THE UNIVERSITY OF CHICAGO

ANALYSIS OF BIG HIGH-DIMENSIONAL DEPENDENT DATA: UNIT-ROOTS AND DISTRIBUTED COMPUTATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
SHUO-CHIEH HUANG

CHICAGO, ILLINOIS
JUNE 2024

To my parents, Shu-Fang Lin and Kuang-Hsien Huang

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Gratitude fills my heart as I reflect on the completion of this thesis, which is only possible because of the unwavering support of numerous individuals whose kindness and guidance have brought this thesis to fruition.

First and foremost, my gratitude extends to my advisor, Ruey S. Tsay, for his timely guidance and advice on the development of the research. I also thank him for being supportive and thoughtful throughout my entire journey at Chicago Booth. His dedication to scholarly pursuit and encouraging attitude towards colleagues have shaped my standards of excellence for leading scholars, which is perhaps the most valuable takeaway in my Ph.D. study.

I am equally indebted to Tengyuan Liang for the intellectual discussions and constant encouragement. Attending his classes and reading group meetings was very enjoyable, and I was inspired a lot by his acute thoughts on so many interesting topics. The keen insights and genuine thoughts he had put in our discussion on my academic career development are also invaluable. I have learned a great deal from his talent and character.

A special note of gratitude goes to my family. My wife, Meng-Yu Tsai, has been my steadfast companion during my Ph.D. journey. Her unbounded patience and love have sustained me through challenging times. My parents, Shu-Fan Lin and Kuang-Hsien Huang, have nourished me with their unconditional love. In loving memory of my mother, whose passing in 2021 during the Covid lockdown weighed heavily on my heart, I dedicate this thesis as a tribute to her enduring influence and spirit.

I extend my heartfelt thanks to those who have provided indispensable assistance along the way: Ching-Kang Ing at the National Tsing Hua University for his inspiring mentorship since my master studies; Chien-Ming Chi at the Institute of Statistical Science, Academia Sinica, for our delightful intellectual conversations; Yu-Chang Chen at the National Taiwan University, for being an agreeable coauthor. I also thank Mladen Kolar, Yuexi Wang, Kai-Jie

Wu, Kuan-Ming Chen, Sung-Ju Wu, and Tsu-En Wang for their generous help during my job search and as dear friends. Finally, I thank all the wonderful assistance provided by the Ph.D. program office that helped me navigate the journey.

ABSTRACT

This dissertation consists of two parts that address problems arising frequently in the analysis of big, high-dimensional dependent data. The first part proposes a new approach for model selection when the response variable contains unit roots with unknown multiplicities at unknown locations. The proposed method, FHTD, is based on the forward stepwise regression technique. Despite of unit roots, high-dimensional predictors, and conditionally heteroscedastic errors, FHTD is shown to select exactly the correct model with probability tending to one. Thus, our approach is applicable to a wide range of data without recourse to any delicate unit-root tests.

The second part tackles the computational problem when the data are vertically partitioned and stored across computing nodes. To jointly learn a multivariate linear model, we propose a two-stage relaxed greedy algorithm so that communication between computing nodes is minimized and hence the algorithm is computationally efficient. Throughout, we supply simulation studies and real data examples to demonstrate the performance of the proposed methods.

CHAPTER 1

INTRODUCTION

The analysis of big, high-dimensional dependent data has become central to many scientific fields as the advancement in information technology has enabled unprecedented data collection and computational power. However, despite of the rich information that one can exploit from the newly available data, whether it is more variables being monitored (high-dimension) or longer periods of observations (sample size), big dependent data present unique challenges in contrast to independent data.

First, in practice, long time series is often nonstationary, which has a very different statistical nature compared to independent or stationary data. In particular, for many statistics, the sample covariance matrix for example, the "population counterpart" is not well-defined, and the sample statistic often lacks a deterministic limit. Hence, most existing tools for estimating high-dimensional models are invalid in dealing with nonstationary data. Second, to facilitate computation via distributed computing, it is more natural for high-dimensional dependent data to partition the data vertically and store each part in distinct computing nodes. That is, each node only owns some predictors of the whole data set. While conceptually appealing and in some applications it is normal to deal with such data, the vertical partition scheme means each node, on its own, is unable to learn the dependence structure among the predictors. Therefore, carefully designed distributed algorithm is in need to make use of such data for statistical analysis. In this dissertation, we tackle these two challenges and propose novel methods to enhance the analysis of big dependent data.

In Chapter 2, we study the variable selection in the autoregressive model with exogenous inputs (ARX) for general unit-root time series when many predictors are present. A new model selection algorithm called FHTD that leverages the advantages of forward stepwise regression (FSR), a high-dimensional information criterion (HDIC), a backward elimination method based on HDIC, and a data-driven thresholding (DDT) approach, is proposed. By

applying a new functional central limit theorem for multivariate linear processes, along with a uniform lower bound for the minimum eigenvalue of the sample covariance matrices of the series under study, we establish the sure screening property of FSR and the selection consistency of FHTD under some mild assumptions that allow for unknown locations and multiplicities of the characteristic roots on the unit circle of the time series and conditional heteroscedasticity in the predictors and errors. The method is applied to U.S. monthly housing starts and unemployment data and it is found to be more preferable to commonly used benchmarks.

In Chapter 3, we propose a two-stage relaxed greedy algorithm (TSRGA) for applying multivariate linear regression to the feature-distributed data, referred to data partitioned by features and stored across multiple computing nodes. The main advantage of TSRGA is that its communication complexity does not depend on the feature dimension, making it highly scalable to very large data sets. For multivariate response variables, TSRGA can be used to yield low-rank coefficient estimates. In addition, we offer a simple modification of TSRGA which can be used to estimate the generalized linear model (GLM) with high-dimensional predictors. Finally, we apply the proposed TSRGA in a financial application that leverages unstructured data from the 10-K reports, demonstrating its usefulness in applications with many dense large-dimensional matrices.

CHAPTER 2

MODEL SELECTION FOR UNIT-ROOT TIME SERIES WITH MANY PREDICTORS

2.1 Introduction

With the widespread availability of large-scale and fine-grained datasets, researchers analyzing time series data now have a plethora of predictors available for constructing informative and interpretable models. Regularization techniques (Tibshirani, 1996; Zou, 2006; Candes and Tao, 2007; Zhang, 2010; Zheng et al., 2014), which select a few relevant features in a sparse model for prediction, have thus been adapted from the independent framework to time series data (Medeiros and Mendes, 2016; Han and Tsay, 2020). In addition, greedy forward selection algorithms (Bühlmann, 2006; Wang, 2009; Fan and Lv, 2008; Ing and Lai, 2011) have also proved useful for a similar task involving dependent data (Ing, 2020).

However, the aforementioned methods are generally not applicable to unit-root nonstationary time series, prevalent in economics, finance, and environmental sciences. To apply these methods to unit-root time series, one must carefully transform the series under study into stationary ones. This step often involves multiple intricate unit-root tests since the underlying unit-root structure is typically unknown. In addition, it becomes even more challenging to transform the series into stationary ones when the data are driven by complex unit roots that exhibit persistent cyclic behavior. Diagnosing the order of integration and the frequency at which the series is integrated are far from straightforward and are sometimes sensitive to model specifications. Yet, persistent cyclic (or seasonal) time series are widely encountered in applications, such as the unemployment rate (Bierens, 2001), spot exchange rates (Al-Zoubi, 2008), entrepreneurship series (Faria et al., 2009), firms' capital structure (Al-Zoubi et al., 2018), sunspot numbers (Gil-Alana, 2009; Maddanu and Proietti, 2022), oil prices (Gil-Alana and Gupta, 2014), tourist arrivals (del Barrio Castro et al., 2022), and

CO_2 concentrations (Proietti and Maddanu, 2024), to name a few examples.

In this chapter, we study model selection for an autoregressive model with exogenous variables, known as the ARX model, when the dependent variable is a general unit-root nonstationary time series and the number of exogenous variables is large. The model considered is

$$(1-B)^a(1+B)^b\prod_{k=1}^{l}(1-2\cos\vartheta_k B+B^2)^{d_k}\psi_n(B)y_{t,n}=\sum_{j=1}^{p_n}\sum_{l=1}^{r_j^{(n)}}\beta_{l,n}^{(j)}x_{t-l,j}^{(n)}+\epsilon_{t,n}, \quad (2.1)$$

$t=1,\ldots,n$, where n is the sample size, B denotes the back-shift operator, a, b, l, and d_k are unknown nonnegative integers, ϑ_k are unknown real numbers in $(0,\pi)$, and $\psi_n(z)=1+\sum_{s=1}^{\iota_n}a_{s,n}z^s\neq 0$ for all $|z|\leq 1$, with $a_{s,n}$ being unknown real numbers and ι_n being an unknown nonnegative integer. In model (2.1), $\{\epsilon_{t,n}\}$ denotes a sequence of random errors with mean zero, $\{x_{t-l,j}^{(n)}\}$ and $\{\beta_{l,n}^{(j)}\}, 1\leq l\leq r_j^{(n)}, 1\leq j\leq p_n$, are observable exogenous variables and their respective unknown coefficients, and p_n and $r_j^{(n)}$ are known nonnegative integers. We adopt $y_{t,n}=0$ for $t\leq 0$ as the initial conditions, which are widely used in the literature for unit-root series (e.g. Chan and Wei, 1988). Let $d=a+b+2\sum_{k=1}^{l}d_k$. The number of lags of $y_{t,n}$, $m_n=\iota_n+d$, is assumed to be smaller than n, whereas the number of exogenous predictors, $p_n^*=\sum_{j=1}^{p_n}r_j^{(n)}$, can be much greater than n. Last but not least, we allow $\{x_{t,j}^{(n)}, 1\leq t\leq n\}$, for $1\leq j\leq p_n$, and $\{\epsilon_{t,n}\}$ to be conditionally heteroscedastic.

Due to the practical importance of the ARX model (2.1), numerous authors have investigated its model selection for the special cases when $d=0$ or $p_n^*=0$. When $d=0$, $y_{t,n}$ is stationary. In this case, under some strong sparsity conditions, the LASSO (Tibshirani, 1996) and the adaptive LASSO (Zou, 2006) have been shown to achieve model selection consistency (Han and Tsay, 2020; Medeiros and Mendes, 2016). In addition, Ing (2020) proved that the orthogonal greedy algorithm (OGA), used in conjunction with a high-dimensional information criterion (HDIC), is rate-optimal adaptive to unknown sparsity patterns. When

$p_n^* = 0$, model (2.1) reduces to a nonstationary AR model with unit roots. In this case, traditional information criteria such as AIC, BIC, and Fisher information criterion (FIC) can be employed to perform model selection (Ing et al., 2012; Tsay, 1984; Wei, 1992). More recently, Kock (2016) applied the adaptive LASSO to the Dickey-Fuller regression of fixed AR order under the specific case of a single unit root (i.e., $a = 1$, $b = d_1 = \ldots = d_l = 0$, and ι_n is a fixed positive integer).

Although there are methods available for the special cases when $d = 0$ or $p_n^* = 0$, applying them to model (2.1) in a general context remains a challenging task. As pointed out earlier, the existence of unknown ϑ_k makes it difficult to transform $\{y_{t,n}\}$ into an asymptotically stationary time series. Even worse, when applied to nonstationary time series, LASSO performs poorly due to its internal standardization of unit-root variables, which can "wash out the dependence of the stationary part" (Han and Tsay, 2020). In fact, due to the near-perfect correlation of some (or all) lagged variables in model (2.1) when $d > 0$, the strong irrepresentable condition, which is almost necessary and sufficient for LASSO to achieve selection consistency in high-dimensional regression models (Zhao and Yu, 2006), is no longer valid. This issue also undermines the effectiveness of other correlation-based feature selection methods, such as L_2-Boosting and OGA. Indeed, in Sections 2.3.1 and 2.3.2, we prove that both LASSO and OGA can fail to achieve variable selection consistency in the presence of unit roots. While AIC, BIC, and FIC are reliable methods for selecting the AR order when $d > 0$ and $p_n^* = 0$, they involve subset selection and are therefore not suitable for selecting exogenous variables when p_n^* is large, especially when $p_n^* \gg n$.

We address these difficulties by combining the strengths of the *least squares* method in unit-root AR models with *forward stepwise regression* (FSR, defined in Section 2.2) in high-dimensional regression models, and work directly with the observed nonstationary series.

Our procedure starts by rewriting (2.1) as

$$y_t = \sum_{i=1}^{q_n} \alpha_i y_{t-i} + \sum_{j=1}^{p_n} \sum_{l=1}^{r_j^{(n)}} \beta_l^{(j)} x_{t-l,j} + \epsilon_t, \tag{2.2}$$

where $q_n < n$ is a prescribed upper bound for m_n, $1 - \sum_{i=1}^{q_n} \alpha_i z^i = (1-z)^a (1+z)^b \prod_{k=1}^{l} (1 - 2\cos\vartheta_k z + z^2)^{d_k} \psi_n(z)$, and the dependence of y_t, α_i, $\beta_l^{(j)}$, $x_{t-l,j}$, and ϵ_t on n is suppressed for simplicity in notation. Then, FSR is used to sequentially choose the exogenous predictors *after* $y_{t-1}, \ldots, y_{t-q_n}$ are coerced into the model. By fitting an AR(q_n) model by least squares in advance, this approach handles the nonstationarity of $\{y_t\}$ without recourse to any tests for (complex) unit roots, thereby facilitating the implementation of FSR without being encumbered by the highly correlated lagged dependent variables. Next, we use HDIC to guide the stopping rule of FSR, and use a backward elimination method also based on HDIC, which we call Trim, to remove redundant exogenous predictors that have been previously included by FSR. Finally, we introduce a data-driven thresholding method referred to as DDT to weed out irrelevant lagged dependent variables. Throughout the chapter, the combined model selection procedure is called the FHTD algorithm. Under a strong sparsity condition, which assumes that the number of relevant predictors in model (2.2) is smaller than n, we establish the sure screening property of FSR and the selection consistency of FHTD. Since complex unit roots, conditional heteroscedasticity, and high dimensionality are allowed simultaneously, this is one of the most comprehensive results to date on model selection consistency established for the ARX model.

The rest of this chapter is organized as follows. We detail the FSR and FHTD algorithms in Section 2.2. The theoretical properties of these methods are given in Section 2.3; see Theorems 2.3.1–2.3.3. The finite-sample performance of the proposed methods is illustrated using simulated and two U.S. monthly macroeconomic datasets in Sections 2.4 and 2.5, respectively. Section 2.6 concludes. We have moved the proofs and auxiliary results to

Section 2.7. Nevertheless, it is noteworthy that, to tackle the nonstationary series, we employed a novel functional central limit theorem (FCLT) for linear processes driven by $\{\sum_{j=1}^{p_n}\sum_{l=1}^{r_j^{(n)}}\beta_l^{(j)}x_{t-l,j}+\epsilon_t\}$ and a uniform lower bound for the minimum eigenvalue of the sample covariance matrices associated with model (2.2). These theoretical foundations, crucial for Theorems 2.3.1–2.3.3, can be found in a recent paper (Huang et al., 2023).

The following notation is used throughout the chapter. For a matrix $\mathbf{A}$, $\lambda_{\min}(\mathbf{A})$, $\lambda_{\max}(\mathbf{A})$, $\|\mathbf{A}\|$, and $\mathbf{A}^\top$ denote its minimum eigenvalue, maximum eigenvalue, operator norm, and transpose, respectively. For a set J, $\sharp(J)$ is its cardinality. For two sequences of positive numbers, $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means $L < a_n/b_n < U$ for some $0 < L \le U < \infty$. For an event E, its complement and indicator function are denoted by E^c and $\mathbb{I}_E$, respectively. For $k \in \{1,2,\ldots\}$, $[k] = \{1,2,\ldots,k\}$. For $r \in \mathbb{R}$, $\lfloor r \rfloor$ is the largest integer $\le r$. For two real numbers a and b, $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$. For a vector $\mathbf{v}$, $\|\mathbf{v}\|$ denotes its Euclidean norm. For a random variable X, $\|X\|_q = (\mathbb{E}\,|X|^q)^{1/q}$. Generic absolute constants are denoted by C whose value may vary at different places.

2.2 The FHTD algorithm

Let $\mathbf{y}_n = (y_n, y_{n-1}, \ldots, y_{\bar{r}_n+1})^\top$, where $\bar{r}_n = \{\max_{1\le j\le p} r_j^{(n)}\} \vee q_n$. Define

$$\boldsymbol{o}_i = (y_{n-i}, y_{n-i-1}, \ldots, y_{\bar{r}_n-i+1})^\top$$

for $i = 1,2,\ldots,q_n$, and $\boldsymbol{x}_l^{(j)} = (x_{n-l,j}, x_{n-l-1,j}, \ldots, x_{\bar{r}_n-l+1,j})^\top$, $l = 1,2,\ldots,r_j^{(n)}$, $j = 1,2,\ldots,p_n$. Then, it follows from (2.2) that $\mathbf{y}_n = \mathbf{O}_n\boldsymbol{\alpha} + \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n := \boldsymbol{\mu}_n + \boldsymbol{\varepsilon}_n$, where $\mathbf{O}_n = (\boldsymbol{o}_1,\ldots,\boldsymbol{o}_{q_n})$, $\mathbf{X}_n = (\boldsymbol{x}_1^{(1)},\ldots,\boldsymbol{x}_{r_1^{(n)}}^{(1)},\ldots,\boldsymbol{x}_1^{(p_n)},\ldots,\boldsymbol{x}_{r_{p_n}^{(n)}}^{(p_n)})$, $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_{q_n})^\top$, $\boldsymbol{\varepsilon}_n = (\epsilon_n,\ldots,\epsilon_{\bar{r}_n+1})^\top$, and $\boldsymbol{\beta} = (\beta_1^{(1)},\ldots,\beta_{r_1^{(n)}}^{(1)},\ldots,\beta_1^{(p_n)},\ldots,\beta_{r_{p_n}^{(n)}}^{(p_n)})^\top$. Note that $\boldsymbol{\mu}_n$ can be

expressed as $(\mu_n, \ldots, \mu_{\bar{r}_n+1})^\top$, with $\mu_t = \sum_{i=1}^{q_n} \alpha_i y_{t-i} + v_{t,n}$ and

$$v_{t,n} = \sum_{j=1}^{p_n} \sum_{l=1}^{r_j^{(n)}} \beta_l^{(j)} x_{t-l,j}. \tag{2.3}$$

Let the candidate variable, $x_{t-l,j}$, be indexed by (j,l). FSR is an iterative algorithm that greedily chooses variables from $\bar{J} := \{(j,l) : j \in [p_n], l \in [r_j^{(n)}]\}$ after $y_{t-1}, \ldots, y_{t-q_n}$ are included in the regression model. Specifically, the algorithm begins with $\hat{J}_0 = \emptyset$ and generates $\hat{J}_m \subset \bar{J}$ via $\hat{J}_m = \hat{J}_{m-1} \cup \{(\hat{j}_m, \hat{l}_m)\}$, where $m \geq 1$ and

$$(\hat{j}_m, \hat{l}_m) = \arg\max_{(j,l) \in \bar{J} \setminus \hat{J}_{m-1}} \frac{n^{-1} |\mathbf{y}_n^\top (\mathbf{I} - \mathbf{H}_{[q_n] \oplus \hat{J}_{m-1}}) \boldsymbol{x}_l^{(j)}|}{(n^{-1} \boldsymbol{x}_l^{(j)\top} (\mathbf{I} - \mathbf{H}_{[q_n] \oplus \hat{J}_{m-1}}) \boldsymbol{x}_l^{(j)})^{1/2}}, \tag{2.4}$$

where $\mathbf{H}_{Q \oplus J}$ is the orthogonal projection matrix associated with the linear space spanned by $\{\boldsymbol{o}_l : l \in Q \subseteq [q_n]\} \cup \{\boldsymbol{x}_l^{(j)} : (j,l) \in J \subseteq \bar{J}\}$. In the sequel, we also use $Q \oplus J$ to denote a candidate model consisting of predictor variables $\{y_{t-i}, i \in Q\}$ and $\{x_{t-l,j}, (j,l) \in J\}$.

When m reaches a prescribed upper bound $K_n \leq p_n^*$, the algorithm stops and outputs the index set $\hat{J}_{K_n}$. Because the effects of the unit root have been neutralized by including the lagged dependent variables $y_{t-1}, \ldots, y_{t-q_n}$ beforehand, the algorithm is expected to exhibit reliable performance in including the set of *relevant* exogenous variables

$$\mathcal{J}_n = \{(j,l) : \beta_l^{(j)} \neq 0, l \in [r_j^{(n)}], j \in [p_n]\}.$$

However, $[q_n] \oplus \hat{J}_{K_n}$ may contain some irrelevant variables, in particular, when q_n or K_n is large compared to $\sharp(\mathcal{Q}_n)$ or $\sharp(\mathcal{J}_n)$, where $\mathcal{Q}_n = \{q : \alpha_q \neq 0, q \in [q_n]\}$ is the set of relevant lagged dependent variables. To alleviate the overfitting problem with FSR, we propose eliminating the irrelevant exogenous variables in $\hat{J}_{K_n}$ using HDIC and Trim, followed by DDT to remove the redundant lagged dependent variables in $[q_n]$. Given a candidate model

$Q \oplus J$, its HDIC value is given by

$$\text{HDIC}(Q \oplus J) = n \log \hat{\sigma}^2_{Q \oplus J} + [\sharp(J) + \sharp(Q)] w_{n,p_n}, \tag{2.5}$$

where $\hat{\sigma}^2_{Q \oplus J} = n^{-1} \mathbf{y}_n^\top (\mathbf{I} - \mathbf{H}_{Q \oplus J}) \mathbf{y}_n$ and w_{n,p_n}, penalty for the model complexity $\sharp(J) + \sharp(Q)$, depends on the sample size n as well as the number of candidate exogenous variables p_n^*.

Our approach is to first find a "promising" subset $\hat{J}_{\hat{k}_n}$ of $\hat{J}_{K_n}$ that minimizes the HDIC values along the FSR path $\{\hat{J}_1, \ldots, \hat{J}_{K_n}\}$, where

$$\hat{k}_n = \arg\min_{1 \le m \le K_n} \text{HDIC}([q_n] \oplus \hat{J}_m) \tag{2.6}$$

is an early stopping rule in which w_{n,p_n} is an increasing function of p_n^*. We then refine $\hat{J}_{\hat{k}_n}$ by comparing the HDIC values of $[q_n] \oplus \hat{J}_{\hat{k}_n}$ and $[q_n] \oplus (\hat{J}_{\hat{k}_n} \setminus \{(\hat{j}_i, \hat{l}_i)\}), 1 \le i \le \hat{k}_n$, to judge whether the marginal contribution of $(\hat{j}_i, \hat{l}_i)$ is significant enough to warrant its inclusion in the final model. The resultant refinement of $\hat{J}_{\hat{k}_n}$ is

$$\hat{\mathcal{J}}_n = \{(\hat{j}_i, \hat{l}_i) : 1 \le i \le \hat{k}_n, \text{HDIC}([q_n] \oplus (\hat{J}_{\hat{k}_n} \setminus \{(\hat{j}_i, \hat{l}_i)\})) > \text{HDIC}([q_n] \oplus \hat{J}_{\hat{k}_n})\}, \tag{2.7}$$

and the method is called "Trim."

For $J \in \bar{J}$, define $\mathbf{x}_t(J) = (x_{t-l,j} : (j,l) \in J)^\top$ and $\mathbf{w}_t(J) = (y_{t-1}, \ldots, y_{t-q_n}, \mathbf{x}_t^\top(J))^\top$. Then the least squares estimates of the regression coefficients for model $[q_n] \oplus \hat{\mathcal{J}}_n$ is

$$(\hat{\alpha}_1(\hat{\mathcal{J}}_n), \ldots, \hat{\alpha}_{q_n}(\hat{\mathcal{J}}_n), \hat{\boldsymbol{\beta}}^\top(\hat{\mathcal{J}}_n))^\top = \left(\sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(\hat{\mathcal{J}}_n) \mathbf{w}_t^\top(\hat{\mathcal{J}}_n) \right)^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(\hat{\mathcal{J}}_n) y_t.$$

With the estimated AR coefficients, $\hat{\alpha}_i(\hat{\mathcal{J}}_n), 1 \le i \le q_n$, we suggest using a data-driven

thresholding (DDT) method,

$$\hat{\mathcal{Q}}_n = \{1 \le q \le q_n : |\hat{\alpha}_q(\hat{\mathcal{J}}_n)| \ge \hat{H}_n\}, \tag{2.8}$$

to weed out redundant AR variables, where $\hat{H}_n$ is a data-driven thresholding value depending on $\hat{\mathcal{J}}_n$ and q_n; see Section 2.3.2. Note that identifying $\hat{\mathcal{Q}}_n$ is crucial for accurate prediction because an overfitted model tends to have a larger mean squared prediction error, especially in tackling nonstationary time series where the cost of overfitting is more prominent (see Example 2.3.3). The final estimated model is $\hat{N}_n = \hat{\mathcal{Q}}_n \oplus \hat{\mathcal{J}}_n$. The above procedure, which combines FSR, HDIC, Trim, and DDT, is referred to as FHTD.

Some notable existing methods related to FHTD are worth mentioning. First, Chudik et al. (2018) have also employed a forward selection method similar to (2.4) in the One Covariate at a Time Multiple Testing (OCMT) procedure, which can control the false positive rate and the false discovery rate in high-dimensional linear regression models. However, their analysis of OCMT does not account for the scenario in which the pre-selected covariates exhibit near-perfect correlations. Note also that (2.4) simplifies to the forward regression algorithm in Wang (2009) when $[q_n]$ becomes an empty set, and it further simplifies to OGA in Ing and Lai (2011) if the orthogonal projection matrix in the denominator is removed. Second, HDIC becomes BIC if $w_{n,p_n} = \log n$ and AIC if $w_{n,p_n} = 2$. However, failing to account for potential spuriousness of the greedily chosen variables among p_n^* candidate variables, AIC and BIC may result in serious overfitting in the case of $p_n^* \gg n$. Third, in the context of independent observations, (2.5) and (2.7) have been employed by Ing and Lai (2011) to eliminate the redundant variables introduced by OGA for high-dimensional regression models. This combined technique is called OGA+HDIC+Trim by the authors. Indeed, under an appropriate "beta-min" condition, it can be derived from an argument in Ing (2020) that, with probability approaching 1, OGA+HDIC+Trim is capable of directly selecting $\mathcal{J}_n$ and $\mathcal{Q}_n$ in *stationary* ARX models without having to fit an AR(q_n) model using

least squares beforehand. However, the effectiveness of OGA+HDIC+Trim in identifying $\mathcal{J}_n$ and $\mathcal{Q}_n$ is significantly compromised under model (2.1); see Examples 2.4.1 and 2.4.2 of Section 2.4. This difficulty is also encountered by LASSO and adaptive LASSO, highlighting the inherent challenges in model selection for high-dimensional nonstationary ARX models with highly correlated lagged dependent variables.

Finally, it is important to point out that the distinctiveness of FHTD does not come from the methods it uses, but rather from the innovative way in which these methods are combined to tackle a highly challenging model selection problem outlined in Section 2.1. This problem is notoriously difficult to overcome and poses a significant obstacle for most existing high-dimensional methods. Furthermore, a comprehensive analysis of these methods in non-standard scenarios is necessary to provide a theoretical justification for FHTD, particularly when some predictors display near-perfect correlations and all of them are conditionally heteroscedastic. In the next section, we show that FSR boasts the sure screening property while $\hat{N}_n$ consistently estimate $N_n = \mathcal{J}_n \oplus \mathcal{Q}_n$.

2.3 Screening and selection consistency

In this section, we present the sure-screening property of FSR and the model selection consistency of FHTD in Sections 2.3.1 and 2.3.2. To this end, we will consider in Section 2.3.3 Assumptions (A1)–(A6) concerning model (2.2). For a detailed discussion of (A1)–(A6), see Section 2.3.3 and Section 2.7.1.

2.3.1 The sure screening property of FSR

In addition to (A1)–(A6), we require a *strong sparsity* condition **(SS$_{\mathrm{X}}$)** on $\{\beta_l^{(j)}\}$ to ensure the sure screening property of FSR. Note that (A1)–(A6) imply that for some $\eta \geq 2$ and

$q_0 > 2$,

$$\sup_t \mathbb{E}\,|\epsilon_t|^{2\eta} < C, \tag{2.9}$$

$$\max_{1\le s\le p_n} \sup_t \mathbb{E}\,|x_{t,s}|^{2\eta q_0} < C, \tag{2.10}$$

and $p_n^* \asymp n^\nu$, where $\nu \in [1, \eta/2]$, so that p_n^* can be greater than n. Now we state $\mathbf{(SS_X)}$.

$\mathbf{(SS_X)}$ $s_0 = \sharp(\mathcal{J}_n)$ and $\min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|$ obey

$$\frac{s_0^{1/2} {p_n^*}^{\bar\theta}}{n^{1/2}} = o(\min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|), \tag{2.11}$$

where $\bar\theta = \max\{2/(q_0\eta), (q_0+1)/(2\eta q_0)\}$.

Note that Medeiros and Mendes (2016) used a similar condition to derive the selection consistency of adaptive LASSO when $\{y_t\}$ is stationary. $\mathbf{(SS_X)}$ is less stringent than their strong sparsity condition. See the discussion of Section 2.7.1.

As the first step toward model selection consistency, Theorem 2.3.1 below shows that FSR asymptotically screens all relevant variables.

Theorem 2.3.1. *Assume that* (A1)–(A6) *and* $\mathbf{(SS_X)}$ *hold. Then, for*

$$K_n \asymp (n/{p_n^*}^{2\bar\theta})^\varsigma, \tag{2.12}$$

where $1/3 < \varsigma < 1/2$,

$$\lim_{n\to\infty} P(\mathcal{J}_n \subseteq \hat{J}_{K_n}) = 1. \tag{2.13}$$

It is important to note that forcing the lagged dependent variables in FSR is essential to the sure-screening property. As illustrated in the following example, the (conventional)

OGA fails to include all relevant variables when it is used to choose the lagged dependent variables and exogenous variables simultaneously in the presence of unit-roots.

Example 2.3.1. Consider a special case of model (2.2),

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \sum_{j=1}^{p_n} \beta_j x_{t-1,j} + \epsilon_t, \tag{2.14}$$

where $\alpha_1 = 1 + a$, $\alpha_2 = -a$, $|a| < 1$, $\beta_j = 0$, $1 \le j \le p_n$, and $\{(x_{t,1}, \ldots, x_{t,p_n}, \epsilon_t)^\top\}$ is a sequence of independent normal random vectors with mean zero and identity covariance matrix. It is easy to see that (2.14) is an AR(2) model with a characteristic polynomial $(1 - az)(1-z)$ which has a unit root of 1. In addition, all $x_{t-1,j}, 1 \le j \le p_n$, are redundant. Under this model, when OGA is directly applied to the set of candidate variables $\{y_{t-1}, y_{t-2}, x_{t-1,j} : 1 \le j \le p_n\}$, one of the relevant variables $\{y_{t-1}, y_{t-2}\}$ will *not* be included in the OGA path.

To see this, let

$$F_{1,n} = (\mathbf{y}_n^\top \boldsymbol{o}_1)^2 / \|\boldsymbol{o}_1\|^2, \quad F_{2,n} = (\mathbf{y}_n^\top \boldsymbol{o}_2)^2 / \|\boldsymbol{o}_2\|^2$$

and

$$Q_{j,n} = (\mathbf{y}_n^\top \boldsymbol{x}^{(j)})^2 / \|\boldsymbol{x}^{(j)}\|^2, \quad 1 \le j \le p_n,$$

where, analogous to Section 2.2, we define $\mathbf{y}_n = (y_n, \ldots, y_3)^\top$, $\boldsymbol{o}_1 = (y_{n-1}, \ldots, y_2)^\top$, $\boldsymbol{o}_2 = (y_{n-2}, \ldots, y_1)^\top$, $\boldsymbol{x}^{(j)} = (x_{n-1,j}, \ldots, x_{2,j})^\top$, $1 \le j \le p_n$. It is not difficult to show that

$$\frac{F_{1,n}}{n^2} \Rightarrow (1-a)^{-2} \int_0^1 w^2(t)dt, \quad \frac{F_{2_n}}{n^2} \Rightarrow (1-a)^{-2} \int_0^1 w^2(t)dt, \tag{2.15}$$

where $w(t)$ is the standard Brownian motion and $\Rightarrow$ denotes convergence in law. Moreover,

we show in Section 2.7.4 that

$$\frac{1}{n}(F_{1,n} - F_{2,n}) \to \frac{1+2a}{1-a^2} \quad \text{in probability.} \tag{2.16}$$

By Bernstein's inequality, it can be shown that $\max_{1\le j\le p_n} Q_{j,n} = O_p(n \log p_n)$. Hence, if $a > -0.5$, then (2.15)–(2.16) imply that with probability tending to 1, y_{t-1} will be selected in the initial iteration of OGA, provided that $\log p_n = o(n)$.

Assuming that y_{t-1} is already included by OGA, define $\boldsymbol{\epsilon} = (\mathbf{I} - \boldsymbol{o}_1\boldsymbol{o}_1^\top/\|\boldsymbol{o}_1\|^2)\mathbf{y}_n$, which is the residual vector obtained by regressing y_t on y_{t-1}. It can be shown that $(\boldsymbol{o}_2^\top\boldsymbol{\epsilon})^2/\|\boldsymbol{o}_2\|^2 = O_p(1)$ and for some small $\underline{c} > 0$, $P(\max_{1\le j\le p_n}[(\boldsymbol{x}^{(j)})^\top\boldsymbol{\epsilon}]^2/\|\boldsymbol{x}^{(j)}\|^2 > \underline{c}\log p_n) \to 1$. Hence, the probability of choosing y_{t-2} in the second OGA iteration approaches 0 provided $\log p_n \to \infty$. By a similar argument, y_{t-2} will not be selected by OGA in the first K_n iterations when $p_n \gg n \gg K_n$ with probability approaching 1. Thus, while y_{t-1} will be selected by OGA with probability tending to 1, it is very difficult for OGA to choose the other relevant lagged dependent variable in the presence of unit roots. If $a < -0.5$, y_{t-2} will enter the model in the first iteration and y_{t-1} will then be neglected by OGA due to the same argument.

