DIMENSION REDUCTION FOR VECTOR AUTOREGRESSIVE (VAR(P)) MODELS VIA SPATIAL QUANTILE REGRESSION

by

Yijiang Wang

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2025

Approved by:

____

Dr. Jiancheng Jiang

____

Dr. Maciej Noras

____

Dr. Eliana Christou

____

Dr. Wenyu Gao

ABSTRACT

YIJIANG WANG. Dimension Reduction for Vector Autoregressive (VAR(P)) Models via Spatial Quantile Regression. (Under the direction of DR. JIANCHENG JIANG)

The Vector Autoregressive (VAR) model is a fundamental tool for analyzing multivariate time series, capturing the dynamic relationships among variables through their own and others' lagged values. we first introduce a novel framework for high-dimensional Vector Autoregressive (VAR) models by incorporating Spatial Quantile Regression (SQR) with adaptive Lasso and SCAD regularization. Unlike traditional quantile regression, SQR extends to the multivariate setting, allowing for more robust inference on conditional quantiles, particularly in the presence of heavy-tailed or non-Gaussian errors. However, as the dimensionality and lag order increase, VAR models often suffer from over-parameterization, leading to potential overfitting. Therefore, to address this issue, we employ adaptive Lasso and SCAD penalties, which facilitate both parameter estimation and automatic variable selection, enforcing sparsity in a data-driven manner. Under mild regularity conditions, we establish the oracle properties of our proposed estimators, proving their ability to recover the true underlying model structure while yielding asymptotically normal estimates for significant parameters. To efficiently solve the resulting non-concave penalized optimization problem, we develop a computationally efficient algorithm based on alternating optimization and the Alternating Direction Method of Multipliers (ADMM), ensuring scalability for large-scale datasets. Extensive simulation studies demonstrate the proposed method's advantages in estimation accuracy, variable selection, and robustness against various error distributions, including normal, mixed-normal, Student's t, and Laplace distributions. Finally, we apply our approach to real-world data, illustrating its practical utility in high-dimensional time series analysis.

Furthermore, to tackle the curse of dimensionality and enhance model interpretability, another novel framework is proposed that integrates tensor decomposition with Spatial Quantile Regression (SQR), restructuring the transition matrices of the VAR model into a tensor representation for simultaneous parameter reduction while enhancing robustness and flexibility by modeling covariate effects across different quantiles of the response distribution. Specifically, we introduce the Multilinear Low-Rank Spatial Quantile Regression (MLRSQR) method for parameter estimation post-tensor decomposition. We establish the asymptotic properties of the MLRSQR estimator and develop an efficient alternating spatial quantile regression algorithm for its implementation. To further refine estimation, we extend our approach by incorporating sparsity. We propose the ℓ_1-penalized Sparse Higher-Order Reduced-Rank Spatial Quantile Regression (SHORRSQR) estimator, which balances dimensionality reduction with sparsity constraints. Theoretical guarantees for its asymptotic behavior are rigorously derived, and we design an ADMM-based algorithm for its efficient computation. Through extensive simulations and an application to real-world data, we demonstrate the advantages of our proposed methods in mitigating over-parameterization, improving robustness, and enhancing interpretability, thereby making them well-suited for high-dimensional time series analysis.

DEDICATION

To my dear father, who taught me the value of perseverance, kindness, and integrity. Though you are no longer here, your wisdom and love continue to guide me. This work is dedicated to you, in gratitude for all that you have given me.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my advisor Dr. Jiancheng Jiang, who has offered patient guidance and support consistently through my graduate life. The completion of my dissertation would not have been possible without his constant help. It was a great fortune to learn from his extensive knowledge and tireless commitment to research, which inspired and motivated me a lot during my research and will benefit me in my future career.

I would also like to extend my gratitude to to the rest of my committee members, Dr. Maciej Noras, Dr. Eliana Christou, and Dr. Wenyu Gao, for their insightful comments and valuable advice on my research work.

Thanks also to Dr. Shaozhong Deng and Dr. Mohammad A. Kazemi for their guidance and help as graduate coordinators.

Last but not least, I'm deeply grateful to my parents, Long Wang and Fang Xin, for making me who I am and for their selfless love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Quantile regression (QR), first introduced by Koenker and Bassett (1978) (1), has become a fundamental tool for estimating conditional quantile functions and conducting inference. Unlike mean regression, QR provides a more comprehensive view of the stochastic relationship between variables by capturing the entire conditional distribution rather than focusing solely on the mean (Chaudhuri, Doksum, and Samarov, 1997 (2); Koenker, 2005 (3)). Moreover, it offers robust and efficient estimation, particularly when the error distribution deviates from normality, making it a valuable alternative to traditional least squares methods (Koenker and Bassett, 1978 (1); Koenker and Zhao, 1996 (4)). Due to these advantages, QR has inspired a vast body of research, playing a pivotal role in statistics and econometrics. While the literature on QR is extensive, this paper focuses on its application in the context of high-dimensional time series modeling.

Quantile regression (QR) has been extensively studied in time series analysis, with notable contributions from Koul and Saleh (1995) (5), Davis and Dunsmuir (1997) (6), Jiang, Zhao, and Hui (2001) (7), and Peng and Yao (2003) (8). However, much of this research has been centered on univariate time series, leaving a significant gap in the theoretical development of QR for vector time series. While maximum likelihood and least squares estimation methods have been widely applied in the multivariate setting—see, for instance, Bollerslev (1990) (9), Engle and Kroner (1995) (10), Chen and Tsay (1993) (11), and Pan and Yao (2008) (12)—the extension of QR to multivariate models remains relatively underexplored. A fundamental challenge in this area is the lack of a universally accepted definition of multivariate quantiles.

To address this, the notion of spatial quantiles, originally introduced by Chaudhuri (1996) (13) and subsequently refined by Koltchinskii (1997) (14), provides a geometric extension of univariate quantiles to multivariate settings through the use of the ℓ_1-norm. Unlike traditional definitions, spatial quantiles define central regions in a multivariate distribution that expand with increasing coverage, thereby offering a valuable volume-based interpretation. This generalization maintains core features of univariate quantiles while enabling a more nuanced representation of distributional structure in higher dimensions. As highlighted by Serfling (2004) (15), spatial quantiles possess a number of favorable properties, including shift equivariance, resistance to outliers, and invariance under orthogonal and homogeneous scaling transformations. These attributes make spatial quantiles a compelling choice for multivariate analysis, motivating their integration into spatial quantile regression (SQR) for vector autoregressive (VAR(P)) models (Sims, 1980 (16)).

According to Chaudhuri (1996) (13) and Koltchinskii (1997) (14), given a sample $\{\mathbf{z}_i\}_{i=1}^n$ of $\mathbf{z}$ in $\mathbb{R}^N$, the $\mathbf{u}$-th spatial quantiles are defined as

$$\hat{\boldsymbol{\alpha}}(\mathbf{u}) = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \sum_{i=1}^{n} \left\{ \|\mathbf{z}_i - \boldsymbol{\alpha}\| + \mathbf{u}^T(\mathbf{z}_i - \boldsymbol{\alpha}) \right\} \tag{1.1}$$

where $\mathbf{u} \in \mathcal{B}^N = \{\mathbf{u} | \mathbf{u} \in \mathbb{R}^N, \|\mathbf{u}\| < 1\}$, and $\| \cdot \|$ is the Euclidean norm. For $N = 1$, the solution to (1.1) reduces to the sample τ-th quantile ($\tau = (1 + \mathbf{u})/2$) based on the real-valued observations $\mathbf{z}_i$'s. Let $Q_u(\mathbf{t}) = \|\mathbf{t}\| + \langle \mathbf{u}, \mathbf{t} \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Then equation (**??**) can be rewritten as

$$\hat{\boldsymbol{\alpha}}(\mathbf{u}) = \arg \min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} Q_u(\mathbf{z}_i - \boldsymbol{\alpha})$$

Define the $\mathbf{u}$-th quantile of the distribution of $\mathbf{z}$ as

$$\boldsymbol{\alpha}(\mathbf{u}) = \arg \min_{\boldsymbol{\alpha}} E \left\{ Q_u(\mathbf{z} - \boldsymbol{\alpha}) - Q_u(\mathbf{z}) \right\}$$

Chaudhuri (1996) (13) established that the spatial quantile estimator $\hat{\boldsymbol{\alpha}}(\mathbf{u})$ is asymptotically normal, specifically, $\sqrt{n}[\hat{\boldsymbol{\alpha}}(\mathbf{u}) - \boldsymbol{\alpha}(\mathbf{u})] \to N(\mathbf{0}, \boldsymbol{\Sigma})$ for some covariance matrix $\boldsymbol{\Sigma}$. Based on the estimate $\hat{\boldsymbol{\alpha}}(\cdot)$, several multivariate descriptive statistics can be constructed. For instance, a trimmed mean of a multivariate random vector $\mathbf{z}$ can be estimated using $\int_S \hat{\boldsymbol{\alpha}}(\mathbf{u})\, \mu(\mathrm{d}\mathbf{u})$, where μ is a properly defined probability measure on the unit ball $\mathcal{B}^N$, and the integration domain is given by $S = \{\mathbf{u} \in \mathbb{R}^N : \|\mathbf{u}\| \leq r\}$ for some $r \in (0, 1)$. Similarly, the multivariate L-estimator can be derived in the same form but with an alternative choice of region S; see Chaudhuri (1996) (13) for details. This spatial quantile framework naturally extends to multivariate regression settings. Consider the multivariate linear model:

$$\mathbf{y}_i = \boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n \tag{1.2}$$

where $\boldsymbol{\beta}$ is a $N \times v$ matrix of unknown parameters and $\mathbf{x}_i$ are $v \times 1$ vector of covariates without intercept.

Let $\{\mathbf{y}_t \in \mathbb{R}^N\}_{t=1}^T$ denote an N-dimensional vector time series of length T. A Pth-order vector autoregressive model, VAR(P), can be expressed as:

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_P\mathbf{y}_{t-P} + \boldsymbol{\varepsilon}_t, \quad \text{for } t = 1, \ldots, T. \tag{1.3}$$

where $\{\mathbf{y}_t\}$ is the observed time series with $\mathbf{y}_t = (y_{1t}, \cdots, y_{Nt})' \in \mathbb{R}^N$, The innovations satisfy $\boldsymbol{\varepsilon}_t = \boldsymbol{\Sigma}^{1/2}\mathbf{a}_t$, and $\boldsymbol{\Sigma}^{1/2}$ are symmetric positive definite matrices and $\mathbf{a}_t$ is a sequence of serially uncorrelated random vectors with mean $\mathbf{0}$ and identity covariance matrix $\mathbf{I}$. $\mathbf{A}_j$'s are $N \times N$ transition matrices of unknown parameters and T is the sample size.

Then, the model in (1.3) can be reformulated as:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \quad \text{for} \quad t = 1, \ldots, T, \tag{1.4}$$

where $\mathbf{x}_t = (\mathbf{y}_{t-1}^{'}, \cdots, \mathbf{y}_{t-P}^{'})'$ and $\mathbf{A} = (\mathbf{A}_1, \cdots, \mathbf{A}_P)$. Then the above model (1.4) has the form of (1.2). It is straightforward to extend the above spatial quantile notion by defining the $\mathbf{u}$-th spatial regression quantiles as

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{q}}_u) = \arg\min_{\mathbf{A}, \mathbf{q}_u} L(\mathbf{A}, \mathbf{q}_u) \tag{1.5}$$

where $L(\mathbf{A}, \mathbf{q}_u) = \sum_{t=1}^T Q_u(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t - \mathbf{q}_u)$, and $\mathbf{q}_u$ is the $\mathbf{u}$-th quantile of $\boldsymbol{\varepsilon}$. Under this formulation, for any direction $\mathbf{u} \in \mathcal{B}^N$, the estimator $\widehat{\mathbf{A}}$ consistently estimates the true coefficient matrix $\mathbf{A}$; see Jiang, Jiang, Li, Liu, and Yan (2017) (17). In the special case where $\mathbf{u} = 0$, the model reduces to the spatial median regression, originally studied by Bai, Chen, Miao, and Rao (1990) (18). When the response is univariate, i.e., $N = 1$, the spatial quantile regression formulation becomes equivalent to the classical quantile regression proposed by Koenker and Bassett (1978) (1). Moreover, to obtain affine equivariant versions of the estimator, one can adopt the transformation-retransformation technique introduced by Chaudhuri (1996) (13), and further developed by Chakraborty and Chaudhuri (1998) (19) and Chakraborty (2003) (20).

A standard VAR(P) model is widely used to characterize temporal dependencies among multiple time series variables, enabling both forecasting and structural analysis. However, as the number of variables N and the lag order P grow, estimating the parameters of VAR models becomes increasingly difficult. Traditional estimation approaches, such as ordinary least squares (OLS), become computationally burdensome and statistically inefficient in high-dimensional settings. This issue has been extensively discussed in the literature (De Mol, Giannone, and Reichlin 2008 (21); Carriero, Kapetanios, and Marcellino 2011 (22); Koop 2013 (23)), where it is shown that even with moderate values of N and P, the large number of parameters can lead to overfitting, reduced estimation accuracy, and unreliable inference—phenomena collectively referred to as the curse of dimensionality.

Vector Autoregressive (VAR) models are widely employed in multivariate time series analysis due to their relative simplicity and ease of estimation. While the Vector Autoregressive Moving Average (VARMA) model offers a more flexible framework to capture autocorrelation dynamics, it is often replaced in practice by the VAR model due to well-known identification issues and numerical instability associated with estimating high-order polynomials (Chan, Eisenstat, and Koop, 2016 (24); Wilms et al., 2017 (25); Dias and Kapetanios, 2018 (26)). Nonetheless, the VAR approximation may require a large lag order P, particularly when the asymptotic conditions $T \to \infty$, $P \to \infty$, and $PT^{-1/3} \to 0$ are considered (Said and Dickey, 1984 (27); Li, Leng, and Tsai, 2014 (28)). This requirement exacerbates the curse of dimensionality, as the number of parameters increases rapidly with both the number of variables N and the lag order P, resulting in N^2P coefficients to estimate (Ravenna, 2007 (29)).

To address these challenges, imposing structural constraints on the parameter space has become an essential strategy in high-dimensional VAR modeling. A primary concern is overfitting, which is exacerbated in finite samples due to the inflated parameter space. Sparsity—where a substantial portion of coefficients are effectively zero—is a prevalent characteristic in high-dimensional data. Regularization methods such as the Lasso and the Dantzig selector provide a principled approach to exploiting this structure, facilitating variable selection and enhancing estimation accuracy (Basu and Michailidis, 2015 (30); Han, Lu, and Liu, 2015 (31); Kock and Callot, 2015 (32); Davis, Zang, and Zheng, 2016 (33); Wu and Wu, 2016 (34)). In addition to sparsity, enforcing stationarity is crucial to avoid explosive dynamics, which can compromise both inference and forecasting performance.

Another promising avenue for dimensionality reduction in high-dimensional VAR models is through low-rank approximations of the transition matrices. Reduced-rank regression (RRR) imposes a low-rank structure on the parameter matrices, effectively controlling model complexity without sacrificing dynamic structure (Yuan et al., 2007 (35); Negahban and Wainwright, 2011 (36); Chen, Dong, and Chan, 2013 (37); Basu, Li, and Michailidis, 2019 (38); Raskutti, Yuan, and Chen, 2019 (39)). In the standard formulation,

$$\mathbf{y}_t = \mathbf{A}^{(C)}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T, \tag{1.6}$$

where $\mathbf{A}^{(C)} = (\mathbf{A}_1, \ldots, \mathbf{A}_P)$, the matrix $\mathbf{A}^{(C)}$ is assumed to have low rank (Velu, Reinsel, and Wichern, 1986 (40); Velu and Reinsel, 2013 (41)). Carriero, Kapetanios, and Marcellino (2011 (22)) proposed a Bayesian framework for forecasting in large-scale macroeconomic systems under this reduced-rank VAR setting. It is worth noting that model (1.6) and (1.4) are similar in form; however, the transition matrix in model (1.6) emphasizes a low-rank structure. Therefore, we denote it as $\mathbf{A}^{(C)}$ to distinguish it from the matrix in model (1.4). Similarly, the above model (1.6) has the form of (1.2).

Therefore, in Chapters 2 and 3, we will explore several methods for constraining the coefficients of VAR models. Chapter 2 focuses on two sparse regularization techniques—adaptive Lasso and SCAD—which impose penalty terms on the coefficients to effectively identify important variables while shrinking insignificant ones toward zero.

This work systematically explores the integration of adaptive penalization techniques within the spatial quantile regression (SQR) framework for VAR models. Due to the lack of closed-form solutions and the presence of non-convex objective functions, the resulting SQR estimators pose both theoretical and computational challenges. In particular, establishing their asymptotic properties requires careful analysis. In this paper, we rigorously derive the theoretical guarantees for the proposed estimators, including consistency and oracle properties. Furthermore, we develop efficient computational algorithms to enable practical implementation of the methodology. Through extensive simulation studies and real-world data applications, the proposed approach demonstrates strong performance in terms of robustness, sparsity, and estimation efficiency.

To better understand the structure of the coefficient matrices, we consider alternative rearrangements: $\mathbf{A}^{(R)} = (\mathbf{A}_1', \ldots, \mathbf{A}_P')$ and $\mathbf{A}^{(L)} = (\text{vec}(\mathbf{A}_1)', \ldots, \text{vec}(\mathbf{A}_P)')'$. These formulations capture the row space and the vectorized coefficient space, respectively. Reinsel (1983 (42)) introduced an autoregressive index model that imposes a low-rank constraint on $\mathbf{A}^{(R)}$, while low-rank dependencies across lags suggest that $\mathbf{A}^{(L)}$ may also exhibit low effective dimensionality. Importantly, the ranks of $\mathbf{A}^{(C)}$, $\mathbf{A}^{(R)}$, and $\mathbf{A}^{(L)}$ each correspond to distinct structural interpretations.

Motivated by this observation, we reformulate the transition matrices as a three-way tensor. In this setting, the mode-1, mode-2, and mode-3 matricizations of the tensor correspond to $\mathbf{A}^{(C)}$, $\mathbf{A}^{(R)}$, and $\mathbf{A}^{(L)}$, respectively (Kolda and Bader, 2009 (43)). We apply Tucker decomposition to the transition tensor, which enables simultaneous rank reduction across all three modes. This leads to the multilinear low-rank VAR model, where the Tucker ranks—also known as multilinear ranks—determine the model's complexity and structure (Wang, Zheng, Lian, and Li, 2022 (44)).

Chapter 3 discusses how to impose structural constraints on the coefficient matrices through tensor decomposition, thereby reducing the dimensionality of the parameter space. These approaches not only offer favorable statistical properties in theory but also demonstrate strong predictive performance and interpretability in practice. By combining tensor decomposition with spatial quantile regression, our approach provides a robust and scalable framework for analyzing high-dimensional VAR models, addressing both over-parameterization and computational challenges. Similar challenges also arise when incorporating sparsity into the tensor-based framework, as it introduces additional non-convexity and complexity into the estimation process. In particular, deriving the asymptotic properties of the proposed estimators becomes analytically demanding due to the high dimensionality of the parameter space and the non-convex nature of the optimization problem. Nevertheless, we provide rigorous theoretical guarantees, establishing both consistency and asymptotic normality under mild regularity conditions. This integration of tensor decomposition, quantile-based inference, and sparse regularization yields a flexible and powerful modeling tool capable of handling large-scale, non-Gaussian time series data. Besides, An efficient ADMM-based algorithm is developed to solve the resulting optimization problem, which enables joint estimation of sparse, low-rank structures in a computationally feasible manner.

The remainder of this dissertation is organized as follows. In addition to methodological development, Chapter 2 and Chapter 3 provide in-depth discussions on the tuning parameter selection for the respective approaches. Each chapter also includes comprehensive simulation studies and real data applications to validate the proposed methods. Proofs of the main results are given in Appendix A and Appendix B.

CHAPTER 2: ORACLE MODEL SELECTION FOR VAR(P) MODELS BASED ON SPATIAL QUANTILE REGRESSION

2.1 Overview

Building upon the challenges outlined in the Introduction, we now turn to methodological developments aimed at improving the estimation of high-dimensional VAR models. In particular, we focus on regularization and dimension reduction techniques that address the curse of dimensionality by imposing structural constraints on the model coefficients. Among these, sparsity-inducing methods have shown great promise, as they enable the identification of a parsimonious subset of relevant parameters while reducing estimation variance. One of the most widely used approaches is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996 (45)), which applies an ℓ_1-norm penalty to encourage sparsity. Despite its effectiveness, LASSO tends to produce biased estimates due to the uniform shrinkage it imposes across all coefficients. To mitigate this drawback, we consider two alternatives: the smoothly clipped absolute deviation (SCAD) and the adaptive LASSO. These methods maintain the sparsity-promoting advantages of LASSO while improving variable selection consistency and reducing estimation bias, making them better suited for capturing complex temporal structures in VAR models.

The smoothly clipped absolute deviation (SCAD) penalty, proposed by Fan and Li (2001) (46), introduces a nonconvex regularization function specifically designed to alleviate the bias associated with large coefficient shrinkage. By applying less penalization to large coefficients, SCAD leads to more accurate parameter estimation compared to convex alternatives. In a similar vein, the adaptive LASSO method developed by Zou (2006) (47) incorporates data-driven weights into the ℓ_1 penalty, allowing it to differentially penalize coefficients based on their magnitudes. This adaptive weighting mechanism improves the identification of truly relevant predictors while maintaining model sparsity. Both SCAD and adaptive LASSO satisfy the oracle property, meaning they can asymptotically recover the true model structure of the underlying VAR system with high probability.

Over the past few decades, model selection techniques that assume a fixed number of parameters have been widely adopted. However, in real-world applications, a large number of variables are often introduced to reduce potential model misspecification. As pointed out by Huber (1973 (48), 1988 (49)), Portnoy (1988) (50)), and Donoho (2000) (51)), the total number of parameters—denoted by N^2P—can be substantial and should be treated as a function of the sample size. Following this perspective, we allow the number of parameters to grow with T and denote it by $N_T^2 P_T$. This increasing-dimensional framework is consistent with the viewpoint of Fan and Peng (2004) (52) and Lam and Fan (2008) (53), who advocated that in many model selection settings, the number of parameters should increase with the sample size.

To explicitly reflect this dependence, we rewrite equations (1.4) and (1.5) as:

$$\mathbf{y}_t = \mathbf{A}_T \mathbf{x}_t + \boldsymbol{\varepsilon}_t, \quad \text{for } t = 1, \ldots, T \tag{2.1}$$

$$(\widehat{\mathbf{A}}_T, \widehat{\mathbf{q}}_u) = \arg \min_{\mathbf{A}_T, \mathbf{q}_u} L_T(\mathbf{A}_T, \mathbf{q}_u) \tag{2.2}$$

where the objective function is defined as $L_T(\mathbf{A}_T, \mathbf{q}_u) = \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \mathbf{A}_T \mathbf{x}_t - \mathbf{q}_u)$.

To emphasize this dependence, we write the parameter vector as $\text{vec}(\mathbf{A}_T)$, which lies in a p_T-dimensional space with $p_T = N_T^2 P_T$. We further partition it as $\text{vec}(\mathbf{A}_T) = (\text{vec}(\mathbf{A}_{T1})', \text{vec}(\mathbf{A}_{T2})')'$, where $\text{vec}(\mathbf{A}_{T1}) \in \mathbb{R}^{s_T}$ represents the subset of relevant parameters and $\text{vec}(\mathbf{A}_{T2}) \in \mathbb{R}^{p_T - s_T}$ corresponds to parameters assumed to be zero. Accordingly, we assume that the true coefficient vector takes the form $\text{vec}(\mathbf{A}_T^*) = (\text{vec}(\mathbf{A}_{T1}^*)', \mathbf{0}')'$, where the s_T nonzero components in $\text{vec}(\mathbf{A}_{T1}^*)$ are associated with the true underlying model. For notational convenience, if we reshape these s_T active components into a matrix, we denote it by $\mathbf{A}_{T1}^* \in \mathbb{R}^{N_{s_T} \times N_{s_T} P_{s_T}}$, implying that $s_T = N_{s_T}^2 P_{s_T}$.

In this Chapter, we address the issue of variable/parameter selection using SQR with adaptive LASSO and SCAD penalties. These penalization methods are well suited for high-dimensional settings, where overparameterization often complicates estimation. The adaptive LASSO dynamically adjusts penalty weights, addressing the bias issues associated with traditional LASSO, while SCAD minimizes estimation bias and enhances oracle model selection properties. These methods allow us to achieve consistent model selection and efficient parameter estimation under mild regularity conditions.

Our approach is especially appealing for applications involving non-Gaussian or heavy-tailed error distributions, as frequently observed in financial and environmental time series. By integrating SQR with sparsity-inducing regularization, the proposed methodology enhances robustness to distributional irregularities while simultaneously achieving model sparsity, thereby facilitating both interpretability and computational tractability in high-dimensional regimes.

We systematically explore the incorporation of adaptive regularization techniques within the SQR framework for VAR modeling. Due to the absence of closed-form solutions and the presence of nonconvex objective functions, the resulting estimators present both analytical and algorithmic challenges. In particular, establishing asymptotic properties requires delicate theoretical treatment. We rigorously derive the theoretical guarantees of the proposed estimators, including consistency and oracle properties, under mild regularity assumptions. To address the computational aspects, we develop efficient numerical algorithms tailored to the nonconvex optimization landscape of the penalized SQR problem. Extensive simulation studies and empirical analyses demonstrate the superior performance of our method in terms of robustness, sparsity, and estimation accuracy.

The rest of this Chapter is organized as follows. In Section 2.2, we develop the oracle model selection framework for high-dimensional VAR(P) models based on spatial quantile regression (SQR), incorporating adaptive LASSO and SCAD penalties. In Section 2.3, we propose a computationally efficient ADMM algorithm to solve the penalized SQR optimization problem and describe the selection of tuning parameters. In Section 2.4, we perform simulation studies under various error distributions and apply the proposed methods to a real-world air pollution dataset to demonstrate their practical effectiveness. For clarity and focus, the regularity conditions and technical proofs of the main theoretical results are deferred to Appendix A.

2.2 Oracle Model Selection Based on SQR

Our methodology is naturally motivated by the model specified in equation (1.4). The spatial quantile regression (SQR) estimator of $\mathbf{A}$ can be obtained by minimizing the empirical loss function in equation (1.5) over both $\mathbf{A}$ and $\mathbf{q}_u$. Since we allow the number of parameters N^2P to grow with the sample size T, estimation is conducted by solving the following optimization problem:

$$L_T(\mathbf{A}_T, \mathbf{q}_u) = \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \mathbf{A}_T\mathbf{x}_t - \mathbf{q}_u), \tag{2.3}$$

where the minimization is taken over $\mathbf{A}_T$ and $\mathbf{q}_u$.

However, the loss function in (2.3) does not account for variable or parameter selection. To address this limitation, we adopt a penalized estimation approach by minimizing the following objective:

$$L_T(\mathbf{A}_T, \mathbf{q}_u) + T\sum_{i=1}^{N_T}\sum_{j=1}^{N_T P_T} p_{\lambda_T}(|A_{Tij}|), \tag{2.4}$$

where $p_{\lambda_T}(\cdot)$ denotes a generic penalty function and $\lambda_T \geq 0$ is a regularization parameter controlling the degree of shrinkage. Various choices for $p_{\lambda_T}(\cdot)$ have been proposed in the literature; in this Chapter, we focus on two widely used and theoretically justified penalties: the adaptive LASSO and SCAD. The proposed framework, however, can be extended to accommodate other penalty functions.

To establish the theoretical properties of the proposed penalized estimators, we first introduce some technical assumptions. A fundamental requirement for the validity of the vector autoregressive model is the stationarity of the underlying process. The following assumption provides a necessary and sufficient condition for strict stationarity of the VAR(P) process:

Assumption 1. All roots of the matrix polynomial $\mathcal{A}(z) = \mathbf{I}_N - \mathbf{A}_1 z - \cdots - \mathbf{A}_P z^P$, $z \in \mathbb{C}$, lie outside the unit circle, where $\mathbb{C}$ denotes the set of complex numbers.

Assumption 1 ensures the existence of a unique strictly stationary solution to the model specified in equation (1.4). This condition forms the basis for deriving the asymptotic properties of the proposed spatial quantile regression estimators.

2.2.1 Model selection with SCAD penalty

The SCAD penalty function $p_\lambda(\cdot)$, proposed by Fan and Li (2001) (46), is defined via its first-order derivative and is symmetric about the origin. For $\theta > 0$, its derivative is given by

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\},$$

where $\lambda > 0$ is a regularization parameter, $a > 2$ is a shape parameter, and $(x)_+ = \max(x, 0)$. This formulation corresponds to a quadratic spline with knots at λ and $a\lambda$, ensuring that large coefficients are not overly penalized, thus reducing bias and yielding continuous solutions.

The associated thresholding rule takes the following closed-form:

$$\hat{\theta} = \begin{cases} \operatorname{sgn}(z)(|z| - \lambda)_+, & \text{if } |z| \le 2\lambda, \\ \frac{(a-1)z - \operatorname{sgn}(z)a\lambda}{a-2}, & \text{if } 2\lambda < |z| \le a\lambda, \\ z, & \text{if } |z| > a\lambda, \end{cases}$$

which highlights the adaptive nature of the SCAD penalty in treating small and large values differently.

The SCAD-penalized spatial quantile regression estimator is obtained by minimizing the penalized objective function:

$$(\widehat{\mathbf{A}}_T^{SC}, \widehat{\mathbf{q}}_u) = \arg \min_{\mathbf{A}_T, \mathbf{q}_u} Q_T^{SC}(\mathbf{A}_T, \mathbf{q}_u) \tag{2.5}$$

where $Q_T^{SC}(\mathbf{A}_T, \mathbf{q}_u) = L_T(\mathbf{A}_T, \mathbf{q}_u) + T \sum_{i=1}^{N_T} \sum_{j=1}^{N_T P_T} p_{\lambda_T}(|A_{Tij}|)$. We refer to this estimation procedure as the SQRSCAD method.

We establish both the consistency and asymptotic normality of the SCAD-penalized estimator under a set of mild regularity conditions. For clarity and conciseness, all technical assumptions and theoretical details are deferred to the Appendix.

To facilitate theoretical analysis, we vectorize the transition matrix in the model. Let $\boldsymbol{\beta}_T = \operatorname{vec}(\mathbf{A}'_T)$ denote the vectorized form of the transition matrix. Using this notation, the objective function in equation (2.3) can be rewritten as:

$$L_T(\boldsymbol{\beta}_T, \mathbf{q}_u) = \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \mathbf{z}_t\boldsymbol{\beta}_T - \mathbf{q}_u), \tag{2.6}$$

where $\mathbf{z}_t = (\mathbf{I}_N \otimes \mathbf{x}_t')$ is the regressor vector formed via Kronecker product. Define the full parameter vector as $\boldsymbol{\theta}_T = (\boldsymbol{\beta}_T, \mathbf{q}_u)$, and let $\mathbf{V}_{Tt} = (\mathbf{z}_t, \mathbf{y}_t)$ denote the observable random vector at time t.

The penalized estimation problem in equation (2.4) can then be equivalently expressed as:

$$L_T(\boldsymbol{\theta}_T) + T\sum_{j=1}^{p_T} p_{\lambda_T}(|\beta_{Tj}|), \tag{2.7}$$

where $p_T = N_T^2 P_T$ is the total number of parameters in $\boldsymbol{\beta}_T$.

Theorem 1. (Consistency) Suppose that the regression function $L_T(\boldsymbol{\theta}_T)$ satisfies conditions (B_1)–(B_3) in Appendix A, and the penalty function $p_{\lambda_T}(\cdot)$ satisfies conditions (A_2)–(A_4) in Appendix A. If $p_T^4/T \to 0$ as $T \to \infty$, then there exists a local minimizer $\text{vec}(\widehat{\mathbf{A}}_T^{SC})$ of (2.5) such that $\|\text{vec}(\widehat{\mathbf{A}}_T^{SC}) - \text{vec}(\mathbf{A}_T^*)\| = O_p\left(\sqrt{p_T}\left(T^{-1/2} + a_T\right)\right)$, where a_T is defined in Appendix A.

$$\begin{aligned}
\text{Let}\quad \mathbf{b}_T &= \left\{p'_{\lambda_T}(|\beta^*_{T1}|)\,\text{sgn}(\beta^*_{T1}), \ldots, p'_{\lambda_T}(|\beta^*_{Ts_T}|)\,\text{sgn}(\beta^*_{Ts_T})\right\}', \\
\mathbf{b}_T &= \mathbf{P}\tilde{\mathbf{b}}_T, \\
\boldsymbol{\Sigma}_{\lambda_T} &= \text{diag}\left\{p''_{\lambda_T}(\beta^*_{T1}), \ldots, p''_{\lambda_T}(\beta^*_{Ts_T})\right\}.
\end{aligned}$$

where $\tilde{\mathbf{b}}_T = \left\{p'_{\lambda_T}(|\phi^*_{T1}|)\,\text{sgn}(\phi^*_{T1}), \ldots, p'_{\lambda_T}(|\phi^*_{Ts_T}|)\,\text{sgn}(\phi^*_{Ts_T})\right\}'$, $\boldsymbol{\phi}_T = \text{vec}(\mathbf{A}_T)$, and $\mathbf{P}$ is the permutation matrix. Then we have $\boldsymbol{\beta}_T = \mathbf{P}\boldsymbol{\phi}_T$.

Theorem 2. (Oracle property) Suppose that conditions (A_1)–(B_4) in Appendix A hold. If $\lambda_T \to 0$, $\sqrt{T/p_T}\lambda_T \to \infty$, and $p_T^5/T \to 0$ as $T \to \infty$, then, with probability tending to 1, the $\sqrt{T/p_T}$-consistent local minimizer $\text{vec}(\widehat{\mathbf{A}}_T^{SC}) = (\text{vec}(\widehat{\mathbf{A}}_{T1}^{SC})', \text{vec}(\widehat{\mathbf{A}}_{T2}^{SC})')'$ obtained in Theorem 1 satisfies:

(i) Sparsity: $\text{vec}(\widehat{\mathbf{A}}_{T2}^{SC}) = 0$;

(ii) Asymptotic normality:

$$\sqrt{T}\mathbf{C}_T\mathbf{D}_{T1}^{-1/2}\{\mathbf{D}_{T1}+\boldsymbol{\Sigma}_{\lambda_T}\}$$
$$\times\,[\mathrm{vec}(\widehat{\mathbf{A}}_{T1}^{SC})-\mathrm{vec}(\mathbf{A}_{T1}^{*})+\{\mathbf{D}_{T1}+\boldsymbol{\Sigma}_{\lambda_T}\}^{-1}\tilde{\mathbf{b}}_T]\xrightarrow{D}\mathcal{N}(0,\boldsymbol{\Sigma}).$$

Remark 1. When the sample size T is finite and sufficiently large, $\boldsymbol{\Sigma}_{\lambda_T}=\mathbf{0}$ and $\tilde{\mathbf{b}}_T=\mathbf{0}$. Under this condition, the asymptotic normality result in part (ii) of Theorem 2 simplifies to $\sqrt{T}\mathbf{C}_T\mathbf{D}_{T1}^{1/2}\left(\mathrm{vec}(\widehat{\mathbf{A}}_{T1}^{SC})-\mathrm{vec}(\mathbf{A}_{T1}^{*})\right)\xrightarrow{D}\mathcal{N}(0,\boldsymbol{\Sigma})$. This result implies that $\mathrm{vec}(\widehat{\mathbf{A}}_{T1}^{SC})$ attains the same asymptotic efficiency as the spatial quantile regression (SQR) estimator of $\mathrm{vec}(\widehat{\mathbf{A}}_{T1})$ under the oracle scenario where $\mathrm{vec}(\mathbf{A}_{T2})=\mathbf{0}$ is known a priori.

2.2.2 Model selection with adaptive-LASSO penalty

As a widely used variable selection technique, the LASSO method, introduced by Tibshirani (1996) (45), imposes an L_1 penalty to encourage sparsity in regression coefficients. Building on this, Zou (2006) (47) proposed the adaptive LASSO, which improves upon the traditional LASSO by assigning adaptive, data-driven weights to each parameter, thereby achieving the oracle property.

In this section, we extend the adaptive LASSO framework to the SQR estimation of model (2.3). Let $\widetilde{\mathbf{A}}_T$ denote the solution to the unpenalized problem $\min_{\mathbf{A}_T,\mathbf{q}_u} L_T(\mathbf{A}_T,\mathbf{q}_u)$. By applying similar arguments as those used in the proof of Theorem 1, it can be shown that $\mathrm{vec}(\widetilde{\mathbf{A}}_T)$ is $\sqrt{T/p_T}$-consistent.

This preliminary estimate $\widetilde{\mathbf{A}}_T$ serves as the basis for constructing adaptive weights for the penalty function. Specifically, define the weight for each coefficient as $\tilde{w}_{Tij}=|\widetilde{A}_{Tij}|^{-\gamma}$ for some $\gamma>0$. The adaptive LASSO penalized SQR estimator is then obtained by solving:

$$(\widehat{\mathbf{A}}_T^{AL},\widehat{\mathbf{q}}_u)=\arg\min_{\mathbf{A}_T,\mathbf{q}_u} Q_T^{AL}(\mathbf{A}_T,\mathbf{q}_u), \tag{2.8}$$

where $Q_T^{AL}(\mathbf{A}_T,\mathbf{q}_u)=L_T(\mathbf{A}_T,\mathbf{q}_u)+Th_T\sum_{i=1}^{N_T}\sum_{j=1}^{N_TP_T}\tilde{w}_{Tij}|A_{Tij}|)$, and h_T is a non-negative regularization parameter. For ease of reference, we refer to this estimator as the SQRADLASSO method.

