Modeling and Estimating Multi-Block Interactions for High-Dimensional Stationary Time Series

by

Jiahe Lin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Moulinath Banerjee, Co-Chair
Professor George Michailidis, Co-Chair
Professor Amiyatosh Purnanandam
Professor Ji Zhu

Jiahe Lin

jiahelin@umich.edu

ORCID iD: 0000-0001-9523-0981

ACKNOWLEDGEMENTS

I would like to start by expressing my gratitude to my advisor Professor George Michailidis for his guidance and support in the development of this work. His incisive perception and broad perspective on research have profoundly influenced my way of approaching statistical problems; more importantly, he has been a role model who always encourages and reminds me to be dedicated. I am appreciative of him sharing with me so many intellectually stimulating discussions, of his patience in guiding me throughout my once-in-a-lifetime journey as a Ph.D. student, and feel fortunate for having him as my mentor.

I would also like to thank Professor Ji Zhu, whom I have the pleasure to work with during my years at Michigan. He introduced me to a different set of problems which enriched my research experience, and his great ideas and insightful comments have broadened my horizon. Thanks also to my co-advisor Professor Moulinath Banerjee for taking me under his wing over the past four years, as well as his instructions and suggestions on the completion and improvement of this dissertation. I also owe credit to Dr. Sumanta Basu, who guided and gave me an enjoyable collaborative time at the beginning of my Ph.D. research. Moreover, I wish to thank Professor Amiyatosh Purnanandam for his invaluable input as a dissertation committee member.

Finally, I would like to acknowledge my beloved parents for their unwavering support at every step of my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Modeling and estimating interactions amongst multiple groups of variables is an important task for understanding the structure of complex systems. In particular, for time series, the interdependence structure can be either on contemporaneous correlations, or on lead-lag cross-relations. Both structures are of interest in diverse applications in economics, finance, functional genomics, neuroscience and control theory.

This thesis addresses a number of topics related to such interdependence structures, under high-dimensional scaling, namely when the number of time series under consideration becomes larger than the number of available time points. The first part of the thesis considers modeling and estimating interactions between observable blocks of variables, as well as their respective within-block dependence structures, in high-dimensional independent and identically distributed (iid), as well as temporal dependent settings. In the iid case, we model the blocks of variables of interest through a multi-layered Gaussian graphical model, and introduce a penalized maximum likelihood (MLE) procedure that provides both statistical and algorithmic guarantees, leveraging the structure of the log-likelihood function and its bi-convex nature. For the case where the data exhibit temporal dependence, the blocks are modeled through a stable Vector Autoregressive (VAR) system with group Granger-causal ordering. Building upon the work for the iid case, we estimate their lead-lag relationships, as well as the contemporaneous dependence structure using a penalized MLE criterion, under different structural assumptions of the transition matrices — sparse or low rank. We establish theoretical properties for the estimates analogous to the iid case, modulo an additional cost due to the temporal dependence in the data. Moreover, we devise a testing procedure for the presence of such group Granger causality, tailoring it to the posited structural assumptions on the transition matrix that couples the blocks. The devised estimation and testing procedure are assessed via numerical experiments, and further illustrated on a real data example from economics that examines the impact of the stock market on major macroeconomic indicators.

However, large stable VAR systems have the inherent limitation that the transition matrix

needs to be very sparse or has small averaged magnitude to satisfy the stationary constraint. This further raises the issue of whether VAR model is the appropriate modeling framework for ultra large number of time series. To this end, we consider systems of time series that can be summarized by a small set of latent factors. In the second part of this thesis, we focus on estimating the interaction between an observable process and a dynamically evolving latent factor process. Specifically, we extend the popular in applied economics work, factor-augmented vector autoregressive (FAVAR) model to high dimensions and study estimation of the model parameters by formulating an optimization problem that involves a low-rank-plus-sparse type decomposition. Moreover, we investigate model identifiability issues and establish theoretical properties for the proposed estimator. The performance of the proposed method is evaluated through synthetic data, and the model is further illustrated on an economic data set that examines interlinkages between commodity prices and macroeconomic variables. Along a slightly different line of inquiry where the contemporaneous dependence is of prime interest rather than lead-lag relationships, we extend the approximate factor model where correlations amongst the idiosyncratic (error) component are assumed to be weak, to the case where moderate-to-strong correlations are allowed. Using a formulation similar to the FAVAR problem, we propose an algorithm to estimate the model parameters and investigate its statistical and algorithmic properties. The model and the quality of the resulting estimates are illustrated on log-returns of stock prices of large financial institutions from the banking, brokers & dealers and insurance sectors.

CHAPTER I

Introduction

Technological advances enable the collection of large number of time series that in turn creates a need for novel modeling, inference and forecasting methods. Of particular interest is how such a large number of time series interact with each other, either contemporaneously or across time or both. There are two popular paradigms in modeling a large panel of stationary time series: (i) vector autoregressions (VAR) that examine the lead-lag relationships of the time series under consideration, by modeling the linear dependence of current values of each time series on past values of itself as well as other time series; and (ii) dynamic factor models (DFM), which primarily investigate the contemporaneous correlation structure amongst the time series of interest.

This thesis focuses on developing modeling frameworks for systems involving multiple interacting blocks of time series. In addition, computationally efficient algorithms are introduced to obtain estimates of the model parameters, and their statistical properties established under high-dimensional scaling. Within the VAR framework, we specifically consider a special instance of VAR models, where blocks of time series are naturally partitioned into interacting blocks with Granger-causal ordering; within the DFM framework, we extend the current approximate factor models framework in which idiosyncratic components are required to be weakly correlated to a new modeling regime where strong correlations are permitted. Further, we consider the estimation of high-dimensional factor-augmented vector regressive (FAVAR) models, which can be viewed as the bridge connecting the two paradigms, while partially addressing some of their respective limitations.

1.1 Existing Work and Challenges.

VAR models constitute a popular framework for modeling multiple time series due to their analytical tractability. Further, they have been the subject of extensive theoretical and empirical work. Ever since their introduction, VAR models have been used in various areas

such as macroeconomic modeling [**?**], finance [**?**], control theory [**?**], and more recently neuroscience [**?**]. In particular, VAR models represent a standard tool for macroeconomic forecasting and investigating the interdependences of multiple macro-economic indicators, while further allowing policy makers to examine the effect of structural shocks through impulse response function analysis. However, since the number of parameters in VAR models grows quadratically with the number of time series under study and the issue is additionally compounded with the number of lags involved, the actual size (number of time series) of the VAR system is typically limited, due to sample size considerations.

Recent work, through imposing (structured) sparsity constraints on the model parameters, has enabled the estimation of large scale VAR models even in the presence of relatively small sample sizes. Such sparse VAR models have been employed in diverse application domains; for example, **?**] examined connectivity patterns across brain regions, while **? ?**] the interactions amongst genes. Further, theoretical properties of sparse VAR model parameter estimates were also established; specifically, **?**] examined Lasso penalized Gaussian VAR models and proved consistency results, while providing technical tools useful for analysis of sparse models involving temporally dependent data.

In contrast to VAR models that focus on temporal cross-correlations across time series, DFMs investigate contemporaneous relationships by aggregating their cross-correlation information into a few latent factors that exhibit temporal dynamics themselves, and express each variable as a linear combination of such factors plus an idiosyncratic component. Such models have been widely used in reducing the dimension of large datasets; in particular, in the field of economic forecasting [**?**] and financial econometrics [**?**]. By focusing on the factor model equation which decomposes the large number of series into factors, it results in a factor model, whose properties have been thoroughly investigated in **?**] under the assumption that the correlations amongst the idiosyncratic components are weak.

Factor-augmented vector autoregressive models (FAVAR), initially proposed by **?**], as a combination of the above two paradigms provides a middle ground in terms of modeling. On one hand, it summarizes the information of a large panel of times series into latent factors through the calibration equation which resembles the information aggregation within the DFM scheme; on the other, it further investigates the temporal effect of the latent factors on another set of core variables of interest by jointly modeling the two as a VAR system. The framework has been employed in a large body of empirical work [e.g. **? ?**], and its estimation and theoretical properties in the low-dimensional setting have been investigated by **?**].

Despite the long history and diverse application of the aforementioned models, a considerable number of interesting questions and challenges still remain regarding their statistical

properties under high-dimensional scaling, with selected ones addressed in this thesis. First, within the VAR modeling framework, estimation and inference of a VAR system comprising of blocks with Granger-causal ordering, under structural assumptions on the parameters. The concept of Granger causality [**?**] and its related testing problems have been thoroughly studied in low-dimensional settings, yet is missing in the high-dimensional time series context. Second, within the DFM framework, the relaxation of the weakly correlated assumption amongst the idiosyncratic components. In particular, due to its theoretically appealing properties in establishing consistency results for the estimates, the weakly correlated assumption is prevalent in the literature analyzing such models, yet is excessively stringent and often fails to hold in real applications. In fact, practitioners have detected such issues [e.g., **? ?**], yet failed to address them in a principled way due to the lack of appropriate technical tools. Finally, comes the extension of FAVAR models to the high-dimensional setting. Currently the investigation of such models lies solely in the low-dimensional context with the core block of interest consisting of only a few variables. There is a need to extend them to the high-dimensional setting, allowing the inclusion of a larger number of time series, so that their bridging role between VAR and DFM can be fully explored and utilized.

1.2 Thesis Contributions.

This thesis makes the following contributions to the existing literature. As a general algorithmic and theoretical development, we consider the estimation of both the regression and error covariance parameters for a multivariate regression model through penalized maximum likelihood. We establish the algorithmic convergence of the optimization procedure and the consistency properties of the estimators, leveraging the bi-convexity of the objective function and appropriately bounding the estimation error through all iterations of the algorithm. The key results and employed proof techniques are broadly applicable to other settings, where the objective function of the underlying statistical estimation problem is bi-convex.

Within the high-dimensional VAR modeling paradigm, we consider the VAR-X model (X standing for the inclusion of exogenous variables), with the two blocks (endogenous and exogenous variables) being components of a joint VAR system with Granger-causal ordering. Specifically, we focus on the estimation of model parameters under certain structural assumptions (e.g. sparsity, low-rankness) and provide estimators with statistical guarantees. Moreover, we devise a testing procedure that addresses the hypothesis testing problem with the null hypothesis being that a group of variables does not collectively Granger-cause another, while taking into consideration the structural characteristics of the parameter in question under the alternative. Since both the estimation and the testing procedure are

for time series data, the effect of temporal dependence introduces a number of technical challenges that need to be accounted for and properly handled.

We extend the FAVAR model to the high-dimensional setting and enable the estimation of its model parameters, so that a much larger set of variables can be studied. Since all of the currently existing model identification constraints that are feasible under the low-dimensional setting fail to remain effective when we deal with a high-dimensional setting, we propose a novel model identification scheme applicable to that setting. Moreover, it can be seamlessly incorporated in the optimization problem, that yields estimators with good statistical properties. Further, we establish a high-probability error bound of the estimated transition matrix of the underlying VAR model, when one block of variables (the factors) is contaminated with non-random errors, due to it being estimated from other time series data.

Finally for the DFM paradigm, it is worth pointing out that the current framework already accommodates a large number of time series, since to obtain consistent estimates for the factors the size of the panel of variables is required to grow to infinity [**?**]. However as previously mentioned, all current results rely on the assumption that the correlation among the idiosyncratic components is weak, but often fails to hold in practice. Therefore, we relax the weakly correlated assumption for idiosyncratic components in approximate factor models and propose a new model which allows for moderate-to-strong correlations. Specifically, without deviating too much from the grand DFM paradigm, the proposed model provides a principled way of dealing with datasets with such features, while having sufficient modeling and theoretical justifications. Building upon the formulation in estimating high-dimensional FAVAR models, we consider an optimization problem which is further convexified so that the corresponding alternating minimization algorithm has convergence guarantees; meanwhile, its global optimizer, which corresponds to the model parameter estimates, possesses certain statistical guarantees.

For all proposed formulations and corresponding estimators in the above settings, we also consider their empirical implementation issues and devise computationally efficient procedures to obtain the estimates. Further, these models are used in several real applications involving financial and economic data, and yield interpretable results that uncover some of the interactions among the variables of interest.

1.3 Organization of the Thesis.

The main body of this thesis consists of four chapters (Chapters II to V), with each addressing the issues outlined above sequentially, and is concluded with Chapter VI.

In Chapter II, through a multi-layered Gaussian graphical model, we consider the estimation of inter- and intra- block structure under the setting where the data are independently identically distributed (iid). The proposed formulation is based on penalized maximum likelihood, and we consider a block-coordinate descent algorithm to obtain an estimate that is guaranteed to be a global optimizer and has statistical guarantees, while leveraging the bi-convexity of the objective function and the descent property of the algorithm. The performance of the proposed estimator is evaluated on synthetic data, in conjunction with its two-step competitors that effectively terminates the alternating procedure after one iteration.

In Chapter III, we consider the estimation of a multi-block VAR system whose components have Granger-causal ordering, under certain structural assumptions on the model parameters. With a similar penalized ML formulation to that of the iid setting, we provide estimates corresponding to both the transition matrices and the inverse covariance matrix, with the former capturing the lead-lag relationship between and within blocks and the latter capturing the contemporaneous conditional interdependence. Further, we devise a procedure for testing the existence of group Granger-causality, tailoring to the structural characteristic of the parameter in question under the alternative hypothesis. The performance of both the estimation and the testing procedures are assessed via simulation studies, and the model is illustrated on a motivating real data application involving both stock prices and macroeconomic variables.

Next, in Chapter IV, we extend the FAVAR model to high-dimensional settings. With the proposed model identification constraint, we enable the estimation of the FAVAR model under such settings. Specifically, for the calibration equation, we formulate an optimization problem based on least squares loss and the structural assumptions of the parameters, while further compactifying the formulation based on the identification constraint; then for the VAR equation, we estimate the transition matrix based on the estimated factors from the calibration equation and the observed samples of the core variables. The obtained estimates are proved to have good statistical properties. Further, we consider the empirical implementation of the proposed formulation, whose estimation and forecasting performance are evaluated based on synthetic data. Finally, we employ the model to study the interlinkage among commodity prices while taking into account the effect of global economic activities.

Finally in Chapter V, we consider a relaxation of the weak correlation assumption in approximate factor models. Specifically, by modeling the dynamics of the idiosyncratic component of the classical factor model through a sparse VAR, then de-correlating such serial correlation, we obtain a new model which automatically allows for a stronger correlation among the idiosyncratic component and simultaneously resolves the endogeneity issue of the original model. Building upon the formulation in estimating the high-dimensional FAVAR

models, we further convexify the formulation, so that the estimator given by the global optimizer possesses both algorithmic and theoretical guarantees. Again, the performance of the estimation procedure is evaluated through synthetic data, and the model is further applied to a financial dataset comprising of the stock prices of large financial institutions across banking, broker & dealers and insurance companies to investigate their connectivity pattern over time.

CHAPTER II

Penalized Maximum Likelihood Estimation of Multi-layered Gaussian Graphical Models

2.1 Introduction.

The estimation of directed and undirected graphs from high-dimensional data has received a lot of attention in the machine learning and statistics literature [e.g., see **?** , and references therein], due to their importance in diverse applications including the understanding of biological processes and disease mechanisms, financial systems stability and social interactions, just to name a few [**?** **?** **?**]. In the case of undirected graphs, the edges capture conditional dependence relationships between the nodes, while for directed graphs they are used to model causal relationships [**?**].

However, in a number of applications the nodes can be *naturally partitioned* into sets that exhibit interactions both between them and amongst them. As an example, consider an experiment where one has collected data for both genes and metabolites for the same set of patient specimens. In this case, we have three types of interactions between genes and metabolites: regulatory interactions between the two of them and co-regulation within the gene and within the metabolic compartments. The latter two types of relationships can be expressed through undirected graphs within the sets of genes and metabolites, respectively, while the regulation of metabolites by genes corresponds to directed edges. Note that in principle there are feedback mechanisms from the metabolic compartment to the gene one, but these are difficult to detect and adequately estimate in the absence of carefully collected time course data. Another example comes from the area of financial economics, where one collects data on returns of financial assets (e.g. stocks, bonds) and also on key macroeconomic indicators (e.g. interest rate, prices indices, various measures of money supply and various unemployment indices). Once again, over short time periods there is influence from the economic variables to the returns (directed edges), while there are co-dependence relationships between the asset returns and the macroeconomic variables, respectively, that can

be modeled as undirected edges.

Technically, such *layered* network structures correspond to multi-partite graphs that possess undirected edges and exhibit a directed acyclic graph structure between the layers, as depicted in Figure 2.1, where we use directed solid edges to denote the dependencies across layers and dashed undirected edges to denote within-layer conditional dependencies.

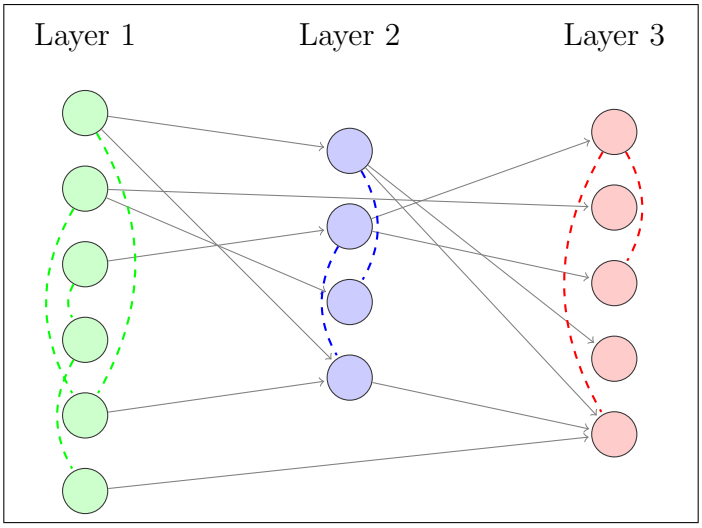

Figure 2.1: Diagram for a three-layered network.

Selected properties of such so-called *chain graphs* have been studied in the work of **?**], with an emphasis on two alternative Markov properties including the LWF Markov property [**?** **?**] and the AMP Markov property [**?**].

While layered networks being interesting from a theoretical perspective and having significant scope for applications, their estimation has received little attention in the literature. Note that for a 2-layered structure, the directed edges can be obtained through a multivariate regression procedure, while the undirected edges in both layers through existing procedures for graphical models (for more technical details see Section 2.2.2). This is the strategy leveraged in the work of **?**], where for a 2-layered network structure they proposed a multivariate regression with covariance estimation (MRCE) method for estimating the undirected edges in the second layer and the directed edges between them. A block coordinate descent algorithm was introduced to estimate the directed edges, while the popular glasso estimator [**?**] was used for the undirected edges. However, this method does not scale well according to the simulation results presented and no theoretical properties of the estimates were provided.

In follow-up work, **?**] used a cyclic block coordinate descent algorithm and claimed convergence to a stationary point leveraging a result in **?**] (see Proposition 2 in the Supplemental material). Unfortunately, a key assumption in **?**] —namely, that a corresponding coordinate wise optimization problem that is given by a high-dimensional lasso regression

has unique minimum- fails and hence the convergence result does not go through.

In related work, **?**] proposed the Plug-in Joint Weighted Lasso (PWL) and the Plug-in Joint Graphical Weighted Lasso (PWGL) estimator for estimating the same 2-layered structure, where they use a weighted version of the algorithm in **?**] and also provide theoretical results for the low dimensional setting, where the number of samples exceeds the number of potential directed and undirected edges to be estimated. Finally, **?**] proposed a method for estimating the same 2-layered structure and provided corresponding theoretical results in the high dimensional setting. The Dantzig-type estimator [**?**] was used for the regression coefficients and the corresponding residuals were used as surrogates, for obtaining the precision matrix through the CLIME estimator [**?**]. In another line of work [**?** **?** **?**], structured sparsity of directed edges was considered and the edges were estimated with a different parametrization of the objective function. We further elaborate on the connections of our work with these three papers in Section 2.5.

The above work assumed a Gaussian distribution for the data, in more recent work by **?**], the authors constructed the model under a general *mixed graphical model* framework, which allows each node-conditional distribution to belong to a potentially different univariate exponential family. In particular, with an underlying *mixed MRF* graph structure, instead of maximizing the joint likelihood, the authors proposed to estimate the homogeneous and heterogeneous neighborhood for each node, by obtaining the ℓ_1 regularized M-estimator of the node-conditional distribution parameters, using traditional approaches [e.g. **?**] for neighborhood estimation. However, rather than estimating directed edges directly, the directed edges are obtained from a nonlinear transformation of the estimated homogeneous and heterogeneous neighborhood, whose sparsity pattern gets compromised during the process.

In this work, we obtain the regularized maximum likelihood estimator under a sparsity assumption on both directed and undirected parameters for multi-layered Gaussian graphical models and establish its consistency properties in a high-dimensional setting. As discussed in Section 2.3, the problem is *not jointly convex* on the parameters, but convex on selected subsets of them. Further, it turns out that the problem is *biconvex* if we consider a recursive multi-stage estimation approach that at each stage involves only regression parameters (directed edges) from preceding layers and precision matrix parameters (undirected edges) for the *last layer considered* in that stage. Hence, we decompose the multi-layer network structure estimation into a sequence of 2-layer problems that allows us to establish the desired results. Leveraging the biconvexity of the 2-layer problem, we establish the convergence of the iterates to the maximum-likelihood estimator, which under certain regularity conditions is arbitrarily close to the true parameters. The theoretical guarantees provided require a *uniform control* of the precision of the regression and precision matrix parameters, which

poses a number of theoretical challenges resolved in Section 2.3.

In summary, despite the lack of overall convexity, we are able to provide theoretical guarantees for the MLE in a high dimensional setting. We believe that the proposed strategy is generally applicable to other non-convex statistical estimation problems that can be decomposed to two biconvex problems. Further, to enhance the numerical performance of the MLE in finite (and small) sample settings, we introduce a screening step that selects active nodes for the iterative algorithm used and that leverages recent developments in the high-dimensional regression literature [e.g., **? ? ?**]. We also post-process the final MLE estimate through a stability selection procedure. As mentioned above, the screening and stability selection steps are beneficial to the performance of the MLE in finite samples and hence recommended for similarly structured problems.

The remainder of the chapter is organized as follows. In Section 2.2, we introduce the proposed methodology, with an emphasis on how the multi-layered network estimation problem is decomposed into a sequence of two-layered network estimation problem(s). In Section 2.3, we provide theoretical guarantees for the estimation procedure posited. In particular, we show consistency of the estimates and convergence of the algorithm, under a number of common assumptions in high-dimensional settings. In Section 2.4, we show the performance of the proposed algorithm with simulation results under different simulation settings, and introduce several acceleration techniques which speed up the convergence of the algorithm and reduce the computing time in practical settings. Finally in Section 2.5, we briefly discuss the connections between different parametrizations of the layered network estimation problem.

2.2 Problem Formulation.

Consider an M-layered Gaussian graphical model. Suppose there are p_m nodes in Layer m, denoted by

$$\vec{X}^m = (X_1^m, \cdots, X_{p_m}^m)^\top, \quad \text{for } m = 1, \cdots, M.$$

The structure of the model is given as follows:

- Layer 1. $\vec{X}^1 = (X_1^1, \cdots, X_{p_1}^1)^\top \sim \mathcal{N}(0, \Sigma^1)$.
- Layer 2. For $j = 1, \cdots, p_2$: $X_j^2 = (B_j^{12})^\top \vec{X}^1 + \epsilon_j^2$, with $B_j^{12} \in \mathbb{R}^{p_1}$, and $\vec{\epsilon}^2 = (\epsilon_1^2, \cdots, \epsilon_{p_2}^2)^\top \sim \mathcal{N}(0, \Sigma^2)$.

  $\vdots$

– Layer M. For $j = 1, 2, \cdots, p_M$:

$$X_j^M = \sum_{m=1}^{M-1} \{(B_j^{mM})^\top \vec{X}^m\} + \epsilon_j^M, \quad \text{where } B_j^{mM} \in \mathbb{R}^{p_m} \text{ for } m = 1, \cdots, M-1,$$

and $\vec{\epsilon}^M = (\epsilon_1^M, \cdots, \epsilon_{p_M}^M)^\top \sim \mathcal{N}(0, \Sigma^M)$.

The parameters of interest are *all directed edges* that encode the dependencies across layers, that is,

$$B^{st} := \left[B_1^{st} \ \cdots \ B_{p_t}^{st}\right], \quad \text{for } 1 \le s < t \le M,$$

and *all undirected edges* that encode the conditional dependencies within layers after adjusting for the effects from directed edges, that is:

$$\Theta^m := (\Sigma^m)^{-1}, \quad \text{for } m = 1, \cdots, M.$$

It is assumed that B^{st} and Θ^m are *sparse* for all $1, \ldots, M$ and $1 \le s < t \le M$.

Given centered data for all M layers with each layer $m = 1, \cdots, M$ denoted by $\mathbf{X}^m \in \mathbb{R}^{n \times p_m}$ whose rows are iid realizations of $\vec{X}^m$, we aim to obtain the MLE for all $B^{st}, 1 \le s < t \le M$ and all $\Theta^m, m = 1, \cdots, M$ parameters. Henceforth, we use $\vec{X}^m$ to denote the random vector of Layer m nodes, and $\mathbf{X}_j^m$ to denote the jth column in the data matrix $\mathbf{X}^m \in \mathbb{R}^{n \times p_m}$ whenever there is no ambiguity.

Through Markov factorization [**?**], the full log-likelihood function can be decomposed as

$$\begin{aligned}
&\ell(\mathbf{X}^m; B^{st}, \Theta^m, 1 \le s < t \le M, 1 \le m \le M) \\
&\qquad = \ell(\mathbf{X}^M | \mathbf{X}^{M-1}, \cdots, \mathbf{X}^1; B^{1M}, \cdots, B^{M-1,M}, \Theta^M) \\
&\qquad\quad + \ell(\mathbf{X}^{M-1} | \mathbf{X}^{M-2}, \cdots, \mathbf{X}^1; B^{1M-1}, \cdots, B^{M-2,M-1}, \Theta^{M-1}) \\
&\qquad\quad + \cdots + \ell(\mathbf{X}^2 | \mathbf{X}^1; B^{12}, \Theta^2) + \ell(\mathbf{X}^1; \Theta^1) \\
&\qquad = \ell(\mathbf{X}^1; \Theta^1) + \sum\nolimits_{m=2}^{M} \ell(\mathbf{X}^m | \mathbf{X}^1, \cdots, \mathbf{X}^{m-1}; B^{1m}, \cdots, B^{m-1,m}, \Theta^m).
\end{aligned}$$

Note that the summands share no common parameters, which enables us to maximize the likelihood with respect to individual parameters in the M terms separately. More importantly, by conditioning Layer m nodes on nodes in its previous $(m-1)$ layers, we can treat Layer m nodes as the"response" layer, and all nodes in the previous $(m-1)$ layers combined as a super "parent" layer. If we ignore the structure within the bottom layer $(\vec{X}^1)$ for the moment, the M-layered network can be viewed as $(M-1)$ two-layered networks, each comprising a response layer and a parent layer. Thus, the network structure in Figure 2.1 can be viewed as a 2 two-layered network: for the first network, Layer 3 is the response layer, while

Layers 1 and 2 combined form the "parent" layer; for the second network, Layer 2 is the response layer, and Layer 1 is the "parent" layer. Therefore, the problem for estimating all $\binom{M}{2}$ coefficient matrices and M precision matrices can be translated into estimating $(M-1)$ two-layered network structures with directed edges from the parent layer to the response layer, and undirected edges within the response layer, and finally estimating the undirected edges within the bottom layer separately.

Since all estimation problems boil down to estimating the structure of a 2-layered network, we focus the technical discussion on introducing our proposed methodology for a 2-layered network setting[1]. The theoretical results obtained extend in a straightforward manner to an M-layered Gaussian graphical model.

Remark 2.1. For the M-layer network structure, we impose certain identifiability-type condition on the largest "parent" layer (encompassing $M-1$ layers), so that the directed edges of the entire network are estimable. The imposed condition translates into a minimum eigenvalue-type condition on the population precision matrix within layers, and conditions on the magnitude of dependencies across layers. Intuitively, consider a three-layered network: if $\vec{X}^1$ and $\vec{X}^2$ are highly correlated, then the proposed (as well as any other) method will exhibit difficulties in distinguishing the effect of $\vec{X}^1$ on $\vec{X}^3$ from that of $\vec{X}^2$ on $\vec{X}^3$. The (group) identifiability-type condition is thus imposed to obviate such circumstances. An in-depth discussion on this issue is provided in Section 2.3.4.

2.2.1 A two-layered network setup.

Consider a two-layered Gaussian graphical model with p_1 nodes in the first layer, denoted by $X = (X_1, \cdots, X_{p_1})'$, and p_2 nodes in the second layers, denoted by $Y = (Y_1, \cdots, Y_{p_2})'$. The model is defined as

- $X = (X_1, \cdots, X_{p_1})^\top \sim \mathcal{N}(0, \Sigma_X)$.
- For $j = 1, 2, \cdots, p_2$: $Y_j = B_j^\top X + \epsilon_j$, $B_j \in \mathbb{R}^{p_1}$ and $\epsilon = (\epsilon_1, \cdots, \epsilon_{p_2})^\top \sim \mathcal{N}(0, \Sigma_\epsilon)$.

The parameters of interest are: $\Theta_X := \Sigma_X^{-1}, \Theta_\epsilon := \Sigma_\epsilon^{-1}$ and $B := [B_1, \cdots, B_{p_2}]$. As with most estimation problems in the high dimensional setting, we assume these parameters to be sparse.

Now given data $\mathbf{X} \in \mathbb{R}^{n \times p_1}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p_2}$ with their rows being iid random samples of X and Y respectively, both centered, we would like to use the penalized maximum likelihood

[1] In Appendix A.4 we give a detail example on how our proposed method works under a 3-layered network setting.

approach to obtain estimates for Θ_X, Θ_ϵ and B. The full log-likelihood can be written as

$$\ell(\mathbf{X}, \mathbf{Y}; B, \Theta_\epsilon, \Theta_X) = \ell(\mathbf{Y}|\mathbf{X}; \Theta_\epsilon, B) + \ell(\mathbf{X}; \Theta_X). \tag{2.1}$$

Note that the first term only involves Θ_ϵ and B, and the second term only involves Θ_X. Hence, (2.1) can be maximized by maximizing $\ell(\mathbf{Y}|\mathbf{X})$ w.r.t. (Θ_ϵ, B), and maximizing $\ell(\mathbf{X})$ w.r.t. Θ_X, respectively. $\widehat{\Theta}_X$ can be obtained using traditional methods for estimating undirected graphs, e.g., the Graphical Lasso [**?**] or the Nodewise Regression prcoedure [**?**]. Therefore, the rest of this paper will mainly focus on obtaining estimates for Θ_ϵ and B. In the next subsection, we introduce our estimation procedure for obtaining the MLE for Θ_ϵ and B.

Remark 2.2. Our proposed method is targeted towards maximizing $\ell(\mathbf{Y}|\mathbf{X}; \Theta_\epsilon, B)$ (with proper penalization) in (2.1) only, which gives the estimates for across-layers dependencies between the response layer and the parent layer, as well as estimates for the conditional dependencies within the response layer each time we solve a 2-layered network estimation problem. For an M-layered estimation problem, the maximization regarding $\ell(\mathbf{X}; \Theta_X)$ occurs only when we are estimating the within-layer conditional dependencies for the bottom layer.

2.2.2 Estimation algorithm.

The conditional likelihood for response $\mathbf{Y}$ given $\mathbf{X}$ can be written as

$$\ell(\mathbf{Y}|\mathbf{X}) \quad = (\tfrac{1}{\sqrt{2\pi}})^{np_2} |\Sigma_\epsilon \otimes \mathrm{I}_n|^{-1/2} \exp\left\{ -\tfrac{1}{2} (\mathcal{Y} - \mathcal{X}\beta)^\top (\Sigma_\epsilon \otimes \mathrm{I}_n)^{-1} (\mathcal{Y} - \mathcal{X}\beta) \right\},$$

where $\mathcal{Y} = \mathrm{vec}(\mathbf{Y})$, $\mathcal{X} = \mathrm{I}_{p_2} \otimes \mathbf{X}$ and $\beta = \mathrm{vec}(B)$. After writing out the Kronecker product, the log-likelihood can be written as

$$\ell(\mathbf{Y}|\mathbf{X}) = \text{constant} + \frac{n}{2} \log\det \Theta_\epsilon - \frac{1}{2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_2} \sigma_\epsilon^{ij} (\mathbf{Y}_i - \mathbf{X}B_i)^\top (\mathbf{Y}_j - \mathbf{X}B_j).$$

Here, $\mathbf{Y}_j$ is the j-th column of $\mathbf{Y}$ and σ_ϵ^{ij} denotes the ij-th entry of Θ_ϵ. With ℓ_1 penalization which induces sparsity, we formulate the following optimization problem using penalized log-likelihood, which was initially proposed in **?**], and has also been examined in **?**]:

$$\min_{\substack{B \in \mathbb{R}^{p_1 \times p_2} \\ \Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}} \left\{ \frac{1}{n} \sum_{j=1}^{p_2} \sum_{i=1}^{p_2} \sigma_\epsilon^{ij} (\mathbf{Y}_i - \mathbf{X}B_i)^\top (\mathbf{Y}_j - \mathbf{X}B_j) - \log\det \Theta_\epsilon + \lambda_n \sum_{j=1}^{p_2} \|B_j\|_1 + \rho_n \|\Theta_\epsilon\|_{1,\text{off}} \right\}, \tag{2.2}$$

and the first term in (2.2) can be equivalently written as

$$\mathrm{tr}\left\{\frac{1}{n}\begin{bmatrix}(\mathbf{Y}_1-\mathbf{X}B_1)^\top\\ \vdots \\ (\mathbf{Y}_{p_2}-\mathbf{X}B_{p_2})^\top\end{bmatrix}\begin{bmatrix}(\mathbf{Y}_1-\mathbf{X}B_1) & \cdots & (\mathbf{Y}_{p_2}-\mathbf{X}B_{p_2})\end{bmatrix}\Theta_\epsilon\right\} := \mathrm{tr}(S\Theta_\epsilon).$$

where S is defined as the sample covariance matrix of $\mathbf{E} := \mathbf{Y} - \mathbf{X}B$. This gives rise to the following optimization problem:

$$\min_{\substack{B\in\mathbb{R}^{p_1\times p_2}\\ \Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}} \left\{\mathrm{tr}(S\Theta_\epsilon) - \log\det\Theta_\epsilon + \lambda_n\sum_{j=1}^{p_2}\|B_j\|_1 + \rho_n\|\Theta_\epsilon\|_{1,\mathrm{off}}\right\} =: f(B,\Theta_\epsilon), \tag{2.3}$$

where $\|\Theta\|_{1,\mathrm{off}}$ is the absulote sum of the off-diagonal entries in Θ, λ_n and ρ_n are both positive tuning parameters.

Note that the objective function (2.3) is *not jointly convex* in (B,Θ_ϵ), but only convex in B for fixed Θ_ϵ and in Θ_ϵ for fixed B; hence, it is bi-convex, which in turn implies that the proposed algorithm may fail to converge to the global optimum, especially in settings where $p_1 > n$, as pointed out by **?**]. As is the case with most non-convex problems, good initial parameters are beneficial for fast convergence of the algorithm, a fact supported by our numerical work on the present problem. Further, a good initialization is critical in establishing convergence of the algorithm for this problem (see Section 2.3.1). To that end, we introduce a *screening step* for obtaining a good initial estimate for B. The theoretical justification for employing the screening step is provided in Section 2.3.3.

An outline of the computational procedure is presented in Algorithm II.1, while the details of each step involved are discussed next.

Screening. For each variable $\mathbf{Y}_j, j = 1,\cdots,p_2$ in the response layer, regress $\mathbf{Y}_j$ on $\mathbf{X}$ via the de-biased Lasso procedure proposed by **?**]. The output consists of the p-value(s) for each predictor in each regression, denoted by P_j, with $P_j \in [0,1]^{p_1}$. To control the family-wise error rate of the estimates, we do a Bonferroni correction at level α: define $\alpha^\star = \alpha/(p_1p_2)$ and set $B_{j,k} = 0$ if the p-value obtained for the k'th predictor in the j'th regression $P_{j,k}$ exceeds $\alpha^\star$. Further, let

$$\mathcal{B}_j = \{B_j \in \mathbb{R}^{p_1} : B_{j,k} = 0 \text{ if } k \in \widehat{S}_j^c\} \subseteq \mathbb{R}^{p_1}, \tag{2.4}$$

where $\widehat{S}_j$ is the collection of indices for those predictors deemed "active" for response Y_j:

$$\widehat{S}_j = \{k : P_{j,k} < \alpha^\star\}, \quad \text{for } j = 1,\cdots,p_2.$$

Algorithm 2.1: Computational procedure for estimating B and Θ_ϵ.

Input: Data from the parent layer $\mathbf{X}$ and the response layer $\mathbf{Y}$.
Screening:
  for $j = 1, \cdots, p_2$ **do**
    regress $\mathbf{Y}_j$ on $\mathbf{X}$ using the de-biased Lasso procedure in **?**] and obtain the corresponding vector of p-values P_j
  end
  obtain adjusted p-values $\widetilde{P}_j$ by applying Bonferroni correction to $\text{vec}(P_1, \cdots, P_j)$ determine the support set $\mathcal{B}_j$ for each regression using (2.4).
Initialization:
  Initialize column $j = 1, \cdots, p_2$ of $\widehat{B}^{(0)}$ by solving (2.5).
  Initialize $\widehat{\Theta}_\epsilon^{(0)}$ by solving (2.2.2) using the graphical lasso [**?**].
Alternating Search:
  while $|f(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}) - f(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k+1)})| \geq \epsilon$ **do**
    update $\widehat{B}$ with (2.6) update $\widehat{\Theta}_\epsilon$ with (2.8)
  end
Refitting B and Θ_ϵ
  for $j = 1, \cdots, p_2$ **do**
    Obtain the refitted $\widetilde{B}_j$ using (2.9)
  end
  re-estimate $\widetilde{\Theta}_\epsilon$ using (2.10) with W coming from stability selection.
Output: Final Estimates $\widetilde{B}$ and $\widetilde{\Theta}_\epsilon$.

Therefore, subsequent estimation of the elements of B will be restricted to $\mathcal{B}_1 \times \cdots \times \mathcal{B}_{p_2}$.

Alternating search. In this step, we utilize the bi-convexity of the problem and estimate B and Θ_ϵ by minimizing in an iterative fashion the objective function with respect to (w.r.t.) one set of parameters, while holding the other set fixed within each iteration.

As with most iterative algorithms, we need an initializer; for $\widehat{B}^{(0)}$ it corresponds to a Lasso/Ridge regression estimate with a small penalty, while for $\widehat{\Theta}_\epsilon$ we use the Graphical Lasso procedure applied to the residuals obtained from the first stage regression. That is, for each $j = 1, \cdots, p_2$,

$$\widehat{B}_j^{(0)} = \underset{B_j \in \mathcal{B}_j}{\arg\min} \Big\{ \|\mathbf{Y}_j - \mathbf{X}B_j\|_2^2 + \lambda_n^0 \|B_j\|_1 \Big\}, \tag{2.5}$$

where λ_n^0 is some small tuning parameter for initialization, and set $\widehat{\mathbf{E}}_j^{(0)} := \mathbf{Y}_j - \mathbf{X}\widehat{B}_j^{(0)}$. An initial estimate for $\widehat{\Theta}_\epsilon$ is then given by solving for the following optimization problem with the graphical lasso [**?**] procedure:

$$\widehat{\Theta}_\epsilon^{(0)} = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\arg\min} \Big\{ \log\det\Theta_\epsilon - \text{tr}(\widehat{S}^{(0)}\Theta_\epsilon) + \rho_n \|\Theta_\epsilon\|_{1,\text{off}} \Big\},$$

where $\widehat{S}^{(0)}$ is the sample covariance matrix based on $(\widehat{E}_1^{(0)}, \cdots, \widehat{E}_{p_2}^{(0)})$.

Next, we use an alternating block coordinate descent algorithm with ℓ_1 penalization to reach a stationary point of the objective function (2.3).

– Update B as

$$\widehat{B}^{(k+1)} = \underset{B \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_{p_2}}{\arg\min} \Big\{ \frac{1}{n} \sum_{i=1}^{p_2} \sum_{j=1}^{p_2} (\widehat{\sigma}_\epsilon^{ij})^{(k)} (\mathbf{Y}_i - \mathbf{X}B_i)^\top (\mathbf{Y}_j - \mathbf{X}B_j) + \lambda_n \sum_{j=1}^{p_2} \|B_j\|_1 \Big\}, \quad (2.6)$$

which can be obtained by cyclic coordinate descent w.r.t each column B_j of B, that is, update each column B_j by:

$$\widehat{B}_j^{(t+1)} = \underset{B_j \in \mathcal{B}_j}{\arg\min} \left\{ \tfrac{(\widehat{\sigma}_\epsilon^{jj})^{(k)}}{n} \|\mathbf{Y}_j + \mathbf{r}_j^{(t+1)} - \mathbf{X}B_j\|_2^2 + \lambda_n \|B_j\|_1 \right\}, \quad (2.7)$$

where $\mathbf{r}_j^{(t+1)} = \frac{1}{(\widehat{\sigma}_\epsilon^{jj})^{(k)}} \big[\sum_{i=1}^{j-1} (\widehat{\sigma}_\epsilon^{ij})^{(k)} (\mathbf{Y}_i - \mathbf{X}\widehat{B}_i^{(t+1)}) + \sum_{i=j+1}^{p_2} (\widehat{\sigma}_\epsilon^{ij})^{(k)} (\mathbf{Y}_i - \mathbf{X}\widehat{B}_i^{(t)}) \big]$, and iterate over all columns until convergence. Here, we use k to index the outer iteration while minimizing w.r.t. B or Θ_ϵ, and use t to index the inner iteration while cyclically minimizing w.r.t. each column of B.

– Update Θ_ϵ as

$$\widehat{\Theta}_\epsilon^{(k+1)} = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\arg\min} \big\{ \log\det\Theta_\epsilon - \mathrm{tr}(\widehat{S}^{(k+1)}\Theta_\epsilon) + \rho_n \|\Theta_\epsilon\|_{1,\mathrm{off}} \big\}, \quad (2.8)$$

where $\widehat{S}^{(k+1)}$ is the sample covariance matrix based on $\widehat{\mathbf{E}}_j^{(k+1)} = \mathbf{Y}_j - \mathbf{X}\widehat{B}_j^{(k+1)}, j = 1, \cdots, p_2$.

Refitting and stabilizing. As noted in the introduction, this step is beneficial in applications, especially when one deals with large scale multi-layer networks and relatively smaller sample sizes. Denote the solution obtained by the above iterative procedure by B^∞ and Θ_ϵ^∞. For each $j = 1, \cdots, p_2$, set $\widetilde{\mathcal{B}}_j = \{B_j : B_{j,i} = 0 \text{ if } B_{j,i}^\infty = 0, B_j \in \mathbb{R}^{p_1}\}$ and the final estimate for B_j is given by ordinary least squares:

$$\widetilde{B}_j = \underset{B_j \in \widetilde{\mathcal{B}}_j}{\arg\min} \ \|\mathbf{Y}_j - \mathbf{X}B_j\|^2. \quad (2.9)$$

For Θ_ϵ, we obtain the final estimate by a combination of stability selection [**?**] and graphical lasso [**?**]. That is, after obtaining the refitted residuals $\widetilde{\mathbf{E}}_j := \mathbf{Y}_j - \mathbf{X}\widetilde{B}_j, j = 1, \cdots, p_2$, based on the stability selection procedure with the graphical lasso, we obtain the stability path, or probability matrix W for each edge, which records the proportion of each edge being

selected based on bootstrapped samples of $\widetilde{\mathbf{E}}_j$'s. Then, using this probability matrix W as a weight matrix, we obtain the final estimate of $\widetilde{\Theta}_\epsilon$ as follow:

$$\widetilde{\Theta}_\epsilon = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\arg\min} \left\{ \log\det\Theta_\epsilon - \mathrm{tr}(\widetilde{S}\Theta_\epsilon) + \widetilde{\rho}_n \|(1-W) * \Theta_\epsilon\|_{1,\mathrm{off}} \right\}, \tag{2.10}$$

where we use $*$ to denote the element-wise product of two matrices, and $\widetilde{S}$ is the sample covariance matrix based on the refitted residuals $\widetilde{E}$. Again, (2.10) can be solved by the graphical lasso procedure [**?**], with $\widetilde{\rho}_n$ properly chosen.

2.2.3 Tuning parameter selection.

To select the tuning parameters (λ_n, ρ_n), we use the Bayesian Information Criterion(BIC), which is the summation of a goodness-of-fit term (log-likelihood) and a penalty term. The explicit form of BIC (as a function of B and Θ_ϵ) in our setting is given by

$$\mathrm{BIC}(B, \Theta_\epsilon) = -\log\det\Theta_\epsilon + \mathrm{tr}(S\Theta_\epsilon) + \frac{\log n}{n}\left(\frac{\|\Theta_\epsilon\|_0 - p_2}{2} + \|B\|_0\right)$$

where

$$S := \frac{1}{n}\begin{bmatrix} (\mathbf{Y}_1 - \mathbf{X}B_1)^\top \\ \vdots \\ (\mathbf{Y}_{p_2} - \mathbf{X}B_{p_2})^\top \end{bmatrix}\begin{bmatrix} (\mathbf{Y}_1 - \mathbf{X}B_1) & \cdots & (\mathbf{Y}_{p_2} - \mathbf{X}B_{p_2}) \end{bmatrix},$$

and $\|\Theta_\epsilon\|_0$ is the total number of nonzero entries in Θ_ϵ. Here we penalize the non-zero elements in the upper-triangular part of Θ_ϵ and the non-zero ones in B. We choose the combination (λ_n^*, ρ_n^*) over a grid of (λ, ρ) values, and (λ_n^*, ρ_n^*) should minimize the BIC evaluated at $(B^\infty, \Theta_\epsilon^\infty)$.

2.3 Theoretical Results.

In this section, we establish a number of theoretical results for the proposed iterative algorithm. We focus the presentation on the two-layer structure, since as explained in the previous section the multi-layer estimation problem decomposes to a series of two-layer ones. As mentioned in the introduction, one key challenge for establishing the theoretical results comes from the fact that the objective function (2.3) is not jointly convex in B and Θ_ϵ. Consequently, if we simply used properties of block-coordinate descent algorithms, we would not be able to provide the necessary theoretical guarantees for the estimates we obtain. On the other hand, the biconvex nature of the objective function allows us to establish convergence of the alternating algorithm to a stationary point, provided it is initialized from a point close enough to the true parameters. This can be accomplished using a Lasso-based

initializer for B and Θ_ϵ as previously discussed. The details of algorithmic convergence are presented in Section 2.3.1.

Another technical challenge is that each update in the alternating search step relies on estimated quantities—namely the regression and precision matrix parameters—rather than the raw data, whose estimation precision needs to be controlled *uniformly* across all iterations. The details of establishing consistency of the estimates for both fixed and random realizations are given in Section 2.3.2.

Next, we outline the structure of this section. In Section 2.3.1 Theorem 2.1, we show that for any fixed set of realization of $\mathbf{X}$ and $\mathbf{E}$ [2], the iterative algorithm is guaranteed to converge to a stationary point if estimates for all iterations lie in a compact ball around the true value of the parameters. In Section 2.3.2, we show in Theorem 2.4 that for any random $\mathbf{X}$ and $\mathbf{E}$, with high probability, the estimates for all iterations lie in a compact ball around the true value of the parameters. Then in Section 2.3.3, we show that asymptotically with $\log(p_1 p_2)/n \to 0$, while keeping the family-wise type I error under some pre-specified level, the screening step correctly identifies the true support set for each of the regressions, based upon which the iterative algorithm is provided with an initializer that is close to the true value of the parameters. Finally in Section 2.3.4, we provide sufficient conditions for both directed and undirected edges to be identifiable (estimable) for multi-layered network.

To aid the readability of the main results, we only present statements of theorems and propositions, while all proofs are relegated to Appendix A (Sections A.1 and A.2).

Throughout this section, to distinguish the estimates from the true values, we use $B^\star$ and $\Theta_\epsilon^\star$ to denote the true values.

2.3.1 Convergence of the iterative algorithm.

In this subsection, we prove that the proposed block relaxation algorithm converges to a stationary point for any fixed set of data, provided that the estimates for all iterations lie in a compact ball around the true value of the parameters. This requirement is shown to be satisfied with high probability in the next subsection 2.3.2.

Decompose the optimization problem in (2.3) as follows:

$$\min_{\substack{B\in\mathbb{R}^{p_1\times p_2}\\ \Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}} f(B,\Theta_\epsilon) := f_0(B,\Theta_\epsilon) + f_1(B) + f_2(\Theta_\epsilon),$$

[2] We actually observe $\mathbf{X}$ and $\mathbf{Y}$, which is given by a corresponding set of realization in $\mathbf{X}$ and $\mathbf{E}$ based on the model.

where

$$f_0(B,\Theta_\epsilon)=\frac{1}{n}\sum_{j=1}^{p_2}\sum_{i=1}^{p_2}\sigma_\epsilon^{ij}(\mathbf{Y}_i-\mathbf{X}B_i)^\top(\mathbf{Y}_j-\mathbf{X}B_j)-\log\det\Theta_\epsilon=\mathrm{tr}(S\Theta_\epsilon)-\log\det\Theta_\epsilon,$$

$$f_1(B)=\lambda_n\|B\|_1,\quad f_2(\Theta_\epsilon)=\rho_n\|\Theta_\epsilon\|_{1,\mathrm{off}},$$

and $\mathbb{S}_{++}^{p_2\times p_2}$ is the collection of $p_2\times p_2$ symmetric positive definite matrices. Further, denote the limit point (if there is any) of $\{\widehat{B}^{(k)}\}$ and $\{\widehat{\Theta}_\epsilon^{(k)}\}$ by $B^\infty=\lim_{k\to\infty}\widehat{B}^{(k)}$ and $\Theta_\epsilon^\infty=\lim_{k\to\infty}\widehat{\Theta}_\epsilon^{(k)}$, respectively.

Definition 2.1 (stationary point [**?**] pp.479)**.** Define z to be a stationary point of f if $z\in\mathrm{dom}(f)$ and $f'(z;d)\geq 0,\forall$ direction $d=(d_1,\cdots,d_N)$ where d_t is the t^{th} coordinate block.

Definition 2.2 (Regularity [**?**] pp.479)**.** f is regular at $z\in\mathrm{dom}(f)$ if $f'(z;d)\geq 0$ for all $d=(d_1,\cdots,d_N)$ such that

$$f'(z;(0,\cdots,d_t,\cdots,0))\geq 0,\qquad t=1,2,\cdots,N.$$

Definition 2.3 (Coordinate-wise minimum)**.** Define $(B^\infty,\Theta_\epsilon^\infty)$ to be a coordinate-wise minimum if

$$\begin{aligned} f(B^\infty,\Theta_\epsilon) &\geq f(B^\infty,\Theta_\epsilon^\infty),\quad \forall\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2},\\ f(B,\Theta_\epsilon^\infty) &\geq f(B^\infty,\Theta_\epsilon^\infty),\quad \forall B\in\mathbb{R}^{p_1\times p_2}.\end{aligned}$$

Note for our iterative algorithm, we only have two blocks, hence with the above notation, $N=2$.

Remark 2.3. **?**] proved that if the level set $\{x:f(x)\leq f(x^0)\}$ is compact and f satisfies certain conditions [**?** , see Theorem 4.1 (a), (b) and (c) for details], the limit point given by the general block-coordinate descent algorithm (with $N\geq 2$ blocks) is a stationary point of f. However, the conditions given in Theorem 4.1 (a), (b) and (c) are not satisfied for the objective function at hand. Hence, for the problem under consideration, a different strategy is needed to prove convergence of the 2-block alternating algorithm to a stationary point, and the resulting statements hold true for all problems that use a 2-block coordinate descent algorithm.

Since $\mathrm{dom}(f_0)$ is open and f_0 is Gâteaux-differentiable on the $\mathrm{dom}(f_0)$, by **?**] Lemma 3.1, f is regular in the $\mathrm{dom}(f)$. From the discussion on Page 479 of [**?**], we then have:

Fact 1: Every coordinate-wise minimum is a stationary point of f.

The following theorem shows that any limit point $(B^\infty, \Theta_\epsilon^\infty)$ of the iterative algorithm described in Section 2.2.2 is a stationary point of f, as long as all the iterates are within a closed ball around the truth.

Theorem 2.1 (Convergence for fixed design)**.** *Suppose for any fixed realization of* $\mathbf{X}$ *and* $\mathbf{E}$*, the estimates* $\left\{(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})\right\}_{k=1}^\infty$ *obtained by implementing the alternating search step satisfy the following bound for some* $R > 0$ *that only depends on* p_1*,* p_2 *and* n*:*

$$\left|\left|\left|(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}) - (B^\star, \Theta^\star)\right|\right|\right|_F \leq R(p_1, p_2, n), \qquad \forall k \geq 1.$$

Then any limit point $(B^\infty, \Theta_\epsilon^\infty)$ *of the iterative algorithm is a stationary point of* f*.*

Remark 2.4. Recall that in classical parametric statistics, MLE-type asymptotics are derived after establishing that with probability tending to 1 as the sample size n goes to infinity, the likelihood equation has a sequence of roots (hence stationary points of the likelihood function) that converges in probability to the true value. Any such sequence of roots is shown to be asymptotically normal and efficient. Note that such (a sequence of) roots may not be global maximizers since parametric likelihoods are not globally log-concave [see Chapter 6 **?**]. Here we show that the $(B^\infty, \Theta_\epsilon^\infty)$ obtained by the iterative algorithm is a stationary point which satisfies the first-order condition for being a maximizer of the penalized log-likelihood function (which is just the negative of the penalized least-squares function). Moreover, if we let n go to infinity, $(B^\infty, \Theta_\epsilon^\infty)$ converges to the true value in probability (shown in Theorem 2.4), and therefore behaves the same as the sequence of roots in the classical parametric problem alluded to above. Thus, while $(B^\infty, \Theta_\epsilon^\infty)$ may not be the global maximizer, it can, nevertheless, to all intents and purposes, be deemed as the MLE.

Remark 2.5. The above convergence result is based upon solving the optimization problem on the "entire" space, that is, we don't restrict B to live in any subspace. However, when actually implementing the proposed computational procedure, the optimization of the B coordinate is restricted to $\mathcal{B}_1 \times \cdots \times \mathcal{B}_{p_2}$ (as defined in eqn (2.4)). It should be noted that the same convergence property still holds, since for all $k \geq 1$, the following bound holds, for some $R' > 0$:

$$\left|\left|\left|(\widehat{B}^{(k)}_{\text{restricted}}, \widehat{\Theta}_\epsilon^{(k)}) - (B^\star, \Theta_\epsilon^\star)\right|\right|\right|_{\text{F}} \leq R'(p_1, p_2, n). \tag{2.11}$$

Consequently, the rest of the derivation in Theorem 2.1 follows, leading to the convergence property. The bound in eqn (2.11) will be shown at the end of Section 2.3.2.

2.3.2 Estimation consistency.

In this subsection, we show that given a random realization of X and E, with high probability, the sequence $\big\{(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})\big\}_{k=1}^{\infty}$ lies in a non-expanding ball around $(B^\star, \Theta_\epsilon^\star)$, thus satisfying the condition of Theorem 2.1 for convergence of the alternating algorithm.

It should be noted that for the alternating search procedure, we restrict our estimation on a subspace identified by the screening step. However, for the remaining of this subsection, the main propositions and theorems are based on the procedure without such restriction, i.e., we consider "generic" regressions on the entire space of dimension $p_1 \times p_2$. Notwithstanding, it can be easily shown that the theoretical results for the regression parameters on a restricted domain follow easily from the generic case, as explained in Remark 2.9.

Before providing the details of the main theorem statements and proofs, we first introduce additional notations. Let $\beta = \mathrm{vec}(B)$ be the vectorized version of the regression coefficient matrix. Correspondingly, we have $\widehat{\beta}^{(k)} = \mathrm{vec}(\widehat{B}^{(k)})$ and $\beta^\star = \mathrm{vec}(B^\star)$. Moreover, we drop the superscripts and use $\widehat{\beta}$ and $\widehat{\Theta}_\epsilon$ to denote the generic estimators given by (2.12) and (2.13), as opposed to those obtained in any specific iteration:

$$\widehat{\beta} \;:=\; \underset{\beta\in\mathbb{R}^{p_1p_2}}{\arg\min}\; \big\{ -2\beta^\top\widehat{\gamma} + \beta^\top\widehat{\Gamma}\beta + \lambda_n\|\beta\|_1\big\}, \tag{2.12}$$

$$\widehat{\Theta}_\epsilon \;:=\; \underset{\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}{\arg\min}\; \big\{ -\log\det\Theta_\epsilon + \mathrm{tr}\big(\widehat{S}\Theta_\epsilon\big) + \rho_n\|\Theta_\epsilon\|_{1,\mathrm{off}}\big\}, \tag{2.13}$$

where

$$\widehat{\Gamma} = \big(\widehat{\Theta}_\epsilon \otimes \frac{\mathbf{X}^\top\mathbf{X}}{n}\big),\ \widehat{\gamma} = \big(\widehat{\Theta}_\epsilon \otimes \mathbf{X}^\top\big)\mathrm{vec}(\mathbf{Y})/n,\ \widehat{S} = \frac{1}{n}\big(\mathbf{Y} - \mathbf{X}\widehat{B}\big)^\top\big(\mathbf{Y} - \mathbf{X}\widehat{B}\big).$$

Remark 2.6. As opposed to (2.12) and (2.13), if $\widehat{\gamma}$ and $\widehat{\Gamma}$ are replaced by plugging in the true values of the parameters, the two problems in (2.12)) and (2.13) become

$$\bar{\beta} \;:=\; \underset{\beta\in\mathbb{R}^{p_1p_2}}{\arg\min}\; \big\{ -2\beta^\top\bar{\gamma} + \beta^\top\bar{\Gamma}\beta + \lambda_n\|\beta\|_1\big\}, \tag{2.14}$$

$$\bar{\Theta}_\epsilon \;:=\; \underset{\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}{\arg\min}\; \big\{ -\log\det\Theta_\epsilon + \mathrm{tr}\big(S\Theta_\epsilon\big) + \rho_n\|\Theta_\epsilon\|_{1,\mathrm{off}}\big\}, \tag{2.15}$$

where

$$\bar{\Gamma} = \big(\Theta_\epsilon^\star \otimes \frac{\mathbf{X}^\top\mathbf{X}}{n}\big),\ \ \bar{\gamma} = \big(\Theta_\epsilon^\star \otimes \mathbf{X}^\top\big)\mathrm{vec}(\mathbf{Y})/n,\ \ S = \frac{1}{n}\big(\mathbf{Y} - \mathbf{X}B^\star\big)^\top\big(\mathbf{Y} - \mathbf{X}B^\star\big) =: \widehat{\Sigma}_\epsilon.$$

In (2.14), we obtain β using a penalized maximum likelihood regression estimate, and (2.15) corresponds to the generic setting for using the graphical Lasso. A key difference between

the estimation problems in (2.12) and (2.13) versus those in (2.14) and (2.15) is that to obtain $\widehat{\beta}$ and $\widehat{\Theta}_\epsilon$ we use *estimated quantities* rather than the raw data. This is exactly how we implement our iterative algorithm, namely, we obtain $\widehat{\beta}^{(k)}$ using $\widehat{S}^{(k-1)}$ as a surrogate for the sample covariance of the true error (which is unavailable), then estimate $\widehat{\Theta}_\epsilon^{(k)}$ using the information in $\widehat{\beta}^{(k)}$. This adds complication for establishing the consistency results. Original consistency results for the estimation problem in (2.14) and (2.15) are available in **?**] and **?**], respectively. Here we borrow ideas from corresponding theorems in those two papers, but need to tackle concentration bounds of relevant quantities with additional care. This part of the result and its proof are shown in Theorem 2.4.

As a road map toward our desired result established in Theorem 2.4, we first show in Theorem 2.2 that for any fixed realization of $\mathbf{X}$ and $\mathbf{E}$, under a number of conditions on (or related to) $\mathbf{X}$ and $\mathbf{E}$, when $\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star\|_\infty$ is small (up to a certain order), the error of $\widehat{\beta}$ is well-bounded. We then verify in Propositions 2.1 and 2.2 that for random $\mathbf{X}$ and $\mathbf{E}$, the above-mentioned conditions hold with high probability. Similarly in Theorem 2.3, we show that for fixed realizations in $\mathbf{X}$ and $\mathbf{E}$, under certain conditions (verified for random $\mathbf{X}$ and $\mathbf{E}$ in Proposition 2.3), the error of $\widehat{\Theta}_\epsilon$ is also well-bounded, given $\|\widehat{\beta} - \beta^\star\|_1$ being small. Finally in Theorem 2.4, we show that for random $\mathbf{X}$ and $\mathbf{E}$, with high probability, the iterative algorithm gives $\{(\widehat{\beta}^{(k)}, \Theta_\epsilon^{(k)})\}$ that lies in a small ball centered at $(\beta^\star, \Theta_\epsilon^\star)$, whose radius depends on p_1, p_2, n and the sparsity levels.

Next, for establishing the main propositions and theorems, we introduce some additional notations.

- Sparsity level of $\beta^\star$: $s^{\star\star} := \|\beta^\star\|_0 = \sum_{j=1}^{p_2} \|B_j^\star\|_0 = \sum_{j=1}^{p_2} s_j^\star$. As a reminder of the previous notation, we have $s^\star = \max_{j=1,\cdots,p_2} s_j^\star$.
- True edge set of $\Theta_\epsilon^\star$: $S_\epsilon^\star$, and let $s_\epsilon^\star := |S_\epsilon^\star|$ be its cardinality.
- Hessian of the log-determinant barrier $\log\det\Theta$ evaluated at $\Theta_\epsilon^\star$:

$$H^\star := \frac{\mathrm{d}^2}{\mathrm{d}\Theta^2} \log\Theta\big|_{\Theta_\epsilon^\star} = \Theta_\epsilon^{\star -1} \otimes \Theta_\epsilon^{\star -1}.$$

- Matrix infinity norm of the true error covariance matrix $\Sigma_\epsilon^\star$:

$$\kappa_{\Sigma_\epsilon^\star} := |||\Sigma_\epsilon^\star|||_\infty = \max_{i=1,2,\cdots,p_2} \sum_{j=1}^{p_2} |\Sigma_{\epsilon,ij}^\star|.$$

- Matrix infinity norm of the Hessian restricted to the true edge set:

$$\kappa_{H^\star} := \Big|\Big|\Big|(H_{S_\epsilon^\star S_\epsilon^\star}^\star)\Big|\Big|\Big|_\infty = \max_{i=1,2,\cdots,p_2} \sum_{j=1}^{p_2} \big|H_{S_\epsilon^\star S_\epsilon^\star,ij}^\star\big|.$$

– Maximum degree of $\Theta_\epsilon^\star$: $d := \max_{i=1,2,\cdots,p_2} \|\Theta^\star_{\epsilon,i\cdot}\|_0$.
– We write $A \gtrsim B$ if there exists some absolute constant c that is independent of the model parameters such that $A \geq cB$.

Definition 2.4 (Incoherence condition [**?**])**.** $\Theta_\epsilon^\star$ satisfies the incoherence condition if:

$$\max_{e \in (S_\epsilon^\star)^c} \|H^\star_{eS_\epsilon^\star}(H^\star_{S_\epsilon^\star S_\epsilon^\star})^{-1}\|_1 \leq 1 - \xi, \quad \text{for some } \xi \in (0,1).$$

Definition 2.5 (Restricted eigenvalue (RE) condition [**?**])**.** A symmetric matrix $A \in \mathbb{R}^{m\times m}$ satisfies the RE condition with curvature $\varphi > 0$ and tolerance $\phi > 0$, denoted by $A \sim \mathrm{RE}(\varphi, \phi)$ if

$$\theta' A\theta \geq \varphi\|\theta\|^2 - \phi\|\theta\|_1^2, \quad \forall \theta \in \mathbb{R}^m.$$

Definition 2.6 (Diagonal dominance)**.** A matrix $A \in \mathbb{R}^{m\times m}$ is strictly diagonally dominant if

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|, \quad \forall i = 1, \cdots, m.$$

Based on the model in Section 2.2.1, since we are assuming $X = (X_1, \cdots, X_{p_1})^\top$ and $\epsilon = (\epsilon_1, \cdots, \epsilon_{p_2})^\top$ come from zero-mean Gaussian distributions, it follows that X and ϵ are zero-mean sub-Gaussian random vectors with parameters (Σ_X, σ_x^2) and $(\Sigma_\epsilon^\star, \sigma_\epsilon^2)$, respectively. Moreover, throughout this section, all results are based on the assumption that $\Theta_\epsilon^\star$ is diagonally dominant.

Remark 2.7. Before moving on to the main statements of Theorem 2.2, we would like to point out that with a slight abuse of notation, for Theorem 2.2 and its related propositions and corollaries, the statements and analyses are based on equation (2.12) only, with *any deterministic symmetric matrix* $\widehat{\Theta}_\epsilon$ within a small ball around $\Theta_\epsilon^\star$. Similarly in Theorem 2.3, Proposition 2.3 and Corollary 2.2, the analyses are based on equation (2.13) only, for *any given deterministic* $\widehat{\beta}$ within a small ball around $\beta^\star$. The randomness of $\widehat{\beta}$ and $\widehat{\Theta}_\epsilon$ during the iterative procedure will be taken into consideration comprehensively in Theorem 2.4.

Theorem 2.2 (Error bound for $\widehat{\beta}$ with fixed realizations of $\mathbf{X}$ and $\mathbf{E}$)**.** *Consider* $\widehat{\beta}$ *given by (2.12). For any fixed pair of realizations of* $\mathbf{X}$ *and* $\mathbf{E}$ *, assume the following:*

A1. $\widehat{\Theta}_\epsilon$ *is a deterministic matrix satisfying the bound* $\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star\|_\infty \leq \nu_\Theta$ *where* $\nu_\Theta = \eta_\Theta\left(\sqrt{\frac{\log p_2}{n}}\right)$ *and* η_Θ *is some constant depending only on* $\Theta_\epsilon^\star$*;*

A2. $\widehat{\Gamma} \sim RE(\varphi, \phi)$*, with* $s^{\star\star}\phi \leq \varphi/32$*;*

A3. $(\widehat{\Gamma}, \widehat{\gamma})$ *satisfies the deviation bound*

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^\star\|_\infty \leq \mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1 p_2)}{n}},$$

where $\mathbb{Q}(\nu_\Theta)$ *is some quantity depending on* ν_Θ*.*

Then, for any $\lambda_n \geq 4\mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1 p_2)}{n}}$*, the following bound holds:*

$$\|\widehat{\beta} - \beta^\star\|_1 \leq 64 s^{\star\star}\lambda_n/\varphi.$$

The following two propositions verify the RE condition for $\widehat{\Gamma}$ and deviation bound for $(\widehat{\Gamma}, \widehat{\gamma})$ hold with high probability for a random pair $(\mathbf{X}, \mathbf{E})$, given any symmetric, matrix $\widehat{\Theta}_\epsilon$ satisfying (A1).

Proposition 2.1 (Verification of RE condition for random $\mathbf{X}$ and $\mathbf{E}$)**.** *Consider any deterministic matrix* $\widehat{\Theta}_\epsilon$ *satisfying (A1). Let the sample size satisfy* $n \gtrsim \max\{s^{\star\star}\log p_1, d^2 \log p_2\}$*. With probability at least* $1 - 2\exp(-c_3 n)$ *for some constant* $c_3 > 0$*,* $\widehat{\Gamma}$ *satisfies the following RE condition:*

$$\widehat{\Gamma} := \widehat{\Theta}_\epsilon \otimes (\mathbf{X}^\top \mathbf{X}/n) \sim RE\Big(\varphi^\star(\min_i \psi^i - d\nu_\Theta), \phi^\star \max_i(\psi^i + d\nu_\Theta)\Big),$$

where $\varphi^\star = \frac{\Lambda_{\min}(\Sigma_X^\star)}{2}$*,* $\phi^\star = (\varphi^\star \log p_1)/n$*, and* ψ^i *is defined as:*

$$\psi^i := \sigma_\epsilon^{ii} - \sum_{j \neq i}^{p_2} \sigma_\epsilon^{ij},$$

where σ_ϵ^{ij}*'s denote the entries in* $\Theta_\epsilon^\star$ *hence* ψ^i *is the gap between its diagonal entry and the sum of off-diagonal entries for row* i*.*

Proposition 2.2 (Deviation bound for $(\widehat{\Gamma}, \widehat{\gamma})$ for random $\mathbf{X}$ and $\mathbf{E}$)**.** *Consider any deterministic matrix* $\widehat{\Theta}_\epsilon$ *satisfying (A1). Let sample size* n *satisfy* $n \gtrsim \log(p_1 p_2)$*. With probability at least*

$$1 - 12c_1 \exp\{-(c_2^2 - 1)\log(p_1 p_2)\} \quad \text{for some } c_1 > 0, c_2 > 1,$$

the following bound holds:

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^\star\|_\infty = \frac{1}{n}\|\mathbf{X}^\top \mathbf{E}\widehat{\Theta}_\epsilon\|_\infty \leq \mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1 p_2)}{n}},$$

where

$$\mathbb{Q}(\nu_\Theta) = c_2\Big\{d\nu_\Theta\Big[\Lambda_{\max}(\Sigma_X^\star)\Lambda_{\max}(\Sigma_\epsilon^\star)\Big]^{1/2} + \Big[\frac{\Lambda_{\max}(\Sigma_X^\star)}{\Lambda_{\min}(\Sigma_\epsilon^\star)}\Big]^{1/2}\Big\}. \tag{2.16}$$

Remark 2.8. In Proposition 2.1, the quantity $d^2 \log p_2$ that shows up in the sample size requirement is a result of $\nu_\Theta = O(\sqrt{\log p_2/n})$, which is the common order of error in a generic graphical Lasso problem. Hence here we explicitly list it for the purpose of showing results for the generic graphical Lasso estimation problem. In our iterative algorithm, the order of $\nu_\Theta^{(k)}$ depends on the relative order of p_1 and p_2, which may potentially make the sample size requirement more stringent. This will be discussed in more detail in the proof of Theorem 2.4.

Given the results in Theorem 2.2, Proposition 2.1 and Proposition 2.2, next we provide Corollary 2.1, which gives the error bound for $\widehat{\beta}$ for random realizations of $\mathbf{X}$ and $\mathbf{E}$.

Corollary 2.1 (Error Bound for $\widehat{\beta}$ for random $\mathbf{X}$ and $\mathbf{E}$)**.** *Consider any deterministic $\widehat{\Theta}_\epsilon$ satisfying the following element-wise ℓ_∞-bound:*

$$\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star\|_\infty \le \nu_\Theta,$$

with $\nu_\Theta = \eta_\Theta\sqrt{\frac{\log p_2}{n}}$. Then for sample size $n \gtrsim \log(p_1p_2)$ and for any regularization parameter $\lambda_n \ge 4\mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1p_2)}{n}}$ with the expression of $\mathbb{Q}(\cdot)$ given in (2.16)*, there exists $c_1 > 0$ and $c_2 > 1$ such that with probability at least:*

$$1 - 12c_1\exp\{-(c_2^2-1)\log(p_1p_2)\} - 2\exp(-c_3 n),$$

the following bound holds:

$$\|\widehat{\beta} - \beta^\star\|_1 \le 64 s^{\star\star}\lambda_n/\varphi, \tag{2.17}$$

where $\varphi = \frac{1}{2}\Lambda_{\min}(\Sigma_\epsilon^\star)(\min_i \psi^i - d\nu_\Theta)$.

Next, we move onto analyzing the error bound of the other component, for a fixed given $\widehat{\beta}$.

Theorem 2.3 (Error bound for $\widehat{\Theta}_\epsilon$ for fixed realizations of $\mathbf{X}$ and $\mathbf{E}$)**.** *Consider $\widehat{\Theta}_\epsilon$ given by* (2.13)*. For any fixed pair of realization $(\mathbf{X}, \mathbf{E})$, assume the following:*

- *B1. $\widehat{\beta}$ is a deterministic vector satisfying $\|\widehat{\beta} - \beta^\star\|_1 \le \nu_\beta$, where $\nu_\beta = \eta_\beta\Big(\sqrt{\frac{\log(p_1p_2)}{n}}\Big)$, with$\eta_\beta$ being some constant depending only on $\beta^\star$;*
- *B2. $\|\widehat{S} - \Sigma_\epsilon^\star\|_\infty \le g(\nu_\beta)$ where $\widehat{S} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{B})^\top(\mathbf{Y} - \mathbf{X}\widehat{B})$, and $g(\nu_\beta)$ is some quantity depending on ν_β;*

B3. Incoherence condition holds for $\Theta_\epsilon^\star$.

Then, for $\rho_n = (8/\xi)g(\nu_\beta)$ *and sample size* n *satisfying* $n \gtrsim \log(p_1 p_2)$, *the following error bound for* $\widehat{\Theta}_\epsilon$ *holds:*

$$\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star\|_\infty \leq \{2(1+8\xi^{-1})\kappa_{H^\star}\} g(\nu_\beta), \tag{2.18}$$

where ξ *is the incoherence parameter as defined in Definition 2.4.*

Proposition 2.3 gives an explicit expression for $g(\nu_\beta)$ under condition (B1). Specifically, it shows how well $\widehat{S}$ concentrates around $\Sigma_\epsilon^\star$ for random $\mathbf{X}$ and $\mathbf{E}$, given some $\widehat{B}$ exhibiting a small error from its true value (or $\widehat{\beta}$, equivalently),

Proposition 2.3. *Consider any deterministic* $\widehat{\beta}$ *satisfying (B1). Then for sample size* n *satisfying* $n \gtrsim \log(p_1 p_2)$, *with probability at least*

$$1 - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 6c_1 \exp\{-(c_2^2 - 1)\log(p_1 p_2)\}, \quad \textit{for some } c_1 > 0, c_2 > 1, \tau_1, \tau_2 > 2,$$

the following bound holds:

$$\|\widehat{S} - \Sigma_\epsilon^\star\|_\infty \leq g(\nu_\beta),$$

where

$$\begin{aligned} g(\nu_\beta) = &\sqrt{\frac{\log 4 + \tau_2 \log p_2}{c_\epsilon^\star n}} + \nu_\beta^2 \Big(\sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^\star n}} + \max_i(\Sigma_{X,ii}^\star) \Big) \\ &+ 2c_2 \nu_\beta \Big[\Lambda_{\max}(\Sigma_X^\star) \Lambda_{\max}(\Sigma_\epsilon^\star) \Big]^{1/2} \sqrt{\frac{\log(p_1 p_2)}{n}}, \end{aligned} \tag{2.19}$$

$c_\epsilon^\star$ *and* $c_X^\star$ *are population quantities given in* (A.30) *and* (A.35), *respectively.*

Given Theorem 2.3 and Proposition 2.3, we provide Corollary 2.2, which gives the error bound for $\widehat{\Theta}_\epsilon$ for random realizations of $\mathbf{X}$ and $\mathbf{E}$:

Corollary 2.2 (Error bound for $\widehat{\Theta}$ for random $\mathbf{X}$ and $\mathbf{E}$)**.** *Consider any deterministic* $\widehat{\beta}$ *satisfying the following bound*

$$\|\widehat{\beta} - \beta^\star\|_1 \leq \nu_\beta,$$

with $\nu_\beta = \eta_\beta \sqrt{\frac{\log(p_1 p_2)}{n}}$. *Also suppose the incoherence condition (B3) is satisfied. Then, for sample size* $n \gtrsim \log(p_1 p_2)$ *and regularization parameter* $\rho_n = (8/\xi)g(\nu_\beta)$ *with* $g(\nu_\beta)$ *given in* (2.19), *with probability at least*

$$1 - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 6c_1 \exp\{-(c_2^2 - 1)\log(p_1 p_2)\}, \quad \textit{for some } c_1 > 0, c_2 > 1, \tau_1, \tau_2 > 2,$$

the following bound holds:

$$\|\widehat{\Theta}_\epsilon - \Theta^\star_\epsilon\|_\infty \leq \{2(1+8\xi^{-1})\kappa_{H^\star}\} g(\nu_\beta).$$

After providing the error bound for (2.12) and (2.13), in Theorem 2.4 we establish that with high probability, the error of the sequence of estimates obtained in the alternating search step of the algorithm described in Section 2.2.2 is *uniformly* bounded; that is, the sequence of estimates lie in a non-expanding ball around the true value of the parameters uniformly with a radius that does not depend on the iteration number k.

Theorem 2.4 (Error bound for $\{\widehat{\beta}^{(k)}\}$ and $\{\widehat{\Theta}^{(k)}_\epsilon\}$)**.** *Consider the iterative algorithm given in Section 2.2.2 that gives rise to sequences of $\{\widehat{\beta}^{(k)}\}$ and $\{\widehat{\Theta}^{(k)}_\epsilon\}$ alternately. For random realization of $\mathbf{X}$ and $\mathbf{E}$, we assume the following:*

C1. The incoherence condition holds for $\Theta^\star_\epsilon$.

C2. $\Theta^\star_\epsilon$ is diagonally dominant.

C3. The maximum sparsity level for all p_2 regression $s^\star$ satisfies $s^\star = o(n/\log p_1)$.

(I) For sample size satisfying $n \gtrsim \log(p_1p_2)$, there exist constants $c_1 > 0, c_2 > 1, c_3 > 0$ such that for any

$$\lambda^0_n \geq 4c_2\Big[\Lambda_{\max}(\Sigma^\star_X)\Lambda_{\max}(\Sigma^\star_\epsilon)\Big]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}},$$

with probability at least $1 - 2\exp(-c_3n) - 6c_1\exp\{-(c_2^2-1)\log(p_1p_2)\}$, the initial estimate $\widehat{\beta}^{(0)} := vec(\widehat{B}^{(0)})$ satisfies the following bound

$$\|\widehat{\beta}^{(0)} - \beta^\star\|_1 \leq 64 s^{\star\star}\lambda^0_n/\varphi^\star := \nu^{(0)}_\beta, \tag{2.20}$$

where $\varphi^\star = \Lambda_{\min}(\Sigma^\star_X)/2$. Moreover, by choosing $\rho^0_n = (\frac{8}{\xi})g(\nu^{(0)}_\beta)$ where the expression for $g(\cdot)$ is given in (2.19)*, with probability at least*

$$1 - 1/p_1^{\tau_1-2} - 1/p_2^{\tau_2-2} - 2\exp(-c_3n) - 6c_1\exp\{-(c_2^2-1)\log(p_1p_2)\}, \quad \text{for some } \tau_1, \tau_2 > 2,$$

the following bound holds:

$$\|\widehat{\Theta}^{(0)}_\epsilon - \Theta^\star_\epsilon\|_\infty \leq \{2(1+8\xi^{-1})\kappa_{H^\star}\} g(\nu^{(0)}_\beta) := \nu^{(0)}_\Theta. \tag{2.21}$$

(II) For sample size satisfying $n \gtrsim d^2\log(p_1p_2)$, for any iteration $k \geq 1$, with probability at least

$$1 - 1/p_1^{\tau_1-2} - 1/p_2^{\tau_2-2} - 12c_1\exp\{-(c_2^2-1)\log(p_1p_2)\} - 2\exp(-c_3n),$$

the following bounds hold for all $\widehat{\beta}^{(k)}$ *and* $\widehat{\Theta}_\epsilon^{(k)}$*:*

$$\|\widehat{\beta}^{(k)} - \beta^\star\|_1 \leq C_\beta\Big(s^{\star\star}\sqrt{\frac{\log(p_1p_2)}{n}}\Big), \qquad \|\widehat{\Theta}_\epsilon^{(k)} - \Theta_\epsilon^\star\|_\infty \leq C_\Theta\Big(\sqrt{\frac{\log(p_1p_2)}{n}}\Big),$$

where $s^{\star\star}$ *is the sparsity of* $\beta^\star$*,* C_β *and* C_Θ *are constants depending only on* $\beta^\star$ *and* $\Theta_\epsilon^\star$*, respectively.*

As a direct result of Proposition 1 in **?**] and Corollary 3 in **?**], the following bound also holds:

Corollary 2.3. *Under the same set of conditions C1, C2 and C3 in Theorem 2.4, there exists* $\tau_1, \tau_2 > 2$*,* $c_1 > 0, c_2 > 1, c_3 > 0$ *and constants* C'_β *and* C'_Θ *such that for all iterations* k*, the following bound holds:*

$$\|\widehat{\beta}^{(k)} - \beta^\star\|_2 \leq C'_\beta\Big(\sqrt{\frac{s^{\star\star}\log(p_1p_2)}{n}}\Big), \quad \|\!|\widehat{\Theta}_\epsilon^{(k)} - \Theta_\epsilon^\star|\!\|_F \leq C'_\Theta\Big(\sqrt{\frac{(s_\epsilon^\star + p_2)\log(p_1p_2)}{n}}\Big),$$

with probability at least

$$1 - 1/p_1^{\tau_1-2} - 1/p_2^{\tau_2-2} - 12c_1\exp\{-(c_2^2-1)\log(p_1p_2)\} - 2\exp(-c_3n),$$

where $s^{\star\star}$ *and* $s_\epsilon^\star$ *are the sparsity for* $\beta^\star$ *and* $\Theta_\epsilon^\star$*, respectively.*

Remark 2.9. As mentioned earlier in this subsection, the actual implementation of the alternating search step is restricted to a subspace of $\mathbb{R}^{p_1\times p_2}$. Next, we outline the corresponding theoretical results for this specific scenario in which for each regression j, some *fixed superset* of the indices of true covariates is given, and the regressions are restricted to these supersets, respectively. Note that we need to make sure that the restricted subspace contains all the true covariates for the results below to be valid.

Let S_j denote the given *fixed superset* for each regression j, and we consider regressing the response on $\mathbf{X}_{S_j}$. We use $\widehat{\beta}_{\mathrm{R}}^{(k)}$ to denote the corresponding vectorized estimator of iteration k, that is,

$$\widehat{\beta}_{\mathrm{R}}^{(k)} = (\widehat{B}_{1,\mathrm{Restricted}}^{(k)\prime}, \cdots, \widehat{B}_{p_2,\mathrm{Restricted}}^{(k)\prime})^\top,$$

where $\widehat{B}_{j,\mathrm{Restricted}}^{(k)\prime}$ is obtained by doing the regression in (2.7), however with the indices of covariates restricted to S_j. Also, we let $\beta_{\mathrm{R}}^\star$ be the corresponding true value of $\widehat{\beta}_{\mathrm{R}}^{(k)}$. Note that it always holds that

$$\|\widehat{\beta}_{\mathrm{R}}^{(k)} - \beta_{\mathrm{R}}^\star\| = \|\widehat{\beta}^{(k)} - \beta^\star\|.$$

Now let

$$\bar{S} = \bigcup_{j \in \{1, \cdots, p_2\}} S_j,$$

and let $\bar{s}$ be its cardinality. It can be shown that the best achievable error bound for $\widehat{\beta}_{\mathrm{R}}^{(k)}$ is identical to $\widehat{\beta}_{\bar{S}}^{(k)}$, where $\widehat{\beta}_{\bar{S}}^{(k)}$ is obtained by considering covariates $\mathbf{X}_{\bar{S}}$ for all p_2 regressions, instead of the entire $\mathbf{X}$. For this specific reason, formally, we state the theoretical results for the case where we consider regressing on $\mathbf{X}_{\bar{S}}$, which is almost identical to the generic case.

Suppose conditions C1, C2 and C3 in Theorem 4 hold, then there exists constants $c_1 > 0, c_2 > 1, c_3 > 0, \tau_1 > 2, \tau_2 > 2$ such that: (I) for sample size satisfying $n \gtrsim \log(\bar{s}p_2)$, w.p. at least $1 - 2\exp(-c_3 n) - 6c_1 \exp\{-(c_2^2 - 1)\log(\bar{s}p_2)\}$, for any

$$\lambda_n^0 \geq 4c_2 \Big[\Lambda_{\max}(\Sigma^\star_{X_{\bar{S}}})\Lambda_{\max}(\Sigma^\star_\epsilon)\Big]^{1/2} \sqrt{\frac{\log(\bar{s}p_2)}{n}},$$

the initial estimate $\widehat{\beta}_{\bar{S}}^{(0)}$ satisfies the following bound:

$$\|\widehat{\beta}_{\bar{S}}^{(0)} - \beta^\star_{\bar{S}}\|_1 \leq 64 s^{\star\star} \lambda_n^0 / \varphi^\star_{\bar{S}} := \nu^{(0)}_{\beta_{\bar{S}}},$$

where $\varphi^\star_{\bar{S}} = \Lambda_{\min}(\Sigma^\star_{X_{\bar{S}}})/2$ and $\Sigma^\star_{X_{\bar{S}}}$ is the population covariance of the random vector X restricted to $\bar{S}$. Moreover, by choosing $\rho_n^0 = (\frac{8}{\xi}) g(\nu^{(0)}_{\beta_{\bar{S}}})$ where the expression for $g(\cdot)$ is given in (2.19), with probability at least

$$1 - 1/\bar{s}^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 2\exp(-c_3 n) - 6c_1 \exp\{-(c_2^2 - 1)\log(\bar{s}p_2)\},$$

the following bound holds:

$$\|\widehat{\Theta}_\epsilon^{(0)} - \Theta^\star_\epsilon\|_\infty \leq \{2(1 + 8\xi^{-1})\kappa_{H^\star}\} g(\nu^{(0)}_{\beta_{\bar{S}}}) := \nu^{(0)}_\Theta.$$

(II) For sample size satisfying $n \gtrsim d^2 \log(\bar{s}p_2)$, for any iteration $k \geq 1$, with probability at least

$$1 - 1/\bar{s}^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 12c_1 \exp\{-(c_2^2 > 1)\log(\bar{s}p_2)\} - 2\exp(-c_3 n),$$

the following bound hold for all $\widehat{\beta}_{\bar{S}}^{(k)}$ and $\widehat{\Theta}_\epsilon^{(k)}$:

$$\|\widehat{\beta}_{\bar{S}}^{(k)} - \beta^\star\|_1 \leq C_\beta \Big(s^{\star\star}\sqrt{\tfrac{\log(\bar{s}p_2)}{n}}\Big), \quad \|\widehat{\beta}_{\bar{S}}^{(k)} - \beta^\star\|_2 \leq C'_\beta \Big(\sqrt{\frac{s^{\star\star}\log(\bar{s}p_2)}{n}}\Big),$$
$$\|\widehat{\Theta}_\epsilon^{(k)} - \Theta^\star_\epsilon\|_\infty \leq C_\Theta \Big(\sqrt{\tfrac{\log(\bar{s}p_2)}{n}}\Big), \quad \|\|\widehat{\Theta}_\epsilon^{(k)} - \Theta^\star_\epsilon\|\|_{\mathrm{F}} \leq C'_\Theta \Big(\sqrt{\frac{(s^\star_\epsilon + p_2)\log(\bar{s}p_2)}{n}}\Big),$$

where $s^{\star\star}$ is the sparsity of $\beta^\star$, C_β, C'_β, C_Θ and C'_θ are all constants that do not depend on $n, \bar{S}, p_2$.

2.3.3 Family-wise error rate control of the screening step.

As mentioned in the Introduction, for the iterative algorithm to work effectively, it is crucial to initialize from points that are close to the true parameters. Our screening step provides such guarantees *asymptotically*. Based on the screening step described in Section 2.2.2, initial estimates for each column of the regression matrix are obtained by Lasso or Ridge regression with the support set restricted to the one identified by the screening step. It is desirable for the screening step to correctly identify the true support set. In particular, we would like to retain as many true positive predictor variables as possible without discovering too many false positive ones. The following theorem states that as long as $\log(p_1 p_2)/n = o(1)$ and the sparsity is not beyond a specified level, the screening step will be able to recover all true positive predictors, while keeping the family-wise type I error under control.

Theorem 2.5. *Let $S_j^\star$ denote the true support set of the jth regression and $s_j^\star$ be its cardinality. Suppose that $\log(p_1 p_2)/n \to 0$ and the following condition for sparsity holds:*

$$\max\{s_j^\star, j = 1, \cdots, p_2\} = o(\sqrt{n}/\log p_1).$$

Then, the screening step described in Section 2.2 will correctly recover $S_j^\star$ for all $j = 1, \cdots, p_2$ with probability approaching to 1, while keeping the family-wise type I error rate under the pre-specified level α.

Remark 2.10. The specified level for sparsity is necessary for the de-biased Lasso procedure in **?**] to produce unbiased estimates for the regression coefficients. In terms of support recovery for the screening step, with $\log(p_1 p_2)/n = o(1)$, we only require $s^\star = o(p_1)$, which is much weaker and easily satisfied.

The following corollary connects the screening step with the alternating search step, under the discussed asymptotic regime :

Corollary 2.4. *Consider the model setup given in Section 2.2.1. Let $s^\star$ denote the maximum sparsity for all $B_j^\star, j = 2, \cdots, p_2$, and d denote the maximum degree of $\Theta_\epsilon^\star$. Also, let $s^{\star\star}$ denote the sparsity for $\beta^\star$ and $s_\epsilon^\star$ denote the sparsity for $\Theta_\epsilon^\star$. Assume there exist positive constants $c_{s^\star}, c_{s^{\star\star}}, c_d, c_{\bar{s}}, c_{p_2}$ satisfying*

$$0 < c_{s^\star} + c_{\bar{s}} < 1/2; \quad 0 < c_{s^{\star\star}} + c_{\bar{s}} < 1; \quad 0 < 2c_d + c_{\bar{s}} < 1; \quad 0 < \max\{c_{s_\epsilon^\star}, c_{p_2}\} + c_{\bar{s}} < 1$$

such that

$$s^\star = O(n^{c_s}); \quad s^{\star\star} = O(n^{c_{s^{\star\star}}}); \quad s^\star_\epsilon = O(n^{c_{s^\star_\epsilon}}); \quad d = O(n^{c_d}); \quad \bar{s} = O(e^{n^{c_{p_1}}}); \quad p_2 = O(n^{c_{p_2}}).$$

As $n \to \infty$,

$$\mathbb{P}\Big(\{\textit{The screening step correctly recovers the true support set for all } B_j, j = 1, \cdots, p\}\Big) \to 1,$$

and for all iterations k*:*

$$\max_{k \geq 1} \Big|\Big|\Big|(\widehat{\beta}_{\mathrm{R}}, \widehat{\Theta}^{(k)}_\epsilon) - (\beta^\star_{\mathrm{R}}, \Theta^\star_\epsilon)\Big|\Big|\Big|_F \xrightarrow{p} 0.$$

The proof of this corollary follows along the same lines as Theorem 2.4, and we leave the details to the reader.

2.3.4 Estimation error and identifiability.

In this subsection, we discuss in detail the conditions needed for the parameters in our multi-layered network to be identifiable (estimable). We focus the presentation for ease of exposition on a three-layer network and then discuss the general M-layer case.

Consider a 3-layer graphical model. Let $\widetilde{X} = [(\vec{X}^1)^\top, (\vec{X}^2)^\top]^\top$ be the (p_1+p_2) dimensional random vector, which represents the "super-layer" on which we regress $\vec{X}^3$ to estimate B^{13}, B^{23} and Σ^3. As shown in Theorem 2.2, the estimation error for $\widehat{\beta}$ takes the following form:

$$\|\widehat{\beta} - \beta^\star\|_1 \leq 64 s^{\star\star} \lambda_n / \varphi,$$

where φ is the curvature parameter for RE condition that scales with $\Lambda_{\min}(\Sigma_{\widetilde{X}})$ (see Proposition 2.1). Therefore, the error of estimating these regression parameters is higher when $\Lambda_{\min}(\Sigma_{\widetilde{X}})$ is smaller. In this section, we derive a lower bound on this quantity to demonstrate how the estimation error depends on the underlying structure of the graph.

For the undirected subgraph within a layer k, we denote its maximum node capacity by $\mathbf{v}(\Theta^k) := \max_{1 \leq i \leq p_k} \sum_{j=1}^{p_k} |\Theta_{ij}|$. For the directed bipartite subgraph consisting of Layer $s \to t$ edges ($s < t$), we similarly define the maximum incoming and outgoing node capacities by $\mathbf{v}_{\mathrm{in}}(B^{st}) := \max_{1 \leq j \leq p_t} \sum_{i=1}^{p_s} |B^{st}_{ij}|$ and $\mathbf{v}_{\mathrm{out}}(B^{st}) := \max_{1 \leq i \leq p_s} \sum_{j=1}^{p_t} |B^{st}_{ij}|$. The following proposition establishes the lower bound in terms of these node capacities

Proposition 2.4. *The following lower bound holds for the minimum eigenvalue of* $\Sigma_{\widetilde{X}}$*:*

$$\Lambda_{\min}(\Sigma_{\widetilde{X}}) \geq \mathbf{v}(\Theta^1)^{-1} \mathbf{v}(\Theta^2)^{-1} \Big[1 + \Big(\mathbf{v}_{in}(B^{12}) + \mathbf{v}_{out}(B^{12})\Big)/2\Big]^{-2}.$$

The three components in the lower bound demonstrate how the structure of Layers 1 and 2 impact the accurate estimation of directed edges to Layer 3. Essentially, the bound suggests that accurate estimation is possible when the total effect (incoming and outgoing edges) at every node of each of the three subgraphs is not very large. This is inherently related to the identifiability of the multi-layered graphical models and our ability to distinguish between the parents from different layers. For instance, if a node in Layer 2 has high partial correlation with nodes of Layer 1, i.e., a node in Layer 2 has parents from many nodes in Layer 1 and yields a large $\mathbf{v}_{\text{in}}(B^{12})$; or similarly, a node in Layer 1 is the parent of many nodes in Layer 2, yielding a large $\mathbf{v}_{\text{out}}(B^{12})$. In either case, we end up with some large lower bound for $\Lambda_{\min}(\Sigma_{\widetilde{X}})$ and it can be hard to distinguish Layer $1 \to 3$ edges from Layer $2 \to 3$ edges.

For a general M-layer network, the argument in the proof of Proposition 2.4 (see Section A.2 for details) can be generalized in a straightforward manner. In the 2-layer network setting, with the notation defined in Section 2.2, by setting $\vec{\epsilon}^{\,1} = \vec{X}^1$, we have

$$\begin{bmatrix} \vec{\epsilon}^{\,1} \\ \vec{\epsilon}^{\,2} \end{bmatrix} = P \begin{bmatrix} \vec{X}^1 \\ \vec{X}^2 \end{bmatrix}, \quad \text{where} \quad P = \begin{bmatrix} \mathrm{I} & O \\ -(B^{12})^\top & \mathrm{I} \end{bmatrix}.$$

For an M-layer network, a modified P is given in the following form:

$$P = \begin{bmatrix} \mathrm{I} & 0 & \cdots & O \\ -(B^{12})^\top & \mathrm{I} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ -(B^{1,M-1})^\top & -(B^{2,M-1})^\top & \cdots & \mathrm{I} \end{bmatrix},$$

and combines node capacities for different layers. The conclusion is qualitatively similar, i.e., the estimation error of an M-layer graphical model is small as long as the maximum node capacities of different inter-layer and intra-layer subgraphs are not too large.

