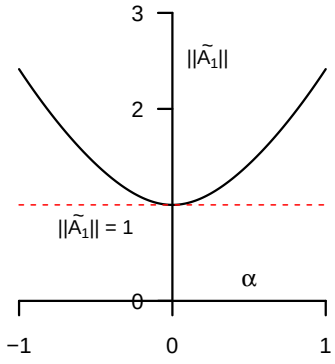
(a) VAR(1) model with $p = 2$



(b) VAR(2) model with $p = 1$



(c) Stability and $\|A_1\| < 1$



(d) Stability and $\|\tilde{A}_1\| < 1$

Figure 4.1: In the left panel, we consider a VAR(1) model with $p = 2$, $X^t = A_1 X^{t-1} + \epsilon^t$, where $A_1 = [\alpha\ 0; \beta\ \alpha]$. The unbounded set (dotted) denotes the values of (α, β) for which the process is stable. The bounded region (solid) represents the VAR models that satisfy $\|A_1\| < 1$. In the right panel, we consider a VAR(2) model with $p = 1$, $X^t = 2\alpha X^{t-1} - \alpha^2 X^{t-2} + \epsilon^t$. Equivalent formulation of this model as VAR(1) is: $Y^t = \tilde{A}_1 Y^{t-1} + \tilde{\epsilon}^t$, where $Y^t = [X^t, X^{t-1}]'$, $\tilde{A}_1 = [2\alpha\ \ -\alpha^2; 1\ 0]$, and $\tilde{\epsilon}^t = [\epsilon^t, 0]'$. The model is stable whenever $|\alpha| < 1$ but $\|\tilde{A}_1\|$ is always greater than or equal to 1.

Michailidis, 2010b; *Fujita et al.*, 2007b) and in neuroscience for understanding effective connectivity patterns between brain regions (*Smith*, 2012; *Friston*, 2009; *Seth et al.*, 2013).

Formally, for a p-dimensional vector-valued stationary time series $\{X^t\} = \{(X_1^t, \ldots, X_p^t)\}$,

a VAR model of lag d (VAR(d)) with serially uncorrelated Gaussian errors takes the form

$$X^t = A_1 X^{t-1} + \ldots + A_d X^{t-d} + \epsilon^t,\ \epsilon^t \overset{i.i.d}{\sim} N(\mathbf{0}, \Sigma_\epsilon) \tag{4.3}$$

where $A_1, \ldots, A_d$ are $p \times p$ matrices and ϵ^t is a p-dimensional vector of possibly correlated innovation shocks. The main objective in VAR models is to estimate the transition matrices $A_1, \ldots, A_d$, together with the order of the model d, based on realizations $\{X^0, X^1, \ldots, X^T\}$. The structures of the transition matrices provide insight into the complex temporal relationships amongst the p time series and lead to efficient forecasting strategies.

VAR estimation is a natural high-dimensional problem because the dimensionality of the parameter space (dp^2) grows quadratically with p. For example, estimating a VAR(10) model with $p = 10$ time series requires estimating $dp^2 = 1000$ parameters. However, a comparable number of stationary observations are rarely available in practice. In the low dimensional setting, VAR estimation is carried out by reformulating it as a multivariate regression problem (*Lütkepohl*, 2005). Under high-dimensional scaling and sparsity assumptions on the transition matrices, a natural strategy is to resort to ℓ_1-penalized least squares or log-likelihood based methods (*Song and Bickel*, 2011; *Davis et al.*, 2012). Compared to stochastic regression, the analysis of large VAR problems requires addressing two important issues. First, since the response variable is multivariate, the choice of the loss function (least squares, negative log-likelihood) plays an important role in forecasting problems, especially when the error process has correlated components. Second, correlation of the error process with the process of predictors $\text{Cov}(X^{t-1}, \epsilon^{t-1}) \neq 0$ makes the theoretical analysis more involved. Existing work on high-dimensional VAR models requires stringent assumptions on the dependence structure (*Song and Bickel*, 2011), or on the transition matrix (*Negahban and Wainwright*, 2011), which are violated by many stable VAR models, as discussed above. Our results show that consistent estimation is possible with both ℓ_1-penalized

least squares and log-likelihood based estimates under high-dimensional scaling for *any* stable VAR models. As in the case of stochastic regression, we establish the validity of suitable restricted eigenvalue and deviation conditions using the stability measures introduced in our work. The results rely on some novel techniques involving the spectral properties of the predictor and the error process to handle the intricate dependence structure (see Proposition IV.10).

Covariance Estimation. The third problem considered is that of sparse covariance matrix estimation by thresholding, originally proposed by *Bickel and Levina* (2008) and studied further by *Cai and Liu* (2011); *Cai and Zhou* (2012b,a). High-dimensional covariance estimation is useful in finance for analyzing large volatility matrices (*Fan et al.*, 2011), in neuroscience for studying functional connectivity amongst different regions of human brain (*Smith*, 2012). The theoretical works mentioned above assume that the samples are independent. In recent work, *Chen et al.* (2013) developed an asymptotic theory in the time series context under a suitable decay assumption on the functional dependence measure of the stationary, causal process. Our results do not require specific decay assumptions on the temporal dependence and are applicable to non-causal processes. We assume that the data $\{X^t\}$, for $t = 1, \ldots, n$, were generated according to a stationary Gaussian process with sparse covariance matrix $\Gamma_X(0) = \mathbb{E}\left[(X^1)(X^1)'\right]$. Under the stability assumption, we establish consistency of a thresholded estimate under operator and Frobenius norms. The convergence rates are the same as those obtained for independent samples (*Bickel and Levina*, 2008), modulo a "price" of dependence expressed by its measure of stability.

The rest of the chapter is organized as follows. In Section 4.2 we introduce the measure of stability, discuss its properties for stable, invertible ARMA systems and present some deviation inequalities, used in the subsequent analyses. In Section 4.3 we derive non-asymptotic upper bounds on the estimation and prediction error of lasso in stochastic regression with serially correlated errors. Section 4.4 is devoted to

the modeling, estimation and theoretical analysis of sparse VAR models. We discuss least squares and likelihood based regularized estimation of VAR models and their consistency properties. In Section 4.6 we study the problem of covariance estimation from time series data by adaptive thresholding. We defer the technical proofs to Section 4.8.

Notations. Throughout this chapter, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$ will denote the sets of integers, real numbers and complex numbers, respectively. We denote the cardinality of a set J by $|J|$. For a vector $v \in \mathbb{R}^p$, we denote ℓ_q norms by $\|v\|_q := \left(\sum_{j=1}^p |v_j|^q\right)^{1/q}$, for $q > 0$. We use $\|v\|_0$ to denote $|supp(v)| = \sum_{i=1}^p \mathbf{1}[v_j \neq 0]$ and $\|v\|_\infty$ to denote $\max_j |v_j|$. Unless mentioned otherwise, we always use $\|.\|$ to denote ℓ_2-norm of a vector v. For a matrix A, $\|A\|$ and $\|A\|_F$ will denote its operator norm $\sqrt{\Lambda_{\max}(A'A)}$ and Frobenius norm $\sqrt{tr(A'A)}$, respectively. We will also use $\|A\|_{\max}$, $\|A\|_1$ and $\|A\|_\infty$ to denote the coordinate-wise maximum (in absolute value), maximum absolute row sum and maximum absolute column sum of a matrix, respectively. For any $p \geq 1$, $q \geq 0$, $r > 0$, we denote the unit balls by $\mathbb{B}_q(r) := \{v \in \mathbb{R}^p : \|v\|_q \leq r\}$. For any $J \subset \{1, \ldots, p\}$ and $\kappa > 0$, we define the cone set $\mathcal{C}(S, \kappa) = \{v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leq \kappa \|v_S\|_1\}$ and the sparse set $\mathcal{K}(s) = \mathbb{B}_0(s) \cap \mathbb{B}_2(1)$, for any $s \geq 1$. For any set V, we denote its closure and convex hull by $cl\{V\}$ and $conv\{V\}$. For a symmetric or Hermitian matrix A, we denote its maximum and minimum eigenvalues by $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$. We use e_i to denote the i^{th} unit vector in $\mathbb{R}^p$. Throughout the chapter, we write $A \gtrsim B$ if there exists an absolute constant c, independent of the model parameters, such that $A \geq cB$. We use $A \asymp B$ to denote $A \gtrsim B$ and $B \gtrsim A$.

4.2 Main Results

In this section, we first discuss the connection between the spectral density and the autocovariance function and introduce our measure of stability. Then, we present the key deviation inequalities used in subsequent analyses.

4.2.1 Measure of Stability

Consider a p-dimensional discrete time, centered, covariance-stationary process $\{X^t\}_{t\in\mathbb{Z}}$ with autocovariance function $\Gamma_X(h) = \mathrm{Cov}(X^t, X^{t+h})$, $t, h \in \mathbb{Z}$.

Assumption IV.1. *The spectral density function*

$$f_X(\theta) := \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma_X(\ell) e^{-i\ell\theta}, \ \theta \in [-\pi, \pi] \tag{4.4}$$

exists and is continuous.

We will often write f instead of f_X and Γ instead of Γ_X, when the underlying process is clear from the context. Existence of the spectral density is guaranteed if $\sum_{l=0}^{\infty} \|\Gamma(l)\| < \infty$. The assumption of continuity is satisfied by a large class of general linear processes, including stable, invertible ARMA processes (*Priestley*, 1981). Further, the spectral density has a closed form expression for these processes, as shown in the following example.

Example. An ARMA(d, ℓ) process $\{X^t\}$

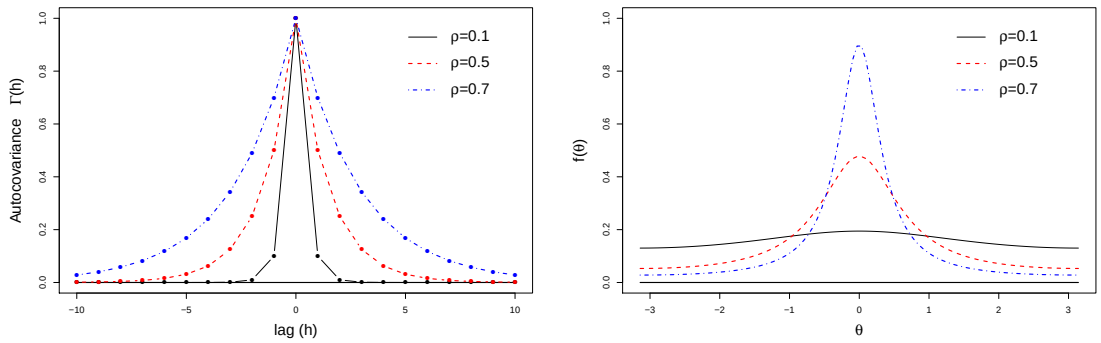
$$\begin{aligned} X^t &= A_1 X^{t-1} + A_2 X^{t-2} + \ldots + A_d X^{t-d} \\ &\quad + \epsilon^t - B_1 \epsilon^{t-1} - B_2 \epsilon^{t-2} - \ldots - B_\ell \epsilon^{t-\ell} \end{aligned} \tag{4.5}$$

is stable, invertible if the matrix valued polynomials $\mathcal{A}(z) := I_p - \sum_{t=1}^{d} A_t z^t$ and $\mathcal{B}(z) := I_p - \sum_{t=1}^{\ell} B_t z^t$ satisfy $\det(\mathcal{A}(z)) \neq 0$ and $\det(\mathcal{B}(z)) \neq 0$ on the unit circle of the complex plane $\{z \in \mathbb{C} : |z| = 1\}$.

For a stable, invertible ARMA process, the spectral density takes the form

$$f_X(\theta) = \frac{1}{2\pi} \left(\mathcal{A}^{-1}(e^{-i\theta})\right) \mathcal{B}(e^{-i\theta}) \, \Sigma_\epsilon \, \mathcal{B}^*(e^{-i\theta}) \left(\mathcal{A}^{-1}(e^{-i\theta})\right)^* \tag{4.6}$$

Existence of the spectral density ensures the following representation of the auto-



(a) Autocovariance of $AR(1)$ (b) Spectral Density of $AR(1)$

Figure 4.2: Autocovariance $\Gamma(h)$ and spectral density $f(\theta)$ of a univariate $AR(1)$ process $X^t = \rho X^{t-1} + \epsilon^t$, $0 < \rho < 1$, $\Gamma_X(0) = 1$. Processes with stronger temporal dependence, i.e., with larger ρ, have flatter Γ and more spiky f. For $\rho = 1$, the process is unstable and the spectral density does not exist.

covariance matrices:

$$\Gamma_X(\ell) = \int_{-\pi}^{\pi} f_X(\theta) e^{i\ell\theta}\, d\theta, \quad \text{for all } \ell \in \mathbb{Z} \tag{4.7}$$

Since the spectral density characterizes the autocovariance function, it can be used to study the temporal and cross-sectional dependence of the process. In particular, the spectral density provides insight into the stability of the process. In Figure 4.2, we illustrate this using the autocovariance function $\Gamma_X(h)$ and the spectral density $f_X(\theta)$ of a univariate $AR(1)$ process $X^t = \rho\, X^{t-1} + \epsilon^t$, $0 < \rho < 1$, $\Gamma_X(0) = 1$. Note that processes with stronger temporal dependence (larger ρ) have a narrower spectral density, with a higher peak. As ρ approaches 1, the peak of the spectral density $\mathcal{M}(f_X) := \max_{\theta\in[-\pi,\pi]} f_X(\theta)$ diverges. For $\rho = 1$, the process is not stable, and the spectral density does not exist. This indicates that the peak of the spectral density can be used as a measure of stability of the process.

More generally, for a p-dimensional time series $\{X^t\}$, a natural analogue of the "peak" is the maximum eigenvalue of the (matrix-valued) spectral density function

over the unit circle:

$$\mathcal{M}(f_X) := \max_{\theta \in [-\pi, \pi]} \Lambda_{\max} \left(f_X(\theta) \right) \tag{4.8}$$

In our analysis of high-dimensional time series, we will use $\mathcal{M}(f_X)$ as a **measure of stability** of the process. Processes with larger $\mathcal{M}(f_X)$ will be considered less stable.

For any k-dimensional subset J of $\{1, \ldots, p\}$, we can similarly measure the stability of the subprocess $\{X(J)\} = \{(X_j^t) : j \in J\}_{t \in \mathbb{Z}}$ as $\mathcal{M}(f_{X(J)})$. We will measure the stability of all k-dimensional subprocesses of $\{X^t\}$ using

$$\mathcal{M}(f_X, k) := \max_{J \subseteq \{1, \ldots, p\}, |J| \leq k} \mathcal{M}(f_{X(J)}) \tag{4.9}$$

Clearly, $\mathcal{M}(f_X) = \mathcal{M}(f_X, p)$. For completeness, we define $\mathcal{M}(f_X, k)$ to be $\mathcal{M}(f_X)$, for all $k \geq p$. It follows from the definitions that

$$\mathcal{M}(f_X, 1) \leq \mathcal{M}(f_X, 2) \leq \ldots \leq \mathcal{M}(f_X, p) = \mathcal{M}(f_X) \tag{4.10}$$

If $\{X^t\}$ and $\{Y^t\}$ are independent p-dimensional time series satisfying assumption IV.1 and $Z^t = X^t + Y^t$, then $f_Z = f_X + f_Y$. Consequently, we have

$$\mathcal{M}(f_Z) \leq \mathcal{M}(f_X) + \mathcal{M}(f_Y) \tag{4.11}$$

For studying stochastic regression and autoregression problems, we will also use the minimum eigenvalue of the spectral density over the unit circle:

$$\mathfrak{m}(f_X) := \min_{\theta \in [-\pi, \pi]} \Lambda_{\min} \left(f_X(\theta) \right) \tag{4.12}$$

$\mathfrak{m}(f_X)$ captures the dependence among the components of the vector-valued time series. In our analysis of high-dimensional regression problems, $\mathfrak{m}(f_X)$ plays a crucial role in quantifying dependence among the columns of the design matrix.

The quantities $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$ are well-defined because of the continuity of eigenvalues and the compactness of the unit circle $\{z \in \mathbb{C} : |z| = 1\}$.

$\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$ may not have closed form expressions for general stationary processes. However, for a stationary ARMA process (4.5), we have the following bounds

$$\mathfrak{m}(f_X) \geq \frac{1}{2\pi} \frac{\Lambda_{\min}(\Sigma_\epsilon)\mu_{\min}(\mathcal{B})}{\mu_{\max}(\mathcal{A})}, \ \mathcal{M}(f_X) \leq \frac{1}{2\pi} \frac{\Lambda_{\max}(\Sigma_\epsilon)\mu_{\max}(\mathcal{B})}{\mu_{\min}(\mathcal{A})} \tag{4.13}$$

where

$$\mu_{\min}(\mathcal{A}) := \min_{|z|=1} \Lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z)), \ \mu_{\max}(\mathcal{A}) := \max_{|z|=1} \Lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)) \tag{4.14}$$

and $\mu_{\min}(\mathcal{B})$, $\mu_{\max}(\mathcal{B})$ are defined accordingly.

It is often easier to work with $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$ instead of $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$. In particular, we have the following bounds:

Proposition IV.2. *Consider a polynomial $\mathcal{A}(z) = I_p - \sum_{t=1}^d A_t z^t$, $z \in \mathbb{C}$, satisfying $det(\mathcal{A}(z)) \neq 0$ for all $|z| \leq 1$.*

(i) For any $d \geq 1$, $\mu_{\max}(\mathcal{A}) \leq [1 + (v_{in} + v_{out})/2]^2$, where

$$v_{in} = \sum_{h=1}^{d} \max_{1 \leq i \leq p} \sum_{j=1}^{p} |A_h(i,j)|, \ v_{out} = \sum_{h=1}^{d} \max_{1 \leq j \leq p} \sum_{i=1}^{p} |A_h(i,j)|$$

(ii) If $d = 1$ and A_1 is diagonalizable, then

$$\mu_{\min}(\mathcal{A}) \geq (1 - \rho(A_1))^2 \, \|P\|^{-2} \|P^{-1}\|^{-2}$$

where $\rho(A_1)$ is the spectral radius (maximum absolute eigenvalue) of A_1 and the columns of P are eigenvectors of A_1.

Proposition IV.2, together with (4.13), shows that $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$ are bounded away from zero and infinity as long as the noise covariance structure is well-conditioned, the eigenvalues of A_1 are bounded away from 1 and the entries of A_t and B_t do not concentrate on a single row or column.

4.2.2 Deviation Bounds

Based on realizations $\{X^t\}_{t=1}^n$ generated according to a stationary process satisfying assumption (IV.1), we construct the data matrix $\mathcal{X} = [X^n : \cdots : X^1]'$ and the sample Gram matrix $S = \mathcal{X}'\mathcal{X}/n$. Deriving suitable concentration bounds on S is a key step for studying regression and covariance estimation problems in high-dimension. In the time series context, this is particularly challenging, since both the rows and columns of the data matrix $\mathcal{X}$ are dependent on each other. When the underlying process is Gaussian, this dependence can be expressed using the covariance matrix of the random vector $vec(\mathcal{X}')$. We denote this covariance matrix by $\Upsilon_n^X := \mathrm{Cov}(vec(\mathcal{X}'), vec(\mathcal{X}'))_{np\times np}$.

The next proposition provides bounds on the extreme eigenvalues of Υ_n^X. A similar result under slightly different conditions can be found in *Parter* (1961). Note that these bounds depend only on the spectral density f_X and are independent of the sample size n.

Proposition IV.3. *For any $n \geq 1$, $p \geq 1$,*

$$2\pi\mathfrak{m}(f_X) \leq \Lambda_{\min}\left(\Upsilon_n^X\right) \leq \Lambda_{\max}\left(\Upsilon_n^X\right) \leq 2\pi\mathcal{M}(f_X) \tag{4.15}$$

In particular, for $n = 1$,

$$2\pi\mathfrak{m}(f_X) \leq \Lambda_{\min}\left(\Gamma_X(0)\right) \leq \Lambda_{\max}\left(\Gamma_X(0)\right) \leq 2\pi\mathcal{M}(f_X) \tag{4.16}$$

In the next proposition, we establish two important deviation bounds on $S - \Gamma(0)$

for Gaussian time series. These bounds serve as the starting point for analyzing regression and covariance estimation problems. The first deviation bound is about the concentration of $\|\mathcal{X}v\|^2/n\|v\|^2$ around its expectation, where $v \in \mathbb{R}^p$ is a fixed vector. This will be used to verify restricted eigenvalue assumptions for stochastic regression and VAR estimation problems. The second deviation bound is about the concentration of the entries of S around their expectations. This will be useful for estimating sparse covariance matrices.

Proposition IV.4. *For a stationary, centered Gaussian time series $\{X^t\}_{t\in\mathbb{Z}}$ satisfying Assumption IV.1, there exists a constant $c > 0$ such that for any k-sparse vectors $u, v \in \mathbb{R}^p$ with $\|u\| \leq 1$, $\|v\| \leq 1$, $k \geq 1$, and any $\eta \geq 0$,*

$$\mathbb{P}\left[|v'\left(S - \Gamma_X(0)\right)v| > 2\pi\mathcal{M}(f_X, k)\eta\right] \leq 2\exp\left[-cn\min\{\eta^2, \eta\}\right] \quad (4.17)$$

$$\mathbb{P}\left[|u'\left(S - \Gamma_X(0)\right)v| > 6\pi\mathcal{M}(f_X, 2k)\eta\right] \leq 6\exp\left[-cn\min\{\eta^2, \eta\}\right] \quad (4.18)$$

In particular, for any $i, j \in \{1, \ldots, p\}$, we have

$$\mathbb{P}\left[|S_{ij} - \Gamma_{ij}(0)| > 6\pi\mathcal{M}(f_X, 2)\eta\right] \leq 6\exp\left[-cn\min\{\eta^2, \eta\}\right] \quad (4.19)$$

We give the proofs of the these two key propositions next, that employ techniques in spectral theory of multivariate time series and non-asymptotic random matrix theory results.

PROOF OF PROPOSITION IV.3. For $1 \leq r, s \leq n$, the $(r, s)^{th}$ block of the $np \times np$ matrix Υ_n^X is a $p \times p$ matrix

$$\Gamma_X(r - s) = \mathrm{Cov}\left(X^{n-r+1}, X^{n-s+1}\right)$$

For any $x \in \mathbb{R}^{np}$, $\|x\| = 1$, write x as $x = \{(x^1)', (x^2)', \ldots, (x^p)'\}'$, where each $x^i \in \mathbb{R}^p$.