The above example not only highlights the constraint of OGA in handling nonstationary time series but also suggests that proving Theorem 3.1 requires a distinct skill set compared to the methodologies employed by Bühlmann (2006), Ing and Lai (2011), and Ing (2020). While these previous works have been instrumental in analyzing greedy-type methods in high-dimensional stationary time series, they rely heavily on the convergence rates of the "population" OGA and its "semi-population" version (see Section 6 of Bühlmann (2006), Sections 2 and 3 of Ing and Lai (2011), or Section 2 and Appendix A of Ing (2020)). However, the population OGA can hardly be defined for nonstationary time series due to the varying covariances between the input variables and between the input and output variables over time.

To resolve this dilemma, we introduce the "noiseless" FSR, which is FSR of (2.4) with $\mathbf{y}_n$

in the numerator replaced by its noiseless counterpart $\boldsymbol{\mu}_n$; see Section 2.7.2 for details. We derive in Section 2.7.2 the rate of convergence of the corresponding "noiseless" mean squared error

$$\hat{a}_m = n^{-1}\boldsymbol{\mu}_n^\top(\mathbf{I} - \mathbf{H}_{[q_n]\oplus J_m})\boldsymbol{\mu}_n, \tag{2.17}$$

as the number of iteration, m, increases, where J_m (defined in Eq. (24) of Section 2.7.2) is the set of exogenous variables determined by the noiseless FSR after m iterations. The rate of convergence of $\hat{a}_m$, together with a probability bound for

$$\max_{\sharp(J)\le \bar{K}_n} \lambda_{\min}^{-1}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{w}_t(J)\mathbf{w}_t^\top(J)\right), \tag{2.18}$$

developed in Theorem 4.1 of Huang et al. (2023), leads to a convergence rate of $\hat{a}_m$'s "semi-noiseless" counterpart,

$$\hat{s}_m = n^{-1}\boldsymbol{\mu}_n^\top(\mathbf{I} - \mathbf{H}_{[q_n]\oplus \hat{J}_m})\boldsymbol{\mu}_n, \tag{2.19}$$

where $\bar{K}_n$ in (2.18) satisfies $\bar{K}_n = o(n^{1/2}/p_n^{*\bar{\theta}})$ and $\bar{K}_n \gg K_n + s_0$, and the sole distinction between $\hat{a}_m$ and $\hat{s}_m$ is that the infeasible J_m in the former is replaced with the data-driven $\hat{J}_m$ in the latter. As we will see later, the rate of convergence of $\hat{s}_m$ serves as the key vehicle for us to prove (2.13).

In sharp contrast to conventional high-dimensional models where the sample covariance matrix of the explanatory variables can be accurately approximated by a non-random and positive definite matrix, the sample covariance matrix, $n^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{w}_t(J)\mathbf{w}_t^\top(J)$, in (2.18) lacks a non-random limit due to the presence of highly correlated lagged dependent variables. Therefore, our probability bound for (2.18) highlights another interesting aspect of our analysis. This probability bound is based on an FCLT for linear processes driven by

$\delta_t = \delta_{t,n} = v_{t,n} + \epsilon_t$ and moment bounds for quadratic forms in linear processes driven by δ_t, as detailed in Huang et al. (2023), where $v_{t,n}$ is defined in (2.3). It is noteworthy that despite accounting for high dimensionality and general conditional heteroscedasticity, our findings regarding (2.18) align with (3.10) of Lai and Wei (1982), where a finite-order nonstationary AR model with a conditionally homogeneous error is considered.

2.3.2 Selection consistency

This section starts by establishing the selection consistency of $\hat{\mathcal{J}}_n$ defined in (2.7), which is a backward elimination method based on a refinement, $\hat{J}_{\hat{k}_n}$, of $\hat{J}_{K_n}$; see (2.5) and (2.6). To this end, we impose a sparsity condition slightly stronger than **(SS$_{\mathrm{X}}$)**.

(SS) There exists $d_n/\log n \to \infty$ such that

$$\frac{s_0^{1/2} {p_n^*}^{\bar{\theta}} d_n^{1/2}}{n^{1/2}} = o(\min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|). \tag{2.20}$$

Note that the left-hand side in (2.20) is larger than that of (2.11) by a factor of about $(\log n)^{1/2}$. Further discussion of **(SS)** is deferred to Section 2.7.1. Based on **(SS)**, among other conditions, Theorem 2.3.2 ensures the consistency of Trim in selecting the exogenous variables.

Theorem 2.3.2. *Assume that the assumptions of Theorem 2.3.1 hold with* **(SS$_{\mathrm{X}}$)** *replaced by* **(SS)**. *Let the w_{n,p_n} in* (2.5) *satisfy*

$$\frac{w_{n,p_n}}{{p_n^*}^{2\bar{\theta}}} \to \infty \quad \textit{and} \quad \frac{w_{n,p_n}}{{p_n^*}^{2\bar{\theta}}} = O((d_n/\log n)^{1-\delta}) \textit{ for any } 0 < \delta < 1. \tag{2.21}$$

Then, $\hat{k}_n$ and $\hat{\mathcal{J}}_n$ defined in (2.6) *and* (2.7) *satisfy*

$$\lim_{n\to\infty} P(\mathcal{J}_n \subseteq \hat{J}_{\hat{k}_n}) = 1, \tag{2.22}$$

$$\lim_{n\to\infty} P(\hat{\mathcal{J}}_n = \mathcal{J}_n) = 1. \tag{2.23}$$

As an early stopping rule for FSR, $\hat{J}_{\hat{k}_n}$ not only preserves $\hat{J}_{K_n}$'s sure screening property (see (2.22)), but also substantially suppresses the impact of spurious variables greedily chosen by FSR, resulting in reliable performance of Trim. With the help of (2.23), we are now in a position to develop the consistency of DDT in selecting the AR variables. Likewise, we rely on a strong sparsity condition on the AR coefficients.

(SS$_{\mathrm{A}}$) $\min_{q\in\mathcal{Q}_n} |\alpha_q|$, s_0, and $\underline{s}_0$ obey

$$\frac{\max\{q_n^{3/2}/\sqrt{n},\ [(s_0+q_n)^{1/2} \wedge (\underline{s}_0^{1/2} q_n^{1/\eta})]\}}{n^{1/2}} = o(\min_{q\in\mathcal{Q}_n} |\alpha_q|), \tag{2.24}$$

where $\underline{s}_0 = \sharp(\mathcal{D}_0)$ and $\mathcal{D}_0 = \{j : (j,l) \in \mathcal{J}_n\}$.

Compared with **(SS)** or **(SS$_{\mathrm{X}}$)**, a distinct feature of **(SS$_{\mathrm{A}}$)** is that it allows for a smaller lower bound for the non-zero coefficients, enabling detection of weaker signals. In particular, since the spurious exogenous variables chosen by FSR from among p_n^* candidates have been (asymptotically) eliminated after the HDIC and Trim steps, p_n^* is now removed from the lower bound for $\min_{q\in\mathcal{Q}_n} |\alpha_q|$ in which the much smaller q_n is used instead.

Theorem 2.3.3. *Assume that the assumptions of Theorem 2.3.2,* (2.21)*, and* **(SS$_{\mathrm{A}}$)** *hold. Then, the DDT procedure, $\hat{\mathcal{Q}}_n$, defined in* (2.8)*, satisfies*

$$\lim_{n\to\infty} P(\hat{\mathcal{Q}}_n = \mathcal{Q}_n) = 1, \tag{2.25}$$

provided that the data-driven threshold satisfies

$$\hat{H}_n = \frac{\max\{q_n^{3/2}/\sqrt{n},\ [(q_n+\hat{s}_0)^{1/2} \wedge (\hat{\underline{s}}_0^{1/2} q_n^{1/\eta})]\}}{n^{1/2}} \tilde{d}_n, \tag{2.26}$$

where $\tilde{d}_n$ diverges to ∞ at an arbitrarily slow rate, $\hat{s}_0 = \sharp(\hat{\mathcal{J}}_n)$, and $\underline{\hat{s}}_0 = \sharp(\{i : (i,l) \in \hat{\mathcal{J}}_n\})$.

Note that if $q_n = o(n^{1/3})$, then the $\hat{H}_n$ in (2.26) simplifies to

$$\frac{[(q_n + \hat{s}_0)^{1/2} \wedge (\underline{\hat{s}}_0^{1/2} q_n^{1/\eta})]\tilde{d}_n}{n^{1/2}}. \tag{2.27}$$

We stress that the key to the success of DDT is to turn the infeasible thresholding value in **($\mathbf{SS}_{\mathrm{A}}$)** into the feasible one, $\hat{H}_n$, by replacing the unknown s_0 and $\underline{s}_0$ with their consistent estimates $\hat{s}_0$ and $\underline{\hat{s}}_0$. Combining Theorems 2.3.2 and 2.3.3 yields that FHTD asymptotically captures the relevant AR and exogenous covariates despite complex unit roots, conditional heteroscedasticity, and a large pool of candidate variables, resolving the difficulties described in the Introduction.

In Section 2.3.1, we have demonstrated that OGA overlooks certain relevant variables when applied directly to model (2.2). The following example highlights the contributions of Theorems 2.3.2 and 2.3.3 by illustrating that, regardless of the penalty sequence $\{\lambda_n\}$, the LASSO method lacks selection consistency in the presence of a unit root.

Example 2.3.2. Consider the model $y_t = \beta_1^* y_{t-1} + \beta_2^* y_{t-2} + \beta_3^* x_{t-1} + \epsilon_t$, $t = 1, 2, \ldots, n$, where $(\epsilon_t, x_t)^\top$ are i.i.d. Gaussian with zero mean and an identity covariance matrix, and $\beta_1^* = \beta_3^* = 1$ and $\beta_2^* = 0$. If LASSO is applied to estimate $(\beta_1^*, \beta_2^*, \beta_3^*)$, and $\hat{\boldsymbol{\beta}}^{(\lambda_n)} = (\hat{\beta}_1^{(\lambda_n)}, \hat{\beta}_2^{(\lambda_n)}, \hat{\beta}_3^{(\lambda_n)})^\top$ is its estimate corresponding to the penalty λ_n; namely,

$$\hat{\boldsymbol{\beta}}^{(\lambda_n)} \in \arg\min_{\{\beta_j\}_{j=1}^3} \sum_{t=3}^{n} (y_t - \beta_1 y_{t-1} - \beta_2 y_{t-2} - \beta_3 x_{t-1})^2 + \lambda_n \sum_{j=1}^{3} |\beta_j|,$$

then LASSO will *not* exhibit model selection consistency, as described in equations (2.23) and (2.25). This lack of consistency holds true whether the sequence $\{\lambda_n\}$ is chosen such that (a) $\limsup_{n\to\infty} \lambda_n/n = \infty$, (b) $\liminf_{n\to\infty} \lambda_n/n = 0$, or (c) $\lambda_n \asymp n$. In fact, one can

demonstrate that for any sequence $\{\lambda_n\}$ satisfying (a), (b), or (c), we always have

$$\liminf_{n\to\infty} P(\hat{\beta}_1^{(\lambda_n)} \neq 0, \hat{\beta}_2^{(\lambda_n)} = 0, \hat{\beta}_3^{(\lambda_n)} \neq 0) \leq \frac{1}{2}. \tag{2.28}$$

The proof of (2.28) can be found in Section 2.7.4.

It is also worth noting that (2.25), achieving consistency for subset selection, is more desirable for prediction than order selection consistency, whose corresponding model may still contain redundant AR variables. To the best of our knowledge, this type of consistency has not been reported elsewhere, even when q_n is bounded, and $v_{t,n}$ (see (2.3)) is dropped from model (2.2). In the following example, we elucidate why achieving consistency in subset selection can offer significantly greater advantages compared to order selection from a predictive standpoint.

Example 2.3.3. Consider the model

$$y_t = \sum_{j=1}^{k} \beta_j y_{t-j} + \epsilon_t, \quad t = 1, \ldots, n, \tag{2.29}$$

where $k \geq 1$ is an integer, $\beta_1 = \cdots = \beta_{k-1} = 0$, $\beta_k = 1$, and ϵ_t are i.i.d. random variables with a mean of zero and a constant variance of $0 < \sigma^2 < \infty$. It is evident that model (2.29) is a nonstationary AR(k) model containing k unit roots. If k is known or can be consistently estimated by an order selection criterion such as BIC, then it is natural to predict y_{n+1} using the least squares predictor, $\hat{y}_{n+1}(k) = \mathbf{y}_n^\top(k)\hat{\boldsymbol{\beta}}(k)$, where $\mathbf{y}_t(k) = (y_t, \ldots, y_{t-k+1})^\top$ and $\hat{\boldsymbol{\beta}}(k) = (\sum_{t=k}^{n-1} \mathbf{y}_t(k)\mathbf{y}_t^\top(k))^{-1} \sum_{t=k}^{n-1} \mathbf{y}_t(k) y_{t+1}$. The performance of $\hat{y}_{n+1}(k)$ can be evaluated using its mean squared prediction error (MSPE), defined as $\text{MSPE}_k = \mathbb{E}(y_{n+1} - \hat{y}_{n+1}(k))^2$. Assume $\mathbb{E}\,|\epsilon_1|^s < \infty$ for some $s > 4$, and a smoothness condition on ϵ_t described in Section 2 of Ing et al. (2010). Then, by extending an argument

used in Ing (2001), Ing et al. (2010), and Ing and Yang (2014), it can be shown that

$$\lim_{n\to\infty} n(\mathrm{MSPE}_k - \sigma^2) = \sigma^2 \mathrm{plim}_{n\to\infty} \frac{\log\det(\sum_{t=k}^{n-1} \mathbf{y}_t(k)\mathbf{y}_t^\top(k))}{\log n} = 2k\sigma^2, \tag{2.30}$$

where the second equality is ensured by Theorem 5 of Wei (1987).

Alternatively, if a method can consistently select the non-zero coefficient β_k while excluding the redundant ones, such as the FHTD, the least squares predictor,

$$\tilde{y}_{n+1}(k) = y_{n+1-k}\tilde{\beta}_k = \frac{y_{n+1-k}\sum_{t=k}^{n-1} y_{t-k+1}y_{t+1}}{\sum_{t=k}^{n-1} y_{t-k+1}^2},$$

would emerge as another appropriate predictor for y_{n+1}, where $\tilde{\beta}_k$ is the least squares estimate of β_k obtained from regressing y_t on y_{t-k}. By an argument similar to that used to prove (2.30), it can be shown that $\widetilde{\mathrm{MSPE}}_k = \mathbb{E}(y_{n+1} - \tilde{y}_{n+1}(k))^2$ obeys

$$\lim_{n\to\infty} n(\widetilde{\mathrm{MSPE}}_k - \sigma^2) = \sigma^2 \mathrm{plim}_{n\to\infty} \frac{\log\det(\sum_{t=1}^{n-k} y_t^2)}{\log n} = 2\sigma^2. \tag{2.31}$$

Equations (2.30) and (2.31) reveal that the least squares predictor constructed from a consistent order selection method could indeed lead to significantly higher MSPE than the one derived from a consistent subset selection method, especially when the underlying unit-root model contains many irrelevant lagged dependent variables. Moreover, in the stationary case ($0 < |\beta_k| < 1$), the constant **2** on the right-hand side of (2.30) and (2.31) reduces to **1**; see Ing (2003). Therefore, excessive fitting in a unit-root time series could result in a notably larger MSPE than in a stationary series.

2.3.3 Model assumptions

Considering model (2.2), let $x_{t,s}$ for $1 \le s \le p_n$ and ϵ_t be $\mathcal{F}_t$-measurable random variables, where $\{\mathcal{F}_t\}$ is an increasing sequence of σ-fields representing available information up to

time t. We impose the following assumptions.

(A1) $\{\epsilon_t, \mathcal{F}_t\}$ is a martingale difference sequence (m.d.s.) with $\mathbb{E}\,\epsilon_t^2 = \sigma^2$ and

$$\epsilon_t^2 - \sigma^2 = \sum_{j=0}^{\infty} \theta_j^\top e_{t-j}, \tag{2.32}$$

where θ_j are l_0-dimensional real vectors such that

$$\sum_{j=0}^{\infty} \|\theta_j\| \leq C, \tag{2.33}$$

with l_0 being a fixed positive integer, and $\{e_t, \mathcal{F}_t\}$ is an l_0-dimensional m.d.s. with

$$\sup_t \mathbb{E}\,\|e_t\|^\eta \leq C, \text{ for some } \eta \geq 2. \tag{2.34}$$

(A2) For each $1 \leq s \leq p_n$, $\{x_{t,s}\}_{-\infty<t<\infty}$ is a covariance stationary time series with mean zero and admits a one-sided moving average representation,

$$x_{t,s} = \sum_{k=0}^{\infty} p_{k,s}\pi_{t-k,s}, \tag{2.35}$$

where $p_{0,s} = 1$, $\{\pi_{t,s}, \mathcal{F}_t\}$ is an m.d.s. and

$$\sum_{k=0}^{\infty} \max_{1\leq s\leq p_n} \sqrt{k}|p_{k,s}| \leq C. \tag{2.36}$$

Moreover, for $0 \leq s_1 \leq s_2 \leq p_n$ and $s_1 + s_2 \geq 1$,

$$\pi_{t,s_1}\pi_{t,s_2} - \sigma_{s_1,s_2} = \sum_{j=0}^{\infty} \theta_{j,s_1,s_2}^\top e_{t-j,s_1,s_2}, \tag{2.37}$$

where $\pi_{t,0} = \epsilon_t$, $\sigma_{s_1,s_2} = \mathbb{E}(\pi_{t,s_1}\pi_{t,s_2})$, θ_{j,s_1,s_2} are l_{s_1,s_2}-dimensional real vectors, with

l_{s_1,s_2} being a fixed positive integer, such that

$$\sum_{j=0}^{\infty} \|\theta_{j,s_1,s_2}\| \leq C, \tag{2.38}$$

and $\{e_{t,s_1,s_2}, \mathcal{F}_t\}$ is a l_{s_1,s_2}-dimensional m.d.s. satisfying for some $q_0 > 2$,

$$\sup_t \mathbb{E} \|e_{t,s_1,s_2}\|^{q_0\eta} \leq C, \text{ if } \min\{s_1, s_2\} > 0, \tag{2.39}$$

$$\sup_t \mathbb{E} \|e_{t,s_1,s_2}\|^{2q_0\eta/(1+q_0)} \leq C, \text{ if } \min\{s_1, s_2\} = 0, \tag{2.40}$$

where η is defined in (2.34). Note that C does not depend on s_1, s_2 in the above.

(A3) There exists a positive definite sequence, $\{\gamma_h\}_{-\infty<h<\infty}$, of real numbers such that

$$\lim_{n\to\infty} \sum_{h=0}^{\infty} |\gamma_{h,n} - \gamma_h| = 0, \tag{2.41}$$

where $\gamma_{h,n} = \mathbb{E}(\delta_t \delta_{t+h})$ and $\delta_t = \delta_{t,n} = v_{t,n} + \epsilon_t$, noting that $v_{t,n}$ is defined in (2.3).

(A4) $\sum_{j=1}^{p_n} \sum_{l=1}^{r_j^{(n)}} |\beta_l^{(j)}| \leq C$ and $\sum_{j=1}^{\iota_n} j|a_j| \leq C$, where $a_j = a_{j,n}$ is defined after (2.1).

(A5) $\max_{1\leq j\leq p_n} r_j^{(n)} = o(n^{1/2})$, $p_n^* \asymp n^{\nu}$, and $q_n = o(n^{1/2-\theta_o})$, where $\nu \in [1, \eta/2]$ and $\theta_o = \nu(1+q_0)/(2\eta q_0)$.

Assumption (A1) implies that (2.9) holds and is satisfied by many conditionally heteroscedastic processes, such as the stationary GJR-GARCH model with a finite 2η-th moment. Assumption (A2) necessitates that $\{x_{t,s}\}$ follows an MA(∞) process driven by the conditionally heteroscedastic innovations $\{\pi_{t,s}\}$, while also requiring that the process possesses a finite $2\eta q_0$-th moment; see (2.10). Furthermore, it accommodates the possibility that $(\epsilon_t, \pi_{t,1} \ldots, \pi_{t,p_n})^\top$ constitutes a multivariate GARCH process, with the diagonal VEC

model introduced by Bollerslev et al. (1988) being a particular instance within this framework. Assumption (A3) is used to derive the FCLT for the multivariate linear process driven by $\{\delta_t\}$, and Assumption (A4), known as the weak sparsity condition, frequently finds application in high-dimensional data analysis literature. Finally, Assumption (A5) allows that the covariate dimension is at least of the same order as n, and can be much larger than n if $\eta > 2$. It also permits that q_n, the prescribed upper bound of the number of AR variables, increases to ∞ at a rate slower than $n^{1/2}$. For a more detailed and comprehensive exploration of (A1)–(A5), readers are referred to Section 2.7.1 and Section 2.1 of Huang et al. (2023).

Apart from (A1)–(A5), we also require (A6), which assumes the covariance structures of $x_{t,j}$ and the stationary component, $z_t = [\psi^{-1}(B)\phi(B)]y_t$, of y_t, where

$$\phi(z) = (1-z)^a(1+z)^b \prod_{k=1}^{l}(1 - 2\cos\vartheta_k z + z^2)^{d_k}\psi(z),$$

with $\psi(z) = \psi_n(z)$ defined after (2.1). Since $\psi(z) \neq 0$ for all $|z| \leq 1$, by the second part of (A4) and Theorem 3.8.4 of Brillinger (1975), z_t can be expressed as $z_t = \sum_{j=0}^{t-1} b_j \delta_{t-j}$, with $b_0 = 1$, $\sum_{j=0}^{\infty} b_j z^j \neq 0, |z| \leq 1$, and $\sum_{j=0}^{\infty} |jb_j| \leq C$. Define $z_{t,\infty} = \sum_{j=0}^{\infty} b_j \delta_{t-j}$, $\mathbf{z}_{t,\infty}^{\top}(k) = (z_{t-1,\infty}, \ldots, z_{t-k,\infty})$,

$$\mathbf{\Gamma}_n(J) = \mathbb{E}\left\{ \begin{pmatrix} \mathbf{z}_{t,\infty}(q_n - d) \\ \mathbf{x}_t(J) \end{pmatrix} \left(\mathbf{z}_{t,\infty}^{\top}(q_n - d), \quad \mathbf{x}_t^{\top}(J) \right) \right\}, \quad J \subseteq \bar{J},$$

and for $(i,l) \notin J$, $\mathbf{g}_J^{\top}(i,l) = (\mathbb{E}(\mathbf{z}_{t,\infty}^{\top}(q_n - d)x_{t-l,i}), \mathbb{E}(\mathbf{x}_t^{\top}(J)x_{t-l,i}))$. Now, (A6) is presented as follows:

(A6)

$$\max_{\sharp(J) \leq K_n} \lambda_{\min}^{-1}(\mathbf{\Gamma}_n(J)) \leq C, \tag{2.42}$$

and

$$\sum_{s=1}^{q_n-d} \max_{\sharp(J)\le K_n,(i,l)\notin J} |a_{s,J}(i,l)| + \max_{\sharp(J)\le K_n,(i,l)\notin J} \sum_{(i^*,l^*)\in J} |a_{(i^*,l^*)}(i,l)| \le C, \tag{2.43}$$

where $(a_{1,J}(i,l),\ldots,a_{q_n-d,J}(i,l),(a_{(i^*,l^*)}(i,l):(i^*,l^*)\in J))^\top = \mathbf{\Gamma}_n^{-1}(J)\mathbf{g}_J(i,l)$.

In Section 2.7.1, we provide illustrative examples to demonstrate the applicability of (A6). Specifically, we establish that (2.43) remains valid, even when model (2.2) contains highly correlated lag-dependent variables and exogenous variables with strong correlations.

2.4 Simulation studies

In this section, we examine the model selection performance of FHTD using data generated from model (2.1), with coefficients, covariates, and error terms specified below. For the purpose of comparison, we also employ several existing high-dimensional model selection methods, such as LASSO, adaptive LASSO (ALasso), and OGA+HDIC+Trim (OGA-3), where the names in the parentheses are shorthands used throughout the chapter. Since FHTD first coerces all candidate AR variables into the model, we modify ALasso and OGA-3 accordingly and consider the analogous methods, AR-ALasso and AR-OGA-3. For AR-ALasso, the AR variables are not penalized in the first-stage LASSO and the resulting coefficients are used as the initial weights (weighted inversely) for the second-stage LASSO. AR-OGA-3 also forces the AR variables into the base model when implementing OGA-3.

According to Theorem 2.3.2, the penalty term, w_{n,p_n}, in HDIC can be taken to be $t_n p_n^{*2\bar{\theta}}$, where $\bar{\theta}$ is defined in $\mathbf{(SS_X)}$ and $\{t_n\}$ diverges to ∞ arbitrarily slowly. Here, we approximate $2\bar{\theta}$ using $1/\eta$ because the exogenous variables are often allowed to have finite higher-order moments. On the other hand, we set $\eta = 2$ to include GARCH-type errors with relatively

heavy tails. As a result, for the FSR- and OGA-based methods,

$$\mathrm{HDIC}(Q \oplus J) = n \log \hat{\sigma}^2_{Q \oplus J} + c p^{*1/\eta} (\sharp(Q) + \sharp(J)), \quad \eta = 2, \tag{2.44}$$

is used throughout all simulations, where $c > 0$ is a tuning parameter. Similarly, in view of Theorem 2.3.3 and (2.27), the $\hat{H}_n$ in DDT is set to

$$\frac{[(q_n + \hat{s}_0)^{1/2} \wedge (\hat{s}_0^{1/2} q_n^{1/2})]d}{n^{1/2}}, \tag{2.45}$$

where d is also subject to fine-tuning. In practice, one may use a hold-out validation set to determine c and d. To reduce the computational burden, we set $c = d = 0.5$ in all simulation examples and leave the problem of tuning c and d to Section 2.5. The number of iterations, K_n, of FSR and OGA is set to 40. The tuning parameters for LASSO-type methods are selected using BIC as in Medeiros and Mendes (2016). Finally, q_n and $r_j^{(n)}$ are set to $q_n = \lfloor 2n^{0.25} \rfloor$ and $r_j^{(n)} = r^{(n)}$ for all $1 \le j \le p_n$, where $(n, p_n, r^{(n)}) = (200, 100, 4), (400, 200, 5)$, and (800, 500, 6). Note that $p_n^* = p_n r^{(n)} > n$ in all cases.

Let $\tilde{\mathcal{Q}}_i$ and $\tilde{\mathcal{J}}_i$ denote the sets of the AR and exogenous variables chosen by a model selection method in the i-th simulation. Then its performance is measured by the frequencies of selecting exactly the relevant variables (E) and including all relevant variables (SS) as well as the average numbers of true positives (TP) and false positives (FP), namely,

$$\mathrm{E} = \sum_{i=1}^{1000} \mathbb{I}_{\{\tilde{\mathcal{Q}}_i = \mathcal{Q}_n\}} \mathbb{I}_{\{\tilde{\mathcal{J}}_i = \mathcal{J}_n\}}, \ \mathrm{SS} = \sum_{i=1}^{1000} \mathbb{I}_{\{\mathcal{Q}_n \subseteq \tilde{\mathcal{Q}}_i\}} \mathbb{I}_{\{\mathcal{J}_n \subseteq \tilde{\mathcal{J}}_i\}},$$
$$\mathrm{TP} = \frac{1}{1000} \sum_{i=1}^{1000} (\sharp\{\tilde{\mathcal{Q}}_i \cap \mathcal{Q}_n\} + \sharp\{\tilde{\mathcal{J}}_i \cap \mathcal{J}_n\}), \mathrm{FP} = \frac{1}{1000} \sum_{i=1}^{1000} (\sharp\{\tilde{\mathcal{Q}}_i \cap \mathcal{Q}_n^c\} + \sharp\{\tilde{\mathcal{J}}_i \cap \mathcal{J}_n^c\}),$$

where $\mathcal{Q}^c = [q_n] \setminus \mathcal{Q}_n$ and $\mathcal{J}_n^c = \bar{J} \setminus \mathcal{J}_n$. All simulation results are based on 1,000 replicates.

Example 2.4.1. In this example, we generate n observations from

$$(1 - 0.45B^4 - 0.45B^5)(1 - B)y_t = \sum_{j=1}^{5} \beta_1^{(j)} x_{t-1,j} + \sum_{j=6}^{10} \beta_2^{(j)} x_{t-2,j} + \epsilon_t, \tag{2.46}$$

where ϵ_t is independently drawn from a $t(6)$ distribution. The candidate covariates are generated according to the AR(1) model, $x_{t,j} = 0.8x_{t-1,j} + 2w_t + v_{t,j}$, $j = 1, 2, \ldots, p_n$, where $\{w_t\}$ and $\{v_{t,j}\}$ are independent standard Gaussian white noise processes and are independent of $\{\epsilon_t\}$. The coefficients are given by $(\beta_1^{(1)}, \beta_1^{(2)}, \beta_1^{(3)}, \beta_1^{(4)}, \beta_1^{(5)}) = (3, 3.75, 4.5, 5.25, 6)$, and $(\beta_2^{(6)}, \beta_2^{(7)}, \beta_2^{(8)}, \beta_2^{(9)}, \beta_2^{(10)}) = (6.75, 7.5, 8.25, 9, 9.25)$. Since a unit-root is introduced in (2.46), $\{y_t\}$ is nonstationary and the model contains three lagged dependent variables, y_{t-1}, y_{t-4}, y_{t-6}, and ten exogenous variables. In addition, the candidates $x_{t-l,j}$ are highly correlated because $\text{Corr}(x_{t,i}, x_{t,j}) = 0.8$, for $i \neq j$.

Simulation results for Example 2.4.1 are summarized in Table 2.1. Clearly, the LASSO-type methods fail to identify the correct model. Their TP values are only slightly larger than 1, meaning on average they detect only one relevant variable. A closer look at the results reveals that y_{t-1} is always included by these methods. However, they include only another one or two variables at most, which are usually irrelevant, resulting in a low FP value. OGA-3 performs equally poorly in terms of TP values, and tends to select more irrelevant variables. AR-OGA-3 has much higher TP values than OGA-3 though its performance in variable screening and selection is unsatisfactory. This inferior performance of AR-OGA-3 is mainly ascribed to OGA's relatively poor selection path, which falls short of including all relevant exogenous variables after adding all candidate AR variables in the model. By contrast, FSR successfully includes all relevant exogenous variables. Based on the reliable selection path of FSR, HDIC, Trim, and DDT further remove all redundant variables and identify the true ARX model over 90% of the time when $n \geq 400$.

Table 2.1: Values of E, SS, TP, and FP in Example 2.4.1, where E denotes selecting exactly the relevant variables and SS including all relevant variables, and TP and FP are the average numbers of true positives and false positives. Results are based on 1000 replications.

	LASSO	ALasso	OGA-3	AR-ALasso	AR-OGA-3	FHTD
$(n, p_n^*, p_n, r^{(n)}, q_n) = (200, 400, 100, 4, 7)$						
E	0	0	0	0	1	431
SS	0	0	0	0	1	1000
TP	1.02	1.02	1.16	1.12	6.67	13.00
FP	0.73	0.39	3.36	0.50	12.49	0.98
$(n, p_n^*, p_n, r^{(n)}, q_n) = (400, 1000, 200, 5, 8)$						
E	0	0	0	0	78	919
SS	0	0	0	0	78	1000
TP	1.01	1.00	1.12	1.07	10.46	13.00
FP	0.22	0.09	4.39	0.63	11.12	0.09
$(n, p_n^*, p_n, r^{(n)}, q_n) = (800, 3000, 500, 6, 10)$						
E	0	0	0	0	229	998
SS	0	0	0	0	229	1000
TP	1.03	1.00	1.32	1.06	11.87	13.00
FP	0.13	0.00	5.62	0.66	9.29	0.00

Example 2.4.2. In this example, we generate data from

$$(1 - 0.3B)(1 - 2\cos(0.1)B + B^2)y_t = \sum_{j=1}^{5} \beta_1^{(j)} x_{t-1,j} + \sum_{j=6}^{10} \beta_2^{(j)} x_{t-2,j} + \epsilon_t, \tag{2.47}$$

where $\{\epsilon_t\}$ is a GARCH(1,1) process satisfying

$$\epsilon_t = \sigma_t Z_t, \quad \sigma_t^2 = 5 \times 10^{-2} + 0.05\epsilon_{t-1}^2 + 0.9\sigma_{t-1}^2,$$

in which $\{Z_t\}$ is a sequence of i.i.d. standard Gaussian random variables. Using Theorem 2.2 of Ling and McAleer (2002), one can verify that ϵ_t has a finite sixth moment. Let $\mathbf{w}_t = \mathbf{A}\boldsymbol{\pi}_t$, where $\mathbf{A} = (a_{ij})_{1 \le i,j \le p_n}$, with $a_{ij} = 0.6^{|i-j|}$ if $|i - j| \le 7$ and $a_{ij} = 0$ otherwise, and $\{\boldsymbol{\pi}_t\}$, independent of $\{Z_t\}$, is a sequence of i.i.d. random vectors whose entries are independently drawn from a $t(13)$ distribution. We then generate $x_{t,j}$ by $(1 - 0.1B + 0.7B^2)x_{t,j} = (1 + 0.7B)w_{t,j}$, $1 \le j \le p_n$, where $w_{t,j}$ is the j-th component of $\mathbf{w}_t$. Note that $\{x_{t,j}\}$ is an

ARMA(2,1) process. Moreover, the relevant coefficients are $(\beta_1^{(1)}, \beta_1^{(2)}, \beta_1^{(3)}, \beta_1^{(4)}, \beta_1^{(5)}) = (0.82, -1.03, 1.92, -2.21, 2.42)$, and $(\beta_2^{(6)}, \beta_2^{(7)}, \beta_2^{(8)}, \beta_2^{(9)}, \beta_2^{(10)}) = (-2.57, 3.28, -3.54, 3.72, -3.90)$.