2.4 Performance Evaluation and Implementation Issues.

In this section, we present selected simulation results for our proposed method, in two-layer and three-layer network settings. Further, we introduce some acceleration techniques that can speed up the algorithm and reduce computing time.

2.4.1 Simulation results.

For the 2-layer network, as mentioned in Section 2.2.1, since the main target of our proposed algorithm is to provide estimates for $B^\star$ and $\Theta_\epsilon^\star$ (since Θ_X can be estimated separately), we only present evaluation results for $B^\star$ and $\Theta_\epsilon^\star$ estimates. Similarly, for the 3-layer network, we only present evaluation results involving Layer 3, using the notation in Section

2.3.4, that is, $B^\star_{XZ}, B^\star_{YZ}$ and $\Theta^\star_{\epsilon,Z}$ estimates, which is sufficient to show how our proposed algorithm works in the presence of a "super-layer", taking advantage of the separability of the log-likelihood.

2-layer network. To compare the proposed method with the most recent methodology that also provides estimates for the regression parameters and the precision matrix (CAPME, [**?**]), we use the exact same model settings that have been used in that paper. Specifically, we consider the following two models:

- Model A: Each entry in $B^\star$ is nonzero with probability $5/p_1$, and off-diagonal entries for $\Theta^\star_\epsilon$ are nonzero with probability $5/p_2$.
- Model B: Each entry in $B^\star$ is nonzero with probability $30/p_1$, and off-diagonal entries for $\Theta^\star_\epsilon$ are nonzero with probability $5/p_2$.

As in **?**], for both models, nonzero entries of $B^\star$ and $\Theta^\star_\epsilon$ are generated from $\mathsf{Unif}\,[(-1,-0.5)\cup(0.5,1)]$, and diagonals of $\Theta^\star_\epsilon$ are set identical such that the condition number of $\Theta^\star_\epsilon$ is p_2.

	(p_1, p_2, n)		(p_1, p_2, n)
Model A	$(30, 60, 100)$	Model B	$(200, 200, 100)$
	$(60, 30, 100)$		$(200, 200, 200)$
	$(200, 200, 150)$		
	$(300, 300, 150)$		

Table 2.1: Model Dimensions for Model A and B.

To evaluate the selection performance of the algorithm, we use sensitivity (SEN), specificity (SPE) and Mathews Correlation Coefficient (MCC) as criteria:

$$\text{SEN} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \quad \text{SPE} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}.$$

Further, to evaluate the accuracy of the magnitude of the estimates, we use the relative error in Frobenius norm:

$$\text{rel-Fnorm} = |||\widetilde{B} - B^\star|||_{\text{F}} / |||B^\star|||_{\text{F}} \quad \text{or} \quad |||\widetilde{\Theta}_\epsilon - \Theta^\star_\epsilon|||_{\text{F}} / |||\Theta^\star_\epsilon|||_{\text{F}}.$$

Tables 2.2 and 2.3 show the results for both the regression matrix and the precision matrix. For the precision matrix estimation, we compare our result with those available in **?**], denoted as CAPME.

As it can be seen from Tables 2.2 and 2.3, the sample size is a key factor that affects the performance. Our proposed algorithm performs extremely well in its selection properties on

	(p_1, p_2, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(30,60,100)	0.96(.018)	0.99(.004)	0.93(.014)	0.22(.029)
	(60,30,100)	0.99(.009)	0.99(.003)	0.93(.017)	0.18(.021)
	(200,200,150)	0.99(.001)	0.99(.001)	0.88(.009)	0.18(.007)
	(300,300,150)	1.00(.001)	0.99(.001)	0.84(.010)	0.21(.007)
Model B	(200,200,200)	0.97(.004)	0.98(.001)	0.92(.002)	0.19(.009)
	(200,200,100)	0.32(.010)	0.99(.001)	0.49(.009)	0.85(.006)

Table 2.2: Performance evaluation for the estimated regression matrix over 50 replications.

	(p_1, p_2, n)		SEN	SPE	MCC	rel-Fnorm
Model A	(30,60,100)		0.77(.031)	0.92(.007)	0.56(.030)	0.51(.017)
		CAPME	0.58(.030)	0.89(.010)	0.45(.030)	
	(60,30,100)		0.76(.041)	0.89(.015)	0.59(.039)	0.49(.014)
	(200,200,150)		0.78(.019)	0.97(.001)	0.55(.012)	0.60(.007)
	(300,300,150)		0.71(.017)	0.98(.001)	0.51(.011)	0.59(.005)
Model B	(200,200,200)		0.73(.023)	0.94(.003)	0.39(.017)	0.62(.011)
		CAPME	0.36(.020)	0.97(.000)	0.35(.010)	
	(200,200,100)		0.57(.027)	0.44(.007)	0.04(.008)	0.84(.002)
		CAPME	0.19(.010)	0.87(.000)	0.04(.010)	

Table 2.3: Performance evaluation for the estimated precision matrix over 50 replications.

B and strikes a good balance between sensitivity and specificity in estimating Θ_ϵ.[3] For most settings, it provides substantial improvements over the CAPME estimator.

3-layer network. For a 3-layer network, we consider the following data generation mechanism: for all three models A, B and C, each entry in B_{XY} is nonzero with probability $5/p_1$, each entry in B_{XZ} and B_{YZ} is nonzero with probability $5/(p_1+p_2)$, and off-diagonal entries in $\Theta_{\epsilon,Z}$ are nonzero with probability $5/p_3$. Similar to the 2-layered set-up, the nonzero entries in $\Theta_{\epsilon,Z}$ are generated from $\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)]$ with its diagnals set identical such that its condition number is p_3. For the regression matrices in the three models, nonzeros in B_{XY} are generated from $\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)]$, and nonzeros in B_{XZ} and B_{YZ} are generated from $\{\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)] * \text{Signal.Strength}\}$, where the signal strength in the three models are given by 1, 1.5 and 2, respectively. More specifically, for Model A, B and C, nonzeros in B_{XZ} or B_{YZ} are generated from $\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)]$, $\mathsf{Unif}[(-1.5,-0.75)\cup(0.75,1.5)]$ and $\mathsf{Unif}[(-2,-1)\cup(1,2)]$, respectively.

As mentioned in the beginning of this subsection, we only evaluate the algorithm's performance on B_{XZ}, B_{YZ} and $\Theta_{\epsilon,Z}$. Based on the results shown in Tables 2.5, 2.6 and 2.7, the signal strength across layers affects the accuracy of the estimation, which is in accordance with what has been discussed regarding identifiability. Overall, the MLE estimator performs satisfactorily across a fairly wide range of settings and in many cases achieving very high

[3]In practice, for the debias Lasso procedure, we use the default choice of tuning parameters suggested in the implementation of the code provided in **?**]; for FWER, we suggest using $\alpha = 0.1$ as the thresholding level; for tuning parameter selection, we suggest doing a grid search for (λ_n, ρ_n) on $[0, 0.5\sqrt{\log p_1/n}] \times [0, 0.5\sqrt{\log p_2/n}]$ with BIC.

	Layer 3 Signal Strength	(p_1, p_2, p_3, n)
Model A	1	(50,50,50,200)
Model B	1.5	(50,50,50,200)
Model C	2	(50,50,50,200)
		(20,80,50,200)
		(80,20,50,200)
		(100,100,100,200)

Table 2.4: Model Dimensions and Signal Strength for Model A, B and C.

	(p_1, p_2, p_3, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(50,50,50,200)	0.51(.065)	0.99(.001)	0.69(.049)	0.68(.050)
Model B	(50,50,50,200)	0.85(.043)	0.99(.001)	0.898(.025)	0.36(.056)
Model C	(50,50,50,200)	0.97(.018)	0.99(.002)	0.96(.016)	0.16(.040)
	(20,80,50,200)	0.55(.078)	0.99(.001)	0.72(.059)	0.63(.066)
	(80,20,50,200)	0.99(.006)	0.99(.002)	0.94(.017)	0.08(.032)
	(100,100,100,200)	1.00(.001)	0.99(.001)	0.87(.016)	0.07(.007)

Table 2.5: Performance evaluation for estimated regression matrix B_{XZ} over 50 replications.

	(p_1, p_2, p_3, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(50,50,50,200)	0.53(.051)	1.00(.000)	0.72(.036)	0.65(.041)
Model B	(50,50,50,200)	0.90(.033)	1.00(.000)	0.95(.019)	0.25(.049)
Model C	(50,50,50,200)	0.98(.013)	1.00(.000)	0.99(.007)	0.12(.042)
	(20,80,50,200)	0.95(.013)	1.00(.000)	0.98(.007)	0.19(.030)
	(80,20,50,200)	0.96(.027)	0.99(.001)	0.97(.022)	0.14(.063)
	(100,100,100,200)	1.00(.000)	1.00(.000)	0.99(.002)	0.025(.002)

Table 2.6: Performance evaluation for estimated regression matrix B_{YZ} over 50 replications.

	(p_1, p_2, p_3, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(50,50,50,200)	0.69(.044)	0.638(.032)	0.20(.036)	0.82(.017)
Model B	(50,50,50,200)	0.77(.050)	0.82(.036)	0.42(.071)	0.68(.040)
Model C	(50,50,50,200)	0.88(.041)	0.91(.019)	0.63(.059)	0.56(.034)
	(20,80,50,200)	0.72(.041)	0.80(.028)	0.36(.050)	0.72(.021)
	(80,20,50,200)	0.90(.028)	0.92(0.011)	0.68(.039)	0.58(.018)
	(100,100,100,200)	0.96(.014)	0.96(0.003)	0.68(.016)	0.49(.010)

Table 2.7: Performance evaluation for estimated precision matrix $\Theta_{\epsilon,Z}$ over 50 replications.

values for the MCC criterion.

2.4.1.1 Simulation results for non-Gaussian data.

In many applications, the data may not be exactly Gaussian, but approximately Gaussian. Next, we present selected simulation results when the data comes from some distribution that deviates from Gaussian. Specifically, we consider two types of deviations based on the following transformations: (i) a truncated empirical cumulative distribution function and (ii) a shrunken empirical cumulative distribution functions as discussed in **?**]. In both simulation settings, we consider Model A with $(p_1, p_2, n) = (30, 60, 100)$ under the two-layer setting, and the transformation is applied to errors in Layer 2. Table 2.8 shows the simulation results for these two scenarios over 50 replications.

Based on the results in Table 2.8, relatively small deviations from the Gaussian distribu-

tion do not affect the performance of the MLE estimates under the examined settings that are comparable to those obtained with Gaussian distributed data.

Setting	Parameter	SEN	SPE	MCC	rel-Fnorm
Model A $(30, 60, 100)$	B	0.96(.017)	0.99(.003)	0.94(.012)	0.20(.028)
shrunken	Θ_ϵ	0.76(.031)	0.91(.008)	0.55(.030)	0.51(.019)
Model A $(30, 60, 100)$	B	0.96(.021)	0.98(.004)	0.93(.015)	0.21(.034)
truncation	Θ_ϵ	0.76(.033)	0.92(.008)	0.56(.035)	0.52(.023)

Table 2.8: Simulation results for B and Θ_ϵ over 50 replications under npn transformation.

2.4.2 A comparison with the two-step estimator in ?].

Next, we present a comparison between the MLE estimator and the two-step estimator of **?**]. Specifically, we use the CAPME estimate as an initializer for the MLE procedure and examine its evolution over successive iterations. We evaluate the value of the objective function at each iteration, and also compare it to the value of the objective function evaluated at our initializer (screening + Lasso/Ridge) and the estimates afterward. For illustration purposes, we only show the results for a single realization under Model A with $p_1 = 30, p_2 = 60, n = 100$, although similar results were obtained in other simulation settings. Figure 2.2 shows the value of the objective function as a function of the iteration under both initialization procedures, while Table 2.9 shows how the cardinality of the estimates changes over iterations for both initializers. It can be seen that the iterative MLE algorithm significantly improves the value of the objective function over the CAPME initialization and also that the set of directed and undirected edges stabilizes after a couple iterations.

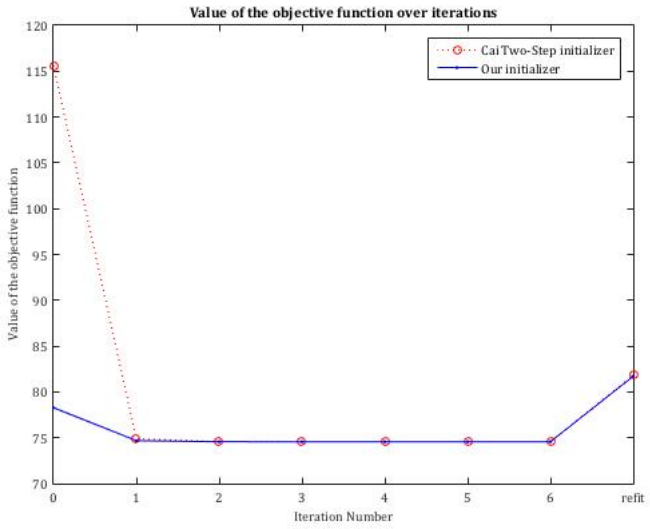

Figure 2.2: Comparison between Cai's estimate and our estimate.

		0	1	2	3	4	5	6	refit
Our initializer	$\widehat{B}^{(k)}$	275	275	275	275	275	275	275	275
	$\widehat{\Theta}_\epsilon^{(k)}$	282	255	247	247	248	248	248	260
CAPME initializer	$\widehat{B}^{(k)}$	433	275	275	275	275	275	275	275
	$\widehat{\Theta}_\epsilon^{(k)}$	979	267	250	249	249	248	248	260

Table 2.9: Change in cardinality over iterations for B and Θ_ϵ.

Based on Figure 2.2 and Table 2.9, we notice that Cai et. al's two-step estimator yields larger value of the objective function compared with our initializer that is obtained through screening followed by Lasso. However, over subsequent iterations, both initializers yield the same value in the objective function, which keeps decreasing according to the nature of block-coordinate descent.

2.4.3 Implementation issues.

Next, we introduce some acceleration techniques for the MLE algorithm aiming to reduce computing time, yet maintaining estimation accuracy over iterations.

(p_2+1)-block update. In Section 2.2, we update B and Θ_ϵ by (2.6) and (2.8), respectively, and within each iteration, the updated B is obtained by an application of cyclic p_2-block coordinate descent with respect to each of its columns until convergence. As shown in Section 2.3.1, the outer 2-block update guarantees the MLE iterative algorithm to converge to a stationary point. However in practice, we can speed up the algorithm by updating B without waiting for it to reach the minimizer for every iteration other than the first one. More precisely, for the alternating search step, we take the following steps when actually implementing the proposed algorithm with initializer $\widehat{B}^{(0)}$ and $\widehat{\Theta}_\epsilon^{(0)}$:

– Iteration 1: update B and Θ_ϵ as follows, respectively:

$$\widehat{B}^{(1)} = \underset{B\in\mathcal{B}_1\times\cdots\times\mathcal{B}_{p_2}}{\arg\min} \big\{\frac{1}{n}\sum_{i=1}^{p_2}\sum_{j=1}^{p_2}(\sigma_\epsilon^{ij})^{(0)}(\mathbf{Y}_i - \mathbf{X}B_i)^\top(\mathbf{Y}_j - \mathbf{X}B_j) + \lambda_n\sum_{j=1}^{p_2}\|B_j\|_1\big\},$$

and

$$\widehat{\Theta}_\epsilon^{(1)} = \underset{\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}{\arg\min}\Big\{\log\det\Theta_\epsilon - \mathrm{tr}(\widehat{S}^{(1)}\Theta_\epsilon) + \rho_n\|\Theta_\epsilon\|_{1,\mathrm{off}}\Big\},$$

where $\widehat{S}^{(1)}$ is the sample covariance matrix of $\widehat{\mathbf{E}}^{(1)} := \mathbf{Y} - \mathbf{X}\widehat{B}^{(1)}$.

– For Iteration $k \geq 2$, while not converged:

· For $j = 1, \cdots, p_2$, update B_j once by

$$\widehat{B}_j^{(k)} = \underset{B_j \in \mathcal{B}_j}{\arg\min} \ \big\{ \frac{(\sigma_\epsilon^{jj})^{(k-1)}}{n} \|\mathbf{Y}_j + \mathbf{r}_j^{(k)} - \mathbf{X} B_j\|_2^2 + \lambda_n \|B_j\|_1 \big\},$$

where

$$\mathbf{r}_j^{(k)} = \frac{1}{(\sigma_\epsilon^{jj})^{(k-1)}} \big[\sum_{i=1}^{j-1} (\sigma_\epsilon^{ij})^{(k-1)} (\mathbf{Y}_i - \mathbf{X}\widehat{B}_i^{(k)}) + \sum_{i=j+1}^{p_2} (\sigma_\epsilon^{ij})^{(k-1)} (\mathbf{Y}_i - \mathbf{X}\widehat{B}_i^{(k-1)}) \big]. \quad (2.22)$$

· Update Θ_ϵ by

$$\widehat{\Theta}_\epsilon^{(k)} = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\arg\min} \ \Big\{ \log\det \Theta_\epsilon - \mathrm{tr}(\widehat{S}^{(k)} \Theta_\epsilon) + \rho_n \|\Theta_\epsilon\|_{1,\mathrm{off}} \Big\},$$

where $\widehat{S}^{(k)}$ is defined similarly.

Intuitively, for the first iteration, we wait for the algorithm to complete the whole cyclic p_2 block-coordinate descent step, as the first iteration usually achieves a big improvement in the value of the objective function compared to the initialization values, as depicted in Figure 2.2. However, in subsequent iterations, the changes in the objective function become relatively small, so we do $(p_2 + 1)$ successive block-updates in every iteration, and start to update Θ_ϵ once a full p_2 block update in B is completed, instead of waiting for the update in B proceeds cyclically until convergence. In practice, this way of updating B and Θ_ϵ leads to faster convergence in terms of total computing time, yet yields the same estimates compared with the exact 2-block update shown in Section 2.2.

Parallelization. A number of steps of the MLE algorithm is parallelizable. In the screening step, when applying the de-biased Lasso procedure [**?**] to obtain the p-values, we need to implement p_2 separate regressions, which can be distributed to different compute nodes and carried out in parallel. So does the refitting step, in which we refit each column in B in parallel.

Moreover, according to **? ? ?**] and a series of similar studies, though the block update in the alternating search step is supposed to be carried out sequentially, we can implement the update in parallel to speed up convergence, yet empirically yield identical estimates. This parallelization can be applied to either the minimization with respect to B within the 2-block update method, or the minimization with respect to each column of B for the $(p_2 + 1)$-block

update method. Either way, $\mathbf{r}_j^{(k)}$ in (2.22) is substituted by

$$\mathbf{r}_{j,\text{parallel}}^{(k)} = \frac{1}{(\sigma_\epsilon^{jj})^{(k-1)}} \sum_{i \neq j}^{p_2} (\sigma_\epsilon^{ij})^{(k-1)} (\mathbf{Y}_i - \mathbf{X}\widehat{B}_i^{(k-1)}),$$

which is not updated until we have updated B_j's once for all $j = 1, \cdots, p_2$ in parallel.

Table 2.10 shows the elapsed time for carrying out our proposed algorithm using 2-block/(p_2+1) -block update with/without parallelization, under the simulation setting where we have $p_1 = p_2 = 200, n = 150$. The screening step and refitting step are both carried out in parallel for all four different implementations. [4]

	2-block	(p_2+1)-block	2-block in parallel	(p_2+1)-block in parallel
elapsed time (sec)	5074	2556	848	763

Table 2.10: Computing time with different update methods.

As shown in the table, using $(p_2 + 1)$-block update and parallelization both can speed up convergence and reduce computing time, which takes only 1/7 of the computing time compared with using 2-block update without parallelization.

Remark 2.11. The total computing time depends largely on the number of bootstrapped samples we choose for the stability selection step. For the above displayed results, we used 50 bootstrapped samples to obtain the weight matrix. Nevertheless, one can select the number of bootstrap samples judiciously and reduce them if performance would not be seriously impacted.

2.5 Discussion.

In this chapter, we examined multi-layered Gaussian networks, proposed a provably converging algorithm for obtaining the estimates of the key model parameters and established their theoretical properties in high-dimensional settings. Note that we focused on ℓ_1 penalties for both the directed and undirected edges, since it was assumed that the multi-layer network was sparse both between layers and within layers. In many scientific applications, external information may require imposing group penalties, primarily on the directed edge parameters (B). For example, in a gene-protein 2-layer network, genes can be grouped according to their function in pathways and one may be interested in assessing the pathway's impact on proteins. In that case, a group lasso penalty can be imposed. In general, the proposed framework can easily accommodate other types of penalties in accordance to the underlying data generating procedure. The exact form of the error bounds established

[4]For parallelization, we distribute the computation on 8 cores.

would be different, depending on the exact choice of penalty selected. Nevertheless, as long as the penalty is convex, all arguments regarding bi-convexity and convergence follow, and we can use similar strategies to bound the statistical error of the estimators, obtained via the iterative algorithm.

Next, we discuss connections of this work to that in **? ? ?**]. In these papers, an alternative parameterization of the 2-layer network is adopted. Specifically, all nodes in layers 1 and 2 are considered jointly and assumed to be drawn from the following Gaussian distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Omega_X & \Omega_{XY} \\ \Omega_{YX} & \Omega_Y \end{pmatrix}^{-1}\right),$$

and by conditioning Y on X, one obtains

$$Y|X \sim \mathcal{N}\left(-\Omega_Y^{-1}\Omega_{XY}' X, \Omega_Y^{-1}\right). \tag{2.23}$$

Compare (2.23) with our model set-up in Section 2.2.1, the following correspondence holds:

$$B = -\Omega_{XY}\Omega_Y^{-1}, \qquad \Omega_Y = \Theta_\epsilon. \tag{2.24}$$

Note that the correspondence in (2.24) is only guaranteed to hold in selective settings. Specifically, at the population level, the correspondence between (Ω_{XY}, Ω_Y) and (B, Θ_ϵ) holds in the absence of any sparsity penalization. Further, in a low-dimensional data setting without penalty terms in the objective function, the estimates from the two parameterizations would be similar provided that the problem is well-conditioned and the sample size reasonably large.

However, the situation is different in high-dimensional settings and in the presence of sparsity penalties. Specifically, given data X and Y, instead of parametrizing the model in terms of (B, Θ_ϵ), the authors in [**? ? ?**] consider the following optimization problem, parametrized in (Ω_{XY}, Ω_Y):

$$\min_{\Omega_{XY}, \Omega_Y} g(\Omega_{XY}, \Omega_Y) \equiv g_0(\Omega_{XY}, \Omega_Y), + \mathcal{R}(\Omega_{XY}, \Omega_Y) \tag{2.25}$$

where $g_0(\Omega_{XY}, \Omega_Y) = -\log\det\Omega_Y + \frac{1}{n}\text{tr}\left[(Y + \Omega_{XY}\Omega_Y^{-1}X)'\Omega_Y(Y + \Omega_{XY}\Omega_Y^{-1}X)\right]$ is jointly convex in (Ω_{XY}, Ω_Y), and $\mathcal{R}(\Omega_{XY}, \Omega_Y)$ is some regularization term. In particular, the element-wise ℓ_1 norm on Ω_Y, and the element-wise ℓ_1 or column-wise ℓ_1 norm (matrix $2,1$ norm) on Ω_{XY} are the main penalties under consideration in those papers.

Despite the convex formulation in (2.25), we would like to point out that in general, the

sparsity pattern in B and Ω_{XY} are not transferable through the regularization term, which underlies a major difference between the formulation in (2.25) and the one presented in this paper. Given the correspondence in (2.24), there are two cases where B and Ω_{XY} share the same sparsity pattern: 1) Ω_Y (or Θ_ϵ, equivalently) is diagonal, or 2) both the i^{th} row in B and Ω_{XY} are identically zero, for an arbitrary $i = 1, \cdots, p_1$. However, both settings are fairly restrictive and unlike to occur in many applications.

Note that the linear model represents a natural modeling tool for a number of problems and the regression coefficients have a specific scientific interpretation. This is easily accomplished through the (B, Θ_ϵ)-parametrization, by adding proper regularization to B (e.g., penalty which enforces element-wise sparsity or group-Lasso type of sparsity, etc) if necessary. However, with the (Ω_{XY}, Ω_Y)-parametrization, the underlying sparsity in the true data generating procedure, encoded by B, will not be easily incorporated, and to add a regularization term on Ω_{XY} may lose the scientific interpretability, and may also lead to an estimated B whose sparsity pattern is completely mis-specified, obtained from (2.24) with $\widehat{\Omega}_{XY}, \widehat{\Omega}_Y$ plugged in.

Another difference we would like to point out is that once we add penalty terms to the objective function in the low dimensional setting, or switch to the high dimensional setting (as considered in **?**] and **?**]), the correspondence between the optimizer(s) of (2.1) and the optimizer(s) of (2.25) become difficult to connect analytically.

CHAPTER III

Regularized Estimation and Testing of High-dimensional Multi-block Vector Autoregressive Models

3.1 Introduction.

The study of linear dynamical systems has a long history in control theory [**?**] and economics [**?**] due to their analytical tractability and ease to estimate their parameters. Such systems in their so-called *reduced form* give rise to Vector Autoregressive (VAR) models [**?**] that have been widely used in macroeconomic modeling for policy analysis [**? ? ?**], in financial econometrics [**?**], and more recently in functional genomics [**?**], financial systemic risk analysis [**?**] and neuroscience [**?**].

In many applications, the components of the system under consideration can be naturally partitioned into *interacting blocks*. For example, [**?**] studied the impact of monetary policy in a small open economy, where the economy under consideration is modeled as one block, while variables in other (foreign) economies as the other. Both blocks have their own autoregressive structure, and the inter-dependence between blocks is unidirectional: the foreign block *influences* the small open economy, but *not* the other way around, thus effectively introducing a *linear ordering* amongst blocks. Another example comes from the connection between the stock market and employment macroeconomic variables [**? ? ?**] that focuses on the impact through a wealth effect mechanism of the former on the latter. Once again, the underlying hypothesis of interest is that the stock market influences employment, but not the other way around. In another application domain, molecular biologists conduct time course experiments on cell lines or animal models and collect data across multiple molecular compartments (transcripotmics, proteomics, metabolomics, lipidomics) in order to delineate mechanisms for disease onset and progression or to study basic biological processes. In this case, the interactions amongst the blocks (molecular compartments) are clearly delineated (transciptomic compartment influencing the proteomic and metabolomic ones), thus leading again to a linear ordering of the blocks [see **?**].

The proposed model also encompasses the popular in marketing, regional science and growth theory VAR-X model, provided that the temporal evolution of the *exogenous* block of variables "X" exhibits autoregressive dynamics. For example, **?**] examine the sensitivity of over 500 product prices to various marketing promotion strategies (the exogenous block), while **?**] examine changes in subscription rates, search engine referrals and marketing efforts of customers when switched from a free account to a fee-based structure, the latter together with customer characteristics representing the exogenous block. **?**] examine regional inter-dependencies, building a model where country specific macroeconomic indicators evolve according to a VAR model and they are influenced exogenously by key macroeconomic variables from neighboring countries/regions. Finally, **?**] studies the impact of the price of oil on Gross Domestic Product growth rates for a number of countries, while controlling for other exogenous variables such as the country's consumption and investment expenditures along with its trade balance.

The proposed model gives rise to a network structure that in its most general form corresponds to a multi-partite graph, depicted in Figure 3.1 for 3 blocks, that exhibits a *directed acyclic structure* between the constituent blocks, and can also exhibit additional dependence between the nodes in each block. Selected properties of such multi-block structures, known as *chain graphs* [**?**], have been studied in the literature. Further, their maximum likelihood estimation for *independent and identically distributed* Gaussian data under a high-dimensional *sparse* regime is thoroughly investigated in **?**], where a provably convergent estimation procedure is introduced and its theoretical properties are established.

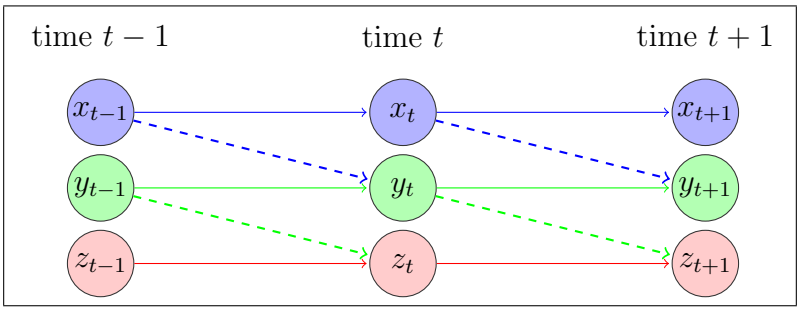

Figure 3.1: Diagram for a dynamic system with three groups of variables

Given the wide range of applications of multi-block VAR models, which in addition encompass the widely used VAR-X model, the key contributions of the current paper are fourfold: (i) formulating the model as a recursive dynamical system and examining its stability properties; (ii) developing a provably convergent algorithm for obtaining the regularized maximum likelihood estimates (MLE) of the model parameters under high-dimensional scaling; (iii) establishing theoretical properties of the ML estimates; and (iv) devising a testing

procedure for the parameters that connect the constituent blocks of the model: if the null hypothesis is not rejected, then one is dealing with a set of independently evolving VAR models, otherwise with the posited multi-block VAR model. Finally, the model, estimation and testing procedures are illustrated on an important problem in macroeconomics, as gleaned by the background of the problem and discussion of the results provided in Section 3.6.

For the multi-block VAR model, we assume that the time series within each block are generated by a Gaussian VAR process. Further, the transition matrices within and across blocks can be either *sparse* or *low rank*. The posited regularized Gaussian likelihood function is not *jointly convex* in all the model parameters, which poses a number of technical challenges that are compounded by the presence of temporal dependence. These are successfully addressed and resolved in Section 3.3, where we provide a numerically convergent algorithm and establish the theoretical properties of the resulting ML estimates, that constitutes a key contribution in the study of multi-block VAR models.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce the model setup and the corresponding estimation procedure. In Section 3.3, we provide consistency properties of the obtained ML estimates under a high-dimensional scaling. In Section 3.4, we introduce the proposed testing framework, both for low-rank and sparse interaction matrices between the blocks. Section 3.5 contains selected numerical results that assess the performance of the estimation and testing procedures. Finally, an application to financial and macroeconomic data that was previously discussed as motivation for the model under consideration is presented in Section 3.6.

Notations. Throughout this chapter, we use $|||A|||_1$ and $|||A|||_\infty$ respectively to denote the matrix induced 1-norm and infinity norm of $A \in \mathbb{R}^{m\times n}$, that is, $|||A|||_1 = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|$, $|||A|||_\infty = \max_{1\le i\le m}\sum_{j=1}^{n}|a_{ij}|$, and use $\|A\|_1$ and $\|A\|_\infty$ respectively to denote the elementwise 1-norm and infinity norm: $\|A\|_1 = \sum_{i,j}|a_{ij}|$, $\|A\|_\infty = \max_{i,j}|a_{ij}|$. Moreover, we use $|||A|||_*$, $|||A|||_F$ and $|||A|||_{\text{op}}$ to denote the nuclear, Frobenius and operator norms of A, respectively. For two matrices A and B of commensurate dimensions, denote their inner product by $\langle\!\langle A, B\rangle\!\rangle = \text{trace}(A'B)$. Finally, we write $A \gtrsim B$ if there exists some absolute constant c that is independent of the model parameters such that $A \ge cB$.

3.2 Problem Formulation.

To convey the main ideas and the key technical contributions, we consider a recursive linear dynamical system comprising of two blocks of variables, whose structure is given by:

$$\begin{aligned} X_t &= AX_{t-1} + U_t, \\ Z_t &= BX_{t-1} + CZ_{t-1} + V_t, \end{aligned} \tag{3.1}$$

where $X_t \in \mathbb{R}^{p_1}, Z_t \in \mathbb{R}^{p_2}$ are the variables in groups 1 and 2, respectively. The temporal intra-block dependence is captured by transition matrices A and C, while the inter-block dependence by B. Noise processes $\{U_t\}$ and $\{V_t\}$, respectively, capture additional contemporaneous intra-block dependence of X_t and Z_t, after conditioning on their respective past values. Further, we assume that U_t and V_t follow mean zero Gaussian distributions with covariance matrices given by Σ_u and Σ_v, i.e.,

$$U_t \sim \mathcal{N}(0, \Sigma_u), \quad \text{and} \quad V_t \sim \mathcal{N}(0, \Sigma_v).$$

With the above model setup, the parameters of interest are transition matrices $A \in \mathbb{R}^{p_1 \times p_1}$, $B \in \mathbb{R}^{p_2 \times p_1}$ and $C \in \mathbb{R}^{p_2 \times p_2}$, as well as the covariances Σ_u, Σ_v. In high-dimensional settings, different combinations of structural assumptions can be imposed on these transition matrices to enable their estimation from limited time series data. In particular, the intra-block transition matrices A and C are sparse, while the inter-block matrix B can be either sparse or low rank. Note that the block of X_t variables acts as an *exogenous* effect to the evolution of the Z_t block [e.g., **? ?**]. Further, we assume $\Omega_u := \Sigma_u^{-1}$ and $\Omega_v := \Sigma_v^{-1}$ are sparse.

Remark 3.1. For ease of exposition, we posit a VAR(1) modeling structure. Extensions to general multi-block structures akin to the one depicted in Figure 3.1 and VAR(d) specifications are rather straightforward and briefly discussed in Section 3.7.

The triangular (recursive) structure of the system enables a certain degree of separability between X_t and Z_t. In the posited model, X_t is a stand-alone VAR(1) process, and the time series in block Z_t is "Granger-caused" by that in block X_t, but not vice versa. The second equation in (3.1), as mentioned in the introductory section, also corresponds to the so-called "VAR-X" model in the econometrics literature [e.g., **? ? ?**], that extends the standard VAR model to include influences from lagged values of *exogenous* variables. Consider the joint process $W_t = (X_t', Z_t')'$, it corresponds to a VAR(1) model whose transition matrix G has a block triangular form:

$$W_t = GW_{t-1} + \varepsilon_t, \qquad \text{where} \quad G := \begin{bmatrix} A & O \\ B & C \end{bmatrix}, \quad \varepsilon_t = \begin{bmatrix} U_t \\ V_t \end{bmatrix}. \tag{3.2}$$

The model in (3.2) can also be viewed from a Structural Equations Modeling viewpoint involving time series data, and also has a Moving Average representation corresponding to a structural VAR representation with Granger causal ordering [**?**]. As mentioned in the introductory section, the focus of this paper is model parameter estimation under high-dimensional scaling, rather than their cause and effect relationship. For a comprehensive discourse of causality issues for VAR models, we refer the reader to **? ?**], and references

therein.

Next, we introduce the notion of stability and spectrum with respect to the system.

System Stability. To ensure that the joint process $\{W_t\}$ is stable [**?**], we require the spectral radius, denoted by $\rho(\cdot)$, of the transition matrix G to be smaller than 1, which is guaranteed by requiring that $\rho(A) < 1$ and $\rho(C) < 1$, since

$$|\lambda \mathrm{I}_{p_1 \times p_2} - G| = \begin{vmatrix} \lambda \mathrm{I}_{p_1} - A & O \\ -B & \lambda \mathrm{I}_{p_2} - C \end{vmatrix} = |\lambda \mathrm{I}_{p_1} - A||\lambda \mathrm{I}_{p_2} - C|,$$

implying that the set of eigenvalues of G is the union of the sets of eigenvalues of A and C, hence

$$\rho(A) < 1 \ , \rho(C) < 1, \qquad \Rightarrow \quad \rho(G) = \max\{\rho(A), \rho(C)\} < 1.$$

The latter relation implies that the stability of such a recursive system imposes spectrum constraints only on the diagonal blocks that govern the intra-block evolution, whereas the off-diagonal block that governs the inter-block interaction is left unrestricted.

Spectrum of the joint process. Throughout, we assume that the spectral density of $\{W_t\}$ exists, which then possesses a special structure as a result of the block triangular transition matrix G. Formally, we define the spectral density of $\{W_t\}$ as

$$f_W(\theta) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_W(h) e^{-ih\theta}, \qquad \theta \in [-\pi, \pi],$$

where $\Gamma_W(h) := \mathbb{E} W_t W'_{t+h}$. For two (generic) processes $\{X_t\}$ and $\{Z_t\}$, define their cross-covariance as $\Gamma_{X,Z}(h) = \mathbb{E} X_t Z'_{t+h}$ and $\Gamma_{Z,X}(h) = \mathbb{E} Z_t X'_{t+h}$. In general, $\Gamma_{X,Z}(h) \neq \Gamma_{Z,X}(h)$. The cross-spectra are defined as:

$$f_{X,Z}(\theta) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_{X,Z}(h) e^{-ih\theta}, \quad \text{and} \quad f_{Z,X}(\theta) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_{Z,X}(h) e^{-ih\theta}, \ \ \theta \in [-\pi, \pi].$$

For the model given in (3.2), by writing out the dynamics of Z_t, the cross-spectra between X_t and Z_t are given by

$$f_{X,Z}(\theta)(\mathrm{I}_{p_2} - C^\top e^{-i\theta}) = f_X(\theta) B^\top e^{-i\theta}, \quad \text{and} \quad (\mathrm{I}_{p_2} - C e^{i\theta}) f_{Z,X}(\theta) = B e^{i\theta} f_X(\theta). \tag{3.3}$$

Similarly, we have

$$(\mathrm{I}_{p_2} - C e^{i\theta}) f_Z(\theta) = B e^{i\theta} f_{X,Z}(\theta) + f_{V,Z}(\theta). \tag{3.4}$$

Combining (3.3) and (3.4), and after some algebra, the spectrum of the joint process W_t is given by

$$f_W(\theta) = [H_1(e^{i\theta})]^{-1}\big(H_2(e^{i\theta})[\mathbf{1}_{2\times 2} \otimes f_X(\theta)]H_2^*(e^{i\theta}) + \left[\begin{smallmatrix} O & O \\ O & \Sigma_v \end{smallmatrix}\right]\big)[H_1^*(e^{i\theta})]^{-1}, \tag{3.5}$$

where $\mathbf{1}_{2\times 2}$ is a 2×2 matrix with all entries being 1, $*$ denotes the conjugate transpose, and

$$H_1(x) := \left[\begin{smallmatrix} \mathrm{I}_{p_1} & O \\ O & \mathrm{I}_{p_2} - Cx \end{smallmatrix}\right] \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}, \quad H_2(x) := \left[\begin{smallmatrix} \mathrm{I}_{p_1} & O \\ O & Bx \end{smallmatrix}\right] \in \mathbb{R}^{(p_1+p_2)\times(2p_1)}.$$

Equation (3.5) implies that the spectrum of the joint process $\{W_t\}$ can be decomposed into the sum of two parts: the first, is a function of $f_X(\theta)$, while the second part involves the embedded idiosyncratic error process $\{V_t\}$ of $\{Z_t\}$, which only affects the right-bottom block of the spectrum. Note that since $\{W_t\}$ is a VAR(1) process, its matrix-valued characteristic polynomial is given by $\mathcal{G}(\theta) := \mathrm{I}_{(p_1+p_2)} - G\theta$, and its spectral density also takes the following form [c.f. **?** **?**]:

$$f_W(\theta) = \frac{1}{2\pi}\big[\mathcal{G}^{-1}(e^{i\theta})\big]\Sigma_\varepsilon\big[\mathcal{G}^{-1}(e^{i\theta})\big]^*, \quad \text{where} \quad \mathcal{G}(x) = \left[\begin{smallmatrix} \mathrm{I}_{p_1} - Ax & O \\ -Bx & \mathrm{I}_{p_2} - Cx \end{smallmatrix}\right], \quad \Sigma_\varepsilon = \left[\begin{smallmatrix} \Sigma_u & O \\ O & \Sigma_v \end{smallmatrix}\right].$$

One can easily reach the same conclusion as in (3.5) by multiplying each term, followed by some algebraic manipulations.

3.2.1 Estimation.

Next, we outline the algorithm for obtaining the ML estimates of the transition matrices A, B and C and inverse covariance matrices Σ_u^{-1} and Σ_v^{-1} from time series data. We allow for a potential high-dimensional setting, where the ambient dimensions p_1 and p_2 of the model exceed the total number of observations T.

Given centered times series data $\{x_0, \cdots, x_T\}$ and $\{z_0, \cdots, z_T\}$, we use $\mathbf{X}_T$ and $\mathbf{Z}_T$ respectively, to denote the "response" matrix from time 1 to T, that is:

$$\mathbf{X}_T = \begin{bmatrix} x_1 & x_2 & \dots & x_T \end{bmatrix}^\top \quad \text{and} \quad \mathbf{Z}_T = \begin{bmatrix} z_1 & z_2 & \dots & z_T \end{bmatrix}^\top,$$

and use $\mathbf{X}$ and $\mathbf{Z}$ without the subscript to denote the "design" matrix from time 0 to $T-1$:

$$\mathbf{X} = \begin{bmatrix} x_0 & x_1 & \dots & x_{T-1} \end{bmatrix}^\top \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} z_0 & z_1 & \dots & z_{T-1} \end{bmatrix}^\top.$$

We use $\mathbf{U}$ and $\mathbf{V}$ to denote the error matrices. To obtain estimates for the parameters of interest, we formulate optimization problems using a penalized log-likelihood function, with regularization terms corresponding to the imposed structural assumptions on the model

parameters–sparsity and/or low-rankness. To solve the optimization problems, we employ block-coordinate descent algorithms, akin to those described in **?**], to obtain the solution.

As previously mentioned, $\{X_t\}$ is not "Granger-caused" by Z_t and hence it is a stand-alone VAR(1) process; this enables us to separately estimate the parameters governing the X_t process (A and Σ_u^{-1}) from those of the Z_t process (B, C, and Σ_v^{-1}).

Estimation of A and Σ_u^{-1}. Conditional on the initial observation x_0, the likelihood of $\{x_t\}_{t=1}^T$ is given by:

$$\begin{aligned} L(x_T, x_{T-1}, \cdots, x_1 | x_0) &= L(x_T | x_{T-1}, \cdots, x_0) L(x_{T-1} | x_{T-2}, \cdots, x_0) \cdots L(x_1 | x_0) \\ &= L(x_T | x_{T-1}) L(x_{T-1} | x_{T-2}) \cdots L(x_1 | x_0), \end{aligned}$$

where the second equality follows from the Markov property of the process. The log-likelihood function is given by:

$$\ell(A, \Sigma_u^{-1}) = \frac{T}{2} \log\det(\Sigma_u^{-1}) - \frac{1}{2} \sum_{t=1}^{T} (x_t - A x_{t-1})^\top \Sigma_u^{-1} (x_t - A x_{t-1}) + \text{constant}.$$

Letting $\Omega_u := \Sigma_u^{-1}$, then the penalized maximum likelihood estimator can be written as

$$\begin{aligned} (\widehat{A}, \widehat{\Omega}_u) = \underset{A \in \mathbb{R}^{p_1 \times p_2}, \Omega_u \in \mathbb{S}_{p_1 \times p_1}^{++}}{\arg\min} \Big\{ & \mathrm{tr}\big[\Omega_u (\mathbf{X}_T - \mathbf{X} A^\top)^\top (\mathbf{X}_T - \mathbf{X} A^\top)/T\big] - \log\det \Omega_u \\ & + \lambda_A \|A\|_1 + \rho_u \|\Omega_u\|_{1,\text{off}} \Big\}. \end{aligned} \tag{3.6}$$

Algorithm 3.1 describes the key steps for obtaining $\widehat{A}$ and $\widehat{\Omega}_u$.

Algorithm 3.1: Computational procedure for estimating A and Σ_u^{-1}.

Input: Time series data $\{x_t\}_{t=1}^T$, tuning parameter λ_A and ρ_u.
Initialization: Initialize with $\widehat{\Omega}_u^{(0)} = \mathrm{I}_{p_1}$, then

$$\widehat{A}^{(0)} = \arg\min_A \big\{ \tfrac{1}{T} \|\!| \mathbf{X}_T - \mathbf{X} A^\top |\!\|_{\mathrm{F}}^2 + \lambda_A \|A\|_1 \big\}.$$

Iterate until convergence:
(1) Update $\widehat{\Omega}_u^{(k)}$ by graphical Lasso [**?**] on the residuals with the plug-in estimate $\widehat{A}^{(k)}$.
(2) Update $\widehat{A}^{(k)}$ with the plug-in $\widehat{\Omega}_u^{(k-1)}$ and cyclically update each row with a Lasso penalty, which solves

$$\min_A \big\{ \tfrac{1}{T} \mathrm{tr}\big[\widehat{\Omega}_u^{(k-1)} (\mathbf{X}_T - \mathbf{X} A^\top)^\top (\mathbf{X}_T - \mathbf{X} A^\top)/T\big] + \lambda_A \|A\|_1 \big\}.$$

Output: Estimated sparse transition matrix $\widehat{A}$ and sparse inverse covariance matrix $\widehat{\Omega}_u$.

Specifically, for fixed $\widehat{\Omega}_u$, each row $j = 1, \ldots, p_1$ of $\widehat{A}$ is cyclically updated by:

$$\widehat{A}_{j\cdot}^{[s+1]} = \underset{\beta \in \mathbb{R}^{p_1}}{\arg\min} \ \big\{ \tfrac{\widehat{\omega}_u^{jj}}{T} \|\mathbf{X}_{T;\cdot j} + \mathbf{r}_j^{[s+1]} - \mathbf{X}\beta\|_2^2 + \lambda_A \|\beta\|_1 \big\},$$

where

$$\mathbf{r}_j^{[s+1]} = \tfrac{1}{\widehat{\omega}_u^{jj}} \big[\sum_{i=1}^{j-1} \widehat{\omega}_u^{ij} \big(\mathbf{X}_{T;\cdot j} - \mathbf{X}(\widehat{A}_{i\cdot}^{[s+1]})^\top \big) + \sum_{i=j+1}^{p_1} \widehat{\omega}_u^{ij} \big(\mathbf{X}_{T;\cdot j} - \mathbf{X}(\widehat{A}_{i\cdot}^{[s]})^\top \big) \big].$$

Here $\widehat{\Omega}_u^{(k)} = \big[(\widehat{\omega}_u^{ij})^{(k)} \big]$ is the estimate from the previous iteration, and for notation convenience we drop the index (k) for the outer iteration and use $[s]$ to denote the index for the inner iteration, for each round of cyclic update of the rows.

Estimation of B, C and Σ_v^{-1}. Similarly, to obtain estimates of B, C and $\Omega_v := \Sigma_v^{-1}$, we formulate the optimization problem as follows:

$$\begin{aligned}(\widehat{B}, \widehat{C}, \widehat{\Omega}_v) := \underset{\substack{B \in \mathbb{R}^{p_2 \times p_1}, C \in \mathbb{R}^{p_2 \times p_2} \\ \Omega_v \in \mathbb{S}_{p_2 \times p_2}^{++}}}{\arg\min} & \big\{ \operatorname{tr}\big[\tfrac{1}{T} \Omega_v (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)^\top (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top) \big] \\ & - \log\det \Omega_v + \lambda_B \mathcal{R}(B) + \lambda_C \|C\|_1 + \rho_v \|\Omega_v\|_{1,\mathrm{off}} \big\}, \end{aligned} \tag{3.7}$$

where the regularizer $\mathcal{R}(B) = \|B\|_1$ if B is assumed to be sparse, and $\mathcal{R}(B) = |||B|||_*$ if B is assumed to be low rank. Algorithm 3.2 outlines the procedure for obtaining estimates $\widehat{B}$, $\widehat{C}$ and $\widehat{\Omega}_v$. Note that $\widehat{B}^{(k)}$ and $\widehat{C}^{(k)}$ need to be treated as a "joint block" in the outer update and convergence of the "joint block" is required before moving on to updating Ω_v.

In many real applications, B is low rank and C is sparse, while in other settings both are sparse. In the first case, X_t "Granger-causes" Z_t and the information can be compressed to a lower dimensional space spanned by a relative small number of bases compared to the dimension of the blocks, and Z_t is autoregressive through a subset of its components. Next, we give details for updating B and C under this model specification.

For fixed $\widehat{\Omega}_v^{(k)}$, with B being low rank and C sparse, the updated $\widehat{B}^{(k)}$ and $\widehat{C}^{(k)}$ satisfies

$$\begin{aligned}(\widehat{B}^{(k)}, \widehat{C}^{(k)}) = \underset{B,C}{\arg\min} \ \big\{ & \tfrac{1}{T} \operatorname{tr}\big[\widehat{\Omega}_v^{(k-1)} (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)^\top (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top) \big] \\ & + \lambda_B |||B|||_* + \lambda_C \|C\|_1 \big\}. \end{aligned}$$

To obtain the solution to the above optimization problem, B and C need to be updated alternately, and within the update of each block, an iterative algorithm is required. Here we drop the superscript (k) that denotes the outer iterations, and use $[s]$ as the inner iteration index for the alternate update between B and C, and use $\{t\}$ as the index for the within

Algorithm 3.2: Computational procedure for estimating B, C and Σ_v^{-1}.

Input: Time series data $\{x_t\}_{t=1}^T$ and $\{z_t\}_{t=1}^T$, tuning parameters λ_B, λ_C, ρ_v.
Initialization: Initialize with $\widehat{\Omega}_v^{(0)} = \mathrm{I}_{p_2}$, then

$$(\widehat{B}^{(0)}, \widehat{C}^{(0)}) = \underset{(B,C)}{\arg\min} \ \big\{ \tfrac{1}{T} \| \mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top \|_{\mathrm{F}}^2 + \lambda_B \mathcal{R}(B) + \lambda_C \|C\|_1 \big\}.$$

Iterate until convergence:
(1) Update $\widehat{\Omega}_v^{(k)}$ by graphical Lasso on the residuals with the plug-in estimates $\widehat{B}^{(k)}$ and $\widehat{C}^{(k)}$
(2) For fixed $\widehat{\Omega}_v^{(k)}$, $(\widehat{B}^{(k+1)}, \widehat{C}^{(k+1)})$ solves

$$\min_{B,C} \big\{ \tfrac{1}{T} \mathrm{tr}\big[\widehat{\Omega}_v^{(k)} (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C')^\top (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top) \big] + \lambda_B \mathcal{R}(B) + \lambda_C \|C\|_1 \big\}.$$

- Fix $\widehat{C}^{[s]}$, update $\widehat{B}^{[s+1]}$ by Lasso or singular value thresholding, which solves

$$\min_{B} \big\{ \tfrac{1}{T} \mathrm{tr}\big[\widehat{\Omega}_v^{(k)} (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}\widehat{C}^{[s]\top})^\top (\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}\widehat{C}^{[s]\top}) \big] + \lambda_B \mathcal{R}(B) \big\};$$

- Fix $\widehat{B}^{[s]}$, update $\widehat{C}^{[s]}$ by Lasso, which solves

$$\min_{C} \big\{ \tfrac{1}{T} \mathrm{tr}\big[\widehat{\Omega}_v^{(k)} (\mathbf{Z}_T - \mathbf{X}\widehat{B}^{[s]\top} - \mathbf{Z}C^\top)^\top (\mathbf{Z}_T - \mathbf{X}\widehat{B}^{[s]\top} - \mathbf{Z}C^\top) \big] + \lambda_C \|C\|_1 \big\}.$$

Output: Estimated transition matrices $\widehat{B}$, $\widehat{C}$ and sparse $\widehat{\Omega}_v$.

block iterative update.

- Update $\widehat{B}^{[s+1]}$ for fixed $\widehat{C}^{[s]}$: instead of directly updating B, update $\widetilde{B} := \widehat{\Omega}_v^{1/2} B$ with

$$\widetilde{B}^{\{t+1\}} = \mathcal{T}_{\alpha_{t+1}\lambda_B}\big(\widetilde{B}^{\{t\}} - \alpha_{t+1} \nabla g(\widetilde{B}^{\{t\}}) \big), \tag{3.8}$$

  where $\mathcal{T}_\tau(\cdot)$ is the singular value thresholding operator with thresholding level τ,

$$\begin{aligned} g(\widetilde{B}) :&= \tfrac{1}{T} \mathrm{tr}\big[\widetilde{B}^\top \widetilde{B} \mathbf{X}^\top \mathbf{X} - 2\mathbf{X}^\top (\mathbf{Z}_T - \mathbf{Z}\widehat{C}^{[s]\top}) \widehat{\Omega}_v^{1/2} \widetilde{B} \big], \\ \nabla g(\widetilde{B}) &= \tfrac{2}{T} \big[\widetilde{B} \mathbf{X}^\top \mathbf{X} - \widehat{\Omega}_v^{1/2} (\mathbf{Z}_T - \mathbf{Z}\widehat{C}^{[s]\top})^\top \mathbf{X} \big]. \end{aligned}$$

  Denote the convergent solution by $\widetilde{B}^{\{\infty\}}$, and $\widehat{B}^{[s+1]} = \widehat{\Omega}_v^{-1/2} \widetilde{B}^{\{\infty\}}$.

- Update $\widehat{C}^{[s]}$ for fixed $\widehat{B}^{[s]}$: each row $j = 1, \ldots, p_2$ of $\widehat{C}^{[s]}$ is cyclically updated by

$$\widehat{C}_{j\cdot}^{\{t+1\}} = \underset{\beta \in \mathbb{R}^{p_2}}{\arg\min} \ \big\{ \tfrac{\widehat{\omega}_v^{jj}}{T} \| (\mathbf{Z}_T - \mathbf{X}\widehat{B}^{[s]\top})_{\cdot j} + \mathbf{r}_j^{\{t+1\}} - \mathbf{Z}\beta \|_2^2 + \lambda_C \|\beta\|_1 \big\},$$

  where

$$\mathbf{r}_j^{\{t+1\}} = \tfrac{1}{\widehat{\omega}_v^{jj}} \Big[\sum_{i=1}^{j-1} \widehat{\omega}_v^{ij} \big[(\mathbf{Z}_T - \mathbf{X}\widehat{B}^{[s]\top})_{\cdot j} - \mathbf{Z}(\widehat{C}_{i\cdot}^{\{t+1\}})^\top \big] + \sum_{i=j+1}^{p_2} \widehat{\omega}_v^{ij} \big[(\mathbf{Z}_T - \mathbf{X}\widehat{B}^{[s]\top})_{\cdot j} - \mathbf{Z}(\widehat{C}_{i\cdot}^{\{t\}})^\top \big] \Big],$$

and $\widehat{\omega}_v^{ij}$ are entries of $\widehat{\Omega}_v$ coming from the previous outer iteration.

Although based on the outlined procedure, a number of iterative steps are required to obtain the final estimate, we have empirically observed that the number of iterations between the B, C and Ω_v blocks is usually rather small. Specifically, based on large number of simulation settings (selected ones presented in Section 3.5), for fixed $\widehat{\Omega}_v^{(k)}$, the alternate update for B and C usually converges within 20 iterations, while the update involving (B, C) and Ω_v takes less than 10 iterations.

In case, B is sparse, it can be updated by Lasso regression as outlined in Algorithm 2 (details omitted).

Remark 3.2. In the low dimensional setting where the Gram matrix $\mathbf{X}^\top\mathbf{X}$ is invertible, the update of B when $\mathcal{R}(B) = |||B|||_*$ can be obtained by a one-shot SVD and singular value thresholding; that is, we first obtain the generalized least squares estimator, then threshold the singular values at a certain level. On the other hand, in the high dimensional setting, an iterative algorithm is required. Note that the singular value thresholding algorithm corresponds to a proximal gradient algorithm, and thus a number of acceleration schemes are available [see **?** **?** **?** **?**], whose theoretical properties have been thoroughly investigated in **?**]. We recommend using the acceleration scheme proposed by **?**], in which the "momentum" is carried over to the next iteration, as an extension of **?**] to composite functions. Instead of updating $\widetilde{B}$ with (3.8), an "accelerated" update within the SVT step is given by:

$$\widetilde{B}^{\{t+1\}} = \mathcal{T}_{\alpha_{t+1}\lambda_B}\big(y - \alpha_{t+1}\nabla g(\widetilde{B}^{\{t\}})\big), \quad \text{where} \quad y = \widetilde{B}^{\{t\}} - \frac{t-1}{t+2}\big(\widetilde{B}^{\{t\}} - \widetilde{B}^{\{t-1\}}\big).$$

Note that the objective function in (3.6) is not *jointly convex* in both parameters, but *biconvex*. Similarly in (3.7), the objective function is biconvex in $\big[(B, C), \Omega_v\big]$. Consequently, convergence to a stationary point is guaranteed, as long as estimates from all iterations lie within a ball around the true value of the parameters, with the radius of the ball upper bounded by a universal constant that only depends on model dimensions and sample size [**?** , Theorem 4.1]. This condition is satisfied upon the establishment of consistency properties of the estimates.

To establish consistency properties of the estimates requires the existence of good initial values for the model parameters (A, Ω_u), and (B, C, Ω_v), respectively, in the sense that they are sufficiently close to the true parameters. For the (A, Ω_u) parameters, the results in **?**] guarantee that for random realizations of $\{X_t, E_t\}$, with sufficiently large sample size, the errors of $\widehat{A}^{(0)}$ and $\widehat{\Omega}_u^{(0)}$ are bounded with high probability, which provides us with good initialization values. Yet, additional handling of the bounds is required to ensure that estimates from subsequent iterations are also uniformly close to the true value (see Section 3.3.2 Theorems 3.1). A similar property for $(\widehat{B}^{(0)}, \widehat{C}^{(0)}, \widehat{\Omega}_v^{(0)})$ and subsequent iterations is established

in Section 3.3.2 Theorems 3.2 (see also Theorem B.2 in Appendix B.1).

3.3 Theoretical Properties.

In this section, we investigate the theoretical properties of the penalized maximum likelihood estimation procedure proposed in Section 3.2, with an emphasis on the error bounds for the obtained estimates. We focus on the model specification in which the inter-block transition matrix B is *low rank*, which is of interest in many applied settings. Specifically, we consider the consistency properties of $\widehat{A}$ and $(\widehat{B}, \widehat{C})$ that are solutions to the following two optimization problems:

$$(\widehat{A}, \widehat{\Omega}_u) = \underset{A,\Omega_u}{\arg\min} \Big\{ \mathrm{tr}\Big[\Omega_u(\mathbf{X}_T - \mathbf{X}A^\top)^\top(\mathbf{X}_T - \mathbf{X}A^\top)/T\Big] - \log\det\Omega_u + \lambda_A\|A\|_1 + \rho_u\|\Omega_u\|_{1,\mathrm{off}}\Big\}, \tag{3.9}$$

and

$$\begin{aligned}(\widehat{B}, \widehat{C}, \widehat{\Omega}_v) = \underset{B,C,\Omega_v}{\arg\min} \Big\{ &\mathrm{tr}\big[\Omega_v(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)^\top(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)/T\big] - \log\det\Omega_v \\ &+\lambda_B|||B|||_* + \lambda_C\|C\|_1 + \rho_v\|\Omega_v\|_{1,\mathrm{off}}\Big\}.\end{aligned} \tag{3.10}$$

The case of a sparse B can be handled similarly to that of A and/or C with minor modifications (details shown in Appendix B.5).

3.3.1 A road map for establishing the consistency results.

Next, we outline the main steps followed in establishing the theoretical properties for the model parameters. Throughout, we denote with a superscript "$\star$" the true value of the corresponding parameters.

The following key concepts, widely used in high-dimensional regularized estimation problems, are needed in subsequent developments.

Definition 3.1 (Restricted Strong Convexity (RSC))**.** For some generic operator $\mathfrak{X} : \mathbb{R}^{m_1\times m_2} \mapsto \mathbb{R}^{T\times m_1}$, it satisfies the RSC condition with respect to norm Φ with curvature $\alpha_{\mathrm{RSC}} > 0$ and tolerance $\tau > 0$ if

$$\frac{1}{2T}|||\mathfrak{X}(\Delta)|||_F^2 \geq \alpha_{\mathrm{RSC}}|||\Delta|||_F^2 - \tau\Phi^2(\Delta), \qquad \text{for some } \Delta \in \mathbb{R}^{m_1\times m_2}.$$

Note that the choice of the norm Φ is context specific. For example, in sparse regression problems, $\Phi(\Delta) = \|\Delta\|_1$ corresponds to the element-wise ℓ_1 norm of the matrix (or the usual vector ℓ_1 norm for the vectorized version). The RSC condition becomes equivalent to the *restricted eigenvalue (RE) condition* [see **? ?** , and references therein] imposed on $\Gamma_X := \Omega_u \otimes \frac{\mathbf{X}'\mathbf{X}}{T}$[1]. This is

[1]We say Γ_X satisfies the RE condition if $\forall\theta$, $\theta'\Gamma_X\theta \geq \alpha_{\mathrm{RE}}\|\theta\|^2 - \tau\|\theta\|_1^2$ for some curvature α_{RE} and tolerance τ. If we define the operator $\mathfrak{X}_{\Omega_u}$ (induced jointly by $\mathbf{X}$ and Ω_u) as: $\mathfrak{X}_{\Omega_u}(\Delta) := \mathbf{X}\Delta'\Omega_u^{1/2}$, then

the case for the problem of estimating transition matrix A. For estimating B and C, define $\mathcal{Q}$ to be the weighted regularizer $\mathcal{Q}(B,C) := |||B|||_* + \frac{\Lambda_C}{\lambda_B}\|C\|_1$, and the associated norm Φ in this setting is defined as

$$\Phi(\Delta) := \inf_{B_{\text{aug}}+C_{\text{aug}}=\Delta} \mathcal{Q}(B,C),$$

where $B_{\text{aug}} := [B, O_{p_2\times p_2}]$ and $C_{\text{aug}} := [O_{p_2\times p_1}, C]$.

Definition 3.2 (Diagonal dominance)**.** A matrix $\Omega \in \mathbb{R}^{p\times p}$ is strictly diagonally dominant if

$$|\Omega_{ii}| > \sum_{j\neq i} |\Omega_{ij}|, \quad \forall \;\; i = 1, \cdots, p.$$

Definition 3.3 (Incoherence condition [**?**])**.** A matrix $\Omega \in \mathbb{R}^{p\times p}$ satisfies the incoherence condition if:

$$\max_{e\in(S_\Omega)^c} \|H_{eS_\Omega}(H_{S_\Omega S_\Omega})^{-1}\|_1 \leq 1-\xi, \quad \text{for some } \xi \in (0,1),$$

where $H_{S_\Omega S_\Omega}$ denotes the Hessian of the log-determinant barrier $\log\det\Omega$ restricted to the true edge set of Ω denoted by S_Ω, and H_{eS} is similarly defined.

The above two conditions are associated with the inverse covariance matrices Ω_u and Ω_v. Specifically, the diagonal dominance condition is required for $\Omega_u^\star$ and $\Omega_v^\star$ as we build the consistency properties for $\widehat{A}$ and $(\widehat{B},\widehat{C})$ with the penalized maximum likelihood formulation. The incoherence condition is primarily required for establishing the consistency of $\widehat{\Omega}_u$ and $\widehat{\Omega}_v$.

We additionally introduce the upper and lower extremes of the spectrum, defined as

$$\mathcal{M}(f_X) := \operatorname*{esssup}_{\theta\in[-\pi,\pi]} \Lambda_{\max}(f_X(\theta)) \qquad \text{and} \qquad \mathfrak{m}(f_X) := \operatorname*{essinf}_{\theta\in[-\pi,\pi]} \Lambda_{\min}(f_X(\theta)).$$

Analogously, the upper extreme for the cross-spectrum is given by:

$$\mathcal{M}(f_{X,Z}) := \operatorname*{esssup}_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\max}(f^*_{X,Z}(\theta) f_{X,Z}(\theta))},$$

with $f^*_{X,Z}(\theta)$ being the conjugate transpose of $f_{X,Z}(\theta)$. With this definition, $\mathcal{M}(f_{X,Z}) = \mathcal{M}(f_{Z,X})$.

Next, consider the solution to (3.9) that is obtained by the alternate update of A and Ω_u. If Ω_u is held fixed, then A solves (3.11), and we denote the solution by $\bar{A}$ and its corresponding vectorized version as $\bar{\beta}_A := \text{vec}(\bar{A})$:

$$\bar{\beta}_A := \arg\min_{\beta\in\mathbb{R}^{p_1^2}} \big\{ -2\beta^\top\gamma_X + \beta^\top\Gamma_X\beta + \lambda_A\|\beta\|_1\big\}, \tag{3.11}$$

Γ_X satisfying the RE condition implies $\mathfrak{X}_{\Omega_u}$ satisfying the RSC condition. For the rest of this section, we will loosely refer to "Γ_X satisfying the RE condition" (or equivalently, $\mathfrak{X}_{\Omega_u}$ satisfying the RSC condition) as "Γ_X satisfying the RSC condition", whenever there is no ambiguity.

where

$$\Gamma_X := \Omega_u \otimes \frac{\mathbf{X}^\top \mathbf{X}}{T}, \qquad \gamma_X := \frac{1}{T}\left(\Omega_u \otimes \mathbf{X}^\top\right) \mathrm{vec}(\mathbf{X}_T). \tag{3.12}$$

Using a similar notation, if A is held fixed, then Ω_u solves (3.13):

$$\bar{\Omega}_u := \underset{\Theta \in \mathbb{S}^{++}_{p_1 \times p_1}}{\arg\min} \ \big\{ \log\det \Omega_u - \mathrm{trace}\,(S_u \Omega_u) + \rho_u \|\Omega_u\|_{1,\mathrm{off}} \big\}, \tag{3.13}$$

where

$$S_u = \tfrac{1}{T}(\mathbf{X}_T - \mathbf{X}A^\top)^\top (\mathbf{X}_T - \mathbf{X}A^\top). \tag{3.14}$$

For *fixed* realizations of $\mathbf{X}$ and $\mathbf{U}$, by [**?**], the error bound of $\bar{\beta}_A$ relies on (1) Γ_X (or the operator $\mathfrak{X}_{\Omega_u}$) satisfying the RSC condition; and (2) the tuning parameter λ_A is chosen in accordance with a deviation bound condition associated with $\|\mathbf{X}^\top \mathbf{U} \Omega_u / T\|_\infty$. By **?**], the error bound of $\bar{\Omega}_u$ relies on how well S_u concentrates around $\Sigma_u^\star$, that is, $\|S_u - \Sigma_u^\star\|_\infty$. Specifically, for (3.12) and (3.14), with $\Omega_u^\star$ and $A^\star$ plugged in respectively, for *random* realizations of $\mathbf{X}$ and $\mathbf{U}$, these conditions hold with high probability. In the actual implementation of the algorithm, however, quantities in (3.12) and (3.14) are substituted by estimates so that at iteration k, $\widehat{\beta}_A^{(k)}$ and $\widehat{\Omega}_u^{(k)}$ solve

$$\begin{aligned}
\widehat{\beta}_A^{(k)} &:= \underset{\beta \in \mathbb{R}^{p_1^2}}{\arg\min} \ \big\{ -2\beta^\top \widehat{\gamma}_X^{(k)} + \beta^\top \widehat{\Gamma}_X^{(k)} \beta + \lambda_A \|\beta\|_1 \big\}, \\
\widehat{\Omega}_u^{(k)} &:= \underset{\Omega_u \in \mathbb{S}^{++}_{p_1 \times p_1}}{\arg\min} \ \big\{ \log\det \Omega_u - \mathrm{trace}\big(\widehat{S}_u^{(k)} \Omega_u\big) + \rho_u \|\Omega_u\|_{1,\mathrm{off}} \big\},
\end{aligned}$$

where

$$\widehat{\Gamma}_X^{(k)} = \widehat{\Omega}_u^{(k-1)} \otimes \frac{\mathbf{X}^\top \mathbf{X}}{T}, \ \ \widehat{\gamma}_X^{(k)} = \tfrac{1}{T}\big(\widehat{\Omega}_u^{(k-1)} \otimes \mathbf{X}^\top\big)\mathrm{vec}(\mathbf{X}_T), \ \ \widehat{S}_u^{(k)} = \tfrac{1}{T}\big[\mathbf{X}_T - \mathbf{X}(\widehat{A}^{(k)})^\top\big]^\top \big[\mathbf{X}_T - \mathbf{X}(\widehat{A}^{(k)})^\top\big].$$

As a consequence, to establish the finite-sample bounds of $\widehat{A}$ and $\widehat{\Omega}_u$ given in (3.9), we need $\widehat{\Gamma}_X^{(k)}$ to satisfy the RSC condition, a bound on $\|\mathbf{X}^\top \mathbf{U} \widehat{\Omega}_u^{(k-1)}\|_\infty$ and a bound on $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty$ for all k. Toward this end, we prove that for random realizations of $\mathbf{X}$ and $\mathbf{U}$, with high probability, the RSC condition for $\widehat{\Gamma}_X^{(k)}$ and the universal bounds for $\|\mathbf{X}^\top \mathbf{U} \widehat{\Omega}_u^{(k-1)}\|_\infty$ and $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty$ hold for *all iterations* k, albeit the quantities of interest rely on estimates from the previous or current iterations. Consistency results of $\widehat{A}$ and $\widehat{\Omega}_u$ then readily follow.

Next, consider the solution to (3.10) that alternately updates (B, C) and Ω_v. As the regularization term involves both the nuclear norm penalty and the ℓ_1 norm penalty, additional handling of the norms is required which leverages the idea of decomposable regularizers [**?**]. Specifically, if

Ω_v and (B, C) are respectively held fixed, then

$$(\bar{B}, \bar{C}) := \arg\min_{B,C} \ \big\{\tfrac{1}{T}\mathrm{tr}\big[\Omega_v(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)^\top(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)\big] + \lambda_B|||B|||_* + \lambda_C\|C\|_1\big\},$$
$$\bar{\Omega}_v := \arg\min_{\Omega_v} \ \big\{\log\det\Omega_v - \mathrm{trace}\big(S_v\Omega_v\big) + \rho_v\|\Omega_v\|_{1,\mathrm{off}}\big\},$$

where $S_v = \frac{1}{T}(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)^\top(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)$. If we let $\mathbf{W} := [\mathbf{X}, \mathbf{Z}] \in \mathbb{R}^{T\times(p_1+p_2)}$, and define the operator $\mathfrak{W}_{\Omega_v} : \mathbb{R}^{p_2\times(p_1+p_2)} \mapsto \mathbb{R}^{T\times p_2}$ induced jointly by $\mathbf{W}$ and Ω_v as

$$\mathfrak{W}_{\Omega_v}(\Delta) := \mathbf{W}\Delta^\top\Omega_v^{1/2} \ \text{ for } \Delta \in \mathbb{R}^{p_2\times(p_1+p_2)}, \tag{3.15}$$

then $\bar{B}_{\mathrm{aug}} := [\bar{B}, O_{p_2\times p_2}]$ and $\bar{C}_{\mathrm{aug}} := [O_{p_2\times p_1}, \bar{C}]$ are equivalently given by

$$(\bar{B}_{\mathrm{aug}}, \bar{C}_{\mathrm{aug}}) = \arg\min_{B,C} \ \Big\{\tfrac{1}{T}\big|\big|\big|\mathbf{Z}_T\Omega_v^{1/2} - \mathfrak{W}_{\Omega_v}(B_{\mathrm{aug}} + C_{\mathrm{aug}})\big|\big|\big|_{\mathrm{F}}^2 + \lambda_B|||B|||_* + \lambda_C\|C\|_1\Big\}, \tag{3.16}$$

where $B_{\mathrm{aug}} := [B, O_{p_2\times p_2}], C_{\mathrm{aug}} := [O_{p_2\times p_1}, C] \in \mathbb{R}^{p_2\times(p_1+p_2)}$. Then, for fixed realizations of $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{V}$, with an extension of **?**] the error bound of $(\bar{B}_{\mathrm{aug}}, \bar{C}_{\mathrm{aug}})$ relies on (1) the operator $\mathfrak{W}_{\Omega_v}$ satisfying the RSC condition; and (2) tuning parameters λ_B and λ_C are respectively chosen in accordance with the deviation bound conditions associated with

$$\big|\big|\big|\mathbf{W}^\top\mathbf{V}\Omega_v/T\big|\big|\big|_{\mathrm{op}} \qquad \text{and} \qquad \big|\big|\big|\mathbf{W}^\top\mathbf{V}\Omega_v/T\big|\big|\big|_{\infty}. \tag{3.17}$$

The error bound of $\bar{\Omega}_v$ again relies on $\|S_v - \Sigma_v^\star\|_\infty$. In an analogous way, for the actual alternate update,

$$(\widehat{B}_{\mathrm{aug}}^{(k)}, \widehat{C}_{\mathrm{aug}}^{(k)}) = \arg\min_{B,C} \ \big\{\tfrac{1}{T}\big|\big|\big|\mathbf{Z}_T\big[\widehat{\Omega}_v^{(k-1)}\big]^{1/2} - \mathfrak{W}_{\widehat{\Omega}_v^{(k-1)}}(B_{\mathrm{aug}} + C_{\mathrm{aug}})\big|\big|\big|_{\mathrm{F}}^2 + \lambda_B|||B|||_* + \lambda_C\|C\|_1\big\},$$
$$\widehat{\Omega}_v^{(k)} := \arg\min_{\Omega_v} \ \big\{\log\det\Omega_v - \mathrm{trace}\big(\widehat{S}_v^{(k)}\Omega_v\big) + \rho_v\|\Omega_v\|_{1,\mathrm{off}}\big\},$$

and the error bound of $(\widehat{B}, \widehat{C}, \widehat{\Omega}_v)$ defined in (3.10) depends on the properties of $\mathfrak{W}_{\widehat{\Omega}_v^{(k)}}$, $\big|\big|\big|\mathbf{W}^\top\mathbf{V}\Omega_v^{(k)}/T\big|\big|\big|_{\mathrm{op}}$ and $\big|\big|\big|\mathbf{W}^\top\mathbf{V}\Omega_v^{(k)}/T\big|\big|\big|_{\infty}$ for all k. Specifically, when Ω_v and (B, C) (in (3.15) and (3.17), resp.) are substituted by estimated quantities, we prove that the RSC condition and bounds hold with high probability for random realizations of $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{V}$, for *all iterations* k, which then establishes the consistency properties of $(\widehat{B}, \widehat{C})$ and $\widehat{\Omega}_v$.

3.3.2 Consistency results for the Maximum Likelihood estimators.

Theorems and 3.2 below give the error bounds for the estimators in (3.9) and (3.10) obtained through Algorithms III.1 and III.2, using random realizations coming from the stable

VAR system defined in (3.1). As previously mentioned, to establish error bounds for both the transition matrices and the inverse covariance matrix obtained from alternating updates, we need to take into account that the quantities associated with the RSC condition and the deviation bound condition are based on *estimated quantities* obtained from the previous iteration. On the other hand, the sources of randomness contained in the observed data are fixed, hence errors from observed data stop accumulating once all sources of randomness are considered after a few iterations, which govern both the leading term of the error bounds and the probability for the bounds to hold.

Specifically, using the same notation as defined in Section 3.3.1, we obtain the error bounds of the estimated transition matrices and inverse covariance matrices iteratively, building upon that for all iterations k:

(1) Operator $\mathfrak{X}_{\widehat{\Omega}_u^{(k)}}$ and Operator $\mathfrak{W}_{\widehat{\Omega}_v^{(k)}}$ satisfy the RSC condition;
(2) deviation bounds hold for $\|\mathbf{X}^\top \mathbf{U}\widehat{\Omega}_u^{(k)}/T\|_\infty$, $\|\mathbf{W}^\top \mathbf{V}\widehat{\Omega}_v^{(k)}/T\|_\infty$, and $\|\|\mathbf{W}^\top \mathbf{V}\widehat{\Omega}_v^{(k)}/T\|\|_{\mathrm{op}}$;
(3) a good concentration given by $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty$ and $\|\widehat{S}_v^{(k)} - \Sigma_v^\star\|_\infty$.

We keep track of how the bounds change in each iteration until convergence, by properly controlling the norms and track the rate of the error bound that depends on p_1, p_2 and T, and reach the conclusion that the error bounds *hold uniformly for all iterations*, for the estimates of both the transition matrices A, B and C and the inverse covariance matrices Ω_u and Ω_v.

Theorem 3.1. *Consider the stable Gaussian VAR process defined in* (3.1) *in which $A^\star$ is assumed to be $s_A^\star$-sparse. Further, assume the following:*

C1. The incoherence condition holds for $\Omega_u^\star$.

C2. $\Omega_u^\star$ is diagonally dominant.

C3. The maximum node degree of $\Omega_u^\star$ satisfies $d^{\max}_{\Omega_u^\star} = o(p_1)$.

Then, for random realizations of $\{X_t\}$ and $\{U_t\}$, and the sequence $\{\widehat{A}^{(k)}, \widehat{\Omega}_u^{(k)}\}_k$ returned by Algorithm III.1 outlined in Section 3.2.1, there exist constants $c_1, c_2, \tilde{c}_1, \tilde{c}_2 > 0, \tau > 0$ such that for sample size $T \gtrsim \max\{(d^{\max}_{\Omega_u^\star})^2, s_A^\star\} \log p_1$, with probability at least

$$1 - c_1 \exp(-c_2 T) - \tilde{c}_1 \exp(-\tilde{c}_2 \log p_1) - \exp(-\tau \log p_1),$$

the following hold for all $k \geq 1$ for some $C_0, C_0' > 0$ that are functions of the upper and lower extremes $\mathcal{M}(f_X), \mathfrak{m}(f_X)$ of the spectrum $f_X(\theta)$ and do not depend on p_1, T or k:

(i) $\mathfrak{X}_{\widehat{\Omega}_u^{(k)}}$ satisfies the RSC condition;

(ii) $\|\mathbf{X}^\top \mathbf{U}\widehat{\Omega}_u^{(k)}/T\|_\infty \leq C_0\sqrt{\frac{\log p_1}{T}}$;

(iii) $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty \leq C_0'\sqrt{\frac{\log p_1}{T}}$.

As a consequence, the following bounds hold uniformly for all iterations $k \geq 1$*:*

$$|||\widehat{A}^{(k)} - A^\star|||_F = O\Big(\sqrt{\tfrac{s_A^\star \log p_1}{T}}\Big), \qquad |||\widehat{\Omega}_u^{(k)} - \Omega_u^\star|||_F = O\Big(\sqrt{\tfrac{(s_{\Omega_u}^\star + p_1)\log p_1}{T}}\Big).$$

It should be noted that the above result establishes the *consistency for the ML estimates* of the model presented in **?**].

Theorem 3.2. *Consider the stable Gaussian VAR system defined in* (3.1) *in which* $B^\star$ *is assumed to be low rank with rank* $r_B^\star$ *and* $C^\star$ *is assumed to be* $s_C^\star$*-sparse. Further, assume the following*

C1. The incoherence condition holds for $\Omega_v^\star$*.*

C2. $\Omega_v^\star$ *is diagonally dominant.*

C3. The maximum node degree of $\Omega_v^\star$ *satisfies* $d_{\Omega_v^\star}^{\max} = o(p_2)$*.*

Then, for random realizations of $\{X_t\}$*,* $\{Z_t\}$ *and* $\{V_t\}$*, and the sequence* $\{(\widehat{B}^{(k)}, \widehat{C}^{(k)}), \widehat{\Omega}_v^{(k)}\}_k$ *returned by Algorithm III.2 outlined in Section 3.2.1, there exist constants* $\{c_i, \tilde{c}_i\}, i = (0, 1, 2)$ *and* $\tau > 0$ *such that for sample size* $T \gtrsim (d_{\Omega_v^\star}^{\max})^2(p_1 + 2p_2)$*, with probability at least*

$$1 - c_0\exp\{-\tilde{c}_0(p_1+p_2)\} - c_1\exp\{-\tilde{c}_1(p_1+2p_2)\} - c_2\exp\{-\tilde{c}_2\log[p_2(p_1+p_2)]\} - \exp\{-\tau\log p_2\},$$

the following hold for all $k \geq 1$ *for* $C_0, C_0', C_0'' > 0$ *that are functions of the upper and lower extremes* $\mathcal{M}(f_W), \mathfrak{m}(f_W)$ *of the spectrum* $f_W(\theta)$ *and of the upper extreme* $\mathcal{M}(f_{W,V})$ *of the cross-spectrum* $f_{W,V}(\theta)$ *and do not depend on* p_1, p_2 *or* T*:*

(i) $\mathfrak{W}_{\widehat{\Omega}_v^{(k)}}$ *satisfies the RSC condition;*

(ii) $\|\mathbf{W}^\top \mathbf{V}\widehat{\Omega}_v^{(k)}/T\|_\infty \leq C_0\sqrt{\frac{(p_1+p_2)+p_2}{T}}$ *and* $|||\mathbf{W}^\top \mathbf{V}\widehat{\Omega}_v^{(k)}/T|||_{op} \leq C_0'\sqrt{\frac{(p_1+p_2)+p_2}{T}}$;

(iii) $\|\widehat{S}_v^{(k)} - \Sigma_v^\star\|_\infty \leq C_0''\sqrt{\frac{(p_1+p_2)+p_2}{T}}$.

As a consequence, the following bounds hold uniformly for all iterations $k \geq 1$*:*

$$|||\widehat{B}^{(k)} - B^\star|||_F^2 + |||\widehat{C}^{(k)} - C^\star|||_F^2 = O\Big(\tfrac{(r_B^\star + s_C^\star)(p_1+2p_2)}{T}\Big),$$
$$|||\widehat{\Omega}_v^{(k)} - \Omega_v^\star|||_F^2 = O\Big(\tfrac{(s_{\Omega_v}^\star + p_2)(p_1+2p_2)}{T}\Big).$$

Remark 3.3. It is worth pointing out that the initializers $\widehat{A}^{(0)}$ and $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ are slightly different from those obtained in successive iterations, as they come from the penalized least square formulation where the inverse covariance matrices are temporarily assumed diagonal. Consistency results for these initializers under deterministic realizations are established in Theorems B.1 and B.2 (see Appendix B.1), and the corresponding conditions are later verified for random realizations in Lemmas B.1 to B.4 (see Appendix B.2). These theorems and lemmas serve as the stepping stone toward the proofs of Theorems 3.1 and 3.2.

Further, the constants C_0, C_0', C_0'' reflect both the temporal dependence among X_t and Z_t blocks, as well as the cross-sectional dependence within and across the two blocks.

3.3.3 The effect of temporal and cross-dependence on the established bounds.

We conclude this section with a discussion on the error bounds of the estimators that provides additional insight into the impact of temporal and cross dependence within and between the blocks; specifically, how the exact bounds depend on the underlying processes through their spectra when explicitly taking into consideration the triangular structure of the joint transition matrix.

First, we introduce additional notations needed in subsequent technical developments. The definition of the spectral densities and the cross-spectrum are the same as previously defined in Section 3.2 and their upper and extremes are defined in Section 3.3.1. For $\{X_t\}$ defined in (3.1), let $\mathcal{A}(\theta) = \mathrm{I}_{p_1} - A\theta$ denote the characteristic matrix-valued polynomial of $\{X_t\}$ and $\mathcal{A}^*(\theta)$ denote its conjugate. We further define its upper and lower extremes by:

$$\mu_{\max}(\mathcal{A}) = \max_{|\theta|=1} \Lambda_{\max}\left(\mathcal{A}^*(\theta)\mathcal{A}(\theta)\right), \quad \mu_{\min}(\mathcal{A}) = \min_{|\theta|=1} \Lambda_{\min}\left(\mathcal{A}^*(\theta)\mathcal{A}(\theta)\right).$$

The same set of quantities for the joint process $\{W_t = (X_t', Z_t')'\}$ are analogously defined, that is,

$$\mu_{\max}(\mathcal{G}) = \max_{|\theta|=1} \Lambda_{\max}\left(\mathcal{G}^*(\theta)\mathcal{G}(\theta)\right), \quad \mu_{\min}(\mathcal{G}) = \min_{|\theta|=1} \Lambda_{\min}\left(\mathcal{G}^*(\theta)\mathcal{G}(\theta)\right).$$

Using the result in Theorem 3.2 as an example, we show how the error bound depends on the underlying processes $\{(X_t', Z_t')'\}$. Specifically, we note that the bounds for $(\widehat{B}^{(k)}, \widehat{C}^{(k)})$ can be equivalently written as

$$\left|\left|\left|\widehat{B}^{(k)} - B^\star\right|\right|\right|_F^2 + \left|\left|\left|\widehat{C}^{(k)} - C^\star\right|\right|\right|_F^2 \leq \bar{C}\Big(\tfrac{(r_B^\star + s_C^\star)(p_1+2p_2)}{T}\Big),$$

which holds for all k and some constant $\bar{C}$ that does not depend on p_1, p_2 or T. Specifically,

by Theorem B.2, Lemmas B.3 and B.4,

$$C_0 \propto \left[\mathcal{M}(f_W) + \tfrac{1}{2\pi}\Lambda_{\max}(\Sigma_v) + \mathcal{M}(f_{W,V^+})\right] / \mathfrak{m}(f_W).$$

where $\{V_t^+\} := \{V_{t+1}\}$ denotes the shifted V_t process. This indicates that the exact error bound depends on $\mathfrak{m}(f_W), \mathcal{M}(f_W)$ and $\mathcal{M}(f_{W,V^+})$. Next, we provide bounds on these quantities. The joint process W_t as we have noted in (3.2), is a VAR(1) process with characteristic polynomial $\mathcal{G}(\theta)$ and spectral density $f_W(\theta)$. The bounds for $\mathfrak{m}(f_W)$ and $\mathcal{M}(f_W)$ are given by **?** , Proposition 2.1], that is,

$$\mathfrak{m}(f_W) \geq \frac{\min\{\Lambda_{\min}(\Sigma_u), \Lambda_{\min}(\Sigma_v)\}}{(2\pi)\mu_{\max}(\mathcal{G})} \quad \text{and} \quad \mathcal{M}(f_W) \leq \frac{\max\{\Lambda_{\max}(\Sigma_u), \Lambda_{\max}(\Sigma_v)\}}{(2\pi)\mu_{\min}(\mathcal{G})}. \quad (3.18)$$

Consider the bound for $\mathcal{M}(f_{W,V^+})$. First, we note that $\{V_t\}$ is a sub-process of the joint error process $\{\varepsilon_t\}$, where $\varepsilon_t = (U_t', V_t')'$ and we additionally let $\{\varepsilon_t^+\} = \{\varepsilon_{t+1}\}$. Then, by Lemma B.12,

$$\mathcal{M}(f_{W,V^+}) \leq \mathcal{M}(f_{W,\varepsilon^+}) \leq \mathcal{M}(f_W)\mu_{\max}(\mathcal{G}),$$

where the second inequality follows from **?** , Proof of Proposition 2.4].

What are left to be bounded are $\mu_{\min}(\mathcal{G})$ and $\mu_{\max}(\mathcal{G})$. By Proposition 2.2 in **?**], these two quantities are bounded by:

$$\mu_{\max}(\mathcal{G}) \leq \left[1 + \frac{|||G|||_\infty + |||G|||_1}{2}\right]^2 \quad (3.19)$$

and

$$\mu_{\min}(\mathcal{G}) \geq (1 - \rho(G))^2 \cdot |||P|||_{\text{op}}^{-2} \cdot \left|||P^{-1}|||_{\text{op}}^{-2},\right.$$

where $G = P\Lambda_G P^{-1}$ with Λ_G being a diagonal matrix consisting of the eigenvalues of G. Since $|||P^{-1}|||_{\text{op}} \geq |||P|||_{\text{op}}^{-1}$, it follows that

$$|||P|||_{\text{op}}^{-2} \cdot \left|||P^{-1}|||_{\text{op}}^{-2} \geq \left|||P^{-1}|||_{\text{op}}^{2} \cdot \left|||P^{-1}|||_{\text{op}}^{-2} = 1,\right.\right.$$

and therefore

$$\mu_{\min}(\mathcal{G}) \geq \left(1 - \max\{\rho(A), \rho(C)\}\right)^2. \quad (3.20)$$

Remark 3.4. *The impact of the system's lower-triangular structure on the established bounds.* Consider the bounds in (3.19) and (3.20). An upper bound of $\mu_{\max}(\mathcal{G})$ depends on $|||G|||_\infty$ and $|||G|||_1$, whereas a lower bound of $\mu_{\min}(\mathcal{G})$ involves only the spectral radius of G. Combined with (3.18), this suggests that the lower extreme of the spectral density is associated with the average of the maximum weighted in-degree and out-degree of the system, whereas the

upper extreme is associated with the stability condition: the *less the system is intra- and inter-connected*, the *tighter the bound* for the lower extreme will be; similarly, the *more stable* (exhibits smaller temporal dependence) the system is, the *tighter the bound* for the upper extreme will be. Finally, we note that an upper bound for $(|||G|||_\infty + |||G|||_1)$ is given by

$$\max\{|||A|||_\infty + |||B|||_\infty, |||C|||_\infty\} + \max\{|||A|||_1, |||B|||_1 + |||C|||_1\}.$$

The presence of $|||B|||_\infty$ and $|||B|||_1$ depicts the role of the inter-connectedness between $\{X_t\}$ and $\{Z_t\}$ on the lower extreme of the spectrum, which is associated with the overall curvature of the joint process.

The impact of the system's lower-triangular structure on the system capacity. With G being a lower-triangular matrix, we only require $\rho(A) < 1$ and $\rho(C) < 1$ to ensure the stability of the system. This enables the system to have "larger capacity" (can accommodate more cross-dependence within each block), since the two sparse components A and C can exhibit larger average weighted in- and out-degrees compared with a system where G does not possess such triangular structure. In the case where G is a complete matrix, one deals with a $(p_1 + p_2)$-dimensional VAR system and $\rho(G) < 1$ is required to ensure its stability. As a consequence, the average weighted in- and out-degree requirements for each time series become more restrictive.

3.4 Testing Group Granger-Causality.

In this section, we develop a procedure for testing the hypothesis $H_0 : B = 0$. Note that without the presence of B, the blocks X_t and Z_t in the model become *decoupled* and can be treated as two separate VAR models, whereas with a nonzero B, the group of variables in Z_t is collectively "Granger-caused" by those in X_t. Moreover, since we are testing whether or not the entire block of B is zero, we do not need to rely on the exact distribution of its individual entries, but rather on the properly measured correlation between the responses and the covariates. To facilitate presentation of the testing procedure, we illustrate the proposed framework via a simpler model setting with $Y_t = \Pi X_t + \epsilon_t$ and testing whether $\Pi = 0$; subsequently, we translate the results to the actual setting of interest, namely, whether or not $B = 0$ in the model $Z_t = BX_{t-1} + CZ_{t-1} + V_t$.

The testing procedure focuses on the following sequence of tests for the rank of B:

$$H_0 : \text{rank}(B) \le r, \qquad \text{for an arbitrary} \quad r < \min(p_1, p_2). \tag{3.21}$$

Note that the hypothesis of interest, $B = 0$ corresponds to the special case with $r = 0$. To

test for it, we develop a procedure associated with *canonical correlations*, which leverages ideas present in the literature [see **?**].

As mentioned above, we consider a simpler setting similar to that in **? ?**], given by

$$Y_t = \Pi X_t + \epsilon_t,$$

where $Y_t \in \mathbb{R}^{p_2}$, $X \in \mathbb{R}^{p_1}$ and ϵ_t is independent of X_t. At the population level, let

$$\mathbb{E}Y_t Y_t^\top = \Sigma_Y, \qquad \mathbb{E}X_t X_t^\top = \Sigma_X, \qquad \mathbb{E}Y_t X_t^\top = \Sigma_{YX} = \Sigma_{XY}^\top.$$

The population canonical correlations between Y_t and X_t are the roots of

$$\begin{vmatrix} -\rho\Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & -\rho\Sigma_X \end{vmatrix} = 0,$$

i.e., the nonnegative solutions to

$$|\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} - \rho^2\Sigma_Y| = 0, \tag{3.22}$$

with ρ being the unknown. By the results in **? ?**], the number of positive solutions to (3.22) is equal to the rank of Π, and indicates the "degree of dependency" between processes Y_t and X_t. This suggests that if $\text{rank}(\Pi) \leq r < p$, we would expect $\sum_{k=r+1}^{p} \lambda_k$ to be small, where the λ's solve the eigen-equation

$$|S_{YX}S_X^{-1}S_{XY} - \lambda S_Y| = 0, \qquad \text{with} \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p,$$

and S_X, S_{XY} and S_Y are the sample counterparts corresponding to Σ_X, Σ_{XY} and Σ_Y, respectively.

With this background, we switch to our model setting given by

$$Z_t = BX_{t-1} + CZ_{t-1} + V_t, \tag{3.23}$$

where V_t is assumed to be independent of X_{t-1} and Z_{t-1}, B encodes the canonical correlation between Z_t and X_{t-1}, conditional on Z_{t-1}. We use the same notation as in Section 3.3; that is, let $\Gamma_X(h) = \mathbb{E}X_t X_{t+h}'$, $\Gamma_Z(h) = \mathbb{E}Z_t Z_{t+h}'$, and $\Gamma_{X,Z}(h) = \mathbb{E}X_t Z_{t+h}'$, with (h) omitted whenever $h = 0$. At the population level, under the Gaussian assumption,

$$\begin{bmatrix} Z_t \\ X_{t-1} \\ Z_{t-1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Gamma_Z & \Gamma_{X,Z}'(1) & \Gamma_Z(1) \\ \Gamma_{X,Z}(1) & \Gamma_X & \Gamma_{X,Z} \\ \Gamma_Z'(1) & \Gamma_{X,Z}' & \Gamma_Z \end{bmatrix} \right),$$

which suggests that conditionally,

$$Z_t \big| Z_{t-1} \sim \mathcal{N}\left(\Gamma_Z(1)\Gamma_Z^{-1} Z_{t-1}, \Sigma_{00}\right) \qquad \text{and} \qquad X_{t-1} \big| Z_{t-1} \sim \mathcal{N}\left(\Gamma_{X,Z}\Gamma_Z^{-1} Z_{t-1}, \Sigma_{11}\right),$$

where

$$\Sigma_{00} := \Gamma_Z - \Gamma_Z(1)\Gamma_Z^{-1}\Gamma_Z'(1) \qquad \text{and} \qquad \Sigma_{11} := \Gamma_X - \Gamma_{X,Z}\Gamma_Z^{-1}\Gamma_{X,Z}'. \tag{3.24}$$

Then, we have that jointly

$$\begin{bmatrix} Z_t \\ X_{t-1} \end{bmatrix} \Big| Z_{t-1} \sim \mathcal{N}\left(\begin{bmatrix} \Gamma_Z(1) \\ \Gamma_{X,Z} \end{bmatrix} \Gamma_Z^{-1} Z_{t-1} \, , \, \begin{bmatrix} \Gamma_Z & \Gamma_{XZ}'(1) \\ \Gamma_{XZ}(1) & \Gamma_Z \end{bmatrix} - \begin{bmatrix} \Gamma_Z(1) \\ \Gamma_{XZ} \end{bmatrix} \Gamma_Z^{-1} \begin{bmatrix} \Gamma_Z'(1) & \Gamma_{ZX} \end{bmatrix} \right),$$

so the partial covariance matrix between Z_t and X_{t-1} conditional on Z_{t-1} is given by

$$\Sigma_{10} = \Sigma_{01}' := \Gamma_{X,Z}(1) - \Gamma_Z(1)\Gamma_Z^{-1}\Gamma_{X,Z}. \tag{3.25}$$

The population canonical correlations between Z_t and X_{t-1} conditional on Z_{t-1} are the non-negative roots of

$$\left|\Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10} - \rho^2\Sigma_{00}\right| = 0,$$

and the number of positive solutions corresponds to the rank of B; see **?**] for a discussion in which the author is interested in estimating and testing linear restrictions on regression coefficients. Therefore, to test $\text{rank}(B) \leq r$, it is appropriate to examine the behavior of $\Psi_r := \sum_{k=r+1}^{\min(p_1,p_2)} \phi_k$, where ϕ's are ordered non-increasing solutions to

$$|S_{01}S_{11}^{-1}S_{10} - \phi S_{00}| = 0, \tag{3.26}$$

and S_{01}, S_{11} and S_{00} are the empirical surrogates for the population quantities Σ_{01}, Σ_{11} and Σ_{00}. For subsequent developments, we make the very mild assumption that $p_1 < T$ and $p_2 < T$ so that $\mathbf{Z}^\top\mathbf{Z}$ is invertible.