Define $G(\theta) = \sum_{r=1}^{n} x^r e^{-ir\theta}$, for $\theta \in [-\pi, \pi]$. Note that

$$\begin{aligned} \int_{-\pi}^{\pi} G^*(\theta)G(\theta)\, d\theta &= \sum_{r=1}^{n}\sum_{s=1}^{n}\int_{-\pi}^{\pi} (x^r)'(x^s)e^{i(r-s)\theta}\, d\theta \\ &= \sum_{r=1}^{n} \|x^r\|^2\, 2\pi = 2\pi \end{aligned} \tag{4.20}$$

Also

$$\begin{aligned} x'\Upsilon_n^X x &= \sum_{r=1}^{n}\sum_{s=1}^{n}(x^r)'\Gamma_X(r-s)(x^s) \\ &= \sum_{r=1}^{n}\sum_{s=1}^{n}\int_{-\pi}^{\pi}(x^r)' f_X(\theta)e^{i(r-s)\theta}(x^s)\, d\theta \text{ using (4.7)} \\ &= \int_{-\pi}^{\pi} G^*(\theta)f_X(\theta)G(\theta)\, d\theta \end{aligned}$$

Since $f_X(\theta)$ is Hermitian, $G^*(\theta)f_X(\theta)G(\theta)$ is real, for all $\theta \in [-\pi, \pi]$, and

$$\mathfrak{m}(f_X)G^*(\theta)G(\theta) \le G^*(\theta)f_X(\theta)G(\theta) \le \mathcal{M}(f_X)G^*(\theta)G(\theta)$$

This, together with (4.20), implies

$$2\pi\mathfrak{m}(f_X) \le x'\Upsilon_n^X x \le 2\pi\mathcal{M}(f_X)$$

for all $x \in \mathbb{R}^{np}$, $\|x\| = 1$. □

PROOF OF PROPOSITION IV.4. We will establish the deviation bounds using a version of the Hansen-Wright inequality presented in Lemma 4.8.5 which says that for any n-dimensional centered Gaussian vector $Y \sim N(0, Q)$, and any $\eta \ge 0$, we

have

$$\mathbb{P}\left[\frac{1}{n}\left|\|Y\|^2 - tr(Q)\right| > \eta\|Q\|\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\}\right]$$

for some constant $c > 0$.

First, note that it is enough to prove (4.17) for $\|v\| = 1$. For any $v \in \mathbb{R}^p$, $\|v\| = 1$, let J denote its support $supp(v)$ so that $|J| = k$. define $Y = \mathcal{X}v = \mathcal{X}_J v_J$. Then $Y \sim N(0_{n\times 1}, Q_{n\times n})$ with

$$Q_{rs} = v_J' \mathrm{Cov}(X_J^{n-r+1}, X_J^{n-s+1}) v_J = v_J' \Gamma_{X(J)}(r-s) v_J, \quad \text{for all } 1 \leq r, s \leq n$$

Then $tr(Q) = nv_J'\Gamma_{X(J)}(0)v_J = v'\Gamma_X(0)v$ and $v'(S - \Gamma(0))v = \frac{1}{n}\left|\|Y\|^2 - tr(Q)\right|$. Also, for any $w \in \mathbb{R}^n$, $\|w\| = 1$, we have

$$\begin{aligned} w'Qw &= \sum_{r=1}^{n}\sum_{s=1}^{n} w_r w_s Q_{rs} = \sum_{r=1}^{n}\sum_{s=1}^{n} w_r w_s v_J' \Gamma_{X(J)}(r-s) v_J \\ &= (w \otimes v)' \Upsilon_n^{X(J)} (w \otimes v) \\ &\leq \Lambda_{\max}\left(\Upsilon_n^{X(J)}\right), \quad \text{since } \|w \otimes v\| = 1 \\ &\leq 2\pi \mathcal{M}(f_{X(J)}) \leq 2\pi \mathcal{M}(f_X, k) \end{aligned}$$

This establishes an upper bound on the operator norm $\|Q\| \leq 2\pi\mathcal{M}(f_X, k)$. The result then follows from Hansen-Wright inequality.

To prove (4.18), note that

$$\begin{aligned} 2\left|u'\left(S - \Gamma_X(0)\right)v\right| &\leq \left|u'(S - \Gamma_X(0))u\right| + \left|v'(S - \Gamma_X(0))v\right| \\ &+ \left|(u+v)'(S - \Gamma_X(0))(u+v)\right| \end{aligned}$$

and $u + v$ is $2k$-sparse with $\|u + v\| \leq 2$. The result follows by applying (4.17) separately on each of the three terms on the right.

The element-wise deviation bound (4.19) is obtained by choosing $u = e_i$, $v = e_j$.

□

4.3 Stochastic Regression

Stochastic regression with exogenous predictors and serially correlated errors is a canonical problem in classical time series analysis. As is well known, the standard errors of Ordinary Least Squares (OLS) estimates are affected in the presence of serially correlated errors, so that one resorts to employing Generalized Least Squares (GLS) estimates. However, the first step in GLS estimation with unknown error covariance is to come up with consistent estimates of the regression coefficient vector β^*, which are subsequently used to analyze the serial correlation in the residuals (*Hamilton*, 1994). In a low-dimensional setting (p fixed, $n \to \infty$), a natural choice of $\hat{\beta}$ is the OLS estimates. In high-dimensional setting under sparsity assumption on β^*, we establish that lasso based estimates are consistent for β^*, as long as the predictor and noise processes are stable.

We consider the lasso estimate (4.2) for the stochastic regression model (4.1). Further, we assume that both f_X and f_ϵ satisfy Assumption IV.1 and β^* is k-sparse, with support J, i.e., $|J| = k$.

In the low-dimensional regime, consistent estimation relies on the following assumptions:

(a) $\mathcal{X}'\mathcal{X}/n$ converges to a non-singular matrix ($\lim_{n\to\infty} \Lambda_{\min}\left(\frac{\mathcal{X}'\mathcal{X}}{n}\right) > 0$)

(b) $\mathcal{X}'E/n$ converges to zero

In the high-dimensional regime ($n \ll p$), the first assumption is never true since the design matrix is rank-deficient (more variables than observations). The second assumption is also very stringent, since the dimension of $\mathcal{X}'E$ grows with n and p.

Interestingly, consistent estimation in the high-dimensional regime can be ensured under two analogous sufficient conditions. The first one comes from a class of conditions commonly referred to as **Restricted Eigenvalue** (RE) condition. Different variants of the RE condition have been proposed in the literature (*Bickel et al.*, 2009; *van de Geer and Bühlmann*, 2009b). Roughly speaking these assumptions require that $\|\mathcal{X}(\hat{\beta}-\beta^*)\|$ is small only when $\|\hat{\beta}-\beta^*\|$ is small. If $\hat{\beta}, \beta^*$ are any arbitrary vectors in $\mathbb{R}^p$, this assumption is never true since $\mathcal{X}$ is singular. However, if β^* is sparse and λ_N is appropriately chosen, it is now well-understood that the vectors $v = \hat{\beta}-\beta^*$ only vary on a small subset of the high-dimensional space $\mathbb{R}^p$ (*Negahban et al.*, 2012). As shown in the proof of Proposition IV.7, the error vectors v in stochastic regression lie in a low-dimensional cone

$$\mathcal{C}(J,3) = \{v \in \mathbb{R}^p : \|v_{J^c}\|_1 \leq 3\|v_J\|_1\}$$

whenever $\lambda_n \geq 4\|\mathcal{X}'E/n\|_\infty$. This indicates that the RE condition may not be very stringent after all, even though $\mathcal{X}$ is singular. Note however that verifying that the assumption indeed holds with high probability is a non-trivial task.

The next proposition shows that a restricted eigenvalue (RE) condition holds with high probability when the sample size is sufficiently large and the process of predictors $\{X^t\}$ is stable, with a full-rank spectral density.

Proposition IV.5 (Restricted Eigenvalue)**.** *If* $\mathfrak{m}(f_X) > 0$*, then there exist constants* $c_i > 0$ *such that for* $n \gtrsim \max\{1, \omega^2\} \min\{k\log(c_0 p/k), k\log p\}$*,*

$$\mathbb{P}\left[\inf_{v\in\mathcal{C}(J,3)\setminus\{0\}} \frac{\|\mathcal{X}v\|^2}{n\|v\|^2} \geq \alpha_{RE}\right] \geq 1 - c_1\exp\left[-c_2 n\min\{1,\omega^{-2}\}\right]$$

where $\alpha_{RE} = \pi\mathfrak{m}(f_X)$*,* $\omega = c_3\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X)$*.*

REMARKS. (a) The assumption $\mathfrak{m}(f_X) > 0$ is fairly mild and holds for stable,

invertible ARMA processes. However, the conclusion holds under weaker assumptions like $\Lambda_{\min}(\Gamma_X(0)) > 0$ or a RE condition on $\Gamma_X(0)$, replacing $2\pi\mathfrak{m}(f_X)$ by the minimum (or restricted) eigenvalue of $\Gamma_X(0)$.

(b) For large k, $k\log(c_0 p/k)$ can be much smaller than $k\log p$, the sample size required for consistent estimation with lasso.

(c) The factor $\omega \asymp \mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X)$ captures the effect of temporal and cross-sectional dependence in the data. Larger values of $\mathcal{M}(.)$ and smaller values of $\mathfrak{m}(.)$ indicate stronger dependence in the data and more samples are required to ensure RE holds with high probability. We demonstrate this on three special types of dependence in the design matrix $\mathcal{X}$ - independent entries, independent rows and independent columns.

(i) If the entries of $\mathcal{X}$ are independent $N(0, \sigma^2)$, we have $\Gamma_X(0) = \sigma^2 I$ and $\Gamma_X(h) = \mathbf{0}$ for $h \neq 0$. In this case, $f_X(\theta) \equiv (1/2\pi)\,\sigma^2 I$ and $\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X) = 1$.

(ii) If the rows of $\mathcal{X}$ are independent and identically distributed as $N(0, \Sigma_X)$, i.e., $\Gamma_X(0) = \Sigma_X$, $\Gamma_X(h) = \mathbf{0}$ for $h \neq 0$, the spectral density takes the form $f_X(\theta) \equiv (1/2\pi)\Sigma_X$, and $\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X)$ can be at most $\Lambda_{\max}(\Sigma_X)/\Lambda_{\min}(\Sigma_X)$.

(iii) If the columns of $\mathcal{X}$ are independent, i.e., all the univariate components of $\{X^t\}$ are independently generated according to a common stationary process with spectral density f, then the spectral density of $\{X^t\}$ is $f_X(\theta) = f(\theta)\, I$ and we have

$$\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X) = \max_{\theta\in[-\pi,\pi]} f(\theta) \,/\, \min_{\theta\in[-\pi,\pi]} f(\theta)$$

The ratio on the right can be viewed as a measure of narrowness of f. Since narrower spectral densities correspond to processes with flatter autocovariance, it shows that more samples are needed when the dependence is stronger.

The second sufficient condition for consistency of lasso requires that the coordinates of $\mathcal{X}'E/n$ uniformly concentrate around 0. In the next proposition, we establish

a deviation bound on $\|\mathcal{X}'E/n\|_\infty$ that holds with high probability. Similar results were established in *Loh and Wainwright* (2012) for VAR(1) process with serially uncorrelated errors, under the assumption $\|A_1\| < 1$. Our result relies on different techniques, holds for a much larger class of stationary processes and allows for serial correlation in the noise term, as well.

Proposition IV.6 (Deviation Condition)**.** *For $n \gtrsim \log p$, there exist constants $c_i > 0$ such that*

$$\mathbb{P}\left[\frac{1}{n}\|\mathcal{X}'E\|_\infty > c_0 2\pi\left[\mathcal{M}(f_X,1)+\mathcal{M}(f_\epsilon)\right]\sqrt{\frac{\log p}{n}}\right] \le c_1\exp\left[-c_2\log p\right] \qquad (4.21)$$

REMARKS. (a) The deviation inequality suggests that the coordinates of $\mathcal{X}'E/n$ uniformly concentrate around 0, as long as $\mathcal{M}(f_X,1)$ and $\mathcal{M}(f_\epsilon)$ are not large, i.e., the univariate components of the predictor process and the noise process are stable.

Using the above propositions, we can establish error rates of estimation and prediction in stochastic regression with exogenous predictors and serially correlated errors.

Proposition IV.7 (Estimation and Prediction Error)**.** *Consider the stochastic regression setup of* (4.1)*. If β^* is k-sparse, $n \gtrsim [\mathcal{M}(f_X,k)/\mathfrak{m}(f_X)]^2 k\log p$, then there exist constants $c_i > 0$ such that for*

$$\lambda_n \ge c_0 2\pi\left[\mathcal{M}(f_X,1)+\mathcal{M}(f_\epsilon)\right]\sqrt{(\log p)/n}$$

any solution $\hat\beta$ of (4.2) *satisfies, with probability at least $1 - c_1\exp[-c_2\log p]$,*

$$\begin{aligned}
\left\|\hat\beta-\beta^*\right\| &\le \frac{2\lambda_n\sqrt{k}}{\alpha_{RE}}\\
\left\|\hat\beta-\beta^*\right\|_1 &\le \frac{8\lambda_n k}{\alpha_{RE}}\\
\frac{1}{n}\left\|\mathcal{X}(\hat\beta-\beta^*)\right\|^2 &\le \frac{4\lambda_n^2 k}{\alpha_{RE}}
\end{aligned}$$

where the restricted eigenvalue $\alpha_{RE} = \pi\mathfrak{m}(f_X)$.

Further, a thresholded variant of lasso $\tilde{\beta}$*, defined as* $\tilde{\beta}_j = \{\hat{\beta}_j \mathbf{1}_{|\hat{\beta}_j| > \lambda_n}\}$*, for* $1 \leq j \leq p$*, satisfies, with the same probability,*

$$\left| supp(\tilde{\beta}) \backslash supp(\beta^*) \right| \leq \frac{24k}{\alpha_{RE}} \tag{4.22}$$

REMARKS. (a) The convergence rates of ℓ_2-estimation and prediction $\sqrt{k \log p \,/n}$ are of the same order as the rates for regression with i.i.d. samples. Dependence contributes the additional term $[\mathcal{M}(f_X, 1) + \mathcal{M}(f_\epsilon)] \,/\mathfrak{m}(f_X)$ in the error rates and $[\mathcal{M}(f_X, 2k)/\mathfrak{m}(f_X)]^2$ in the sample size requirement. This ensures fast convergence rates of lasso under high-dimensional scaling as long as the processes of predictors and noise are stable.

(b) A thresholded version of lasso enjoys small false positive rates, as shown in (4.22). Note that we do not assume any "beta-min" condition, i.e., a lower bound on the minimum signal strength. It is possible to control the false negatives under suitable "beta-min" conditions, as shown in (*Zhou*, 2010).

4.4 Transition Matrix Estimation in Sparse Vector Autoregressive Models

The problem of estimating sparse VAR models under ℓ_1-penalized regression has been considered by several authors in recent years (*Song and Bickel*, 2011; *Davis et al.*, 2012; *Kock and Callot*, 2012; *Han and Liu*, 2013). Most of these studies consider a least squares based objective function or estimating equation to derive the estimates. An important aspect of this approach is that it is agnostic to the presence of cross-correlation among the error components (non-diagonal Σ_ϵ). *Davis et al.* (2012) provided numerical evidence that the forecasting performance can be improved by using a log-likelihood based loss function that incorporates knowledge

about the error correlations. In this section, we consider both least squares and log-likelihood estimates and study their theoretical properties.

A key contribution of our theoretical analysis is to verify suitable RE and deviation conditions for the entire class of stable VAR(d) models. Existing works either assume such conditions without verification, or use a stringent condition on the model parameters, such as $\|A\| < 1$, as discussed in Section 4.1.

We consider a single realization of $\{X^0, X^1, \ldots, X^T\}$ generated according to the VAR model (4.3). We will assume the error covariance matrix Σ_ϵ is positive definite so that $\Lambda_{\min}(\Sigma_\epsilon) > 0$ and $\Lambda_{\max}(\Sigma_\epsilon) < \infty$. We will also assume that the VAR process is *stable*, i.e., $det(\mathcal{A}(z)) \neq 0$ on the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. For stable VAR(d) processes, the spectral density (4.6) simplifies to

$$f_X(\theta) = \frac{1}{2\pi} \left(\mathcal{A}^{-1}(e^{-i\theta})\right) \Sigma_\epsilon \left(\mathcal{A}^{-1}(e^{-i\theta})\right)^* \tag{4.23}$$

To deal with dependence in the VAR estimation problem, we will work with $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$ and the extreme eigenvalues of Σ_ϵ instead of $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$. For a VAR(d) process with serially uncorrelated errors, equation (4.13) simplifies to

$$\mathcal{M}(f_X) \leq \frac{1}{2\pi} \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}, \ \mathfrak{m}(f_X) \geq \frac{1}{2\pi} \frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})} \tag{4.24}$$

This factorization helps provide better insight into the temporal and contemporaneous dependence in VAR models. A graphical representation of a stable VAR(d) model (4.3) is provided in Figure 4.3. The transition matrices $A_1, \ldots, A_d$ encode the temporal dependence of the process. When the components of the error process $\{\epsilon^t\}$ are correlated, Σ_ϵ^{-1} captures the additional contemporaneous dependence structure. Expressing the estimation and prediction errors in terms of $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$, $\Lambda_{\min}(\Sigma_\epsilon)$ and $\Lambda_{\max}(\Sigma_\epsilon)$ instead of $\mathfrak{m}(f_X)$ and $\mathcal{M}(f_X)$ help separate the effect of the two sources of dependence.

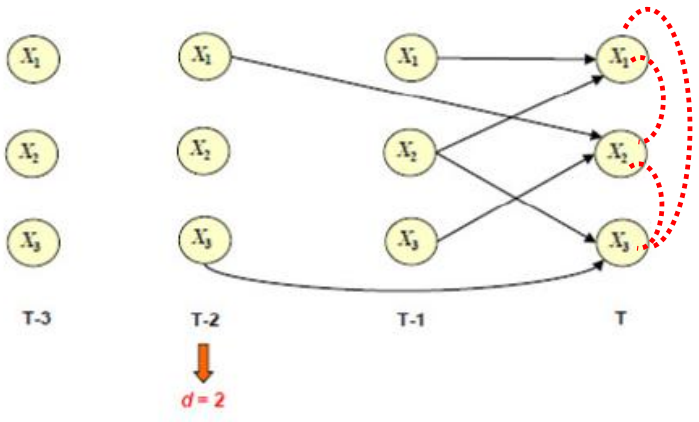


Figure 4.3: Graphical representation of the VAR model (4.3): directed edges (solid) correspond to the entries of the transition matrices, undirected edges (dashed) correspond to the entries of Σ_ϵ^{-1}

We will often use the following alternative representation of a p-dimensional VAR(d) process (4.3) as a dp-dimensional VAR(1) process $\tilde{X}^t = \tilde{A}_1 \tilde{X}^{t-1} + \tilde{\epsilon}^t$ with

$$\tilde{X}^t = \begin{bmatrix} X^t \\ X^{t-1} \\ \vdots \\ X^{t-d+1} \end{bmatrix}_{dp\times 1} \quad \tilde{A}_1 = \begin{bmatrix} A_1 & A_2 & \cdots & A_{d-1} & A_d \\ I_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_p & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_p & \mathbf{0} \end{bmatrix}_{dp\times dp} \quad \tilde{\epsilon}^t = \begin{bmatrix} \epsilon^t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}_{dp\times 1} \tag{4.25}$$

The process $\tilde{X}^t$ with reverse characteristic polynomial $\tilde{\mathcal{A}}(z) := I_{dp} - \tilde{A}_1 z$ is stable if and only if the process X^t is stable (*Lütkepohl*, 2005). However, the quantities $\mu_{\min}(\mathcal{A})$, $\mu_{\max}(\mathcal{A})$ are not necessarily the same as $\mu_{\min}(\tilde{\mathcal{A}})$, $\mu_{\max}(\tilde{\mathcal{A}})$.

4.4.1 Estimation Procedure

Based on the data $\{X^0, \ldots, X^T\}$, we construct autoregression

$$\underbrace{\begin{bmatrix} (X^T)' \\ \vdots \\ (X^d)' \end{bmatrix}}_{\mathcal{Y}} = \underbrace{\begin{bmatrix} (X^{T-1})' & \cdots & (X^{T-d})' \\ \vdots & \ddots & \vdots \\ (X^{d-1})' & \cdots & (X^0)' \end{bmatrix}}_{\mathcal{X}} \underbrace{\begin{bmatrix} A_1' \\ \vdots \\ A_d' \end{bmatrix}}_{B^*} + \underbrace{\begin{bmatrix} (\epsilon^T)' \\ \vdots \\ (\epsilon^d)' \end{bmatrix}}_{E}$$

$$\begin{aligned} vec(\mathcal{Y}) &= vec(\mathcal{X}\, B^*) + vec(E) \\ &= (I \otimes \mathcal{X})\, vec(B^*) + vec(E) \\ \underbrace{Y}_{Np\times 1} &= \underbrace{Z}_{Np\times q}\underbrace{\beta^*}_{q\times 1} + \underbrace{vec(E)}_{Np\times 1}\ N = (T-d+1),\ q = dp^2 \end{aligned} \tag{4.26}$$

This is a linear regression problem with $N = T - d + 1$ samples and $q = dp^2$ variables. We will assume that β^* is a k-sparse vector, i.e., $\sum_{t=1}^{d} \|vec(A_t)\|_0 = k$.

We consider two different estimates for the transition matrices $A_1, \ldots, A_d$, or equivalently, for β^*. The first one is an ℓ_1-penalized least squares estimate of VAR coefficients (ℓ_1-LS). It is defined as

$$\underset{\beta\in\mathbb{R}^q}{\operatorname{argmin}} \frac{1}{N} \|Y - Z\beta\|^2 + \lambda_N \|\beta\|_1 \tag{4.27}$$

This estimate does not exploit the error covariance structure Σ_ϵ.

The second one uses an ℓ_1-penalized log-likelihood estimation (ℓ_1-LL) (*Davis et al.*, 2012). This is defined as

$$\underset{\beta\in\mathbb{R}^q}{\operatorname{argmin}} \frac{1}{N} (Y - Z\beta)' \left(\Sigma_\epsilon^{-1} \otimes I\right) (Y - Z\beta) + \lambda_N \|\beta\|_1 \tag{4.28}$$

This gives the maximum likelihood estimate of β, assuming the error covariance Σ_ϵ is known. In practice, Σ_ϵ is often unknown and needs to be estimated from the data.

4.4.2 Theoretical Properties

We analyze the two procedures (4.27) and (4.28) under a general penalized M-estimation framework proposed in *Loh and Wainwright* (2012). To motivate this general framework, note that the VAR estimation problem with ordinary least squares is equivalent to the following optimization

$$\underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} \; -2\beta' \hat{\gamma} + \beta' \hat{\Gamma} \beta, \tag{4.29}$$

where $\hat{\Gamma} = \left(I \otimes \mathcal{X}'\mathcal{X}/N\right)$, $\hat{\gamma} = \left(I \otimes \mathcal{X}'\right) Y/N$ are unbiased estimates for their population analogues. A more general choice of $(\hat{\gamma}, \hat{\Gamma})$ in the penalized version of the objective function leads to the following optimization problem

$$\begin{gathered} \underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} \; -2\beta' \hat{\gamma} + \beta' \hat{\Gamma} \beta + \lambda_N \left\|\beta\right\|_1, \\ \hat{\Gamma} = \left(W \otimes \mathcal{X}'\mathcal{X}/N\right), \; \hat{\gamma} = \left(W \otimes \mathcal{X}'\right) Y/N \end{gathered} \tag{4.30}$$

where W is a symmetric, positive definite matrix of weights. The optimization problems (4.27) and (4.28) are special cases of (4.30) with $W = I$ and $W = \Sigma_\epsilon^{-1}$, respectively.