Table 2.2 reports the performance of the same methods as those in 2.4.1. In addition to conditionally heteroscedastic errors, the major challenge in this example lies in the fact that the AR component on the left-hand side of (2.47) contains complex unit roots; thus, $\{y_t\}$ cannot be made stationary through simple difference transforms. As observed in Table 2.2, this challenge hinders the performance of the OGA- and LASSO-type methods, all of which have zero SS and E values and low TP values even when $n = 800$. In contrast, FHTD still works well under the challenge. Specifically, it detects all relevant variables over 94% of the time for $n \geq 200$. In addition, its E value rapidly increases from 493 to over 840 when n increases from 200 to 400 (or 800).

Table 2.2: Values of E, SS, TP, and FP in Example 2.4.2, where E, SS, TP and FP are defined similarly as thos of Table 2.1. Results are also based on 1000 replications.

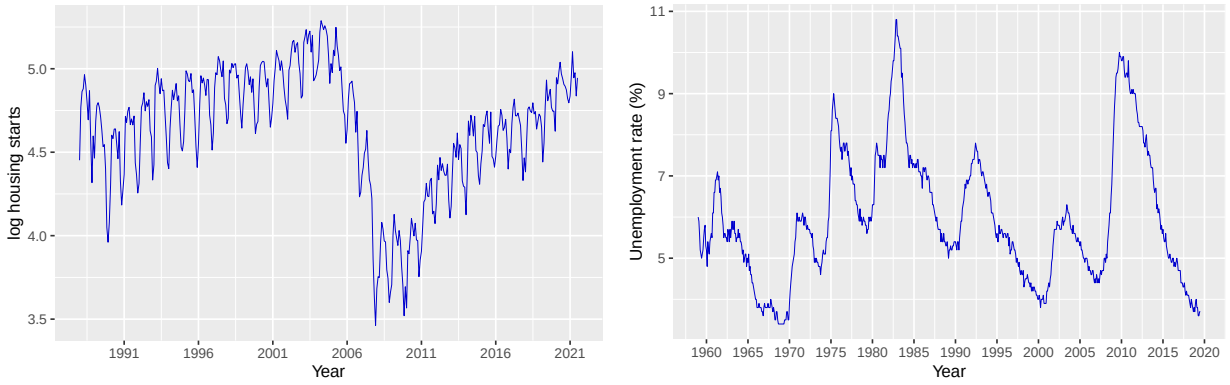
	LASSO	ALasso	OGA-3	AR-ALasso	AR-OGA-3	FHTD
$(n, p_n^*, p_n, r^{(n)}, q_n) = (200, 400, 100, 4, 7)$						
E	0	0	0	0	0	493
SS	0	0	0	0	1	943
TP	1.45	1.24	1.33	1.19	5.40	12.93
FP	3.04	2.22	2.09	1.73	6.33	0.80
$(n, p_n^*, p_n, r^{(n)}, q_n) = (400, 1000, 200, 5, 8)$						
E	0	0	0	0	0	845
SS	0	0	0	0	0	999
TP	1.21	1.10	1.83	1.05	5.63	13.00
FP	1.93	1.58	3.19	1.50	6.40	0.24
$(n, p_n^*, p_n, r^{(n)}, q_n) = (800, 3000, 500, 6, 10)$						
E	0	0	0	0	0	850
SS	0	0	0	0	0	1000
TP	1.04	1.01	1.94	1.00	6.19	13.00
FP	1.32	1.18	3.34	1.18	6.66	0.33

We also considered another challenging example, where the error term and all candidate exogenous variables are conditionally heteroscedastic in addition to two unit roots in the

AR component. FHTD still substantially outperforms the other methods in this example. Details are deferred to Section 2.7.5.

2.5 Applications

In this section, we apply the proposed FHTD to the U.S. monthly housing starts and unemployment series.



(a) Logarithm of housing starts. (b) Unemployment rates, seasonally adjusted.

Figure 2.1: Time plots of U.S. monthly housing starts and unemployment series

2.5.1 Housing starts in the U.S.

In this application, we are interested in modeling the logarithm of U.S. monthly housing starts. As depicted in Figure 2.1a, the series exhibits an apparent seasonal pattern along with a drastic level change around subprime financial crisis of 2008. For covariates, we collect the monthly new private housing units authorized by building permits for each state[1] and the 30-year fixed rate mortgage averages from the Economic Data of St. Louis Federal Reserve[2]. After removing series with missing values, we have 49 housing permits series

1. For instance, data for Illinois are retrieved from `https://fred.stlouisfed.org/series/ILBP1FH`

2. Freddie Mac, 30-Year Fixed Rate Mortgage Average in the United States [MORTGAGE30US], retrieved from FRED, Federal Reserve Bank of St. Louis; `https://fred.stlouisfed.org/series/MORTGAGE30US`, October 27, 2022.

$\{x_{t,j}, j = 1, 2, \ldots, 49\}$ and the mortgage rate series r_t from January 1988 through August 2022. We also remove the seasonality and unit root by taking $\tilde{x}_{t,j} = (1-B^{12})(1-B)\log x_{t,j}$, $j = 1, 2, \ldots, 49$, and $\tilde{r}_t = r_t - r_{t-1}$. Consequently, we have 403 observations for each series.

Then we employ the following predictive model

$$h_t = \sum_{l=1}^{18} \alpha_l h_{t-l} + \sum_{j=1}^{49} \sum_{k=1}^{18} \beta_k^{(j)} \tilde{x}_{t-k,j} + \sum_{k=1}^{18} \beta_k^{(50)} \tilde{r}_{t-k} + \epsilon_t, \tag{2.48}$$

where h_t denotes the logarithm of U.S. housing starts at month t. Note that there are 918 potential predictors. We also consider the model with a drift,

$$h_t = \beta_0 + \sum_{l=1}^{18} \alpha_l h_{t-l} + \sum_{j=1}^{49} \sum_{k=1}^{18} \beta_k^{(j)} \tilde{x}_{t-k,j} + \sum_{k=1}^{18} \beta_k^{(50)} \tilde{r}_{t-k} + \epsilon_t. \tag{2.49}$$

In implementing FHTD, we estimate (2.49) via the following procedure. First, subtract from each variable (including the dependent variable) its own sample average. Then, perform model selection with the transformed data using FHTD. Finally, estimate (2.49) by OLS with the selected variables and an intercept.

We perform rolling-window one-step-ahead prediction using FHTD as well as the other methods described in Section 2.4. We reserve the last 18 years of data as the test set; that is, there are $W = 216$ windows. Each window contains 169 observations as training data. Figure 2.2 plots some selected windows. As shown in the figure, the methods are challenged to forecast the sharp dip around 2008 and the following recovery. Since the true model is unknown, the performance of the methods under consideration is measured by the root mean squared prediction error (RMSE) and the median absolute prediction error (MAE), where RMSE $= \{W^{-1} \sum_{w=1}^{W} (h_{T-w+1} - \hat{h}_{T-w+1})^2\}^{1/2}$ and MAE is the median of $\{|h_{T-w+1} - \hat{h}_{T-w+1}| : w = 1, 2 \ldots, W\}$, in which T is the time index for the last data point and $\hat{h}_{T-w+1}$ is the predicted value of h_{T-w+1}. In implementing FHTD and AR-OGA-3, we use HDIC in (2.44), $\hat{H}_n$ in (2.45), and choose c and d therein over a grid of values between

0.1 and 0.7 via a hold-out validation set consisting of the last 20% of the training data in each window. The BIC is used to select the penalty parameters for LASSO-type methods.

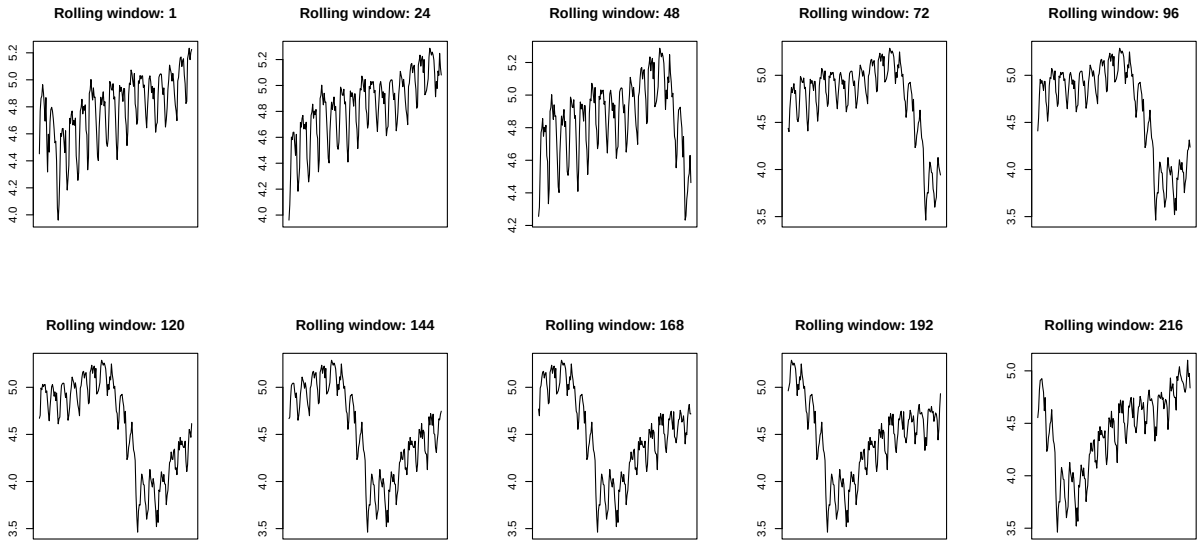


Figure 2.2: Time plots of logarithim of monthly U.S. Housting Starts, h_t, of selected windows

The prediction results are recorded in Table 2.3. We observe that LASSO-type methods are highly sensitive to the specification of the intercept. They performed poorly when the intercept is omitted. In view of Figure 2.2, fitting the drift term to the upward trend appeared in the first few windows may help alleviate the unit-root property in the data in finite sample, and without the drift, LASSO-type methods are unable to adapt to the unit-root behavior in the data. On the contrary, FHTD remains stable whether or not an intercept is included, and its prediction errors are substantially lower than the other methods. This example indicates that the proposed method can reliably select relevant predictors in predicting unit-root time series.

Table 2.3: Out-of-sample RMSEs and MAEs of competing methods applied to (2.48) and (2.49)

	FHTD	OGA-3	AR-OGA-3	LASSO	ALasso	AR-ALasso
Model (2.48)						
RMSE($\times$10)	1.08	1.34	1.32	2.14	2.13	1.43
MAE($\times$10)	0.72	0.82	0.82	1.21	1.17	0.97
Model (2.49)						
RMSE($\times$10)	1.11	1.31	1.29	1.24	1.21	1.14
MAE($\times$10)	0.71	0.82	0.91	0.86	0.88	0.81

2.5.2 U.S. unemployment rate

Next, we consider the prediction of U.S. monthly unemployment rate $\{u_t\}$, shown in Figure 2.1b. In some empirical studies, the unemployment rate is considered as difference-stationary. Nevertheless, Bierens (2001) has found some evidence that the fluctuations may be due to complex unit roots. In forecasting unemployment rate, Montgomery et al. (1998) have also noted the possibility of complex unit roots. Regardless of such complications, we can directly apply FHTD and the other methods discussed in the previous section to select a model for $\{u_t\}$ and to predict its future values. The data used are from the FRED-MD dataset[3], which contains 128 U.S. monthly macroeconomic variables from January 1959 to July 2019.

We use data from January 1973 to June 2019 and again consider rolling-window one-step-ahead predictions. After discarding the series with missing values during the time span, there remain 124 macroeconomic time series that can be used to forecast u_t. Following McCracken and Ng (2016), we transform some series by taking logs, differencing, or both, so that all 124 series are considered stationary after the transformations. Denote these series

3. available on `https://research.stlouisfed.org/econ/mccracken/fred-databases/`

by $\{x_{t,j}\}, j = 1, \ldots, 124$. Then we apply FHTD to the following model,

$$u_t = \sum_{i=1}^{6} \alpha_i u_{t-i} + \sum_{j=1}^{124} \sum_{l=1}^{6} \beta_l^{(j)} x_{t-l,j} + \epsilon_t, \tag{2.50}$$

which contains 750 candidate predictors. The last two years of data are reserved as test samples, resulting in a window size of 310 observations.

The results are reported in Table 2.4. In both performance measures, FHTD outperforms OGA-3, AR-OGA-3, AR-ALasso, and LASSO. Its RMSE improves by 6%, 7%, 12%, and 14% over OGA-3, LASSO, AR-OGA-3, and AR-ALasso, respectively. The results are similar when comparing the MAEs. Note that for LASSO, ALasso, and AR-ALasso, we only report their performance when an intercept is included, since, as observed in the previous application, these methods performed poorly when the intercept is omitted. In this particular application, without the intercept the RMSEs and MAEs of LASSO and ALasso can be more than 14 times as large as their counterparts for the AR-AIC. These results, combined with those in Table 2.3, show that FHTD is applicable to general unit-root time series, stable across specifications, and makes best use of the available predictors. Finally, we remark that the performance of ALasso (with intercept) is also quite competitive, implying that both are the most recommendable approaches to forecast $\{u_t\}$.

Table 2.4: Out-of-sample RMSEs and MAEs of competing methods applied to (2.50) for U.S. monthly unemployment rate series.

	FHTD	OGA-3	AR-OGA-3	LASSO	ALasso	AR-ALasso
RMSE(×10)	1.34	1.42	1.50	1.44	1.38	1.52
MAE(×10)	0.88	0.91	0.95	0.96	0.88	1.00

2.6 Concluding remarks

This chapter proposed the FHTD algorithm for variable selection in high-dimensional non-stationary ARX models with heteroscedastic covariates and errors. Under strong sparsity

conditions, we established its selection consistency, valid even when the lagged dependent variables are highly correlated and sample covariance matrices lack deterministic limits. Finally, we point out some potential research directions. First, shifting from **(SS)** and **($\mathbf{SS_X}$)** to weak sparsity assumptions (where coefficients are mostly non-zero but only a few are significant) may prioritize optimal forecasting over variable selection consistency. Addressing this issue remains challenging, particularly in the presence of complex unit roots. Another intriguing avenue is the model selection for cointegrated data, common in economics and environmental studies, within the realm of high-dimensional data analysis.

2.7 Supplementary details

The supplemental material includes five subsections. Subsection 2.7.1 provides comments on (A1)–(A6) and discusses the sparsity conditions **($\mathbf{SS_X}$)** and **(SS)**. Subsection 2.7.2 presents the proofs for the main results, Theorems 2.3.1–2.3.3. Several theoretical results crucial to the proofs are stated in the same section for completeness. Further details related to Subsection 2.7.2 can be found in Subsection 2.7.3. Subsection 2.7.4 contains specific information regarding Examples 2.3.1 and 2.3.2. Finally, Subsection 2.7.5 offers additional simulation results.

2.7.1 Comments on Assumptions (A1)–(A6), (SS_X), and (SS)

In this subsection, we offer a few comments on assumptions (A1)–(A6). The reader is also referred to Huang et al. (2023) for related discussions on assumptions (A1)–(A4). Assumption (A1) is fulfilled by many conditionally heteroscedastic processes, such as the GJR-GARCH model (see Huang et al., 2023, 2022). Assumption (A2) requires that $\{x_{t,s}\}$ is an MA(∞) process driven by the conditionally heteroscedastic innovations $\{\pi_{t,s}\}$. This type of assumption is broadly adopted in time series analysis. In fact, (A2) allows $(\epsilon_t, \pi_{t,1} \ldots, \pi_{t,p_n})^\top$ to be a multivariate GARCH process. By the same argument used in Huang et al. (2022), it can

be shown that the diagonal VEC model of Bollerslev et al. (1988) is a special case of (2.37). Moment conditions (2.39) and (2.40) are more stringent than (2.34). These stronger moment assumptions ensure a reliable screening performance of FSR when the number of exogenous covariates is larger than the sample size, as described in Assumption (A5). Furthermore, in the notable special case where $\{\pi_{t,s}\}$ is a sequence of independent and identically distributed random variables with $\mathbb{E}(\pi_{t,s}) = 0$ and $\mathbb{E}(\pi_{t,s_1}\pi_{t,s_2}) = \sigma_{s_1,s_2}$, (2.37) remains valid, with $e_{t,s_1,s_2} = \pi_{t,s_1}\pi_{t,s_2} - \sigma_{s_1,s_2}$, $\theta_{0,s_1,s_2} = 1$, and $\theta_{j,s_1,s_2} = 0$ for $j > 0$.

Assumption (A3) is used to derive the FCLT for the multivariate linear process driven by $\{\delta_t\}$ (Theorem 2.1, Huang et al., 2023), leading to a uniform lower bound for the minimum eigenvalues of the sample covariance matrices of dimensions less than or equal to $q_n + K_n$; see Theorem 4.1 of Huang et al. (2023). Assumption (A4), referred to as the weak sparsity condition, is commonly used in the literature on high-dimensional data analysis. It follows from (2.35), (2.36), and (A4) that

$$\sup_{n\geq 1} \sum_{h=-\infty}^{\infty} |\gamma_{h,n}| \leq C, \tag{2.51}$$

which, together with (2.41), leads to

$$\sum_{h=-\infty}^{\infty} |\gamma_h| \leq C. \tag{2.52}$$

Assumption (A5) allows that the covariate dimension to be at least of the same order as n, and can be much larger than n if $\eta > 2$. It also permits that q_n, the prescribed upper bound of the number of AR variables, increases to ∞ at a rate slower than $n^{1/2}$.

When the moment conditions are controlled, (A5) appears to be more flexible than the assumptions on model dimensions in Medeiros and Mendes (2016), where $\{y_t\}$ is assumed to be stationary, corresponding to the case of $a = b = d_1 = \cdots = d_l = 0$. To see this, note that (A1) and (A2) imply (2.9) and (2.10) respectively. Moreover, (A1), together with (A2),

yields

$$\sup_t \mathbb{E}\,|y_t|^{2\eta} < C, \tag{2.53}$$

provided $a = b = d_1 = \cdots = d_l = 0$. By (2.9), (2.10), (2.53), and Hölder's inequality,

$$\sup_t \mathbb{E}\,|y_{t-i}\epsilon_t|^{\eta} < C \text{ and } \sup_t \mathbb{E}\,|x_{t-l,j}\epsilon_t|^{2\eta q_0/(q_0+1)} < C,$$

for all $1 \le i \le q_n$, $1 \le l \le r_j^{(n)}$, and $1 \le j \le p_n$. Therefore, the m in Assumption DGP(4) of Medeiros and Mendes (2016) obeys

$$m = \min\{\eta, 2\eta q_0/(q_0+1)\} = \eta. \tag{2.54}$$

Equation (2.54) and the discussion after Assumption (REG) of Medeiros and Mendes (2016) lead to a restriction on the number of candidate variables such that

$$q_n + p_n^* = o(n^{\alpha\eta(\eta-2)/(2\eta+4b)}), \tag{2.55}$$

where $0 < \alpha < 1$ and $b > 0$ are positive numbers defined therein. Equation (2.55) requires p_n^* to be much smaller than n unless $\eta > 4$. In contrast, (A5) allows $p_n^* > n$ even if $\eta = 2$.

We also make a few comments on (A6). For $D \subset \{1, \ldots, p_n\}$, let

$$\begin{aligned} &\boldsymbol{\pi}_t(D) = (\pi_{t,s} : s \in D)^\top,\ \boldsymbol{\mu}_t(D) = (\epsilon_t, \boldsymbol{\pi}_t^\top(D))^\top, \\ &\boldsymbol{\Sigma}_n(D) = \mathbb{E}(\boldsymbol{\mu}_t(D)\boldsymbol{\mu}_t^\top(D)). \end{aligned} \tag{2.56}$$

Then, it can be shown that (2.42) holds if $\{x_{t,j}\}$ admits an infinite-order AR representation

with absolutely summable coefficients and

$$\max_{\sharp(D)\le K_n+\underline{s}_0} \lambda_{\min}^{-1}(\mathbf{\Sigma}_n(D)) \le C, \tag{2.57}$$

where $\underline{s}_0$ is defined in $(\mathbf{SS_A})$. When the AR components are deleted from model (2.2), (2.43) reduces to (3.2) of Ing and Lai (2011), which is closely related to the "exact recovery condition" introduced by Tropp (2004) in the analysis of the orthogonal matching pursuit and plays a role similar to the "restricted eigenvalue assumption" introduced by Bickel et al. (2009) in the study of LASSO. Condition (2.43) is a natural generalization of (3.2) of Ing and Lai (2011) when the (asymptotically) stationary AR component, $\mathbf{z}_t(q_n - d) = (z_{t-1}, \ldots, z_{t-q_n+d})^\top$, is taken into account.

We now present an example that illustrates the validity of (2.43) even when model (2.2) includes highly correlated lagged dependent variables and highly correlated exogenous variables. Assume in model (2.2) that $r_j^{(n)} = 1$ for all $1 \le j \le p_n$ and $\{(x_{t-1,1}, \ldots, x_{t-1,p_n}, \epsilon_t)^\top\}$ is a sequence of white noise vectors obeying $\mathbb{E}(\epsilon_t^2) = \mathbb{E}(x_{t,j}^2) = 1$ for all $1 \le j \le p_n$ and $0 \le \mathbb{E}(x_{t,i}x_{t,j}) = \mathbb{E}(x_{t,l}\epsilon_t) = \lambda < 1$ for all $1 \le i \ne j \le p_n$ and $1 \le l \le p_n$. In this model specification, not only are $y_{t-j}, 1 \le j \le q_n$, highly correlated, but also $x_{t-1,j}, 1 \le j \le p_n$, especially when λ is close to 1. Define $\mathbf{G}(q_n - d) = [\mathbf{G}_{ij}]_{1\le i,j\le q_n-d} = \mathbb{E}^{-1}(\mathbf{z}_{t,\infty}(q_n - d)\mathbf{z}_{t,\infty}^\top(q_n - d))$ and $c_J^2 = \lambda^2\sharp(J)/(1 - \lambda + \sharp(J)\lambda)$. Since $0 < c_J^2 < \lambda$ and $0 < \mathbf{G}_{11} \le 1$, it holds that

$$\max_{\sharp(J)\le K_n,(i,1)\notin J} \sum_{(i^*,1)\in J} |a_{i^*,1}(i,1)| \le \max_{\sharp(J)\le K_n} \frac{(1 - \lambda\mathbf{G}_{11})\lambda\sharp(J)}{(1 - c_J^2\mathbf{G}_{11})(1 - \lambda + \sharp(J)\lambda)} \le 1. \tag{2.58}$$

Moreover, one has

$$\sum_{s=1}^{q_n-d} \max_{\sharp(J)\le K_n,(i,1)\notin J} |a_{s,J}(i,1)| \le \sum_{s=1}^{q_n-d} \max_{\sharp(J)\le K_n} \frac{\lambda - c_J^2}{1 - c_J^2\mathbf{G}_{11}}|\mathbf{G}_{s1}| \le \sum_{s=1}^{q_n-d} |\mathbf{G}_{s1}|. \tag{2.59}$$

Define $a^2 = 1 + \lambda(\sum_{j=1}^{p_n} \beta_1^{(j)})^2 + (1 - \lambda)\sum_{j=1}^{p_n}(\beta_1^{(j)})^2$, $b = \lambda\sum_{j=1}^{p_n} \beta_1^{(j)}$, and $h^2 = (a^2 +$

$(a^4 - 4b^2)^{1/2})/2$, noting that $|b| < a^2/2$ and

$$h^2 > \max\{1, \lambda(\sum_{j=1}^{p_n} \beta_1^{(j)})^2 + (1-\lambda)\sum_{j=1}^{p_n} (\beta_1^{(j)})^2\}. \tag{2.60}$$

Then, it can be shown that $\{z_{t,\infty}\}$ admits an infinite-order AR representation,

$$z_{t,\infty} + \sum_{j=1}^{\infty} \phi_j z_{t-j,\infty} = \eta_t, \tag{2.61}$$

where $1+\sum_{j=1}^{\infty} \phi_j z^j \neq 0$, for $|z| \leq 1$, $\sum_{j=0}^{\infty} |\phi_j| \leq C$, and $\{\eta_t\}$ is a white noise sequence with variance h^2. By using a modified Cholesky decomposition (e.g., Ing et al. (2016)), (2.60), (2.61), and Baxter's inequality (Baxter (1962)), one gets $\sum_{s=1}^{q_n - d} |\mathbf{G}_{s1}| \leq Ch^{-2} \sum_{j=0}^{\infty} |\phi_j| \leq C$, which, together with (2.58) and (2.59), leads to (2.43).

Before concluding this subsection, we provide a brief discussion of the strong sparsity conditions **(SS$_{\mathrm{X}}$)** and **(SS)** in Sections 2.3.1 and 2.3.2. As mentioned earlier, a condition similar to **(SS$_{\mathrm{X}}$)** has been utilized by Medeiros and Mendes (2016) to establish the selection consistency of the adaptive LASSO when $\{y_t\}$ is stationary. Specifically, they assume

$$\frac{\lambda s_0^{1/2}}{n^{1-\xi/2}\phi_{\min}} = o(\min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|), \tag{2.62}$$

where $0 < \xi < 1$ is some constant defined in their Assumption (WEIGHTS) and $2\phi_{\min}$ is a lower bound for the minimum eigenvalue of the covariance matrix of the random vector formed by all relevant predictors. Assuming that $\phi_{\min}$ is bounded away from 0 and choosing λ to be the value suggested after Assumption (REG) of Medeiros and Mendes (2016), (2.62) becomes

$$\frac{s_0^{1/2} p_n^{*1/m} n^{\xi/m}}{n^{1/2}} = o(\min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|). \tag{2.63}$$

In view of (2.54) and the definitions of $\bar{\theta}$ and ξ, we conclude that (2.63) is more stringent than (2.11) in **(SS$_{\mathrm{X}}$)**. While the left-hand side of (2.20) in **(SS)** is larger than that of (2.11) by a factor about $(\log n)^{1/2}$, it is still smaller than that of (2.63).

2.7.2 Key theoretical results and main proofs

In this subsection, we present the proofs of the main results in this chapter, namely Theorems 2.3.1–2.3.3. The proofs are proceeded by a few theoretical results developed in Huang et al. (2023) that are key to the proofs. For the sake of completeness, we state the results relevant to this chapter, and refer the readers to Huang et al. (2023) for proofs and detailed discussions. To simplify the exposition and without loss of generality, we assume, in what follows, that l_0 and l_{s_1,s_2} in (A1) and (A2) are equal to 1.

Theoretical tools

In this subsection, we collect some useful theoretical apparatus, including a novel FCLT for the multivariate linear processes, (2.64), driven by $\{\delta_t\}$ under a set of mild conditions (Theorem 2.7.1) and the moment bounds for quadratic forms associated with model (2.2) (Lemma 2.7.1). These results are used to bound from below the minimum eigenvalues of the sample covariance matrices of the candidate models; see Theorem 2.7.2. For the proofs in this subsection, we refer to the companion paper (Huang et al., 2023). Due to the presence of $v_{t,n} = \sum_{j=1}^{p_n} \sum_{l=1}^{r_j^{(n)}} \beta_l^{(j)} x_{t-l,j}$ in δ_t, the FCLT is quite different from the classical ones (see, e.g., Chan and Wei (1988) and Ling and Li (1998)), where the linear processes are driven by

$\{\epsilon_t\}$ only. Let

$$\begin{aligned}
&\mathbb{B}_n(t_1, t_2, \ldots, t_{2l+2}) \\
&= \frac{1}{\sqrt{n}} \left(\sum_{k=1}^{\lfloor nt_1 \rfloor} \delta_k, \sum_{k=1}^{\lfloor nt_2 \rfloor} (-1)^k \delta_k, \sum_{k=1}^{\lfloor nt_3 \rfloor} \sqrt{2} \sin(k\vartheta_1)\delta_k, \sum_{k=1}^{\lfloor nt_4 \rfloor} \sqrt{2} \cos(k\vartheta_1)\delta_k, \right. \\
&\quad \left. \ldots, \sum_{k=1}^{\lfloor nt_{2l+2} \rfloor} \sqrt{2} \cos(k\vartheta_l)\delta_k \right).
\end{aligned} \tag{2.64}$$

Note that $\mathbb{B}_n$ is a random element in D^{2l+2}, where D is the Skorohod space $D = D[0,1]$ (Billingsley (1999)).

Theorem 2.7.1 (Huang et al., 2023, Theorem 2.1)**.** *Assume that* (A1)–(A4) *hold with* η*,* $q_0\eta$*, and* $2q_0\eta/(1+q_0)$ *in* (2.34)*,* (2.39)*, and* (2.40)*, respectively, replaced by* η_1 *for some* $\eta_1 > 1$*. In addition, assume*

$$\max_{1 \le j \le p_n} r_j^{(n)} = o(n^{\kappa}), \tag{2.65}$$

where $\kappa = \min\{1/2, 1 - \eta_1^{-1}\}$*. Then*

$$\mathbf{V}^{-1/2} \mathbb{B}_n \Rightarrow \mathbb{W}, \tag{2.66}$$

where $\Rightarrow$ *denotes convergence in law,* $\mathbb{W}$ *is a* $(2l+2)$*-dimensional standard Brownian motion, and* $\mathbf{V}$ *is a* $(2l+2)$*-dimensional diagonal matrix with positive diagonal elements,*

$$v_1^2 = \sum_{h=-\infty}^{\infty} \gamma_h, \; v_2^2 = \sum_{h=-\infty}^{\infty} (-1)^h \gamma_h, \; v_{2k+1}^2 = v_{2k+2}^2 = \sum_{h=-\infty}^{\infty} \cos(h\vartheta_k)\gamma_h,$$

$k = 1, 2, \ldots, l$.

Theorem 2.7.1 is crucial for the uniform lower bound for the minimum eigenvalues of the sample covariance matrices. To state this result, we need to introduce some notations.

Recall $\phi(B)$ in Section 2.3.3. Inspired by Chan and Wei (1988), we define

$$\begin{aligned} u_t(j) &= [(1-B)^{-j}\phi(B)]y_t, \\ v_t(j) &= [(1+B)^{-j}\phi(B)]y_t, \\ g_t(k,j) &= [(1-2\cos\vartheta_k B + B^2)^{-j}\phi(B)]y_t, \end{aligned} \tag{2.67}$$

where $k = 1, \ldots, l$. For $k = 1, \ldots, l$, it can be shown that

$$\begin{aligned} g_t(k,1) =& \frac{1}{\sin\vartheta_k}\sum_{s=1}^{t}\sin[(t-s+1)\vartheta_k]\delta_s \\ =& \sum_{s=0}^{t-1}\frac{\sin[(s+1)\vartheta_k]}{\sin\vartheta_k}\delta_{t-s} := \sum_{s=0}^{t-1}\kappa_s(k,1)\delta_{t-s}, \end{aligned}$$

where $|\kappa_s(k,1)| \le C$ for all $s \ge 0$. By induction it follows that

$$g_t(k,j) = \sum_{s=0}^{t-1}\kappa_s(k,j)\delta_{t-s},$$

where

$$|\kappa_s(k,j)| \le C(s+1)^{j-1}, \tag{2.68}$$

for all $s \ge 0$, $1 \le k \le l$, and $1 \le j \le d_k$. Similarly,

$$u_t(j_1) = \sum_{s=0}^{t-1}\iota_s(j_1)\delta_{t-s},\ v_t(j_2) = \sum_{s=0}^{t-1}\vartheta_s(j_2)\delta_{t-s},$$

where

$$|\iota_s(j_1)| \le C(s+1)^{j_1-1},\ |\vartheta_s(j)| \le C(s+1)^{j_2-1}, \tag{2.69}$$

for all $s \geq 0$, $1 \leq j_1 \leq a$, and $1 \leq j_2 \leq b$.