Proposition 3.1 gives the tail behavior of the eigenvalues and Corollary 3.1 gives the testing procedure for block "Granger-causality" as a direct consequence.

Proposition 3.1. *Consider the model setup given in* (3.23)*, where* $B \in \mathbb{R}^{p_2 \times p_1}$*. Further, assume all positive eigenvalues* μ *of the following eigen-equation are of algebraic multiplicity one:*

$$\left|\Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10} - \mu\Sigma_{00}\right| = 0, \tag{3.27}$$

where Σ_{00}, Σ_{11} *and* Σ_{01} *are given in* (3.24) *and* (3.25)*. The test statistic for testing*

$$H_0 : \textit{rank}(B) \leq r, \qquad \textit{for an arbitrary} \quad r < \min(p_1, p_2),$$

is given by

$$\Psi_r := \sum_{k=r+1}^{\min(p_1,p_2)} \phi_k,$$

where ϕ_k's are ordered decreasing solutions to the eigen-equation $|S_{01}S_{11}^{-1}S_{10} - \phi S_{00}| = 0$ where

$$S_{11} = \tfrac{1}{T}\mathbf{X}^\top(I - P_z)\mathbf{X}, \quad S_{00} = \tfrac{1}{T}\left(\mathbf{Z}_T\right)^\top\left(I - P_z\right)\left(\mathbf{Z}_T\right), \quad S_{10} = S_{01}' = \tfrac{1}{T}\mathbf{X}^\top(I - P_z)\left(\mathbf{Z}_T\right),$$

and $P_z = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$. Moreover, the limiting behavior of Ψ_r is given by

$$T\Psi_r \sim \chi^2_{(p_1-r)(p_2-r)}.$$

Remark 3.5. We provide a short comment on the assumption that the positive solutions to (3.27) have algebraic multiplicity one in Proposition 3.1. This assumption is imposed on the eigen-equation associated with population quantities, to exclude the case where a positive root has algebraic multiplicity greater than one and its geometric multiplicity does not match the algebraic one, and hence we would fail to obtain r mutually independent canonical variates and the rank-r structure becomes degenerate. With the imposed assumption which is common in the canonical correlation analysis literature [e.g. **?** **?**], such a scenario is automatically excluded. Specifically, this condition is not stringent, as for ϕ's that are solutions to the eigen-equation associated with sample quantities, the distinctiveness amongst roots is satisfied with probability 1 [see **?** , Proof of Lemma 3].

Corollary 3.1 (Testing group Granger-causality)**.** *Under the model setup in* (3.23)*, the test statistic for testing $B = 0$ is given by*

$$\Psi_0 := \sum_{k=1}^{\min(p_1,p_2)} \phi_k,$$

with ϕ_k being the ordered decreasing solutions of

$$\left|S_{01}\left[\operatorname{diag}(S_{11})\right]^{-1}S_{10} - \phi S_{00}\right| = 0.$$

Asymptotically, $T\Psi_0 \sim \chi^2_{p_1p_2}$. To conduct a level α test, we reject the null hypothesis H_0 : $B = 0$ if

$$\Psi_0 > \frac{1}{T}\chi^2_{p_1p_2}(\alpha),$$

where $\chi^2_d(\alpha)$ is the upper α quantile of the χ^2 distribution with d degrees of freedom.

Remark 3.6. Corollary 3.1 is a special case of Proposition 3.1 with the null hypothesis being $H_0 : r = 0$, which corresponds to the Granger-causality test. Under this particular setting, we are able to take the inverse with respect to $\text{diag}(S_{11})$, yet maintain the same asymptotic distribution due to the fact that $S_{01} = S_{10} = 0$ under the null hypothesis $B = 0$. This enables us to perform the test even with $p_1 > T$.

The above testing procedure takes advantage of the fact that when $B = 0$, the canonical correlations among the partial regression residuals after removing the effect of Z_{t-1} are very close to zero. However, the test may not be as powerful under a sparse alternative, i.e., $H_A : B$ is sparse. In Appendix B.4, we present a testing procedure that specifically takes into consideration the fact that the alternative hypothesis is sparse, and the corresponding performance evaluation is shown in Section 3.5.3 under this setting.

3.5 Performance Evaluation.

Next, we present the results of numerical studies to evaluate the performance of the developed ML estimates (Section 3.2.1) of the model parameters, as well as that of the testing procedure (Section 3.4).

3.5.1 Simulation results for the estimation procedure.

A number of factors may potentially influence the performance of the estimation procedure; in particular, the model dimension p_1 and p_2, the sample size T, the rank of $B^\star$ and the sparsity level of $A^\star$ and $C^\star$, as well as the spectral radius of $A^\star$ and $C^\star$. Hence, we consider several settings where these parameters vary.

For all settings, the data $\{x_t\}_t$ and $\{z_t\}_t$ are generated according to the model

$$x_t = A^\star x_{t-1} + u_t,$$
$$z_t = B^\star x_{t-1} + C^\star z_{t-1} + v_t.$$

For the sparse components, each entry in $A^\star$ and $C^\star$ is nonzero with probability $2/p_1$ and $1/p_2$ respectively, and the nonzero entries are generated from $\mathsf{Unif}\left([-2.5, -1.5] \cup [1.5, 2.5]\right)$, then scaled down so that the spectral radii $\rho(A)$ and $\rho(C)$ satisfy the stability condition. For the low rank component, each entry in $B^\star$ is generated from $\mathsf{Unif}(-10, 10)$, followed by singular value thresholding, so that $\text{rank}(B^\star)$ conforms with the model setup. For the contemporaneous dependence encoded by $\Omega_u^\star$ and $\Omega_v^\star$, both matrices are generated according to an Erdös-Rényi random graph, with sparsity being 0.05 and condition number being 3.

Table 3.1 depicts the values of model parameters under different model settings. Specifically, we consider three major settings in which the size of the system, the rank of the cross-dependence component, and the stability of the system vary. The sample size is fixed at $T = 200$ unless otherwise specified. Additional settings examined (not reported due to space considerations) are consistent with the main conclusions presented next.

		model parameters				
		p_1	p_2	rank(B^*)	ρ_A	ρ_C
model dimension	A.1	50	20	5	0.5	0.5
	A.2	100	50	5	0.5	0.5
	A.3	200	50	5	0.5	0.5
	A.4	50	100	5	0.5	0.5
rank	B.1	100	50	10	0.5	0.5
	B.2	100	50	20	0.5	0.5
spectral radius	C.1	50	20	5	0.8	0.5
	C.2	50	20	5	0.5	0.8
	C.3	50	20	5	0.8	0.8

Table 3.1: Model parameters under different model settings.

We use sensitivity (SEN), specificity (SPC) and relative error in Frobenius norm (Error) as criteria to evaluate the performance of the estimates of transition matrices A, B and C. Tuning parameters are chosen based on BIC. Since the exact contemporaneous dependence is not of primary concern, we omit the numerical results for $\widehat{\Omega}_u$ and $\widehat{\Omega}_v$.

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{SPE} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \quad \text{Error} = \frac{|||\text{Est.} - \text{Truth}|||_{\text{F}}}{|||\text{Truth}|||_{\text{F}}}.$$

Table 3.2 illustrates the performance for each of the parameters under different simulation settings considered. The results are based on an average of 100 replications and their standard deviations are given in parentheses.

	performance of $\widehat{A}$			performance of $\widehat{B}$		performance of $\widehat{C}$		
	SEN	SPC	Error	rank($\widehat{B}$)	Error	SEN	SPC	Error
A.1	0.98(.014)	0.99(.004)	0.34(.032)	5.2(.42)	0.11(.008)	1.00(.000)	0.97(.008)	0.15(.074)
A.2	0.97(.014)	0.99(.001)	0.38(.015)	5.2(.42)	0.31(.011)	0.97(.008)	0.97(.004)	0.28(.033)
A.3	0.99(.005)	0.96(.002)	0.87(.011)	5.8(.92)	0.54(.022)	0.98(.000)	0.92(.009)	0.28(.028)
A.4	0.96(.026)	0.99(.002)	0.36(.034)	5.2(.42)	0.32(.012)	0.95(.009)	0.98(.001)	0.37(.010)
B.1	0.97(.008)	0.99(.001)	0.37(.017)	11.4(1.17)	0.15(.008)	1.00(.000)	0.99(.001)	0.09(.021)
B.2	0.98(.008)	0.99(.001)	0.38(.016)	21.2(.91)	0.12(.006)	1.00(.000)	0.99(.001)	0.08(.018)
C.1	1.00(.000)	0.97(.005)	0.25(.015)	5.6(.52)	0.23(.006)	1.00(.000)	0.92(.021)	0.11(.072)
C.2	0.99(.007)	0.95(.004)	0.45(.022)	5.0(.00)	0.31(.014)	1.00(.000)	0.92(.019)	0.04(.011)
C.3	1.00(.000)	0.96(.004)	0.18(.013)	6.7(1.16)	0.19(.011)	1.00(.000)	0.87(.029)	0.14(.067)
C.3'	1.00(.000)	0.99(.002)	0.13(.016)	5.2(.42)	0.23(.005)	1.00(.000)	0.90(.021)	0.06(.023)

Table 3.2: Performance evaluation of $\widehat{A}$, $\widehat{B}$ and $\widehat{C}$ under different model settings.

Overall, the results are highly satisfactory and all the parameters are estimated with a high degree of accuracy. Further, all estimates were obtained in less than 20 iterations, thus indicating that the estimation procedure is numerically stable. As expected, when the the spectral radii of A and C increase thus leading to less stable $\{X_t\}$ and $\{Z_t\}$ processes, a

larger sample size is required for the estimation procedure to match the performance of the setting with same parameters but smaller $\rho(A)$ and $\rho(C)$. This is illustrated in row C.3' of Table 3.2, where the sample size is increased to $T = 500$, which outperforms the results in row C.3 in which $T = 200$ and broadly matches that of row A.1.

Next, we investigate the robustness of the algorithm in settings where the marginal distributions of $\{X_t\}$ and $\{Z_t\}$ deviate from the Gaussian assumption posited and may be more heavy-tailed. Specifically, we consider the following two distributions that have been studied in **?**]:

- t-distribution: the idiosyncratic error processes $\{u_t\}$ and $\{v_t\}$ are generated from multivariate t-distributions with degree of freedom 3, and covariance matrices $(\Omega_u^\star)^{-1}$ and $(\Omega_v^\star)^{-1}$, respectively.

- elliptical distribution: $(u_1', \ldots, u_T')$ and $(v_1', \ldots, v_T')'$ are generated from an elliptical distribution [e.g. **?**] with a log-normal generating variate $\log \mathcal{N}(0, 2)$ and covariance matrices $\widetilde{\Sigma}_u$ and $\widetilde{\Sigma}_v$ – both are block-diagonal with $\Sigma_u^\star = (\Omega_u^\star)^{-1}$ and $\Sigma_v^\star = (\Omega_v^\star)^{-1}$ respectively on the diagonals.

For both scenarios, transition matrices, $\Omega_u^\star$ and $\Omega_v^\star$ are generated analogously to those in the Gaussian setting. We present the results for $\widehat{A}, \widehat{B}$ and $\widehat{C}$ under model settings A.2, B.1, C.1 and C.2 (see Table 3.1).

		performance of $\widehat{A}$			performance of $\widehat{B}$		performance of $\widehat{C}$		
		SEN	SPC	Error	rank($\widehat{B}$)	Error	SEN	SPC	Error
A.2	t(df=3)	0.99(.005)	0.95(.013)	0.60(.062)	6.00(1.45)	0.24(.019)	0.96(.013)	0.96(.005)	0.27(.038)
	elliptical	0.97(.014)	0.99(.001)	0.36(.016)	5.1(.30)	0.34(.009)	1.00(.000)	0.85(.026)	0.15(.033)
B.1	t(df=3)	0.98(.008)	0.95(.014)	0.61(.083)	10.4(.49)	0.34(.026)	0.99(.015)	0.95(.004)	0.25(.091)
	elliptical	0.95(.015)	0.99(.001)	0.37(.024)	10.1(.22)	0.40(.013)	1.00(.000)	0.90(.013)	0.09(.001)
C.1	t(df=3)	0.99(.001)	0.92(.011)	0.22(.03)	6.0(1.13)	0.09(.014	1.00(.000)	0.93(.016)	0.10(.068)
	elliptical	1.00(.000)	0.90(.006	0.32(.013)	5.2(.44)	0.13(.007)	1.00(.000)	0.92(.020)	0.07(.041)
C.2	t(df=3)	0.99 (.002)	0.95(.023)	0.37(.056)	5.1(.22)	0.22(.017)	1.00(.000)	0.89(.017)	0.10(.029)
	elliptical	0.88(.029)	0.97(.001)	0.43(.032)	5.1(.14)	0.40(.020)	1.00(.000)	0.86(.026)	0.10(.046)

Table 3.3: Performance evaluation of $\widehat{A}$, $\widehat{B}$ and $\widehat{C}$ under non-Gaussian settings.

Based on Table 3.3, the performance of the estimates under these heavy-tailed settings is comparable in terms of sensitivity and specificity for A and C, as well as for rank selection for B to those under Gaussian settings. However, the estimation error exhibits some deterioration which is more pronounced for the t-distribution case. In summary, the estimation procedure proves to be very robust for support recovery and rank estimation even in the presence of more heavy-tailed noise terms.

Lastly, we examine performance with respect to one-step-ahead forecasting. Recall that VAR models are widely used for forecasting purposes in many application areas [**?**]. The

performance metric is given by the relative error as measured by the ℓ_2 norm of the out-of-sample points x_{T+1} and z_{T+1}, where the predicted values are given by $\widehat{x}_{T+1} = \widehat{A}x_T$ and $\widehat{z}_{T+1} = \widehat{B}x_T + \widehat{C}z_T$, respectively. It is worth noting that both $\{X_t\}$ and $\{Z_t\}$ are mean-zero processes. However, since the transition matrix of $\{X_t\}$ is subject to the spectral radius constraints to ensure the stability of the corresponding process, the magnitude of the realized value x_t's is small; whereas for $\{Z_t\}$, since no constraints are imposed on the B coefficient matrix that encodes the inter-dependence, z_t's has the capacity of having relative large values in magnitude. Consequently, the relative error of $\widehat{x}_{T+1}$ is significantly larger than that of $\widehat{z}_{T+1}$, partially due to the small total magnitude of the denominator.

The results show that an increase in the spectral radius (keeping the other structural parameters fixed) leads to a decrease of the relative error, since future observations become more strongly correlated over time. On the other hand, an increase in dimension leads to a deterioration in forecasting, since the available sample size impacts the quality of the parameter estimates. Finally, an increase in the rank of the B matrix is beneficial for forecasting, since it plays a stronger role in the system's temporal evolution.

		$\frac{\|\widehat{x}_{T+1}-x_{T+1}\|_2}{\|x_{T+1}\|_2}$	$\frac{\|\widehat{z}_{T+1}-z_{T+1}\|_2}{\|z_{T+1}\|_2}$
baseline	A.1	0.89(.066)	0.23(.075)
spectral radius	C.1	0.62(.100)	0.10(.035)
	C.2	0.93(.062)	0.17(.059)
	C.3	0.68(.096)	0.10(.045)
rank	B.1	0.92(.044)	0.14(.038)
	B.2	0.94(.042)	0.14(.025)
dimension	A.2	0.87(.051)	0.24(.073)
	A.3	0.96(.040)	0.44(.139)
	A.4	0.89(.059)	0.274(.068)

Table 3.4: One-step-ahead relative forecasting error.

3.5.2 A comparison between the two-step and the ML estimates.

We briefly compare the ML estimates to the ones obtained through the following two-step procedure:

– Step 1: estimate transition matrices through penalized least squares:

$$\widehat{A}^{\text{t-s}} = \arg\min_{A} \ \left\{ \tfrac{1}{T} \left\| \mathbf{X}_T - \mathbf{X}A^\top \right\|_{\text{F}}^2 + \lambda_A \|A\|_1 \right\},$$
$$(\widehat{B}^{\text{t-s}}, \widehat{C}^{\text{t-s}}) = \arg\min_{(B,C)} \ \left\{ \tfrac{1}{T} \left\| \mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top \right\|_{\text{F}}^2 + \lambda_B \left\| B \right\|_* + \lambda_C \|C\|_1 \right\}.$$

– Step 2: estimate the inverse covariance matrices applying the graphical Lasso algorithm [**?**] to the residuals calculated based on the Step 1 estimates.

Note that the two-step estimates coincide with our ML estimates at iteration 0, and they yield the same error rate in terms of the relative scale of p_1, p_2 and T. We compare the two sets of estimates under setting A.1 with $B^\star$ being low rank and setting A.2 with $B^\star$ being sparse, whose entries are nonzero with probability $1/p_1$.

In Table 3.5, we present the performance evaluation of the two-step estimates and the ML estimates under setting A.1. Additionally in Tables 3.6 and 3.7, we track the value of the objective function, the relative error (in $\|\cdot\|_F$) and the cardinality (or rank) of the estimates along iterations, with iteration 0 corresponding to the two-step estimates. A similar set of results is shown in Tables 3.8 to 3.10 for setting A.2, but with a sparse $B^\star$. All other model parameters are identically generated according to the procedure described in Section 3.5.1.

As the results show, the ML estimates clearly outperform their two-step counterparts, in terms of the relative error in Frobenius norm. On the other hand, both sets of estimates exhibit similar performance in terms of sensitivity and specificity and rank specification. More specifically, when estimating A, the ML estimate decreases the false positive rate (higher SPC). Under setting A.1, while estimating B and C, both estimates correctly identify the rank of B, and the ML estimate provides a more accurate estimate in terms of the magnitude of C, at the expense of incorrectly including a few more entries in its support set; under setting A.2 with a sparse $B^\star$, improvements in both the relative error of B and C are observed. In particular, due to the descent nature of the algorithm, we observe a sharp drop in the value of the objective function at iteration 1, as well as the most pronounced change in the estimates.

	performance of $\widehat{A}$			performance of $\widehat{B}$		performance of $\widehat{C}$		
	SEN	SPC	Error	Error	rank($\widehat{B}$)	SEN	SPC	Error
two-step estimates	0.97	0.95	0.52	0.27	5	1.00	0.98	0.12
ML estimates	0.97	0.97	0.36	0.24	5	1.00	0.95	0.05

Table 3.5: Performance comparison under A.1 with a low-rank B.

iteration	0	1	2	3	4	5
Rel.Error	0.521	0.408	0.376	0.360	0.359	0.359
Cardinality	227	169	160	155	155	155
Value of Obj	128.14	41.74	37.94	37.85	37.70	37.70

Table 3.6: Relative error of $\widehat{A}$ and the values of the objective function under A.1.

iteration	0	1	2	3	4	5	6	7	8	9	10
Rel.Error of $\widehat{B}$	0.274	0.235	0.236	0.237	0.237	0.237	0.237	0.237	0.237	0.237	0.237
Rank of $\widehat{B}$	5	5	5	5	5	5	5	5	5	5	5
Rel.Error of $\widehat{C}$	0.119	0.050	0.049	0.049	0.048	0.048	0.048	0.047	0.047	0.047	0.047
Cardinality of $\widehat{C}$	30	34	38	41	39	39	39	39	39	39	39
Value of Obj	160.41	134.26	131.90	131.48	131.38	131.17	131.01	130.96	130.96	130.85	130.8

Table 3.7: Relative error of $\widehat{B}$,$\widehat{C}$ and the values of the objective function under A.1.

	performance of $\widehat{A}$			performance of $\widehat{B}$			performance of $\widehat{C}$		
	SEN	SPC	Error	SEN	SPC	Error	SEN	SPC	Error
two-step estimates	0.97	0.95	0.44	0.96	0.98	0.45	1	0.99	0.35
ML estimates	0.97	0.98	0.35	0.99	0.95	0.34	1	0.98	0.30

Table 3.8: Performance comparison under A.2 with a sparse B.

iteration	0	1	2	3	4
Rel.Error	0.438	0.346	0.351	0.351	0.351
Cardinality	479	350	325	324	324
Value of Obj	156.58	48.25	38.16	36.97	36.93

Table 3.9: Relative error of $\widehat{A}$ and the values of the objective function under A.2.

iteration	0	1	2	3	4
Rel.Error of $\widehat{B}$	0.454	0.340	0.337	0.337	0.337
Cardinality of $\widehat{B}$	301	325	323	323	323
Rel.Error of $\widehat{C}$	0.35	0.304	0.302	0.302	0.301
Cardinality of $\widehat{C}$	63	70	74	74	74
Value of Obj	143.942	59.63	41.46	41.87	41.87

Table 3.10: Changes over iteration under A.2.

3.5.3 Simulation results for the block Granger-causality test.

Next, we illustrate the empirical performance of the testing procedure introduced in Section 3.4, together with the alternative one (in Appendix B.4) when B is sparse, with the null hypothesis being $B^\star = 0$ and the alternative being $B^\star \neq 0$, either low rank or sparse. Specifically, when the alternative hypothesis is true and has a low-rank structure, we use the general testing procedure proposed in Section 3.4, whereas when the alternative is true and sparse, we use the testing procedure presented in Appendix B.4. We focus on evaluating the type I error (empirical false rejection rate) when $B^\star = 0$, as well as the power of the test when $B^\star$ has nonzero entries.

For both testing procedures, the transition matrix $A^\star$ is generated with each entry being nonzero with probability $2/p_1$, and the nonzeros are generated from $\mathsf{Unif}\left([-2.5,-1.5]\cup[1.5,2.5]\right)$, then further scaled down so that $\rho(A^\star) = 0.5$. For transition matrix $C^\star$, each entry is nonzero with probability $1/p_2$, and the nonzeros are generated from $\mathsf{Unif}\left([-2.5,-1.5]\cup[1.5,2.5]\right)$, then further scaled down so that $\rho(C^\star) = 0.5$ or 0.8, depending on the simulation setting. Finally, we only consider the case where v_t and u_t have diagonal covariance matrices.

We use sub-sampling as in **?**] and **?**] with the number of subsamples set to 3,000; an alternative would have been a block bootstrap procedure [e.g. **? ? ?**]. Note that the length of the subsamples varies across simulation settings in order to gain insight on how sample size impacts the type I error or the power of the test.

Low-rank testing. To evaluate the type I error control and the power of the test, we primarily consider the case where $\text{rank}(B^\star) = 0$, with the data alternatively generated based on $\text{rank}(B^\star) = 1$. We test the hypothesis $H_0 : \text{rank}(B) = 0$ and tabulate the empirical proportion of falsely rejecting H_0 when $\text{rank}(B^*) = 0$ (type I error) and the probability that we reject H_0 when $\text{rank}(B^*) = 1$ (power). In addition, we also show how the testing procedure performs when the underlying $B^\star$ has rank $r \geq 0$. In particular, when $\text{rank}(B^\star) = r^\star$, the type I error of the test corresponds to the empirical proportion of rejections of the null hypothesis $H_0 : r \leq r^\star$, while the power of the test to the empirical proportion of rejections of the null hypothesis set to $H_0 : r \leq (r^\star - 1)$. The latter resembles the sequential test in **?**].

Empirically, we expect that when $B^\star = 0$, the value of the proposed test statistic mostly falls below the cut-off value (upper α quantile), while when $\text{rank}(B^\star) = 1$, the value of the proposed test statistic mostly falls beyond the critical value $\chi^2(\alpha)_{p_1p_2}/T$ with T being the sample size, hence leading to a detection. Table 3.11 gives the type I error of the test when setting $\alpha = 0.1, 0.05, 0.1$, and the power of the test using the upper 0.01 quantile of the reference distribution as the cut-off, for different combinations of model dimensions (p_1, p_2) and sample size.

	(p_1, p_2)	sample size	type I error ($B^\star = 0$) $\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	power ($\text{rank}(B^\star) = 1$) cut-off $\chi^2(0.01)_{p_1p_2}/T$
$\rho(C^\star) = 0.5$	$(20, 20)$	$T = 500$	0.028	0.123	0.227	1
		$T = 1000$	0.015	0.073	0.137	1
		$T = 2000$	0.011	0.059	0.118	1
	$(50, 20)$	$T = 500$	0.070	0.228	0.355	1
		$T = 1000$	0.026	0.125	0.226	1
		$T = 2000$	0.013	0.094	0.163	1
	$(20, 50)$	$T = 500$	0.484	0.751	0.857	1
		$T = 1000$	0.089	0.246	0.375	1
		$T = 2000$	0.020	0.088	0.164	1
	$(100, 50)$	$T = 500$	0.997	0.999	1	1
		$T = 1000$	0.608	0.828	0.908	1
		$T = 2000$	0.166	0.374	0.511	1
$\rho(C^\star) = 0.8$	$(20, 50)$	$T = 500$	0.533	0.789	0.880	1
		$T = 1000$	0.130	0.306	0.452	1
		$T = 2000$	0.045	0.145	0.252	1
	$(50, 20)$	$T = 500$	0.083	0.250	0.382	1
		$T = 1000$	0.039	0.133	0.234	1
		$T = 2000$	0.019	0.096	0.174	1
			type I error ($H_0 : r \leq 5$) $\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	power ($H_0 : r \leq 4$) cut-off $\chi^2(0.01)_{(p_1-4)(p_2-4)}/T$
$\rho(C^\star) = 0.5$ $\text{rank}(B^\star) = 5$	$(20, 50)$	$T = 500$	0.092	0.274	0.400	1
		$T = 1000$	0.034	0.140	0.236	1
		$T = 2000$	0.022	0.096	0.178	1
	$(50, 20)$	$T = 500$	0.454	0.722	0.829	1
		$T = 1000$	0.126	0.313	0.452	1
		$T = 2000$	0.062	0.184	0.284	1

Table 3.11: Empirical type I error and power for low-rank testing.

Based on the results shown in Table 3.11, it can be concluded that the proposed low-rank testing procedure accurately detects the presence of "Granger causality" across the

two blocks, when the data have been generated based on a truly multi-layer VAR system. Further, when $B^\star = 0$, the type I error is close to the nominal α level for sufficiently large sample sizes, but deteriorates for increased model dimensions. In particular, relatively large values of p_2 and larger spectral radius $\rho(C^\star)$ negatively impact the empirical false rejection proportion, which deviates from the desired control level of the type I error. In the case where $\text{rank}(B^\star) = r > 0$, the testing procedure provides satisfactory type I error control for larger sample sizes and excellent power.

Sparse testing. Since the rejection rule of the HC-statistic is based on empirical process theory [**?**] and its dependence on α is not explicit, we focus on illustrating how the empirical proportion of false rejections (type I error) varies with the sample size T, the model dimensions (p_1, p_2) and the spectral radius of $C^\star$. To show the power of the test, each entry in $B^\star$ is nonzero with probability $q \in (0, 1)$ such that $q(p_1p_2) = (p_1p_2)^\theta$ with $\theta = 0.6$, to ensure the overall sparsity of $B^\star$ satisfies the sparsity requirement posited in Proposition B.1. The magnitude is set such that the signal-to-noise ratio is 1.2. Note that the actual number of parameters is p_1p_2, while the total number of subsamples used is 3000 with the length of subsamples varying according to different simulation settings to demonstrate the dependence of type I error and power on sample sizes.

		type I error ($B^\star = 0$)				power (SNR($B^\star$) = 0.8)			
	(p_1, p_2)	200	500	1000	2000	200	500	1000	2000
$\rho(C^\star) = 0.5$	$(20, 20)$	0.244	0.097	0.074	0.055	1	1	1	1
	$(50, 20)$	0.393	0.131	0.108	0.074	1	1	1	1
	$(20, 50)$	0.996	0.351	0.153	0.093	1	1	1	1
	$(100, 50)$	1.000	0.963	0.270	0.115	1	1	1	1
$\rho(C^\star) = 0.8$	$(50, 20)$	0.402	0.158	0.112	0.075	0.829	0.996	1	1
	$(20, 50)$	0.999	0.430	0.166	0.111	1	1	1	1

Table 3.12: Empirical type I error and power for sparse testing.

Based on the results shown in Table 3.12, when $B^\star = 0$, the proposed testing procedure can effectively detect the absence of block "Granger causality", provided that the sample size is moderately large compared to the total number of parameters being tested. However, if the model dimension is large, whereas the sample size is small, the test procedure becomes problematic and fails to provide legitimate type I error control, as desired. When $B^\star$ is nonzero, empirically the test is always able detect its presence, as long as the effective signal-to-noise ratio is beyond the detection threshold.

3.6 Real Data Analysis Illustration.

We employ the developed framework and associated testing procedures to address one of the motivating applications. Specifically, we analyze the temporal dynamics of the log-

returns of stocks with large market capitalization and key macroeconomic variables, as well as their cross-dependence. Specifically, using the notation in (3.1), we assume that the X_t block consists of the stock log-returns, while the macroeconomic variables form the Z_t block. With this specification, we assume that the macroeconomic variables are "Granger-caused" by the stock market, but not vice versa. Note that our framework allows us to pose and test a more general question than previous work in the economics literature considered. For example, **?**] building on previous work by [**?** **?**] tests only the relationship between the employment index and the composite stock index, using a bivariate VAR model. On the other hand, our framework enables us to consider the components of the S&P 100 index and the "medium" list of macroeconomic variables considered in the work of **?**].

Next, we provide a brief description of the data used. The stock data consist of monthly observations of 71 stocks corresponding to a stable set of historical components comprising the S&P 100 index for the 2001-2016 period. The macroeconomic variables are chosen from the "medium" list in **?**]; that is, the 3 core economic indicators (Fed Funds Rate, Consumer Price Index and Gross Domestic Product Growth Rate), plus 17 additional variables with aggregate information (e.g., exchange rate, employment, housing, etc.). However, in our study, we exclude variables that exhibit a significant change after the financial crisis of 2008 (e.g. total reserves/reserves of depository institutions). We process the macroeconomic variables to ensure stationarity following the recommendations in **?**]. As a general guideline, for real quantitative variables (e.g., GDP, money supply M2), we use the relative first difference, which corresponds to their growth rate, while for rate-related variables (e.g., Federal Funds Rate, unemployment rate), we use their first differences directly. The complete lists of stocks and macroeconomic variables used in this study are given in Appendix B.7.

We start the analysis by using the VAR model for the stock log-returns to study their evolution over the 2001-2016 period. Analogously to the strategy employed by **?**], we consider 36-month long rolling-windows for fitting the model $X_t = AX_{t-1} + U_t$, for a total of 143 estimates of the transition matrix A. VAR models involving more than 1 lag were also fitted to the data, but did not indicate temporal dependence beyond lag 1.

To obtain the final estimates across all 143 subsamples, we employ *stability selection* [**?**], with the threshold set at 0.6 for including an edge in A.[2] Figure 3.2 depicts the global clustering coefficient [**?**] of the skeleton of the estimated A over all 143 rolling windows, with the time stamps on the horizontal axis specifying the starting time of the corresponding window.

The results clearly indicate strong connectivity in lead-lag stock relationships during the financial crisis period March 2007-June 2009. It is of interest that the data exhibit such

[2]The threshold is set at a relatively low level to compensate for the relative small rolling window size.

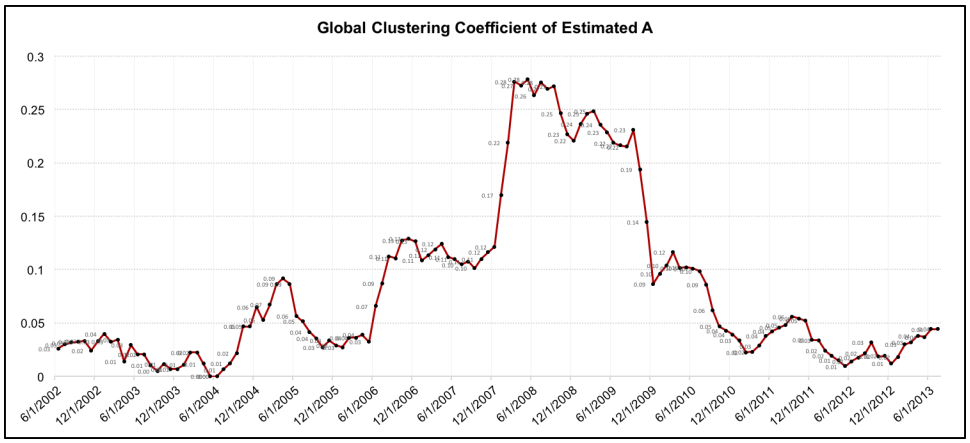

Figure 3.2: Global clustering coefficient of estimated A over different periods

sharp changes at time points that correspond to well documented events in the literature; namely, March 2007 when several subprime lenders declared bankruptcy, put themselves up for sale or reported significant losses and June 2009 that the National Bureau of Economic Research declared as the end point of the crisis. Similar patterns were broadly observed in **?** **?**], albeit for a different set of stocks and using a different statistical methodology. Specifically, **?**] considered financial sector stocks (banks, brokerages, insurance companies), while **?**] considered European banking stocks, and both studies used bivariate VAR models to obtain the results, thus ignoring the influence of all other components on the pairwise Granger causal relationship estimated and hence producing potentially biased estimates of connectivity.

Next, we present the analysis based on the VAR-X component of our model, given by $Z_t = BX_{t-1} + CZ_{t-1} + V_t$ with the stock log-returns corresponding to the X_t block and the (stationary) macroeconomic variables to the Z_t block. As before, we fit the data within each rolling window, with the tuning parameters based on a search over a 10×10 lattice (with $(\lambda_B, \lambda_C) \in [0.5, 4] \times [0.2, 2]$, equal-spaced) using the BIC. It should be noted that for the majority of the rolling windows, the rank of B is 1 (data not shown). The sparsity level of the estimated C over the 143 rolling windows is depicted in Figure 3.3. The connectivity patterns in C show more complex and nuanced patterns than for stocks. Several local peaks depicted correspond to the following events: (i) March-April 2003, when the Federal Reserve cut the Effective Federal Funds Rate aggressively driving it down to 1%, the lowest level in 45 years up to that point, (ii) January-March 2008, a significant decline in the major US stock indexes, coupled with the freezing up of the market for auctioning rate securities with investors declining to bid, (iii) January-April 2009, characterizes the unfolding of the European debt crisis with a rescue package put together for Greece and significant downgrades of Spanish and Portuguese sovereign debt by rating agencies and (iv) July 2010, that correspond to the

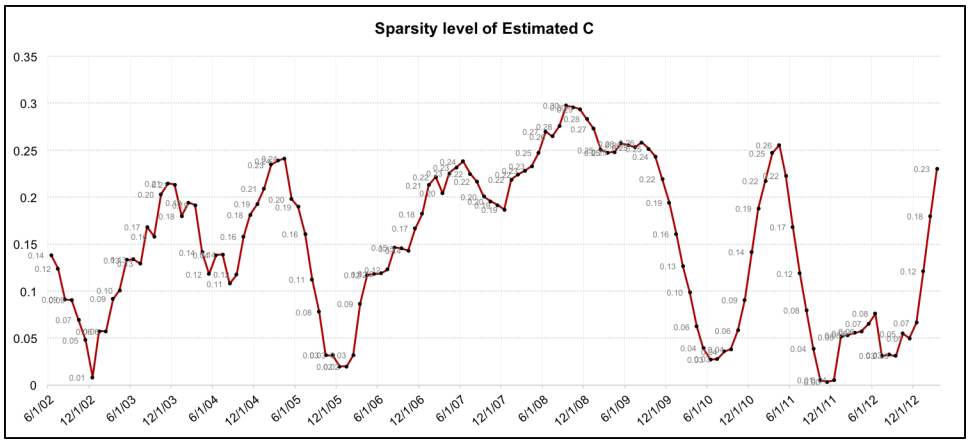

Figure 3.3: Sparsity of estimated C over different periods

enactment of the Dodd-Frank Wall Street Reform and Consumer Protection Act and the acceleration of quantitative easing by the Federal Reserve Board.

Based on the previous findings, we partition the time frame spanning 2001-2016 into the following periods: pre- (2001/07–2007/03), during- (2007/01–2009/12) and post-crisis (2010/01-2016/06) one. We estimate the model parameters using the data within the entire sub-period(s).

The estimation procedure of the transition matrix A for different periods is identical to that described above using subsamples over rolling-windows. For the pre- and post- crisis periods, since we have 76 and 77 samples respectively, the stability selection threshold is set at 0.75, whereas for the during-crisis period, at 0.6 to compensate for the small sample size (36). Table 3.13 shows the average R-square for all 71 stocks, as well as its standard deviation, which is calculated based on in-sample fit; i.e.,the proportion of variation explained by using the VAR(1) model to fit the data. The overall sparsity level and the spectral radius of the estimated transition matrices A are also presented. The results are consistent with the previous finding of increased connectivity during the crisis. Further, for all periods the estimate of the spectral radius is fairly large, indicating strong temporal dependence of the log-returns.

	2001/07–2007/03	2007/01–2009/12	2010/01–2016/06
Averaged R sq	0.31	0.72	0.28
Sd of R sq	0.103	0.105	0.094
Sparsity level of $\widehat{A}$	0.17	0.23	0.19
Spectral radius of $\widehat{A}$	0.67	0.90	0.75

Table 3.13: Summary for estimated A within different periods.

Figures 3.7 to 3.9 depict the estimated transition matrices A for different periods, as a network, with edges thresholded based on their magnitude for ease of presentation. The

node or edge coloring red/blue indicates the sign positive/negative of the corresponding entry in the transition matrix. Further, node size is proportional to the out-degree, thus indicating which stocks influence other stocks in the next time period. The most striking feature is the outsize influence exerted by the insurance company AIG and the investment bank Goldman Sachs, whose role during the financial crisis has been well documented [**?**]. On the other hand, the pre- and post-crisis periods are characterized by more sparse and balanced networks, in terms of in- and out-degree magnitude.

Next, we focus on the key motivation for developing the proposed modeling framework, namely the inter-dependence of stocks and macroeconomic variables over the specified three sub-periods. The p-value for testing the hypothesis of lack of block "Granger causality" $H_0 : B = 0$, together with the spectral radius and the sparsity level for the estimated C transition matrices are listed in Table 3.14. Specifically, for all three periods, the rank of estimated B is 1, indicating that the stock market as captured by its leading stocks, "Granger-causes" the temporal evolution of the macroeconomic variables. The fact that the rank of B is 1, indicates that the inter-block influence can be captured as a single portfolio acting in unison. To investigate the relative importance of each sector in the portfolio, we group the stocks by sectors. The proportion of each sector (up to normalization) is obtained by summing up the loadings (first right singular vector of the estimated B) of the stocks within this sector, weighted by their market capitalization. Further, the estimated transition matrices C's are depicted in network form, in Figures 3.4 to 3.6. It is worth noting that the temporal correlation of the macroeconomic variables significantly increased during the crisis.

Note that the proportion of various sectors in the portfolio is highly consistent with their role in stock market. For example, before crisis the financial sector had a large market capitalization (roughly 20%), while it shrunk (to roughly 12%) after the crisis. Also, the Information Technology (IT) and Financial (FIN) sectors are the ones exhibiting highest volatility (high beta) relative to the market, while the Utilities is the one with low volatility (low beta) , a well established stylized fact in the literature for the time period under consideration.

Next, we discuss some key relationships emerging from the model. We start with total employment (ETTL), whose dynamics are only influenced by its own past values as seen by the lack of an incoming arrow in Figure 3.5. Further, an examination of the left singular vector (see Table 3.15) of B strongly indicates the impact of the stock market on total employment. This finding is consistent with the analysis in **?**], which argues that the crash of the stock market provides a plausible explanation for the great recession. However, the analysis in **?**] is based on bivariate VAR models involving only employment and the stock index. Therefore, there is a possibility that the stock market is reacting to some other

information captured by other macroeconomic variables, such as GDP, capital spending, inflation, interest rates, etc. However, our high-dimensional VAR model simultaneously analyzes a key set of macroeconomic variables and also accounts for the influence of the largest stocks in the market. Hence, it automatically overcomes the criticism leveraged by [**?**] about misinterpretations of findings from small scale VAR models due to the omission of important variables, and further echoed in the discussion in [**?**].

Another interesting finding is the strong influence of the stock market on GDP in the pre- and post-crisis period, consistent with the popular view of being a leading indicator for GDP growth. Further, capital utilization is positively impacted during the crisis period by GDP growth and total employment—which are both falling and hence reducing capital utilization—and further accentuated by the impact of the stock market—also falling—thus reinforcing the lack of available capital goods and resources.

In summary, the brief analysis presented above provides interesting insights into the interaction of the stock market with the real economy, identifies a number of interesting developments during the crisis period and reaffirms a number of findings studied in the literature, while ensuring that a much larger information set is utilized (a larger number of variables included) than in previous analysis. Therefore, high-dimensional multi-block VAR models are useful for analyzing complex temporal relationships and provide insights into their dynamics.

	2001/07–2007/03	2007/01–2009/12	2010/01–2016/16
p-value for testing $H_0 : B = 0$	0.075	0.009	0.044
Sparsity level of $\widehat{C}$	0.06	0.25	0.06
Spectral radius of $\widehat{C}$	0.35	0.76	0.40

Table 3.14: Summary for estimated B and C within different periods.

	Pre-Crisis	During-Crisis	Post-Crisis
FFR	-0.24	-0.26	-0.23
T10yr	-0.09	0.14	0.16
UNEMPL	-0.07	0.01	-0.07
IPI	-0.43	0.34	0.26
ETTL	0.33	0.24	0.13
M1	0.23	-0.12	-0.47
AHES	-0.01	0.30	0.17
CU	-0.49	0.32	0.27
M2	0.10	-0.04	-0.32
HS	0.51	-0.02	-0.02
EX	-0.18	0.41	0.06
PCEQI	-0.07	-0.18	0.41
GDP	0.10	-0.02	0.05
PCEPI	0.00	0.14	-0.01
PPI	-0.15	0.00	0.06
CPI	0.01	0.15	-0.31
SP.IND	-0.06	-0.53	0.38

Table 3.15: Left singular vectors of estimated B for different periods.

Remark 3.7. We also applied our multi-block model with the first block X_t corresponding to the macro-economic variables and the second block Z_t the stocks variables (results not shown). The key question is whether there is also "Granger causality" from the broader economy to the stock market. The results are inconclusive due to sample size issues that do not allow us to properly test for the key hypothesis whether $B = 0$ or not. Specifically, the length of the sub-periods is short compared to the dimensionality required for the test procedure. A similar issue arises, which is related to the detection boundary for the sparse testing procedure during the crisis period. Further, for a sparse B, an examination of its entries shows that Employment Total did not impact the stock market, which is in line with the conclusion reached at the aggregate level by **?**]. On the other hand, GDP negatively impacts stock log-returns, which may act as a leading indicator for suppressed investment and business growth and hence future stock returns.

3.7 Discussion.

We briefly discuss generalizations of the model to the case of more than two blocks, as mentioned in the introductory section. For the sake of concreteness, consider a triangular recursive linear dynamical system given by:

$$\begin{aligned} X_t^{(1)} &= A_{11}X_{t-1}^{(1)} + \epsilon_t^{(1)}, \\ X_t^{(2)} &= A_{12}X_{t-1}^{(1)} + A_{22}X_{t-1}^{(2)} + \epsilon_t^{(2)}, \\ X_t^{(3)} &= A_{13}X_{t-1}^{(1)} + A_{23}X_{t-1}^{(2)} + A_{33}X_{t-1}^{(3)} + \epsilon_t^{(3)}, \\ &\vdots \end{aligned} \tag{3.28}$$

where $X^{(j)} \in \mathbb{R}^{p_j}$ denotes the variables in group j, A_{ij} $(i < j)$ encodes the dependency of $X^{(j)}$ on the past values of variables in group i, and A_{jj} encodes the dependency on its own past values. Further, $\{\epsilon_t^{(j)}\}$ is the innovation process that is neither temporally, nor cross-sectionally correlated, i.e.,

$$\mathrm{Cov}(\epsilon_t^{(j)}, \epsilon_s^{(j)}) = 0 \ (s \neq t), \quad \mathrm{Cov}(\epsilon_t^{(i)}, \epsilon_s^{(j)}) = 0 \ (i \neq j, \ \forall \ (s,t)), \quad \mathrm{Cov}(\epsilon_t^{(j)}, \epsilon_t^{(j)}) = \left(\Omega^{(j)}\right)^{-1},$$

with $\Omega^{(j)}$ capturing the conditional contemporaneous dependency of variables within group j. The model in (3.28) can also be viewed from a multi-layered time-varying network perspective: nodes in each layer are "Granger-caused" by nodes from its previous layers, and are also dependent on its own past values. As previously mentioned, in various real applications, it is of interest to obtain estimates of the transition matrices, and/or test if "Granger-causality" is present between interacting blocks; i.e., to test $A_{ij} = 0$ for some $i \neq j$.

The triangular structure of the system decouples the estimation of the transition matrices from each equation, and hence a straightforward extension of the estimation procedure presented in Section 3.2.1 becomes applicable. Specifically, to obtain estimates of the transition matrices A_{ij}'s for fixed j and $1 \leq i \leq j$, and the inverse covariance $\Omega^{(j)}$, the optimization problem is formulated as follows:

$$(\{\widehat{A}_{ij}\}_{i\leq j}, \widehat{\Omega}^{(j)}) = \underset{A_{ij},\Omega^{(j)}}{\arg\min} \Big\{ -\log\det\Omega^{(j)} + \frac{1}{T}\sum_{t=1}^{T}\big(x_t^{(j)} - \sum_{i=1}^{j} A_{ij}x_{t-1}^{(i)}\big)^{\top}\Omega^{(j)}\big(x_t^{(j)} - \sum_{1\leq i\leq j} A_{ij}x_{t-1}^{(i)}\big) + \sum_{i=1}^{j}\mathcal{R}(A_{ij}) + \rho^{(j)}\|\Omega^{(j)}\|_{1,\text{off}}\Big\}, \tag{3.29}$$

where the exact expression for the $\mathcal{R}(A_{ij})$ adapts to the structural assumption imposed on the corresponding transition matrix (sparse/low-rank). Solving (3.29) again requires an iterative algorithm involving the alternate update between transition matrices and the inverse covariance matrices. Further, for updating the values of the transition matrices, a cyclic block-coordinate updating procedure is used.

Consistency results can be established analogously to those provided in Section 3.3, under the posited conditions of restricted strong convexity (RSC) and a deviation bound. With a larger number of interacting blocks of variables, lower bounds for the lower extremes of the spectra involve all corresponding transition matrices. The error rates that can be obtained are as follows: (i) if equation k only involves sparse transition matrices, then the finite-sample bounds of the transition matrices in this layer in Frobenius norm are of the order $O\big(\sqrt{\frac{\log p_k + \log\sum_{i\leq k} p_k}{T}}\big)$, while (ii) if some of the transition matrices are assumed low rank, then the corresponding finite sample bounds are of the order $O\big(\sqrt{\frac{p_k + \sum_{i\leq k} p_k}{T}}\big)$.

Another generalization that can be handled algorithmically with the same estimation procedure discussed above is the presence of d-lags in the specification of the linear dynamical system. Based on the consistency results developed in this work, together with the theoretical findings for VAR(d) models presented in **?**], we expect all the established theoretical properties of the transition matrices estimates to go through under appropriate RSC and deviation bound conditions.

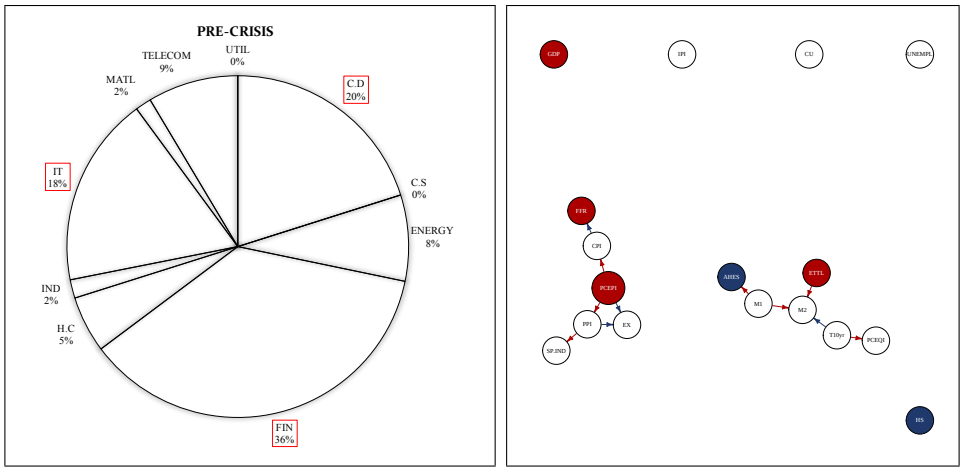

Figure 3.4: Sector proportion and estimated C for pre-crisis period.

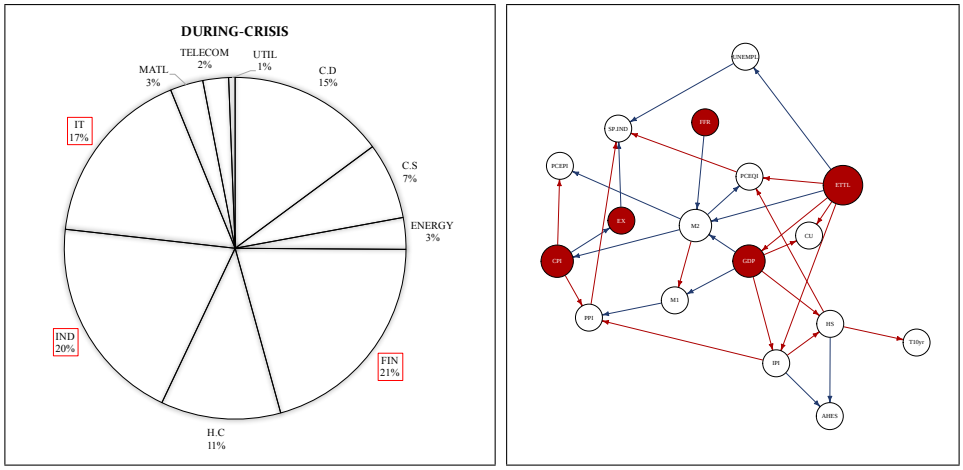

Figure 3.5: Sector proportion and estimated C for during-crisis period.

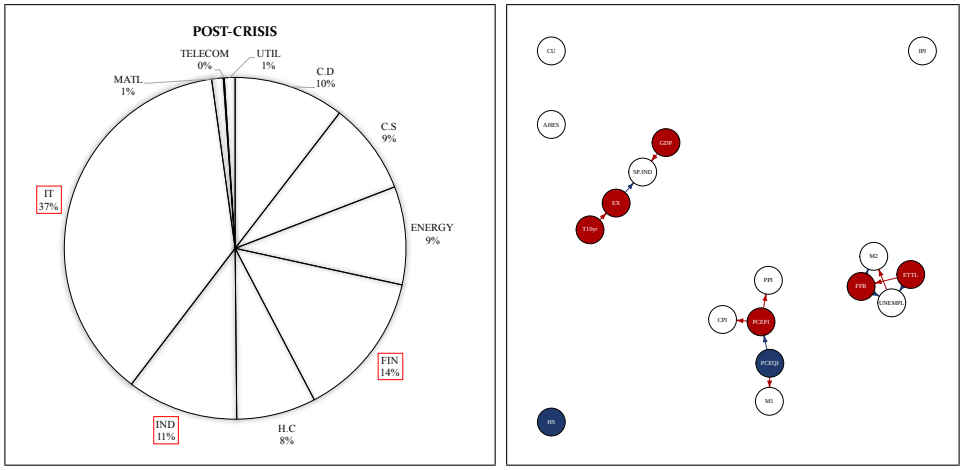

Figure 3.6: Sector proportion and estimated C for post-crisis period.

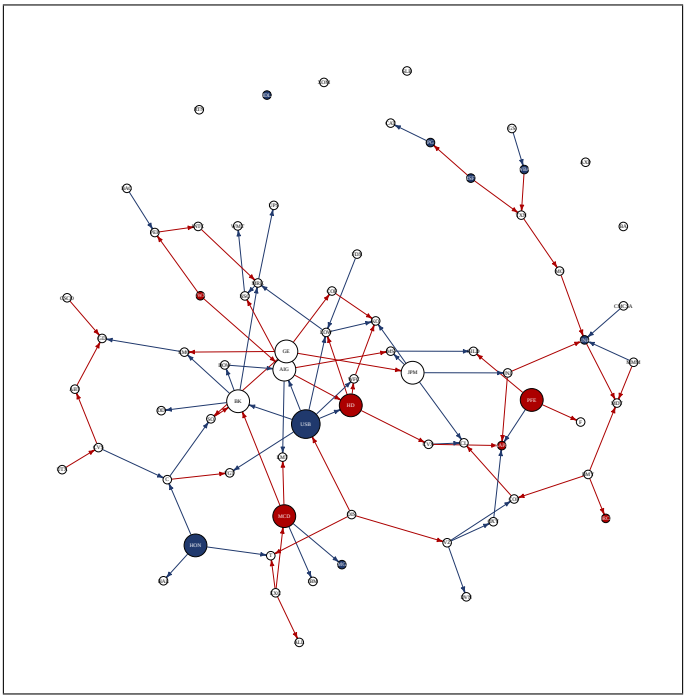

Figure 3.7: Estimated transition matrix for stock dynamics between 2001 to 2007.

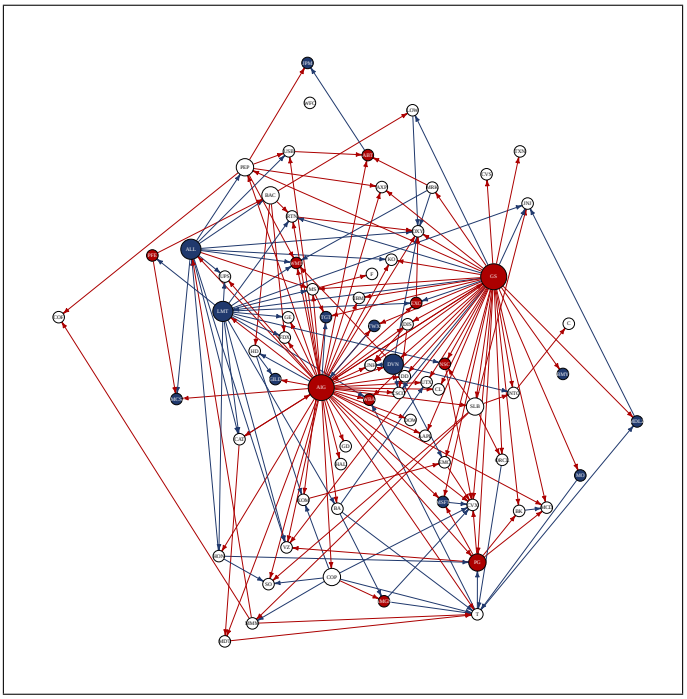

Figure 3.8: Estimated transition matrix for stock dynamics between 2007 to 2009.

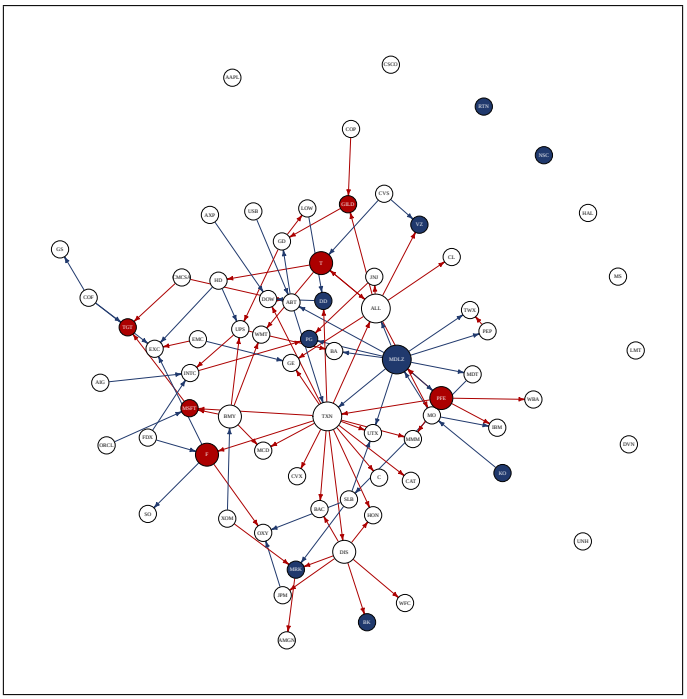

Figure 3.9: Estimated transition matrix for stock dynamics between 2010 to 2016.

CHAPTER IV

Regularized Estimation of High-dimensional Factor-Augmented Vector Autoregressive (FAVAR) Models

4.1 Introduction.

There is a growing need in employing a large set of time series (variables) for modeling social or physical systems. For example, economic policy makers have concluded based on extensive empirical evidence [e.g. **?** **?** **?**] that large scale models of economic indicators provide improved forecasts, together with better estimates of how current economic shocks propagate into the future, which produces better guidance for policy actions. Another reason for considering large number of time series in social sciences is that key variables implied by theoretical models for policy decisions[1] are not directly observable, but related to a large number of other variables that collectively act as a good proxy of the unobservable key variables. In other domains such as genomics and neuroscience, advent of high throughput technologies have enabled researchers to obtain measurements on hundreds of genes from functional pathways of interest [**?**] or brain regions [**?**], thus allowing a more comprehensive modeling to gain insights into biological mechanisms of interest. There are two popular modeling paradigms for such large panel of time series, with the first being the Vector Autoregressive (VAR) model [**?**] and the second being the Dynamic Factor Model (DFM) [**?** **?**].

The VAR model has been the subject of extensive theoretical and empirical work primarily in econometrics, due to its relevance in macroeconomic and financial modeling. However, the number of model parameters increases quadratically with the number of time series included for each lag period considered, and this feature has limited its applicability since in many applications it is hard to obtain adequate number of time points for accurate estimation. Nevertheless, there is a recent body of technical work that leveraging *structured*

[1] such as the concept of output gap for monetary policy, the latter defined as the difference between the actual output of an economy and its potential output

sparsity and the corresponding regularized estimation framework has established results for consistent estimation of the VAR parameters under high dimensional scaling. **?**] examined Lasso penalized Gaussian VAR models and proved consistency results, while providing technical tools useful for analysis of sparse models involving temporally dependent data. [**?**] extended the results to other regularizers, **?**] to the inclusion of exogenous variables (the so-called VAR-X model in the econometrics literature), [**?**] to models for count data and **?**] to the simultaneous estimation of time lags and model parameters. However, a key requirement for the theoretical developments is a spectral radius constraint that ensures the *stability* of the underlying VAR process [see **?** **?** , for details]. For large VAR models, this constraint implies a smaller magnitude on average for all model parameters, which makes their estimation more challenging, unless one compensates with a higher level of sparsity. Nevertheless, very sparse VAR models may not be adequately informative, while their estimation requires larger penalties that in turn induce higher bias due to shrinkage, when the sample size stays fixed.

The DFM model aims to decompose a large number of time series into a few common latent factors and idiosyncratic components. The premise is that these common factors are the key drivers of the observed data, which themselves can exhibit temporal dynamics. They have been extensively used for forecasting purposes in economics [**?**], while their statistical properties have been studied in depth [see **?** , and references therein]. Despite their ability to handle very large number of time series, theoretically appealing properties and extensive use in empirical work in economics, DFMs aggregate the underlying time series and hence are not suitable for examining their individual cross-dependencies. Since in many applications researchers are primarily interested in understanding the interactions between key variables [**?** **?**], while accounting for the influence of many others so as to avoid model misspecification that leads to biased results, DFMs may not be the most appropriate model.

To that end, **?**] came up with a compromise model, called the Factor Augmented VAR, that aims to summarize the information contained in a large set of time series by a small number of factors and include those in a standard VAR. Specifically, let $\{F_t\} \in \mathbb{R}^{p_1}$ be the latent factor and $\{X_t\} \in \mathbb{R}^{p_2}$ the observed sets of variables, they jointly form a VAR system given by

$$\begin{bmatrix} F_t \\ X_t \end{bmatrix} = A^{(1)} \begin{bmatrix} F_{t-1} \\ X_{t-1} \end{bmatrix} + \cdots + A^{(d)} \begin{bmatrix} F_{t-d} \\ X_{t-d} \end{bmatrix} + \begin{bmatrix} w_t^F \\ w_t^X \end{bmatrix}. \tag{4.1}$$

In addition, there is a large panel of observed time series $Y_t \in \mathbb{R}^q$, whose current values are influenced by both X_t and F_t:

$$Y_t = \Lambda F_t + \Gamma X_t + e_t. \tag{4.2}$$

Hence, the primary variables of interest X_t together with the unobserved factors F_t—both are assumed to have small and fixed dimensions—drives the dynamics of the system, and the factors are inferred from (4.2).

Note that there is very limited theoretical work [e.g. **?**] on the FAVAR model and some work on identification restrictions for the model parameters [e.g. **?**]. However, the fixed dimensionality $(p_1 + p_2)$ assumption is rather restrictive in many applications as discussed next. The model has been extensively used in empirical work in economics and finance [e.g. **?** **?**], where customarily a very small size block X_t is considered. For example, in the paper that introduced the FAVAR model [**?**] X_t comprises of three "core" economic indicators (industrial production, consumer price index and the federal funds rate) and Y_t of 120 other economic indicators. The VAR model considered is augmented by one factor summarizing the macroeconomic indicators and its dependence over time involves 7 lags, thus increasing the sample size requirement for its estimation. In a recent application, **?**] applies the FAVAR model to macroeconomics effects of oil supply shocks, the VAR model comprises of 8 times series (observed and latent), but due to the limitation in sample size to avoid non-stationarities ($T = 120$) the lag of the model is fixed to 1. Hence, as argued in **?**] there is a growing need for large scale FAVAR models and this papers aims to examine their estimation in high-dimensions, leveraging sparsity constraints on key model parameters.

The key contributions of this chapter are the investigation of the theoretical properties of estimates of the FAVAR model parameters under high-dimensional scaling, together with the introduction of an identifiability constraint compatible with the high-dimensional nature of the model. At the technical level there are two sets of challenges that are successfully resolved: (i) the calibration equation involves both an observed set of predictor variables and a set of latent factors, and their interactions require careful handling to enable accurate estimation of the factors, which is crucial for estimating the transition matrix since they constitute part of the input to the VAR system; and (ii) the presence of a block of variables in the VAR system that are subject to error due to it being an estimated quantity introduces a number of technical challenges, which are compounded by the presence of temporal dependence.

Outline of the chapter. The remainder of the chapter is organized as follows. In Section 4.2, the model identifiability constraint is introduced, followed by formulation of the objective function to be optimized that obtains estimates of the model parameters. Theoretical properties of the proposed estimators, specifically, their high probability finite-sample error bounds, are investigated in Section 4.3. Subsequently in Section 4.4, we introduce an empirical implementation procedure for obtaining the estimates and present its performance

evaluation based on synthetic data. An application of the model on interlinkages of commodity prices and the influence of world macroeconomic indicators on them is presented in Section 4.5, while Section 4.6 provides some concluding remarks. All proofs and other supplementary materials are deferred to Appendix C.

Notations. Throughout this chapter, we use $|||A|||_{\cdot}$ to denote matrix norms for some generic matrix $A \in \mathbb{R}^{m\times n}$. For example, $|||A|||_1$ and $|||A|||_\infty$ respectively denote the matrix induced 1-norm and infinity norm, $|||A|||_{\text{op}}$ the matrix operator norm and $|||A|||_{\text{F}}$ the Frobenius norm. Moreover, We use $\|A\|_1$ and $\|A\|_\infty$ respectively to denote the element-wise 1-norm and infinity norm. For two matrices A and B of commensurate dimensions, denote their inner product by $\langle\!\langle A, B\rangle\!\rangle = \text{tr}(A^\top B)$. Finally, we write $A \gtrsim B$ if there exists some absolute constant c that is independent of the model parameters such that $A \geq cB$; and $A \asymp B$ if $A \gtrsim B$ and $B \gtrsim A$ hold simultaneously.

4.2 Model Identification and Problem Formulation.

The FAVAR model proposed in **?**] has the following two components, as seen in Section 4.1: a system given in (4.1) that describes the dynamics of the latent block $F_t \in \mathbb{R}^{p_1}$ and the observed block $X_t \in \mathbb{R}^{p_2}$ that jointly follow a stationary VAR(d) model (the "VAR equation"); and the model in (4.2) that characterizes the contemporaneous dependence of the large observed informational series $Y_t \in \mathbb{R}^q$ as a linear function of X_t and F_t (the "calibration equation"). Further, w_t^F, w_t^X and e_t are all noise terms that are independent of the predictors, and we assume they are serially uncorrelated mean-zero Gaussian random vectors: $w_t^F \sim \mathcal{N}(0, \Sigma_w^F)$, $w_t^X \sim \mathcal{N}(0, \Sigma_w^X)$ and $e_t \sim \mathcal{N}(0, \Sigma_e)$. In this study we consider a potentially large VAR system that has many coordinates, hence in contrast to **?**] and **?**] where both p_1 and p_2 are fixed and small, we allow the size of the observed block, p_2, to be large[2] and to grow with the sample size; yet the size of the latent block, p_1, can not be too large and is still assumed fixed. Moreover, the size of the informational series, q, can also be large and grow with the sample size. Further, we assume that the transition matrices $\{A^{(i)}\}_{i=1}^d$ and the regression coefficient matrix Γ are *sparse*. Finally, the factor loading matrix Λ is assumed to be dense.

[2]We do not impose the restriction that p_2 is smaller than the available sample size.

4.2.1 Model identification considerations.

The latent nature of F_t leads to the following observational equivalence across the following two models: for any invertible matrix $Q_1 \in \mathbb{R}^{p_1 \times p_1}$ and $Q_2 \in \mathbb{R}^{p_1 \times p_2}$,

$$Y_t = \Lambda F_t + \Gamma X_t + e_t \;\equiv\; \widetilde{\Lambda}\widetilde{F}_t + \widetilde{\Gamma} X_t + e_t,$$

where

$$\widetilde{\Lambda} := \Lambda Q_1, \qquad \widetilde{F}_t := Q_1^{-1} F_t - Q_1^{-1} Q_2 X_t, \qquad \widetilde{\Gamma} := \Gamma + \Lambda Q_2. \tag{4.3}$$

Hence, the key model parameters (Λ, Γ) and the latent factors F_t are *not uniquely* identified, a known problem even in classical factor analysis [**?**]. Thus, additional restrictions are required to overcome this indeterminacy, since there is an equivalence class indexed by (Q_1, Q_2) within which individual models are not mutually distinguishable based on observational data. For the FAVAR model, a total number of $p_1^2 + p_1 p_2$ restrictions are needed for unique identification of Λ, Γ and F_t.

Various schemes have been proposed in the literature to address this issue. Specifically, **?**] imposes the necessary restrictions through the coefficient matrices of the calibration equation, requiring $\Lambda = \left[\begin{smallmatrix} \mathrm{I}_{p_1} \\ * \end{smallmatrix}\right]$ and $\Gamma_{[1:p_1],\cdot} = 0$; that is, the upper $p_1 \times p_1$ block of Λ is set to the identity matrix and the first p_1 rows of Γ to zero. **?**] considers different sets of restrictions (respectively labeled as IRa, IRb and IRc), all involving parameters from both the calibration and the VAR equations; in particular, $p_1 p_2$ of the total restrictions required are imposed through $\mathrm{Cov}(w_t^X, w_t^F) = O$ and the remaining p_1^2 ones are imposed in an analogous fashion to those in classical factor analysis.

In the low-dimensional setting (p_2 fixed), one can proceed to estimate the parameters subject to these restrictions. For example, **?**] uses a single-step Bayesian likelihood approach that fully incorporates their proposed identifiability restrictions, yet is computationally intensive. The procedure in **?**] requires the projection onto the orthogonal space spanned by samples of X_t as the very first step and the inverse matrix associated with of the sample covariance of w_X^t for further rotation. However, in high-dimensional settings, the growing dimension p_2 of the observed block X_t will further exacerbate the computational inefficiency of the aforementioned Bayesian approach. Further, neither the projection step nor the matrix inversion one are possible, which automatically renders the estimation procedure proposed in **?**] infeasible[3].

Next, we introduce an alternative identification scheme **(IR+)** that is compatible with the model specification and can also be seamlessly incorporated in the estimation procedure.

[3]For a full account of the estimation procedure in **?**] and why it fails to go through in high dimensions, see Appendix C.4

First, we require:

(IR) $\Lambda = \left[\begin{smallmatrix} \mathrm{I}_{p_1} \\ * \end{smallmatrix}\right]$: the upper $p_1 \times p_1$ block of Λ is an identity matrix, while the bottom block is left unconstrained.

Note that (IR) only involves p_1^2 constraints and yields uniquely identifiable Λ and F, for any given product ΛF_t. Note that the latent factor under (IR) remains completely unrestricted which is desirable given its use in the VAR system. The (IR) constraint corresponds to a commonly employed identifiability scheme in classical factor analysis [e.g. **?**]. Specifically, with (IR), the indeterminacy incurred by $Q_1 \in \mathbb{R}^{p_1 \times p_1}$ in (4.3) vanishes; however, the issue is not fully resolved, since for any $Q_2 \in \mathbb{R}^{p_1 \times p_2}$, the following relationship holds:

$$Y_t = \Lambda F_t + \Gamma X_t + e_t \equiv \Lambda \check{F}_t + \check{\Gamma} X_t + e_t,$$

where

$$\check{F}_t = F_t - Q_2 X_t \qquad \check{\Gamma} := \Gamma + \Lambda Q_2. \tag{4.4}$$

All such models encoded by $(\check{F}_t, \check{\Gamma})$, form an equivalence class indexed by Q_2 that specifies the transformation. We denote this equivalence class by $\mathcal{C}(Q_2)$ and the magnitude of Q_2 can be interpreted as a rough measure of discordance between the true data-generating model encoded by (F_t, Γ) and those encoded by $(\check{F}_t, \check{\Gamma})$. In particular, such discordance becomes zero when $Q_2 = O$ and $\mathcal{C}(Q_2)$ degenerates to a singleton that contains only the true data-generating model, which requires the imposition of $p_1 p_2$ restrictions on primary model quantities. For example, as previously mentioned, **?**] impose the restrictions through Γ by constraining its first p_1 rows to be equal to zero. Nevertheless, from a model perspective it translates to expressing the first p_1 coordinates of Y_t as noisy versions of F_t, which in turn makes it difficult to appropriately choose those coordinates in applications. **?**] requires $\mathrm{Cov}(w_t^X, w_t^F) = O$ which resolves the identifiability issue at the population level, but this constraint can not be operationalized in the high-dimensional setting as explained above.

An applicable constraint to high-dimensional settings is given by $\mathrm{Cov}(F_t, X_t) = O$ which yields the necessary $p_1 p_2$ restrictions. Yet, it is excessively stringent and limits the appeal of the FAVAR model, while also being challenging to operationalize. Therefore as a good working alternative, we address the identifiability issue through a weaker constraint that effectively limits sufficiently the size of the $\mathcal{C}(Q_2)$.

To this end, we first let $\mathbf{X} \in \mathbb{R}^{n \times p_2}$, $\mathbf{Y} \in \mathbb{R}^{n \times q}$ and $\mathbf{F} \in \mathbb{R}^{n \times p_1}$ be centered data matrices whose rows are samples of X_t, Y_t and the latent process F_t respectively, and $\check{\mathbf{F}}$ is analogously defined. The characterization of $\mathcal{C}(Q_2)$ is through the sample versions of the underlying

processes. Specifically, define the set of *factor hyperplanes* induced by $\mathcal{C}(Q_2)$ by

$$\mathcal{S}(\check{\Theta}) := \{\check{\Theta} := \check{\mathbf{F}}\Lambda^\top \mid \check{\mathbf{F}} \text{ are samples of } \check{F}_t \text{ defined through (4.4)}\}.$$

Further, let Θ (without the check) denote the factor hyperplane associated with the true data-generating model, to distinguish it from some generic element in $\mathcal{S}(\check{\Theta})$ that is denoted by $\check{\Theta}$. Note that $\Theta \in \mathcal{S}(\check{\Theta})$ and $\check{\Theta}$ coincides with Θ when $Q_2 = 0$. In addition, we require that all elements in $\mathcal{S}(\check{\Theta})$ to satisfy the following constraint:

(Compactness) $\|\check{\Theta}/\sqrt{n}\|_{\text{op}} \leq \phi$, that is, the largest singular value of $(\check{\Theta}/\sqrt{n})$ does not exceed a pre-specified value ϕ.

(Compactness) limits the spikiness of all possible $\check{\Theta}$'s by imposing a *box constraint* on their eigen-spectra, and restricts the factor hyperplane set induced by $\mathcal{C}(Q_2)$ to its ϕ-radius subset $\mathcal{S}_\phi(\check{\Theta})$, where

$$\mathcal{S}_\phi(\check{\Theta}) := \{\|\check{\Theta}/\sqrt{n}\|_{\text{op}} \leq \phi \mid \check{\Theta} \in \mathcal{S}(\check{\Theta})\}.$$

This in turn limits the size of the equivalence class $\mathcal{C}(Q_2)$ under consideration, since there is a one-to-one correspondence at the set level between $\mathcal{C}(Q_2)$ and the factor hyperplane set induced by it. Since $\Theta \in \mathcal{S}(\check{\Theta})$, $\phi \geq \phi_0 := \|\Theta/\sqrt{n}\|_{\text{op}}$. The $\sqrt{n}$ factor is introduced to reflect proper scaling with respect to the available number of samples. Note that this constraint also indirectly limits the magnitude of Q_2, since by singular value inequalities[4] and (4.4), we get

$$\|\mathbf{X}Q_2^\top\Lambda^\top/\sqrt{n}\|_{\text{op}} - \|\Theta/\sqrt{n}\|_{\text{op}} \leq \|\check{\Theta}/\sqrt{n}\|_{\text{op}} \leq \phi.$$

The above gives that

$$\|Q_2\|_{\text{op}} \leq \frac{\phi + \phi_0}{\sigma_{\min}(\mathbf{X}/\sqrt{n})\sigma_{\min}(\Lambda)}, \tag{4.5}$$

where $\sigma_{\min}$ denotes the smallest nonzero singular value that comes from the reduced SVD of the corresponding matrix. Even though the bound in (4.5) may not be the tightest, it nevertheless imposes an effective constraint on Q_2, since it no longer allows Q_2 to take arbitrary values in the set of $p_1 \times p_2$ matrices. Consequently, the size of the equivalence class $\mathcal{C}(Q_2)$ is also limited, which implies that although the models encoded by (F_t, Γ) and $(\check{F}_t, \check{\Gamma})$ may not be perfectly distinguishable based on observational data, at the population level the discordance between the two models can not be too large.

In summary, our proposed identification scheme (IR+) entails two parts: (IR) and (Compactness). The former provides exact identification within the factor hyperplane and narrows

[4]For two generic matrices A and B of commensurate dimensions, let $\sigma_1 \geq \sigma_2 \geq \cdots$ denote their singular values in decreasing order, then the following inequality holds: $\sigma_i(A+B) \geq \sigma_i(A) - \sigma_1(B)$. This can be derived from Theorem 3.4.1 in **?**].

the scope of observationally equivalent models to $\mathcal{C}(Q_2)$, while the latter limits its size. Hence, (IR+) can be viewed as an *approximate identification* scheme of the true data generating model.

Thus, for estimation purposes henceforth, it becomes adequate to focus on this restricted equivalence class, rather than its individual elements. The (IR+) constraint is suitable for the high-dimensional nature of the problem and can easily be incorporated in the formulation of the optimization problem for parameter estimation (see Section 4.2.2), which in turn yields estimates with tight error bounds (see Section 4.3).

Remark 4.1. It is worth pointing out that the sparsity requirement on Γ further limits the size of the equivalence class $\mathcal{C}(Q_2)$. To see this, note that for an arbitrary element in $\mathcal{C}(Q_2)$, $\check{\Gamma}$ satisfies $\check{\Gamma} = \Gamma + \Lambda Q_2$. In order for Γ and $\check{\Gamma}$ to have the same support, Q_2 needs to be further restricted. However, since the support set of Γ is unknown, this further implicit restriction on structural equivalence can not be enforced or verified, and the effective equivalence class can not be characterized through the support set either.

4.2.2 Proposed formulation.

Without loss of generality, we focus on the case where $d = 1$ in subsequent technical developments, so that $Z_t := (F_t^\top, X_t^\top)^\top$ follows a VAR(1) model $Z_t = AZ_{t-1} + W_t$:

$$\begin{bmatrix} F_t \\ X_t \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} F_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} w_t^F \\ w_t^X \end{bmatrix}. \tag{4.6}$$

The generalization to the VAR(d) ($d > 1$) case is straightforward since for any generic VAR(d) process satisfying $\mathcal{A}_d(L)Z_t = w_t$ where $\mathcal{A}_d(L) := \mathrm{I} - A^{(1)}L - \cdots - A^{(d)}L^d$, it can always be written in the form of a VAR(1) model for some dp-dimensional process $\widetilde{Z}_t$ [see **?** , for details].

Based on the introduced model identification scheme (IR+), we propose the following procedure to estimate the FAVAR model with a sparse coefficient matrix Γ and a dense loading matrix Λ, together with a sparse transition matrix A. Observed data matrices $\mathbf{X}$ and $\mathbf{Y}$ are identical to what have been previously defined, and to distinguish the responses from their lagged predictors when considering the VAR system, we let $\mathbf{X}_{n-1} := [x_1, \ldots, x_{n-1}]^\top$ denote the predictor matrix and $\mathbf{X}_n := [x_2, \ldots, x_n]^\top$ the response one; $\mathbf{F}_n, \mathbf{F}_{n-1}, \mathbf{Z}_n, \mathbf{Z}_{n-1}$ are analogously defined. Based on these notations, the sample versions of the VAR system and the calibration equation in (4.6) and (4.2) can be written as

$$\mathbf{Z}_n = \mathbf{Z}_{n-1}A^\top + \mathbf{W}, \quad \text{and} \quad \mathbf{Y} = \mathbf{F}\Lambda^\top + \mathbf{X}\Gamma^\top + \mathbf{E} =: \Theta + \mathbf{X}\Gamma^\top + \mathbf{E}.$$

We propose the following estimators obtained from a two-stage procedure for the coefficient matrices Λ, Γ and subsequently the transition matrices $\{A_{ij}\}_{i,j=1,2}$.

– Stage I: estimation of the calibration equation under (IR+). We formulate the following *constrained optimization* problem using a least squares loss function and incorporating the sparsity-induced ℓ_1 regularization of the sparse block Γ, the rank constraint on the hyperplane Θ, and (Compactness):

$$\begin{aligned}(\widehat{\Theta}, \widehat{\Gamma}) \quad &:= \underset{\Theta \in \mathbb{R}^{n\times q}, \Gamma \in \mathbb{R}^{q\times p_2}}{\arg\min} \big\{ \frac{1}{2n} \|\!| \mathbf{Y} - \Theta - \mathbf{X}\Gamma^\top \|\!|_{\mathrm{F}}^2 + \lambda_\Gamma \|\Gamma\|_1 \big\}, \\ &\qquad \text{subject to} \quad \operatorname{rank}(\Theta) \le r, \quad \|\!| \Theta/\sqrt{n} \|\!|_{\mathrm{op}} \le \phi. \end{aligned} \tag{4.7}$$

Once $\widehat{\Theta}$ is obtained, under (IR), the estimated factors $\widehat{\mathbf{F}}$ and the corresponding loading matrix $\widehat{\Lambda}$ are extracted as follows:

$$\widehat{\mathbf{F}} = \widehat{\mathbf{F}}^{\mathrm{PC}} (\widehat{\Lambda}_1^{\mathrm{PC}})^\top, \qquad \widehat{\Lambda} = \widehat{\Lambda}^{\mathrm{PC}} (\widehat{\Lambda}_1^{\mathrm{PC}})^{-1}, \tag{4.8}$$

where $\widehat{\Lambda}_1^{\mathrm{PC}}$ is the upper p_1 sub-block of $\widehat{\Lambda}^{\mathrm{PC}}$, with $\widehat{\mathbf{F}}^{\mathrm{PC}}$ and $\widehat{\Lambda}^{\mathrm{PC}}$ being the PC estimators [**?**] given by $\widehat{\mathbf{F}}^{\mathrm{PC}} := \sqrt{n}\widehat{U}$ and $\widehat{\Lambda}^{\mathrm{PC}} := \widehat{V}\widehat{D}/\sqrt{n}$. The estimates $\widehat{U}, \widehat{D}$ and $\widehat{V}$ are obtained from the SVD of $\widehat{\Theta} = \widehat{U}\widehat{\Theta}\widehat{V}^\top$. Note that after these algebra, $\widehat{\mathbf{F}}$ is the first p_1 columns of $\widehat{\Theta}$.

– Stage II: estimation of the VAR equation based on $\mathbf{X}$ and $\widehat{\mathbf{F}}$. With the estimated factor $\widehat{\mathbf{F}}$ as the surrogate for the true latent factor $\mathbf{F}$, the transition matrix A can be estimated by solving

$$\widehat{A} \quad := \underset{A \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}}{\arg\min} \big\{ \frac{1}{2n} \|\!| \widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1} A^\top \|\!|_{\mathrm{F}}^2 + \lambda_A \|A\|_1 \big\}, \tag{4.9}$$

where $\widehat{\mathbf{Z}}_n := [\widehat{\mathbf{F}}_n, \mathbf{X}_n]$ and $\widehat{\mathbf{Z}}_{n-1}$ is analogously defined. The ℓ_1-norm penalty induces sparsity on A according to the model assumption.

In principle, there may be additional contemporaneous dependence amongst the coordinates of the error processes e_t, w_t^X, w_t^F, respectively. In that case, one has to make additional assumptions on the structure of the inverses of covariance matrices Σ_e, Σ_w^X and Σ_w^F (e.g. sparsity) and modify the loss function accordingly. The complete estimation procedure for a VAR system whose error process exhibits contemporaneous dependence is discussed in detail in **?**] and an analogous strategy can be adopted for this model. We do not further elaborate in this study, since our prime interest is estimating the coefficient/transition matrices of the FAVAR model.

The formulation in (4.9) based on the least squares loss function and the surrogate $\widehat{\mathbf{F}}$ is straightforward. However, the formulation for the calibration equation merits additional discussion. First, note that the factor hyperplane Θ has at most rank p_1 and therefore has low rank structure relative to its size $n \times q$. We impose a rank constraint in the estimation procedure to enforce such structure. Together with the (IR+) constraint introduced above, the objective then becomes to estimate accurately the parameters of a model within the equivalence class $\mathcal{C}(Q_2)$, in the sense that the estimate of an arbitrary $\check{\Theta}$ ($\check{\Theta} \in \mathcal{C}(Q_2)$) is close to the true data generating Θ. Once this goal is achieved, this would enable accurate estimation of the transition matrix of the VAR system.

From an optimization perspective, the objective function admits a low-rank-plus-sparse decomposition and compactification is necessary for establishing the statistical properties of the global optima in the absence of explicitly specifying the interaction structure between the low rank and the sparse blocks (or the spaces they live in). Note that the form of the compactness constraint is dictated by the statistical problem under consideration. For example, **?**] studies a multivariate regression problem, where the coefficient is decomposed to a sparse and a low rank block. In that setting, a compactness constraint is imposed through the entry-wise infinity norm bound of the low rank block. **?**] studies a graphical model with latent variables where the conditional concentration matrix is the parameter of interest. The marginal concentration matrix is decomposed to a sparse and a low rank block via the alignment of the Schur complement, and the compactness constraint is imposed on both blocks and manifests through the corresponding regularization terms in the resulting optimization problem. Hence, the compactness constraint takes different forms but ultimately serves the same goal, namely, to introduce an upper bound on the magnitude of the low rank–sparse block interaction, with the latter being an important component in analyzing the estimation errors. The compacteness constraint adopted for the FAVAR model serves a similar purpose, although the presence of temporal dependence introduces a number of additional technical challenges compared to the two aforementioned settings that consider independent and identically distributed data.

Finally, we remark that the model identification scheme (IR+) incorporated in the optimization problem as a constraint, enables us to establish high-probability error bounds (relative to the true data generating parameters/factors) for the proposed estimators, as shown next in Section 4.3. Therefore, although (IR+) does not encompass the full $p_1^2 + p_1 p_2$ restrictions, it provides sufficient identifiability for estimation purposes.

4.3 Theoretical Properties.

In this section, we investigate the theoretical properties of the estimators proposed in Section 4.2.2. We focus on the formulation (4.7) and (4.9), whose global optima correspond to $(\widehat{\Theta}, \widehat{\Gamma})$ and $\widehat{A}$, respectively.

Since (4.9) relies not only on prime observable quantities (namely X_t), but also on estimated quantities from Stage I (namely $\widehat{F}_t$), the analysis requires a careful examination of how the estimation error in the factor propagates to that for A. We start by outlining a road map of our proof strategy together with a number of regularity conditions needed in subsequent developments. Section 4.3.1 establishes error bounds for $\widehat{\Gamma}$, $\widehat{\Theta}$ [5]and $\widehat{A}$ under certain regularity conditions and employing suitable choices of the tuning parameters, for *deterministic realizations* from the underlying observable processes. Specifically when considering the error bound of A, the error of the plug-in estimate $\widehat{\mathbf{F}}$ is assumed non-random and given. Subsequently, Section 4.3.2 examines the probability of the events in which the regularity conditions are satisfied for *random realizations*, and further establishes high-probability upper bounds for quantities to which the tuning parameters need to conform. Finally, the high-probability finite sample error bounds for the estimates obtained based on random realizations of the data generating processes readily follow after properly aligning the conditioning arguments, and the results are presented in Section 4.3.3.

Throughout, we use superscript $\star$ to denote the true value of the parameters of interest, and Δ for errors of the estimators; e.g., $\Delta_A = \widehat{A} - A^\star$. All proofs are deferred to the relevant Appendices.

A road map for establishing the consistency results. As previously mentioned, the key steps are:

- Part 1: analyses based on deterministic realizations using the optimality of the estimators, assuming the parameters of the objective function (e.g., the Hessian and the penalty parameter) satisfy certain regularity conditions;
- Part 2: analyses based on random realizations that the probability of the regularity conditions being satisfied, primarily involving the utilization of concentration inequalities.

In Part 1, note that the first-stage estimators obtained from the calibration equation are based on observed data and thus the regularity conditions needed are imposed on (functions of) the observed samples. On the other hand, the second-stage estimator relies on the

[5]Consequently, the error bounds of $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$ under IR are also obtained.

plugged-in first-stage estimates that have bounded error; therefore, the analysis is carried out in an analogous manner to problems involving error-in-variables. Specifically, the required regularity conditions on quantities appearing in the optimization (4.9) involve the error of the first stage estimates, with the latter assumed fixed. In Part 2, the focus shifts to the probability of the regularity conditions being satisfied under random realizations, again starting from the first stage estimates, with the aid of Gaussian concentration inequalities and proper accounting for temporal dependence. Once the required regularity conditions are shown to hold with high probability, combining the results established in Part 1 for deterministic realizations, provide the high-probability error bounds for $\widehat{\Theta}$ and $\widehat{\Gamma}$. The high-probability error bound of the estimated factors is subsequently established, which ensures that the variables which Stage II estimates rely upon are sufficiently accurate with high probability. Based on the latter result, the regularity conditions required for the Stage II estimates are then verified to hold with high probability at a certain rate. In the FAVAR model, since the estimation of the VAR equation is based on quantities among which one block is subject to error, to obtain an accurate estimate of the transition matrix requires more stringent conditions on population quantities (e.g., extremes of the spectrum), so that the regularity conditions hold with high probability. In essence, the joint process Z_t need to be adequately "regular" in order to get good estimates of the transition matrix , vis-a-vis the case of the standard VAR model where all variables are directly observed. Next, we introduce the following key concepts that are widely used in establishing theoretical properties of high-dimensional regularized M-estimators [e.g. **?** **?**], as well as quantities that are related to processes exhibiting temporal dependence [see also **?**].

Definition 4.1 (Restricted Strong Convexity (RSC))**.** A matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ satisfies the RSC condition with respect to norm Φ with curvature $\alpha_{\text{RSC}} > 0$ and tolerance $\tau_n \geq 0$, if

$$\frac{1}{2n}|||\mathbf{X}\Delta|||_{\text{F}}^2 \geq \frac{\alpha_{\text{RSC}}}{2}|||\Delta|||_{\text{F}}^2 - \tau_n\Phi^2(\Delta), \qquad \forall\ \Delta \in \mathbb{R}^{p\times p}.$$

In our setting, we consider the norm $\Phi(\Delta) = \|\Delta\|_1$.

Definition 4.2 (Deviation condition)**.** For a regularized M-estimator given in the form of

$$\widehat{A} := \min_A \big\{\frac{1}{2n}|||\mathbf{Y} - \mathbf{X}A^\top|||_{\text{F}}^2 + \lambda_A\|A\|_1\big\},$$

with $\mathcal{H}_A := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ denoting the Hessian and $\mathcal{G}_A := \frac{1}{n}\mathbf{Y}^\top\mathbf{X}$ denoting the gradient, we define the tuning parameter λ_A to be selected in accordance with the deviation condition, if

$$\lambda_A \geq c_0\|\mathcal{H}_A - \mathcal{G}_A(A^\star)^\top\|_\infty.$$

Definition 4.3 (Spectrum and its extremes)**.** For a p-dimensional stationary process X_t, its spectral density $f_X(\omega)$ is defined as $f_X(\omega) := \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\Sigma_X(h)e^{i\omega h}$, where $\Sigma_X(h) := \mathbb{E}(X_t X_{t+h}^\top)$. Its upper and lower extremes are defined as

$$\mathcal{M}(f_X) := \operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]} \Lambda_{\max}(f_X(\omega)), \qquad \text{and} \qquad \mathfrak{m}(f_X) := \operatorname*{ess\,inf}_{\omega\in[-\pi,\pi]} \Lambda_{\min}(f_X(\omega)).$$

The cross-spectrum for two generic stationary processes X_t and Y_t is defined as

$$f_{X,Y}(\omega) := \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\Sigma_{X,Y}(h)e^{i\omega h},$$

where $\Sigma_{X,Y}(h) := \mathbb{E}(X_t Y_{t+h}^\top)$, and its upper extreme is defined as

$$\mathcal{M}(f_{X,Y}) := \operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]} \sqrt{\Lambda_{\max}\big(f_{X,Y}^*(\omega) f_{X,Y}(\omega)\big)}, \quad \text{where } * \text{ denotes the conjuate transpose.}$$

Additionally, we let $S_{\mathbf{X}} := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ denote the sample covariance matrix, and similar quantities (e.g., $S_{\mathbf{E}}$) are analogously defined. Denote the density level of $\Gamma^\star$ by $s_{\Gamma^\star} := \|\Gamma^\star\|_0$, and that of $A^\star$ by $s_{A^\star}$.

We start by providing error bounds for $\widehat{\Gamma}$ and $\widehat{\Theta}$, as well as those of the corresponding $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$ extracted under IR. For the optimization problem given in (4.7), we assume that $r \geq p_1$ and ϕ is always compatible with the true data generating mechanism, so that $\Theta^\star$ is always feasible.

The error bounds of $\widehat{\Theta}$ and $\widehat{\Gamma}$ for deterministic realizations rely on: (i) $\mathbf{X}$ satisfying the RSC condition with curvature $\alpha_{\mathrm{RSC}}^{\mathbf{X}}$; and (ii) the tuning parameter λ_Γ being chosen in accordance with the deviation bound condition that is associated with the interaction between $\mathbf{X}$ and $\mathbf{E}$, the strength of the noise, and the interaction between the space spanned by the factor hyperplane and the observed $\mathbf{X}$. Upon the satisfaction of these conditions, the error bounds of $\widehat{\Theta}$ and $\widehat{\Gamma}$ are given by

$$\|\!|\Delta_\Gamma|\!\|_{\mathrm{F}}^2 + \|\!|\Delta_\Theta/\sqrt{n}|\!\|_{\mathrm{F}}^2 \leq C_1\lambda_\Gamma^2\big((p_1+r) + (2\sqrt{s_{\Gamma^\star}}+1)^2\big)/\min\{\alpha_{\mathrm{RSC}}^{\mathbf{X}},1\}^2,$$

and these conditions hold with high probability for random realizations of X_t and Y_t. Since $\widehat{\mathbf{F}}$ is the first p_1 columns of $\widehat{\Theta}$, it possesses an error bound of the similar form.

Next, we briefly sketch the error bounds of $\widehat{A}$. For the optimization in (4.9), for deterministic realizations, the results in **?**] can be applied with the corresponding RSC condition and deviation condition imposed on quantities associated with $\widehat{\mathbf{Z}}_n$ and $\widehat{\mathbf{Z}}_{n-1}$, and the error

for $\widehat{A}$ is in the form of

$$\|\!|\Delta_A|\!\|_{\mathrm{F}}^2 \leq C_2 s_{A^\star} \lambda_A^2 / (\alpha_{RSC}^{\widehat{\mathbf{Z}}})^2.$$

Then, for random realizations, assuming $\Delta_{\mathbf{F}}$ known and non-random, to satisfy the corresponding regularity conditions, we additionally require that the following functional involing the spectral density of the underlying joint process Z_t exhibits adequate curvature, that is, $\mathfrak{m}(f_Z)/\sqrt{\mathcal{M}(f_Z)} > c_0 h_1(\Delta_{\mathbf{F}_{n-1}})$ for constant c_0 and some function h_1 of the error $\Delta_{\mathbf{F}_{n-1}}$ that captures its strength. Moreover, the deviation bound is of the form $h_2(\Delta_{\mathbf{F}})$, which can be viewed as another function of the error[6]. Further, since $\Delta_{\mathbf{F}}$ is bounded with high probability from the analysis in Stage I, it will be established that $h_1(\Delta_{\mathbf{F}})$ and $h_2(\Delta_{\mathbf{F}})$ are both upper bounded at a certain rate, thus ensuring that the RSC condition and the deviation conditions can both be satisfied unconditionally, by properly choosing the required constants.

4.3.1 Statistical error bounds with deterministic realizations.

Proposition 4.1 below gives the error bounds for the estimators in (4.7), assuming certain regularity conditions hold for deterministic realizations of the processes X_t and Y_t, upon suitable choice of the regularization parameters.