As in the analysis of stochastic regression in Section 4.3, we establish consistency of VAR estimates under two sufficient conditions - Restricted Eigenvalue (RE) and Deviation Condition. Then we show that all stable VAR models satisfy these assumptions with high probability, as long as the sample size is of the same order as required for consistency. Although similar in spirit, these assumptions take different forms than the ones used in stochastic regression due to the different choice of loss function. We work with the RE condition proposed in *Loh and Wainwright* (2012). For a detailed discussion of the curvature and the tolerance parameters we refer the readers to the above paper.

(A1) Restricted Eigenvalue (RE): a symmetric matrix $\hat{\Gamma}_{q\times q}$ satisfies restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau > 0$ ($\hat{\Gamma} \sim RE(\alpha, \tau)$) if

$$\theta'\hat{\Gamma}\theta \geq \alpha \|\theta\|^2 - \tau \|\theta\|_1^2, \ \forall\, \theta \in \mathbb{R}^q \tag{4.31}$$

The deviation condition ensures that $\hat{\gamma}$ and $\hat{\Gamma}$ are well-behaved in the sense that they concentrate nicely around their population means. As $\hat{\gamma}$ and $\hat{\Gamma}\beta^*$ have the same expectation, this assumption requires an upper bound on their difference. Note that in the low-dimensional context of (4.29), $\hat{\gamma} - \hat{\Gamma}\beta^*$ is precisely $vec(\mathcal{X}'E)/N$.

(A2) Deviation Condition: There exists a deterministic function $\mathbb{Q}(\beta^*, \Sigma_\epsilon)$ such that

$$\left\|\hat{\gamma} - \hat{\Gamma}\beta^*\right\|_\infty \leq \mathbb{Q}(\beta^*, \Sigma_\epsilon)\sqrt{\frac{\log d + 2\log p}{N}} \tag{4.32}$$

The following proposition establishes non-asymptotic upper bounds on the estimation and prediction errors when the above conditions are satisfied.

Proposition IV.8 (Estimation and prediction error)**.** *Consider the penalized M-estimation problem* (4.30) *with* $W = I$ *or* $W = \Sigma_\epsilon^{-1}$*. Suppose* $\hat{\Gamma}$ *satisfies RE condition (4.31) with* $k\tau \leq \alpha/32$ *and* $(\hat{\Gamma}, \hat{\gamma})$ *satisfies the deviation bound (4.32). Then, for any* $\lambda_N \geq 4\mathbb{Q}(\beta^*, \Sigma_\epsilon)\sqrt{(\log d + 2\log p)/N}$*, any solution* $\hat{\beta}$ *of* (4.30) *satisfies*

$$\|\hat{\beta} - \beta^*\|_1 \leq 64\, k\, \lambda_N/\alpha \tag{4.33}$$

$$\|\hat{\beta} - \beta^*\| \leq 16\sqrt{k}\, \lambda_N/\alpha \tag{4.34}$$

$$(\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*) \leq 128\, k\, \lambda_N^2/\alpha \tag{4.35}$$

Further, a thresholded variant of lasso $\tilde{\beta} = \{\hat{\beta}_j \mathbf{1}_{|\hat{\beta}_j| > \lambda_N}\}$ *satisfies*

$$\left|supp(\tilde{\beta}) \backslash supp(\beta^*)\right| \leq \frac{192k}{\alpha_{RE}}$$

Remark. (a) $\|\hat{\beta} - \beta^*\|$ is precisely $\sum_{t=1}^{d} \|\hat{A}_t - A_t\|_F$, the ℓ_2-error in estimating the transition matrices. For ℓ_1-LS, $(\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*)$ is a measure of in-sample prediction error under ℓ_2-norm, defined as $\sum_{t=d}^{T} \|\sum_{h=1}^{d}(\hat{A}_h - A_h)X^{t-h}\|^2/N$. For ℓ_1-LL, $(\hat{\beta} - \beta^*)'\hat{\Gamma}(\hat{\beta} - \beta^*)$ takes the form $\sum_{t=d}^{T} \|\sum_{h=1}^{d}(\hat{A}_h - A_h)X^{t-h}\|^2_{\Sigma_\epsilon}/N$, where $\|v\|_\Sigma := \sqrt{v'\Sigma^{-1}v}$. This can be viewed as a measure of in-sample prediction error under a Mahalanobis type distance on $\mathbb{R}^p$ induced by Σ_ϵ .

(b) The convergence rates are governed by two sets of parameters: (i) dimensionality parameters - dimension of the process (p), order of the process (d), number of parameters (k) in the transition matrices A_i and sample size ($N = T - d + 1$); (ii) internal parameters - curvature (α), tolerance (τ) and the deviation bound $\mathbb{Q}(\beta^*, \Sigma_\epsilon)$. The squared ℓ_2-errors of estimation and prediction scale with the dimensionality parameters as $k(2\log p + \log d)/N$, similar to the rates obtained when the observations are independent (*Bickel et al.*, 2009). The temporal and cross-sectional dependence affect the rates only through the internal parameters. Typically, the rates are better when α is large and $\mathbb{Q}(\beta^*, \Sigma_\epsilon), \tau$ are small. In propositions IV.9 and IV.10, we investigate in detail how these quantities are related to the dependence structure of the process.

(c) Although the above proposition is derived under the assumption that d is the true order of the VAR process, the results hold even if d is replaced by any upper bound $\bar{d}$ on the true order. This follows from the fact that a VAR(d) model can also be viewed as VAR($\bar{d}$), for any $\bar{d} > d$, with transition matrices $A_1, \ldots, A_d, 0_{p\times p}, \ldots, 0_{p\times p}$. Note that the convergence rates change from $\sqrt{(\log p + 2\log d)/N}$ to $\sqrt{(\log p + 2\log \bar{d})/N}$.

Proposition IV.8 is deterministic, i.e., it assumes a fixed realization of $\{X^0, \ldots, X^T\}$. To show that these error bounds hold with high probability, one needs to verify that the assumptions (A1-2) are satisfied with high probability when $\{X^0, \ldots, X^T\}$ is a random realization from the VAR(d) process. This is accomplished in the next two propositions.

Proposition IV.9 (Verifying RE for $\hat{\Gamma}$)**.** *Consider a random realization* $\{X^0, \ldots, X^T\}$ *generated according to a stable VAR(d) process* (4.3)*. Then there exist constants* $c_i > 0$ *such that for all* $N \gtrsim \max\{\omega^2, 1\}k(\log d + \log p)$*, with probability at least* $1 - c_1 \exp(-c_2 N \min\{\omega^{-2}, 1\})$*, the matrix*

$$\hat{\Gamma} = I_p \otimes (\mathcal{X}'\mathcal{X}/N) \sim RE(\alpha, \tau),$$

where

$$\omega = c_3 \frac{\Lambda_{\max}(\Sigma_\epsilon)/\mu_{\min}(\tilde{\mathcal{A}})}{\Lambda_{\min}(\Sigma_\epsilon)/\mu_{\max}(\mathcal{A})},\ \alpha = \frac{\Lambda_{\min}(\Sigma_\epsilon)}{2\mu_{\max}(\mathcal{A})},\ \tau = \alpha \max\{\omega^2, 1\} \frac{\log d + \log p}{N}.$$

Further, if Σ_ϵ^{-1} *satisfies* $\bar{\sigma}_\epsilon^i := \sigma_\epsilon^{ii} - \sum_{j \neq i} \sigma_\epsilon^{ij} > 0$*, for* $i = 1, \ldots, p$*, then, with the same probability as above, the matrix*

$$\hat{\Gamma} = \Sigma_\epsilon^{-1} \otimes (\mathcal{X}'\mathcal{X}/N) \sim RE\left(\alpha \min_i \bar{\sigma}_\epsilon^i,\ \tau \max_i \bar{\sigma}_\epsilon^i\right)$$

This proposition provides insight into the effect of temporal and cross-sectional dependence on the convergence rates obtained in Proposition IV.8. As mentioned earlier, the convergence rates are faster for larger α and smaller τ. From the expressions of ω, α and τ, it is clear that the VAR estimates have lower error bounds when $\Lambda_{\max}(\Sigma_\epsilon), \mu_{\max}(\mathcal{A})$ are smaller and $\Lambda_{\min}(\Sigma_\epsilon), \mu_{\min}(\tilde{\mathcal{A}})$ are larger. We defer the proof to Section 4.8.2.

Proposition IV.10 (Deviation Bound)**.** *There exist constants* $c_i > 0$ *such that for* $N \gtrsim (\log d + 2\log p)$*, with probability at least* $1 - c_1 \exp\left[-c_2(\log d + 2\log p)\right]$*, we have*

$$\left\| \hat{\gamma} - \hat{\Gamma}\beta^* \right\|_\infty \leq \mathbb{Q}(\beta^*, \Sigma_\epsilon) \sqrt{\frac{\log d + 2\log p}{N}},$$

where, for ℓ_1-LS,

$$\mathbb{Q}(\beta^*, \Sigma_\epsilon) = c_0 \left[\Lambda_{\max}(\Sigma_\epsilon) + \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} + \frac{\Lambda_{\max}(\Sigma_\epsilon)\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})} \right]$$

and for ℓ_1-LL,

$$\mathbb{Q}(\beta^*, \Sigma_\epsilon) = c_0 \left[\frac{1}{\Lambda_{\min}(\Sigma_\epsilon)} + \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} + \frac{\Lambda_{\max}(\Sigma_\epsilon)\mu_{\max}(\mathcal{A})}{\Lambda_{\min}(\Sigma_\epsilon)\mu_{\min}(\mathcal{A})} \right]$$

As before, this proposition shows that the VAR estimates have lower error bounds when $\Lambda_{\max}(\Sigma_\epsilon)$, $\mu_{\max}(\mathcal{A})$ are smaller and $\Lambda_{\min}(\Sigma_\epsilon)$, $\mu_{\min}(\mathcal{A})$ are larger.

4.5 Implementation

The optimization problem ℓ_1-LS in (4.27) can be expressed as p *separate* penalized regression problems:

$$\begin{aligned} & \underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} \frac{1}{N} \|Y - Z\beta\|^2 + \lambda_N \|\beta\|_1 \\ \equiv \quad & \underset{B_1,\ldots,B_p}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{p} \|\mathcal{Y}_i - \mathcal{X} B_i\|^2 + \lambda_N \sum_{i=1}^{p} \|B_i\|_1 \end{aligned}$$

This amounts to running p separate lasso programs, each with dp predictors: $\mathcal{Y}_i \sim \mathcal{X}$, $i = 1, \ldots, p$. For large d and p, the p programs can be solved in parallel.

In the optimization problem ℓ_1-LL, the above regressions are coupled through Σ_ϵ^{-1}. One way to solve the problem, as mentioned in *Davis et al.* (2012), is to reformulate it into a single penalized regression problem:

$$\begin{aligned} & arg \quad \min_{\beta \in \mathbb{R}^q} \frac{1}{N} (Y - Z\beta)' \left(\Sigma_\epsilon^{-1} \otimes I\right) (Y - Z\beta) + \lambda_N \|\beta\|_1 \\ \equiv \quad & arg \quad \min_{\beta \in \mathbb{R}^q} \frac{1}{N} \left\| \left(\Sigma_\epsilon^{-1/2} \otimes I\right) Y - \left(\Sigma_\epsilon^{-1/2} \otimes \mathcal{X}\right) \beta \right\|^2 + \lambda_N \|\beta\|_1 \end{aligned}$$

This amounts to running a single lasso program with dp^2 predictors: $\left(\Sigma_\epsilon^{-1/2} \otimes I\right) Y \sim \Sigma_\epsilon^{-1/2} \otimes \mathcal{X}$. This is computationally expensive for large d and p. Unlike ℓ_1-LL, this algorithm is not parallelizable.

We propose an alternative algorithm based on blockwise coordinate descent to estimate the ℓ_1-LL coefficients. To this end, we first observe that the objective function in (4.28) can be simplified to

$$\frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_\epsilon^{ij} \left(\mathcal{Y}_i - \mathcal{X} B_i\right)' \left(\mathcal{Y}_j - \mathcal{X} B_j\right) + \lambda_N \sum_{k=1}^{p} \|B_k\|_1$$

Minimizing the above objective function cyclically with respect to each B_i leads to the following algorithm for ℓ_1-LL:

1. pre-select d. Run ℓ_1-LS to get $\hat{B}$, $\hat{\Sigma}_\epsilon^{-1}$.

2. iterate till convergence:

   (a) For $i = 1, \ldots, p$,

   - set $r_i := (1/2\,\hat{\sigma}_\epsilon^{ii}) \sum_{j \neq i} \hat{\sigma}_\epsilon^{ij} \left(\mathcal{Y}_j - \mathcal{X} \hat{B}_j\right)$
   - update $\hat{B}_i = \underset{B_i}{\operatorname{argmin}} \; \frac{\hat{\sigma}_\epsilon^{ii}}{N} \|(\mathcal{Y}_i + r_i) - \mathcal{X} B_i\|^2 + \lambda_N \|B_i\|_1$

In this algorithm, a single iteration amounts to running p *separate* lasso programs, each with dp predictors: $\mathcal{Y}_i + r_i \sim \mathcal{X}$, $i = 1, \ldots, p$. As in ℓ_1-LS, these p programs can be solved in parallel.

4.6 Sparse Covariance Estimation in Time Series

We consider a p-dimensional centered Gaussian stationary time series $\{X^t\}_{t \in \mathbb{Z}}$ satisfying assumption IV.1. Based on realizations $\{X^1, \ldots, X^n\}$ generated according to the above stationary process, we aim to estimate the contemporaneous covariance matrix $\Sigma = \Gamma(0)$. The sample covariance matrix $\hat{\Gamma}(0) = \frac{1}{n} \sum_{t=1}^{n} (X^t - \bar{X})(X^t -$

$\bar{X})'$ is known to be inconsistent when p grows faster than n (*Marčenko and Pastur*, 1967; *Johnstone*, 2001). *Bickel and Levina* (2008) showed that when the samples are generated independently from a centered Gaussian or subgaussian distribution, a thresholded version of the sample covariance matrix $T_u(\hat{\Gamma}(0)) = \{\hat{\Gamma}_{ij}(0)\mathbf{1}_{|\hat{\Gamma}_{ij}(0)|>u}\}$ can perform consistent estimation, if $\Gamma(0)$ belongs to the following uniformity class of approximately sparse matrices

$$\mathcal{U}_\tau(q, c_0(p), M) := \left\{\Sigma : \sigma_{ii} \leq M, \sum_{j=1}^{p} |\sigma_{ij}|^q \leq c_0(p), \text{ for all } i\right\} \tag{4.36}$$

In this section, we show that consistent estimation is possible in the time series context, as long as the underlying process is stable.

Proposition IV.11. *Let $\{X^t\}_{t=1}^n$ be generated according to a p-dimensional stationary centered Gaussian process with spectral density f_X, satisfying Assumption IV.1. Then, uniformly on $\mathcal{U}_\tau(q, c_0(p), M)$, for sufficiently large M', if $u_n = \mathcal{M}(f_X, 2)M'\sqrt{\log p/n}$ and $n \gtrsim \mathcal{M}^2(f_X, 2)\log p$, then*

$$\left\|T_{u_n}(\hat{\Gamma}(0)) - \Gamma(0)\right\| = O_p\left(c_0(p)\left(\mathcal{M}^2(f_X, 2)\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right) \tag{4.37}$$

$$\frac{1}{p}\left\|T_{u_n}(\hat{\Gamma}(0)) - \Gamma(0)\right\|_F = O_p\left(c_0(p)\left(\mathcal{M}^2(f_X, 2)\frac{\log p}{n}\right)^{1-\frac{q}{2}}\right) \tag{4.38}$$

REMARK. (a) The errors of estimation in operator and Frobenius norm scale with $\log p/n$, with an additional "price" of dependence $\mathcal{M}^2(f_X, 2)$. Interestingly, only the stability measure of bivariate subprocesses of $\{X^t\}$ appear in the bounds. For large p, this can be substantially smaller than the stability measure of the entire process $\mathcal{M}(f_X)$.

(b) *Chen et al.* (2013) established consistency of thresholding procedures for covariance estimation in stationary time series using the framework of functional de-

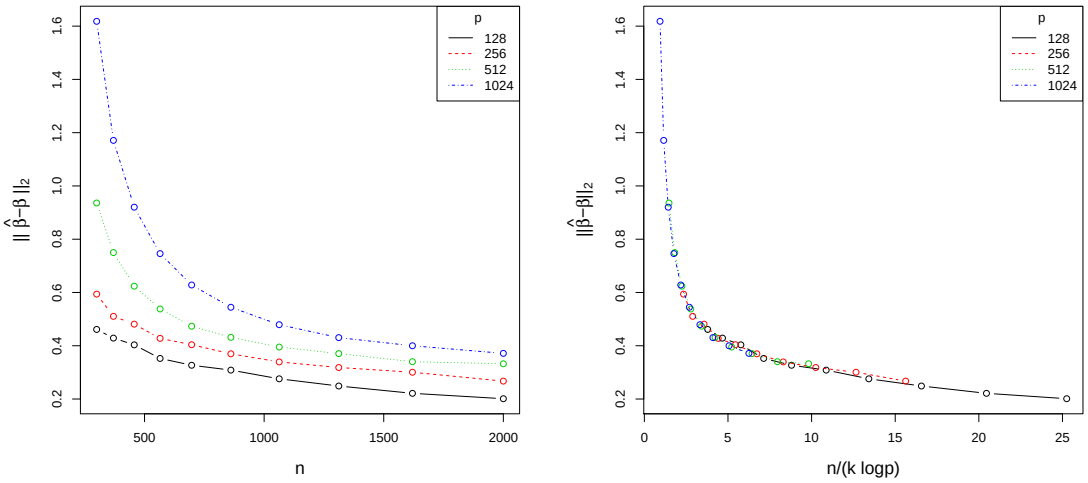
pendence measure (*Wu*, 2005). Proposition IV.11 relies on different structural and distributional assumptions on the underlying time series. On one hand, the results in the above paper are applicable on causal processes and require a specific decay assumption on the functional dependence measure, even for stationary linear processes (cf. Example 2.2, *Chen et al.* (2013)). Our results are applicable for non-causal processes as well, and do not assume any specific decay on the temporal dependence. On the other hand, their results are applicable under a mild moment condition on the distribution of the random variables, while our results are derived under stronger assumption of normality.

4.7 Numerical Experiments

We conduct numerical experiments to demonstrate the properties of ℓ_1-regularized estimates for stochastic regression and VAR estimation in finite samples. In the first subsection, we study the estimation error of lasso for stochastic regression, when the noise process is serially correlated. In the next subsection, we compare the performance of ℓ_1-penalized least squares and log-likelihood based estimates with different correlation structures of the error process Σ_ϵ.

4.7.1 Stochastic Regression

In the first experiment we demonstrate how the estimation error of lasso scales with n and p. We simulated observations from a p-dimensional ($p = 128, 256, 512, 1024$) stationary predictor process $\{X^t\}$ with independent components generated according to AR(2) processes $X_i^t = 0.4X_i^{t-1} - 0.16X_i^{t-2} + \xi^t$, where $\xi^t \sim N(0,1)$. We generated the errors $\{\epsilon^t\}$ according to a univariate MA(2) process $\epsilon^t = 0.4\epsilon^{t-1} - 0.16\epsilon^{t-2} + \eta^t$, where $\eta^t \sim N(0,1)$. For different values of p, we generated sparse vectors β^* with $k \approx \sqrt{p}$ non-zero entries, with a signal-to-noise ratio of 1.2. With a choice of tuning parameter $\lambda_n = \sqrt{\log p/n}$, we applied lasso on simulated samples of size



(a) $\|\hat{\beta} - \beta^*\|$ vs. n (b) $\|\hat{\beta} - \beta^*\|$ vs. $n/k \log p$

Figure 4.4: Estimation error of lasso $\|\hat{\beta} - \beta^*\|$ in stochastic regression with serially correlated error. Predictors $\{X_i^t\}$, $i = 1, \ldots, p$ are generated according to AR(2) processes and the errors are generated from MA(2) process. In the left panel, errors are plotted against sample size (n). For the same sample size, errors are higher for larger p. In the right panel, the errors are plotted against the rescaled sample size $n/k \log p$. The error curves align perfectly, showing the errors scale as $\sqrt{k \log p/n}$.

$n \in (100, 3000)$. The ℓ_2-error of estimation $\|\hat{\beta} - \beta^*\|$ is plotted in 4.4. The left panel displays the errors for different values of p, plotted against the sample size n. As expected, the errors are larger for larger p. The right panel displays the estimation errors against the rescaled sample size $n/k \log p$. The error curves for different values of p now align very well. This demonstrates that lasso can achieve an estimation error rate of $\sqrt{k \log p/n}$, even with stochastic predictors and serially correlated errors.