Let Q_n be defined implicitly by

$$Q_n \mathbf{w}_t(J) = (\mathbf{u}_t^\top, \mathbf{v}_t^\top, \mathbf{g}_t^\top(1), \ldots, \mathbf{g}_t^\top(l), \mathbf{z}_t^\top(q_n - d), \mathbf{x}_t^\top(J))^\top,$$

where

$$\begin{aligned}
\mathbf{u}_t &= (u_{t-1}(a), \ldots, u_{t-1}(1))^\top, \\
\mathbf{v}_t &= (v_{t-1}(b), \ldots, v_{t-1}(1))^\top, \\
\mathbf{g}_t(k) &= (g_{t-1}(k,1), g_{t-2}(k,1), \ldots, g_{t-1}(k,d_k), g_{t-2}(k,d_k))^\top,\ 1 \leq k \leq l,
\end{aligned}$$

and recall that $\mathbf{w}_t(J) = (y_{t-1}, \ldots, y_{t-q_n}, \mathbf{x}_t^\top(J))^\top$, $\mathbf{x}_t(J) = (x_{t-l,j} : (j,l) \in J)^\top$, and $\mathbf{z}_t(k) = (z_{t-1}, \ldots, z_{t-k})^\top$. Consider a normalized version,

$$\mathbf{s}_t(J) = (\tilde{y}_{t,1}, \ldots, \tilde{y}_{t,d}, \mathbf{z}_t^\top(q_n - d), \mathbf{x}_t^\top(J))^\top = G_n Q_n \mathbf{w}_t(J),$$

of $Q_n \mathbf{w}_t(J)$, where

$$G_n = \operatorname{diag}(G_{n,u}, G_{n,v}, G_{n,g}(1), \ldots, G_{n,g}(l), \mathbf{I}_{q_n + \sharp(J) - d}) \in \mathbb{R}^{(q_n+\sharp(J))\times(q_n+\sharp(J))},$$

with

$$\begin{aligned}
&G_{n,u} = \operatorname{diag}(n^{-a+1/2}, \ldots, n^{-1/2}),\ G_{n,v} = \operatorname{diag}(n^{-b+1/2}, \ldots, n^{-1/2}), \\
&G_{n,g}(k) = \operatorname{diag}(\underbrace{n^{-1/2}, n^{-1/2}, n^{-3/2}, n^{-3/2}, \ldots, n^{-d_k+1/2}, n^{-d_k+1/2}}_{2d_k}),\ k = 1, \ldots, l.
\end{aligned}$$

The following moment bounds imply useful concentration inequalities for quadratic forms involving the covariates and the lagged dependent variables that lend a helping hand throughout this chapter.

Lemma 2.7.1 (Huang et al., 2023, Corollary 3.1)**.** *Assume that* (A1), (A2)*, and* (A4) *hold. Then,*

$$\max_{\substack{1\le j_1,j_2\le p_n\\ 1\le l_1\le r_{j_1}^{(n)},\, 1\le l_2\le r_{j_2}^{(n)}}} \mathbb{E}\left| n^{-1/2}\sum_{t=\bar{r}_n+1}^{n}\left\{x_{t-l_1,j_1}x_{t-l_2,j_2}-\mathbb{E}(x_{t-l_1,j_1}x_{t-l_2,j_2})\right\}\right|^{\eta q_0} = O(1),$$

$$\max_{1\le i,j\le q_n-d} \mathbb{E}\left| n^{-1/2}\sum_{t=\bar{r}_n+1}^{n}\left\{z_{t-i}z_{t-j}-\mathbb{E}(z_{t-i,\infty}z_{t-j,\infty})\right\}\right|^{\eta} = O(1),$$

$$\max_{\substack{1\le j\le p_n, 1\le l\le r_j^{(n)}\\ 1\le k\le q_n-d}} \mathbb{E}\left| n^{-1/2}\sum_{t=\bar{r}_n+1}^{n}\left\{x_{t-l,j}z_{t-k}-\mathbb{E}(x_{t-l,j}z_{t-k,\infty})\right\}\right|^{\frac{2\eta q_0}{q_0+1}} = O(1),$$

$$\max_{\substack{1\le j\le p_n, 1\le l\le r_j^{(n)}\\ 1\le i\le d}} \mathbb{E}\left| n^{-1/2}\sum_{t=\bar{r}_n+1}^{n} x_{t-l,j}\tilde{y}_{t,i}\right|^{\frac{2\eta q_0}{q_0+1}} = O(1),$$

$$\max_{\substack{1\le k\le q_n-d\\ 1\le i\le d}} \mathbb{E}\left| n^{-1/2}\sum_{t=\bar{r}_n+1}^{n} z_{t-k}\tilde{y}_{t,i}\right|^{\eta} = O(1).$$

Now we can state the last piece of tools we need to prove our main results in the next subsection.

Theorem 2.7.2 (Huang et al., 2023, Theorem 4.1)**.** *Assume* (A1)–(A5) *and* (2.42)*. Then, for*

$$\bar{K}_n = o(n^{1/2}/p_n^{*\bar{\theta}}), \tag{2.70}$$

where $\bar{\theta}$ is defined in $(\mathbf{SS}_{\mathrm{X}})$*,*

$$\max_{\sharp(J)\leq \bar{K}_n} \lambda_{\min}^{-1} \left(n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(J)\mathbf{s}_t^{\top}(J) \right) = O_p(1). \tag{2.71}$$

Moreover,

$$\max_{\sharp(J)\leq \bar{K}_n} \lambda_{\min}^{-1} \left(n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(J)\mathbf{w}_t^{\top}(J) \right) = O_p(1). \tag{2.72}$$

Theorem 2.7.2 highlights one of the most intriguing subtleties of our analysis. Since our predictors contain highly correlated lagged dependent variables, it is not known a priori whether they lead to asymptotically ill-conditioned (or even singular) sample covariance matrices. The delicacy lies in the fact that $\mathbf{P}_n := n^{-1} \sum_{t=\bar{r}_n+1}^{n} \tilde{\mathbf{y}}_t \tilde{\mathbf{y}}_t^{\top}$ does not converge in probability to a deterministic limit, and hence the analysis of its minimum eigenvalue is much more involved. The problem is resolved through a novel FCLT (Theorem 2.7.1) that ensures $\mathbf{P}_n$'s weak limit exists and is a.s. positive definite. Consequently, as long as the size of a candidate model is equal to (or less than) $\bar{K}_n + q_n$, Theorem 2.7.2 guarantees that the corresponding sample covariance matrix is well-behaved. It also suggests that model selection criteria based on least squares estimation can differentiate between candidate predictors, albeit containing q_n highly correlated AR variables. Equations (2.72) and (2.71), respectively, are also aligned with (3.10) of Lai and Wei (1982) and (3.5.1) of Chan and Wei (1988), in which model (2.2) is simplified to a finite-order nonstationary AR model with a conditionally homogeneous error.

Proofs of Theorems 2.3.1–2.3.3

Define

$$\psi_{J,(i,l)} = \frac{n^{-1}\boldsymbol{\mu}_n^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)}}{(n^{-1}\boldsymbol{x}_l^{(i)\top}(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)})^{1/2}} \text{ and } \hat{\psi}_{J,(i,l)} = \frac{n^{-1}\mathbf{y}_n^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)}}{(n^{-1}\boldsymbol{x}_l^{(i)\top}(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)})^{1/2}}.$$

The main distinction between $\psi_{J,(i,l)}$ and $\hat{\psi}_{J,(i,l)}$ is that the $\mathbf{y}_n$ in the latter is substituted with its noiseless counterpart $\boldsymbol{\mu}_n$ in the former. Recalling that $\hat{J}_0 = \emptyset$, according to (2.4), FSR chooses

$$(\hat{j}_m, \hat{l}_m) = \arg\max_{(j,l)\in\bar{J}\setminus\hat{J}_{m-1}} \hat{\psi}_{\hat{J}_{m-1},(i,l)},\ m \geq 1,$$

at the m-th iteration, and then updates $\hat{J}_{m-1}$ by $\hat{J}_m = \hat{J}_{m-1} \cup \{(\hat{j}_m, \hat{l}_m)\}$. To analyze the asymptotic performance of FSR, we introduce the weak noiseless FSR. This algorithm is initialized with $J_0 = \emptyset$ and selects (j_m, l_m) satisfying

$$|\psi_{J_{m-1},(j_m,l_m)}| \geq \xi \max_{(j,l)\in\bar{J}\setminus J_{m-1}} |\psi_{J_{m-1},(j,l)}|, \quad m \geq 1, \tag{2.73}$$

where $0 < \xi \leq 1$ is a prescribed constant. It subsequently updates J_{m-1} by $J_m = J_{m-1} \cup \{(j_m, l_m)\}$. Note that when $\xi = 1$, the algorithm is referred to as noiseless FSR.

The performance of the weak noiseless FSR is evaluated by the "noiseless" mean squared error $\hat{a}_m$ defined in (2.17). In (2.109) of Section 2.7.3, we derive a convergence rate of $\hat{a}_m$ as m increases. The convergence of $\hat{a}_m$, along with Theorem 2.7.2, enables us to establish in (2.81) the convergence rate of $\hat{a}_m$'s semi-noiseless counterpart, $\hat{s}_m$, defined in (2.19). As we will see later, (2.81) serves as the key vehicle for us to develop the surely screening property of $\hat{J}_m$.

PROOF OF THEOREM 2.3.1. By (2.11), there exists $l_n \to \infty$ such that

$$\frac{l_n s_0 p_n^{*2\bar{\theta}}}{n \min_{(j,l)\in\mathcal{J}_n} {\beta_l^{(j)}}^2} = o(1).$$

Define

$$\mathcal{A}_n(K_n) = \left\{ \max_{\substack{\sharp(J)\le K_n-1 \\ (i,l)\notin J}} |\psi_{J,(i,l)} - \hat{\psi}_{J,(i,l)}| \le \frac{l_n^{1/2} p_n^{*(q_0+1)/(2\eta q_0)}}{n^{1/2}} \right\}$$

and

$$\mathcal{B}_n(K_n) = \left\{ \min_{0\le m\le K_n-1} \max_{(j,l)\notin \hat{J}_m} |\psi_{\hat{J}_m,(j,l)}| > \tilde{\xi} \frac{l_n^{1/2} p_n^{*(q_0+1)/(2\eta q_0)}}{n^{1/2}} \right\},$$

where $\tilde{\xi} > 2$ is some constant. On $\mathcal{A}_n(K_n) \cap \mathcal{B}_n(K_n)$, it holds that for all $1 \le m \le K_n$,

$$\begin{aligned}
|\psi_{\hat{J}_{m-1},(\hat{j}_m,\hat{l}_m)}| &\ge -|\hat{\psi}_{\hat{J}_{m-1},(\hat{j}_m,\hat{l}_m)} - \psi_{\hat{J}_{m-1},(\hat{j}_m,\hat{l}_m)}| + |\hat{\psi}_{\hat{J}_{m-1},(\hat{j}_m,\hat{l}_m)}| \\
&\ge - \max_{\substack{\sharp(J)\le m-1 \\ (j,l)\notin J}} |\hat{\psi}_{J,(j,l)} - \psi_{J,(j,l)}| + \max_{(j,l)\notin \hat{J}_{m-1}} |\hat{\psi}_{\hat{J}_{m-1},(j,l)}| \\
&\ge -2 l_n^{1/2} p_n^{*(q_0+1)/(2\eta q_0)} n^{-1/2} + \max_{(j,l)\notin \hat{J}_{m-1}} |\psi_{\hat{J}_{m-1},(j,l)}| \\
&\ge \xi \max_{(j,l)\notin \hat{J}_{m-1}} |\psi_{\hat{J}_{m-1},(j,l)}|,
\end{aligned} \tag{2.74}$$

where $0 < \xi = 1 - 2/\tilde{\xi} < 1$. By (2.74), we show in Section 2.7.3 that for all $1 \le m \le K_n$,

$$\hat{s}_m \mathbb{I}_{\mathcal{A}_n(K_n)\cap\mathcal{B}_n(K_n)} \le C_n \exp(-m\xi^2 D_n/s_0), \tag{2.75}$$

where

$$
\begin{aligned}
C_n &= n^{-1} \sum_{t=\bar{r}_n+1}^{n} v_{t,n}^2, \\
D_n &= \frac{\min_{1 \le \sharp(J) \le K_n} \lambda_{\min}(n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(\mathcal{J}_n \cup J)\mathbf{w}_t^\top(\mathcal{J}_n \cup J))}{\max_{(j,l)\in\mathcal{J}_n} n^{-1}\|\boldsymbol{x}_l^{(j)}\|^2},
\end{aligned}
\tag{2.76}
$$

recalling that $v_{t,n}$ is defined in (2.3). We also show in Section 2.7.3 that for all $1 \le m \le K_n$,

$$
\hat{s}_m \mathbb{I}_{\mathcal{B}_n^c(K_n)} \le \frac{s_0 l_n \tilde{\xi}^2 p_n^{*(q_0+1)/(\eta q_0)}}{n D_n},
\tag{2.77}
$$

and

$$
\lim_{n\to\infty} P(\mathcal{A}_n^c(K_n)) = 0.
\tag{2.78}
$$

By (A4) and (2.20),

$$
s_0 = o((n/p_n^{*2\bar{\theta}})^{1/3}),
\tag{2.79}
$$

which, together with (2.12), yields $s_0 u_n \log n = o(K_n)$ for some $u_n \to \infty$. It follows from (2.72), (A2), (A4), and (2.79) that

$$
C_n = O_p(1) \text{ and } D_n^{-1} = O_p(1).
\tag{2.80}
$$

According to (2.75)–(2.78) and (2.80),

$$
\max_{1\le m\le K_n} \frac{\hat{s}_m}{\exp(-m\xi^2 D_n/s_0) + (s_0 l_n p_n^{*(q_0+1)/(\eta q_0)}/n)} = O_p(1).
\tag{2.81}
$$

Let $\tilde{m}_n = s_0 u_n \log n$. The second equation of (2.80) implies for any $\bar{M} > 0$,

$$\exp(-\tilde{m}_n \xi^2 D_n / s_0) = O_p(n^{-\bar{M}}),$$

and hence (2.81) leads to

$$\hat{s}_{\tilde{m}_n} = O_p(s_0 l_n p_n^{*(q_0+1)/(\eta q_0)}/n). \tag{2.82}$$

On the set $\{\mathcal{J}_n \not\subseteq \hat{J}_{\tilde{m}_n}\}$, one has

$$\begin{aligned} \hat{s}_{\tilde{m}_n} &\geq \lambda_{\min}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(\mathcal{J}_n \cup \hat{J}_{\tilde{m}_n})\mathbf{w}_t^\top(\mathcal{J}_n \cup \hat{J}_{\tilde{m}_n})\right) \min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|^2 \\ &\geq \min_{\sharp(J)\leq K_n} \lambda_{\min}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(J)\mathbf{w}_t^\top(J)\right) \min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|^2. \end{aligned} \tag{2.83}$$

Combining (2.83), (2.82), (2.72), and (2.20) leads to

$$P(\mathcal{J}_n \not\subseteq \hat{J}_{K_n}) \leq P(\mathcal{J}_n \not\subseteq \hat{J}_{\tilde{m}_n}) \leq P\left(O_p(s_0 l_n p_n^{*(q_0+1)/(\eta q_0)}/n) \geq \min_{(j,l)\in\mathcal{J}_n} |\beta_l^{(j)}|^2\right) = o(1). \tag{2.84}$$

Thus, the desired conclusion (2.13) follows. □

PROOF OF THEOREM 2.3.2. Let $\tilde{k}_n = \min\{1 \leq k \leq K_n : \mathcal{J}_n \subseteq \hat{J}_k\}$ if $\mathcal{J}_n \subseteq \hat{J}_{K_n}$ and K_n if $\mathcal{J}_n \not\subseteq \hat{J}_{K_n}$. We start by showing that

$$\lim_{n\to\infty} P(\hat{k}_n = \tilde{k}_n) = 1. \tag{2.85}$$

By Theorem 2.3.1, (2.22) is an immediate consequence of (2.85). In the rest of the proof, we suppress the dependence on n and write $\hat{k}$ and $\tilde{k}$ instead of $\hat{k}_n$ and $\tilde{k}_n$. Let $\tilde{m}_n^* =$

$s_0 \log n \min\{u_n, (d_n/\log n)^{\delta}\}$, where u_n is defined after (2.79) and δ is defined in (2.21). By an argument similar to that used to prove (2.84), it holds that

$$\lim_{n\to\infty} P(\mathcal{D}_n) = 1, \tag{2.86}$$

where $\mathcal{D}_n = \{\mathcal{J}_n \subseteq \hat{J}_{\tilde{m}_n^*}\} = \{\tilde{k}_n \leq \tilde{m}_n^*\}$. Therefore, (2.85) is ensured by

$$P(\hat{k} < \tilde{k}, \mathcal{D}_n) = o(1) \tag{2.87}$$

and

$$P(\tilde{k} < \hat{k}, \mathcal{D}_n) = o(1). \tag{2.88}$$

By the definition of HDIC,

$$\hat{\sigma}^2_{M_{\tilde{k}-1}} - \hat{\sigma}^2_{M_{\tilde{k}}} \leq \hat{\sigma}^2_{M_{\tilde{k}}} \left\{ \exp\left(\frac{(\tilde{k} - \hat{k}) w_{n,p_n}}{n} \right) - 1 \right\} \quad \text{on } \{\hat{k} < \tilde{k}\}, \tag{2.89}$$

where $M_k = [q_n] \oplus \hat{J}_k$. Straightforward calculations give

$$\begin{aligned} &\hat{\sigma}^2_{M_{\tilde{k}-1}} - \hat{\sigma}^2_{M_{\tilde{k}}} \\ &= n^{-1} (\beta^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}} \boldsymbol{x}^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}} + \boldsymbol{\varepsilon}_n)^{\top} (\mathbf{H}_{M_{\tilde{k}}} - \mathbf{H}_{M_{\tilde{k}-1}}) (\beta^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}} \boldsymbol{x}^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}} + \boldsymbol{\varepsilon}_n) \\ &= {\beta^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}}}^2 \hat{A}_n + 2\beta^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}} \hat{B}_n + \hat{A}_n^{-1} \hat{B}_n^2 \quad \text{on } \mathcal{D}_n, \end{aligned} \tag{2.90}$$

in which

$$\begin{aligned} \hat{A}_n =& n^{-1} {\boldsymbol{x}^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}}}^{\top} (\mathbf{H}_{M_{\tilde{k}}} - \mathbf{H}_{M_{\tilde{k}-1}}) \boldsymbol{x}^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}}, \\ \hat{B}_n =& n^{-1} {\boldsymbol{x}^{(\hat{j}_{\tilde{k}})}_{\hat{l}_{\tilde{k}}}}^{\top} (\mathbf{H}_{M_{\tilde{k}}} - \mathbf{H}_{M_{\tilde{k}-1}}) \boldsymbol{\varepsilon}_n. \end{aligned}$$

In view of (2.89), (2.90), and

$$\frac{w_{n,p_n}\tilde{m}_n^*}{n} = o(\min_{(j,l)\in\mathcal{J}_n}(\beta_l^{(j)})^2) = o(1) \tag{2.91}$$

(which is ensured by (2.20), the second part of (2.21), and the definition of $\tilde{m}_n^*$), we have for all large n,

$$\beta_{\hat{l}_{\tilde{k}}}^{(\hat{j}_{\tilde{k}})^2}\hat{A}_n + 2\beta_{\hat{l}_{\tilde{k}}}^{(\hat{j}_{\tilde{k}})}\hat{B}_n + \hat{A}_n^{-1}\hat{B}_n^2 \le (\hat{C}_n + \sigma^2)\lambda_1\tilde{m}_n^* w_{n,p_n}/n \ \text{ on } \mathcal{D}_n \cap \{\hat{k} < \tilde{k}\}, \tag{2.92}$$

where $\hat{C}_n = \hat{\sigma}^2_{M_{\tilde{k}}} - \sigma^2$ and $\lambda_1 > 1$ is some constant. In addition, by making use of Theorem 2.7.2 and Lemma 2.7.1, we show in Section 2.7.3 that

$$\hat{A}_n^{-1} = O_p(1), \tag{2.93}$$

$$|\hat{B}_n| = O_p\left(\left(\frac{p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}}\right)\right), \tag{2.94}$$

$$|\hat{C}_n| = O_p\left(\frac{n^{1/2} + q_n + \tilde{m}_n^* p_n^{*\frac{q_0+1}{\eta q_0}}}{n}\right) = o_p(1). \tag{2.95}$$

As a result, (2.87) follows from (2.91)–(2.95).

On the other hand,

$$\hat{\sigma}^2_{M_{\tilde{k}}} - \hat{\sigma}^2_{M_{\hat{k}}} \ge \hat{\sigma}^2_{M_{\tilde{k}}}\{1 - \exp(n^{-1}w_{n,p_n}(\tilde{k} - \hat{k}))\} \text{ on } \{\tilde{k} < \hat{k}\}. \tag{2.96}$$

Let $F_{\hat{k},\tilde{k}} = (\boldsymbol{x}_l^{(j)} : (j,l) \in \hat{J}_{\hat{k}} \cap \hat{J}_{\tilde{k}}^c)$. Then on $\{\tilde{k} < \hat{k}\} \cap \mathcal{D}_n$,

$$\begin{aligned}
&\hat{\sigma}^2_{M_{\tilde{k}}} - \hat{\sigma}^2_{M_{\hat{k}}} = n^{-1}\boldsymbol{\varepsilon}_n^\top (\mathbf{H}_{M_{\hat{k}}} - \mathbf{H}_{M_{\tilde{k}}})\boldsymbol{\varepsilon}_n \\
&\leq 2\left\|\left\{n^{-1}F_{\hat{k},\tilde{k}}^\top(\mathbf{I}-\mathbf{H}_{M_{\tilde{k}}})F_{\hat{k},\tilde{k}}\right\}^{-1}\right\| \left\{\left\|n^{-1}F_{\hat{k},\tilde{k}}^\top \boldsymbol{\varepsilon}_n\right\|^2 + \left\|n^{-1}F_{\hat{k},\tilde{k}}^\top \mathbf{H}_{M_{\tilde{k}}}\,\boldsymbol{\varepsilon}_n\right\|^2\right\} \qquad (2.97) \\
&\leq 2(\hat{k}-\tilde{k})(\hat{a}_n + \hat{b}_n),
\end{aligned}$$

where

$$\begin{aligned}
\hat{a}_n =& \lambda_{\min}^{-1}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(\hat{J}_{\hat{k}})\mathbf{w}_t^\top(\hat{J}_{\hat{k}})\right) \max_{\substack{1\leq j\leq p_n \\ 1\leq l\leq r_j^{(n)}}} \left| n^{-1}\sum_{t=\bar{r}_n+1}^{n} x_{t-l,j}\epsilon_t\right|^2, \\
\hat{b}_n =& \lambda_{\min}^{-1}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n} \mathbf{w}_t(\hat{J}_{\hat{k}})\mathbf{w}_t^\top(\hat{J}_{\hat{k}})\right) \max_{\substack{\sharp(J)\leq \tilde{m}_n^* \\ (j,l)\notin J}} \left| n^{-1}\sum_{t=\bar{r}_n+1}^{n} \hat{x}_{t-l,j;J}\epsilon_t\right|^2,
\end{aligned}$$

with

$$\begin{aligned}
\hat{x}_{t-l,i;J} &:= \mathbf{w}_t^\top(J)\left(\sum_{t=\bar{r}_n+1}^{n}\mathbf{w}_t(J)\mathbf{w}_t^\top(J)\right)^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{w}_t(J)x_{t-l,i} \\
&= \mathbf{s}_t^\top(J)\left(\sum_{t=\bar{r}_n+1}^{n}\mathbf{s}_t(J)\mathbf{s}_t^\top(J)\right)^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{s}_t(J)x_{t-l,i}.
\end{aligned}$$

Combining (2.96) and (2.97), we have

$$2(\hat{k}-\tilde{k})(\hat{a}_n+\hat{b}_n) \geq \hat{\sigma}^2_{M_{\tilde{k}}}\{1-\exp(n^{-1}w_{n,p_n}(\tilde{k}-\hat{k}))\} \text{ on } \{\tilde{k}<\hat{k}\}\cap\mathcal{D}_n. \qquad (2.98)$$

With the help of the first part of (2.21) and Theorem 2.7.2, we also show in Section 2.7.3 that for any $\delta > 0$,

$$P\{(\hat{k}-\tilde{k})(\hat{a}_n+\hat{b}_n) \geq \delta[1-\exp(n^{-1}w_{n,p_n}(\tilde{k}-\hat{k}))], \tilde{k}<\hat{k}\} = o(1). \qquad (2.99)$$

As a consequence of (2.96)–(2.99) and (2.95), (2.88) follows. Thus, the proof of (2.85) is complete. Moreover, by an argument similar to that used to prove (2.85), it can be shown that (2.23) holds true. The details are omitted here for brevity. □

PROOF OF THEOREM 2.3.3 By Theorem 2.3.2,

$$\lim_{n\to\infty} P(\hat{s}_0 = s_0) = \lim_{n\to\infty} P(\hat{\underline{s}}_0 = \underline{s}_0) = 1. \tag{2.100}$$

In view of (2.100), Theorem 2.3.2, and $\mathbf{(SS_A)}$, it suffices for Theorem 2.3.3 to show that

$$\lim_{n\to\infty} P(\max_{1\le i\le q_n} |\hat{\alpha}_i(\mathcal{J}_n) - \alpha_i| \ge H_n) = 0, \tag{2.101}$$

where H_n is $\hat{H}_n$ with $\hat{s}_0$ and $\hat{\underline{s}}_0$ replaced by s_0 and $\underline{s}_0$, respectively.

Straightforward calculations yield

$$\max_{1\le i\le q_n} |\hat{\alpha}_i(\mathcal{J}_n) - \alpha_i| \le C(J_{1,n} + J_{2,n} + J_{3,n}), \tag{2.102}$$

where

$$\begin{aligned}
J_{1,n} &= \|(n^{-1}\sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(\mathcal{J}_n)\mathbf{s}_t^\top(\mathcal{J}_n))^{-1} - \mathbf{S}_n^{-1}(\mathcal{J}_n))\| \|n^{-1}\sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(\mathcal{J}_n)\epsilon_t\|,\\
J_{2,n} &= n^{-1/2}\|(n^{-1}\sum_{t=\bar{r}_n+1}^{n} \tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t^\top)^{-1} n^{-1}\sum_{t=\bar{r}_n+1}^{n} \tilde{\mathbf{y}}_t\epsilon_t\|,\\
J_{3,n} &= \max_{1\le i\le q_n-d} |\boldsymbol{\nu}_i^\top n^{-1}\sum_{t=\bar{r}_n+1}^{n} (\mathbf{z}_t^\top(q_n-d), \mathbf{x}_t^\top(\mathcal{J}_n))^\top \epsilon_t|,
\end{aligned}$$

where $\boldsymbol{\nu}_i$ is the i-th column (row) of $\mathbf{\Gamma}^{-1}(\mathcal{J}_n)$.

By Lemma 2.7.1, Theorem 2.7.2, (2.79), and (2.112) and (2.118) in Section 2.7.3, it can

be shown that

$$J_{1,n} = O_p\left(\frac{(s_0+q_n)^{3/2}}{n}\right) = O_p\left(\frac{n^{1/2}+q_n^{3/2}}{n}\right) \text{ and } J_{2,n} = O_p(n^{-1}). \tag{2.103}$$

Since (2.42) ensures

$$\max_{1\le i\le q_n-d} \|\boldsymbol{\nu}_i\| \le C, \tag{2.104}$$

it follows from (2.112) that

$$J_{3,n} \le C\|n^{-1}\sum_{t=\bar{r}_n+1}^{n} (\mathbf{z}_t(q_n-d), \mathbf{x}_t^\top(\mathcal{J}_n))^\top \epsilon_t\| = O_p\left(\frac{(s_0+q_n)^{1/2}}{n^{1/2}}\right). \tag{2.105}$$

In addition, we write

$$\boldsymbol{\nu}_i^\top(\mathbf{z}_t^\top(q_n-d), \mathbf{x}_t^\top(\mathcal{J}_n))^\top = \sum_{m=0}^{\infty} (\boldsymbol{c}_m^{(i)})^\top \boldsymbol{\mu}_{t-1-m}(\mathcal{D}_0),$$

where $\mathcal{D}_0$ and $\boldsymbol{\mu}_t(\cdot)$ are defined in $\mathbf{(SS_A)}$ and (2.56), respectively, and $\{\boldsymbol{c}_m^{(i)}\}$ is a sequence of $(\underline{s}_0+1)$-dimensional vectors depending on $\boldsymbol{\nu}_i$, $\{p_{m,j}\}, 1\le j\le p_n$, $\{b_m\}$, and $\beta_j^{(l)}, 1\le j\le r_j^{(n)}, 1\le j\le p_n$. By (2.42) and (2.104),

$$\max_{1\le i\le q_n-d}\sum_{m=0}^{\infty}\|(c_{m,1}^{(i)},\ldots,c_{m,\underline{s}_0+1}^{(i)})^\top\|^2 \equiv \max_{1\le i\le q_n-d}\sum_{m=0}^{\infty}\|\boldsymbol{c}_m^{(i)}\|^2 \le C,$$

which, together with Theorem 3.1 of Huang et al. (2023), gives,

$$\begin{aligned}
&\max_{1\le i\le q_n-d}\mathbb{E}\left|n^{-1/2}\sum_{t=\bar{r}_n+1}^{n}[\sum_{m=0}^{\infty}(\boldsymbol{c}_m^{(i)})^\top\boldsymbol{\mu}_{t-1-m}(\mathcal{D}_0)]\epsilon_t\right|^\eta\\
&\le C\max_{1\le i\le q_n-d}\left\{\sum_{j=1}^{\underline{s}_0+1}(\sum_{m=0}^{\infty}{c_{m,j}^{(i)}}^2)^{1/2}\right\}^\eta \le C(\max_{1\le i\le q_n-d}\sum_{m=0}^{\infty}\|\boldsymbol{c}_m^{(i)}\|^2)^{\eta/2}(\underline{s}_0+1)^{\eta/2}\\
&\le C\underline{s}_0^{\eta/2},
\end{aligned}$$

yielding

$$J_{3,n}\le\max_{1\le i\le q_n-d}\left|n^{-1}\sum_{t=\bar{r}_n+1}^{n}[\sum_{m=0}^{\infty}(\boldsymbol{c}_m^{(i)})^\top\boldsymbol{\mu}_{t-1-m}(\mathcal{D}_0)]\epsilon_t\right| = O_p\left(\frac{q_n^{1/\eta}\underline{s}_0^{1/2}}{n^{1/2}}\right). \quad (2.106)$$

Consequently, (2.101) follows from (2.102), (2.103), (2.105), and (2.106). Thus, the proof of Theorem 2.3.3 is complete. □

2.7.3 Proofs of (2.75), (2.77), (2.78), (2.93)–(2.95), and (2.99)

PROOF OF (2.75). Let's recall $\hat{a}_m$ as defined in (2.17). It follows from (2.73) that

$$\hat{a}_m=\hat{a}_{m-1}-\psi^2_{J_{m-1},(j_m,l_m)}\le\hat{a}_{m-1}-\xi^2\max_{(j,l)\notin J_{m-1}}\psi^2_{J_{m-1},(j,l)}. \quad (2.107)$$

Moreover, since for $1\le m\le K_n$,

$$\begin{aligned}
\hat{a}_{m-1}&=n^{-1}\boldsymbol{\mu}_n^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J_{m-1}})\boldsymbol{\mu}_n\\
&\ge(\sum_{(j,l)\in\mathcal{J}_n-J_{m-1}}{\beta_l^{(j)}}^2)\min_{1\le\sharp(J)\le K_n}\lambda_{\min}(n^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{w}_t(\mathcal{J}_n\cup J)\mathbf{w}_t^\top(\mathcal{J}_n\cup J)),
\end{aligned}$$

it holds that

$$\begin{aligned}\hat{a}_{m-1} &= \sum_{(j,l)\in\mathcal{J}_n} \beta_l^{(j)} n^{-1} \boldsymbol{\mu}_n^\top (\mathbf{I} - \mathbf{H}_{[q_n]\oplus J_{m-1}}) \boldsymbol{x}_l^{(j)} \\ &\le \max_{(j,l)\in\mathcal{J}_n - J_{m-1}} n^{-1} |\boldsymbol{\mu}_{\boldsymbol{n}}^\top (\mathbf{I} - \mathbf{H}_{[q_n]\oplus J_{m-1}}) \boldsymbol{x}_l^{(j)}| s_0^{1/2} \Big(\sum_{(j,l)\in\mathcal{J}_n - J_{m-1}} {\beta_l^{(j)}}^2 \Big)^{1/2} \\ &\le \max_{(j,l)\notin J_{m-1}} |\psi_{J_{m-1},(j,l)}| \hat{a}_{m-1}^{1/2} s_0^{1/2} D_n^{-1/2},\end{aligned} \tag{2.108}$$

where D_n is defined in (2.76). Equations (2.107) and (2.108) imply for $1 \le m \le K_n$,

$$\hat{a}_m \le \hat{a}_{m-1} \left(1 - \frac{\xi^2 D_n}{s_0}\right),$$

noting that D_n is bounded by 1. Thus, as long as a selection path obeying (2.73) is chosen, the resultant noiseless mean squared error satisfies

$$\hat{a}_m \le \hat{a}_0 \exp(-\xi^2 m D_n / s_0) \le C_n \exp(-\xi^2 m D_n / s_0),\ 1 \le m \le K_n, \tag{2.109}$$

where C_n is also defined (2.76).

Now since (2.74) ensures that on $\mathcal{A}_n(K_n) \cap \mathcal{B}_n(K_n)$, $\{\hat{J}_1, \ldots, \hat{J}_{K_n}\}$ obeys (2.73), with $0 < \xi < 1$ defined after (2.74), we conclude that (2.109) holds with $\hat{a}_m$ replaced by $\hat{s}_m$ on $\mathcal{A}_n(K_n) \cap \mathcal{B}_n(K_n)$. This completes the proof of (2.75). □

PROOF OF (2.77). By an argument similar to (2.108), one has for $1 \le m \le K_n$,

$$\begin{aligned} n^{-1} \boldsymbol{\mu}_n^\top (\mathbf{I} - \mathbf{H}_{[q_n]\oplus \hat{J}_m}) \boldsymbol{\mu}_n &\le \min_{0\le k\le m-1} n^{-1} \boldsymbol{\mu}_n^\top (\mathbf{I} - \mathbf{H}_{[q_n]\oplus \hat{J}_k}) \boldsymbol{\mu}_n \\ &\le \min_{0\le k\le m-1} \max_{(j,l)\notin \hat{J}_k} \psi^2_{\hat{J}_k,(j,l)} s_0 D_n^{-1}. \end{aligned} \tag{2.110}$$

Consequently, (2.77) follows from (2.110) and

$$\min_{0\le k\le m-1}\max_{(j,l)\notin\hat{J}_k}\psi^2_{\hat{J}_k,(j,l)}\le\frac{\tilde{\xi}^2 l_n p_n^{*(q_0+1)/(\eta q_0)}}{n}$$

on $\mathcal{B}_n^c(m)$. □

To prove (2.78), we need an auxiliary lemma.