To facilitate theoretical analysis and derivation of asymptotic properties, we express the transition matrix in vectorized form. Following the same transformation as earlier, equation (2.4) can be reformulated as:

$$L_T(\boldsymbol{\theta}_T) + Th_T \sum_{j=1}^{p_T} \tilde{w}_{Tj}(|\beta_{Tj}|), \tag{2.9}$$

where $\boldsymbol{\theta}_T = (\boldsymbol{\beta}_T, \mathbf{q}_u)$, $\boldsymbol{\beta}_T = \mathrm{vec}(\mathbf{A}_T')$, and $\tilde{w}_{Tj} = |\tilde{\beta}_{Tj}|^{-\gamma}$ denotes the adaptive weight corresponding to the j-th element of $\boldsymbol{\beta}_T$.

Theorem 3. (Consistency) Suppose that the regression function $L_T(\boldsymbol{\theta}_T)$ satisfies conditions (B_1)–(B_3) in Appendix A. If $p_T^4/T \to 0$ and $\sqrt{T}h_T \to 0$ as $T \to \infty$, then there exists a local minimizer $\mathrm{vec}(\widehat{\mathbf{A}}_T^{AL})$ of (2.8) such that $\|\mathrm{vec}(\widehat{\mathbf{A}}_T^{AL}) - \mathrm{vec}(\mathbf{A}_T^*)\| = O_p(\sqrt{p_T/T})$.

$$\begin{aligned} \text{Let} \quad \mathbf{d}_T &= \left\{\mathrm{sgn}(\beta_{T1}^*)/|\tilde{\beta}_{T1}|^{\gamma}, \ldots, \mathrm{sgn}(\beta_{Ts_T}^*)/|\tilde{\beta}_{Ts_T}|^{\gamma})\right\}', \\ \mathbf{d}_T &= \mathbf{P}\tilde{\mathbf{d}}_T \end{aligned}$$

where $\tilde{\mathbf{d}}_T = \left\{\mathrm{sgn}(\phi_{T1}^*)/|\tilde{\phi}_{T1}|^{\gamma}, \ldots, \mathrm{sgn}(\phi_{Ts_T}^*)/|\tilde{\phi}_{Ts_T}|^{\gamma})\right\}'$ and $\tilde{\boldsymbol{\phi}}_T = \mathrm{vec}(\widetilde{\mathbf{A}}_T)$.

Theorem 4. (Oracle property) Suppose that the conditions of Theorem 3 and condition (B_5) in Appendix A hold. If $(T/p_T)^{(\gamma+1)/2}h_T \to \infty$, then, with probability tending to 1, the $\sqrt{T/p_T}$-consistent local minimizer $\mathrm{vec}(\widehat{\mathbf{A}}_T^{AL}) = (\mathrm{vec}(\widehat{\mathbf{A}}_{T1}^{AL})', \mathrm{vec}(\widehat{\mathbf{A}}_{T2}^{AL})')'$ obtained in Theorem 3 satisfies:

(i) Sparsity: $\mathrm{vec}(\widehat{\mathbf{A}}_{T2}) = 0$;

(ii) Asymptotic normality:

$$\sqrt{T}\mathbf{C}_T\mathbf{D}_{T1}^{1/2}[\mathrm{vec}(\widehat{\mathbf{A}}_{T1}^{AL}) - \mathrm{vec}(\mathbf{A}_{T1}^*) + \mathbf{D}_{T1}^{-1}h_T\tilde{\mathbf{d}}_T] \xrightarrow{D} \mathcal{N}(0, \boldsymbol{\Sigma}).$$

Remark 2. It is important to note that when T is finite and sufficiently large, the bias term $\tilde{\mathbf{d}}_T$ in Theorem 4 is generally nonzero and therefore cannot be ignored. However, under condition (B_5) in Appendix A, we have $\sqrt{T}h_T\tilde{\mathbf{d}}_T \to \mathbf{0}$ as $T \to \infty$. Thus, the asymptotic bias vanishes in the limit, and part (ii) of Theorem 4 simplifies to

$\sqrt{T}\mathbf{C}_T\mathbf{D}_{T1}^{1/2}\left(\text{vec}(\widehat{\mathbf{A}}_{T1}^{AL}) - \text{vec}(\mathbf{A}_{T1}^*)\right) \xrightarrow{D} \mathcal{N}(0, \mathbf{\Sigma})$. When this result is combined with the conclusion in Remark 1, it confirms that both the SQRADLASSO and SQRSCAD estimators achieve the oracle property asymptotically.

2.3 ADMM-Based Optimization and Tuning Parameter Selection

2.3.1 Using ADMM Algorithm to estimate $\mathbf{A}_T$ and $\mathbf{q}_u$

In this section, we introduce the Alternating Direction Method of Multipliers (ADMM) to estimate $\mathbf{A}_T$ and $\mathbf{q}_u$ in the penalized spatial quantile regression (SQR) model.

To facilitate efficient computation, we introduce an auxiliary variable $\mathbf{Z}_T$ and reformulate the optimization problem in equation (2.4) as:

$$\min_{\mathbf{A}_T, \mathbf{q}_u, \mathbf{Z}_T} \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \mathbf{A}_T\mathbf{x}_t - \mathbf{q}_u) + T\sum_{i=1}^{N_T}\sum_{j=1}^{N_T P_T} p_{\lambda_T}(|Z_{Tij}|) \quad \text{subject to} \quad \mathbf{A}_T = \mathbf{Z}_T,$$

where $p_{\lambda_T}(\cdot)$ is a penalty function such as SCAD or adaptive LASSO.

The corresponding augmented Lagrangian function is given by:

$$\mathcal{L}(\mathbf{A}_T, \mathbf{q}_u, \mathbf{Z}_T, \mathbf{\Lambda}) = \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \mathbf{A}_T\mathbf{x}_t - \mathbf{q}_u) + T\sum_{i=1}^{N_T}\sum_{j=1}^{N_T P_T} p_{\lambda_T}(|Z_{Tij}|) + \frac{\rho}{2}\|\mathbf{A}_T - \mathbf{Z}_T + \mathbf{\Lambda}\|_F^2,$$

where $\mathbf{\Lambda}$ is the dual variable and $\rho > 0$ is a tuning parameter that controls the penalty on the constraint violation. This reformulation leads to efficient update steps for $\mathbf{A}_T$ and $\mathbf{q}_u$, which are iteratively implemented in Algorithm 1.

Note that the $\mathbf{A}_T$-update and $\mathbf{q}_u$-update steps can be efficiently solved using iterative optimization techniques such as iterative reweighted least squares (IRLS), gradient descent, or Newton's method, depending on the structure of the loss function. The $\mathbf{Z}_T$-update corresponds to a proximal operator problem, which admits a closed-form solution via element-wise soft-thresholding.

As a result, the ADMM algorithm alternately updates $\mathbf{A}_T$, $\mathbf{q}_u$, $\mathbf{Z}_T$, and the dual variable to iteratively minimize the augmented Lagrangian. This approach not only

Algorithm 1 ADMM algorithm

```
Initialize: $\mathbf{A}_T^{(0)}, \mathbf{q}_u^{(0)}, \mathbf{Z}_T^{(0)}, \mathbf{\Lambda}^{(0)}$
repeat
    $\mathbf{A}_T^{(k+1)}, \mathbf{q}_u^{(k+1)} \leftarrow \underset{\mathbf{A}_T, \mathbf{q}_u}{\arg\min} \left\{ \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \mathbf{A}_T \mathbf{x}_t - \mathbf{q}_u) + \frac{\varrho}{2} \|\mathbf{A}_T - \mathbf{Z}_T^{(k)} + \mathbf{\Lambda}^{(k)}\|_F^2 \right\}$
    $\mathbf{Z}_T^{(k+1)} \leftarrow \underset{\mathbf{Z}_T}{\arg\min} \left\{ T \sum_{i=1}^{N_T} \sum_{j=1}^{N_T P_T} p_\lambda(|Z_{Tij}|) + \frac{\varrho}{2} \|\mathbf{A}_T^{(k+1)} - \mathbf{Z}_T + \mathbf{\Lambda}^{(k)}\|_F^2 \right\}$
    $\mathbf{\Lambda}^{(k+1)} \leftarrow \mathbf{\Lambda}^{(k)} + (\mathbf{A}_T^{(k+1)} - \mathbf{Z}_T^{(k+1)})$
    Check Convergence
    If $\|\mathbf{A}_T^{(k+1)} - \mathbf{Z}_T^{(k+1)}\|_F < \epsilon$ and $\|\mathbf{Z}_T^{(k+1)} - \mathbf{Z}_T^{(k)}\|_F < \epsilon$, then stop.
until convergence
Return: Optimal $\mathbf{A}_T, \mathbf{q}_u, \mathbf{Z}_T, \mathbf{\Lambda}$
```

enhances computational efficiency but also facilitates the enforcement of sparsity through penalization. Moreover, the framework is flexible and can be extended to accommodate alternative penalty functions or additional structural constraints as needed.

2.3.2 Choice of the tuning parameters

For penalized SQR estimators, selecting appropriate tuning parameters λ_T and h_T is essential for SCAD and adaptive LASSO penalties, respectively. These parameters can be selected using similar procedures. Here, we focus on the choice of λ_T for illustration. Common approaches include the generalized cross-validation (GCV) criterion (Wang, Li, and Tsai, 2007 (54)) and the Schwarz Information Criterion (SIC), as discussed in Koenker, Ng, and Portnoy (1994) (55) and Zou and Yuan (2008b) (56).

Since the resulting estimators depend on the tuning parameter λ_T, we denote them by $(\widehat{\mathbf{A}}_{\lambda_T}, \widehat{\mathbf{q}}_{u_{\lambda_T}})$ to emphasize this dependence. Using the SIC method, we propose selecting λ_T by minimizing the following criterion:

$$SIC(\lambda_T) = \log \left\{ \frac{1}{T} L_T(\widehat{\mathbf{A}}_{\lambda_T}, \widehat{\mathbf{q}}_{u_{\lambda_T}}) \right\} + \frac{\log(T)}{2T} df(\lambda_T),$$

where $df(\lambda_T)$ is the effective degrees of freedom of the fitted model that calibrates the complexity of model.

Following Koenker, Ng, and Portnoy (1994) (55), for each given λ_T we define the index set

$$\mathcal{E}_{\lambda_T} = \left\{ t : \mathbf{y}_t - \widehat{\mathbf{A}}_{\lambda_T} \mathbf{x}_t - \widehat{\mathbf{q}}_{u_{\lambda_T}} = \mathbf{0} \right\},$$

and estimate $df(\lambda_T)$ using the cardinality $|\mathcal{E}_{\lambda_T}|$, as suggested by Jiang, Jiang, and Song (2012) (57). Consequently, the tuning parameter λ_T is estimated by

$$\hat{\lambda}_T = \arg\min_{\lambda_T} \left\{ \log\left(\frac{1}{T} L_T(\widehat{\mathbf{A}}_{\lambda_T}, \widehat{\mathbf{q}}_{u_{\lambda_T}}) \right) + \frac{\log(T)}{2T} |\mathcal{E}_{\lambda_T}| \right\}.$$

2.4 Simulation Studies and Real Data Application

2.4.1 Simulation studies

In this section, we present simulation results to evaluate the finite-sample performance of the SQR estimators and their associated model selection capabilities. Specifically, we consider a stationary VAR(2) model given by:

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t.$$

which can be rewritten in compact form as:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t.$$

where $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$ is the coefficient matrix, and $\mathbf{x}_t = (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2})'$ denotes the stacked lagged vectors. The innovation process $\boldsymbol{\varepsilon}_t$ follows a specified distribution, to be detailed later.

The true coefficient matrix is specified as:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2] = \begin{bmatrix} -0.8 & 0.6 & 0 & 0 \\ -0.7 & 0.5 & 0 & 0 \end{bmatrix}$$

ensuring the stationarity of the VAR(2) process.

We investigate the performance of two penalization methods: with $\gamma = 1$, defined by $h_T \sum_{i=1}^{N_T} \sum_{j=1}^{N_T P_T} |A_{Tij}|/|\tilde{A}_{Tij}|$; the SCAD penalty, defined by $\sum_{i=1}^{N_T} \sum_{j=1}^{N_T P_T} p_{\lambda_T}(|A_{ij}|)$, where h_T and λ_T are tuning parameters and $\tilde{A}_{Tij}$'s are consistent estimators of A_{Tij}'s. For the adaptive LASSO, we adopt $\gamma = 1$, corresponding to the nonnegative garrote (Breiman, 1995 (58)), as discussed in Zou (2006) (59). Although other choices for γ are possible, no universally optimal value exists.

The tuning parameters are selected via the Schwarz Information Criterion (SIC), as described in the previous section.

We generated synthetic data from the working VAR(2) model for sample sizes $T = 300, 400, 500,$ and 600. For each value of T, we conducted 500 Monte Carlo replications. In each simulation, the time series data were independently generated according to the VAR(2) specification described earlier.

To assess the robustness of the proposed estimation methods under different distributional settings, we considered four types of error distributions for the innovation term:

i. $\boldsymbol{\varepsilon}_t$ follows a bivariate normal distribution with mean $\mu = (0,0)'$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$.

ii. $\boldsymbol{\varepsilon}_t$ follows a bivariate t-distribution with degree of freedom 3 and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$.

iii. 95% of data points follow a bivariate normal distribution with mean $\mu = (0,0)'$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$. The remaining 5% follow a normal distribution with mean $\mu = (0,0)'$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 25.0 & 0.0 \\ 0.0 & 25.0 \end{bmatrix}$.

iv. $\boldsymbol{\varepsilon}_t$ follows a bivariate laplace distribution with mean $\mu = (0,0)'$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$.

These different error structures allow us to evaluate the performance of the penalized SQR estimators under both light-tailed and heavy-tailed scenarios.

We compared five estimation methods: the penalized SQR estimation (SCAD, AD-LASSO, and LASSO), the oracle maximum likelihood (OML) estimator, and the oracle unpenalized spatial quantile regression (SQR-Oracle) estimator. In each simulation, the "root of mean squared errors (RMSE)" for different coefficient estimators were calculated, and their average over simulations is reported in Tables 2.1 – 2.4, where Σ denotes the sum of RMSE for all components in $\mathbf{A}$.

To assess variable selection accuracy, we computed the true positives (TP) and false positives (FP). Specifically, TP denotes the average number of non-zero coefficients correctly identified, while FP represents the average number of zero coefficients incorrectly estimated as non-zero, see Tables 2.1−2.4. A coefficient was considered zero if its estimate was smaller than 10^{-8} in absolute value.

Table 2.1: RMSE (multiplied by 10^3) of penalized estimators under the normal error.

	$n = 300$							$n = 400$						
Estimates	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP
OML	63	67	62	67	259			55	57	56	57	224		
SQR-Oracle	72	77	71	77	297			62	64	62	65	252		
SC-SQR	81	80	87	84	396	4	0.030	67	67	73	68	317	4	0.012
AD-SQR	78	78	86	84	405	4	0.038	66	65	72	68	336	4	0.028
LA-SQR	79	78	81	78	637	4	2.446	67	65	70	65	543	4	2.210
	$n = 500$							$n = 600$						
OML	48	50	50	51	200			43	45	45	47	179		
SQR-Oracle	53	56	56	59	224			47	51	50	54	202		
SC-SQR	59	58	63	61	252	4	0.002	54	53	57	55	219	4	0.000
AD-SQR	57	57	63	61	254	4	0.004	52	52	57	55	216	4	0.000
LA-SQR	58	58	62	60	485	4	2.076	53	53	55	55	445	4	1.964

Note: SC-SCAD, AD-Adaptive LASSO, LA-LASSO.

Table 2.2: RMSE (multiplied by 10^3) of penalized estimators under the t(3) error.

	$n = 300$							$n = 400$						
Estimates	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP
OML	48	51	47	48	194			39	43	39	42	163		
SQR-Oracle	50	53	48	51	202			40	45	40	44	169		
SC-SQR	54	56	55	53	244	4	0.008	45	47	45	45	190	4	0.002
AD-SQR	54	56	55	52	250	4	0.008	44	47	45	45	199	4	0.008
LA-SQR	54	56	55	53	449	4	1.948	45	47	45	45	373	4	1.646
	$n = 500$							$n = 600$						
OML	35	38	36	37	145			31	34	32	33	130		
SQR-Oracle	36	39	37	39	151			32	36	33	35	135		
SC-SQR	39	41	41	39	169	4	0.002	36	36	36	35	143	4	0.000
AD-SQR	39	41	41	39	175	4	0.004	35	36	36	35	143	4	0.000
LA-SQR	39	41	41	40	330	4	1.476	35	36	36	35	292	4	1.322

Note: SC-SCAD, AD-Adaptive LASSO, LA-LASSO.

The simulation results show that the OML estimator achieves the best overall performance, with the lowest RMSE across all settings, serving as a natural performance

Table 2.3: RMSE (multiplied by 10^3) of penalized estimators under the mixed normal error.

	$n = 300$							$n = 400$						
Estimates	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP
OML	34	36	33	35	137			29	32	28	31	121		
SQR-Oracle	39	40	38	39	156			33	36	32	34	136		
SC-SQR	51	51	48	41	210	4	0.006	42	45	39	36	171	4	0.002
AD-SQR	48	48	48	41	214	4	0.014	38	43	39	36	167	4	0.004
LA-SQR	47	50	49	45	414	4	1.902	40	43	40	39	344	4	1.642
	$n = 500$							$n = 600$						
OML	26	28	25	27	107			23	26	23	24	97		
SQR-Oracle	29	32	28	30	119			26	29	26	27	108		
SC-SQR	37	38	34	33	142	4	0.000	33	34	31	30	129	4	0.000
AD-SQR	35	37	34	33	147	4	0.004	33	33	31	30	132	4	0.002
LA-SQR	36	37	35	35	298	4	1.362	33	34	32	32	269	4	1.208

Note: SC-SCAD, AD-Adaptive LASSO, LA-LASSO.

Table 2.4: RMSE (multiplied by 10^3) of penalized estimators under the laplace error.

	$n = 300$							$n = 400$						
Estimates	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP	$\widehat{A}_{1,11}$	$\widehat{A}_{1,12}$	$\widehat{A}_{1,21}$	$\widehat{A}_{1,22}$	Σ	TP	FP
OML	44	48	43	47	181			38	42	37	40	157		
SQR-Oracle	45	49	44	48	187			39	43	38	41	161		
SC-SQR	51	52	50	49	228	4	0.010	43	45	41	41	171	4	0.000
AD-CQR	49	51	47	49	233	4	0.012	42	45	41	42	176	4	0.002
LA-CQR	49	51	49	49	405	4	1.746	42	45	41	42	339	4	1.414
	$n = 500$							$n = 600$						
OML	33	35	32	35	134			30	32	28	31	122		
SQR-Oracle	34	36	32	35	136			30	33	29	32	124		
SC-SQR	37	38	34	36	145	4	0.000	33	35	31	32	132	4	0.000
AD-SQR	36	37	34	36	151	4	0.002	33	34	31	32	138	4	0.002
LA-SQR	36	37	35	37	293	4	1.236	33	34	31	33	263	4	1.098

Note: SC-SCAD, AD-Adaptive LASSO, LA-LASSO.

benchmark. SQR-Oracle estimator provides a strong reference point, achieving RMSE close to OML and illustrating the potential of spatial quantile regression when the true sparsity structure is known. The SCAD and ADLASSO penalized SQR estimators perform comparably to the oracle estimator, demonstrating both low RMSE and strong variable selection ability. In contrast, the LASSO-based SQR method performs the worst, both in terms of estimation error and variable selection accuracy. As the sample size increases, the number of false positives consistently decreases across all methods, indicating improved variable selection in larger samples.

In summary, the SCAD and ADLASSO penalized SQR methods demonstrate strong performance in terms of robustness, estimation accuracy, and sparsity recovery, nearly matching the oracle estimator under large sample sizes. The results confirm the advantages of incorporating penalization into spatial quantile regression, particularly in high-dimensional and heavy-tailed settings.

2.4.2 Real data applications

In this section, we analyze an air pollution dataset comprising multivariate time series with dimension $N = 5$. The series includes four key air pollutants—carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO_2), and ozone (O_3)—along with solar radiation intensity (R). These variables were measured hourly throughout the year 2006 at a monitoring station in Azusa, California. The data were obtained from the Air Quality and Meteorological Information System (AQMIS), resulting in a total of $T = 8370$ observations.

To facilitate an initial exploration, we compute the daily averages of each component over the one-year period. The resulting time series plots are presented in Figure 2.1, providing a visual overview of the temporal patterns and potential interactions among the variables.

When incorporating regularization techniques such as Adaptive LASSO or SCAD into the VAR model, standardization becomes an essential preprocessing step. These methods impose penalties on model coefficients to control complexity and prevent overfitting; however, the penalties are scale-sensitive. If variables differ substantially in magnitude or

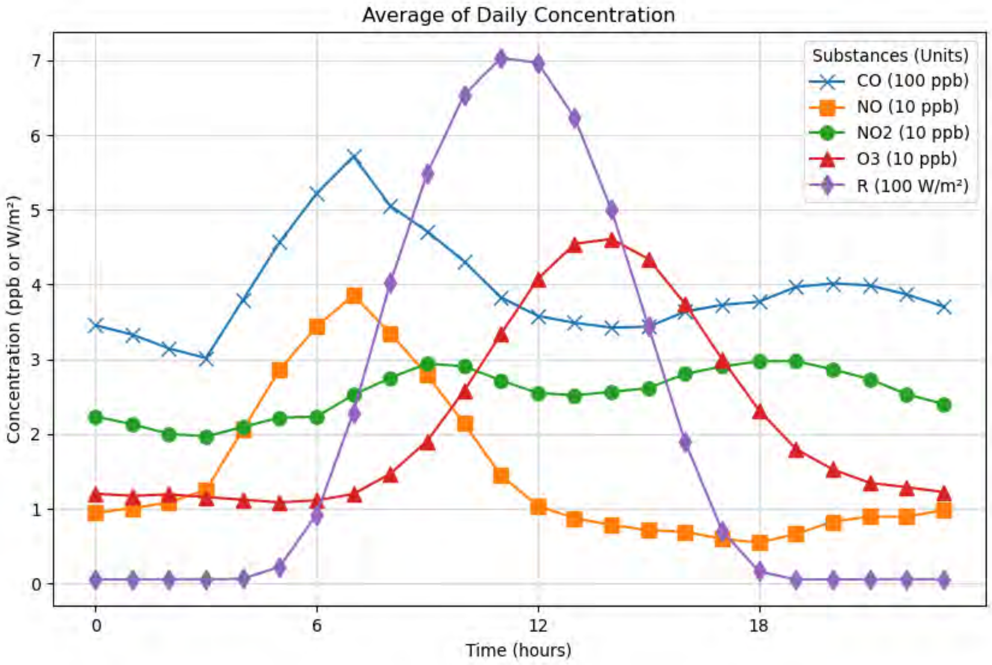


Figure 2.1: Average of Daily Concentration of CO, NO, NO_2, and O_3, and the solar radiation R.

units, the regularization may disproportionately penalize those with larger scales, leading to biased estimates and distorted model interpretation.

Standardization addresses this issue by rescaling all variables to have zero mean and unit variance, thereby placing them on a comparable scale. This ensures that the penalty terms are applied equitably across variables, allowing for a fair assessment of each predictor's relative importance. Without standardization, predictors with larger scales may dominate the penalty term, undermining the effectiveness of regularization and potentially resulting in suboptimal estimates.

By standardizing the variables, we enhance both the stability and interpretability of the regularized VAR model. Therefore, standardization was applied to the air pollution dataset described earlier. The standardized time series are shown in Figure 2.2.

We applied four estimation methods—SQR, SQRSCAD, SQRADLASSO, and SQRLASSO—to analyze the standardized dataset. Following the approach in Davis, Zang, and Zheng (2016) (60), we set the lag order to $P = 4$. After estimating the model coefficients using each method, we computed daily predictions from hour 4 to hour 23.

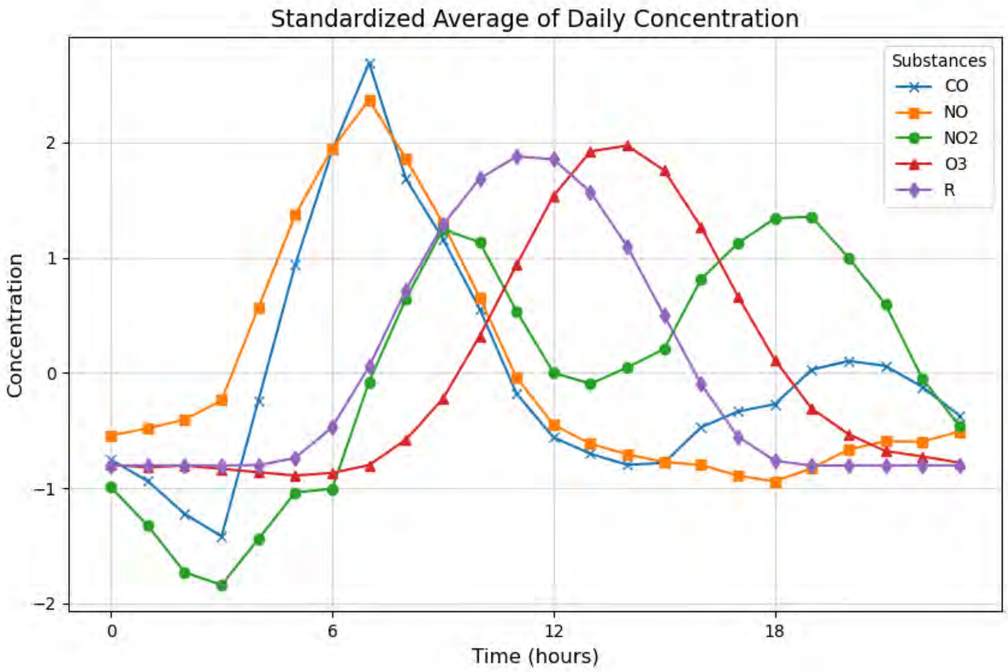


Figure 2.2: Standardized Average of Daily Concentration of CO, NO, NO_2, and O_3, and the solar radiation R.

The comparison between the predicted and observed values for each method is presented in Figures 2.3–2.6. Additionally, Table 2.5 reports the root mean squared errors (RMSE) corresponding to the different estimation techniques.

Table 2.5: RMSE under Different Estimation Methods.

	SQR	SQRSCAD	SQRADLASSO	SQRLASSO
RMSE	1.0	1.2	1.2	1.6

In addition, Figure 2.7 presents the heatmaps of the estimated coefficient matrices for the four estimation methods. These visualizations provide insight into the structure of the estimated parameters, particularly in terms of sparsity and the relative importance of individual coefficients. The color intensity reflects the magnitude of the estimates: darker shades of blue indicate larger absolute values, whereas lighter shades (closer to white) correspond to smaller or near-zero estimates.

The four estimation methods, SQR, SQRADLASSO, SQRSCAD, and SQRLASSO, exhibit distinct characteristics in balancing model sparsity and estimation accuracy. The SQR method delivers the highest estimation accuracy, with predicted values closely

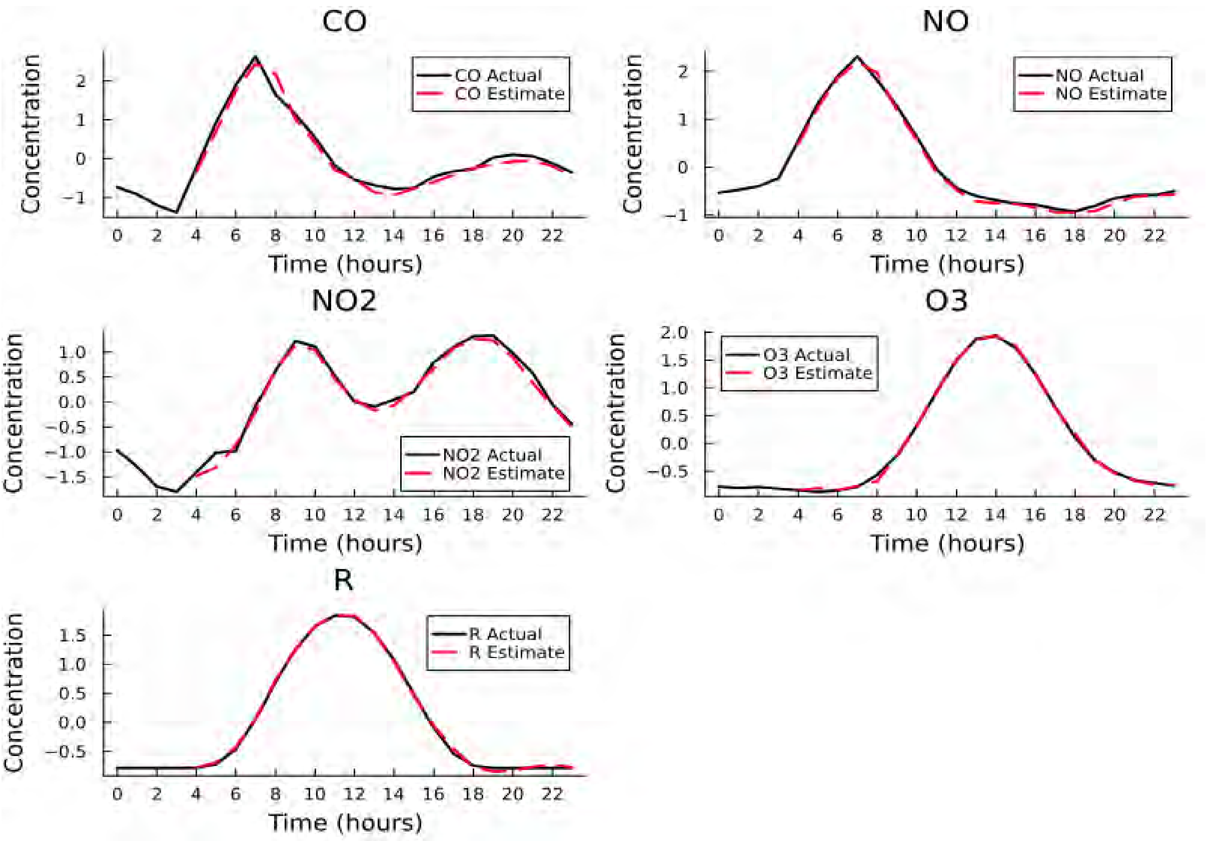


Figure 2.3: Comparison between the estimated and true values for SQR.

matching the true observations, as illustrated in Figure 2.3. However, its lack of sparsity limits interpretability in high-dimensional contexts, as shown in Figure 2.7(a).

To address this issue, SQRSCAD and SQRADLASSO incorporate regularization to induce sparsity, effectively setting many coefficients to zero while preserving estimation accuracy. As evidenced by their RMSE values, both methods achieve prediction performance comparable to that of SQR. Among them, SQRADLASSO offers a balanced trade-off between sparsity and accuracy through adaptive penalization, while SQRSCAD further reduces the risk of over-shrinkage, resulting in slightly more retained nonzero coefficients.

In contrast, SQRLASSO yields the sparse model but tends to over-penalize, leading to the omission of relevant variables and consequently higher estimation errors. The heatmaps of the estimated coefficients in Figure 2.7 visually underscore these trade-offs. Overall, both SQRSCAD and SQRADLASSO emerge as the most effective methods, striking an balance between model interpretability and predictive performance.

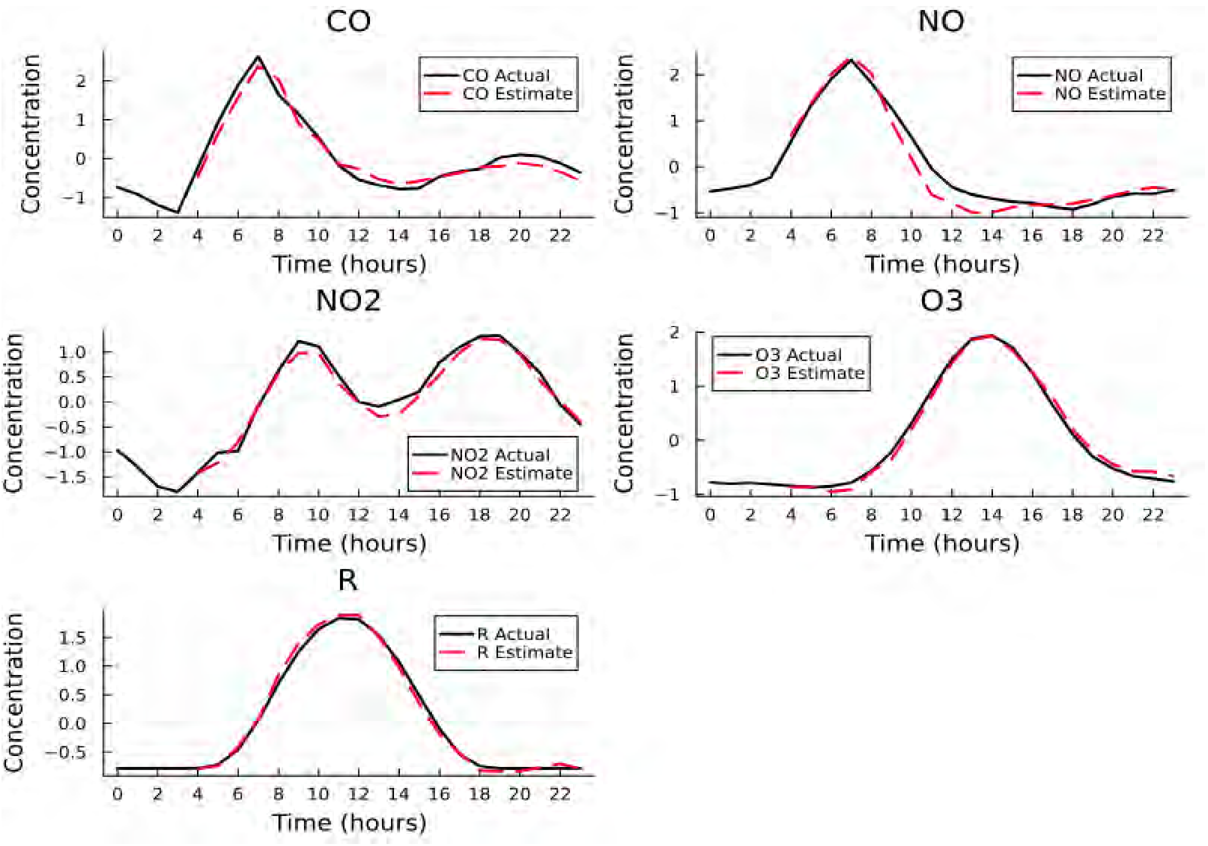


Figure 2.4: Comparison between the estimated and true values for SQRSCAD.

In summary, the empirical analysis demonstrates that while SQR achieves the highest predictive accuracy, it lacks the sparsity required for interpretability in high-dimensional settings. Both SQRSCAD and SQRADLASSO effectively introduce sparsity without sacrificing much in estimation accuracy, making them well-suited for practical applications where variable selection is critical. Among these, SQRADLASSO offers a more balanced trade-off, whereas SQRSCAD provides slightly more flexibility in retaining informative variables. Although SQRLASSO enforces the sparsity, it may lead to increased estimation error due to over-penalization. Overall, SQRSCAD and SQRADLASSO emerge as the preferred choices for achieving both interpretability and predictive performance in penalized spatial quantile regression for VAR models.

An important extension of this work is to incorporate tensor decomposition structures into the regularized VAR modeling framework. Specifically, representing the coefficient matrices of a high-dimensional VAR model as low-rank tensors using Tucker or CP decompositions can significantly reduce model complexity while preserving the multidimensional

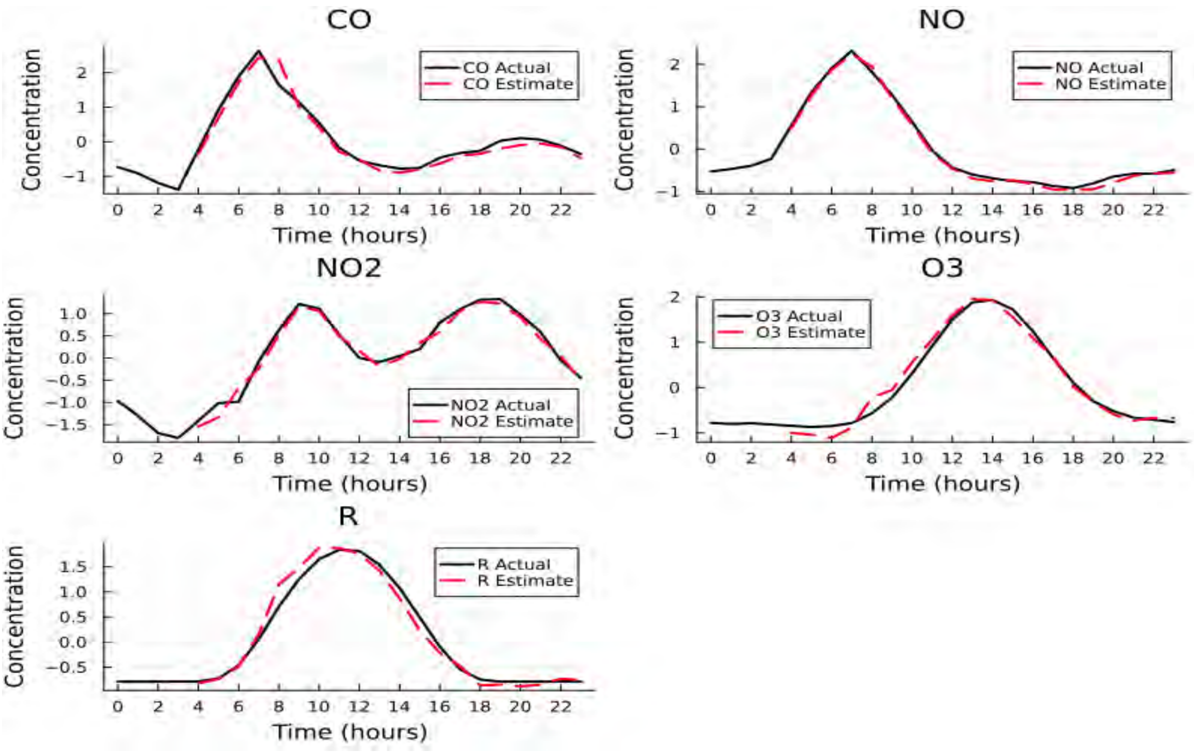


Figure 2.5: Comparison between the estimated and true values for SQRADLASSO.

dependency structure. By integrating such tensor representations with sparsity-inducing penalties—such as adaptive LASSO, SCAD, or group penalties—one can achieve simultaneous dimensionality reduction and variable selection. This hybrid approach would be particularly beneficial in settings where the number of variables and lags is very large, leading to severe overparameterization.