The second numerical experiment demonstrates how the estimation error changes with the dependence in data. We have simulated samples of size $n \in (100, 2000)$ from a $p = 500$-dimensional stationary process $\{X^t\}$ with independent components generated according to AR(2) process $X_i^t = 2\rho\, X_i^{t-1} - \rho^2\, X_i^{t-2} + \xi^t$, $\xi^t \sim N(0, 1)$, for $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. Larger values of ρ correspond to stronger temporal dependence in the data. The process is unstable for $\rho = 1$. We generated a sparse

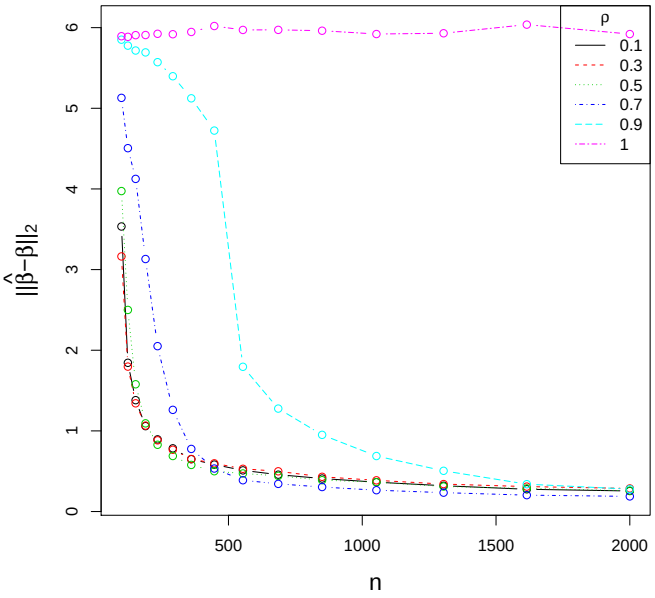


Figure 4.5: Estimation error $\|\hat{\beta} - \beta^*\|$ of lasso, for different degree of dependence in the data. $p = 500$ predictors $\{X_i^t\}$, $i = 1, \ldots, p$ are generated according to AR(2) process $X_i^t = 2\rho X_i^{t-1} - \rho^2 X_i^{t-2} + \xi^t$, $\xi^t \sim N(0, 1)$. With the same sample size n, the estimates have larger error for stronger dependence in the data, i.e., for larger ρ. The process of predictors is unstable for $\rho = 1$ and lasso is inconsistent.

signal β^* with $k = 25$ non-zero entries, and simulated serially correlated errors $\{\epsilon^t\}$ according to AR(2) process $\epsilon^t = 2\gamma\,\epsilon^{t-1} - \gamma^2\,\epsilon^{t-2} + \xi^t$, $\gamma = 0.2$, $\xi^t \sim N(0, 1)$. The signal-to-noise ratio was set to 1.2. With the response process $Y^t = \langle \beta^*, X^t \rangle + \epsilon^t$, we applied lasso and plotted the estimation errors $\|\hat{\beta} - \beta^*\|$ for different values of ρ and n in Figure 4.5. As expected, the errors are larger for stronger temporal dependence, i.e., larger values of ρ. For $\rho = 1$, the process of predictors is not stable and lasso estimate is no longer consistent. Interestingly, the estimation error of lasso changes with ρ in a highly non-linear fashion. The error curves for $\rho = 0.1$, 0.3, 0.5 are very close. The error curve for $\rho = 0.7$ is slightly higher and the error curve for $\rho = 0.9$ is farther apart from the rest.

4.7.2 VAR Estimation

We evaluate the performance of ℓ_1-LS and ℓ_1-LL on simulated data and compare it with the performance of ordinary least squares (OLS) and Ridge estimates. Implementing ℓ_1-LL requires an estimate of Σ_ϵ in the first step. For this, use the residuals from ℓ_1-LS to construct a plug-in estimate $\hat{\Sigma}_\epsilon$. To evaluate the effect of error correlation on the transition matrix estimates more precisely, we also implement an oracle version, ℓ_1-LL-O, which uses the true Σ_ϵ in the estimation. Next, we describe the simulation settings, choice of performance metrics and discuss the results.

We design two sets of numerical experiments - (a) SMALL VAR ($p = 10, d = 1, T = 30, 50$) and (b) MEDIUM VAR ($p = 30, d = 1, T = 80, 120, 160$). In each setting, we generate an adjacency matrix A_1 with $5 \sim 10\%$ non-zero edges selected at random and rescale to ensure that the process is stable with $SNR = 2$. We generate three different error processes with covariance matrix Σ_ϵ from one of the following families:

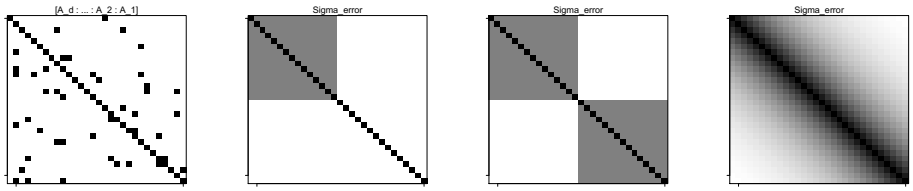
1. Block-I: $\Sigma_\epsilon = ((\sigma_{\epsilon,ij}))_{1\leq i,j\leq p}$ with $\sigma_{\epsilon,ii} = 1$, $\sigma_{\epsilon,ij} = \rho$ if $1 \leq i \neq j \leq p/2$, $\sigma_{\epsilon,ij} = 0$ otherwise;

2. Block-II: $\Sigma_\epsilon = ((\sigma_{\epsilon,ij}))_{1\leq i,j\leq p}$ with $\sigma_{\epsilon,ii} = 1$, $\sigma_{\epsilon,ij} = \rho$ if $1 \leq i \neq j \leq p/2$ or $p/2 < i \neq j \leq p$, $\sigma_{\epsilon,ij} = 0$ otherwise;

3. Toeplitz: $\Sigma_\epsilon = ((\sigma_{\epsilon,ij}))_{1\leq i,j\leq p}$ with $\sigma_{\epsilon,ij} = \rho^{|i-j|}$.

We let ρ vary in $\{0.5, 0.7, 0.9\}$. Larger values of ρ indicate that the error processes are more strongly correlated. Figure 4.6 illustrates the structure of a random transition matrix used in our simulation and the three different types of error covariance structure.

We compare the different methods for VAR estimation (OLS, ℓ_1-LS, ℓ_1-LL, ℓ_1-LL-O, Ridge) based on the following performance metrics:



(a) A_1 (b) Σ_ϵ: Block-I (c) Σ_ϵ: Block-II (d) Σ_ϵ: Toeplitz

Figure 4.6: Adjacency matrix A_1 and error covariance matrix Σ_ϵ of different types used in the simulation studies

1. *Model Selection:* Area under ROC curve (AUROC)

2. *Estimation error:* Relative estimation accuracy $\|\hat{A}_1 - A_1\|_F / \|A_1\|_F$

Table 4.1: VAR(1) model with $p = 10$, $T = 30$

		BLOCK-I			BLOCK-II			Toeplitz		
	ρ	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
AUROC	ℓ_1-LS	0.77	0.74	0.7	0.79	0.76	0.74	0.82	0.79	0.77
	ℓ_1-LL	0.77	0.75	0.73	0.79	0.77	0.77	0.81	0.8	0.81
	ℓ_1-LL-O	0.8	0.79	0.76	0.82	0.8	0.81	0.85	0.84	0.84
Estimation	OLS	1.24	1.39	1.77	1.29	1.63	2.36	1.32	1.56	2.58
Error	ℓ_1-LS	0.68	0.72	0.76	0.64	0.67	0.7	0.63	0.66	0.69
	ℓ_1-LL	0.66	0.66	0.66	0.57	0.59	0.53	0.59	0.56	0.49
	ℓ_1-LL-O	0.61	0.62	0.62	0.53	0.54	0.47	0.53	0.51	0.42
	ridge	0.72	0.74	0.75	0.7	0.71	0.72	0.7	0.71	0.72

We report the results for small VAR with $T = 30$ and medium VAR with $T = 120$ (averaged over 50 replicates) in Tables 4.1 and 4.2. The results in the other settings are qualitatively similar, although the overall accuracy changes with the sample size. We find that the regularized VAR estimates outperform ordinary least squares uniformly in all the cases.

In terms of model selection, the ℓ_1-penalized estimates perform fairly well, as reflected in their AUROC. Ordinary least squares and ridge regression do not perform any model selection. Further, for all three choices of error covariance, the two variants of ℓ_1-LL outperform ℓ_1-LS. The difference in their performances is more prominent for larger values of ρ. Among the three covariance structures, the difference between

least squares and log-likelihood based methods is more prominent in Block-II and Toeplitz family since the error processes are more strongly correlated. Finally, in all the cases, the accuracy of ℓ_1-LL is somewhere in between ℓ_1-LS and ℓ_1-LL-O, which suggests that a more accurate estimation of Σ_ϵ might improve the model selection performance of regularized VAR estimates.

In terms of estimation error, the conclusions are broadly the same. The effect of over-fitting is reflected in the performance of ordinary least squares. In many settings, the estimation error of ordinary least squares is even twice as large as the signal strength. The performance of ordinary least squares deteriorates when the error processes are more strongly correlated (see, for example, $\rho = 0.9$ for block-II). Ridge regression performs better than ordinary least squares as it applies shrinkage on the coefficients. However the ℓ_1-penalized estimates show higher accuracy than Ridge in almost all the cases. This is somewhat expected as the data were simulated from a sparse model with strong signals, whereas Ridge regression tend to favor a non-sparse model with many small coefficients.

Table 4.2: VAR(1) model with $p = 30$, $T = 120$

		BLOCK-I			BLOCK-II			Toeplitz		
	ρ	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
AUROC	ℓ_1-LS	0.89	0.85	0.77	0.87	0.81	0.69	0.91	0.87	0.76
	ℓ_1-LL	0.89	0.87	0.82	0.9	0.89	0.88	0.91	0.91	0.89
	ℓ_1-LL-O	0.92	0.9	0.84	0.93	0.92	0.9	0.94	0.93	0.92
Estimation	OLS	1.73	2	2.93	1.95	2.53	4.28	1.82	2.28	3.88
Error	ℓ_1-LS	0.72	0.76	0.85	0.74	0.82	0.93	0.69	0.73	0.86
	ℓ_1-LL	0.71	0.71	0.72	0.68	0.68	0.65	0.67	0.63	0.6
	ℓ_1-LL-O	0.66	0.66	0.68	0.64	0.63	0.59	0.63	0.59	0.54
	Ridge	0.81	0.83	0.85	0.82	0.85	0.88	0.81	0.82	0.86

4.8 Technical Results

4.8.1 Results on Stochastic Regression

Proof of Proposition IV.5. Let us recall that $S = \mathcal{X}'\mathcal{X}/n$, $J = supp(\beta^*)$ with $|J| = k$, $\mathcal{C}(J, \kappa) = \{v \in \mathbb{R}^p : \|v_{J^c}\|_1 \leq \kappa \|v_J\|_1\}$ and $\mathcal{K}(s) = \mathbb{B}_0(s) \cap \mathbb{B}_2(1)$, for any $s \geq 1$. We

need a positive lower bound on $v'Sv/\|v\|^2$, uniformly over all $v \in \mathcal{C}(J,3)\backslash\{0\}$, that holds with high probability. Assuming $\|v\| = 1$ does not result in any loss of generality since $v \in \mathcal{C}(J,3)\backslash\{0\}$ if and only if $v/\|v\| \in \mathcal{C}(J,3)\backslash\{0\}$. If $\mathfrak{m}(f_X) > 0$, the lower bound in Proposition IV.3 ensures that

$$\inf_{v\in\mathcal{C}(J,3),\,\|v\|=1} v'\Gamma_X(0)v \geq 2\pi\mathfrak{m}(f_X) > 0 \tag{4.39}$$

So it remains to show that $v'(S - \Gamma_X(0))v$ is sufficiently small, uniformly for all $v \in \mathcal{C}(J,3)$ with $\|v\| = 1$. We start with the single deviation bound of (4.17) with a $2k$-sparse v, $\|v\| = 1$:

$$\mathbb{P}\left[|v'\left(S - \Gamma_X(0)\right)v| > 2\pi\mathcal{M}(f_X, 2k)\eta\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\}\right]$$

Using a discretization argument presented in Lemma 4.8.7, we can extend it to the following uniform lower bound on all $2k$-sparse vectors v of unit norm:

$$\begin{aligned} &\mathbb{P}\left[\sup_{v\in\mathcal{K}(2k)} |v'\left(S - \Gamma_X(0)\right)v| > 2\pi\mathcal{M}(f_X, 2k)\eta\right] \\ \leq\ & 2\exp\left[-cn\min\{\eta, \eta^2\} + 2k\min\{\log p,\, \log\left(21ep/2k\right)\}\right] \end{aligned} \tag{4.40}$$

In the next step, we use Lemma 4.8.6 to conclude that the set $\mathcal{C}(J,3) \cap \mathbb{B}_2(1)$ is contained in a closed, convex hull of k-sparse vectors $5cl\{conv\{\mathcal{K}(k)\}\}$. This, together with the approximation of Lemma 4.8.8, leads to the following upper bound

$$\begin{aligned} \sup_{v\in\mathcal{C}(J,3),\,\|v\|=1} |v'(S - \Gamma_X(0))v| &\leq \sup_{v\in 5cl\{conv\{\mathcal{K}(k)\}\}} |v'(S - \Gamma_X(0))v| \\ &= 25 \sup_{v\in cl\{conv\{\mathcal{K}(k)\}\}} |v'(S - \Gamma_X(0))v| \\ &\leq 75 \sup_{v\in\mathcal{K}(2k)} |v'(S - \Gamma_X(0))v| \end{aligned}$$

Using the deviation bound of (4.40) and $\min\{\eta, \eta^2\} \leq \min\{1, \eta^2\}$, we have

$$\begin{aligned} & \mathbb{P}\left[\sup_{v \in \mathcal{C}(J,3),\, \|v\|=1} |v'(S - \Gamma_X(0))v| > 150\pi \mathcal{M}(f_X, 2k)\eta\right] \\ \leq \;& 2\exp\left[-cn\min\{1, \eta^2\} + 2k\min\{\log p, \log(21ep/2k)\}\right] \end{aligned}$$

Setting $\eta = \mathfrak{m}(f_X)/150\mathcal{M}(f_X, 2k)$ and combining this deviation bound with (4.39), we obtain the final result.

Note that similar lower bounds can be derived if, instead of assuming $\mathfrak{m}(f_X) > 0$, one assumes $\Lambda_{\min}(\Gamma_X(0)) > 0$, or $\alpha := \inf_{v \in \mathcal{C}(J,3)\setminus\{0\}} v'\Gamma_X(0)v/\|v\|^2 > 0$. In these cases, $2\pi\mathfrak{m}(f_X)$ is replaced by $\Lambda_{\min}(\Gamma_X(0))$ or α in (4.39). □

Proof of Proposition IV.6. We need an upper bound on $\|\mathcal{X}'E/n\|_\infty$ that holds with high probability. To this end, note that for any $j \in \{1, \ldots, p\}$,

$$\begin{aligned} 2\mathcal{X}_j'E/n &= \frac{1}{n}\left[\|\mathcal{X}_j + E\|^2 - n\mathrm{Var}(X_j^1 + \epsilon^1)\right] \\ &\quad + \frac{1}{n}\left[\|\mathcal{X}_j\|^2 - n\mathrm{Var}(X_j^1)\right] + \frac{1}{n}\left[\|E\|^2 - n\mathrm{Var}(\epsilon^1)\right] \end{aligned}$$

So it suffices to derive deviation bounds for each of the terms on the right.

Term III: The deviation bound (4.17) for the time series $\{\epsilon^t\}$ with $p = 1$, $v = 1$, $k = 1$ gives

$$\mathbb{P}\left[\frac{1}{n}\left|\|E\|^2 - n\mathrm{Var}(\epsilon^1)\right| > 2\pi\mathcal{M}(f_\epsilon)\eta\right] \leq 2\exp\left[-cn\min\{\eta^2, \eta\}\right]$$

Term II: Applying the deviation bound (4.17) for the time series $\{X_j^t\}$ with $p = 1$, $v = 1$, $k = 1$ and using (4.9), we have

$$\mathbb{P}\left[\frac{1}{n}\left|\|\mathcal{X}_j\|^2 - n\mathrm{Var}(X_j^1)\right| > 2\pi\mathcal{M}(f_X, 1)\eta\right] \leq 2\exp\left[-cn\min\{\eta^2, \eta\}\right]$$

Term I: Setting $Z^t = X_j^t + \epsilon^t$ and using (4.11) with the deviation bound (4.17), we conclude

$$\mathbb{P}\left[\frac{1}{n}\left|\|\mathcal{X}_j + E\|^2 - n\mathrm{Var}(X_j^1 + \epsilon^1)\right| > 2\pi\left[\mathcal{M}(f_X, 1) + \mathcal{M}(f_\epsilon)\right]\eta\right]$$

is at most $2\exp\left[-cn\min\{\eta^2, \eta\}\right]$.

Putting the three concentration bounds together, we obtain, for any $j \in \{1, \ldots, p\}$,

$$\mathbb{P}\left[\frac{1}{n}\left|\mathcal{X}_j' E\right| > 2\pi\eta\left[\mathcal{M}(f_X, 1) + \mathcal{M}(f_\epsilon)\right]\right] \leq 6\exp\left[-cn\min\{\eta^2, \eta\}\right]$$

Taking an union bound over all j, we have:

$$\mathbb{P}\left[\max_{1\leq j\leq p}\frac{1}{n}\left|\mathcal{X}_j' E\right| > 2\pi\eta\left[\mathcal{M}(f_X, 1) + \mathcal{M}(f_\epsilon)\right]\right] \leq 6p\exp\left[-cn\min\{\eta^2, \eta\}\right]$$

Setting $\eta = c_0\sqrt{\frac{\log p}{n}}$ and using the fact that $n \gtrsim \log p$, we have the required result. □

Proof of Proposition IV.7. The events of Propositions IV.5 and IV.6 hold with probability $1 - c_1\exp\left[-c_2\log p\right]$ for some $c_i > 0$, under the assumptions on n and λ_n. Denote $v = \hat{\beta} - \beta^*$ and $J = supp(\beta^*)$ so that $|J| = k$. Then we have,

$$\frac{1}{n}\|Y - \mathcal{X}\hat{\beta}\|^2 + \lambda_n\|\hat{\beta}\|_1 \leq \frac{1}{n}\|Y - \mathcal{X}\beta^*\|^2 + \lambda_n\|\beta^*\|_1$$

After some algebra, this reduces to

$$v'Sv - \frac{2}{n}v'\left(\mathcal{X}'E\right) \leq \lambda_n\|\beta^*\|_1 - \lambda_n\|\beta^* + v\|_1$$

With the proposed choice of λ_n, we have

$$\begin{aligned}
0 \leq v'Sv &\leq \frac{\lambda_n}{2}\|v\|_1 + \lambda_n\|\beta^*\|_1 - \lambda_n\|\beta^* + v\|_1 \\
&\leq \frac{\lambda_n}{2}\|v\|_1 + \lambda_n\left(\|\beta^*_J\|_1 - \|\beta^*_J + v_J\|_1 - \|v_{J^c}\|_1\right), \text{ since } \beta^*_{J^c} = 0 \\
&\leq \frac{\lambda_n}{2}\left(\|v_J\|_1 + \|v_{J^c}\|_1\right) + \lambda_n\left(\|v_J\|_1 - \|v_{J^c}\|_1\right), \text{ by triangle inequality} \\
&\leq \frac{3\lambda_n}{2}\|v_J\|_1 - \frac{\lambda_n}{2}\|v_{J^c}\|_1
\end{aligned}$$

This ensures $\|v_J\|_1 \leq 3\|v_{J^c}\|_1$, i.e., $v \in \mathcal{C}(J,3)$ and $v'Sv \leq 2\lambda_n\|v_J\|_1 \leq 2\lambda_n\sqrt{k}\|v\|$. Using RE condition, we have

$$\alpha_{RE}\|v\|^2 \leq v'Sv \leq 2\lambda_n\sqrt{k}\|v\|$$

This implies

$$\begin{aligned}
\|v\| &\leq \frac{2\lambda_n\sqrt{k}}{\alpha_{RE}} \\
\|v\|_1 &\leq 4\|v_J\|_1 \leq 4\sqrt{k}\|v_J\| \leq \frac{8\lambda_n k}{\alpha_{RE}} \\
\|v'Sv\| &\leq \frac{4\lambda_n^2 k}{\alpha_{RE}}
\end{aligned}$$

To derive the upper bound on the number of false positives selected by the thresholded lasso, note that

$$\begin{aligned}
\left|supp(\tilde{\beta})\backslash supp(\beta^*)\right| &= \sum_{j\notin J}\mathbf{1}_{\{|\hat{\beta}_j|>\lambda_n\}} \leq \sum_{j\notin J}\left|\hat{\beta}_j\right|/\lambda_n \\
&\leq \frac{1}{\lambda_n}\sum_{j\notin J}|v_j| \leq \frac{3}{\lambda_n}\sum_{j\in J}|v_j| \leq \frac{3\left\|v\right\|_1}{\lambda_n} \leq \frac{24k}{\alpha_{RE}}
\end{aligned}$$

□

4.8.2 Results on VAR Estimation

Proof of Proposition IV.8. Since $\hat{\beta}$ is a minimizer of (4.30), for all $\beta \in \mathbb{R}^q$ we have

$$-2\hat{\beta}'\hat{\gamma} + \hat{\beta}'\hat{\Gamma}\hat{\beta} + \lambda_N\|\hat{\beta}\|_1 \leq -2\beta'\hat{\gamma} + \beta'\hat{\Gamma}\beta + \lambda_N\|\beta\|_1$$

For $\beta = \beta^*$, the above inequality reduces to

$$v'\hat{\Gamma}v \leq 2v'(\hat{\gamma} - \hat{\Gamma}\beta^*) + \lambda_N\left\{\|\beta^*\|_1 - \|\beta^* + v\|_1\right\} \tag{4.41}$$

where $v = \hat{\beta} - \beta^*$.

The first term on the right hand side of (4.41) is at most $2\|v\|_1\mathbb{Q}(\beta^*, \Sigma_\epsilon)\sqrt{\log q/N}$. The second term, by triangle inequality, is at most $\lambda_N\{\|v_J\|_1 - \|v_{J^c}\|_1\}$, where J denotes the support of β^*. Together with the proposed choice of λ_N, this leads to the following inequality

$$\begin{aligned} 0 \leq v'\hat{\Gamma}v &\leq \frac{\lambda_N}{2}\left\{\|v_J\|_1 + \|v_{J^c}\|_1\right\} + \lambda_N\left\{\|v_J\|_1 - \|v_{J^c}\|_1\right\} \\ &\leq \frac{3\lambda_N}{2}\|v_J\|_1 - \frac{\lambda_N}{2}\|v_{J^c}\|_1 \leq 2\lambda_N\|v\|_1 \end{aligned} \tag{4.42}$$

In particular, this ensures $\|v_{J^c}\|_1 \leq 3\|v_J\|_1$ so that $\|v\|_1 \leq 4\|v_J\|_1 \leq 4\sqrt{k}\|v\|$. From the restricted eigenvalue assumption and the upper bound on $k\tau(N, q)$, we have

$$v'\hat{\Gamma}v \geq \alpha\|v\|^2 - \tau(N,q)\|v\|_1^2 \geq (\alpha - 16k\tau(N,q))\|v\|^2 \geq \frac{\alpha}{2}\|v\|^2$$

Together, the upper and lower bounds on $v'\hat{\Gamma}v$ guarantee that

$$\frac{\alpha}{4}\|v\|^2 \leq \lambda_N\|v\|_1 \leq 4\sqrt{k}\lambda_N\|v\|$$

This implies

$$\|v\| \leq 16\sqrt{k}\lambda_N/\alpha$$
$$\|v\|_1 \leq 4\sqrt{k}\lambda_N\|v\| \leq 64k\lambda_N/\alpha$$
$$v'\hat{\Gamma}v \leq 2\lambda_N\|v\|_1 \leq 128k\lambda_N^2/\alpha$$

To derive the upper bound on the number of false positives selected by thresholded lasso, note that

$$\begin{aligned}\left|supp(\tilde{\beta})\backslash supp(\beta^*)\right| &= \sum_{j\notin J}\mathbf{1}_{\{|\hat{\beta}_j|>\lambda_N\}} \leq \sum_{j\notin J}\left|\hat{\beta}_j\right|/\lambda_N \\ &\leq \frac{1}{\lambda_N}\sum_{j\notin J}|v_j| \leq \frac{3}{\lambda_N}\sum_{j\in J}|v_j| \leq \frac{3\|v\|_1}{\lambda_N} \leq \frac{192k}{\alpha}\end{aligned}$$

□

Proof of Proposition IV.9. Note that the matrix $\hat{\Gamma}$ takes the form $I_p \otimes (\mathcal{X}'\mathcal{X}/N)$ and $\Sigma_\epsilon^{-1}\otimes(\mathcal{X}'\mathcal{X}/N)$ for ℓ_1-LS and ℓ_1-LL, respectively. To prove that $\hat{\Gamma}$ satisfies RE, we first show that the random matrix $S = \mathcal{X}'\mathcal{X}/N$ satisfies RE(α, τ) with high probability, for some $\alpha > 0$, $\tau > 0$. Then we invoke Lemma 4.8.1 to extend the result to $\hat{\Gamma}$.