Lemma 2.7.2. *Assume that* (A1), (A2), (A4), *and* (A5) *hold. Then,*

$$\begin{aligned}&\max_{1\le l\le r_j^{(n)},1\le j\le p_n}\left|n^{-1}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t x_{t-l,j}\right|+\max_{1\le k\le q_n}\left|n^{-1}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t z_{t-k}\right|\\&=O_p\left(\frac{p_n^{*(q_0+1)/(2\eta q_0)}}{n^{1/2}}+\frac{q_n^{1/\eta}}{n^{1/2}}\right)=O_p\left(\frac{p_n^{*(q_0+1)/(2\eta q_0)}}{n^{1/2}}\right).\end{aligned}\tag{2.111}$$

PROOF. The first identity of (2.111) is ensured by

$$\max_{1\le l\le r_j^{(n)},1\le j\le p_n}\mathbb{E}\left|n^{-1/2}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t x_{t-l,j}\right|^{\frac{2\eta q_0}{q_0+1}}+\max_{1\le k\le q_n-d}\mathbb{E}\left|n^{-1/2}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t z_{t-k}\right|^{\eta}<C,\tag{2.112}$$

which can be proved using Burkholder's inequality, Jensen's inequality, Hölder's inequality, $\mathrm{E}|x_{t-l,j}|^{2\eta q_0}<C$, and $\mathrm{E}|z_{t-k}|^{2\eta}<C$ for all $-\infty<t<\infty, 1\le l\le r_j^{(n)}, 1\le j\le p_n$, and $1\le k\le q_n-d$. The second identity of (2.111) follows from $q_n=o(n^{1/2})$ and $p_n^*\asymp n^\nu$ with $\nu\ge 1$. □

PROOF OF (2.78). It suffices for (2.78) to show that

$$\begin{aligned}&\max_{\sharp(J)\le K_n-1\atop(i,l)\notin J}|\psi_{J,(i,l)}-\hat{\psi}_{J,(i,l)}|=\max_{\sharp(J)\le K_n-1\atop(i,l)\notin J}\frac{n^{-1}|\boldsymbol{\varepsilon}_n^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)}|}{(n^{-1}{\boldsymbol{x}_l^{(i)}}^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)})^{1/2}}\\&=O_p\left(\frac{p_n^{*(q_0+1)/(2\eta q_0)}}{n^{1/2}}\right),\end{aligned}\tag{2.113}$$

which is, in turn, ensured by

$$\max_{\substack{\sharp(J)\le K_n-1\\(i,l)\notin J}} n^{-1}|\boldsymbol{\varepsilon}_n^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)}| = O_p\left(\frac{p_n^{*(q_0+1)/(2\eta q_0)}}{n^{1/2}}\right) \tag{2.114}$$

and

$$\max_{\substack{\sharp(J)\le K_n-1\\(i,l)\notin J}} (n^{-1}{\boldsymbol{x}_l^{(i)}}^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)})^{-1/2} = O_p(1). \tag{2.115}$$

Note that (2.115) is an immediate consequence of

$$\max_{\substack{\sharp(J)\le K_n-1\\(i,l)\notin J}} |n^{-1}{\boldsymbol{x}_l^{(i)}}^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)}|^{-1/2} \le \max_{\sharp(J)\le K_n} \lambda_{\min}^{-1/2}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{w}_t(J)\mathbf{w}_t^\top(J)\right)$$

and Theorem 2.7.2. Hence, it remains to prove (2.114). Since

$$\begin{aligned}\max_{\substack{\sharp(J)\le K_n-1\\(i,l)\notin J}} \frac{1}{n}|\boldsymbol{\varepsilon}_n^\top(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(i)}| \le & \max_{\substack{1\le i\le p_n\\1\le l\le r_i^{(n)}}}\left|\frac{1}{n}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t x_{t-l,i}\right| \\ & + \max_{\substack{\sharp(J)\le K_n-1\\(i,j)\notin J}}\left|\frac{1}{n}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t\hat{x}_{t-l,i;J}\right|,\end{aligned}$$

(2.114) follows from

$$\max_{\substack{\sharp(J)\le K_n-1\\(i,l)\notin J}}\left|\frac{1}{n}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t\hat{x}_{t-l,i;J}\right| = O_p\left(\frac{p_n^{*(q_0+1)/(2\eta q_0)}}{n^{1/2}}\right) \tag{2.116}$$

in light of Lemma 2.7.2.

For $(i,l) \notin J$

$$\begin{aligned}\left| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t \hat{x}_{t-l,i;J} \right| \leq & \left\| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t \mathbf{s}_t(J) \right\| \times \\ & \left\| \left(n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(J) \mathbf{s}_t^\top(J) \right)^{-1} \right\| \left\| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(J) x^{\perp}_{t-l,i;J} \right\| \\ & + \left| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t \mathbf{s}_t^\top(J) \mathbf{b}_J(i,l) \right|, \end{aligned} \tag{2.117}$$

where $x^{\perp}_{t-l,i;J} = x_{t-l,i} - \mathbf{s}_t^\top(J)\mathbf{b}_J(i,l)$ and $\mathbf{b}_J(i,l) = (\underbrace{0,\ldots,0}_{d}, \mathbf{g}_J^\top(i,l)\mathbf{\Gamma}_n^{-1}(J))^\top$. By Lemma 2.7.2, (2.43), (2.112), and

$$\max_{1\leq k\leq d} \mathrm{E}|n^{-1/2} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t \tilde{y}_{t,k}|^\eta < C, \tag{2.118}$$

one obtains

$$\max_{\substack{\sharp(J)\leq K_n-1 \\ (i,j)\notin J}} \left| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t \mathbf{s}_t^\top(J) \mathbf{b}_J(i,l) \right| = O_p\left(\frac{p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}} \right) \tag{2.119}$$

and

$$\max_{\sharp(J)\leq K_n-1} \left\| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t \mathbf{s}_t(J) \right\| = O_p\left(\frac{K_n^{1/2} p_n^{*\frac{q_0+1}{2\eta q_0}} + q_n^{1/2}}{n^{1/2}} \right). \tag{2.120}$$

Define

$$\text{(I)} = \left\| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(J) x_{t-l,i} - (\underbrace{0,\ldots,0}_{d}, \mathbf{g}_J^\top(i,l))^\top \right\|,$$

$$\text{(II)} = \left\| (n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(J)\mathbf{s}_t^\top(J) - \mathbf{S}_n(J))\mathbf{b}_J(i,l) \right\|.$$

Then,

$$\left\| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(J) x_{t-l,i;J}^{\perp} \right\| \le \text{(I)} + \text{(II)}. \tag{2.121}$$

It follows from Lemma 2.7.1 that

$$\begin{aligned}
&\max_{\sharp(J)\le K_n-1,(i,l)\notin J} \text{(I)} \le \sqrt{d} \max_{\substack{1\le l\le r_i^{(n)} \\ 1\le i\le p_n, 1\le k\le d}} |n^{-1} \sum_{t=\bar{r}_n+1}^{b} \tilde{y}_{t,k} x_{t-l,i}| \\
&+ \left\{ \sum_{k=1}^{q_n-d} \max_{1\le l\le r_i^{(n)}, 1\le i\le p_n} \left| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \{z_{t-k}x_{t-l,i} - \mathbb{E}(z_{t-k}x_{t-l,i})\} \right|^2 \right\}^{1/2} \\
&+ \left\{ K_n \max_{\substack{1\le l_1\le r_{i_1}^{(n)}, 1\le l\le r_i^{(n)} \\ 1\le i_1, i\le p_n}} \left| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \{x_{t-l_1,i_1}x_{t-l,i} - \mathbb{E}(x_{t-l_1,i_1}x_{t-l,i})\} \right|^2 \right\}^{1/2} \qquad (2.122) \\
&= O_p\left(\frac{p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}} \right) + O_p\left(\frac{q_n p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}} \right) + O_p\left(\frac{K_n^{1/2} p_n^{*\frac{2}{\eta q_0}}}{n^{1/2}} \right) \\
&= O_p\left(\frac{K_n^{1/2} p_n^{*\frac{2}{\eta q_0}} + q_n p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}} \right).
\end{aligned}$$

By Lemma 2.7.1 and (2.43), we also show below that

$$\max_{\sharp(J)\leq K_n-1,(i,l)\notin J}(\mathrm{II}) = O_p\left(\frac{K_n^{1/2}{p_n^*}^{\frac{2}{\eta q_0}} + (K_n^{1/2}+q_n^{1/2}){p_n^*}^{\frac{q_0+1}{2\eta q_0}}}{n^{1/2}}\right). \tag{2.123}$$

According to (2.121)–(2.123),

$$\begin{aligned}&\max_{\sharp(J)\leq K_n-1,(i,l)\notin J}\left\|n^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{s}_t(J)x_{t-l,i;J}^{\perp}\right\| \\ &= O_p\left(\frac{K_n^{1/2}{p_n^*}^{\frac{2}{\eta q_0}} + (K_n^{1/2}+q_n^{1/2}){p_n^*}^{\frac{q_0+1}{2\eta q_0}}}{n^{1/2}}\right).\end{aligned} \tag{2.124}$$

Consequently, (2.117), (2.119), (2.120), (2.124), and (2.71) imply

$$\begin{aligned}&\max_{\sharp(J)\leq K_n-1,(i,l)\notin J}\left|n^{-1}\sum_{t=\bar{r}_n+1}^{n}\epsilon_t\hat{x}_{t-l,i;J}\right| = O_p\left(\frac{K_n^{1/2}{p_n^*}^{\frac{2}{\eta q_0}} + (K_n^{1/2}+q_n^{1/2}){p_n^*}^{\frac{q_0+1}{2\eta q_0}}}{n^{1/2}}\right) \\ &\times O_p\left(\frac{K_n^{1/2}{p_n^*}^{\frac{q_0+1}{2\eta q_0}} + q_n^{1/2}}{n^{1/2}}\right),\end{aligned}$$

which, together with (2.12) and (A5), leads to (2.116). Thus, the proof is complete. □

Proof of (2.123). Note first that

$$
\begin{aligned}
&\max_{\sharp(J)\le K_n-1,(i,l)\notin J} (\mathrm{II}) \\
&\le \max_{\sharp(J)\le K_n-1,(i,l)\notin J} \Bigg\{ \sum_{k=1}^{d} \Big[\sum_{s=1}^{q_n-d} a_{s,J}(i,l)(n^{-1} \sum_{t=\bar{r}_n+1}^{n} \tilde{y}_{t,k} z_{t-s}) \\
&+ \sum_{(i^*,l^*)\in J} a_{(i^*,l^*)}(i,l)(n^{-1} \sum_{t=\bar{r}_n+1}^{n} \tilde{y}_{t,k} x_{t-l^*,i^*}) \Big]^2 \Bigg\}^{1/2} \\
&+ \max_{\sharp(J)\le K_n-1,(i,l)\notin J} \Bigg\{ \sum_{k=1}^{q_n-d} \Big[\sum_{s=1}^{q_n-d} a_{s,J}(i,l)(n^{-1} \sum_{t=\bar{r}_n+1}^{n} z_{t-k} z_{t-s} - \mathbb{E}(z_{t-k} z_{t-s})) \\
&+ \sum_{(i^*,l^*)\in J} a_{(i^*,l^*)}(i,l)(n^{-1} \sum_{t=\bar{r}_n+1}^{n} z_{t-k} x_{t-l^*,i^*} - \mathbb{E}(z_{t-k} x_{t-l^*,i^*})) \Big]^2 \Bigg\}^{1/2} \\
&+ \max_{\sharp(J)\le K_n-1,(i,l)\notin J} \Bigg\{ \sum_{(\tilde{i},\tilde{l})\in J} \Big[\sum_{s=1}^{q_n-d} a_{s,J}(i,l)(n^{-1} \sum_{t=\bar{r}_n+1}^{n} x_{t-\tilde{l},\tilde{i}} z_{t-s} - \mathbb{E}(x_{t-\tilde{l},\tilde{i}} z_{t-s})) \\
&+ \sum_{(i^*,l^*)\in J} a_{(i^*,l^*)}(i,l)(n^{-1} \sum_{t=\bar{r}_n+1}^{n} x_{t-\tilde{l},\tilde{i}} x_{t-l^*,i^*} - \mathbb{E}(x_{t-\tilde{l},\tilde{i}} x_{t-l^*,i^*})) \Big]^2 \Bigg\}^{1/2} \\
&\equiv (\mathrm{III}) + (\mathrm{IV}) + (\mathrm{V}).
\end{aligned}
\tag{2.125}
$$

By Lemma 2.7.1 and (2.43), it can be shown that

$$
(\mathrm{III}) = O_p\left(\frac{p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}} \right) \tag{2.126}
$$

and

$$
(\mathrm{IV}) = O_p\left(\frac{q_n^{2^{-1}+\eta^{-1}+\delta I_{\{\eta=2\}}}}{n^{1/2}} + \frac{q_n^{1/2} p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}} \right) = O_p\left(\frac{q_n^{1/2} p_n^{*\frac{q_0+1}{2\eta q_0}}}{n^{1/2}} \right), \tag{2.127}
$$

where $\delta > 0$ is arbitrarily small and the second equality is ensured by (A5). Moreover, it follows that

$$
\begin{aligned}
(\mathrm{V})^2 &\le C(K_n-1)\sum_{s_1=1}^{q_n-d}\sum_{s_2=1}^{q_n-d} b_{s_1}b_{s_2}A_{s_1}A_{s_2} \\
&+ C \max_{\sharp(J)\le K_n-1,(i,l)\notin J} \sum_{(i^*,l^*)\in J} |a_{(i^*,l^*)}(i,l)| \\
&\times \max_{\substack{1\le \tilde{i},i^*\le p_n \\ 1\le \tilde{l}\le r_{\tilde{i}}^{(n)},\, 1\le l^*\le r_{i^*}^{(n)}}} \left| n^{-1}\sum_{t=\bar{r}_n+1}^{n} x_{t-\tilde{l},\tilde{i}}x_{t-l^*,i^*} - \mathbb{E}(x_{t-\tilde{l},\tilde{i}}x_{t-l^*,i^*}) \right|^2 \qquad (2.128) \\
&= O_p\left(\frac{K_n(p_n^{*\frac{q_0+1}{\eta q_0}} + p_n^{*\frac{4}{\eta q_0}})}{n} \right),
\end{aligned}
$$

where

$$
\begin{aligned}
b_s &= \max_{\sharp(J)\le K_n-1,(i,l)\notin J} |a_{s,J}(i,l)|, \\
A_s &= \max_{1\le i\le p_n, 1\le l\le r_i^{(n)}} |n^{-1}\sum_{t=\bar{r}_n+1}^{n} x_{t-l,i}z_{t-s} - \mathbb{E}(x_{t-l,i}z_{t-s})|.
\end{aligned}
$$

Combining (2.125)–(2.128) yields (2.123). □

Proof of (2.93)–(2.95). Since

$$
\begin{aligned}
\hat{A}_n^{-1} &\le \max_{\substack{\sharp(J)\le K_n-1 \\ (j,l)\notin J}} \{n^{-1}\boldsymbol{x}_l^{(j)\top}(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{x}_l^{(j)}\}^{-1}, \\
|\hat{B}_n| &\le \max_{\substack{\sharp(J)\le K_n-1 \\ (j,l)\notin J}} |n^{-1}\boldsymbol{x}_l^{(j)\top}(\mathbf{I}-\mathbf{H}_{[q_n]\oplus J})\boldsymbol{\varepsilon}_n|,
\end{aligned}
$$

(2.93) and (2.94) follow directly from (2.114) and (2.115), respectively. To show (2.95), note

first that

$$|\hat{C}_n| \leq \left| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t^2 - \sigma^2 \right| + n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{H}_{[q_n] \oplus \hat{J}_{\tilde{k}}} \boldsymbol{\varepsilon}_n. \tag{2.129}$$

By Assumption (A1), it is easy to show that

$$|n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t^2 - \sigma^2| = O_p(n^{-1/2}). \tag{2.130}$$

In addition,

$$n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{H}_{[q_n] \oplus \hat{J}_{\tilde{k}}} \boldsymbol{\varepsilon}_n \leq \max_{\sharp(J) \leq m_n^*} \lambda_{\min}^{-1} \left(n^{-1} \sum_{t=\bar{r}_n+1}^{n} \mathbf{s}_t(J) \mathbf{s}_t^\top(J) \right)$$
$$\times \max_{\sharp(J) \leq m_n^*} \left\| n^{-1} \sum_{t=\bar{r}_n+1}^{n} \epsilon_t \mathbf{s}_t(J) \right\|^2 \text{ on } \mathcal{D}_n,$$

which, together with (2.129), (2.130), (2.120), (2.71), and (2.86), gives (2.95). □

Proof of (2.99). Note first that for some $c_1 > 0$,

$$\frac{1 - \exp(-w_{n,p_n} n^{-1} (\hat{k} - \tilde{k}))}{\hat{k} - \tilde{k}} \geq c_1 \Big\{ \frac{w_{n,p_n}}{n} \wedge \frac{1}{\hat{k} - \tilde{k}} \Big\} \geq c_1 \Big\{ \frac{w_{n,p_n}}{n} \wedge K_n^{-1} \Big\} \text{ on } \{\tilde{k} < \hat{k}\}.$$

Define $B_{n,p_n} = (w_{n,p_n}/n) \wedge K_n^{-1}$. Then, it follows from (2.12) and the first part of (2.21) that

$$p_n^{*\bar{\theta}} / n^{1/2} = o(B_{n,p_n}^{1/2}). \tag{2.131}$$

Now, for any $\delta > 0$,

$$
\begin{aligned}
&P\{(\hat{k}-\tilde{k})(\hat{a}_n+\hat{b}_n) \geq \delta[1-\exp(-n^{-1}w_{n,p_n}(\hat{k}-\tilde{k})]), \tilde{k} < \hat{k}\} \\
\leq & P\left(\lambda_{\min}^{-1}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{s}_t(\hat{J}_{\hat{k}})\mathbf{s}_t^{\top}(\hat{J}_{\hat{k}})\right)\max_{\substack{1\leq j\leq p_n \\ 1\leq l\leq r_j^{(n)}}}\left|n^{-1}\sum_{t=\bar{r}_n+1}^{n}x_{t-l,j}\epsilon_t\right|^2 \geq c_1\delta B_{n,p_n}\right) \\
&+P\left(\lambda_{\min}^{-1}\left(n^{-1}\sum_{t=\bar{r}_n+1}^{n}\mathbf{s}_t(\hat{J}_{\hat{k}})\mathbf{s}_t^{\top}(\hat{J}_{\hat{k}})\right)\max_{\substack{\sharp(J)\leq K_n-1 \\ (j,l)\notin J}}\left|n^{-1}\sum_{t=\bar{r}_n+1}^{n}\hat{x}_{t-l,j;J}\epsilon_t\right|^2 \geq c_1\delta B_{n,p_n}\right) \\
:=&\text{(I)} + \text{(II)}.
\end{aligned}
$$

By (2.111), (2.116), Theorem 2.7.2, and (2.131), (I) + (II) = $o(1)$. Thus (2.99) is proved. □

2.7.4 Some technical details about Examples 2.3.1 and 2.3.2 in Section 2.3.1

Proof of (2.16) in Example 2.3.1

In this subsection, all summations are understood as summing from $t = 3$ to $t = n$. Let $z_t = y_t - y_{t-1}$. Clearly, $z_t = az_{t-1} + \epsilon_t$ for $t = 1, 2, \ldots, n$. Note that with some algebraic manipulation and using the AR definition, we can express

$$
\begin{aligned}
F_{2,n} =& \frac{(\sum y_t y_{t-1})^2}{\sum y_{t-1}^2} + \frac{(\sum y_t y_{t-1})^2}{\sum y_{t-1}^2}\left(\frac{2\sum y_{t-2}z_{t-1} + \sum z_{t-1}^2}{\sum y_{t-2}^2}\right) \\
&- \frac{2(\sum y_t y_{t-1})(\sum y_t z_{t-1})}{\sum y_{t-1}^2}\left(1 + \frac{2\sum y_{t-2}z_{t-1} + \sum z_{t-1}^2}{\sum y_{t-2}^2}\right) \\
&+ \frac{(\sum y_t z_{t-1})^2}{\sum y_{t-1}^2}\left(1 + \frac{2\sum y_{t-2}z_{t-1} + \sum z_{t-1}^2}{\sum y_{t-2}^2}\right).
\end{aligned}
$$

By a similar argument used in Lemma 2.7.1 and Theorem 2.7.2, we have

$$\begin{aligned}\frac{1}{n}(F_{1,n}-F_{2,n}) =& -\frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2}\frac{\sum y_t y_{t-1}}{\sum y_{t-2}^2}(2n^{-1}\sum y_{t-2}z_{t-1}+n^{-1}\sum z_{t-1}^2)\\ &+2\frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2}(n^{-1}\sum y_t z_{t-1})\left(1+\frac{2\sum y_{t-2}z_{t-1}+\sum z_{t-1}^2}{\sum y_{t-2}^2}\right)\\ &-\frac{(\sum y_t z_{t-1})^2}{\sum y_{t-1}^2}\left(\frac{1}{n}+\frac{2\sum y_{t-2}z_{t-1}+\sum z_{t-1}^2}{n\sum y_{t-2}^2}\right)\\ =&(-1+O_p(n^{-1}))\left(2n^{-1}\sum y_{t-2}z_{t-1}+n^{-1}\sum z_{t-1}^2\right)\\ &+2(n^{-1}\sum y_t z_{t-1})(1+O_p(n^{-1}))+O_p(n^{-1})\\ =&n^{-1}\sum z_{t-1}^2+2n^{-1}\sum z_t z_{t-1}+O_p(n^{-1}),\end{aligned}$$

which implies

$$\frac{1}{n}(F_{1,n}-F_{2,n})\to\frac{1}{1-a^2}+\frac{2a}{1-a^2}\text{ in probability.}$$

□

Proof of (2.28) in Example 2.3.2

Note that

$$\begin{aligned}\mathcal{A}_n &= \{\hat{\boldsymbol{\beta}}^{(\lambda_n)}\text{ selects the correct model }\} = \{\hat{\beta}_1^{(\lambda_n)}\neq 0,\hat{\beta}_3^{(\lambda_n)}\neq 0,\hat{\beta}_2^{(\lambda_n)}=0\}\\ &= \big\{\boldsymbol{s}_n(1)=(\text{sign}(\hat{\beta}_1^{(\lambda_n)}),\text{sign}(\hat{\beta}_3^{(\lambda_n)}))^\top\in\{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_4\}\text{ and }\hat{\beta}_2^{(\lambda_n)}=0\big\},\end{aligned}\tag{2.132}$$

where $\boldsymbol{a}_1^\top = (1,1)$, $\boldsymbol{a}_2^\top = (1,-1)$, $\boldsymbol{a}_3^\top = (-1,1)$, and $\boldsymbol{a}_4^\top = (-1,-1)$. Define

$$\mathbf{C}_{11} = \sum_{t=3}^{n} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} \begin{pmatrix} y_{t-1} & x_{t-1} \end{pmatrix}, \quad \mathbf{c}_{21} = \sum_{t=3}^{n} \begin{pmatrix} y_{t-1}y_{t-2} & x_{t-1}y_{t-2} \end{pmatrix},$$
$$\hat{\mathbf{u}}_n = \begin{pmatrix} \hat{\beta}_1^{(\lambda_n)} - 1 \\ \hat{\beta}_3^{(\lambda_n)} - 1 \end{pmatrix}, \quad \mathbf{w}_n(1) = \sum_{t=3}^{n} \epsilon_t \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix}, \quad w_n(2) = \sum_{t=3}^{n} \epsilon_t y_{t-2}.$$

Then by an argument used in Zhao and Yu (2006),

$$\mathcal{A}_n \subseteq \bigcup_{i=1}^{4} \mathcal{E}_n(i), \tag{2.133}$$

where

$$\mathcal{E}_n(i) = \{\mathbf{C}_{11}\hat{\mathbf{u}}_n - \mathbf{w}_n(1) = -\frac{\lambda_n \boldsymbol{a}_i}{2},\ -\frac{\lambda_n}{2} \le \mathbf{c}_{21}\hat{\mathbf{u}}_n - w_n(2) \le \frac{\lambda_n}{2},\ \boldsymbol{s}_n(1) = \boldsymbol{a}_i\}.$$

In the following, we will show that regardless of whether $\{\lambda_n\}$ satisfies (a') $\lambda_n/n \to \infty$, (b') $\lambda_n/n \to 0$, or (c') $0 < \lim_{n\to\infty} \lambda_n/n = d^* < \infty$,

$$\limsup_{n\to\infty} P(\mathcal{E}_n(1)) \le 1/2, \tag{2.134}$$

and

$$\lim_{n\to\infty} P(\mathcal{E}_n(i)) = 0, i = 2,3,4. \tag{2.135}$$

By (2.132)–(2.135), the desired conclusion (2.28) follows.

We commence by proving (2.134). Straightforward calculations give

$$\mathcal{E}_n(1) \subseteq \left\{ \mathbf{c}_{21}\mathbf{C}_{11}^{-1}\mathbf{w}_n(1) - w_n(2) \ge -\frac{\lambda_n}{2}\left(1 - \mathbf{c}_{21}\mathbf{C}_{11}^{-1}\boldsymbol{a}_1\right) \right\}, \tag{2.136}$$

$$\mathbf{c}_{21}\mathbf{C}_{11}^{-1} = \left(1 + O_p(n^{-1}) \quad -n^{-1}\sum_{t=3}^{n} x_{t-1}\delta_{t-1} + O_p(n^{-1})\right),$$

where $\delta_t = x_{t-1} + \epsilon_t$,

$$\mathbf{c}_{21}\mathbf{C}_{11}^{-1}\mathbf{w}_n(1) - w_n(2) = \sum_{t=3}^{n} \delta_{t-1}\epsilon_t + O_p(1), \tag{2.137}$$

$$1 - \mathbf{c}_{21}\mathbf{C}_{11}^{-1}\boldsymbol{a}_1 = \frac{1}{n}\sum_{t=3}^{n} x_{t-1}\delta_{t-1} + O_p(1/n), \tag{2.138}$$

and

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{t=3}^{n} \delta_{t-1}\epsilon_t \\ \frac{1}{\sqrt{n}}\sum_{t=3}^{n} x_{t-1}\delta_{t-1} \end{pmatrix} \Rightarrow \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}, \tag{2.139}$$

where N_1 and N_2 are two independent Gaussian random variables with mean zero and variance 2. By (2.136)–(2.139), it holds that

$$\begin{aligned}
P(\mathcal{E}_n(1)) &\le P\left(\frac{1}{\sqrt{n}}\sum_{t=3}^{n} \delta_{t-1}x_{t-1} \ge o_p(1)\right) \to 1/2 \text{ in case } (a'), \\
P(\mathcal{E}_n(1)) &\le P\left(\frac{1}{\sqrt{n}}\sum_{t=3}^{n} \delta_{t-1}\epsilon_t \ge o_p(1)\right) \to 1/2 \text{ in case } (b'), \\
P(\mathcal{E}_n(1)) &\le P\left(\frac{1}{\sqrt{n}}\sum_{t=3}^{n} \delta_{t-1}\epsilon_t \ge \frac{-\lambda_n}{2n}\frac{1}{\sqrt{n}}\sum_{t=3}^{n} \delta_{t-1}x_{t-1} + o_p(1)\right) \\
&\to P\left(N_1 \ge -\frac{d^*}{2}N_2\right) = 1/2 \text{ in case } (c').
\end{aligned} \tag{2.140}$$

Thus, (2.134) follows.

For $i = 2, 3$, and 4,

$$\begin{aligned}\mathcal{E}_n(i) &\subseteq \{\hat{\mathbf{u}}_n = \mathbf{C}_{11}^{-1}[\mathbf{w}_n(1) - (\lambda_n \boldsymbol{a}_i)/2],\ \boldsymbol{s}_n(1) = \boldsymbol{a}_i\} \\ &= \{(\hat{\beta}_1^{(\lambda_n)}, \hat{\beta}_3^{(\lambda_n)})^\top = (1,1)^\top + \mathbf{C}_{11}^{-1}[\mathbf{w}_n(1) - (\lambda_n \boldsymbol{a}_i)/2],\ \boldsymbol{s}_n(1) = \boldsymbol{a}_i\}.\end{aligned} \tag{2.141}$$

By an argument similar to that used to prove (2.137) and (2.138),

$$\begin{aligned}&\mathbf{C}_{11}^{-1}[\mathbf{w}_n(1) - (\lambda_n \boldsymbol{a}_i)/2] \\ &= \left(O_p\big(n^{-1} + \frac{\lambda_n}{n^2}\big), O_p(n^{-1/2}) + \frac{\lambda_n}{2n}(I_{\{i=2,4\}} - I_{\{i=3\}})(1 + o_p(1)).\right)^\top\end{aligned} \tag{2.142}$$

Combining (2.141) and (2.142) yields that there exist an arbitrarily small positive constant $\varepsilon > 0$ and an arbitrarily large positive constant $M < \infty$ such that for all sufficiently large n,

$$\begin{aligned}P(\mathcal{E}_n(j)) &\le P(\hat{\beta}_3^{(\lambda_n)} > 1 - \varepsilon, \hat{\beta}_3^{(\lambda_n)} < 0) + o(1) \\ &= o(1) \text{ in all cases of } (a'), (b'), \text{ and } (c'),\end{aligned} \tag{2.143}$$

where $j = 2, 4$, and

$$\begin{aligned}&P(\mathcal{E}_n(3)) \le P(\hat{\beta}_3^{(\lambda_n)} < -M, \hat{\beta}_3^{(\lambda_n)} > 0) + o(1) = o(1) \text{ in case } (a'), \\ &P(\mathcal{E}_n(3)) \le P(\hat{\beta}_1^{(\lambda_n)} > 1 - \varepsilon, \hat{\beta}_1^{(\lambda_n)} < 0) + o(1) = o(1) \text{ in cases } (b') \text{ and } (c').\end{aligned} \tag{2.144}$$

Consequently, (2.135) is ensured by (2.143) and (2.144). This completes the proof of (2.28).

□

2.7.5 Complementary simulation results

We generate data from

$$(1 + 0.4B)(1 - B)^2 y_t = \sum_{j=1}^{2} \sum_{l=1}^{4} \beta_l^{(j)} x_{t-l,j} + \epsilon_t, \tag{2.145}$$

where $\{\epsilon_t\}$ is a GARCH(1,1) model,

$$\begin{aligned}\epsilon_t =&\sigma_t Z_t,\\ \sigma_t^2 =&5\times 10^{-2}+0.5\epsilon_{t-1}^2+0.1\sigma_{t-1}^2,\end{aligned}$$

in which $\{Z_t\}$ is a sequence of i.i.d. standard Gaussian random variables. Let $\{\pi_{t,1}\}$ and $\{\pi_{t,2}\}$ be two independent ARCH(1) processes such that for $j=1$ and 2,

$$\begin{aligned}\pi_{t,j} =&h_{t,j}G_{t,j},\\ h_{t,j}^2 =&1+0.2\pi_{t-1,j}^2,\end{aligned}$$

where $\{G_{t,1}\}$ and $\{G_{t,2}\}$ have the same probabilistic structure as that of $\{Z_t\}$ and these three sequences are independent of each other. Also let $v_{t,j}$, $1\le t\le n, 1\le j\le p_n$, be independent standard Gaussian random variables and independent of $\{G_{t,1}\}$, $\{G_{t,2}\}$, and $\{Z_t\}$. Define $w_{t,j}=\pi_{t,1}+v_{t,j}$ if j is odd, $w_{t,j}=\pi_{t,2}+v_{t,j}$ if j is even. Then, $x_{t,j}$ are MA(2) processes satisfying $x_{t,j}=0.8w_{t,j}+0.1w_{t-1,j}$ if j is odd and $x_{t,j}=0.2w_{t,j}+0.6w_{t-1,j}$ otherwise. The coefficients are set to

$$\begin{aligned}(\beta_1^{(1)},\beta_2^{(1)},\beta_3^{(1)},\beta_4^{(1)}) =&(-7.62,6.72,-5.55,3.77),\\ (\beta_1^{(2)},\beta_2^{(2)},\beta_3^{(2)},\beta_4^{(2)}) =&(6.89,-6.18,4.47,-3.10).\end{aligned}$$

Using Theorem 2.2 of Ling and McAleer (2002) again, one can verify that ϵ_t only has a finite fourth moment and $x_{t,j}$ has a finite twelfth moment. Moreover, it is easy to show that (A1) and (A2) in Section 2.3.3 are fulfilled by the above model specification.

One distinct feature of this example is that the error term and all candidate covariates are conditionally heteroscedastic. Table 2.5 records the performance of the methods introduced in Section 2.4 based on 1000 replications and $(n,p_n,r^{(n)})=(800,250,4),(1000,275,5)$, and

(1500, 300, 6). The table reveals that FHTD is the only method that efficiently identifies the correct ARX model. More specifically, it successfully chooses the correct ARX model over 89% of the time, in all cases of n considered in this example.