Proposition 4.1 (Bound for Δ_Θ and Δ_Γ under fixed realizations)**.** *Suppose the fixed realizations $\mathbf{X} \in \mathbb{R}^{n \times p_2}$ of process $\{X_t \in \mathbb{R}^{p_2}\}$ satisfies the RSC condition with curvature $\alpha_{RSC}^{\mathbf{X}} > 0$ and a tolerance $\tau_{\mathbf{X}}$ for which*

$$\tau_{\mathbf{X}} \cdot \big(p_1 + r + 4 s_{\Gamma^\star}\big) < \min\{\alpha_{RSC}^{\mathbf{X}}, 1\}/16.$$

Then, for any matrix pair $(\Theta^\star, \Gamma^\star)$ satisfying $\|\!|\Theta^\star/\sqrt{n}|\!\|_{op} \leq \phi$ that generates Y_t, for estimators $(\widehat{\Theta}, \widehat{\Gamma})$ obtained by solving (4.7) with regularization parameters λ_Γ satisfying

$$\lambda_\Gamma \geq \max\big\{2\|\mathbf{X}^\top \mathbf{E}/n\|_\infty,\ \Lambda_{\max}^{1/2}(S_{\mathbf{E}}),\ (p_1 + r)\phi \Lambda_{\max}^{1/2}(S_{\mathbf{X}})\big\},$$

the following bound holds:

$$\|\!|\Delta_\Gamma|\!\|_F^2 + \|\!|\Delta_\Theta/\sqrt{n}|\!\|_F^2 \leq \frac{16\lambda_\Gamma^2 \big(p_1 + r + (2\sqrt{s_{\Gamma^\star}} + 1)^2\big)}{\min\{\alpha_{RSC}^{\mathbf{X}}, 1\}^2}. \tag{4.10}$$

Based on Proposition 4.1, under fixed realizations of X_t and Y_t, the error bounds of $\widehat{\Gamma}$ and $\widehat{\Theta}$ are established. Using these Stage I estimates and the IR condition, estimates of

[6]note the deviation bound in principle also depends on other population quantities such as $\mathfrak{m}(f_Z)$, $\mathcal{M}(f_Z)$, $\Lambda_{\max}(\Sigma_w)$ etc.

the factors and their loadings can be calculated. In particular, since $\Delta_{\mathbf{F}}$ corresponds to the first p_1 columns of Δ_Θ, the above bound automatically holds for $\Delta_{\mathbf{F}}$. Further, the following lemma provides the relative error of the estimated Λ under IR and the condition $\Lambda_{\max}^{1/2}(S_{\mathbf{F}})$, which translates to the requirement that the leading signal of $\mathbf{F}$ overrules the averaged row error of Δ_Θ.

Lemma 4.1 (Bound of Δ_Λ)**.** *The following error bound holds for $\widehat{\Lambda}$, provided that $\Lambda_{\max}^{1/2}(S_{\mathbf{F}}) > |||\Delta_\Theta/\sqrt{n}|||_F$:*

$$\frac{|||\Delta_\Lambda|||_F}{|||\Lambda^\star|||_F} \leq \frac{\sqrt{p_1}\cdot |||\Delta_\Theta/\sqrt{n}|||_F}{\Lambda_{\max}^{1/2}(S_{\mathbf{F}}) - |||\Delta_\Theta/\sqrt{n}|||_F}\Big(1 + 1/|||\Lambda^\star|||_F\Big). \tag{4.11}$$

Up to this point, error bounds have been obtained for all the parameters in the calibration equation. The following proposition establishes the error bound for the estimator obtained from solving (4.9), based on observed $\mathbf{X}$ and estimated $\widehat{\mathbf{F}}$, and assuming $\Delta_{\mathbf{F}}$ is fixed.

Proposition 4.2 (Bound for Δ_A under fixed realization and a non-random $\Delta_{\mathbf{F}}$.)**.** *Consider the estimator $\widehat{A}$ obtained by solving* (4.9)*. Suppose the following conditions hold:*

A1. $\widehat{\mathbf{Z}}_{n-1} := [\widehat{\mathbf{F}}_{n-1}, \mathbf{X}_{n-1}]$ *satisfies the RSC condition with curvature $\alpha_{RSC}^{\widehat{\mathbf{Z}}}$ and tolerance $\tau_{\mathbf{Z}}$ for which $s_{A^\star}\tau_{\mathbf{Z}} < \alpha_{RSC}^{\widehat{\mathbf{Z}}}/64$;*

A2. $\|\widehat{\mathbf{Z}}_{n-1}^\top\big(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^\star)^\top\big)/n\|_\infty \leq C(n, p_1, p_2)$ *where $C(n, p_1, p_2)$ is some function that depends on n, p_1 and p_2.*

Then, for any $\lambda_A \geq 4C(n, p_1, p_2)$, the following error bound holds for $\widehat{A}$:

$$|||\Delta_A|||_F \leq 16\sqrt{s_{A^\star}}\lambda_A/\alpha_{RSC}^{\widehat{\mathbf{Z}}}.$$

Note that Proposition 4.2 applies the results in **?** , Proposition 4.1] to the setting in this study, where Stage II estimation of the transition matrix is based on $\widehat{\mathbf{Z}}_n$ and $\widehat{\mathbf{Z}}_{n-1}$; consequently, the regularity conditions should be imposed on corresponding quantities associated with $\widehat{\mathbf{Z}}_n$ and $\widehat{\mathbf{Z}}_{n-1}$.

Propositions 4.1 and 4.2 give finite sample error bounds for the estimators of the parameters obtained by solving optimization problems (4.7) and (4.9) based on fixed realizations of the observable processes X_t and Y_t, and the regularity conditions outlined. Next, we examine and verify these conditions for random realizations of the processes, to establish high probability error bounds for these estimators.

4.3.2 High probability bounds under random realizations.

We provide high probability bounds or concentrations for the quantities associated with the required regularity conditions, for random realizations of X_t and Y_t. Specifically, we note

that when X_t is considered separately from the joint system, it follows a high-dimensional VAR-X model [**?**]

$$X_t = A_{22}X_{t-1} + A_{21}F_{t-1} + w_t^X,$$

whose spectrum $f_X(\omega)$ satisfies

$$f_X(\omega) = \big[\mathcal{A}_X^{-1}(e^{-i\omega})\big]\big(A_{21}f_F(\omega)A_{21}^\top + f_{w^X}(\omega) + f_{w^X,F}A_{21}^\top + A_{21}f_{w^X}(\omega)\big)\big[\mathcal{A}_X^{-1}(e^{-i\omega})\big]^*,$$

with $\mathcal{A}_X(z) := \mathrm{I} - A_{22}z$. Similar properties hold for F_t. Throughout, we assume $\{X_t\}, \{F_t\}$ and $\{Y_t\}$ are all mean-zero stable Gaussian processes.

Lemmas 4.2 to 4.5 respectively verify the RSC condition associated with $\mathbf{X}$ and establish the high probability bounds for $\|\mathbf{X}^\top\mathbf{E}/n\|_\infty$, $\Lambda_{\max}(S_{\mathbf{X}})$ and $\Lambda_{\max}(S_{\mathbf{E}})$.

Lemma 4.2 (Verification of the RSC condition for $\mathbf{X}$)**.** *Consider $\mathbf{X} \in \mathbb{R}^{n\times p_2}$ whose rows correspond to a random realization $\{x_1, \ldots, x_n\}$ of the stable Gaussian $\{X_t\}$ process, and its dynamics is governed by* (4.6)*. Then, there exist positive constants c_i $(i = 1, 2)$ such that with probability at least $1 - c_1\exp(-c_2 n\min\{\gamma^{-2}, 1\})$ where $\gamma := 54\mathcal{M}(g_X)/\mathfrak{m}(g_X)$, the RSC condition holds for $\mathbf{X}$ with curvature $\alpha_{RSC}^{\mathbf{X}}$ and tolerance $\tau_{\mathbf{X}}$ satisfying*

$$\alpha_{RSC}^{\mathbf{X}} = \pi\mathfrak{m}(f_X), \qquad \tau_{\mathbf{X}} = \alpha_{RSC}\gamma^2\Big(\frac{\log p_2}{n}\Big)/2\ ,$$

provided that $n \gtrsim \log p_2$.

Lemma 4.3 (High probability bound for $\|\mathbf{X}^\top\mathbf{E}/n\|_\infty$)**.** *There exist positive constants c_i $(i = 0, 1, 2)$ such that for sample size $n \gtrsim \log(p_2 q)$, with probability at least $1-c_1\exp(-c_2\log(p_2 q))$, the following bound holds:*

$$\|\mathbf{X}^\top\mathbf{E}/n\|_\infty \le c_0\Big(2\pi\mathcal{M}(f_X) + \Lambda_{\max}(\Sigma_e)\Big)\sqrt{\frac{\log p_2 + \log q}{n}}. \tag{4.12}$$

Lemma 4.4 (High probability bound for $\Lambda_{\max}(S_{\mathbf{X}})$)**.** *Consider $\mathbf{X} \in \mathbb{R}^{n\times p_2}$ whose rows correspond to a random realization $\{x_1, \ldots, x_n\}$ of the stable Gaussian $\{X_t\}$ process, and its dynamics is governed by* (4.6)*. Then, there exist positive constants c_i $(i = 0, 1, 2)$ such that for sample size $n \gtrsim p_2$, with probability at least $1 - c_1\exp(-c_2 p_2)$, the following bound holds for the eigen-spectrum of $S_{\mathbf{X}}$:*

$$\Lambda_{\max}(S_{\mathbf{X}}) \le c_0\mathcal{M}(f_X).$$

Lemma 4.5 (High probability bound for $\Lambda_{\max}(S_{\mathbf{E}})$)**.** *Consider $\mathbf{E} \in \mathbb{R}^{n\times q}$ whose rows are independent realizations of the mean zero Gaussian random vector e_t with covariance Σ_e. Then, for sample size $n \gtrsim q$, with probability at least $1 - \exp(-n/2)$, the following bound*

holds:

$$\Lambda_{\max}(S_{\mathbf{E}}) \leq 9\Lambda_{\max}(\Sigma_e).$$

In the next two lemmas, we verify the RSC condition for random realizations of $\widehat{\mathbf{Z}}_{n-1}$ and obtain the high probability bound $C(n, p_1, p_2)$ for $\|\widehat{\mathbf{Z}}_{n-1}^\top\big(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^\star)^\top\big)/n\|_\infty$, with the underlying truth $\mathbf{F}$ being random but the error $\Delta_{\mathbf{F}}$ non-random. Note that this can be equivalently viewed as a *conditional* RSC condition and deviation bound, when conditioning on some fixed $\Delta_{\mathbf{F}}$.

Lemma 4.6 (Verification of RSC for $\widehat{\mathbf{Z}}_{n-1}$)**.** *Consider $\widehat{\mathbf{Z}}_{n-1}$ given by*

$$\widehat{\mathbf{Z}}_{n-1} = \mathbf{Z}_{n-1} + \Delta_{\mathbf{Z}_{n-1}} = [\mathbf{F}_{n-1}, \mathbf{X}_{n-1}] + [\mathbf{\Delta}_{\mathbf{F}_{n-1}}, O],$$

with rows of $[\mathbf{F}_{n-1}, \mathbf{X}_{n-1}]$ being a random realization drawn from process $\{Z_t\}$ whose dynamics are given by (4.6)*. Suppose the lower and upper extremes of its spectral density $f_Z(\omega)$ satisfy*

$$\mathfrak{m}(f_Z)/\mathcal{M}^{1/2}(f_Z) > c_0 \cdot \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big) \qquad \textit{where} \quad S_{\Delta_{\mathbf{F}_{n-1}}} := \Delta_{\mathbf{F}_{n-1}}^\top \Delta_{\mathbf{F}_{n-1}}/n,$$

for some constant $c_0 \geq 6\sqrt{165\pi}$. Then, with probability at least $1 - c_1\exp(-c_2 n)$, $\widehat{\mathbf{Z}}_{n-1}$ satisfies the RSC condition with curvature

$$\alpha_{RSC}^{\widehat{\mathbf{Z}}} = \pi\mathfrak{m}(f_Z) - 54\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + \pi\mathfrak{m}(f_Z)/27}, \tag{4.13}$$

and tolerance

$$\tau_n = \Big(\frac{\pi}{2}\mathfrak{m}(f_Z) + 27\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + \pi\mathfrak{m}(f_Z)/27}\Big)\omega^2\sqrt{\frac{\log(p_1+p_2)}{n}},$$

where $\omega = 54\frac{\mathcal{M}(f_Z)}{\mathfrak{m}(f_Z)}$, provided that the sample size $n \gtrsim \log(p_1 + p_2)$.

Lemma 4.7 (Deviation bound for $\|\widehat{\mathbf{Z}}_{n-1}^\top\big(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^\star)^\top\big)/n\|_\infty$)**.** *There exist positive constants c_i $(i = 1, 2)$ and C_i $(i = 1, 2, 3)$ such that with probability at least $1 - c_1\exp\big(-$*

$c_2 \log(p_1+p_2)\big)$ *we have*

$$
\begin{aligned}
C(n,p_1,p_2) \leq\; & C_1\big[\mathcal{M}(f_Z) + \frac{\Lambda_{\max}(\Sigma_w)}{2\pi} + \mathcal{M}(f_{Z,W^+})\big]\sqrt{\frac{\log(p_1+p_2)}{n}} \\
& + \; C_2\big[\mathcal{M}^{1/2}(f_Z) \max_{j\in\{1,\ldots,p_1\}} \|\Delta_{\mathbf{F}_{n,\cdot j}}/\sqrt{n}\|\big]\sqrt{\frac{\log p_1 + \log(p_1+p_2)}{n}} \\
& + \; C_3\big[\Lambda^{1/2}_{\max}(\Sigma_w) \max_{j\in\{1,\ldots,(p_1+p_2)\}} \|\varepsilon_{n,\cdot j}/\sqrt{n}\|\big]\sqrt{\frac{\log(p_1+p_2)}{n}} \\
& + \; \frac{1}{n}\|\Delta^\top_{\mathbf{F}_{n-1}}\Delta_{\mathbf{F}_n}\|_\infty + \frac{1}{n}\|\Delta^\top_{\mathbf{F}_{n-1}}\Delta_{\mathbf{F}_{n-1}}(A^\star_{11})^\top\|_\infty,
\end{aligned}
\tag{4.14}
$$

where $\varepsilon_n := \Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top = [\Delta_{\mathbf{F}_n} - \Delta_{\mathbf{F}_{n-1}}(A^\star_{11})^\top, -\Delta_{\mathbf{F}_{n-1}}(A^\star_{21})^\top]$, *and* $\{W^+_t\} := \{W_{t+1}\}$ *is the shifted* W_t *process.*

Remark 4.2. Before moving to the high probability error bounds of the estimates, we discuss the conditions and the various quantities appearing in Lemmas 4.6 and 4.7 that determine the error bound of the estimated transition matrix and underlie the differences between the original VAR estimation problem based on primal observed quantities ("Original Problem" henceforth), and the present one in which one block of the variables enters the VAR system with errors. Note that the statements in the two lemmas are under the assumption that the error in the F_t block is pre-determined and non-random.

As previously mentioned, due to the presence of the error of the latent factor block, the corresponding regularity conditions need to be imposed and verified on quantities with the error incorporated, namely, $\widehat{\mathbf{Z}}$, instead of the original true random realizations $\mathbf{Z}$. Lemma 4.6 shows that with high probability, the random design matrix although exhibits error-in-variables, will still satisfy the RSC condition with some positive curvature as long as the spectrum of the process Z_t has sufficient regularity relative to the magnitude of the error, with the former determined by $\mathfrak{m}(f_X)/\mathcal{M}^{1/2}(f_X)$ and the latter by $\Lambda^{1/2}_{\max}(S_{\Delta_{\mathbf{F}_{n-1}}})$. In particular, the RSC curvature is pushed toward zero compared with that in the Original Problem, due to the presence of the second term in (4.13) that would be 0 if $\Delta_{\mathbf{F}_{n-1}} = 0$, i.e., there were no estimation errors. This curvature affects the constant scalar part of the ultimate high probability error bound obtained for the transition matrix.

Lemma 4.7 gives the deviation bound associated with the Hessian and the gradient (both random), which comprises of three components attributed to the random samples observed, the non-random error, and their interactions, respectively. Further, it is the relative order of these components that determines the error rate (as a function of model dimensions and the sample size). In particular, for the Original Problem, only the first term in (4.14) exists and yields an error rate of $\mathcal{O}(\sqrt{\log(p_1+p_2)/n})$ [see also **?**]. For the current setting, as it

is later shown in Theorem 4.1, since $\|\|\Delta_{\mathbf{F}}/\sqrt{n}\|\|_{\mathrm{F}} \asymp \mathcal{O}(1)$, the dominating term of the three components is the one attributed to the non-random error[7] and it ultimately determines the error rate of $\widehat{A}$, which will also be $\mathcal{O}(1)$.

4.3.3 High probability error bounds for the estimators.

Given the results in Sections 4.3.1 and 4.3.2, we provide next high probability error bounds for the estimates, obtained by solving the optimization problems in (4.7) and (4.9) based on random snapshots from the underlying processes X_t and Y_t.

Theorem 4.1 combines the results in Proposition 4.1 and Lemmas 4.2 to 4.5 and provides the high probability error bound of the estimates, when $\widehat{\Theta}$ and $\widehat{\Gamma}$ are estimated based on random realizations from the observable processes X_t and Y_t, with the latter driven by both X_t and the latent F_t.

Theorem 4.1 (High probability error bounds for $\widehat{\Theta}$ and $\widehat{\Gamma}$)**.** *Suppose we are given some randomly observed snapshots $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ obtained from the stable Gaussian processes X_t and Y_t, whose dynamics are described in* (4.6) *and* (4.2)*. Suppose the following conditions hold for some $(C_{X,l}, C_{X,u})$ and $(C_{e,l}, C_{e,u})$:*

C1. $C_{X,l} \leq \mathfrak{m}(f_X) \leq \mathcal{M}(f_X) \leq C_{X,u}$*;*

C2. $C_{e,l} \leq \Lambda_{\min}(\Sigma_e) \leq \Lambda_{\max}(\Sigma_e) \leq C_{e,u}$*.*

Then, there exist universal constants $\{C_i\}$ and $\{c_i\}$ such that for sample size $n \gtrsim q$, by solving (4.7) *with regularization parameter*

$$\lambda_\Gamma = \max\Big\{C_1(2\pi\mathcal{M}(f_X) + \Lambda_{\max}(\Sigma_e))\sqrt{\frac{\log(p_2 q)}{n}},\ C_2(p_1+r)\phi\mathcal{M}^{1/2}(f_X),\ C_3\Lambda^{1/2}_{\max}(\Sigma_e)\Big\}, \tag{4.15}$$

the solution $(\widehat{\Theta}, \widehat{\Gamma})$ has the following bound with probability at least $1 - c_1\exp(-c_2\log(p_2 q))$:

$$\|\|\Delta_\Theta/\sqrt{n}\|\|_F^2 + \|\|\Delta_\Gamma\|\|_F^2 \lesssim C(\mathfrak{m}(f_X), \mathcal{M}(f_X), \Lambda_{\max}(\Sigma_e)) \cdot \kappa(s_{\Gamma^\star}, p_1^3, r^3, \phi) =: K_1, \tag{4.16}$$

for some function $C(\mathfrak{m}(f_X), \mathcal{M}(f_X), \Lambda_{\max}(\Sigma_e))$ that does not depend on n, p_2, q, and $\kappa(\cdot)$ that depends linearly on $s_{\Gamma^\star}$, p_1^3, r^3 and the box constraint ϕ.

Note that the above bound also holds if we replace Δ_Θ by $\Delta_{\mathbf{F}}$ under IR. Next, using the results in Proposition 4.2, Lemmas 4.6 and 4.7 and combine the bound in Theorem 4.1, we establish a high probability error bound for the estimated $\widehat{A}$ in Theorem 4.2.

[7] with the implicit assumption that $\log(p_1+p_2)/n \asymp o(1)$ which is indeed the case for this study.

Theorem 4.2 (High probability error bound for $\widehat{A}$)**.** *Under the settings and with the procedures in Theorem 4.1, we additionally assume the following condition holds for the spectrum of the joint process Z_t:*

C3. $\mathfrak{m}(f_Z)/\mathcal{M}^{1/2}(f_Z) > C_Z$ *for some constant* C_Z*.*

Then there exists universal constants $\{c_i\}$*,* $\{c_i'\}$ *and* $\{C_i\}$ *such that for sample size* $n \gtrsim q$*, such that the estimator* $\widehat{A}$ *obtained by solving for* (4.9) *with* λ_A *satisfying*

$$\lambda_A = C_1\big(\mathcal{M}(f_Z) + \frac{\Sigma_w}{2\pi} + \mathcal{M}(f_{Z,W^+})\big)\sqrt{\frac{\log(p_1+p_2)}{n}} + C_2\mathcal{M}^{1/2}(f_Z)\sqrt{\frac{\log(p_1+p_2)+\log p_1}{n}} + C_3\Lambda_{\max}^{1/2}(\Sigma_w)\sqrt{\frac{\log(p_1+p_2)}{n}} + C_4,$$

with probability at least

$$\Big(1 - c_1\exp\{-c_2\log(p_2 q)\}\Big)\Big(1 - c_1'\exp\{-c_2'\log(p_1+p_2)\}\Big), \tag{4.17}$$

the following bound holds for Δ_A*:*

$$|||\Delta_A|||_F^2 \le \check{C}(K_1, \mathfrak{m}(f_Z), \mathcal{M}(f_Z)) \cdot \check{\kappa}(s_{A^\star}),$$

for some function $\check{C}(K_1, \mathfrak{m}(f_Z), \mathcal{M}(f_Z))$ *that does not depend on* n, p_2, q *and* $\check{\kappa}(\cdot)$ *that depends linearly on* $s_{A^\star}$*. Here* K_1 *denotes the upper bound of the first stage error shown in* (4.16)*.*

Remark 4.3. Note that to establish the high probability finite-sample error bound of the transition matrix estimate $\widehat{A}$, the sample size requirement $n \gtrsim q$ for the proposed estimation procedure is more stringent compared to that for the Original Problem, with the latter given by $n \gtrsim \sqrt{\log(p_1+p_2)}$. The root of this discrepancy is due to the estimated factor, whose accurate recovery from the calibration equation requires the concentration of $\Lambda_{\max}(S_{\mathbf{E}})$ that provides adequate control over $\Delta_{\mathbf{F}}$, which in turn places the tightest condition on the sample size.

Remark 4.4. As a straightforward generalization, for a VAR(d), $d > 1$ system $Z_t = (F_t^\top, X_t^\top)^\top$, a similar error bound holds by considering the augmented process $\widetilde{Z}_t^\top := (Z_t, Z_{t-1}, \ldots, Z_{t-d+1})$ that satisfies

$$\widetilde{Z}_t = \widetilde{A}\widetilde{Z}_{t-1} + \widetilde{W}_t, \qquad \text{where} \qquad \widetilde{A} := \begin{bmatrix} A^{(1)} & A^{(2)} & \cdots & A^{(d)} \\ \mathrm{I}_p & O & O & O \\ \vdots & \ddots & \vdots & \vdots \\ O & O & \mathrm{I}_p & O \end{bmatrix}, \quad \widetilde{W}_t = \begin{bmatrix} W_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

In particular, with probability at least $\big(1 - c_1\exp\{-c_2\log(p_2 q)\}\big)\big(1 - c_1'\exp\{-c_2'\log(d(p_1+$

$p_2))\}\big)$, the following bound holds for the estimate of $\widetilde{A}$:

$$\|\Delta_{\widetilde{A}}\|_{\mathrm{F}}^2 \leq \widetilde{C}(K_1, \mathfrak{m}(f_{\widetilde{Z}}), \mathcal{M}(f_{\widetilde{Z}})) \cdot \widetilde{\kappa}(s_{\widetilde{A}^\star}).$$

However, note that although the error bound is still of the same form, the stronger temporal dependence yields a larger $\widetilde{C}(K_1, \mathfrak{m}(f_{\widetilde{Z}}), \mathcal{M}(f_{\widetilde{Z}})$ through the RSC curvature parameter; specifically, a smaller value of $\mathfrak{m}(f_{\widetilde{Z}})$. Its impact on the deviation bound will not manifest itself in terms of the order of the error, since it only affects the constants in front of lower order terms in the expression of choosing λ_A.

4.4 Implementation and Performance Evaluation.

We first discuss implementation issues of the proposed problem formulation for the high-dimensional FAVAR model. Specifically, the formulation requires imposing the compactness constraint for identifiability purposes and for obtaining the necessary statistical guarantees for the estimates of the model parameters. However, the value ϕ in the compactness constraint is hard to calibrate in any real data set. Hence, in the implementation we relax this constraint and assess the performance of the algorithm. Due to its importance in constraining the size of the equivalence class $\mathcal{C}(Q_2)$, we examine in Appendix C.3 certain relative extreme settings where the proposed relaxation fails to provide accurate estimates of the model parameters.

Implementation. The following relaxation of (4.7) is used in practice:

$$\min_{\Theta,\Gamma} f(\Theta, \Gamma) := \big\{\frac{1}{2n}\|\mathbf{Y} - \Theta - \mathbf{X}\Gamma^\top\|_{\mathrm{F}}^2 + \lambda_\Gamma \|\Gamma\|_1\big\}, \quad \text{subject to} \quad \mathrm{rank}(\Theta) \leq r, \tag{4.18}$$

which leads to Algorithm 4.1.

The implementation of Stage I requires the pair of tuning parameters (λ_Γ, r) as input, and the choice of r is particularly critical since it determines the effective size of the latent block. In our implementation, we select the optimal pair based on the Panel Information Criterion (PIC) proposed in **?**], which searches for (λ_Γ, r) over a lattice that minimizes

$$\mathrm{PIC}(\lambda_\Gamma, r) := \frac{1}{nq}\Big\|\mathbf{Y} - \widehat{\Theta} - \mathbf{X}\widehat{\Gamma}^\top\Big\|_{\mathrm{F}}^2 + \widehat{\sigma}^2\Big[\frac{\log n}{n}\|\widehat{\Gamma}\|_0 + r(\frac{n+p}{nq})\log(nq)\Big],$$

where $\widehat{\sigma}^2 = \frac{1}{nq}\|\mathbf{Y} - \widehat{\Theta} - \mathbf{X}\widehat{\Gamma}^\top\|_{\mathrm{F}}^2$. Analogously, the implementation of Stage II requires λ_A as input, and we select λ_A over a grid of values that minimizes the Bayesian Information

Algorithm 4.1: Computational procedure for estimating A, Γ and Λ.

Input: Time series data $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, (λ_Γ, r), and λ_A.
Stage I: recover the latent factors by solving (4.18), through iterating between (1.1) and (1.2) until $|f(\Theta^{(m)}, \Gamma^{(m)}) - f(\Theta^{(m-1)}, \Gamma^{(m-1)})| <$ tolerance:

(1.1) Update $\widehat{\Theta}^{(m)}$ by singular value thresholding (SVT): do SVD on the lagged value-adjusted hyperplane, i.e., $\mathbf{Y} - \mathbf{X}(\widehat{\Gamma}^{(m-1)})^\top = UDV^\top$, where $D := \operatorname{diag}\big(d_1, \ldots, d_{\min(n,q)}\big)$, and construct $\widehat{\Theta}^{(m)}$ by

$$\widehat{\Theta}^{(m)} = UD_rV, \qquad \text{where} \quad D_r := \operatorname{diag}\big(d_1, \ldots, d_r, 0, \ldots, 0\big).$$

(1.2) Update $\widehat{\Gamma}^{(m)}$ with the plug-in $\widehat{\Theta}^{(m)}$ so that each row j is obtained with Lasso regression (in parallel) and solves

$$\min_\beta \big\{ \frac{1}{2n} \|(\mathbf{Y} - \widehat{\Theta}^{(m)})_{\cdot j} - \mathbf{X}\beta\|^2 + \lambda_A \|\beta_1\|_1 \big\}.$$

Stage I output: $\widehat{\Theta}$ and $\widehat{\Gamma}$; the estimated factor $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$ via (4.8) under (IR).
Stage II: estimate the transition matrix by solving (4.9): update each row of A (in parallel) by solving the Lasso problem:

$$\min_\beta \big\{ \frac{1}{2n} \|(\widehat{\mathbf{Z}}_n)_{\cdot j} - \widehat{\mathbf{Z}}_{n-1}\beta\|^2 + \lambda_A \|\beta\|_1 \big\}.$$

Stage II output: $\widehat{A}$.
Output: Estimates $\widehat{\Gamma}$, $\widehat{\Lambda}$, $\widehat{A}$ and the latent factor $\widehat{\mathbf{F}}$.

Criterion (BIC):

$$\mathrm{BIC}(\lambda_A) = \sum_{i=1}^q \log \mathrm{RSS}_i + \frac{\log n}{n} \|\widehat{A}\|_0,$$

where $\mathrm{RSS}_i := \|(\mathbf{X}_n)_{\cdot i} - \mathbf{X}_{n-1}\widehat{A}_{i\cdot}^\top\|^2$ is the residual sum of square of the i-th regression. Extensive numerical work shows that these two criteria select very satisfactory values for the tuning parameters, which in turn yield highly accurate estimates of the model parameters.

Simulation setup. Throughout, we assume Σ_w^X, Σ_X^F and Σ_e are all diagonal matrices, and the sample size is fixed at 200, unless otherwise specified. We first generate samples of $F_t \in \mathbb{R}^{p_1}$ and $X_t \in \mathbb{R}^{p_2}$ recursively according to the VAR(d) model in (4.1), and then the samples of $Y_t \in \mathbb{R}^q$ are generated according to the linear model given in (4.2). In particular, (IR) is imposed on the true value of the parameter, hence $\Lambda^\star$ that is used for generating Y_t always satisfies the restriction $\Lambda = \left[\begin{smallmatrix} \mathrm{I}_{p_1} \\ * \end{smallmatrix}\right]$.

For the calibration equation, the density level of the sparse coefficient matrix $\Gamma \in \mathbb{R}^{q \times p_2}$ is fixed at $5/p_2$ for each regression; thus, each Y_t coordinate is affected by 5 series (coordinates) from the X_t block on average. The bottom $(q - p_1) \times p_1$ block of the loading matrix $\Lambda \in \mathbb{R}^{q \times p_1}$

is dense. The magnitude of nonzero entries of Γ and that of entries of Λ may vary to capture different levels of signal contributions to Y_t, and we adjust the standard deviation of e_t to maintain the desired level of the signal-to-noise ratio for Y_t (averaged across all coordinates).

For the transition matrix A of the VAR equation, the sparsity for each of its component block $\{A_{ij}\}_{i,j=1,2}$ varies across settings, so as to capture different levels of the influence from the lagged value of the latent block F_t on the observed X_t. Note that to ensure stability of the VAR system, the spectral radius of A, $\varrho(A)$, needs to be smaller than 1. In particular, when a VAR(d) ($d > 1$) system is considered, we need to ensure that the spectral radius of $\widetilde{A}$ is smaller than 1[8], where we let $p = p_1 + p_2$ and

$$\widetilde{A} := \begin{bmatrix} A^{(1)} & A^{(2)} & \cdots & A^{(d)} \\ \mathrm{I}_p & O & O & O \\ \vdots & \ddots & \vdots & \vdots \\ O & O & \mathrm{I}_p & O \end{bmatrix}.$$

Table 4.1 lists the simulation settings and their parameter setup.

	q	p_1	p_2	$s_{A_{11}}$	$s_{A_{12}}$	$s_{A_{21}}$	$s_{A_{22}}$	$\mathrm{SNR}(Y_t)$
A1	100	5	50	$s_A = 3/(p_1+p_2)$				1.5
A2	200	10	100	$s_A = 3/(p_1+p_2)$				1.5
A3	200	5	100	$3/p_1$	$2/p_2$	$2/p_1$	$2/p_2$	1.5
A4	300	5	500	$3/p_1$	$2/p_2$	0.8	$2/p_2$	1.5
B1 ($d=2$)	200	5	100	$s_{A^{(1)}} = 3/(p_1+p_2)$				2
				$s_{A^{(2)}} = 2/(p_1+p_2)$				
B2 ($d=4$)	200	5	100	0.5	$3/p_2$	0.5	$3/p_2$	2
				0.2	$2/p_2$	0.25	$2/p_2$	
				$s_{A^{(3)}} = 2/(p_1+p_2)$				
				$s_{A^{(4)}} = 2/(p_1+p_2)$				
B3 ($d=4$)	100	5	25	0.5	$2/p_2$	0.5	$2/p_2$	2
				0.2	$1.5/p_2$	0.1	$1.5/p_2$	
				$s_{A^{(3)}} = 1/(p_1+p_2)$				
				$s_{A^{(4)}} = 0.8/(p_1+p_2)$				

Table 4.1: Parameter setup for different simulation settings for the VAR equation.

Specifically, in settings A1–A4, $(F_t^\top, X_t^\top)^\top$ jointly follows a VAR(1) model. The (average) signal-to-noise ratio for each regression of Y_t is 1.5. For settings A1 and A2, the transition matrix A is uniformly sparse, with A2 corresponding to a larger system; for settings A3 and A4, we increase the density level (the proportion of nonzero entries) for the transition matrices that govern the effect of F_{t-1} on F_t and X_t. In particular, for setting A4, we consider a large system with 500 coordinates in X_t, and the factor effect is almost pervasive on these coordinates (through the lags), as the density level of A_{21} is set at 0.8. Settings B1, B2 and B3 consider settings with more lags ($d = 2$ and $d = 4$, respectively), and to compensate for the higher level of correlation between F_t and X_t, we elevate the signal-to-noise for each

[8]In practice, this can be achieved by first generating $A^{(1)}, \ldots, A^{(d)}$, align them in $\widetilde{A}_{\text{initial}}$ and obtain the scale factor $\zeta := \varrho_{\text{target}}/\varrho(\widetilde{A}_{\text{initial}})$, then scale $A^{(i)}$ by ζ^i. The validity of this procedure follows from simple algebraic manipulations.

regression of Y_t to 2. For B1, the transition matrices for both lags ($A^{(1)}$ and $A^{(2)}$) have uniform sparsity patterns, with $A^{(2)}$ being slightly more sparse compared to $A^{(1)}$; for B2, the transition matrices for the first two lags have higher sparsity in the component that governs the $F_{t-i} \to X_t$ cross effect, and those for the last two lags have uniform sparsity. B3 has approximately the same scale as observed in real data, and due to a small p_2, the system exhibits a higher sparsity level in general.

Performance evaluation. We consider both the estimation and forecasting performance of the proposed estimation procedure. The performance metrics used for estimation are sensitivity (SEN), specificity (SPC) and the relative error in Frobenius norm (Err) for the sparse components (transition matrices A and the coefficient matrix Γ), defined as

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{SPC} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \qquad \text{Err} = |||\Delta_M|||_{\text{F}} / |||M^\star|||_{\text{F}} \text{ (for some generic matrix } M).$$

We also track the estimated size of the latent component (i.e., the rank constraint in (4.7), jointly with λ_Γ is selected by PIC), as well as the relative errors of $\widehat{\Theta}$, $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$. For forecasting, we focus on evaluating the h-step-ahead predictions for the X_t block. Specifically, for settings A1–A4, we consider $h = 1$; for settings B1–B3, we consider $h = 1, 2$. We use the same benchmark model as in **?**] which is based on a special case of the Minnesota prior distribution [**?**], so that the for any generic time series $X_t \in \mathbb{R}^p$, each of its coordinates $j = 1, \ldots, p$ follows a centered random walk:

$$X_{t,j} = X_{t-1,j} + u_{t,j}, \qquad u_{t,j} \sim \mathcal{N}(0, \sigma_u^2). \tag{4.19}$$

For each forecast $\widehat{x}_{T+h}$, its performance is evaluated based on the following two measures:

$$\text{RE} = \|\widehat{x}_{T+h} - x_{T+h}\|_2^2 / \|x_{T+h}\|_2^2, \qquad \text{RER} = \frac{\frac{1}{p_2}\sum_{j=1}^{p_2} \left| \frac{\widehat{x}_{T+h,j} - x_{T+h,j}}{x_{T+h,j}} \right|}{\frac{1}{p_2}\sum_{j=1}^{p_2} \left| \frac{\widetilde{x}_{T+h,j} - x_{T+h,j}}{x_{T+h,j}} \right|},$$

where RE measures the ℓ_2 norm of the relative error of the forecast to the true value; whereas for RER, it measures the ratio between the relative error of the forecast and the above described benchmark. In particular, its numerator and denominator respectively capture the averaged relative error of all coordinates of the forecast $\widehat{x}_{T+h}$ and that of the benchmark $\widetilde{x}_{T+h}$ that evolves according to (4.19), while the ratio measures how much the forecast based on the proposed FAVAR model outperforms (< 1) or under-performs (> 1) compared to the benchmark.

All tabulated results are based on the average of 50 replications. Table 4.2, 4.3 and 4.4,

respectively, depict the performance of the estimates of the parameters in the calibration and the VAR equations, as well as the forecasting performance under the settings considered.

	PIC-selected r	Err($\widehat{\Theta}$)	Err($\widehat{\mathbf{F}}$)	Err($\widehat{\Lambda}$)	SEN($\widehat{\Gamma}$)	SPC($\widehat{\Gamma}$)	Err($\widehat{\Gamma}$)
A1	4.8(.40)	0.32(.010)	0.56(.074)	0.67(.345)	0.99(.007)	0.98(.003)	0.45(.013)
A2	9.96(.19)	0.32(.008)	0.90(.065)	2.54(1.30)	0.99(.005)	0.98(.001)	0.52(.010)
A3	4.78(.54)	0.33(.048)	0.73(.103)	2.59(1.59)	0.99(.003)	0.99(.001)	0.57(.009)
A4	4.42(.49)	0.38(.040)	0.84(.100)	2.66(2.14)	0.97(.009)	0.99(.001)	0.59(.015)
B1	5(0)	0.23(.004)	0.41(.043)	0.54(.020)	1.00(.000)	0.97(.011)	0.27(.014)
B2	5(0)	0.26(.007)	0.38(.047)	0.42(.087)	1.00(.000)	0.99(.002)	0.37(.007)
B3	5(0)	0.25(.007)	0.34(.031)	0.34(.080)	1.00(.000)	0.99(.001)	0.32(.012)

Table 4.2: Performance evaluation of the parameters in the calibration equation.

	lag	SEN($\widehat{A}$)	SPC($\widehat{A}$)	Err($\widehat{A}$)	SEN($\widehat{A}_{22}$)	SPC($\widehat{A}_{22}$)	Err($\widehat{A}_{22}$)
A1		0.99(.003)	0.95(.012)	0.35(.019)	0.99(.001)	0.96(.013)	0.31(.022)
A2		0.98(.008)	0.97(.004)	0.46(.018)	0.99(.001)	0.98(.003)	0.39(.017)
A3		0.86(.050)	0.98(.006)	0.73(.029)	0.93(.032)	0.98(.005)	0.65(.034)
A4		0.75(.046)	0.92(.002)	0.71(0.024)	0.99(.001)	0.92(.002)	0.60(.018)
B1	$A^{(1)}$	0.99(.003)	0.98(.002)	0.47(.017)	0.99(.002)	0.98(.002)	0.46(.017)
	$A^{(2)}$	0.97(.010)	0.98(.002)	0.55(.017)	0.98(.011)	0.98(.003)	0.55(.018)
B2	$A^{(1)}$	0.89(.017)	0.88(.003)	0.71(.014)	0.90(.017)	0.99(.003)	0.70(.014)
	$A^{(2)}$	0.75(.028)	0.88(.003)	0.89(.020)	0.77(0.032)	0.88(.003)	0.90(.021)
	$A^{(3)}$	0.84(.025)	0.88(.003)	0.85(.015)	0.85(.027)	0.88(.004)	0.84(.018)
	$A^{(4)}$	0.72(.022)	0.88(.003)	0.99(.017)	0.73(.025)	0.88(.003)	0.98(.017)
B3	$A^{(1)}$	0.93(.034)	0.96(.010)	0.61(.043)	0.94(.035)	0.97(.009)	0.60(.045)
	$A^{(2)}$	0.77(.078)	0.96(.010)	0.74(.044)	0.78(.084)	0.97(.010)	0.74(.046)
	$A^{(3)}$	0.80(.098)	0.96(.012)	0.75(.052)	0.81(.102)	0.97(.010)	0.74(.056)
	$A^{(4)}$	0.74(.122)	0.97(.011)	0.78(.059)	0.72(.134)	0.97(.009)	0.79(.065)

Table 4.3: Perofrmance evaluation of the estimated transition matrices in the VAR equation.

		A1	A2	A3	A4	B1	B2	B3
$h=1$	RE	0.53(.117)	0.60(.075)	0.80(.075)	0.56(.109)	0.62(.060)	0.89(.091)	0.81(.094)
	RER	0.38(.065)	0.38(.046)	0.45(.064)	0.40(.055)	0.35(.171)	0.42(.217)	0.32(.129)
$h=2$	RE					0.66(.127)	0.94(.173)	0.90(.402)
	RER					0.24(.071)	0.29(.118)	0.26(.174)

Table 4.4: Evaluation of forecasting performance.

Based on the results listed in Tables 4.2 and 4.3, we notice that in all settings, the parameters in the calibration equation $\widehat{\Theta}$ and $\widehat{\Gamma}$ are well estimated, while the rank slightly underestimated. Further, the SEN and SPC measures of $\widehat{\Gamma}$ show excellent performance regarding support recovery. It is worth pointing out that the estimation accuracy of the parameters in the calibration equation strongly depends on the signal-to-noise ratio of Y_t. In particular, if the signal-to-noise ratio in A1-A4 is increased to 1.8, the rank is always correctly selected by PIC, and the estimation relative error of $\widehat{\Theta}$ further decreases(results omitted for space considerations)[9]. Under the given IR, we decompose the estimated factor hyperplane into the factor block and its loadings. The results show that both quantities exhibit a higher relative error compared to that of the factor hyperplane. Of note, the loadings estimates

[9]This also comes up when comparing the relative error of $\widehat{\Theta}$ in the A1-A4 settings to that in the B1-B2 ones, where the latter two have a higher SNR.

exhibit a lot of variability as indicated by the high standard deviation in the Table.

Regarding the estimates in the VAR equation, for settings A1, A2 and B1 that are characterized by an adequate degree of sparsity, the recovery of the skeleton of the transition matrices is very good. However, performance deteriorates if the latent factor becomes "more pervasive" (settings A3 and A4), which translates to the A_{21} block having lower sparsity. On the other hand, this does not have much impact on the recovery of the A_{22} sub-block, as for these two settings, SEN and SPC of A_{22} still remain at a high level. For settings with more lags, performance deteriorates (as expected) although SEN and SPC remain fairly satisfactory. On the other hand, the relative error of the transition matrices increases markedly. Nevertheless, the estimates of the first lag transition matrix is better than the remaining ones. Further, the results indicate that smaller size VAR systems (B3) exhibit better performance than larger ones. Finally, in terms of forecasting (results depicted in Table 4.4), the one-step-ahead forecasting value yields approximately 50% to 90% RE (compared to the truth), depending on the specific setting and the actual SNR, while it outperforms the forecast of the benchmark by around 40% (based on the RER measure). Of note, the 2-step-ahead forecasting value for settings with more lags outperforms the benchmark by an even wider margin with the RER ratio decreasing to less than 0.3.

4.5 Application to Commodity Price Interlinkages.

Interlinkages between commodity prices represent an active research area in economics, together with a source of concern for policymakers. Commodity prices, unlike stocks and bonds, are determined more strongly by global demand and supply considerations. Nevertheless, other factors are also at play as outlined next. The key ones are: (i) the state of the global macro-economy and the state of the business cycle that manifest themselves as direct demand for commodities; (ii) monetary policy, specifically, interest rates that impact the opportunity cost for holding inventories, as well as having an impact on investment and hence production capacity that subsequently contribute to changes in supply and demand in the market; and (iii) the relative performance of other asset classes through portfolio allocation [see **? ?** , and references therein]. We employ the FAVAR model and the proposed estimation method to investigate interlinkages amongst major commodity prices. The X_t block corresponds to the set of commodity prices of interest, while the Y_t block contains representative indicators for the global economic environment. We extract the factors F_t based on the calibration equation and then consider the augmented VAR system of (F_t, X_t), so that the estimated interlinkages amongst commodity prices are based on a larger information set that takes into account broader economic activities.

Data. The commodity price data (X_t) are retrieved from the International Monetary Fund, comprising of 16 commodity prices in the following categories: Metal, Energy (oil) and Agricultural. The set of economic indicators (Y_t) contain core macroeconomic variables and stock market composite indices from major economic entities including China, EU, Japan, UK and US, with a total number of 54 indicators. Specifically, the macroeconomic variables primarily account for: Output & Income (e.g. industrial production index), Labor Market (unemployment), Money & Credit (e.g. M2), Interest & Exchange Rate (e.g. Fed Funds Rate and the effective exchange rate), and Price Index (e.g. CPI). For variables that reflect interest rates, we use both the short-term interest rate such as 6-month LIBOR, and the 10-year T-bond yields from the secondary market. Further, to ensure stationarity of the time series, we take the difference of the logarithm for X_t; for Y_t, we apply the same transformation as proposed in **?**]. A complete list of the commodity prices and economic indicators used in this study is provided in Appendix C.5. For all time series considered, we use monthly data spanning the January 2001 to December 2016 period. Further, based on previous empirical findings in the literature related to the global financial crisis of 2008 [**?**], we break the analysis into the following three sub-periods [**?**]: pre-crisis (2001–2006), crisis (2007–2010) and post-crisis (2011–2016), each having sample size (available time points) 72, 48, and 72, respectively.

We apply the same estimation procedure for each of the above three sub-periods. Starting with the calibration equation, we estimate the factor hyperplane Θ and the sparse regression coefficient matrix Γ, then extract the factors based on the estimated factor hyperplane under the (IR) condition. For each of the three sub-periods, 4, 3, and 3 factors are respectively identified based on the PIC criterion, with the key variable loadings (collapsed into categories) on each extracted factor listed in Table 4.5, after adjusting for ΓX_t. Based on the

	pre-crisis				crisis			post-crisis		
	F1	F2	F3	F4	F1	F2	F3	F1	F2	F3
bond return	−		+	+	−	+				−
economic output	+						+		+	
equity return	+				−	−		−		+
interest/exchange rate			∗					∗		
labor		+			−		−			
money & credit			+		+				+	
price index		+					+			−
trade		−				∗		∗		

Table 4.5: Composition of the factors identified for three sub-periods. +, − and ∗ respectively stand for positive (all economic indicators have a positive sign in Λ), negative and mixed (sign) contribution.

composition of the factors, we note that the factors summarize both the macroeconomic environment and also capture information from the secondary market (bond & equity return), as suggested by economic analysis of potential contributors to commodity price movements [**?** **?**]. Hence, the obtained factors summarize the necessary information to include in

the VAR system that examines commodity price interlinkages over time. Further, across all three periods considered, Economic Output and Money & Credit indicators contribute positively to the factor composition. In particular, the positive contribution from the M2 measure of money supply for the US during the crisis period and that from the Fed Funds Rate post crisis are pronounced; hence, the estimated factors strongly reflect the effect of the Quantitative Easing policy adopted by the US central bank. The contribution of the other categories are mixed, with that from bond returns being noteworthy due to their role as a proxy for long-term interest rates, which impact both the cost of investment in increasing production capacity and on holding inventories, as well as on the composition of asset portfolios across a range of investment possibilities (stocks, bonds, commodities, etc.).

Next, using these estimated factors, we fit a sparse VAR(2) model to the augmented $(\widehat{F}_t^\top, X_t^\top)^\top$ system. The estimated transition matrices are depicted in Figures 4.1 to 4.3 as networks. It is apparent that the factors play an important role, both as emitters and receivers. The effects from the first lag are generally stronger to that from the second one. In particular, focusing on the first lag, the dominant nodes in the system have shifted over time from (OIL, SOYBEANS, ZINC) pre crisis to (SUGAR, WHEAT, COPPER) during the crisis, then to (OIL, SOYBEANS, RICE) post crisis. Based on node weighted degree, the role of OIL is dominant in both pre- and post-crisis periods, but is much weaker during the crisis.

Another key feature of the interlinkage networks is their increased connectivity during the crisis period, vis-a-vis the pre- and post-crisis periods. The same empirical finding has been noted for stock returns [see **?** , and references therein]. Before the global financial crisis of 2008, commodity prices were fast rising primarily due to increased demand from China. Specifically, as Chinese industrial production quadrupled between 2001 and 2011, its consumption of industrial metals (Copper, Zinc, Aluminum, Lead) increased by 330%, while its oil consumption by 98%. This strong demand shock led to a sharp rise in these commodity prices, particularly accentuated beginning in 2006 (the onset of the crisis period considered in our analysis), briefly disrupted with a quick plunge of commodity prices in 2008 and their subsequent recovery in the ensuing period until late 2010, when demand from China subsided, which coupled with weak demand from the EU, Japan and the US in the aftermath of the crisis created an oversupply that put downward pressure on prices. These events induce strong inter-temporal and cross-temporal correlations amongst commodity prices, and hence are reflected in their estimated interlinkage network.

4.6 Discussion.

This chapter considered the estimation of FAVAR model under the high-dimensional scaling. It introduced an identifiability constraint (IR+) that is suitable for high-dimensional settings, and when such a constraint is incorporated in the optimization problem based upon the calibration equation, the global optimizer corresponds to model parameter estimates with bounded statistical errors. This development also allows for accurate estimation of the transition matrices of the VAR system, despite the plug-in factor block contains error due to the fact that it is an estimated quantity. Extensive numerical work illustrates the overall good performance of the proposed empirical implementation procedure, but also illustrates that the (IR+) constraint is not particularly stringent, especially in settings where the coefficient matrix Γ of the observed predictor variables in the calibration equation exhibits sufficient level of sparsity.

Recall that the nature of the FAVAR model results in estimating the transition matrix of a VAR system with one block of the observations (factors) being an estimated quantity, rather than conducting the estimation based on observed samples. This introduces a problem of independent interest, namely what statistical guarantees can be established for the estimates of the transition matrix of a VAR system under high-dimensional scaling when one block (or even all) of the variables are subject to error. Similar problems have been examined in the high-dimensional iid setting [e.g. **?**], as well as low dimensional time series settings; for example, **?**] examines parameter estimation of a univariate autoregressive process with error-in-variables and in more recent work **?**] investigates parameter identification of VAR-X and dynamic panel VAR models subject to measurement errors.

The results obtained in this paper provide some initial insights, based on the roadmap used to establish them. Building on the discussion in Remark 4.2, consider the following setting where one is interested in estimating a VAR system with one block of variables contaminated by some *non-random* $\mathbf{Z}$, so that the transition matrix is obtained by solving

$$\min_{A} \quad \Big\{ \frac{1}{2n} \Big|\Big|\Big| \begin{bmatrix} \mathbf{x}_n^{(1)} \\ \mathbf{x}_n^{(2)} + \mathbf{z}_n \end{bmatrix} - \begin{bmatrix} \mathbf{x}_{n-1}^{(1)} \\ \mathbf{x}_{n-1}^{(2)} + \mathbf{z}_{n-1} \end{bmatrix} A^\top \Big|\Big|\Big|_{\mathrm{F}}^2 + \lambda_A \|A\|_1 \Big\},$$

whereas the true data generating mechanism is that $\binom{X_t^{(1)}}{X_t^{(2)}}$ jointly follow a VAR(1) model. Then, based on Lemma 4.6 and 4.7, as long as the RSC condition on the corresponding quantity is satisfied with high probability and the tuning parameter is chosen in accordance with the deviation bound condition, the error of the estimated transition matrix is still well-bounded. In particular, if the magnitude of $\mathbf{Z}$ satisfies $\|\mathbf{Z}/\sqrt{n}\|_{\mathrm{F}} \asymp o(1)$, then the error of the estimated transition matrix would still be $\mathcal{O}(\sqrt{\log(p_1+p_2)/n})$, which is identical to that

of a VAR model without error-in-variables, despite the fact that the estimation is based on contaminated quantities rather than uncontaminated samples. In addition, the presence of the contaminating $\mathbf{Z}$ does not affect the sample size requirement with the latter remaining at $n \gtrsim \sqrt{\log(p_1 + p_2)}$, although it does affect the exact error bound through both the deviation bound and the curvature in the RSC condition. Thus, it is of interest to investigate the conditions required on a *random* $\mathbf{Z}$, so that the VAR estimates exhibit similar rates to those without contamination and this constitutes a topic of future research.

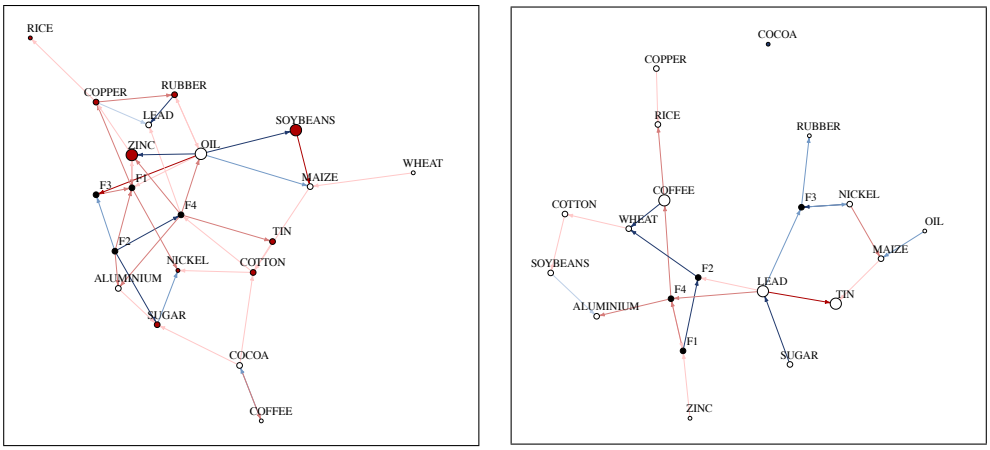

Figure 4.1: Estimated transition matrices for Pre-crisis period.

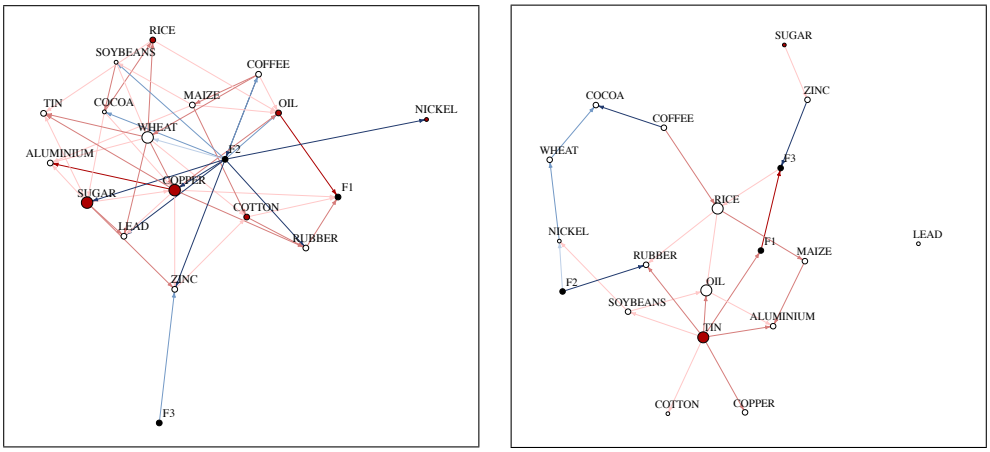

Figure 4.2: Estimated transition matrices for the Crisis period.

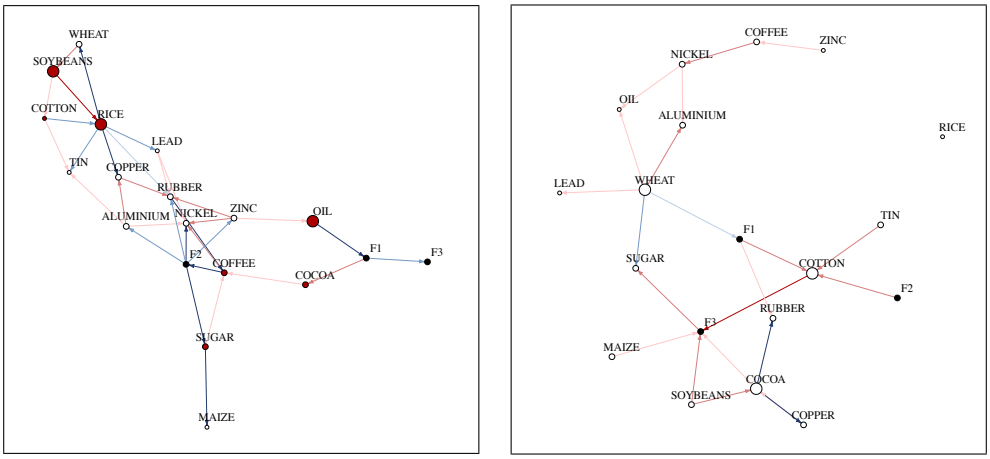

Figure 4.3: Estimated transition matrices for Post-crisis period

Left panel: $\widehat{A}^{(1)}$; right panel: $\widehat{A}^{(2)}$. Node sizes are proportional to node weighted degrees. Positive edges are in red and negative edges are in blue. Edges with higher saturation have larger magnitudes.

CHAPTER V

Approximate Factor Models with Strongly Correlated Idiosyncratic Errors

5.1 Introduction.

Factor models are widely used in a number of scientific fields for reducing the dimension of data sets comprising of a large number of variables [**?**]. A factor model assumes that each variable under consideration can be expressed as a linear combination of a small number of *latent* factors plus an idiosyncratic component (error term). Co-movements between the variables can be accounted for by these few factors, thus aiding their interpretation. When *exact* factor models are used in the analysis of cross-sectional data, it is assumed that the idiosyncratic components are mutually uncorrelated [**?**]. However, for time series data such assumptions are often too restrictive, especially if a large number of them are considered. In that case, it is of interest to examine *approximate* factor models that allow for correlations within the idiosyncratic components, or equivalently the common factors do not fully capture all relationships among the observed time series.

Such an approximate factor model was introduced in **?**] for the analysis of portfolios comprising of a large number of assets. Since then a number of papers have appeared in the literature investigating properties of such approximate factor models, under the assumption that the correlations between the common factors and the idiosyncratic component, as well as those amongst the idiosyncratic components are *weak*. Formally, the approximate factor model is defined as

$$X_t = \Lambda F_t + u_t, \quad t = 1, \cdots, n, \tag{5.1}$$

where X_t is a vector of p-dimensional time series, F_t a K-dimensional latent factor process, Λ a $p \times K$ matrix of *factor loadings* and u_t the vector of idiosyncratic components. It is often further assumed that the factor process exhibits Vector Autoregressive dynamics, namely $F_t = \sum_{i=1}^{q} \Phi_i F_{t-i} + \eta_t$, where η_t is an independent and identically distributed error process

and Φ_i are $K \times K$ transition matrices. The model in (5.1) is typically estimated through principal component (PC) decomposition, which operates under the assumption that as the time series panel size $p \to \infty$, the leading K eigenvalues of $\Sigma_X := \mathbb{E}(X_t X_t^\top)$ diverge whereas all eigenvalues of $\Sigma_u := \mathbb{E}(u_t u_t^\top)$ are bounded, thus enabling the separation between the common factors and the idiosyncratic components. Some key theoretical results for this model are given in **?** **?**], where the asymptotic normality of the estimated factors and factor loadings[1] obtained from PC analysis is established, under a $\sqrt{p}/n \to 0$ scaling for the first result and a $\sqrt{n}/p \to 0$ scaling for the latter one. Further, in order that the maximum deviation across time of the estimated factors relative to the true factors vanishes, it is required that $n/p \to 0$.

In later work, **?**] consider the same factor model representation, but u_t is allowed to exhibit serial correlation within each coordinate and is assumed to be uncorrelated with F_t across all time leads and lags. By decorrelating the coordinates of the u_t error process, the model can be rewritten as

$$X_t = \Lambda F_t + D(L) X_{t-1} + \epsilon_t, \tag{5.2}$$

where $D(L) = \mathrm{diag}(\delta_1(L), \ldots, \delta_p(L))$ is a *diagonal* matrix with each entry being the autoregressive polynomial corresponding to coordinates of u_t, while ϵ_t is a pure noise term that is neither cross-sectionally nor serially correlated. The model implies that given F_t, the lagged value of X_{it} does not help in predicting X_{jt} for $i \neq j$, since $D(L)$ is diagonal. A detailed review of variants of approximate factor models and their applications in macroeconomics and finance are provided in **?** **?**].

?] note that when the number of time series under consideration is not too large, inference based on the PC estimator is distorted, if large values are present in the idiosyncratic components (e.g. substantial jumps). As a remedy in practical settings, the authors propose to use the PC decomposition on $X^{\perp}_{t|t-}$ instead of the original X_t, where $X^{\perp}_{t|t-}$ is the orthogonal projection of the observed process X_t on the space spanned by its time lags. This approach implicitly assumes that $dp < n$ where d is the number of lags on which the orthogonal projection is carried out, otherwise a proper projection operator is not readily available. A similar issue has been observed in **?**] which points out that even with moderate serial dependence amongst the idiosyncratic components, the number of factors will be overestimated when the available sample size is not sufficiently large. This issue is also present in the study of the sub-prime mortgage crisis by **?**] that uses the model in (5.2) to assess the contribution of common factors to the spreads of Credit Default Swaps for a set of global banks. Specifically, the authors remark that the presence of strongly correlated idiosyncratic

[1] up to some invertible transformation

components renders the criterion proposed in **?**] for accurately selecting the number of common factors ineffective. To address this problem and further test for a "spillover effect" from lagged terms of other banks, the authors incorporate lagged terms of the X_t process in the model. The analysis proceeds by estimating the posited model for each pair of banks separately, resulting in running a large number of such models. However, a more principled and informative analysis would be based on estimating a single model for all the banks, but the latter would entail incorporating sparsity constraints due to the limited sample size (time points) available. Note that procedures associated with filtering the impact of the temporal dependence in the idiosyncratic component have also been considered in other applied work. For instance, **?**] use a two-stage procedure, wherein a VAR model is first estimated and subsequently a factor model fitted to the residuals. **?**] specify the lag terms explicitly and estimate the factor and the transition matrix of lags through Kalman filter, which makes the permissible time series panel fairly restrictive. In their empirical analysis, they only consider three time series.

This brief literature review shows that on the theory front, the "weak correlation" assumption is prevalent in the literature, since in essence it provides formal justification for the use of PC analysis for estimating the space spanned by the common factors. Further, by imposing additional regularity conditions it also enables establishing consistency of both the factors and their respective loadings. However, once the idiosyncratic component exhibits stronger dependence and thus the "weak correlation" assumption breaks down, existing approaches that solely focus on estimating the factors lead to incorrect selection of them and thus inaccurate inference [**?** **?**]. Hence, due to lack of appropriate analytical tools, researchers undertaking empirical work resort to either filtering the data for temporal dependence before using a factor model, or ignoring the issue altogether. It is thus highly desirable to model the latent factors and past lags of the process *simultaneously*. To this end, following **?**] we write the approximate factor model in the form given in (5.2), but allow for $D(L)$ to exhibit cross-correlation structure; i.e. $D(L)$ is not assumed to be diagonal, but merely *sparse*. Hence, the dynamics of the p time series in X_t can be written in the form of a lag-adjusted static factor model, with the lag term impacting the current values through sparse transition matrices of past values, and the noise term now being completely uncorrelated with the factor structure. Note that the common factors together with the lag term play the role of mean structure of the observable process X_t in this formulation. In contrast to the **?**] formulation, in our proposed model, for the observable process conditional on the latent factors, the lagged value of the i^{th} time series can be predictive of the present value of time series j, if the corresponding entry in the sparse transition matrices is nonzero.

Based on the proposed formulation, two key quantities in the model are the space spanned

by the common factors (henceforth called factor hyperplane) and the sparse transition matrices of the time lags of the observable process. To obtain their estimates, we posit a penalized least squares objective function, introduce an iterative algorithm to minimize it and establish finite sample high-probability error bounds for the convergent solution estimates. It is of interest to point out that although the issue of the contribution of lagged values of the X_t process is noted in **?**], it is never further explored in their analysis. On the other hand, since the transition matrices of the autoregressive components constitute key parameters of our model formulation, we show in our application study that they provide useful and interpretable information.

Finally, note that in recent work **?**], the authors consider a covariate-adjusted factor model with unknown grouping structures among the multivariate response, which shares a similar form with the posited model, when viewed as a regression model, modulo the absence of the group-specific factors. Consistency results are established for the estimated regression coefficients of that model and the factor loadings under rather stringent assumptions on the deterministic realizations of the underlying stochastic processes, which however are hardly satisfied for random realizations drawn from the assumed distributions. In our work, we impose significantly weaker conditions on these deterministic sample quantities, which are later verified to be satisfied with high probability for random realizations. Therefore, a by-product of our results, with rather minor modifications, is the consistency of the model in **?**] under significantly weaker assumptions.

Thus, the key contributions of this chapter are the formulation and consistent estimation of factor models that include past lags of the observed process, which explicitly accounts for strong cross-correlations amongst the coordinates of the idiosyncratic component. As previously mentioned, such a setting shows up often in empirical work, but due to lack of approaches to properly handle it, the contribution of the idiosyncratic component is largely ignored. The technical developments in this paper provide insights of how to handle the interaction of the *latent* factor space with the past history of the observed process, and the strategy used to establish consistency is broadly applicable to other models in the literature [e.g. **?**] with similar structure. Of particular interest, are the verification of the Restricted Strong Convexity and a Deviation Condition for Gaussian processes X_t, F_t, ϵ_t that are at the heart of establishing consistency for the model under study. While the verification of such conditions for high-dimensional VAR models [**?**] and their variants [**?**] is challenging, the presence of a latent process and its strong interaction with the error process poses additional technical complications, successfully resolved in Theorem 5.2.

The remainder of this chapter is organized as follows. In Section 5.2, we introduce our model setup and the estimation procedure for the parameters of interest—the sparse tran-

sition matrix (assuming for ease of presentation a single lag) and the common factor space. Theoretical properties of the proposed estimator are established in Section 5.3, including its high-probability statistical error bound, convergence property and its connections to other possible model formulations and estimators in related literature. In Section 5.4, we introduce an empirical implementation procedure and present the performance evaluation of the estimates based on synthetic data. In Section 5.5, an application of our model to weekly stock return data of large US financial institutions for the 2001 to 2016 period is considered. An extension to serial dependence with more lags for the idiosyncratic component is briefly discussed in Section 5.6. Finally, Section 5.7 concludes the paper.

Notation. Throughout this chapter, for some generic matrix A, we use $|||\cdot|||$ to denote its matrix norms, including the operator norm $|||A|||_{\text{op}}$, the Frobenius norm $|||A|||_{\text{F}}$, the nuclear norm $|||A|||_*$, $|||A|||_1 = \max_{1\le j\le n}\sum_{i=1}^m |a_{ij}|$, and $|||A|||_\infty = \max_{1\le i\le m}\sum_{j=1}^n |a_{ij}|$. We use $\|A\|_1 = \sum_{i,j}|a_{ij}|$ and $\|A\|_\infty = \max_{i,j}|a_{ij}|$ to denote the elementwise 1-norm and infinity norm. Additionally, we use $\varrho(A)$ to denote its spectral radius ($\max|\lambda(A)|$). For two matrices A and B of commensurate dimensions, denote their inner product by $\langle\!\langle A, B\rangle\!\rangle = \text{trace}(A'B)$. Finally, we write $A \gtrsim B$ if there exists some absolute constant c that is independent of the model parameters such that $A \ge cB$.

5.2 Problem Formulation and Estimation.

We introduce the model setup by assuming the idiosyncratic component follows the aforementioned sparse VAR(d) model, which simultaneously incorporates the cross-sectional and serial structure among its coordinates. To convey the main arguments, we assume without loss of generality that $d = 1$ and present the extension to the general lag case in Section 5.6.

Toward this end, starting from the dynamic factor representation of the observable process $X_t = \tilde{\lambda}(L)f_t + u_t$, where f_t is the common factor and u_t the idiosyncratic component, the dynamics of u_t satisfy $\mathcal{B}(L)u_t = \epsilon_t$ with $\mathcal{B}(L) = \text{I}_p - BL$ being the lagged matrix polynomial for some sparse B. Multiplying $\mathcal{B}(L)$ on both sides leads to the dynamic factor model that consists of (5.3) and (5.4), where F_t collects the lags of f_t so that it only enters the dynamics of X_t contemporaneously, and is additionally assumed to follow a VAR model:

$$X_t = \Lambda F_t + BX_{t-1} + \epsilon_t, \tag{5.3}$$

$$F_t = \Phi(L)F_{t-1} + \eta_t. \tag{5.4}$$

Note that $X_t \in \mathbb{R}^p$ is observable, whereas $F_t \in \mathbb{R}^K$ ($K \ll p$) is a latent VAR(q) process with $\Phi(L) := \Phi_1 + \Phi_2 L + \cdots + \Phi_q L^{q-1}$ for some q. Further, ϵ_t is the mean zero noise process that satisfies $\mathbb{E}(\epsilon_{it}\epsilon_{js}) = 0\ \forall i, j, s, t$; $\text{Cov}(X_t, \epsilon_{t+h}) = 0\ \forall\ h \ge 1$, and is additionally assumed

to be uncorrelated with F_t for its present and future values, i.e., $\text{Cov}(F_t, \epsilon_{t+h}) = 0 \ \forall h \geq 0$. The parameters of interest are the factor hyperplane (to be specified later) and the sparse transition matrix B. Although in principle, the transition matrices corresponding to the dynamics of F_t in (5.4) can be recovered once the factors are identified, in this study in accordance with work on this subject in the literature [e.g. **?**], we focus solely on the recovery of the common factors. It is worth pointing out that due to the latency of the factors, $\Lambda R^{-1} R F_t = \Lambda F_t$ always holds for any rotation matrix $R \in \mathbb{R}^{K \times K}$; therefore given ΛF_t, to separately identify (Λ, F_t) from its observationally equivalent counterpart $(\Lambda R^{-1}, RF_t)$, a total number of $K \times K$ restrictions are required to resolve such indeterminacies.

Further, to ensure that X_t is covariance stationary, we only require that the spectral radius of B satisfies $\varrho(B) < 1$, with no further restrictions needed on Λ. Throughout this paper, we assume X_t is covariance stationary and its spectral density exists, defined next. Define the auto-covariance function of some generic process X_t as $\Gamma_X(h) = \mathbb{E}(X_t X_{t+h}^\top)$, and its spectral density $g_X(\omega)$ is given by

$$g_X(\omega) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_X(h) e^{i\omega h}.$$

To obtain the explicit expression of $g_X(\omega)$ where X_t satisfies (5.3), we start from the filtered process $Z_t := \mathcal{B}(L) X_t = \Lambda F_t + \epsilon_t$, whose spectral density satisfies

$$g_Z(\omega) = \Lambda g_F(\omega) \Lambda + g_\epsilon(\omega) + g_{\epsilon,F}(\omega) \Lambda^\top + \Lambda g_{F,\epsilon}(\omega),$$

where we additionally define $\Gamma_{\epsilon,F}(h) := \mathbb{E}(\epsilon_t F_{t+h}^\top)$ and $\Gamma_{F,\epsilon}(h) := \mathbb{E}(F_{t+h} \epsilon_{t+h}^\top)$, with $g_{\epsilon,F}(\omega)$ and $g_{F,\epsilon}(\omega)$ accordingly defined. As a consequence, the spectral density of X_t is expressed as

$$g_X(\omega) = \left[\mathcal{B}^{-1}(e^{-i\omega})\right] \Big(\Lambda g_F(\omega) \Lambda^\top + g_\epsilon(\omega) + g_{\epsilon,F}(\omega) \Lambda^\top + \Lambda g_{F,\epsilon}(\omega) \Big) \left[\mathcal{B}^{-1}(e^{-i\omega})\right]^*.$$

5.2.1 Estimation.

Given a snapshot of the p-dimensional observable process X_t, denoted by $\{x_0, x_1, \ldots, x_n\}$, let

$$\mathbf{X}_n := \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}^\top, \quad \mathbf{X}_{n-1} := \begin{bmatrix} x_0 & x_1 & \ldots & x_{n-1} \end{bmatrix}^\top, \quad \mathbf{F} := \begin{bmatrix} F_1 & F_2 & \ldots & F_n \end{bmatrix}^\top,$$

where $\mathbf{X}_n \in \mathbb{R}^{n \times p}$ and $\mathbf{X}_{n-1} \in \mathbb{R}^{n \times p}$ respectively denote the contemporaneous response matrix and the lagged predictor matrix, and $\mathbf{F} \in \mathbb{R}^{n \times K}$ denotes the latent factor matrix with the latent factor F_t at time point t stacked in its rows. The noise matrix $\mathbf{E}$ is analogously

defined. We additionally define the *factor hyperplane* associated with the latent factor $\mathbf{F}$ as $\Theta := \mathbf{F}\Lambda^\top \in \mathbb{R}^{n\times p}$, and note that Θ has rank at most K. With the above notations, the model in (5.3) for the observed samples can be written as follows:

$$\mathbf{X}_n = \Theta + \mathbf{X}_{n-1}B^\top + \mathbf{E}.$$

To estimate the transition matrix $B \in \mathbb{R}^{p\times p}$, as well as the factor hyperplane $\Theta = \mathbf{F}\Lambda^\top \in \mathbb{R}^{n\times p}$, we consider the following constrained optimization problem

$$\begin{aligned} &\min_{B,\Theta} \ \Big\{\frac{1}{2n}\big|\big|\big|\mathbf{X}_n - \Theta - \mathbf{X}_{n-1}B^\top\big|\big|\big|_{\mathrm{F}}^2\Big\}, \\ &\text{subject to} \quad \mathrm{rank}(\Theta) \le r, \quad \|B\|_1 \le t, \qquad \text{for some } r \text{ and } t, \end{aligned} \tag{5.5}$$

with the feasible region determined through a rank constraint imposed on the factor hyperplane Θ and a sparsity-inducing norm constraint imposed on the transition matrix B. However, note that the rank constraint makes the feasible region *non-convex*. Without additional assumptions on the sequence of iterative updates of Θ and B obtained through an alternating minimization algorithm, convergence to a stationary point of the sequence is not guaranteed. Consequently, it becomes analytically intractable to characterize the resulting solution obtained by terminating the computational procedure, subject to some empirical convergence criterion. For this reason, we consider an alternative formulation based on the tight convex relaxation of the rank constraint, whose optimal solution has convergence guarantees and shares similar statistical properties vis-a-vis its non-convex counterpart (see Section 5.3.4 Remark 5.2 for a detailed discussion).

Formally, we focus on analyzing the convex program in (5.6), which can be obtained from (5.5) by alternatively considering the nuclear norm constraint for the factor hyperplane and the ℓ_1 norm constraint for the sparse transition matrix B in the Lagrangian form:

$$\begin{aligned} (\widehat{B}, \widehat{\Theta}) = \arg\min_{B,\Theta} \ &\Big\{\frac{1}{2n}\big|\big|\big|\mathbf{X}_n - \Theta - \mathbf{X}_{n-1}B^\top\big|\big|\big|_{\mathrm{F}}^2 + \lambda_B\|B\|_1 + \lambda_\Theta|||\Theta/\sqrt{n}|||_*\Big\}, \\ &\text{subject to} \quad \Theta \in \mathbb{B}_n(\phi), \end{aligned} \tag{5.6}$$

where $\mathbb{B}_n(\phi) := \{\Theta \in \mathbb{R}^{n\times p} \mid |||\Theta/\sqrt{n}|||_* \le \phi\}$ is a nuclear norm ball of radius ϕ, and λ_B and λ_Θ are tuning parameters. The constraint on the feasible region of Θ is to incur additional compactness on the low rank component, as later explained in greater detail in Section 5.3.1. Note that $(\widehat{B}, \widehat{\Theta})$ falls into the class of *regularized M-estimators*, whose properties have been extensively studied in the statistical literature for diverse settings [e.g., **? ?**]. In particular, a polynomial-time computation procedure outlined in Algorithm V.1 can be used to obtain

$(\widehat{B}, \widehat{\Theta})$, which involves the alternating minimization with respect to B and Θ in each outer update, and the composite gradient update [**?**] for each inner update of Θ.

Algorithm 5.1: An alternate minimizing algorithm for estimating B and Θ based on (5.6).

Input: Time series data $\{x_i\}_{i=0}^n$, tuning parameter λ_B, λ_Θ, feasible region constraint ϕ.
Initialization: Initialize with $\widehat{\Theta}^{(0)} = \mathrm{O}_{p\times p}$.
Iterate until convergence:
(1) Update $\widehat{B}^{(m)}$ with the plug-in $\widehat{\Theta}^{(m-1)}$ so that each row j is obtained with Lasso regression (in parallel) and solves

$$\widehat{B}_{j\cdot} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \Big\{ \frac{1}{2n} \Big\| \big[\mathbf{X}_n - \widehat{\Theta}^{(m)}\big]_{\cdot j} - \mathbf{X}_{n-1}\beta \Big\|^2 + \lambda_B \|\beta\|_1 \Big\}. \tag{5.7}$$

(2) Update $\widehat{\Theta}^{(m)}$ with the plug-in $\widehat{B}^{(m)}$ by

$$\widehat{\Theta}^{(m)} = \underset{\Theta\in\mathbb{B}_n(\phi)}{\arg\min} \Big\{ \frac{1}{2n} |\!|\!| \mathbf{X}_n - \mathbf{X}_{n-1}(\widehat{B}^{(m)})^\top - \Theta |\!|\!|_{\mathrm{F}}^2 + \lambda_\Theta |\!|\!| \Theta/\sqrt{n} |\!|\!|_* \Big\}, \tag{5.8}$$

where each inner update involves two singular value thresholding (SVT) operations:

$$\widehat{\Theta}^{(m,t+1)} = \underset{\Theta\in\mathbb{B}_n(\phi)}{\arg\min} \Big\{ \langle\!\langle \Theta, \nabla\mathcal{L}_m(\widehat{\Theta}^{(m,t)}) \rangle\!\rangle + \frac{\eta}{2} |\!|\!| \Theta - \Theta^{(m,t)} |\!|\!|_{\mathrm{F}}^2 + \lambda_\Theta |\!|\!| \Theta/\sqrt{n} |\!|\!|_* \Big\},$$

for some stepsize η and $\nabla\mathcal{L}_m(\Theta) := -\frac{1}{n}\big(\mathbf{X}_n - \mathbf{X}_{n-1}(\widehat{B}^{(m)})^\top - \Theta\big)$.
Output: Estimated sparse transition matrix $\widehat{B}$ and the low rank hyperplane $\widehat{\Theta}$.

Reconstruction of the factors. The solution to (5.6) provides an estimate of the factor hyperplane, based on which realizations of the K-dimensional latent factors process can be reconstructed under certain identifiability restrictions. As mentioned in Section 5.1, for any invertible matrix $R \in \mathbb{R}^{K\times K}$, the following equality holds

$$\Theta = \mathbf{F}\Lambda^\top = \big[\mathbf{F}R^\top\big]\big[\Lambda R^{-1}\big]^\top := \check{\mathbf{F}}\check{\Lambda}^\top,$$

hence, given a factor hyperplane and the latency of the factors, to fully identify the factors and the corresponding loading matrix $(\mathbf{F}, \Lambda)$ from their observationally equivalent counterpart $(\check{\mathbf{F}}, \check{\Lambda})$, a total number of K^2 restrictions is required to address their indeterminancy. Various choices for the identification restrictions have been discussed in the literature [e.g., **?** , and references therein], including the most popular PC estimator [**?**] which assumes orthogonality for both the factors and the loadings, as well as the ones that implicitly assume certain ordering of the factors and impose specific structural restrictions on the loading matrix [see PC2 and PC3 identification restrictions in **?**]. Under these restrictions, the factors

and the loading matrix can always be uniquely identified[2] and obtained based on the SVD of the estimated hyperplane $\widehat{\Theta}$. For example, the PC estimator that assumes $\frac{1}{n}\mathbf{F}^\top\mathbf{F} = \mathrm{I}_K$ and $\Lambda^\top\Lambda$ is diagonal can be obtained by letting $\widehat{\mathbf{F}}_{\mathrm{PC}} = \sqrt{n}\widehat{U}$ and $\widehat{\Lambda}_{\mathrm{PC}} = \frac{1}{\sqrt{n}}\widehat{V}\widehat{D}$, where $\widehat{U}$, $\widehat{V}$ and $\widehat{D}$ come from the singular value decomposition of $\widehat{\Theta} = \widehat{U}\widehat{D}\widehat{V}^\top$. Estimators subject to other restrictions can be obtained by transforming the PC estimator accordingly. It is worth noting that regardless of the identification restrictions that lead to different versions of the estimated factors, the space spanned by the estimated factors is invariant once $\widehat{\Theta}$ is obtained. In other words, the estimated factor hyperplane already contains all the information regarding the space spanned by the factors..

5.3 Theoretical Properties.

Next, we investigate the theoretical properties of the proposed estimator in Section 5.2.1. The road map of the main steps in establishing these properties is: first, in Section 5.3.1 we derive statistical error bounds of $\widehat{\Theta}$ and $\widehat{B}$ under certain regularity conditions, when the proposed estimation procedure is used on a *deterministic realization* of the observable process X_t. In particular, the regularity conditions assumed primarily entail the *restricted strong convexity* (RSC) condition [**?**] and that the choice of λ_B and λ_Θ are in accordance with some *deviation bound* [**?**]. Moreover, we also provide an error bound for the estimated factor space measured by its $\sin\theta$ distance relative to the true factor space, based upon the magnitude of the error of the estimated factor hyperplane $\widehat{\Theta}$. Subsequently, in Section 5.3.2, we analyze the probability of the required conditions being satisfied as well as the high probability bounds of relevant quantities, for *random realizations* drawn from the underlying observable process X_t and the latent process F_t, under the Gaussianity assumption. From a numerical perspective, we establish the convergence of the proposed iterative algorithm to a stationary point in Section 5.3.3 and finally we discuss connections between the convex formulation adopted in (5.6) and its non-convex original counterpart, as well as those between the proposed framework and related work in the literature in Section 5.3.4. All proofs are deferred to Appendices A and B.

Throughout our exposition, we use superscript $\star$ to denote the true value of the parameters of interest, and denote the errors of the estimators by $\Delta_\Theta := \widehat{\Theta} - \Theta^\star$ and $\Delta_B := \widehat{B} - B^\star$,

[2]For the PC estimator or under the PC2 restriction, where $\mathbf{F}'\mathbf{F}/n = \mathrm{I}_K$ and Λ is assumed lower-triangular, the identification is up to sign rotation; under the PC3 one, where the upper $K \times K$ upper sub-matrix of Λ is assumed an identity matrix and $\mathbf{F}$ is left unrestricted, the identification is exact [see **?**].

respectively. We focus on the estimator obtained through the convex program in (5.6), i.e.,

$$(\widehat{B}, \widehat{\Theta}) = \underset{B,\Theta}{\arg\min} \ \big\{ f_0(B, \Theta) + \lambda_B \|B\|_1 + \lambda_\Theta |||\Theta/\sqrt{n}|||_* \big\},$$
$$\text{subject to} \quad |||\Theta/\sqrt{n}|||_* \leq \phi,$$

where $f_0(B, \Theta) = \frac{1}{2n} \big|\big|\big| \mathbf{X}_n - \Theta - \mathbf{X}_{n-1} B^\top \big|\big|\big|_{\mathrm{F}}^2$. The true value of the parameters $(B^\star, \Theta^\star)$ is assumed feasible.

5.3.1 Statistical error bounds with deterministic realizations.

We first discuss the extra constraint imposed for the feasible region of the above shown convex program. The constraint is on the radius of the nuclear norm ball of Θ and aims to "limit" the total signal of the low rank component through its eigen-spectrum, which is closely associated with the amount of interaction between the latent factor space and the observable space (spanned by X_{t-1}). In particular, since the basis of the factor space is latent, it becomes contrived and impractical to impose any restrictions directly on their interaction. Notwithstanding, such an interaction can be properly bounded by restricting the product of the signals from the two spaces. As it is later shown in the proof of Theorem 5.1 and Remark 5.1, the limited interaction between these two space acts as a relaxed surrogate of the model identifiability restriction, with which the latent and the observable spaces become distinguishable. A larger ϕ will potentially lead to a looser error bound since conceptually the problem becomes more difficult due to the increased degree of interaction between the two spaces allowed. Note that this constraint is in the same spirit as a similar one in **?**] that limits the spikiness of the signal by imposing an ℓ_∞ norm constraint on the low rank regression coefficient in that problem formulation. However, the problem in their setting is fundamentally different from that in the current one, in that the low rank and sparse regression coefficients both operate in the same space whose basis is observed, and the constraint in the form of ℓ_∞ norm arises from being the dual norm of the ℓ_1 norm associated with the sparse component. For our problem, since the basis for the low rank component is non-observed and the factor hyperplane is treated as an "intercept" term, it is no longer sufficient to limit the magnitude of individual entries; rather, a global constraint on the eigen-spectrum proves necessary.

Next, we introduce additional notations needed in the ensuing technical developments. We use K to denote the true dimension of factors, that is, $\mathrm{rank}(\Theta^\star) = K$ and use s to denote the cardinality of the support set of $B^\star$, i.e., $s := \|B^\star\|_0$. Let $S_{\mathbf{X}} := \mathbf{X}_{n-1}^\top \mathbf{X}_{n-1}/n$ denote the sample covariance matrix of the predictors and let $\Lambda_{\max}(S_{\mathbf{X}})$ be its maximum eigenvalue. $S_{\mathbf{E}}$

and $\Lambda_{\max}(S_{\mathbf{E}})$ are analogously defined for the noise process $\mathbf{E}$. Note that $S_{\mathbf{X}}$ corresponds to $\nabla_B^2 f_0$, with the gradient $\nabla_B f_0$ given by $\Upsilon_{\mathbf{X}} := (\mathbf{X}_n - \Theta)^\top \mathbf{X}_{n-1}/n$.

Before stating the main results, we formally define the *RSC condition* [c.f. **?** **?**]:

Definition 5.1 (Restricted Strong Convexity (RSC))**.** For some generic data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, it satisfies the RSC condition with respect to norm Φ with curvature $\alpha_{\text{RSC}} > 0$ and tolerance $\tau_n \geq 0$ if

$$\frac{1}{2n}|||\mathbf{X}\Delta|||_{\text{F}}^2 \geq \frac{\alpha_{\text{RSC}}}{2}|||\Delta|||_{\text{F}}^2 - \tau_n \Phi^2(\Delta), \qquad \forall\ \Delta \in \mathbb{R}^{p\times p}.$$

In our context, we consider the elementwise ℓ_1 norm $\Phi(\Delta) = \|\Delta\|_1$.

Additionally, for the B direction, we say the tuning parameter λ_B is chosen in accordance with the *deviation condition* [**?**] if

$$\lambda_B \geq c_0 \|\Upsilon_{\mathbf{X}} - S_{\mathbf{X}}(B^\star)^\top\|_\infty = c_0 \|\mathbf{X}_{n-1}^\top \mathbf{E}/n\|_\infty, \qquad \text{for some constant } \ c_0 > 0.$$

Theorem 5.1 (Error bound for $(\widehat{B}, \widehat{\Theta})$ under fixed realizations)**.** *Suppose the fixed realizations $\mathbf{X}_{n-1} \in \mathbb{R}^{n\times p}$ of process $X_t \in \mathbb{R}^p$ satisfy the RSC condition with curvature $\alpha_{RSC} > 0$ and a tolerance τ_n such that*

$$\tau_n\Big(s + (2K)\big(\frac{\lambda_\Theta}{\lambda_B}\big)^2\Big) < \min\{\alpha_{RSC}, 1\}/16. \tag{5.9}$$

Then, for any matrix pair $(B^\star, \Theta^\star)$ that generates the evolution of the X_t process, for estimators $(\widehat{B}, \widehat{\Theta})$ obtained by solving the optimization (5.6) *with regularization parameters λ_B and λ_Θ satisfying*

$$\lambda_B \geq \max\Big\{2\|\mathbf{X}_{n-1}^\top \mathbf{E}/n\|_\infty,\ \ 2\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})\Big\} \qquad \textit{and} \qquad \lambda_\Theta \geq \Lambda_{\max}^{1/2}(S_{\mathbf{E}}), \tag{5.10}$$

the following error bound holds:

$$|||\Delta_B|||_F^2 + |||\Delta_\Theta/\sqrt{n}|||_F^2 \leq \frac{64\Big(\lambda_B^2(\sqrt{s}+1)^2 + \lambda_\Theta^2(2K)\Big)}{\min\{\alpha_{RSC}, 1\}^2}. \tag{5.11}$$

The proof of this theorem is in Appendix D.1.

Regarding the quantities appearing in (5.9), (5.10) and (5.11), we make the following comments. First note that the tolerance $\tau_n \geq 0$ measures the extent to which the Hessian of the B direction given by $\big(\mathbf{X}_{n-1}^\top \mathbf{X}_{n-1}/n\big)$ deviates from strong convexity. The smaller τ_n is, the closer the Hessian gets to being strongly convex. For (5.9) to be satisfied, neither

the rank of $\Theta^\star$ nor the cardinality of $\text{supp}(B^\star)$ can be too large. As τ_n decreases (e.g., as sample size increases), the Hessian becomes "more convex" so that conceptually we have more degrees of freedom and get closer to a "low-dimensional regime", so that the model can handle a denser $B^\star$ and a larger number of factors, as manifested by larger permissible values for K and s, respectively. Moving on to (5.10), for the choice of λ_B, the cross-product term $\|\mathbf{X}_{n-1}^\top \mathbf{E}/n\|_\infty$ measures the maximum interaction between the design matrix $\mathbf{X}_{n-1}$ and the noise $\mathbf{E}$, which according to the model assumption should center around 0; and the second term $\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})$ is an upper bound for the latent-observable spaces interaction, presented in the form of product signals. The choice of λ_Θ indicates that it needs to be stronger than the maximum signal coming from the noise in the form of $\Lambda_{\max}^{1/2}(S_{\mathbf{E}})$. Based on (5.10), we further require the regularization parameters to overcome the maximum deviation from zero of $\mathbf{X}^\top\mathbf{E}$, the latent-observable space interaction, and the maximum noise level. A smaller λ_B is required when interactions between associated terms are weaker and a smaller λ_Θ is required if the noise is weaker, thus leading to a tighter error bound for the estimators. Lastly as shown in (5.11), the final error bound depends on the overall curvature of the objective function, as well as the sparsity level of the sparse transition matrix and the true number of factors. In summary, the quantities determining the tuning parameters and the error bound clearly reflect how the information summarized by the factor hyperplane and by the past history of the process itself interact, as well as their balance in order to be able to estimate the model parameters consistently.

Note that Theorem 5.1 establishes the error bound for the estimated factor hyperplane through the quantity Δ_Θ. However, the prime quantity of interest is that of the estimation of the space spanned by the latent factors vis-a-vis the true underlying one. To that end we derive an error bound for $\sin\theta$ that measures the distance between the estimated factor space and the true factor space. In particular, we focus on analyzing the error between the leading rank-K subspace spanned by Θ and $\widehat{\Theta}$, although potentially $\widehat{\Theta}$ could span an r-dimensional subspace (whenever $r \neq K$) that depends on the value of the selected λ_Θ.

Note that regardless of the identification restrictions imposed on the factors, once $\widehat{\Theta}$ is obtained, the space spanned by the factors becomes invariant, since $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{F}}R^\top$ always span the same space for any $K \times K$ rotation matrix R. Therefore, it is sufficient to examine the $\sin\theta$ distance between $\widehat{U}_K$ and $U_K^\star$, where $\widehat{U}_K$ and $U_K^\star$ are the first K left singular vectors corresponding to $\widehat{\Theta}$ and $\Theta^\star$, respectively. Specifically, the angle between the spaces they span is defined as

$$\theta(\widehat{\mathbf{F}}_K, \mathbf{F}) = \theta(\widehat{U}_K, U_K^\star) := \text{diag}\Big(\cos^{-1}(\bar{\sigma}_1), \cos^{-1}(\bar{\sigma}_2), \ldots, \cos^{-1}(\bar{\sigma}_K)\Big), \tag{5.12}$$

where $\bar{\sigma}_1 \geq \bar{\sigma}_2 \geq \cdots \geq \bar{\sigma}_K \geq 0$ are singular values of $\widehat{U}_K^\top U_K^\star$. The following proposition associates the error of $\sin\theta$ to that of Δ_Θ, and its proof is available in Appendix D.1.

Proposition 5.1 ($\sin\theta$ error of the estimated factor space)**.** *Suppose the estimated factor hyperplane $\widehat{\Theta} \in \mathbb{R}^{n\times p}$ is obtained by solving (5.6), whose error is given by $\Delta_\Theta = \widehat{\Theta} - \Theta^\star$. Let σ_1 and σ_K be the leading and the smallest nonzero singular values of $\Theta^\star$. The following bound holds for the $\sin\theta$ distance between the estimated and the true factor spaces:*

$$\left\| \sin\theta(\widehat{\mathbf{F}}_K, \mathbf{F}) \right\|_F^2 \leq \frac{2(2\sigma_1 + \|\Delta_\Theta\|_{op}) \min\left\{ \sqrt{K}\|\Delta_\Theta\|_{op}, \|\Delta_\Theta\|_F \right\}}{\sigma_K^2}. \tag{5.13}$$

The bound in (5.13) is obtained by considering $\widehat{\Theta}$ as a Δ_Θ-perturbation of $\Theta^\star$, and the size of the perturbation is upper bounded in Frobenius norm given by $\|\Delta_\Theta/\sqrt{n}\|_{\mathrm{F}}$ given in Theorem 5.1. The stronger the minimum signal is for the true space (i.e., σ_K), the tighter the $\sin\theta$ error bound will be. Note that for the true space spanned by $\mathbf{F}$, although it is not observable, it can nevertheless be interpreted as a random (but fixed for this specific part of the analysis) realization drawn from the specified VAR model driving the dynamics of F_t, which in turn directly influences the evolution of the observable X_t process.

5.3.2 High probability bounds under random realizations.

Next, we provide high probability bounds/concentrations for the key quantities associated with the derived error bound in Section 5.3.1, for random realizations of the underlying factor and error processes. Specifically, this involves the verification of the RSC condition, as well as the examination of quantities associated with the deviation bound condition to which the choice of $(\lambda_B, \lambda_\Theta)$ needs to conform, as shown in (5.10).