To prove that $S = \mathcal{X}'\mathcal{X}/N$ satisfies RE condition, note that the rows of the design matrix $\mathcal{X}$ are sequentially generated according to a stable VAR(1) process $\{\tilde{X}^t\}$, as defined in (4.25). In particular, each row of $\mathcal{X}$ is centered Gaussian with covariance $\Gamma_{\tilde{X}}(0)$. Now $\Gamma_{\tilde{X}}(0) = \Upsilon_1^{\tilde{X}} = \Upsilon_d^X$, where Υ_n^X is the covariance of the vectorized data matrix containing n observations generated according to the process $\{X^t\}$, as defined in Section 4.2.2. Hence, from Proposition IV.3 and the bounds in (4.24), we have

$$\Lambda_{\min}\left(\Gamma_{\tilde{X}}(0)\right) \geq \frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})} \tag{4.43}$$

Also, from Proposition IV.4 and (4.24), we have, for any $v \in \mathbb{R}^{dp}$, $\|v\| \leq 1$, and any

$\eta > 0$,

$$\mathbb{P}\left[|v'\left(S - \Gamma_{\tilde{X}}(0)\right)v| > \eta\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})}\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\}\right] \tag{4.44}$$

The next step is to extend the deviation bound (4.44) for a single v to an appropriate set of sparse vectors $\mathcal{K}(2s) := \{v \in \mathbb{R}^{dp} : \|v\| \leq 1, \|v\|_0 \leq 2s\}$, for an integer $s \geq 1$ to be specified later. Using the discretization argument of Lemma 4.8.7, we have,

$$\mathbb{P}\left[\sup_{v\in\mathcal{K}(2s)}|v'\left(S - \Gamma_{\tilde{X}}(0)\right)v| > \eta\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})}\right]$$

is at most $2\exp\left[-cN\min\{\eta, \eta^2\} + 2s\min\{\log(dp), \log(21e\,dp/2s)\}\right]$.

Next, we set $\eta = \omega^{-1}$ with $c_3 = 54$ and note that $\min\{\eta, \eta^2\} \geq \min\{1, \eta^2\}$. Applying Supplementary Lemma 12 in *Loh and Wainwright* (2012) with $\delta = \Lambda_{\min}(\Sigma_\epsilon)/54\mu_{\max}(\mathcal{A})$ and $\Gamma = S - \Gamma_{\tilde{X}}(0)$, we have

$$v'Sv \geq \alpha\|v\|^2 - \frac{\alpha}{s}\|v\|_1^2, \quad \text{for all } v \in \mathbb{R}^{dp}$$

with probability at least $1 - 2\exp\left[-cN\min\{\omega^{-2}, 1\} + 2s\log(dp)\right]$.

Finally, we set $s = \lceil cN\ \min\{\omega^{-2}, 1\}/4\log(dp)\rceil$ [note that $s \geq 1$ with the required choice of N] to conclude that $S \sim RE(\alpha, \tau)$ with high probability. □

Lemma 4.8.1 (RE condition for $\hat{\Gamma}$)**.** *If $\mathcal{X}'\mathcal{X}/N \sim RE(\alpha, \tau)$, then so does $I_p \otimes \mathcal{X}'\mathcal{X}/N$. Further, if Σ_ϵ^{-1} satisfies $\bar{\sigma}_\epsilon^i := \sigma_\epsilon^{ii} - \sum_{j\neq i}\sigma_\epsilon^{ij} > 0$, for $i = 1, \ldots, p$, then*

$$\Sigma_\epsilon^{-1} \otimes \mathcal{X}'\mathcal{X}/N \sim RE\left(\alpha\ \min_i \bar{\sigma}_\epsilon^i, \tau\ \max_i \bar{\sigma}_\epsilon^i\right)$$

Proof. $S = \mathcal{X}'\mathcal{X}/N \sim RE(\alpha, \tau)$. Consider $\hat{\Gamma} = I_p \otimes S$. For any $\theta \in \mathbb{R}^{dp^2}$ with $\theta' = (\theta_1', \ldots, \theta_p')'$, each $\theta_i \in \mathbb{R}^{dp}$, we have

$$\theta'(I_p \otimes S)\theta = \sum_{r=1}^{p}\theta_r' S\theta_r \geq \alpha\sum_{r=1}^{p}\|\theta_r\|^2 - \tau\sum_{r=1}^{p}\|\theta_r\|_1^2 \geq \alpha\|\theta\|^2 - \tau\|\theta\|_1^2$$

proving the first part. To prove the second part, note that

$$\theta'(\Sigma_\epsilon^{-1} \otimes S)\theta = \sum_{r,s=1}^{p} \sigma_\epsilon^{rs} \theta_r' S \theta_s = \sum_{r=1}^{p} \sigma_\epsilon^{rr} \theta_r' S \theta_r + \sum_{r \neq s}^{p} \sigma_\epsilon^{rs} \theta_r' S \theta_s$$

Since the matrix S is non-negative definite, $\theta_r' S \theta_s \geq -\frac{1}{2}(\theta_r' S \theta_r + \theta_s' S \theta_s)$ for every $r \neq s$. This implies

$$\begin{aligned}
\theta'(\Sigma_\epsilon^{-1} \otimes S)\theta &\geq \sum_{r=1}^{p} \sigma_\epsilon^{rr} \theta_r' S \theta_r - \sum_{r<s} \sigma_\epsilon^{rs} (\theta_r' S \theta_r + \theta_s' S \theta_s) \\
&= \sum_{r=1}^{p} \left(\sigma_\epsilon^{rr} - \sum_{r \neq s} \sigma_\epsilon^{rs} \right) \theta_r' S \theta_r = \sum_{r=1}^{p} \bar{\sigma}_\epsilon^{r} \theta_r' S \theta_r \\
&\geq \alpha \sum_{r=1}^{p} \bar{\sigma}_\epsilon^{r} \|\theta_r\|^2 - \tau \sum_{r=1}^{p} \bar{\sigma}_\epsilon^{r} \|\theta_r\|_1^2 \\
&\geq \left(\alpha \min_i \bar{\sigma}_\epsilon^{i} \right) \|\theta\|^2 - \left(\tau \max_i \bar{\sigma}_\epsilon^{i} \right) \|\theta\|_1^2
\end{aligned}$$

□

Proof of Proposition IV.10. We want to establish an upper bound on $\|\hat{\gamma} - \hat{\Gamma}\beta^*\|_\infty$ that holds with high probability. To this end, we first note that in the context of (4.30),

$$\begin{aligned}
\hat{\gamma} &= (W \otimes \mathcal{X}')\,(I_p \otimes \mathcal{X})\,\beta^*/N + (W \otimes \mathcal{X}')\,vec(E)/N \\
\hat{\Gamma}\beta^* &= (W \otimes \mathcal{X}'\mathcal{X}/N)\,\beta^*
\end{aligned}$$

which implies $\hat{\gamma} - \hat{\Gamma}\beta^* = (W \otimes \mathcal{X}')\,vec(E)/N = vec(\mathcal{X}'EW)/N$.

For ℓ_1-LS, $EW = E$, a matrix with independent rows, each row $\sim N(\mathbf{0}, \Sigma_\epsilon)$. For ℓ_1-LL, $EW = \bar{E}$, a matrix with independent rows, each row $\sim N(\mathbf{0}, \Sigma_\epsilon^{-1})$. Note that in both cases i^{th} row of E or $\bar{E}$ is independent of the i^{th} row of $\mathcal{X}$. First, we present the argument for ℓ_1-LS.

For any $l \in \{1, \ldots, dp\}$, any $k \in \{1, \ldots, p\}$, we first derive an upper bound on

$\mathbb{P}(|\mathcal{X}_l' E_k/N| > t)$. Taking union bound over all l, k then leads to the final upper bound on the tail probability of the maximum

$$\mathbb{P}\left(\frac{1}{N}\|\mathcal{X}'E\|_{\max} > t\right) = \mathbb{P}\left(\max_{1\le l\le dp, 1\le k\le p} \frac{1}{N}|\mathcal{X}_l' E_k| > t\right)$$

Note that for any given l, $1 \le l \le dp$, there exist unique j, $h > 0$ such that $l = p(h-1)+j$, $1 \le h \le d$, $1 \le j \le p$. The l^{th} column of $\mathcal{X}$ and the k^{th} column of E are precisely

$$\mathcal{X}(j,h) := \mathcal{X}_l = \begin{bmatrix} X_j^{T-h} \\ X_j^{T-1-h} \\ \vdots \\ X_j^{d-h} \end{bmatrix}, \; E_k = \begin{bmatrix} \epsilon_k^T \\ \epsilon_k^{T-1} \\ \vdots \\ \epsilon_k^d \end{bmatrix}$$

We will use $\mathcal{X}(j,h)$ and $\mathcal{X}_l$ interchangeably for notational convenience. First note that

$$\begin{aligned} \frac{2}{N}\mathcal{X}(j,h)'E_k &= \frac{1}{N}\left[\|\mathcal{X}(j,h)+E_k\|^2 - N\mathrm{Var}(X_j^{T-h}+\epsilon_k^T)\right] \\ &\quad -\frac{1}{N}\left[\|\mathcal{X}(j,h)\|^2 - N\mathrm{Var}(X_j^{T-h})\right] - \frac{1}{N}\left[\|E_k\|^2 - N\mathrm{Var}(\epsilon_k^T)\right] \end{aligned}$$

Next we establish deviation bound for each of the three terms on the right.

Term III: $E_k \sim N(\mathbf{0}, Q)$ with $Q = (e_k'\Sigma_\epsilon e_k) I_N$, so that $\|Q\| \le \Lambda_{\max}(\Sigma_\epsilon)$. So, by Hansen-Wright inequality of Lemma 4.8.5, we have

$$\mathbb{P}\left(\frac{1}{N}\left|\|E_k\|^2 - N\,\mathrm{Var}(\epsilon_k^T)\right| > \eta\Lambda_{\max}(\Sigma_\epsilon)\right) \le 2\exp\left[-cN\min\{\eta,\eta^2\}\right]$$

Term II: $\mathcal{X}(j,h) \sim N(\mathbf{0}, Q)$ with $Q_{rs} = e_j'\Gamma_X(r-s)e_j$, so that $tr(Q) = N\mathrm{Var}(X_j^{T-h})$.

Also, for any $u \in \mathbb{R}^N$ with $\|u\| = 1$,

$$u'Qu = \sum_{r=1}^{N}\sum_{s=1}^{N} u_r u_s e_j' \Gamma_X(r-s) e_j = (u \otimes e_j)' \Upsilon_N^X (u \otimes e_j) \leq \Lambda_{\max}\left(\Upsilon_N^X\right) \leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}$$

so that $\|Q\| \leq \Lambda_{\max}(\Sigma_\epsilon)/\mu_{\min}(\mathcal{A})$. Again, by Hansen-Wright inequality of Lemma 4.8.5, we have

$$\mathbb{P}\left(\frac{1}{N}\left|\|\mathcal{X}(j,h)\|^2 - N\,\mathrm{Var}(X_j^{T-h})\right| > \eta \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}\right) \leq 2\exp\left[-cN\min\{\eta,\eta^2\}\right]$$

Term I: The N-dimensional random vector $\mathcal{X}(j,h) + E_k$ is centered Gaussian with covariance matrix Q, where

$$Q_{rs} = \mathrm{Cov}(X_j^{T-h-r+1} + \epsilon_k^{T-r+1}, X_j^{T-h-s+1} + \epsilon_k^{T-s+1}),\ 1 \leq r,s \leq N$$

To apply Lemma 4.8.5, we need an upper bound on $\|Q\|$. To this end, note that

$$\begin{aligned} Q_{rs} &= \mathrm{Cov}(X_j^{T-h-r+1}, X_j^{T-h-s+1}) + \mathrm{Cov}(X_j^{T-h-s+1}, \epsilon_k^{T-r+1}) \\ &+ \mathrm{Cov}(X_j^{T-h-r+1}, \epsilon_k^{T-s+1}) + \mathrm{Cov}(\epsilon_k^{T-r+1}, \epsilon_k^{T-s+1}) \\ &= e_j'\Gamma_X(r-s)e_j + e_j'\Delta(s,r)e_k + e_j'\Delta(r,s)e_k + \mathbf{1}_{\{r=s\}}\, e_k'\Sigma_\epsilon e_k \end{aligned}$$

where $\Delta(r,s)_{p\times p}$ is the $(r,s)^{th}$ block of the covariance matrix

$$\Delta := \mathrm{Cov}\left(\begin{bmatrix} X^{T-h} \\ X^{T-1-h} \\ \vdots \\ X^{d-h} \end{bmatrix}, \begin{bmatrix} \epsilon^{T} \\ \epsilon^{T-1} \\ \vdots \\ \epsilon^{d} \end{bmatrix}\right), \quad \Delta(r,s) = \mathrm{Cov}(X^{T-h-r+1}, \epsilon^{T-s+1}) \quad 1 \leq r,s \leq N \tag{4.45}$$

This implies that for any $u \in \mathbb{R}^N$, $\|u\| = 1$,

$$\begin{aligned} u'Qu &= \sum_{r=1}^{N}\sum_{s=1}^{N} u_r u_s Q_{rs} \\ &= (u \otimes e_j)' \Upsilon_N^X (u \otimes e_j) + 2 (u \otimes e_j)' \Delta (u \otimes e_k) + e_k' \Sigma_\epsilon e_k \end{aligned}$$

Since $\|u \otimes e_j\| = \|u \otimes e_k\| = 1$, it follows from the upper bounds in Lemma 4.8.2 and Proposition IV.3 that

$$\|Q\| \leq \Lambda_{\max}(\Sigma_\epsilon) \left[1 + (1 + 2\mu_{\max}(\mathcal{A}))/\mu_{\min}(\mathcal{A})\right]$$

Once again, using Hansen-Wright inequality of Lemma 4.8.5, we have

$$\mathbb{P}\left(\frac{1}{N}\left|\|\mathcal{X}(j,h) + E_k\|^2 - N\mathrm{Var}(X_j^{T-h} + \epsilon_k^T)\right| > \eta\Lambda_{\max}(\Sigma_\epsilon)\left[1 + \frac{1 + 2\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right]\right)$$

is at most $2\exp\left[-cN\min\{\eta, \eta^2\}\right]$.

Putting together the deviation bounds for Terms *I* - *III*, we have

$$\mathbb{P}\left(\frac{1}{N}\left|\mathcal{X}(j,h)'E\right| > \eta\Lambda_{\max}(\Sigma_\epsilon)\left[1 + \frac{1 + \mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right]\right) \leq 6\exp\left[-cN\min\{\eta, \eta^2\}\right]$$

Taking an union bound over all j, h and setting $\eta = c_0\sqrt{\log q\,/N}$ yields the final result.

The proof for ℓ_1-LL can be derived exactly along the same line. For term III, we have $\|Q\| \leq 1/\Lambda_{\min}(\Sigma_\epsilon)$. For term II, $\|Q\|$ remains the same. For term I,

$$\|Q\| \leq \frac{1}{\Lambda_{\min}(\Sigma_\epsilon)} + \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} + 2\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\Lambda_{\min}(\Sigma_\epsilon)}\frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}$$

The additional $\Lambda_{\min}(\Sigma_\epsilon)$ in the denominator of the third expression appears because

the $(r,s)^{th}$ block of Δ in lemma 4.8.2 now changes to

$$\begin{aligned}\Delta(r,s) &= \mathrm{Cov}(X^{T-h-r+1}, \Sigma_\epsilon^{-1}\epsilon^T) \\ &= \left[(\Gamma_X(r-s+h) - \ldots - \Gamma_X(r-s+h-d)A_d')\right]\Sigma_\epsilon^{-1}\end{aligned}$$

Setting η as before and taking union bounds over all j, h yield the final result. □

Lemma 4.8.2 (Bounding $u'\Delta v$)**.** *Consider* Δ*, as defined in* (4.45)*. For any* u, $v \in \mathbb{R}^{Np}$ *with* $\|u\| = \|v\| = 1$*,* $|u'Qv| \leq \Lambda_{\max}(\Sigma_\epsilon)\mu_{\max}(\mathcal{A})/\mu_{\min}(\mathcal{A})$*.*

Proof.

$$\begin{aligned}\Delta &= \mathrm{Cov}\left(\begin{bmatrix} X^{T-h} \\ \vdots \\ X^{d-h} \end{bmatrix}, \begin{bmatrix} X^T - A_1X^{T-1} - \ldots - A_dX^{T-d} \\ \vdots \\ X^d - A_1X^{d-1} - \ldots - A_dX^0 \end{bmatrix}\right) \\ \Delta(r,s) &= \left[(\Gamma(r-s+h) - \Gamma(r-s+h-1)A_1' - \ldots - \Gamma(r-s+h-d)A_d')\right],\end{aligned}$$

for any r, s, $1 \leq r, s \leq N$.

For any u, $v \in \mathbb{R}^{Np}$ with $\|u\| = 1, \|v\| = 1$, define $G(\theta) = \sum_{r=1}^N u^r e^{-ir\theta}$, $H(\theta) = \sum_{r=1}^N v^r e^{-ir\theta}$. It is easy to check that $\int_{-\pi}^{\pi} G^*(\theta)G(\theta)\,d\theta = 2\pi$, $\int_{-\pi}^{\pi} H^*(\theta)H(\theta)\,d\theta = 2\pi$. Then, using the representation of (4.7), we can write

$$\begin{aligned}u'\,\Delta\,v &= \sum_{r,s=1}^N (u^r)' \left(\int_{-\pi}^{\pi} f(\theta)e^{i(r-s+h)\theta}\mathcal{A}^*(e^{i\theta})\,d\theta\right) v^s \\ &= \int_{-\pi}^{\pi} \left[\sum_{r=1}^N (u^r)'e^{ir\theta}\right] f(\theta)e^{ih\theta}\mathcal{A}^*(e^{i\theta}) \left[\sum_{s=1}^N v^s e^{-is\theta}\right] d\theta \\ &= \int_{-\pi}^{\pi} G^*(\theta)\, f(\theta)e^{ih\theta}\mathcal{A}^*(e^{i\theta})\, H(\theta)\,d\theta\end{aligned}$$

By Cauchy-Schwarz inequality,

$$\left| \int_{-\pi}^{\pi} G^*(\theta) I(\theta)\, d\theta \right| \leq \left(\int_{-\pi}^{\pi} G^*(\theta) G(\theta)\, d\theta \right)^{1/2} \left(\int_{-\pi}^{\pi} I^*(\theta) I(\theta)\, d\theta \right)^{1/2}$$

This leads to the following upper bound on the quadratic form.

$$\begin{aligned} |u' \Delta\, v| &\leq \left(\int_{-\pi}^{\pi} G^*(\theta) G(\theta)\, d\theta \right)^{1/2} \left(\int_{-\pi}^{\pi} H^*(\theta) \mathcal{A}(e^{i\theta}) f^2(\theta) \mathcal{A}^*(e^{i\theta}) H(\theta)\, d\theta \right)^{1/2} \\ &\leq 2\pi \max_{\theta \in [-\pi, \pi]} \Lambda_{\max}^{1/2} \left(\mathcal{A}(e^{i\theta}) f^2(\theta) \mathcal{A}^*(e^{i\theta}) \right) \\ &\leq \frac{\Lambda_{\max}(\Sigma_\epsilon) \mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})} \end{aligned}$$

where the last inequality follows from the expression of $f(\theta)$ in (4.23). □

4.8.3 Results on Covariance Estimation

Proof of Proposition IV.11. The sample covariance matrix $\hat{\Gamma}(0)$ can be expressed as $\hat{\Gamma}(0) = S - \bar{X}\bar{X}'$ where $S = \mathcal{X}'\mathcal{X}/n$ and $\bar{X} = \mathcal{X}'\mathbf{1}/n$, $\mathbf{1}_{n\times 1} = (1, 1, \ldots, 1)'$. First, we derive element-wise concentration bound for $\hat{\Gamma}(0)$ around $\Gamma(0)$. To this end, note that for any $i, j \in \{1, \ldots, p\}$,

$$\left| \hat{\Gamma}_{ij}(0) - \Gamma_{ij}(0) \right| \leq |S_{ij} - \Gamma_{ij}(0)| + \left| \bar{X}_i \bar{X}_j \right| \tag{4.46}$$

Taking maximum over all i, j, we have

$$\max_{1 \leq i,j \leq p} \left| \hat{\Gamma}_{ij}(0) - \Gamma_{ij}(0) \right| \leq \max_{1 \leq i,j \leq p} |S_{ij} - \Gamma_{ij}(0)| + \max_{1 \leq i \leq p} \left| \bar{X}_i \right|^2$$

Equation (4.19) provides a concentration bound on the first term. To concentrate the second term, note that $\bar{X}_i = \mathbf{1}'\mathcal{X}e_i/n$. Set $Y = \mathcal{X}e_i$. Then $Y_{n\times 1}$ can be viewed as the data matrix consisting of n observations from the i^{th} subprocess of $\{X^t\}$.