Table 2.5: Values of E, SS, TP, and FP in Example 2.145

	LASSO	ALasso	OGA-3	AR-ALasso	AR-OGA-3	FHTD
$(n, p_n^*, p_n, r^{(n)}, q_n) = (800, 1000, 250, 4, 10)$						
E	0	0	0	0	137	926
SS	0	0	0	0	138	1000
TP	1.08	1.00	3.81	1.00	9.27	11.00
FP	0.13	0.00	0.17	0.00	5.39	0.09
$(n, p_n^*, p_n, r^{(n)}, q_n) = (1000, 1375, 275, 5, 11)$						
E	0	0	0	0	181	891
SS	0	0	0	0	183	932
TP	1.05	1.00	3.56	1.00	9.32	10.83
FP	0.17	0.00	0.35	0.00	6.31	0.32
$(n, p_n^*, p_n, r^{(n)}, q_n) = (1500, 1800, 300, 6, 12)$						
E	0	0	0	0	299	960
SS	0	0	0	0	301	989
TP	1.02	1.00	3.19	1.00	9.57	10.99
FP	0.19	0.01	0.24	0.00	5.83	0.08

CHAPTER 3

SCALABLE HIGH-DIMENSIONAL MULTIVARIATE LINEAR REGRESSION FOR FEATURE-DISTRIBUTED DATA

3.1 Introduction

A computational strategy often adopted for tackling high-dimensional big data is to employ feature-distributed analysis: to partition the data by features and to store them across multiple computing nodes. For instance, when the data have an extremely large number of features that do not fit in a single computer, this strategy is used to circumvent storage constraints or to accelerate computation (Heinze et al., 2016; Wang et al., 2017; Richtárik and Takáč, 2016; Gao and Tsay, 2023). In addition, feature-distributed data may be inevitable when the data are collected and maintained by multiple parties. Because of bandwidth or administrative reasons, merging them in a central computing node from those sources might not be feasible (Hu et al., 2019). In some applications, data come naturally feature-distributed, such as the wireless sensor networks (Bertrand and Moonen, 2010, 2014, 2015).

A challenge in estimating statistical models with feature-distributed data is to avoid the high *communication complexity*, which is the amount of data that are transmitted across the nodes. Indeed, because distributed computing systems typically operate under limited bandwidth, sending voluminous data significantly slows down the algorithm. Unfortunately, data transmission is often a necessary evil with feature-distributed data: each node by itself is unable to learn about the parameters associated with the features it does not own. Thus, algorithms that have lower communication complexities are preferred in practice.

Based on the rationale that the empirical minimizers of certain optimization problems are desirable statistical estimators, prior works have proposed various optimization algorithms with feature-distributed data. Richtárik and Takáč (2016) and Fercoq et al. (2014) employed randomized coordinate descent to solve ℓ_1-regularized problems and to exploit par-

allel computation from the distributed computing system. In addition, random projection techniques were used in Wang et al. (2017) and Heinze et al. (2016) for ℓ_2-regularized convex problems. However, for estimating linear models, the existing approaches usually incur a high communication complexity for very large data sets. To illustrate, consider the Lasso problem. The Hydra algorithm of Richtárik and Takáč (2016) requires $O(np\log(1/\epsilon))$ bytes of communication to reach ϵ-close to the optimal loss, where n is the sample size and p is the number of features. For data with extremely large p and n that do not fit in a single modern computer, such communication complexity appears prohibitively expensive. Similarly, the distributed iterative dual random projection (DIDRP) algorithm of Wang et al. (2017) needs $O(n^2 + n\log(1/\epsilon))$ bytes of total communication for estimating the ridge regression, where the dominating n^2 factor comes from each node sending the sketched data matrix to a coordinator node. Thus it incurs not only a high communication cost but also a storage bottleneck.

This chapter proposes a two-stage relaxed greedy algorithm (TSRGA) for feature-distributed data to mitigate the high communication complexity. TSRGA first applies the conventional relaxed greedy algorithm (RGA) to feature-distributed data. But we terminate the RGA with the help of a just-in-time stopping criterion, which aims to save excessive communication via reducing RGA iterations. In the second stage, we employ a modification of RGA to estimate the coefficient matrices associated with the selected predictors from the first stage. The modified second-stage RGA yields low-rank coefficient matrices, that exploit information across tasks and improve statistical performance.

Instead of treating TSRGA as merely an optimization means, we directly analyze the convergence of TSRGA to the unknown parameters, which in turn implies the communication costs of TSRGA. The key insight of the proposed method is that the conventional RGA often incurs a high communication cost because it takes many iterations to minimize its loss function, but it tends to select relevant predictors in its early iterations. Therefore, one

should decide when the RGA has done screening the predictors *before* it iterates too many steps. To this end, the just-in-time stopping criterion tracks the reduction in training error in each step, and calls for halting the RGA as soon as the reduction becomes smaller than some threshold. With the potential predictors narrowed down in the first stage, the second-stage employs a modified RGA and focuses on the more amenable problem of estimating the coefficient matrices of the screened predictors. The two-stage design enables TSRGA to substantially cut down the communication costs and produce even more accurate estimates than the original RGA.

Our theoretical results show that the proposed TSRGA enjoys a communication complexity of $O_p(\mathfrak{s}_n(n + d_n))$ bytes, up to a multiplicative term depending logarithmically on the problem dimensions, where d_n is the dimension of the response vector (or the number of tasks), and $\mathfrak{s}_n$ is a sparsity parameter defined later. This communication complexity improves that of Hydra by a factor of $p/\mathfrak{s}_n$, and is much smaller than that of DIDRP and other one-shot algorithms (for example, Wang et al. 2016 and Heinze et al. 2016) if $\mathfrak{s}_n \ll n$. The RGA was also employed by Bellet et al. (2015) as a solver for ℓ_1-constrained problems, but it requires $O(n/\epsilon)$ communication since it only converges at a sub-linear rate (see also Jaggi, 2013 and Garber, 2020), where ϵ is again the optimization tolerance. Hence TSRGA offers a substantial speedup for estimating sparse models compared to the conventional RGA.

To validate the performance of TSRGA, we apply it to both synthetic and real-world data sets and show that TSRGA converges much faster than other existing methods. In the simulation experiments, TSRGA achieved the smallest estimation error using the least number of iterations. It also outperforms other centralized iterative algorithms both in speed and statistical accuracy. In a large-scale simulation experiment, TSRGA can effectively estimate the high-dimensional multivariate linear regression model with more than 16 GB data in less than 5 minutes. For an empirical application, we apply TSRGA to predict simultaneously some financial outcomes (volatility, trading volume, market beta, and returns) of the S&P

500 component companies using textual features extracted from their 10-K reports. The results show that TSRGA efficiently utilizes the information provided by the texts and works well with high dimensional feature matrices.

Finally, we propose some extensions of TSRGA. First, we also considered applying TSRGA to big feature-distributed data which have not only many features but also a large number of observations. Thus, in addition to separately storing each predictors in different computing nodes, it is also necessary to partition the observations of each feature into chunks that could fit in one node. In this case, the computing nodes shall coordinate both horizontally and vertically, and we show that the communication cost to carry out TSRGA in this setting is still free of p, but could be larger than that of the purely feature-distributed case. Second, the idea of TSRGA can be extended beyond linear regression models. In Section 3.8.4, we show how TSRGA can be applied to the generalized linear models.

For ease in reading, we collect the notations used throughout the chapter here. The transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^\top$ and that of a vector $\mathbf{v}$ is $\mathbf{v}^\top$. The inner product between two vectors $\mathbf{u}$ and $\mathbf{v}$ is denoted interchangeably as $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$. If $\mathbf{A}, \mathbf{B}$ are $\mathbb{R}^{m \times n}$, $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ denotes their trace inner product. The minimum and maximum eigenvalues of a matrix $\mathbf{A}$ are denoted by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$, respectively. We also denote by $\sigma_l(\mathbf{A})$ the l-th singular value of $\mathbf{A}$, in descending order. When the argument is a vector, $\|\cdot\|$ denotes the usual Euclidean norm and $\|\cdot\|_p$ the ℓ_p norm. If the argument is a matrix, $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_{op}$ the operator norm, and $\|\cdot\|_*$ the nuclear norm. For a set J, $\sharp(J)$ denotes its cardinality. For an event $\mathcal{E}$, its complement is denoted as $\mathcal{E}^c$ and its associated indicator function is denoted as $\mathbf{1}\{\mathcal{E}\}$. For two positive (random) sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = o_p(y_n)$ if $\lim_{n\to\infty} \mathbb{P}(x_n/y_n < \epsilon) = 1$ for any $\epsilon > 0$ and write $x_n = O_p(y_n)$ if for any $\epsilon > 0$ there exists some $M_\epsilon < \infty$ such that $\limsup_{n\to\infty} \mathbb{P}(x_n/y_n > M_\epsilon) < \epsilon$.

3.2 Distributed framework and two-stage relaxed greedy algorithm

In this section, we first introduce the multivariate linear regression model considered in the chapter and show how the data are distributed across the nodes. Then we lay out the implementation details of the proposed TSRGA, which consists of two different implementations of the conventional RGA and a just-in-time stopping criterion to guide the termination of the first-stage RGA. The case of needing horizontal partition will be discussed in Section 3.6.

3.2.1 Model and distributed framework

Consider the following multivariate linear regression model:

$$\mathbf{y}_t = \sum_{j=1}^{p_n} \mathbf{B}_j^{*\top} \mathbf{x}_{t,j} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, n, \tag{3.1}$$

where $\mathbf{y}_t \in \mathbb{R}^{d_n}$ is the response vector, $\mathbf{x}_{t,j} \in \mathbb{R}^{q_{n,j}}$ a multivariate predictor, for $j = 1, 2, \ldots, p_n$, and $\mathbf{B}_j^*$ is the $(q_{n,j} \times d_n)$ unknown coefficient matrix, for $j = 1, \ldots, p_n$. In particular, we are most interested in the case $p_n \gg n$ and $q_{n,j} < n$. Clearly, when $d_n = q_{n,1} = \ldots = q_{n,p_n} = 1$, (3.1) reduces to the usual multiple linear regression model. Without loss of generality, we assume $\mathbf{y}_t$, $\mathbf{x}_{t,j}$ and $\boldsymbol{\epsilon}_t$ are mean zero.

There are several motivations for considering general d_n and $q_{n,j}$'s. First, imposing group-sparsity can be advantageous when the predictors display a natural grouping structure (e.g. Lounici et al. 2011). This advantage is inherited by (3.1) when only a limited number of $\mathbf{B}_j^*$'s are non-zero. Second, it is not uncommon that we are interested in modeling more than one response variable ($d_n > 1$). In this case, one can gain statistical accuracy if the prediction tasks are related, which is often embodied by the assumption that $\mathbf{B}_j^*$'s are of low rank (see, e.g., Reinsel et al. 2022). In modern machine learning, some predictors may be

constructed from unstructured data sources. For instance, for functional data, $\mathbf{x}_{t,j}$'s may be the first few Fourier coefficients (Fan et al., 2015). On the other hand, for textual data, $\mathbf{x}_{t,j}$'s may be topic loading or outputs from some pre-trained neural networks (Kogan et al., 2009; Yeh et al., 2020; Bybee et al., 2021). Finally, model (3.1) can also accommodate the so-called multi-view of multi-modal data, which have also received considerable attention in recent years.

Next, we specify how the data are distributed across computing nodes. In matrix notations, we can write (3.1) as

$$\mathbf{Y} = \sum_{j=1}^{p_n} \mathbf{X}_j \mathbf{B}_j^* + \mathbf{E}, \tag{3.2}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$, $\mathbf{X}_j = (\mathbf{x}_{1,j}, \ldots, \mathbf{x}_{n,j})^\top \in \mathbb{R}^{n \times q_{n,j}}$, for $j = 1, 2, \ldots, p_n$, and $\mathbf{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\top$. As discussed in the Introduction, since pooling the large matrices $\mathbf{X}_1, \ldots, \mathbf{X}_{p_n}$ in a central node may not be feasible, a common strategy is to store them across nodes. In the following, we suppose that M nodes are available. Furthermore, the i-th node contains the data $\{\mathbf{Y}, \mathbf{X}_j : j \in \mathcal{I}_i\}$, for $i = 1, 2, \ldots, M$, where $\cup_{i=1}^M \mathcal{I}_i = \{1, 2, \ldots, p_n\} := [p_n]$. For ease in exposition, we assume a master node coordinates the other computing nodes. In particular, each worker node is able to send and receive data from the master node.

3.2.2 First-stage relaxed greedy algorithm and a just-in-time stopping criterion

We now introduce the first-stage RGA and describe how it can be applied to feature-distributed data. First, initialize $\hat{\mathbf{G}}^{(0)} = \mathbf{0}$ and $\hat{\mathbf{U}}^{(0)} = \mathbf{Y}$. For iteration $k = 1, 2, \ldots,$

RGA finds $(\hat{j}_k, \tilde{\mathbf{B}}_{\hat{j}_k})$ such that

$$(\hat{j}_k, \tilde{\mathbf{B}}_{\hat{j}_k}) \in \arg \max_{\substack{1 \le j \le p_n \\ \|\mathbf{B}_j\|_* \le L_n}} \langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_j \mathbf{B}_j \rangle, \tag{3.3}$$

where $L_n = d_n^{1/2} L_0$ for some large constant $L_0 > 0$. Then RGA constructs updates by

$$\begin{aligned} \hat{\mathbf{G}}^{(k)} =& (1 - \hat{\lambda}_k) \hat{\mathbf{G}}^{(k-1)} + \hat{\lambda}_k \mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k}, \\ \hat{\mathbf{U}}^{(k)} =& \mathbf{Y} - \hat{\mathbf{G}}^{(k)}, \end{aligned} \tag{3.4}$$

where $\hat{\lambda}_k$ is determined by

$$\hat{\lambda}_k \in \arg \min_{0 \le \lambda \le 1} \|\mathbf{Y} - (1 - \lambda) \hat{\mathbf{G}}^{(k-1)} - \lambda \mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k}\|_F. \tag{3.5}$$

RGA has important computational advantages that are attractive for big data computation. First, for a fixed j, the maximum in (3.3) is achieved at $\mathbf{B}_j = L_n \mathbf{u}\mathbf{v}^\top$, where $(\mathbf{u}, \mathbf{v})$ is the leading pair of singular vectors (i.e., corresponding to the largest singular value) of $\mathbf{X}_j^\top \hat{\mathbf{U}}^{(k-1)}$. Since computing the leading singular vectors is much cheaper than full SVD, RGA is computationally lighter than algorithms using singular value soft-thresholding, such as the alternating direction method of multipliers (ADMM). This feature has already been exploited in Zheng et al. (2018) and Zhuo et al. (2020) for nuclear-norm constrained optimization. Second, $\hat{\lambda}_k$ is easy to compute and has the closed-form $\hat{\lambda}_k = \max\{\min\{\hat{\lambda}_{k,uc}, 1\}, 0\}$, where

$$\hat{\lambda}_{k,uc} = \frac{\langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)} \rangle}{\|\mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}$$

is the unconstrained minimizer of (3.5).

When applied to feature-distributed data, we can leverage these advantages. Observe

from (3.3)-(3.5) that the history of RGA is encoded in $\hat{\mathbf{G}}^{(k)}$. That is, to construct $\hat{\mathbf{G}}^{(k+1)}$, which predictors were chosen and the order in which they were chosen are irrelevant, provided $\hat{\mathbf{G}}^{(k)}$ is known. In particular, each node only needs $\hat{\lambda}_{k+1}$ and $\mathbf{X}_{\hat{j}_{k+1}}\tilde{\mathbf{B}}_{\hat{j}_{k+1}}$ to construct $\hat{\mathbf{G}}^{(k+1)}$. As argued in the previous paragraph, $\mathbf{X}_{\hat{j}_{k+1}}\tilde{\mathbf{B}}_{\hat{j}_{k+1}}$ is a rank-one matrix. Thus transmitting this matrix only requires $O(n + d_n)$ bytes of communication, which are much lighter than that of the full matrix with $O(nd_n)$ bytes. In addition, each node requires only the extra memory to store $\hat{\mathbf{G}}^{(k)}$ throughout the training. This is less burdensome than random projection techniques, which require at least one node to make extra room to store the sketched matrix of size $O(n^2)$.

The above discussions are summarized in Algorithm 1, detailing how workers and the master node communicate to implement RGA with feature-distributed data. Clearly, each node sends and receives data of size $O(n + d_n)$ bytes (line 4 and 15) in each iteration. We remark that Algorithm 1 asks each node to send the potential updates to the master (line 15). This is for reducing rounds of communications, which can be a bottleneck in practice. If bandwidth limit is more stringent, one can instead first ask the workers to send ρ_c to the master. After master decides c^*, it only asks the c^*-th node to send the update, so that only one node is transmitting the data.

Although the per-iteration communication complexity is low for RGA, the total communication can still be costly if the required number of iteration is high. Indeed, RGA converges to $\arg\min_{\sum_{j=1}^{p_n}\|\mathbf{B}_j\|_*\le L_n}\|\mathbf{Y}-\sum_{j=1}^{p_n}\mathbf{X}_j\mathbf{B}_j\|_F^2$ at the rate $O(k^{-1})$, where k is the number of iterations (Jaggi, 2013; Temlyakov, 2015). There are many attempts to design variants of RGA that converge faster (see Jaggi and Lacoste-Julien, 2015; Lei et al., 2019; Garber, 2020 and references therein). Instead of adapting these increasingly sophisticated optimization schemes with feature-distributed data, we propose to terminate RGA early with the help of a just-in-time stopping criterion. The key insight, as to be shown in Theorem 3.3.1, is that RGA is capable of screening relevant predictors in the early iterations. The stopping

Algorithm 1: Feature-distributed relaxed greedy algorithm (RGA)

Input: Number of maximum iterations K_n; $L_n > 0$.

Output: Each worker $1 \le c \le M$ obtains the coefficient matrices $\{\hat{\mathbf{B}}_j : j \in \mathcal{I}_c\}$.

Initialization: $\hat{\mathbf{B}}_j = \mathbf{0}$ for all j and $\hat{\mathbf{G}}^{(0)} = \mathbf{0}$

for $k = 1, 2, \ldots, K_n$ **do**

  Workers $c = 1, 2, \ldots, M$ **in parallel do**

    if $k > 1$ **then**

      Receive $(c^*, \hat{\lambda}_{k-1}, \sigma_{\hat{j}_{k-1}}, \mathbf{u}_{\hat{j}_{k-1}}, \mathbf{v}_{\hat{j}_{k-1}})$ from the master.

      $\hat{\mathbf{G}}^{(k-1)} = (1 - \hat{\lambda}_{k-1})\hat{\mathbf{G}}^{(k-2)} + \hat{\lambda}_{k-1}\sigma_{\hat{j}_{k-1}}\mathbf{u}_{\hat{j}_{k-1}}\mathbf{v}_{\hat{j}_{k-1}}^\top$.

      $\hat{\mathbf{B}}_j = (1 - \hat{\lambda}_{k-1})\hat{\mathbf{B}}_j$ for $j \in \mathcal{I}_c$.

      if $c = c^*$ **then**

        $\hat{\mathbf{B}}_{\hat{j}^{(c)}_{k-1}} = \hat{\mathbf{B}}_{\hat{j}^{(c)}_{k-1}} + \hat{\lambda}_{k-1}\tilde{\mathbf{B}}_{\hat{j}^{(c)}_{k-1}}$

      end

    end

    $\hat{\mathbf{U}}^{(k-1)} = \mathbf{Y} - \hat{\mathbf{G}}^{(k-1)}$

    $(\hat{j}^{(c)}_k, \tilde{\mathbf{B}}_{\hat{j}^{(c)}_k}) \in \arg\max_{\substack{j \in \mathcal{I}_c \\ \|\mathbf{B}_j\|_* \le L_n}} |\langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_j \mathbf{B}_j \rangle|$

    $\rho_c = |\langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_{\hat{j}^{(c)}_k}\tilde{\mathbf{B}}_{\hat{j}^{(c)}_k}\rangle|$

    Find the leading singular value decomposition: $\mathbf{X}_{\hat{j}^{(c)}_k}\tilde{\mathbf{B}}_{\hat{j}^{(c)}_k} = \sigma_{\hat{j}^{(c)}_k}\mathbf{u}_{\hat{j}^{(c)}_k}\mathbf{v}_{\hat{j}^{(c)}_k}^\top$

    Send $(\sigma_{\hat{j}^{(c)}_k}, \mathbf{u}_{\hat{j}^{(c)}_k}, \mathbf{v}_{\hat{j}^{(c)}_k}, \rho_c)$ to the master.

  end

  Master do

    Receives $\{(\sigma_{\hat{j}^{(c)}_k}, \mathbf{u}_{\hat{j}^{(c)}_k}, \mathbf{v}_{\hat{j}^{(c)}_k}, \rho_c) : c = 1, 2, \ldots, M\}$ from the workers.

    $c^* = \arg\max_{1 \le c \le N} \rho_c$

    $\sigma_{\hat{j}_k} = \sigma_{\hat{j}^{(c^*)}_k}, \mathbf{u}_{\hat{j}_k} = \mathbf{u}_{\hat{j}^{(c^*)}_k}, \mathbf{v}_{\hat{j}_k} = \mathbf{v}_{\hat{j}^{(c^*)}_k}$

    $\hat{\mathbf{G}}^{(k)} = (1 - \hat{\lambda}_k)\hat{\mathbf{G}}^{(k-1)} + \hat{\lambda}_k\sigma_{\hat{j}_k}\mathbf{u}_{\hat{j}_k}\mathbf{v}_{\hat{j}_k}^\top$, where $\hat{\lambda}_k$ is determined by

$$\hat{\lambda}_k \in \arg\min_{0 \le \lambda \le 1} \|\mathbf{Y} - (1 - \lambda)\hat{\mathbf{G}}^{(k-1)} - \lambda\sigma_{\hat{j}_k}\mathbf{u}_{\hat{j}_k}\mathbf{v}_{\hat{j}_k}^\top\|_F^2.$$

    Broadcasts $(c^*, \hat{\lambda}_k, \sigma_{\hat{j}_k}, \mathbf{u}_{\hat{j}_k}, \mathbf{v}_{\hat{j}_k})$ to all workers.

  end

end

criterion is defined as follows. Let $\hat{\sigma}_k^2 = (nd_n)^{-1}\|\mathbf{Y} - \hat{\mathbf{G}}^{(k)}\|_F^2$. We terminate the first-stage

RGA at step $\hat{k}$, defined as

$$\hat{k} = \min\left\{1 \le k \le K_n : \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2} \ge 1 - t_n\right\}, \tag{3.6}$$

and $\hat{k} = K_n$ if $\hat{\sigma}_k^2/\hat{\sigma}_{k-1}^2 < 1-t_n$, for all $1 \le k \le K_n$, where t_n is some threshold specified later and K_n is a prescribed maximum number of iterations. Intuitively, $\hat{k}$ is determined based on whether the current iteration provides sufficient improvement in reducing the training error. Note that $\hat{k}$ is determined just-in-time without fully iterating K_n steps. The algorithm is halted once the criterion is triggered, thereby saving excessive communication costs. This is in sharp contrast to the model selection criteria used in prior works to terminate greedy-type algorithms that compare all K_n models, such as the information criteria (Ing and Lai, 2011; Ing, 2020).

3.2.3 Second-stage relaxed greedy algorithm

After the first-stage RGA is terminated, the second-stage RGA focuses on estimation of the coefficient matrices. In this stage, we implement a modified version of RGA so that the coefficient estimates are of low rank.

For predictors with "large" coefficient matrices, failing to account for their low-rank structure may result in statistical inefficiency. To see this, let $\hat{J} := \hat{J}_{\hat{k}}$ be the predictors selected by the first-stage RGA, and let $\hat{\mathbf{B}}_j$, $j \in \hat{J}$, be the corresponding coefficient estimates produced by the first-stage RGA. Assume for now $q_{n,j} = q_n$. If $\min\{q_n, d_n\} > \hat{r} = \sum_{j\in\hat{J}} \hat{r}_j$, where $\hat{r}_j = \operatorname{rank}(\hat{\mathbf{B}}_j)$, then estimating this coefficient matrix alone without regularization amounts to estimating $d_n q_n$ parameters. It will be shown later in Theorem 3.3.1 that $\hat{r}_j \ge \operatorname{rank}(\mathbf{B}_j^*)$ with probability tending to one. Since $d_n q_n \asymp \min\{d_n, q_n\}(q_n + d_n) > \hat{r}(q_n + d_n)$, estimating this coefficient matrix would cost us more than the best achievable degrees of freedom (Reinsel et al., 2022).

To avoid loss in efficiency for these large coefficient estimators, we impose a constraint on the space in which our final estimators reside. Suppose the j-th predictor, $j \in \hat{J}$, satisfies $\min\{q_{n,j}, d_n\} > \hat{r}$. We require its coefficient estimator to be of the form $\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{S}\mathbf{V}_j^\top$, where $\hat{\mathbf{\Sigma}}_j = n^{-1}\mathbf{X}_j^\top\mathbf{X}_j$; $\mathbf{U}_j = (\mathbf{u}_{1,j}, \ldots, \mathbf{u}_{\hat{r},j})$ and $\mathbf{V}_j = (\mathbf{v}_{1,j}, \ldots, \mathbf{v}_{\hat{r},j})$ form the leading $\hat{r}$ pairs of singular vectors of $\mathbf{X}_j^\top\mathbf{Y}$, and $\mathbf{S}$ is an $\hat{r} \times \hat{r}$ matrix to be optimized.

The second-stage RGA proceeds as follows. Initialize again $\hat{\mathbf{G}}^{(0)} = \mathbf{0}$ and $\hat{\mathbf{U}}^{(0)} = \mathbf{Y}$. For $k = 1, 2, \ldots$, choose

$$(\hat{j}_k, \hat{\mathbf{S}}_k) \in \arg \max_{\substack{j \in \hat{J} \\ \|\mathbf{S}\|_* \le L_n}} \langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{S}\mathbf{V}_j^\top \rangle, \tag{3.7}$$

where the maximum is searching over $\mathbf{S} \in \mathbb{R}^{\hat{r}\times\hat{r}}$ if $\hat{r} < \min\{q_{n,j}, d_n\}$. For j such that $\hat{r} \ge \min\{q_{n,j}, d_n\}$, we define $\mathbf{U}_j$ and $\mathbf{V}_j$ to be the full set of singular vectors and the maximum is searching over $\mathbf{S} \in \mathbb{R}^{q_{n,j}\times d_n}$. Next, we construct the update by

$$\begin{aligned} \hat{\mathbf{G}}^{(k)} =& (1-\hat{\lambda}_k)\hat{\mathbf{G}}^{(k-1)} + \hat{\lambda}_k \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \\ \hat{\mathbf{U}}^{(k)} =& \mathbf{Y} - \hat{\mathbf{G}}^{(k)}, \end{aligned} \tag{3.8}$$

where $\hat{\lambda}_k$ is, again, determined by

$$\hat{\lambda}_k \in \arg \min_{0\le\lambda\le 1} \|\mathbf{Y} - (1-\lambda)\hat{\mathbf{G}}^{(k-1)} - \lambda \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top\|_F^2. \tag{3.9}$$

At first glance, the updating scheme (3.7)-(3.9) may appear similar to those proposed by Ding et al. (2021) or Ding et al. (2020), but we note one important difference here: the matrices $\mathbf{U}_j$ and $\mathbf{V}_j$ are fixed at the onset of the second stage. Thus our estimators' ranks remain controlled, which is not the case in the aforementioned works. More comparisons between TSRGA and these works will be made in Section 3.3.2.

We briefly comment on the computational aspects of the second-stage RGA. First, simi-

larly to the first-stage, for a fixed j the maximum in (3.7) is attained at $\mathbf{S} = L_n \mathbf{u}\mathbf{v}^\top$, where $(\mathbf{u}, \mathbf{v})$ is the leading pair of singular vectors of $\mathbf{U}_j^\top \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{X}_j^\top \hat{\mathbf{U}}^{(k-1)} \mathbf{V}_j$, which can be computed locally by each node. As a result, the per-iteration communication is still $O(n + d_n)$ for each node. For $j \in \hat{J}$ with $\hat{r} \geq \min\{q_{n,j}, d_n\}$, since $\mathbf{U}_j$ and $\mathbf{V}_j$ are non-singular, the parameter space is not limited except for the bounded nuclear norm constraint. Indeed, it is not difficult to see that for such j,

$$\max_{\|\mathbf{S}\|_* \leq L_n} \langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{S} \mathbf{V}_j^\top \rangle$$

is equivalent to

$$\max_{\|\mathbf{B}\|_* \leq L_n} \langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{B} \rangle \tag{3.10}$$

with the correspondence $\mathbf{B} = \mathbf{U}_j \mathbf{S} \mathbf{V}_j^\top$. Thus, for such j, it is not necessary to compute the singular vectors $\mathbf{U}_j$ and $\mathbf{V}_j$. Instead, one can directly solve (3.10). Finally, it is straightforward to modify Algorithm 1 to implement the second-stage RGA with feature-distributed data. We defer the details to Section 3.8.1.

It is worth mentioning that the idea of two-stage RGA can be employed beyond the linear regression setup. For example, by replacing the squared loss with log likelihood function, we can use TSRGA to estimate generalized linear models, which include logistic regression for classification tasks and Poisson regression for modeling count data. The details of the modified algorithm are deferred to Section 3.8.4, where we also examine its performance through simulations.

3.2.4 Related algorithms

In this subsection, we consider TSRGA in several contexts and compare it with related algorithms. By viewing TSRGA as either a novel feature-distributed algorithm, an improve-

ment over the Frank-Wolfe algorithm, a new method to estimate the integrative multi-view regression (Li et al., 2019), or a close relative of the greedy-type algorithms (Temlyakov, 2000), we highlight both its computational ease in applying to feature-distributed data and its theoretical applicability in estimating high-dimensional linear models.

Over the last decade, a few methods for estimating linear regression with feature-distributed data have been proposed. For instance, Richtárik and Takáč (2016) and Fercoq et al. (2014) use randomized coordinate descent to solve ℓ_1-regularized optimization problem, and Hu et al. (2019) proposes an asynchronous stochastic gradient descent algorithm, to name just a few. These methods either require a communication complexity that scales with p_n, or converge only at sub-linear rates, both of which translate to high communication costs. The screen-and-clean approach of Yang et al. (2016), similar in spirit to TSRGA, first applies sure independence screening (SIS, Fan and Lv, 2008) to identify a subset of potentially relevant predictors. Then it uses an iterative procedure similar to the iterative Hessian sketch (Pilanci and Wainwright, 2016) to estimate the associated coefficients. While SIS does not require communication, it imposes stronger assumptions on the predictors and the error term. In contrast, the proposed TSRGA can be applied at low communication complexity without succumbing to those assumptions.

TSRGA also adds to the line of studies that attempt to modify the conventional Frank-Wolfe algorithm (Frank and Wolfe, 1956). RGA, more often called the Frank-Wolfe algorithm in the optimization literature, has been widely adopted in big data applications for its computational simplicity. Recently, various modifications of the Frank-Wolfe algorithm have been proposed to attain a linear convergence rate that does not depend on the feature dimension p_n (Lei et al., 2019; Garber, 2020; Ding et al., 2021, 2020). However, strong convexity or quadratic growth of the loss function is typically assumed in these works, which precludes high-dimensional data ($n \ll p_n$). Frank-Wolfe algorithm has also been found useful in distributed systems, though most prior works employed the horizontally-partitioned

data (Zheng et al., 2018; Zhuo et al., 2020). That is, data are partitioned and stored across nodes by observations instead of by features. A notable exception is Bellet et al. (2015), who found that Frank-Wolfe outperforms ADMM in communication and wall-clock time for sparse scalar regression with feature-distributed data, despite that Frank-Wolfe still suffers from sub-linear convergence. In this chapter, we neither assume strong convexity (or quadratic growth) nor limit ourselves to scalar regression, and TSRGA demands much less computation than the usual Frank-Wolfe algorithm.

Model (3.1) was also employed by Li et al. (2019), and they termed it the integrative multi-view regression. They propose an ADMM-based algorithm, integrative reduced-rank regression (iRRR), for optimization in a centralized computing framework. The major drawback, as discussed earlier, is a computationally-expensive step of singular value soft-thresholding. Thus, TSRGA can serve as a computationally attractive alternative. In Section 3.4, we compare their empirical performance and find that TSRGA is much more efficient.

Other closely related greedy algorithms such as the orthogonal greedy algorithm (OGA) have also been applied to high-dimensional linear regression. OGA, when used in conjunction with an information criterion, attains the optimal prediction error (Ing, 2020) under various sparsity assumptions. However, it is computationally less adaptable to feature-distributed data. To keep the per-iteration communication low, the sequential orthogonalization scheme of Ing and Lai (2011) can be used with feature-distributed data, but the individual nodes would not have the correct coefficients to use at the prediction time when new data, possibly not orthogonalized, become available. Alternatively, one needs to allocate extra memory in each node to store the history of the OGA path to compute the projection in each iteration.

3.3 Communication complexity of TSRGA

In this section, we derive theoretical guarantees on the communication complexity of TSRGA. Specifically, we show that the communication complexity of TSRGA does not scale with the feature dimension p_n, but instead depends on the sparsity of the underlying problem.

3.3.1 Assumptions

For the theoretical analysis, we maintain the following mild assumptions of model (3.1).

(C1) There exists some $\mu < \infty$ such that with probability approaching one,

$$\mu^{-1} \leq \min_{1 \leq j \leq p_n} \lambda_{\min}(\hat{\mathbf{\Sigma}}_j) \leq \max_{1 \leq j \leq p_n} \lambda_{\max}(\hat{\mathbf{\Sigma}}_j) \leq \mu,$$

where $\hat{\mathbf{\Sigma}}_j = n^{-1}\mathbf{X}_j^\top \mathbf{X}_j$ with $\mathbf{X}_j$ being defined in (3.2).