We introduce additional notations for the subsequent technical development. For some generic process $\{X_t\}$, in addition to the auto-covariance function $\Gamma_X(h)$ and its spectral density $g_X(\omega)$, we define its maximum and minimum eigenvalue associated with the spectral density $g_X(\omega)$ introduced in Section 5.2 as follows [**?**]:

$$\mathcal{M}(g_X) := \operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]} \Lambda_{\max}(g_X(\omega)), \qquad \mathfrak{m}(g_X) := \operatorname*{ess\,inf}_{\omega\in[-\pi,\pi]} \Lambda_{\min}(g_X(\omega)).$$

For two generic centered processes $\{X_t\}$ and $\{Y_t\}$ that are assumed jointly covariance stationary, whose spectral density is given by $g_{X,Y}(\omega) := \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Gamma_{X,Y}(h)e^{i\omega h}$ where

$\Gamma_{X,Y}(h) = \mathbb{E}(X_t Y_{t+h}^\top)$, the upper extreme for $g_{X,Y}(\omega)$ is analogously defined as

$$\mathcal{M}(g_{X,Y}) := \operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]} \sqrt{\Lambda_{\max}\big(g_{X,Y}^*(\omega) g_{X,Y}(\omega)\big)}.$$

In general $g_{X,Y}(\omega) \neq g_{Y,X}(\omega)$, but $\mathcal{M}(g_{X,Y}) = \mathcal{M}(g_{Y,X})$.

For the processes involved in our proposed model, we assume that $\{X_t\}$, $\{\epsilon_t\}$ and $\{F_t\}$ are mean zero Gaussian processes. In particular, $\{\epsilon_t\}$ is a noise process that does not exhibit temporal nor cross-sectional dependence, hence it is effectively a Gaussian random vector with covariance $\Sigma_\epsilon = \sigma_\epsilon^2 \mathrm{I}_p$, and its spectral density simplifies to $g_\epsilon(\omega) = \frac{\Sigma_\epsilon}{2\pi}$. Further, we define the shifted process $\{\widetilde{\epsilon}_t := \epsilon_{t+1}\}$ for notation convenience.

The following lemma verifies that with high probability, for random realizations of the process $\{X_t\}$, the RSC condition is satisfied provided that the sample size is sufficiently large:

Lemma 5.1 (verification of the RSC condition)**.** *Consider $\mathbf{X} \in \mathbb{R}^{n\times p}$ whose rows are some random realization $\{x_0, \ldots, x_{n-1}\}$ of the stable $\{X_t\}$ process with dynamic given in* (5.3)*. Then there exist positive constants c_i $(i = 0, 1, 2, 3)$ such that with probability at least $1 - c_1 \exp(-c_2 n)$, the RSC condition holds for $\mathbf{X}$ with curvature α_{RSC} and tolerance τ_n satisfying*

$$\alpha_{RSC} = \pi \mathfrak{m}(g_X), \qquad \text{and} \qquad \tau_n = \alpha_{RSC}\gamma^2\Big(\frac{\log p}{n}\Big)/2 \quad \text{where } \gamma := 54\mathcal{M}(g_X)/\mathfrak{m}(g_X),$$

provided that $n \gtrsim s \log p$.

The next lemma establishes a high probability bound for the interaction term $\mathbf{X}_{n-1}^\top \mathbf{E}/n$ that influences the choice of λ_n through its elementwise ℓ_∞ norm.

Lemma 5.2 (High probability bound for $\|\mathbf{X}_{n-1}^\top \mathbf{E}/n\|_\infty$)**.** *There exist positive constants c_i $(i = 0, 1, 2)$ such that for sample size $n \gtrsim \log p$, with probability at least $1 - c_1 \exp(-c_2 \log p)$, the following bound holds:*

$$\|\mathbf{X}_{n-1}^\top \mathbf{E}/n\|_\infty \leq c_0\Big(\mathcal{M}(g_X) + \mathcal{M}(g_\epsilon) + \mathcal{M}(g_{X,\widetilde{\epsilon}})\Big)\sqrt{\frac{\log p}{n}}. \tag{5.14}$$

Note that with the definition of the shifted processes $\{\widetilde{\epsilon}_t\}$, we have $g_{X,\widetilde{\epsilon}}(\omega) = e^{-ih\omega} g_{X,\epsilon}(\omega)$, which implies $\mathcal{M}(g_{X,\widetilde{\epsilon}}) = \mathcal{M}(g_{X,\epsilon})$. Hence, the term that measures the upper extreme of the cross-spectrum between X_t and the shifted process in (5.14) can be replaced by its unshifted counterpart. Moreover, since $g_\epsilon(\omega) = \frac{\sigma_\epsilon}{2\pi}$, its upper extreme is given by $\mathcal{M}(g_\epsilon) = \Lambda_{\max}(\Sigma_\epsilon)/(2\pi)$.

In the next two lemmas, we provide bounds for the extremes of the eigen-spectra of the sample covariance matrices $S_{\mathbf{X}}$ and $S_{\mathbf{E}}$ for random realizations of $\{X_t\}$ process and the noise vector ϵ_t.

Lemma 5.3 (High probability concentration for $\Lambda_{\max}(S_{\mathbf{X}})$)**.** *Consider $\mathbf{X} \in \mathbb{R}^{n\times p}$ whose rows constitute random realizations $\{x_0, \ldots, x_{n-1}\}$ of the stable $\{X_t\}$ process whose dynamics are given in* (5.3)*. Then, there exist positive constants c_i $(i = 0, 1, 2)$ such that for sample size $n \gtrsim p$, with probability at least $1 - c_1 \exp(-c_2 p)$, the following bound holds:*

$$\Lambda_{\max}(S_{\mathbf{X}}) \leq c_0 \mathcal{M}(g_X).$$

Lemma 5.4 (High probability concentration for $\Lambda_{\max}(S_{\mathbf{E}})$)**.** *Consider $\mathbf{E} \in \mathbb{R}^{n\times p}$ whose rows are independent realizations of the mean zero Gaussian random vector ϵ_t with covariance Σ_ϵ. Then, for sample size $n \gtrsim p$, with probability at least $1 - \exp(-n/2)$, the following bound holds:*

$$\Lambda_{\max}(S_{\mathbf{E}}) \leq 9\Lambda_{\max}(\Sigma_\epsilon).$$

Proofs for Lemmas 5.1 to 5.4 can be found in Appendix D.1. Up to this stage, we have verified the RSC condition and obtained the high probability bounds for quantities that are associated with the choice of $(\lambda_B, \lambda_\Theta)$, for random realizations from the underlying processes. Theorem 5.2 combines the results in Theorem 5.1 and Lemmas 5.1 to 5.4, and provides a high probability error bound of the estimates when the data are random realizations from the underlying processes, as stated next.

Theorem 5.2 (high probability error bound with random realizations)**.** *Suppose we are given a snapshot of length $(n + 1)$ $\{x_0, \ldots, x_n\}$ from the p-dimensional observable process $\{X_t\}$, whose dynamics are described in* (5.3)*. Then, there exist universal positive constants c_i $(i = 1, 2, 3)$ and c_i' $(i = 1, 2)$ such that for sample size $n \gtrsim p$, by solving convex problem* (5.6) *with regularization parameters*

$$\lambda_B = \max\Big\{c_1 \mathbb{Q}_{X,\epsilon}\sqrt{\tfrac{\log p}{n}},\ c_2\phi\mathcal{M}^{1/2}(g_X)\Big\} \quad \textit{and} \quad \lambda_\Theta = c_3\Lambda^{1/2}_{\max}(\Sigma_\epsilon),$$

where $\mathbb{Q}_{X,\epsilon} := (2\pi)\big(\mathcal{M}(g_X) + \mathcal{M}(g_\epsilon) + \mathcal{M}(g_{X,\epsilon})\big)$, the solution $(\widehat{B}, \widehat{\Theta})$ has the following bound with probability at least $1 - c_1' \exp\big(-c_2' \log p\big)$:

$$|||\Delta_B|||_F^2 + |||\Delta_\Theta/\sqrt{n}|||_F^2 \leq \max\Big\{\mathfrak{m}^{-2}(g_X), \pi^2\Big\}\Big(C_1 \cdot s \cdot \max\big\{\frac{\log p}{n}, 1\big\} + C_2 \cdot K\Big),$$

for some positive constants C_i $(i = 1, 2)$ that are independent of n and p.

Remark 5.1. We illustrate the result established in Theorem 5.2, which is inherently associated with model identifiability. First, note that Theorem 5.2 implies the following rate of convergence, since we require $n \gtrsim p$ for adequate concentration in the eigen-spectrum:

$$\|\!|\Delta_B|\!\|_{\mathrm{F}}^2 + \|\!|\Delta_\Theta/\sqrt{n}|\!\|_{\mathrm{F}}^2 = O\Big(\frac{\log p}{n}\Big) + O(1) = O(1).$$

The requirement $n \gtrsim p$ is standard in the factor analysis literature [e.g., **?**]. This result implies that with $p = O(n)$ and as n goes to infinity, the error bound is always bounded by some constant, but not vanishing. Such a non-vanishing bound is the consequence of F_t being latent and the lack of exact identifiability restrictions between the space spanned by F_t and that by X_{t-1}. Consider the full identification of the given model $X_t = \Lambda F_t + BX_{t-1} + \epsilon_t$. Similar to the analysis in **?**] for the informational series of a factor-augmented VAR model, there exist invertible matrices $M_{11} \in \mathbb{R}^{K\times K}$ and $M_{12} \in \mathbb{R}^{K\times p}$ such that

$$X_t = \Lambda F_t + BX_{t-1} + \epsilon_t = \underbrace{(\Lambda M_{11})}_{\bar{\Lambda}} \underbrace{(M_{11}^{-1}F_t - M_{11}^{-1}M_{12}X_{t-1})}_{\bar{F}_t} + \underbrace{(B + M_{12})}_{\bar{B}} X_{t-1} + \epsilon_t, \quad (5.15)$$

which are observationally equivalent to the original model. So for the model to be fully identifiable (including the factors), a total number of $K^2 + Kp$ restrictions is required. If exact identification of the factors is not required, then Kp restrictions are required to separate the space spanned by F_t from that by X_{t-1}. In low dimensional settings with a different model setup, an estimation procedure based on (5.15) that takes into consideration these Kp restrictions can be carried out in the following three steps: (1) projecting on the orthogonal space of the observed variable so that it is profiled out, by multiplying $\mathbb{P}_{\mathbf{X}}^{\perp} := \mathrm{I}_n - \mathbf{X}_{n-1}(\mathbf{X}_{n-1}^{\top}\mathbf{X}_{n-1})^{-1}\mathbf{X}_{n-1}^{\top}$; (2) doing a one-shot estimation as in standard factor analysis, based on the "profiled" model $\mathbb{P}_{\widehat{X}}^{\perp}\mathbf{X}_n = \mathbb{P}_{\widehat{X}}^{\perp}\mathbf{F}\Lambda^{\top} + \mathbf{V} = \underline{\mathbf{F}}\underline{\Lambda}^{\top} + \mathbf{V}$ where $\mathbf{V}$ collects the error term; (3) rotating the intermediate estimates $(\widehat{\underline{\mathbf{F}}}, \widehat{\underline{\Lambda}})$ subject to the restrictions by operating on the inverse with the aid of additional modeling assumptions on the dynamics [see **?**]. In the high-dimensional setting, neither the projection operator, nor its inverse are available and hence the above strategy can not be operationalized. Without imposing additional model assumptions that would be stringent and only made for the sake of mathematical convenience, we formulate an optimization problem instead, and implicitly incorporate the Kp restrictions through the assumption that the amount of interaction between the latent factor space and the past lags of the observable process is appropriately controlled, which manifests itself in the technical developments as the product of the total signal present in these two spaces. Hence, with properly selected tuning parameters, the global minimizer of the convex problem exhibits good statistical behavior in terms of its error that does *not grow*

with p or n, so that there is adequate control over the performance of the estimator, even though this upper bound of the error does not vanish asymptotically. This represents the price to be paid for handling strongly correlated idiosyncratic components in approximate factor models, under minimal identifiability restrictions.

5.3.3 Convergence analysis.

The convergence property of Algorithm V.1 that solves the optimization problem (5.6) can be established using familiar arguments and exploiting its convex nature. As stated in Section 5.2.1, the objective function

$$f(B, \Theta) := f_0(B, \Theta) + \lambda_B \|B\|_1 + \lambda_\Theta |||\Theta|||_*$$

is *jointly convex* in (B, Θ), with a convex feasible region given by

$$\mathrm{dom}(f) = \Big\{ \Theta \in \mathbb{R}^{n \times p}, B \in \mathbb{R}^{p \times p} \;\Big|\; |||\Theta/\sqrt{n}|||_* \leq \phi \Big\}.$$

Thus, it directly follows from **?**] that the alternating minimization that generates the sequence $\{(\widehat{B}^{(k)}, \widehat{\Theta}^{(k)})\}$ converges to a stationary point which is also a global optimum, though the global optimum is not necessarily unique.

5.3.4 Notes on model connections.

To conclude this section, we discuss connections between different formulations of the problem, and compare and contrast the property of our proposed estimator with those obtained in related work in the literature.

Remark 5.2. *Connections between different formulations.* First we note that if the rank constraint in (5.5) were not relaxed to its convex counterpart, it would be natural to consider the following optimization problem:

$$\begin{aligned} (\widetilde{B}, \widetilde{\Theta}) = \underset{B \in \mathbb{R}^{p \times p}, \Theta \in \mathbb{R}^{n \times p}}{\arg\min} & \Big\{ \frac{1}{2n} \big|\big|\big| \mathbf{X}_n - \Theta - \mathbf{X}_{n-1} B^\top \big|\big|\big|_{\mathrm{F}}^2 + \lambda_B \|B\|_1 \Big\}, \\ \text{subject to} \quad & \mathrm{rank}(\Theta) \leq r, \quad |||\Theta/\sqrt{n}|||_* \leq \phi, \end{aligned} \tag{5.16}$$

where we keep the nuclear norm ball constraint as in (5.6) to limit the overall signal in Θ. Derivations along the lines of those in Sections 5.3.1 show that the global optimum $(\widetilde{B}, \widetilde{\Theta})$ possesses a similar statistical error bound with the same rate as the global optimum $(\widehat{B}, \widehat{\Theta})$ of the convex program (5.6), as shown in Theorem D.1 in Appendix D.3. This indicates that the convex relaxation with respect to the Θ block is tight even for the joint problem

involving both sets of parameters (Θ, B). However, the established statistical properties only hold for the *global optimum* of (5.16). On the other hand, due to the non-convex nature of the rank constraint, it is hard to devise an algorithm that provably ensures convergence to this global optimum. Note that it is natural to investigate a penalized (non-convex) Lagrangian formulation of (5.16), and devise a majorization-minimization algorithm that is guaranteed to converge to some stationary point as a result of **?** , Theorem 4], since the regularity conditions needed in terms of the continuity of the gradient and the compactness of the point-to-set map[3] are satisfied. However, beyond asserting that this stationary point satisfies the first-order optimality condition in a local neighborhood, we are not able to further characterize this stationary point in terms of its statistical properties, since such local optima may be far from the global optimum. The analysis of the global optimum of the original problem formulation and discussion of the majorization-minimization procedure are provided for the sake of completeness in Appendix D.3 and to illustrate the delicate nature of the problem under consideration.

Remark 5.3. *Discussion on obtaining the error bound.* Note that for the approximate factor model, a large panel size (large p) is helpful, since the estimated factors are obtained through cross-sectional aggregation. In particular, as discussed in **?**] and subsequent work, by assuming that the leading K eigenvalues of Σ_X diverge, whereas all eigenvalues of Σ_u are bounded, separation between the common factors and the idiosyncratic components is achieved as the panel size p goes to infinity. On the other hand, the Stock-Watson formulation [**?**] adopted in our work which accounts explicitly for strong correlations amongst the coordinates of the idiosyncratic component, leads to a high-dimensional sparse regression modeling framework. Hence, the estimates for the time-lags of the X_t process suffer from the curse of dimensionality, if we do not compensate appropriately by an increase in the sample size. Hence, we need to strike a balance between these two competing forces. Specifically, when updating the estimate of the factor hyperplane by aggregating cross-sectional information and compress it to a subspace with reduced dimension through the SVD, a larger panel p is helpful. On the other hand, when updating the estimate of the sparse transition matrix, a very high p is detrimental, unless appropriately compensated by a larger sample size n. In addition, the temporal dependence of the coordinates of the X_t process along with the presence of the latent factors add further complications. Thus, careful balancing of these competing issues is needed to obtain estimates of the model parameters with adequate error control.

Remark 5.4. *On the error bound in* **?** *].* As briefly mentioned in Section 5.1, **?**] consider a regression model for time series data involving latent factors that is broadly related to our

[3]See definition of the "point-to-set map" in **?**].

posited model, as explained next. Specifically, the following formulation is considered: the i^{th} coordinate of time evolving response Y_t belonging to group g_i is modeled as

$$Y_{it} = \beta_i^\top X_{it} + \lambda_{c,i}^\top F_{c,t} + \lambda_{g_i,t}^\top F_{g,t} + \varepsilon_{it}, \tag{5.17}$$

where $Y_{it} \in \mathbb{R}^p$ is the response, $X_{it} \in \mathbb{R}^{q_i}$ is the observable covariate, and in addition there is a common factor $F_{c,t} \in \mathbb{R}^r$ impacting all coordinates/responses, and a group-specific factor $F_{g_i,t} \in \mathbb{R}^{r_{g_i}}$ impacting certain coordinates/response, with the group membership being latent. The paper provides consistency results for both a low-dimensional regime with $q := \max q_i$ fixed, and a high-dimensional one with q being a function of sample size n and thus diverging. However, we note that to achieve the consistency results for a high-dimensional q, most of the key assumptions imposed on the paper are for the realized samples. On the other hand, these assumptions would fail to hold with high probability in the high-dimensional regime with a diverging number of predictors q, since the paper requires the existence of the generalized inverse of $\mathbf{X}^\top(\mathrm{I}_n - \mathbf{F}(\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top)\mathbf{X}$ (e.g., assumption $D3$), which can be a highly unstable quantity. Such unaccounted randomness potentially affects the rate of convergence of the estimators for the case of random realizations, an issue not addressed in the paper. Further, the sample size requirement is rather stringent. Note that our technical developments and the error bound obtained in our analysis are directly applicable to the estimated parameters of (5.17), with all randomness properly controlled and the sample size requirement compatible with other high-dimensional literature settings [e.g., **? ?**].

5.4 Implementation and Performance Evaluation.

The actual implementation of Algorithm V.1 requires as inputs λ_B, λ_Θ, as well as a choice of ϕ for the constraint on the feasible region corresponding to the nuclear norm ball of Θ. In practice, it is usually difficult to properly choose ϕ that is always compatible with the feasibility assumption on $\Theta^\star$. Notwithstanding, the computation procedure designed for solving the convex program in (5.6) suggests that to obtain the estimates effectively involves alternating between the following two steps: (1) a Lasso update on the rows of B; and (2) SVT updates on Θ. This naturally motivates an empirical algorithm that can be used to obtain the estimates in practice, outlined next in Algorithm 5.2.

Algorithm 5.2 can be viewed as an alternating minimization algorithm that solves

$$\begin{aligned}(\widehat{\Theta}_{\text{emp}}, \widehat{B}_{\text{emp}}) := \arg\min\ & \big\{\frac{1}{2n}\big|\big|\big|\mathbf{X}_n - \Theta - \mathbf{X}_{n-1}B^\top\big|\big|\big|_{\mathrm{F}}^2 + \lambda_B\|B\|_1\big\}, \\ \text{subject to}\ & \operatorname{rank}(\Theta) \le r,\end{aligned} \tag{5.18}$$

Algorithm 5.2: An empirical AM algorithm to obtain $\widehat{B}_{\text{emp}}$ and $\widehat{\Theta}_{\text{emp}}$.

Input: Time series data $\{x_i\}_{i=0}^n$, tuning parameter λ_B, rank constraint r.
Initialization: Initialize with $\bar{\Theta}^{(0)} = \text{SVT}(\mathbf{X}_n)$
Iterate until convergence:

(1) Update $\bar{B}^{(m)}$ with the plug-in $\bar{\Theta}^{(m-1)}$ so that each row j is obtained with Lasso regression (in parallel) and solves

$$\bar{B}_{j\cdot} = \min_{\beta} \Big\{ \frac{1}{2n} \Big\| \big[\mathbf{X}_n - \bar{\Theta}^{(m-1)}\big]_{\cdot j} - \mathbf{X}_{n-1}\beta \Big\|^2 + \lambda_B \|\beta\|_1 \Big\}.$$

(2) Update $\bar{\Theta}^{(m)}$ by singular value thresholding (SVT): do SVD on the lagged value-adjusted hyperplane, i.e.,

$$\mathbf{X}_n - \mathbf{X}_{n-1}\bar{B}^{(m)} = UDV, \qquad \text{where} \quad D := \text{diag}\big(d_1, \ldots, d_r, d_{r+1}, \ldots, d_{\min(n,p)}\big),$$

and construct $\bar{\Theta}^{(m)}$ by

$$\bar{\Theta}^{(m)} = UD_rV, \qquad \text{where} \quad D_r := \text{diag}\big(d_1, \ldots, d_r, 0, \ldots, 0\big).$$

Output: Estimated sparse transition matrix $\widehat{B}_{\text{emp}} = \bar{B}^{(\infty)}$ and the low rank hyperplane $\widehat{\Theta}_{\text{emp}} = \bar{\Theta}^{(\infty)}$.

which is the penalized reformulation corresponding to (5.5). For each update, the partial minimization step with respect to Θ or B ensures that the value of the objective function is always non-ascending, which together with the fact that the objective function is bounded below guarantees convergence of the objective function iterates. In practice, the algorithm is terminated when the descent magnitude of the objective function between successive iterations is smaller than some pre-specified tolerance level. Note that this algorithm does not provide guarantees of convergence to a stationary point of the sequence of $(\bar{\Theta}^{(k)}, \bar{B}^{(k)})$ iterates, which requires stronger assumptions, either that of convexity of the objective function and the constraint region, or uniform compactness of the generated sequence of iterates. Note that (5.18) can be viewed as an empirical relaxation of (5.16) by removing the constraint associated with the nuclear norm ball.

Choice of the tuning parameter λ_B and the rank constraint r. The implementation of Algorithm V.2 requires a specific pair of (λ_B, r) as input. We consider choosing the optimal pair of (λ_B, r) based on the information criterion proposed in **?**], called the Panel Information Criterion (PIC) and defined as:

$$\text{PIC}(\lambda_B, r) := \frac{1}{np} \Big\| \Big\| \mathbf{X}_n - \widehat{\Theta}_{\text{emp}} - \mathbf{X}_{n-1}\widehat{B}_{\text{emp}}^\top \Big\| \Big\|_{\text{F}}^2 + \widehat{\sigma}^2 \Big[\frac{\log n}{n} \|\widehat{B}_{\text{emp}}\|_0 + r\big(\frac{n+p}{np}\big) \log(np) \Big], \quad (5.19)$$

where $\widehat{\sigma}^2 = \frac{1}{np} \| \| \mathbf{X}_n - \widehat{\Theta}_{\text{emp}} - \mathbf{X}_{n-1}\widehat{B}_{\text{emp}}^\top \| \|_{\text{F}}^2$ and $(\widehat{B}_{\text{emp}}, \widehat{\Theta}_{\text{emp}})$ are solutions to (5.18) with the specific pair of plug-in (λ_B, r). We choose the optimal pair (λ_B, r) over the lattice

$\mathcal{G}_{\lambda_B} \times \mathcal{G}_r := \{\lambda_B^{(1)}, \ldots, \lambda_B^{(j_1)}\} \times \{r^{(1)}, \ldots, r^{(j_2)}\}$ that minimizes PIC.

Next, we present selected results of numerical studies to demonstrate the performance of our proposed method, divided into two parts: in Section 5.4.1, we show the performance of estimates obtained by implementing Algorithm V.2, under various model settings that capture different features of the data generating procedure; in Section 5.4.2, we briefly compare the performance of the single-iterate estimator that is obtained by terminating the iterative procedure at iteration 1, with varying factor-lag strength ratios (to be defined later).

Data generating mechanism. To examine different facets of the model settings that impact the performance of the estimates, we generate data according to the lag-adjusted factor model representation $X_t = \Lambda F_t + B X_{t-1} + \epsilon_t$. Since the interaction between the latent factor space and the predictor space based on X_{t-1} is fundamental to both model specification and estimation accuracy, we explicitly model their joint distribution. Specifically, consider the joint distribution of $(X_{t-1}^\top, F_t^\top)^\top$ under the Gaussianity assumption:

$$\begin{pmatrix} X_{t-1} \\ F_t \end{pmatrix} \sim \mathcal{N}\Big(0, \Sigma := \begin{bmatrix} \Sigma_X & \Sigma_{XF} \\ \Sigma_{FX} & \Sigma_F \end{bmatrix}\Big).$$

For the latent factor F_t, we assume $\Sigma_F = \mathrm{I}_K$; for the observed variable X_t, we assume Σ_X is either Toeplitz with exponential decay or has an equal correlation structure, depending on the scenario under consideration:

$$[\Sigma_X]_{ij} = \rho_X^{|i-j|} \ \text{(Toeplitz)} \qquad \text{or} \qquad [\Sigma_X]_{ij\ (i\neq j)} = \rho_X, \mathrm{diag}(\Sigma_X) = 1 \ \text{(equal correlation)}.$$

Entries in Σ_{XF} are closely associated with the factor-lag correlation level, and we generate each entry in Σ_{XF} from $\mathsf{Unif}(-\rho, \rho)$, with $\rho > 0$ specified at different levels. Finally, after Σ_X, Σ_F and Σ_{XF} are generated, we inflate the diagonals of Σ to ensure its positive-definiteness and also satisfying $\Lambda_{\max}(\Sigma)/\Lambda_{\min}(\Sigma) = 10$, then renormalize it so that Σ is a correlation matrix. Due to the recursive nature of the model, to generate data with the designated correlation structure, F_t needs to be generated according to its distribution conditional on X_{t-1}, that is,

$$\big(F_t | X_{t-1} = x\big) \sim \mathcal{N}\big(\Sigma_{FX}\Sigma_X^{-1}x,\ \Sigma_F - \Sigma_{FX}\Sigma_X^{-1}\Sigma_{XF}\big).$$

This procedure imposes the following empirical restriction on the model parameters, so that the generated time series $\{x_0, \ldots, x_n\}$ is stationary:

$$\varrho := \varrho\big(\Lambda\Sigma_{FX}\Sigma_X^{-1} + B\big) < 1.$$

Finally, we generate the noise term according to $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathrm{I}_p)$, for all $t = 0, \ldots, n$.

To generate the sparse transition matrix B, we assume its sparsity level is $2/p$ for each row that corresponds to the coefficients of each single time series regression, and further nonzero entries are generated from $\mathsf{Unif}\big([-\kappa_B - 0.1, -\kappa_B + 0.1] \cup [\kappa_B - 0.1, \kappa_B + 0.1]\big)$. Each entry of the dense factor loading matrix Λ is generated from $\mathsf{Unif}\big([-\kappa_\Lambda - 0.1, -\kappa_\Lambda + 0.1] \cup [\kappa_\Lambda - 0.1, \kappa_\Lambda + 0.1]\big)$. Note that the specification of κ_B and κ_Λ determines the factor-lag strength ratio ($\mathrm{SNR}_F/\mathrm{SNR}_X$), in terms of their contribution to the overall signal. In particular, we empirically calculate the overall SNR, SNR of F_t, and SNR of X_t as (averaged across the p coordinates of the time series panel):

$$\mathrm{SNR} = \frac{1}{p}\sum_{j=1}^{p} \frac{\mathrm{Var}\big([\mathbf{F}\Lambda^\top + \mathbf{X}_{n-1}B^\top]_{\cdot j}\big)}{\mathrm{Var}(\mathbf{E}_{\cdot j})}, \quad \mathrm{SNR}_F = \frac{1}{p}\sum_{j=1}^{p} \frac{\mathrm{Var}\big([\mathbf{F}\Lambda^\top]_{\cdot j}\big)}{\mathrm{Var}(\mathbf{E}_{\cdot j})}, \quad \mathrm{SNR}_X = \frac{1}{p}\sum_{j=1}^{p} \frac{\mathrm{Var}\big([\mathbf{X}_{n-1}B^\top]_{\cdot j}\big)}{\mathrm{Var}(\mathbf{E}_{\cdot j})}.$$

In practice, we need to adjust the values of κ_B, κ_Λ and σ_ϵ jointly to get the desired level of SNR and different allocations of the signal.

To measure the accuracy of the obtained estimates, for the sparse transition matrix B, we use sensitivity (SEN), specificity (SPC) and relative error in Frobenius norm (RErr_B) as performance criteria:

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{SPC} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}, \qquad \mathrm{RErr}_B = |||\Delta_B|||_{\mathrm{F}} / |||B^\star|||_F.$$

For the factor hyperplane, since we don't separately identify the factors and the factor space is invariant to the identification restrictions, we measure the error of the space based on the $\sin\theta$ distance relative to the true factor space ($\sin\theta.\mathrm{Err}$) and the relative error in Frobenius norm of the hyperplane (RErr_Θ):

$$\sin\theta.\mathrm{Err} = |||\sin\theta(U, U^\star)|||_F^2, \qquad \mathrm{RErr}_\Theta = |||\Delta_\Theta|||_{\mathrm{F}} / |||\Theta^\star|||_F,$$

where $\theta(U, U^\star)$ is defined in (5.12).

Throughout all numerical experiments presented in this section, the sample size n is fixed at 200. Moreover, for our proposed iterative estimator, tuning parameters (λ_B, r) are chosen according to PIC, with r ranging between $[\max\{K-2, 1\}, K+2]$. For each parameter setting, the reported results are based on the average of 50 replications.

5.4.1 Performance evaluation of the proposed estimator.

We evaluate the performance of our proposed estimator under the simulation settings listed in Table 5.1. For all settings, we also consider different levels of correlation between F_t and X_{t-1}, with ρ taking values in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Table 5.2 presents various per-

	p	K	ϱ	Σ_X structure	ρ_X	factor-lag SNR ratio
(baseline) A0	100	2	0.5	eqcorr	0.2	2:1
A1	100	2	0.5	Toeplitz	0.5	2:1
A2	300	2	0.5	eqcorr	0.2	2:1
B1	100	2	0.8	eqcorr	0.2	1:1
B2	500	2	0.5	Toeplitz	0.5	2:1
C1	100	5	0.8	eqcorr	0.2	1:1
D1	300	10	0.8	Toeplitz	0.5	2:1
D2	300	5	0.8	eqcorr	0.2	10:1

Table 5.1: Simulation settings for performance evaluation. Settings that vary by 1, 2, 3 and 4 parameters compared with the baseline setting A0 are indexed by A, B, C and D respectively.

formance metrics for $(\widehat{\Theta}_{\text{emp}}, \widehat{B}_{\text{emp}})$ (the subscript is dropped for notation convenience henceforth). It can be seen that the estimates obtained from Algorithm V.2 exhibit good performance in estimating both the factor hyperplane and identifying the skeleton of $\widehat{B}$. In particular, (i) the covariance structure amongst the coordinates of the observed predictor X_{t-1} does not affect the performance, since A0 and A1 yield similar results; (ii) a larger panel size p favors the factor hyperplane estimation as manifested in the form of a smaller $\text{RErr}_{\widehat{\Theta}}$ and $\sin\theta.\text{Err}$, but requires the sparsity of the transition matrix to decrease accordingly (recall that it is set to $2/p$) for it to have comparable performance (A2, B2 vs. A0); (iii) as we decrease the factor-lag SNR ratio so that the autoregressive structure becomes stronger, $\widehat{B}$ has a better $\text{RErr}_{\widehat{B}}$ as a result of larger SNR_X, whereas the factor hyperplane estimation gets compromised (B1, C1 vs. A0). In particular, if the signal is distributed among more factors (higher rank K) with the total SNR_F fixed, the factor hyperplane estimation is worse (B1 vs. C1). In general, the proposed estimator is robust to model stability and can handle systems with many time series, although we have observed that for larger ϱ or p, the algorithm takes more iterations to converge.

5.4.2 Comparison to single-iterate estimates.

To illustrate how the optimal solution to (5.18) obtained by iterating helps in improving the accuracy compared with a single-iterate (SI) estimate, we briefly compare the performance of the two sets of estimates under various settings. Specifically, the SI estimator can be equivalently obtained by first applying a SVD on $\mathbf{X}_n$, and then fitting a sparse VAR(1) model on the residuals. Note that this corresponds to a natural strategy of obtaining estimates for the model under consideration based on past work in the literature and in the absence of all technical developments presented in this paper. In practice, since this procedure also requires as inputs the number of factors r and the penalty parameter λ_B^{si} (which may not necessarily coincide with the optimal choice based on the iterative procedure), during the SVD step we choose the number of factors according to the IC_{p_1} criterion discussed

			$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
A0	$\widehat{B}$	SEN	0.93	0.99	0.98	0.98	0.96
		SPC	0.98	0.95	0.95	0.96	0.96
		RErr	0.86	0.83	0.86	0.85	0.85
	$\widehat{\Theta}$	Rank	2	2	2	2	2
		$\sin\theta$.Err	0.039	0.033	0.034	0.041	0.025
		RErr	0.17	0.16	0.16	0.17	0.15
A1	$\widehat{B}$	SEN	0.99	0.96	0.96	0.93	0.97
		SPC	0.96	0.97	0.97	0.96	0.96
		RErr	0.78	0.84	0.85	0.86	0.86
	$\widehat{\Theta}$	Rank	2	2	2	2	2
		$\sin\theta$.Err	0.031	0.027	0.032	0.020	0.029
		RErr	0.15	0.14	0.15	0.12	0.14
A2	$\widehat{B}$	SEN	0.98	0.97	0.99	0.99	0.94
		SPC	0.98	0.98	0.98	0.97	0.98
		RErr	0.84	0.86	0.85	0.83	0.84
	$\widehat{\Theta}$	Rank	2	2	2	2	2
		$\sin\theta$.Err	0.013	0.012	0.012	0.006	0.005
		RErr	0.13	0.12	0.12	0.10	0.08
B1	$\widehat{B}$	SEN	1.00	1.00	1.00	0.99	0.98
		SPC	0.95	0.95	0.96	0.97	0.97
		RErr	0.74	0.66	0.66	0.81	0.80
	$\widehat{\Theta}$	Rank	2	2	2	2	2
		$\sin\theta$.Err	0.031	0.032	0.027	0.086	0.080
		RErr	0.16	0.16	0.17	0.27	0.25
B2	$\widehat{B}$	SEN	0.98	0.97	0.99	0.91	0.91
		SPC	0.97	0.99	0.97	0.97	0.98
		RErr	0.83	0.80	0.77	0.85	0.84
	$\widehat{\Theta}$	Rank	2	2	2	2	2
		$\sin\theta$.Err	0.005	0.06	0.015	0.005	0.004
		RErr	0.09	0.09	0.13	0.09	0.08
C1	$\widehat{B}$	SEN	1.00	1.00	1.00	1	0.99
		SPC	0.94	0.94	0.94	0.94	0.96
		RErr	0.47	0.70	0.66	0.68	0.71
	$\widehat{\Theta}$	Rank	5	4.72	5	5	5
		$\sin\theta$.Err	0.185	0.414	0.212	0.189	0.315
		RErr	0.24	0.34	0.25	0.24	0.32
D1	$\widehat{B}$	SEN	1.00	1.00	1.00	1.00	1.00
		SPC	0.97	0.99	0.98	0.99	0.99
		RErr	0.84	0.83	0.82	0.82	0.85
	$\widehat{\Theta}$	Rank	10	10	10	10	10
		$\sin\theta$.Err	0.210	0.172	0.201	0.176	0.208
		RErr	0.26	0.21	0.23	0.22	0.23
D2	$\widehat{B}$	SEN	0.93	0.95	0.95	0.90	0.90
		SPC	0.98	0.97	0.96	0.98	0.98
		RErr	0.82	0.86	0.83	0.90	0.90
	$\widehat{\Theta}$	Rank	5	5	5	5	5
		$\sin\theta$.Err	0.053	0.051	0.061	0.056	0.056
		RErr	0.16	0.16	0.18	0.17	0.17

Table 5.2: Performance evaluation of $\widehat{B}$ and $\widehat{\Theta}$ under settings in Table 5.2, based on the average of 50 replications.

in **?**], i.e., we choose the number of factors r over a sequence that minimizes

$$\mathrm{IC}_{p_1}(r) = \log\big(V(r)\big) + r\Big(\frac{p+n}{pn}\Big)\log\big(\frac{np}{n+p}\big), \qquad \text{where} \quad V(r) := \frac{1}{np}\|\!|\mathbf{X}_n - \widehat{\Theta}^{\mathrm{si}}|\!\|_{\mathrm{F}}^2 \ ;$$

and subsequently during the sparse VAR fitting step, we choose the penalty parameter based on BIC, i.e., we choose the λ_B^{si} that minimizes

$$\mathrm{BIC}(\lambda_B^{\mathrm{si}}) = \sum_{j=1}^{p}\log(\mathrm{RSS}_i) + \frac{\log n}{n}\|\widehat{B}^{\mathrm{si}}\|_0, \qquad \text{where} \quad \mathrm{RSS}_j = \|\widehat{\mathbf{U}}_{n,\cdot j} - \widehat{\mathbf{U}}_{n-1}(\widehat{B}^{\mathrm{si}}_{j\cdot})^\top\|_2^2,$$

with $\widehat{\Theta}^{\mathrm{si}} := \bar{\Theta}^{(0)}$, $\widehat{B}^{\mathrm{si}} = \bar{B}^{(1)}$, $\widehat{\mathbf{U}}_n := \mathbf{X}_n - \widehat{\Theta}^{\mathrm{si}}$ and $\widehat{\mathbf{U}}_{n-1}$ being the matrix of its lags. As a remark, the performance of $\widehat{\Theta}^{\mathrm{si}}$ will not be affected by that of $\widehat{B}^{\mathrm{si}}$, but not vice versa.

To compare the iterative and SI estimators in greater depth, we also include the "oracle SI estimator" denoted by $(\widehat{B}^{\mathrm{si}\star}, \widehat{\Theta}^{\mathrm{si}\star})$ to the comparison, obtained by doing the PC estimation with the *true* number of factors as the rank constraint in the first step so that it is forced to return the correct rank, and proceed to fit a VAR(1) model on the residuals. For these three sets of estimates, we plot the SEN, SPC of $\widehat{B}$, $\widehat{B}^{\mathrm{si}}$ and $\widehat{B}^{\mathrm{si}\star}$ as well as the $\sin\theta$.Err, RErr of $\widehat{\Theta}$, $\widehat{\Theta}^{\mathrm{si}}$ and $\widehat{\Theta}^{\mathrm{si}\star}$, over different ρ's that controls the level of correlation between F_t and X_{t-1}. As a remark, since in our study $\sin\theta$.Err is calculated based on the first K singular vectors of the estimated hyperplane, the ordinary and the oracle SI estimates will always have the same $\sin\theta$.Err. In other words, they will always span the same leading-K subspace, and the difference will lie in the tail which is determined by the rank constraint.

The comparison is performed under selected simulation settings from Section 5.4.1 (A0, A2, B1 and D2) that are also listed in Table 5.3. Note that for all four settings, the structure of Σ_X is fixed to be "equally correlated" with $\rho_X = 0.2$. Setting $\boldsymbol{\delta}.1$ is the baseline setting; Setting $\boldsymbol{\delta}.2$ increases the dimension of the observable process, so that it favors the PC estimation in step 1 whose asymptotics rely on the panel size approaching infinity [**?**]; Setting $\boldsymbol{\delta}.3$ considers a more dependent autoregressive structure, manifested by the larger ϱ which makes $\{X_t\}$ less stable and yields a smaller factor-lag SNR ratio. Finally, setting $\boldsymbol{\delta}.4$ is designed to capture scenarios where the correlation between F_t and the idiosyncratic component u_t is weak, with an exceedingly dominating factor and weak signal in X_{t-1}.

	p	K	ϱ	factor-lag SNR ratio	choices of ρ for $\|\mathrm{Cov}(F_t, X_{t-1})\|_\infty$
$\boldsymbol{\delta}.1$	100	2	0.5	2:1	$\{0.1, 0.3, 0.5, 0.7, 0.9\}$
$\boldsymbol{\delta}.2$	300	2	0.5	2:1	
$\boldsymbol{\delta}.3$	100	2	0.8	1:1	
$\boldsymbol{\delta}.4$	300	5	0.8	10:1	

Table 5.3: Simulation settings for comparing the iterative estimator and the one-shot estimator.

As depicted in Figures 5.1, 5.2 and 5.3, the iterative estimator outperforms the both the ordinary and oracle SI versions across all performance metrics, and the discrepancy in

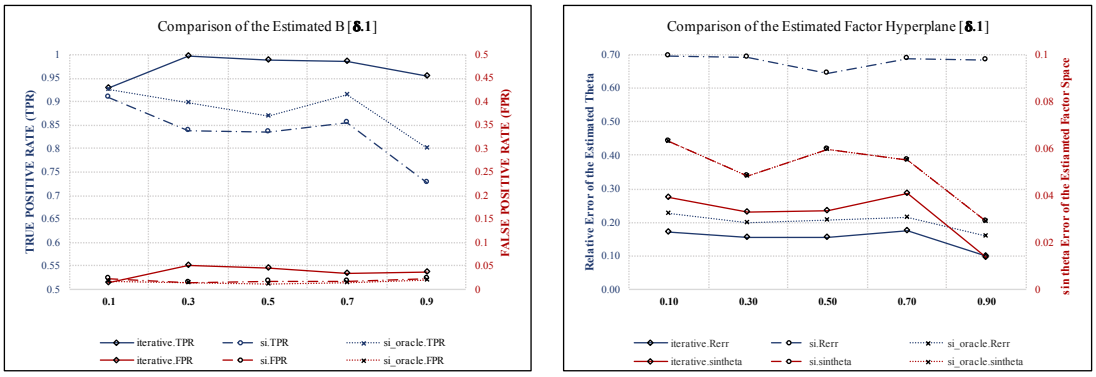

Figure 5.1: Comparison for Setting $\boldsymbol{\delta}$.1. Left panel: the true positive rate (sensitivity) (left axis, in blue) and the false positive rate (1-specificity) (right axis, in red) of $\widehat{B}$, $\widehat{B}^{\mathrm{si}}$ and $\widehat{B}^{\mathrm{si}\star}$. Right panel: relative error in Frobenius norm (left axis, in blue) and the sinθ error (right axis, in red) of $\widehat{\Theta}$, $\widehat{\Theta}^{\mathrm{si}}$ and $\widehat{\Theta}^{\mathrm{si}\star}$. Note: $\mathrm{rank}(\widehat{\Theta}) \equiv 2$, $\mathrm{rank}(\widehat{\Theta}^{\mathrm{si}}) \equiv 1$.

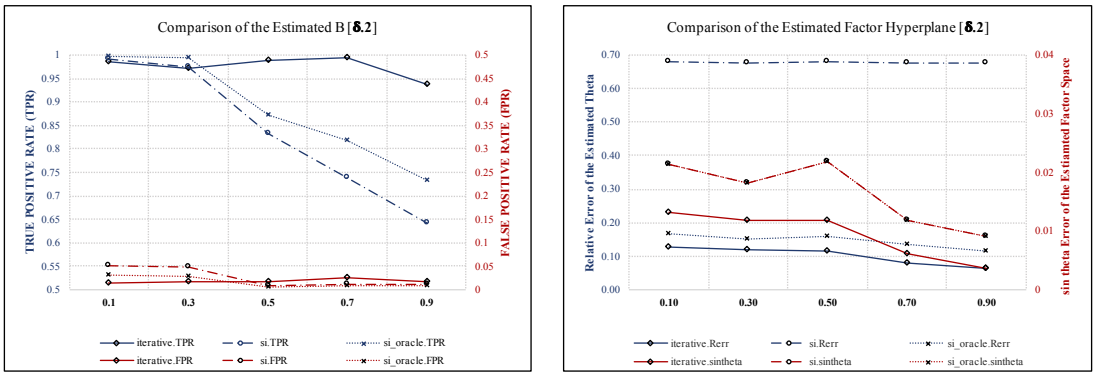

Figure 5.2: Comparison for Setting $\boldsymbol{\delta}$.2. Left panel: the true positive rate (sensitivity) (left axis, in blue) and the false positive rate (1-specificity) (right axis, in red) of $\widehat{B}$, $\widehat{B}^{\mathrm{si}}$ and $\widehat{B}^{\mathrm{si}\star}$. Right panel: relative error in Frobenius norm (left axis, in blue) and the sinθ error (right axis, in red) of $\widehat{\Theta}$, $\widehat{\Theta}^{\mathrm{si}}$ and $\widehat{\Theta}^{\mathrm{si}\star}$. Note: $\mathrm{rank}(\widehat{\Theta}) \equiv 2$, $\mathrm{rank}(\widehat{\Theta}^{\mathrm{si}}) \equiv 1$.

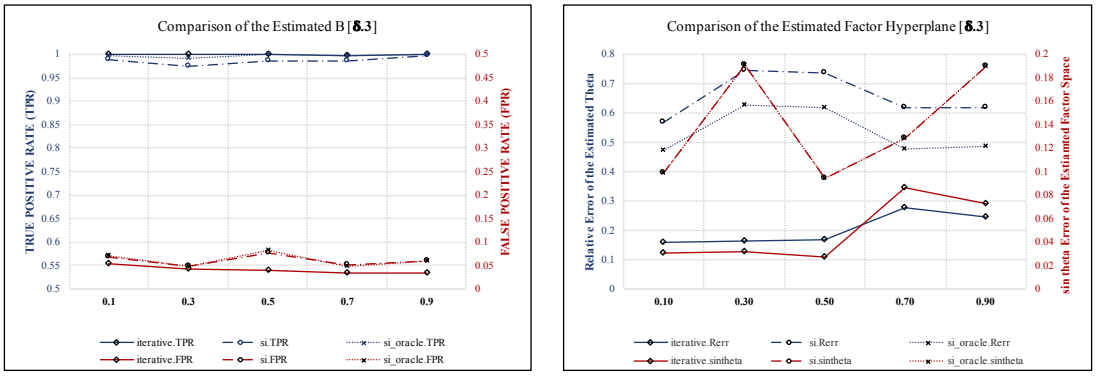

Figure 5.3: Comparison for Setting $\boldsymbol{\delta}$.3. Left panel: the true positive rate (sensitivity) (left axis, in blue) and the false positive rate (1-specificity) (right axis, in red) of $\widehat{B}$, $\widehat{B}^{\mathrm{si}}$ and $\widehat{B}^{\mathrm{si}\star}$. Right panel: relative error in Frobenius norm (left axis, in blue) and the sinθ error (right axis, in red) of $\widehat{\Theta}$, $\widehat{\Theta}^{\mathrm{si}}$ and $\widehat{\Theta}^{\mathrm{si}\star}$. Note: $\mathrm{rank}(\widehat{\Theta}) \equiv 2$, $\mathrm{rank}(\widehat{\Theta}^{\mathrm{si}}) \equiv 3$.

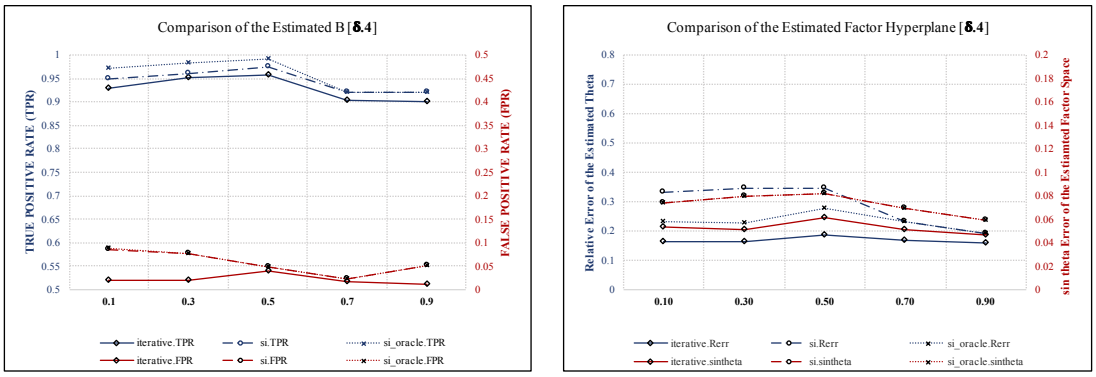

Figure 5.4: Comparison for Setting $\boldsymbol{\delta}.4$. Left panel: the true positive rate (sensitivity) (left axis, in blue) and the false positive rate (1-specificity) (right axis, in red) of $\widehat{B}$, $\widehat{B}^{\text{si}}$ and $\widehat{B}^{\text{si}\star}$. Right panel: relative error in Frobenius norm (left axis, in blue) and the sinθ error (right axis, in red) of $\widehat{\Theta}$, $\widehat{\Theta}^{\text{si}}$ and $\widehat{\Theta}^{\text{si}\star}$. Note: rank$(\widehat{\Theta}) \equiv 5$, rank$(\widehat{\Theta}^{\text{si}}) = 7$ for $\rho \in \{0.1, 0.3, 0.5\}$, rank$(\widehat{\Theta}^{\text{si}}) = 5$ for $\rho \in \{0.7, 0.9\}$.

the accuracy of estimating $\widehat{B}$ becomes larger as we increase the level of correlation between F_t and X_{t-1}. Under settings $\boldsymbol{\delta}.1$ and $\boldsymbol{\delta}.2$ where the signal for the factor is strong and that for the lag is relatively weak, for the estimated B (left panels), when ρ is small, all estimators exhibit good performance as indicated by a high TPR and a low FPR. As ρ increases, TPR of the iterative estimator (solid line) stays above 0.9, where that of the SI estimators drops significantly, where $\widehat{B}^{\text{si}\star}$ (dotted line) has a slightly higher TPR compared with $\widehat{B}^{\text{si}}$ (dashed line). For all three estimators, FPR stays at a relative low level ($0.01 \leq$ FPR ≤ 0.05) in all settings. For the estimated hyperplane, (i) the iterative estimator always correctly estimates the number of factors (rank), whereas the SI one systematically under-estimates the rank; (ii) the inaccuracy of the rank determination leads to a large relative error for the estimated hyperplane ($\text{RErr}_{\widehat{\Theta}^{\text{si}}}$ is around 0.7 versus $\text{RErr}_{\widehat{\Theta}^{\text{si}\star}}$, $\text{RErr}_{\widehat{\Theta}}$ that are both around 0.2); (iii) despite the rank being correctly specified for the oracle version, $\widehat{\Theta}^{\text{si}\star}$ still exhibits worse performance compared with $\widehat{\Theta}$ in both the relative error in magnitude and the $\sin\theta$ error regarding the space it spans; this result is in accordance with the conclusion of Proposition 5.1. Under setting $\boldsymbol{\delta}.3$ where the factor and lag signals are comparable, all three estimators show good performance in estimating the transition matrices in all settings, with the iterative estimator exhibiting a slightly smaller FPR. The major problem for the SI estimators lies in the factor hyperplane estimation, which systematically over-estimates the true rank and has a significantly higher relative error, even for the oracle $\widehat{\Theta}^{\text{si}\star}$. Note that $\text{RErr}_{\widehat{\Theta}}$ again stays around 0.2, whereas $\text{RErr}_{\widehat{\Theta}^{\text{si}}}$ and $\text{RErr}_{\widehat{\Theta}^{\text{si}\star}}$ can be as high as 0.6 to 0.7. Turning to Figure 5.4 which corresponds to Setting $\boldsymbol{\delta}.4$, where we have an exceedingly strong factor and a weak auto-regressive structure, both the SI and iterative estimators exhibit similar performance. In particular, the ordinary SI estimator obtains the correct rank for $\rho = 0.7$ and $\rho = 0.9$, which makes it identical to the oracle SI estimator and the

two have the same overall performance.

We also examined the performance of the SI estimates when the filtering VAR step is applied first and then the FA model on the residuals (results not included) for settings $\boldsymbol{\delta}.3$ and $\boldsymbol{\delta}.4$. For the former, where the factor contributes equally to the VAR component, this SI estimate identifies the rank correctly most of the time, but the specificity of the B estimates suffer (values around 0.8) together with the $\sin\theta$.Err (values around 0.45). For the latter setting, where the factor is the dominant contributor, the specificity of the B estimates further deteriorate (values around 0.6), as well as those for $\sin\theta$.Err (values ranging from 1.2-2.8).

The simulation results clearly demonstrate our earlier point that the proposed model formulation is capable of incorporating the correlation between the common factors and the idiosyncratic error, as well as that amongst the coordinates of the idiosyncratic component. After re-writing the model as a lag-adjusted factor model, the devised estimation procedure yields a more accurate reconstruction of the factor hyperplane vis-a-vis the traditional factor analysis through PC estimation (or SVD, equivalently), thus overcoming the difficulties noted by **?**] in their empirical work. In addition, an estimated transition matrix with good selection properties is available for further analysis even with moderate signal strength of the corresponding component in the model (that translates into relatively strong correlations amongst the coordinates of the idiosyncratic component).

5.5 Application to Log-Returns of US Financial Assets.

Factor models have been widely used in financial applications. In particular, they have been employed in analyzing the dynamics of asset returns, either for the purpose of identifying risk factors, or for estimating the covariance structure amongst assets for better portfolio diversification and asset allocation [e.g., **?**]. We applied the proposed modeling framework to a set of stocks return data corresponding to 75 large US financial institutions, which also exhibit strong (serial) correlation in the error terms. Specifically, we analyze the log-returns of 25 banks, 25 insurance companies and 25 broker/dealer firms for the period of 2001-16. Note that this time period contains a number of significant events for the financial industry, including the growth of mortgage bank securities [**?**] in the early 2000s, rapid changes in monetary policy in 2005-06, the great financial crisis [**?**] in 2008-09 and the European debt crisis in 2011-12 and their aftermath. Our analysis identifies a number of interesting patterns, especially around the period 2007-09 encompassing the beginning, height and immediate aftermath of the US financial crisis, both through changes in the factor structure and the partial autocorrelation one governed by the VAR model transition matrix of the log-returns

of these financial assets.

Data. The data consist of weekly stock return data corresponding to 75 large financial institutions in terms of market capitalization, for the period of January 2001 to December 2016 and were obtained from the Center for Research in Security Prices (CRSP) database. The 75 companies are categorized into three sectors: banks (SIC code 6000–6199), broker/dealers (SIC code 6200–6299) and insurance companies (SIC code 6300–6499), with 25 in each sector [see also **?**]. As we require that the data be available for the entire time span under consideration, there are 56 firms that are kept for further analysis, since the remaining ones either went bankrupt or were forced to merge with financially healthier companies (e.g. Lehman Brothers and Merill Lynch in 2008, respectively). To get an overview of the correlation structure amongst the stocks after accounting for the first principal component that captures the average return of the portfolio comprising of the 56 stocks under consideration [**?**], we plot the correlation among the principal component regression residuals. We consider the entire period as well as three sub-periods that have been considered in the literature [c.f. **?**]: 2001–2006 (pre-crisis), 2007–2009 (crisis), 2010-2016 (post-crisis), and plot the corresponding correlation map. As Figure 5.5 demonstrates, overall, we observe positive correlation within each sector and negative correlation across the three sectors (top left panel). Such a structural pattern is particularly predominant in the pre-crisis period (top right panel), and is significantly less pronounced in the post crisis one (bottom right panel), but gets disrupted during the crisis (bottom left panel). Specifically, apart from weakened within-block positive correlations, strong negative and positive correlations across blocks are observed. This suggests that different factor and auto-regressive structures emerge during the crisis period. Further, note that similar results hold if we examine the residuals after removing a second principal component, so as to capture a larger percentage of variance of the stock returns.

The analysis is based on 104-week-long rolling windows to avoid issues with non-stationarity due to length of the period under consideration. This strategy has also been used in [**?** **?**] and allows monitoring change in the number of factors over time, as well as the sparsity level of B which measures the connectivity of the partial autocorrelation network across these financial institutions. We also track the change in R-squared and the R-squared attributed to the factor over time, as a surrogate for the quality of the model fit. We fit the proposed lag-adjusted factor model in each time window, with tuning parameters selected according to PIC described in Section 5.4.

As Figure 5.6 shows, sharp changes are observed in both the factor structure and the temporal dependence of stock returns during the crisis period. In particular, two change

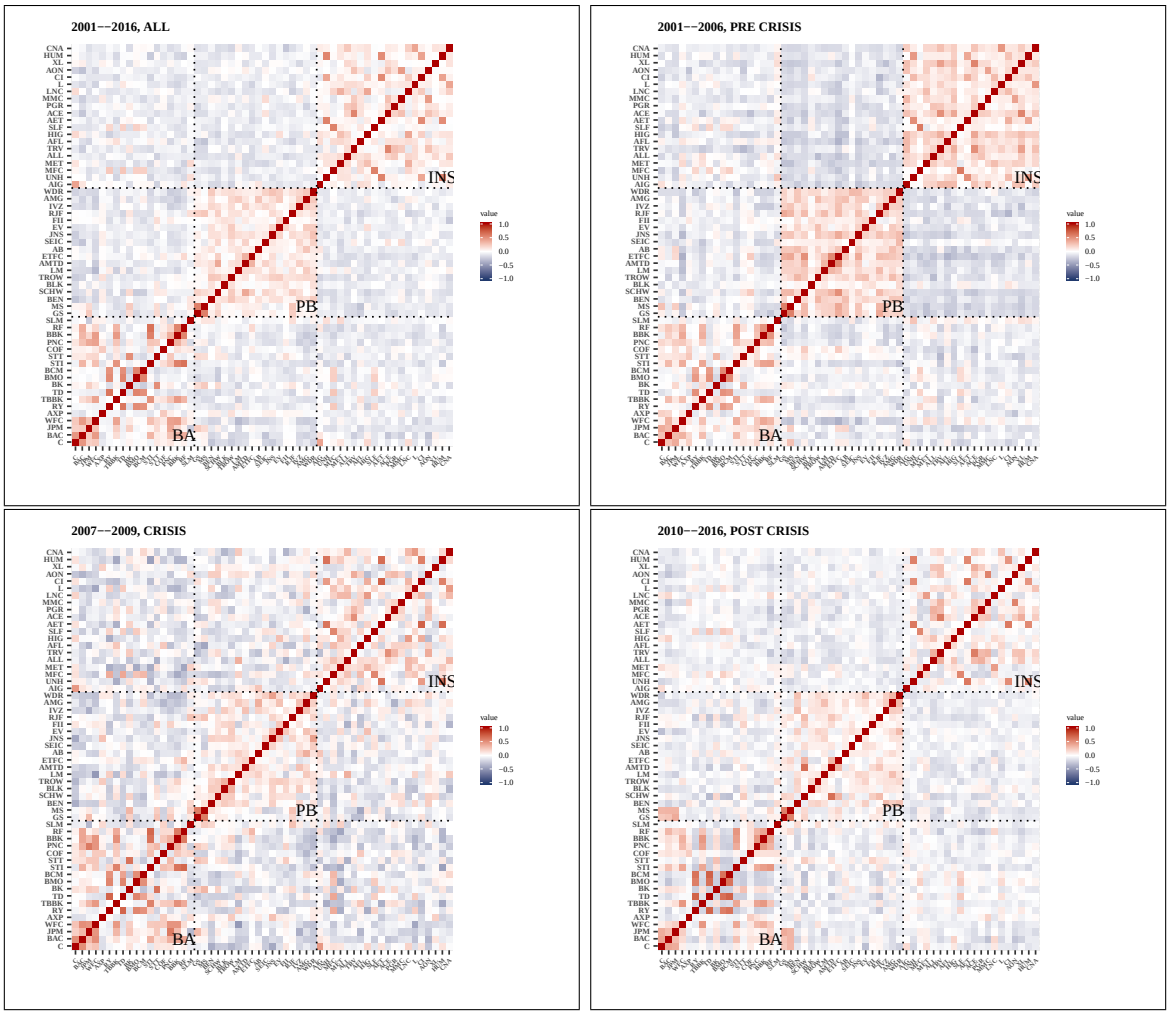

Figure 5.5: Correlation map for PCR residuals with the number of PC fixed at 1. Top left panel: 2001–2016; top right panel: 2001–2006, pre-crisis; bottom left panel: 2007-2009, crisis; bottom right panel: 2010–2016, post-crisis. Red to blue corresponds to correlation from 1 to -1.

points respectively correspond to the beginning of the 2007 sub-prime mortgage crisis and the ending of the 2008–2009 global financial crisis. Specifically, during the pre- and post-crisis periods, only 1 factor is detected most of the time with the density of the transition matrix being close to 0, suggesting that not much serial correlation exists in the idiosyncratic component after the common factor (surrogate for market portfolio) is accounted for. Nevertheless, our analysis identifies sharp changes in the connectivity of $\widehat{B}$ from 2005 to 2006 that is concordant with the period during which the Federal Reserve frequently adjusted monetary policy by rapidly raising the Federal Funds Rate from 2.5% as of 2/2/05 to 5.5% on 6/29/06. Since financial stock prices are sensitive and fast responding to changes in interest rates, these successive increases in the Fed Funds Rate are reflected through the autoregressive component of the model. During the crisis period, up to 5 factors are detected, together with a sharp increase in the connectivity pattern in $\widehat{B}$ reaching its maximum to-

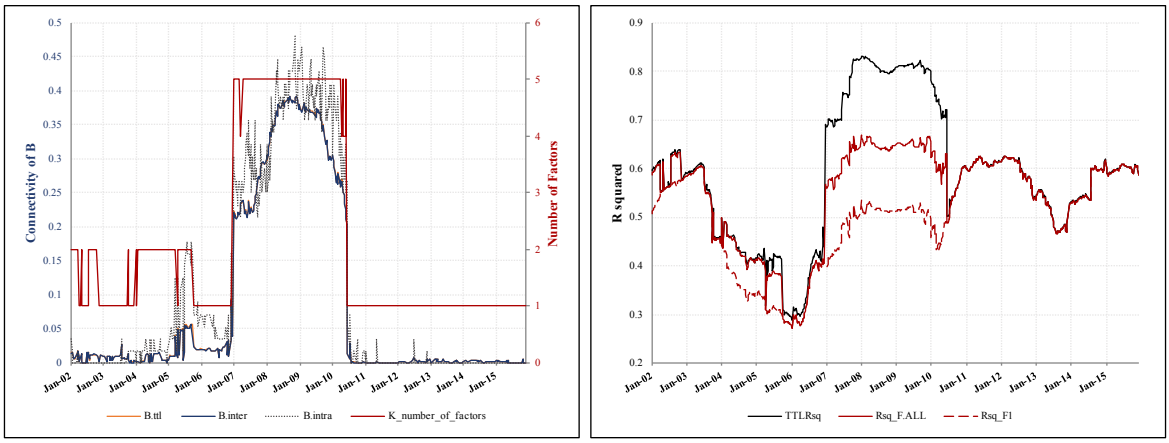

Figure 5.6: Results after fitting the model to the real data based on 104-week-long rolling windows over time. Left panel: number of factors (right axis, in red) and the connectivity level of $\widehat{B}$ (left axis, in blue). Right panel: overall R-squared (solid line) and the R-squared attributed to the factor (dotted line).

wards the end of 2008. In addition the R-squared jumps to 0.80, with 0.65 attributed to the factor hyperplane and accompanied by the largest "R-squared gap". Specifically, under normal market conditions, most of the fit of the model (R-squared) comes from the factor hyperplane (as the solid and dotted lines almost overlap); whereas during the crisis period, the lag term explains a significant proportion of the R-squared, indicating the presence of significant cross autocorrelations in the lag returns. Further, the proposed model provides a much better fit to the data than the standard approximate factor model available in the literature.

We further investigate the composition of the factors and the major emitters/receivers for the network (transition matrix) during the crisis period. The singular values of the estimated factor hyperplane correspond to the strength of each factor identified. Not surprisingly, we have one dominant factor that accounts for around 50% of the signal in the eigen-spectrum, with the rest distributed uniformly amongst the other four factors. Further, assuming orthogonality amongst the factors for reconstruction purposes, their loadings is retrieved from the right singular vector of $\widehat{\Theta}$, up to sign rotation. As depicted in Figure 5.7, all financial institutions contribute positively to the first factor, with Citigroup (C) and Bank of America (BAC) being the top contributors. The composition of the second factor shows an interesting pattern: negative contributors primarily belong to brokers/dealers and insurers, whereas commercial banks contribute positively, albeit weakly. However, AIG (an insurance company) contributes positively to the factor, unlike its peers and is consistent with other findings that it played a prominent role during the crisis.[4] The composition of the remaining three factors is relatively unstructured. In Figure 5.8, we plot the partial

[4]According to an estimate as of January 2010, AIG accounted for 38% of the total losses incurred by insurance companies (98.2 out of 261.0 billions) since 2007. Source: Bloomberg, see also **?**].

autocorrelation network of the firms during the crisis after properly thresholding the entries that have small magnitudes, with red denoting a positive link and with blue a negative one. A careful examination of the node weighted in/out-degrees shows that the top emitters are relatively uniform, in the sense that their weighted out-degrees do not differ by much; whereas the top receivers are dominant, since the weighted in-degrees for HIG and AIG (two insurance companies) are significantly higher compared with the rest. It is worth noting that all top emitters concentrate in the insurance sector. Meanwhile, some of the top receivers are also major contributors to the factors' composition, e.g., HIG to the 2nd factor, AIG to the 3rd, and ETFC to the 4th. This is an interesting finding, given the role that many insurance companies played in magnifying the impact of the crisis on the overall stability of the financial system, due to their large insurance underwriting of Credit Default Swaps and subsequent exposure to accentuated risks [**?**]. However, this analysis points to the importance of insurance companies based on publicly available data and before their role in the crisis was fully revealed and understood.

To conclude this section, we compare and contrast our results with those obtained in **?**], in which the authors consider 100 financial institutions comprising of the largest 25 among each of the four categories: hedge funds, broker/dealers, banks and insurers; thus, that data set is enhanced by the inclusion of big hedge funds for which publicly available stock quotes are not accessible. From the systemic risk standpoint, the authors measure the connectedness of the system based on principal component analysis (PCA) and Granger-causality network analysis during the 1994–2008 period, and identify increased level of interconnectedness during the crisis period and the asymmetry in the degree of connectedness amongst different sectors. Our results are qualitatively similar to these results, and the conclusions broadly match. However, we would like to highlight some key differences in both modeling and in the empirical results obtained. From the modeling perspective, **?**] consider two separate modeling strategies: (i) a Principal Components Analysis (akin to a static factor model) and (ii) a Granger-causality based analysis through fitting a VAR model for each pair of stocks returns. The PCA analysis examines a *fixed* number of principal components/factors and the authors argue that the increasing proportion of variation explained by them is an indication of the systematic response of the financial system to the crisis. Their pairwise based Granger-causal network also reveals increased connectivity during the crisis period. Our model considers latent factors and lead-lag relationships among stock returns *simultaneously*, thus gaining better and more informative insights. In addition, the lead-lag relationships are considered across all firms simultaneously rather than in a pairwise fashion. By incorporating the strong correlations present in the idiosyncratic component (see also Figure D.1), our model is more parsimonious. Specifically, during the crisis, **?**] uses 10 principal components to account

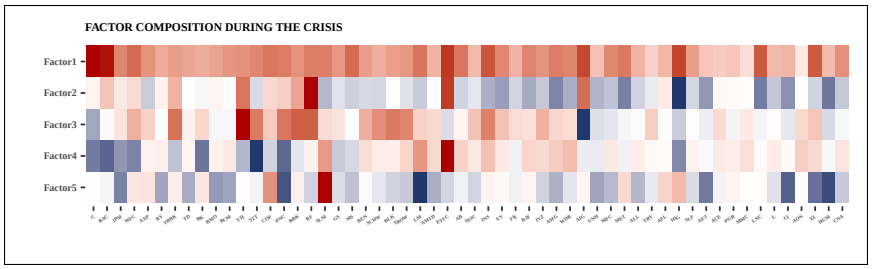

Figure 5.7: Composition of the 5 factors identified during the crisis period.

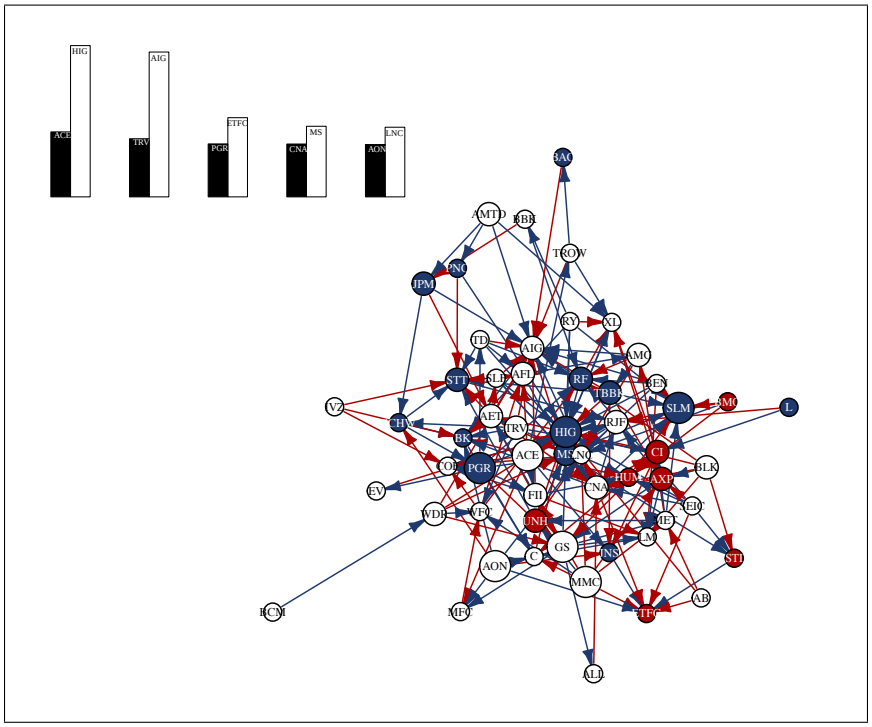

Figure 5.8: Partial autocorrelation network during the crisis, after proper thresholding of entries with small magnitudes. Top emitters (in black bars): ACT, TRV, PGR, CNA, AON. Top receivers (in white bars): HIG, AIG, ETFC, MS, LNC.

for 85% of the returns variance, whereas only 5 suffice in our model; further, the leading factor in their analysis only accounts 37% of the variance, compared to 50% in our model. Finally, extending the analysis period to 2016 shows that after 2011 the influence of banks and insurance companies on stock returns waned, as the marker slowly returned to normalcy. However, we are in broad agreement with the **?**] conclusion on the heightened role of banks and insurers up to 2009 (see figures in the Supplementary Material).

5.6 Extensions.

The proposed modeling framework and estimation procedure are easily generalizable to cases where the idiosyncratic error u_t exhibits further into the past temporal dependence, and come with similar theoretical guarantees. Specifically, we use a sparse VAR(d) model

to account for such dependency, that is,

$$\mathcal{B}_d(L)u_t = \epsilon_t, \qquad \text{where} \qquad \mathcal{B}_d(L) = \mathrm{I}_p - B_1 L - \cdots - B_d L^d, \quad \epsilon_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2 \mathrm{I}_p).$$

with B_k's assumed sparse. By stacking the lagged values of the factors and the corresponding loading matrices, the dynamic of the observable process X_t can be written in the following form, in terms of the latent static factor $F_t \in \mathbb{R}^K$:

$$X_t = \Lambda F_t + B_1 X_{t-1} + B_2 X_{t-2} + \cdots + B_d X_{t-d} + \epsilon_t. \tag{5.20}$$

Similar to the VAR(1) case, the condition required for stationarity is the same as cases where X_t were a VAR(d) process, that is, all roots of $\det(\mathcal{B}_d(z))$ should lie outside the unit circle: $\det(\mathcal{B}_d(z)) \neq 0$ for all $|z| \leq 1$.

Estimation and theoretical guarantees. Given a snapshot of the realizations from $\{X_t\}$, denoted by $\{x_0, \ldots, x_n\}$, we can estimate $\{B_k, k = 1, \ldots, d\}$ and the factor hyperplane in an analogous way. Specifically, let the contemporaneous response and the lagged predictor matrices be $\mathbf{X}_n^d \in \mathbb{R}^{n_d \times p}$ and $\mathbf{X}_{n-1}^d \in \mathbb{R}^{n_d \times dp}$ by stacking the observations in their rows with $n_d = n - d + 1$ being the sample size. $\mathbf{E}_n^d$ is similarly defined to $\mathbf{X}_n^d$. Further, letting $B := [B_1, B_2, \ldots, B_d] \in \mathbb{R}^{p \times dp}$, then with $\mathbf{F}$ and Θ identically defined to those in Section 5.2.1, we can write

$$\mathbf{X}_n^d = \mathbf{F}\Lambda^\top + \mathbf{X}_{n-1}^d B^\top + \mathbf{E}_n^d.$$

$\widehat{B}$ and $\widehat{\Theta}$ can be obtained by solving an analogously formulated optimization, that is

$$\begin{aligned}(\widehat{B}, \widehat{\Theta}) := \underset{B \in \mathbb{R}^{p \times dp}, \Theta \in \mathbb{R}^{n_d \times K}}{\arg\min} \Big\{ \frac{1}{2n_d} |||\mathbf{X}_n^d - \mathbf{X}_{n-1}^d B^\top - \Theta|||_F^2 + \lambda_B \|B\|_1 + \lambda_\Theta |||\Theta|||_* \Big\}, \\ \text{subject to} \quad |||\Theta/\sqrt{n_d}|||_* \leq \phi.\end{aligned} \tag{5.21}$$

Empirically at each iteration, Θ is updated by SVT with hard-thresholding and each row of B is updated via Lasso regression.

With deterministic realizations based on which we solve the optimization problem, we can obtain essentially the same error bound, with the conditions imposed on the corresponding augmented quantities. Formally, the error bound is given in the next corollary, with a superscript $\star$ associated with the true value of the parameters, $\underline{s} := \sum_{k=1}^d \|B_k^\star\|_0$ being the overall sparsity, and $S_{\mathbf{X}}$ being the sample covariance matrix corresponding to $\mathbf{X}_{n-1}^d$.

Corollary 5.1 (Error bound under VAR(d) dependence)**.** *Suppose the observations stacked in $\mathbf{X}_{n-1}^d$ are deterministic realizations from $\{X_t\}$ process with dynamic given in* (5.20)*, and*

$\mathbf{X}^d_{n-1}$ *satisfies the RSC condition with curvature* $\alpha_{RSC} > 0$ *and a tolerance* τ_{n_d} *such that* $\tau_{n_d}\Big(\underline{s} + (2K)\big(\frac{\lambda_\Theta}{\lambda_B}\big)^2\Big) < \min\{\alpha_{RSC}, 1\}/16$. *Then for any matrix pair* $(B^\star, \Theta^\star)$ *that drives the dynamic of* X_t, *for estimators* $(\widehat{B}, \widehat{\Theta})$ *obtained by solving* (5.21) *with* λ_B *and* λ_Θ *chosen such that*

$$\lambda_B \geq \max\big\{2\|(\mathbf{X}^d_{n-1})^\top \mathbf{E}^d_n/n_d\|_\infty,\ 2\phi\Lambda^{1/2}_{\max}(S_{\mathbf{X}})\big\} \quad \textit{and} \quad \lambda_\Theta \geq \Lambda^{1/2}_{\max}(S_{\mathbf{E}}),$$

the following error bound holds:

$$|||\Delta_B|||_F^2 + |||\Delta_\Theta/\sqrt{n_d}|||_F^2 \leq \frac{64\Big(\lambda_B^2(\sqrt{\underline{s}}+1)^2 + \lambda_\Theta^2(2K)\Big)}{\min\{\alpha_{RSC}, 1\}^2}. \tag{5.22}$$

5.7 Discussion.

In this chapter, we introduced a novel modeling framework that generalizes the classical approximate factor model to include lags of the observable process, so that stronger correlations among the idiosyncratic component are allowed. The autoregressive structure is assumed to be sparse, which enables its estimation for large time series panels. Estimation of the model parameters is based on a convex optimization problem, and the resulting estimates have high probability error bounds that can be expressed in terms of key structural parameters (n, p, K, s, etc.), and exhibit superior empirical performance in synthetic data.

In addition to generalizing the model in **?**], our proposed model can also be perceived as a robust treatment of endogeneity. Specifically, as noted by **?**], in the presence of large values in u_t and for a relatively small panel size p, the factor estimates will be distorted as a result of this endogeneity. Here, by explicitly taking into consideration the lag term X_{t-1} in the dynamic evolution of X_t, the noise term ϵ_t becomes strictly exogenous. Our proposed model and estimation procedure has the capacity of handling much stronger correlation between F_t and u_t, although ultimately we do require $\mathrm{Cov}(F_t, u_t)$ to be indirectly bounded in some appropriate way.

CHAPTER VI

Conclusion

The main contributions of this thesis are: (i) the estimation and testing of the high-dimensional VAR-X model; (ii) the extension-to-high-dimensional-setting of the FAVAR model; and (iii) the relaxation of the weak correlation assumption of the approximate factor model.

On the modeling front, as a special instance of the high-dimensional VAR models, we first considered a VAR-X model with the endogenous and exogenous blocks being components of a VAR system and having group Granger-causal ordering. We provided estimates of its model parameters under structural assumptions on the transition matrices and the inverse covariance matrices, and devised a procedure for testing the existence of such group Granger-causality. Moving to FAVAR models, we extended them to the high-dimensional setting and enabled their model parameter estimation under such setting, by investigating their model identifiability issues and formulating an optimization problem that incorporates the newly proposed identification constraint. Finally within the DFM scheme, we relaxed the weak correlation assumption, proposed a new model that accommodates stronger correlations, and provided estimates for the model parameters.

On the theoretical front, we addressed several technical challenges that are further compounded by the presence of temporal dependence among the data. First, we proved the consistency properties of the estimators obtained from the penalized maximum likelihood formulation which jointly estimates the regression coefficient and the covariance matrix. Moreover, we established the algorithmic convergence of the alternating minimization procedure, leveraging the bi-convexity of the objective function and the descent property of each update. Further, we investigated the statistical properties of the estimators corresponding to a compactified low-rank-plus-sparse formulation that is based upon the least squares loss, and established high-probability finite sample error bounds for the estimators.

On the application front, we employed the aforementioned models to several economic datasets involving stock prices, commodity prices and major macroeconomic indicators. The

results have revealed interesting connectivity patterns amongst or across these sets of variables, and have pointed to future exploration and refinements.

APPENDICES

APPENDIX A

Supplementary Materials to "Penalized Maximum Likelihood Estimation of Multi-layered Gaussian Graphical Models."

A.1 Proofs for main theorems.

Proof of Theorem 2.1. We initialize the algorithm at $(\widehat{B}^{(0)}, \widehat{\Theta}_\epsilon^{(0)}) \in \mathrm{dom}(f)$. Then for all $k \geq 1$,

$$\begin{aligned} \widehat{B}^{(k)} &= \arg\min_{B} \ f(B, \widehat{\Theta}_\epsilon^{(k-1)}), \\ \widehat{\Theta}_\epsilon^{(k)} &= \arg\min_{\Theta_\epsilon} \ f(\widehat{B}^k, \Theta_\epsilon). \end{aligned} \tag{A.1}$$

Now, consider a limit point $(B^\infty, \Theta_\epsilon^\infty)$ of the sequence $\{(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\geq 1}$. Note that such limit point exists by Bolzano-Weierstrass theorem since the sequence $\{(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\geq 1}$ is bounded. Consider a subsequence $\mathcal{K} \subseteq \{1, 2, \cdots\}$ such that $(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})_{k\in\mathcal{K}}$ converges to $(B^\infty, \Theta_\epsilon^\infty)$. Now for the bounded sequence $\{(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\in\mathcal{K}}$, without loss of generality,[1] we can say that

$$\{(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\in\mathcal{K}} \to (\widetilde{B}^\infty, \widetilde{\Theta}_\epsilon^\infty), \quad \text{for some } (\widetilde{B}^\infty, \widetilde{\Theta}_\epsilon^\infty) \in \mathrm{dom}(f).$$

By (A.1) it follows immediately that $\widetilde{\Theta}_\epsilon^\infty = \Theta_\epsilon^\infty$. Also, the following inequality holds:

$$f(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k+1)}) \leq f(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k)}) \leq f(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}).$$

Thus, by letting $k \to \infty$ over $\mathcal{K}$, we have

$$f(B^\infty, \Theta_\epsilon^\infty) \leq f(\widetilde{B}^\infty, \Theta_\epsilon^\infty) \leq f(B^\infty, \Theta_\epsilon^\infty),$$

[1] switching to some further subsequence of $\mathcal{K}$ if necessary.

since f is continuous. This implies that

$$f(\widetilde{B}^{\infty}, \Theta_{\epsilon}^{\infty}) = f(B^{\infty}, \Theta_{\epsilon}^{\infty}). \tag{A.2}$$

Next, since $f(\widehat{B}^{(k+1)}, \widehat{\Theta}_{\epsilon}^{(k)}) \leq f(B, \widehat{\Theta}_{\epsilon}^{(k)})$, for all $B \in \mathbb{R}^{p_1 \times p_2}$, let k grow along $\mathcal{K}$, and we obtain the following:

$$f(\widetilde{B}^{\infty}, \Theta_{\epsilon}^{\infty}) \leq f(B, \Theta_{\epsilon}^{\infty}), \quad \forall B \in \mathbb{R}^{p_1 \times p_2}.$$

It then follows from (A.2) that

$$f(B^{\infty}, \Theta_{\epsilon}^{\infty}) \leq f(B, \Theta_{\epsilon}^{\infty}), \quad \forall B \in \mathbb{R}^{p_1 \times p_2}. \tag{A.3}$$

Finally, note that $f(\widehat{B}^{(k)}, \widehat{\Theta}_{\epsilon}^{(k)}) \leq f(\widehat{B}^{(k)}, \Theta_{\epsilon})$, for all $\Theta \in \mathbb{S}_{++}^{p_2 \times p_2}$. As before, let k grow along $\mathcal{K}$ and with the continuity of f, we obtain:

$$f(B^{\infty}, \Theta_{\epsilon}^{\infty}) \leq f(B^{\infty}, \Theta_{\epsilon}), \quad \forall \Theta_{\epsilon} \in \mathbb{S}_{++}^{p_2 \times p_2}. \tag{A.4}$$

Now, (A.3) and (A.4) together imply that $(B^{\infty}, \Theta_{\epsilon}^{\infty})$ is a coordinate-wise minimum of f and by Fact 1, also a stationary point of f. □

Proof of Theorem 2.2. The statement of Theorem 2.2 is a variation of Proposition 4.1 in **?**], and its proof follows directly from the proof of the proposition in **?** , Appendix B]. We only outline how the statement differs. In the original statement of Proposition 4.1 in **?**], the authors provide the error bound for $\bar{\beta}$, obtained as per (2.14) whose dimension is qp^2 with q denoting the true lag of the vector-autoregressive process, under an RE condition for $\bar{\Gamma}$ and a deviation bound for $(\bar{\gamma}, \bar{\Gamma})$. For our problem, we impose a similar RE condition on $\widehat{\Gamma}$ and deviation bound on $(\widehat{\gamma}, \widehat{\Gamma})$, so as to yield a bound on $\widehat{\beta}$ that lies in a $p_1 p_2$-dimensional space. □

Proof of Theorem 2.3. The statement of this theorem is a variation of Theorem 1 in **?**], so here, instead of providing a complete proof of the theorem, we only outline how the estimation problem differs in our setting, as well as the required changes in its proof.

In **?**], the authors consider the optimization problem in (2.15), and show that for a random realization, with certain sample size requirement and choice of the regularization parameter, the following bound for $\bar{\Theta}_{\epsilon}$ holds with probability at least $1 - 1/p_2^{\tau}$ for some $\tau > 2$:

$$\|\bar{\Theta}_{\epsilon} - \Theta_{\epsilon}^{\star}\|_{\infty} \leq \{2(1 + 8\xi^{-1})\kappa_{H^{\star}}\}\bar{\delta}_f(p_2^{\tau}, n), \tag{A.5}$$

where $\bar{\delta}(r,n)$ is defined as

$$\bar{\delta}(r,n) := 8(1+4\sigma^2)\max_i(\Sigma^\star_{\epsilon,ii})\sqrt{\frac{2\log(4r)}{n}}. \tag{A.6}$$

The quantity $\bar{\delta}(p_2^\tau, n)$ that shows up in expression (A.5) is the bound for $\|S-\Sigma^\star_\epsilon\|_\infty \equiv \|\widehat{\Sigma}_\epsilon - \Sigma^\star_\epsilon\|_\infty$. In particular, in Lemma 8 [**?**], they show that with probability at least $1-1/p_2^\tau$, $\tau>2$, the following bound holds:

$$\|S-\Sigma^\star_\epsilon\|_\infty \le \bar{\delta}(p_2^\tau, n).$$

In our optimization problem (2.13), we are using $\widehat{S}$ instead of S, hence a bound for $\|\widehat{S}-\Sigma^\star_\epsilon\|_\infty$ is necessary, and the remaining argument in the proof of Theorem 1 [**?**] will follow through.

Therefore in our theorem statement, we use $g(\nu_\beta)$ as a bound for $\|\widehat{S}-\Sigma^\star_\epsilon\|_\infty$ then yield the bound for $\|\widehat{\Theta}_\epsilon - \Theta^\star_\epsilon\|_\infty$, since we are using the surrogate error $\widehat{E} = Y - X\widehat{B}$ in the estimation, instead of the true error E. □

Proof of Theorem 2.4. We first consider part (I) of the theorem. Note that by (2.5), $\widehat{\beta}^{(0)}$ can be equivalently written as

$$\widehat{\beta}^{(0)} \equiv \underset{\beta\in\mathbb{R}^{p_1\times p_2}}{\arg\min}\ \big\{-2\beta^\top\gamma^0 + \beta^\top\Gamma^0\beta + \lambda_n^0\|\beta\|_1\big\}, \tag{A.7}$$

where

$$\Gamma^{(0)} = \mathrm{I}\otimes\frac{\mathbf{X}^\top\mathbf{X}}{n}, \quad \gamma^{(0)} = (\mathrm{I}\otimes\mathbf{X}^\top)\mathrm{vec}(\mathbf{Y})/n.$$

Consider the following events:

E1. $\big\{\frac{\mathbf{X}^\top\mathbf{X}}{n} \sim \mathrm{RE}(\varphi^\star,\phi^\star)\big\}$,

E2. $\big\{\frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\big\|_\infty \le c_2\big[\Lambda_{\max}(\Sigma^\star_X)\Lambda_{\max}(\Sigma^\star_\epsilon)\big]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}\big\}$.

Note that **E1** ∩ **E2** implies the following events:

$$\Gamma^{(0)} = \mathrm{I}\otimes\frac{\mathbf{X}^\top\mathbf{X}}{n} \sim \mathrm{RE}(\varphi^\star,\phi^\star), \quad \text{where } \varphi^\star = \Lambda_{\min}(\Sigma^\star_X)/2.$$

and

$$\|\gamma^{(0)} - \Gamma^{(0)}\beta^\star\|_\infty = \frac{1}{n}\|\mathbf{X}^\top\mathbf{E}\|_\infty \le c_2\big[\Lambda_{\max}(\Sigma^\star_X)\Lambda_{\max}(\Sigma^\star_\epsilon)\big]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}. \tag{A.8}$$

Hence, by Proposition 4.1 of **?**], the bound (2.20) holds on **E1** ∩ **E2**.

By Lemmas A.1 and A.2, $\mathbb{P}(\mathbf{E1})$ is at least $1-2\exp(-c_3 n)$, for some $c_3 > 0$. By Lemma A.3, $\mathbb{P}(\mathbf{E2})$ is at least $1 - 6c_1\exp[-(c_2^2-1)\log(p_1p_2)]$ for some $c_1 > 0$, $c_2 > 1$. Hence, with probability at least

$$\mathbb{P}\big(\mathbf{E1}\cap\mathbf{E2}\big) \geq 1 - \mathbb{P}\big(\mathbf{E1}^c\big) - \mathbb{P}\big(\mathbf{E2}^c\big),$$

the bound in (2.20) holds, which proves the first part of (I). In particular, we have $\|\hat{\beta}^0 - \beta^\star\|_1 \leq \nu_\beta^{(0)} \sim O(\sqrt{\log(p_1p_2)/n})$ on $\mathbf{E1}\cap\mathbf{E2}$.

To prove the second part of (I), note that by Theorem 2.3 the bound in (2.21) holds when B1-B3 are satisfied. Now, from the argument above, B1 holds on the event $\mathbf{E1}\cap\mathbf{E2}$. Also, from the proof of Proposition 2.3, B2 is satisfied, i.e.,

$$\big\|\widehat{S}^{(0)} - \Sigma_\epsilon^\star\big\|_\infty \leq g(\nu_\beta^{(0)}), \quad \text{where } \widehat{S}^{(0)} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{B}^{(0)})^\top(\mathbf{Y} - \mathbf{X}\widehat{B}^{(0)}), \tag{A.9}$$

on $\mathbf{E1}\cap\mathbf{E2}\cap\mathbf{E3}\cap\mathbf{E4}$, where the events **E3** and **E4** are given by:

E3. $\Big\{\big\|\frac{\mathbf{E}^\top\mathbf{E}}{n} - \Sigma_\epsilon^\star\big\|_\infty \leq \sqrt{\frac{\log 4+\tau_2\log p_2}{c_\epsilon^\star n}}\Big\}$ for some $\tau_2 > 2$ and $c_\epsilon^\star > 0$ that depends on $\Sigma_\epsilon^\star$,

E4. $\Big\{\big\|\frac{\mathbf{X}^\top\mathbf{X}}{n} - \Sigma_X^\star\big\|_\infty \leq \sqrt{\frac{\log 4+\tau_1\log p_1}{c_X^\star n}}\Big\}$ for some $\tau_1 > 2$ and $c_X^\star > 0$ that depends on $\Sigma_X^\star$.

Therefore, the probability of the bound for $\widehat{\Theta}_\epsilon^{(0)}$ in (2.21) to hold is at least

$$\mathbb{P}\big(\mathbf{E1}\cap\mathbf{E2}\cap\mathbf{E3}\cap\mathbf{E4}\big), \tag{A.10}$$

By Lemma A.2, Lemma A.3 and the proof of Proposition 2.3, the probability in (A.10) is lower bounded by:

$$1 - 2\exp(-c_3 n) - 6c_1\exp[-(c_2^2-1)\log(p_1p_2)] - 1/p_1^{\tau_1-2} - 1/p_2^{\tau_2-2}.$$

Consider the following two cases where the relative order of p_1 and p_2 differ. Case 1: $p_2 \gtrsim p_1$, then $\nu_\Theta^{(0)} \sim O(\sqrt{\log p_2/n})$; case 2: $p_1 \gtrsim p_2$, then $\nu_\Theta^{(0)} \sim O\big(\log(p_1p_2)/n\big)$. In either case, since we are assuming $\log(p_1p_2)/n$ to be a small quantity and it follows that $\sqrt{\log(p_1p_2)/n} \gtrsim \log(p_1p_2)/n$, the following bound always holds:

$$\nu_\Theta^{(0)} \leq C_\Theta\sqrt{\frac{\log(p_1p_2)}{n}} \equiv M_\Theta,$$

where C_Θ is some large fixed constant that bounds the constant terms in front of $\sqrt{\log(p_1p_2)/n}$.

Now we consider part (II) of the theorem. Note that for each $k \geq 1$, $\widehat{\beta}^{(k)}$ and $\widehat{\Theta}_\epsilon^{(k)}$ are

obtained via solving the following two optimizations:

$$\widehat{\beta}^{(k)} = \underset{\beta\in\mathbb{R}^{p_1\times p_2}}{\arg\min}\ \big\{ -2\beta^\top\widehat{\gamma}^{(k-1)} + \beta^\top\widehat{\Gamma}^{(k-1)}\beta + \lambda_n\|\beta\|_1\big\}, \tag{A.11}$$

$$\widehat{\Theta}_\epsilon^{(k)} = \underset{\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}{\arg\min}\ \big\{ \log\det\Theta_\epsilon - \mathrm{tr}(\widehat{S}^{(k)}\Theta_\epsilon) + \rho_n\|\Theta_\epsilon\|_{1,\mathrm{off}}\big\}, \tag{A.12}$$

where

$$\widehat{\gamma}^{(k)} = \widehat{\Theta}^{(k)}\otimes\frac{\mathbf{X}^\top\mathbf{Y}}{n},\quad \widehat{\Gamma}^{(k)} = \widehat{\Theta}^{(k)}\otimes\frac{\mathbf{X}^\top\mathbf{X}}{n},\quad \widehat{S}^{(k)} = \frac{1}{n}(\mathbf{Y}-\mathbf{X}\widehat{B}^{(k)})^\top(\mathbf{Y}-\mathbf{X}\widehat{B}^{(k)}).$$

Consider the bound on $\hat{\beta}^{(k)}$ for $k=1$. The argument is similar to that of $\hat{\beta}^{(0)}$, with appropriate modifications to account for the fact that the objective function now involves log likelihood instead of least squares. Formally, we consider the event $\mathbf{E1}\cap\mathbf{E2}\cap\mathbf{E3}\cap\mathbf{E4}\cap\mathbf{E5}$, where

E5. $\big\{\frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\Theta_\epsilon^\star\big\|_\infty \le c_2\big[\frac{\Lambda_{\max}(\Sigma_X^\star)}{\Lambda_{\min}(\Sigma_\epsilon^\star)}\big]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}\big\}$.

Note that $\{\|\widehat{\Theta}_\epsilon^{(0)}-\Theta_\epsilon^\star\|_\infty \le \nu_\Theta^{(0)}\}$ holds on this event. By Lemma A.3, $\mathbb{P}(\mathbf{E5})\ge 1-6c_1\exp[-(c_2^2-1)\log(p_1p_2)]$. Combining this with the lower bound on (A.10) and the sample size requirement (note this sample size requirement can be relaxed to $n\gtrsim\log(p_1p_2)$ if $p_1\prec p_2$), we obtain that with probability at least

$$1-1/p_1^{\tau_1-2}-1/p_2^{\tau_2-2}-12c_1\exp[-(c_2^2-1)\log(p_1p_2)]-2\exp[-c_3n],$$

the following three events hold simultaneously:

A1' $\|\widehat{\Theta}_\epsilon^{(0)}-\Theta_\epsilon^\star\|_\infty\le\nu_\Theta^{(0)}\precsim O(\sqrt{\log(p_1p_2)/n})$;

A2' $\widehat{\Gamma}^{(0)}\sim\mathrm{RE}(\varphi^{(0)},\phi^{(0)})$ where

$$\varphi^{(0)}\ge\frac{\Lambda_{\min}(\Sigma_X^\star)}{2}(\min_i\psi^i - dM_\Theta)\quad\text{and}\quad \phi^{(0)}\le\frac{\log p_1}{n}\frac{\Lambda_{\min}(\Sigma_X^\star)}{2}(\max_j\psi^j+dM_\Theta);$$

A3' $\|\widehat{\gamma}^{(0)}-\widehat{\Gamma}^{(0)}\beta^\star\|_\infty\le\mathbb{Q}(\nu_\Theta^{(0)})\sqrt{\frac{\log(p_1p_2)}{n}}$ with the expression for $\mathbb{Q}(\cdot)$ given in (2.16).