This approach is particularly appealing in large-scale time series analysis, where the dimensionality of the coefficient space grows rapidly with the number of variables and lags. Developing efficient algorithms for penalized tensor-based SQR estimation and studying their theoretical properties—such as estimation consistency and variable selection consistency—would be a valuable direction for future research.

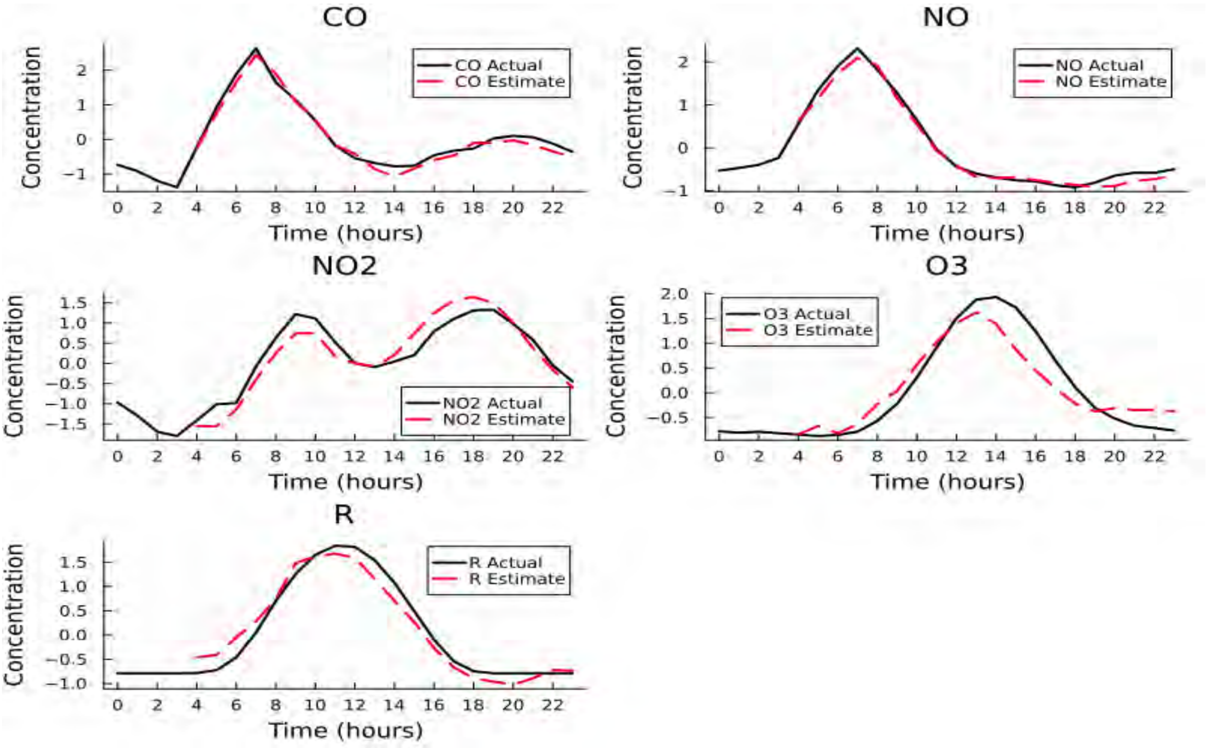
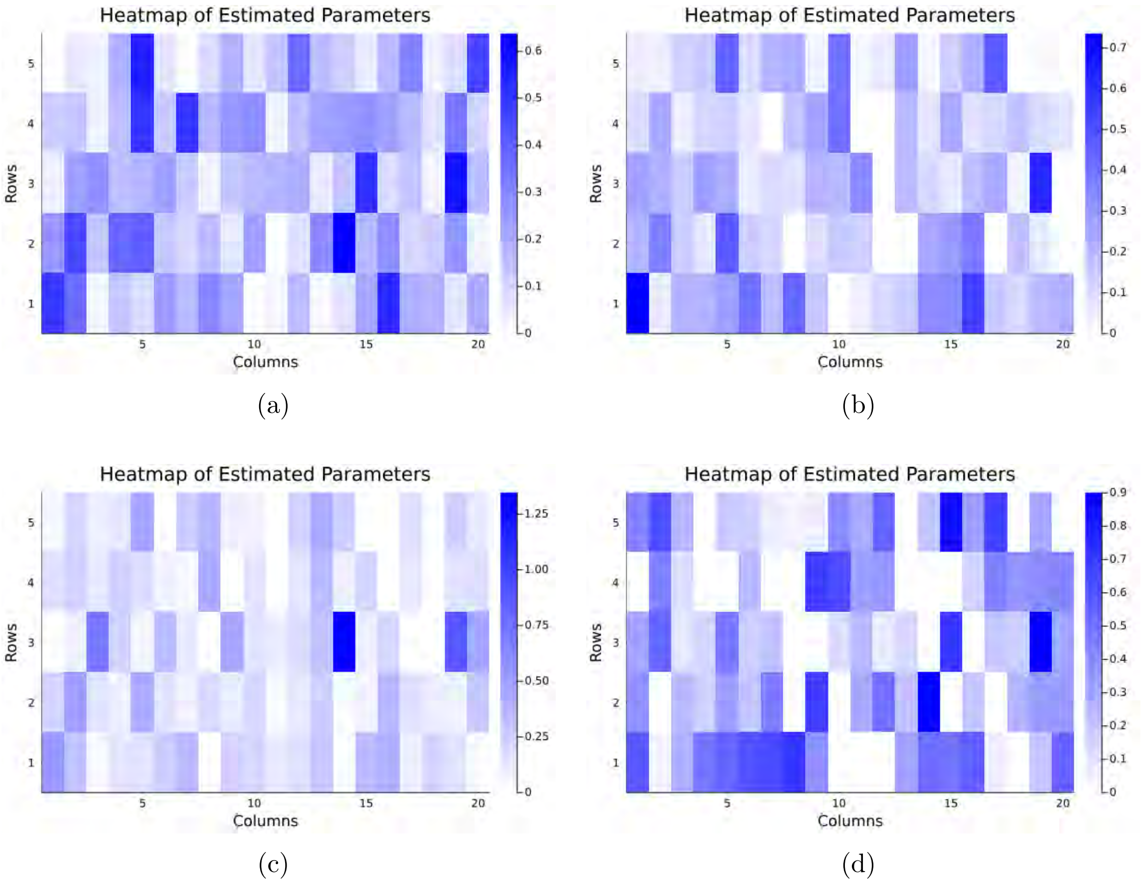


Figure 2.6: Comparison between the estimated and true values for SQRLASSO.



(a)

(b)

(c)

(d)

Figure 2.7: Heatmap of (a) SQR, (b) SARSCAD, (c) SQRADLASSO, and (d) SQR-LASSO.

CHAPTER 3: SPATIAL QUANTILE ESTIMATION OF HIGH-DIMENSIONAL VAR(P) VIA TENSOR DECOMPOSITION

3.1 Overview

As mentioned in the Introduction, it is natural to consider integrating the transition matrix of the VAR model with a tensor structure. Therefore, in this chapter, we propose an estimation method that combines spatial quantile regression with tensor decomposition.

In addition to multilinear structure mentioned in the Introduction, we incorporate sparsity into the tensor decomposition for further interpretability and estimation efficiency. Prior work has explored sparsity in matrix decomposition through various strategies. For example, Chen and Huang (2012 (61)) and Bunea, She, and Wegkamp (2012 (62)) considered row-wise sparsity, while Lian, Feng, and Zhao (2015 (63)) imposed element-wise sparsity on the coefficient matrix. Chen, Chan, and Stenseth (2012 (64)) proposed a sparse singular value decomposition by relaxing orthogonality constraints, whereas Uematsu et al. (2019 (65)) simultaneously enforced both sparsity and strict orthogonality. Our method extends these ideas to the tensor framework, enabling a structured and interpretable representation of the transition matrices.

Our approach is further motivated by recent advances in tensor regression for high-dimensional data (Zhou, Li, and Zhu, 2013 (66); Li and Zhang, 2017 (67); Sun and Li, 2017 (68); Li et al., 2018 (69); Raskutti, Yuan, and Chen, 2019 (39)). By integrating tensor decomposition with spatial quantile regression, we propose a novel framework that achieves both robustness and scalability in high-dimensional VAR analysis.

The proposed model offers the following key advantages:

(i) Multidirectional Dimensionality Reduction: The Tucker decomposition reduces the parameter space along three structural dimensions, preserving essential dynamic dependencies and improving estimation robustness.

(ii) Sparse and Structured Estimation: ℓ_1-regularization on the factor matrices induces sparsity, enhancing model interpretability and efficiency.

(iii) Efficient Optimization via ADMM: The Alternating Direction Method of Multipliers effectively incorporates both sparsity and orthogonality constraints, ensuring computational efficiency and accurate estimation.

In summary, our framework provides a robust and interpretable approach to high-dimensional VAR modeling by combining multilinear low-rank tensor decomposition with spatial quantile regression, effectively addressing both over-parameterization and estimation instability.

The remainder of this article is organized as follows. Section 3.2 provides a detailed overview of tensor decomposition, and introduces the proposed modeling framework. In Section 3.3, we establish the asymptotic properties of the MLRSQR estimator and develop an alternating optimization algorithm for its implementation. Section 3.4 extends the methodology to incorporate sparse higher-order reduced-rank estimation, simultaneously addressing orthogonality and sparsity constraints, and derives the corresponding asymptotic theory. An efficient ADMM-based algorithm is also introduced for estimation. Section 3.5 presents a consistent procedure for rank selection. Section 3.6 conducts extensive simulation studies to assess the empirical performance of the proposed estimators, while Section 3.7 illustrates the practical utility of our method through an application to real-world data. All technical proofs are provided in the Appendix B.

3.2 Multilinear Low-Rank VAR models with Spatial Quantile Regression

3.2.1 Tensor Decomposition

Tensors, or multidimensional arrays, generalize vectors and matrices to higher dimensions, providing a natural framework for representing multiway data. A tensor of order K, denoted by $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, consists of elements indexed by K dimensions, commonly referred to as modes. This work focuses primarily on third-order tensors ($K = 3$), which form the foundation of our proposed modeling approach. For a comprehensive review of tensor operations and notation, we refer readers to Kolda and Bader (2009) (43).

Throughout this article, we adopt the following notational conventions: lowercase boldface letters (e.g., $\mathbf{x}, \mathbf{y}$) represent vectors, uppercase boldface letters (e.g., $\mathbf{X}, \mathbf{Y}$) denote matrices, and Euler script letters (e.g., $\mathcal{X}, \mathcal{Y}$) are used for tensors. The ℓ_1, ℓ_2, and ℓ_∞ norms of a vector $\mathbf{x}$ are defined as $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_\infty$, respectively. For a matrix $\mathbf{X}$, we define $\|\mathbf{X}\|_F$ as the Frobenius norm, $\|\mathbf{X}\|_1 = \|\text{vec}(\mathbf{X})\|_1$ as the vectorized ℓ_1 norm, $\|\mathbf{X}\|_0$ as the count of nonzero entries (the ℓ_0 norm), $\|\mathbf{X}\|_2$ as the spectral norm, $\|\mathbf{X}\|_{op}$ as the operator norm, and $\|\mathbf{X}\|_*$ as the nuclear norm. The notation $\text{vec}(\mathbf{X})$ denotes the column-wise vectorization of $\mathbf{X}$, and $\mathbf{X}'$ or $\mathbf{X}^T$ represents its transpose. We denote the j-th largest singular value of a matrix $\mathbf{X}$ by $\sigma_j(\mathbf{X})$.

For third-order tensors $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, we define the Frobenius norm as

$$\|\mathcal{X}\|_F = \left(\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} \mathcal{X}_{ijk}^2 \right)^{1/2},$$

and the ℓ_0 norm as

$$\|\mathcal{X}\|_0 = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} \mathbf{1}(\mathcal{X}_{ijk} \neq 0),$$

which counts the number of nonzero elements in $\mathcal{X}$. These definitions establish a unified notation framework that facilitates clarity and consistency throughout the theoretical development of the proposed methodology.

To facilitate tensor algebra and analysis, we frequently employ matricization, the process of unfolding a tensor into a matrix. For a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, its mode-1 matricization $\mathcal{X}_{(1)}$ arranges the mode-1 fibers as columns of a matrix in $\mathbb{R}^{p_1 \times (p_2 p_3)}$, where the entry at position $\{i, (k-1)p_2 + j\}$ corresponds to $\mathcal{X}_{ijk}$. Similarly, the mode-2 and mode-3 matricizations, denoted $\mathcal{X}_{(2)}$ and $\mathcal{X}_{(3)}$, are formed by reshaping along the second and third modes, respectively. These matricizations provide a crucial bridge between tensor operations and classical matrix analysis, enabling efficient computation and theoretical derivation.

A key operation in tensor analysis is mode-wise multiplication. For instance, the mode-

1 product between a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ and a matrix $\mathbf{Y} \in \mathbb{R}^{q_1 \times p_1}$ is defined by

$$\mathcal{X} \times_1 \mathbf{Y} = \left(\sum_{i=1}^{p_1} \mathcal{X}_{ijk} \mathbf{Y}_{si} \right)_{1 \le s \le q_1,\ 1 \le j \le p_2,\ 1 \le k \le p_3} .$$

Mode-2 and mode-3 multiplications, denoted by $\times_2$ and $\times_3$, are defined analogously. These multilinear products form the computational foundation of tensor decomposition methods.

The multilinear rank of a tensor characterizes its low-dimensional structure across different modes. For a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, the multilinear ranks (r_1, r_2, r_3) are defined as the ranks of its respective matricizations:

$$\begin{aligned} r_1 &= \operatorname{rank}(\mathcal{X}_{(1)}) = \dim \left(\operatorname{span} \left\{ \mathcal{X}_{[:,j,k]} \mid 1 \le j \le p_2,\ 1 \le k \le p_3 \right\} \right), \\ r_2 &= \operatorname{rank}(\mathcal{X}_{(2)}), \quad r_3 = \operatorname{rank}(\mathcal{X}_{(3)}). \end{aligned}$$

Unlike matrices, where row and column ranks are equal, the ranks across tensor modes can differ, providing a richer structural characterization. These multilinear ranks serve as the basis for Tucker decomposition, which enables compact representation and efficient inference in tensor models.

The Tucker decomposition (Tucker, 1966 (70); De Lathauwer, De Moor, and Vandewalle, 2000 (71)) expresses a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ as a multilinear transformation of a lower-dimensional core tensor $\mathcal{Y} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$:

$$\mathcal{X} = \mathcal{Y} \times_1 \mathbf{Y}_1 \times_2 \mathbf{Y}_2 \times_3 \mathbf{Y}_3,$$

where each factor matrix $\mathbf{Y}_j \in \mathbb{R}^{p_j \times r_j}$ for $j = 1, 2, 3$ encodes the basis vectors spanning the mode-j subspace. This decomposition projects the original tensor onto a low-dimensional multilinear subspace, preserving its essential structure while substantially reducing its dimensionality.

A more compact notation for the Tucker decomposition is given by

$$\mathcal{X} = [[\mathcal{Y}; \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3]],$$

which highlights the joint transformation of the core tensor by the factor matrices along each mode. The Tucker framework plays a central role in high-dimensional modeling, offering both structural interpretability and computational efficiency. In the context of our proposed model, it serves as a powerful tool for representing the transition dynamics of high-dimensional VAR processes under multilinear constraints.

3.2.2 Multilinear Low-Rank Vector Autoregression

To accommodate the complex dynamics of high-dimensional time series, we reformulate the traditional VAR model using a multilinear low-rank structure. This is achieved by organizing the set of transition matrices into a third-order tensor and applying Tucker decomposition for dimensionality reduction and structural regularization.

Consider the VAR(P) model in (1.1). The P transition matrices $\mathbf{A}_1, \ldots, \mathbf{A}_P$ can be naturally rearranged into a third-order tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times P}$, where the first two modes correspond to spatial dependencies, and the third mode encodes the temporal lag structure. This tensor representation facilitates the application of tensor decomposition techniques and offers a compact, structured formulation of the coefficient space. An illustration of this construction is provided in Figure 3.1.

Let $\mathcal{A}_{(j)}$ denote the mode-j matricization of $\mathcal{A}$ for $j = 1, 2, 3$. These matricizations offer alternative perspectives on the transition structure:

- $\mathcal{A}_{(1)} = (\mathbf{A}_1, \ldots, \mathbf{A}_P)$ captures the column space across lags.
- $\mathcal{A}_{(2)} = (\mathbf{A}_1', \ldots, \mathbf{A}_P')$ captures the row space.
- $\mathcal{A}_{(3)} = (\text{vec}(\mathbf{A}_1)', \ldots, \text{vec}(\mathbf{A}_P)')'$ encodes the vectorized transition matrices.

Under this formulation, the VAR model can be equivalently written as:

$$\mathbf{y}_t = \mathcal{A}_{(1)}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{3.1}$$

where $\mathbf{x}_t = (\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-P}')'$ stacks the lagged responses. To further reduce dimensionality, we assume that $\mathcal{A}$ admits a Tucker decomposition with multilinear ranks (r_1, r_2, r_3):

$$\mathcal{A} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 = [[\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]],$$

where $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, and $\mathbf{U}_1 \in \mathbb{R}^{N \times r_1}$, $\mathbf{U}_2 \in \mathbb{R}^{N \times r_2}$, and $\mathbf{U}_3 \in \mathbb{R}^{P \times r_3}$ are the factor matrices. Substituting this into (3.1) yields:

$$\mathbf{y}_t = (\mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)} \mathbf{x}_t + \boldsymbol{\varepsilon}_t. \tag{3.2}$$

This formulation defines the multilinear low-rank VAR model, which captures complex spatio-temporal dependencies with far fewer parameters than conventional VAR or reduced-rank models (Wang, Zheng, Lian, and Li, 2022 (44)). Leveraging the property that $(\mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)} = \mathbf{U}_1 \mathcal{G}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)'$, where $\otimes$ denotes the Kronecker product, we obtain an equivalent, matrix-based representation:

$$\mathbf{y}_t = \mathbf{U}_1 \mathcal{G}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)' \mathbf{x}_t + \boldsymbol{\varepsilon}_t = \mathbf{U}_1 \mathcal{G}_{(1)} \mathrm{vec}(\mathbf{U}_2' \mathbf{X}_t \mathbf{U}_3) + \boldsymbol{\varepsilon}_t, \tag{3.3}$$

where $\mathbf{X}_t = (\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-P}) \in \mathbb{R}^{N \times P}$ is the lagged observation matrix.

To ensure a well-defined decomposition, we adopt the higher-order singular value decomposition (HOSVD) (De Lathauwer et al., 2000 (71)), which imposes orthonormality on the factor matrices and mutual orthogonality on the core tensor slices. Specifically, $\mathbf{U}_j$ contains the top r_j left singular vectors of $\mathcal{A}_{(j)}$, and the core tensor is recovered via:

$$\mathcal{G} = \mathcal{A} \times_1 \mathbf{U}_1' \times_2 \mathbf{U}_2' \times_3 \mathbf{U}_3'.$$

This formulation ensures that the core tensor $\mathcal{G}$ possesses the all-orthogonal property, which means that for $j = 1, 2, 3$, the rows of $\mathcal{G}_{(j)}$ are mutually orthogonal.

Remark 3. By leveraging the higher-order singular value decomposition (HOSVD), the multilinear low-rank VAR model in (2.7) achieves substantial dimensionality reduction. Specifically, the total number of parameters required by the model is $r_1 r_2 r_3 + (N - r_1) r_1 +$

$(N-r_2)r_2+(P-r_3)r_3$, which grows linearly with both N and P; see Zhang (2019) (72). This is in stark contrast to the standard VAR model in (1.1), which entails estimating N^2P parameters. Even when compared to the reduced-rank VAR model in (1.2), which requires $(NP+N-r_1)r_1$ parameters where $r_1 = \text{rank}(\mathcal{A}_{(1)})$, the multilinear formulation offers a significantly more parsimonious parameterization. These results highlight the efficiency of the multilinear low-rank structure in capturing the essential dynamic dependencies while substantially mitigating the curse of dimensionality inherent in high-dimensional VAR modeling.

Moreover, $\mathbf{U}_1$ is an orthonormal matrix, multiplying both sides of (3.3) by $\mathbf{U}_1'$ yields a latent factor representation:

$$\mathbf{U}_1'\mathbf{y}_t = \mathcal{G}_{(1)}(\mathbf{U}_3 \otimes \mathbf{U}_2)'\mathbf{x}_t + \mathbf{U}_1'\boldsymbol{\varepsilon}_t = \mathcal{G}_{(1)}\text{vec}(\mathbf{U}_2'\mathbf{X}_t\mathbf{U}_3) + \mathbf{U}_1'\boldsymbol{\varepsilon}_t. \tag{3.4}$$

Model (2.9) reveals an interpretable dynamic factor structure. The transformed response $\mathbf{U}_1'\mathbf{y}_t$ corresponds to r_1 latent response factors, while the bilinear form $\mathbf{U}_2'\mathbf{X}_t\mathbf{U}_3$ captures interactions among spatial and temporal predictor factors.

Specifically, the rows of $\mathbf{U}_1$ serve as response loadings, mapping the original N-dimensional outcomes to r_1 latent components. If an entry of $\mathbf{U}_1$ is zero, the corresponding response variable does not contribute to the associated factor. The matrix $\mathbf{U}_2$ similarly defines loadings for spatial predictors, and $\mathbf{U}_3$ captures temporal dynamics across lags. For simplicity, refer to r_1, r_2, and r_3 as the response rank, predictor rank, and temporal rank, respectively. Together, this structure provides both interpretability and parsimony, paralleling developments in matrix variate regression and dynamic factor models (Zhao and Leng, 2014 (73); Ding and Cook, 2018 (74)).

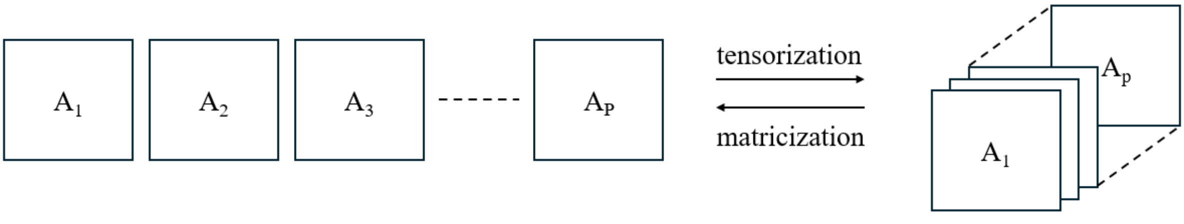


Figure 3.1: Tensor representation of P transition matrices in a VAR model.

3.3 Spatial Quantile Regression Estimation via Tensor Decomposition

3.3.1 Multilinear Low-Rank Spatial Quantile Regression Estimation

We now introduce the Multilinear Low-Rank Spatial Quantile Regression (MLRSQR) estimator for the VAR model reformulated in (3.2). Under the assumption that the transition tensor $\mathcal{A}$ admits a multilinear low-rank Tucker structure with known ranks (r_1, r_2, r_3), we define the MLRSQR estimator as the minimizer of a quantile loss function over the core tensor, the factor matrices, and the conditional quantile shift:

$$\widehat{\mathcal{A}}_{\mathrm{MLRSQR}} \equiv \{[[\widehat{\mathcal{G}}; \widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3]], \widehat{\mathbf{q}}_u\} = \underset{\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{q}_u}{\arg\min} L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{q}_u), \tag{3.5}$$

where the objective function is defined as

$$L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{q}_u) = \sum_{t=1}^{T} Q_u\big(\mathbf{y}_t - (\mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)}\mathbf{x}_t - \mathbf{q}_u\big). \tag{3.6}$$

Here, $\mathbf{q}_u$ represents the $\mathbf{u}$-th quantile of the error term $\boldsymbol{\varepsilon}_t$, which is estimated as

$$\mathbf{q}_u = \arg \min_{\mathbf{q}_u \in \mathbb{R}^k} E\big[Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}) - Q_u(\boldsymbol{\varepsilon}_t)\big]. \tag{3.7}$$

While the optimization problem in (3.5) is unconstrained, it inherits the intrinsic non-uniqueness of Tucker decompositions due to rotational and scaling indeterminacy among the factor matrices. This non-identifiability does not affect the consistency of the estimated transition tensor $\widehat{\mathcal{A}}_{\mathrm{MLRSQR}}$ but may influence the interpretability of individual components.

To establish the asymptotic properties of the MLRSQR estimator under the assumption that both N and P are fixed and the true multilinear ranks (r_1, r_2, r_3) are known, we define the overparameterized vector of parameters as $\boldsymbol{\phi} = (\mathrm{vec}(\mathcal{G}_{(1)})', \mathrm{vec}(\mathbf{U}_1)', \mathrm{vec}(\mathbf{U}_2)', \mathrm{vec}(\mathbf{U}_3)')'$, and denote its estimator by $\widehat{\boldsymbol{\phi}}_{\mathrm{MLRSQR}}$. Let the mapping $\mathbf{h}(\boldsymbol{\phi}) = \mathrm{vec}(\mathcal{A}_{(1)}) = \mathrm{vec}(\mathbf{U}_1\mathcal{G}_{(1)}(\mathbf{U}_3 \otimes \mathbf{U}_2)')$ represent the vectorized mode-1 matricization of the Tucker-form transition tensor.

We also define the block autocovariance matrix $\mathbf{\Gamma}^*$ as:

$$\mathbf{\Gamma}^* = \begin{bmatrix} \mathbf{\Gamma}_0 & \mathbf{\Gamma}_1 & \cdots & \mathbf{\Gamma}_{P-1} \\ \mathbf{\Gamma}_1' & \mathbf{\Gamma}_0 & \cdots & \mathbf{\Gamma}_{P-2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Gamma}_{P-1}' & \mathbf{\Gamma}_{P-2}' & \cdots & \mathbf{\Gamma}_0 \end{bmatrix}, \quad \text{where } \mathbf{\Gamma}_j = \text{Cov}(\mathbf{y}_{t+j}, \mathbf{y}_t) \quad \text{for} \quad j \geq 0.$$

Let $\mathbf{J} = [\mathbf{D}_1(\mathbf{u})\mathbf{D}_2^{-1}(\mathbf{u})\mathbf{D}_1(\mathbf{u})] \otimes \mathbf{\Gamma}^*$, and denote $\mathbf{H}$ as the Jacobian matrix of $\mathbf{h}(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$.

Theorem 5. Suppose the time series $\{\mathbf{y}_t\}$ follows model (2.7), and Assumption 1 and condition C_4 in Appendix B hold. Assume further that (r_1, r_2, r_3) are known and both N and P are fixed. Then,

$$\sqrt{T}\left\{\text{vec}((\widehat{\mathcal{A}}_{\text{MLRSQR}})_{(1)}) - \text{vec}(\mathcal{A}_{(1)})\right\} \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{MLRSQR}}), \tag{3.8}$$

where $\mathbf{\Sigma}_{\text{MLRSQR}} = \mathbf{H}(\mathbf{H}'\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}'$, and $\dagger$ denotes the Moore–Penrose inverse.

The proof of Theorem 5 builds on the asymptotic analysis of overparameterized models via techniques from Shapiro (1986) (75). Crucially, it does not rely on the uniqueness of the Tucker components $\mathcal{G}$ and $\mathbf{U}_j$, thereby ensuring robustness of the result without imposing identification constraints.

However, if the objective is to consistently recover the true HOSVD components $\mathcal{G}$ and $\mathbf{U}_j$ of the transition tensor $\mathcal{A}$, it becomes necessary to ensure the uniqueness of the Tucker decomposition. This requirement can be satisfied under the following assumption:

Assumption 2. For $j = 1, 2, 3$, (i) the singular values of $\mathcal{A}_{(j)}$ are distinct; (ii) the first element in every column of $\mathbf{U}_j$ is positive.

Remark 4. Condition (i) in Assumption 2 ensures uniqueness up to permutation and scaling of the factor matrices, while (ii) provides a standard sign constraint to fix orientation, as commonly adopted in low-rank matrix decomposition (Li et al., 2016 (76)).

Using constraints, one can reconstruct the HOSVD-based estimators as follows: extract the top r_j left singular vectors of the mode-j matricization of $\widehat{\mathcal{A}}_{\text{MLRSQR}}$ to form $\widehat{\mathbf{U}}_j$, ensuring that the first element of each column is positive; then compute the core tensor by projecting $\widehat{\mathcal{G}} = \widehat{\mathcal{A}}_{\text{MLRSQR}} \times_1 \widehat{\mathbf{U}}_1' \times_2 \widehat{\mathbf{U}}_2' \times_3 \widehat{\mathbf{U}}_3'$. As a result, both estimators $\widehat{\mathcal{G}}$ and $\widehat{\mathbf{U}}_j$ are consistent and asymptotically normal, thus ensuring statistically efficient estimation of the underlying multilinear low-rank structure.

Corollary 1. Suppose the conditions in Theorem 5 and Assumption 2 hold, Then, the estimators of the core tensor and factor matrices satisfy the asymptotic normality properties: $\sqrt{T}(\text{vec}(\widehat{\mathcal{G}}) - \text{vec}(\mathcal{G})) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{G}})$, $\sqrt{T}(\text{vec}(\widehat{\mathbf{U}}_1) - \text{vec}(\mathbf{U}_1)) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{U}_1})$, $\sqrt{T}(\text{vec}(\widehat{\mathbf{U}}_2) - \text{vec}(\mathbf{U}_2)) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{U}_2})$, and $\sqrt{T}(\text{vec}(\widehat{\mathbf{U}}_3) - \text{vec}(\mathbf{U}_3)) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{U}_3})$.

The result in Corollary 2 demonstrates that the estimator $\widehat{\mathcal{A}}_{\text{MLRSQR}}$ is asymptotically more efficient than the spatial quantile regression (SQR) estimator $\widehat{\mathcal{A}}_{\text{SQR}}$, which is defined as

$$(\widehat{\mathbf{A}}_{\text{SQR}}, \widehat{\mathbf{q}}_u) = \arg\min_{\mathbf{A}, \mathbf{q}_u} Q_u \left(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t - \mathbf{q}_u\right) \tag{3.9}$$

where $\mathbf{A} = (\mathbf{A}_1, \dots, \mathbf{A}_P)$ denotes the full transition matrix such that $\mathbf{A} \in \mathbb{R}^{N \times NP}$ in the original model (1.4). We denote by $\widehat{\mathcal{A}}_{\text{SQR}}$ the transition tensor constructed from $\widehat{\mathbf{A}}_{\text{SQR}}$.

Corollary 2. Under the condition of Theorem 1, $\sqrt{T}\{\text{vec}((\widehat{\mathcal{A}}_{\text{SQR}})_{(1)}) - \text{vec}(\mathcal{A}_{(1)})\} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{SQR}})$, such that $\boldsymbol{\Sigma}_{\text{MLRSQR}} \le \boldsymbol{\Sigma}_{\text{SQR}}$.

The results of Corollary 1 and Corollary 2 highlight both the statistical efficiency and structural advantages of the proposed MLRSQR estimator over classical methods that neglect multilinear structure. All technical proofs are provided in Appendix B.

3.3.2 Alternating SQR Algorithm

Let $\mathcal{F}_t = \sigma(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_{t-1}, \dots)$ denote the σ-field generated by the history of the error process up to time t, and recall that $\mathbf{X}_t = (\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-P})$ denotes the lagged predictor matrix. Although the objective function in (3.5) is nonlinear in the parameters $\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$, and $\mathbf{q}_u$, a closer inspection of the model representation in (3.2) reveals a valuable structural property.

Specifically, the conditional quantile function $\mathbf{Q}(\mathbf{y}_t|\mathcal{F}_{t-1})$ can be expressed as

$$\begin{aligned} \mathbf{Q}(\mathbf{y}_t|\mathcal{F}_{t-1}) &= \left((\mathbf{x}_t'(\mathbf{U}_3 \otimes \mathbf{U}_2)\mathcal{G}_{(1)}') \otimes \mathbf{I}_N\right) \operatorname{vec}(\mathbf{U}_1) + \mathbf{q}_u \\ &= \mathbf{U}_1\mathcal{G}_{(1)} \left((\mathbf{U}_3'\mathbf{X}_t') \otimes \mathbf{I}_{r_2}\right) \operatorname{vec}(\mathbf{U}_2') + \mathbf{q}_u \\ &= \mathbf{U}_1\mathcal{G}_{(1)} \left(\mathbf{I}_{r_3} \otimes (\mathbf{U}_2'\mathbf{X}_t)\right) \operatorname{vec}(\mathbf{U}_3) + \mathbf{q}_u \\ &= \left(((\mathbf{U}_3 \otimes \mathbf{U}_2)'\mathbf{x}_t)' \otimes \mathbf{U}_1\right) \operatorname{vec}(\mathcal{G}_{(1)}) + \mathbf{q}_u, \end{aligned} \tag{3.10}$$

which demonstrates that the loss function is linear in each of the parameters $\mathcal{G}$, $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$ when the others are held fixed. This property facilitates the use of an alternating optimization procedure.

Given the multilinear ranks (r_1, r_2, r_3), we estimate $\widehat{\mathcal{A}}_{\text{MLRSQR}}$ using the iterative algorithm described in Algorithm 2, which follows an alternating spatial quantile regression (SQR) scheme. At each iteration, one parameter block is updated while others are fixed, and the subproblem reduces to a convex quantile regression problem. Numerical optimization techniques, such as gradient descent or Newton-type methods, can be employed to efficiently solve each subproblem.

Algorithm 2 Alternating SQR algorithm for $\widehat{\mathcal{A}}_{\text{MLRSQR}}$

```
Initialize: $\mathcal{A}^{(0)}$
HOSVD: $\mathcal{A}^{(0)} \approx \mathcal{G}^{(0)} \times_1 \mathbf{U}_1^{(0)} \times_2 \mathbf{U}_2^{(0)} \times_3 \mathbf{U}_3^{(0)}$ with multilinear ranks $(r_1, r_2, r_3)$
repeat $k = 0, 1, 2, \ldots$
    $\mathbf{U}_1^{(k+1)} \leftarrow \arg\min_{\mathbf{U}_1} \sum_{t=1}^T Q_u \left(\mathbf{y}_t - \left((\mathbf{x}_t'(\mathbf{U}_3^{(k)} \otimes \mathbf{U}_2^{(k)})\mathcal{G}_{(1)}^{(k)'}) \otimes \mathbf{I}_N\right) \operatorname{vec}(\mathbf{U}_1) - \mathbf{q}_u^k\right)$
    $\mathbf{U}_2^{(k+1)} \leftarrow \arg\min_{\mathbf{U}_2} \sum_{t=1}^T Q_u \left(\mathbf{y}_t - \mathbf{U}_1^{(k+1)}\mathcal{G}_{(1)}^{(k)} \left((\mathbf{X}_t\mathbf{U}_3^{(k)})' \otimes \mathbf{I}_{r_2}\right) \operatorname{vec}(\mathbf{U}_2') - \mathbf{q}_u^k\right)$
    $\mathbf{U}_3^{(k+1)} \leftarrow \arg\min_{\mathbf{U}_3} \sum_{t=1}^T Q_u \left(\mathbf{y}_t - \mathbf{U}_1^{(k+1)}\mathcal{G}_{(1)}^{(k)}(\mathbf{I}_{r_3} \otimes (\mathbf{U}_2^{(k+1)'}\mathbf{X}_t))\operatorname{vec}(\mathbf{U}_3) - \mathbf{q}_u^k\right)$
    $\mathcal{G}^{(k+1)} \leftarrow \arg\min_{\mathcal{G}} \sum_{t=1}^T Q_u \left(\mathbf{y}_t - (((\mathbf{U}_3^{(k+1)} \otimes \mathbf{U}_2^{(k+1)})'\mathbf{x}_t)' \otimes \mathbf{U}_1^{(k+1)})\operatorname{vec}(\mathcal{G}_{(1)}) - \mathbf{q}_u^k\right)$
    $\mathbf{q}_u^{(k+1)} \leftarrow \arg\min_{\mathbf{q}_u} \sum_{t=1}^T Q_u \left(\mathbf{y}_t - (\mathcal{G}^{(k+1)} \times_1 \mathbf{U}_1^{(k+1)} \times_2 \mathbf{U}_2^{(k+1)} \times_3 \mathbf{U}_3^{(k+1)})_{(1)}\mathbf{x}_t - \mathbf{q}_u\right)$
        $$\mathcal{A}^{(k+1)} \leftarrow \mathcal{G}^{(k+1)} \times_1 \mathbf{U}_1^{(k+1)} \times_2 \mathbf{U}_2^{(k+1)} \times_3 \mathbf{U}_3^{(k+1)}$$
until convergence
Finalize: $\widehat{\mathbf{U}}_i \leftarrow$ top $r_i$ left left singular vectors of $\widehat{\mathcal{A}}_{(i)}$ with positive first elements, $1 \leq i \leq 3$
        $$\widehat{\mathcal{G}} \leftarrow [[\widehat{\mathcal{A}}; \widehat{\mathbf{U}}_1', \widehat{\mathbf{U}}_2', \widehat{\mathbf{U}}_3']]$$
```

In practice, the initialization of the algorithm plays a critical role in achieving good performance. When the sample size T is large, the ordinary least squares (OLS) estimator

$$\widehat{\mathbf{A}}_{\text{OLS}} = \underset{\mathbf{A}\,\in\,\mathbb{R}^{N\times NP}}{\arg\min} \sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{A}\mathbf{x}_t\|_2^2 \tag{3.11}$$

can be used to construct the initial tensor estimate $\mathcal{A}^{(0)}$. When smaple size T is small, alternative initialization strategies—such as the reduced-rank regression (RRR) estimator,

$$\widehat{\mathbf{A}}_{\text{RRR}} = \underset{\substack{\mathbf{A}\in\mathbb{R}^{N\times NP}\\ \text{rank}(\mathbf{A})\leq r_1}}{\arg\min} \sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{A}\mathbf{x}_t\|_2^2 \tag{3.12}$$

or the nuclear norm penalized estimator introduced in Section 3.5—may yield improved results. Note that, $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_P)$ in models (3.11) and (3.12) correspond to models (1.4) and (1.6), respectively. Denote by $\widehat{\mathcal{A}}_{\text{OLS}}$ and $\widehat{\mathcal{A}}_{\text{RRR}}$ the transition tensors formed by $\widehat{\mathbf{A}}_{\text{OLS}}$ and $\widehat{\mathbf{A}}_{\text{RRR}}$, respectively. On the other hand, selecting the multilinear ranks consistently is also crucial, and the details of this selection processare discussed in Section 3.5.