(C2) Put $\xi_E = \max_{1 \leq j \leq p_n} \|\mathbf{X}_j^\top \mathbf{E}\|_{op}$. There exists a sequence of $K_n \to \infty$ such that $K_n \xi_E = O_p(n d_n^{1/2})$.

(C3)

$$\lim_{n \to \infty} \mathbb{P}\left(\min_{\sharp(J) \leq 2K_n} \lambda_{\min}(n^{-1}\mathbf{X}(J)^\top \mathbf{X}(J)) > \mu^{-1} \right) = 1,$$

where $\mathbf{X}(J) = (\mathbf{X}_j : j \in J) \in \mathbb{R}^{n \times (\sum_{j \in J} q_{n,j})}$.

(C4) There exists some large $L < \infty$ such that $d_n^{-1/2} \sum_{j=1}^{p_n} \|\mathbf{B}_j^*\|_* \leq L$. Moreover, there exists a non-decreasing $\{s_n\}$ such that $s_n^2 = o(K_n)$ and

$$\min_{j \in J_n} \sigma_{r_j^*}^2 \left(d_n^{-1/2} \mathbf{B}_j^* \right) \geq s_n^{-1},$$

where $J_n = \{1 \leq j \leq p_n : \mathbf{B}_j^* \neq \mathbf{0}\}$ is the set of indices corresponding the relevant predictors, and $r_j^* = \text{rank}(\mathbf{B}_j^*)$.

These assumptions are quite standard. (C1) requires the variances of the predictors to be on the same order of magnitude, which is often the case if the predictors are normalized. ξ_E in (C2) is typically regarded as the effect size of the noise. Through auxiliary concentration inequalities in the literature, we will verify (C2) in the examples following the main result. (C3) assumes a lower bound on the minimum eigenvalue of the covariance matrices formed by small subsets of predictors. Note that (C3) could hold even when $p_n \gg n$ and the observations are dependent; we refer to Ing and Lai (2011) and Ing (2020) for related discussions on (C3). s_n in (C4) imposes a lower bound on the minimum non-zero singular value of the (normalized) coefficient matrices $d_n^{-1/2}\mathbf{B}_j^*$. Since (C4) implies $\sharp(J_n) \leq s_n^{1/2}L$, it can be interpreted as a measure of sparsity of the underlying model.

Next, we introduce two assumptions that are important to the feature-distributed problem. Let $\tilde{\mathbf{Y}} = \sum_{j=1}^{p_n} \mathbf{X}_j\mathbf{B}_j^*$ be the noiseless part of $\mathbf{Y}$.

(C5) Let $\bar{r}_j = \text{rank}(\mathbf{X}_j^\top\tilde{\mathbf{Y}})$ and $J_o = J_n \cap \{j : \min\{q_{n,j}, d_n\} > \bar{r}_j\}$. There exists $\delta_n > 0$ such that $\xi_E = o_p(n\delta_n)$ and with probability approaching one,

$$\min_{j\in J_o} \sigma_{\bar{r}_j}(\mathbf{X}_j^\top\tilde{\mathbf{Y}}) \geq n\delta_n.$$

(C6) (Local revelation) If the column vectors of $\tilde{\mathbf{U}}_j \in \mathbb{R}^{q_{n,j}\times\bar{r}_j}$ and $\tilde{\mathbf{V}}_j \in \mathbb{R}^{d_n\times\bar{r}_j}$ are the leading pairs of singular vectors corresponding to the non-zero singular values of $\mathbf{X}_j^\top\tilde{\mathbf{Y}}$, then with probability approaching one, there exists an $\bar{r}_j \times \bar{r}_j$ matrix $\mathbf{\Lambda}_j$ such that

$$\hat{\mathbf{\Sigma}}_j\mathbf{B}_j^* = \tilde{\mathbf{U}}_j\mathbf{\Lambda}_j\tilde{\mathbf{V}}_j^\top \tag{3.11}$$

for all $j \in J_o$.

(C5) and (C6) are assumptions that endow the local nodes sufficient information in the feature-distributed setting. Both assumptions concern relevant predictors that are "large" such that their dimensions $q_{n,j} \times d_n$ satisfy $\min\{q_{n,j}, d_n\} > \bar{r}_j$. Intuitively, (C5) requires, for relevant predictors which are of large dimension, the marginal correlations between these predictors and $\tilde{\mathbf{Y}}$ are sufficiently large. The local revelation condition (C6) assumes each node could use its local data to re-construct $\hat{\mathbf{\Sigma}}_j \mathbf{B}_j^*$ for $j \in J_o$. This would simplify information sharing between the nodes. Although they are key assumptions used to derive a fast convergence rate for the second-stage RGA, they are not needed for establishing the sure-screening property of the just-in-time stopping criterion (see Theorem 3.3.1). In addition, these two assumptions are vacuous when all predictors are of small dimensions. For instance, for scalar group-sparse linear regression, $\min\{d_n, q_{n,j}\} = \min\{1, q_{n,j}\} = 1 \leq \bar{r}_j$. Hence $J_o = \emptyset$ and the two assumptions are immaterial.

To better understand (3.11), consider the following example.

$$\mathbf{y}_t = \mathbf{B}_1^{*\top} \mathbf{x}_{t,1} + \mathbf{B}_2^{*\top} \mathbf{x}_{t,2} + \boldsymbol{\epsilon}_t,$$

where $\mathbf{B}_1^*, \mathbf{B}_2^*$ are rank-1 matrices such that $\mathbf{B}_1^* = \mathbf{u}_1^* \mathbf{v}_1^{*\top}$ and $\mathbf{B}_2^* = \mathbf{u}_2^* \mathbf{v}_2^{*\top}$. In matrix notation, we write $\mathbf{Y} = \mathbf{X}_1 \mathbf{B}_1^* + \mathbf{X}_2 \mathbf{B}_2^* + \mathbf{E}$. Suppose $q_{n,1} = q_{n,2} > 2$, and consider

$$\mathbf{X}_1^\top \tilde{\mathbf{Y}} = \underbrace{\begin{pmatrix} \mathbf{X}_1^\top \mathbf{X}_1 \mathbf{u}_1^* & \mathbf{X}_1^\top \mathbf{X}_2 \mathbf{u}_2^* \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} \mathbf{v}_1^{*\top} \\ \mathbf{v}_2^{*\top} \end{pmatrix}}_{\mathbf{B}}.$$

It is not difficult to show that (3.11) holds (for $j = 1$) if $\mathbf{A}$ and $\mathbf{B}$ are of full rank. Since $\mathbf{y}_t = (\mathbf{x}_{t,1}^\top \mathbf{u}_1^*)\mathbf{v}_1^* + (\mathbf{x}_{t,2}^\top \mathbf{u}_2^*)\mathbf{v}_2^*$, one can interpret $f_{t,j} = \mathbf{x}_{t,j}^\top \mathbf{u}_j^*$ as the predictive factor associated with predictor j, for $j = 1, 2$. $f_{t,j}$ has differential effects on each element of $\mathbf{y}_t$, which are determined by $\mathbf{v}_j^*$. Hence, that $\mathbf{B}$ has full rank translates to that the two factors $f_{t,1}$ and $f_{t,2}$ have distinct impacts on $\mathbf{y}_t$. On the other hand, $\mathbf{A}$ has full rank if and only if

$\mathbf{u}_1^* \neq \alpha(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \mathbf{X}_2 \mathbf{u}_2^*$ for any $\alpha \neq 0$. This implies the factor $f_{t,1}$ must not be equal to the projection of $f_{t,2}$ onto the space spanned by $\mathbf{X}_1$. Therefore, (3.11) can be interpreted as requiring the factors $f_{t,1}$ and $f_{t,2}$ are truly distinct and make distinguishable contributions to the response vector. Moreover, if (3.11) fails, the marginal product $\mathbf{X}_1^\top \tilde{\mathbf{Y}}$ may no longer be useful, because the signals are contaminated by possible collinearity.

3.3.2 Main results

We now present some theoretical properties of TSRGA, with proofs relegated to Section 3.8.2. In the following, we assume L_n, the hyperparameter input to the TSRGA algorithm, is chosen to be $L_n = d_n^{1/2} L_0$ with $L_0 \geq L/(1-\epsilon_L)$, where $1 - \epsilon_L \leq \mu^{-2}/4$.

Our first result proves that RGA, coupled with the just-in-time stopping criterion, can screen the relevant predictors. Moreover, it provides an upper bound on the rank of the corresponding coefficient matrices.

Theorem 3.3.1. *Assume (C1)-(C4) hold. Suppose there exists an $M_o < \infty$ such that $M_o^{-1} \leq (nd_n)^{-1}\|\mathbf{E}\|_F^2 \leq M_o$ with probability tending to one. Write $\hat{\mathbf{G}}^{(k)} = \sum_{j=1}^{p_n} \mathbf{X}_j \hat{\mathbf{B}}_j^{(k)}$, $k = 1, 2, \ldots, K_n$, for the iterates of the first-stage RGA. If $\hat{k}$ is defined by (3.6) with $t_n = Cs_n^{-2}$ for some sufficiently small $C > 0$, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\operatorname{rank}(\mathbf{B}_j^*) \leq \operatorname{rank}(\hat{\mathbf{B}}_j^{(\hat{k})}) \text{ for all } j\right) = 1. \tag{3.12}$$

Although Theorem 3.3.1 only provides an upper bound for the ranks of $\mathbf{B}_j^*$'s, it renders a useful diagnosis for the rank of the coefficient matrices for model (3.1). When $p_n = 1$, Bunea et al. (2011) proposed a rank selection criterion (RSC) to select the optimal reduced rank estimator, which is shown to be a consistent estimator of the effective rank. However, rank selection for model (3.1) with $p_n > 1$ is less investigated. Moreover, we can bound $\hat{k}$ by the following lemma.

Lemma 3.3.1. *Under the assumptions of Theorem 3.3.1,* $\hat{k} = O_p(s_n^2)$.

Lemma 3.3.1 ensures the just-in-time stopping criterion is triggered in no more than $O(s_n^2)$ iterations, which is much smaller than $O(K_n)$ by (C4). Thus compared to the model selection rules using information criteria that iterate K_n steps in full, the proposed method greatly reduces communication costs.

Next, we derive the required number of iterations for TSRGA to converge near the unknown parameters, which translates to its communication costs. With a slight abuse of notation, we also write the second-stage RGA iterates as $\hat{\mathbf{G}}^{(k)} = \sum_{j\in\hat{J}} \mathbf{X}_j \hat{\mathbf{B}}_j^{(k)}$.

Theorem 3.3.2. *Assume the assumptions of Theorem 3.3.1 hold, and additionally (C5) and (C6) also hold. If* $\xi_E = O_p(\xi_n)$ *and* $m_n = \lceil \rho\kappa_n \log(n^2 d_n/\xi_n^2) \rceil$ *for some sequence* $\{\xi_n\}$ *of positive numbers, where* $\rho = 64\mu^5/\tau^2$ *with* $0 < \tau < 1$ *being arbitrary, and*

$$\kappa_n = \sharp(\hat{J}) \max\left\{ \max_{j\in\hat{J}-\hat{J}_o} (q_{n,j} \wedge d_n), \hat{r}\mathbf{1}\{\hat{J}_o \neq \emptyset\} \right\},$$

with $a \wedge b = \min\{a, b\}$ *and* $\hat{J}_o = \{j \in \hat{J} : \hat{r} < \min\{q_{n,j}, d_n\}\}$, *then the proposed second-stage RGA satisfies*

$$\sup_{m\geq m_n} \frac{1}{d_n} \sum_{j=1}^{p_n} \|\mathbf{B}_j^* - \hat{\mathbf{B}}_j^{(m)}\|_F^2 = O_p\left(\frac{\kappa_n \xi_n^2}{n^2 d_n} \log \frac{n^2 d_n}{\xi_n^2} + \frac{\xi_n^2}{n^2 \delta_n^2} \mathbf{1}\{J_o \neq \emptyset\} \right).$$

Since the per-iteration communication cost of TSRGA is $O(n + d_n)$, Theorem 3.3.2, together with Lemma 3.3.1, directly imples the communication complexity of TSRGA, which we state as the following corollary.

Corollary 3.3.1. *If* $\kappa_n = O_p(\mathfrak{s}_n)$ *for some sequence* $\{\mathfrak{s}_n\}$ *of positive numbers, then TSRGA achieves an error of order*

$$O_p\left(\frac{\mathfrak{s}_n \xi_n^2}{n^2 d_n} \log \frac{n^2 d_n}{\xi_n^2} + \frac{\xi_n^2}{n^2 \delta_n^2} \mathbf{1}\{J_o \neq \emptyset\} \right),$$

with a communication complexity of order

$$O_p\left((n+d_n)\mathfrak{s}_n \log \frac{n^2 d_n}{\xi_n^2}\right).$$

Thus, the communication complexity, up to a logarithmic factor, scales mainly with $\mathfrak{s}_n$. In general, Lemma 3.3.1 implies $\mathfrak{s}_n = O_p(s_n^4)$. Thus $\mathfrak{s}_n$ is also a measure of the sparsity of the underlying model. Moreover, in the important special case when the response is a scalar, $\mathfrak{s}_n = O_p(s_n^2)$ since $d_n = 1$ and $\hat{J}_o = \emptyset$. To demonstrate this result more concretely, we discuss the communication complexity of TSRGA when applied to several well-known models below.

Example 3.3.1 (High-dimensional sparse linear regression)**.** *Consider the model* $y_t = \sum_{j=1}^{p_n} \beta_j x_{t,j} + \epsilon_t$. *Under suitable conditions, such as* $\{\epsilon_t\}$ *being i.i.d. sub-Gaussian random variables, it can be shown that* $\xi_E = O_p(\sqrt{n \log p_n})$ *(see, for example, Ing and Lai, 2011 and Ing, 2020). Then TSRGA achieves an error of order*

$$\sum_{j=1}^{p_n} |\beta_j - \hat{\beta}_j|^2 = O_p\left(\frac{s_n^2 \log p_n}{n}\right) \tag{3.13}$$

with a communication complexity of

$$O_p\left(n s_n^2 \log \frac{n}{\log p_n}\right).$$

To reach ϵ-close to the minimizer of the Lasso problem, the communication complexity of the Hydra algorithm (Richtárik and Takáč, 2016) is

$$O\left(\frac{n p_n}{M\tau} \log \frac{1}{\epsilon}\right),$$

where M is the number of nodes and τ is the number of coordinates to update in each

iteration. Given limited computational resources, τM may still be of order smaller than p_n. Thus the communication complexity of TSRGA, which does not scale with p_n, is more favorable for large data sets with huge p_n. In our simulation studies, we also observe that TSRGA converges near $(\beta_1, \ldots, \beta_{p_n})$ much more faster than Hydra-type algorithms.

Example 3.3.2 (Multi-task linear regression with common relevant predictors)**.** *Suppose we are interested in modeling T tasks simultaneously. Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$ be the vectors of n observations of the T responses, and $\mathbf{X}$ be the $n \times p$ design matrix consisting of p predictors. Consider the system of linear regressions*

$$\mathbf{y}_t = \mathbf{X}\mathbf{b}_t + \mathbf{e}_t, \quad t = 1, \ldots, T, \tag{3.14}$$

where $\mathbf{b}_i = (\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,p})^T$, for $i = 1, 2, \ldots, T$, and $\mathbf{e}_i$, for $1 \le i \le T$, are independent standard Gaussian random vectors. Let $\mathbf{x}_j$ be the j-th column vector of $\mathbf{X}$. Then we may rearrange (3.14) *as*

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \sum_{j=1}^{p} \mathbf{X}_j \mathbf{B}_j + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_T \end{pmatrix}, \tag{3.15}$$

where $\mathbf{B}_j = (\beta_{1,j}, \beta_{2,j}, \ldots, \beta_{T,j})^T$ and $\mathbf{X}_j = \mathbf{I}_T \otimes \mathbf{x}_j$, where $\mathbf{I}_T$ is the $T \times T$ identity matrix and $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$. Now (3.15) *falls under our general model* (3.1)*. Sparsity of the $\mathbf{B}_j$'s promotes that each task is driven by the same small set of predictors, or equivalently, $\mathbf{b}_j$'s in* (3.14) *have a common support. By a similar argument used in Lemma 3.1 of Lounici et al. (2011), it can be shown that $\xi_E = O_p(\sqrt{nT(1 + T^{-1} \log p)})$. Hence Corollary 3.3.1 implies that TSRGA applied to* (3.15)

achieves an error of order

$$\sum_{j=1}^{p} \|\mathbf{B}_j - \hat{\mathbf{B}}_j\|^2 = O_p\left(\frac{s_n^2}{nT}(1 + \frac{\log p}{T})\right) \tag{3.16}$$

with the communication complexity

$$O_p\left(nTs_n^2 \log \frac{nT}{1 + T^{-1}\log p}\right).$$

Notice again that the iteration complexity scales primarily with the strong sparsity parameter s_n, not with p. As illustrated by Lounici et al. (2011), (3.14) can be motivated from a variety of applications, such as the seemingly unrelated regressions (SUR) in econometrics and the conjoint analysis in marketing research.

Example 3.3.3 (Integrative multi-view regression)**.** *Consider the general model* (3.1)*, which is called the integrative multi-view regression by Li et al. (2019). Assume* $\mathbf{E}$ *has i.i.d. Gaussian entries, and for simplicity that* $q_{n,1} = q_{n,2} = \ldots = q_{n,p_n} = q_n$*. Then by a similar argument used by Li et al. (2019) it follows that* $\xi_E = O_p(\sqrt{n \log p_n}(\sqrt{d_n} + \sqrt{q_n}))$*. Suppose the predictors* $\mathbf{X}_j$*, for* $j = 1, 2, \ldots, p_n$*, are distributed across computing nodes. TSRGA achieves*

$$\frac{1}{d_n}\sum_{j=1}^{p_n} \|\mathbf{B}_j^* - \hat{\mathbf{B}}_j\|_F^2 = O_p\left(\frac{s_n^4(d_n + q_n)\log p_n}{nd_n} + \frac{(d_n + q_n)\log p_n}{n\delta_n}\right) \tag{3.17}$$

with a communication complexity of

$$O_p\left((n + d_n)s_n^4 \log \frac{nd_n}{(d_n + q_n)\log p_n}\right).$$

Although Li et al. (2019) did not consider the feature-distributed data, they offer an ADMM-based algorithm, iRRR, for estimating (3.1). However, updating many parameters

in each iteration causes significant computational bottleneck. In our Monte Carlo simulation, iRRR is unable to run efficiently with $p_n \geq 50$ even with centralized computing and a moderate sample size, whereas TSRGA can handle such data sizes easily.

In general, the statistical errors of TSRGA in the above examples ((3.13), (3.16), and (3.17)) are sub-optimal compared to the minimax rates unless $s_n = O(1)$, in which case the model is strongly sparse with a fixed number of relevant predictors. One reason is that Theorem 3.3.1 only guarantees sure-screening instead of predictor and rank selection consistency. In Examples 1 and 2, the statistical error could be improved if one applies hard-thresholding after the second-stage RGA, and then estimates the coefficients associated with the survived predictors again. This would not hurt the communication complexity in terms of the order of magnitude since this step takes even less number of iterations. Nevertheless, in our simulation studies, TSRGA performs on par with and in many cases even outperforms strong benchmarks in the finite-sample case.

Another reason for the sub-optimality comes from the dependence on δ_n in the error. In the second-stage, TSRGA relies on the sample SVD of the (scaled) marginal covariance $\mathbf{X}_j^\top \mathbf{Y}$ to estimate the singular subspaces of the unknown coefficient matrices. How well these sample singular vectors recover their noiseless counterparts depends on the strength of the marginal covariance, which is controlled by δ_n in Assumption (C5). This is needed because we try to avoid searching for the singular subspaces of the coefficient matrices, a challenging task for greedy algorithms. Unlike the scalar case, for the multivariate linear regression the dictionary for RGA contains all rank-one matrices and therefore the geometric structure is more intricate to exploit. For example, the argument used in Ing (2020) will not work with this dictionary.

Recently, Ding et al. (2020) and Ding et al. (2021) proposed new modifications of the Frank-Wolfe algorithm that directly search within the nuclear norm ball, under the assumptions of strict complementarity and quadratic growth. These algorithms rely on solving more

complicated sub-problems. To illustrate one main difference between these modifications and TSRGA, note that for the usual reduced rank regression where $\min\{d_n, q_{n,1}\} > 1$ and $p_n = 1$, one of the leading examples in Ding et al. (2020) and Ding et al. (2021), our theoretical results for TSRGA still hold (though in this case the data are not feature-distributed because p_n is only one). In this case, (C5) and (C6) automatically hold with $\delta_n \leq d_n^{1/2}/(\mu s_n^{1/2})$. Consequently, Corollary 3.3.1 implies the error is of order $O_p(\frac{s_n^2\xi_n^2}{n^2 d_n}\log\frac{n^2 d_n}{\xi_n^2})$ using $O_p(s_n^2\log\frac{n^2 d_n}{\xi_n^2})$ iterations, regardless of whether strict complementarity holds. This advantage precisely comes from that TSRGA uses the singular vectors of $\mathbf{X}_1^\top\mathbf{Y}$ in its updates in the second stage instead of searching over the intricate space of nuclear norm ball in each iteration.

3.4 Simulation experiments

In this section, we apply TSRGA to synthetic data sets and compare its performance with some existing distributed as well as centralized methods. We first examine how well TSRGA and other algorithms estimate the unknown parameters. Then we apply TSRGA to a large-scale feature-distributed data to measure its prowess in speed. In both experiments, TSRGA delivered superior performance.

3.4.1 Statistical performance of TSRGA

In this subsection, we compare the effectiveness of TSRGA in estimating the parameters. Specifically, it is applied to the well-known high-dimensional linear regression and the general multi-view regression (3.2).

Consider first the high-dimensional linear regression model:

$$y_t = \sum_{j=1}^{p_n} \beta_j^* x_{t,j} + \epsilon_t, \quad t = 1, \ldots, n,$$

which is sparse with only $a_n = \lfloor p_n^{1/3} \rfloor$ non-zero β_j^*'s, where $\lfloor x \rfloor$ denotes the integer part of x.

We also generate $\{\epsilon_t\}$ as i.i.d. t-distributed random variables with five degrees of freedom.

To estimate this model, we employ the Hydra (Richtárik and Takáč, 2016) and Hydra^2 (Fercoq et al., 2014) algorithms to solve the Lasso problem, namely,

$$\min_{\{\beta_j\}_{j=1}^{p_n}} \left\{ \frac{1}{2n} \sum_{t=1}^{n} \left(y_t - \sum_{j=1}^{p_n} \beta_j x_{t,j} \right)^2 + \lambda \sum_{j=1}^{p_n} |\beta_j| \right\}. \tag{3.18}$$

The predictors are divided into 10 groups at random; each of the groups is owned by one node in the Hydra-type algorithm. The step size of the Hydra-type algorithms is set to the lowest value so that we observe convergence of the algorithms instead of divergence. As a benchmark, we also solve the Lasso problem with 5-fold cross validation using `glmnet` package in R. To further reduce the computational burden, we use the λ selected by the cross-validated Lasso in implementing Hydra-type algorithms.

Choosing the hyperparameter for RGA-type methods is more straightforward, but there is one subtlety. It is well-known that the Lasso problem corresponds to the constrained minimization problem

$$\min_{\{\beta_j\}_{j=1}^{p_n}} \frac{1}{2n} \sum_{t=1}^{n} \left(y_t - \sum_{j=1}^{p_n} \beta_j x_{t,j} \right)^2 \text{ subject to } \sum_{j=1}^{p_n} |\beta_j| \leq L_n.$$

Moreover, setting L_n to $\sum_{j=1}^{p_n} |\beta_j^*|$, which is nonetheless unknown in practice, would yield the usual Lasso statistical guarantee (see, e.g., Theorem 10.6.1 of Vershynin, 2018). However, our theoretical results in Section 3.3.2 recommend setting L_n to a larger value than this conventionally recommended value. To illustrate the advantage of a larger L_n, we employ two versions of RGA: one with $L_n = 500$ and the other with $L_n = \sum_{j=1}^{p_n} |\beta_j^*|$. For TSRGA, we simply set $L_n = 500$ and $t_n = 1/(10 \log n)$, and the performance is not too sensitive to these choices.

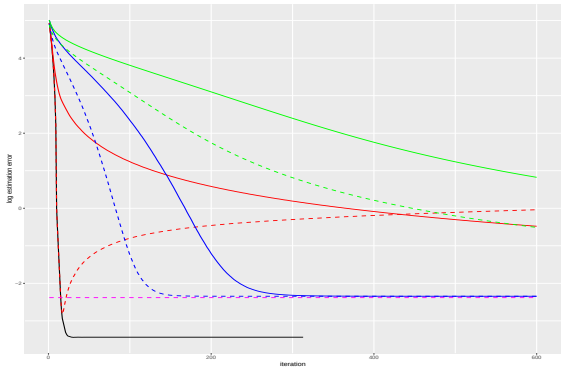
For Specifications 1 and 2 below, we consider three cases with $(n, p_n) \in \{(800, 1200),$

$(1200, 2000)$, $(1500, 3000)\}$. In Specification 1, we simulate the predictors as independent, t-distributed data. Together with the t-distributed errors, this specification simulates the situation where heavy-tailed data are frequently observed.

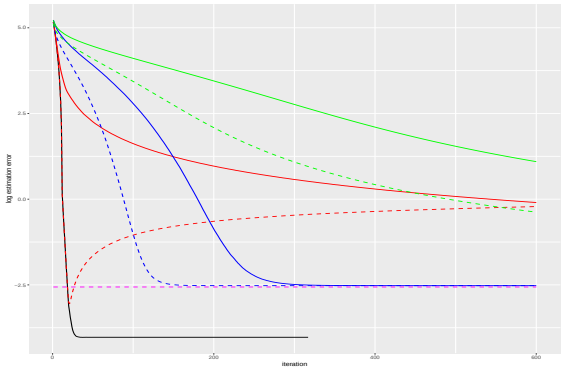
Specification 1. In the first experiment, we generate $x_{t,j}$ as i.i.d. $t(6)$ random variables, for all $t = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p_n$. Hence the predictors have heavy tails with only 6 finite moments. The nonzero coefficients are generated independently by $\beta_j^* = z_j u_j$, where z_j is uniform over $\{-1, +1\}$ and u_j is uniform over $[2.5, 5.5]$. The coefficients are drawn at the start of each of the 100 Monte Carlo simulations.

Figure 3.1 plots the logarithm of the parameter estimation error against the number of iterations. The parameter estimation error is defined as $\sum_{j=1}^{p_n}(\beta_j^* - \hat{\beta}_j)^2$, where $\{\hat{\beta}_j\}$ are the estimates made by the aforementioned methods. In the plot, the trajectories are averaged across 100 simulations. TSRGA (black) converges using the least number of iterations. Since the per-iteration communication costs of TSRGA and Hydra-type algorithms are similar ($O(n)$ bytes), this serves as a proxy for a smaller communication overhead of TSRGA. In addition, the parameter estimation error of TSRGA is also the smallest among the employed methods. RGA with $L_n = 500$ (dashed red) follows the same trajectories as TSRGA in the first few iterations, but without the two-step design, it suffers from over-fitting in later iterations and hence an increasing parameter estimation error. On the other hand, RGA with oracle $L_n = \sum_{j=1}^{p_n} |\beta_j^*|$ (solid red) converges much slower than TSRGA due to a sub-linear convergence rate. For Hydra (blue lines) and Hydra^2 (green lines) algorithms, we consider two implementations: updating 25% of the coordinates in each node (solid), and updating 50% of the coordinates in each node (dashed). Hydra converges to the centralized Lasso (dashed magenta) at a faster rate if 50% of the coordinates were updated in each iterations than the 25% counterparts. However, Hydra^2 converges much slower.

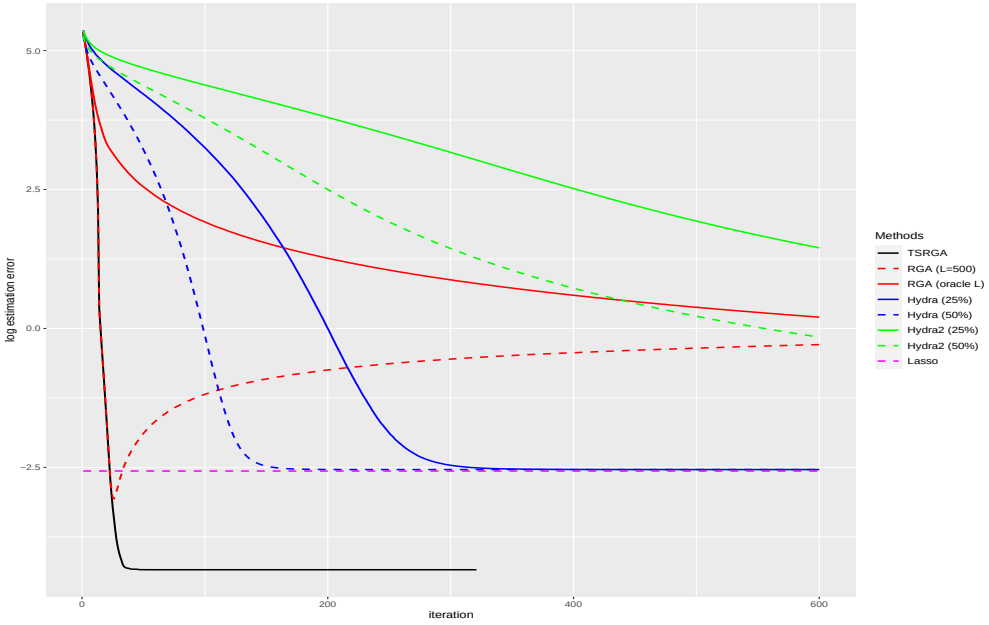
In the next specification, we generate the predictors so that they are correlated and the correlations are the same between any two predictors. This simulates the situation where



(a) $n = 800$, $p_n = 1200$



(b) $n = 1200$, $p_n = 2000$



(c) $n = 1500$, $p_n = 3000$

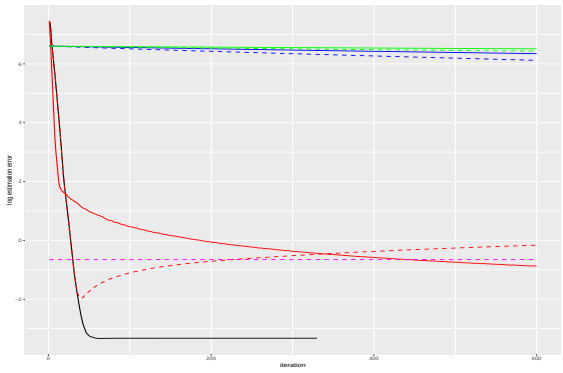
Figure 3.1: Logarithm of parameter estimation errors of various methods under Specification 1, where n is the sample size and p_n is the dimension of predictors. The results are averages of 100 simulations.

one cannot simply divide groups of variables that have weak inter-group dependence into different computing nodes to alleviate the difficulties caused by feature-distributed data.

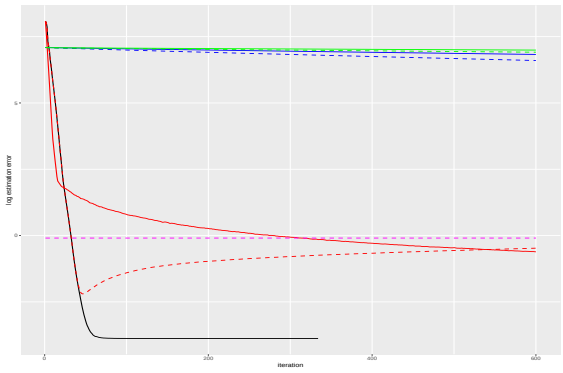
Specification 2. In this experiment, we generate the predictors by

$$x_{t,j} = \nu_t + w_{t,j}, \quad t = 1, \ldots, n; \quad j = 1, \ldots, p_n,$$

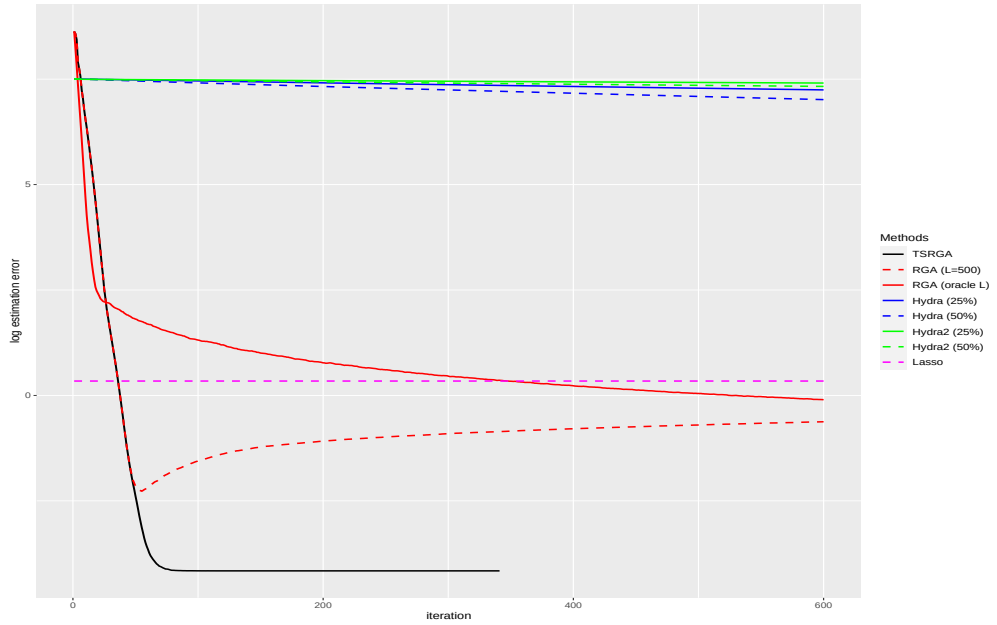
where $\{\nu_t\}$ and $\{w_{t,j}\}$ are independent $N(0,1)$ random variables. Consequently, $\text{Cor}(x_{t,k}, x_{t,j}) = 0.5$, for $k \neq j$. The coefficients are set to $\beta_j^* = 2.5 + 1.2(j-1)$ for $j = 1, 2, \ldots, a_n = \lfloor p_n^{1/3} \rfloor$. The rest of the specification is the same as that of Specification 1.