By Theorem 2.2, by choosing $\lambda_n\ge 4\mathbb{Q}(M_\Theta)\sqrt{\frac{\log(p_1p_2)}{n}}$, the following bound holds:

$$\|\widehat{\beta}^{(1)}-\beta^\star\|_1\le 64s^{\star\star}\lambda_n/\varphi^{(0)}. \tag{A.13}$$

The error bound for $\widehat{\Theta}_\epsilon^{(1)}$ can now be established using the same argument for $\hat{\Theta}_\epsilon^{(0)}$, with the only difference that now we consider the event $\mathbf{E1} \cap \ldots \cap \mathbf{E5}$ instead of $\mathbf{E1} \cap \ldots \cap \mathbf{E4}$ and use (A.13) instead of (2.20).

Note that an upper bound for the leading term of the right hand side of (A.13) is at most of the order $O(\sqrt{\log(p_1 p_2)/n})$, and can be written as

$$C_\beta\big(s^{\star\star}\sqrt{\frac{\log(p_1 p_2)}{n}}\big) \equiv M_\beta,$$

with C_β being some potentially large number that bounds the constant term. Notice that M_β is of the same order as $\nu_\beta^{(0)}$; thus, for $\widehat{\Theta}_\epsilon^{(1)}$, we can also achieve the following bound:

$$\|\widehat{\Theta}_\epsilon^{(1)} - \Theta_\epsilon^\star\|_\infty \leq M_\Theta,$$

with high probability since we are assuming C_Θ to be some potentially large number.

Note that the events $\mathbf{E1}, \ldots, \mathbf{E5}$ rely only on the parameters and not on the estimated quantities, and on their intersection we have uniform upper bounds on the errors of $\widehat{\beta}^{(k)}$ and $\widehat{\Theta}_\epsilon^k$ for $k = 0, 1$. Hence the error bounds for $k = 1$ can be used to invoke Theorems 2 and 3 inductively on realizations X and E from the set $\mathbf{E1} \cap \ldots \cap \mathbf{E5}$ to provide high probability error bounds for all subsequent iterates as well. This leads to the uniform error bounds of part (II) with the desired probability. □

Proof of Theorem 2.5. First, we note that with a Bonferroni correction, the family-wise type I error will be automatically controlled at level α. Hence, we will focus on the power of the screening step. Also, from Theorem 7 of **?**], it is easy to see that all the arguments below hold for a large set of random realizations of X, whose probability approaches 1 under the specified asymptotic regime when the eigenvalues of Σ_X are bounded away from 0 and infinity.

Let $B^\star = \begin{bmatrix} B_1^\star & \cdots & B_{p_2}^\star \end{bmatrix}$ denote the true value of the regression coefficients and $\check{B}_j, j = 1, \cdots, p_2$ denote the estimates given by the de-biased Lasso procedure in **?**]. With the given level for sparsity, by Theorem 8 in **?**], each $\check{B}_j$ satisfies the following:

$$\sqrt{n}(\check{B}_j - B_j^\star) = Z + \Delta,$$

where $Z \sim \mathcal{N}\big(0, \sigma^2 M_j \widehat{\Sigma}_X M_j'\big)$ and Δ vanishes asymptically. Here $\widehat{\Sigma}_X$ is the sample covariance matrix of the predictors X, σ is the population noise level of the error term ϵ_j, and M_j is the matrix corresponding to the jth regression, produced by the procedure described in **?**

]^2. Let $\check{B}_{j,i}$ denote the ith coordinate of the jth regression coefficient vector $\check{B}_j$ and $\check{\Sigma}_j$ be the covariance matrix of the estimator $\check{B}_j$, then

$$\check{\Sigma}_j = \frac{\sigma^2}{n} M_j \widehat{\Sigma}_X M_j^\top,$$

and in particular, the variance of $\check{B}_{j,i}$ is $\check{\Sigma}_{j,ii} := \check{\sigma}_{ii}^j$. Using these notations, for a prespecified level α, the test statistics for testing $H_0^{ji} : B_{j,i}^\star = 0$ vs. $H_A^{ji} : B_{j,i}^\star \neq 0$, for all $i = 1, \cdots, p_1; j = 1, \cdots, p_2$ can be equivalently written as

$$\widehat{T}_{j,i} = \begin{cases} 1 & \text{if } |\check{B}_{j,i}|/\check{\sigma}_{ii}^j > z_{\alpha/(2p_1p_2)}, \\ 0 & \text{otherwise.} \end{cases}$$

where z_α denotes the upper α quantiles of $\mathcal{N}(0,1)$.

Define the "family-wise" power as follows:

$$\begin{aligned} \mathbb{P}\big(\text{all true alternatives are detected}\big) &= \mathbb{P}\big(\bigcap_{1\le j\le p_2} \bigcap_{k\in S_j^\star} \{\widehat{T}_{j,k} = 1\}\big) \\ &= 1 - \mathbb{P}\big(\bigcup_{1\le j\le p_2} \bigcup_{k\in S_j^\star} \{\widehat{T}_{j,k} = 0\}\big). \end{aligned}$$

Correspondingly, the family-wise type II error can be written as

$$\mathbb{P}\big(\bigcup_{1\le j\le p_2} \bigcup_{k\in S_j^\star} \{\widehat{T}_{j,k} = 0\}\big) \le \sum_{j=1}^{p_2} \sum_{k\in S_j^\star} \mathbb{P}\big(\widehat{T}_{j,k} = 0\big). \tag{A.14}$$

By Theorem 16 in **?**], asymptotically, $\forall k \in S_j, j = 1, \cdots, p_2$,

$$\mathbb{P}\big(\widehat{T}_{j,k} = 0\big) \le 1 - G\big(\frac{\alpha}{p_1p_2}, \frac{\sqrt{n}\gamma}{\sigma[\Sigma_{k,k}^{-1}]^{1/2}}\big); \qquad 0 < \gamma \le \min|B_{j,k}^\star|, \quad \forall k \in S_j, j = 1, \cdots, p_2. \tag{A.15}$$

Here

$$G(\alpha, u) \equiv 2 - \mathbb{P}(\Phi < z_{\alpha/2} + u) - \mathbb{P}(\Phi < z_{\alpha/2} - u),$$

where we use Φ to denote the random variable following a standard Gaussian distribution

[2]Details of the procedure is described in p.2871 in **?**], with M being an intermediate quantity obtained by solving an optimization problem.

and the choice of n in (A.15) doesn't depend on k. Hence, (A.15) can be rewritten as

$$\begin{aligned}\mathbb{P}\big(\widehat{T}_{j,k}=0\big) &\le 1-G\big(\frac{\alpha}{p_1p_2},\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\big)\\ &=\mathbb{P}\big(\Phi<z_{\alpha/(2p_1p_2)}-\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\big)-\mathbb{P}\big(\Phi>z_{\alpha/(2p_1p_2)}+\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\big) \qquad (A.16)\\ &\le \mathbb{P}\big(\Phi>\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}-z_{\alpha/(2p_1p_2)}\big),\end{aligned}$$

where we use Φ to denote the random variable following a standard Gaussian distribution.

Note that the following inequality holds for standard Normal percentiles:

$$2e^{-t^2}\le\mathbb{P}(|\Phi|>t)\le e^{-t^2/2},$$

and by taking the inverse function, the following inequality holds:

$$\sqrt{-\log\frac{y}{2}}\le z_{y/2}\le\sqrt{-2\log y}.$$

Letting $y=\frac{\alpha}{p_1p_2}$, it follows that

$$\big(-\log\frac{\alpha}{2p_1p_2}\big)^{1/2}\le z_{\alpha/(2p_1p_2)}\le\big(-2\log\frac{\alpha}{p_1p_2}\big)^{1/2},$$

hence

$$\mathbb{P}\big(\Phi>\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}-z_{\alpha/(2p_1p_2)}\big)\le\mathbb{P}\big(\Phi>\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}-\sqrt{-2\log\frac{\alpha}{p_1p_2}}\big).$$

Now given

$$\frac{\log(p_1p_2)}{n}\to 0,$$

it follows that

$$\frac{\sqrt{2\log\big(\frac{p_1p_2}{\alpha}\big)}}{\sqrt{n}/\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\to 0,$$

indicating that for sufficiently large n, the following lower bound holds for some constant $c_0>0$:

$$\big(\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}-\sqrt{-2\log\frac{\alpha}{p_1p_2}}\big)\ge c_0\sqrt{n}.$$

Note that c_0 is universal for all choices of k, since this lower bound can be achieved by substituting $\Sigma^{-1}_{k,k}$ by $(1/\Lambda_{\min}(\Sigma_X))$, which is assumed to be bounded away from infinity. Combined with the fact that $\mathbb{P}(\Phi>t)\le e^{-t^2/2}$, the last expression in (A.16) can thus be

bounded by

$$\mathbb{P}\big(\Phi > \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - z_{\alpha/(2p_1p_2)}\big) \le \exp\big[-\frac{1}{2}\big(\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - \sqrt{-2\log\frac{\alpha}{p_1p_2}}\big)^2\big] \le e^{-c_1 n}, \quad \text{(A.17)}$$

for some universal constant $c_1 > 0$, and the bound in (A.17) holds uniformly for all $k \in S_j, \forall j$. Combine (A.14), (A.15) and (A.17), it follows that

$$\mathbb{P}\big(\bigcup_{1\le j\le p_2}\bigcup_{k\in S_j^\star}\{\widehat{T}_{j,k}=0\}\big) \le s^\star p_2 \exp(-c_1 n). \quad \text{(A.18)}$$

Now with $\log(p_1p_2)/n = o(1)$ and the given sparsity level, that is, $s^\star = o(\sqrt{n}/\log p_1)$, it follows that

$$s^\star p_2 \exp(-c_1 n) = o(1),$$

and by (A.18), we have:

$$\mathbb{P}\big(\text{family-wise type II error}\big) \to 0, \quad \Leftrightarrow \quad \mathbb{P}\big(\text{family-wise power}\big) \to 1.$$

This is equivalent to establishing that, given $\log(p_1p_2)/n \to 0$, the screening step recovers the true support sets $S_j^\star$ for all $j = 1, 2, \cdots, p_2$ with high probability, while keeping the family-wise type I error rate under control. □

A.2 Proofs for propositions and auxiliary lemmas.

In this subsection, we provide proofs for the propositions presented in Section 2.3, which requires several auxiliary lemmas, whose proofs are presented along the context.

To prove Proposition 2.1, we need the following two lemmas. Lemma A.1 was originally provided as Lemma B.1 in **?**], which states that if the sample covariance matrix of X satisfies the RE condition and Θ is diagonally dominant, then $(\mathbf{X}^\top\mathbf{X}/n) \otimes \Theta$ also satisfies the RE condition. Here we omit its proof and only state the main result. Lemma A.2 verifies that with high probability, the sample covariance matrix of the design matrix X satisfies the RE condition.

Lemma A.1. *If $\mathbf{X}^\top\mathbf{X}/n \sim RE(\varphi^\star, \phi^\star)$, and Θ is diagonally dominant, that is, $\psi^i := \sigma^{ii} - \sum_{j\neq i}\sigma^{ij} > 0$ for all $i = 1, 2, \cdots, p_2$, where σ^{ij} is the ijth entry in Θ, then*

$$\Theta \otimes \mathbf{X}^\top\mathbf{X}/n \sim RE\big(\varphi^\star \min_i \psi^i, \phi^\star \max_i \psi^i\big).$$

Lemma A.2. *With probability at least* $1-2\exp(-c_3 n)$*, for a zero-mean sub-Gaussian random design matrix* $\mathbf{X}\in\mathbb{R}^{n\times p_1}$*, its sample covariance matrix* $\widehat{\Sigma}_X$ *satisfies the RE condition with parameter* $\varphi^\star$ *and* $\phi^\star$*, i.e.,*

$$\widehat{\Sigma}_X \sim RE(\varphi^\star, \phi^\star), \tag{A.19}$$

where $\widehat{\Sigma}_X = \mathbf{X}^\top\mathbf{X}/n$*,* $\varphi^\star = \Lambda_{\min}(\Sigma_X^\star)/2$*,* $\phi^\star = \varphi^\star \log p_1/n$*.*

Proof. To prove this lemma, we first use Lemma 15 in **?**], which states that if $\mathbf{X}\in\mathbb{R}^{n\times p}$ is zero-mean sub-Gaussian with parameter (Σ,σ^2), then there exists a universal constant $c>0$ such that

$$\mathbb{P}\big(\sup_{v\in\mathbb{K}(2s)}\big|\frac{\|\mathbf{X}v\|_2^2}{n}-\mathbb{E}\big[\frac{\|\mathbf{X}v\|_2^2}{n}\big]\big|\geq t\big)\leq 2\exp\big(-cn\min(\frac{t^2}{\sigma^4},\frac{t}{\sigma^2})+2s\log p\big), \tag{A.20}$$

where $\mathbb{K}(2s)$ is a set of $2s$ sparse vectors, defined as

$$\mathbb{K}(2s) := \{v\in\mathbb{R}^p : \|v\|\leq 1, \|v\|_0\leq 2s\}.$$

By taking $t=\frac{\Lambda_{\min}(\Sigma_X^\star)}{54}$, with probability at least $1-2\exp\big(-c'n+2s\log p_1\big)$ for some $c'>0$, the following bound holds:

$$|v^\top(\widehat{\Sigma}_X-\Sigma_X^\star)v|\leq\frac{\Lambda_{\min}(\Sigma_X^\star)}{54},\quad \forall v\in\mathbb{K}(2s). \tag{A.21}$$

Then applying supplementary Lemma 13 in **?**], for an estimator $\widehat{\Sigma}_X$ of $\Sigma_X^\star$ satisfying the deviation condition in (A.21), the following RE condition holds:

$$v^\top S_x v\geq\frac{\Lambda_{\min}(\Sigma_X^\star)}{2}\|v\|_2^2-\frac{\Lambda_{\min}(\Sigma_X^\star)}{2s}\|v\|_1^2.$$

Finally, set $s=c''n/4\log p_1$, then with probability at least $1-2\exp(-c_3 n)$ $(c_3>0)$, $\widehat{\Sigma}_X\sim$ RE$(\varphi^\star,\phi^\star)$ with $\varphi^\star=\Lambda_{\min}(\Sigma_X^\star)/2$, $\phi^\star=\varphi^\star\log p_1/n$. □

With the above two lemmas, we are ready to prove Proposition 2.1.

Proof of Proposition 2.1. We first show that if $\Theta_\epsilon^\star$ is diagonally dominant, then $\widehat{\Theta}_\epsilon$ is also diagonally dominant provided that the error of $\widehat{\Theta}_\epsilon$ is of the given order and n is sufficiently large. Define

$$\widehat{\psi}^i=\widehat{\sigma}_\epsilon^{ii}-\sum_{j\neq i}\widehat{\sigma}_\epsilon^{ij},$$

where $\widehat{\sigma}_\epsilon^{ij}$ is the ijth entry of $\widehat{\Theta}_\epsilon$, then $\widehat{\psi}^i$ is the gap between the diagonal entry and the

off-diagonal entries of row i in matrix $\widehat{\Theta}_\epsilon$. We can decompose $\widehat{\psi^i}$ into the following:

$$\widehat{\psi^i} = \big[\sigma_\epsilon^{ii} - \sum_{j\neq i}\sigma_\epsilon^{ij}\big] + \big[(\widehat{\sigma}_\epsilon^{ii} - \sigma_\epsilon^{ii}) + \sum_{j\neq i}(\sigma_\epsilon^{ij} - \widehat{\sigma}_\epsilon^{ij})\big].$$

Recall that we define ψ_i as $\psi^i = \sigma_\epsilon^{ii} - \sum_{j\neq i}^{p_2}\sigma_\epsilon^{ij}$. Hence

$$\begin{aligned}
\min\widehat{\psi^i} &\geq \min_i \psi^i - |||\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star|||_\infty \geq \min_i(\sigma_\epsilon^{ii} - \sum_{j\neq i}\sigma_\epsilon^{ij}) - d\nu_\Theta = \min\psi^i - d\nu_\Theta,\\
\max\widehat{\psi^i} &\leq \max_i \psi^i + |||\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star|||_\infty \leq \max_i(\sigma_\epsilon^{ii} - \sum_{j\neq i}\sigma_\epsilon^{ij}) + d\nu_\Theta = \max\psi^i + d\nu_\Theta.
\end{aligned} \tag{A.22}$$

Now given $\nu_\Theta = \eta_\Theta\frac{\log p_2}{n} = O(\sqrt{\log p_2/n})$, with $n \gtrsim d^2\log p_2$, $d\nu_\Theta = o(1)$, and it follows that

$$\min_i \psi^i - d\nu_\Theta \geq 0.$$

Now by Lemma A.2, $\mathbf{X}^\top\mathbf{X}/n \sim RE(\varphi^\star, \phi^\star)$ with high probability. Combine with Lemma A.1 and inequality (A.22), with probability at least $1 - 2\exp(-c_3 n)$ for some $c_3 > 0$, $\widehat{\Gamma}$ satisfies the following RE condition:

$$\widehat{\Gamma} = \widehat{\Theta}_\epsilon \otimes (\mathbf{X}^\top\mathbf{X}/n) \sim \mathrm{RE}\big(\varphi^\star(\min_i \psi^i - d\nu_\Theta), \phi^\star \max_i(\psi^i + d\nu_\Theta)\big), \tag{A.23}$$

where $\varphi^\star = \Lambda_{\min}(\Sigma_X^\star)/2$, $\phi^\star = \varphi^\star \log p_1/n$. □

To prove Proposition 2.2, we first prove Lemma A.3.

Lemma A.3. *Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a zero-mean sub-Gaussian matrix with parameter (Σ_X, σ_X^2) and $\mathbf{E} \in \mathbb{R}^{n\times p_2}$ be a zero-mean sub-Gaussian matrix with parameters $(\Sigma_\epsilon, \sigma_\epsilon^2)$. Moreover, $\mathbf{X}$ and $\mathbf{E}$ are independent. Let $\Theta_\epsilon := \Sigma_\epsilon^{-1}$, then if $n \gtrsim \log(p_1p_2)$, the following two expressions hold with probability at least $1 - 6c_1\exp[-(c_2^2-1)\log(p_1p_2)]$ for some $c_1 > 0, c_2 > 1$, respectively:*

$$\frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\big\|_\infty \leq c_2\big[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon)\big]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}, \tag{A.24}$$

and

$$\frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\Theta_\epsilon\big\|_\infty \leq c_2\big[\frac{\Lambda_{\max}(\Sigma_X)}{\Lambda_{\min}(\Sigma_\epsilon)}\big]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}. \tag{A.25}$$

Proof. The proof of this lemma uses Lemma 14 in **?**], in which they show that if $\mathbf{X} \in \mathbb{R}^{n\times p_1}$ is a zero-mean sub-Gaussian matrix with parameters (Σ_x, σ_x^2) and $\mathbf{Y} \in \mathbb{R}^{n\times p_2}$ is a zero-mean

sub-Gaussian matrix with parameters (Σ_y, σ_y^2), then if $n \gtrsim \log(p_1 p_2)$,

$$\mathbb{P}\big(\big\|\frac{\mathbf{Y}^\top \mathbf{X}}{n} - \mathrm{Cov}(Y_i, X_i)\big\|_\infty \geq t\big) \leq 6p_1p_2 \exp\big(-cn \min\big\{\frac{t^2}{(\sigma_x\sigma_y)^2}, \frac{t}{\sigma_x\sigma_y}\big\}\big),$$

where X_i and Y_i are the ith row of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Here, we replace $\mathbf{Y}$ by $\mathbf{E}$, and since $\mathbf{E}$ and $\mathbf{X}$ are independent, $\mathrm{Cov}(X_i, E_i) = 0$. Let $t = c_2\sigma_X\sigma_\epsilon\sqrt{\log(p_1p_2)/n}$, $c_2 > 1$ we get

$$\mathbb{P}\big(\big\|\frac{\mathbf{X}^\top \mathbf{E}}{n}\big\|_\infty \geq c_2\sigma_X\sigma_\epsilon\sqrt{\frac{\log(p_1p_2)}{n}}\big) \leq 6c_1(p_1p_2)^{1-c_2^2} = 6c_1 \exp\big[-(c_2^2-2)\log(p_1p_2)\big].$$

Note that the sub-Gaussian parameter satisfies $\sigma_X^2 \leq \max_i(\Sigma_{X,ii}) \leq \Lambda_{\max}(\Sigma_X)$. This directly gives the bound in (A.24).

To obtain the bound in (A.25), we note that if $\mathbf{E}$ is sub-Gaussian with parameters $(\Sigma_\epsilon, \sigma_\epsilon^2)$, then $\mathbf{E}\Theta$ is sub-Gaussian with parameter $(\Theta, \theta_\epsilon^2)$, where

$$\theta_\epsilon^2 \leq \max_i(\Theta_{\epsilon,ii}) \leq \Lambda_{\max}(\Theta_\epsilon) = \frac{1}{\Lambda_{\min}(\Sigma_\epsilon)}.$$

Then we replace $\mathbf{Y}$ by $\mathbf{E}\Theta_\epsilon$ and yield the bound in (A.25). □

As a remark, here we note that the event in (A.24) and (A.25) may not be independent. However, the two events hold simultaneously with probability at least $1-2c_2\exp[-c_2\log(p_1p_2)]$, with this crude bound for probability hold for sure.

Now we are ready to prove Proposition 2.2.

Proof of Proposition 2.2. First we note that

$$\mathbf{X}^\top\mathbf{E}\widehat{\Theta}_\epsilon = \mathbf{X}^\top\mathbf{E}\Theta_\epsilon + \mathbf{X}^\top\mathbf{E}(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star),$$

which directly gives the following inequality:

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^\star\|_\infty = \frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\widehat{\Theta}_\epsilon\big\|_\infty \leq \frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\Theta_\epsilon^\star\big\|_\infty + \frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star)\big\|_\infty. \tag{A.26}$$

Now we would like to bound the two terms separately.

The first term can be bounded by (A.25) in Lemma A.3, that is,

$$\frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\Theta_\epsilon^\star\big\|_\infty \leq c_2\big[\frac{\Lambda_{\max}(\Sigma_X)}{\Lambda_{\min}(\Sigma_\epsilon^\star)}\big]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}.$$

w.p. at least $1 - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)]$.

For the second term, first we note that

$$\begin{aligned} \frac{1}{n}\big\|\mathbf{X}^\top \mathbf{E}(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star)\big\|_\infty &= \frac{1}{n} \max_{\substack{1\le i\le p_1 \\ 1\le j\le p_2}} \big|e_i' \mathbf{X}^\top \mathbf{E}(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star) e_j\big| \\ &\le \frac{1}{n} \max_i \big\|e_i' \mathbf{X}^\top \mathbf{E}\big\|_\infty \max_j \big\|(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star) e_j\big\|_1, \end{aligned} \tag{A.27}$$

where we have $e_i \in \mathbb{R}^{p_1}$ and $e_j \in \mathbb{R}^{p_2}$, and the inequality comes from the fact that $|a'b| \le \|a\|_\infty \|b\|_1$. Note that

$$\max_i \big\|e_i' \mathbf{X}^\top \mathbf{E}\big\|_\infty = \|\mathbf{X}^\top \mathbf{E}\|_\infty,$$

since $\|e_i' \mathbf{X}^\top \mathbf{E}\|_\infty$ gives the largest element (in absolute value) of the ith row of $\mathbf{X}^\top \mathbf{E}$, and taking the maximum over all i's gives the largest element of $\mathbf{X}^\top \mathbf{E}$ over all entries. And for $\max_j \big\|(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star) e_j\big\|_1$, it holds that

$$\max_j \big\|(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star) e_j\big\|_1 = |||\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star|||_1 = |||\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star|||_\infty,$$

where $|||A|||_1 := \max_{\|x\|_1 = 1} \|Ax\|_1$ is the ℓ_1-operator norm, and the last equality follows from the fact that $|||A|||_1 = |||A'|||_\infty$. As a result, (A.27) can be re-written as:

$$\frac{1}{n}\big\|\mathbf{X}^\top \mathbf{E}(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star)\big\|_\infty \le \big(\frac{1}{n}\|\mathbf{X}^\top \mathbf{E}\|_\infty\big)\big(|||\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star|||_\infty\big). \tag{A.28}$$

Now, using (A.24), w.p. at least $1 - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)]$, we have

$$\frac{1}{n}\big\|\mathbf{X}^\top \mathbf{E}\big\|_\infty \le c_2 \big[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon^\star)\big]^{1/2} \sqrt{\frac{\log(p_1 p_2)}{n}},$$

and since $\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star\|_\infty \le \nu_\Theta$, it directly follows that $|||\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star|||_\infty \le d\nu_\Theta$. Therefore, with probability at least $1 - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)]$,

$$\frac{1}{n}\big\|\mathbf{X}^\top \mathbf{E}(\widehat{\Theta}_\epsilon - \Theta_\epsilon^\star)\big\|_\infty \le c_2 d\nu_\Theta \big[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon^\star)\big]^{1/2} \sqrt{\frac{\log(p_1 p_2)}{n}}. \tag{A.29}$$

Combine the two terms, we obtain the conclusion in Proposition 2.2. □

Proof of Proposition 2.3. First we note the following decomposition:

$$\|\widehat{S} - \Sigma_\epsilon^\star\|_\infty \le \|S - \Sigma_\epsilon\|_\infty + \|\widehat{S} - S\|_\infty := \|W_1\|_\infty + \|W_2\|_\infty,$$

where S is the sample covariance matrix of the true errors E.

For W_1, by Lemma 8 in **?**], for sample size

$$n \geq 512(1+4\sigma_\epsilon^2)^4 \max_i(\Sigma^\star_{\epsilon,ii})^4 \log(4p_2^{\tau_2}),$$

the following bound holds w.p. at least $1 - 1/p_2^{\tau_2-2}(\tau_2 > 2)$,

$$\|W_1\|_\infty \leq \sqrt{\frac{\log 4 + \tau_2 \log p_2}{c_\epsilon^\star n}}, \quad \text{where } c_\epsilon^\star = \big[128(1+4\sigma_\epsilon^2)^2 \max_i(\Sigma^\star_{\epsilon,ii})^2\big]^{-1}. \tag{A.30}$$

For W_2, rewrite it as:

$$W_2 = \frac{2}{n}\mathbf{E}^\top \mathbf{X}(B^\star - \widehat{B}) + (B^\star - \widehat{B})^\top \big(\frac{\mathbf{X}^\top \mathbf{X}}{n}\big)(B^\star - \widehat{B}). \tag{A.31}$$

The first term in (A.31) can be bounded as:

$$\big\|\frac{2}{n}\mathbf{E}^\top \mathbf{X}(B^\star - \widehat{B})\big\|_\infty \leq 2|||B^\star - \widehat{B}|||_1 \big\|\frac{1}{n}\mathbf{X}^\top \mathbf{E}\big\|_\infty \leq 2\|\beta^\star - \widehat{\beta}\|_1 \cdot \big\|\frac{1}{n}\mathbf{X}^\top \mathbf{E}\big\|_\infty. \tag{A.32}$$

By Lemma A.3, with probability at least $1 - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)]$, the following bound holds:

$$\big\|\frac{2}{n}\mathbf{E}^\top \mathbf{X}(B^\star - \widehat{B})\big\|_\infty \leq 2c_2 \nu_\beta \big[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon^\star)\big]^{1/2} \sqrt{\frac{\log(p_1 p_2)}{n}}, \tag{A.33}$$

with the sample size requirement being $n \gtrsim \log(p_1 p_2)$.

For the second term in (A.31), we consider the following bound:

$$\begin{aligned} \|(B^\star - \widehat{B})^\top \big(\frac{\mathbf{X}^\top \mathbf{X}}{n}\big)(B^\star - \widehat{B})\|_\infty &\leq |||B^\star - \widehat{B}|||_1 \big\|\big(\frac{\mathbf{X}^\top \mathbf{X}}{n}\big)(B^\star - \widehat{B})\big\|_\infty \\ &\leq |||B^\star - \widehat{B}|||_1^2 \big\|\big(\frac{\mathbf{X}^\top \mathbf{X}}{n}\big)\big\|_\infty. \end{aligned} \tag{A.34}$$

Here, we apply Lemma 8 in **?**] to the design matrix X, for sample size

$$n \geq 512(1+4\sigma_x^2)^4 \max_i(\Sigma_{X,ii})^4 \log(4p_1^{\tau_1}),$$

the following bound holds w.p. at least $1 - 1/p_1^{\tau_1-2}(\tau_1 > 2)$,

$$\big\|\big(\frac{\mathbf{X}^\top \mathbf{X}}{n}\big) - \Sigma_X\big\|_\infty \leq \sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^* n}}, \quad \text{where } \; c_X^\star = \big[128(1+4\sigma_x^2)^2 \max_i(\Sigma_{X,ii})^2\big]^{-1}. \tag{A.35}$$

This indicates that with this choice of n, the following bound holds with probability at least

$1 - 1/p_1^{\tau_1 - 2} (\tau_1 > 2)$,

$$\big\| (\frac{\mathbf{X}^\top \mathbf{X}}{n}) \big\|_\infty \leq \sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^\star n}} + \max_i (\Sigma_{X,ii}).$$

Combine with the bound in (A.34), with probability at least $1 - 1/p_1^{\tau_1 - 2} (\tau_1 > 2)$, the following bound holds:

$$\|(B^\star - \widehat{B})^\top \big(\frac{\mathbf{X}^\top \mathbf{X}}{n}\big)(B^\star - \widehat{B})\|_\infty \leq \nu_\beta^2 \big(\sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^\star n}} + \max_i (\Sigma_{X,ii}) \big). \tag{A.36}$$

Now combine (A.32), (A.33) and (A.36), we reach the conclusion of Proposition 3, with the leading term in the sample size requirement being $n \gtrsim \log(p_1 p_2)$. □

Proof for Proposition 2.4. From the structural equations of a multi-layered graph introduced in Section 2.2.1, and setting $\vec{\epsilon}^1 := \vec{X}^1$, we can write

$$\begin{bmatrix} \vec{\epsilon}^1 \\ \vec{\epsilon}^2 \end{bmatrix} = \begin{bmatrix} \mathrm{I} & 0 \\ -(B^{12})^\top & \mathrm{I} \end{bmatrix} \begin{bmatrix} \vec{X}^1 \\ \vec{X}^2 \end{bmatrix}. \tag{A.37}$$

Define $P = [\mathrm{I}, O; -(B^{12})^\top, \mathrm{I}]$. Then, $P\widetilde{X}$ is a centered Gaussian random vector with a block diagonal variance-covariance matrix $\mathrm{diag}(\Sigma^1, \Sigma^2)$. Hence, the concentration matrix of $\widetilde{X}$ takes the form

$$\Theta_{\tilde{X}} = \Sigma_{\tilde{X}}^{-1} = \begin{bmatrix} \mathrm{I} & -B^{12} \\ O & \mathrm{I} \end{bmatrix} \begin{bmatrix} \Theta^1 & O \\ O & \Theta^2 \end{bmatrix} \begin{bmatrix} \mathrm{I} & O \\ -(B^{12})^\top & \mathrm{I} \end{bmatrix}.$$

This leads to an upper bound

$$|||\Theta_{\tilde{X}}|||_{\mathrm{op}} \leq |||\Theta^1|||_{\mathrm{op}} |||\Theta^2|||_{\mathrm{op}} |||P|||_{\mathrm{op}}^2.$$

The result then follows by using the matrix norm inequality $|||A|||_{\mathrm{op}} \leq \sqrt{|||A|||_1 |||A|||_\infty}$ [**?**], where $|||A|||_1$ and $|||A|||_\infty$ denote the maximum absolute row and column sums of A, and the fact that $\Lambda_{\min}(\Sigma_{\tilde{X}}) = |||\Theta_{\tilde{X}}|||_{\mathrm{op}}^{-1}$. □

A.3 Numerical comparisons between different parametrizations.

In this subsection, we provide some numerical evidence to substantiate the point we made in Section 2.5, that the two parametrizations are not always equivalent. This is a point also mentioned in the original work on AMP graphs by **?**], the framework adopted in this paper. The other parametrization which we referred to as the (Ω_{XY}, Ω_Y)-*parametrization*

corresponds to the LWF framework [see **?** , p.34-35]. In the presence of sparsity penalization, a specific sparsity pattern for the (B, Θ_ϵ)-*parameterization* may not be recoverable through the (Ω_{XY}, Ω_Y)-*parametrization* and vice versa.

Consider the following two simulation settings, in which the data are generated from the AMP framework ((B, Θ_ϵ)-parameterization) and the LWF framework ((Ω_{XY}, Ω_Y)-*parametrization*) respectively.

- AMP framework. The data are generated according to the model $\mathbf{Y} = \mathbf{X}B^\star + \mathbf{E}$, similar to Model A described in Section 2.4; that is, each entry in $B^\star$ is nonzero with probability $5/p_1$, and off-diagonal entries for $\Theta_\epsilon^\star$ are nonzero with probability $5/p_2$. Nonzero entries of $B^\star$ and $\Theta_\epsilon^\star$ are generated from $\mathsf{Unif}\big[(-1,-0.5)\cup(0.5,1)\big]$, and diagonals of $\Theta_\epsilon^\star$ are set identical, such that the condition number of $\Theta_\epsilon^\star$ is p_2. Table A.1 shows the performance of estimated B using different methods that are designed for different parameterizations: the node-conditional method (mixed MRF) and the proposed method in this study (PML).

(p_1, p_2, n)	Method	SEN	SPC	MCC
$(30, 60, 100)$	mixed MRF (th)	0.86	0.71	0.45
	PML	0.96	0.99	0.93
$(60, 30, 100)$	mixed MRF (th)	0.96	0.76	0.70
	PML	0.99	0.99	0.93
$(200, 200, 150)$	mixed MRF (th)	0.80	0.99	0.70
	PML	0.99	0.99	0.88

Table A.1: Performance for $\widehat{B}$ using different methods for different parameterizations.

- LWF framework. The data are generated based on the multivariate Gaussian specification:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\Big(0, \begin{pmatrix} \Omega_X & \Omega_{XY} \\ \Omega_{YX} & \Omega_Y \end{pmatrix}^{-1}\Big).$$

  Specifically, Ω_X is banded with 1 on the diagonal and 0.2 on the upper and lower first diagonal, Ω_Y is also banded with 1 on the diagonal and 0.3 on the upper and lower first diagonal. Each entry in $\Omega_X Y$ is nonzero with probability $5/p_1$, and the nonzero entries are generated from $\mathsf{Unif}\big[(-1,-0.8)\cup(0.8,1)\big]$. Further, we bump up the diagonal of the joint precision matrix $\left[\begin{smallmatrix} \Omega_X & \Omega_{XY} \\ \Omega'_{XY} & \Omega_Y \end{smallmatrix}\right]$ such that it is positive definite. Table A.2 depicts the selection property of the estimated Ω_{XY} using different methods that are designed for different parameterizations.

Note that to retrieve stable and meaningful results, for the AMP framework, the estimates using mixed MRF are thresholded at a proper level, and for the LWF framework, the estimates using PML are also thresholded.

(p_1, p_2, n)	Method	SEN	SPC	MCC
$(30, 60, 100)$	mixed MRF	0.84	0.88	0.63
	PML-th	0.99	0.52	0.39
$(60, 30, 100)$	mixed MRF	0.847	0.95	0.70
	PML-th	1	0.80	0.52
$(200, 200, 150)$	mixed MRF	0.89	0.93	0.70
	PML-th	1	0.79	0.30

Table A.2: Performance for $\widehat{\Omega}_{XY}$ using different methods for different parameterizations

It can be seen that the method compatible with the data generation mechanism exhibits superior performance, vis-a-vis its competitor that was designed for another parameterization. Further, the mixed MRF method suffers in terms of both sensititvity and specificity under the AMP parameterization, while the PML method suffers in terms of specificity only under the LWF parameterization.

A.4 An example for multi-layered network estimation.

As mentioned at the beginning of Section 2.2, the proposed methodology is designed for obtaining MLEs for multi-layer Gaussian networks, but the problem breaks down into a sequence of 2-layered estimation problems. Here we give an detailed example to illustrate how our proposed methodology proceeds for a 3-layered network.

Suppose there are p_1, p_2 and p_3 nodes in Layers 1, 2 and 3, respectively. This three-layered network is modeled as follows:

- $X \sim \mathcal{N}(0, \Sigma_X)$, $X \in \mathbb{R}^{p_1}$.
- For $j = 1, \cdots, p_2$: $Y_j = X^\top B_j^{xy} + \epsilon_j^Y$, $B_j^{xy} \in \mathbb{R}^{p_1}$. $(\epsilon_1^Y \cdots \epsilon_{p_2}^Y)' \sim \mathcal{N}(0, \Sigma_{\epsilon,Y})$.
- For $l = 1, 2, \cdots, p_3$: $Z_l = X^\top B_l^{xz} + Y^\top B_l^{yz} + \epsilon_l^Z$, $B_l^{xz} \in \mathbb{R}^{p_1}$ and $B_l^{yz} \in \mathbb{R}^{p_2}$. $(\epsilon_1^Z \cdots \epsilon_{p_3}^Z)' \sim \mathcal{N}(0, \Sigma_{\epsilon,Z})$.

The parameters of interest are : Θ_X, $\Theta_{\epsilon,Y} := \Sigma_{\epsilon,Y}^{-1}$, $\Theta_{\epsilon,Z} := \Sigma_{\epsilon,Z}^{-1}$, which denote the within-layer conditional dependencies, and

$$B_{XY} = \left[B_1^{xy} \cdots B_{p_2}^{xy} \right], \quad B_{XZ} = \left[B_1^{xz} \cdots B_{p_3}^{xz} \right] \text{ and } B_{YZ} = \left[B_1^{yz} \cdots B_{p_3}^{yz} \right],$$

which encode the across-layer dependencies.

Now given data $\mathbf{X} \in \mathbb{R}^{n \times p_1}$, $\mathbf{Y} \in \mathbb{R}^{n \times p_2}$ and $\mathbf{Z} \in \mathbb{R}^{n \times p_3}$, all centered, the full log-likelihood can be written as:

$$\ell(\mathbf{Z}, \mathbf{Y}, \mathbf{X}) = \ell(\mathbf{Z}|\mathbf{Y}, \mathbf{X}; \Theta_{\epsilon,Z}, B_{YZ}, B_{XZ}) + \ell(\mathbf{Y}|\mathbf{X}; \Theta_{\epsilon,Y}, B_{XY}) + \ell(\mathbf{X}; \Theta_X). \tag{A.38}$$

The separability of the log-likelihood enables us to ignore the inner structure of the combined layer $\widetilde{\mathbf{X}} := [\mathbf{X}, \mathbf{Y}]$ when trying to estimate the dependencies between Layer 1 and Layer 3, Layer 2 and Layer 3, as well as the conditional dependencies within Layer 3. As a consequence, the optimization problem minimizing the negative log-likelihood can be decomposed into three separate problems, i.e., solving for $\{\Theta_{\epsilon,Z}, B_{XZ}, B_{YZ}\}$, $\{\Theta_{\epsilon,Y}, B_{XY}\}$ and $\{\Theta_X\}$, respectively.

The estimation procedure described in Section 2.2.2 can thus be carried out in a recursive way in a sense of what follows. To obtain estimates for $\{B_{XZ}, B_{YZ}, \Theta_{\epsilon,Z}\}$, based on the formulation in (2.2), we solve the following optimization problem:

$$\min_{\substack{\Theta_{\epsilon,Z}\in\mathbb{S}_{++}^{p_3\times p_3}\\ B_{XZ},B_{YZ}}} \Big\{ -\log\det\Theta_{\epsilon,Z} + \frac{1}{n}\sum_{j=1}^{p_3}\sum_{i=1}^{p_3}\sigma_Z^{ij}(\mathbf{Z}_i - \mathbf{X}B_i^{xz} - \mathbf{Y}B_i^{yz})^\top(\mathbf{Z}_j - \mathbf{X}B_j^{xz} - \mathbf{Y}B_j^{yz}) \\ +\lambda_n(\|B_{XZ}\|_1 + \|B_{YZ}\|_1) + \rho_n\|\Theta_{\epsilon,Z}\|_{1,\text{off}}\Big\},$$

which can be solved by treating the combined design matrix $\widetilde{\mathbf{X}} := [\mathbf{X}, \mathbf{Y}]$ as a single super layer and $\mathbf{Z}$ as the response layer, then apply each step described in Section 2.2.2. To obtain estimates for B_{XY} and $\Theta_{\epsilon,Y}$, we can ignore the 3rd layer for now and apply the exact procedure all over again, by treating $\mathbf{Y}$ as the response layer and $\mathbf{X}$ as the design layer. The estimate for the precision matrix of the bottom layer Θ_X can be obtained by graphical lasso [**?**] or the nodewise regression [**?**].

APPENDIX B

Supplementary Materials to "Regularized Estimation and Testing of High-dimensional Multi-block Vector Autoregressive Models."

B.1 Additional Theorems and Proofs for Theorems.

In this section, we introduce two additional theorems that respectively establish the consistency properties for the initializers $\widehat{A}^{(0)}$ and $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$, for *fixed* realizations of the processes $\{X_t\}$ and $\{Z_t\}$. Specifically, $\widehat{A}^{(0)}$ and $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ are solutions to the following optimization problems:

$$\widehat{A}^{(0)} := \arg\min_{A} \ \big\{ \tfrac{1}{T} |||\mathbf{X}_T - \mathbf{X}A^\top|||_{\mathrm{F}} + \lambda_A \|A\|_1 \big\}, \tag{B.1}$$

$$(\widehat{B}^{(0)}, \widehat{C}^{(0)}) := \arg\min_{B,C} \ \big\{ \tfrac{1}{T} |||\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top|||_{\mathrm{F}} + \lambda_B |||B|||_* + \lambda_C \|C\|_1 \big\}. \tag{B.2}$$

Note that they also correspond to estimators of the setting where there is no contemporaneous dependence among the idiosyncratic error processes. If we additionally introduce operators $\mathfrak{X}_0$ and $\mathfrak{W}_0$ defined as

$$\begin{aligned} \mathfrak{X}_0 : &\quad \mathfrak{X}_0(\Delta) = \mathbf{X}\Delta^\top, \ \text{ for } \Delta \in \mathbb{R}^{p_1 \times p_1}, \\ \mathfrak{W}_0 : &\quad \mathfrak{W}_0(\Delta) = \mathbf{W}\Delta^\top, \ \text{ for } \Delta \in \mathbb{R}^{p_2 \times (p_1+p_2)} \quad \text{where } \mathbf{W} := [\mathbf{X}, \mathbf{Z}], \end{aligned}$$

then (B.1) and (B.2) can be equivalently written as

$$\begin{aligned} \widehat{A}^{(0)} &:= \arg\min_{A} \ \big\{ \tfrac{1}{T} |||\mathbf{X}_T - \mathfrak{X}_0(A)|||_{\mathrm{F}} + \lambda_A \|A\|_1 \big\}, \\ (\widehat{B}^{(0)}, \widehat{C}^{(0)}) &:= \arg\min_{B,C} \ \big\{ \tfrac{1}{T} |||\mathbf{Z}_T - \mathfrak{W}_0(B_{\mathrm{aug}}, C_{\mathrm{aug}})|||_{\mathrm{F}} + \lambda_B |||B|||_* + \lambda_C \|C\|_1 \big\}, \end{aligned}$$

where $B_{\mathrm{aug}} := [B, O_{p_2 \times p_2}], C_{\mathrm{aug}} := [O_{p_2 \times p_1}, C]$.

Theorem B.1 (Error bounds for $\widehat{A}^{(0)}$)**.** *Suppose the operator $\mathfrak{X}_0$ satisfies the RSC condition with norm $\Phi(\Delta) = \|\Delta\|_1$, curvature $\alpha_{RSC} > 0$ and tolerance $\tau > 0$, so that*

$$s_A^\star \tau \leq \alpha_{RSC}/32.$$

Then, with regularization parameter λ_A satisfying $\lambda_A \geq 4\|\mathbf{X}^\top \mathbf{U}/T\|_\infty$, the solution to (B.1) *satisfies the following bounds:*

$$\|\!|\widehat{A}^{(0)} - A^\star|\!\|_F \leq 12\sqrt{s_A^\star}\lambda_A/\alpha_{RSC} \qquad \text{and} \qquad \|\widehat{A} - A^\star\|_1 \leq 48 s_A^\star \lambda_A/\alpha_{RSC}.$$

Theorem B.2 (Error bound for $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$)**.** *Let $\mathcal{J}_{C^\star}$ be the support set of $C^\star$ and $s_C^\star$ denote its cardinality. Let $r_B^\star$ be the rank of $B^\star$. Assume that $\mathfrak{W}_0$ satisfies the RSC condition with norm*

$$\Phi(\Delta) := \inf_{B_{aug}+C_{aug}=\Delta} \mathcal{Q}(B,C), \qquad \text{where} \quad \mathcal{Q}(B,C) := |\!|\!|B|\!|\!|_* + \tfrac{\lambda_C}{\lambda_B}\|C\|_1,$$

curvature α_{RSC} and tolerance τ such that

$$128\tau r_B^\star < \alpha_{RSC}/4 \qquad \text{and} \qquad 64\tau s_C^\star(\lambda_C/\lambda_B)^2 < \alpha_{RSC}/4.$$

Then, with regularization parameters λ_B and λ_C satisfying

$$\lambda_B \geq 4|\!|\!|\mathbf{W}^\top \mathbf{V}/T|\!|\!|_{op} \quad \text{and} \quad \lambda_C \geq 4\big\|\mathbf{W}^\top \mathbf{V}/T\big\|_\infty,$$

the solution to (B.2) *satisfies the following bounds:*

$$|\!|\!|\widehat{B}^{(0)} - B^\star|\!|\!|_F^2 + |\!|\!|\widehat{C}^{(0)} - C^\star|\!|\!|_F^2 \leq 4\left(2r_B^\star\lambda_B^2 + s_C^\star\lambda_C^2\right)/\alpha_{RSC}^2.$$

In the rest of this subsection, we first prove Theorem B.1 and B.2, then prove Theorem 3.1 and 3.2, whose statements are given in Section 3.3.2.

Proof of Theorem B.1. For the ease of notation, in this proof, we use $\widehat{A}$ to refer to $\widehat{A}^{(0)}$ whenever there is no ambiguity. Let $\beta_A^\star = \mathrm{vec}(A^\star)$ and denote the residual matrix and its vectorized version by $\Delta_A = \widehat{A} - A^\star$ and $\Delta_{\beta_A} = \widehat{\beta}_A - \beta_A^\star$, respectively. By the optimality of $\widehat{A}$ and the feasibility of $A^\star$, the following *basic inequality* holds:

$$\tfrac{1}{T}|\!|\!|\mathfrak{X}_0(\Delta_A)|\!|\!|_{\mathrm{F}}^2 \leq \tfrac{2}{T}\langle\!\langle \Delta_A, \mathbf{X}^\top \mathbf{U}\rangle\!\rangle + \lambda_A\left\{\|A^\star\|_1 - \|A^\star + \Delta_A\|_1\right\},$$

which is equivalent to:

$$\Delta_{\beta_A}^\top \widehat{\Gamma}_X^{(0)} \Delta_{\beta_A} \leq \tfrac{2}{T}\langle \Delta_{\beta_A}, \mathrm{vec}(\mathbf{X}^\top \mathbf{U})\rangle + \lambda_A \left\{ \|\beta_A^\star\|_1 - \|\beta_A^\star + \Delta_{\beta_A}\|_1 \right\}, \tag{B.3}$$

where $\widehat{\Gamma}_X^{(0)} = \mathrm{I}_{p_1} \otimes \frac{\mathbf{X}'\mathbf{X}}{T}$. By Hölder's inequality and the triangle inequality, an upper bound for the right-hand-side of (B.3) is given by

$$\tfrac{2}{T}\|\Delta_{\beta_A}\|_1 \left\|\mathbf{X}^\top \mathbf{U}\right\|_\infty + \lambda_A \|\Delta_{\beta_A}\|_1. \tag{B.4}$$

Now with the specified choice of λ_A, by Lemma B.5, $\|\Delta_{\beta_A|\mathcal{J}_{A^\star}}\|_1 \leq 3\|\Delta_{\beta_A|\mathcal{J}^c_{A^\star}}\|_1$ i.e., $\Delta_{\beta_A} \in \mathcal{C}(\mathcal{J}_{A^\star}, 3)$, hence $\|\Delta_{\beta_A}\|_1 \leq 4\|\Delta_{\beta_A|\mathcal{J}_{A^\star}}\|_1 \leq 4\sqrt{s_A^\star}\|\Delta_{\beta_A}\|$. By choosing $\lambda_A \geq 4\|\mathbf{X}^\top \mathbf{U}/T\|_\infty$, (B.4) is further upper bounded by

$$\frac{3}{2}\lambda_A \|\Delta_{\beta_A}\|_1 \leq 6\sqrt{s_A^\star}\lambda_A \|\Delta_{\beta_A}\|. \tag{B.5}$$

Combined with the RSC condition and the upper bound given in (B.5), we have

$$\begin{aligned}
&\frac{\alpha_{\mathrm{RSC}}}{2}\|\Delta_{\beta_A}\|^2 - \frac{\tau}{2}\|\Delta_{\beta_A}\|_1^2 \leq \frac{1}{2}\Delta_{\beta_A}^\top \widehat{\Gamma}_X^{(0)} \Delta_{\beta_A} \leq 3\sqrt{s_A^\star}\lambda_A \|\Delta_{\beta_A}\|, \\
&\frac{\alpha_{\mathrm{RSC}}}{4}\|\Delta_{\beta_A}\|^2 \leq \left(\frac{\alpha_{\mathrm{RSC}}}{2} - \frac{16 s_A^\star \tau}{4}\right)\|\Delta_{\beta_A}\|^2 \leq 3\sqrt{s_A^\star}\lambda_A \|\Delta_{\beta_A}\|,
\end{aligned}$$

which implies

$$\|\Delta_{\beta_A}\| \leq 12\sqrt{s_A^\star}\lambda_A/\alpha_{\mathrm{RSC}} \qquad \text{and} \qquad \|\Delta_{\beta_A}\|_1 \leq 48 s_A^\star \lambda_A/\alpha_{\mathrm{RSC}}.$$

It is easy to see that these bounds also hold for $\|\Delta_A\|_{\mathrm{F}}$ and $\|\Delta_A\|_1$, respectively. □

Next, to prove Theorem B.2, we introduce the following two sets of subspaces $\{\mathcal{S}_\Theta, \mathcal{S}_\Theta^\perp\}$ and $\{\mathcal{R}_\Theta, \mathcal{R}_\Theta^c\}$ associated with some generic matrix $\Theta \in \mathbb{R}^{m_1 \times m_2}$, in which the nuclear norm and the ℓ_1-norm are decomposable, respectively [see **?**]. Specifically, let the singular value decomposition of Θ be $\Theta = U\Sigma V'$ with U and V being orthogonal matrices. Let $r = \mathrm{rank}(\Theta)$, and we use U^r and V^r to denote the first r columns of U and V associated with the r singular values of Θ. Further, define

$$\begin{aligned}
\mathcal{S}_\Theta &:= \left\{\Delta \in \mathbb{R}^{m_1 \times m_2} | \mathrm{row}(\Delta) \subseteq V^r \quad \text{and} \quad \mathrm{col}(\Delta) \subseteq U^r \right\}, \\
\mathcal{S}_\Theta^\perp &:= \left\{\Delta \in \mathbb{R}^{m_1 \times m_2} | \mathrm{row}(\Delta) \perp V^r \quad \text{and} \quad \mathrm{col}(\Delta) \perp U^r \right\}.
\end{aligned} \tag{B.6}$$

Then, for an arbitrary (generic) matrix $M \in \mathbb{R}^{m_1 \times m_2}$, its restriction on the subspace $\mathcal{S}(\Theta)$

and $\mathcal{S}^{\perp}(\Theta)$, denoted by $M_{\mathcal{S}(\Theta)}$ and $M_{\mathcal{S}^{\perp}(\Theta)}$ respectively, are given by:

$$M_{\mathcal{S}_\Theta} = U \begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{21} & O \end{bmatrix} V' \qquad \text{and} \qquad M_{\mathcal{S}_\Theta^\perp} = U \begin{bmatrix} O & O \\ O & \widetilde{M}_{22} \end{bmatrix} V',$$

where $\Theta = U\Sigma V'$ and $\widetilde{M}$ is defined and partitioned as

$$\widetilde{M} = U'MV = \begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{21} & \widetilde{M}_{22} \end{bmatrix}, \quad \text{where } \widetilde{M}_{11} \in \mathbb{R}^{r\times r}.$$

Note that by Lemma B.7, $M_{\mathcal{S}_\Theta} + M_{\mathcal{S}_\Theta^\perp} = M$. Moreover, when Θ is restricted to the subspace induced by itself $\Theta_{\mathcal{S}_\Theta}$ (and we write $\Theta_{\mathcal{S}}$ for short for this specific case), the following decomposition for the nuclear norm holds:

$$|||\Theta|||_* = |||\Theta_{\mathcal{S}} + \Theta_{\mathcal{S}^\perp}|||_* = |||\Theta_{\mathcal{S}}|||_* + |||\Theta_{\mathcal{S}^\perp}|||_*.$$

Let $\mathcal{J}_\Theta$ be the set of indexes in which Θ is nonzero. Analogously, we define

$$\begin{aligned} \mathcal{R}_\Theta &:= \left\{\Delta \in \mathbb{R}^{m_1\times m_2} | \Delta_{ij} = 0 \ \text{ for } \ (i,j) \notin \mathcal{J}_\Theta\right\}, \\ \mathcal{R}_\Theta^c &:= \left\{\Delta \in \mathbb{R}^{m_1\times m_2} | \Delta_{ij} = 0 \ \text{ for } \ (i,j) \in \mathcal{J}_\Theta\right\}. \end{aligned} \tag{B.7}$$

Then, for an arbitrary matrix M, $M_{\mathcal{J}_\Theta} \in \mathcal{R}_\Theta$ is obtained by setting the entries of M whose indexes are not in $\mathcal{J}_\Theta$ to 0, and $M_{\mathcal{J}_\Theta^c} \in \mathcal{R}_\Theta^c$ is obtained by setting the entries of M whose indexes are in $\mathcal{J}_\Theta$ to 0. Then, the following decomposition holds:

$$\left\|M_{\mathcal{J}_\Theta} + M_{\mathcal{J}_\Theta^c}\right\|_1 = \|M_{\mathcal{J}_\Theta}\|_1 + \left\|M_{\mathcal{J}_\Theta^c}\right\|_1.$$

Proof of Theorem B.2. Again for the ease of notation, in this proof, we drop the superscript and use $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ to denote $(\widehat{B}, \widehat{C})$ whenever there is no ambiguity. Define $\mathcal{Q}$ to be the weighted regularizer:

$$\mathcal{Q}(B, C) = |||B|||_* + \tfrac{\lambda_C}{\lambda_B}\|C\|_1.$$

Note that $(B^\star, C^\star)$ is always feasible, and by the optimality of $(\widehat{B}, \widehat{C})$, the following inequality holds:

$$\frac{1}{T}|||\mathbf{Z}_T - \mathfrak{W}_0(\widehat{B}_{\text{aug}} + \widehat{C}_{\text{aug}})|||_{\text{F}}^2 + \lambda_B|||\widehat{B}|||_* + \lambda_C\|\widehat{C}\|_1 \le \frac{1}{T}|||\mathbf{Z}_T - \mathfrak{W}_0(B^\star + C^\star)|||_{\text{F}}^2 + \lambda_B|||B^\star|||_* + \lambda_C\|C^\star\|_1,$$

By defining $\Delta_{\text{aug}}^B = \widehat{B}_{\text{aug}} - B_{\text{aug}}^\star = [\Delta^B, O]$, $\Delta_{\text{aug}}^C = \widehat{C}_{\text{aug}} - C_{\text{aug}}^\star = [O, \Delta^C]$, we obtain the

following *basic inequality*:

$$\tfrac{1}{T}|||\mathfrak{W}_0(\Delta^B_{\text{aug}} + \Delta^C_{\text{aug}})|||^2_{\text{F}} \leq \tfrac{2}{T}\langle\!\langle \Delta^B_{\text{aug}} + \Delta^C_{\text{aug}}, \mathbf{W}^\top \mathbf{V} \rangle\!\rangle + \lambda_B \mathcal{Q}(B^\star, C^\star) - \lambda_B \mathcal{Q}(\widehat{B}, \widehat{C}). \quad \text{(B.8)}$$

By Hölder's inequality and Lemma B.8, we have

$$\begin{aligned}\tfrac{1}{T}|||\mathfrak{W}_0(\Delta^B_{\text{aug}} + \Delta^C_{\text{aug}})|||^2_{\text{F}} \leq \tfrac{2}{T}\big(|||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + |||\Delta^B_{\mathcal{S}^\perp_{B^\star}}|||_*\big)|||\mathbf{W}^\top\mathbf{V}|||_{\text{op}} + \tfrac{2}{T}\big(\|\Delta^C_{\mathcal{J}^c_{C^\star}}\|_1 + \|\Delta^C_{\mathcal{J}^c_{C^\star}}\|_1\big)\left\|\mathbf{W}^\top\mathbf{V}\right\|_\infty \\ + \lambda_B \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \lambda_B \mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}). \quad \text{(B.9)}\end{aligned}$$

With the specified choice of λ_B and λ_C, after some algebra, (B.9) is further bounded by

$$\frac{3\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}).$$

By Lemma B.9, and using this upper bound, we obtain

$$\frac{\alpha_{\text{RSC}}}{2}(|||\Delta^B|||^2_{\text{F}} + |||\Delta^C|||^2_{\text{F}}) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B, \Delta^C) \leq \frac{3\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}).$$

By the triangle inequality, $\mathcal{Q}(\Delta^B, \Delta^C) \leq \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) + \mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}})$, rearranging gives

$$\frac{\alpha_{\text{RSC}}}{2}(|||\Delta^B|||^2_{\text{F}} + |||\Delta^C|||^2_{\text{F}}) \leq 2\lambda_B \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}). \quad \text{(B.10)}$$

By Lemma B.7, with $N = B^\star$, $M_1 = \Delta^B_{\mathcal{S}_{B^\star}}$, $M_2 = \Delta^B_{\mathcal{S}^\perp_{B^\star}}$, we get

$$\text{rank}(\Delta^B_{\mathcal{S}_{B^\star}}) \leq 2r^\star_B \qquad \text{and} \qquad \langle\!\langle \Delta^B_{\mathcal{S}_{B^\star}}, \Delta^B_{\mathcal{S}^\perp_{B^\star}} \rangle\!\rangle = 0,$$

which implies $|||\Delta^B_{\mathcal{S}_{B^\star}}|||_* \leq (\sqrt{2r^\star_B})|||\Delta^B_{\mathcal{S}_{B^\star}}|||_{\text{F}} \leq (\sqrt{2r^\star_B})|||\Delta^B|||_{\text{F}}$. Since $\Delta^C_{\mathcal{J}_{C^\star}}$ has at most $s^\star_C$ nonzero entries, it follows that $\|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 \leq \sqrt{s^\star_C}|||\Delta^C_{\mathcal{J}_{C^\star}}|||_{\text{F}} \leq \sqrt{s^\star_C}|||\Delta^C|||_{\text{F}}$. Therefore,

$$\mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) = \lambda_B|||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \lambda_C\big\|\Delta^C_{\mathcal{J}_{C^\star}}\big\|_1 \leq \lambda_B(\sqrt{2r^\star_B})|||\Delta^B|||_{\text{F}} + \lambda_C(\sqrt{s^\star_C})|||\Delta^C|||_{\text{F}}$$

With an application of the Cauchy-Schwartz inequality, (B.10) yields:

$$\frac{\alpha_{\text{RSC}}}{2}(|||\Delta^B|||^2_{\text{F}} + |||\Delta^C|||^2_{\text{F}}) \leq \sqrt{2r^\star_B\lambda^2_B + s^\star_C\lambda^2_C} * \sqrt{|||\Delta^B|||^2_{\text{F}} + |||\Delta^C|||^2_{\text{F}}}$$

and we obtain the following bound:

$$|||\Delta^B|||^2_{\text{F}} + |||\Delta^C|||^2_{\text{F}} \leq 4\left(2r^\star_B\lambda^2_B + s^\star_C\lambda^2_C\right)/\alpha^2_{\text{RSC}}.$$

□

Proof of Theorem 3.1. At iteration 0, $\widehat{A}^{(0)}$ solves the following optimization problem:

$$\widehat{A}^{(0)} = \underset{A\in\mathbb{R}^{p_1\times p_1}}{\arg\min} \ \big\{\tfrac{1}{T}\|\!|\mathbf{X}_T - \mathbf{X}A^\top|\!\|_{\mathrm{F}}^2 + \lambda_A|\!|\!|A|\!|\!|_*\big\}.$$

By Theorem B.1, its error bound is given by

$$\|\widehat{A}^{(0)} - A^\star\|_1 \le 48 s_A^\star \lambda_A/\alpha_{\mathrm{RSC}},$$

provided that $\widehat{\Gamma}_X^{(0)} = \mathrm{I}_{p_1}\otimes \mathbf{X}^\top\mathbf{X}/T$ satisfies the RSC condition, and the regularization parameter λ_A satisfies $\lambda_A \ge 4\|\mathbf{X}^\top\mathbf{U}/T\|_\infty$. For random realizations $\mathbf{X}$ and $\mathbf{U}$, by Lemma B.1 and Lemma B.2, there exist constants c_i and c_i' such that for sample size $T \gtrsim s_A^\star \log p_1$, with probability at least $1 - c_1\exp(-c_2 T\min\{1,\omega^{-2}\})$, where $\omega = c_3\mu_{\max}(\mathcal{A})/\mu_{\min}(\mathcal{A})$

$$(\mathbf{E_1}) \qquad \widehat{\Gamma}_X^{(0)} \text{ satisfies RSC condition with } \alpha_{\mathrm{RSC}} = \Lambda_{\min}(\Sigma_u^\star)/(2\mu_{\max}(\mathcal{A}))\ ,$$

and with probability at least $1 - c_1'\exp(-c_2'\log p_1)$,

$$(\mathbf{E_2}) \qquad \|\mathbf{X}^\top\mathbf{U}/T\|_\infty \le C_0\sqrt{\frac{\log p_1}{T}}, \qquad \text{for some constant } C_0.$$

Hence with probability at least $1 - c_1\exp(-c_2 T) - c_1'\exp(-c_2'\log p_1)$,

$$\|\widehat{A}^{(0)} - A^\star\|_1 = O\Big(s_A^\star\sqrt{\tfrac{\log p_1}{T}}\Big).$$

Moving onto $\widehat{\Omega}_u^{(0)}$, which is given by

$$\widehat{\Omega}_u^{(0)} = \underset{\Omega_u\in\mathbb{S}_{++}^{p_1\times p_1}}{\arg\min} \ \Big\{\log\det\Omega_u - \mathrm{trace}\big(\widehat{S}_u^{(0)}\Omega_u\big) + \rho_u\|\Omega_u\|_{1,\mathrm{off}}\Big\},$$

where $\widehat{S}_u^{(0)} = \frac{1}{T}(\mathbf{X}_T - \mathbf{X}\widehat{A}^{(0)\top})^\top(\mathbf{X}_T - \mathbf{X}\widehat{A}^{(0)\top})$. By Theorem 1 in **?**], the error bound for $\widehat{\Omega}_u^{(0)}$ relies on how well $\widehat{S}_u^{(0)}$ concentrates around $\Sigma_u^\star$, more specifically, $\|\widehat{S}_u^{(0)} - \Sigma_u^\star\|_\infty$. Note that

$$\|\widehat{S}_u^{(0)} - \Sigma_u^\star\|_\infty \le \|S_u - \Sigma_u^\star\|_\infty + \|\widehat{S}_u^{(0)} - S_u\|_\infty,$$

where $S_u = \mathbf{U}^\top\mathbf{U}/T$ is the sample covariance based on true errors. For the first term, by **?**], there exists constant $\tau_0 > 2$ such that with probability at least $1 - 1/p_1^{\tau_0-2} =$

$1-\exp(-\tau\log p_1)$ $(\tau>0)$, the following bound holds:

$$(\mathbf{E_3}) \qquad \|S_u-\Sigma_u^\star\|_\infty \le C_1\sqrt{\frac{\log p_1}{T}}, \qquad \text{for some constant } C_1.$$

For the second term,

$$\widehat{S}_u^{(0)}-S_u=\tfrac{2}{T}\mathbf{U}^\top\mathbf{X}(A^\star-\widehat{A}^{(0)})^\top+(A^\star-\widehat{A}^{(0)})\left(\tfrac{\mathbf{X}'\mathbf{X}}{T}\right)(A^\star-\widehat{A}^{(0)})':=I_1+I_2,$$

For I_1, based on the analysis of $\|A^\star-\widehat{A}^{(0)}\|_1$ and $\|\mathbf{X}^\top\mathbf{U}/T\|_\infty$,

$$\|I_1\|_\infty\le 2\|\|A^\star-\widehat{A}^{(0)}\|\|_\infty\|\tfrac{1}{T}\mathbf{X}^\top\mathbf{U}\|_\infty\le 2\|A^\star-\widehat{A}^{(0)}\|_1\|\tfrac{1}{T}\mathbf{X}^\top\mathbf{U}\|_\infty=O\Big(\tfrac{s_A^\star\log p_1}{T}\Big)$$

For I_2,

$$\begin{aligned}\|(A^\star-\widehat{A}^{(0)})\left(\tfrac{\mathbf{X}^\top\mathbf{X}}{T}\right)(A^\star-\widehat{A}^{(0)})^\top\|_\infty &\le \|\|A^\star-\widehat{A}^{(0)}\|\|_\infty\|\|A^\star-\widehat{A}^{(0)}\|\|_1\|\tfrac{\mathbf{X}^\top\mathbf{X}}{T}\|_\infty\\ &\le \|A^\star-\widehat{A}^{(0)}\|_1^2\|\tfrac{\mathbf{X}'\mathbf{X}}{T}\|_\infty,\end{aligned}$$

where by Proposition 2.4 in **?**] and then taking a union bound, with probability at least $1-c_1''\exp(-c_2''\log p_1)$ $(c_1'',c_2''>0)$,

$$(\mathbf{E_4}) \qquad \|\tfrac{\mathbf{X}^\top\mathbf{X}}{T}\|_\infty\le C_2\sqrt{\frac{\log p_1}{T}}+\Lambda_{\max}(\Gamma_X), \qquad \text{for some constant } C_2.$$

Hence,

$$\|I_2\|_\infty=O\Big((s_A^\star)^2\big(\tfrac{\log p_1}{T}\big)^{3/2}\Big)+O\Big((s_A^\star)^2\tfrac{\log p_1}{T}\Big)$$

Combining all terms, and since we assume that $T^{-1}\log p_1$ is small, $O(\sqrt{T^{-1}\log p_1})$ becomes the leading term, and the following bound holds with probability at least $1-c_1\exp(-c_2T)-c_1'\exp(-c_2'\log p_1)-c_1''\exp(-c_2''\log p_1)-\exp(-\tau\log p_1)$:

$$\|\widehat{S}_u^{(0)}-\Sigma_u^\star\|_\infty=O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big).$$

Consequently,

$$\|\widehat{\Omega}_u^{(0)}-\Omega_u^\star\|_\infty=O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big).$$

At iteration 1, the vectorized $\widehat{A}^{(1)}$ solves

$$\widehat{\beta}_A^{(1)}=\underset{\beta\in\mathbb{R}^{p_1^2}}{\arg\min}\ \big\{-2\beta^\top\widehat{\gamma}_X^{(1)}+\beta'\widehat{\Gamma}_X^{(1)}\beta+\lambda_A\|\beta\|_1\big\},$$

where

$$\widehat{\gamma}_X^{(1)} = \tfrac{1}{T}\big(\widehat{\Omega}_u^{(0)} \otimes \mathbf{X}^\top\big)\mathrm{vec}(\mathbf{X}_T), \qquad \widehat{\Gamma}_X^{(1)} = \widehat{\Omega}_u^{(0)} \otimes \tfrac{\mathbf{X}^\top\mathbf{X}}{T}.$$

The error bound for $\widehat{\beta}_A^{(1)}$ relies on (1) $\widehat{\Gamma}_X^{(1)}$ satisfying the RSC condition, which holds for sample size $T \gtrsim (d_{\Omega_u^\star}^{\max})^2 \log p_1$ upon $\|\widehat{\Omega}_u^{(0)} - \Omega_u^\star\|_\infty = O(\sqrt{T^{-1}\log p_1})$; and (2) a bound for $\|\mathbf{X}^\top\mathbf{U}\widehat{\Omega}_u^{(0)}/T\|_\infty$. For $\|\mathbf{X}^\top\mathbf{U}\widehat{\Omega}_u^{(0)}/T\|_\infty$,

$$\tfrac{1}{T}\mathbf{X}^\top\mathbf{U}\widehat{\Omega}_u^{(0)} = \tfrac{1}{T}\mathbf{X}^\top\mathbf{U}\Omega_u^\star + \tfrac{1}{T}\mathbf{X}^\top\mathbf{U}(\widehat{\Omega}_u^{(0)} - \Omega_u^\star) := I_3 + I_4.$$

For I_3, by Lemma 3 in **?**] and with the aid of Proposition 2.4 in **?**], again with probability at least $1 - c_1'''\exp(-c_2'''\log p_1)$ we get

$$(\mathbf{E_5}) \qquad \big\|\tfrac{1}{T}\mathbf{X}^\top\mathbf{U}\Omega_u^\star\big\|_\infty \le C_3\sqrt{\frac{\log p_1}{T}}, \qquad \text{for some constant } C_3.$$

For I_4, by Corollary 3 in **?**], we get

$$\Big\|\tfrac{1}{T}\mathbf{X}^\top\mathbf{U}(\widehat{\Omega}_u^{(0)} - \Omega_u^\star)\Big\|_\infty \le d_{\Omega_u^\star}^{\max}\|\tfrac{1}{T}\mathbf{X}'\mathbf{U}\|_\infty\|\widehat{\Omega}_u^{(0)} - \Omega_u^\star\|_\infty = O\Big(\tfrac{\log p_1}{T}\Big).$$

Combining all terms and taking the leading one, once again we have

$$\|\widehat{A}^{(1)} - A^\star\|_1 = O\Big(s_A^\star\sqrt{\tfrac{\log p_1}{T}}\Big),$$

which holds with probability at least $1 - c_1\exp(-c_2T) - \tilde{c}_1\exp(-\tilde{c}_2\log p_1) - \exp(-\tau\log p_1)$, by letting $\tilde{c}_1 = \max\{c_1', c_1'', c_1'''\}$ and $\tilde{c}_1 = \min\{c_2', c_2'', c_2'''\}$. It should be noted that up to this step, all sources of randomness from the random realizations have been captured by events from $\mathbf{E_1}$ to $\mathbf{E_5}$; thus, for $\widehat{\Omega}_u^{(1)}$ and iterations thereafter, the probability for which the bounds hold will no longer change, and the same holds for the error bounds for $\widehat{A}^{(k)}$ and $\widehat{\Omega}_u^{(k)}$ in terms of the relative order with respect to the dimension p_1 and sample size T. Therefore, we conclude that with high probability, for all iterations k,

$$\|\mathbf{X}^\top\mathbf{U}\widehat{\Omega}_u^{(k)}/T\|_\infty = O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big), \qquad \|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty = O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big).$$

With the aid of Theorem B.1, it then follows that

$$|||\widehat{A}^{(k)} - A^\star|||_{\mathrm{F}} = O\Big(\sqrt{\tfrac{s_A^\star\log p_1}{T}}\Big), \qquad |||\widehat{\Omega}_u^{(k)} - \Omega_u^\star|||_{\mathrm{F}} = O\Big(\sqrt{\tfrac{(s_{\Omega_u^\star} + p_1)\log p_1}{T}}\Big).$$

□

Proof of Theorem 3.2. At iteration 0, $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ solves the following optimization:

$$(\widehat{B}^{(0)}, \widehat{C}^{(0)}) = \underset{(B,C)}{\arg\min} \left\{ \tfrac{1}{T} |||\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top|||_{\mathrm{F}}^2 + \lambda_B |||B|||_* + \lambda_C \|C\|_1 \right\}.$$

Let $W_t = (X_t^\top, Z_t^\top)^\top \in \mathbb{R}^{p_1+p_2}$ be the joint process and $\mathbf{W}$ be the realizations, with operators $\mathfrak{W}_0$ identically defined to that in Theorem B.2 . By Theorem B.2,

$$|||\widehat{B}^{(0)} - B^\star|||_{\mathrm{F}}^2 + |||\widehat{C}^{(0)} - C^\star|||_{\mathrm{F}}^2 \le 4(2r_B^\star \lambda_B^2 + s_C^\star \lambda_C^2)/\alpha_{\mathrm{RSC}}^2,$$

provided that $\mathfrak{W}$ satisfies the RSC condition and λ_B, λ_C respectively satisfy

$$\lambda_B \ge 4|||\mathbf{W}^\top \mathbf{V}/T|||_{\mathrm{op}} \quad \text{and} \quad \lambda_C \ge 4\big\|\mathbf{W}^\top \mathbf{V}/T\big\|_\infty.$$

In particular, by Lemma B.3 for random realizations of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{V}$, for sample size $T \gtrsim (p_1 + 2p_2)$, with probability at least $1 - c_1 \exp\{-c_2(p_1+p_2)\}$,

$$(\mathbf{E_1'}) \qquad \mathfrak{W}_0 \text{ satisfies the RSC condition.}$$

By Lemma B.4, for sample size $T \gtrsim (p_1 + 2p_2)$ and some constant $C_1, C_2 > 0$,

$$(\mathbf{E_2'}) \qquad |||\mathbf{W}^\top \mathbf{V}/T|||_{\mathrm{op}} \le C_1 \sqrt{\frac{p_1+2p_2}{T}} \quad \text{and} \quad \big\|\mathbf{W}^\top \mathbf{V}/T\big\|_\infty \le C_2 \sqrt{\frac{\log(p_1+p_2) + \log p_2}{T}},$$

with probability at least $1 - c_1' \exp\{-c_2'(p_1+2p_2)\}$ and $1 - c_1'' \exp\{-c_2'' \log[p_2(p_1+p_2)]\}$, respectively. Hence, with probability at least

$$1 - c_1 \exp\{-c_2(p_1+p_2)\} - c_1' \exp\{-c_2'(p_1+2p_2)\} - c_1'' \exp\{-c_2'' \log[p_2(p_1+p_2)]\},$$

the following bound holds for the initializers as long as sample size $T \gtrsim (p_1 + 2p_2)$:

$$|||\widehat{B}^{(0)} - B^\star|||_{\mathrm{F}}^2 + |||\widehat{C}^{(0)} - C^\star|||_{\mathrm{F}}^2 = O\Big(\tfrac{p_1+2p_2}{T}\Big) + O\Big(\tfrac{\log(p_1+p_2)+\log p_2}{T}\Big). \tag{B.11}$$

Considering the estimation of $\widehat{\Omega}_v^{(0)}$, it solves a graphical Lasso problem:

$$\widehat{\Omega}_v^{(0)} = \underset{\Omega_v \in \mathbb{S}_{++}^{p_2 \times p_2}}{\arg\min} \left\{ \log\det \Omega_v - \mathrm{trace}\big(\widehat{S}_u^{(0)} \Omega_v\big) + \rho_v \|\Omega_v\|_{1,\mathrm{off}} \right\},$$

where $\widehat{S}_v^{(0)} = \frac{1}{T}(\mathbf{Z}_T - \mathbf{X}\widehat{B}^{(0)\top} - \mathbf{Z}\widehat{C}^{(0)\top})^\top (\mathbf{Z}_T - \mathbf{X}\widehat{B}^{(0)\top} - \mathbf{Z}\widehat{C}^{(0)\top})$. Similar to the proof of

Theorem 3.1, the error bound for $\widehat{\Omega}_v^{(0)}$ depends on $\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty$, which can be decomposed as

$$\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty \le \|S_v - \Sigma_v^\star\|_\infty + \|\widehat{S}_v^{(0)} - S_v\|_\infty,$$

where $S_v = \mathbf{V}^\top \mathbf{V}/T$ is the sample covariance based on the true errors. For the first term, by Lemma 1 in **?**], there exists constant $\tau_0 > 2$ such that with probability at least $1 - 1/p_2^{\tau_0 - 2} = 1 - \exp(-\tau \log p_2)$ $(\tau > 0)$, the following bound holds:

$$(\mathbf{E_3'}) \qquad \|S_v - \Sigma_v^\star\|_\infty \le C_3 \sqrt{\frac{\log p_1}{T}}, \qquad \text{for some constant } C_3.$$

For the second term, let $\Pi = [B, C] \in \mathbb{R}^{p_2 \times (p_1 + p_2)}$, then

$$\widehat{S}_v^{(0)} - S_v = \frac{2}{T} \mathbf{V}^\top \mathbf{W} (\Pi^\star - \widehat{\Pi}^{(0)})^\top + (\Pi^\star - \widehat{\Pi}^{(0)}) \left(\frac{\mathbf{W}^\top \mathbf{W}}{T} \right) (\Pi^\star - \widehat{\Pi}^{(0)})^\top := I_1 + I_2,$$

For I_1, we have

$$\|\tfrac{2}{T} \mathbf{V}^\top \mathbf{W} (\Pi^\star - \widehat{\Pi}^{(0)})^\top\|_\infty \le \|\tfrac{2}{T} \mathbf{V}^\top \mathbf{W} (\Pi^\star - \widehat{\Pi}^{(0)})^\top\|_{\mathrm{F}} \le 2 |||\tfrac{1}{T} \mathbf{W}^\top \mathbf{V}|||_{\mathrm{op}} |||\Pi^\star - \widehat{\Pi}^{(0)}|||_{\mathrm{F}}.$$

Consider the leading term of $|||\Pi^\star - \widehat{\Pi}^{(0)}|||_{\mathrm{F}}$ as in (B.11), whose rate is $O(\sqrt{T^{-1}(p_1 + 2p_2)})$. We therefore obtain

$$\|I_1\|_\infty \le \|I_1\|_{\mathrm{F}} = O\Big(\tfrac{p_1 + 2p_2}{T} \Big).$$

Similarly for I_2,

$$\|I_2\|_\infty \le \|I_2\|_{\mathrm{F}} \le |||\Pi^\star - \widehat{\Pi}^{(0)}|||_{\mathrm{F}}^2 |||\tfrac{\mathbf{W}^\top \mathbf{W}}{T}|||_{\mathrm{op}},$$

where with a similar derivation to that in Lemma B.10, for sample size $T \gtrsim (p_1 + p_2)$, with probability at least $1 - c_1''' \exp\{-c_2'''(p_1 + p_2)\}$, we get

$$(\mathbf{E_4'}) \qquad |||\tfrac{\mathbf{W}^\top \mathbf{W}}{T}|||_{\mathrm{op}} \le C_4 \sqrt{\frac{p_1 + 2p_2}{T}} + \Lambda_{\max}(\Gamma_W), \qquad \text{for some constant } C_4.$$

Hence,

$$\|I_2\|_\infty \le \|I_2\|_{\mathrm{F}} \le = O\Big(\big(\frac{p_1 + 2p_2}{T} \big)^{3/2} \Big).$$

Combining all terms and then taking the leading one, with probability at least

$$1 - c_1 \exp\{-c_2(p_1 + p_2)\} - c_1' \exp\{-c_2'(p_1 + 2p_2)\} - c_1'' \exp\{-c_2'' \log[p_2(p_1 + p_2)]\} \\ - c_1''' \exp\{-c_2'''(p_1 + p_2)\} - \exp(-\tau \log p_2),$$

we obtain

$$\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty = O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big).$$

Note that here with the required sample size, $(p_1+2p_2)/T$ is a small quantity, and therefore

$$O\Big(\big(\tfrac{p_1+2p_2}{T}\big)^{3/2}\Big) \le O\Big(\tfrac{p_1+2p_2}{T}\Big) \le O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big).$$

At iteration 1, the bound of $\|\widehat{B}^{(1)} - B^\star\|_{\mathrm{F}}^2 + \|\widehat{C}^{(1)} - C^\star\|_{\mathrm{F}}^2$ relies on the following two quantities:

$$\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\widehat{\Omega}_v^{(0)}\|_{\mathrm{op}} \qquad \text{and} \qquad \Big\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\widehat{\Omega}_v^{(0)}\Big\|_\infty.$$

Using a similar derivation to that in the proof of Theorem 3.1,

$$\Big\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\widehat{\Omega}_v^{(0)}\Big\|_\infty \le \Big\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\Big\|_\infty + \|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\Omega_v^\star\|_\infty, \tag{B.12}$$

where by viewing $\mathbf{V}\Omega_v^\star$ as some random realization coming from a certain sub-Gaussian process, with probability at least $1 - \bar{c}_1'' \exp\{-\bar{c}_2'' \log[p_2(p_1+p_2)]\}$, we get

$$(\mathbf{E_5'}) \qquad \|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\Omega_v^\star\|_\infty \le C_5\sqrt{\frac{\log(p_1+p_2)+\log p_2}{T}}, \qquad \text{for some constant } C_5,$$

and

$$\begin{aligned}\Big\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\Big\|_\infty &\le d_{\max}^{\Omega_v^\star}\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\|_\infty\|\widehat{\Omega}_v^{(0)} - \Omega_v^\star\|_\infty \\ &= O\Big(\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}}\Big)\cdot O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big).\end{aligned}$$

For $\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\widehat{\Omega}_v^{(0)}\|_{\mathrm{op}}$, similarly we have

$$\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\widehat{\Omega}_v^{(0)}\|_{\mathrm{op}} \le \|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\|_{\mathrm{op}} + \|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\Omega_v^\star\|_{\mathrm{op}}, \tag{B.13}$$

where with probability at least $1 - \bar{c}_1' \exp\{-\bar{c}_2'(p_1+p_2)\}$,

$$(\mathbf{E_6'}) \qquad \|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\Omega_v^\star\|_{\mathrm{op}} \le C_6\sqrt{\frac{p_1+2p_2}{T}} \qquad \text{for some constant } C_6,$$

and

$$\begin{aligned}\|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\|_{\mathrm{op}} &\le \|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\|_{\mathrm{op}}\|\widehat{\Omega}_v^{(0)} - \Omega_v^\star\|_{\mathrm{op}} \\ &\le \|\tfrac{1}{T}\mathbf{W}^\top\mathbf{V}\|_{\mathrm{op}}\Big[d_{\max}^{\Omega_v^\star}\|\widehat{\Omega}_v^{(0)} - \Omega_v^\star\|_\infty\Big] = O\Big(\tfrac{p_1+2p_2}{T}\Big),\end{aligned}$$

where the second inequality follows from Corollary 3 of **?**]. Combining all terms from (B.12) and (B.13), the leading term gives the following bound:

$$\|\!|\widehat{B}^{(1)} - B^\star|\!\|_{\mathrm{F}}^2 + \|\!|\widehat{C}^{(1)} - C^\star|\!\|_{\mathrm{F}}^2 \leq C_7\Big(\frac{p_1+2p_2}{T}\Big) \qquad \text{for some constant } C_7,$$

and this error rate coincides with that in the bound of $\|\!|\widehat{B}^{(0)} - B^\star|\!\|_{\mathrm{F}}^2 + \|\!|\widehat{C}^{(0)} - C^\star|\!\|_{\mathrm{F}}^2$. This implies that for $\widehat{\Omega}_v^{(1)}$ and iterations thereafter, the error rate remains unchanged. Moreover, all sources of randomness have been captured up to this step in events $\mathbf{E}_\mathbf{1}'$ to $\mathbf{E}_\mathbf{6}'$, and therefore the probability for the bounds to hold no longer changes. Consequently, the following bounds hold for all iterations k:

$$\|\mathbf{W}^\top \mathbf{V}\widehat{\Omega}_v^{(k)}/T\|_\infty = \|\!|\mathbf{W}'\mathbf{V}\widehat{\Omega}_v^{(k)}/T|\!\|_{\mathrm{op}} = O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big)$$

and

$$\|\widehat{S}_v^{(k)} - \Sigma_v^\star\|_\infty = O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big),$$

with probability at least

$$1-c_0\exp\{-\tilde{c}_0(p_1+p_2)\}-c_1\exp\{-\tilde{c}_1(p_1+2p_2)\}-c_2\exp\{-\tilde{c}_2\log[p_2(p_1+p_2)]\}-\exp\{-\tau\log p_2\}.$$

for some new positive constants $c_i, \tilde{c}_i$ $(i = 0, 1, 2)$ and τ.[1] The above bounds directly imply the bound in the statement in Theorem 3.2, with the aid of Theorem B.2. □

B.2 Key Lemmas and Their Proofs.

In this section, we verify the conditions that are required for establishing the consistency results in Theorem B.1 and B.2, under random realizations of $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{U}$ and $\mathbf{V}$.

The following two lemmas verify the conditions for establishing the consistency properties for $\widehat{A}^{(0)}$. Specifically, Lemma B.1 establishes that with high probability, $\mathfrak{X}_0$ satisfies the RSC condition. Further, Lemma B.2 gives a high probability upper bound for $\|\mathbf{X}'\mathbf{U}/T\|_\infty$ for random $\mathbf{X}$ and $\mathbf{U}$.

Lemma B.1 (Verification of the RSC condition)**.** *For the VAR*(1) *model* $\{X_t\}$ *posited in* (3.1)*, there exist* $c_i > 0$ $(i = 1, 2, 3)$ *such that for sample size* $T \gtrsim \max\{\omega^2, 1\}s_A^\star \log p_1$*,*

[1] Here we slightly abuse the notations and redefine $c_0 := \max\{c_1, c_1'''\}$, $c_1 := \max\{c_1', \bar{c}_1'\}$, $\tilde{c}_1 := \min\{c_2', \bar{c}_2'\}$, $c_2 = \max\{c_1'', \bar{c}_1''\}$, $\tilde{c}_2 := \min\{c_2'', \bar{c}_2''\}$.

with probability at least

$$1 - c_1 \exp\left[-c_2 T \min\{1, \omega^{-2}\}\right], \qquad \omega = c_3 \frac{\Lambda_{\max}(\Sigma_u)\mu_{\max}(\mathcal{A})}{\Lambda_{\min}(\Sigma_u)\mu_{\min}(\mathcal{A})},$$

the following inequality holds

$$\frac{1}{2T}|||\mathfrak{X}_0(\Delta)|||_F^2 \geq \alpha_{RSC}|||\Delta|||_F^2 - \tau\|\Delta\|_1^2, \qquad \textit{for } \Delta \in \mathbb{R}^{p_1 \times p_1},$$

where $\alpha_{RSC} = \frac{\Lambda_{\min}(\Sigma_u)}{\mu_{\max}(\mathcal{A})}$, $\tau = 4\alpha_{RSC}\max\{\omega^2, 1\}\log p_1/T$.

Proof of Lemma B.1. For the specific VAR(1) process $\{X_t\}$ given in (3.1), using Proposition 4.2 in **?**] with $d = 1$ directly gives the result. Specifically, we note that by letting $\theta = \mathrm{vec}(\Delta)$,

$$\tfrac{1}{T}|||\mathfrak{X}_0(\Delta)|||_{\mathrm{F}}^2 = \theta^\top \widehat{\Gamma}_X^{(0)}\theta,$$

where $\widehat{\Gamma}_X^{(0)} = \mathrm{I}_{p_1} \otimes (\mathbf{X}^\top\mathbf{X}/T)$, and $\|\theta\|_2^2 = |||\Delta|||_{\mathrm{F}}^2$, $\|\theta\|_1 = \|\Delta\|_1$. □

Lemma B.2 (Verification of the deviation bound)**.** *For the model in* (3.1)*, there exist constants* $c_i > 0$, $i = 0, 1, 2$ *such that for* $T \gtrsim 2\log p_1$*, with probability at least* $1 - c_1\exp(-2c_2\log p_1)$*, the following bound holds:*

$$\|\mathbf{X}^\top\mathbf{U}/T\|_\infty \leq c_0\Lambda_{\max}(\Sigma_u)\left[1 + \frac{1}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right]\sqrt{\frac{2\log p_1}{T}}. \tag{B.14}$$

Proof of Lemma B.2. First, we note that,

$$\left\|\mathbf{X}^\top\mathbf{U}/T\right\|_\infty = \max_{\substack{1\leq i\leq p_1\\ 1\leq j\leq p_1}} \left|e_i^\top\left(\mathbf{X}^\top\mathbf{U}/T\right)e_j\right|.$$

Applying Proposition 2.4(b) in **?**] for an arbitrary pair of (i, j) gives:

$$\mathbb{P}\left(\left|e_i^\top\left(\mathbf{X}^\top\mathbf{U}/T\right)e_j\right| > \eta\left[\Lambda_{\max}(\Sigma_u)\left(1 + \frac{1}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right)\right]\right) \leq 6\exp[-cT\min\{\eta, \eta^2\}].$$

Setting $\eta = c_0\sqrt{2\log p_1/T}$ and taking a union bound over all $1 \leq i \leq p_1, 1 \leq j \leq p_1$, we get that for some $c_1, c_2 > 0$, with probability at least $1 - c_1\exp[-2c_2\log p_1]$,

$$\max_{\substack{1\leq i\leq p_1\\ 1\leq j\leq p_1}} \left|e_i^\top\left(\mathbf{X}^\top\mathbf{U}/T\right)e_j\right| \leq c_0\Lambda_{\max}(\Sigma_u)\left[1 + \frac{1}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right]\sqrt{\frac{2\log p_1}{T}}.$$

□

In the next two lemmas, Lemma B.3 gives an RSC curvature that holds with high probability for $\mathfrak{W}$ induced by a random $\mathbf{W}$, and Lemma B.4 gives a high probability upper bound for $\left|\left|\left|\mathbf{W}^\top \mathbf{V}/T\right|\right|\right|_{\text{op}}$ and $\left\|\mathbf{W}^\top \mathbf{V}/T\right\|_\infty$.

Lemma B.3 (Verification of the RSC condition)**.** *Consider the covariance stationary process $W_t = (X_t^\top, Z_t^\top)^\top \in \mathbb{R}^{p_1+p_2}$ whose spectral density exists. Suppose $\mathfrak{m}(f_W) > 0$. There exist constants $c_i > 0, i = 1, 2$ such that with probability at least $1 - 2c_1 \exp(-c_2(p_1 + p_2))$, the RSC condition for $\mathfrak{W}$ induced by a random $\mathbf{W}$ holds for α_{RSC} and tolerance 0, where*

$$\alpha_{RSC} = \pi \mathfrak{m}(f_W)/4,$$

whenever $T \gtrsim (p_1 + p_2)$.

Proof of Lemma B.3. First ,we note that the following inequality holds, for any $\mathbf{W}$:

$$\frac{1}{2T}\left|\left|\left|\mathfrak{W}_0(\Delta)\right|\right|\right|_{\text{F}}^2 = \frac{1}{2T}\left|\left|\left|\mathbf{W}\Delta^\top\right|\right|\right|_{\text{F}}^2 = \frac{1}{2T}\sum_{j=1}^{p_2}\left\|[\mathbf{W}^\top\Delta]_j\right\|_2^2 \geq \frac{1}{2}\Lambda_{\min}\left(\widehat{\Gamma}_W^{(0)}\right)\left|\left|\left|\Delta\right|\right|\right|_{\text{F}}^2. \tag{B.15}$$

where $\widehat{\Gamma}_W^{(0)} = \mathbf{W}^\top\mathbf{W}/T$. Applying Lemma 4 in **?**] on $\mathbf{W}$ together with Proposition 2.3 in **?**], the following bound holds with probability at least $1 - 2c_1 \exp[-c_2(p_1 + p_2)]$, as long as $T \gtrsim (p_1 + p_2)$:

$$\Lambda_{\min}\left(\widehat{\Gamma}_W^{(0)}\right) \geq \frac{\Lambda_{\min}(\Gamma_W(0))}{4} \geq \frac{\pi}{2}\mathfrak{m}(f_W),$$

where $\Gamma_W(0) = \mathbb{E}W_t W_t^\top$. Combining with (B.15), the RSC condition holds with $\kappa(\mathfrak{W}) = \pi\mathfrak{m}(f_W)/4$. □

Lemma B.4 (Verification of the deviation bound)**.** *There exist constants $c_i > 0$ and $c_i' > 0, i = 1, 2$ such that the following statements hold:*

(a) With probability at least $1 - c_1 \exp[-c_2(p_1 + 2p_2)]$, as long as $T \gtrsim (p_1 + 2p_2)$,

$$\left|\left|\left|\mathbf{W}^\top\mathbf{V}/T\right|\right|\right|_{op} \leq c_0\left[\mathcal{M}(f_W) + \tfrac{1}{2\pi}\Lambda_{\max}(\Sigma_v) + \mathcal{M}(f_{W,V^+})\right]\sqrt{\frac{p_1 + 2p_2}{T}}. \tag{B.16}$$

(b) With probability at least $1 - c_1' \exp(-c_2' \log(p_1 + p_2) - c_2' \log p_2)$, as long as $T \gtrsim c_3' \log[(p_1 + p_2)p_2]$,

$$\left\|\mathbf{W}^\top\mathbf{V}/T\right\|_\infty \leq c_0'\left[\mathcal{M}(f_W) + \tfrac{1}{2\pi}\Lambda_{\max}(\Sigma_v) + \mathcal{M}(f_{W,V^+})\right]\sqrt{\frac{\log(p_1 + p_2) + \log p_2}{T}}. \tag{B.17}$$

Proof of Lemma B.4. (a) is a direct application of Lemma B.10 on processes $\{W_t\} \in \mathbb{R}^{(p_1+p_2)}$ and $\{V_t^+\} \in \mathbb{R}^{p_2}$, and (b) is a direct application of Lemma B.2. □

B.3 Auxiliary Lemmas and Their Proofs.

Lemma B.5. *Let $\widehat{\beta}$ be the solution to the following optimization problem:*

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \tfrac{1}{2n} \|Y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 \right\},$$

where the data is generated according to $Y = X\beta^\star + E$ with $X \in \mathbb{R}^{n \times p}$ and $E \in \mathbb{R}^n$. The errors E_i are assumed to be i.i.d. for all $i = 1, \cdots, n$. Then, by choosing $\lambda_n \geq 2\|X'E/n\|_\infty$, the error vector $\Delta := \widehat{\beta} - \beta^\star$ satisfies $\|\Delta_{\mathcal{J}}\|_1 \leq 3\|\Delta_{\mathcal{J}^c}\|_1$.