To further guard against local optima, we recommend a randomized initialization strategy: $\mathcal{A}^{(0)} = \widehat{\mathcal{A}}_{\text{pre}} + T^{-1/2}\mathcal{T}$, where $\widehat{\mathcal{A}}_{\text{pre}}$ is a preliminary estimate such as $\widehat{\mathcal{A}}_{\text{OLS}}$ or $\widehat{\mathcal{A}}_{\text{RRR}}$, and $\mathcal{T} \in \mathbb{R}^{N\times N\times P}$ is a tensor with i.i.d. $N(0,1)$ entries. Among multiple initializations, the one that yields the lowest objective value is selected.

Importantly, the alternating estimation algorithm does not impose explicit constraints on the Tucker components $\mathcal{G}$ and $\mathbf{U}_j$, and thus the orthogonality conditions are not enforced during optimization. Nevertheless, as discussed in Proposition 1 of Wang, Zheng, Lian, and Li (2022) (44), the convergence of the algorithm is guaranteed under mild regularity conditions, even without identification constraints or uniqueness assumptions on the Tucker decomposition.

Remark 5. The final estimates $\widehat{\mathcal{G}}$ and $\widehat{\mathbf{U}}_j$ obtained from Algorithm 2 correspond to an unconstrained minimizer of the MLRSQR objective. As established in Corollary 1, these estimators are consistent and asymptotically normal. Furthermore, the use of alternating optimization without enforcing identifiability is widely adopted in the literature on tensor decomposition (e.g. Zhou, Li and Zhu, 2013 (66); Li et al., 2018 (69)) and has proven effective in practice.

3.4 Regularized Spatial Quantile Regression Estimation via Tensor Decomposition

3.4.1 Sparse Higher-Order Reduced-Rank Spatial Quantile Regression Estimation

Section 3.2.2 highlights the model's capacity to capture multidimensional dynamic dependencies through the factor loading matrices $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$, which correspond to the response, predictor, and temporal dimensions, respectively. In high-dimensional settings, where either the number of variables (N) or the number of lags (P) is large, it is often observed that many entries in the estimated loading matrices are close to zero. This empirical pattern suggests that a substantial proportion of variables or time lags contribute marginally to the factor structure. For example, when the (i, j)th entry of $\mathbf{U}_1$ is near zero, the ith response variable y_{it} plays a negligible role in forming the jth response factor (for $1 \leq i \leq N$ and $1 \leq j \leq r_1$). Similar interpretations apply to $\mathbf{U}_2$ and $\mathbf{U}_3$.

To explicitly exploit this empirical sparsity, we introduce a structured regularization framework that encourages sparse solutions for the factor loading matrices. This strategy not only reduces the effective number of parameters but also enables data-driven variable selection at the level of individual factor contributions, thereby enhancing interpretability and estimation efficiency. The benefits of incorporating sparsity into reduced-rank frameworks are well-documented in the literature; see, for instance, Chen, Chan, and Stenseth (2012) (64) and Uematsu et al. (2019) (65).

Motivated by these insights, we propose the Sparse Higher-Order Reduced-Rank Spatial Quantile Regression (SHORRSQR) estimator, defined as the solution to the following penalized optimization problem:

$$\begin{aligned}\widehat{\mathcal{A}}_{\text{SHORRSQR}} \equiv \{[[\widehat{\mathcal{G}}; \widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3]], \widehat{\mathbf{q}}_u\} = \underset{\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{q}_u}{\arg\min} \{&L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{q}_u) \\ &+ \lambda \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1\}\end{aligned} \tag{3.13}$$

subject to

$$\mathcal{G} \in \text{AO}(r_1, r_2, r_3) \quad \text{and} \quad \mathbf{U}_i'\mathbf{U}_i = \mathbf{I}_{r_i}, \quad i = 1, 2, 3, \tag{3.14}$$

where the loss function $L(\cdot)$ is defined in (3.6), and $\text{AO}(r_1, r_2, r_3)$ denotes the set of core tensors whose mode-i matricizations are row-orthogonal for each i, i.e., $\text{AO}(r_1, r_2, r_3) = \{\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \mathcal{G}_{(i)} \text{ is row-orthogonal}, i = 1, 2, 3\}$. These orthogonality conditions are essential for ensuring identifiability of the sparsity structure in $\mathbf{U}_i$, distinguishing this formulation from the unconstrained MLRSQR estimator introduced earlier in (3.5). Our analysis of $\widehat{\mathcal{A}}_{\text{SHORRSQR}}$ and its asymptotic properties assumes fixed dimensions for both N and P, along with known true multilinear ranks (r_1, r_2, r_3). Section 3.5 will further elaborate on a consistent methodology for rank selection.

A distinguishing feature of SHORRSQR is its use of a unified sparsity penalty $\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1 = \|\mathbf{U}_1\|_1 \|\mathbf{U}_2\|_1 \|\mathbf{U}_3\|_1$ across all three factor loading matrices, as opposed to applying separate ℓ_1 penalties to each matrix (e.g., $\sum_{i=1}^{3} \lambda_i \|\mathbf{U}_i\|_1$). This unified formulation avoids the complexity of tuning multiple penalty parameters and provides a parsimonious structure that integrates all three factor dimensions. The estimator can be efficiently computed via the alternating algorithm described in Section 3.4.2.

It is worth noting that the joint penalization strategy employed here is conceptually aligned with the joint Lasso penalties proposed by Zhao and Leng (2014) (73) and the coordinated penalties for left and right singular vectors in sparse SVD investigations by Chen, Chan, and Stenseth (2012) (64). Moreover, in cases where the number of lags P is relatively small, it may be computationally advantageous to apply the sparsity constraint only to $\mathbf{U}_1$ and $\mathbf{U}_2$. In such settings, the penalty term $\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1$ can be replaced by the reduced form $\|\mathbf{U}_2 \otimes \mathbf{U}_1\|_1$ without loss of modeling fidelity.

Lastly, it is important to distinguish SHORRSQR from conventional row-sparse reduced-rank regression models. Unlike approaches that enforce row-level sparsity—potentially excluding entire variables or responses from the model—our method encourages elementwise sparsity. This preserves the flexibility of the VAR model and ensures that all time series remain forecastable, even when their influence on certain factors is weak. Such a design choice promotes generalizability while maintaining interpretability in high-dimensional environments.

To establish the asymptotic properties of the SHORRSQR estimator, we need the sparse assumption:

Assumption 3. (*Sparsity*) Each column of the factor matrices $\mathbf{U}_i$ has at most s_i nonzero entries, for $i = 1, 2, 3$.

Assumption 3 imposes a structured sparsity constraint on each factor matrix, which ensures that the model remains interpretable and identifiable in high dimensions.

Theorem 6. Suppose that the time series $\{\mathbf{y}_t\}$ is generated by model (3.2) with conditions C_4 in Appendix B, both N and P are fixed, and (r_1, r_2, r_3) are known. Then, under Assumption 1, and Assumption 3,

$$\sqrt{T}\{\text{vec}((\widehat{\mathcal{A}}_{\text{SHORRSQR}})_{(1)}) - \text{vec}(\mathcal{A}_{(1)})\} \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{SHORRSQR}}). \tag{3.15}$$

The proof of Theorem 6 is provided in Appendix B. This result demonstrates that the SHORRSQR estimator attains $\sqrt{T}$-consistency and asymptotic normality under appropriate sparsity and identifiability conditions, thereby establishing its statistical efficiency in high-dimensional, low-rank, and sparse VAR settings.

3.4.2 ADMM Algorithm

Developing an efficient estimation algorithm for the SHORRSQR model introduces two significant computational challenges. First, the core tensor $\mathcal{G}$ is required to satisfy the all-orthogonal constraint in (3.14), which is not straightforward to handle. Second, the factor loading matrices $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$ are simultaneously subject to both ℓ_1-regularization and orthogonality constraints. The former induces nonsmoothness in the objective function, while the latter introduces non-convexity, thereby complicating the optimization process.

To overcome these difficulties, we adopt an alternating direction method of multipliers (ADMM) framework (Boyd et al., 2011 (77)) that updates $\mathcal{G}$ and $\mathbf{U}_i$ alternately; see Algorithm 3. To facilitate the update of $\mathcal{G}$, we reformulate its all-orthogonal constraint as three independent constraints on its matricizations. Specifically, for $i = 1, 2, 3$, each matricization $\mathcal{G}_{(i)}$ is expressed as $\mathbf{D}_i\mathbf{V}_i'$, where $\mathbf{D}_i$ is a diagonal matrix and $\mathbf{V}_i$ is an orthonormal matrix such that $\mathbf{V}_i'\mathbf{V}_i = \mathbf{I}_{r_i}$. This decomposition allows us to formulate the augmented Lagrangian for the objective function in (3.13) as:

$$\begin{aligned}&\mathcal{L}_\varrho(\mathcal{G}, \{\mathbf{U}_i\}, \{\mathbf{D}_i\}, \{\mathbf{V}_i\}, \mathbf{q}_u; \{\mathcal{C}_i\})\\&= L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{q}_u) + \lambda\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1\\&+ 2\sum_{i=1}^{3} \varrho_i \langle (\mathcal{C}_i)_{(i)}, \mathcal{G}_{(i)} - \mathbf{D}_i\mathbf{V}_i'\rangle + \sum_{i=1}^{3} \varrho_i\|\mathcal{G}_{(i)} - \mathbf{D}_i\mathbf{V}_i'\|_F^2,\end{aligned} \tag{3.16}$$

where $\mathcal{C}_i$ denotes the dual variable for $i = 1, 2, 3$, and ϱ_i is the corresponding penalty parameter. With this reparameterization, the constraint on $\mathcal{G}$ is transferred to the update of $\mathbf{V}_i$ in line of Algorithm 3, enabling an unconstrained update of $\mathcal{G}$ in line 7 of Algorithm 3.

The update of each factor loading matrix $\mathbf{U}_i$ poses additional challenges due to the combined presence of the quantile loss, the ℓ_1 penalty, and the orthogonality constraint. Each $\mathbf{U}_i$-update becomes an instance of the following optimization problem:

$$\min_{\mathbf{B}} \left\{ n^{-1}Q_u(\mathbf{y} - \mathbf{X}\text{vec}(\mathbf{B}) - \mathbf{q}_u) + \lambda\|\mathbf{B}\|_1 \right\}, \quad \text{s.t. } \mathbf{B}'\mathbf{B} = \mathbf{I}. \tag{3.17}$$

where the nonsmoothness of the ℓ_1 term and the non-convexity of the orthogonality constraint complicate direct optimization. To decouple these components, we apply a nested ADMM subroutine that introduces an dual variable $\mathbf{W}$ to enforce $\mathbf{B} = \mathbf{W}$, reformulating problem (3.17) into the following equivalent form:

$$\min_{\mathbf{B},\mathbf{W}} \left\{ n^{-1}Q_u(\mathbf{y} - \mathbf{X}\text{vec}(\mathbf{B}) - \mathbf{q}_u) + \lambda\|\mathbf{W}\|_1 \right\}, \quad \text{s.t. } \mathbf{B}'\mathbf{B} = \mathbf{I} \quad \text{and} \quad \mathbf{B} = \mathbf{W}. \tag{3.18}$$

Then the corresponding augmented Lagrangian formulation is

$$\min_{\mathbf{B},\mathbf{W}} \left\{ n^{-1}Q_u(\mathbf{y} - \mathbf{X}\text{vec}(\mathbf{B}) - \mathbf{q}_u) + \lambda\|\mathbf{W}\|_1 + 2\kappa\langle \mathbf{M}, \mathbf{B} - \mathbf{W}\rangle + \kappa\|\mathbf{B} - \mathbf{W}\|_F^2 \right\}. \tag{3.19}$$

where κ is a regularization parameter. The ADMM subroutine for (3.19) is presented in Algorithm 4. This yields solutions to the $\mathbf{U}_i$-update subproblems in Algorithm 3.

Note that the $\mathbf{B}$-update step in Algorithm 4 and the $\mathbf{V}_i$-update step in line 11 of Algorithm 3 are formulated as SQR and least squares problems with an orthogonality constraint. These steps can be efficiently solved using the splitting orthogonality con-

Algorithm 3 ADMM algorithm for SHORRSQR estimator

```
Initialize: $\mathcal{A}^{(0)}$
HOSVD: $\mathcal{A}^{(0)} \approx \mathcal{G}^{(0)} \times_1 \mathbf{U}_1^{(0)} \times_2 \mathbf{U}_2^{(0)} \times_3 \mathbf{U}_3^{(0)}$ with multilinear ranks $(r_1, r_2, r_3)$.
repeat
    $\mathbf{U}_1^{(k+1)} \leftarrow \underset{\mathbf{U}_1'\mathbf{U}_1 = \mathbf{I}_{r_1}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \mathbf{U}_1, \mathbf{U}_2^{(k)}, \mathbf{U}_3^{(k)}) + \lambda \|\mathbf{U}_1\|_1 \|\mathbf{U}_2^{(k)}\|_1 \|\mathbf{U}_3^{(k)}\|_1 \right\}$
    $\mathbf{U}_2^{(k+1)} \leftarrow \underset{\mathbf{U}_2'\mathbf{U}_2 = \mathbf{I}_{r_2}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \mathbf{U}_1^{(k+1)}, \mathbf{U}_2, \mathbf{U}_3^{(k)}) + \lambda \|\mathbf{U}_1^{(k+1)}\|_1 \|U_2\|_1 \|\mathbf{U}_3^{(k)}\|_1 \right\}$
    $\mathbf{U}_3^{(k+1)} \leftarrow \underset{\mathbf{U}_3'\mathbf{U}_3 = \mathbf{I}_{r_3}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \mathbf{U}_1^{(k+1)}, \mathbf{U}_2^{(k+1)}, \mathbf{U}_3) + \lambda \|\mathbf{U}_1^{(k+1)}\|_1 \|\mathbf{U}_2^{(k+1)}\|_1 \|\mathbf{U}_3\|_1 \right\}$
    $\mathcal{G}^{(k+1)} \leftarrow \arg\min \left\{ L(\mathcal{G}, \mathbf{U}_1^{(k+1)}, \mathbf{U}_2^{(k+1)}, \mathbf{U}_3^{(k+1)}) + \sum_{i=1}^{3} \varrho_i \|\mathcal{G}_{(i)} - \mathbf{D}_i^{(k)} \mathbf{V}_i^{(k)\prime} + (\mathcal{C}_i^{(k)})_{(i)}\|_F^2 \right\}$
    $\mathbf{q}_u^{(k+1)} \leftarrow \arg\min \left\{ L(\mathcal{G}^{(k+1)}, \mathbf{U}_1^{(k+1)}, \mathbf{U}_2^{(k+1)}, \mathbf{U}_3^{(k+1)}) + \lambda \|\mathbf{U}_1^{(k+1)}\|_1 \|\mathbf{U}_2^{(k+1)}\|_1 \|\mathbf{U}_3^{(k+1)}\|_1 \right\}$
    for $i \in \{1, 2, 3\}$ do
        $\mathbf{D}_i^{(k+1)} \leftarrow \underset{\mathbf{D}_i = \mathrm{diag}(\mathbf{d}_i)}{\arg\min} \|\mathcal{G}_{(i)}^{(k+1)} - \mathbf{D}_i \mathbf{V}_i^{(k)\prime} + (\mathcal{C}_i^{(k)})_{(i)}\|_F^2$
        $\mathbf{V}_i^{(k+1)} \leftarrow \underset{\mathbf{V}_i'\mathbf{V}_i = \mathbf{I}_{r_i}}{\arg\min} \|\mathcal{G}_{(i)}^{(k+1)} - \mathbf{D}_i^{(k+1)} \mathbf{V}_i' + (\mathcal{C}_i^{(k)})_{(i)}\|_F^2$
        $(\mathcal{C}_i^{(k+1)})_{(i)} \leftarrow (\mathcal{C}_i^{(k)})_{(i)} + \mathcal{G}_{(i)}^{(k+1)} - \mathbf{D}_i^{(k+1)} \mathbf{V}_i^{(k+1)\prime}$
    end for
    $\mathcal{A}^{(k+1)} \leftarrow \mathcal{G}^{(k+1)} \times_1 \mathbf{U}_1^{(k+1)} \times_2 \mathbf{U}_2^{(k+1)} \times_3 \mathbf{U}_3^{(k+1)}$
until convergence
```

Algorithm 4 ADMM subroutine for sparse and orthogonal regression

```
Initialize: $\mathbf{B}^{(0)} = \mathbf{W}^{(0)}, \mathbf{M}^{(0)} = 0$
repeat
    $\mathbf{B}^{(k+1)} \leftarrow \underset{\mathbf{B}'\mathbf{B} = \mathbf{I}}{\arg\min} \left\{ n^{-1} Q_u(\mathbf{y} - \mathbf{X}\,\mathrm{vec}(\mathbf{B}) - \mathbf{q}_u) + \kappa \|\mathbf{B} - \mathbf{W}^{(k)} + \mathbf{M}^{(k)}\|_F^2 \right\}$
    $\mathbf{W}^{(k+1)} \leftarrow \underset{\mathbf{W}}{\arg\min} \left\{ \kappa \|\mathbf{B}^{(k+1)} - \mathbf{W} + \mathbf{M}^{(k)}\|_F^2 + \lambda \|\mathbf{W}\|_1 \right\}$
    $\mathbf{M}^{(k+1)} \leftarrow \mathbf{M}^{(k)} + \mathbf{B}^{(k+1)} - \mathbf{W}^{(k+1)}$
until convergence
```

straint (SOC) method (Lai and Osher 2014 (78)). The $\mathbf{W}$-update step in Algorithm 4 involves ℓ_1-regularized minimization, which can be solved explicitly via soft-thresholding. The $\mathcal{G}$- and $\mathbf{D}_i$-update steps in lines 7 and 10 of Algorithm 3 correspond to SQR and simple least squares problems, respectively.

For general nonconvex problems, it is well established that ADMM algorithms may fail to converge, and even when they do, there is no guarantee of convergence to an optimal solution. Conducting a comprehensive convergence analysis for Algorithm 3 is particularly challenging due to the presence of the nested ADMM subroutine, Algorithm 4, and its interaction with the outer loop of Algorithm 3.

Wang, Yin, and Zeng (2019) (72) provided a rigorous convergence analysis for multi-block ADMMs applied to nonconvex nonsmooth optimization problems with linear equality constraints. Their theoretical results would apply to Algorithm 4 if the $\mathbf{B}$-update step in line 3 were performed exactly. Extending this analysis to the inexact $\mathbf{B}$-update step would require a detailed examination of the optimization error introduced by the SOC method. In this article, we do not further explore the development of the convergence theory. Nevertheless, following a similar approach to the analysis by Uematsu et al. (2019) (65), and under certain high-level assumptions on $\mathcal{L}_{\varrho}(\cdot)$, we can still establish a convergence result for Algorithm 3, as stated in Proposition 2 of Wang, Zheng, Lian, and Li (2022) (44). Specifically, the sequence generated by Algorithm 3 converges to a local solution of problem (3.13).

Remark 6. The initial value $\mathcal{A}^{(0)}$ for Algorithm 3 can be chosen as the nuclear norm (NN) estimator $\widehat{\mathcal{A}}_{\text{NN}}$ for low-rank VAR models (Negahban and Wainwright 2011 (36)). This estimator satisfies $\|\widehat{\mathcal{A}}_{\text{NN}} - \mathcal{A}\|_F = O_p(\sqrt{r_1 NP/T})$; see also Section 3.5. As a result, if the SHORRSQR estimator is searched within a neighborhood of radius $O(\sqrt{r_1 NP/T})$ around $\widehat{\mathcal{A}}_{\text{NN}}$, all iterates $\mathcal{A}^{(k)}$ will satisfy $\|\mathcal{A}\|_F \leq \|\mathcal{A}^{(k)} - \widehat{\mathcal{A}}_{\text{NN}}\|_F + \|\widehat{\mathcal{A}}_{\text{NN}} - \mathcal{A}\|_F = O_p(\sqrt{r_1 NP/T})$. A similar convex relaxation-based initialization approach has been used by Uematsu et al. (2019) (65) for nonconvex optimization problems involving jointly imposed sparsity and orthogonality constraints. Since Algorithm 3 does not guarantee convergence to a global solution, a practical alternative is to use randomized initial values. Specifically, one can initialize with $\mathcal{A}^{(0)} = \widehat{\mathcal{A}}_{\text{NN}} + (NP/T)^{1/2}\mathcal{T}$, where the entries of the perturbation $\mathcal{T} \in \mathbb{R}^{N\times N\times P}$ are independently drawn from $N(0, (N^2P)^{-1})$ such that $\|\mathcal{T}\|_F = O_p(1)$. The final solution is then selected as the one that minimizes the objective function.

Finally, while the algorithm is developed under the assumption of known multilinear ranks and fixed λ, practical implementation requires data-driven selection of these parameters. To this end, we recommend a two-stage procedure: first estimate the ranks via the consistent method outlined in Section 3.5, then fix these ranks and tune λ using a model selection criterion such as BIC. Although the degrees of freedom in sparse orthonormal structures are difficult to quantify exactly, the total number of nonzero entries in $\mathcal{G}$ and $\mathbf{U}_i$ offers a reasonable approximation.

3.5 Rank Selection

The theoretical guarantees for both the MLRSQR and SHORRSQR estimators rely critically on the correctness of the specified multilinear ranks. In this section, we adopt a rank selection procedure proposed by Xia, Xu, and Zhu (2015) (79) to consistently estimate the true multilinear ranks (r_1, r_2, r_3) in a data-driven manner.

Suppose that a preliminary estimator $\widehat{\mathcal{A}}$ of the transition tensor $\mathcal{A}$ is available and satisfies a suitable convergence rate. We employ the ridge-type ratio estimator to determine the multilinear ranks from the singular values of the mode-i matricizations $\widehat{\mathcal{A}}_{(i)}$. Specifically, for $1 \le i \le 3$, we define:

$$\hat{r}_i = \underset{1\le j\le p_i-1}{\arg\min} \frac{\sigma_{j+1}(\hat{\mathcal{A}}_{(i)}) + c}{\sigma_j(\hat{\mathcal{A}}_{(i)}) + c}, \tag{3.20}$$

where $p_1 = p_2 = N$, $p_3 = P$, and $c > 0$ is a ridge parameter that must satisfy certain regularity conditions for consistent rank recovery.

To ensure the validity of this method, the ridge parameter c must satisfy two conditions:

(i) $\|\widehat{\mathcal{A}} - \mathcal{A}\|_F / c = o_p(1)$,

(ii) $c \cdot \max_{1\le i\le 3} \varsigma_i = o(1)$,

where ς_i is defined as

$$\varsigma_i = \frac{1}{\sigma_{r_i}(\mathcal{A}_{(i)})} \cdot \max_{1\le j\le r_i} \frac{\sigma_j(\mathcal{A}_{(i)})}{\sigma_{j+1}(\mathcal{A}_{(i)})}. \tag{3.21}$$

Condition (i) specifies that the estimation error is bounded by c, while Condition (ii) requires that c grows significantly slower than the ς_i's. Broadly speaking, Condition (ii) may fail if the smallest nonzero singular value of $\mathcal{A}_{(i)}$ is too small or if there is a significant drop between $\sigma_j(\mathcal{A}_{(i)})$ and $\sigma_{j+1}(\mathcal{A}_{(i)})$ for some $1 \leq j < r_i$ and $1 \leq i \leq 3$. In either scenario, the ridge-type ratio may struggle to select the rank correctly. If all nonzero singular values are bounded above and away from zero, Condition (ii) reduces to $c = o(1)$. Under above conditions and the conditions stated in Theorem 6, the probability of correctly identifying the ranks $(\widehat{r}_1, \widehat{r}_2, \widehat{r}_3)$ as (r_1, r_2, r_3) converges to 1 as $T \to \infty$, i.e., $\mathbb{P}(\widehat{r}_1 = r_1, \widehat{r}_2 = r_2, \widehat{r}_3 = r_3) \to 1$ as $T \to \infty$ (Wang, Zheng, Lian, and Li, 2022 (44)).

For the initial estimator, we use the nuclear norm (NN) estimator for low-rank VAR models defined as

$$\widehat{\mathcal{A}}_{\text{NN}} = \arg\min \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{y}_t - \mathcal{A}_{(1)}\mathbf{x}_t\|_2^2 + \lambda \|\mathcal{A}_{(1)}\|_*.$$

The estimation error rate derived by Negahban and Wainwright (2011) (36) for VAR(1) models can be straightforwardly extended to VAR(P) models, resulting in $\|\widehat{\mathcal{A}}_{\text{NN}} - \mathcal{A}\|_F = O_p(\sqrt{r_1 NP/T})$. Consequently, rank selection consistency can be achieved over a relatively wide range of c. In practice, we recommend setting $c = \sqrt{NP\log(T)/10T}$ (Wang, Zheng, Lian, and Li, 2022 (44)), which has been shown to perform well in the first simulation experiment presented in Section 3.6.1.

3.6 Simulation Experiments

3.6.1 Rank Selection Consistency

We adopt the rank selection procedure introduced in Section 3.5 for all subsequent simulation studies and real data analyses. To assess its empirical performance, we first conduct a dedicated experiment to examine the consistency of this procedure under various scenarios.

The data are simulated according to the multilinear low-rank VAR model specified in model (3.2) with dimensions $(N, P) = (10, 5)$ and true multilinear ranks $(r_1, r_2, r_3) = (3, 3, 3)$. To investigate consistency under different distributional settings of the innovation term $\boldsymbol{\varepsilon}_t$, we consider three types of error distributions: Case (i) multivariate normal distribution, $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_1)$; Case (ii) multivariate t distribution with 3 degrees of freedom, $\boldsymbol{\varepsilon}_t \sim t_3(\mathbf{0}, \boldsymbol{\Sigma}_2)$; and Case (iii) multivariate mixed normal distribution, $\boldsymbol{\varepsilon}_t \sim 0.9 \cdot N(\mathbf{0}, \boldsymbol{\Sigma}_3) + 0.1 \cdot N(\mathbf{0}, \boldsymbol{\Sigma}_4)$, where $\boldsymbol{\Sigma}_4$ has substantially larger variance.

To evaluate the influence of singular value structure on rank selection performance, we consider four cases of the core tensor $\mathcal{G}$, specified as a diagonal tensor with superdiagonal entries $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333})$ given by: (case a) $(2, 2, 2)$, (case b) $(4, 3, 2)$, (case c) $(1, 1, 1)$, and (case d) $(1.5, 1, 1)$. These settings directly determine the nonzero singular values of the matricizations $\mathcal{A}_{(i)}$ for $1 \leq i \leq 3$. The factor matrices $\mathbf{U}_i$ are constructed as the top r_i left singular vectors of independent Gaussian random matrices, and the generated VAR processes are verified to satisfy the stationarity condition described in Assumption 1.

For rank estimation, we apply the ridge-type ratio method with a tuning parameter $c = \sqrt{NP \log(T)/10T}$, as recommended in Wang, Zheng, Lian, and Li (2022) (44). The performance metric is the proportion of simulations in which the estimated ranks exactly match the true ranks, i.e., $\{(\widehat{r}_1, \widehat{r}_2, \widehat{r}_3) = (r_1, r_2, r_3)\}$. This is computed across a range of sample sizes $T \in [50, 600]$, with 1000 replications per setting. The results are visualized in Figure 3.2.

The simulation findings indicate a clear trend: as T increases, the accuracy of rank improves steadily across all error distributions. In particular, the correct selection rate approaches unity around $T = 400$ for the multivariate normal distribution (Case (i)), $T = 600$ for the multivariate t distribution (Case (ii)), and $T = 500$ for the multivariate mixed normal distribution (Case (iii)). These results demonstrate the consistency of the proposed selection procedure.

Additionally, the simulation reveals how the singular value structure of $\mathcal{A}_{(i)}$ influences rank consistency. As discussed in Section 3.5, rank selection becomes more challenging when the smallest nonzero singular value $\sigma_{r_i}(\mathcal{A}_{(i)})$ is small or when the gap between successive singular values is large. Accordingly, for all types of errors, cases a and b demonstrate better performance due to their larger $\sigma_{r_i}(\mathcal{A}(i))$ compared to cases c and d. Furthermore, cases a and c outperform cases b and d, likely because the singular values

$\mathcal{G}_{111}, \mathcal{G}_{222}$, and $\mathcal{G}_{333}$ are equal in the former cases, contributing to more stable performance.

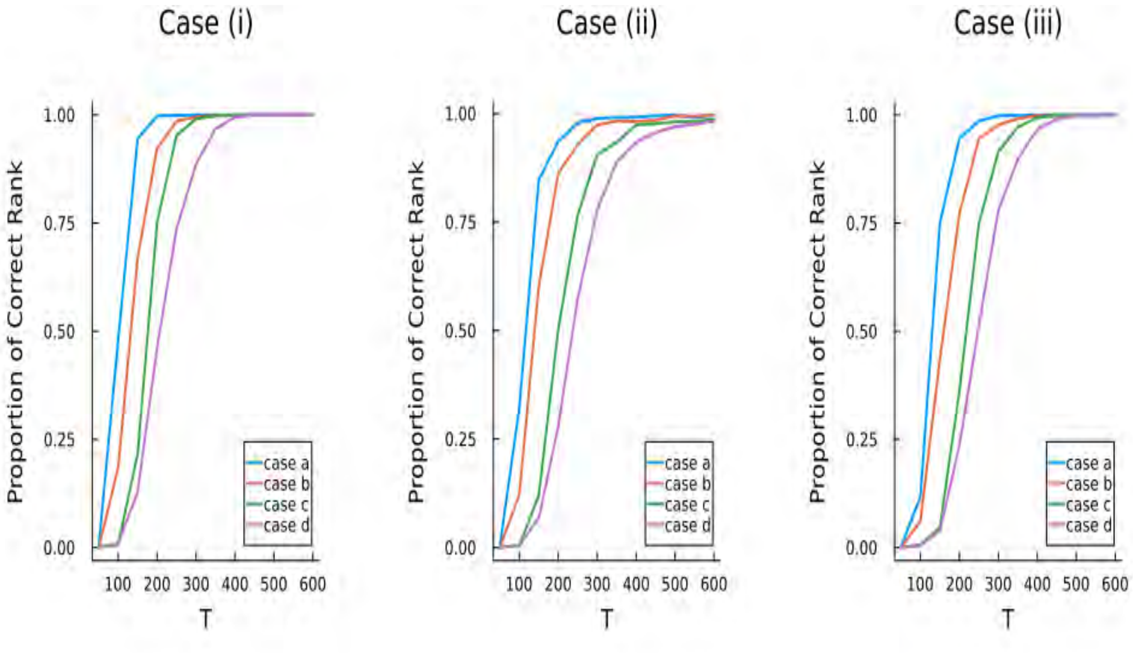


Figure 3.2: Proportion of correct rank selection under different singular value structures: (a) $(2, 2, 2)$, (b) $(4, 3, 2)$, (c) $(1, 1, 1)$, and (d) $(1.5, 1, 1)$, across three types of innovation errors.

3.6.2 Performance of MLRSQR Estimators

We now conduct a series of simulation experiments to examine the finite-sample performance and verify the asymptotic properties of the proposed MLRSQR estimator under various settings. In particular, we evaluate its behavior across different spatial quantile directions $\mathbf{u}$ under three different types of innovations, i.e., Case (i), Case (ii), and Case (iii), are generated as the same as in Section 3.6.1.

The synthetic data are generated from model (3.2) with dimensions $(N, P) = (10, 5)$, and multilinear ranks fixed at $(r_1, r_2, r_3) = (3, 3, 3)$. For the core tensor $\mathcal{G}$, we first draw entries from i.i.d. standard normal distributions and scale the tensor so that $\min_{1 \le i \le 3} \sigma_{r_i}(\mathcal{G}_{(i)}) = 1$. The factor matrices $\mathbf{U}_i$ are constructed using the top r_i left singular vectors of Gaussian random matrices, with the stationarity condition in Assumption 1 being satisfied.

The experiment includes 500 Monte Carlo replications per setting. The ranks are selected using the consistent rank selection method introduced in Section 3.5.

Throughout this and all the following experiments, the multilinear ranks are selected by the method in Section 3.5. We first assess the sensitivity of MLRSQR to the choice of $\mathbf{u}$. We consider the three different cases of $\mathbf{u}$: Case (1) is spatial median regression, and there are four sub-cases for Case (2) and Case (3), respectively. The details are as follows:

We consider the following cases for the direction vector $\mathbf{u}$ in the spatial quantile regression framework:

Case (1): $\mathbf{u} = \mathbf{0}$, i.e., spatial median regression.

Case (2.1): $\mathbf{u} = \frac{1}{\sqrt{10}} \times [0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]'$.

Case (2.2): $\mathbf{u} = \frac{1}{\sqrt{10}} \times [-0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5]'$.

Case (2.3): $\mathbf{u} = \frac{1}{\sqrt{10}} \times [0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5]'$.

Case (2.4): $\mathbf{u} = \frac{1}{\sqrt{10}} \times [-0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5]'$.

Case (3.1): $\mathbf{u} = \frac{1}{\sqrt{110}} \times [0.9, 0.9, 1.8, 1.8, 2.7, 2.7, 3.6, 3.6, 4.5, 4.5]'$.

Case (3.2): $\mathbf{u} = \frac{1}{\sqrt{110}} \times [-0.9, -0.9, -1.8, -1.8, -2.7, -2.7, -3.6, -3.6, -4.5, -4.5]'$.

Case (3.3): $\mathbf{u} = \frac{1}{\sqrt{110}} \times [0.9, -0.9, 1.8, -1.8, 2.7, -2.7, 3.6, -3.6, 4.5, -4.5]'$.

Case (3.4): $\mathbf{u} = \frac{1}{\sqrt{110}} \times [-0.9, 0.9, -1.8, 1.8, -2.7, 2.7, -3.6, 3.6, -4.5, 4.5]'$.

Note that, $\|\mathbf{u}\| = 0$ for Case (1), $\|\mathbf{u}\| = 0.5$ for Case (2), and $\|\mathbf{u}\| = 0.9$ for Case (3) are satisfied with $\|\mathbf{u}\| < 1$. Then we compute the mean squared error (MSE) for all entries in $\widehat{\mathcal{A}}$ and for all replications. See Figure 3.3.

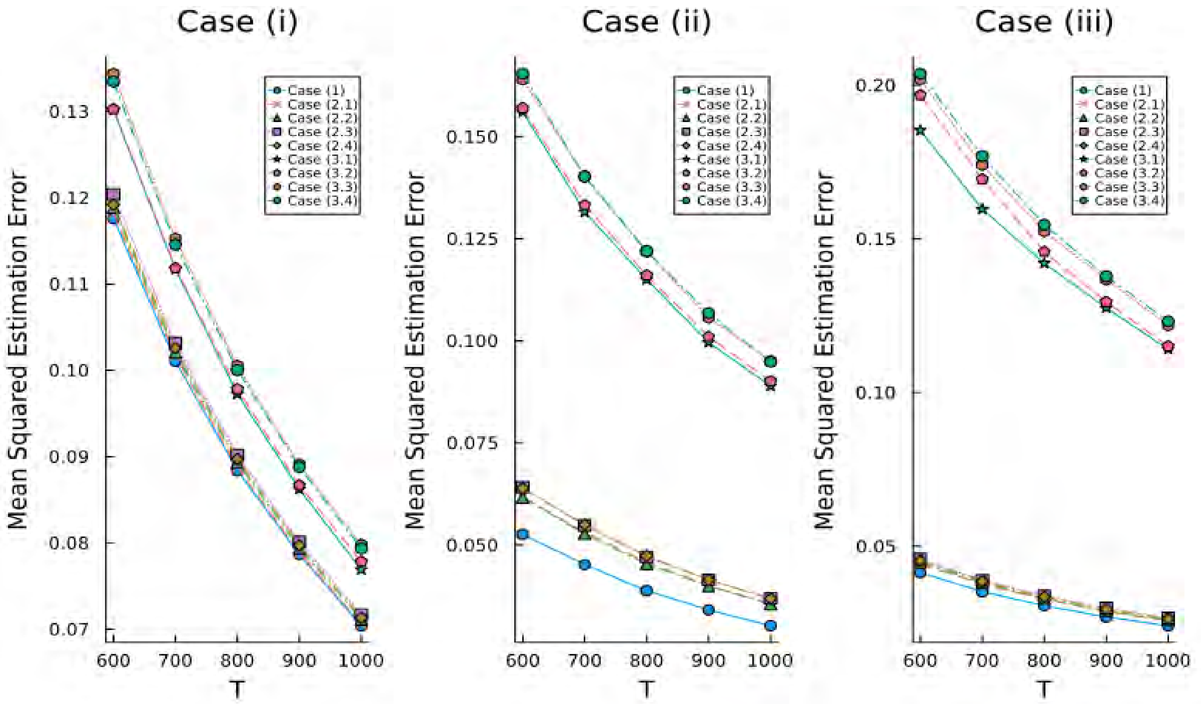


Figure 3.3: Mean Squared Errors of MLRSQR estimator for different **u** under different type of errors.