Thus, $Y \sim N(0, Q)$ with $\|Q\| \leq 2\pi\mathcal{M}(f_X, 1)$, using Proposition IV.3. Now, for $Z = \mathbf{1}'Y/\sqrt{n}$, we have $\text{Var}(Z) = u'Qu \leq 2\pi\mathcal{M}(f_X, 1)$, since $u = \mathbf{1}/\sqrt{n}$ is an unit norm vector. Using this upper bound on $\text{Var}(Z)$ together with the standard Gaussian tail bound, we have, for any $\eta \geq 0$,

$$\begin{aligned}\mathbb{P}\left(|\bar{X}_i|^2 > 4\pi\mathcal{M}(f_X, 1)\eta\right) &\leq \mathbb{P}\left(|Z| > \sqrt{4\pi\mathcal{M}(f_X, 1)\eta}\sqrt{n}\right) \\ &\leq 2\exp\left[-\frac{4\pi\mathcal{M}(f_X, 1)\eta n}{2\text{Var}(Z)}\right] \leq 2\exp\left[-n\min\{\eta, \eta^2\}\right]\end{aligned}$$

Combining the concentration bounds for the two terms and setting $\eta = \sqrt{\frac{\log p}{n}} = o_P(1)$, we conclude

$$\max_{i,j}\left|\hat{\Gamma}_{ij}(0) - \Gamma_{ij}(0)\right| = O_P\left(\mathcal{M}(f_X, 1)\sqrt{\frac{\log p}{n}}\right) \tag{4.47}$$

This provides element-wise concentration bounds similar to equation (12) in *Bickel and Levina* (2008). Rest of the proof follows exactly along the lines of Theorems 1 and 2 in that paper. □

4.8.4 Measure of Stability

In this section we discuss some properties of the stability measure introduced in Section 4.2 and its connection with the assumption $\|A_1\| < 1$. In particular, we show that the assumption $\|A_1\| < 1$ guarantees stability of the process, but not the other way. If, however, the transition matrix A_1 is symmetric, the assumption $\|A_1\| < 1$ is necessary for stability. We also show that this assumption is violated for all stable VAR(d) models, whenever $d > 1$. We conclude the section with the proof of Proposition IV.2, where we derive upper and lower bounds on the quantities $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$.

Lemma 4.8.3. *A VAR(1) process is stable if $\|A_1\| < 1$. If A_1 is symmetric, then a*

VAR(1) process is stable only if $\|A_1\| < 1$.

Proof. If $\|A_1\| < 1$, then all the eigenvalues of A_1 lie inside the unit circle $\{z \in \mathbb{C} : |z| \leq 1\}$. So the process is stable.

If the process is stable, then all the eigenvalues of A_1 lie inside the unit circle. In addition, if A_1 is symmetric, then this implies that $\|A_1\| = \sqrt{\Lambda_{\max}(A_1' A_1)} < 1$. □

Lemma 4.8.4. *Consider the VAR(1) representation of a VAR(d) process in* (4.25). $\|\tilde{A}_1\| \not< 1$ *whenever* $d > 1$.

Proof.

$$\tilde{A}_1 \tilde{A}_1' = \begin{bmatrix} \sum_{t=1}^{d} A_t A_t' & A_1 & \dots & A_{d-1} \\ A_1' & I_p & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ A_{d-1}' & \mathbf{0} & \dots & I_p \end{bmatrix}_{dp \times dp}$$

So for any $v \in \mathbb{R}^{dp}$ with $v' = (v_1', \dots, v_d')$, each $v_t \in \mathbb{R}^p$, we have

$$v' \tilde{A}_1 \tilde{A}_1' v = v_1' \left(\sum_{t=1}^{d} A_t A_t' \right) v_1 + 2 v_1' \sum_{t=2}^{d} A_{t-1} v_t + \sum_{t=2}^{d} v_t^2$$

This implies

$$\Lambda_{\max}\left(\tilde{A}_1 \tilde{A}_1'\right) = \max_{\|v\|=1} v' \tilde{A}_1 \tilde{A}_1' v \geq \max_{\substack{\|v\| = 1 \\ v_1 = \mathbf{0}}} v' \tilde{A}_1 \tilde{A}_1' v = \max_{\substack{\|v\| = 1 \\ v_1 = \mathbf{0}}} \sum_{t=2}^{d} v_t^2 = 1$$

□

Proof of Proposition IV.2. $\mathcal{A}(z) = I_p - A_1 z - A_2 z^2 - \ldots - A_d z^d$

(i) Using $|z| = 1$ together with the matrix norm inequality $\|A\| \leq \|A\|_1 \|A\|_\infty$ (cf.

Cor. 2.3.2, *Golub and Van Loan* (1996)), we have

$$\begin{aligned}
\mu_{\max}(\mathcal{A}) &= \max_{|z|=1} \left\| I - A_1 z - \ldots - A_d z^d \right\| \\
&\leq 1 + \sum_{h=1}^{d} \|A_h\| \leq 1 + \sum_{h=1}^{d} \sqrt{\|A_h\|_1 \, \|A_h\|_\infty} \\
&\leq 1 + \sum_{h=1}^{d} \left(\max_{1 \leq i \leq p} \sum_{j=1}^{p} |A_{h,ij}| + \max_{1 \leq j \leq p} \sum_{i=1}^{p} |A_{h,ij}| \right) /2
\end{aligned}$$

(ii) For $d = 1$, $\mathcal{A}(z) = I_p - A_1 z$. First note that

$$\mu_{\min}(\mathcal{A}) = \min_{|z|=1} \Lambda_{\min} \left((I - A_1 z)^* (I - A_1 z) \right) = \min_{|z|=1} \Lambda_{\min} \left((zI - A_1)^* (zI - A_1) \right)$$

If A_1 is diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_p$ and corresponding eigenvectors $w_1, \ldots, w_p$, we have the decomposition $A_1 = PDP^{-1}$, where D is a diagonal matrix with entries λ_i and $P = [w_1 : \ldots : w_p]$. So, $zI - A_1 = PD_zP^{-1}$, where D_z is diagonal with entries $(z - \lambda_i)$, $i = 1, \ldots, p$. The condition $det(\mathcal{A}(z)) \neq 0$ ensures all the eigenvalues of A_1 are inside the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. This implies D_z is invertible, for all $|z| = 1$ and the eigenvalues of $D_z^* D_z$ are $|z - \lambda_i|^2 \geq (1 - \rho(A_1))^2$, for all $|z| = 1$ and $i = 1, \ldots, p$. Hence,

$$\begin{aligned}
\mu_{\min}(\mathcal{A}) &= \min_{|z|=1} \left[\left\| P D_z^{-1} P^{-1} (P')^{-1} (D_z^*)^{-1} P' \right\| \right]^{-1} \\
&\geq \|P\|^{-2} \|P^{-1}\|^{-2} \left(1 - \rho(A_1)\right)^{-2}
\end{aligned}$$

□

4.8.5 Auxiliary Lemmas

Lemma 4.8.5 (Hansen-Wright Inequality)**.** *If $Y \sim N(0_{n\times 1}, Q_{n\times n})$, then there exists an universal constant $c > 0$ such that for any $\eta \geq 0$,*

$$\mathbb{P}\left[\frac{1}{n}\left|\|Y\|^2 - tr(Q)\right| > \eta\|Q\|\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\}\right]$$

Proof. The proof follows from Theorem 1.1 in *Rudelson and Vershynin* (2013). Write $Y = Q^{1/2}X$, where $X \sim N(0, I)$ and $(Q^{1/2})'(Q^{1/2}) = Q$. Note that each component X_i of X is independent $N(0, 1)$, so that $\|X_i\|_{\psi_2} \leq 1$. Then, by the above theorem,

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\left|\|Y\|^2 - tr(Q)\right| > \eta\|Q\|\right] &= \mathbb{P}\left[\frac{1}{n}\left|X'QX - \mathbb{E}[X'QX]\right| > \eta\|Q\|\right] \\
&\leq 2\exp\left[-c\min\left\{\frac{n^2\eta^2\|Q\|^2}{\|Q\|_F^2}, \frac{n\eta\|Q\|}{\|Q\|}\right\}\right] \\
&\leq 2\exp\left[-cn\min\{\eta^2, \eta\}\right] \text{ since } \|Q\|_F^2 \leq n\|Q\|^2
\end{aligned}$$

□

Lemma 4.8.6 (Approximating cone sets by sparse sets)**.** *For any $S \subset \{1, \ldots, p\}$ with $|S| = s$ and $\kappa > 0$,*

$$\mathcal{C}(S, \kappa) \cap \mathbb{B}_2(1) \subseteq \mathbb{B}_1((\kappa + 1)\sqrt{s}) \cap \mathbb{B}_2(1) \subseteq (\kappa + 2)cl\{conv\{\mathcal{K}(s)\}\}$$

Proof. The first inequality follows from the fact that for any $v \in \mathcal{C}(S, \kappa)$,

$$\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1 \leq (\kappa + 1)\|v_S\|_1 \leq (\kappa + 1)\sqrt{s}\|v_S\| \leq (\kappa + 1)\sqrt{s}$$

Both $A := \mathbb{B}_1((\kappa + 1)\sqrt{s}) \cap \mathbb{B}_2(1)$ and $B := (\kappa + 2)cl\{conv\{\mathcal{K}(s)\}\}$ are closed convex sets. We will show that the support function of A is dominated by the support function of B.

The support function of A is $\phi_A(z) = \sup_{\theta \in A} \langle \theta, z \rangle$. For a given $z \in \mathbb{R}^p$, let J denote the set of coordinates of z with the s largest absolute values, so that $||z_{J^c}||_\infty \leq ||z_J||_1/s \leq ||z_J||/\sqrt{s}$. Also note that for any $\theta \in A$, $\|\theta_{J^c}\|_1 \leq (\kappa+1)\sqrt{s}$. Then we have, for any $\theta \in A$, $z \in \mathbb{R}^p$,

$$\langle \theta, z \rangle = \sum_{i \in J^c} \theta_i z_i + \sum_{i \in J} \theta_i z_i \leq ||z_{J^c}||_\infty \|\theta_{J^c}\|_1 + \|z_J\| \|\theta_J\| \leq (\kappa+1)\|z_J\| + \|z_J\|$$

so that $\phi_A(z) \leq (\kappa+2)\|z_J\|$.

On the other hand, $\phi_B(z) := \sup_{\theta \in B} \langle \theta, z \rangle = \sup_{|U|=s} \sum_{i \in U} \theta_i z_i = (\kappa+2)\|z_J\|$. □

Lemma 4.8.7. *Consider a symmetric matrix $D_{p \times p}$. If, for any vector $v \in \mathbb{R}^p$ with $\|v\| \leq 1$, and any $\eta \geq 0$,*

$$\mathbb{P}\left[|v'Dv| > C\eta\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\}\right]$$

then, for any integer $s \geq 1$, we have

$$\mathbb{P}\left[\sup_{v \in \mathcal{K}(s)} |v'Dv| > C\eta\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\} + s\min\{\log p,\, \log(21ep/s)\}\right]$$

Proof. Choose $U \subset \{1, \ldots, p\}$ with $|U| = s$. Define $S_U = \{v \in \mathbb{R}^p : \|v\| \leq 1,\ supp(v) \subseteq U\}$. Then $\mathcal{K}(s) = \cup_{|U| \leq s} S_U$. Choose $\mathcal{A} = \{u_1, \ldots, u_m\}$, a 1/10-net of S_U. By Lemma 3.5 of *Vershynin* (2009), $|\mathcal{A}| \leq 21^s$. For every $v \in S_U$, there exists some $u_i \in \mathcal{A}$ such that $\|\Delta v\| \leq 1/10$, where $\Delta v = v - u_i$. Then we have,

$$\gamma := \sup_{v \in S_U} |v'Dv| \leq \max_i |u_i'Du_i| + 2\sup_{v \in S_U} \left|\max_i u_i'D(\Delta v)\right| + \sup_{v \in S_U} |(\Delta v)'D(\Delta v)|$$

Since $10(\Delta v) \in S_U$, the third term is bounded above by $\gamma/100$. The second term is

bounded above by $6\gamma/10$, as shown below:

$$\begin{aligned} 2\sup_{v\in S_U}\left|\max_i u_i' D(\Delta v)\right| &\leq \frac{1}{10}\sup_{v\in S_U}|(u_i+10\Delta v)'D(u_i+10\Delta v)| \\ &\quad +\frac{1}{10}\sup_{v\in S_U}|u_i'Du_i| + \frac{1}{10}\sup_{v\in S_U}|(10\Delta v)'D(10\Delta v)| \\ &\leq \frac{4\gamma}{10}+\frac{\gamma}{10}+\frac{\gamma}{10} \end{aligned}$$

Readjusting, we have $\gamma \leq 3\max_i |u_i'Du_i|$. Taking an union bound over all $u_i \in \mathcal{A}$, we have

$$\mathbb{P}\left[\sup_{v\in S_U}|v'Dv| > 3C\eta\right] \leq 2\exp\left[-cn\min\{\eta,\eta^2\}+s\log(21)\right]$$

Taking another union bound over $\begin{pmatrix} p \\ s \end{pmatrix} \leq \min\{p^s, (ep/s)^s\}$ choices of U, we obtain the required result. □

Lemma 4.8.8.

$$\sup_{v\in cl\{conv\{\mathcal{K}(s)\}\}}|v'Dv| \leq 3\sup_{v\in\mathcal{K}(2s)}|v'Dv|$$

Proof. Let $v \in conv\{\mathcal{K}(s)\}$. Then $v = \sum_{i=1}^k \alpha_i v_i$, for some $k \geq 1$, $v_i \in \mathcal{K}(s)$ and $0 \leq \alpha_i \leq 1$, for all $1 \leq i \leq k$, such that $\sum_i \alpha_i = 1$. Then

$$\begin{aligned} 2|v'Dv| &\leq 2\sum_{i,j=1}^k \alpha_i\alpha_j |v_i'Dv_j| \\ &\leq \sum_{i,j=1}^k \alpha_i\alpha_j\left[|(v_i+v_j)'D(v_i+v_j)| + |v_i'Dv_i| + \left|v_j'Dv_j\right|\right] \\ &\leq 6\sum_{i,j=1}^k \alpha_i\alpha_j \sup_{v\in\mathcal{K}(2s)}|v'Dv| \end{aligned}$$

By continuity of quadratic forms, the result follows. □

CHAPTER V

Low-Rank and Sparse VAR modeling

5.1 Introduction

An important challenge in autoregressive modeling of multivariate time series stems from the fact that failure to include relevant variables in the model can introduce spurious correlations among the individual time series, resulting in incorrect estimation of the edge set of the underlying Granger causal network. This is also a major critique against causal interpretation of Granger-causality. This problem in VAR modeling is well-known in the economics literature. For instance, *Christiano et al.* (1999) argue that a positive response of prices to monetary tightening in the post-war US economy, commonly known as the "price puzzle", is an artefact of not including forward looking variables in the model (*Bańbura et al.*, 2010). The high-dimensional VAR framework with sparity based regularizers like lasso resolves this problem to a certain extent by allowing many variables in the model. However, in many macroeconomic applications it is not possible to observe all the relevant variables driving the market economy. A popular strategy is factor modeling, where the key idea is that there are a few latent factors driving the major co-movements of many time series (*Stock and Watson*, 2005). Indeed, empirical evidence suggests that the co-movement of many macroeconomic time series in the US economy can be explained by a small number of unobserved factors extracted from the data.

Failure to account for the presence of unobseved common factors can negatively impact high-dimensional sparse VAR modeling in two ways. First, the correlation among the time series that is driven by underlying factors introduces spurious connectivities among the observed time series. Second, even if the true Granger causal network is sparse, failure to account for hidden factors can result in a non-sparse VAR representation of the process and lasso estimates become inaccurate (cf Examples 1 and 2 below).

In this chapter, we propose to deal with this issue with a low-rank and sparse modeling strategy. Low-rank approximation and low-rank+sparse decomposition of Hankel matrices, which represent the input-output structure of a linear time invariant system (LTI), have appeared in the literature (*Fazel et al.*, 2003). A low-rank representation of the Hankel matrix corresponds to a system of small order or dimension and a sparse Hankel matrix represents sparse input-output system (*Chandrasekaran et al.*, 2011). In the context of high-dimensional stationary time series, we show that a low-rank or a sparse+low-rank structure in the transition matrix arises naturally, if the components of the observed process are affected by some latent factors. We demonstrate this using two examples.

Example 1. We consider a p-dimensional stationary process $\{X^t\}$ with the entire dynamics driven by a r-dimensional ($r \ll p$) unobserved process of factors $\{F^t\}$, which itself follows a $VAR(1)$ process

$$X^t = \Lambda F^t + \xi^t,\ \xi^t \sim N(0, \Sigma_\xi),\ \mathrm{Cov}(\xi^t, \xi^s) = 0 \text{ if } t \neq s \tag{5.1}$$

$$F^t = HF^{t-1} + \eta^t,\ \eta^t \sim N(0.\Sigma_\eta),\ \mathrm{Cov}(\eta^t, \eta^s) = 0, \text{ if } t \neq s \tag{5.2}$$

This is a simple example of a static factor model used in economics. We assume the matrix of factor loadings $\Lambda_{p\times r}$ has full column rank r, so that it has a left inverse Λ^-

satisfying $\Lambda^{-}\Lambda = I_r$. It then readily follows that

$$\begin{aligned} X^t &= \Lambda\left[HF^{t-1} + \eta^t\right] + \xi^t \\ &= \Lambda\left[H\Lambda^{-}(X^{t-1} - \xi^{t-1}) + \eta^t\right] + \xi^t \\ &= \Lambda H\Lambda^{-}\, X^{t-1} + \left[\Lambda\eta^t + \xi^t - \Lambda H\Lambda^{-}\xi^{t-1}\right] \\ &= LX^{t-1} + \epsilon^t \end{aligned}$$

where $L = \Lambda H\Lambda^{-}$ has rank at most r and the new error process $\epsilon^t = \Lambda\eta^t + \xi^t - L\xi^{t-1}$ has a MA(1) component.

Example 2. Consider the same process $\{X^t\}$, but assume that its dynamics is governed by two sources: an underlying process of factors $\{F^t\}$ as before and the interaction among its components, as captured by a VAR(1) process with a sparse transition matrix S

$$X^t = \Lambda F^t + SX^{t-1} + \xi^t,\ S \text{ sparse} \tag{5.3}$$

$$F^t = HF^{t-1} + \eta^t \tag{5.4}$$

A similar calculation shows

$$\begin{aligned} X^t &= \Lambda H\Lambda^{-}\left[X^{t-1} - SX^{t-2} - \xi^{t-1}\right] + SX^{t-1} + \xi^t + \Lambda\eta^t \\ &= (L+S)X^{t-1} - LSX^{t-2} + \epsilon^t \\ &\approx (L+S)X^{t-1} + \epsilon^t, \text{ assuming the second order effects in } LS \text{ are small} \end{aligned}$$

Model. Motivated by the above connections, we propose to model the process $\{X^t\}$ as a stable VAR(1) process with the transition matrix having a low-rank and a

sparse component. Formally, we consider the class of models

$$X^t = AX^{t-1} + \epsilon^t, \epsilon^t \text{ i.i.d. } N(0, \Sigma_\epsilon) \tag{5.5}$$

$$A = L^0 + S^0,\ rank(L^0) = r,\ \|S^0\|_0 = s,\ r \ll p,\ s \ll p^2 \tag{5.6}$$

In this model, the matrix L^0 captures the effects of latent variables and S^0 encodes the dynamics among the individual time series, after accounting for the latent effects. The goal is to estimate S^0 and L^0 with high accuaracy using moderate sample sizes. In this chapter we restrict our analysis to only models with serially uncorrelated errors. A general model with serially correlated error structure, although more well-suited for the examples described above, poses significant technical challenges due to endogeneity (correlation between predictors and the noise in the regression) and we intend to pursue it as a separate problem.

Stability. As shown in Section 4.2, the VAR(1) models considered in (5.5), under the assumption of stability, has a spectral density satisfying assumption (IV.1). Proposition IV.2 provides a lower bound on $\mu_{\min}(\mathcal{A})$. Further, for the special structure of the models considered here, one can get an improved upper bound on $\mu_{\max}(\mathcal{A})$, as shown in the following lemma:

Lemma 5.1.1. *For a stable VAR(1) model of the class* (5.5)*, we have*

$$\mu_{\max}(\mathcal{A}) \leq [1 + l + (v_{in} + v_{out})/2]^2 \tag{5.7}$$

where l is the largest singular value of L^0, $v_{in} = \max_{1\leq j\leq p} |S^0_{ij}|$ and $v_{out} = \max_{1\leq i\leq p} |S^0_{ij}|$.

Proof. $\|\mathcal{A}(z)\| = \|I - (L^0 + S^0)z\| \leq \|I\| + \|L^0\| + \|S^0\|$ for any $z \in \mathbb{C}$ with $|z| = 1$. The result follows from the fact that $\mu_{\max}(\mathcal{A}) = \max_{|z|=1} \|\mathcal{A}(z)\|^2$. □

Notations. We reserve the symbol $\|.\|$ to denote the ℓ_2-norm of a vector and the spectral norm of a matrix. The symbol $\|A\|_*$ is used to denote the nuclear norm,

i.e., sum of the singular values, of a matrix. A^* denotes the conjugate transpose of a matrix A. For any matrix A, we use the notations $\|A\|_0$ to denote $card(vec(A))$, $\|A\|_1$ to denote $\|vec(A)\|_1$ and $\|A\|_{\max}$ to denote $\|vec(A)\|_\infty$. Throughout the chapter, $\Lambda_{\max}(.)$, $\Lambda_{\min}(.)$ are used to denote the maximum and minimum eigenvalues of a symmetric or Hermitian matrix. For any integer $p \geq 1$, we use $\mathbb{S}^{p-1}$ to denote the unit ball $\{v \in \mathbb{R}^p : \|v\| = 1\}$. We use $\{e_1, e_2, \ldots\}$ generically to denote unit vectors in $\mathbb{R}^p$, when p is clear from the context. Throughout the chapter, we write $A \gtrsim B$ if there exists an absolute constant c, independent of the model parameters, such that $A \geq cB$.

5.2 Related Work

Factor models have a long history in the statistics and econometrics literature as a popular technique for dimension reduction. *Bai and Ng* (2008) provide a comprehensive review of the theoretical and empirical work on factor models.

The problem that we consider in this chapter, however, is considerably different in nature. We are interested in learning both the effect of the latent variables on the system and the Granger causal estimates or interactions among the system components, *after* accounting for the effects of latent factors. *Bernanke et al.* (2005b) considered a similar problem using factor augmented vector autoregressive (FAVAR) models. The authors proposed to model the joint process $[(F^t)', (X^t)']'$ as a vector autoregression, with the restriction that there is no effect from $\{X^t\}$ to $\{F^t\}$. Since the process of factors is unobserved, the modeling strategy amounts to iteratively estimating the factors and using them in the VAR model. The method relies on consistent estimation of the number of factors and testing the restrictions imposed by the factor structure. Our approach of modeling the transition matrix as a combination of sparse and low-rank component does not require estimating the number of factors or the factor process separately and provides a framework for jointly estimating

the common effects of the market and interactions among the system components. Further, the theory presented in the subsequent section can be easily generalized for appoximately sparse and low-rank matrices, capturing a broader model class.