Proof. Note that $\beta^\star$ is always feasible. By the optimality of $\widehat{\beta}$, we have

$$\frac{1}{2n}\|Y - X\widehat{\beta}\|_2^2 + \lambda_n \|\widehat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta^\star\|_2^2 + \lambda_n \|\beta^\star\|_1$$

which after some algebra, can be simplified to

$$\begin{aligned} \frac{1}{2}\Delta^\top \left(\frac{X^\top X}{n}\right) \Delta &\leq \langle \Delta^\top, \frac{1}{n} X^\top E \rangle + \lambda_n \|\beta^\star\|_1 - \lambda_n \|\beta^\star + \Delta\|_1 \\ &\leq \|\Delta\|_1 \|\frac{1}{n} X'E\|_\infty + \lambda_n \|\beta^\star\|_1 - \lambda_n \|\beta^\star + \Delta\|_1, \end{aligned}$$

with the second inequality obtained through an application of from Hölder's inequality. With the specified choice of λ_n, if follows that

$$\begin{aligned} 0 &\leq \frac{\lambda_n}{2}\|\Delta\|_1 + \lambda_n \left\{ \|\beta^\star_{\mathcal{J}}\|_1 - \|\beta^\star_{\mathcal{J}} + \Delta_{\mathcal{J}} + \beta^\star_{\mathcal{J}^c} + \Delta_{\mathcal{J}^c}\|_1 \right\} \\ &= \frac{\lambda_n}{2}(\|\Delta_{\mathcal{J}}\|_1 + \|\Delta_{\mathcal{J}^c}\|_1) + \lambda_n \{ \|\beta^\star_{\mathcal{J}}\|_1 - \|\beta^\star_{\mathcal{J}} + \Delta_{\mathcal{J}}\|_1 - \|\Delta_{\mathcal{J}^c}\|_1 \} \quad (\beta^\star_{\mathcal{J}^c} = 0, \ell_1 \text{ norm decomposable}) \\ &\leq \frac{\lambda_n}{2}(\|\Delta_{\mathcal{J}}\|_1 + \|\Delta_{\mathcal{J}^c}\|_1) + \lambda_n (\|\Delta_{\mathcal{J}}\|_1 - \|\Delta_{\mathcal{J}^c}\|) \qquad (\text{triangle inequality}) \\ &= \frac{3\lambda_n}{2}\|\Delta_{\mathcal{J}}\|_1 - \frac{\lambda_n}{2}\|\Delta_{\mathcal{J}^c}\|_1, \end{aligned}$$

which implies $\|\Delta_{\mathcal{J}}\|_1 \leq 3\|\Delta_{\mathcal{J}^c}\|_1$. □

Lemma B.6. *Consider two centered stationary Gaussian processes $\{X_t\}$ and $\{Z_t\}$. Further, assume that the spectral density of the joint process $\{(X_t', Z_t')'\}$ exists. Denote their cross-*

covariance by $\Gamma_{X,Z}(\ell) := \mathrm{Cov}(X_t, Z_{t+\ell})$*, and their cross-spectral density is defined as*

$$f_{X,Z}(\theta) := \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma_{X,Z}(\ell) e^{-i\ell\theta}, \qquad \theta \in [-\pi, \pi],$$

whose upper extreme is given by:

$$\mathcal{M}(f_{X,Z}) = \mathrm{esssup}_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\max}\left(f^*_{X,Z}(\theta) f_{X,Z}(\theta)\right)}.$$

Let $\mathbf{X}$ *and* $\mathbf{Z}$ *be data matrices with sample size* n*. Then, there exists a constant* $c > 0$*, such that for any* $u, v \in \mathbb{R}^p$ *with* $\|u\| \le 1$*,* $\|v\| \le 1$*, we have*

$$\mathbb{P}\left[\left|u^\top \left(\frac{\mathbf{X}'\mathbf{Z}}{T} - \mathrm{Cov}(X_t, Z_t)\right) v\right| > 2\pi \left(\mathcal{M}(f_X) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z})\right) \eta\right] \le 6 \exp\left(-cT \min\{\eta, \eta^2\}\right).$$

Proof. Let $\xi_t = \langle u, X_t\rangle$, $\eta_t = \langle v, Z_t\rangle$. Let $f_X(\theta), f_Z(\theta)$ denote the spectral density of $\{X_t\}$ and $\{Z_t\}$, respectively. Then, the spectral density of $\{\xi_t\}$ and $\{\eta_t\}$, respectively, is $f_\xi(\theta) = u' f_X(\theta) u$, $f_\eta(\theta) = v' f_Z(\theta) v$. Also, we note that $\mathcal{M}(f_\xi) \le \mathcal{M}(f_X)$, $\mathcal{M}(f_\eta) \le \mathcal{M}(f_Z)$. Then,

$$\begin{aligned}\frac{2}{T}\left[\sum_{t=0}^{T} \xi_t\eta_t - \mathrm{Cov}(\xi_t, \eta_t)\right] &= \left[\frac{1}{T}\sum_{t=0}^{T}(\xi_t + \eta_t)^2 - \mathrm{Var}(\xi_t + \eta_t)\right] \\ &\quad - \left[\frac{1}{T}\sum_{t=0}^{T}(\xi_t)^2 - \mathrm{Var}(\xi_t)\right] - \left[\frac{1}{T}\sum_{t=0}^{T}(\eta_t)^2 - \mathrm{Var}(\eta_t)\right].\end{aligned} \tag{B.18}$$

By Proposition 2.7 in [**?**],

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T}(\xi_t)^2 - \mathrm{Var}(\xi_t)\right| > 2\pi\mathcal{M}(f_X)\eta\right) \ge 2\exp\left[-cn\min(\eta, \eta^2)\right],$$

and

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T}(\eta_t)^2 - \mathrm{Var}(\eta_t)\right| > 2\pi\mathcal{M}(f_Z)\eta\right) \ge 2\exp\left[-cn\min(\eta, \eta^2)\right].$$

What remains to be considered is the first term in (B.18), whose spectral density is given by

$$f_{\xi+\eta}(\theta) = u^\top f_X(\theta) u + v^\top f_Z(\theta) z + u^\top f_{X,Z}(\theta) v + v^\top f^*_{X,Z}(\theta) u,$$

and its upper extreme satisfies

$$\mathcal{M}(f_{\xi+\eta}) \le \mathcal{M}(f_X) + \mathcal{M}(f_Z) + 2\mathcal{M}(f_{X,Z}).$$

Hence, we get:

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T}(\xi_t+\eta_t)^2-\mathrm{Var}(\xi_t+\eta_t)\right|>2\pi[\mathcal{M}(f_X)+\mathcal{M}(f_Z)+2\mathcal{M}(f_{X,Z})]\eta\right)\geq 2\exp[-cn\min(\eta,\eta^2)].$$

Combining all three terms yields the desired result. □

Lemma B.7. *Let N and M be matrices of the same dimension. Then, there exists a decomposition $M = M_1 + M_2$, such that*

(a) rank$(M_1) \leq$ 2rank(N);

(b) $\langle\!\langle M_1, M_2 \rangle\!\rangle = 0$.

Proof. Let the SVD of N be given by $N = U\Sigma V'$, where both U and V are orthogonal matrices and assume $\mathrm{rank}(N) = r$. Define $\widetilde{M}$ and do partition it as follows:

$$\widetilde{M} = U^\top M V = \begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{21} & \widetilde{M}_{22} \end{bmatrix}.$$

Next, let

$$M_1 = U\begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{21} & O \end{bmatrix}V^\top \qquad \text{and} \qquad M_2 = U\begin{bmatrix} O & O \\ O & \widetilde{M}_{22} \end{bmatrix}V^\top, \qquad \widetilde{M}_{11}\in\mathbb{R}^{r\times r}.$$

Then, $M_1 + M_2 = M$ and

$$\mathrm{rank}(M_1)\leq \mathrm{rank}\left(U\begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ O & O \end{bmatrix}V^\top\right)+\mathrm{rank}\left(U\begin{bmatrix} \widetilde{M}_{11} & O \\ \widetilde{M}_{21} & O \end{bmatrix}V^\top\right)\leq 2r.$$

Moreover,

$$\langle\!\langle M_1, M_2 \rangle\!\rangle = \mathrm{tr}\left[M_1 M_2^\top\right] = 0.$$

□

Lemma B.8. *Define the error matrix by $\Delta^B = \widehat{B} - B^\star$ and $\Delta^C = \widehat{C} - C^\star$, and let the weighted regularizer $\mathcal{Q}$ be defined as*

$$\mathcal{Q}(B,C) = |\!|\!|B|\!|\!|_* + \frac{\lambda_C}{\lambda_B}\|C\|_1.$$

*With the subspaces defined in (*B.*6) and (*B.*7), the following inequality holds:*

$$\mathcal{Q}(B^\star,C^\star)-\mathcal{Q}(\widehat{B},\widehat{C})\leq \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}},\Delta^C_{\mathcal{J}_{C^\star}})-\mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}},\Delta^C_{\mathcal{J}^c_{C^\star}}).$$

Proof. First, from definitions (B.6) and (B.7), we know that $B^\star_{\mathcal{S}^\perp} = 0$ and $C^\star_{\mathcal{J}^c_{C^\star}} = 0$. Using the definition of $\mathcal{Q}$, we obtain

$$\mathcal{Q}(B^\star, C^\star) = |||B^\star_{\mathcal{S}} + B^\star_{\mathcal{S}^\perp}|||_* + \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}_C} + C^\star_{\mathcal{J}^c_{C^\star}}\right\|_1 = |||B^\star_{\mathcal{S}}|||_* + \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}_C}\right\|_1,$$

and

$$\begin{aligned}
\mathcal{Q}(\widehat{B}, \widehat{C}) &= \mathcal{Q}(B^\star + \Delta^B, C^\star + \Delta^C) \\
&= |||B^\star_{\mathcal{S}} + \Delta^B_{\mathcal{S}^\perp_{B^\star}} + \Delta^B_{\mathcal{S}_{B^\star}} + B^\star_{\mathcal{S}^\perp}|||_* + \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}_C} + \Delta^C_{\mathcal{J}_{C^\star}} + C^\star_{\mathcal{J}^c_{C^\star}} + \Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1 \\
&\geq |||B^\star_{\mathcal{S}} + \Delta^B_{\mathcal{S}^\perp_{B^\star}}|||_* - |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\left(\left\|C^\star_{\mathcal{J}_C} + \Delta^C_{\mathcal{J}_{C^\star}}\right\|_1 + \left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1\right) \\
&\geq |||B^\star_{\mathcal{S}}|||_* + |||\Delta^B_{\mathcal{S}^\perp_{B^\star}}|||_* - |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\left(\left\|C^\star_{\mathcal{J}_C}\right\|_1 + \|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 - \left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1\right).
\end{aligned}$$

The decomposition of the first term comes from the construction of $\Delta^B_{\mathcal{S}^\perp_{B^\star}}$. It then follows that

$$\begin{aligned}
\mathcal{Q}(B^\star, C^\star) - \mathcal{Q}(\widehat{B}, \widehat{C}) &\leq \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}_C}\right\|_1 + |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* - |||\Delta^B_{\mathcal{S}^\perp_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\left(\|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 - \left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1 - \left\|C^\star_{\mathcal{J}_C}\right\|_1\right) \\
&= |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 - \left(|||\Delta^B_{\mathcal{S}^\perp_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1\right) \\
&= \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}).
\end{aligned}$$

□

Lemma B.9. *Under the conditions of Theorem B.2, the following bound holds:*

$$\frac{1}{T}|||\mathfrak{W}_0(\Delta^B_{aug} + \Delta^C_{aug})|||^2_F \geq \frac{\alpha_{RSC}}{2}(|||\Delta^B|||^2_F + |||\Delta^C|||^2_F) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B, \Delta^C).$$

Proof. This lemma directly follows from Lemma 2 in **?**], by setting $\Theta^\star = B^\star$, $\Gamma^\star = C^\star$, with the regularizer $\mathbf{R}(\cdot)$ being the element-wise ℓ_1 norm. Note that $\sigma_j(B^\star) = 0$ for $j = r+1, \cdots, \min\{p_1, p_2\}$ since $\mathrm{rank}(B) = r$. For our problem, it suffices to set $\mathbb{M}^\perp$ as $\mathcal{J}^c_{C^\star}$, and therefore $\|C^\star_{\mathcal{J}^c_{C^\star}}\|_1 = 0$. □

Lemma B.10. *Consider the two centered Gaussian processes $\{X_t\} \in \mathbb{R}^{p_1}$ and $\{Z_t\} \in \mathbb{R}^{p_2}$, and denote their cross covariance matrix by $\Gamma_{X,Z}(h) = (X_t, Z_{t+h}) = \mathbb{E}(X_t Z^\top_{t+h})$. Let $\mathbf{X}$ and $\mathbf{Z}$ denote the data matrix. There exist positive constants $c_i > 0$ such that whenever $T \gtrsim c_3(p_1 + p_2)$, with probability at least*

$$1 - c_1 \exp[-c_2(p_1 + p_2)],$$

the following bound holds:

$$\frac{1}{T}\|\|\mathbf{X}^\top\mathbf{Z}\|\|_{op} \leq \mathbb{Q}_{X,Z}\sqrt{\frac{p_1+p_2}{T}} + 4\|\|\Gamma_{X,Z}(0)\|\|_{op},$$

where

$$\mathbb{Q}_{X,Z} = c_0\left[\mathcal{M}(f_X) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z})\right].$$

Proof. The main structure of this proof follows from that of Lemma 3 in **?**], and here we focus on how to handle the temporal dependency present in our problem. Let $S^p = \{u \in \mathbb{R}^p | \|u\| = 1\}$ denote the p-dimensional unit sphere. The operator norm has the following variational representation form:

$$\frac{1}{T}\|\|\mathbf{X}^\top\mathbf{Z}\|\|_{\text{op}} = \frac{1}{n}\sup_{u\in S^{p_1}}\sup_{v\in S^{p_2}} u^\top\mathbf{X}^\top\mathbf{Z}v.$$

For positive scalars s_1 and s_2, define

$$\Psi(s_1,s_2) = \sup_{u\in s_1 S^{p_1}}\sup_{v\in s_2 S^{p_2}}\langle\mathbf{X}u,\mathbf{Z}v\rangle,$$

and the goal is to establish an upper bound for $\Psi(1,1)/T$. Let $\mathcal{A} = \{u^1,\cdots,u^A\}$ and $\mathbf{B} = \{v^1,\cdots,v^B\}$ denote the $1/4$ coverings of S^{p_1} and S^{p_2}, respectively. **?**] showed that

$$\Psi(1,1) \leq 4\max_{u^a\in\mathcal{A},v^b\in\mathbf{B}}\langle\mathbf{X}u^a,\mathbf{Z}v^b\rangle,$$

and by **?**] and **?**], there exists a $1/4$ covering of S^{p_1} and S^{p_2} with at most $A \leq 8^{p_1}$ and $B \leq 8^{p_2}$ elements, respectively. Consequently,

$$\mathbb{P}\left[\left|\frac{1}{T}\Psi(1,1)\right| \geq 4\delta\right] \leq 8^{p_1+p_2}\max_{u^a,v^b}\mathbb{P}\left[\frac{|(u^a)^\top\mathbf{X}\mathbf{Z}(v^b)|}{T} \geq \delta\right].$$

What remains to be bounded is

$$\frac{1}{T}u^\top\mathbf{X}^\top\mathbf{Z}v, \qquad \text{for an arbitrary fixed pair of } (u,v)\in S^{p_1}\times S^{p_2}.$$

By Lemma B.6, we have

$$\mathbb{P}\left[\left|u^\top\left(\frac{\mathbf{X}^\top\mathbf{Z}}{T}\right)v\right| > 2\pi\left(\mathcal{M}(f_X)+\mathcal{M}(f_Z)+\mathcal{M}(f_{X,Z})\right)\eta + \|\|\Gamma_{X,Z}(0)\|\|_{\text{op}}\right] \leq 6\exp\left(-cT\min\{\eta,\eta^2\}\right).$$

Therefore, we have

$$\mathbb{P}\Big\{|\frac{1}{T}\Psi(1,1)| \geq 8\pi\left(\mathcal{M}(f_X)+\mathcal{M}(f_Z)+\mathcal{M}(f_{X,Z})\right)\eta + 4\|\!|\Gamma_{X,Z}(0)|\!\|_{\text{op}}\Big\} \\ \leq 6\exp\left[(p_1+p_2)\log 8 - cT\min\{\eta,\eta^2\}\right].$$

With the specified choice of sample size T, the probability vanishes by choosing $\eta = c_0\sqrt{\frac{p_1+p_2}{T}}$, for c_0 large enough, and we yield the conclusion in Lemma B.10. □

Lemma B.11. *Consider some generic matrix $A \in \mathbb{R}^{m\times n}$ and let $\gamma = \{\gamma_1,\ldots,\gamma_p\}$ $(p<n)$ denote the set of column indices of interest. Then, the following inequalities hold*

$$\Lambda_{\min}(A^\top A) \leq \Lambda_{\min}(A_\gamma^\top A_\gamma) \leq \Lambda_{\max}(A_\gamma^\top A_\gamma) \leq \Lambda_{\max}(A^\top A).$$

Proof. Let

$$\mathcal{V} := \{v=(v_1,\cdots,v_n)\in\mathbb{R}^n | v^\top v = 1\}.$$

and

$$\mathcal{V}_\gamma := \{v=(v_1,\cdots,v_n)\in\mathbb{R}^n | v^\top v = 1 \text{ and } v_j = 0\ \forall j\notin\gamma\}.$$

It is obvious that $\mathcal{V}_\gamma \subseteq \mathcal{V}$. By the definition of eigenvalues through their *Rayleigh quotient* characterization,

$$\Lambda_{\min}(A_\gamma^\top A_\gamma) = \min_{u^\top u=1, u\in\mathbb{R}^p} u^\top(A_\gamma^\top A_\gamma)u = \min_{v^\top v=1, v\in\mathcal{V}_\gamma} v^\top(A^\top A)v \geq \min_{v^\top v=1, v\in\mathcal{V}} v'(A^\top A)v = \Lambda_{\min}(A^\top A).$$

Similarly,

$$\Lambda_{\max}(A_\gamma^\top A_\gamma) = \max_{u^\top u=1, u\in\mathbb{R}^p} u^\top(A_\gamma^\top A_\gamma)u = \max_{v^\top v=1, v\in\mathcal{V}_\gamma} v^\top(A^\top A)v \leq \max_{v^\top v=1, v\in\mathcal{V}} v^\top(A^\top A)v = \Lambda_{\max}(A^\top A).$$

□

Lemma B.12. *Let $\{X_t\}$ and $\{\varepsilon_t\}$ be two generic processes, where $\varepsilon_t = (U_t^\top, V_t^\top)^\top$. Suppose the spectral density of the joint process (X_t', ε_t') exists. Then, the following inequalities hold*

$$\mathfrak{m}(f_{X,V}) \geq \mathfrak{m}(f_{X,\varepsilon}), \qquad \mathcal{M}(f_{X,V}) \leq \mathcal{M}(f_{X,\varepsilon}).$$

Proof. By definition, the spectral density $f_{X,\varepsilon}(\theta)$ can be written as

$$f_{X,\varepsilon}(\theta) = \left(\frac{1}{2\pi}\right)\sum_{\ell=-\infty}^{\infty}\Gamma_{X,\varepsilon}(\ell)e^{-i\ell\theta} = \left(\frac{1}{2\pi}\right)\sum_{\ell=-\infty}^{\infty}\left[\mathbb{E}X_tU_{t+\ell}^\top,\ \mathbb{E}X_tV_{t+\ell}^\top\right]e^{-i\ell\theta} \\ = (f_{X,U}(\theta),\quad f_{X,V}(\theta)), \qquad \theta\in[-\pi,\pi].$$

It follows that

$$\mathcal{M}(f_{X,\varepsilon}) = \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\max}(H(\theta))},$$

where

$$H(\theta) = \begin{bmatrix} f^*_{X,U}(\theta) \\ f^*_{X,V}(\theta) \end{bmatrix} \begin{bmatrix} f_{X,U}(\theta) & f_{X,V}(\theta) \end{bmatrix} = \begin{bmatrix} f^*_{X,U}(\theta) f_{X,U}(\theta) & f^*_{X,U}(\theta) f_{X,V}(\theta) \\ f^*_{X,V}(\theta) f_{X,U}(\theta) & f^*_{X,V}(\theta) f_{X,V}(\theta) \end{bmatrix}.$$

Note that

$$\mathcal{M}(f_{X,V}) = \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\max}(f^*_{X,V}(\theta) f_{X,V}(\theta))}.$$

By Lemma B.11, $\forall\theta$, $\Lambda_{\min}(f^*_{X,V}(\theta) f_{X,V}(\theta)) \geq \Lambda_{\min}(H(\theta))$ and $\Lambda_{\max}(f^*_{X,V}(\theta) f_{X,V}(\theta)) \leq \Lambda_{\max}(H(\theta))$, hence

$$\mathfrak{m}(f_{X,V}) \geq \mathfrak{m}(f_{X,\varepsilon}), \qquad \mathcal{M}(f_{X,V}) \leq \mathcal{M}(f_{X,\varepsilon}).$$

□

B.4 Testing group Granger-causality under a sparse alternative.

In this section, we develop a testing procedure to test the null hypothesis against its sparse alternatives, that is, H_0: $B = 0$ vs. H_A: B is nonzero and sparse. Throughout, we impose assumptions on the sparsity level of B (to be specified later), and use the *higher criticism* framework [c.f. **? ? ?**] as the building block of the testing procedure.

Once again, we start with testing sparse alternatives in a simpler model setting

$$Y_t = \Pi X_t + \epsilon_t,$$

where $Y_t \in \mathbb{R}^{p_2}$, $X_t \in \mathbb{R}^{p_1}$, and $\epsilon_t \in \mathbb{R}^{p_2}$ with each component being independent and identically distributed (i.i.d) and also independent of X_t. We would like to test the null hypothesis $H_0 : \Pi = 0$. Written in a compact form, the model is given by

$$\mathbf{Y} = \mathbf{X}\Pi^\top + \mathbf{E}, \tag{B.19}$$

where $\mathbf{Y} \in \mathbb{R}^{T\times p_2}$, $\mathbf{X} \in \mathbb{R}^{T\times p_1}$, and $\mathbf{E}$ are both contemporaneously and temporally independent. The latter shares similarities to the setting in **?**], with the main difference being that here we have a multi-response $\mathbf{Y}$. By rewriting (B.19) using Kronecker products, we have

$$\operatorname{vec}(\mathbf{Y}) = (\mathrm{I}_{p_2} \otimes \mathbf{X}) \operatorname{vec}(\Pi^\top) + \operatorname{vec}(\mathbf{E}) \qquad \text{i.e.,} \qquad \mathcal{Y} = \mathcal{X}\operatorname{vec}(\Pi^\top) + \mathcal{E},$$

where $\mathcal{Y} = \text{vec}(\mathbf{Y}) \in \mathbb{R}^{Tp_2}, \mathcal{X} = \mathrm{I}_{p_2} \otimes \mathbf{X} \in \mathbb{R}^{Tp_2 \times p_1 p_2}$. Each coordinate in $\mathcal{E}$ is iid. In this form, using the higher criticism [**?** **?** **?**] with proper scaling, the test statistic is given by:

$$\text{HC}^*(\mathcal{X}, \mathcal{Y}) = \sup_{t>0} H(t, \mathcal{X}, \mathcal{Y}) := \sqrt{\frac{p_1 p_2}{2\bar{\Phi}(t)(1-2\bar{\Phi}(t))}}\Big[\frac{1}{p_1 p_2}\sum_{k=1}^{p_1 p_2} \mathbf{1}\left(\frac{\sqrt{T}\cdot|\mathcal{X}_k^\top \mathcal{Y}|}{\|\mathcal{X}_k\|_2 \|\mathcal{Y}\|_2} > t\right) - 2\bar{\Phi}(t)\Big], \quad \text{(B.20)}$$

where $\mathcal{X}_k$ is the k^{th} column of $\mathcal{X}$ and $\bar{\Phi}(t) = 1 - \Phi(t)$ with $\Phi(t)$ being the cumulative distribution function of a standard Normal random variable. Intuitively,

$$\big(\tfrac{1}{p_1 p_2}\big)\sum_{k=1}^{p_1 p_2} \mathbf{1}\big\{\sqrt{T}\mathcal{X}_k^\top \mathcal{Y}/(\|\mathcal{X}_k\|_2 \|\mathcal{Y}\|_2) > t\big\}$$

is the fraction of significance beyond a given level t, after scaling for the vector length and the noise level. To conduct a level α test, H_0 is rejected when $\text{HC}^*(\mathcal{X}, \mathcal{Y}) > h(p_1 p_2, \alpha_{p_1 p_2})$ where $h(p_1 p_2, \alpha_{p_1 p_2}) \approx \sqrt{2\log\log(p_1 p_2)}$, provided that $\alpha_{p_1 p_2} \to 0$ slowly enough in the sense that $h(p_1 p_2, \alpha_{p_1 p_2}) = 2\sqrt{\log\log(p_1 p_2)}(1 + o(1))$ [see **?**]. The effectiveness of the test relies on a number of assumptions on the design matrix and the sparse vector to be tested. Next, we introduce the three most relevant definitions for subsequent developments, originally mentioned in **?**].

Definition B.1 (Bayes risk)**.** Following **?**], the Bayes risk of a test $\mathcal{T}$ for testing $\text{vec}(\Pi^\top) = 0$ vs. $\text{vec}(\Pi^\top) \sim \pi$, when H_0 and H_1 occur with the same probability, is defined as the sum of type I error and its average probability of type II error; i.e.,

$$\text{Risk}_\pi(\mathcal{T}) = \mathbb{P}_0(\mathcal{T} = 1) + \pi[\mathbb{P}_{\text{vec}}(\Pi^\top)(\mathcal{T} = 0)],$$

where π is a prior on the set of alternatives Ω. When no prior is specified, the risk is defined as the worst-case risk:

$$\text{Risk}(\mathcal{T}) = \mathbb{P}_0(\mathcal{T} = 1) + \max_{\text{vec}(\Pi^\top)\in\Omega}[\mathbb{P}_{\text{vec}}(\Pi^\top)(\mathcal{T} = 0)].$$

Definition B.2 (Asymptotically powerful)**.** We use $\mathcal{T}_{n,p}$ to denote the dependency of the test on the sample size n and the parameter dimension p. With $p \to \infty$ and $n = n(p) \to \infty$, a sequence of tests $\{\mathcal{T}_{n,p}\}$ is said to be *asymptotically powerful* if

$$\lim_{p\to\infty} \text{Risk}(\mathcal{T}_{n,p}) = 0.$$

Definition B.3 (Weakly correlated)**.** Let $\mathcal{S}_p(\gamma, \Delta)$ denote the set of $p \times p$ correlation matrices

$C = [c_{jk}]$ satisfying the weakly correlated assumption: for all $j = 1, \ldots, p$,

$$|c_{jk}| < 1 - (\log p)^{-1} \qquad \text{and} \qquad \{k : |c_{jk}| > \gamma\} \leq \Delta, \qquad \text{for some } \gamma \leq 1, \Delta \geq 1.$$

With the above definitions, **?**] establishes that using the test based on higher criticism is asymptotically powerful, provided that (1) $\text{vec}(\Pi^\top)$ satisfies the *strong sparsity assumption*, that is, the total number of nonzeros $s^\star_{\text{vec}(\Pi^\top)} = (p_1 p_2)^\theta$ with $\theta \in (1/2, 1)$; (2) the correlation matrix of $\mathcal{X}$ belongs to $\mathcal{S}(\gamma, \Delta)$ with γ and Δ satisfying certain assumptions in terms of their relative order with respect to parameter dimension and sample size; and (3) the minimum magnitude of the nonzero elements of $\text{vec}(\Pi^\top)$ exceeds a certain lower detection threshold.

Switching to our model setting in which

$$Z_t = BX_{t-1} + CZ_{t-1} + V_t, \qquad B \in \mathbb{R}^{p_2 \times p_1},$$

where B encodes the dependency between Z_t and X_{t-1}, conditional on Z_{t-1}, the above discussion suggests that we can use higher criticism on the residuals $\mathbf{R}_1$ and $\mathbf{R}_0$, where $\mathbf{R}_1$ and $\mathbf{R}_0$ are identically defined to those in the low-rank testing; that is, $\mathbf{R}_1$ is the residual after regressing $\mathbf{X}$ on $\mathbf{Z}$, and $\mathbf{R}_0$ is the residual after regressing $\mathbf{Z}_T$ on $\mathbf{Z}$:

$$\mathbf{R}_1 = (I - P_z)\mathbf{X} \qquad \text{and} \qquad \mathbf{R}_0 = (I - P_z)\mathbf{Z}_T,$$

where $P_z = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top$. Writing the model in terms of $\mathbf{R}_1$ and $\mathbf{R}_0$, we have

$$\mathbf{R}_0 = \mathbf{R}_1 B^\top + \mathbf{V}, \qquad \text{i.e.,} \qquad \mathcal{R}_0 = \mathcal{R}_1 \beta_B + \mathcal{V},$$

where $\mathcal{R}_0 = \text{vec}(\mathbf{R}_0), \mathcal{R}_1 = \mathrm{I} \otimes \mathbf{R}_1$, $\mathcal{V} = \text{vec}(\mathbf{V})$, and $\beta_B = \text{vec}(B^\top) \in \mathbb{R}^{p_1 p_2}$. To test $H_0 : \beta_B = 0$, the higher criticism is given by

$$\begin{aligned} \text{HC}^*(\mathcal{R}_1, \mathcal{R}_0) &= \sup_{t>0} H(t, \mathcal{R}_1, \mathcal{R}_0) := \sqrt{\frac{p_1 p_2}{2\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}} \Big[\frac{1}{p_1 p_2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_1} \mathbf{1}\Big(\frac{\sqrt{T}|\mathcal{R}_{1k}^\top \mathcal{R}_0|}{\|\mathcal{R}_{1k}\|_2 \|\mathcal{R}_0\|_2} > t \Big) - 2\bar{\Phi}(t) \Big] \\ &= \sup_{t>0} \sqrt{\frac{p_1 p_2}{2\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}} \Big[\frac{1}{p_1 p_2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_1} \mathbf{1}\Big(\frac{\sqrt{T}|S_{10,ij}|}{\sqrt{S_{11,ii} S_{00,jj}}} > t \Big) - 2\bar{\Phi}(t) \Big] \end{aligned} \tag{B.21}$$

where $S_{10} = \mathbf{R}_1^\top \mathbf{R}_0 / T$, $S_{11} = \mathbf{R}_1^\top \mathbf{R}_1 / T$ and $S_{00} = \mathbf{R}_0^\top \mathbf{R}_0 / T$. The second equality is due to the block-diagonal structure of $\mathcal{R}_1$. We reject the null hypothesis if

$$\text{HC}^*(\mathcal{R}_1, \mathcal{R}_0) > 2\sqrt{\log \log (p_1 p_2)}.$$

Empirically, t can be chosen from $\{[1, \sqrt{5\log(p_1p_2)}] \cap \mathbb{N}\}$ [**?**].

Next, we analyze the theoretical properties of the above testing procedure. If the parameter dimension is fixed, then classical consistency results in terms of convergence (in probability or almost surely) hold when letting $T \to \infty$, and everything follows trivially, as long as the corresponding population quantities satisfy the posited assumptions.

In the remainder, we allow the parameter dimension p_1p_2 to slowly vary with the sample size T. Let $S_{\mathcal{R}_1} = \mathcal{R}_1^\top \mathcal{R}_1/T$ be the sample covariance matrix based on the residuals $\mathcal{R}_1$, and let $C_{\mathcal{R}_1}$ be the corresponding correlation matrix. The following proposition directly follows from Theorem 4 in **?**].

Proposition B.1 (An asymptotically powerful test)**.** *Under the following conditions, the testing procedure associated with the Higher Criticism statistics defined in* (B.21) *is asymptotically powerful, provided that the smallest magnitude of nonzero entries of $B^\star$ exceeds the lower detection boundary.*[2]

(a) Strong sparsity: let $p_B = p_1p_2$ be the dimension of $\beta_B^\star$, then the total number of nonzeros satisfies $s_B^\star = p_B^\theta$, where $\theta \in (1/2, 1)$.

(b) Weakly correlated design: $C_{\mathcal{R}_1} \in \mathcal{S}(\gamma, \Delta)$ with the parameters satisfying $\Delta = O(p_1^\epsilon)$, $\gamma = O(p_1^{-1/2+\epsilon})$, $\forall\ \epsilon > 0$.

Note that $S_{\mathcal{R}_1} = \mathrm{I} \otimes S_{11}$, where $S_{11} = \mathbf{R}_1^\top \mathbf{R}_1/T$; hence, $C_{\mathcal{R}_1} = \mathrm{I} \otimes C_{11}$, with C_{11} being the sample correlation matrix based on $\mathbf{R}_1$. The weakly correlated design assumption is thus effectively imposed on C_{11}, with the parameters γ and Δ satisfying the same condition. The weakly correlated design assumption on C_{11} in Proposition B.1 is for a deterministic realization of $\mathbf{R}_1$. The following corollary states that for a random realization of $\mathbf{R}_1$, obtained by regressing a random $\mathbf{X}$ on $\mathbf{Z}$, to satisfy the weakly correlated design assumption with high probability, it is sufficient that the population counterparts of the associated quantities satisfy the required assumptions.

Corollary B.1. *Consider residual $\mathbf{R}_1$ obtained by regressing a random realization of $\mathbf{X}$ on that of $\mathbf{Z}$. Let $\Sigma_{11} := \Gamma_X - \Gamma_{X,Z}\Gamma_Z^{-1}\Gamma_{X,Z}^\top$ be the covariance of X_t conditional on Z_t, and ρ_{11} be the corresponding correlation matrix. Suppose $\rho_{11} \in S(\gamma, \Delta)$ with γ and Δ satisfying the same condition as in Proposition B.1. Then with high probability, the sample correlation matrix based on $\mathbf{R}_1$ belongs to $\mathcal{S}(\gamma', \Delta')$, where γ' and Δ' respectively satisfy the same condition as γ and Δ, provided that the same condition imposed on γ holds for $\sqrt{T^{-1}\log(p_1p_2)}$. Moreover, the conclusion in Proposition B.1 holds.*

[2]For a thorough discussion on the lower detection boundary, we refer the reader to **? ?**] and references therein.

Remark B.1. In the work of **?**] and **?**], the authors focus their analysis primarily on the multiple regression setting, where the regression coefficient matrix directly encodes the relationship between the response variable and the covariates, in an iid data setting. We consider a more complicated model setting in which the regression coefficient matrix of interest encodes the partial auto-correlations between a multivariate response and a set of exogenous variables, while the data exhibit temporal dependence. It is worth pointing out that with the presence of temporal dependence, the rate with respect to the model dimension p and sample size T stays the same, as in the case where the data are iid [e.g., **?** **?**]; specifically, it is $\sqrt{\log p/T}$ in terms of the element-wise infinity norm, whereas the associated constant is a function of the lower and upper extremes of the spectral density, which intricately controls the exact coverage and power of the testing procedures. Therefore, as long as the rate constraint on p and T is satisfied (as in Corollary B.1), the main conclusion is compatible with previous work, and asymptotically, we either obtain the distribution of the test statistic (low rank testing), or have a powerful test (sparse testing).

Remark B.2. To solve the global testing problem for the sparse setting, a possible alternative is to construct a test statistic based on estimates of the regression coefficients, then perform a global or max test on the estimated coefficients. A key issue for such a test is that the estimated entries of B are biased due to the use of Lasso; therefore, a debiasing procedure [e.g. **?**] would be required to obtain valid marginal distributions for the entries of the B matrix. In contrast, the higher criticism test statistic is based on the correlation between the response and the covariates (see Equation (B.20)), and here we employ the idea on the residuals so that the effect of Z_t block is removed. We do not directly deal with the estimates of the B matrix and thus avoid the complications induced by the potentially biased estimates of B.

B.5 Estimation and Consistency for an Alternative Model Specification.

In this section, we consider the finite-sample error bound for the case where both B and C are sparse. We assume the presence of a sparse contemporaneous conditional dependence, hence the alternate between the estimation of transition matrices and that of the covariance matrix is required. In what follows, we briefly outline the estimation procedure and the error bounds of the estimates. All notations follow from those in Section 3.3.

The joint optimization problem is given by

$$(\widehat{B},\widehat{C},\widehat{\Omega}_v) = \underset{B,C,\Omega_v}{\arg\min} \Big\{ \mathrm{tr}\big[\Omega_v(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)^\top(\mathbf{Z}_T - \mathbf{X}B^\top - \mathbf{Z}C^\top)/T\big] - \log\det\Omega_v + \lambda_B\|B\|_1 + \lambda_C\|C\|_1 + \rho_v\|\Omega_v\|_{1,\mathrm{off}}\Big\}. \tag{B.22}$$

For every iteration, with a fixed $\widehat{\Omega}_v^{(k)}$, $\widehat{B}^{(k+1)}$ and $\widehat{C}^{(k+1)}$ are both updated via Lasso; for fixed $(\widehat{B}^{(k)},\widehat{C}^{(k)})$, $\widehat{\Omega}_v^{(k)}$ is updated by the graphical Lasso.

Corollary B.2. *Consider the stable Gaussian VAR system defined in* (3.1) *in which $B^\star$ is assumed to be $s_B^\star$-sparse and $C^\star$ is assumed to be $s_C^\star$-sparse. Further, assume the following*

C1. The incoherence condition holds for $\Omega_v^\star$.

C2. $\Omega_v^\star$ is diagonally dominant.

C3. The maximum node degree of $\Omega_v^\star$ satisfies $d^{\max}_{\Omega_v^\star} = o(p_2)$.

Then, for random realizations of $\{X_t\}$, $\{Z_t\}$ and $\{V_t\}$, and the sequence $\{(\widehat{B}^{(k)},\widehat{C}^{(k)}),\widehat{\Omega}_v^{(k)}\}_k$ returned by Algorithm III.2 outlined in Section 3.2.1, with high probability, the following bounds hold for all iterations k for sufficiently large sample size T:

$$|||\widehat{B}^{(k)} - B^\star|||_F^2 + |||\widehat{C}^{(k)} - C^\star|||_F^2 = O\Big(\frac{(s_B^\star + s_C^\star)\big(\log(p_1+p_2) + \log p_2\big)}{T}\Big),$$

and

$$|||\widehat{\Omega}_v^{(k)} - \Omega_v^\star|||_F^2 = O\Big(\frac{(s_{\Omega_v}^\star + p_2)(\log(p_1+p_2) + \log p_2)}{T}\Big).$$

Note that when no contemporaneous dependence is present, $(\widehat{B},\widehat{C})$ solves

$$(\widehat{B},\widehat{C}) = \underset{(B,C)}{\arg\min} \Big\{ \tfrac{1}{T}|||\mathbf{Z}_T - \mathfrak{W}_0(B_{\mathrm{aug}} + C_{\mathrm{aug}})|||_{\mathrm{F}}^2 + \lambda_B\|B\|_1 + \lambda_C\|C\|_1\Big\}, \tag{B.23}$$

whose error bound is given by

$$|||\widehat{B} - B^\star|||_{\mathrm{F}}^2 + |||\widehat{C} - C^\star|||_{\mathrm{F}}^2 \le 4(\lambda_B^2 + \lambda_C^2)/\alpha_{\mathrm{RSC}}^2, \tag{B.24}$$

provided that the RSC condition holds and the regularization parameters are chosen properly. By setting the weighted regularizer as $\mathcal{Q}(B,C) = \|B\|_1 + \frac{\lambda_C}{\lambda_B}\|C\|_1$ and $\Delta^B := \widehat{B} - B^\star$ can be decomposed as (see equation (B.7))

$$\|\Delta^B_{\mathcal{J}_B} + \Delta^B_{\mathcal{J}_B^c}\|_1 = \|\Delta^B_{\mathcal{J}_B}\|_1 + \|\Delta^B_{\mathcal{J}_B^c}\|_1.$$

The rest of the proof is similar to that of Theorem B.2 hence is omitted here.

B.6 Proof of Propositions and Corollaries.

Proof of Proposition 3.1. The joint process $W_t = \{(X_t^\top, Z_t^\top)^\top\}$ is a stationary VAR(1) process, and it follows that

$$S_w(h) := \begin{bmatrix} S_x(h) & S_{x,z}(h) \\ S_{z,x}(h) & S_z(h) \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T} w_t w_{t+h}^\top \xrightarrow{p} \Gamma_W(h) := \mathbb{E} W_t W_{t+h}^\top, \quad \text{as } T \to \infty,$$

which implies

$$S_x \xrightarrow{p} \Gamma_X, \quad S_z \xrightarrow{p} \Gamma_Z, \quad S_{x,z} \xrightarrow{p} \Gamma_{X,Z}, \quad S_{x,z}(1) \xrightarrow{p} \Gamma_{X,Z}(1).$$

Note that sample partial regression residual covariances can be obtained by

$$S_{00} = S_z - S_z(1)S_z^{-1}S_z^\top(1), \qquad S_{11} = S_x - S_{x,z}S_z^{-1}S_{x,z}^\top, \qquad S_{10} = S_{x,z}(1) - S_z(1)S_z^{-1}S_{x,z}^\top.$$

An application of the Continuous Mapping Theorem yields

$$S_{00} \xrightarrow{p} \Sigma_{00}, \qquad S_{10} \xrightarrow{p} \Sigma_{10}, \qquad S_{11} \xrightarrow{p} \Sigma_{11}.$$

By **? ?**], the limiting behavior of $T\Psi_r$ is given by

$$T\Psi_r \sim \chi^2_{(p_1-r)(p_2-r)}, \qquad \text{as} \quad T \to \infty.$$

Note that since μ is of multiplicity one and the ordered eigenvalues are continuous functions of the matrices, the following holds:

$$\phi_k \xrightarrow{p} \mu_k, \qquad \forall\, k = 1, \ldots, \min(p_1, p_2).$$

□

Proof of Corollary B.1. First, we note that $\mathbf{R}_1$ effectively comes from the following stochastic regression:

$$X_t = QZ_t + R_t, \qquad \text{for some regression matrix } Q, \tag{B.25}$$

with $\mathbf{R}_1 = \mathbf{X} - \mathbf{Z}\widehat{Q}$ being the sample residual. The population covariance of R_t is the

conditional covariance of X_t on Z_t, given by

$$\Sigma_{11} = \Sigma_R^\star := \Gamma_X - \Gamma_{X,Z}\Gamma_Z^{-1}\Gamma_{X,Z}^\top.$$

Σ_{11} is identical to that defined in equation (3.25). Writing the model in terms of data, we have

$$\mathbf{X} = \mathbf{Z}Q + \mathbf{R}^\star,$$

where we use $\mathbf{R}^\star$ to denote the true error term, for the purpose of distinguishing it from the residuals by regressing $\mathbf{X}$ on $\mathbf{Z}$. Note that $\mathbf{R}^\star$ is also sub-Gaussian. First, we would like to obtain a bound for $\|S_{11} - \Sigma_{11}\|_\infty$. Let S_R be the sample covariance matrix based on the actual errors, i.e., $S_{R^\star} = (\mathbf{R}^\star)'(\mathbf{R}^\star)/T$, then

$$\|S_{11} - \Sigma_{11}\|_\infty \leq \|S_{R^\star} - \Sigma_{11}\|_\infty + \|S_{11} - S_{R^\star}\|_\infty$$

The first term can be directed bounded by Lemma 1 in **?**], that is, there exists some constant $\tau > 2$, such that for large enough sample size T, with probability at least $1 - 1/p_2^{\tau-2}$,

$$\|S_{R^\star} - \Sigma_{11}\|_\infty \leq C_0\sqrt{\log p_2/T}, \qquad \text{for some constant } C_0 > 0.$$

Consider the second term. Rewrite it as

$$S_{11} - S_{R^\star} = \tfrac{2}{T}(\mathbf{R}^\star)^\top\mathbf{Z}(Q^\star - \widehat{Q}) + (Q^\star - \widehat{Q})^\top\Big(\tfrac{\mathbf{Z}^\top\mathbf{Z}}{T}\Big)(Q^\star - \widehat{Q}) := I_1 + I_2,$$

then for I_1,

$$I_1 \leq 2\|\!|Q^\star - \widehat{Q}|\!\|_1\|\frac{1}{T}(\mathbf{R}^\star)^\top\mathbf{Z}\|_\infty \leq 2\Big\|\mathrm{vec}(Q^\star) - \mathrm{vec}(\widehat{Q})\Big\|_1\|\frac{1}{T}(\mathbf{R}^\star)^\top\mathbf{Z}\|_\infty.$$

By Lemma B.4, there exist constants $c_i > 0$ such that with probability at least $1 - c_1\exp(-c_2\log(p_1p_2))$, for sufficiently large sample size T, we get

$$\|\frac{1}{T}(\mathbf{R}^\star)^\top\mathbf{Z}\|_\infty \leq C_1\sqrt{\log(p_1p_2)/T}, \qquad \text{for some constant } C_1 > 0. \tag{B.26}$$

For I_2, we have that

$$I_2 \leq \|\!|Q^\star - \widehat{Q}|\!\|_1^2\Big\|\frac{\mathbf{Z}^\top\mathbf{Z}}{T}\Big\|_\infty \leq \Big\|\mathrm{vec}(Q^\star) - \mathrm{vec}(\widehat{Q})\Big\|_1^2\Big\|\frac{\mathbf{Z}^\top\mathbf{Z}}{T}\Big\|_\infty.$$

By Proposition 2.4 in **?**] and taking the union bound, there exist some constants c_1' and c_2'

such that with probability at least $1 - c_1' \exp(-c_2' \log p_2)$, for sufficiently large sample size T, we obtain

$$\left\| \frac{\mathbf{Z}^\top \mathbf{Z}}{T} - \Gamma_Z \right\|_\infty \leq C_2 \sqrt{\log p_2 / T}, \qquad \text{for some constant } C_2 > 0,$$

which implies

$$\left\| \frac{\mathbf{Z}^\top \mathbf{Z}}{T} \right\|_\infty \leq C_2 \sqrt{\log p_2 / n} + \max_i (\Gamma_{Z,ii}). \tag{B.27}$$

By assuming that $\left\| \mathrm{vec}(Q^\star) - \mathrm{vec}(\widehat{Q}) \right\|_1 \leq \varepsilon_Q$, it follows that

$$I_1 + I_2 \leq C_1' \varepsilon_Q \sqrt{\frac{\log(p_1 p_2)}{T}} + C_2' \varepsilon_Q^2 \sqrt{\frac{\log p_2}{T}},$$

hence

$$\|S_{11} - \Sigma_{11}\|_\infty \leq C_0 \sqrt{\frac{\log p_2}{T}} + C_1' \varepsilon_Q \sqrt{\frac{\log(p_1 p_2)}{T}} + C_2' \varepsilon_Q^2 \sqrt{\frac{\log p_2}{T}}. \tag{B.28}$$

Regardless of the relative order of p_1 and p_2, one can easily verify that

$$\|S_{11} - \Sigma_{11}\|_\infty = O\Big(\sqrt{\tfrac{\log(p_1 p_2)}{T}} \Big). \tag{B.29}$$

by assuming $\log(p_1 p_2)/T$ being a small quantity. Since

$$C_{11} = (\mathrm{diag}(S_{11}))^{-1/2} \, S_{11} \, (\mathrm{diag}(S_{11}))^{-1/2}$$

and letting $\widetilde{R}_t = \mathrm{diag}(\Sigma_{11})^{-1/2} R_t$, we then have that C_{11} is simply the sample covariance matrix based on residual surrogates of $\widetilde{R}_t$, whose error rate stays unchanged by scaling, i.e, $\|C_{11} - \rho_{11}\|_\infty = O(\sqrt{T^{-1} \log(p_1 p_2)})$. The latter fact further implies that if $\rho_{11} \in \mathcal{S}(\gamma, \Delta)$, then $C_{11} \in \mathcal{S}(\gamma', \Delta')$ with $\Delta' \geq \Delta - (\mathrm{const}) \sqrt{\log(p_1 p_2)/T}$ and $\gamma' \geq \gamma + (\mathrm{const}) \sqrt{\log(p_1 p_2)/T}$.

It then follows that as long as $\sqrt{T^{-1} \log(p_1 p_2)}$ satisfies the same condition imposed on $\gamma = O(p_1^{-1/2+\epsilon})$, that is,

$$p_1^{1-2\epsilon} \log(p_1 p_2) = O(T), \qquad \text{for all } \epsilon > 0,$$

with high probability, the sample covariance matrix based on the residuals $\mathbf{R}_1$ satisfies the weakly correlated design assumption, for a random realization $\mathbf{X}$ and $\mathbf{Z}$. □

Finally, we briefly outline the main steps of how to obtain the error bound in Corollary B.2.

Proof of Corollary B.2. Let $W_t = (X_t^\top, Z_t^\top)^\top$. At iteration 0, $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ solves (B.23), and

the following bound holds:

$$\|\!|\widehat{B}^{(0)} - B^\star|\!\|_{\mathrm{F}}^2 + \|\!|\widehat{C}^{(0)} - C^\star|\!\|_{\mathrm{F}}^2 \le 4(\lambda_B^2 + \lambda_C^2)/\alpha_{\mathrm{RSC}}^2,$$

provided that $\mathfrak{W}$ satisfies the RSC condition, and λ_B, λ_C both satisfy

$$\lambda_B \ge 4\|\mathbf{W}^\top \mathbf{V}/T\|_\infty, \qquad \lambda_C \ge 4\|\mathbf{W}^\top \mathbf{V}/T\|_\infty.$$

In particular, by Lemma B.3 and Lemma B.4, for random realizations of $\{X_t\}$, $\{Z_t\}$ and $\{V_t\}$, for sufficiently large sample size, with high probability

$$\mathfrak{W} \text{ satisfies the RSC condition,}$$

and

$$\big\|\mathbf{W}^\top \mathbf{V}/T\big\|_\infty \le C_1 \sqrt{\frac{\log(p_1+p_2) + \log p_2}{T}},$$

for some constant C_1. Hence, with high probability,

$$\|\!|\widehat{B}^{(0)} - B^\star|\!\|_{\mathrm{F}}^2 + \|\!|\widehat{C}^{(0)} - C^\star|\!\|_{\mathrm{F}}^2 = O\Big(\tfrac{\log(p_1+p_2)+\log p_2}{T}\Big). \tag{B.30}$$

For $\widehat{\Omega}_v^{(0)}$, it solves a graphical Lasso problem:

$$\widehat{\Omega}_v^{(0)} = \mathop{\arg\min}_{\Omega_v \in \mathbb{S}_{++}^{p_2 \times p_2}} \Big\{ \log\det\Omega_v - \mathrm{trace}\big(\widehat{S}_u^{(0)}\Omega_v\big) + \rho_v \|\Omega_v\|_{1,\mathrm{off}} \Big\},$$

where $\widehat{S}_v^{(0)} = \frac{1}{T}(\mathbf{Z}_T - \mathbf{X}\widehat{B}^{(0)\top} - \mathbf{Z}\widehat{C}^{(0)\top})^\top(\mathbf{Z}_T - \mathbf{X}\widehat{B}^{(0)\top} - \mathbf{Z}\widehat{C}^{(0)\top})$. Similar to the proof of Theorem 3.2, its error bound depends on $\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty$. With the same decomposition and consider only the leading term,

$$\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty = O\Big(\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}}\Big), \qquad \Rightarrow \quad \|\widehat{\Omega}_v^{(0)} - \Omega_v^\star\|_\infty = O\Big(\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}}\Big).$$

At iteration 1, the bound of $\|\!|\widehat{B}^{(1)} - B^\star|\!\|_{\mathrm{F}}^2 + \|\!|\widehat{C}^{(1)} - C^\star|\!\|_{\mathrm{F}}^2$ relies on

$$\begin{aligned}
\|\frac{1}{T}\mathbf{W}^\top \mathbf{V}\widehat{\Omega}_v^{(0)}\|_\infty &\le \|\tfrac{1}{T}\mathbf{W}^\top \mathbf{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\|_\infty + \|\tfrac{1}{T}\mathbf{W}^\top \mathbf{V}\Omega_v^\star\|_\infty, \\
&\le C_2\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}} + d_{\max}^{\Omega_v^\star}\|\tfrac{1}{T}\mathbf{W}^\top \mathbf{V}\|_\infty \|\widehat{\Omega}_v^{(0)} - \Omega_v^\star\|_\infty \\
&= O\Big(\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}}\Big),
\end{aligned} \tag{B.31}$$

hence $\|\|\widehat{B}^{(1)} - B^\star\|\|_{\mathrm{F}}^2 + \|\|\widehat{C}^{(1)} - C^\star\|\|_{\mathrm{F}}^2 = O(\frac{\log(p_1+p_2)+\log p_2}{T})$, which coincides with the bound of the estimator of iteration 0, implying the error rate remains unchanged henceforth. Up to this step, all sources of randomness have been captured. Consequently, the following bounds hold with high probability for all iterations k:

$$\|\mathbf{W}^\top \mathbf{V} \widehat{\Omega}_v^{(k)}/T\|_\infty = O\Big(\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}}\Big),$$

and

$$\|\widehat{S}_v^{(k)} - \Sigma_v^\star\|_\infty = O\Big(\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}}\Big),$$

which imply the bounds in Corollary B.2. □

B.7 List of Stock and Macroeconomic Variables

Stock Symbol	Name	Stock Symbol	Company Name
AAPL	Apple Inc.	JNJ	Johnson & Johnson Inc
AIG	American International Group Inc.	JPM	JP Morgan Chase & Co
ALL	Allstate Corp.	KO	The Coca-Cola Company
AMGN	Amgen Inc.	LMT	Lockheed-Martin
AXP	American Express Inc.	LOW	Lowe's
BA	Boeing Co.	MCD	McDonald's Corp
BAC	Bank of America Corp	MDLZ	Mondel International
BK	Bank of New York Mellon Corp	MDT	Medtronic Inc.
BMY	Bristol-Myers Squibb	MMM	3M Company
C	Citigroup Inc	MO	Altria Group
CAT	Caterpillar Inc	MRK	Merck & Co.
CL	Colgate-Palmolive Co.	MS	Morgan Stanley
CMCSA	Comcast Corporation	MSFT	Microsoft
COF	Capital One Financial Corp.	NSC	Norfolk Southern Corp
COP	ConocoPhillips	ORCL	Oracle Corporation
CSCO	Cisco Systems	OXY	Occidental Petroleum Corp.
CVS	CVS Caremark	PEP	Pepsico Inc.
CVX	Chevron	PFE	Pfizer Inc
DD	DuPont	PG	Procter & Gamble Co
DIS	The Walt Disney Company	RTN	Raytheon Company
DOW	Dow Chemical	SLB	Schlumberger
DVN	Devon Energy Corp	SO	Southern Company
EMC	EMC Corporation	T	AT&T Inc
EXC	Exelon	TGT	Target Corp.
F	Ford Motor	TWX	Time Warner Inc.
FCX	Freeport-McMoran	TXN	Texas Instruments
FDX	FedEx	UNH	UnitedHealth Group Inc.
GD	General Dynamics	UPS	United Parcel Service Inc
GE	General Electric Co.	USB	US Bancorp
GILD	Gilead Sciences	UTX	United Technologies Corp
GS	Goldman Sachs	VZ	Verizon Communications Inc
HAL	Halliburton	WBA	Walgreens Boots Alliance
HD	Home Depot	WFC	Wells Fargo
HON	Honeywell	WMT	Wal-Mart
IBM	International Business Machines	XOM	Exxon Mobil Corp
INTC	Intel Corporation		

Table B.1: List of stocks used in the analysis.

Symbol	Description	Transformation
FFR	Federal Funds Rate	abs diff
T10yr	10-Year Treasury Yield with Constant Maturity	abs diff
UNEMPL	Unemployment Rate for 16 and above	abs diff
IPI	Industrial Production Index	relative diff
ETTL	Employment Total	relative diff
M1	M1 Money Stock	relative diff
AHES	Average Hourly Earnings of Production and Nonsupervisory Employees	relative diff
CU	Capital Utilization	relative diff
M2	M2 Money Stock	relative diff
HS	Housing starts	relative diff
EX	US Exchange Rate	abs diff
PCEQI	Personal Consumption Expenditures Quantity Index	relative diff
GDP	real Gross Domestic Product	relative diff
PCEPI	Personal Consumption Expenditures Price Index	relative diff
PPI	Producer Price Index	relative diff
CPI	Consumer Price Index	relative diff
SP.IND	S&P Industrial Sector index	relative diff

*abs diff: $x_t - x_{t-1}$, relative diff: $\frac{x_t - x_{t-1}}{x_{t-1}}$

Table B.2: List of macroeconomic variables and the transformation used in the analysis.

APPENDIX C

Supplementary Materials to "Regularized Estimation of High-dimensional Factor-Augmented Vector Autoregressive (FAVAR) Models."

C.1 Proofs for Theorems and Propositions.

This section is divided into two parts. In the first part, we provide proofs for the proposition and theorem related to Stage I estimates, i.e., $\widehat{\Theta}$ and $\widehat{\Gamma}$. In the second part, we give proofs for the statements related to Stage II estimates, namely $\widehat{A}$, with an emphasis on how to obtain the final high probability error bound through properly conditioning on related events.

Part 1. Proofs for the $\widehat{\Theta}$ and $\widehat{\Gamma}$ estimates.

Proof of Proposition 4.1. Using the optimality of $(\widehat{\Gamma}, \widehat{\Theta})$ and the feasibility of $(\Gamma^\star, \Theta^\star)$, the following *basic inequality* holds:

$$\frac{1}{2n}|||\mathbf{X}\Delta_\Gamma^\top + \Delta_\Theta|||_{\mathrm{F}}^2 \leq \frac{1}{n}\Big(\langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) + \lambda_\Gamma\Big(\|\Gamma^\star\|_1 - \|\widehat{\Gamma}\|_1\Big), \tag{C.1}$$

which after rearranging terms gives

$$\begin{aligned}\frac{1}{2n}|||\mathbf{X}\Delta_\Gamma^\top|||_{\mathrm{F}}^2 + \frac{1}{2}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 \leq \frac{1}{n}\langle\!\langle \mathbf{X}\Delta_\Gamma^\top, \widehat{\Theta} - \Theta^\star\rangle\!\rangle + \frac{1}{n}\Big(\langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) \\ +\lambda_\Gamma\Big(\|\Gamma^\star\|_1 - \|\widehat{\Gamma}\|_1\Big).\end{aligned} \tag{C.2}$$

The remainder of the proof proceeds in three steps: in Step (i), we obtain a lower bound for the left-hand side (LHS) leveraging the RSC condition; in Step (ii), an upper bound for the right hand side (RHS) based on the designated choice of λ_Γ is derived; in Step (iii), the two sides are aligned to yiled the desired error bound after rearranging terms.

To complete the proof, we first define a few quantities that are associated with the support set of Γ and its complement:

$$\mathbb{S} := \left\{\Delta \in \mathbb{R}^{q\times p_2} | \Delta_{ij} = 0 \text{ for } (i,j) \notin S_{\Gamma^\star}\right\},$$
$$\mathbb{S}^c := \left\{\Delta \in \mathbb{R}^{q\times p_2} | \Delta_{ij} = 0 \text{ for } (i,j) \in S_{\Gamma^\star}\right\},$$

where $S_{\Gamma^\star}$ is the support of $\Gamma^\star$. Further, define $\Delta_{\mathbb{S}}$ and $\Delta_{\mathbb{S}^c}$ as

$$\Delta_{\mathbb{S},ij} = 1\{(i,j) \in S_{\Gamma^\star}\}\Delta_{ij}, \qquad \Delta_{\mathbb{S}^c,ij} = 1\{(i,j) \in S^c_{\Gamma^\star}\}\Delta_{ij},$$

and note that they satisfy

$$\Delta = \Delta_{\mathbb{S}} + \Delta_{\mathbb{S}^c}, \qquad \|\Delta\|_1 = \|\Delta_{\mathbb{S}}\|_1 + \|\Delta_{\mathbb{S}^c}\|_1$$

and

$$\|\Delta_{\mathbb{S}}\|_1 \le \sqrt{s}|||\Delta_{\mathbb{S}}|||_{\mathrm{F}} \le \sqrt{s_{\Gamma^\star}}|||\Delta|||_{\mathrm{F}}. \tag{C.3}$$

Step (i). Since $\mathbf{X}$ satisfies the RSC condition, the first term on the LHS of (C.2) is lower bounded by

$$\frac{\alpha^{\mathbf{X}}_{\mathrm{RSC}}}{2}|||\Delta_\Gamma|||^2_{\mathrm{F}} - \tau_{\mathbf{X}}\|\Delta_\Gamma\|_1^2. \tag{C.4}$$

To get a lower bound for (C.4), consider an upper bound for $\|\Delta_\Gamma\|_1$ with the aid of (C.1). Specifically, for the first two terms in the RHS of (C.1), by Hölder's inequality, the following inequalities hold for the inner products:

$$\langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\mathbf{E}\rangle\!\rangle \le \|\Delta_\Gamma\|_1\|\mathbf{X}^\top\mathbf{E}\|_\infty, \qquad \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle \le |||\Delta_\Theta|||_*|||\mathbf{E}|||_{op} = n|||\Delta_\Theta|||_*\Lambda^{1/2}_{\max}(S_{\mathbf{E}}); \tag{C.5}$$

for the last term, since

$$\|\widehat{\Gamma}\|_1 = \|\Gamma^\star_{\mathbb{S}} + \Gamma^\star_{\mathbb{S}^c} + \Delta_{\Gamma|\mathbb{S}} + \Delta_{\Gamma|\mathbb{S}^c}\|_1 = \|\Gamma^\star_{\mathbb{S}} + \Delta_{\Gamma|\mathbb{S}}\|_1 + \|\Delta_{\mathbb{S}|\mathbb{S}^c}\|_1 \ge \|\Gamma^\star_{\mathbb{S}}\|_1 - \|\Delta_{\Gamma|\mathbb{S}}\|_1 + \|\Delta_{\Gamma|\mathbb{S}^c}\|_1,$$

the following inequality holds:

$$\|\Gamma^\star\|_1 - \|\widehat{\Gamma}\|_1 \le \|\Delta_{\Gamma|\mathbb{S}}\|_1 - \|\Delta_{\Gamma|\mathbb{S}^c}\|_1. \tag{C.6}$$

Using the non-negativity of the RHS in (C.1), by choosing $\lambda_\Gamma \ge \max\left\{2\|\mathbf{X}^\top\mathbf{E}/n\|_\infty,\ \Lambda^{1/2}_{\max}(S_{\mathbf{E}})\right\}$, the following inequality holds:

$$0 \le \frac{\lambda_\Gamma}{2}\|\Delta_\Gamma\|_1 + \lambda_\Gamma|||\Delta_\Theta/\sqrt{n}|||_* + \lambda_\Gamma(\|\Delta_{\Gamma|\mathbb{S}}\|_1 - \|\Delta_{\Gamma|\mathbb{S}^c}\|_1) = \frac{3\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1 - \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}^c}\|_1 + \lambda_\Gamma|||\Delta_\Theta/\sqrt{n}|||_*.$$

Since $\Delta_\Theta = \widehat{\Theta} - \Theta^\star$ has rank at most $p_1 + r$, $|||\Delta_\Theta/\sqrt{n}|||_* \le \sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}$. It follows that

$$\begin{aligned}
&\frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}^c}\|_1 \le \lambda_\Gamma\sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + \frac{3\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1,\\
&\frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1 + \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}^c}\|_1 \le \lambda_\Gamma\sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + \frac{3\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1 + \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1,\\
&\|\Delta_\Gamma\|_1 \le \sqrt{4(p_1+r)}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + 4\|\Delta_{\Gamma|\mathbb{S}}\|_1 \le \sqrt{4(p_1+r)}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + 4\sqrt{s_{\Gamma^\star}}|||\Delta_\Gamma|||_{\mathrm{F}},
\end{aligned}$$

where the second line is obtained by adding $\frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1$ on both sides, and the last inequality uses (C.3). Further, by the Cauchy-Schwartz inequality, we have

$$\|\Delta_\Gamma\|_1 \le \sqrt{(\sqrt{4(p_1+r)})^2 + (4\sqrt{s})^2}\sqrt{|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2},$$

that is,

$$\|\Delta_\Gamma\|_1^2 \le 4\big(p_1 + r + 4s\big)\Big[|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2\Big]. \tag{C.7}$$

Combine (C.4) and (C.7), a lower bound for the LHS of (C.2) is given by

$$\Big(\frac{\alpha_{\mathrm{RSC}}^{\mathbf{X}}}{2} - 4\tau_{\mathbf{X}}\big(p_1 + r + 4s\big)\Big)|||\Delta_\Gamma|||_{\mathrm{F}}^2 + \Big(\frac{1}{2} - 4\tau_{\mathbf{X}}\big(p_1 + r + 4s\big)\Big)|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2. \tag{C.8}$$

Step (ii). For the first term in the RHS of (C.2), using the duality of the nuclear-operator norm pair, the following inequality holds:

$$\begin{aligned}
\frac{1}{n}|\langle\!\langle \mathbf{X}\Delta_\Gamma^\top, \widehat{\Theta} - \Theta^\star\rangle\!\rangle| &\le \frac{1}{n}|\langle\!\langle \mathbf{X}\Delta_\Gamma^\top, \widehat{\Theta}\rangle\!\rangle| + \frac{1}{n}|\langle\!\langle \mathbf{X}\Delta_\Gamma^\top, \Theta^\star\rangle\!\rangle| &\text{(C.9)}\\
&\le |||\mathbf{X}\Delta_\Gamma^\top/\sqrt{n}|||_{\mathrm{op}}|||\widehat{\Theta}/\sqrt{n}|||_* + |||\mathbf{X}\Delta_\Gamma^\top/\sqrt{n}|||_{\mathrm{op}}|||\Theta^\star/\sqrt{n}|||_*. &\text{(C.10)}
\end{aligned}$$

For $|||\mathbf{X}\Delta_\Gamma^\top/\sqrt{n}|||_{\mathrm{op}}$, we have

$$|||\mathbf{X}\Delta_\Gamma^\top/\sqrt{n}|||_{\mathrm{op}} \le |||\mathbf{X}/\sqrt{n}|||_{\mathrm{op}}|||\Delta_\Gamma^\top|||_{\mathrm{op}} \le |||\mathbf{X}/\sqrt{n}|||_{\mathrm{op}}|||\Delta_\Gamma^\top|||_{\mathrm{F}} = \Lambda_{\max}^{1/2}(S_{\mathbf{X}})|||\Delta_\Gamma|||_{\mathrm{F}}, \tag{C.11}$$

where the first inequality comes from the sub-multiplicativity of the nuclear norm. Combining with (IR+) box constraint on the eigen-spectrum and the feasibility of $\Theta^\star$, we obtain $|||\Theta^\star/\sqrt{n}|||_* \le p_1\phi$ and $|||\widehat{\Theta}/\sqrt{n}|||_* \le r\phi$, thus (C.9) is upper bounded by

$$(p_1 + r)\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})|||\Delta_\Gamma|||_{\mathrm{F}}.$$

Further, combining with (C.5) and (C.6), as long as $\lambda_\Gamma \ge \big\{\|\mathbf{X}^\top\mathbf{E}/n\|_\infty, \Lambda^{1/2}(S_{\mathbf{E}}), (p_1 +$

$r)\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})\big\}$, the following upper bound holds for the RHS of (C.2):

$$\begin{aligned}
&\lambda_\Gamma|||\Delta_\Gamma|||_{\mathrm{F}} + \lambda_\Gamma\|\Delta_\Gamma\|_1 + \lambda_\Gamma\sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + \lambda_\Gamma(\|\Delta_{\Gamma|\mathbb{S}}\|_1 - \|\Delta_{\Gamma|\mathbb{S}^c}\|_1) \\
\leq\ & \lambda_\Gamma\Big((2\sqrt{s_{\Gamma^\star}}+1)|||\Delta_\Gamma|||_{\mathrm{F}} + \sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}\Big) \qquad \text{(C.12)} \\
\leq\ & \lambda_\Gamma\sqrt{(2\sqrt{s_{\Gamma^\star}}+1)^2+(p_1+r)^2}\sqrt{|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2}.
\end{aligned}$$

Step (iii). Combine (C.8) and (C.12), by rearranging terms and requiring $\tau_{\mathbf{X}}$ to satisfy $\tau_{\mathbf{X}}(p_1+r+4s_{\Gamma^\star}) < \min\{\alpha_{\mathrm{RSC}}^{\mathbf{X}},1\}/16$, the following inequality holds:

$$\frac{\min\{\alpha_{\mathrm{RSC}}^{\mathbf{X}},1\}}{4}\Big(|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2\Big) \leq \lambda_\Gamma\sqrt{(2\sqrt{s_{\Gamma^\star}}+1)^2+(p_1+r)^2}\sqrt{|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2},$$

which gives

$$|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 \leq \frac{16\lambda_\Gamma^2\Big((p_1+r)+(2\sqrt{s_{\Gamma^\star}}+1)^2\Big)}{\min\{\alpha_{\mathrm{RSC}}^{\mathbf{X}},1\}^2}.$$

□

Proof sketch for Theorem 4.1. First we note that the requirement on the tuning parameter λ_Γ determines the leading term in the ultimate high probability error bound. By Lemma 4.4 and 4.5, to have adequate concentration for the leading eigenvalue $\Lambda_{\max}(\cdot)$ of the sample covariance matrices, the requirement imposed on the sample size makes $\sqrt{\log(p_2q)/n}$ a lower order term relative to $\mathcal{M}^{1/2}(f_X)$ and $\Lambda_{\max}^{1/2}(\Sigma_e)$, with the latter two being $\mathcal{O}(1)$ terms. Consequently, the choice of the tuning parameter effectively becomes

$$\lambda_\Gamma \asymp \mathcal{O}(1),$$

and by conditions C1 and C2, there exists some constant C such that $\lambda_\Gamma^2 \leq C$. The conclusion readily follows as a result of Proposition 4.1. □

Part 2. This part contains the proofs for the results related to $\widehat{A}$.

Proof sketch for Proposition 4.2. The result follows along the lines of **?** , Proposition 4.1]. In particular, in **?**], the authors consider estimation of A based on the directly observed samples of the X_t process, with the restricted eigenvalue (RE) condition imposed on the corresponding Hessian matrix and the tuning parameter selected in accordance to the deviation bound defined in Definition 4.2.

On the other hand, in the current setting, estimation of the transition matrix is based on quantities that are surrogates for the true sample quantities. Consequently, as long as

the required conditions are imposed on their counterparts associated with these surrogate quantities, the conclusion directly follows.

Finally, we would like to remark that the RSC condition used is in essence identical to the RE condition required in **?**] in the setting under consideration. □

Proof of Theorem 4.2. First, we note that under (IR), by Theorem 4.1, there exists some constant K_1 that is independent of n, p_1, p_2 and q such that the following event holds with probability at least $P_1 := 1 - c_1 \exp(-c_2 \log(p_2 q))$:

$$\mathcal{E}_1 := \Big\{ |||\Delta_{\mathbf{F}}/\sqrt{n}|||_{\mathrm{F}} \leq K_1 \Big\}.$$

Conditional on $\mathcal{E}_1$, by Proposition 4.2, Lemmas 4.6 and 4.7, with high probability, the following event holds:

$$\mathcal{E}_2 := \Big\{ |||\Delta_A|||_{\mathrm{F}} \leq \varphi(n, p_1, p_2, K_1) \Big\},$$

for some function $\varphi(\cdot)$ that not only depends on sample size and dimensions, but also on K_1, provided that the "conditional" RSC condition is satisfied. What are left to be examined are: (i) what does $\mathcal{E}_1$ imply in terms of the RSC condition being satisfied *unconditionally*; and (ii) what does $\mathcal{E}_1$ imply in terms of the bound in $\mathcal{E}_2$,

Towards this end, for (i), we note that since

$$\Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}}) = |||\Delta_{\mathbf{F}}/\sqrt{n}|||_{op} \leq |||\Delta_{\mathbf{F}}/\sqrt{n}|||_{\mathrm{F}} \leq K_1,$$

then as long as C_Z in condition C3 satisfies $C_Z \geq c_0 K_1$ with the specified $c_0 \geq 6\sqrt{165\pi}$, with probability at least $P_1 P_{2,\mathrm{RSC}}$ where we define $P_{2,\mathrm{RSC}} := 1 - c_1' \exp(-c_2' n)$, by Lemma 4.6 the required RSC condition is guaranteed to be satisfied with a positive curvature. For (ii), with the aid of Lemma 4.7, with probability at least $P_1 P_{2,\mathrm{DB}}$ where we define $P_{2,\mathrm{DB}} := 1 - c_1' \exp(-c_2' \log(p_1 + p_2))$, the following bound holds for the deviation bound $C(n, p_1, p_2)$ *unconditionally*: [1]

$$\begin{aligned} C(n, p_1, p_2) \leq\ & C_1 \big(\mathcal{M}(f_Z) + \frac{\Sigma_w}{2\pi} + \mathcal{M}(f_{Z,W}) \big) \sqrt{\frac{\log(p_1 + p_2)}{n}} + C_2 \mathcal{M}^{1/2}(f_Z) \sqrt{\frac{\log(p_1 + p_2) + \log p_1}{n}} \\ & + C_3 \Lambda_{\max}^{1/2}(\Sigma_w) \sqrt{\frac{\log(p_1 + p_2)}{n}} + C_4, \end{aligned}$$

where the constants $\{C_i\}$ have already absorbed the upper error bound K_1 of the Stage I estimates, compared with the original expression in Proposition 4.2. With the required sample size, the constant becomes the leading term, so that there exists some constant K_2

[1] Note that it can be shown that $|||\varepsilon_n|||_{\mathrm{F}}^2 = O(|||\Delta_{\mathbf{F}}|||_{\mathrm{F}}^2)$

such that *unconditionally*:

$$C(n, p_1, p_2) \leq K_2 \asymp \mathcal{O}(1).$$

Combine (i) and (ii), and with probability at least $\min\{P_1 P_{2,\mathrm{RSC}}, P_1 P_{2,\mathrm{DB}}\}$, the bound in Theorem 4.2 holds. □

C.2 Proof for Lemmas.

In this section, we provide proofs for the lemmas in Section 4.3.2.

Proof of Lemma 4.1. Note that

$$\begin{aligned}\widehat{\Theta} = \Theta^\star + \Delta_\Theta &= (\mathbf{F} + \Delta_\mathrm{F})(\Lambda^\star + \Delta_\Lambda)^\top \\ \Delta_\Theta &= \Delta_\mathrm{F}(\Lambda^\star)^\top + \widehat{\mathbf{F}}\Delta_\Lambda^\top.\end{aligned}$$

Multiply the left inverse of $\widehat{\mathbf{F}}$ which gives

$$\Delta_\Lambda^\top = \big(\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}\big)^{-1}\widehat{\mathbf{F}}^\top\Delta_\Theta + \big(\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}\big)^{-1}\widehat{\mathbf{F}}^\top\Delta_{\mathbf{F}}(\Lambda^\star)^\top.$$

Since for some generic matrix M, we have $|||M^{-1}|||_\mathrm{F} \geq (|||M|||_\mathrm{F})^{-1}$, an application of the triangle inequality gives

$$\begin{aligned}|||\Delta_\Lambda|||_\mathrm{F} &\leq \frac{|||\widehat{\mathbf{F}}|||_\mathrm{F}}{|||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F}}\Big(|||\Delta_\Theta|||_\mathrm{F} + |||\Delta_{\mathbf{F}}(\Lambda^\star)^\top|||_\mathrm{F}\Big) = \frac{|||\widehat{\mathbf{F}}/\sqrt{n}|||_\mathrm{F}}{|||\frac{1}{n}\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F}}(\frac{1}{\sqrt{n}})\Big(|||\Delta_\Theta|||_\mathrm{F} + |||\Delta_{\mathbf{F}}(\Lambda^\star)^\top|||_\mathrm{F}\Big) \\ &\leq \sqrt{p_1}\Lambda_{\max}^{-1/2}(S_{\widehat{\mathbf{F}}})|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F}\Big(1 + |||\Lambda^\star|||_\mathrm{F}\Big),\end{aligned}$$

where $S_{\widehat{\mathbf{F}}} := \frac{1}{n}\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}$, and after relaxing the numerator and the denominator of $\frac{|||\widehat{\mathbf{F}}|||_\mathrm{F}}{|||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F}}$ respectively by

$$|||\widehat{\mathbf{F}}|||_\mathrm{F} \leq \sqrt{p_1}|||\widehat{\mathbf{F}}|||_\mathrm{op}, \qquad |||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F} \geq |||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{op}.$$

Further, note that $|||\widehat{\mathbf{F}}/\sqrt{n}|||_\mathrm{op}^2 = \Lambda_{\max}(S_{\widehat{\mathbf{F}}}) = |||S_{\widehat{\mathbf{F}}}|||_\mathrm{op}$. What remains is to obtain a lower bound for

$$\Lambda_{\max}^{1/2}\big(S_{\widehat{\mathbf{F}}}\big) = |||(\mathbf{F} + \Delta_{\mathbf{F}})/\sqrt{n}|||_\mathrm{op}.$$

One such bound is given by

$$\begin{aligned}|||(\mathbf{F} + \Delta_{\mathbf{F}})/\sqrt{n}|||_\mathrm{op} \geq |||\mathbf{F}/\sqrt{n}|||_\mathrm{op} - |||\Delta_{\mathbf{F}}/\sqrt{n}|||_\mathrm{op} &\geq |||\mathbf{F}/\sqrt{n}|||_\mathrm{op} - |||\Delta_{\mathbf{F}}/\sqrt{n}|||_\mathrm{F} \\ &\geq \Lambda_{\max}^{1/2}(S_{\mathbf{F}}) - |||\Delta_\Theta/\sqrt{n}|||_\mathrm{F},\end{aligned}$$

which leads to the following bound for $|||\Delta_\Lambda|||_{\mathrm{F}}$, provided that the RHS is positive:

$$\frac{|||\Delta_\Lambda|||_{\mathrm{F}}}{|||\Lambda^\star|||_{\mathrm{F}}} \leq \sqrt{p_1}\frac{|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}}{\Lambda_{\max}^{1/2}(S_{\mathbf{F}}) - |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}}\Big(1 + 1/|||\Lambda^\star|||_{\mathrm{F}}\Big).$$

□

Proof of Lemma 4.2. First, suppose we have

$$\frac{1}{2}v'S_{\mathbf{X}}v = \frac{1}{2}v'\Big(\frac{\mathbf{X}'\mathbf{X}}{n}\Big)v \geq \frac{\alpha_{\mathrm{RSC}}}{2}\|v\|_2^2 - \tau_n\|v\|_1^2, \quad \forall\, v \in \mathbb{R}^p; \tag{C.13}$$

then, for all $\Delta \in \mathbb{R}^{p\times p}$, and letting Δ_j denote its jth column, the RSC condition automatically holds since

$$\frac{1}{2n}|||\mathbf{X}\Delta|||_{\mathrm{F}}^2 = \frac{1}{2}\sum_{j=1}^q \Delta_j'\big(\tfrac{\mathbf{X}'\mathbf{X}}{n}\big)\Delta_j \geq \frac{\alpha_{\mathrm{RSC}}}{2}\sum_{j=1}^q\|\Delta_j\|_2^2 - \tau_n\sum_{j=1}^q\|\Delta_j\|_1^2 \geq \frac{\alpha_{\mathrm{RSC}}}{2}|||\Delta|||_{\mathrm{F}}^2 - \tau_n\|\Delta\|_1^2.$$

Therefore, it suffices to verify that (C.13) holds. In **?** , Proposition 4.2], the authors prove a similar result under the assumption that X_t is a VAR(d) process. Here, we adopt the same proof strategy and state the result for a *more general process* X_t.

Specifically, by **?** , Proposition 2.4(a)], $\forall v \in \mathbb{R}^p, \|v\| \leq 1$ and $\eta > 0$,

$$\mathbb{P}\Big[\big|v^\top\big(S_{\mathbf{X}} - \Gamma_X(h)\big)v\big| > 2\pi\mathcal{M}(g_X)\eta\Big] \leq 2\eta\exp\Big(-cn\min\{\eta^2,\eta\}\Big).$$

Applying the discretization in **?** , Lemma F.2] and taking the union bound, define $\mathbb{K}(2s) := \{v \in \mathbb{R}^p, \|v\| \leq 1, \|v\|_0 \leq 2k\}$, and the following inequality holds:

$$\mathbb{P}\Big[\sup_{v\in\mathbb{K}(2k)}\big|v^\top\big(S_{\mathbf{X}}-\Gamma_X(h)\big)v\big| > 2\pi\mathcal{M}(g_X)\eta\Big] \leq 2\exp\Big(-cn\min\{\eta,\eta^2\}+2k\min\{\log p, \log(21ep/2k)\}\Big).$$

With the specified $\gamma = 54\mathcal{M}(g_X)/\mathfrak{m}(g_X)$, set $\eta = \gamma^{-1}$, then apply results from **?** , Lemma 12] with $\Gamma = S_{\mathbf{X}} - \Gamma_X(0)$ and $\delta = \pi\mathfrak{m}(g_X)/27$, so that the following holds

$$\frac{1}{2}v^\top S_{\mathbf{X}}v \geq \frac{\alpha_{\mathrm{RSC}}}{2}\|v\|^2 - \frac{\alpha_{\mathrm{RSC}}}{2k}\|v\|_1^2,$$

with probability at least $1-2\exp\big(-cn\min\{\gamma^{-2},1\}+2k\log p\big)$ and note that $\min\{\gamma^{-2},1\} = \gamma^{-2}$ since $\gamma > 1$. Finally, let $k = \min\{cn\gamma^{-2}/(c'\log p), 1\}$ for some $c' > 2$, and conclude that with probability at least $1 - c_1\exp(-c_2 n)$, the inequality in (C.13) holds with

$$\alpha_{\mathrm{RSC}} = \pi\mathfrak{m}(g_X), \qquad \tau_n = \alpha_{\mathrm{RSC}}\gamma^2\frac{\log p}{2n},$$

and so does also the RSC condition. □

Proof of Lemma 4.3. We note that

$$\frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\big\|_\infty = \max_{1\le i,j\le p}\big|e_i^\top\big(\mathbf{X}^\top\mathbf{E}/n\big)e_j\big|,$$

where e_i is the p-dimensional standard basis with its i-th entry being 1. Applying **?** , Proposition 2.4(b)], for an arbitrary pair of (i,j), the following inequality holds:

$$\mathbb{P}\Big[\big|e_i^\top\big(\mathbf{X}^\top\mathbf{E}/n\big)e_j\big| > 2\pi\big(\mathcal{M}(g_X)+\frac{\Lambda_{\max}(\Sigma_e)}{2\pi}\big)\eta\Big] \le 6\exp\Big(-cn\min\{\eta^2,\eta\}\Big),$$

and note that e_t is a pure noise term that is assumed to be independent of X_t; hence, there is no cross-dependence term to consider. Take the union bound over all $1\le i\le p_2, 1\le j\le q$, and the following bound holds:

$$\mathbb{P}\Big[\max_{1\le i\le p_2,1\le j\le q}\big|e_i^\top\big(\mathbf{X}^\top\mathbf{E}/n\big)e_j\big| > 2\pi\big(\mathcal{M}(g_X)+\frac{\Lambda_{\max}(\Sigma_e)}{2\pi}\big)\eta\Big] \le 6\exp\Big(-cn\min\{\eta^2,\eta\}+\log(p_2q)\Big).$$

Set $\eta = c'\sqrt{\log p/n}$ for $c' > (1/c)$ and with the choice of $n \gtrsim \log(p_2q)$, $\min\{\eta^2,\eta\} = \eta^2$, then with probability at least $1-c_1\exp(-c_2\log p_2q)$, there exists some c_0 such that the following bound holds:

$$\frac{1}{n}\big\|\mathbf{X}^\top\mathbf{E}\big\|_\infty \le c_0\big(2\pi\mathcal{M}(g_X)+\Lambda_{\max}(\Sigma_e)\big)\sqrt{\frac{\log(p_2q)}{n}}.$$

□

Before proving Lemma 4.4, we first state Lemma C.1 which provides a concentration inequality in the operator norm.

Lemma C.1. *Consider the stationary centered Gaussian process $\{X_t\}\in\mathbb{R}^p$, whose spectral density function $g_X(\omega)$ exits and the maximum eigenvalue is bounded a.e. on $[-\pi,\pi]$. Then, for $\mathbf{X}$ whose rows are random realizations $\{x_0,\ldots,x_{n-1}\}$ of $\{X_t\}$, the following bound holds for $S_\mathbf{X} = \mathbf{X}^\top\mathbf{X}/n$, for some $c>0$:*

$$\mathbb{P}\Big[\|\!|S_\mathcal{X}-\Gamma_X(0)|\!\|_{op} > 4\pi\mathcal{M}(g_X)\eta\Big] \le 2\exp(-cn\min\{\eta,\eta^2\}+p\log 8).$$

Proof of Lemma C.1. First, we note that by **?** , Proposition 2.4], the following inequality holds for any fixed $v\in S^p$, where $S^p := \{v\in\mathbb{R}^p : \|v\|=1\}$ is the p-dimensional unit sphere:

$$\mathbb{P}\Big[\big|v'\big(S_\mathbf{X}-\Gamma_X(0)\big)v\big| > 2\pi\mathcal{M}(g_X)\eta\Big] \le 2\exp(-cn\min\{\eta,\eta^2\}). \tag{C.14}$$

Additionally, by **?** , Lemma 5.4],

$$\|\|\|S_{\mathbf{X}} - \Gamma_X(0)\|\|\|_{op} = \sup_{v \in S^p} \left|v'\big(S_{\mathbf{X}} - \Gamma_X(0)\big)v\right| \leq (1-2\delta)^{-1} \sup_{v \in \mathcal{N}_\delta} v'\big[S_{\mathbf{X}} - \Gamma_X(0)\big]v,$$

where $\mathcal{N}_\delta$ is a δ-net of S^p for some $\delta \in [0,1)$, which guarantees that the sphere can essentially be replaced by its δ-net whose cardinality is finite. Towards this end, based upon (C.14), take the union bound over all vectors v in the $\frac{1}{4}$-net of S^p, whose cardinality is at most 8^p [e.g. **?**], we have

$$\begin{aligned}\mathbb{P}\Big[\|\|\|\tfrac{1}{n}X'X - \Gamma_X(0)\|\|\|_{op} > 4\pi\mathcal{M}(g_X)\eta\Big] &\leq \mathbb{P}\Big[\sup_{v \in \mathcal{N}_\delta}\left|v'\big(S - \Gamma_X(0)\big)v\right| > 4\pi\mathcal{M}(g_X)\eta\Big] \\ &\leq 8^p \cdot 2\exp\Big(-cn\min\{\eta, \eta^2\}\Big).\end{aligned}$$

□

Proof of Lemma 4.4. The result follows in a straightforward manner based on Lemma C.1. Specifically, by letting $\eta = c'\sqrt{p_2/n}$ for $c' > (\log 8/c)$ and with $n \gtrsim p$ so that $\min\{\eta^2, \eta\} = \eta^2$, then if we relax $\Lambda_{\max}(\Gamma_X(0))$ by its upper bound $2\pi\mathcal{M}(g_X)$ [**?** , Proposition 2.3], with probability at least $1 - c_1\exp(-c_2 p_2)$, the following bound holds for some c_0:

$$\Lambda_{\max}(S_{\mathbf{X}}) \leq c_0\mathcal{M}(g_X).$$

□

Proof of Lemma 4.5. For $\mathbf{E}$ whose rows are iid realizations of a sub-Gaussian random vector e_t, by **?** , Lemma 9], the following bound holds:

$$\mathbb{P}\Big[\|\|\|S_{\mathbf{E}} - \Sigma_e\|\|\|_{op} \geq \Lambda_{\max}(\Sigma_e)\delta(n,q,\eta)\Big] \leq 2\exp(-n\eta^2/2),$$

where $\delta(n,q,\eta) := 2\big(\sqrt{\frac{q}{n}} + \eta\big) + \big(\sqrt{\frac{q}{n}} + \eta\big)^2$. In particular, by triangle inequality, with probability at least $1 - 2\exp(-n\eta^2/2)$,

$$\|\|\|S_{\mathbf{E}}\|\|\|_{op} \leq \|\|\|\Sigma_\epsilon\|\|\|_{op} + \|\|\|S_{\mathbf{E}} - \Sigma_\epsilon\|\|\|_{op} \leq \Lambda_{\max}(\Sigma_\epsilon) + \Lambda_{\max}(\Sigma_\epsilon)\delta(n,q,t).$$

So for $n \gtrsim q$, by setting $\eta = 1$, which yields $\delta(n,q,\eta) \leq 8$ so that with probability at least $1 - 2\exp(-n/2)$, the following bound holds:

$$\Lambda_{\max}(S_{\mathbf{E}}) \leq 9\Lambda_{\max}(\Sigma_\epsilon).$$

□

Proof of Lemma 4.6. It suffices to show that the following inequality holds with high probability for some curvature $\alpha^{\widehat{\mathbf{Z}}}_{\mathrm{RSC}} > 0$ and tolerance $\tau_{\mathbf{Z}}$, where we define $\widehat{\Gamma}_{\mathbf{Z}} := \frac{1}{n}\widehat{\mathbf{Z}}^{\top}_{n-1}\widehat{\mathbf{Z}}_{n-1}$:

$$\frac{1}{2}\theta^{\top}\widehat{\Gamma}_{\mathbf{Z}}\theta \geq \frac{\alpha^{\widehat{\mathbf{Z}}}_{\mathrm{RSC}}}{2}\|\theta\|^2 - \tau_{\mathbf{Z}}\|\theta\|_1^2, \qquad \forall\ \theta \in \mathbb{R}^p.$$

Define $\Gamma_{\mathbf{Z}} := \frac{1}{n}\mathbf{Z}^{\top}_{n-1}\mathbf{Z}_{n-1}$, then $\widehat{\Gamma}_{\mathbf{Z}}$ can be written as

$$\widehat{\Gamma}_{\mathbf{Z}} = \Gamma_{\mathbf{Z}} + \Big(\frac{1}{n}\mathbf{Z}^{\top}_{n-1}\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big) + \Big(\frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\Delta_{\mathbf{Z}_{n-1}}\Big), \tag{C.15}$$

First, notice that the last term satisfies the following natural lower bound *deterministically*, since $\Delta_{\mathbf{F}}$ is assumed non-random and $\Delta_{\mathbf{Z}} = [\Delta_{\mathbf{F}}, O]$:

$$\theta^{\top}\Big(\frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\Delta_{\mathbf{Z}_{n-1}}\Big)\theta \geq 0 \qquad \forall\ \theta \in \mathbb{R}^p,$$

which however, does not contribute to the "positive" part of curvature. For the first two terms, we adopt the following strategy, using Lemma 12 in **?**] as an intermediate step. Specifically, **?** , Lemma 12] proves that for any fixed generic matrix $\Gamma \in \mathbb{R}^{p\times p}$ that satisfies $|\theta^{\top}\Gamma\theta| \leq \delta$ for any $\theta \in \mathbb{K}(2s)$[2], the following bound holds

$$|\theta^{\top}\Gamma\theta| \leq 27\delta\big(\|\theta\|_2^2 + \frac{1}{s}\|\theta\|_1^2\big), \qquad \forall\ \theta \in \mathbb{R}^p. \tag{C.16}$$

Then, based on (C.16), consider $\Gamma = \widehat{\Gamma} - \Sigma$ then rearrange terms, so that $\theta^{\top}\widehat{\Gamma}\theta \geq \theta^{\top}\Sigma\theta - \frac{27\delta}{2}\big(\|\theta\|_2^2 + \frac{1}{2}\|\theta\|_1^2\big)$. The RE condition follows by setting δ to be some quantity related to $\Lambda_{\min}(\Sigma)$.