The results presented in Figure 3.3 clearly illustrate the asymptotic behavior of the proposed estimator. As the sample size T increases, the mean squared error (MSE) consistently decreases, indicating improved estimation accuracy with larger samples. This observation is consistent with the fundamental asymptotic property in statistical estimation, in which the estimation error tends to diminish as T increases. Moreover, while notable differences in performance across various configurations of $\mathbf{u}$ are observed for smaller sample sizes, these discrepancies gradually decrease as T increases. This suggests that the estimator becomes increasingly robust to the specific structure of $\mathbf{u}$ in the asymptotic regime. Another observation from Figure 3.3 is that, within each group of cases sharing the same norm $\|\mathbf{u}\|$, the corresponding mean squared error (MSE) curves are nearly indistinguishable. This indicates that the estimation performance is largely insensitive to the specific structure or pattern of the direction vector $\mathbf{u}$, such as the arrangement or signs of its components. Instead, the estimation error is primarily governed by the magnitude of $\|\mathbf{u}\|$. This finding highlights a desirable property of the proposed estimator: its robustness with respect to variations in the structure of $\mathbf{u}$, as long as the overall norm remains unchanged. Such behavior implies that the estimator is more influenced by the overall intensity of directional perturbation rather than its orientation, enhancing its stability and generalizability across different spatial directions. In general, the convergence of MSE and the reduced sensitivity to $\mathbf{u}$ highlight both the consistency and robustness of the model as it approaches its theoretical performance bounds.

Next, we compare the MLRSQR estimator with four alternatives: OLS, SQR, RRR, and the multilinear low-rank least squares (MLRLS) estimator. The MLRLS estimator is defined by the following:

$$\widehat{\mathcal{A}}_{\text{MLRLS}} \equiv [[\widehat{\mathcal{G}}; \widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3]] = \arg\min L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3), \tag{3.22}$$

where

$$L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \frac{1}{T} \sum_{t=1}^{T} \left\| \mathbf{y}_t - (\mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)} \mathbf{x}_t \right\|_2^2 . \tag{3.23}$$

The data are generated from model (2.7) with $(N, P) = (10, 5)$ and multilinear ranks $(r_1, r_2) = (3, 3)$, with $r_3 \in \{2, 3, 4\}$. We consider three different types of innovation distributions—Case (i), Case (ii), and Case (iii)—as defined in Section 3.6.1. The core tensor $\mathcal{G}$ is generated by scaling a randomly generated tensor with i.i.d. standard normal entries such that $\min_{1 \le i \le 3} \sigma_{r_i}(\mathcal{G}_{(i)}) = 1$. The factor matrices $\mathbf{U}_i$ are generated using the same procedure as described in Section 3.6.1. Each setting is replicated 500 times.

For each estimator, namely $\widehat{\mathcal{A}}_{\text{OLS}}$, $\widehat{\mathcal{A}}_{\text{SQR}}$, $\widehat{\mathcal{A}}_{\text{RRR}}$, $\widehat{\mathcal{A}}_{\text{MLRLS}}$, and $\widehat{\mathcal{A}}_{\text{MLRSQR}}$—we compute the mean squared error (MSE) across all entries of $\widehat{\mathcal{A}}$ and across all replications. See Figure 3.4, Figure 3.5, and Figure 3.6.

The MSEs are plotted against the sample size $T \in [600, 1000]$ in Figure 3.4, Figure 3.5, and Figure 3.6, corresponding to the three innovation types, which collectively demonstrate that all estimators exhibit improved performance as the sample size T increases, consistent with their asymptotic properties. Across all scenarios, the relative performance of the estimators remains consistent for different values of r_3, further confirming the stability and adaptability of MLRSQR across varying tensor dimensions and innovation structures.

On the other hand, under the standard normal errors (Case (i)), the estimator $\widehat{\mathcal{A}}_{\text{MLRLS}}$ achieves the smallest MSE. However, the MSE of $\widehat{\mathcal{A}}_{\text{MLRSQR}}$ is nearly identical to that of

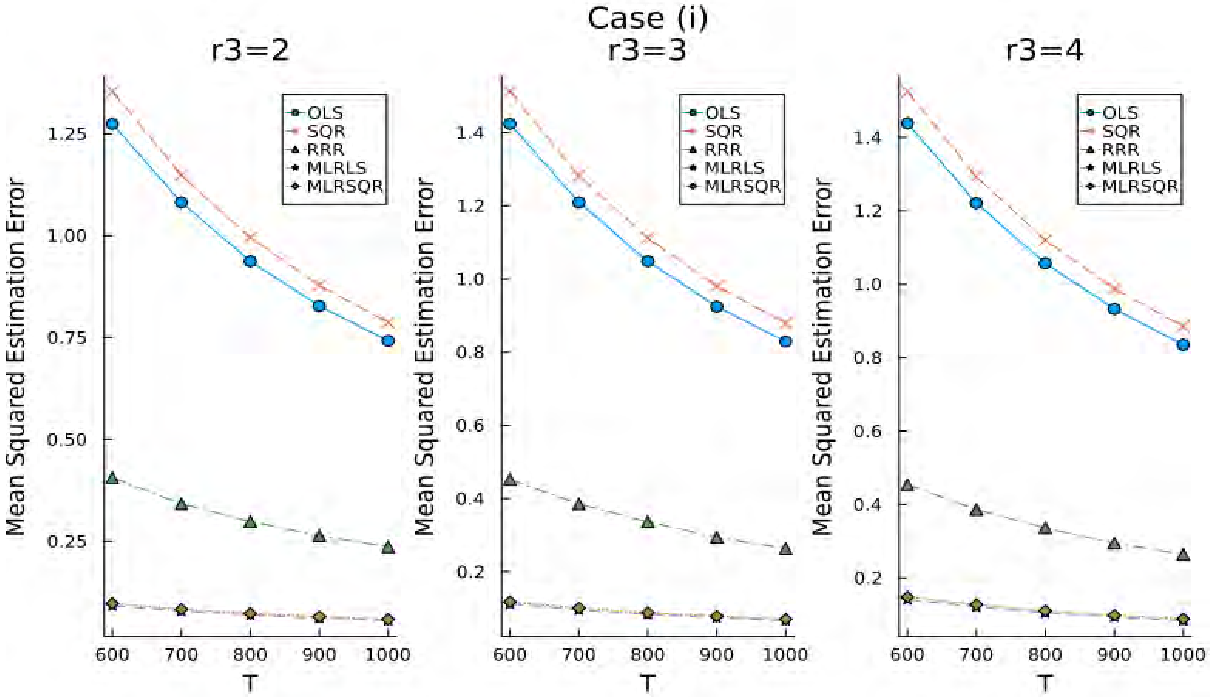


Figure 3.4: Case (i): MSE comparison across estimators under standard normal errors.

$\widehat{\mathcal{A}}_{\text{MLRLS}}$ across different values of T, indicating that MLRSQR maintains competitive performance even under Gaussian innovations. In contrast, under the t-distribution (Case (ii)) and the mixed normal distribution with outliers (Case (iii)), $\widehat{\mathcal{A}}_{\text{MLRSQR}}$ clearly outperforms all other estimators, exhibiting substantially lower MSEs. These results demonstrate the robustness of MLRSQR to heavy-tailed and contaminated error distributions, while still retaining near-optimal efficiency in the Gaussian setting.

3.6.3 Performance of SHORRSQR Estimators

To investigate the effect of sparsity constraints on the factor loading matrices, we conduct a simulation study comparing the SHORRSQR method—which incorporates regularization to enforce sparsity—with the MLRSQR method, which does not impose any sparsity constraints. Specifically, we generate multivariate time series data with known multilinear low-rank structures and varying degrees of sparsity in the factor loading matrices.

The data are generated from model (2.7) with $(N, P) = (10, 5)$, and the multilinear ranks are set as $(r_1, r_2, r_3) = (3, 3, 3)$. To introduce sparsity, we set $(s_1, s_2, s_3) = (2, 2, 2)$, where s_i denotes the number of nonzero rows in the corresponding factor matrix $\mathbf{U}_i$, for

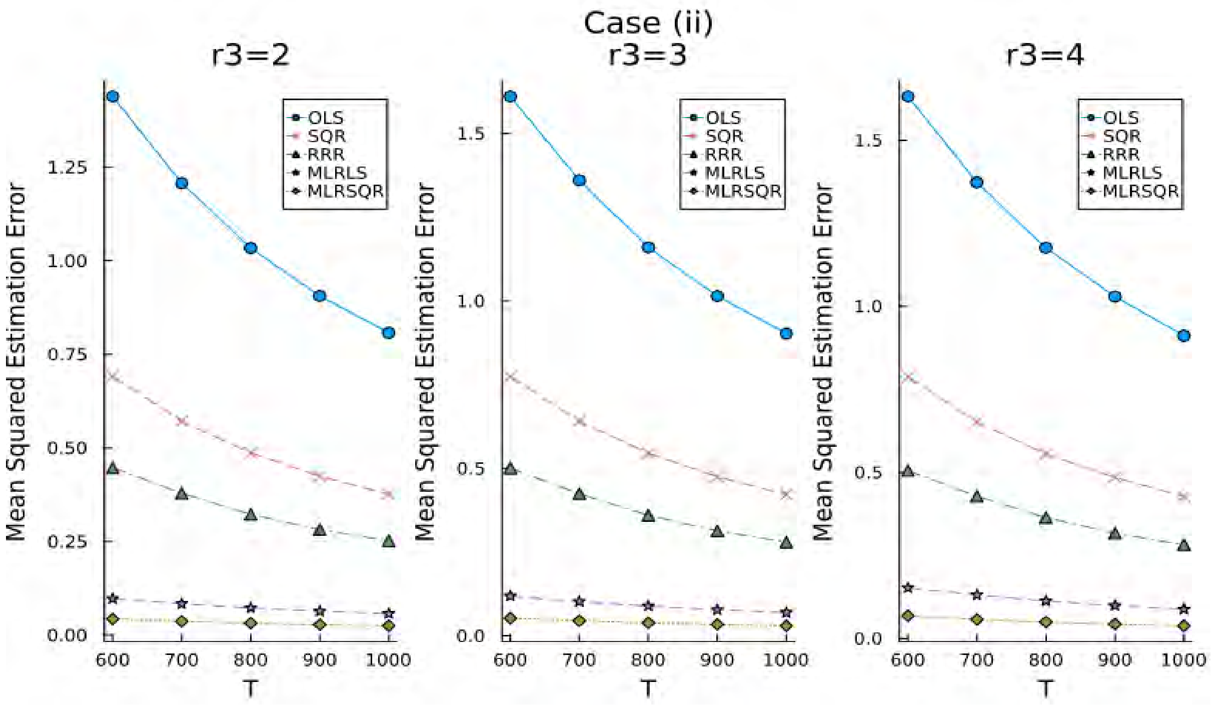


Figure 3.5: Case (ii): MSE comparison across estimators under t_3-distributed errors.

$i = 1, 2, 3$. This setting ensures that each $\mathbf{U}_i$ is row-sparse.

We consider three types of innovation distributions, as defined in Section 3.6.1. The core tensor $\mathcal{G}$ is generated in the same manner as in the previous simulation, and the sparse orthonormal factor matrices $\mathbf{U}_i$ are generated using the procedure described below:

$$\mathbf{U}_1 = \begin{bmatrix} \mathbf{a}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} \\ \mathbf{0}_{2\times1} & \mathbf{b}_{2\times1} & \mathbf{0}_{2\times1} \\ \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{c}_{2\times1} \\ \mathbf{0}_{4\times1} & \mathbf{0}_{4\times1} & \mathbf{0}_{4\times1} \end{bmatrix} \in \mathbb{R}^{10\times3}, \quad \mathbf{U}_2 = \begin{bmatrix} \mathbf{d}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} \\ \mathbf{0}_{2\times1} & \mathbf{e}_{2\times1} & \mathbf{0}_{2\times1} \\ \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{f}_{2\times1} \\ \mathbf{0}_{4\times1} & \mathbf{0}_{4\times1} & \mathbf{0}_{4\times1} \end{bmatrix} \in \mathbb{R}^{10\times3},$$

$$\mathbf{U}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} & \mathbf{0}_{2\times1} \end{bmatrix} \in \mathbb{R}^{5\times3}.$$

Both methods are applied to estimate the underlying tensor parameters, and their performance is evaluated based on the mean squared error (MSE) of the estimated tensor.

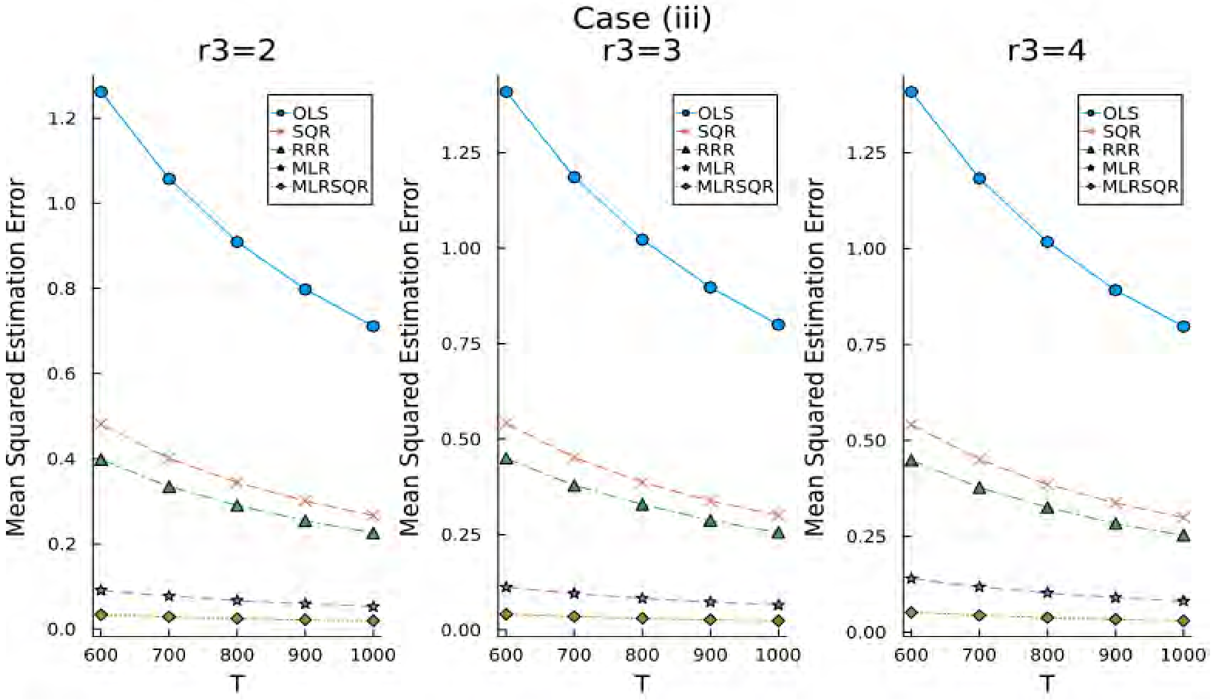


Figure 3.6: Case (iii): MSE comparison across estimators under mixture normal errors.

The simulation is repeated 500 times for each sample size, and the average MSE is reported in Figure 7. Unless otherwise specified, the data generation procedures for the following two simulation experiments are identical to those used in this experiment.

The results presented in Figure 7 demonstrate that the SHORRSQR method not only enhances estimation accuracy by incorporating sparsity constraints, but also effectively reduces the number of parameters, making it particularly suitable for high-dimensional data analysis. Notably, SHORRSQR consistently outperforms MLRSQR across all cases and sample sizes T, achieving substantially lower mean squared estimation errors (MSEs).

Both methods exhibit decreasing MSE as the sample size T increases, reflecting improved estimation accuracy with more data. However, the decline in MSE is more pronounced for SHORRSQR, underscoring the advantage of leveraging sparsity in the estimation process. Furthermore, the superior performance of SHORRSQR is robust across different innovation types and settings, highlighting its effectiveness and reliability in achieving accurate and efficient parameter estimation in sparse multilinear models.

Next, to validate the asymptotic properties discussed in Section 4.2, we conduct a simulation study comparing the proposed SHORRSQR estimator, denoted by $\widehat{\mathcal{A}}_{\text{SHORRSQR}}$,

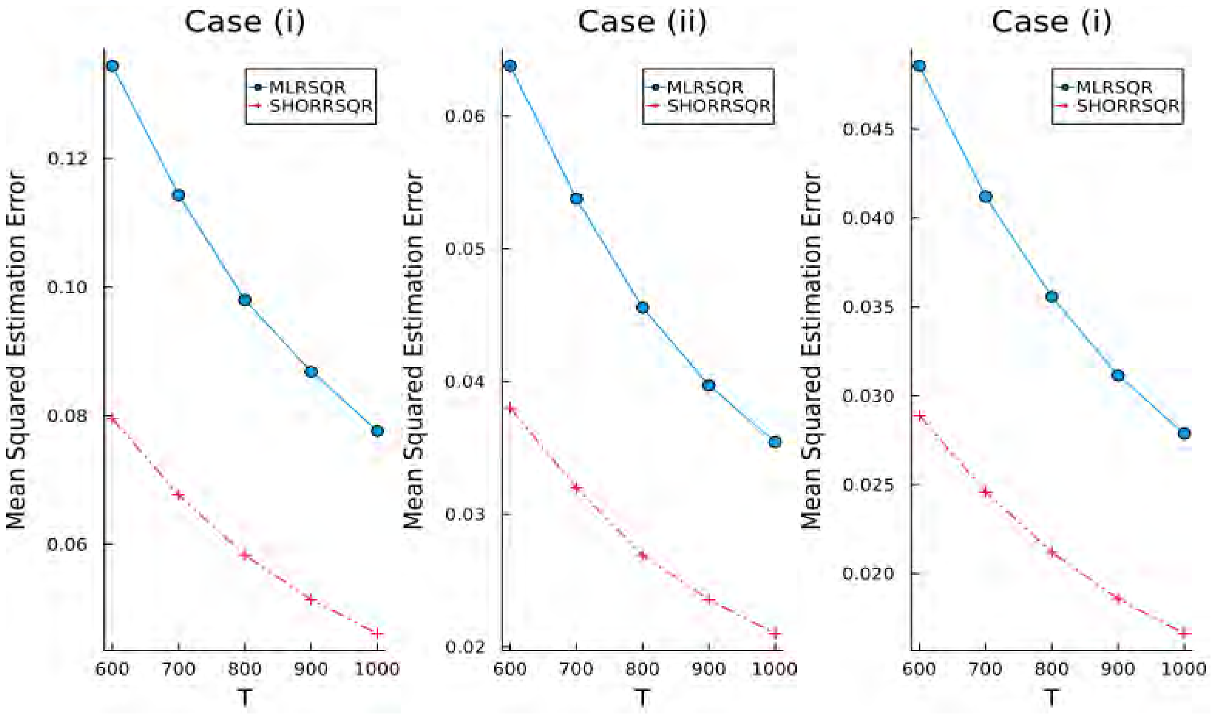


Figure 3.7: Mean Squared Errors of SHORRSQR estimator and MLRSQR estimator under different type of errors.

with five alternative estimators under comparable settings. The competing methods are as follows: (i) The Lasso estimator, $\widehat{\mathcal{A}}_{\text{LASSO}}$, which employs an ℓ_1-penalty for variable selection and regularization (Tibshirani, 1996 (80); Basu and Michailidis, 2015 (81)). (ii) The nuclear norm estimator, $\widehat{\mathcal{A}}_{\text{NN}}$, based on low-rank matrix recovery via nuclear norm minimization (Negahban and Wainwright, 2011 (36)). (iii) The spatial quantile regression via Lasso (SQRLASSO) estimator, $\widehat{\mathcal{A}}_{\text{SQRLASSO}}$, combining spatial quantile regression with an ℓ_1-penalty for sparsity. (iv) The ℓ_1-penalized sparse higher-order reduced-rank least squares (SHORRLS) estimator, $\widehat{\mathcal{A}}_{\text{SHORRLS}}$ (Wang, Zheng, Lian, and Li, 2022 (44)).

In this experiment, we fix the vector $\mathbf{u} = \mathbf{0}$ to eliminate directional effects and focus solely on evaluating the estimation accuracy and sparsity-inducing capability of each estimator. All methods are applied under the same data-generating settings described previously to ensure a fair and consistent comparison.

The spatial quantile regression via Lasso estimator is defined as

$$(\widehat{\mathbf{A}}_{\text{SQRLASSO}}, \widehat{\mathbf{q}}_u) = \arg\min L(\mathbf{A}, \mathbf{q}_u) \tag{3.24}$$

where

$$L(\mathbf{A}, \mathbf{q}_u) = \sum_{t=1}^{T} \{ \|\mathbf{y}_t - \mathbf{A}\mathbf{x}_t - \mathbf{q}_u\| + \mathbf{u}^T(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t - \mathbf{q}_u) \} + \lambda \|\mathbf{A}\|_1. \tag{3.25}$$

Note that, $\|\mathbf{A}\|_1 = \|\text{vec}(\mathbf{A})\|_1$. Denote by $\widehat{\mathcal{A}}_{\text{SQRLASSO}}$ transition tensors formed by $\widehat{\mathbf{A}}_{\text{SQRLASSO}}$.

The ℓ_1-penalized sparse higher-order reduced-rank least square (SHORRLS) estimator is defined as

$$\begin{aligned} \widehat{\mathcal{A}}_{\text{SHORRLS}} \equiv [[\widehat{\mathcal{G}}; \widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3]] = \underset{\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3}{\arg\min} \{ & L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) \\ & + \lambda \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1 \} \end{aligned} \tag{3.26}$$

subject to

$$\mathcal{G} \in \text{AO}(r_1, r_2, r_3) \quad \text{and} \quad \mathbf{U}_i' \mathbf{U}_i = \mathbf{I}_{r_i}, \quad i = 1, 2, 3, \tag{3.27}$$

where $L(\mathcal{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ is defined as in (**??**), and $\text{AO}(r_1, r_2, r_3) = \{\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \mathcal{G}_{(i)}$ is row-orthogonal, $i = 1, 2, 3\}$. Unlike the unconstrained estimation in (**??**), the orthogonality constraints in (3.27) are necessary; otherwise, the sparsity patterns of $\mathbf{U}_i$ cannot be identified.

Figure 8 presents the mean squared estimation errors (MSEs) of five estimators—LASSO, SQRLASSO, NN, SHORRLS, and SHORRSQR—across varying sample sizes T under three different innovation settings: standard normal (Case (i)), t_3-distribution (Case (ii)), and mixed normal distribution (Case (iii)).

In most cases, SHORRSQR consistently achieves the lowest MSE among all methods, with the performance gap becoming more pronounced as T increases. Notably, under the heavy-tailed (Case (ii)) and mixed (Case (iii)) error settings, SHORRSQR substantially outperforms all competing methods, highlighting its robustness to non-Gaussianity and outliers. Even under the standard normal distribution (Case (i)), SHORRSQR maintains highly competitive performance, only marginally higher than SHORRLS in some instances.

SHORRLS performs well under Case (i) but its performance degrades in Cases (ii) and (iii), indicating sensitivity to distributional assumptions. SQRLASSO consistently im-

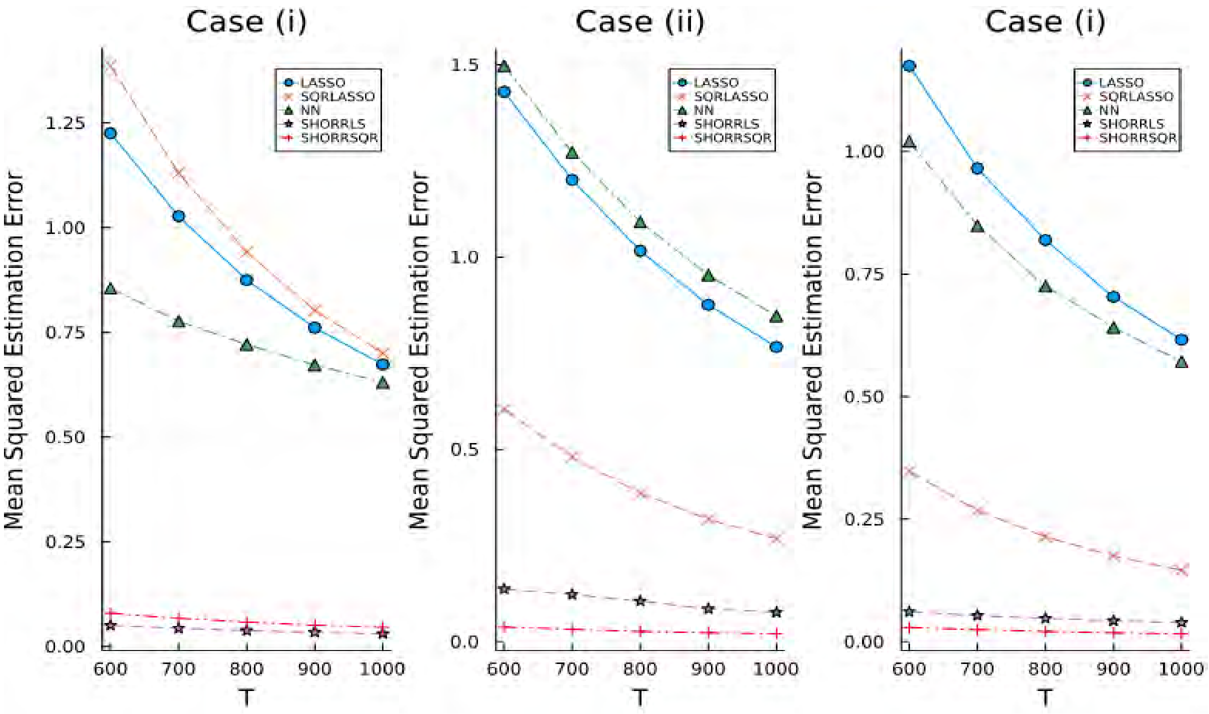


Figure 3.8: Mean Squared Errors for $\widehat{\mathcal{A}}_{\text{LASSO}}$, $\widehat{\mathcal{A}}_{\text{SQRLASSO}}$, $\widehat{\mathcal{A}}_{\text{NN}}$, $\widehat{\mathcal{A}}_{\text{SHORRLS}}$, and $\widehat{\mathcal{A}}_{\text{SHORRSQR}}$ under different type of errors.

proves over standard LASSO under Cases (ii) and (iii), benefiting from the incorporation of spatial quantile structure, but remains less effective than the SHORR-based methods. The performance of the NN estimator is sensitive to the type of error distribution. It performs relatively poorly under the t_3 and mixed normal distributions but demonstrates moderate accuracy when averaged across all settings.

Overall, these results confirm that SHORRSQR provides the most accurate and reliable estimates, particularly in settings with heavy-tailed or heterogeneous errors, while retaining competitive accuracy in the standard Gaussian setting.

Then, we verify the asymptotic results in Section 4.1 for the proposed SHORRSQR estimator on different spatial quantiles $\mathbf{u}$. The data is generated in the same way as above.

The results of Figure 9 clearly demonstrate the asymptotic behavior of the estimation. What's more, Figure 9 plots compare the mean squared estimation error (MSE) across three types of innovations: standard normal distribution (Case i), t_3-distribution with heavy tails (Case ii), and a mixture of normal distributions with differing variances (Case iii). In all cases, MSE decreases asymptotically as the sample size T increases, demon-

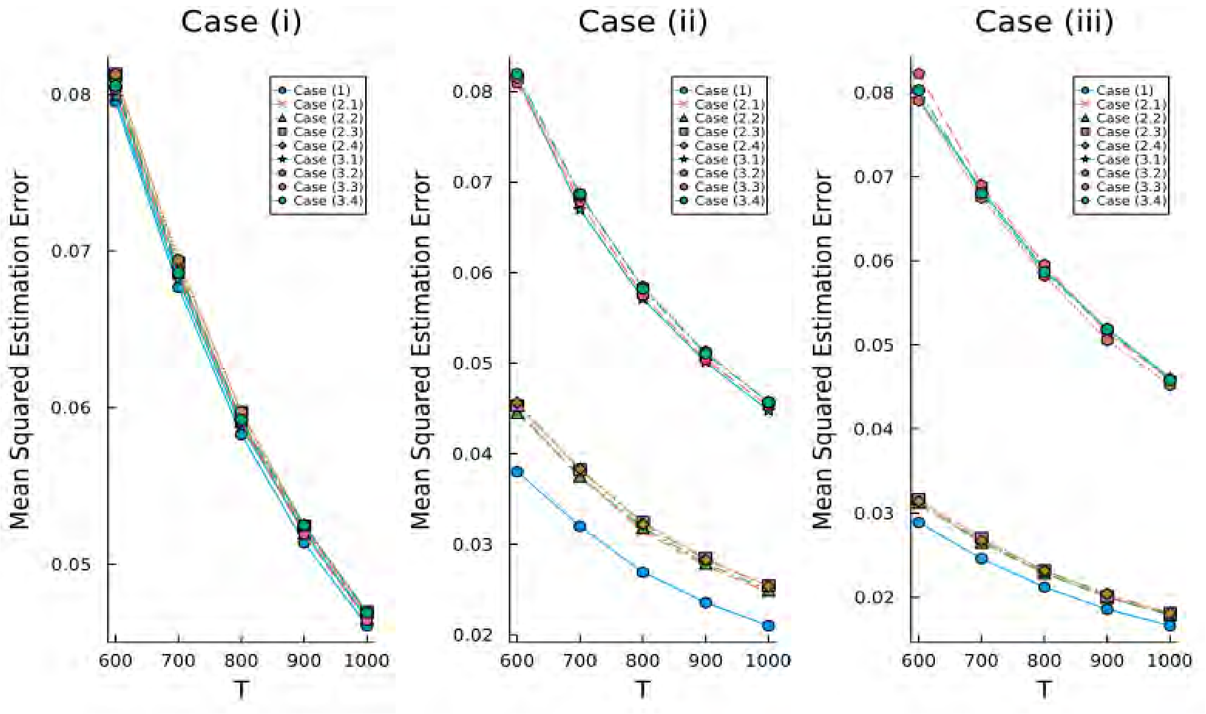


Figure 3.9: Mean Squared Errors of SHORRSQR estimator for different **u** under different type of errors.

strating improved estimation accuracy with larger samples. The results also compare different **u** -th quantiles, the different cases are the same as in Section 6.2, showing that the spatial quantile regression (SQR) method consistently achieves stable performance across quantiles, even in the presence of outliers. In particular, in cases (ii) and (iii), which introduce heavy-tailed errors and heterogeneous variance structures, SQR demonstrates robustness by maintaining a lower MSE compared to methods sensitive to outliers. These results highlight the strong asymptotic property, robustness, and flexibility of SQR in handling outliers, complex error distributions, and different quantile levels, making it a reliable approach in non-Gaussian innovation settings.

3.7 Real Data Analysis

In this section, we apply the proposed estimation methods to jointly model 40 quarterly U.S. macroeconomic time series spanning from 1959 to 2007, with 194 observed values for each variable (Koop, 2013 (23)). Except for financial variables, all series are seasonally adjusted, transformed to achieve stationarity, and standardized to have zero mean and unit variance. These variables capture a broad range of economic activity and are categorized into eight groups: (i) GDP and its components, (ii) National Association of Purchasing Managers (NAPM) indices, (iii) industrial production, (iv) housing, (v) money, credit, and interest rates, (vi) employment, (vii) prices and wages, and (viii) other indicators. The vector autoregressive (VAR) model has been extensively used to analyze and forecast such macroeconomic data in empirical studies; see, for example, Stock and Watson (2009) (82) and Koop (2013) (23). Detailed descriptions of the macroeconomic variables are provided in Table 1.

Table 3.1: Forty quarterly macroeconomic variables belonging to eight categories.

Short name	**C**	**T**	**Description**	**Short name**	**C**	**T**	**Description**
GDP251	1	5	Real GDP, quantity index (2000=100)	FM2	5	6	Money stock: M2 (bil$)
GDP252	1	5	Real personal cons exp, quantity index	FMRNBA	5	3	Depository inst reserves: nonborrowed (mil$)
GDP253	1	5	Real personal cons exp: durable goods	FMRRA	5	6	Depository inst reserves: total (mil$)
GDP256	1	5	Real gross private domestic investment	FSPIN	5	5	S&P's common stock price index: industrials
GDP263	1	5	Real exports	FYFF	5	2	Interest rate: federal funds (% per annum)
GDP264	1	5	Real imports	FYGT10	5	2	Interest rate: US treasury const. mat., 10-yr
GDP265	1	5	Real govt cons expenditures & gross investment	SEYGT10	5	1	Spread btwn 10-yr and 3-mth T-bill rates
GDP270	1	5	Real final sales to domestic purchasers	CES002	6	5	Employees, nonfarm: total private
PMCP	2	1	NAPM commodity price index (%)	LBMNU	6	5	Hrs of all persons: nonfarm business sector
PMDEL	2	1	NAPM vendor deliveries index (%)	LBOUT	6	5	Output per hr: all persons, business sec
PMI	2	1	Purchasing managers' index	LHEL	6	2	Index of help-wanted ads in newspapers
PMNO	2	1	NAPM new orders index (%)	LHUR	6	2	Unemp. rate: All workers, 16 and over (%)
PMNV	2	1	NAPM inventories index (%)	CES275R	7	5	Real avg hrly earnings, nonfarm prod. workers
PMP	2	1	NAPM production index (%)	CPIAUCSL	7	6	CPI all items
IPS10	3	5	Industrial production index: total	GDP273	7	6	Personal consumption exp.: price index
UTL11	3	1	Capacity utilization: manufacturing (SIC)	GDP276	7	6	Housing price index
HSFR	4	4	Housing starts: Total (thousands)	PSCCOMR	7	5	Real spot market price index: all commodities
BUSLOANS	5	6	Comm. and industrial loans at all comm. Banks	PWFSA	7	6	Producer price index: finished goods
CCINRV	5	6	Consumer credit outstanding: nonrevolving	EXRUS	8	5	US effective exchange rate: index number
FM1	5	6	Money stock: M1 (bil$)	HHSNTN	8	2	Univ of Mich index of consumer expectations

NOTE: Category code (C) represents: 1 = GDP and its decomposition, 2 = national association of purchasing managers (NAPM) indices, 3 = industrial production, 4 = housing, 5 = money, credit, interest rates, 6 = employment, 7 = prices and wages, 8 = others. Variables are seasonally adjusted except for those in category 5. All variables are transformed to stationarity with the following transformation codes (T): 1 = no transformation, 2 = first difference, 3 = second difference, 4 = log, 5 = first difference of logged variables, 6 = second difference of logged variables.

To assess forecasting performance, we compare the proposed estimators $\widehat{\boldsymbol{A}}_{\text{MLRSQR}}$ and $\widehat{\boldsymbol{A}}_{\text{SHORRSQR}}$ with the competing methods introduced in Section 6. A rolling one-step-ahead forecasting procedure is implemented as follows: the models are fitted using historical data with the training endpoint rolling from Q4 2000 to Q3 2007, and forecasts are generated for the subsequent quarter. The lag order is fixed at $P = 4$ for all fitted VAR models, as suggested by Koop (2013) (23). The selected multilinear ranks and tuning parameters for $\widehat{\boldsymbol{A}}_{\text{SHORRSQR}}$ are inherited from the full-sample analysis and set as $(r_1, r_2, r_3) = (4, 3, 2)$. For consistency, the same rank configuration is applied to

MLRSQR, MLRLS, SHORRLS, and RRR estimators in the comparison.

The average ℓ_2 and ℓ_∞ norms of the forecast errors for various methods are reported in Table 2. It is evident that the proposed MLRSQR and SHORRSQR estimators yield substantially smaller forecast errors compared to the competing approaches. This superior performance can be attributed to their ability to simultaneously reduce dimensionality along all three tensor modes, thereby capturing the underlying low-rank structure more effectively.

Table 3.2: Forecasting error for 40 quarterly macroeconomic sequences of the United States from 1959 to 2007.

	Unregularized methods					Regularized methods				
Criterion	OLS	SQR	RRR	MLRLS	MLRSQR	SHORRSQR	SHORRLS	SQRLASSO	LSLASSO	NN
ℓ_2 norm	20.13	19.22	12.61	5.80	**5.74**	**5.34**	5.43	8.22	8.24	8.30
ℓ_∞ norm	8.30	8.18	4.54	2.55	**2.53**	**2.48**	2.50	3.28	3.59	3.59

Note: The best cases among (un)regularized methods are marked in bold.

In particular, the advantage of spatial quantile regression (SQR) over ordinary least squares (OLS) becomes especially prominent in the presence of heavy-tailed, asymmetric, or heteroscedastic errors. Unlike OLS, which relies on assumptions of homoscedasticity and symmetric error distributions and is highly sensitive to outliers, SQR minimizes a quantile loss function, offering greater robustness to non-Gaussian and contaminated data. Moreover, SQR allows for the modeling of heterogeneous relationships across different quantiles, capturing variation in the tails of the conditional distribution that OLS fails to detect. It is particularly effective in high-dimensional or spatially dependent settings, especially when integrated with sparsity-inducing regularization.

Among all methods, the SHORRSQR estimator achieves the best overall forecasting performance. By enforcing sparsity in the factor loading matrices, it not only improves interpretability but also mitigates overfitting, making it particularly well-suited for macroeconomic forecasting and other complex, high-dimensional applications involving structural heterogeneity.

REFERENCES

[1] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica.* **46**, 33-50.

[2] Chaudhuri,P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression. *The Annals of Statistics* **25**, 715-744.

[3] Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

[4] Koenker, R. and Zhao Q. (1996). Conditional quantile estimation and inference for ARCH models. *Econometric Theory* **12**, 793-813.

[5] Koul, H. L. and Saleh, A. K. Md. E. (1995). Autoregression quantiles and related rank scores processes. *Ann. Statist.* **23**, 670-689.

[6] Davis, R. A. and Dunsmuir, W. T. M. (1997). Least absolute deviation estimation for regression with ARMA errors. *J. Theor. Prob.* **10**, 481-497.

[7] Jiang, J., Zhao, Q. and Hui, Y. V. (2001). Robust modlling of ARCH models. *Journal of Forecasting.* **20**, 111-133.

[8] Peng, L. and Yao, Q. (2003). Least absolute deviation estimation for ARCH and GARCH models. *Biometrika*, **90**, 967-975.

[9] Bollerslev, T. (1990). Modeling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH approach. *Review of Economics and Statistics* **72**, 498-505.

[10] Engle R. F. and Kroner, K. (1995). Multivariate simultaneous GARCH. *Econometric Theory* **11**, 122-150.

[11] Chen, R. and Tsay, R. S. (1993). Functional coefficient autoregressive models. *Journal of American Statistical Association* **88**, 298-308.