Low rank approximation of a given matrix is a popular technique of dimension reduction in many areas of science and engineering (*Fazel*, 2002), including matrix completion problems, principal component analysis and factor analysis. In recent years, decomposing a given matrix into sparse and low-rank component has gained considerable interest, with applications in video surveillance (*Candès et al.*, 2011), neuroimaging and recommender systems. Finding the best low-rank plus sparse representation of an observed matrix via rank constrained optimization is computationally expensive due to the nonconvex nature of the problem. A tractable alternative commonly used in practice is the convex relaxation

$$\min_{(L,S):\ L+S=A} \|L\|_* + \gamma\|S\|_1,\ \gamma > 0 \tag{5.8}$$

where the nuclear/trace norm (sum of singular values of a matrix) serves as a surrogate for the rank constraint and the ℓ_1 norm serves as a surrogate for the sparsity constraint. Several algorithms for solving the above optimization problem have been propsed in the literature, including semidefinite programming (*Chandrasekaran et al.*, 2011) and alternating direction method of multipliers (*Yuan and Yang*, 2009).

In many noisy settings such as ours, the matrix A is not observed and needs to be estimated from data. An example closely related to our problem is the problem of Gaussian graphical model selection in the presence of latent variables from independent samples (*Chandrasekaran et al.*, 2012). Some other applications in factor analysis and multi-task regression has been covered in *Agarwal et al.* (2012). To the best of our knowledge, the properties of these estimators have not been studied in the context of time series and dependent data.

5.3 Estimation Procedure

Based on the data $\{X^0, \ldots, X^T\}$ generated according to the model (5.5), we form the autoregressive design

$$\underbrace{\begin{bmatrix} (X^T)' \\ \vdots \\ (X^1)' \end{bmatrix}}_{\mathcal{Y}} = \underbrace{\begin{bmatrix} (X^{T-1})' \\ \vdots \\ (X^0) \end{bmatrix}}_{\mathcal{X}} A' + \underbrace{\begin{bmatrix} (\epsilon^T)' \\ \vdots \\ (\epsilon^1)' \end{bmatrix}}_{E} \tag{5.9}$$

This is a linear regression problem with $N = T$ samples and $q = p^2$ variables. The goal is to estimate L^0 and S^0 with high accuracy when $N \ll p^2$.

There is an inherent identifiability issue in the estimation of (5.5). Suppose the low-rank component L^0 itself is s-sparse and the sparse component S^0 is of rank r. In that scenario, we cannot hope for any method to estimate L^0 and S^0 separately without imposing any further constraints. So, a minimal condition for low-rank and sparse recovery is that the low rank part should not be too sparse and the sparse part should not be low-rank.

This issue has been addressed in the literature by several authors (*Chandrasekaran et al.*, 2011; *Candès et al.*, 2011). By and large, all the authors propose to ensure the above identifiability under some form of incoherence type condition. These conditions serve as sufficient conditions for *exact* recovery of the low rank and the sparse component by solving the convex program (5.8). In a recent paper, *Agarwal et al.* (2012) showed that in a noisy setting where exact recovery of the two components is impossible, it is still possible to achieve good approximation under comparatively mild assumption. In particular, they formulated a general measure of the *radius of nonidentifiability* of the problem and established a non-asymptotic upper bound on the approximation error

$$\|\hat{L} - L^0\|_F^2 + \|\hat{S} - S^0\|_F^2 \tag{5.10}$$

which depend on this radius. The key idea is to allow for sparse and low-rank matrices in the model, but controlling for the error introduced. We refer the readers to the above paper for a more detailed discussion on this notion of non-identifiability. The low-rank and sparse decomposition problem under restrictions on the radius of nonidentifiability takes the form

$$(\hat{L}', \hat{S}') = \underset{L,S \in \mathbb{R}^{p \times p} : \|S\|_{\max} \leq \alpha/p}{\operatorname{argmin}} \frac{1}{2} \|\mathcal{Y} - \mathcal{X}(L+S)\|_F^2 + \lambda_N \|L\|_* + \mu_N \|S\|_1 \tag{5.11}$$

where λ_N, μ_N are non-negative tuning parameters controlling the regularization of sparse and low-rank part. The parameter α controls for degree of non-identifiable matrices allowed in the model class.

5.4 Theoretical Properties

In this section, we derive a non-asymptotic upper bound on the estimation error of the low-rank and sparse components of the transition matrix. The main result shows that consistent estimation is possible with a sample size of the order $N \sim p\,\mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)$, as long as the process $\{X^t\}$ is stable, stationary and the radius of nonidentifiability, as measured by $\|S^0\|_{\max}$ is small in an appropriate sense.

We build upon the results of *Agarwal et al.* (2012) for fixed $\mathcal{X}$ and E. In particular, it follows from Corollary 1 of the above paper that for a single realization of $\{X^0, \ldots, X^T\}$, for any $\alpha \geq \|S^0\|_{\max}$, if $\gamma_N := \Lambda_{\min}(\mathcal{X}'\mathcal{X}) > 0$, then any solution $(\hat{L}, \hat{S})$ of the convex program (5.11) with

$$\lambda_N \geq 4\|\mathcal{X}'E\|, \;\; \mu_N \geq 4\|\mathcal{X}'E\|_{\max} + \frac{4\gamma_N \alpha}{p} \tag{5.12}$$

satisfies, for some universal positive constants $c_i > 0$,

$$\|\hat{L} - L^0\|_F^2 + \|\hat{S} - S^0\|_F^2 \leq c_1 \frac{\lambda_N^2}{\gamma_N^2} r + c_2 \frac{\mu_N^2}{\gamma_N^2} s \tag{5.13}$$

In order to obtain meaningful results in the context of our problem, we need upper bounds on $\|\mathcal{X}'E\|$ and $\|\mathcal{X}'E\|_{\max}$ and a lower bound on $\Lambda_{\min}(\mathcal{X}'\mathcal{X})$ that hold with high probability. In the context of time series where all the entries of the matrix $\mathcal{X}$ are dependent on each other, it is a non-trivial task to establish such deviation bounds. The main technical contribution of this chapter is to derive these deviation bounds, which lead to meaningful analysis in the context of VAR. The results rely on the measure of stability defined in Chapter IV and an analysis of the joint spectrum of $\{X^{t-1}\}$ and $\{\epsilon^t\}$.

Proposition V.1. *Consider a random realization of* $\{X^0, \ldots, X^T\}$ *generated according to a stable VAR(1) process* (5.5) *and form the autoregressive design* (5.9)*. Define*

$$\phi(A, \Sigma_\epsilon) = \Lambda_{\max}(\Sigma_\epsilon) \left[1 + \frac{1 + \mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right]$$

Then there exist universal positive constants $c_i > 0$ *such that*

1. *for* $N \gtrsim p$,

$$\mathbb{P}\left[\|\mathcal{X}'E/N\| > c_0 \phi(A, \Sigma_\epsilon)\sqrt{p/N}\right] \leq c_1 \exp\left[-c_2 \log p\right]$$

and for any $N \gtrsim \log p$,

$$\mathbb{P}\left[\|\mathcal{X}'E/N\|_{\max} > c_0 \phi(A, \Sigma_\epsilon)\sqrt{\log p/N}\right] \leq c_1 \exp\left[-c_2 \log p\right]$$

2. for $N \gtrsim p\mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)$,

$$\mathbb{P}\left[\Lambda_{\min}(\mathcal{X}'\mathcal{X}/N) > \frac{\Lambda_{\min}(\Sigma_\epsilon)}{2\mu_{\max}(\mathcal{A})}\right] \leq c_1 \exp\left[-c_2 \log p\right]$$

Using the above deviation bounds in the non-asymptotic error (5.13), we obtain the final result for approximate recovery of the low-rank and the sparse components using nuclear and ℓ_1 norm relaxation, as shown next.

Proposition V.2. *Consider the setup of Proposition V.1. There exist universal positive constants* $c_i > 0$ *such that for* $N \gtrsim p\mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)$, *for any* S^0 *with* $\|S^0\|_{\max} \leq \alpha$, *any solution* $(\hat{L}, \hat{S})$ *of the program* (5.11) *satisfies, with probability at least* $1 - c_1 \exp[-c_2 \log p]$,

$$\|\hat{S} - S^0\|_F^2 + \|\hat{L} - L^0\|_F^2 \leq \frac{c_0 \phi^2(A, \Sigma_\epsilon)\mu_{\max}^2(\mathcal{A})}{\Lambda_{\min}^2(\Sigma_\epsilon)} \frac{(rp + s\log p)}{N} + \frac{32\Lambda_{\min}^2(\Sigma_\epsilon)}{\mu_{\max}^2(\mathcal{A})} \frac{s\alpha^2}{p^2} \quad (5.14)$$

Remarks. The error bound presented in the above proposition consists of two key terms. The first term is the error of estimation emanating from randomness in the data and limited sample capacity. For a given model, this error goes to zero as the sample size increases. The second term represents the error due to the unidentifiability of the problem. This is more fundamental to the structure of the true low-rank and sparse components, depends only on the model parameters and does not change with sample size.

The error in estimation again consists of two terms - the second term $(rp + s\log p)/N$ consists of the dimensionality parameters and matches the parametric convergence rates for independent observations. The effect of dependence in the data is captured through the first part of the term: $\frac{c_0\phi^2(A,\Sigma_\epsilon)\mu_{\max}^2(\mathcal{A})}{\Lambda_{\min}^2(\Sigma_\epsilon)}$. As we discussed in chapter IV, this term is larger when the spectral density is more spiky, indicating a stronger temporal and cross-sectional dependence in the data.

Estimation Error	$\|\hat{A}_{OLS} - A\|_F / \|A\|_F$	$\|\hat{A}_{lasso} - A\|_F / \|A\|_F$	$\|(\hat{L} + \hat{S}) - A\|_F / \|A\|_F$	$\|[\hat{L} : \hat{S}] - [L : S]\|_F$
p=30, N=50	7.20(1.16)	1.17(0.09)	0.96(0.07)	0.96(0.10)
p=30, N=100	3.66(0.63)	1.04(0.06)	0.89(0.07)	0.90(0.15)
p=30, N=200	2.30(0.26)	0.93(0.05)	0.76(0.07)	0.77(0.10)
p=30, N=300	1.83(0.21)	0.87(0.06)	0.69(0.06)	0.73(0.08)
p=30, N=500	1.39(0.21)	0.79(0.05)	0.62(0.06)	0.62(0.09)
p=50, N=50	-	1.27(0.11)	1.01(0.05)	1.10(0.08)
p=50, N=100	6.52(0.52)	1.12(0.08)	0.96(0.05)	1.05(0.08)
p=50, N=200	3.80(0.38)	1.02(0.06)	0.87(0.06)	0.93(0.09)
p=50, N=300	2.90(0.23)	0.97(0.03)	0.80(0.04)	0.89(0.06)
p=50, N=500	2.14(0.21)	0.90(0.06)	0.73(0.06)	0.80(0.09)
p=100, N=50	-	1.37(0.14)	1.03(0.04)	1.33(0.10)
p=100, N=100	-	1.23(0.13)	1.00(0.02)	1.29(0.09)
p=100, N=200	7.83(0.54)	1.14(0.07)	0.96(0.03)	1.22(0.08)
p=100, N=300	5.48(0.44)	1.08(0.04)	0.92(0.04)	1.20(0.09)
p=100, N=500	3.84(0.30)	1.01(0.03)	0.86(0.04)	1.11(0.06)

Table 5.1: Estimation Error $\|\hat{A} - A\|_F / \|A\|_F$ of OLS, lasso and low-rank+sparse estimates of a VAR(1) model $X^t = AX^{t-1} + \epsilon^t$. The transition matrix $A = L + S$ has a low rank component L of rank 2 and a sparse component S with $2 - 3\%$ non-zero entries.

5.5 Numerical Experiments

In this section we conduct numerical experiments to assess the performance of low rank and sparse modeling in VAR analysis and compare it with the performances of ordinary least squares (OLS) and lasso estimates.

We consider three different VAR(1) models with $p = 30$, 50 and 100 variables. For each of these models, we generate $N = 50, 100, 200, 300$ and 500 observations from a Gaussian VAR(1) process $X^t = AX^{t-1} + \epsilon^t$, where $A = L + S$ can be decomposed into a low-rank matrix L of rank 2 and a sparse matrix S with $2 - 3\%$ non-zero entries. We rescale the entries of A to ensure stability of the process (the spectral radius is set to $\rho(A) = 0.7$) and rescale the error variance so that $SNR = 2$. We compare the estimation and in-sample prediction error of the different estimates using two performance metrics:

1. Estimation Error: $\|\hat{A} - A\|_F / \|A\|_F$
2. In-sample Prediction Error: $\|\hat{\mathcal{Y}} - \mathcal{Y}\|_F^2 / \|\mathcal{Y}\|_F^2$

The tuning parameters for lasso and low-rank plus sparse estimates are chosen according to Proposition IV.8 and Equation (5.12). We report median and IQR of the

Prediction Error	OLS	lasso	low-rank+sparse
p=30, N=50	3.50(0.38)	1.00(0.02)	0.95(0.03)
p=30, N=100	1.53(0.08)	0.99(0.01)	0.96(0.02)
p=30, N=200	1.15(0.04)	0.99(0.01)	0.96(0.02)
p=30, N=300	1.08(0.02)	0.98(0.01)	0.95(0.02)
p=30, N=500	1.03(0.02)	0.98(0.01)	0.95(0.01)
p=50, N=50	-	1.01(0.02)	0.96(0.02)
p=50, N=100	2.51(0.16)	1.00(0.01)	0.97(0.01)
p=50, N=200	1.39(0.04)	1.00(0.00)	0.97(0.01)
p=50, N=300	1.21(0.02)	0.99(0.00)	0.97(0.01)
p=50, N=500	1.10(0.01)	0.99(0.01)	0.97(0.01)
p=100, N=50	-	1.02(0.01)	0.98(0.01)
p=100, N=100	-	1.01(0.01)	0.98(0.01)
p=100, N=200	2.50(0.05)	1.00(0.00)	0.99(0.01)
p=100, N=300	1.66(0.02)	1.00(0.00)	0.98(0.01)
p=100, N=500	1.29(0.01)	1.00(0.00)	0.98(0.00)

Table 5.2: In-sample prediction error $\|\hat{\mathcal{Y}} - \mathcal{Y}\|_F^2/\|\mathcal{Y}\|_F^2$ of OLS, lasso and low-rank+sparse estimates of a VAR(1) model $X^t = AX^{t-1} + \epsilon^t$. The transition matrix $A = L + S$ has a low rank component L of rank 2 and a sparse component S with $2 - 3\%$ non-zero entries.

performance metrics from 50 iterations of the above experiments.

The estimation errors are reported in Table 5.1. In all the three settings, we find that the low-rank plus sparse VAR estimates outperform the estimates using ordinary least-squares and lasso. Among the three estimates, OLS has the worst estimation error with a high IQR, whereas the two regularized estimates produce lower estimation error with low IQR. For $p = 30$ and $N = 50$, the median estimation error of OLS is 7.20 with an IQR of 1.16, whereas lasso has an estimation error of 1.17 with an IQR of 0.09. The sparse plus low rank estimate has the lowest estimation error of 0.96 with an IQR of 0.07. The estimation errors of all three methods decrease with increase in sample sizes. We also report the error in estimating separately the low rank and the sparse components in the last column of Table 5.1.

The in-sample prediction errors of the three estimation methods are reported in Table 5.2. As in the case with estimation error, we see that low-rank plus sparse VAR estimates outperform OLS and lasso estimates in terms of prediction error in nearly all the settings. The prediction error of OLS for $p = 30$ and $N = 50$ is 3.50 with an IQR of 0.38, which indicates that the OLS prediction errors are 3.50 times larger than the errors from fitting a white noise model to the data (i.e., assuming $A = 0$). This

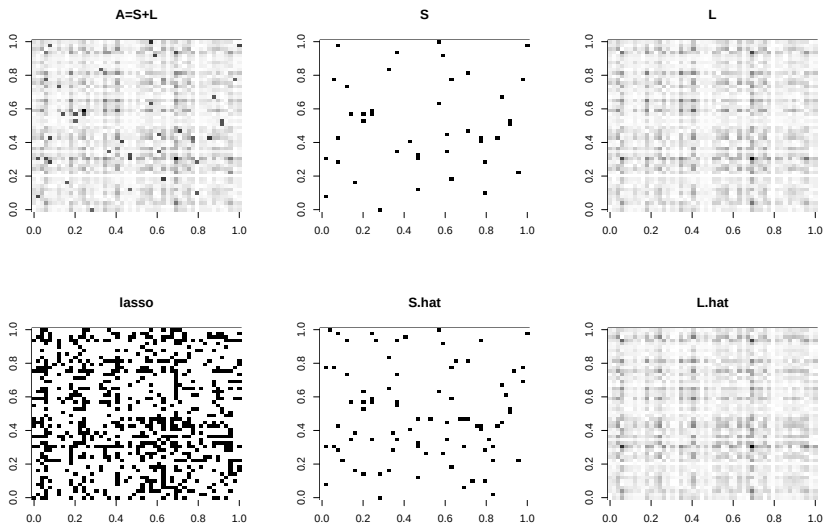


Figure 5.1: Estimated Granger causal networks using lasso and low-rank plus sparse VAR estimates. The top panel displays the true transition matrix A, its low-rank component L and the structure of its sparse component S. The bottom panel displays the structure of the Granger causal networks estimated by lasso ($\hat{A}_{lasso}$), the low-rank plus sparse modeling strategy ($\hat{S}$) and the estimated low-rank component ($\hat{L}$).

effect of overfitting is lower in the lasso regularized estimates, where the prediction error (1.00) is of the same order of the white noise model with an IQR of 0.02. By accounting for a latent low-rank structure of the transition matrix, the low-rank plus sparse estimates produce a lower prediction error of 0.95 with an IQR of 0.03. The results of the other settings are qualitatively similar.

In addition to its improved estimation and prediction performance, the low-rank plus sparse modeling strategy help recover the underlying Granger causal network *after* accounting for the latent structure. In Figure 5.1, we demonstrate this using a VAR(1) model with $p = 50$ and $N = 500$. The top panel displays the true transition matrix A, its low-rank component L and the strucure of its sparse component S. The bottom panel displays the structure of the Granger causal networks estimated by lasso ($\hat{A}_{lasso}$), the low-rank plus sparse modeling strategy ($\hat{S}$) and the estimated low-rank component ($\hat{L}$). As predicted by the theory, we see that the lasso estimate of the Granger causal network, $\hat{A}_{lasso}$, selects many false positives due to its failure to

account for the latent structure. On the other hand, the low-rank plus sparse estimate $\hat{S}$ provides a sparser estimate of the network with significantly less false positives.

It is interesting to note that the estimation performance of the regularized estimates in low-rank plus sparse VAR models is worse than the performance of lasso in sparse VAR models presented in Chapter IV, even for the same sample sizes. This is in line with the error bounds presented in Proposition V.2. The estimation error in low-rank plus sparse models is of the order of $O(rp + s \log p)/N$ while the error of lasso in sparse VAR models scales at a faster rate of $O(s \log p/N)$. This can also be viewed in the factor model examples of Section 5.1. Using the notation of (5.1) and (5.3), a s-sparse VAR requires estimating s parameters in S while the presence of r factors introduces an additional rp parameters in the loading matrix Λ.

5.6 Technical Results

Proof of Proposition V.1. 1. We want to find upper bounds on $\|\mathcal{X}'E/N\|_{\max}$ and $\|\mathcal{X}'E/N\|$ that hold wih high probability. Note that such an upper bound for $\|\mathcal{X}'E/N\|_{\max}$ has already been derived in Proposition IV.10. Here we adopt a different technique that takes a unified approach to provide upper bounds on both quantities. To this end, note that the two norms have the following representations

$$\frac{1}{N}\|\mathcal{X}'E\| = \sup_{u,v\in\mathbb{S}^{p-1}} \frac{1}{N}u'\mathcal{X}'Ev, \ \frac{1}{N}\|\mathcal{X}'E\|_{\max} = \sup_{u,v\in\{e_1,\ldots,e_p\}} \frac{1}{N}u'\mathcal{X}'Ev$$

For any given $u, v \in \mathbb{S}^{p-1}$, we first provide a bound on $u'(\mathcal{X}'E/N)v$. Note that

$$\frac{1}{N}u'\mathcal{X}'Ev = \frac{1}{2N}\left[\|\mathcal{X}u + Ev\|^2 - \|\mathcal{X}u\|^2 - \|Ev\|^2\right]$$

Consider the univariate stochastic processes $\{u'X^{t-1}\}$, $\{v'\epsilon^t\}$ and $\{u'X^{t-1} + v'\epsilon^t\}$. The vectors $\mathcal{X}u$, Ev and $\mathcal{X}u + Ev$ can be viewed as data matrices (see Section 4.2) with N consecutive observations from the above three processes. Also $\mathrm{Var}(u'X^{t-1} + v'\epsilon^t) = \mathrm{Var}(u'X^{t-1}) + \mathrm{Var}(v'\epsilon^t)$ [since $\mathrm{Cov}(X^{t-1}, \epsilon^t) = 0$]. This implies

$$\begin{aligned}\left|\frac{2}{N}u'\mathcal{X}'Ev\right| &\leq \left|\frac{1}{N}\|\mathcal{X}u + Ev\|^2 - \mathrm{Var}(u'X^{t-1} + \epsilon^t)\right| \\ &+ \left|\frac{1}{N}\|\mathcal{X}u\|^2 - \mathrm{Var}(u'X^{t-1})\right| + \left|\frac{1}{N}\|Ev\|^2 - \mathrm{Var}(\epsilon^t)\right|\end{aligned}$$

So it is enough to derive deviation bounds for each of the three terms on the right.

We will use Proposition IV.4 to derive these deviation bounds. For this, we will need the spectral densities of the three processes. By Lemma 5.6.1 and the fact that $f_{u'X}(\theta) = u'f_X(\theta)u$ for any p-dimensional stationary process $\{X^t\}$ satisfying assumption IV.1 and any $u \in \mathbb{R}^p$, we have

$$\begin{aligned}&\mathcal{M}(f_{u'X^{t-1}}) \leq \mathcal{M}(f_X) \leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} \\ &\mathcal{M}(f_{v'\epsilon^t}) \leq \mathcal{M}(f_\epsilon) \leq \Lambda_{\max}(\Sigma_\epsilon) \\ &\mathcal{M}(f_{u'X^{t-1}+v'\epsilon^t}) \leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} + \Lambda_{\max}(\Sigma_\epsilon) + 2\frac{\mu_{\max}(\mathcal{A})\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}\end{aligned}$$

Applying the first inequality of IV.4 on each of the three terms on the right leads to the following deviation bound

$$\mathbb{P}\left[|u'(\mathcal{X}'E/N)v| > 2\pi\eta\phi(A, \Sigma_\epsilon)\right] \leq 6\exp\left[-cN\min\{\eta, \eta^2\}\right] \tag{5.15}$$

for any $u, v \in \mathbb{S}^{p-1}$ and any $\eta > 0$.