In light of this, for the first two terms in (C.15), let

$$\Psi := \Gamma_{\mathbf{Z}} + \Big(\frac{1}{n}\mathbf{Z}^{\top}_{n-1}\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big),$$

denote their sum, in order to obtain an upper bound for $\big|\theta^{\top}\big(\Psi - \Gamma_Z(0)\big)\theta\big|$, so that Lemma 12 in **?**] can be applied. To this end, since

$$\Big|\theta^{\top}\big[\Psi - \Gamma_Z(0)\big]\theta\Big| \leq \Big|\theta'(\Gamma_{\mathbf{Z}} - \Gamma_Z(0))\theta\Big| + \Big|\theta'\Big(\frac{1}{n}\mathbf{Z}'_{n-1}\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta'_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big)\theta\Big|,$$

[2] $\mathbb{K}(2s) := \{\theta : \|\theta\|_0 = 2s\}$ is the set of $2s$-sparse vectors.

we consider getting upper bounds for each of the two terms:

$$\text{(i)} \quad \Big|\theta^\top(\Gamma_{\mathbf{Z}} - \Gamma_Z(0))\theta\Big| \,, \qquad \text{(ii)} \quad \Big|\theta^\top\Big(\frac{1}{n}\mathbf{Z}_{n-1}^\top\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta'_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big)\theta\Big|.$$

For (i), we follow the derivation in **?** , Proposition 2.4(a)], that is, for all $\|\theta\| \le 1$,

$$\mathbb{P}\Big[\Big|\theta'\big(\Gamma_{\mathbf{Z}} - \Gamma_Z(0)\big)\theta\Big| > 2\pi\mathcal{M}(f_Z)\eta\Big] \le 2\exp\big[-cn\min\{\eta^2,\eta\}\big],$$

and further with probability at least $1-2\exp\big(-cn\min\{\eta^2,\eta\}+2s\min\{\log p, \log(21ep/2s)\}\big)$, the following bound holds:

$$\sup_{\theta\in\mathbb{K}(2s)}\Big|\theta^\top\big(\Gamma_{\mathbf{Z}} - \Gamma_Z(0)\big)\theta\Big| < 2\pi\mathcal{M}(f_Z)\eta. \tag{C.17}$$

For (ii), the two terms are identical, with either one given by

$$\frac{1}{n}(\mathbf{Z}_{n-1}\theta)^\top(\Delta_{\mathbf{Z}_{n-1}}\theta).$$

To obtain its upper bound, consider the following inequality, based on which we bound the two terms in the product separately:

$$\sup_{\theta\in\mathbb{K}(2s)}\Big|\frac{1}{n}\langle\mathbf{Z}_{n-1}\theta, \Delta_{\mathbf{Z}_{n-1}}\theta\rangle\Big| \le \Big(\sup_{\theta\in\mathbb{K}(2s)}\|\frac{\mathbf{Z}_{n-1}\theta}{\sqrt{n}}\|\Big)\Big(\sup_{\|\theta\|\le 1}\|\frac{\Delta_{\mathbf{Z}_{n-1}}\theta}{\sqrt{n}}\|\Big). \tag{C.18}$$

For the first term in (C.18), since rows of $\mathbf{Z}_{n-1}$ are time series realizations from (4.6), then if we let $\xi := \mathbf{Z}_{n-1}\theta$, $\xi \sim \mathcal{N}(0_{n\times 1}, Q_{n\times n})$ is Gaussian with $Q_{st} = \theta'\Gamma_Z(t-s)\theta$. To get its upper bound, we bound its square, and use again (C.17), that is,

$$\sup_{\theta\in\mathbb{K}(2s)}\Big|\theta^\top\Big(\frac{1}{n}\mathbf{Z}_{n-1}^\top\mathbf{Z}_{n-1}\Big)\theta\Big| < \sup_{\theta\in\mathbb{K}(2s)}\theta'\Gamma_Z(0)\theta + 2\pi\mathcal{M}(f_Z) \le 2\pi\mathcal{M}(f_Z) + 2\pi\mathcal{M}(f_Z)\eta.$$

For the second term $\|\Delta_{\mathbf{Z}_{n-1}}\theta/\sqrt{n}\|$, this is non-random, and for all $\|\theta\| \le 1$, $\|\Delta_{\mathbf{Z}_{n-1}}\theta/\sqrt{n}\| \le \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{Z}_{n-1}}}\big) = \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)$. Therefore, the following bound holds for (C.18):

$$\sup_{\theta\in\mathbb{K}(2s)}\Big|\frac{1}{n}\langle\mathbf{Z}_{n-1}\theta, \Delta_{\mathbf{Z}_{n-1}}\theta\rangle\Big| \le \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + 2\pi\mathcal{M}(f_Z)\eta}. \tag{C.19}$$

Combine (C.17) and (C.19) that are respectively the bounds for (i) and (ii), and the following

bound holds with probability at least $1-2\exp\big(-cn\min\{\eta^2,\eta\}+2s\min\{\log p,\log(21ep/2s)\}\big)$:

$$\sup_{\theta\in\mathbb{K}(2s)}\Big|\theta^\top\Big(\Psi-\Gamma_Z(0)\Big)\theta\Big|\le 2\pi\mathcal{M}(f_Z)\eta+2\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z)+2\pi\mathcal{M}(f_Z)\eta}. \quad \text{(C.20)}$$

Now applying **?** , Lemma 12] to $\Gamma=\Psi-\Gamma_Z(0)$, and δ being the RHS of (C.20), then the following bound holds:

$$\theta^\top\widehat{\Gamma}_{\mathbf{Z}}\theta\ge 2\pi\mathfrak{m}(f_Z)\|\theta\|_2^2-27\delta(\|\theta\|_2^2+\frac{1}{s}\|\theta\|_1^2)=\big(2\pi\mathfrak{m}(f_Z)-27\delta\big)\|\theta\|^2-\frac{27\delta}{s}\|\theta\|_1^2.$$

By setting $\eta=\omega^{-1}:=\frac{\mathfrak{m}(f_Z)}{54\mathcal{M}(f_Z)}$,

$$\delta=\frac{\pi}{27}\mathfrak{m}(f_Z)+2\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z)+\pi\mathfrak{m}(f_Z)/27}\le\frac{\pi}{27}\mathfrak{m}(f_Z)+2\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{\frac{55\pi}{27}\mathcal{M}(f_Z)}$$

Since we have required that $\mathfrak{m}(f_Z)/\mathcal{M}^{1/2}(f_Z)>c_0\cdot\Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}})$ with $c_0\ge 6\sqrt{165\pi}$, $2\pi\mathfrak{m}(f_Z)-27\delta>0$. Therefore, the RSC condition is satisfied with curvature

$$\alpha_{\mathrm{RSC}}^{\widehat{\mathbf{Z}}}=2\pi\mathfrak{m}(f_Z)-27\delta=\pi\mathfrak{m}(f_Z)-54\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z)+\pi\mathfrak{m}(f_Z)/27}>0,$$

and tolerance $27\delta/(2s)$, with probability at least $1-2\exp\Big(-cn\omega^{-2}+2s\log p\Big)$. Finally, set $s=\big\lceil cn\omega^{-1}/4\log p\big\rceil$, we get the desired conclusion. □

Proof of Lemma 4.7. First, we note that the quantity of interest can be upper bounded by the following four terms:

$$\begin{aligned}\frac{1}{n}\|\widehat{\mathbf{Z}}_{n-1}^\top\Big(\widehat{\mathbf{Z}}_n-\widehat{\mathbf{Z}}_{n-1}(A^\star)^\top\Big)\|_\infty&=\frac{1}{n}\|\Big(\mathbf{Z}_{n-1}+\Delta_{\mathbf{Z}_{n-1}}\Big)^\top\Big(\mathbf{W}+\Delta_{\mathbf{Z}_n}-\Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\Big)\|_\infty\\&\le\|\frac{1}{n}\mathbf{Z}_{n-1}^\top\mathbf{W}\|_\infty+\|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^\top\mathbf{W}\|_\infty+\|\frac{1}{n}\mathbf{Z}_{n-1}^\top\Big(\Delta_{\mathbf{Z}_n}-\Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\Big)\|_\infty\\&\quad+\|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^\top\Big(\Delta_{\mathbf{Z}_n}-\Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\Big)\|_\infty\\&:=T_1+T_2+T_3+T_4. \quad \text{(C.21)}\end{aligned}$$

We provide bounds on each term in (C.21) sequentially. T_1 is the standard Deviation Bound, which according to previous derivations (e.g., **?**] for the expression specifically derived for VAR(1)) satisfies

$$\frac{1}{n}\|\mathbf{Z}_{n-1}^\top\mathbf{W}\|_\infty\le c_0\big[\mathcal{M}(f_Z)+\mathcal{M}(f_W)+\mathcal{M}(f_{Z,W^+})\big]\sqrt{\frac{\log(p_1+p_2)}{n}}$$

with probability at least $1-c_1\exp(-c_2\log(p_1+p_2))$ for some $\{c_i\}$. For T_2, since rows of $\mathbf{W}$

are iid realizations from $\mathcal{N}(0,\Sigma_w)$, then for $\Delta_{\mathbf{Z}_{n-1}}^\top \mathbf{W} \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}$ which has at most $p_1\times(p_1+p_2)$ nonzero entries, each entry (i,j) given by

$$\kappa_{ij} := \big(\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^\top \mathbf{W}\big)_{ij} = \frac{1}{n}\Delta_{\mathbf{Z}_{n-1},\cdot i}^\top \mathbf{W}_{\cdot j}$$

is Gaussian, and the following tail bound holds:

$$\begin{aligned}\mathbb{P}\big[|\kappa_{ij}|\geq t\big] &\leq e\cdot\exp\Big(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w)\max\limits_{i\in\{1,\ldots,p_1+p_2\}}\|\Delta_{\mathbf{Z}\cdot i}/\sqrt{n}\|_2^2}\Big)\\ &= e\cdot\exp\Big(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w)\max\limits_{i\in\{1,\ldots,p_1\}}\|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\|_2^2}\Big).\end{aligned}$$

Taking the union bound over all $p_1\times(p_1+p_2)$ nonzero entries, the following bound holds:

$$\mathbb{P}\Big[\frac{1}{n}\|\Delta_{\mathbf{Z}_{n-1}}^\top\mathbf{W}\|_\infty\geq t\Big]\leq\exp\Big(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w)\max\limits_{i\in\{1,\ldots,p_1\}}\|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\|_2^2}+\log\big(ep_1(p_1+p_2)\big)\Big).$$

Choose $t=c_0\big(\Lambda_{\max}^{1/2}(\Sigma_w)\max\limits_{i=1,\ldots,p_1}\|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\|\big)\sqrt{\frac{\log(p_1(p_1+p_2))}{n}}$, the following bound holds with probability at least $1-\exp\Big(-c_1\log\big(p_1(p_1+p_2)\big)\Big)$:

$$\frac{1}{n}\|\Delta_{\mathbf{Z}_{n-1}}^\top\mathbf{W}\|_\infty\leq c_0\Lambda_{\max}^{1/2}(\Sigma_w)\max_{i=1,\ldots,p_1}\|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\|\sqrt{\frac{\log p_1+\log(p_1+p_2)}{n}}.$$

For T_3, let $\varepsilon_n := \Delta_{\mathbf{Z}_n}-\Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top = [\Delta_{\mathbf{F}_n}-\Delta_{\mathbf{F}_{n-1}}(A_{11}^\star)^\top, -\Delta_{\mathbf{F}_{n-1}}(A_{21}^\star)^\top]$, then each entry of $\frac{1}{n}\mathbf{Z}_{n-1}^\top\varepsilon_n$ is given by

$$\Big(\tfrac{1}{n}\mathbf{Z}_{n-1}^\top\varepsilon_n\Big)_{ij}=\frac{1}{n}\mathbf{Z}_{n-1,\cdot i}^\top\varepsilon_{n,\cdot j},$$

and it has $(p_1+p_2)\times(p_1+p_2)$ entries. Next, note that column i of $\mathbf{Z}_{n-1}\in\mathbb{R}^n$ can be viewed as a mean-zero Gaussian random vector with covariance matrix Q^i where $(Q^i)_{st}=[\Gamma_Z(t-s)]_{ii}$ satisfying $\Lambda_{\max}(Q^i)\leq\Lambda_{\max}(\Gamma_Z(0))\leq 2\pi\mathcal{M}(f_Z)$, so for any (i,j), $\big(\frac{1}{n}\mathbf{Z}_{n-1}^\top\varepsilon_n\big)_{ij}$ satisfies

$$\mathbb{P}\Big[\big|\big(\tfrac{1}{n}\mathbf{Z}_{n-1}^\top\varepsilon_n\big)_{ij}\big|>t\Big]\leq\exp\Big(1-\frac{cnt^2}{\Lambda_{\max}(\Gamma_Z(0))\max\limits_{j\in\{1,\ldots,p_1\}}\|\varepsilon_{n,\cdot j}/\sqrt{n}\|^2}\Big).$$

Again by taking the union bound over all $(p_1+p_2)^2$ entries, and let

$$t=c_0\big(2\pi\mathcal{M}(f_Z)\big)^{1/2}\max_{j\in\{1,\ldots,p_1\}}\|\varepsilon_{n,\cdot j}/\sqrt{n}\|\sqrt{\frac{\log p_1+\log(p_1+p_2)}{n}},$$

the following bound holds w.p. at least $1-\exp(-c_1\log(p_1+p_2))$:

$$\frac{1}{n}\|\mathbf{Z}_{n-1}^\top\Big(\Delta_{\mathbf{Z}_n}-\Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\Big)\|_\infty \le c_0\big(2\pi\mathcal{M}(f_Z)\big)^{1/2}\max_{j\in\{1,\ldots,(p_1+p_2)\}}\|\varepsilon_{n,\cdot j}/\sqrt{n}\|\sqrt{\frac{\log(p_1+p_2)}{n}}.$$

For T_4, it is deterministic, and satisfies

$$\begin{aligned}\frac{1}{n}\|\Delta_{\mathbf{Z}_{n-1}}^\top\Big(\Delta_{\mathbf{Z}_n}-\Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\Big)\|_\infty &\le \|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^\top\Delta_{\mathbf{Z}_n}\|_\infty+\|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^\top\Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\|_\infty\\ &= \|\frac{1}{n}\Delta_{\mathbf{F}_{n-1}}^\top\Delta_{\mathbf{F}_n}\|_\infty+\|\frac{1}{n}\Delta_{\mathbf{F}_{n-1}}^\top\Delta_{\mathbf{F}_{n-1}}(A_{11}^\star)^\top\|_\infty\end{aligned}$$

Combine all terms, and there exist some constant C_1, C_2, C_3 and c_1, c_2 such that with probability at least $1-c_1\exp\big(-c_2\log(p_1+p_2)\big)$, the bound in (4.14) holds. □

C.3 Additional Numerical Studies.

In this section, we investigate selected scenarios where the relaxed implementation on estimating the calibration equation may fail to produce good estimates, due to the absence of the compactness constraint. For illustration purposes, it suffices to consider the setting where X_t and F_t jointly follow a multivariate Gaussian distribution and are independent and identically distributed across samples. Throughout, we set $n=200, p_1=5, p_2=50, q=100$, and $\binom{X_t}{F_t}\sim\mathcal{N}(0,\Sigma)$ with $\Sigma_{ij}=0.25$ $(i\neq j)$ and $\Sigma_{ii}=1$. The noise level is fixed at $\sigma_e=1$.

First, we note that based on the performance evaluation shown in Section 4.4, the estimates demonstrate good performance even without the compactness constraint. Note that the simulation settings are characterized by adequate sparsity in Γ, which in turn limits the size of the equivalence class $\mathcal{C}(Q_2)$ as mentioned in Remark 4.1 in Section 4.2.1. Therefore, we focus on the following two issues: (i) whether sparsity encourages additional "approximate identification"; and (ii) whether a good initializer helps constrain estimates from subsequent iterations to a ball around the true value.

We start by considering a non-sparse Γ. Specifically, for both Λ and Γ, their entries are generated from $\mathsf{Unif}\{(-1.5,-1.2)\cup(1.2,1.5)\}$. Additionally, we specify one model in $\mathcal{C}(Q_2)$ by setting $Q_2=\mathbf{5}_{p_1\times p_2}$, which will generate the corresponding $\breve{\mathbf{F}}$, $\breve{\Theta}$ and $\breve{\Gamma}$. Table C.1 depicts the performance of the estimated Θ based on different initializers:

initializer $\widehat{\Theta}^{(0)}$	$\Theta^\star$	$\mathbf{0}_{n\times q}$	$\Theta^\star+0.1*\mathbf{Z}_{n\times q}$	$\breve{\Theta}$
Rel.Err	0.09	0.63	fail to converge within 5000 iterations	1.82 (0.02, relative to $\breve{\Theta}$)

Table C.1: Performance evaluation of $\widehat{\Theta}$ obtained from different initializers under a non-sparse setting.

The results in Table C.1 show that the algorithm converges (if at all) to different local optima whose values may deviate markedly for the true ones. Specifically, initializer $\Theta^\star+$

$0.1 * \mathbf{Z}_{n\times q}$, where each entry $\Theta^\star$ is perturbed by an iid standard Gaussian random variable scaled by 0.1, fails to converge. Note that the perturbation is small, but the operator norm of the initializer far exceeds ϕ_0. Initializer $\check{\Theta}$ yields an estimate that is far from the true data-generating factor hyperplane, yet close to its observationally equivalent one. This suggests that in non-sparse settings, without imposing the compactness constraint on the equivalence class, a good initializer is required for the actual relaxed implementation to produce a fairly good estimate of the true data generating parameters.

However, this is not the case if there is sufficient sparsity in Γ. Specifically, using the same generating mechanism for Λ and Γ as in Section 4.4, we found that even with different initializers, the algorithm always produces estimates that are close to each other and also exhibit good performance. This finding strongly suggests that sparsity in Γ effectively shrinks the size of the equivalence class and the algorithm after a few iterations produces updates that are close to each other, irrespective of the initializer employed. Hence, the effective equivalence class is constrained to the one whose elements are encoded by $\check{\Gamma}$ that have similar characteristics in terms of the location of the non-zero parameters to Γ.

Finally, we consider a case that lies between the above two settings, that is, there is a structured sparsity patter in Γ. Specifically, we set the last 5 columns of Γ to be dense while the remaining ones are sparse. The overall sparsity level of Γ is fixed at 10%. Note that in this case, the size of the corresponding equivalence class is much larger to the one corresponding to a Γ with 10% uniformly distributed non-zeros entries, due to the presence of the five dense columns. As the results in Table C.2 indicate, when the initializer starts to

initializer $\widehat{\Theta}^{(0)}$	$\Theta^\star$	$\mathbf{0}_{n\times q}$	$\Theta^\star + 0.1 * \mathbf{Z}_{n\times q}$	$\mathbf{20}_{n\times q}$
Rel.Err	0.65	0.65	0.65	0.68

Table C.2: Performance evaluation for $\widehat{\Theta}$ obtained from different initializers under a structured-sparse setting.

deviate from the true value, there exist initializers that would yield inferior estimates.

In summary, in a non-sparse setting without compactification of the equivalence class, different initializers yield drastically different estimates that are not close enough to the true data-generating model, as expected by the approximate (IR+) condition employed. The problem is largely mitigated for sufficiently sparse Γ, which leads to shrinking the equivalence class. However, an exact characterization of the equivalence class is hard to obtain in practice, since the location of the non-zero entries in Γ is unknown.

C.4 An Outline of the Estimation Procedure in Low-dimensional Settings.

For the sake of completeness, we outline the estimation procedure proposed in **?**] and elaborate on the reasons that it is not applicable in high-dimensional settings. Note that the

restriction $\text{Cov}(w_t^X, W_t^F) = O$ is universal for all sets of identifications considered. Given a sample version corresponding to the calibration equation

$$\mathbf{Y} = \mathbf{F}\Lambda^\top + \mathbf{X}\Gamma^\top + \mathbf{E},$$

and that to the VAR equation

$$\mathbf{Z}_n = \mathbf{Z}_{n-1}A^\top + \mathbf{W},$$

the estimation procedure is based on the following steps.

1. Project and estimate a factor model. Specifically, by left multiplying $\mathbb{P}_{\mathbf{X}^\perp} := \mathrm{I}_{p_1} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, the model to estimate becomes

   $$\mathbb{P}_{\mathbf{X}^\perp}\mathbf{Y} = \mathbb{P}_{\mathbf{X}^\perp}\mathbf{F}\Lambda^\top + \mathbb{P}_{\mathbf{X}^\perp}\mathbf{E}.$$

   Proceed by doing factor analysis on $\mathbb{P}_{\mathbf{X}^\perp}\mathbf{Y}$ through a quasi-Maximum Likelihood procedure as detailed in **?**], and obtain *intermediate estimates* denoted by $\widetilde{\Lambda}$, $\widetilde{\mathbf{F}}$, $\widetilde{\Gamma}$ and $\widetilde{\Sigma}_{ee}$.

2. Estimate a VAR model based on $(\widetilde{\mathbf{F}}, \mathbf{X})$, and denote the intermediate estimate of the transition matrix by $\widetilde{A}$ and the residual by $\widetilde{\mathbf{W}}$. Calculate the sample covariance matrix of $\widetilde{\mathbf{W}}$, partitioned as $[\widetilde{\Sigma}_w^{ff}, \widetilde{\Sigma}_w^{fx}; \widetilde{\Sigma}_w^{xf}, \widetilde{\Sigma}_w^{xx}]$.

3. Calculate a rotation matrix depending on the identification restrictions (either IRa, IRb or IRc) so that the one under consideration is satisfied. Specifically, all such rotation matrices involve

   $$\widetilde{\Sigma}_w^{ff\cdot x} := \widetilde{\Sigma}_w^{ff} - \widetilde{\Sigma}_w^{fx}(\widetilde{\Sigma}_w^{xx})^{-1}\widetilde{\Sigma}_w^{xf}.$$

   Apply the rotation matrix (or their related transformations) to all previous intermediate estimates to obtain the final estimates.

To initiate the procedure, $\mathbb{P}_{\mathbf{X}^\perp}$ is required for the first step; yet, this quantity is not readily available in high-dimensional settings where $p_2 \geq n$. Moreover, subsequent calculations of the rotation matrix are based on $\widetilde{\Sigma}_w^{ff\cdot x}$, with the latter relying on $(\widetilde{\Sigma}_w^{xx})^{-1}$, which again is not properly defined under high-dimensional scaling.

C.5 List of Commodities and Macroeconomic Variables.

Commodity	Key	Description
ALUMINUM	PALUM	Aluminum, 99.5% minimum purity, LME spot price
COCOA	PCOCO	Cocoa beans, International Cocoa Organization cash price
COFFEE	PCOFFOTM	Coffee, Other Mild Arabicas, International Coffee Organization New York cash price
COPPER	PCOPP	Copper, grade A cathode, LME spot price
COTTON	PCOTTIND	Cotton, Cotton Outlook 'A Index', Middling 1-3/32 inch staple
LEAD	PLEAD	Lead, 99.97% pure, LME spot price
MAIZE	PMAIZMT	Maize (corn), U.S. No.2 Yellow, FOB Gulf of Mexico, U.S. price
NICKEL	PNICK	Nickel, melting grade, LME spot price
OIL	POILAPSP	Crude Oil (petroleum), simple average of three spot prices
RICE	PRICENPQ	Rice, 5 percent broken milled white rice, Thailand nominal price quote
RUBBER	PRUBB	Rubber, Singapore Commodity Exchange, No. 3 Rubber Smoked Sheets, 1st contract
SOYBEANS	PSOYB	Soybeans, U.S. soybeans, Chicago Soybean futures contract (first contract forward)
SUGAR	PSUGAUSA	Sugar, U.S. import price, contract no.14 nearest futures position
TIN	PTIN	Tin, standard grade, LME spot price
WHEAT	PWHEAMT	Wheat, No.1 Hard Red Winter, ordinary protein
ZINC	PZINC	Zinc, high grade 98% pure

Table C.3: List of commodities considered in this study. Data source: International Monetary Fund.

Name	Description	tCode	Category	Region
IPI_US	IP Index: total	5	Output & Income	US
CUM_US	Capacity Utilization: manufacturing	2	Output & Income	US
UNEMP_US	Civilian unemployment rate: all	2	Labor Market	US
HOUST_US	Housing Starts: ttl new privately owned	4	Housing	US
ISR_US	Total Business: inventories to sales ratio	2	Consumption	US
M2_US	M2 Money Stock	6	Money & Credit	US
BUSLN_US	Commericial and industrial loans	6	Money & Credit	US
REALN_US	Real estate loans at all commercial banks	6	Money & Credit	US
FFR_US	Effective federal funds rate	2	Interest & Exchange Rates	US
TB10Y_US	10-year treasury rate	2	Interest & Exchange Rates	US
BAA_US	Moody's Baa corporate bond yield	2	Interest & Exchange Rates	US
USDI_US	Trade weighted U.S.dollar index	5	Interest & Exchange Rates	US
CPI_US	CPI: all iterms	5	Prices	US
PCEPI_US	Personal Consumption Expenditure: chain index	5	Prices	US
SP500_US	S&P's Common Stock Price Index: composite	5	Stock Market	US
CPI_EU	Consumer Price Indices, percent change	2	Prices	EU
IPI_EU	Industrial Production Index: total industry (excluding construction)	5	Output & Income	EU
IPICP_EU	Industrial Production Index: construction	5	Output & Income	EU
M3_EU	Monetary aggregate M3	6	Money & Credit	EU
LOANRES_EU	Credit to resident sectors, non-MFI excluding gov	6	Money & Credit	EU
LOANGOV_EU	Credit to general government sector	6	Money & Credit	EU
PPI_EU	Producer Price Index: total industry (excluding construction)	6	Prices	EU
UNEMP_EU	Unemployment rate: total	2	Labor Market	EU
IMPORT_EU	Total trade: import value	6	Trade	EU
EXPORT_EU	Total trade: export value	6	Trade	EU
EB1Y_EU	Euribor 1 year	2	Interest & Exchange Rates	EU
TB10Y_EU	10-year government benchmark bond yield	2	Interest & Exchange Rates	EU
EFFEXR_EU	ECB nominal effective exchange rate againt group of trading partners	2	Interest & Exchange Rates	EU
EUROSTOXX50_EU	Euro STOXX composite index	5	Stock Market	EU
IOP_UK	Index of Production	5	Output & Income	UK
CPI_UK	CPI Index	5	Prices	UK
PPI_UK	Output of manufactured products	5	Prices	UK
UNEMP_UK	Unemployment rate: aged 16 and over	2	Labor Market	UK
EFFEXR_UK	Effective exchange rate index, Sterling	2	Interest & Exchange Rates	UK
TB10Y_UK	10-year British government stock, nominal par yield	2	Interest & Exchange Rates	UK
LIBOR6M_UK	6 month interbank lending rate, month end	2	Interest & Exchange Rates	UK
M3_UK	Monetary aggregate M3	6	Money & Credit	UK
CPI_CN	CPI: all iterms	5	Prices	CN
PPI_CN	Producer price index for industrial products (same month last year = 100)	2	Prices	CN
M2_CN	Monetary aggregate M2	6	Money & Credit	CN
EFFEXR_CN	Real broad effective exchange rate	2	Interest & Exchange Rates	CN

EXPORT_CN	Value goods	6	Trade	CN
IMPORT_CN	Value goods	6	Trade	CN
INDGR_CN	Growth rate of industrial value added (last year = 100)	2	Output & Income	CN
SHANGHAI_CN	Shanghai Composite Index	5	Stock Market	CN
TB10Y_JP	10-year government benchmark bond yield	2	Interest & Exchange Rates	JP
EFFEXR_JP	Real broad effective exchange rate	2	Interest & Exchange Rates	JP
CPI_JP	CPI Index: all items	5	Prices	JP
M2_JP	Monetary aggregate M2	6	Money & Credit	JP
UNEMP_JP	Unemployment rate: aged 15-64	2	Labor Market	JP
IPI_JP	Production of Total Industry	5	Output & Income	JP
IMPORT_JP	Import price index: all commodities	6	Trade	JP
EXPORT_JP	Value goods	6	Trade	JP
NIKKEI225_JP	NIKKEI 225 composite index	5	Stock Market	JP

Table C.4: List of macroeconomic variables in this study.

Data source: Fred St.Louis, ECB Statistical Data Warehouse, UK Office for National Statistics, Bank of England, National Bureau of Statistics of China, YAHOO!. tCode: 1: none; 2: ΔX_t; 3: $\Delta^2 X_t$; 4: $\log X_t$; 5: $\Delta \log X_t$; 6: $\Delta^2 \log X_t$; 7: $\Delta(X_t/X_{t-1} - 1)$

APPENDIX D

Supplementary Materials to "Approximate Factor Models with Strongly Correlated Idiosyncratic Errors."

D.1 Proofs for Statistical Error Bounds.

Before presenting the proof of Theorem 5.1, we first define a few quantities associated with the regularizers. Let $S^\star$ denote the support set of $B^\star$, and let the SVD of $\Theta^\star$ be $\Theta^\star = (U^\star)D^\star(V^\star)^\top$, with $U^\star_K$ and $V^\star_K$ respectively denoting the first K columns of $U^\star$ and $V^\star$. Let $\mathbb{S}$, $\mathbb{M}$ and their complements respectively be defined as follows:

$$\mathbb{S} := \big\{\Delta \in \mathbb{R}^{p\times p} \mid \Delta_{ij} = 0 \text{ for } (i,j) \notin S^\star\big\},$$
$$\mathbb{S}^c := \big\{\Delta \in \mathbb{R}^{p\times p} \mid \Delta_{ij} = 0 \text{ for } (i,j) \in S^\star\big\},$$

and

$$\mathbb{M} := \big\{\Delta \in \mathbb{R}^{n\times p} \mid \mathrm{row}(\Delta) \subseteq V^\star_K \text{ and } \mathrm{col}(\Delta) \subseteq U^\star_K\big\},$$
$$\mathbb{M}^\perp := \big\{\Delta \in \mathbb{R}^{n\times p} \mid \mathrm{row}(\Delta) \perp V^\star_K \text{ and } \mathrm{col}(\Delta) \perp U^\star_K\big\}.$$

Further, for some generic matrix $\Delta_1 \in \mathbb{R}^{p\times p}$, we define its projection on $\mathbb{S}$ and $\mathbb{S}^c$ (denoted by $\Delta_{1|\mathbb{S}}$ and $\Delta_{1|\mathbb{S}^c}$, resp.) as

$$\Delta_{1|\mathbb{S},ij} := \mathbf{1}\big\{(i,j) \in S^\star\big\}\Delta_{1,ij} \qquad \text{and} \qquad \Delta_{1|\mathbb{S}^c,ij} := \mathbf{1}\big\{(i,j) \notin S^\star\big\}\Delta_{1,ij}. \tag{D.1}$$

With the above definitions and projections, we can write

$$\Delta_1 = \Delta_{1|\mathbb{S}} + \Delta_{1|\mathbb{S}^c}, \qquad \|\Delta_1\|_1 = \|\Delta_{1|\mathbb{S}}\|_1 + \|\Delta_{1|\mathbb{S}^c}\|_1, \qquad \forall\ \Delta_1 \in \mathbb{R}^{p\times p}, \tag{D.2}$$

and note that the following inequality holds:

$$\|\Delta_{1|\mathbb{S}}\|_1 \le \sqrt{s}\,|||\Delta_{1|\mathbb{S}}|||_{\mathrm{F}} \le \sqrt{s}\,|||\Delta_1|||_{\mathrm{F}}. \tag{D.3}$$

as $\Delta_{1|S}$ has at most s nonzero entries where $s := |S^\star|$. In an analogous way, for some generic matrix $\Delta_2 \in \mathbb{R}^{n\times p}$, its projections on $\mathbb{M}$ and $\mathbb{M}^\perp$ (denoted by $\Delta_{2|\mathbb{M}}$ and $\Delta_{2|\mathbb{M}^\perp}$, resp.) are defined as

$$\Delta_{2|\mathbb{M}} := U^\star \begin{bmatrix} \widetilde{\Delta}_{2,11} & \widetilde{\Delta}_{2,12} \\ \widetilde{\Delta}_{2,21} & O \end{bmatrix} (V^\star)^\top \qquad \text{and} \qquad \Delta_{2|\mathbb{M}^\perp} := U^\star \begin{bmatrix} O & O \\ O & \widetilde{\Delta}_{2,22} \end{bmatrix} (V^\star)^\top, \tag{D.4}$$

where $\widetilde{\Delta}_2$ is defined below and partitioned as:

$$\widetilde{\Delta}_2 = (U^\star)^\top \Delta_2 (V^\star) = \begin{bmatrix} \widetilde{\Delta}_{2,11} & \widetilde{\Delta}_{2,12} \\ \widetilde{\Delta}_{2,21} & \widetilde{\Delta}_{2,22} \end{bmatrix}, \qquad \text{with } \widetilde{\Delta}_{2,11} \in \mathbb{R}^{K\times K}.$$

Note that the following relationships hold

$$\Delta_2 = \Delta_{2|\mathbb{M}} + \Delta_{2|\mathbb{M}^\perp}, \;\; |||\Delta_2|||_* = |||\Delta_{2|\mathbb{M}} + \Delta_{2|\mathbb{M}^\perp}|||_* = |||\Delta_{2|\mathbb{M}}|||_* + |||\Delta_{2|\mathbb{M}^\perp}|||_*, \quad \forall\, \Delta_2 \in \mathbb{R}^{n\times p}. \tag{D.5}$$

Next, we introduce concepts and lemmas regarding *decomposable regularizers* [**?**]. Define the weighted regularizer as

$$\mathcal{R}(B, \Theta) := \|B\|_1 + \frac{\lambda_\Theta}{\lambda_B} |||\Theta/\sqrt{n}|||_*,$$

and let $\Delta_B := \widehat{B} - B^\star$ and $\Delta_\Theta := \widehat{\Theta} - \Theta^\star$.

Lemma D.1. *With the definitions of projections in* (D.1) *and* (D.4)*, the following inequality holds:*

$$\mathcal{R}(B^\star, \Theta^\star) - \mathcal{R}(\widehat{B}, \widehat{\Theta}) \le \mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) - \mathcal{R}(\Delta_{B|\mathbb{S}^c}, \Delta_{\Theta|\mathbb{M}^\perp}).$$

Lemma D.2. *With the definition of* (D.4)*, the following holds for some generic* $\Delta \in \mathbb{R}^{n\times p}$*:*

$$\mathit{rank}(\Delta_{\mathbb{M}}) \le 2 \cdot \mathit{rank}(\Theta^\star).$$

The proofs of these two lemmas are deferred to Supplement D.2. Based on the above preparatory steps, we present next the proof of Theorem 5.1.

Proof of Theorem 5.1. We prove the bound for $\Delta_B := \widehat{B} - B^\star$ and $\Delta_\Theta := \widehat{\Theta} - \Theta^\star$ under the imposed regularity conditions, where $(\widehat{B}, \widehat{\Theta})$ is the solution to the optimization problem (5.6). Using the optimality of $(\widehat{B}, \widehat{\Theta})$ and the feasibility of $(B^\star, \Theta^\star)$, the following *basic inequality*

holds on:

$$\frac{1}{2n}|||\mathbf{X}_{n-1}\Delta_B^\top+\Delta_\Theta|||_{\mathrm{F}}^2 \leq \frac{1}{n}\Big(\langle\!\langle \Delta_B^\top, \mathbf{X}_{n-1}^\top\mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) + \lambda_B\Big(\|B^\star\|_1 - \|\widehat{B}\|_1\Big) + \lambda_\Theta\Big(|||\Theta^\star/\sqrt{n}|||_* - |||\widehat{\Theta}/\sqrt{n}|||_*\Big). \tag{D.6}$$

The LHS can be equivalently written as

$$\frac{1}{2n}|||\mathbf{X}_{n-1}\Delta_B^\top+\Delta_\Theta|||_{\mathrm{F}}^2 = \frac{1}{2n}\Big(|||\mathbf{X}_{n-1}\Delta_B^\top|||_{\mathrm{F}}^2 + |||\Delta_\Theta|||_{\mathrm{F}}^2 + 2\langle\!\langle \mathbf{X}_{n-1}\Delta_B^\top, \widehat{\Theta}-\Theta^\star\rangle\!\rangle\Big),$$

and by rearranging, (D.6) becomes

$$\frac{1}{2n}|||\mathbf{X}_{n-1}\Delta_B^\top|||_{\mathrm{F}}^2 + \frac{1}{2}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 \leq \frac{1}{n}\langle\!\langle \mathbf{X}_{n-1}\Delta_B^\top, \widehat{\Theta}-\Theta^\star\rangle\!\rangle + \frac{1}{n}\Big(\langle\!\langle \Delta_B^\top, \mathbf{X}_{n-1}^\top\mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) + \lambda_B\Big(\|B^\star\|_1 - \|\widehat{B}\|_1\Big) + \lambda_\Theta\Big(|||\Theta^\star/\sqrt{n}|||_* - |||\widehat{\Theta}/\sqrt{n}|||_*\Big). \tag{D.7}$$

Based on (D.7), the rest of the proof is divided into three parts: in part (i), we provide a lower bound for the LHS primarily using the RSC condition; in part (ii), we provide an upper bound for the RHS with the designated choice of λ_B and λ_Θ; in part (iii), we align the two sides and obtain the error bound after some rearrangement.

Part (i). In this part, we obtain a lower bound for the LHS of (D.7). Using the RSC condition for $\mathbf{X}_{n-1}$, the following lower bound holds for the LHS of (D.7):

$$\frac{1}{2n}|||\mathbf{X}_{n-1}\Delta_B^\top|||_{\mathrm{F}}^2 + \frac{1}{2}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 \geq \frac{\alpha_{\mathrm{RSC}}}{2}|||\Delta_B|||_{\mathrm{F}}^2 + \frac{1}{2}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 - \tau_n\|\Delta_B\|_1^2. \tag{D.8}$$

To further lower-bound (D.8), consider an upper bound for $\|\Delta_B\|_1$ with the aid of (D.6). By Hölder's inequality, the following inequalities hold for the inner products:

$$\frac{1}{n}\langle\!\langle \Delta_B^\top, \mathbf{X}_{n-1}^\top\mathbf{E}\rangle\!\rangle \leq \|\Delta_B\|_1\|\mathbf{X}_{n-1}^\top\mathbf{E}/n\|_\infty, \tag{D.9}$$

and

$$\frac{1}{n}\langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle \leq |||\Delta_\Theta/\sqrt{n}|||_*|||\mathbf{E}/\sqrt{n}|||_{op} = |||\Delta_\Theta/\sqrt{n}|||_*\Lambda_{\max}^{1/2}(S_{\mathbf{E}}). \tag{D.10}$$

By choosing $\lambda_B \geq 2\|\mathbf{X}_{n-1}^\top\mathbf{E}/n\|_\infty$ and $\lambda_\Theta \geq \Lambda_{\max}^{1/2}(S_{\mathbf{E}})$, the following inequality can be derived from the non-negativity of the RHS in (D.6):

$$\begin{aligned}0 &\leq \frac{\lambda_B}{2}\|\Delta_B\|_1 + \lambda_\Theta|||\Delta_\Theta/\sqrt{n}|||_* + \lambda_B\mathcal{R}(B^\star,\Theta^\star) - \lambda_B\mathcal{R}(\widehat{B},\widehat{\Theta})\\ &\overset{(1)}{\leq} \frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1 + \frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1 + \lambda_\Theta\Big|\Big|\Big|\frac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{n}}\Big|\Big|\Big|_* + \lambda_\Theta\Big|\Big|\Big|\frac{\Delta_{\Theta|\mathbb{M}^\perp}}{\sqrt{n}}\Big|\Big|\Big|_* + \lambda_B\Big(\mathcal{R}(\Delta_{B|\mathbb{S}},\Delta_{\Theta|\mathbb{M}}) - \mathcal{R}(\Delta_{B|\mathbb{S}^c},\Delta_{\Theta|\mathbb{M}^\perp})\Big),\end{aligned}$$

where the first two terms in (1) come from (D.2), the next two terms come from (D.5) and the last two terms use Lemma D.1. After writing out $\mathcal{R}(\cdot,\cdot)$ and rearranging, we obtain

$$\begin{aligned}\frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1 &\le \frac{3\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1 + 2\lambda_\Theta \Big\|\Big|\frac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{n}}\Big|\Big\|_*,\\ \frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1 + \frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1 &\le 2\lambda_B\|\Delta_{B|\mathbb{S}}\|_1 + 2\lambda_\Theta \Big\|\Big|\frac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{n}}\Big|\Big\|_*,\end{aligned}$$

that is,

$$\|\Delta_B\|_1 \le 2\cdot\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}). \tag{D.11}$$

Note that for $\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}})$, using (D.3) and Lemma D.2,

$$\begin{aligned}\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) &= \|\Delta_{B|\mathbb{S}}\|_1 + \frac{\lambda_\Theta}{\lambda_B}\Big\|\Big|\frac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{n}}\Big|\Big\|_* \le \sqrt{s}\,\|\!|\Delta_B|\mathbb{S}|\!\|_{\mathrm{F}} + \frac{\lambda_\Theta}{\lambda_B}(\sqrt{2K})\Big\|\Big|\frac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{n}}\Big|\Big\|_{\mathrm{F}}\\ &\le \sqrt{s}\,|\!|\!|\Delta_B|\!|\!|_{\mathrm{F}} + \frac{\lambda_\Theta}{\lambda_B}(\sqrt{2K})\,|\!|\!|\Delta_\Theta/\sqrt{n}|\!|\!|_{\mathrm{F}}.\end{aligned} \tag{D.12}$$

Plug (D.12) into (D.11), and by the Cauchy-Schwartz inequality, we have

$$\|\Delta_B\|_1^2 \le 4\big(s + (2K)(\lambda_\Theta/\lambda_B)^2\big)\Big[|\!|\!|\Delta_B|\!|\!|_{\mathrm{F}}^2 + |\!|\!|\Delta_\Theta/\sqrt{n}|\!|\!|_{\mathrm{F}}^2\Big]. \tag{D.13}$$

Combine (D.8) and (D.13), a lower bound for the LHS of (D.7) is given by

$$\Big[\frac{\alpha_{\mathrm{RSC}}}{2} - 4\tau_n\big(s + (2K)(\tfrac{\lambda_\Theta}{\lambda_B})^2\big)\Big]|\!|\!|\Delta_B|\!|\!|_{\mathrm{F}}^2 + \Big[\frac{1}{2} - 4\tau_n\big(s + (2K)(\tfrac{\lambda_\Theta}{\lambda_B})^2\big)\Big]|\!|\!|\Delta_\Theta/\sqrt{n}|\!|\!|_{\mathrm{F}}^2.$$

With the designated choice of τ_n satisfying $4\tau_n\big(s+(2K)(\frac{\lambda_\Theta}{\lambda_B})^2\big) \le \min\{\alpha_{\mathrm{RSC}}, 1\}/4$, the above bound can be further lower bounded by

$$\frac{\min\{\alpha_{\mathrm{RSC}}, 1\}}{4}\Big(|\!|\!|\Delta_B|\!|\!|_{\mathrm{F}}^2 + |\!|\!|\Delta_\Theta/\sqrt{n}|\!|\!|_{\mathrm{F}}^2\Big). \tag{D.14}$$

Part (ii). Next, we obtain an upper bound for the RHS of (D.7). Using the triangle inequality and Hölder's inequality, the first term satisfies

$$\begin{aligned}\frac{1}{n}|\langle\!\langle \mathbf{X}_{n-1}\Delta_B^\top, \widehat{\Theta} - \Theta^\star\rangle\!\rangle| &\le \frac{1}{n}|\langle\!\langle \mathbf{X}_{n-1}\Delta_B^\top, \widehat{\Theta}\rangle\!\rangle| + \frac{1}{n}|\langle\!\langle \mathbf{X}_{n-1}\Delta_B^\top, \Theta^\star\rangle\!\rangle|\\ &\le |\!|\!|\mathbf{X}_{n-1}\Delta_B^\top/\sqrt{n}|\!|\!|_{\mathrm{op}}\,|\!|\!|\widehat{\Theta}/\sqrt{n}|\!|\!|_* + |\!|\!|\mathbf{X}_{n-1}\Delta_B^\top/\sqrt{n}|\!|\!|_{\mathrm{op}}\,|\!|\!|\Theta^\star/\sqrt{n}|\!|\!|_*.\end{aligned}$$

For $|\!|\!|\mathbf{X}_{n-1}\Delta_B^\top/\sqrt{n}|\!|\!|_{\mathrm{op}}$, we have

$$\frac{1}{n}|\!|\!|\mathbf{X}_{n-1}\Delta_B^\top|\!|\!|_{\mathrm{op}} \le |\!|\!|\mathbf{X}_{n-1}/\sqrt{n}|\!|\!|_{\mathrm{op}}\,|\!|\!|\Delta_B^\top|\!|\!|_{\mathrm{op}} \le |\!|\!|\mathbf{X}_{n-1}/\sqrt{n}|\!|\!|_{\mathrm{op}}\,|\!|\!|\Delta_B^\top|\!|\!|_{\mathrm{F}} = \Lambda_{\max}^{1/2}(S_{\mathbf{X}})\,|\!|\!|\Delta_B|\!|\!|_{\mathrm{F}},$$

where the first inequality comes from the sub-multiplicativity of the nuclear norm. Together with the constraint on the feasible region and by choosing $\lambda_B \geq 2\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})$, we have

$$\frac{1}{n}|\langle\!\langle \mathbf{X}_{n-1}\Delta_B^\top, \widehat{\Theta} - \Theta^\star \rangle\!\rangle| \leq 2\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})|||\Delta_B|||_{\mathrm{F}} \leq \lambda_B|||\Delta_B|||_{\mathrm{F}}.$$

With (D.9) and (D.10), by choosing $\lambda_B \geq \max\{\|\mathbf{X}_{n-1}^\top \mathbf{E}/n\|_\infty,\ 2\phi\Lambda_{\max}^{1/2}(S_{\mathbf{E}})\}$ and $\lambda_\Theta \geq \Lambda_{\max}^{1/2}(S_{\mathbf{E}})$, the following upper bound holds for the RHS:

$$\begin{aligned}
& \lambda_B\|\Delta_B\|_1 + \lambda_\Theta\|\Delta_\Theta\|_* + \lambda_B|||\Delta_B|||_{\mathrm{F}} + \lambda_B\big(\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) - \mathcal{R}(\Delta_{B|\mathbb{S}^c}, \Delta_{\Theta|\mathbb{M}^\perp})\big) \\
\overset{(1)}{\leq}\ & \lambda_B\big(\|\Delta_{B|\mathbb{S}}\|_1 + \|\Delta_{B|\mathbb{S}^c}\|_1\big) + \lambda_\Theta\big(|||\tfrac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{n}}|||_* + |||\tfrac{\Delta_{\Theta|\mathbb{M}^\perp}}{\sqrt{n}}|||_*\big) + \lambda_B|||\Delta_B|||_{\mathrm{F}} \\
& \qquad + \lambda_B\big(\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) - \mathcal{R}(\Delta_{B|\mathbb{S}^c}, \Delta_{\Theta|\mathbb{M}^\perp})\big) \\
\overset{(2)}{\leq}\ & 2\lambda_B\|\Delta_{B|\mathbb{S}}\|_1 + 2\lambda_\Theta|||\tfrac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{n}}|||_* + \lambda_B|||\Delta_B|||_{\mathrm{F}} = (2\lambda_B)\cdot\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) + \lambda_B|||\Delta_B|||_{\mathrm{F}} \\
\overset{(3)}{\leq}\ & (2\lambda_B)(\sqrt{s}+1)|||\Delta_B|||_{\mathrm{F}} + (2\lambda_\Theta)\sqrt{2K}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}},
\end{aligned}$$

where (1) uses (D.2) and (D.5); (2) is obtained by writing out $\mathcal{R}(\cdot,\cdot)$ and canceling terms; and (3) uses (D.12). Further by the Cauchy-Schwartz inequality, an upper bounded for the RHS is given by

$$\sqrt{4\lambda_B^2(\sqrt{s}+1)^2 + 4\lambda_\Theta^2(2K)}\sqrt{|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2} \tag{D.15}$$

Part (iii). Combine (D.14) and (D.15), and rearrange terms, the following bound directly follows:

$$\begin{aligned}
\frac{\min\{\alpha_{\mathrm{RSC}}, 1\}}{4}\Big(|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2\Big) &\leq \sqrt{4\lambda_B^2(\sqrt{s}+1)^2 + 4\lambda_\Theta^2(2K)}\sqrt{|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2}\ , \\
|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 &\leq \frac{64\Big[\lambda_B^2(\sqrt{s}+1)^2 + \lambda_\Theta^2(2K)\Big]}{\min\{\alpha_{\mathrm{RSC}}, 1\}^2}\ .
\end{aligned}$$

□

Proof of Proposition 5.1. First, we note that for any given $\widehat{\Theta} = \Theta^\star + \Delta_\Theta$, it can be viewed as a Δ_Θ-perturbation with respect to the true $\Theta^\star$. As mentioned in the main text, as invertible linear transformations preserve the subspace, so does scaling (with a non-zero scale factor), it is equivalent to examining the $\sin\theta$ distance between the first K singular vectors of $\widehat{\Theta}$ and $\Theta^\star$ (denoted by $\widehat{U}$ and $U^\star$, resp.). The rest follows seamlessly from the perturbation theory of singular vectors. Specifically, by applying **?** , Theorem 3] and assuming the singular values of $\Theta^\star$ are given by $\sigma_1 > \sigma_2 > \cdots > \sigma_K > \sigma_{K+1} = \cdots = \sigma_{n,p} = 0$, the following bound

holds for $|||\sin(\widehat{U}, U^\star)|||$:

$$|||\sin\theta(\widehat{U}, U^\star)|||_{\mathrm{F}}^2 \le \frac{2(2\sigma_1 + |||\Delta_\Theta|||_{\mathrm{op}}) \min\Big\{\sqrt{K}|||\Delta_\Theta|||_{\mathrm{op}}, |||\Delta_\Theta|||_{\mathrm{F}}\Big\}}{\sigma_K^2}.$$

Note that the same bound holds for the $\sin\theta$ distance between the factor spaces. □

Proof of Lemma 5.1. First, suppose we have

$$\frac{1}{2}v^\top S_{\mathbf{X}} v = \frac{1}{2}v^\top\Big(\frac{\mathbf{X}^\top\mathbf{X}}{n}\Big)v \ge \frac{\alpha_{\mathrm{RSC}}}{2}\|v\|_2^2 - \tau_n\|v\|_1^2, \quad \forall\ v \in \mathbb{R}^p; \tag{D.16}$$

then, for all $\Delta \in \mathbb{R}^{p\times p}$, and letting Δ_j denote its jth column, the RSC condition automatically holds since

$$\frac{1}{2n}|||\mathbf{X}\Delta|||_{\mathrm{F}}^2 = \frac{1}{2}\sum_{j=1}^q \Delta_j^\top\big(\tfrac{\mathbf{X}^\top\mathbf{X}}{n}\big)\Delta_j \ge \frac{\alpha_{\mathrm{RSC}}}{2}\sum_{j=1}^q\|\Delta_j\|_2^2 - \tau_n\sum_{j=1}^q\|\Delta_j\|_1^2 \ge \frac{\alpha_{\mathrm{RSC}}}{2}|||\Delta|||_{\mathrm{F}}^2 - \tau_n\|\Delta\|_1^2.$$

Therefore, it suffices to verify that (D.16) holds. In **?** , Proposition 4.2], the authors prove a similar result under the assumption that X_t is a VAR(d) process. Here, we adopt the same proof strategy and state the result for a *more general process* X_t.

Specifically, by **?** , Proposition 2.4(a)], $\forall v \in \mathbb{R}^p, \|v\| \le 1$ and $\eta > 0$,

$$\mathbb{P}\Big[\big|v^\top\big(S_{\mathbf{X}} - \Gamma_X(h)\big)v\big| > 2\pi\mathcal{M}(g_X)\eta\Big] \le 2\eta\exp\Big(-cn\min\{\eta^2, \eta\}\Big).$$

Applying the discretization in **?** , Lemma F.2] and taking the union bound, define $\mathbb{K}(2s) := \{v \in \mathbb{R}^p, \|v\| \le 1, \|v\|_0 \le 2k\}$, and the following inequality holds:

$$\mathbb{P}\Big[\sup_{v\in\mathbb{K}(2k)}\big|v^\top\big(S_{\mathbf{X}} - \Gamma_X(h)\big)v\big| > 2\pi\mathcal{M}(g_X)\eta\Big] \le 2\exp\Big(-cn\min\{\eta, \eta^2\} + 2k\min\{\log p, \log(21ep/2k)\}\Big).$$

With the specified $\gamma = 54\mathcal{M}(g_X)/\mathfrak{m}(g_X)$, set $\eta = \gamma^{-1}$, then apply results from **?** , Lemma 12] with $\Gamma = S_{\mathbf{X}} - \Gamma_X(0)$ and $\delta = \pi\mathfrak{m}(g_X)/27$, so that the following holds

$$\frac{1}{2}v' S_{\mathbf{X}} v \ge \frac{\alpha_{\mathrm{RSC}}}{2}\|v\|^2 - \frac{\alpha_{\mathrm{RSC}}}{2k}\|v\|_1^2,$$

with probability at least $1 - 2\exp\big(-cn\min\{\gamma^{-2}, 1\} + 2k\log p\big)$ and note that $\min\{\gamma^{-2}, 1\} = \gamma^{-2}$ since $\gamma > 1$. Finally, let $k = \min\{cn\gamma^{-2}/(c'\log p), 1\}$ for some $c' > 2$, and conclude that with probability at least $1 - c_1\exp(-c_2 n)$, the inequality in (D.16) holds with

$$\alpha_{\mathrm{RSC}} = \pi\mathfrak{m}(g_X), \qquad \tau_n = \alpha_{\mathrm{RSC}}\gamma^2\frac{\log p}{2n},$$

and so does also the RSC condition. □

Proof of Lemma 5.2. We note that

$$\frac{1}{n}\big\|\mathbf{X}_{n-1}^{\top}\mathbf{E}\big\|_{\infty} = \max_{1\le i,j\le p}\big|e_i^{\top}\big(\mathbf{X}_{n-1}^{\top}\mathbf{E}/n\big)e_j\big|,$$

where e_i is the p-dimensional standard basis with the ith entry being 1. Applying **?** , Proposition 2.4(b)], for an arbitrary pair of (i, j), the following inequality holds:

$$\mathbb{P}\Big[\big|e_i^{\top}\big(\mathbf{X}_{n-1}^{\top}\mathbf{E}/n\big)e_j\big| > 2\pi\big(\mathcal{M}(g_X)+\mathcal{M}(g_\epsilon)+\mathcal{M}(g_{X,\tilde{\epsilon}})\big)\eta\Big] \le 6\exp\Big(-cn\min\{\eta^2,\eta\}\Big).$$

Take the union bound over all $1 \le i, j \le p$, and the following bound holds:

$$\mathbb{P}\Big[\max_{1\le i,j\le p}\big|e_i^{\top}\big(\mathbf{X}_{n-1}^{\top}\mathbf{E}/n\big)e_j\big| > 2\pi\big(\mathcal{M}(g_X)+\mathcal{M}(g_\epsilon)+\mathcal{M}(g_{X,\tilde{\epsilon}})\big)\eta\Big] \le 6\exp\Big(-cn\min\{\eta^2,\eta\}+2\log p\Big).$$

Set $\eta = c'\sqrt{\log p/n}$ for $c' > (2/c)$ and with the choice of $n \gtrsim \log p$, $\min\{\eta^2, \eta\} = \eta^2$, then with probability at least $1 - c_1\exp(-c_2\log p)$, the following bound holds:

$$\frac{1}{n}\big\|\mathbf{X}_{n-1}^{\top}\mathbf{E}\big\|_{\infty} \le c_0\big(\mathcal{M}(g_X)+\mathcal{M}(g_\epsilon)+\mathcal{M}(g_{X,\tilde{\epsilon}})\big)\sqrt{\frac{\log p}{n}}.$$

□

Before proving Lemma 5.3, we first state Lemma D.3 which provides a concentration inequality in the operator norm.

Lemma D.3. *Consider the stationary centered Gaussian process $\{X_t\} \in \mathbb{R}^p$, whose spectral density function $g_X(\omega)$ exits and the maximum eigenvalue is bounded a.e. on $[-\pi, \pi]$. Then for $\mathbf{X}$ whose rows are random realizations $\{x_0, \ldots, x_{n-1}\}$ of $\{X_t\}$, the following bound holds for $S_{\mathbf{X}} = \mathbf{X}^{\top}\mathbf{X}/n$, for some $c > 0$:*

$$\mathbb{P}\Big[\|\!|S_{\mathcal{X}} - \Gamma_X(0)|\!\|_{op} > 4\pi\mathcal{M}(g_X)\eta\Big] \le 2\exp(-cn\min\{\eta,\eta^2\} + p\log 8).$$

The proof of this Lemma is deferred to Supplement D.2.

Proof of Lemma 5.3. The result follows in straightforward manner based on Lemma D.3. Specifically, by letting $\eta = c'\sqrt{p/n}$ for $c' > (\log 8/c)$ and with $n \gtrsim p$ so that $\min\{\eta^2, \eta\} = \eta^2$, then if we relax $\Lambda_{\max}(\Gamma_X(0))$ by its upper bound $2\pi\mathcal{M}(g_X)$ [**?** , Proposition 2.3], with probability at least $1 - c_1\exp(-c_2 p)$, the following bound holds for some c_0:

$$\Lambda_{\max}(S_{\mathbf{X}}) \le c_0\mathcal{M}(g_X).$$

□

Proof of Lemma 5.4. For $\mathbf{E}$ whose rows are iid realizations of a sub-Gaussian random vector ϵ_t, by **?** , Lemma 9], the following bound holds:

$$\mathbb{P}\Big[\big|\big|\big|S_{\mathbf{E}} - \Sigma_\epsilon\big|\big|\big|_{op} \geq \Lambda_{\max}(\Sigma_\epsilon)\delta(n,p,\eta)\Big] \leq 2\exp(-n\eta^2/2),$$

where $\delta(n,p,\eta) := 2\big(\sqrt{\frac{p}{n}} + \eta\big) + \big(\sqrt{\frac{p}{n}} + \eta\big)^2$. In particular, by triangle inequality, with probability at least $1 - 2\exp(-n\eta^2/2)$,

$$\big|\big|\big|S_{\mathbf{E}}\big|\big|\big|_{op} \leq \big|\big|\big|\Sigma_\epsilon\big|\big|\big|_{op} + \big|\big|\big|S_{\mathbf{E}} - \Sigma_\epsilon\big|\big|\big|_{op} \leq \Lambda_{\max}(\Sigma_\epsilon) + \Lambda_{\max}(\Sigma_\epsilon)\delta(n,p,t).$$

So for $n \geq p$, by setting $\eta = 1$, which yields $\delta(n,p,\eta) \leq 8$ so that with probability at least $1 - 2\exp(-n/2)$, the following bound holds:

$$\Lambda_{\max}(S_{\mathbf{E}}) \leq 9\Lambda_{\max}(\Sigma_\epsilon).$$

□

D.2 Proofs of Auxiliary Lemmas.

Next, proofs of auxiliary lemmas D.1, D.2 and D.3 are provide. Variations of these Lemmas have been proved in **?**] and **?**]; nevertheless, we provide them also here for the sake of completeness.

Proof of Lemma D.1. First note that with the definition of (D.1) and (D.4), $B^\star_{\mathbb{S}^c} = 0$ and $\Theta^\star_{\mathbb{M}^\perp} = 0$. Then

$$\mathcal{R}(B^\star, \Theta^\star) = \|B^\star_{\mathbb{S}} + B^\star_{\mathbb{S}^c}\|_1 + \frac{\lambda_\Theta}{\lambda_B}\big|\big|\big|\Theta^\star_{\mathbb{M}} + \Theta^\star_{\mathbb{M}^\perp}\big|\big|\big|_* = \|B^\star_{\mathbb{S}}\|_1 + \frac{\lambda_\Theta}{\lambda_B}\big|\big|\big|\Theta^\star_{\mathbb{M}}\big|\big|\big|_*,$$

and

$$\begin{aligned}
\mathcal{R}(\widehat{B},\widehat{\Theta}) &= \mathcal{R}(B^\star+\Delta_B,\Theta^\star+\Delta_\Theta)\\
&= \|B^\star_{\mathbb{S}}+B^\star_{\mathbb{S}^c}+\Delta_{B|\mathbb{S}}+\Delta_{B|\mathbb{S}^c}\|_1+\frac{\lambda_\Theta}{\lambda_B}|||\Theta^\star_{\mathbb{M}}+\Theta^\star_{\mathbb{M}^\perp}+\Delta_{\Theta|\mathbb{M}}+\Delta_{\Theta|\mathbb{M}^\perp}|||_*\\
&\geq \|B^\star_{\mathbb{S}}+\Delta_{B|\mathbb{S}^c}\|_1-\|\Delta_{B|\mathbb{S}}\|_1+\frac{\lambda_\Theta}{\lambda_B}\Big(|||\Theta^\star_{\mathbb{M}}+\Delta_{\Theta|\mathbb{M}^\perp}|||_*-|||\Delta_{\Theta|\mathbb{M}}|||_*\Big)\\
&\overset{(i)}{\geq} \|B^\star_{\mathbb{S}}\|_1+\|\Delta_{B|\mathbb{S}^c}\|_1-\|\Delta_{B|\mathbb{S}}\|_1+\frac{\lambda_\Theta}{\lambda_B}\Big(|||\Theta^\star_{\mathbb{M}}|||_*+|||\Delta_{\Theta|\mathbb{M}^\perp}|||_*-|||\Delta_{\Theta|\mathbb{M}}|||_*\Big)\\
&\geq \mathcal{R}(B^\star_{\mathbb{S}},\Theta^\star_{\mathbb{M}})+\mathcal{R}(\Delta_{B|\mathbb{S}^c},\Delta_{\Theta|\mathbb{M}^\perp})-\mathcal{R}(\Delta_{B|\mathbb{S}},\Delta_{\Theta|\mathbb{M}})\\
&= \mathcal{R}(B^\star,\Theta^\star)+\mathcal{R}(\Delta_{B|\mathbb{S}^c},\Delta_{\Theta|\mathbb{M}^\perp})-\mathcal{R}(\Delta_{B|\mathbb{S}},\Delta_{\Theta|\mathbb{M}}).
\end{aligned}$$

where (i) uses the property of decomposable regularizers. By rearranging, we obtain the desired inequality. □

Proof of Lemma D.2. Let the SVD of $\Theta^\star$ be given by $\Theta^\star=(U^\star)D(V^\star)^\top$, where both $U^\star$ and $V^\star$ are orthogonal matrices. Assume $\mathrm{rank}(\Theta^\star)=K$. For $\Delta\in\mathbb{R}^{n\times p}$, define $\widetilde{\Delta}$ as below and it is partitioned as:

$$\widetilde{\Delta}:=(U^\star)^\top\Delta(V^\star)=\begin{bmatrix}\widetilde{\Delta}_{11} & \widetilde{\Delta}_{12}\\ \widetilde{\Delta}_{21} & \widetilde{\Delta}_{22}\end{bmatrix},\qquad \text{where } \widetilde{\Delta}_{11}\in\mathbb{R}^{K\times K}.$$

Then by further defining

$$\Delta_{\mathbb{M}}:=U^\star\begin{bmatrix}\widetilde{\Delta}_{11} & \widetilde{\Delta}_{12}\\ \widetilde{\Delta}_{21} & O\end{bmatrix}(V^\star)^\top\qquad\text{and}\qquad \Delta_{\mathbb{M}^\perp}:=U^\star\begin{bmatrix}O & O\\ O & \widetilde{\Delta}_{22}\end{bmatrix}(V^\star)^\top,$$

it is straightforward to see that $\Delta_{\mathbb{M}}+\Delta_{\mathbb{M}^\perp}=\Delta$. Moreover,

$$\mathrm{rank}(\Delta_{\mathbb{M}})\leq \mathrm{rank}\left(U^\star\begin{bmatrix}\widetilde{\Delta}_{11} & \widetilde{\Delta}_{12}\\ O & O\end{bmatrix}(V^\star)^\top\right)+\mathrm{rank}\left(U\begin{bmatrix}\widetilde{\Delta}_{11} & O\\ \widetilde{\Delta}_{21} & O\end{bmatrix}(V^\star)^\top\right)\leq 2K.$$

□

Proof of Lemma D.3. First, we note that by **?** , Proposition 2.4], the following inequality holds for any fixed $v\in S^p$, where $S^p:=\{v\in\mathbb{R}^p:\|v\|=1\}$ is the p-dimensional unit sphere:

$$\mathbb{P}\Big[\big|v'\big(S_{\mathbf{X}}-\Gamma_X(0)\big)v\big|>2\pi\mathcal{M}(g_X)\eta\Big]\leq 2\exp(-cn\min\{\eta,\eta^2\}).\tag{D.17}$$

Additionally, by **?** , Lemma 5.4],

$$\|\|\|S_{\mathbf{X}} - \Gamma_X(0)\|\|\|_{op} = \sup_{v \in S^p} \big|v'\big(S_{\mathbf{X}} - \Gamma_X(0)\big)v\big| \leq (1-2\delta)^{-1} \sup_{v \in \mathcal{N}_\delta} v'\big[S_{\mathbf{X}} - \Gamma_X(0)\big]v,$$

where $\mathcal{N}_\delta$ is a δ-net of S^p for some $\delta \in [0,1)$, which guarantees that the sphere can essentially be replaced by its δ-net whose cardinality is finite. Toward this end, based upon (D.17), take the union bound over all vectors v in the $\frac{1}{4}$-net of S^p, whose cardinality is at most 8^p [e.g. **?**], we have

$$\begin{aligned}\mathbb{P}\Big[\|\|\|\tfrac{1}{n}X'X - \Gamma_X(0)\|\|\|_{op} > 4\pi\mathcal{M}(g_X)\eta\Big] &\leq \mathbb{P}\Big[\sup_{v \in \mathcal{N}_\delta} \big|v'\big(S - \Gamma_X(0)\big)v\big| > 4\pi\mathcal{M}(g_X)\eta\Big] \\ &\leq 8^p \cdot 2\exp\Big(-cn\min\{\eta, \eta^2\}\Big).\end{aligned}$$

□

D.3 Analyses for the Non-Convex Formulation.

In this section, we briefly analyze the statistical properties and the computational procedure corresponding to the non-convex formulation in (5.16), that is,

$$\begin{aligned}(\widetilde{B}, \widetilde{\Theta}) = \mathop{\arg\min}_{B \in \mathbb{R}^{p\times p}, \Theta \in \mathbb{R}^{n\times p}} &\Big\{\frac{1}{2n}\|\|\mathbf{X}_n - \Theta - \mathbf{X}_{n-1}B^\top\|\|_{\mathrm{F}}^2 + \lambda_B\|B\|_1\Big\}, \\ \text{subject to} \quad &\mathrm{rank}(\Theta) \leq r, \quad \|\|\Theta/\sqrt{n}\|\|_* \leq \phi.\end{aligned}$$

As a remark, in this formulation, since we have specified the maximum rank allowed for Θ, which is equivalently to specifying the maximum cardinality of the support of the eigenspectrum, the nuclear norm ball constraint can be alternatively changed to a box constraint through the operator norm, namely $\|\|\Theta/\sqrt{n}\|\|_{\mathrm{op}} \leq \phi$, with the choice of the tuning parameter λ_B modified accordingly.

D.3.1 Statistical error bound.

We assume the true value of the parameters $(B^\star, \Theta^\star)$ is always feasible, which automatically imposes the assumption $r \geq K$ for the rank constraint r. The following theorem gives the error bound of $(\widetilde{B}, \widetilde{\Theta})$ with deterministic realizations.

Theorem D.1 (Error bound for $(\widetilde{B}, \widetilde{\Theta})$ under fixed realizations)**.** *Suppose the fixed realizations $\mathbf{X}_{n-1} \in \mathbb{R}^{n\times p}$ of process $X_t \in \mathbb{R}^p$ satisfying the RSC condition with curvature $\alpha_{RSC} > 0$*

and a tolerance τ_n for which

$$\tau_n\big(K+r+4s\big) < \min\{\alpha_{RSC}, 1\}/16.$$

Then, for any matrix pair $(B^\star, \Theta^\star)$ that drives the dynamics of X_t, for estimators $(\widetilde{B}, \widetilde{\Theta})$ obtained by solving the optimization (5.16) *with regularization parameters λ_B satisfying*

$$\lambda_B \geq \max\Big\{2\|\mathbf{X}_{n-1}^\top \mathbf{E}/n\|_\infty,\ \Lambda_{\max}^{1/2}(S_{\mathbf{E}}),\ 2\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})\Big\},$$

the following error bound holds:

$$\big|\!\big|\!\big|\widetilde{B} - B^\star\big|\!\big|\!\big|_F^2 + \big|\!\big|\!\big|(\widetilde{\Theta} - \Theta^\star)/\sqrt{n}\big|\!\big|\!\big|_F^2 \leq \frac{16\lambda_B^2\Big((K+r) + (2\sqrt{s}+1)^2\Big)}{\min\{\alpha_{RSC}, 1\}^2}.$$

Proof sketch of Theorem D.1. For notation convenience and consistency, here we still let $\Delta_B := \widetilde{B} - B^\star$ and $\Delta_\Theta := \widetilde{\Theta} - \Theta^\star$ denote the errors. Using the optimality of $(\widetilde{B}, \widetilde{\Theta})$ and the feasibility of $(B^\star, \Theta^\star)$, we obtain the following *basic inequality*:

$$\frac{1}{2n}\big|\!\big|\!\big|\mathbf{X}_{n-1}\Delta_B^\top + \Delta_\Theta\big|\!\big|\!\big|_{\mathrm{F}}^2 \leq \frac{1}{n}\Big(\langle\!\langle \Delta_B^\top, \mathbf{X}_{n-1}^\top \mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) + \lambda_B\Big(\|B^\star\|_1 - \|\widetilde{B}\|_1\Big). \tag{D.18}$$

which by rearrangement gives

$$\begin{aligned}\frac{1}{2n}\big|\!\big|\!\big|\mathbf{X}_{n-1}\Delta_B^\top\big|\!\big|\!\big|_{\mathrm{F}}^2 + \frac{1}{2}\big|\!\big|\!\big|\Delta_\Theta/\sqrt{n}\big|\!\big|\!\big|_{\mathrm{F}}^2 \leq \frac{1}{n}\langle\!\langle \mathbf{X}_{n-1}\Delta_B^\top, \widetilde{\Theta} - \Theta^\star\rangle\!\rangle + \frac{1}{n}\Big(\langle\!\langle \Delta_B^\top, \mathbf{X}_{n-1}^\top \mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) \\ + \lambda_B\Big(\|B^\star\|_1 - \|\widetilde{B}\|_1\Big).\end{aligned} \tag{D.19}$$

Similar to the proof of Theorem 5.1, based upon (D.19), the rest of the proof is divided into three parts: in part (i), we provide an lower bound for the LHS which primarily uses the RSC condition; in part (ii), we provide an upper bound for the RHS with the designated choice of λ_B; in part (iii), we align the two sides and obtain the error bound after some rearrangement.

Part (i). As RSC holds for $\mathbf{X}_{n-1}$, the first term on the LHS of (D.18) is lower bounded by

$$\frac{\alpha_{\mathrm{RSC}}}{2}\big|\!\big|\!\big|\Delta_B\big|\!\big|\!\big|_{\mathrm{F}}^2 - \tau_n\|\Delta_B\|_1^2. \tag{D.20}$$

Consider an upper bound for $\|\Delta_B\|_1$. Using the non-negativity of the RHS in (D.18) and with the designated choice of λ_B, the following inequality holds:

$$0 \le \frac{\lambda_B}{2}\|\Delta_B\|_1+\lambda_B|||\Delta_\Theta/\sqrt{n}|||_*+\lambda_B(\|\Delta_{B|\mathbb{S}}\|_1-\|\Delta_{B|\mathbb{S}^c}\|_1) = \frac{3\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1-\frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1+\lambda_B|||\Delta_\Theta/\sqrt{n}|||_*.$$

Since $\Delta_\Theta = \widetilde{\Theta} - \Theta^\star$ has rank at most $K+r$, $|||\Delta_\Theta/\sqrt{n}|||_* \le \sqrt{K+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}$. It follows that

$$\begin{aligned}
&\frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1 \le \lambda_B\sqrt{K+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + \frac{3\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1,\\
&\frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1 + \frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1 \le \lambda_B\sqrt{K+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + \frac{3\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1 + \frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1,\\
&\|\Delta_B\|_1 \le \sqrt{4(K+r)}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + 4\|\Delta_{B|\mathbb{S}}\|_1 \le \sqrt{4(K+r)}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + 4\sqrt{s}|||\Delta_B|||_{\mathrm{F}},
\end{aligned}$$

where the second line is obtained by adding $\frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1$ on both sides, and the last inequality uses (D.3). Further, by the Cauchy-Schwartz inequality, we have

$$\|\Delta_B\|_1 \le \sqrt{(\sqrt{4(K+r)})^2+(4\sqrt{s})^2}\sqrt{|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2},$$

that is,

$$\|\Delta_B\|_1^2 \le 4\big(K+r+4s\big)\Big[|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2\Big]. \tag{D.21}$$

Combine (D.20) and (D.21), a lower bound for the LHS of (D.19) is given by

$$\Big(\frac{\alpha_{\mathrm{RSC}}}{2} - 4\tau_n\big(K+r+4s\big)\Big)|||\Delta_B|||_{\mathrm{F}}^2 + \Big(\frac{1}{2} - 4\tau_n\big(K+r+4s\big)\Big)|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2. \tag{D.22}$$

Part (ii). Similar to the derivation in the proof of Theorem 5.1, with the required choice of λ_B, the following upper bound holds for the RHS of (D.19):

$$\begin{aligned}
&\lambda_B|||\Delta_B|||_{\mathrm{F}} + \lambda_B\|\Delta_B\|_1 + \lambda_B\sqrt{K+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + \lambda_n(\|\Delta_{B|\mathbb{S}}\|_1 - \|\Delta_{B|\mathbb{S}^c}\|_1)\\
\le\ & \lambda_B\Big((2\sqrt{s}+1)|||\Delta_B|||_{\mathrm{F}} + \sqrt{K+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}\Big)\\
\le\ & \lambda_B\sqrt{(2\sqrt{s}+1)^2+(K+r)^2}\sqrt{|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2}.
\end{aligned} \tag{D.23}$$

Part (iii). Combine (D.22) and (D.23), by rearranging and requiring τ_n satisfying $\tau_n(K+r+4s) < \min\{\alpha_{\mathrm{RSC}},1\}/16$, the following inequality holds:

$$\frac{\min\{\alpha_{\mathrm{RSC}},1\}}{4}\Big(|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2\Big) \le \lambda_B\sqrt{(2\sqrt{s}+1)^2+(K+r)^2}\sqrt{|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2},$$

which gives

$$\|\!|\Delta_B|\!\|_{\mathrm{F}}^2 + \|\!|\Delta_\Theta/\sqrt{n}|\!\|_{\mathrm{F}}^2 \leq \frac{16\lambda_B^2\Big((K+r)+(2\sqrt{s}+1)^2\Big)}{\min\{\alpha_{\mathrm{RSC}},1\}^2}.$$

□

D.3.2 A majorization-minimization algorithm.

We introduce a majorization-minimization (MM) algorithm that solves the penalized (nonconvex) formulation corresponding to the non-convex program (5.16) with convergence guarantees. Let $\sigma_1 > \cdots > \sigma_r > \sigma_{r+1} > \cdots > \sigma_{\min\{n,p\}}$ be the singular values of Θ, then (5.16) is equivalent to

$$\begin{aligned} \min_{B,\Theta} \quad & \Big\{\frac{1}{2n}\|\!|\mathbf{X}_n - \Theta - \mathbf{X}_{n-1}B^\top|\!\|_{\mathrm{F}}^2 + \lambda_B\|B\|_1\Big\}, \\ \text{subject to} \quad & \|\!|\Theta/\sqrt{n}|\!\|_* \leq \phi, \\ & \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_{\min\{n,p\}} = 0, \end{aligned}$$

whose penalized reformulation can be written as

$$\begin{aligned} \min_{B,\Theta} \quad & \Big\{\frac{1}{2n}\|\!|\mathbf{X}_n - \Theta - \mathbf{X}_{n-1}B^\top|\!\|_{\mathrm{F}}^2 + \lambda_B\|B\|_1 + \frac{\rho}{2}\sum_{k=r+1}^{\min\{n,p\}}\sigma_k^2\Big\}, \\ \text{subject to} \quad & \|\!|\Theta/\sqrt{n}|\!\|_* \leq \phi. \end{aligned} \tag{D.24}$$

Let the SVD of Θ be $\Theta = UDV^\top$ and define $\mathrm{I}_r := \left[\begin{smallmatrix} \mathrm{I}_{r\times r} & O \\ O & O \end{smallmatrix}\right]$; then the following holds:

$$\sum_{k=r+1}^{\min\{n,p\}}\sigma_k^2 = \sum_{k=1}^{\min\{n,p\}}\sigma_k^2 - \sum_{k=1}^{r}\sigma_k^2 = \|\Theta\|_F^2 - \mathrm{Tr}\big[\Theta^\top U I_r U^\top \Theta\big].$$

Therefore, the objective function in (D.24) can be written as

$$H(B,\Theta) := \underbrace{\frac{1}{2n}\|\!|\mathbf{X}_n - \Theta - \mathbf{X}_{n-1}B^\top|\!\|_{\mathrm{F}}^2 + \lambda_B\|B\|_1 + \frac{\rho}{2}\|\!|\Theta|\!\|_{\mathrm{F}}^2}_{h_1(B,\Theta)} - \underbrace{\frac{\rho}{2}\mathrm{Tr}\big[\Theta^\top U I_r U^\top \Theta\big]}_{h_2(\Theta)},$$

where both h_1 and h_2 are convex. In other words, the objective function in (D.24) can be expressed as the difference of two convex functions and has a convex feasible region. At iteration $m+1$, a minorizer for $h_2(\Theta)$ is given by

$$h(\Theta|\Theta^{(m)}) \geq h_2(\Theta^{(m)}) + \nabla h_2(\Theta^{(m)})^\top(\Theta - \Theta^{(m)}),$$

with equality attained at $\Theta = \Theta^{(m)}$, and $\nabla h_2(\Theta) := 2UI_rU^\top\Theta$. As a consequence, at iteration $m+1$, a majarizer for $H(B,\Theta)$ is given by

$$H\big(B,\Theta|(B^{(m)},\Theta^{(m)})\big) \leq h_1(B,\Theta) - \nabla h_2(\Theta^{(m)})^\top\Theta - h_2(\Theta^{(m)}) + \nabla h_2(\Theta^{(m)})^\top\Theta^{(m)}.$$

This motivates the following iterative joint update of the two blocks:

$$(B^{(m+1)},\Theta^{(m+1)}) = \underset{B,\Theta\in\mathbb{B}_n(\phi)}{\arg\min}\Big\{h_1(B,\Theta) - \nabla h_2(\Theta^{(m)})^\top\Theta\Big\} \tag{D.25}$$

where $\mathbb{B}_n(\phi) := |||\Theta/\sqrt{n}|||_* \leq \phi$ and the objective function in (D.25) is jointly convex in the two blocks. Empirically, a search over an increasing sequence of ρ is necessary so that $\Theta^{(\infty)}$ satisfies the exact rank constraint $\text{rank}(\Theta^{(\infty)}) \leq r$, due to the relaxation when considering the Lagrangian formulation.

This MM algorithm is guaranteed to converge to some stationary point by **?** , Theorem 4]; however, there is no guarantee that this stationary point corresponds to a global optimum. Moreover, there exists a duality gap between the formulation in (5.16) and its penalized counterpart in (D.24). In summary, although the statistical error for the desired non-convex problem formulation is the same to that obtained through the convex relaxation given in Section 5.2, we could not come up with an algorithm that *provably converges* to the global minimum. On the other hand, for the convex relaxation we have provided an iterative algorithm (Algorithm IV.1) that does so.

D.4 Supplement to the Real Data Analysis.

In this section, we provide additional details for the real data analysis, which substantiates our model specification that consists of contemporaneous common factors, and a vector-autoregressive idiosyncratic component. Since the crisis period witnesses prominent connectivities within $\widehat{B}$, we primarily focus on this period for illustration purpose.

To examine the correlation and autocorrelation pattern amongst the residuals, we plot in Figure D.1 the correlation and lag-1 autocorrelation amongst the residuals after adjusting for the identified factor hyperplane, that is, $\mathbf{X}_n - \widehat{\Theta}$. According to the model specification, there should not be much structural pattern left in the contemporaneous correlation, which is indeed the case according to the left panel. On the other hand, we postulate that the auto-covariance should exhibit certain patterns, parametrized by $\widehat{B}$. In the right panel, we plot the partial auto-correlation matrix of the residuals after properly scaling the entries for visualization purposes.

Further, we break down the connectivity amongst the nodes by sectors (banks, insurance,

brokers/dealers) to examine the cross-sector interactions, shown in Figure D.2. According to the figure, in general, at the relatively early stage of the crisis, insurance companies played the role of a strong emitter within the financial system, and such a role was transferred to the dealers/brokers toward the end of the crisis.

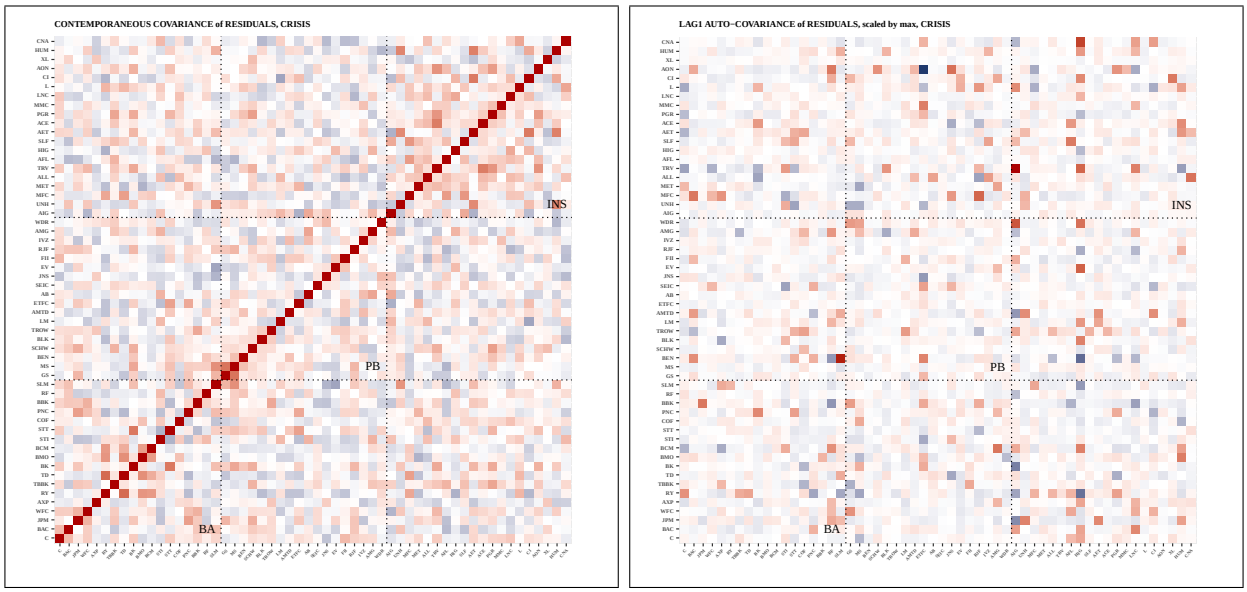

Figure D.1: Left panel: contemporaneous correlation among the residuals after adjusting for the factors. Right panel: partial auto-correlation structure among the residuals for lag = 1.

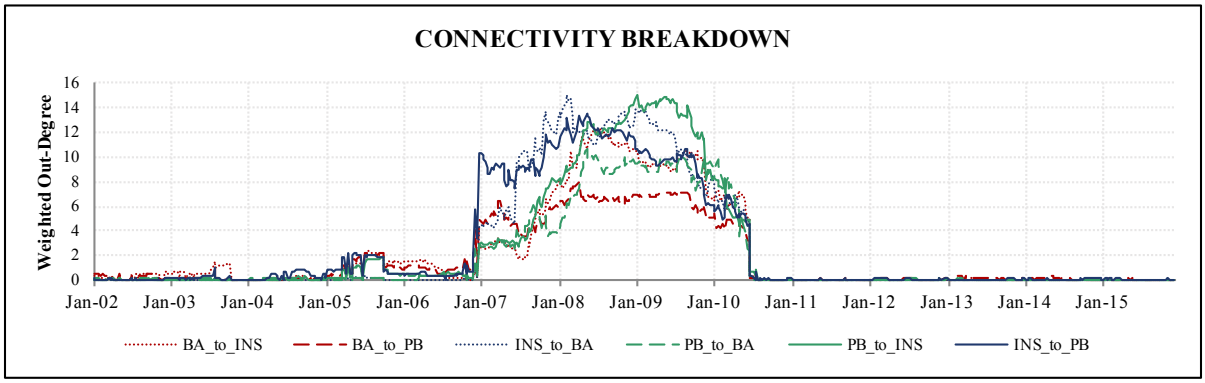

Figure D.2: Breakdown of connectivity by sub-categories. Lines with the same color indicate the same emitter: BA (red), PB (green), INS (blue). Lines with the same type indicate the same (emitter ↔ receiver) pair: (BA ↔ PB)–dashed line, (BA ↔ INS)–dotted line, (PB ↔ INS)–solid line.

As a final sanity check, we compare our results with the ones based on a Fama-French 5 factor model (FF5). Specifically, we fit the following Fama-French 5 factor model for each individual stock i:

$$X_{it} - \mathrm{RF}_t = \alpha_i + \mathrm{MmR}_t + \mathrm{SmB}_t + \mathrm{HmL}_t + \mathrm{RmW}_t + \mathrm{CmA}_t + \epsilon_{it}.$$

The comparison consists of two parts: (i) measuring the principal angle between our identified factor space and that spanned by FF5, and comparing their respective contribution to model fit in terms of R-squared; (ii) comparing the residual structure, which involves the contemporaneous correlation and the lead-lag relationship. Data of FF5 are retrieved from the Kenneth R. French Data Library.

Part (i), factor comparison. We first compare the two factor spaces by measuring their principal angle, defined as

$$\Omega := \arccos(d_1) \qquad \text{where } d_1 \text{ is the leading singular value of } \; Q_{\widehat{\mathbf{F}}}^{\top} Q_{\widetilde{\mathbf{F}}},$$

and here we use $\widetilde{\mathbf{F}}$ to denote the data matrix of FF5. During the crisis period of interest, $\Omega = 0.12$, which is equivalent to a 6.8° angle. Further, for the factor contribution to explaining the return variations, Table D.1 provides a summary of the R squared across the stocks for both models, with the factors identified by our model having an overall higher R squared.

	min	1st quantile	median	mean	3rd quantile	max
Lag-Adjusted Factor Model	0.25	0.59	0.68	0.65	0.76	0.85
Fama-French 5 Factor Model	0.28	0.48	0.58	0.61	0.69	0.84

Table D.1: Summary for R-squared across stocks for Lag-adjusted factor model and Fama-French 5 factor model.

Part (ii), residual comparison. Next, we compare the pattern amongst the residuals. Figure D.3 shows the contemporaneous correlation among the FF5 residuals and the transition matrix after a VAR(1) model is fitted. For the contemporaneous structure, we note that FF5 residuals exhibit detectable sub-clusters within the banking sub-sector and a stronger positive correlation within brokers/dealers. These structural patterns are less evident in the left panel of Figure D.1 and the residuals behave more uniformly, although the sub-clusterings formed by the big national banks and by the mid-size banks still exist. Turning to the transition matrix, the two models are broadly comparable and a similar set of important players have been identified, including AIG, HIG, ETFC etc.

Finally, we briefly compare our results with those obtained through the "single iterate" procedure, i.e., first fit a factor model on $\mathbf{X}_n$ using the PC estimation where the number of factors is chosen according to **?**], then a sparse VAR(1) model is fitted to the residuals, with the tuning parameter that controls the sparsity chosen according to BIC. The PC+VAR estimates differ from our estimates primarily in the following aspects: (1) the PC+VAR estimate gives a rank 1 structure at all times, i.e., $\widehat{K} = 1$ for all rolling windows, whereas our estimate exhibits a more nuanced and informative factor structure during the crisis period;

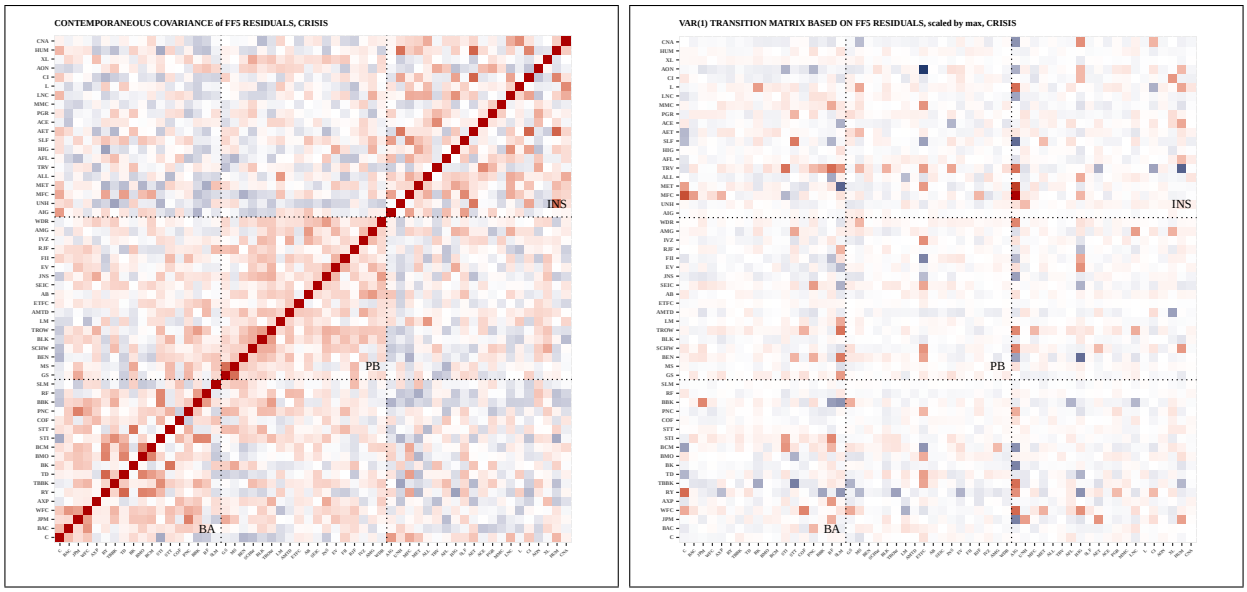

Figure D.3: Left panel: contemporaneous correlation among the FF5 residuals. Right panel: transition matrix for the VAR(1) fitted to FF5 residuals

(2) as Figure D.4 shows, the crisis period witnesses an increase in the connectivity of the sparse transition matrix, which is similar to what we have discovered with our proposed model; (3) leading factors from both estimates explain around 50% of the total variation, and for our estimates, the remaining four factors collectively account for another 15-20% of the total variation. The sharp increase during the crisis period for the estimated B from the PC+VAR estimate suggests that the idiosyncratic component has shown a much higher degree of temporal dependence on its past than at more normal times, whereas the number of factors during this period is likely to be under-estimated. Hence by properly incorporating such dependence in the mean structure specification of the model, one can circumvent the inaccuracy in estimating the number of factors, as pointed out by **?**].

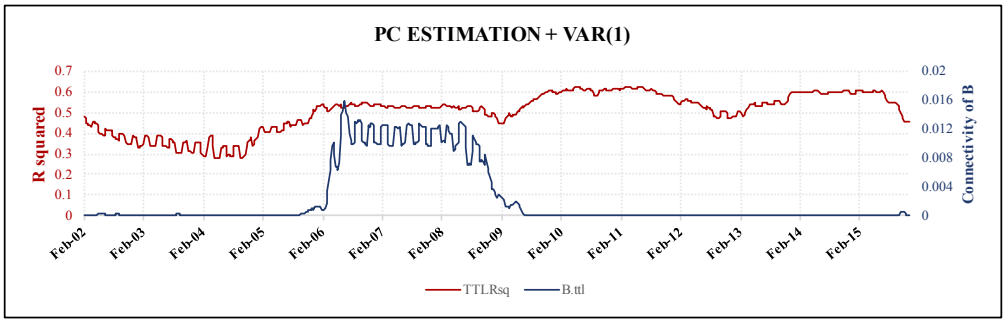

Figure D.4: Total R-squared based on the factor model (red line) and the connectivity of the estimated transition matrix of fitting a sparse VAR(1) on the residuals (blue line).

D.5 Supplement to VAR(d) Dependence.

We give some additional results for the case where u_t exhibits VAR(d) dependence. Note that with the model specification in (5.20), the spectral density of X_t takes the following form:

$$g_X(\omega) = \left[\mathcal{B}_d^{-1}(e^{-i\omega})\right]\Big(\Lambda g_F(\omega)\Lambda^\top + g_\epsilon(\omega) + g_{\epsilon,F}(\omega)\Lambda^\top + \Lambda g_{F,\epsilon}(\omega)\Big)\left[\mathcal{B}_d^{-1}(e^{-i\omega})\right]^*.$$

As Section 5.6 focuses on stating the error bound based on fixed realizations of the process upon certain regularity conditions are satisfied, in the rest of this section, we verify these conditions and provide high probability bounds for relevant quantities when the data are random realizations from the distribution.

Compared with the earlier analyses, a major difference lies in the fact that X_t now has lag-d dependence. However, we note that by considering the stacked transition matrix similar to **?**], each row of $\mathbf{X}_{n-d}^d$ can be viewed as a realization from a dp-dimensional process $\underline{X}_t$, whose dynamic resembles the previous considered model in Section 5.2. Specifically, by letting

$$\underline{X}_t := \begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-d+1} \end{bmatrix} \in \mathbb{R}^{dp}, \qquad \underline{F}_t := \begin{bmatrix} F_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{dK}, \qquad \underline{\epsilon}_t := \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{dp}$$

and

$$\underline{\Lambda} := \begin{bmatrix} \Lambda & \mathrm{O} & \dots & \mathrm{O} & \mathrm{O} \\ \mathrm{O} & & & & \\ \vdots & & \mathrm{O} & & \\ \mathrm{O} & & & & \\ \mathrm{O} & & & & \end{bmatrix} \in \mathbb{R}^{dp\times dK}, \qquad \underline{B} := \begin{bmatrix} B_1 & B_2 & \cdots & B_{d-1} & B_d \\ \mathrm{I}_p & \mathrm{O} & \cdots & \mathrm{O} & \mathrm{O} \\ \mathrm{O} & \mathrm{I}_p & \cdots & \mathrm{O} & \mathrm{O} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{O} & \mathrm{O} & \cdots & \mathrm{I}_p & \mathrm{O} \end{bmatrix} \in \mathbb{R}^{dp\times dp},$$

an alternative representation for (5.20) is given by

$$\underline{X}_t = \underline{\Lambda}\underline{F}_t + \underline{B}\underline{X}_{t-1} + \underline{\epsilon}_t. \tag{D.26}$$

Thus, it suffices to verify the RSC condition in an identical way to that in Lemma 5.1, however with the underlying process substituted by $\underline{X}_t$; for other quantities such as deviation or the

extreme of the eigen-spectrum, the high probability bound should be given based upon $\underline{X}_t$ as well.

Lemma D.4. *For $\mathbf{X}^d \in \mathbb{R}^{n_d \times dp}$ whose rows are random realizations $\{x_0, \ldots, x_{n-1}\}$ of the stable $\{\underline{X}_t\}$ process with dynamic given in* (D.26)*. Then there exist positive constants c_i $(i = 0, 1, 2, 3)$ such that with probability at least $1 - c_1 \exp(-c_2 n)$, the RSC condition holds for $\mathbf{X}^d$ with curvature α_{RSC} and tolerance τ_{n_d} satisfying*

$$\alpha_{RSC} = \pi \mathfrak{m}(g_{\underline{X}}), \qquad \tau_{n_d} = \alpha_{RSC} \gamma^2 \frac{\log dp}{2n_d},$$

where $\gamma := 54\mathcal{M}(g_{\underline{X}})/\mathfrak{m}(g_{\underline{X}})$, provided that $n_d \gtrsim (\sum_{k=1}^{d} \|B_k\|_0) \log(dp)$.

Lemma D.5. *There exist positive constants c_i $(i = 0, 1, 2)$ such that for sample size $n_d \gtrsim \log(dp)$, with probability at least $1 - c_1 \exp(-c_2 \log(dp))$, the following bound holds:*

$$\|(\mathbf{X}_{n-1}^d)^\top \mathbf{E}_n^d / n_d\|_\infty \leq c_0 \Big(\mathcal{M}(g_{\underline{X}}) + \mathcal{M}(g_\epsilon) + \mathcal{M}(g_{\underline{X},\epsilon}) \Big) \sqrt{\frac{\log(dp)}{n}}.$$

Note that with the definition of $\underline{\epsilon}_t$, $\mathcal{M}(g_{\underline{X},\underline{\epsilon}}) = \mathcal{M}(g_{\underline{X},\epsilon})$.

Lemma D.6. *Consider $\mathbf{X} \in \mathbb{R}^{n_d \times dp}$ whose rows are some random realization $\{x_0, \ldots, x_{n-1}\}$ of the stable $\{\underline{X}_t\}$ process whose dynamic is given in* (D.26)*. Then there exist positive constants c_i $(i = 0, 1, 2)$ such that for sample size $n_d \gtrsim dp$, with probability at least $1 - c_1 \exp(-c_2 dp)$, the following bound holds for the spectrum of $S_{\mathbf{X}}$:*

$$\Lambda_{\max}(S_{\mathbf{X}}) \leq c_0 \mathcal{M}(g_{\underline{X}}).$$

With the definition of $\underline{X}_t$, let $\underline{v}_t := \underline{\Lambda} \underline{F}_t + \underline{\epsilon}_t$, then $\underline{X}_t = \underline{B} \underline{X}_t + \underline{v}_t$. The following bounds hold for $\mathcal{M}(g_{\underline{X}})$ and $\mathfrak{m}(g_{\underline{X}})$ [**?**]:

$$\mathcal{M}(g_{\underline{X}}) \leq \frac{1}{2\pi} \frac{\Lambda_{\max}(\Sigma_{\underline{v}})}{\mu_{\min}(\underline{\mathcal{B}})}, \qquad \text{and} \qquad \mathfrak{m}(g_{\underline{X}}) \geq \frac{1}{2\pi} \frac{\Lambda_{\min}(\Sigma_{\underline{v}})}{\mu_{\max}(\mathcal{B}_d)},$$

where we define $\mu_{\max}(\mathcal{B}_d) := \max_{|z|=1} \Lambda_{\max}\Big(\big(\mathcal{B}_d(z)\big)^* \big(\mathcal{B}_d(z)\big) \Big)$, $\mu_{\min}(\underline{\mathcal{B}}) := \min_{|z|=1} \Lambda_{\min}\Big(\underline{\mathcal{B}}(z)^* \underline{\mathcal{B}}(z) \Big)$, with $\underline{\mathcal{B}}(z) = \mathrm{I}_{dp} - \underline{B} z$.

BIBLIOGRAPHY