[12] Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95**, 365-379.

[13] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of American Statistical Association* **91**, 862-871.

[14] Koltchinskii, V. (1997). M-estimation, convexity and quantiles. *The Annals of Statistics* **25**, 435-477.

[15] Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *J. Statist. Plann. Inference*, **123**, 259-278.

[16] Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, **48**, 1-48.

[17] Jiang, J. C., Jiang, X. J., Liu, J. Z., Liu, Y., and Yan, W. F. (2017). Spatial quantile estimation of multivariate threshold time series models. *Physica A: Statistical Mechanics and its Applications*, **486**, 772-781.

[18] Bai, Z. D., Chen, N. R., Miao, B. Q., and Rao, C. R. (1990). Asymptotic theory of least distance estimate in mutilvariate linear models. *Statistics* **21**, 503-519.

[19] Chakraborty, B. and Chaudhuri, P. (1998). On an adaptive transformation-retransformation estimate of multivariate location. *J. R. Statist. Soc. B* **60**, 145-157.

[20] Chakraborty, B. (2003). On multivariate quantile regression. *Journal of Statistical Planning and Inference* **110**, 109-132.

[21] De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, **146**, 318-328.

[22] Carriero, A., Kapetanios, G., and Marcellino, M. (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, **26**, 735-761.

[23] Koop, G. M.(2013). Forecasting with medium and large bayesian VARs. *Journal of Royal Statistical Society*, Series B. **58**, 267-288.

[24] Chan, J. C. C., Eisenstat, E., and Koop, G. (2016). Large bayesian VARMAs. *Journal of Econometrics*, **192**, 374-390.

[25] Wilms, I., Basu, S., Bien, J., and Matteson, D. S. (2017). Sparse identification and estimation of large-scale vector autoregressive moving averages. ArXiv no. 1707.09208.

[26] Dias, G. F., and Kapetanios, G. (2018). Estimation and forecasting in vector autoregressive moving average models for rich datasets. *Journal of Econometrics*, **202**, 75-91.

[27] Said, E. S., and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, **71**, 599-607.

[28] Li, G., Leng, C., and Tsai, C.-L. (2014). A hybrid bootstrap approach to unit root tests. *Journal of Time Series Analysis*, **35**, 299-321.

[29] Ravenna, F. (2007). Vector autoregressions and reduced form representations of DSGE models. *Journal of Monetary Economics*, **54**, 2048-2064.

[30] Basu, S., and Michailidis, G. (2015). Regularized estimation in sparse high-Dimensional time series models. *The Annals of Statistics*, **43**, 1535-1567.

[31] Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, **16**, 3115-3150.

[32] Kock, A. B., and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, **186**, 325-344.

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

[34] Wu, W.-B., and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, **10**, 352-379.

[35] Yuan, M.,Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimationin multivariate linear regression. *Journal of the Royal Statistical Society, Series B*, **69**, 329-346.

[36] Negahban, S., and Wainwright, M. J. (2011). Estimation of (near) low rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, **39**, 1069-1097.

[37] Chen, K., Dong, H.,and Chan,K.-S.(2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, **100**, 901-920.

[38] Basu, S., Li, X., and Michailidis, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, **67**, 1207-1222.

[39] Raskutti, G., Yuan, M., and Chen, H. (2019). Convex regularization for high-dimensional multi-response tensor regression. *The Annals of Statistics*, **47**, 1554-1584.

[40] Velu, R. P., Reinsel, G. C., and Wichern, D. W. (1986). Reduced rank models for multiple time series. *Biometrika*, **73**, 105-118.

[41] Velu, R.P., and Reinsel, G.C.(2013). Multivariate reduced-rank regression: theory and applications. (Vol. 136), New York: Springer-Verlag.

[42] Reinsel, G. (1983). Some results on multivariate autoregressive index models. *Biometrika*, **70**, 145-156.

[43] Kolda, T. G., and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, **51**, 455-500.

[44] Wang, D., Zheng, Y., Lian, H., and Li, G. (2022). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, **117**, 1338-1356.

[45] Tibshirani, R. J.(1996). Regression shrinkage and selection via the lasso. *Journal of Applied Econometrics*, **28**, 177-203.

[46] Fan, J. Q., Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

[47] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

[48] Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics* **1**, 799-821.

[49] Huber, P. J. (1988). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics* **1**, 799-821.

[50] Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics* **16**, 356-366.

[51] Donaho, D. L. (2000). . High-dimensional data analysis: The curses and blessings of dimensionality. *Lecture on August 8, 2000, to the American Mathematical Society on Math Challenges of the 21st Century* Available at http : //www.inma.ucl.ac.be/ francois/these/papers/entry-Donoho-2000.html.

[52] Fan, J., and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parametes. *The Annals of Statistics* **32**, 928-961.

[53] Lam. C., and Fan, J. (2008). Profile-Kernel likelihood inference with diverging number of parameters. *The Annals of Statistics* **36**, 2232-2260.

[54] Wang, H., Li, R., and Tsai, C-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

[55] Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680.

[56] Zou, H., and Yuan, M., (2008b). Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics & Data Analysis* **52**, 5296-5304.

[57] Jiang, J., C., Jiang, X., J., and Song, X., Y. (2012). Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statistica Sinica* **22**, 1479-1506.

[58] Breiman, L. (1995). Better subset regression using the non-negative garotte. *Technometrics* **37**, 373-384.

[59] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

[60] Davis, R., A., Zang, P., F., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics* **25**, 1077-1096.

[61] Chen, L., and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, **107**, 1533-1545.

[62] Bunea, F., She, Y., and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, **40**, 2359-2388.

[63] Lian, H., Feng, S., and Zhao, K. (2015). Parametric and semiparametric reduced-rank regression with flexible sparsity. *Journal of Multivariate Analysis*, **136**, 163-174.

[64] Chen, K., Chan, K.-S., and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society, Series B*, **74**, 203-221.

[65] Uematsu, Y., Fan, Y., Chen, K., Lv, J., and Lin, W. (2019). SOFAR: large scale association network learning. *IEEE Transactions on Information Theory*, **65**, 4924-4939.

[66] Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis, *Journal of the American Statistical Association* **108**, 540-552.

[67] Li, L., and Zhang, X. (2017). Parsimonious tensor response regression, *Journal of the American Statistical Association* **112**, 1131-1146.

[68] Sun, W. W., and Li, L. (2017). Store: sparse tensor response regression and neuroimaging analysis, *The Journal of Machine Learning Research* **18**, 4908-4944.

[69] Li, X., Xu, D., Zhou, H., and Li, L. (2018). Tucker tensor regression and neuroimaging analysis, *Statistics in Biosciences* **10**, 520-545.

[70] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279-311.

[71] De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications* **21**, 1253-1278.

[72] Wang, Y., Yin, W., and Zeng, J. (2019). Global convergence of ADMM in nonconvex nonsmooth optimization, *Journal of Scientific Computing* **78**, 29-63.

[73] Zhao, J., and Leng, C. (2014). Structure lasso for regression with matrix covariates. *Statistica Sinica* **24**, 799-814.

[74] Ding, S., and Cook,R.D.(2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society* **Series B, 80**, 387-408.

[75] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models, *Journal of the American Statistical Association* **81**, 142-149.

[76] Li, G., Yang, D., Nobel, A. B., and Shen, H. (2016). Supervised singular value decomposition and its asymptotic properties, *Journal of Multivariate Analysis* **146**, 7-17.

[77] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends in Machine learning* **3**, 1-122.

[78] Lai, R., and Osher, S. (2014). A splitting method for orthogonality constrained problems, *Journal of Scientific Computing* **58**, 431-449.

[79] Xia, Q., Xu, W., and Zhu, L.(2015). Consistently determining the number of factors in multivariate volatility modelling, *Statistica Sinica* **25**, 1025-1044.

[80] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society* **Series B, 58**, 267-288.

[81] Basu, S., and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models, *The Annals of Statistics* **43**, 1535-1567.

[82] Stock, J. H., and Watson, M. W. (2009). Forecasting in dynamic factor models subject to structural instability, *in The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry,eds.* **J.L.CastleandN.Shep hard, Oxford: Oxford University Press, pp.**, 173-205.

[83] Van Der Vaart, A. W. (1998). Asymptotic Statistics. *Cambridge University Press.*

[84] Niemiro, W., (1992). Asymptotics for M-estimators defined by convex minimization, *The Annals of Statistics*, **20** 1514-1533.

[85] Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Berlin: Springer-Verlag.

[86] GEYER, C., J. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.

[87] Izenman, A., J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, **5** 248-264.

[88] Velu, R., P., and Reinsel, G., C. (2013). Multivariate reduced-rank regression: theory and applications, *Springer Science & Business Media* **volume 136**.

APPENDIX A: CONDITIONS and PROOFS of THEOREMS IN CHAPTER 2

A.1 Regularity conditions

(i) Regularity Conditions on the Penalty Function. Let $a_T = \max_{1 \le j \le p_T}\{p'_{\lambda_T}(|\beta^*_{Tj}|) : \beta^*_{Tj} \neq 0\}$ and $b_T = \max_{1 \le j \le p_T}\{p''_{\lambda_T}(|\beta^*_{Tj}|) : \beta^*_{Tj} \neq 0\}$. The following conditions are imposed on the penalty function $p_{\lambda_T}(\cdot)$:

(A_1) $\liminf_{T \to +\infty} \liminf_{\theta \to 0^+} p'_{\lambda_T}(\theta)/\lambda_T > 0$;

(A_2) $a_T = O(T^{-1/2})$;

(A_2') $a_T = o(1/\sqrt{Tp_T})$;

(A_3) $b_T \to 0$ as $T \to +\infty$;

(A_3') $b_T = o_P(1/\sqrt{p_T})$;

(A_4) There exist constants C and D such that $|p''_{\lambda_T}(\theta_1) - p''_{\lambda_T}(\theta_2)| \le D|\theta_1 - \theta_2|$, where $\theta_1, \theta_2 > C\lambda_T$.

These conditions are adapted from the regularity assumptions commonly used in the literature on penalized estimation, particularly those outlined by Fan and Peng (2004) (52).

(ii) Regularity Conditions on the Regression Function.

(B_1) Let $p'_T = p_T + N$. For each T, assume that the regression function $L_T(\boldsymbol{\theta}_T)$ is differentiable and satisfies:

$$E_{\boldsymbol{\theta}_T}\left\{\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tj}}\right\} = 0,$$
$$E_{\boldsymbol{\theta}_T}\left\{\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tj}}\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tk}}\right\} = -E_{\boldsymbol{\theta}_T}\left\{\frac{\partial^2 L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tj}\partial \theta_{Tk}}\right\}, \quad \text{for } j,k = 1,\ldots,p'_T,$$

which implies

$$E_{\boldsymbol{\beta}_T}\left\{\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}}\right\} = 0,$$
$$E_{\boldsymbol{\beta}_T}\left\{\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}}\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tk}}\right\} = -E_{\boldsymbol{\beta}_T}\left\{\frac{\partial^2 L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}\partial \beta_{Tk}}\right\}, \quad \text{for } j,k = 1,\ldots,p_T.$$

(B_2) The information matrices $I_T(\boldsymbol{\theta}_T)$ and $I_T(\boldsymbol{\beta}_T)$ are defined respectively as:

$$I_T(\boldsymbol{\theta}_T) = E\left[\left(\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \theta_T}\right)\left(\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \theta_T}\right)^T\right],$$
$$I_T(\boldsymbol{\beta}_T) = E\left[\left(\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_T}\right)\left(\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_T}\right)^T\right],$$

with bounded eigenvalues: $0 < C_1 \leq \lambda_{\min}(I_T(\cdot)) \leq \lambda_{\max}(I_T(\cdot)) \leq C_2 < \infty$. Also, for $j, k = 1, \ldots, p_T'$ and $j, k = 1, \ldots, p_T$, we assume:

$$E_{\boldsymbol{\theta}_T}\left\{\left(\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tj}}\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tk}}\right)^2\right\} < C_3 < \infty,$$
$$E_{\boldsymbol{\theta}_T}\left\{\left(\frac{\partial^2 L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tj}\partial \theta_{Tk}}\right)^2\right\} < C_4 < \infty.$$

which implies

$$E_{\boldsymbol{\beta}_T}\left\{\left(\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}}\frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tk}}\right)^2\right\} < C_3 < \infty,$$
$$E_{\boldsymbol{\beta}_T}\left\{\left(\frac{\partial^2 L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}\partial \beta_{Tk}}\right)^2\right\} < C_4 < \infty.$$

(B_3) There exist sufficiently large open subsets $\boldsymbol{\Omega}_T' \subset \mathbb{R}^{p_T+N}$ and $\boldsymbol{\Omega}_T \subset \mathbb{R}^{p_T}$, such that $\boldsymbol{\Omega}_T \in \boldsymbol{\Omega}_T'$, containing the true parameter points $\boldsymbol{\theta}_T$ and $\boldsymbol{\beta}_T$, respectively. For almost all $\mathbf{V}_{Tt}$, the regression function admits third derivatives on these subsets. Additionally, there exist functions $M_{Tjkl}(\mathbf{V}_{Tt})$ such that:

$$\left|\frac{\partial^3 L_T(\boldsymbol{\theta}_T)}{\partial \theta_{Tj}\partial \theta_{Tk}\partial \theta_{Tl}}\right| \leq M_{Tjkl}(\mathbf{V}_{Tt}),$$
$$E_{\boldsymbol{\theta}_T}\{M_{Tjkl}^2(\mathbf{V}_{Tt})\} < C_5 < \infty.$$

which implies

$$\left|\frac{\partial^3 L_T(\boldsymbol{\theta}_T)}{\partial\beta_{Tj}\partial\beta_{Tk}\partial\beta_{Tl}}\right| \le M_{Tjkl}(\mathbf{V}_{Tt}),$$

$$E_{\boldsymbol{\beta}_T}\{M^2_{Tjkl}(\mathbf{V}_{Tt})\} < C_5 < \infty.$$

(B_4) $\beta^*_{T1}, \beta^*_{T2}, \ldots, \beta^*_{Tj}$ satisfy $\min_{1\le j\le s_T} \frac{|\beta^*_{Tj}|}{\lambda_T} \to \infty \quad \text{as} \quad T \to \infty.$

(B_5) $\beta^*_{T1}, \beta^*_{T2}, \ldots, \beta^*_{Tj}$ satisfy $\min_{1\le j\le s_T} \frac{|\beta^*_{Tj}|}{\sqrt{T}h_T} \to \infty \quad \text{as} \quad T \to \infty.$

A.2 Proofs of Theorems

Theorem 3 and Theorem 4 can be established by following similar arguments used in the proofs of Theorem 1 and Theorem 2. Therefore, in what follows, we focus on proving Theorem 1 and Theorem 2.

To distinguish between the SCAD and adaptive LASSO estimators throughout the text, we use superscripts ‘SC’ and ‘AL’, respectively. However, for simplicity, these superscripts maybe omitted in the mathematical derivations and proofs.

Before proceeding to the proofs of the main theorems, we introduce a series of lemmas that will serve as essential building blocks.

Lemma 1. Suppose conditions (A_1) and (B_1)–(B_4) in Appendix A hold. If $\lambda_T \to 0$, $\sqrt{T/p_T}\lambda_T \to \infty$, and $p_T^5/T \to 0$ as $T \to \infty$, then with probability approaching 1, for any $\boldsymbol{\beta}_{T1}$ such that $\|\boldsymbol{\beta}_{T1} - \boldsymbol{\beta}^*_{T1}\| = O_p(\sqrt{p_T/T})$, and any $\mathbf{q}_u$ satisfying $\|\mathbf{q}_u - \mathbf{q}^*_u\| = O_p(1/\sqrt{T})$, the following holds for any fixed constant C:

$$Q_T^{SC}\left((\boldsymbol{\beta}'_{T1}, \mathbf{0}')', \mathbf{q}_u\right) = \min_{\|\boldsymbol{\beta}_{T2}\| \le C\left(\frac{p_T}{T}\right)^{1/2}} Q_T^{SC}\left((\boldsymbol{\beta}'_{T1}, \boldsymbol{\beta}'_{T2})', \mathbf{q}_u\right).$$

Proof. Let $\epsilon_T = C\sqrt{p_T/T}$ for some constant $C > 0$. It suffices to show that, with probability tending to 1 as $T \to \infty$, for any $\boldsymbol{\beta}_{T1}$ such that $\boldsymbol{\beta}_{T1} - \boldsymbol{\beta}^*_{T1} = O_p(\sqrt{p_T/T})$, the following inequalities hold for each $j = s_T + 1, \dots, p_T$:

$$\frac{\partial Q_T^{SC}(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}} < 0, \quad \text{for } 0 < \beta_{Tj} < \epsilon_T, \tag{L.1}$$

$$\frac{\partial Q_T^{SC}(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}} > 0, \quad \text{for } -\epsilon_T < \beta_{Tj} < 0. \tag{L.2}$$

By applying Taylor's expansion, we obtain

$$\begin{aligned}
\frac{\partial Q_T^{SC}(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}} &= \frac{\partial L_T(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}} + T p_{\lambda_T}^{'}(|\beta_{Tj}|)\,\mathrm{sgn}(\beta_{Tj}) \\
&= \frac{\partial L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj}} + \sum_{l=1}^{p_T} \frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} (\beta_{Tl} - \beta_{Tl}^*) \\
&\quad + \sum_{l,k=1}^{p_T} \frac{\partial^3 L_T(\tilde{\boldsymbol{\theta}}_T)}{\partial \beta_{Tj} \partial \beta_{Tl} \partial \beta_{Tk}} (\beta_{Tl} - \beta_{Tl}^*)(\beta_{Tk} - \beta_{Tk}^*) \\
&\quad + T p_{\lambda_T}^{'}(|\beta_{Tj}|)\,\mathrm{sgn}(\beta_{Tj}) \\
&\triangleq I_1 + I_2 + I_3 + I_4,
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_T$ lies between $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}_T^*$. In the following, we analyze the contributions of I_1, I_2, and I_3.

By a standard argument, we have

$$I_1 = O_p(\sqrt{T}) = O_p(\sqrt{T p_T}). \tag{L.3}$$

The term I_2 can be decomposed as follows:

$$\begin{aligned}
I_2 &= \sum_{l=1}^{p_T} \left\{ \frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} - E\left[\frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} \right] \right\} (\beta_{Tl} - \beta_{Tl}^*) \\
&\quad + \sum_{l=1}^{p_T} E\left[\frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} \right] (\beta_{Tl} - \beta_{Tl}^*) \\
&\triangleq K_1 + K_2.
\end{aligned}$$

For the term K_1, by the Cauchy–Schwarz inequality, we obtain

$$|K_1| \leq \|\boldsymbol{\beta}_T - \boldsymbol{\beta}_T^*\| \left[\sum_{l=1}^{p_T} \left\{ \frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} - E\left(\frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} \right) \right\}^2 \right]^{1/2}.$$

According to condition (B_2) in Appendix A, it is easy to show that

$$\left[\sum_{l=1}^{p_T} \left\{ \frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} - E\left(\frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl}} \right) \right\}^2 \right]^{1/2} = O_p(\sqrt{T p_T}).$$

By $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^*\| = O_p(\sqrt{p_T/T})$, it follows that

$$K_1 = O_p(\sqrt{T p_T}). \tag{L.4}$$

As for term K_2, applying the Cauchy–Schwarz inequality and using the fact that $\|\boldsymbol{\beta}_{T1} - \boldsymbol{\beta}_{T1}^*\| = O_p(\sqrt{p_T/T})$, we have

$$\begin{aligned} |K_2| &= \left| T \sum_{l=1}^{p_T} I_T(\boldsymbol{\beta}_T^*)(j,l)(\beta_{Tl} - \beta_{Tl}^*) \right| \\ &\leq T \cdot O_p\left(\sqrt{\frac{p_T}{T}} \right) \left\{ \sum_{l=1}^{p_T} I_T^2(\boldsymbol{\beta}_T^*)(j,l) \right\}^{1/2}. \end{aligned}$$

Since the eigenvalues of the information matrix $I_T(\boldsymbol{\beta}_T^*)$ are bounded by condition (B_2) in Appendix A, we have

$$\sum_{l=1}^{p_T} I_T^2(\boldsymbol{\beta}_T^*)(j,l) = O_p(1).$$

It follows that

$$K_2 = O_p(\sqrt{T p_T}). \tag{L.5}$$

Combining results from (L.4) and (L.5), we obtain

$$I_2 = O_p(\sqrt{T p_T}). \tag{L.6}$$

Next we consider I_3, which can be written as:

$$\begin{aligned} I_3 &= \sum_{l,k=1}^{p_T} \left\{ \frac{\partial^3 L_T(\tilde{\boldsymbol{\theta}}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl} \partial \beta_{Tk}} - E \frac{\partial^3 L_T(\tilde{\boldsymbol{\theta}}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl} \partial \beta_{Tk}} \right\} (\beta_{Tj} - \beta_{Tj}^*)(\beta_{Tk} - \beta_{Tk}^*) \\ &\quad + \sum_{l,k=1}^{p_T} E \frac{\partial^3 L_T(\tilde{\boldsymbol{\theta}}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tl} \partial \beta_{Tk}} (\beta_{Tj} - \beta_{Tj}^*)(\beta_{Tk} - \beta_{Tk}^*) \\ &\triangleq K_3 + K_4. \end{aligned}$$

For the term K_3, by the Cauchy–Schwarz inequality, we have

$$K_3^2 \leq \left\{ \sum_{l,k=1}^{p_T} \left(\frac{\partial^3 L_T(\tilde{\boldsymbol{\theta}}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tk} \partial \beta_{Tl}} - E \frac{\partial^3 L_T(\tilde{\boldsymbol{\theta}}_T^*)}{\partial \beta_{Tj} \partial \beta_{Tk} \partial \beta_{Tl}} \right)^2 \right\} \cdot \|\boldsymbol{\beta}_T - \boldsymbol{\beta}_T^*\|^4.$$

Under conditions (B_3) and (B_4) in Appendix A, it follows that

$$K_3 = O_p\left(\left(T p_T^2 \cdot \frac{p_T^2}{T^2} \right)^{1/2} \right) = o_p(\sqrt{T p_T}). \tag{L.7}$$

By condition (B_3) in Appendix A, we also have

$$|K_4| \leq C_5^{1/2} \cdot T p_T \cdot \|\boldsymbol{\beta}_T - \boldsymbol{\beta}_T^*\|^2 = O_p(p_T^2) = o_p(\sqrt{T p_T}). \tag{L.8}$$

From (L.3) and (L.6)–(L.8), we obtain

$$I_1 + I_2 + I_3 = O_p(\sqrt{T p_T}).$$

Since $\sqrt{p_T/T}/\lambda_T \to 0$ and $\liminf_{T\to\infty} \inf_{\theta \to 0^+} p'_{\lambda_T}(\theta)/\lambda_T > 0$, we deduce that

$$\frac{\partial Q_T^{SC}(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}} = T\lambda_T \left\{ \frac{p'_{\lambda_T}(|\beta_{Tj}|)}{\lambda_T} \operatorname{sgn}(\beta_{Tj}) + O_p\left(\frac{\sqrt{p_T/T}}{\lambda_T} \right) \right\}.$$

It is then straightforward to see that the sign of β_{Tj} exclusively determines the sign of $\frac{\partial Q_T^{SC}(\boldsymbol{\theta}_T)}{\partial \beta_{Tj}}$. Consequently, inequalities (L.1) and (L.2) hold, completing the proof. □

Lemma 2. Under the conditions of Theorem 1, we have

$$\left\| \frac{1}{T}\nabla^2 L_T(\boldsymbol{\beta}_T^*) + I_T(\boldsymbol{\beta}_T^*) \right\| = o_p\left(\frac{1}{p_T}\right), \tag{L.9}$$

and

$$\left\| \left\{ \frac{1}{T}\sum_{i=1}^{T} \frac{\partial L_{Ti}(\boldsymbol{\theta}_{T1}^*)}{\partial \beta_{Tj}} \frac{\partial L_{Ti}(\boldsymbol{\theta}_{T1}^*)}{\partial \beta_{Tk}} \right\} - I_T(\boldsymbol{\beta}_T^*) \right\| = o_p\left(\frac{1}{p_T}\right). \tag{L.10}$$

Proof. For any $\epsilon > 0$, by Chebyshev's inequality, we obtain

$$\begin{aligned} & P\left(\left\| \frac{1}{T}\nabla^2 L_T(\boldsymbol{\beta}_T^*) + I_T(\boldsymbol{\beta}_T^*) \right\| \geq \frac{\epsilon}{p_T} \right) \\ & \leq \frac{p_T^2}{T^2\epsilon^2} E \sum_{i,j=1}^{p_T} \left\{ \frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Ti}\partial \beta_{Tj}} - E\left[\frac{\partial^2 L_T(\boldsymbol{\theta}_T^*)}{\partial \beta_{Ti}\partial \beta_{Tj}} \right] \right\}^2 \\ & = \frac{p_T^4}{T} = o(1). \end{aligned}$$

Therefore, (L.9) holds. The result in (L.10) can be established similarly by applying the same argument to the empirical covariance expression. □

Proof of Theorem 1.

Proof. Let $\alpha_T = \sqrt{p_T}(T^{-1/2} + a_T)$, $\mathbf{u}_T = \alpha_T^{-1}(\boldsymbol{\beta}_T - \boldsymbol{\beta}_T^*)$, $\mathbf{v} = \alpha_T^{-1}(\mathbf{q}_u - \mathbf{q}_u^*)$, and define the set $\mathcal{C}_T = \{(\mathbf{u}_T, \mathbf{v}) : \|(\mathbf{u}_T^{'}, \mathbf{v}^{'})'\| = C\}$, where $\|\cdot\|$ denotes the ℓ_2-norm.

We aim to show that, for any $\delta > 0$, there exists a sufficiently large constant C such that for all sufficiently large T, the following inequality holds:

$$P\left\{ \inf_{(\mathbf{u}_T, \mathbf{v}) \in \mathcal{C}_T} Q_T^{SC}(\boldsymbol{\beta}_T^* + \alpha_T\mathbf{u}_T, \mathbf{q}_u^* + \alpha_T\mathbf{v}) > Q_T^{SC}(\boldsymbol{\beta}_T^*, \mathbf{q}_u^*) \right\} \geq 1 - \delta. \tag{A.1}$$

This implies that, with probability at least $1-\delta$, there exists a local minimizer $\widehat{\boldsymbol{\beta}}_T$ within the ball $\{(\boldsymbol{\beta}_T^* + \alpha_T\mathbf{u}_T, \mathbf{q}_u^* + \alpha_T\mathbf{v}) : \|(\mathbf{u}_T^{'}, \mathbf{v}^{'})'\| \leq C\}$, such that $\|\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_T^*\| = O_p(\alpha_T)$.

Using the fact that $p_{\lambda_T}(0) = 0$, we have

$$\begin{aligned}
D_T^{SC}(\mathbf{u}_T, \mathbf{v}) &= Q_T^{SC}(\boldsymbol{\beta}_T^* + \alpha_T \mathbf{u}_T, \mathbf{q}_u^* + \alpha_T \mathbf{v}) - Q_T^{SC}(\boldsymbol{\beta}_T^*, \mathbf{q}_u^*) \\
&\geq L_T(\boldsymbol{\beta}_T^* + \alpha_T \mathbf{u}_T, \mathbf{q}_u^* + \alpha_T \mathbf{v}) - L_T(\boldsymbol{\beta}_T^*, \mathbf{q}_u^*) \\
&\quad + T \sum_{j=1}^{s_T} \left\{ p_{\lambda_T}(\beta_{Tj}^* + \alpha_T u_{Tj}) - p_{\lambda_T}(\beta_{Tj}^*) \right\} \\
&\triangleq (I) + (II). \qquad \text{(A.2)}
\end{aligned}$$

For notational simplicity, define $\boldsymbol{\theta}_T = (\boldsymbol{\beta}_T', \mathbf{q}_u')'$ and $\mathbf{w}_T = (\mathbf{u}_T', \mathbf{v}')'$. Then, expression (A.2) can be reformulated as

$$\begin{aligned}
D_T^{SC}(\mathbf{w}_T) &= Q_T^{SC}(\boldsymbol{\theta}_T^* + \alpha_T \mathbf{w}_T) - Q_T^{SC}(\boldsymbol{\theta}_T^*) \\
&\geq L_T(\boldsymbol{\theta}_T^* + \alpha_T \mathbf{w}_T) - L_T(\boldsymbol{\theta}_T^*) \\
&\quad + T \sum_{j=1}^{s_T} \left\{ p_{\lambda_T}(\beta_{Tj}^* + \alpha_T u_{Tj}) - p_{\lambda_T}(\beta_{Tj}^*) \right\} \\
&\triangleq (I) + (II).
\end{aligned}$$

Then, by applying Taylor's expansion and the Mean Value Theorem, we obtain

$$\begin{aligned}
(I) &= \alpha_T \nabla^\top L_T(\boldsymbol{\theta}_T^*)\mathbf{w}_T + \frac{1}{2}\alpha_T^2 \mathbf{w}_T^\top \nabla^2 L_T(\boldsymbol{\theta}_T^*)\mathbf{w}_T \\
&\quad + \frac{1}{6}\alpha_T^3 \nabla^\top \left\{ \mathbf{w}_T^\top \nabla^2 L_T(\tilde{\boldsymbol{\theta}}_T)\mathbf{w}_T \right\} \mathbf{w}_T \\
&\triangleq I_1 + I_2 + I_3,
\end{aligned}$$

where the intermediate vector $\tilde{\boldsymbol{\theta}}_T$ lies between $\boldsymbol{\theta}_T^*$ and $\boldsymbol{\theta}_T^* + \alpha_T \mathbf{w}_T$, and the second term (II) can be expressed as

$$\begin{aligned}
(II) &= \sum_{j=1}^{s_T} \left[T\alpha_T p_{\lambda_T}^{'}(|\beta_{Tj}^*|)\operatorname{sgn}(\beta_{Tj}^*)u_{Tj} + T\alpha_T^2 p_{\lambda_T}^{''}(\beta_{Tj}^*)u_{Tj}^2\{1 + o(1)\} \right] \\
&\triangleq I_4 + I_5,
\end{aligned}$$

where I_4 represents the linear term in u_{Tj} and I_5 corresponds to the second-order term involving u^2_{Tj}.

By condition (F) in Appendix A, we obtain the following bound for I_1

$$\begin{aligned} |I_1| &= \left|\alpha_T \nabla^T L_T(\boldsymbol{\theta}^*_T)\mathbf{w}_T\right| \leq \alpha_T \left\|\nabla^T L_T(\boldsymbol{\theta}^*_T)\right\| \|\mathbf{w}_T\| \\ &= O_p(\alpha_T \sqrt{Tp'_T}) \|\mathbf{w}_T\| = O_p(T\alpha^2_T) \|\mathbf{w}_T\| \, . \end{aligned} \tag{A.3}$$

Next, we consider the term I_2. By Lemma 2, we have

$$\begin{aligned} I_2 &= \frac{1}{2}\mathbf{w}'_T \left[\frac{1}{T}\left\{\nabla^2 L_T(\boldsymbol{\theta}^*_T) - E\nabla^2 L_T(\boldsymbol{\theta}^*_T)\right\}\right] \mathbf{w}_T \cdot T\alpha^2_T \\ &\quad - \frac{1}{2}\mathbf{w}'_T I_T(\boldsymbol{\theta}^*_T)\mathbf{w}_T \cdot T\alpha^2_T \\ &= -\frac{T\alpha^2_T}{2}\mathbf{w}'_T I_T(\boldsymbol{\theta}^*_T)\mathbf{w}_T + o_p(1) \cdot T\alpha^2_T \|\mathbf{w}_T\|^2. \end{aligned} \tag{A.4}$$

By the Cauchy–Schwarz inequality and condition (B_3) in Appendix A, we obtain

$$\begin{aligned} |I_3| &= \left| \frac{1}{6} \sum_{i,j,k=1}^{p'_T} \frac{\partial^3 L_T(\boldsymbol{\theta}^*_T)}{\partial\theta_{Ti}\partial\theta_{Tj}\partial\theta_{Tk}} w_{Ti}w_{Tj}w_{Tk}\alpha^3_T \right| \\ &\leq \frac{1}{6}\sum_{t=1}^{T}\left\{\sum_{i,j,k=1}^{p_T} M^2_{Tijk}(\mathbf{V}_{Tt})\right\}^{1/2} \|\mathbf{w}_T\|^3 \alpha^3_T. \end{aligned}$$

Since $p^4_T/T \to 0$ and $p^2_T a_T \to 0$ as $T \to \infty$, we have

$$\begin{aligned} \frac{1}{6}\sum_{t=1}^{T}\left\{\sum_{i,j,k=1}^{p'_T} M^2_{Tijk}(\mathbf{V}_{Tt})\right\}^{1/2} \|\mathbf{w}_T\|^3 \alpha^3_T &= O_p\left((p'_T)^{3/2}\alpha_T\right) \cdot T\alpha^2_T \|\mathbf{w}_T\|^2 \\ &= o_p(T\alpha^2_T)\|\mathbf{w}_T\|^2. \end{aligned}$$

Therefore,

$$I_3 = o_p(T\alpha^2_T)\|\mathbf{w}_T\|^2. \tag{A.5}$$

The remaining terms I_4 and I_5 can be handled as follows. For I_4, we apply the

Cauchy–Schwarz inequality and the definition of a_T:

$$\begin{aligned} |I_4| &\le \sum_{j=1}^{s_T} \left| T\alpha_T p_{\lambda_T}^{'}(|\beta_{Tj}^*|)\,\mathrm{sgn}(\beta_{Tj}^*)u_{Tj} \right| \\ &\le T\alpha_T \cdot \max_{1\le j\le s_T} p_{\lambda_T}^{'}(|\beta_{Tj}^*|) \cdot \sum_{j=1}^{s_T} |u_{Tj}| \\ &\le T\alpha_T a_T \cdot \sqrt{s_T}\|\mathbf{u}_T\| \le T\alpha_T^2\|\mathbf{u}_T\|. \end{aligned} \tag{A.6}$$

For I_5, using the uniform boundedness of the second derivative of the penalty function, we have:

$$\begin{aligned} I_5 &= \sum_{j=1}^{s_T} T\alpha_T^2 p_{\lambda_T}^{''}(\beta_{Tj}^*)u_{Tj}^2(1+o(1)) \\ &\le 2 \cdot \max_{1\le j\le s_T} p_{\lambda_T}^{''}(\beta_{Tj}^*) \cdot T\alpha_T^2\|\mathbf{u}_T\|^2. \end{aligned} \tag{A.7}$$

It follows from equations (A.3)–(A.7) and condition (A_3) in Appendix A that the terms I_1, I_3, I_4, and I_5 are all asymptotically negligible compared to the dominant negative quadratic term I_2, provided that $\|\mathbf{u}_T\|$ and $\|\mathbf{v}\|$ are sufficiently large. Consequently, inequality (A.1) holds, which completes the proof. □

To prove Theorem 2, we leverage the sparsity property of the non-concave penalized estimator, namely that $\widehat{\boldsymbol{\beta}}_{T2} = 0$, as established in Lemma 1. This allows us to focus the asymptotic analysis on the subvector $\widehat{\boldsymbol{\beta}}_{T1}$.

Proof of Theorem 2.

Proof. As established in Theorem 1, there exists a root-(T/p_T)-consistent local minimizer $\widehat{\boldsymbol{\beta}}_T$ of $Q_T^{SC}(\boldsymbol{\beta}_T, \mathbf{q}_u)$. According to Lemma 1, part (i), this estimator satisfies the sparsity property and can be written in the form $(\widehat{\boldsymbol{\beta}}_{T1}^{'}, \mathbf{0}^{'})^{'}$. Therefore, to complete the proof of Theorem 2, it suffices to verify part (ii), which establishes the asymptotic normality of the penalized non-concave estimator $\widehat{\boldsymbol{\beta}}_{T1}$.

If we can show that

$$\{I_T(\boldsymbol{\beta}^*_{T1}) + \boldsymbol{\Sigma}_{\lambda_T}\}\,(\widehat{\boldsymbol{\beta}}_{T1} - \boldsymbol{\beta}^*_{T1}) + \mathbf{b}_T = \frac{1}{T}\nabla L_T(\boldsymbol{\beta}^*_{T1}) + o_p(T^{-1/2}),$$

then it follows that

$$\begin{aligned}&\sqrt{T}\,\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}^*_{T1})\,\{I_T(\boldsymbol{\beta}^*_{T1}) + \boldsymbol{\Sigma}_{\lambda_T}\}\left[\widehat{\boldsymbol{\beta}}_{T1} - \boldsymbol{\beta}^*_{T1} + \{I_T(\boldsymbol{\beta}^*_{T1}) + \boldsymbol{\Sigma}_{\lambda_T}\}^{-1}\,\mathbf{b}_T\right]\\ &= \frac{1}{\sqrt{T}}\,\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}^*_{T1})\nabla L_T(\boldsymbol{\beta}^*_{T1}) + o_p\left(\mathbf{A}_T I_T^{-1/2}(\boldsymbol{\beta}^*_{T1})\right).\end{aligned}$$

Here, $\mathbf{C}_T$ is an $s_T \times s_T$ matrix such that $\mathbf{C}_T\mathbf{C}_T^{'} \to \mathbf{G}$, where $\mathbf{G}$ is an $s_T \times s_T$ symmetric non-negative definite matrix.