To derive the deviation bound on $\|\mathcal{X}'E/N\|_{\max}$, we simply take a union bound over the p^2 possible choices of $u, v \in \{e_1, e_2, \ldots, e_p\}$. This leads to

$$\mathbb{P}\left[\|\mathcal{X}'E/N\|_{\max} > 2\pi\eta\phi(A, \Sigma_\epsilon)\right] \leq 6\exp\left[-cN\min\{\eta, \eta^2\} + 2\log p\right]$$

Since $N \gtrsim p$, we can set $\eta = \sqrt{(2+c_1)\log p/cN}$ so that $\eta < 1$ (i.e., $\eta^2 < \eta$) will be satisfied for large enough N. This implies that

$$\mathbb{P}\left[\|\mathcal{X}'E/N\|_{\max} > c_0\phi(A, \Sigma_\epsilon)\right] \leq c_1\exp\left[-c_2\log p\right]$$

for some universal constants $c_i > 0$.

To derive the deviation bound on the spectral norm, we discretize the unit ball S^{p-1} using an ϵ-net $\mathcal{N}$ of cardinality at most $(1+2/\epsilon)^p$. An argument along the line of Lemma 4.8.7 then shows that for a small enough $\epsilon > 0$,

$$\sup_{u,v\in\mathbb{S}^{p-1}} |u'(\mathcal{X}'E/N)v| \leq K \sup_{u,v\in\mathcal{N}} |u'(\mathcal{X}'E/N)v|$$

for some constant $K > 1$, possibly dependent on ϵ. As before, taking a union bound over the $(1+2/\epsilon)^{2p}$ choices of u and v, we get

$$\mathbb{P}\left[\|\mathcal{X}'E/N\| > 2\pi K\eta\phi(A, \Sigma_\epsilon)\right] \leq 6\exp\left[-cN\min\{\eta, \eta^2\} + 2p\log(1+2/\epsilon)\right]$$

Since $N \gtrsim p$, choosing $\eta = \sqrt{(c_1 + 2\log(1+2/\epsilon))p/cN}$ ensures $\eta < 1$ for large enough N. Setting η as above concludes the proof.

2. We want to obtain a lower bound on the minimum eigenvalue of $\mathcal{X}'\mathcal{X}/N$ that holds with high probability.

   Since $\Lambda_{\min}(\mathcal{X}'\mathcal{X}/N) = \inf_{v\in\mathbb{S}^{p-1}} v'(\mathcal{X}'\mathcal{X}/N)v$, we start with the single deviation

bound of Proposition IV.4

$$\mathbb{P}\left[\left|v'\left(\mathcal{X}'\mathcal{X}/N - \Gamma_X(0)\right)v\right| > 2\pi\eta\mathcal{M}(f_X)\right] \leq 2\exp\left[-cN\min\{\eta,\eta^2\}\right]$$

for any $v \in \mathbb{S}^{p-1}$ and $\eta > 0$.

The next step is to extend this single deviation bound uniformly on the set $\mathbb{S}^{p-1}$. As in the proof of part 1, we construct a ϵ-net of cardinality at most $(1+2/\epsilon)^p$ and approximate the quadratic form using its values on the net. This yields the following deviation bound

$$\mathbb{P}\left[\sup_{v\in\mathbb{S}^{p-1}}\left|v'\left(\frac{\mathcal{X}'\mathcal{X}}{N} - \Gamma_X(0)\right)v\right| > 2K\pi\eta\mathcal{M}(f_X)\right] \leq 2\exp\left[-cN\min\{\eta,\eta^2\} + p\log\left(1+\frac{2}{\epsilon}\right)\right]$$

for some constant $K > 1$. Seting $\eta = \mathfrak{m}(f_X)/4K\pi\mathcal{M}(f_X) < 1$ and noting that $N \gtrsim \mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)p$, we conclude

$$\mathbb{P}\left[\sup_{v\in\mathbb{S}^{p-1}}\left|v'\left(\mathcal{X}'\mathcal{X}/N - \Gamma_X(0)\right)v\right| > \mathfrak{m}(f_X)/2\right] \leq c_0\exp\left[-c_1\log p\right]$$

The result follows from the lower bound on $\mathfrak{m}(f_X)$ presented in (4.24) and the fact that $v'\Gamma_X(0)v \geq \mathfrak{m}(f_X)$ for all $v \in \mathbb{S}^{p-1}$.

□

Lemma 5.6.1. *Consider a stable VAR(1) process $X^t = AX^t + \epsilon^t$ with error process $\{\epsilon^t\}$ satisfying assumption* (IV.1)*. Then*

1. *The spectral density of the joint process $W^t = [(X^{t-1})', (\epsilon^t)']'$ is given by*

$$f_W(\theta) = \begin{bmatrix} f_X(\theta) & e^{2i\theta}f_X(\theta)\mathcal{A}^*(e^{i\theta}) \\ e^{-2i\theta}\mathcal{A}(e^{i\theta})f_X(\theta) & f_\epsilon(\theta) \end{bmatrix}$$

2. *For any $u, v \in \mathbb{S}^{p-1}$, the spectral density of $w^t = u'X^{t-1} + v'\epsilon^t$ satisfies*

$$\mathcal{M}(f_w) \leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} + \Lambda_{\max}(\Sigma_\epsilon) + 2\frac{\mu_{\max}(\mathcal{A})\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} \tag{5.16}$$

Proof. 1. The autocovariance function of the process $\{W^t\}$ is given by

$$\begin{aligned}
\Gamma_W(s) &= \mathrm{Cov}\left(\begin{bmatrix} X^{t-1} \\ \epsilon^t \end{bmatrix}, \begin{bmatrix} X^{t-1+s} \\ \epsilon^{t+s} \end{bmatrix}\right) \\
&= \begin{bmatrix} \Gamma_X(s) & \Gamma_X(s+2) - \Gamma_X(s+1)A' \\ \Gamma_X(s-2) - A\Gamma_X(s-1) & \Gamma_\epsilon(s) \end{bmatrix}
\end{aligned}$$

since $\epsilon^{t+s} = X^{t+s+1} - AX^{t+s}$ and $\epsilon^t = X^{t+1} - AX^t$. Then it is easy to see that the diagonal blocks of the spectral density of $f_W(\theta)$ are precisely $f_X(\theta)$ and $f_\epsilon(\theta)$. The upper off-diagonal block is

$$\begin{aligned}
&\frac{1}{2\pi}\sum_{l=-\infty}^{\infty}\left[\Gamma_X(l+2) - \Gamma_X(l+1)A'\right]e^{-il\theta} \\
&= e^{2i\theta}f_X(\theta) - e^{i\theta}f_X(\theta)A' \\
&= e^{2i\theta}f_X(\theta)\left(I - e^{-i\theta}A'\right) \\
&= e^{2i\theta}f_X(\theta)\mathcal{A}^*(e^{i\theta})
\end{aligned}$$

Since the spectral density matrix is Hermitian, the lower off-diagonal block is the conjugate transpose of the above.

2. Since $w^t = [u'v']W^t$, the spectral density of $\{w^t\}$ is given by

$$\begin{aligned} f_w(\theta) &= \begin{bmatrix} u' & v' \end{bmatrix} f_W(\theta) \begin{bmatrix} u \\ v \end{bmatrix} \\ &= u' f_X(\theta) u + v' f_\epsilon(\theta) v + e^{2i\theta} u' f_X(\theta) \mathcal{A}^*(e^{i\theta}) v + e^{-2i\theta} v' \mathcal{A}(e^{i\theta}) f_X(\theta) u \\ &\leq \mathcal{M}(f_X) + \mathcal{M}(f_\epsilon) + 2\mathcal{M}(f_X)\mu_{\max}(\mathcal{A}) \end{aligned}$$

where the last term comes from applying Cauchy-Schwartz inequality on the cross-product terms. The result follows by substituing the bounds in (4.24).

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Agarwal, A., S. Negahban, and M. J. Wainwright (2012), Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions, *The Annals of Statistics*, *40*(2), 1171–1197, doi:10.1214/12-AOS1000.

Arnold, A., Y. Liu, and N. Abe (2007), Temporal causal modeling with graphical granger methods, in *Proceedings of the 13th ACM SIGKDD*, pp. 66–75.

Bach, F. R. (2008), Consistency of the group lasso and multiple kernel learning, *J. Mach. Learn. Res.*, *9*, 1179–1225.

Bai, J., and S. Ng (2008), *Large dimensional factor analysis*, Now Publishers Inc.

Bańbura, M., D. Giannone, and L. Reichlin (2010), Large bayesian vector auto regressions, *Journal of Applied Econometrics*, *25*(1), 71–92, doi:10.1002/jae.1137.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005a), Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach, *The Quarterly Journal of Economics*, *120*(1), 387–422, doi:10.1162/0033553053327452.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005b), Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach, *The Quarterly Journal of Economics*, *120*(1), 387–422.

Bickel, P., Y. Ritov, and A. Tsybakov (2009), Simultaneous analysis of lasso and dantzig selector, *The Annals of Statistics*, *37*(4), 1705–1732.

Bickel, P. J., and E. Levina (2008), Covariance regularization by thresholding, *Ann. Statist.*, *36*(6), 2577–2604, doi:10.1214/08-AOS600.

Binder, M., C. Hsiao, and M. H. Pesaran (2005), Estimation and inference in short panel vector autoregressions with unit roots and cointegration, *Econometric Theory*, *21*, 795–837, doi:10.1017/S0266466605050413.

Blanchard, O., and R. Perotti (2002), An empirical characterization of the dynamic effects of changes in government spending and taxes on output, *the Quarterly Journal of economics*, *117*(4), 1329–1368.

Breheny, P., and J. Huang (2009), Penalized methods for bi-level variable selection, *Stat. Interface*, *2*(3), 369–380.

Cai, T., and W. Liu (2011), Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association*, *106*(494), 672–684, doi: 10.1198/jasa.2011.tm10560.

Cai, T. T., and H. H. Zhou (2012a), Minimax estimation of large covariance matrices under l1 norm, *Statist. Sinica*, *22*, 1319–1378.

Cai, T. T., and H. H. Zhou (2012b), Optimal rates of convergence for sparse covariance matrix estimation, *The Annals of Statistics*, *40*(5), 2389–2420.

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011), Robust principal component analysis?, *Journal of the ACM (JACM)*, *58*(3), 11.

Cao, B., and Y. Sun (2011), Asymptotic distributions of impulse response functions in short panel vector autoregressions, *Journal of Econometrics*, *163*(2), 127 – 143, doi:10.1016/j.jeconom.2011.03.004.

Chandrasekaran, V., S. Sanghavi, P. A. Parrilo, and A. S. Willsky (2011), Rank-sparsity incoherence for matrix decomposition, *SIAM Journal on Optimization*, *21*(2), 572–596.

Chandrasekaran, V., P. A. Parrilo, A. S. Willsky, et al. (2012), Latent variable graphical model selection via convex optimization, *The Annals of Statistics*, *40*(4), 1935–1967.

Chen, X., M. Xu, and W. B. Wu (2013), Covariance and precision matrix estimation for high-dimensional time series, *Annals of Statistics*, doi:forthcoming.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1999), Monetary policy shocks: What have we learned and to what end?, in *Handbook of Macroeconomics*, *Handbook of Macroeconomics*, vol. 1, edited by J. B. Taylor and M. Woodford, chap. 2, pp. 65–148, Elsevier.

Davis, R. A., P. Zang, and T. Zheng (2012), Sparse Vector Autoregressive Modeling, *ArXiv e-prints*.

De Mol, C., D. Giannone, and L. Reichlin (2008), Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components?, *Journal of Econometrics*, *146*(2), 318–328.

Dobriban, E., and J. Fan (2013), Regularity properties of high-dimensional covariate matrices, *arXiv preprint arXiv:1305.5198*.

Fan, J., J. Lv, and L. Qi (2011), Sparse high-dimensional models in economics, *Annual Review of Economics*, *3*(1), 291–317, doi:10.1146/annurev-economics-061109-080451.

Fazel, M. (2002), Matrix rank minimization with applications, Ph.D. thesis, PhD thesis, Stanford University.

Fazel, M., H. Hindi, and S. P. Boyd (2003), Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices, in *American Control Conference, 2003. Proceedings of the 2003*, vol. 3, pp. 2156–2162, IEEE.

Fournier, J.-D., J. Grimm, J. Leblond, and J. R. Partington (2006), *Harmonic Analysis and Rational Approximation: Their Rôles in Signals, Control and Dynamical Systems*, Springer.

Friedman, N. (2004), Inferring cellular networks using probabilistic graphical models, *Science's STKE*, *303*(5659), 799.

Friston, K. (2009), Causal modelling and brain connectivity in functional magnetic resonance imaging, *PLoS Biol*, *7*(2), e1000,033, doi:10.1371/journal.pbio.1000033.

Fujita, A., J. Sato, H. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. Sogayar, and C. Ferreira (2007a), Modeling gene expression regulatory networks with the sparse vector autoregressive model, *BMC Systems Biology*, *1*(1), 39.

Fujita, A., J. Sato, H. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. Sogayar, and C. Ferreira (2007b), Modeling gene expression regulatory networks with the sparse vector autoregressive model, *BMC Systems Biology*, *1*(1), 39.

Golub, G. H., and C. F. Van Loan (1996), *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, third ed., xxx+698 pp., Johns Hopkins University Press, Baltimore, MD.

Granger, C. (1969a), Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, pp. 424–438.

Granger, C. W. J. (1969b), Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, *37*(3), 424–438.

Hamilton, J. D. (1994), *Time series analysis*, vol. 2, Cambridge Univ Press.

Han, F., and H. Liu (2013), Transition matrix estimation in high dimensional time series, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, *28*(2), 172–180.

Hannan, E. J., and M. Deistler (2012), *The statistical theory of linear systems*, vol. 70, SIAM.

Hiemstra, C., and J. Jones (1994), Testing for linear and nonlinear granger causality in the stock price-volume relation, *Journal of Finance*, pp. 1639–1664.

Huang, J., and T. Zhang (2010), The benefit of group sparsity, *Ann. Statist.*, *38*(4), 1978–2004.

Huang, J., S. Ma, H. Xie, and C.-H. Zhang (2009), A group bridge approach for variable selection, *Biometrika*, *96*(2), 339–355, doi:10.1093/biomet/asp020.

Johnstone, I. M. (2001), On the distribution of the largest eigenvalue in principal components analysis, *The Annals of statistics*, *29*(2), 295–327.

Kim, K., et al. (2005), Nuclear factor of activated t cells c1 induces osteoclast-associated receptor gene expression during tumor necrosis factor-related activation-induced cytokine-mediated osteoclastogenesis, *Journal of Biological Chemistry*, *280*(42), 35,209–35,216.

Kock, A., and L. Callot (2012), Oracle inequalities for high dimensional vector autoregressions.

Kumar, P. R., and P. Varaiya (1986), *Stochastic systems: estimation, identification and adaptive control*, Prentice-Hall, Inc.

Lam, C., and P. Souza (2013), Regularization for spatial panel time series using the adaptive lasso, *Tech. rep.*

Ledoux, M., and M. Talagrand (1991), *Probability in Banach spaces*, *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*, vol. 23, xii+480 pp., Springer-Verlag, Berlin, isoperimetry and processes.

Litterman, R. B. (1986), Forecasting with bayesian vector autoregressionsfive years of experience, *Journal of Business & Economic Statistics*, *4*(1), 25–38.

Loh, P.-L., and M. J. Wainwright (2012), High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity., *Ann. Stat.*, *40*(3), 1637–1664.

Lounici, K., M. Pontil, S. van de Geer, and A. B. Tsybakov (2011), Oracle inequalities and optimal inference under group sparsity, *Ann. Statist.*, *39*(4), 2164–2204.

Lozano, A., N. Abe, Y. Liu, and S. Rosset (2009a), Grouped graphical Granger modeling for gene expression regulatory networks discovery, *Bioinformatics*, *25*(12), i110.

Lozano, A., N. Abe, Y. Liu, and S. Rosset (2009b), Grouped graphical granger modeling for gene expression regulatory networks discovery, *Bioinformatics*, *25*(12), i110.

Lütkepohl, H. (2005), *New introduction to multiple time series analysis*, Springer.

Marčenko, V. A., and L. A. Pastur (1967), Distribution of eigenvalues for some sets of random matrices, *Sbornik: Mathematics*, *1*(4), 457–483.

Meinshausen, N., and B. Yu (2009), Lasso-type recovery of sparse representations for high-dimensional data, *The Annals of Statistics*, *37*(1), 246–270.

Michailidis, G. (2012), Statistical challenges in biological networks, *Journal of Computational and Graphical Statistics*, *21*(4), 840–855, doi: 10.1080/10618600.2012.738614.

Michailidis, G., and F. d'Alché Buc (2013), Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues, *Mathematical Biosciences*, (0), –, doi:http://dx.doi.org/10.1016/j.mbs.2013.10.003.

Mukhopadhyay, N., and S. Chatterjee (2007), Causality and pathway search in microarray time series experiment, *Bioinformatics*, *23*(4), 442.

Murphy, K. (2002), Dynamic Bayesian networks: representation, inference and learning, Ph.D. thesis, University Of California.

Negahban, S., and M. J. Wainwright (2011), Estimation of (near) low-rank matrices with noise and high-dimensional scaling, *Ann. Statist.*, *39*(2), 1069–1097, doi: 10.1214/10-AOS850.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012), A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers, *Statistical Science*, *27*(4), 538–557.

Ong, I., J. Glasner, D. Page, et al. (2002), Modelling regulatory pathways in E. coli from time series expression profiles, *Bioinformatics*, *18*(Suppl 1), S241–S248.

Opgen-Rhein, R., and K. Strimmer (2007), Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process, *BMC bioinformatics*, *8*(Suppl 2), S3.

Parter, S. V. (1961), Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations, *Trans. Amer. Math. Soc.*, *99*, 153–192.

Pearl, J. (2000a), *Causality: Models, Reasoning, and Inference*, Cambridge Univ Press.

Pearl, J. (2000b), *Causality: models, reasoning, and inference*, vol. 47, Cambridge.

Perrin, B., L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche Buc (2003), Gene networks inference using dynamic Bayesian networks, *Bioinformatics*, *19*(suppl 2), 138–148.

Priestley, M. B. (1981), *Spectral analysis and time series. Vol. 2*, i–xviii and 654–890 and Ri–Rxxi and Ii–Ixv pp., Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, multivariate series, prediction and control, Probability and Mathematical Statistics.

Rangel, C., J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. Wild, and F. Falciani (2004), Modeling t-cell activation using gene expression profiling and state-space models, *Bioinformatics*, *20*(9), 1361.

Raskutti, G., M. J. Wainwright, and B. Yu (2010), Restricted eigenvalue properties for correlated Gaussian designs, *J. Mach. Learn. Res.*, *11*, 2241–2259.

Rudelson, M., and R. Vershynin (2013), Hanson-wright inequality and sub-gaussian concentration, *Electron. Commun. Probab.*, *18*.

Rudelson, M., and S. Zhou (2013), Reconstruction from anisotropic random measurements, *Information Theory, IEEE Transactions on*, *59*(6), 3434–3447, doi: 10.1109/TIT.2013.2243201.

Seth, A. K., P. Chorley, and L. C. Barnett (2013), Granger causality analysis of fmri {BOLD} signals is invariant to hemodynamic convolution but not downsampling, *NeuroImage*, *65*(0), 540 – 555, doi: http://dx.doi.org/10.1016/j.neuroimage.2012.09.049.

Shojaie, A., and G. Michailidis (2010a), Discovering graphical Granger causality using the truncating lasso penalty, *Bioinformatics*, *26*(18), i517–i523.

Shojaie, A., and G. Michailidis (2010b), Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs, *Biometrika*, *97*(3), 519–538.

Shojaie, A., and G. Michailidis (2010a), Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs, *Biometrika*, *97*(3), 519–538.

Shojaie, A., and G. Michailidis (2010b), Discovering graphical granger causality using a truncating lasso penalty, *Bioinformatics*, *26*(18), i517–i523.

Sims, C. (1972), Money, income, and causality, *The American Economic Review*, *62*(4), 540–552.

Sims, C. A. (1980), Macroeconomics and reality, *Econometrica*, *48*(1), pp. 1–48.

Smith, S. M. (2012), The future of fmri connectivity, *NeuroImage*, *62*(2), 1257 – 1266, doi:http://dx.doi.org/10.1016/j.neuroimage.2012.01.022.

Song, S., and P. J. Bickel (2011), Large vector auto regressions, *Arxiv preprint arXiv:1106.3915v1*.

Stock, J. H., and M. W. Watson (2005), Implications of dynamic factor models for var analysis, *Working Paper 11467*, National Bureau of Economic Research.

Stock, J. H., and M. W. Watson (2006), Forecasting with many predictors, *Handbook of economic forecasting*, *1*, 515–554.

van de Geer, S., P. Bühlmann, and S. Zhou (2011), The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso), *Electron. J. Stat.*, *5*, 688–749, doi:10.1214/11-EJS624.

van de Geer, S. A., and P. Bühlmann (2009a), On the conditions used to prove oracle results for the Lasso, *Electron. J. Stat.*, *3*, 1360–1392.

van de Geer, S. A., and P. Bühlmann (2009b), On the conditions used to prove oracle results for the Lasso, *Electron. J. Stat.*, *3*, 1360–1392.

Vershynin, R. (2009), *Lectures in Geometric Functional Analysis*, available at http://www-personal.umich.edu/ romanv/papers/GFA-book/GFA-book.pdf.

Wainwright, M. (2009), Sharp thresholds for high-dimensional and noisy sparsity recovery using -constrained quadratic programming (lasso), *Information Theory, IEEE Transactions on*, *55*(5), 2183 –2202, doi:10.1109/TIT.2009.2016018.

Wasserman, L., and K. Roeder (2009), High dimensional variable selection, *Annals of statistics*, *37*(5A), 2178.

Waterman, M., K. Jones, et al. (1990), Purification of tcf-1 alpha, a t-cell-specific transcription factor that activates the t-cell receptor c alpha gene enhancer in a context-dependent manner., *The New biologist*, *2*(7), 621.

Wei, F., and J. Huang (2010), Consistent group selection in high-dimensional linear regression, *Bernoulli*, *16*(4), 1369–1384, doi:10.3150/10-BEJ252.

Wu, W. B. (2005), Nonlinear system theory: Another look at dependence, *Proceedings of the National Academy of Sciences of the United States of America*, *102*(40), 14,150–14,154, doi:10.1073/pnas.0506715102.

Yamaguchi, R., R. Yoshida, S. Imoto, T. Higuchi, and S. Miyano (2007), Finding module-based gene networks with state-space models-Mining high-dimensional and short time-course gene expression data, *IEEE Signal Processing Magazine*, *24*(1), 37–46.

Yuan, X., and J. Yang (2009), Sparse and low-rank matrix decomposition via alternating direction methods, *preprint*.

Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *J. Mach. Learn. Res.*, *7*, 2541–2563.

Zhou, S. (2010), Thresholded lasso for high dimensional variable selection and statistical estimation, *Arxiv preprint arXiv:1002.1583*.