By the conditions of Theorem 2, the last term is $o_p(1)$. Let

$$\mathbf{Y}_{Ti} = \frac{1}{\sqrt{T}}\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}^*_{T1})\nabla L_{Ti}(\boldsymbol{\beta}^*_{T1}), \quad i = 1, 2, \ldots, T.$$

Then, for any $\epsilon > 0$, we have

$$\begin{aligned}\sum_{i=1}^{T} E\|\mathbf{Y}_{Ti}\|^2 \mathbf{1}\{\|\mathbf{Y}_{Ti}\| > \epsilon\} &= TE\|\mathbf{Y}_{T1}\|^2\mathbf{1}\{\|\mathbf{Y}_{T1}\| > \epsilon\}\\ &\leq T\left\{E\|\mathbf{Y}_{T1}\|^4\right\}^{1/2}\{P(\|\mathbf{Y}_{T1}\| > \epsilon)\}^{1/2}.\end{aligned}$$

Using condition (B_2) in Appendix A and the fact that $\mathbf{C}_T\mathbf{C}_T^{'} \to \mathbf{G}$, we obtain

$$P(\|\mathbf{Y}_{T1}\| > \epsilon) \leq \frac{E\|\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}^*_{T1})\nabla L_{T1}(\boldsymbol{\beta}^*_{T1})\|^2}{T\epsilon^2} = O(T^{-1}),$$

and

$$\begin{aligned}E\|\mathbf{Y}_{T1}\|^4 &= \frac{1}{T^2}E\|\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}^*_{T1})\nabla L_{T1}(\boldsymbol{\beta}^*_{T1})\|^4\\ &\leq \frac{1}{T^2}\lambda_{\max}(\mathbf{C}_T\mathbf{C}_T^{'})\lambda_{\max}\{I_n(\boldsymbol{\beta}^*_{T1})\}E\|\nabla^T L_{T1}(\boldsymbol{\beta}^*_{T1})\nabla L_{T1}(\boldsymbol{\beta}^*_{T1})\|^2\\ &= O\left(\frac{p_T^2}{T^2}\right).\end{aligned}$$

Therefore, we conclude

$$\sum_{i=1}^{T} E\|\mathbf{Y}_{Ti}\|^2 \mathbf{1}\{\|\mathbf{Y}_{Ti}\| > \epsilon\} = O\left(T \cdot \frac{p_T}{T} \cdot \frac{1}{\sqrt{T}}\right) = o(1).$$

On the other hand, since $\mathbf{C}_T\mathbf{C}_T^{'} \to \mathbf{G}$, we have

$$\sum_{i=1}^{T} \mathrm{cov}(\mathbf{Y}_{Ti}) = T \cdot \mathrm{cov}(\mathbf{Y}_{T1}) = \mathrm{cov}\left\{\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}_{T1}^*)\nabla L_{T1}(\boldsymbol{\beta}_{T1}^*)\right\} \to \mathbf{G}.$$

Hence, the sequence $\mathbf{Y}_{Ti}$ satisfies the Lindeberg–Feller central limit theorem (see Van Der Vaart (1998) (83)). This implies that

$$\frac{1}{\sqrt{T}}\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}_{T1}^*)\nabla L_T(\boldsymbol{\beta}_{T1}^*)$$

converges in distribution to a multivariate normal distribution.

With a slight abuse of notation, let $Q_T^{SC}(\boldsymbol{\beta}_{T1}) = Q_T^{SC}((\boldsymbol{\beta}_{T1}^{'}, \mathbf{0}^{'})^{'}, \mathbf{q}_u)$. Since the estimator $\widehat{\boldsymbol{\beta}}_{T1}$ satisfies the penalized estimating equation $\nabla Q_T^{SC}(\widehat{\boldsymbol{\beta}}_{T1}) = 0$, we apply Taylor's expansion of $\nabla Q_T^{SC}(\widehat{\boldsymbol{\beta}}_{T1})$ around the true value $\boldsymbol{\beta}_{T1}^*$ to obtain

$$\begin{aligned}
&\frac{1}{T}\left[\left\{\nabla^2 L_T(\boldsymbol{\beta}_{T1}^*) - \nabla^2 P_{\lambda_T}(\tilde{\tilde{\boldsymbol{\beta}}}_{T1})\right\}(\widehat{\boldsymbol{\beta}}_{T1} - \boldsymbol{\beta}_{T1}^*) - \nabla P_{\lambda_T}(\boldsymbol{\beta}_{T1}^*)\right] \\
&= -\frac{1}{T}\left[\nabla L_T(\boldsymbol{\beta}_{T1}^*) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_{T1} - \boldsymbol{\beta}_{T1}^*)^{'}\nabla^2\left\{\nabla L_T(\tilde{\boldsymbol{\beta}}_{T1})\right\}(\widehat{\boldsymbol{\beta}}_{T1} - \boldsymbol{\beta}_{T1}^*)\right],
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{T1}$ and $\tilde{\tilde{\boldsymbol{\beta}}}_{T1}$ lie between $\widehat{\boldsymbol{\beta}}_{T1}$ and $\boldsymbol{\beta}_{T1}^*$.

For convenience, define

$$\begin{aligned}
\mathcal{L} &\triangleq \nabla^2 L_T(\boldsymbol{\beta}_{T1}^*) - \nabla^2 P_{\lambda_T}(\tilde{\tilde{\boldsymbol{\beta}}}_{T1}), \\
\mathcal{C} &\triangleq \frac{1}{2}(\widehat{\boldsymbol{\beta}}_{T1} - \boldsymbol{\beta}_{T1}^*)^{'}\nabla^2\left\{\nabla L_T(\tilde{\boldsymbol{\beta}}_{T1})\right\}(\widehat{\boldsymbol{\beta}}_{T1} - \boldsymbol{\beta}_{T1}^*).
\end{aligned}$$

Under conditions (B_3) and (B_4) in Appendix A, and by the Cauchy–Schwarz inequality, we obtain the following bound for the remainder term $\mathcal{C}$

$$\begin{aligned}\left\|\frac{1}{T}\mathcal{C}\right\|^2 &\le \frac{1}{T^2}\sum_{t=1}^{T} T\|\widehat{\boldsymbol{\beta}}_{T1}-\boldsymbol{\beta}^*_{T1}\|^4 \sum_{j,k,l=1}^{s_T} M^2_{Tjkl}(\mathbf{V}_{Tt})\\ &= O_p\left(\frac{p_T^2}{T^2}\right)\cdot O_p(p_T^3)\\ &= o_p\left(\frac{1}{T}\right). \qquad \text{(A.8)}\end{aligned}$$

At the same time, by Lemma 2 and condition (B_4) in Appendix A, it is easy to show that

$$\lambda_i\left\{\frac{1}{T}\mathcal{L}+I_T(\boldsymbol{\beta}^*_{T1})+\boldsymbol{\Sigma}_{\lambda_T}\right\} = o_p\left(\frac{1}{\sqrt{p_T}}\right), \quad \text{for } i=1,2,\ldots,s_T,$$

where $\lambda_i(\mathbf{A})$ denotes the ith eigenvalue of a symmetric matrix $\mathbf{A}$.

Since $\widehat{\boldsymbol{\beta}}_{T1}-\boldsymbol{\beta}^*_{T1} = O_p\left(\sqrt{p_T/T}\right)$, we then have

$$\left\{\frac{1}{T}\mathcal{L}+I_T(\boldsymbol{\beta}^*_{T1})+\boldsymbol{\Sigma}_{\lambda_T}\right\}(\widehat{\boldsymbol{\beta}}_{T1}-\boldsymbol{\beta}^*_{T1}) = o_p\left(\frac{1}{\sqrt{T}}\right). \qquad \text{(A.9)}$$

Finally, combining (A.8) and (A.9), we obtain

$$\{I_T(\boldsymbol{\beta}^*_{T1})+\boldsymbol{\Sigma}_{\lambda_T}\}(\widehat{\boldsymbol{\beta}}_{T1}-\boldsymbol{\beta}^*_{T1})+\mathbf{b}_T = \frac{1}{T}\nabla L_T(\boldsymbol{\beta}^*_{T1})+o_p\left(\frac{1}{\sqrt{T}}\right). \qquad \text{(A.10)}$$

Following (A.10), Theorem 2 follows.

Since $\boldsymbol{\beta}_T = \operatorname{vec}(\mathbf{A}'_T)$, it is necessary to transform the asymptotic normality result from $\operatorname{vec}(\mathbf{A}'_T)$ to $\operatorname{vec}(\mathbf{A}_T)$. This transformation can be accomplished by introducing a permutation matrix $\mathbf{P}$ such that $\boldsymbol{\beta}_T = \mathbf{P}\operatorname{vec}(\mathbf{A}_T)$. The asymptotic results then hold under this transformation, ensuring consistency in the original parameter space.

The original asymptotic normality result is given by

$$\begin{aligned}&\sqrt{T}\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}^*_{T1})\{I_T(\boldsymbol{\beta}^*_{T1})+\boldsymbol{\Sigma}_{\lambda_T}\}\\ &\qquad\times[\widehat{\boldsymbol{\beta}}_{T1}-\boldsymbol{\beta}^*_{T1}+\{I_T(\boldsymbol{\beta}^*_{T1})+\boldsymbol{\Sigma}_{\lambda_T}\}^{-1}\mathbf{b}_T] \xrightarrow{D} \mathcal{N}(0,\mathbf{G}).\end{aligned}$$

where $\boldsymbol{\beta}_T = \text{vec}(\mathbf{A}_T^{'})$. To express this result in terms of $\text{vec}(\mathbf{A}_T)$, we apply a permutation matrix $\mathbf{P}$ such that

$$\text{vec}(\mathbf{A}_T^{'}) = \mathbf{P}\,\text{vec}(\mathbf{A}_T),$$

where $\mathbf{P}$ is a $s_T \times s_T$ permutation matrix. Similarly, the corresponding penalty gradient vector satisfies $\mathbf{b}_T = \mathbf{P}\tilde{\mathbf{b}}_T$. Substituting into the asymptotic normality result gives

$$\sqrt{T}\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}_{T1}^*)\{I_T(\boldsymbol{\beta}_{T1}^*) + \boldsymbol{\Sigma}_{\lambda_T}\} \\ \times \mathbf{P}[\text{vec}(\widehat{\mathbf{A}}_{T1}^*) - \text{vec}(\mathbf{A}_{T1}^*) + \{I_T(\boldsymbol{\beta}_{T1}^*) + \boldsymbol{\Sigma}_{\lambda_T}\}^{-1}\tilde{\mathbf{b}}_T] \xrightarrow{D} \mathcal{N}(0, \mathbf{G}).$$

Since $\mathbf{P}$ is an orthogonal matrix that preserves asymptotic distributional properties, the result can be rewritten as

$$\sqrt{T}\mathbf{C}_T I_T^{-1/2}(\boldsymbol{\beta}_{T1}^*)\{I_T(\boldsymbol{\beta}_{T1}^*) + \boldsymbol{\Sigma}_{\lambda_T}\} \\ \times [\text{vec}(\widehat{\mathbf{A}}_{T1}^*) - \text{vec}(\mathbf{A}_{T1}^*) + \{I_T(\boldsymbol{\beta}_{T1}^*) + \boldsymbol{\Sigma}_{\lambda_T}\}^{-1}\tilde{\mathbf{b}}_T] \xrightarrow{D} \mathcal{N}(0, \mathbf{P}^{'}\mathbf{G}).$$

Letting $\boldsymbol{\Sigma} = \mathbf{P}^{'}\mathbf{G}$ and $\mathbf{D}_{T1} = I_T^{-1/2}(\boldsymbol{\beta}_{T1}^*)$, the final form becomes

$$\sqrt{T}\,\mathbf{C}_T\mathbf{D}_{T1}^{-1/2}\,\{\mathbf{D}_{T1} + \boldsymbol{\Sigma}_{\lambda_T}\}\left[\text{vec}(\widehat{\mathbf{A}}_{T1}^*) - \text{vec}(\mathbf{A}_{T1}^*) + \{\mathbf{D}_{T1} + \boldsymbol{\Sigma}_{\lambda_T}\}^{-1}\,\tilde{\mathbf{b}}_T\right] \xrightarrow{D} \mathcal{N}(0, \boldsymbol{\Sigma}).$$

□

Fact 1. $\mathbf{P}$ is a permutation matrix that swaps the elements of $\text{vec}(\mathbf{A}_T)$ to obtain $\text{vec}(\mathbf{A}_T^{'})$. It satisfies:

1. $\mathbf{P}$ is an $s_T \times s_T$ square matrix;

2. $\mathbf{P}$ consist of only $0's$ and $1's$;

3. Each row and each column has exactly one 1;

4. $\mathbf{P}$ is orthogonal, meaning $\mathbf{P}\mathbf{P}^{'} = \mathbf{I}$

A standard way to define $\mathbf{P}$ is using the commutation matrix $\mathbf{K}_{m,n}$, which is an $mn \times mn$ matrix satisfying:

$$\mathbf{K}_{m,n}\text{vec}(\mathbf{A}_T) = \text{vec}(\mathbf{A}_T^{'})$$

The explicit construction is:

$$\mathbf{P} = \mathbf{K}_{s,r} = \sum_{i=1}^{s}\sum_{j=1}^{r}\mathbf{E}_{ij} \otimes \mathbf{E}_{ji},$$

where $\mathbf{E}_{ij}$ is the elementary basis matrix, which has a 1 at the (i, j) position and 0 elsewhere.

Alternatively, in index notation, $\mathbf{P}$ rearranges the vector such that:

$$\mathbf{P}\mathbf{e}_k = \mathbf{e}_{\pi(k)},$$

where $\pi(k)$ is the index mapping for the transpose operation.

APPENDIX B: CONDITIONS and PROOFS of THEOREMS IN CHAPTER 3

B.1 Regularity conditions

For any $\mathbf{u} \in \mathcal{B}^N = \{\mathbf{u}|\mathbf{u} \in \mathbb{R}^N, \|\mathbf{u}\| < 1\}$, model (1.1) can be also written as

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{B.1}$$

where $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_P)$, and $\mathbf{x}_t = (\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-P})'$. Then, we define the $\mathbf{u}$th spatial QR estimators of the parameters for model (A.1) by

$$\widehat{\mathbf{A}},\, \widehat{\mathbf{q}}_u = \arg \min_{\mathbf{A},\, \mathbf{q}_u} \sum_{t=p+1}^{T} Q_u \left(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t - \mathbf{q}_u\right) \tag{B.2}$$

For simplicity, we let $\mathbf{Z}_t = (1, \mathbf{x}'_t)' \in \mathbb{R}^{NP+1}$,, and $\boldsymbol{\Phi}(\mathbf{u}) = (\boldsymbol{\Phi}_1(\mathbf{u}), \boldsymbol{\Phi}_2(\mathbf{u})) \in \mathbb{R}^{N\times(NP+1)}$, where $\boldsymbol{\Phi}_1(\mathbf{u}) = \mathbf{q}_u$ and $\boldsymbol{\Phi}_2(\mathbf{u}) = \mathbf{A}$. We define the $\mathbf{u}$-th quantile of $\boldsymbol{\varepsilon}_t$ as

$$\mathbf{q}_u = \arg \min_{\mathbf{q}_u \,\in\, \mathbb{R}^k} \mathbb{E}\left[Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) - Q_u(\boldsymbol{\varepsilon}_t)\right] \tag{B.3}$$

Then

$$E[\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)] = 0. \tag{B.4}$$

where for any $\mathbf{t} \in \mathbb{R}^N$, define $\boldsymbol{\varphi}_{\mathbf{u}}(\mathbf{t}) = \mathbf{t}/\|\mathbf{t}\| + \mathbf{u}$ for $\mathbf{t} \neq 0$, and $\boldsymbol{\varphi}_{\mathbf{u}}(\mathbf{0}) = \mathbf{u}$. Since the first entry in $\mathbf{Z}_t$ is one, the model (B.2) is equivalent to

$$\widehat{\boldsymbol{\Phi}}_T(\mathbf{u}) = \arg \min_{\boldsymbol{\Phi}(\mathbf{u})} \sum_{t=p+1}^{T} Q_u \left(\mathbf{y}_t - \boldsymbol{\Phi}(\mathbf{u})\mathbf{Z}_t\right) \tag{B.5}$$

Then $\widehat{\boldsymbol{\Phi}}_T(\mathbf{u})\mathbf{Z}_t$ is the spatial QR estimator of the $\mathbf{u}$-th conditional quantile of $\mathbf{y}_t$ given $\mathbf{Z}_t$, and $\widehat{\boldsymbol{\Phi}}_T(\mathbf{u})$ is the spatial QR estimator of $\boldsymbol{\Phi}(\mathbf{u})$, which is denoted by $\widehat{\boldsymbol{\Phi}}_T$ to stress dependent on $\mathbf{u}$.

For convenience, we use the following notations throughout the proofs. Let $\boldsymbol{\Psi}(\mathbf{t})$ be the $N \times N$ Hessian matrix, i.e. $\boldsymbol{\Psi}(\mathbf{t}) = \|\mathbf{t}\|^{-1}(\mathbf{I}_N - \mathbf{t}\mathbf{t}^\top\|\mathbf{t}\|^{-2})$ for $\mathbf{t} \neq 0$, and $\boldsymbol{\Psi}(\mathbf{0}) = \mathbf{0}$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Let $\mathbf{D}_1(\mathbf{u}) = E[\boldsymbol{\Psi}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)]$ and $\mathbf{D}_2(\mathbf{u}) = E\left[\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\{\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\}^\top\right]$.

To derive asymptotic distributions of the spatial QR estimators, we need the following regularity conditions:

C_1. $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathbf{Z}_t\mathbf{Z}_t^\top = \mathbf{S}$, where $\mathbf{S}$ is a positive definite matrix.

C_2. The processes $\{\boldsymbol{\varepsilon}_t\}$ are strictly stationary with α-mixing coefficients $\alpha(k)$ such that $\sum_k \alpha(k)^{1-2/\delta} < \infty$ for some $\delta > 2$ and $c > 1 - 2/\delta$.

C_3. $E[\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the σ-field generated by $(\mathbf{y}_{j+1}, \boldsymbol{\varepsilon}_j)$ for $j \le t-1$.

C_4. There exists a positive $\gamma > 0$ such that $E\|\boldsymbol{\varepsilon}_t\|^{2+\gamma} < \infty$.

C_5. $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^\top = \boldsymbol{\Gamma}^*$, where $\boldsymbol{\Gamma}^*$ is a positive definite matrix.

B.2 Proofs of Theorems

Before proceeding to the proofs of the main theorems, we introduce a series of lemmas that will serve as essential building blocks. We first introduce Lemma 3, which is from Chakraborty (2003) (20).

Lemma 3. Let $\boldsymbol{\phi}_n(\boldsymbol{\beta})$, $n = 1, 2, \ldots,$ be a sequence of random functions on $\mathbb{R}^k$ and convex in $\boldsymbol{\beta}$. Let $\boldsymbol{\phi}(\boldsymbol{\beta})$ be a random function such that, for each fixed $\boldsymbol{\beta}$, $\boldsymbol{\phi}_n(\boldsymbol{\beta}) \to \boldsymbol{\phi}(\boldsymbol{\beta})$ in probability. Then for each $M > 0$,

$$\sup_{\|\boldsymbol{\beta}\|\le M} |\boldsymbol{\phi}_n(\boldsymbol{\beta}) - \boldsymbol{\phi}(\boldsymbol{\beta})| \to 0$$

in probability.

The above Lemma 3 follows directly from Lemma 3 of Niemiro (1992) (84).

Lemma 4. For any quadratic function $g(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, where $\mathbf{A}$ is a $p \times p$ positive definite matrix, $\mathbf{b}$ is a $p \times 1$ vector, and c is a constant, we have

(i) $g(\mathbf{x})$ achieves the minimum value $g(\mathbf{x}_0) = c - \frac{1}{4}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$ at $\mathbf{x}_0 = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$;

(ii) $g(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_0)^T\mathbf{A}(\mathbf{x} - \mathbf{x}_0) + g(\mathbf{x}_0)$.

Proof. Routine. □

Lemma 5. Assume that conditions C_1-C_4 hold. Then for any $\mathbf{u} \in \mathcal{B}^N = \{\mathbf{u}|\mathbf{u} \in \mathbb{R}^N, \|\mathbf{u}\| < 1\}$, we have the following Bahadur representation of $\widehat{\mathbf{\Phi}}_T(\mathbf{u})$

$$\sqrt{T}(\widehat{\mathbf{\Phi}}_T(\mathbf{u}) - \mathbf{\Phi}^*(\mathbf{u})) = T^{-1/2}\mathbf{D}_1^{-1}(\mathbf{u}) \left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\mathbf{Z}_t^{'} \right] \mathbf{S}^{-1} + o_p(1). \tag{B.6}$$

Furthermore, $\sqrt{T}(\widehat{\mathbf{\Phi}}_T(\mathbf{u}) - \mathbf{\Phi}^*(\mathbf{u}))$ converges weakly to a $N \times (NP+1)$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma}_{SQR}^* = \left[\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})\right] \otimes \mathbf{S}^{-1}$, where $\mathbf{\Phi}^*(\mathbf{u}) = (\mathbf{\Phi}_1^*(\mathbf{u}), \mathbf{\Phi}_2^*(\mathbf{u}))$ is a $N \times (NP+1)$ matrix. The proof follows the Theorem 4.1 of Chakraborty (2003) (20).

Proof. Fix $\boldsymbol{\gamma} \in \mathbb{R}^{N\times(NP+1)}$ and define the random variables $\mathbf{V}_{Tt}$ as

$$\begin{aligned} \mathbf{V}_{Tt} =& Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) - Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \\ &+ T^{-1/2}[\text{vec}(\boldsymbol{\gamma})]'[\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t]. \end{aligned} \tag{B.7}$$

Since $\mathbf{V}_{T1}, \ldots, \mathbf{V}_{TT}$ are independent, it holds that $\text{Var}\left(\sum_{t=1}^{T} \mathbf{V}_{Tt}\right) \le \sum_{t=1}^{T} E(\mathbf{V}_{Tt}^2)$. Using the convexity of $Q_u(\cdot)$ in $\boldsymbol{\gamma}$, we obtain

$$\begin{aligned} \sum_{t=1}^{T} E(V_{Tt}^2) \le \frac{1}{T} \sum_{t=1}^{T} & E([\text{vec}(\boldsymbol{\gamma})]'\{[\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - n^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) \\ &- \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)] \otimes \mathbf{Z}_t\})^2. \end{aligned} \tag{B.8}$$

The random variable $E(\{\text{vec}(\boldsymbol{\gamma})\}'\{[\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) - \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)] \otimes \mathbf{Z}_t\})$ tends monotonically to zero almost surely. Since $\boldsymbol{\varphi}_{\mathbf{u}}$ is bounded and continuous, the Lebesgue-dominated convergence theorem ensures that the right-hand side of (B.8) tends to zero as $T \to \infty$. Thus, by Chebyshev's inequality,

$$\sum_{t=1}^{T} \mathbf{V}_{Tt} - \sum_{t=1}^{T} E\left[Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) - Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\right] \xrightarrow{P} 0. \tag{B.9}$$

By a Taylor expansion and under condition C_1 in Appendix B, it follows that

$$\sum_{t=1}^{T} E\left[Q_u\left(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t\right) - Q_u\left(\boldsymbol{\varepsilon}_t - \mathbf{q}_u\right)\right] \\ \to \frac{1}{2}\left[\text{vec}(\boldsymbol{\gamma})\right]' \left[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}\right]\left[\text{vec}(\boldsymbol{\gamma})\right]. \qquad \text{(B.10)}$$

Uniform convergence over compact sets follows from Lemma 3. Hence, for any $\varepsilon > 0$ and $M > 0$, there exists sufficiently large T such that with probability at least $1 - \varepsilon/2$,

$$\sup_{\|\boldsymbol{\gamma}\| \le M} \Bigg| \sum_{t=1}^{T} Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) - \sum_{t=1}^{T} Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \\ + T^{-1/2}[\text{vec}(\boldsymbol{\gamma})]' \left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right] \\ - \frac{1}{2}[\text{vec}(\boldsymbol{\gamma})]'[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}][\text{vec}(\boldsymbol{\gamma})] \Bigg| < \varepsilon. \qquad \text{(B.11)}$$

Since, the standardized sums $T^{-1/2}[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}]^{-1}\left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right]$ are bounded in probability, we can select M such that

$$\left\| T^{-1/2}[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}]^{-1}\left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right] \right\| \le M - 1, \qquad \text{(B.12)}$$

with probability exceeding $1 - \varepsilon/2$, too.

Let $K = 2\left(\inf_{\|\mathbf{w}\|=1} \mathbf{w}^T[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}]\mathbf{w}\right)^{-1/2}$. The quadratic function

$$-T^{-1/2}[\text{vec}(\boldsymbol{\gamma})]' \left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right] \\ + \frac{1}{2}[\text{vec}(\boldsymbol{\gamma})]'[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}][\text{vec}(\boldsymbol{\gamma})].$$

has its minimum value equal to

$$-\frac{1}{2T}\left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right]^T [\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}]^{-1}\left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right]$$

at $T^{-1/2}[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}]^{-1}\left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\epsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right]$. Whenever (B.11) and (B.12) hold, the convex function $\sum_{t=1}^{T} Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) - \sum_{t=1}^{T} Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)$ assumes at $T^{-1/2}[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}]^{-1} \left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right]$ a value less than its value on the sphere

$$\left\| \operatorname{vec}(\boldsymbol{\gamma}) - T^{-1/2}[\mathbf{D}_1(\mathbf{u}) \otimes \mathbf{S}]^{-1} \left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \otimes \mathbf{Z}_t\right] \right\| = K\varepsilon^{1/2}.$$

The minimum point of this function is $\sqrt{T}(\hat{\boldsymbol{\Phi}}_T(\mathbf{u}) - \boldsymbol{\Phi}^*(\mathbf{u}))$, so except for an event of probability ε, we have

$$\begin{aligned} \Big\| \sqrt{T}(\widehat{\boldsymbol{\Phi}}_T(\mathbf{u}) - \boldsymbol{\Phi}^*(\mathbf{u})) - T^{-1/2}[\mathbf{D}_1(\mathbf{u})]^{-1} \\ \times \left[\sum_{t=1}^{T} \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\mathbf{Z}_t^{'}\right] \mathbf{S}^{-1} \Big\| < K\sqrt{\varepsilon}. \end{aligned} \tag{B.13}$$

Define $\mathbf{V}_t = \boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\epsilon}_t - \mathbf{q}_u)\mathbf{Z}_t^{'}$. Since function $\boldsymbol{\varphi}$ is bounded and smooth, $\mathbf{Z}_t$ has mean zero. Hence, we know $\{\mathbf{V}_t\}; t = 1, \ldots, T$, is also a strictly stationary and α-mixing process. Since $\boldsymbol{\epsilon}_t$ and $\mathbf{Z}_t$ are uncorrelated, we have $E(\mathbf{V}_t) = E[E(\mathbf{V}_t|\mathbf{Z}_t)] = \mathbf{0}$.

In the following we show the asymptotic normality of $\mathbf{M}_T = \sum_{t=1}^{T} \mathbf{V}_t$, following the Theorem 1 of Jiang (2017) (17). Since $\{\mathbf{V}_t, t = 1, \ldots, T\}$ are dependent, we employ the standard small-block and large-block arguments to complete this task. To this end, we partition the set $\{1, \ldots, T\}$ into $2k_T + 1$ subsets with large blocks of size l_T and small blocks of size s_T. A large block is followed by a small block, and the last remaining set has size $T - k_T(l_T + s_T)$, where l_T and s_T are selected such that $s_T \to \infty$, $s_T/l_T \to 0$, $l_T/T \to 0$, and the number of the blocks $k_T = [T/(l_T + s_T)] = O(s_T)$. Let $l_T = O(T^{(r-1)/r})$ and $s_T = O(T^{1/r})$ for any $r > 2$, then $k_T = O(T^{1/r}) = O(s_T)$.

For $j = 1, \ldots, k_T$, define

$$\boldsymbol{\xi}_j = \sum_{t=(j-1)(l_T+s_T)+1}^{jl_T+(j-1)s_T} \mathbf{V}_t, \quad \boldsymbol{\eta}_j = \sum_{t=jl_T+(j-1)s_T+1}^{j(l_T+s_T)} \mathbf{V}_t, \quad \boldsymbol{\zeta} = \sum_{t=k_T(l_T+s_T)+1}^{T} \mathbf{V}_t.$$

Note that $\alpha(T) = o(T^{-1})$ and $k_T s_T/T \to 0$. It follows from Proposition 2.7 of Fan and

Yao (85) that

$$\frac{1}{T}E\left(\sum_{j=1}^{k_T}\boldsymbol{\eta}_j\right)^2 \to 0 \quad \text{and} \quad \frac{1}{T}E(\boldsymbol{\zeta}^2) \to 0.$$

This means that the summations over the small blocks and the residual block are asymptotically negligible. Thus,

$$\frac{1}{\sqrt{T}}\mathbf{M}_t = \frac{1}{\sqrt{T}}\left(\sum_{j=1}^{k_T}\boldsymbol{\xi}_j + \sum_{j=1}^{k_T}\boldsymbol{\eta}_j + \boldsymbol{\zeta}\right) = \frac{1}{\sqrt{T}}\sum_{j=1}^{k_T}\boldsymbol{\xi}_j + o_p(1).$$

It follows from Proposition 2.6 of Fan and Yao (85) that, as $T \to \infty$,

$$\left| E\left\{\exp\left(\frac{it}{\sqrt{T}}\sum_{j=1}^{k_T}\boldsymbol{\xi}_j\right)\right\} - \prod_{j=1}^{k_T} E\left\{\exp\left(it\boldsymbol{\xi}_j/\sqrt{n}\right)\right\}\right| \le 16(k_T-1)\alpha(s_T) \to 0,$$

which implies the summations over the large blocks $\{\boldsymbol{\xi}_j\}$ are asymptotically independent. Now, by stationarity, we have

$$\frac{1}{T}\mathrm{Var}(\mathbf{M}_T) = \frac{1}{T}\sum_{j=1}^{n}\mathrm{Var}(\mathbf{V}_j) + \frac{2}{T}\sum_{1\le t<j\le T}\mathrm{Cov}(\mathbf{V}_t, \mathbf{V}_j) = \gamma(0) + 2\sum_{l=1}^{T-1}\left(1-\frac{l}{T}\right)\gamma(l),$$

where $\gamma(k) = \mathrm{Cov}(\mathbf{V}_{t+k}, \mathbf{V}_t)$ is the autocovariance function of $\mathbf{V}_t$. Define $\boldsymbol{\Sigma}_1 = \mathbf{D}_2(\mathbf{u})\otimes\mathbf{S}$. It is straightforward to show that $\boldsymbol{\Sigma}_1 = \mathbf{D}_2(\mathbf{u}) \otimes \mathbf{S} = \gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j)$. Applying Theorem 2.20 of Fan and Yao (85), we have $l_T^{-1}E(\boldsymbol{\xi}_1^2) \to \boldsymbol{\Sigma}_1$, which implies the Feller condition

$$\frac{1}{T}\sum_{j=1}^{k_T}E(\boldsymbol{\xi}_j^2) = \frac{k_T l_T}{T}\frac{1}{l_T}E(\boldsymbol{\xi}_1^2) \to \boldsymbol{\Sigma}_1.$$

Note that, for any $\epsilon > 0$,

$$E[\boldsymbol{\xi}_1^2 I(|\boldsymbol{\xi}_1| > \sqrt{T}\epsilon|\boldsymbol{\Sigma}_1|^{1/2})] \le E(\boldsymbol{\xi}_1^4)^{1/2}P[|\boldsymbol{\xi}_j| > \sqrt{T}\epsilon|\boldsymbol{\Sigma}_1|^{1/2}] \le Cl_T\frac{1}{T\epsilon^2}|\boldsymbol{\Sigma}_1|^{-1}E(\boldsymbol{\xi}_1^2) = O(l_T^2/T).$$

It follows that

$$\frac{1}{T}\sum_{j=1}^{k_T}E[\boldsymbol{\xi}_j^2 I(|\boldsymbol{\xi}_j| \ge \sqrt{T}\epsilon|\boldsymbol{\Sigma}_1|^{1/2})] = O(k_T l_T^2/T^2) = O(l_T/T) \to 0,$$

which is the Lindberg condition. Then by the central limit theorem, we have

$$\prod_{j=1}^{k_T} E[\exp(i\mathbf{t}\boldsymbol{\xi}_j/\sqrt{T})] \to \exp\{-\mathbf{t}\boldsymbol{\Sigma}_1\mathbf{t}^{'}/2\},$$

for any $\mathbf{t}$. That is, $T^{-1/2}\mathbf{M}_T \to N(0, \boldsymbol{\Sigma}_1)$ in distribution as $T \to \infty$.

Then, we have $\sqrt{T}(\widehat{\boldsymbol{\Phi}}_T(\mathbf{u}) - \boldsymbol{\Phi}^*(\mathbf{u}))$ converges weakly to a $N \times (NP+1)$-dimensional normal distribution with mean zero and dispersion matrix $\boldsymbol{\Sigma}^*_{SQR} = [\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})] \otimes \mathbf{S}^{-1}$, where $\boldsymbol{\Phi}^*(\mathbf{u}) = (\boldsymbol{\Phi}^*_1(\mathbf{u}), \boldsymbol{\Phi}^*_2(\mathbf{u}))$ is a $N \times (NP+1)$ matrix. □

Fact 2. By the definition of $\boldsymbol{\Phi}^*(\mathbf{u})$, $\boldsymbol{\Phi}^*_1(\mathbf{u})$ is the first column of $\boldsymbol{\Phi}^*(\mathbf{u})$, and $\boldsymbol{\Phi}^*_2(\mathbf{u})$ are the remaining columns of $\boldsymbol{\Phi}^*(\mathbf{u})$ which do not depend on $\mathbf{u}$. We also partition $\widehat{\boldsymbol{\Phi}}_T(\mathbf{u})$ as $\widehat{\boldsymbol{\Phi}}_T(\mathbf{u}) = (\widehat{\boldsymbol{\Phi}}_{1T}(\mathbf{u}), \widehat{\boldsymbol{\Phi}}_{2T}(\mathbf{u}))$ in the same way as that for $\boldsymbol{\Phi}^*(\mathbf{u})$. Then, $\widehat{\boldsymbol{\Phi}}_{2T}(\mathbf{u})$ are always consistent estimators of $\boldsymbol{\Phi}^*_2(\mathbf{u})$ for different $\mathbf{u}$, i.e., $\sqrt{T}(\widehat{\boldsymbol{\Phi}}_{2T}(\mathbf{u}) - \boldsymbol{\Phi}^*_2(\mathbf{u}))$ converges weakly to a $N \times NP$-dimensional normal distribution with mean zero and dispersion matrix $\boldsymbol{\Sigma}_{SQR} = \left[\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})\right] \otimes (\boldsymbol{\Gamma}^*)^{-1}$. Since, $\boldsymbol{\Phi}^*_1(\mathbf{u}) = \mathbf{q}^*_u$ and $\boldsymbol{\Phi}^*_2(\mathbf{u}) = \mathbf{A}^*$, then we have $\sqrt{T}(\text{vec}(\widehat{\mathbf{A}}_{\text{SQR}}) - \text{vec}(\mathbf{A}^*))$ converges weakly to a $N \times NP$-dimensional normal distribution with mean zero and dispersion matrix $\boldsymbol{\Sigma}_{\text{SQR}} = \left[\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})\right] \otimes (\boldsymbol{\Gamma}^*)^{-1}$.

Lemma 6. Suppose that $\alpha \geq 1$. if $\lambda/\sqrt{T} \to \lambda_0 \geq 0$ and under condition C_1 in Appendix B then

$$\sqrt{T}(\widehat{\boldsymbol{\Phi}}_T(\mathbf{u}) - \boldsymbol{\Phi}^*(\mathbf{u})) \xrightarrow{D} \arg\min(\mathbf{V}), \tag{B.14}$$

where, if $\alpha > 1$,

$$\mathbf{V}(\gamma) = T^{-1/2}[\mathbf{D}_1(\mathbf{u})]^{-1}\left[\sum_{t=1}^{T} \varphi_{\mathbf{u}}(\varepsilon_t - \mathbf{q}_u)\mathbf{Z}_t^{'}\right]\mathbf{S}^{-1} + \lambda_0 \sum_{i=1}^{N}\sum_{j=1}^{NP} \gamma_{ij}\text{sgn}(A_{ij})|A_{ij}|^{\alpha-1}$$

if $\alpha = 1$,

$$\mathbf{V}(\boldsymbol{\gamma}) = T^{-1/2}[\mathbf{D}_1(\mathbf{u})]^{-1}\left[\sum_{t=1}^{T}\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\mathbf{Z}_t^{'}\right]\mathbf{S}^{-1}$$
$$+ \lambda_0\sum_{i=1}^{N}\sum_{j=1}^{NP}[\gamma_{ij}\mathrm{sgn}(A_{ij})\mathrm{I}(A_{ij} \neq 0) + |\gamma_{ij}|\mathrm{I}(A_{ij} = 0)]$$

Proof. Define $\mathbf{V}_T(\boldsymbol{\gamma})$ by

$$\mathbf{V}_T(\boldsymbol{\gamma}) = \sum_{t=p+1}^{T}[Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) - Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)]$$
$$+ \lambda\sum_{i=1}^{N}\sum_{j=1}^{NP}[|A_{ij} + \gamma_{ij}/\sqrt{T}|^{\alpha} - |A_{ij}|^{\alpha}].$$

where $\boldsymbol{\gamma} = [\gamma_{ij}]_{N\times(NP+1)}$ and note that $\mathbf{V}_T$ is minimized at $\sqrt{T}(\widehat{\boldsymbol{\Phi}}_T(\mathbf{u}) - \boldsymbol{\Phi}^*(\mathbf{u}))$. First note that

$$\sum_{t=p+1}^{T}[Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\boldsymbol{\gamma}\mathbf{Z}_t) - Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)] \xrightarrow{D} T^{-1/2}[\mathbf{D}_1(\mathbf{u})]^{-1}\left[\sum_{t=1}^{T}\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\mathbf{Z}_t^{'}\right]\mathbf{S}^{-1}.$$

which is already proved in Lemma 5. if $\alpha > 1$ then

$$\lambda\sum_{i=1}^{N}\sum_{j=1}^{NP}[|A_{ij} + \gamma_{ij}/\sqrt{T}|^{\alpha} - |A_{ij}|^{\alpha}] \to \lambda_0\sum_{i=1}^{N}\sum_{j=1}^{NP}\gamma_{ij}\mathrm{sgn}(A_{ij})|A_{ij}|^{\alpha-1},$$

while for $\alpha = 1$, we have

$$\lambda\sum_{i=1}^{N}\sum_{j=1}^{NP}[|A_{ij} + \gamma_{ij}/\sqrt{T}|^{\alpha} - |A_{ij}|^{\alpha}] \to \lambda_0\sum_{i=1}^{N}\sum_{j=1}^{NP}[\gamma_{ij}\mathrm{sgn}(A_{ij})\mathrm{I}(A_{ij} \neq 0) + |\gamma_{ij}|\mathrm{I}(A_{ij} = 0)]$$

Thus $\mathbf{V}_T(\boldsymbol{\gamma}) \xrightarrow{D} \mathbf{V}(\boldsymbol{\gamma})$ (as defined above) with the finite-dimensional convergence holding trivially. Since $\mathbf{V}_T$ is convex and $\mathbf{V}$ has a unique minimum, it follows Geyer (1996) (86) that

$$\mathrm{argmin}(\mathbf{V}_T) = \sqrt{T}(\widehat{\boldsymbol{\Phi}}_T(\mathbf{u}) - \boldsymbol{\Phi}^*(\mathbf{u})) \xrightarrow{D} \mathrm{argmin}(\mathbf{V}).$$

Note that when $\lambda_0 = 0$, $\mathrm{argmin}(\mathbf{V}) = T^{-1/2}\mathbf{D}_1^{-1}(\mathbf{u})\left[\sum_{t=1}^{T}\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\epsilon}_t - \mathbf{q}_u)\mathbf{Z}_t^{'}\right]\mathbf{S}^{-1}$. Then, we have $\sqrt{T}(\widehat{\mathbf{\Phi}}_T(\mathbf{u}) - \mathbf{\Phi}^*(\mathbf{u}))$ converges weakly to a $N \times (NP+1)$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma}^*_{\text{SQRLASSO}} = [\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})] \otimes \mathbf{S}^{-1}$. By the notations from Lemma 5 and explanations from Fact 2, $\mathbf{\Phi}^*_1(\mathbf{u}) = \mathbf{q}^*_u$ and $\mathbf{\Phi}^*_2(\mathbf{u}) = \mathbf{A}^*$, then we have $\sqrt{T}(\text{vec}(\widehat{\mathbf{A}}_{\text{SQRLASSO}}) - \text{vec}(\mathbf{A}^*))$ converges weakly to a $N \times NP$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma}_{\text{SQRLASSO}} = \left[\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})\right] \otimes (\mathbf{\Gamma}^*)^{-1}$.

□

Proof of Theorem 4.

Proof. The proof generally follows from Proposition 4.1 in Shapiro (1986) (75) for over-parameterized models. let $\mathbf{h}(\boldsymbol{\phi})$ denote the true parameter $\text{vec}(\mathcal{A}_{(1)}) = \text{vec}(\mathbf{U}_1\mathcal{G}_{(1)}(\mathbf{U}_3 \otimes \mathbf{U}_2)')$, and let $\widehat{\mathbf{h}}_{\text{SQR}}$ denote the vectorized SQR estimates $\text{vec}(\widehat{\mathbf{A}}_{\text{SQR}})$ without constraint. By the Fact 2, $\sqrt{T}(\widehat{\mathbf{h}}_{\text{SQR}} - \mathbf{h}^*) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{SQR}})$, where $\mathbf{\Sigma}_{\text{SQR}} = \left[\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})\right] \otimes (\mathbf{\Gamma}^*)^{-1}$. Consider the discrepancy function for any $\mathbf{h}(\boldsymbol{\phi})$,

$$F(\widehat{\mathbf{h}}_{\text{SQR}}, \mathbf{h}) = \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \mathbf{h}\mathbf{x}_t - \mathbf{q}_u) - \sum_{t=1}^{T} Q_u(\mathbf{y}_t - \widehat{\mathbf{h}}_{\text{SQR}}\mathbf{x}_t - \widehat{\mathbf{q}}_u)$$

Obviously, $F(\widehat{\mathbf{h}}_{\text{SQR}}, \mathbf{h})$ is a nonnegative and twice continuously differentiable function, and equals to zero if and only if $\widehat{\mathbf{h}}_{\text{SQR}} = \mathbf{h}$.

To calculate the Jacobian matrix $\mathbf{H}$, we define the tensor matricization transformation operator $\mathbf{T}_{ij}(N, N, P)$ which is an $N^2P \times N^2P$ matrix and satisfies that $\text{vec}(\mathcal{A}_{(j)}) = \mathbf{T}_{ij}(N, N, P)\text{vec}(\mathcal{A}_{(i)})$ for any tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times P}$. In fact, $\mathbf{T}_{ij}(N, N, P)$ is a full-rank matrix indicating the corresponding position in $\text{vec}(\mathcal{A}_{(i)})$ of $\mathcal{A}$'s each entry in $\text{vec}(\mathcal{A}_{(j)})$, and can be regarded as the natural extension of the permutation matrix for matrix transpose. Also note that $\mathbf{T}_{ij}(N, N, P)$ only depends on the value of N and P, and since we consider fixed N and P in this part, we simplify it to $\mathbf{T}_{ij}$.

Therefore,

$$\begin{aligned}\text{vec}(\mathcal{A}_{(1)}) &= \text{vec}(\boldsymbol{U}_1\mathcal{G}_{(1)}(\boldsymbol{U}_3\otimes\boldsymbol{U}_2)')\\ &= T_{21}\text{vec}(\boldsymbol{U}_2\mathcal{G}_{(2)}(\boldsymbol{U}_1\otimes\boldsymbol{U}_3)')\\ &= T_{31}\text{vec}(\boldsymbol{U}_3\mathcal{G}_{(3)}(\boldsymbol{U}_1\otimes\boldsymbol{U}_2)'),\end{aligned}$$

and the Jacobian matrix of $\mathbf{h}$ is

$$\begin{aligned}\mathbf{H} = \frac{\partial\mathbf{h}}{\partial\boldsymbol{\phi}} = \{&(\boldsymbol{U}_3\otimes\boldsymbol{U}_2\otimes\boldsymbol{U}_1),\\ &\left[(\boldsymbol{U}_3\otimes\boldsymbol{U}_2)\mathcal{G}'_{(1)}\right]\otimes\boldsymbol{I}_N,\\ &\mathbf{T}_{21}\left\{\left[(\boldsymbol{U}_1\otimes\boldsymbol{U}_3)\mathcal{G}'_{(2)}\right]\otimes\boldsymbol{I}_N\right\},\\ &\mathbf{T}_{31}\left\{\left[(\boldsymbol{U}_1\otimes\boldsymbol{U}_2)\mathcal{G}'_{(3)}\right]\otimes\boldsymbol{I}_P\right\}\right\}.\end{aligned}$$

Then, by Proposition 4.1 in Shapiro (1986) (75), we know that the minimizer of $F(\hat{\mathbf{h}}_{\text{SQR}},\cdot)$, namely the MLRSQR estimator, has the asymptotic normality,

$$\sqrt{T}(\mathbf{h}(\widehat{\boldsymbol{\phi}}_{\text{MLRSQR}})-\mathbf{h}^*)\xrightarrow{D}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_{\text{MLRSQR}}).$$

i.e.,

$$\sqrt{T}\{\text{vec}((\widehat{\mathcal{A}}_{\text{MLRSQR}})_{(1)})-\text{vec}(\mathcal{A}^*_{(1)})\}\xrightarrow{D}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_{\text{MLRSQR}}).$$

and $\boldsymbol{\Sigma}_{\text{MLRSQR}} = \mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^{'}$, where $\mathbf{P} = \mathbf{H}(\mathbf{H}'\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}'\mathbf{J}$ is the projection matrix, $\mathbf{J}$ is the Fisher information matrix of $\mathbf{h}$ as T goes to infinity, $\mathbf{H}$ is the Jacobian matrix of $\mathbf{h}$ with respect to the overparameterized model parameters $\boldsymbol{\phi}$, $\boldsymbol{\Gamma} = \left[\mathbf{D}_1^{-1}(\mathbf{u})\mathbf{D}_2(\mathbf{u})\mathbf{D}_1^{-1}(\mathbf{u})\right]\otimes(\boldsymbol{\Gamma}^*)^{-1}$ is the asymptotic covariance matrix for $\widehat{\mathbf{h}}_{\text{SQR}}$ and † denotes the Moore-Penrose inverse. Since $\boldsymbol{\Gamma} = \mathbf{J}^{-1}$, we have $\boldsymbol{\Sigma}_{\text{MLRSQR}} = \mathbf{H}(\mathbf{H}'\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}'$. □

Fact 3. Consider (B.5), we vectorize the estimated parameters. Then the (B.5) can be written as

$$\text{vec}(\widehat{\boldsymbol{\Phi}}_T^{'}(\mathbf{u})) = \arg\min_{\boldsymbol{\Phi}}\sum_{t=p+1}^{T}Q_u[\mathbf{y}_t-(\mathbf{I}_N\otimes\mathbf{Z}_t^{'})\text{vec}(\boldsymbol{\Phi}^{'}(\mathbf{u}))], \qquad \text{(B.15)}$$

By Lemma 5 and the uniform convergence theory, for every ϵ and $M > 0$, the inequality

$$\sup_{\|\boldsymbol{\gamma}\|\leq M}\left|\sum_{t=1}^{T} Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}(\mathbf{I}_N \otimes \mathbf{Z}_t^{'})\boldsymbol{\gamma}) - \sum_{t=1}^{T} Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u) \right.$$
$$+T^{-1/2}\boldsymbol{\gamma}^{'}\left[\sum_{t=1}^{T}(\mathbf{I}_N \otimes \mathbf{Z}_t)\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\right]$$
$$\left. -\frac{1}{2}\boldsymbol{\gamma}^{'}[(\mathbf{I}_N \otimes \mathbf{Z}_t)\mathbf{D}_1(\mathbf{u})(\mathbf{I}_N \otimes \mathbf{Z}_t^{'})]\boldsymbol{\gamma}\right| < \varepsilon. \tag{B.16}$$

holds with probability at least $1 - \varepsilon/2$ for a fixed $\boldsymbol{\gamma} \in \mathbb{R}^{N\times(NP+1)}$ and for large T. We have the following Bahadur representation of $\widehat{\boldsymbol{\Phi}}_T(\mathbf{u})$

$$\sqrt{T}[\mathrm{vec}(\widehat{\boldsymbol{\Phi}}_T^{'}(\mathbf{u})) - \mathrm{vec}(\boldsymbol{\Phi}^{*'}(\mathbf{u}))] = T^{-1/2}[(\mathbf{I}_N \otimes \mathbf{Z}_t)\mathbf{D}_1(\mathbf{u})(\mathbf{I}_N \otimes \mathbf{Z}_t^{'})]^{-1}$$
$$\times \left[\sum_{t=1}^{T}(\mathbf{I}_N \otimes \mathbf{Z}_t)\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\epsilon}_t - \mathbf{q}_u)\right] + o_p(1). \tag{B.17}$$

Suppose that $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathrm{vec}([\mathbf{I}_N\otimes\mathbf{Z}_t])\mathrm{vec}([\mathbf{I}_N\otimes\mathbf{Z}_t])^\top = \mathbf{Q}^*$, where $\mathbf{Q}^*$ is a positive definite matrix, and $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathrm{vec}([\mathbf{I}_N \otimes \mathbf{x}_t])\mathrm{vec}([\mathbf{I}_N \otimes \mathbf{x}_t])^\top = \mathbf{Q}$, where $\mathbf{Q}$ is a positive definite matrix. Then, we have $\sqrt{T}[\mathrm{vec}(\widehat{\boldsymbol{\Phi}}_T^{'}(\mathbf{u})) - \mathrm{vec}(\boldsymbol{\Phi}^{*'}(\mathbf{u}))]$ converges weakly to a $N\times(NP+1)$-dimensional normal distribution with mean zero and dispersion matrix $\boldsymbol{\Sigma}^* = [(\mathbf{I}_N\otimes\mathbf{Z}_t)\mathbf{D}_1(\mathbf{u})(\mathbf{I}_N\otimes\mathbf{Z}_t^{'})]^{-1}[\mathrm{Tr}[\mathbf{D}_2(\mathbf{u})]\mathbf{Q}^*][(\mathbf{I}_N\otimes\mathbf{Z}_t)\mathbf{D}_1(\mathbf{u})(\mathbf{I}_N\otimes\mathbf{Z}_t^{'})]^{-1}$. By the Fact 2, $\sqrt{T}(\mathrm{vec}(\widehat{\mathbf{A}}^{'})-\mathrm{vec}(\mathbf{A}^{*'}))$ converges weakly to a $N\times NP$-dimensional normal distribution with mean zero and dispersion matrix $\boldsymbol{\Sigma} = [(\mathbf{I}_N\otimes\mathbf{x}_t)\mathbf{D}_1(\mathbf{u})(\mathbf{I}_N\otimes\mathbf{x}_t^{'})]^{-1}[\mathrm{Tr}[\mathbf{D}_2(\mathbf{u})]\mathbf{Q}][(\mathbf{I}_N\otimes \mathbf{x}_t)\mathbf{D}_1(\mathbf{u})(\mathbf{I}_N \otimes \mathbf{x}_t^{'})]^{-1}$.

Proof of Theorem 2.

Proof. Consider the ℓ_1-penalized sparse higher-order reduced-rank with SQR (SHORRSQR) estimator:

$$\widehat{\mathcal{A}}_{\text{SHORRSQR}} \equiv [[\widehat{\mathcal{G}}; \widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3; \widehat{\mathbf{q}}_u]]$$
$$= \underset{\mathcal{G},\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3,\mathbf{q}_u}{\arg\min}\left\{\sum_{t=1}^{T} Q_u\Big(\mathbf{y}_t - \mathcal{A}_{(1)}\mathbf{x}_t - \mathbf{q}_u\Big)\right. \tag{B.18}$$
$$\left. + \lambda\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1\right\}$$

where $\mathcal{A}_{(1)} = (\mathcal{G}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times_3\mathbf{U}_3)_{(1)}$ and note that $\|\mathbf{U}_3\otimes\mathbf{U}_2\otimes\mathbf{U}_1\|_1 = \|\mathbf{U}_3\|_1\|\mathbf{U}_2\|_1\|\mathbf{U}_1\|_1$.

Then, by Section 3.2.2, we have

$$\begin{aligned}\mathbf{y}_t &= \big((\mathbf{x}_t'(\mathbf{U}_3 \otimes \mathbf{U}_2)\mathcal{G}_{(1)}') \otimes \mathbf{I}_N\big)\,\mathrm{vec}(\mathbf{U}_1) + \mathbf{q}_u = \mathbf{A}_t\mathrm{vec}(\mathbf{U}_1) + \mathbf{q}_u \\ &= \mathbf{U}_1\mathcal{G}_{(1)}\left((\mathbf{U}_3'\mathbf{X}_t') \otimes \mathbf{I}_{r_2}\right)\mathrm{vec}(\mathbf{U}_2') + \mathbf{q}_u = \mathbf{B}_t\mathrm{vec}(\mathbf{U}_2') + \mathbf{q}_u \\ &= \mathbf{U}_1\mathcal{G}_{(1)}(\mathbf{I}_{r_3} \otimes (\mathbf{U}_2'\mathbf{X}_t))\mathrm{vec}(\mathbf{U}_3) + \mathbf{q}_u = \mathbf{C}_t\mathrm{vec}(\mathbf{U}_3) + \mathbf{q}_u \\ &= (((\mathbf{U}_3 \otimes \mathbf{U}_2)'\mathbf{x}_t)' \otimes \mathbf{U}_1)\mathrm{vec}(\mathcal{G}_{(1)}) + \mathbf{q}_u = \mathbf{D}_t\mathrm{vec}(\mathcal{G}_{(1)}) + \mathbf{q}_u\end{aligned} \tag{B.19}$$

We show asymptotics using vec($\mathbf{U}_1$) as an example, and a similar process for vec($\mathbf{U}_2$) and vec($\mathbf{U}_3$). Since vec($\mathcal{G}_{(1)}$) does not involve a penalty term, the proof of the asymptotics of vec($\mathcal{G}_{(1)}$) follows the Fact 3.

Define $\mathbf{R}_1$ by

$$\begin{aligned}\mathbf{R}_1(\boldsymbol{\gamma}) = T^{-1/2}[\mathbf{A}_t\mathbf{D}_1(\mathbf{u})\mathbf{A}_t^{'}]^{-1}\left[\sum_{t=1}^{T}\mathbf{A}_t^{'}\varphi_{\mathbf{u}}(\boldsymbol{\epsilon}_t - \mathbf{q}_u)\right] \\ + \frac{\lambda\|\mathbf{U}_3\|_1\|\mathbf{U}_2\|_1}{\sqrt{T}}\sum_{i=1}^{N}\sum_{j=1}^{r_3}[\gamma_{ij}\mathrm{sgn}(\gamma_{ij})\mathrm{I}(U_{1ij} \neq 0) + |\gamma_{ij}|\mathrm{I}(U_{1ij} = 0)\end{aligned}$$

Define $\mathbf{R}_{1T}(\boldsymbol{\gamma})$ by

$$\begin{aligned}\mathbf{R}_{1T}(\boldsymbol{\gamma}) = \sum_{t=p+1}^{T}[Q_u(\boldsymbol{\epsilon}_t - \mathbf{q}_u - T^{-1/2}\mathbf{A}_t\boldsymbol{\gamma}) - Q_u(\boldsymbol{\epsilon}_t - \mathbf{q}_u)] \\ + \lambda\|\mathbf{U}_3\|_1\|\mathbf{U}_2\|_1\sum_{i=1}^{N}\sum_{j=1}^{r_1}[|U_{1ij} + \gamma_{ij}/\sqrt{T}| - |U_{1ij}|].\end{aligned}$$

where $\boldsymbol{\gamma} = [\gamma_{ij}]$ corresponds to the element in vec($\mathbf{U}_1$) and note that $\mathbf{R}_{1T}$ is minimized at $\sqrt{T}(\mathrm{vec}(\widehat{\mathbf{U}}_1) - \mathrm{vec}(\mathbf{U}_1^*))$. First note that

$$\sum_{t=p+1}^{T}[Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u - T^{-1/2}\mathbf{A}_t\boldsymbol{\gamma}) - Q_u(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)] \xrightarrow{D} T^{-1/2}[\mathbf{A}_t\mathbf{D}_1(\mathbf{u})\mathbf{A}_t^{'}]^{-1}\left[\sum_{t=1}^{T}\mathbf{A}_t^{'}\varphi_{\mathbf{u}}(\boldsymbol{\varepsilon}_t - \mathbf{q}_u)\right]$$

According to the Fact 3, we have

$$\lambda\|\mathbf{U}_3\|_1\|\mathbf{U}_2\|_1\sum_{i=1}^{N}\sum_{j=1}^{r_1}[|U_{1ij}+\gamma_{ij}/\sqrt{T}|-|U_{1ij}|]\to$$

$$\frac{\lambda\|\mathbf{U}_3\|_1\|\mathbf{U}_2\|_1}{\sqrt{T}}\sum_{i=1}^{N}\sum_{j=1}^{r_1}[\gamma_{ij}\mathrm{sgn}(\gamma_{ij})\mathrm{I}(U_{1ij}\neq 0)+|\gamma_{ij}|\mathrm{I}(U_{1ij}=0)$$

Let $\frac{\lambda\|\mathbf{U}_3\|_1\|\mathbf{U}_2\|_1}{\sqrt{T}}\to\lambda_1\geq 0$. When $\lambda_1=0$, $\sqrt{T}(\mathrm{vec}(\widehat{\mathbf{U}}_1)-\mathrm{vec}(\mathbf{U}_1^*))\xrightarrow{D}\mathrm{argmin}(\mathbf{R}_1)$. Then $\mathrm{argmin}(\mathbf{R}_1)=T^{-1/2}[\mathbf{A}_t\mathbf{D}_1(\mathbf{u})\mathbf{A}_t^{'}]^{-1}\left[\sum_{t=1}^{T}\mathbf{A}_t^{'}\boldsymbol{\varphi}_{\mathbf{u}}(\boldsymbol{\varepsilon}_t-\mathbf{q}_u)\right]$. By the Fact 3, let $\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\mathbf{A}_t^{'})\mathrm{vec}(\mathbf{A}_t^{'})^{\top}=\mathbf{Q}_1$, where $\mathbf{Q}_1$ is a positive definite matrix. Therefore, $\sqrt{T}(\mathrm{vec}(\widehat{\mathbf{U}}_1)-\mathrm{vec}(\mathbf{U}_1^*))$ converges weakly to a $N\times r_1$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma_1}=[\mathbf{A}_t\mathbf{D}_1(\mathbf{u})\mathbf{A}_t^{'}]^{-1}[\mathrm{Tr}(\mathbf{D}_2(\mathbf{u}))\mathbf{Q}_1][\mathbf{A}_t\mathbf{D}_1(\mathbf{u})\mathbf{A}_t^{'}]^{-1}$.

For $\mathrm{vec}(\mathbf{U}_2)$, by the Fact 3, let $\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\mathbf{B}_t^{'})\mathrm{vec}(\mathbf{B}_t^{'})^{\top}=\mathbf{Q}_2$, where $\mathbf{Q}_2$ is a positive definite matrix, and let $\frac{\lambda\|\mathbf{U}_3\|_1\|\mathbf{U}_1\|_1}{\sqrt{T}}\to\lambda_2\geq 0$, when $\lambda_2=0$, $\sqrt{T}(\mathrm{vec}(\widehat{\mathbf{U}}_2^{'})-\mathrm{vec}(\mathbf{U}_2^{*'}))$ converges weakly to a $N\times r_2$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma_2^*}=[\mathbf{B}_t\mathbf{D}_1(\mathbf{u})\mathbf{B}_t^{'}]^{-1}[\mathrm{Tr}(\mathbf{D}_2(\mathbf{u}))\mathbf{Q}_2][\mathbf{B}_t\mathbf{D}_1(\mathbf{u})\mathbf{B}_t^{'}]^{-1}$. Let $\mathbf{K}$ be the permutation matrix such that $\mathrm{vec}(\mathbf{U}_2)=\mathbf{K}\mathrm{vec}(\mathbf{U}_2^{'})$. $\mathbf{K}$ is a sparse matrix of size $Nr_2\times Nr_2$ with full rank where each row and each column has only one element equal to 1, and all other elements are 0. Then, $\sqrt{T}(\mathrm{vec}(\widehat{\mathbf{U}}_2)-\mathrm{vec}(\mathbf{U}_2^*))$ converges weakly to a $N\times r_2$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma_2}=K\mathbf{\Sigma_2^*}K^{'}$.

For $\mathrm{vec}(\mathbf{U}_3)$, by the Fact 3, let $\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\mathbf{C}_t^{'})\mathrm{vec}(\mathbf{C}_t^{'})^{\top}=\mathbf{Q}_3$, where $\mathbf{Q}_3$ is a positive definite matrix, and let $\frac{\lambda\|\mathbf{U}_2\|_1\|\mathbf{U}_1\|_1}{\sqrt{T}}\to\lambda_3\geq 0$, when $\lambda_3=0$, $\sqrt{T}(\mathrm{vec}(\widehat{\mathbf{U}}_3)-\mathrm{vec}(\mathbf{U}_3^*))$ converges weakly to a $P\times r_3$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma_3}=[\mathbf{C}_t\mathbf{D}_1(\mathbf{u})\mathbf{C}_t^{'}]^{-1}[\mathrm{Tr}(\mathbf{D}_2(\mathbf{u}))\mathbf{Q}_3][\mathbf{C}_t\mathbf{D}_1(\mathbf{u})\mathbf{C}_t^{'}]^{-1}$.

For $\mathrm{vec}(\mathcal{G}_{(1)})$, by the Fact 3, let $\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\mathbf{D}_t^{'})\mathrm{vec}(\mathbf{D}_t^{'})^{\top}=\mathbf{Q}_4$, where $\mathbf{Q}_4$ is a positive definite matrix, we have $\sqrt{T}(\mathrm{vec}(\widehat{\mathcal{G}}_{(1)})-\mathrm{vec}(\mathcal{G}_{(1)}^*))$ converges weakly to a $r_1\times r_2\times r_3$-dimensional normal distribution with mean zero and dispersion matrix $\mathbf{\Sigma_4}=[\mathbf{D}_t\mathbf{D}_1(\mathbf{u})\mathbf{D}_t^{'}]^{-1}[\mathrm{Tr}(\mathbf{D}_2(\mathbf{u}))\mathbf{Q}_4][\mathbf{D}_t\mathbf{D}_1(\mathbf{u})\mathbf{D}_t^{'}]^{-1}$.

By the definition of HOSVD, we have

$$\widehat{\mathcal{A}} = [[\widehat{\mathcal{G}}; \widehat{\mathbf{U}}_1, \widehat{\mathbf{U}}_2, \widehat{\mathbf{U}}_3]], \quad \text{and} \quad \mathcal{A} = [[\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]].$$

Then, we have

$$\begin{aligned}
&\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})\\
=&(\widehat{\mathbf{U}}_3 \otimes \widehat{\mathbf{U}}_2 \otimes \widehat{\mathbf{U}}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)}) - (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\mathcal{G}_{(1)})\\
=&(\widehat{\mathbf{U}}_3 \otimes \widehat{\mathbf{U}}_2 \otimes \widehat{\mathbf{U}}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)}) - (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)})\\
&+ (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)}) - (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\mathcal{G}_{(1)})\\
=&(\widehat{\mathbf{U}}_3 \otimes \widehat{\mathbf{U}}_2 \otimes \widehat{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)}) + (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)} - \mathcal{G}_{(1)})\\
=&[(\widehat{\mathbf{U}}_3 - \mathbf{U}_3) \otimes \mathbf{U}_2 \otimes \mathbf{U}_1]\text{vec}(\widehat{\mathcal{G}}_{(1)}) + [\mathbf{U}_3 \otimes (\hat{\mathbf{U}}_2 - \mathbf{U}_2) \otimes \mathbf{U}_1]\text{vec}(\widehat{\mathcal{G}}_{(1)})\\
&+ [\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes (\widehat{\mathbf{U}}_1 - \mathbf{U}_1)]\text{vec}(\widehat{\mathcal{G}}_{(1)}) + (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)} - \mathcal{G}_{(1)})\\
=&[(\widehat{\mathbf{U}}_3 - \mathbf{U}_3) \otimes \mathbf{U}_2 \otimes \mathbf{U}_1]\text{vec}(\mathcal{G}_{(1)}) + [\mathbf{U}_3 \otimes (\hat{\mathbf{U}}_2 - \mathbf{U}_2) \otimes \mathbf{U}_1]\text{vec}(\mathcal{G}_{(1)})\\
&+ [\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes (\widehat{\mathbf{U}}_1 - \mathbf{U}_1)]\text{vec}(\mathcal{G}_{(1)}) + (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)\text{vec}(\widehat{\mathcal{G}}_{(1)} - \mathcal{G}_{(1)}) + o_p(T^{-1/2})\\
=&[\mathbf{I}_P \otimes ((\mathbf{U}_2 \otimes \mathbf{U}_1)\mathcal{G}_{(3)}^{'})]\text{vec}(\widehat{\mathbf{U}}_3 - \mathbf{U}_3) + [\mathbf{I}_N \otimes ((\mathbf{U}_3 \otimes \mathbf{U}_1)\mathcal{G}_{(2)}^{'})]\text{vec}(\widehat{\mathbf{U}}_2 - \mathbf{U}_2)\\
&+ [\mathbf{I}_N \otimes ((\mathbf{U}_3 \otimes \mathbf{U}_2)\mathcal{G}_{(1)}^{'}]\text{vec}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) + (\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'})\text{vec}(\widehat{\mathcal{G}}_{(1)} - \mathcal{G}_{(1)}) + o_p(T^{-1/2}).
\end{aligned}$$

Therefore, $\sqrt{T}\,\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})$ is also normally distributed with mean zero, as $T \to \infty$. □

B.3 A: Proofs of Corollaries

Proof of Corollary 1.

Proof. We now demonstrate the asymptotic normality for $\text{vec}(\widehat{\mathbf{U}}_1)$, as the same reasoning applies to $\text{vec}(\widehat{\mathbf{U}}_2)$ and $\text{vec}(\widehat{\mathbf{U}}_3)$. In this section, we simplify $\widehat{\mathcal{A}}_{\text{MLRSQR}}$ to $\widehat{\mathcal{A}}$. Note that $\widehat{\mathbf{U}}_1$ and $\mathbf{U}_1$ are the eigenvectors of $\widehat{\mathcal{A}}_{(1)}\widehat{\mathcal{A}}_{(1)}'$ and $\mathcal{A}_{(1)}\mathcal{A}_{(1)}'$, respectively. By Theorem 4, we have $\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{h}})$, where $\boldsymbol{\Sigma}_{\mathbf{h}} = \boldsymbol{\Sigma}_{\text{MLRSQR}}$. Note that

$$\begin{aligned}
&\sqrt{T}(\widehat{\mathcal{A}}_{(1)}\widehat{\mathcal{A}}'_{(1)} - \mathcal{A}_{(1)}\mathcal{A}'_{(1)}) \\
=&\sqrt{T}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})\mathcal{A}'_{(1)} + \sqrt{T}\mathcal{A}_{(1)}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})' + \sqrt{T}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})' \\
=&\sqrt{T}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})\mathcal{A}'_{(1)} + \sqrt{T}\mathcal{A}_{(1)}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})' + O_p(T^{-1/2})
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)}\widehat{\mathcal{A}}'_{(1)} - \mathcal{A}_{(1)}\mathcal{A}'_{(1)}) \\
=&(\mathcal{A}_{(1)} \otimes \mathbf{I}_N)\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)}) + (\mathbf{I}_N \otimes \mathcal{A}_{(1)})\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)}) + O_p(T^{-1/2}) \\
=&[(\mathcal{A}_{(1)} \otimes \mathbf{I}_N) + (\mathbf{I}_N \otimes \mathcal{A}_{(1)})]\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)}) + O_p(T^{-1/2})
\end{aligned}$$

Therefore, $\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)}\widehat{\mathcal{A}}'_{(1)} - \mathcal{A}_{(1)}\mathcal{A}'_{(1)})$ is asymptotically normal.

Using the matrix perturbation theory (Izenman, 1975 (87); Velu and Reinsel, 2013 (88)),

$$\begin{aligned}
&\sqrt{T}(\widehat{\mathbf{U}}_{1k} - \mathbf{U}_{1k}) \\
=&\sum_{i \neq k} \frac{1}{d_k^2 - d_i^2}(\mathbf{U}'_{1i} \otimes \mathbf{U}_{1i}\mathbf{U}'_{1i})\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)}\widehat{\mathcal{A}}'_{(1)} - \mathcal{A}_{(1)}\mathcal{A}'_{(1)}) + O_p(T^{-1/2}) \\
=&\sum_{i \neq k} \frac{1}{d_k^2 - d_i^2}(\mathbf{U}'_{1i} \otimes \mathbf{U}_{1i}\mathbf{U}'_{1i})[(\mathcal{A}_{(1)} \otimes \mathbf{I}_N) + (\mathbf{I}_N \otimes \mathcal{A}_{(1)})]\sqrt{T}\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)}) + O_p(T^{-1/2})
\end{aligned}$$

Let $\mathbf{M}_1 = \sum_{i \neq k} \frac{1}{d_k^2 - d_i^2}(\mathbf{U}'_{1i} \otimes \mathbf{U}_{1i}\mathbf{U}'_{1i})[(\mathcal{A}_{(1)} \otimes \mathbf{I}_N) + (\mathbf{I}_N \otimes \mathcal{A}_{(1)})]$. Then, we have $\boldsymbol{\Sigma}_{\mathbf{U}_1} = \mathbf{M}_1\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{M}_1^{'}$. Thus, $\sqrt{T}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{U}_1})$ in distribution as $T \to \infty$. Similarly, Let $\mathbf{M}_2 = \sum_{i \neq k} \frac{1}{d_k^2 - d_i^2}(\mathbf{U}'_{2i} \otimes \mathbf{U}_{2i}\mathbf{U}'_{2i})[(\mathcal{A}_{(1)} \otimes \mathbf{I}_N) + (\mathbf{I}_N \otimes \mathcal{A}_{(1)})]$. Then, we have $\boldsymbol{\Sigma}_{\mathbf{U}_2} = \mathbf{M}_2\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{M}_2^{'}$. Thus, $\sqrt{T}(\widehat{\mathbf{U}}_2 - \mathbf{U}_2) \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{U}_2})$ in distribution as $T \to \infty$, and let $\mathbf{M}_3 = \sum_{i \neq k} \frac{1}{d_k^2 - d_i^2}(\mathbf{U}'_{3i} \otimes \mathbf{U}_{3i}\mathbf{U}'_{3i})[(\mathcal{A}_{(1)} \otimes \mathbf{I}_P) + (\mathbf{I}_P \otimes \mathcal{A}_{(1)})]$. Then, we have $\boldsymbol{\Sigma}_{\mathbf{U}_3} = \mathbf{M}_3\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{M}_3^{'}$. Thus, $\sqrt{T}(\widehat{\mathbf{U}}_3 - \mathbf{U}_3) \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{U}_3})$ in distribution as $T \to \infty$.

For $\text{vec}(\widehat{\mathcal{G}}_{(1)})$, by the definition of HOSVD, we have

$$\widehat{\mathcal{G}} = [[\widehat{\mathcal{A}}; \widehat{\mathbf{U}}_1^{'}, \widehat{\mathbf{U}}_2^{'}, \widehat{\mathbf{U}}_3^{'}]], \quad \text{and} \quad \mathcal{G} = [[\mathcal{A}; \mathbf{U}_1^{'}, \mathbf{U}_2^{'}, \mathbf{U}_3^{'}]].$$

Then, we have

$$\begin{aligned}
&\text{vec}(\widehat{\mathcal{G}}_{(1)} - \mathcal{G}_{(1)})\\
=&(\widehat{\mathbf{U}}_3^{'} \otimes \widehat{\mathbf{U}}_2^{'} \otimes \widehat{\mathbf{U}}_1^{'})\text{vec}(\widehat{\mathcal{A}}_{(1)}) - (\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'})\text{vec}(\mathcal{A}_{(1)})\\
=&(\widehat{\mathbf{U}}_3^{'} \otimes \widehat{\mathbf{U}}_2^{'} \otimes \widehat{\mathbf{U}}_1^{'} - \mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'})\text{vec}(\mathcal{A}_{(1)}) + (\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'})\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)})\\
=&[(\widehat{\mathbf{U}}_3^{'} - \mathbf{U}_3^{'}) \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'}]\text{vec}(\mathcal{A}_{(1)}) + [\mathbf{U}_3^{'} \otimes (\hat{\mathbf{U}}_2 - \mathbf{U}_2)]^{'} \otimes \mathbf{U}_1^{'}\text{vec}(\mathcal{A}_{(1)})\\
&+ [\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes (\widehat{\mathbf{U}}_1^{'} - \mathbf{U}_1^{'})]\text{vec}(\mathcal{A}_{(1)}) + (\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'})\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)}) + o_p(T^{-1/2})\\
=&[\mathbf{I}_{r_3} \otimes ((\hat{\mathbf{U}}_2^{'} \otimes \widehat{\mathbf{U}}_1^{'})\mathcal{A}_{(3)}^{'})]\text{vec}(\widehat{\mathbf{U}}_3 - \mathbf{U}_3) + [\mathbf{I}_{r_2} \otimes ((\mathbf{U}_3^{'} \otimes \widehat{\mathbf{U}}_1^{'})\mathcal{A}_{(2)}^{'})]\text{vec}(\widehat{\mathbf{U}}_2 - \mathbf{U}_2)\\
&+ [\mathbf{I}_{r_1} \otimes ((\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'})\mathcal{A}_{(1)}^{'})]\text{vec}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) + (\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'})\text{vec}(\widehat{\mathcal{A}}_{(1)} - \mathcal{A}_{(1)}) + o_p(T^{-1/2}).
\end{aligned}$$

Let $\mathbf{L}_1 = [\mathbf{I}_{r_1} \otimes ((\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'})\mathcal{A}_{(1)}^{'})]$, $\mathbf{L}_2 = [\mathbf{I}_{r_2} \otimes ((\mathbf{U}_3^{'} \otimes \widehat{\mathbf{U}}_1^{'})\mathcal{A}_{(2)}^{'})]$, $\mathbf{L}_3 = [\mathbf{I}_{r_3} \otimes ((\widehat{\mathbf{U}}_2^{'} \otimes \widehat{\mathbf{U}}_1^{'})\mathcal{A}_{(3)}^{'})]$, and $\mathbf{L}_4 = (\mathbf{U}_3^{'} \otimes \mathbf{U}_2^{'} \otimes \mathbf{U}_1^{'})$. Therefore, $\mathbf{\Sigma}_{\mathcal{G}} = \mathbf{L}_1\mathbf{\Sigma}_{\mathbf{U}_1}\mathbf{L}_1^{'} + \mathbf{L}_2\mathbf{\Sigma}_{\mathbf{U}_2}\mathbf{L}_2^{'} + \mathbf{L}_3\mathbf{\Sigma}_{\mathbf{U}_3}\mathbf{L}_3^{'} + \mathbf{L}_4\mathbf{\Sigma}_{\mathbf{h}}\mathbf{L}_4^{'}$. Thus, $\sqrt{T}\text{vec}(\widehat{\mathcal{G}}_{(1)} - \mathcal{G}_{(1)}) \to N(\mathbf{0}, \mathbf{\Sigma}_{\mathcal{G}})$ in distribution as $T \to \infty$. □

Proof of Corollary 2.

Proof. From Lemma 4 and Theorem 4, we have $\mathbf{\Sigma}_{\text{SQR}} = \mathbf{\Gamma} = \mathbf{J}^{-1}$. Since $\mathbf{\Sigma}_{\text{MLRSQR}} = \mathbf{P}\mathbf{\Gamma}\mathbf{P}^{'} = \mathbf{P}\mathbf{J}^{-1}\mathbf{P}^{'}$, where $\mathbf{P} = \mathbf{H}(\mathbf{H}'\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}'\mathbf{J}$ is a projection matrix. Note that $\mathbf{J}^{-1} - \mathbf{H}(\mathbf{H}'\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}' = \mathbf{J}^{-1/2}\mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}}\mathbf{J}^{-1/2}$, where $\mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}}$ is the projection matrix onto the orthogonal of span$(\mathbf{J}^{-1/2})$. Then, we have $\mathbf{J}^{-1} \geq \mathbf{H}(\mathbf{H}'\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}'$. □