(a) $n = 800$, $p_n = 1200$



(b) $n = 1200$, $p_n = 2000$



(c) $n = 1500$, $p_n = 3000$

Figure 3.2: Parameter estimation errors of various estimation methods under Specification 2, where n is the sample size and p_n is the number of predictors. The results are averages of 100 simulations.

Figure 3.2 plots the parameter estimation errors under Specification 2. TSRGA remains the most effective method for estimating the unknown parameters, which converges

within 100 iterations in all cases. It is worth noting that the Hydra-type algorithms display a substantially deteriorated rate of convergence compared to the previous specification, highlighting their sensitivity to the dependence between predictors, and potentially high computational expenses in certain scenarios.

It is also important to study the performance of these methods in terms of elapsed time and out-of-sample performance. To save space, we postpone the discussion to Section 3.8.5, as most conclusions drawn above remain valid in examining the elapsed time and prediction performance.

Next we consider the general model:

$$\mathbf{y}_t = \sum_{j=1}^{p_n} \mathbf{B}_j^{*\top} \mathbf{x}_{t,j} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, n, \tag{3.19}$$

where $\mathbf{y}_t \in \mathbb{R}^{d_n}$ and $\mathbf{x}_{t,j} \in \mathbb{R}^{q_n}$, for $j = 1, 2, \ldots, p_n$. We generate $\boldsymbol{\epsilon}_t$ as i.i.d. random vectors with each entry having independent $t(5)$ distributions. In the following cases, the model is sparse with a_n non-zero $\mathbf{B}_j^*$'s, each of which is only of rank r_n. In particular, we generate $\mathbf{B}_j^*$, for $j \leq a_n$, independently by

$$\mathbf{B}_j^* = \sum_{k=1}^{r_n} \sigma_{k,n} \mathbf{u}_{k,j} \mathbf{v}_{k,j}^\top, \tag{3.20}$$

where $\{\mathbf{u}_{k,j}\}_{k=1}^{r_n}$ and $\{\mathbf{v}_{k,j}\}_{k=1}^{r_n}$ are independently drawn (q_n- and d_n-dimensional) orthonormal vectors and $\{\sigma_{k,n}\}$ are i.i.d. uniform over [7,15].

We employ the iRRR method (Li et al., 2019) to estimate (3.19). To select its tuning parameter, we execute iRRR with a grid of tuning parameter values and opt for the one with the lowest mean square prediction error on an independently generated validation set of 500 observations. Although centralized computation is used to implement iRRR, it is too computationally demanding to implement the algorithm for the two cases with $n = 600$ and

$n = 1200$. Therefore, we use the least squares estimator with only the relevant variables as another benchmark. For TSRGA, L_n is set to 10^5, and we hold one third of the training data as validation set to select the tuning parameter t_n for TSRGA over a grid of values[1].

Since iRRR is not a feature-distributed algorithm, we directly report their parameter estimation errors (averaged across 500 Monte Carlo simulations) defined as

$$\sqrt{\sum_{j=1}^{p_n} \|\mathbf{B}_j^* - \hat{\mathbf{B}}_j\|_F^2}, \tag{3.21}$$

where $\{\hat{\mathbf{B}}_j\}$ are the estimated coefficient matrices. Additionally, the out-of-sample prediction performance of these methods are evaluated on an independent test sample of size 500, measured by $(\|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2/(nd_n))^{1/2}$. We consider the cases $(n, d_n, q_n, p_n, a_n, r_n) \in \{(200, 10, 12, 20, 1, 2), (400, 15, 18, 50, 2, 2), (600, 20, 25, 400, 3, 2), (1200, 40, 45, 800, 3, 3)\}$.

Specification 3. In this specification, we consider (3.19) with the predictors generated as in Specification 1. Note that $\{\mathbf{B}_j^* : j \leq a_n\}$ are drawn at the start of each of the 500 Monte Carlo simulations.

Table 3.1 reports the results of the methods averaged over 500 Monte Carlo simulations of data generated under Specification 3. TSRGA achieved the lowest estimation error in all constellations of problem sizes. On the other hand, iRRR yielded larger estimation error than the least squares method using exactly the relevant predictors when $n = 200$, but when n increases, iRRR outperforms the least squares method. However, the computational costs of iRRR became so high that completing 500 simulations would require more than days, even when parallelism with 15 cores is used. TSRGA circumvents such computational overhead and delivers superior estimates. The prediction errors suggest the same conclusions even though the difference is less significant.

1. t_n is selected among $\mathbf{t} = (0.01, 0.07, 1.10, 1.39, 1.61, 1.79, 1.95, 2.08, 2.20, 2.30)/\log n$.

	Parameter estimation			Prediction		
$(n, d_n, q_n, p_n, a_n, r_n)$	TSRGA	iRRR	Oracle LS	TSRGA	iRRR	Oracle LS
(200, 10, 12, 20, 1, 2)	0.666	0.929	0.851	1.138	1.339	1.331
(400, 15, 18, 50, 2, 2)	0.858	1.245	1.287	1.322	1.351	1.355
(600, 20, 25, 400, 3, 2)	1.223	-	1.787	1.361	-	1.381
(1200, 40, 45, 800, 3, 3)	1.388	-	2.378	1.345	-	1.371

Table 3.1: Parameter estimation and prediction errors of various methods under Specification 3. We do not report the results for iRRR with sample sizes of 600 and 1200 since the computation required for these cases is excessively time-consuming. In the table, n, d_n, q_n, p_n, a_n and r_n are the sample size, number of targeted variables, dimension of predictors, number of predictors, number of non-zero coefficient matrices, and rank of coefficient matrices, respectively. The results are based on 500 simulations.

Specification 4. In this specification, we generalize (3.19) to group predictors as follows. Let $\{\boldsymbol{\nu}_t : t = 1, 2, \ldots\}$ and $\{\mathbf{w}_{t,j} : t = 1, 2, \ldots; j = 1, 2, \ldots, p_n\}$ be independent $N(\mathbf{0}, \mathbf{I}_{q_n})$ random vectors. The group predictors are then constructed as $\mathbf{x}_{t,j} = 2\boldsymbol{\nu}_t + \mathbf{w}_{t,j}$, $1 \le t \le n$, $1 \le j \le p_n$. Hence $\mathbb{E}(\mathbf{x}_{t,j}\mathbf{x}_{t,i}^\top) = 4\mathbf{I}_{q_n}$, for $1 \le i < j \le p_n$. Note that $\mathrm{Corr}(x_{t,i,l}, x_{t,j,l}) = 0.8$ for $i \ne j$, $1 \le l \le q_n$, where $\mathbf{x}_{t,i} = (x_{t,i,1}, \ldots, x_{t,i,q_n})^\top$. Hence, the l-th components in each of the group predictors are highly correlated.

	Parameter estimation			Prediction		
$(n, d_n, q_n, p_n, a_n, r_n)$	TSRGA	iRRR	Oracle LS	TSRGA	iRRR	Oracle LS
(200, 10, 12, 20, 1, 2)	0.401	0.616	0.460	1.324	1.337	1.330
(400, 15, 18, 50, 2, 2)	0.562	0.993	1.172	1.345	1.344	1.354
(600, 20, 25, 400, 3, 2)	0.812	-	1.817	1.362	-	1.379
(1200, 40, 45, 800, 3, 3)	0.751	-	2.419	1.310	-	1.371

Table 3.2: Parameter estimation and prediction errors under Specification 4. We do not report the results for iRRR with sample sizes of 600 and 1200 since the computation required for these sample sizes is excessively time-consuming. The same notations as those of Table 3.1 are used. The results are based on 500 simulations.

Table 3.2 reports the results for Specification 4. As in the previous specifications, TSRGA continues to surpass the benchmarks. When $n = 400$, iRRR gains an advantage over the least squares method, despite of a high computational cost. The results in Tables 3.1 and 3.2

suggest that TSRGA is both a fast and a statistically effective tool for parameter estimation for model (3.19).

3.4.2 Large-scale performance of TSRGA

In this subsection, we apply TSRGA to large feature-distributed data. We have an MPI implementation of TSRGA through OpenMPI and the `Python` binding `mpi4py` (Dalcín et al., 2005; Dalcín and Fang, 2021). The algorithm runs on the high-performance computing cluster of the university, which comprises multiple computing nodes equipped with Intel Xeon Gold 6248R processors. We consider again Specification 4 in the previous subsection, with $(n, d_n, q_n, p_n, a_n, r_n) = (20000, 100, 100, 1024, 4, 4)$. In the following experiments we employ $M/4$ nodes, each of which runs 4 processes and each process owns p_n/M predictors, with M varying from 16 to 64. When combined, the data are approximately over 16 GB of size, exceeding the usual RAM capacity on most laptops.

There are two primary goals for the experiments. The first goal is to investigate the wall-clock time required by TSRGA to estimate (3.19). The second goal is to examine the effect of the number of nodes on the required wall-clock time. Each experiment is repeated 10 times, and we average the wall-clock time needed to complete the k-th iteration as well as the parameter estimation error (3.21) at the k-th iteration.

Figure 3.3 plots the (log) estimation errors against the wall-clock time of TSRGA iterations. When using 16 processes, TSRGA took about 16 minutes to estimate (3.19), and the time reduced to less than 5 minutes when 64 processes were employed. The acceleration primarily occurred in the first stage, because solving (3.3) becomes faster when each process handles only a small number of predictors. After screening, there is a spike in estimation error due to re-initialization of the estimators but subsequent second-stage RGA runs extremely fast in all cases and yields accurate estimates. Indeed, Figure 3.4 shows that the estimation error of TSRGA quickly drops below that of the oracle least squares

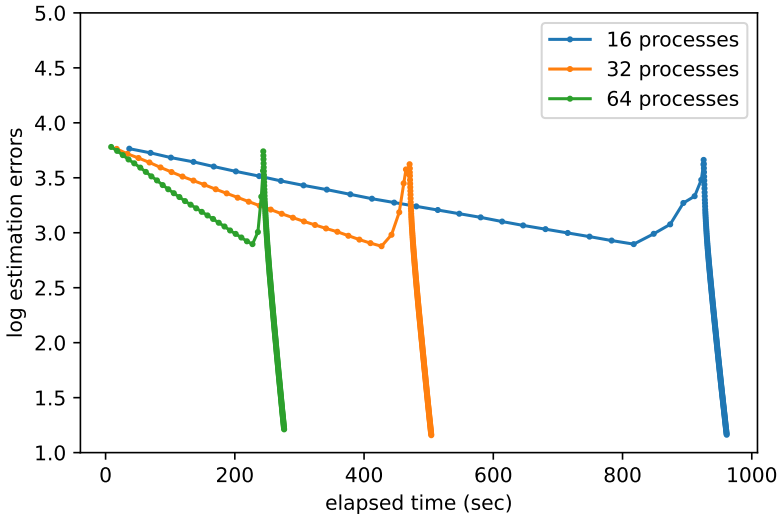


Figure 3.3: Logarithm of the average parameter estimation errors at each iteration of TSRGA, plotted against the average time elapsed at the end of each iteration. Various number of processes are employed for feature-distributed implementation. 10 simulations are used.

in the second stage. We remark that with more diligent programming, one can apply the advanced protocols introduced in Section 6 of Richtárik and Takáč (2016) to TSRGA, using both multi-process and multi-thread techniques. It is anticipated that the required time will be further shortened.

3.5 Empirical application

This section showcases an application of TSRGA to financial data. In addition to the conventional financial data, we further collect the annual 10-K reports of firms under study to extract useful features for augmenting the predictors. Thus, in this application, both the response and predictors are multivariate, and the predictors may consist of large dense matrices, leading to potential computational challenges in practice.

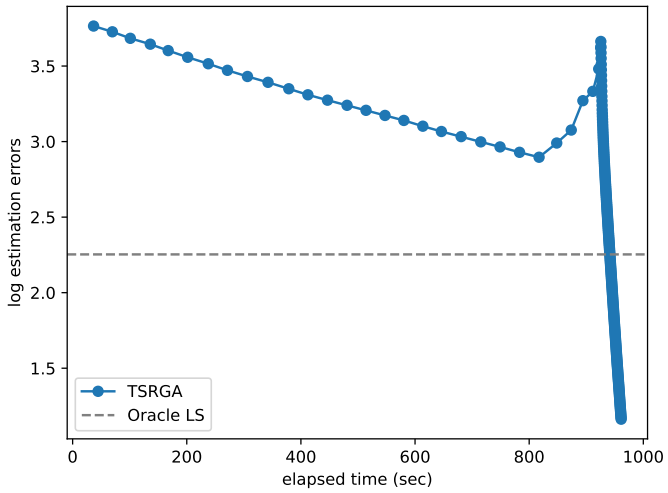


Figure 3.4: Logarithm of the estimation errors of TSRGA (running with 16 processes) and the oracle least squares. The oracle least squares method is performed by applying the second-stage RGA with exactly the relevant predictors and no rank constraints. 10 simulations are used.

3.5.1 Financial data and 10-K reports

We aim to predict four key financial outcomes for companies in the S&P 500 index: volatility, trading volume, market beta, and return. We obtain daily return series for each company from 2010 through 2019, calculate the sample variances of the daily returns in each month, and transform them by taking the logarithm to get the volatility series $\{V_{it}(m) : m = 1, 2, \ldots, 12\}$ for the i-th company in the m-th month of year $t \in \{2010, \ldots, 2019\}$. Next, we regress each company's daily returns on the daily returns of the S&P 500 index for each month and use the slope estimates as market beta, $\{B_{it}(m) : m = 1, 2, \ldots, 12\}$. Finally, we also obtain data of the monthly returns series $\{R_{it}(m) : m = 1, 2, \ldots, 12\}$ and the logarithm of the trading volumes $\{M_{it}(m) : m = 1, 2, \ldots, 12\}$, for the ith company. All series are obtained from Yahoo! Finance via the `tidyquant` package in R.

After obtaining these series, some data cleaning is performed to facilitate subsequent analysis. First, the volume series exhibits a high degree of serial dependence, which could be due to unit-roots caused by the persistence in trading activities. Therefore, we apply a

year-to-year difference, i.e., $\Delta M_{it}(m) = M_{i,t}(m) - M_{i,t-1}(m)$ for all i, $1 \leq m \leq 12$, and $t = 2011, \ldots, 2019$. Additionally, we remove companies that have outlying values in these series.

In addition to these financial time series, we also make use of the information from a pertinent collection of textual data: 10-K reports. Publicly traded companies in the U.S. are required to file these annual reports with the aim of increasing transparency and satisfying the regulation of exchanges. The reports are maintained by the Securities and Exchange Commission (SEC) in the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR), and provide information about a company's risks, liabilities, and corporate agreements and operations. Due to their significance in communicating information to the public, the 10-K reports are an important corpus in finance, economics, and computational social sciences studies (Hanley and Hoberg, 2019; Kogan et al., 2009; Gandhi et al., 2019; Jegadeesh and Wu, 2013).

The corpus utilized in this application is sourced from the EDGAR-CORPUS, originally prepared by Loukas et al. (2021). Our analysis specifically focuses on Section 7, titled "Management's Discussion and Analysis." To process the reports, we preprocess each document using the default functionality in the `gensim` package in Python and discard the documents that consist of fewer than 50 tokens. As a result, we have data of both the financial time series and 10-K reports of 256 companies over the period from 2011 through 2019.

To extract features from the textual data, we employ a technique called Latent Semantic Indexing (LSI, see, e.g., Deerwester et al., 1990). We first construct the term-document matrix as follows. Suppose we have D documents in the training set, and there are V distinct tokens in these documents. The term-document matrix $\boldsymbol{\Theta}$ is a $V \times D$ matrix, whose

entries are given by

$$\begin{aligned}\boldsymbol{\Theta}_{ij} =&(\text{number of times the } i\text{-th token appears in document } j)\times \\ &\log \frac{D}{\sharp\{1 \le k \le D : \text{the } i\text{-th token appears in document } k\}},\end{aligned}$$

for $1 \le i \le V$, $1 \le j \le D$. The entries are known as one form of the term-frequency inverse document frequency (TFIDF, see, e.g., Salton and Buckley, 1988). Then, to extract K features from the text data, LSI uses the singular value decomposition,

$$\boldsymbol{\Theta} = \mathbf{U}_{\boldsymbol{\Theta}} \boldsymbol{\Sigma}_{\boldsymbol{\Theta}} \mathbf{V}_{\boldsymbol{\Theta}}^{\top},$$

and the first K rows of $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} \mathbf{V}_{\boldsymbol{\Theta}}^{\top}$ are used as the features in the training set. For a new document in the test set, we compute its TFIDF representation $\boldsymbol{\theta} \in \mathbb{R}^V$, and then use $\mathbf{x} = \mathbf{U}_K^{\top} \boldsymbol{\theta}$ as its textual features, where $\mathbf{U}_K$ is the sub-matrix of the first K columns of $\mathbf{U}_{\boldsymbol{\Theta}}$.

3.5.2 Results

For each of the four financial response variables, we estimate the following model.

$$\mathbf{y}_{it} = \boldsymbol{\beta}_0 + \mathbf{A}_1^{\top} \mathbf{v}_{i,t-1} + \mathbf{A}_2^{\top} \mathbf{m}_{i,t-1} + \mathbf{A}_3^{\top} \mathbf{b}_{i,t-1} + \mathbf{A}_4^{\top} \mathbf{r}_{i,t-1} + \mathbf{A}_5^{\top} \mathbf{x}_{i,t-1} + \boldsymbol{\epsilon}_{it}, \qquad (3.22)$$

where $\mathbf{y}_{it} = (y_{it}(1), \ldots, y_{it}(12))^{\top}$ is the response variable under study, $\mathbf{v}_{it} = (V_{it}(1), \ldots, V_{it}(12))^{\top}$, $\mathbf{m}_{it} = (\Delta M_{it}(1), \ldots, \Delta M_{it}(12))^{\top}$, $\mathbf{b}_{it} = (B_{it}(1), \ldots, B_{it}(12))^{\top}$, $\mathbf{r}_{it} = (R_{it}(1), \ldots, R_{it}(12))^{\top}$, $\mathbf{x}_{it} \in \mathbb{R}^K$ is the extracted text features, and $\{\boldsymbol{\beta}_0, \mathbf{A}_1, \ldots, \mathbf{A}_5\}$ are unknown parameters. When predicting each of the four financial outcomes, we replace $\mathbf{y}_{it}$ in (3.22) with the corresponding vector ($\mathbf{v}_{it}$, $\mathbf{m}_{it}$, $\mathbf{b}_{it}$, or $\mathbf{r}_{it}$), while keeping the same model structure. Since predicting next-year's financial outcomes in one month is related to predicting the same variable in other months, it is natural to expect low-rank coefficient matrices.

(3.22) can also be viewed as a multi-step ahead prediction model, since we are predicting the next twelve months simultaneously.

When applying TSRGA to (3.22), we use a hold-out validation set from the training sample to select the just-in-time threshold t_n from the grid $(0.1, 0.2, \ldots, 1.0)/\log n$. In addition to TSRGA, we employ several benchmark prediction methods, including the vector autoregression (VAR), reduced rank regression (RR; see, e.g., Chen et al., 2013), the integrative reduced rank regression (iRRR, Li et al., 2019), and the Lasso. For VAR, we concatenate all response variables and estimate the model

$$\mathbf{z}_{it} = \mathbf{A}^\top \mathbf{z}_{i,t-1} + \mathbf{e}_{it},$$

where $\mathbf{z}_{it} = (\mathbf{v}_{it}^\top, \mathbf{m}_{it}^\top, \mathbf{b}_{it}^\top, \mathbf{r}_{it}^\top)^\top \in \mathbb{R}^{48}$. Alternatively, we can implement VAR in a group-wise fashion (gVAR henceforth). Specifically, we separately estimate the model

$$\mathbf{y}_{it} = \mathbf{A}^\top \mathbf{y}_{i,t-1} + \mathbf{e}_{it}, \tag{3.23}$$

for each response variable $\mathbf{y}_{it} \in \{\mathbf{v}_{it}, \mathbf{m}_{it}, \mathbf{b}_{it}, \mathbf{r}_{it}\}$. The reduced rank regression also estimates (3.23) with an intercept term and an additional rank constraint on the coefficient matrix $\mathbf{A}$ in (3.23). We use the generalized cross validation (GCV, Golub et al., 1979) to select the optimal rank. For Lasso, it is applied separately to each row of (3.22); namely, it estimates

$$\begin{aligned} y_{it}(m) = \beta_0 + \sum_{j=1}^{12} \alpha_{j,1} V_{i,t-1}(j) + \sum_{j=1}^{12} \alpha_{j,2} \Delta M_{i,t-1}(j) \\ + \sum_{j=1}^{12} \alpha_{j,3} B_{i,t-1}(j) + \sum_{j=1}^{12} \alpha_{j,4} R_{i,t-1}(j) + \epsilon_{it}, \end{aligned}$$

for $m = 1, 2, \ldots, 12$. Finally, we also apply the iRRR method of Li et al. (2019) to (3.22).

Table 3.3 presents the root mean squared prediction errors (RMSE) for different methods on the test set, for which we reserved the last year of data. The results show that gVAR consistently outperformed the usual VAR in all four financial variables, suggesting using simple least squares could be harmful in prediction when including many financial series as predictors. RR provides a slight improvement in predicting volatility, but performs similarly as VAR and gVAR in predicting other targets. In the case of predicting volatility, the text data proved to be quite useful, and TSRGA, iRRR, and Lasso have all outperformed gVAR by more than 5% with different number of textual features K (except for Lasso with $K = 50$). TSRGA and iRRR, utilizing both the text information and low-rank coefficient estimates, yielded the smallest prediction errors. In some cases, they achieved 10% reduction in RMSE compared with gVAR and RR. For the rest of the targets, the methods did not perform very differently from gVAR and RR.

In addition to the prediction performance, we make two more remarks on the empirical results. First, our finding that textual features are useful in predicting volatility is consistent with previous studies. For instance, Kogan et al. (2009) reported that one-hot text features are already effective in predicting volatility in a scalar linear regression, and Yeh et al. (2020) also observed gains of using neural word embedding to predict volatility. Our results suggest an alternative modeling choice: text data could explain each month's volatility via a low-rank channel. Second, low-rank models may not be suited for the trading volume series. The RR selected a full-rank model and TSRGA iterated more steps before the just-in-time stopping criterion was triggered.

The data set used in the application is relatively small, and can fit in most personal computer's memory. However, incorporating more sections of the 10-K reports or other financial corpus may pose computational challenges due to the increased number of dense text feature matrices. TSRGA can easily handle such cases when feature-distributed data are inevitable.

	Volatility	Volume	Beta	Return
VAR	0.782	0.323	0.583	0.077
gVAR	0.750	0.319	0.556	0.073
RR	0.732	0.325	0.555	0.071
$K = 50$				
Lasso	0.718	0.310	0.574	0.075
iRRR	**0.688**	0.318	0.568	0.072
TSRGA	**0.702**	0.345	0.572	0.072
$K = 100$				
Lasso	**0.700**	0.308	0.574	0.074
iRRR	**0.677**	0.316	0.566	0.072
TSRGA	**0.678**	0.330	0.571	0.072
$K = 150$				
Lasso	**0.693**	0.308	0.571	0.073
iRRR	**0.667**^{a}	0.314	0.566	0.072
TSRGA	**0.681**	0.332	0.573	0.072
$K = 200$				
Lasso	**0.684**	0.309	0.574	0.073
iRRR	**0.663**^{a}	0.314	0.567	0.072
TSRGA	**0.654**^{a,b}	0.345	0.574	0.072

Table 3.3: Root mean squared prediction errors on the test dataset. Entries in boldface are at least 5% below gVAR; ^{a} means 10% below gVAR, and ^{b} means 10% below RR.

3.6 Horizontal partition for big feature-distributed data

In this section, we briefly discuss the usage of TSRGA when the sample size n, in addition to the dimension p_n, is also large so that storing $(\mathbf{Y}, \mathbf{X}_j)$ in one machine is infeasible. In this case, we also horizontally partition the (feature-distributed) data matrices and employ more computing nodes.

To fix ideas, for $h = 1, 2, \ldots, H$, let

$$\mathbf{Y}_{(h)} = (\mathbf{y}_{m_{h-1}+1}, \ldots, \mathbf{y}_{m_h})^\top, \text{ and } \quad \mathbf{X}_{j,(h)} = (\mathbf{x}_{m_{h-1}+1,j}, \ldots, \mathbf{x}_{m_h,j})^\top$$

be horizontal partitions of $\mathbf{Y}$ and $\mathbf{X}_j$, $j = 1, \ldots, p_n$, where $0 = m_0 < m_1 < \ldots < m_H = n$. In the distributed computing system, we label the nodes by (h, c), so that the (h, c)-th node

owns data $\mathbf{Y}_{(h)}$ and $\{\mathbf{X}_{j,(h)} : j \in \mathcal{I}_c\}$, where $h \in [H]$, $c \in [M]$ and $\cup_{c\in[M]}\mathcal{I}_c = [p_n]$. For ease in illustration, we further assume $\{\mathcal{I}_c : c \in [M]\}$ forms a partition of $[p_n]$. Therefore, each computing node only owns a slice of the samples on a subset $\mathcal{I}_c$ of the predictors as well as the same slice of the response variables. Moreover, let $I(j) = \{(h, c) : j \in \mathcal{I}_c\}$ be the indices of the nodes that have some observations of predictor j.

We call the nodes that own the h-th slice of data "segment h". That is, $\{(k, c) : k = h\}$. Note that each segment is essentially the feature-distributed framework discussed in the previous sections. In what follows, quantities computed at nodes in segment h carry a subscript (h). For example, $\hat{\mathbf{\Sigma}}_{j,(h)} = n_h^{-1}\mathbf{X}_{j,(h)}^\top\mathbf{X}_{j,(h)}$, where $n_h = m_h - m_{h-1}$. For simplicity, we also assume $n_1 = \ldots = n_H$ in this section. Finally, we again assume there is at least one master node to coordinate all the computing nodes $\{(h, c) : h \in [H], c \in [M]\}$.

To estimate (3.2) with the horizontally partitioned feature-distributed data described above, we suggest the following procedure. First, we obtain a set of potentially relevant predictors $\hat{J}$ and their respective upper bounds on the coefficient ranks $\hat{r}_j$ by running the first-stage RGA with the just-in-time stopping criterion. This can be done by applying Algorithm 1 to one segment. Alternatively, one can apply it to multiple segments in parallel and set $\hat{J} = \cap_h \hat{J}_{(h)}$ and $\hat{r}_j = \min_h \hat{r}_{j,(h)}$. In either case, Theorem 3.3.1 ensures the sure-screening property as $n_1 \to \infty$ if (C1)-(C4) hold in each of the segments. By Lemma 3.3.1, this step costs $O_p(s_n^2(n_1 + d_n))$ bytes of communication per node in the segment(s) involved.

Next, for each $j \in \hat{J}$, each node $(h, c) \in I(j)$ computes $\mathbf{X}_{j,(h)}^\top\mathbf{X}_{j,(h)}$ and, if $q_{n,j} \wedge d_n > \hat{r} = \sum_j \hat{r}_j$, additionally computes $\mathbf{X}_{j,(h)}^\top\mathbf{Y}_{(h)}$. Then, send these matrices to the master node. The master node computes $\hat{\mathbf{\Sigma}}_j^{-1} = (\sum_{h=1}^H \mathbf{X}_{j,(h)}^\top\mathbf{X}_{j,(h)})^{-1}$ and the leading $\hat{r}$ singular vectors of $\sum_{h=1}^H \mathbf{X}_{j,(h)}^\top\mathbf{Y}_{(h)}$, which form the column vectors of $\mathbf{U}_j$ and $\mathbf{V}_j$. Then $(\hat{\mathbf{\Sigma}}_j^{-1}, \mathbf{U}_j, \mathbf{V}_j)$ (or just $\hat{\mathbf{\Sigma}}_j^{-1}$ if $q_{n,j} \wedge d_n \le \hat{r}$) are sent back to $I(j)$. This step costs $O_p(\sum_{j\in\hat{J}}\{q_{n,j}^2 + (q_{n,j}d_n + \hat{r}(q_{n,j} + d_n))\mathbf{1}\{q_{n,j} \wedge d_n > \hat{r}\}\}$ bytes of communication per node.

Now we can start the second-stage RGA iterations. Initializing $\hat{\mathbf{G}}_{(h)}^{(0)} = \mathbf{0}$ and $\hat{\mathbf{U}}_{(h)}^{(0)} =$

$\mathbf{Y}_{(h)}$ for each computing nodes. At iteration k, for each $j \in \hat{J}$, nodes in $I(j)$ send $\mathbf{U}_j^\top \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{X}_{j,(h)}^\top \hat{\mathbf{U}}_{(h)}^{(k-1)} \mathbf{V}_j$ to the master. The master aggregates the matrices

$$\left\{ \mathbf{P}_j = \sum_{h=1}^{H} \mathbf{U}_j^\top \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{X}_{j,(h)}^\top \hat{\mathbf{U}}_{(h)}^{(k-1)} \mathbf{V}_j : j \in \hat{J} \right\},$$

and decides $\hat{j}_k = \arg\max_{j \in \hat{J}} \sigma_1(\mathbf{P}_j)$ and $\hat{\mathbf{S}}_k = L_n \mathbf{u}\mathbf{v}^\top$, where $(\mathbf{u}, \mathbf{v})$ are the leading singular vectors of $\mathbf{P}_{\hat{j}_k}$. The master node sends $\hat{\mathbf{S}}_k$ to the nodes in $I(\hat{j}_k)$. Sending the matrix $\mathbf{U}_j^\top \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{X}_{j,(h)}^\top \hat{\mathbf{U}}_{(h)}^{(k-1)} \mathbf{V}_j$ requires $O(\hat{r}^2)$ bytes of communication if $q_{n,j} \wedge d_n > \hat{r}$, and $O(q_{n,j} d_n)$ bytes otherwise. Each computing node also receives $O(\hat{r})$ or $O(q_{n,j} + d_n)$ bytes of data from the master, depending on whether $q_{n,\hat{j}_k} \wedge d_n$ is greater than $\hat{r}$.

To compute $\hat{\lambda}_k$, each node $(h, c) \in I(\hat{j}_k)$ computes and sends to the master

$$\mathbf{A}_h = \hat{\mathbf{U}}_{(h)}^{(k-1)\top} \mathbf{X}_{\hat{j}_k,(h)} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{U}}_{(h)}^{(k-1)\top} \hat{\mathbf{G}}_{(h)}^{(k-1)},$$

and

$$a_h = \| \mathbf{X}_{\hat{j}_k,(h)} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}_{(h)}^{(k-1)} \|_F^2.$$

The master then is able to compute $\hat{\lambda}_k = \max\{\min\{\hat{\lambda}_{k,uc}, 1\}, 0\}$, where

$$\hat{\lambda}_{k,uc} = \frac{\mathrm{tr}(\sum_{h=1}^{H} \mathbf{A}_h)}{\sum_{h=1}^{H} a_h}.$$

Subsequently, $\hat{\lambda}_k$ is sent to all nodes. In this step, because $\hat{\mathbf{G}}_h^{(k-1)}$ is of rank at most $k-1$, sending $\mathbf{A}_h$ costs $O(d_n(k \wedge d_n))$ bytes of communication.

Finally, each node $(h, c) \in I(\hat{j}_k)$ updates

$$\begin{aligned}
\hat{\mathbf{G}}_{(h)}^{(k)} =&(1-\hat{\lambda}_k)\hat{\mathbf{G}}_{(h)}^{(k-1)} + \hat{\lambda}_k \mathbf{X}_{\hat{j}_k,(h)} \hat{\boldsymbol{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \\
\hat{\mathbf{U}}_{(h)}^{(k)} =&\mathbf{Y}_{(h)} - \hat{\mathbf{G}}_{(h)}^{(k)}, \\
\hat{\mathbf{B}}_{\hat{j}_k}^{(k)} =&(1-\hat{\lambda}_k)\hat{\mathbf{B}}_{\hat{j}_k}^{(k-1)} + \hat{\lambda}_k \hat{\boldsymbol{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \\
\hat{\mathbf{B}}_{j}^{(k)} =&(1-\hat{\lambda}_k)\hat{\mathbf{B}}_{j}^{(k-1)}, \quad j \in \mathcal{I}_c - \{\hat{j}_k\},
\end{aligned}$$

and also sends (possibly via the master node) the matrix $\mathbf{X}_{\hat{j}_k,(h)} \hat{\boldsymbol{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top$ (which is of rank one and costs $O(n_1 + d_n)$ bytes of communication) to the nodes $\{(h, c') : c' \neq c\}$. Then the node $(h, c') \notin I(\hat{j}_k)$ is able to update $\hat{\mathbf{G}}_{(h)}^{(k)}$, $\hat{\mathbf{U}}_{(h)}^{(k)}$, and $\hat{\mathbf{B}}_{j}^{(k)}$ as above.

It can be verified the above procedure implements the second-stage RGA. Moreover, the communication cost for node (h, c) at the k-th iteration is at most

$$O\left(\sum_{j \in \hat{J} \cap \mathcal{I}_c} \left(\hat{r}^2 \mathbf{1}\{q_{n,j} \wedge d_n > \hat{r}\} + q_{n,j} d_n \mathbf{1}\{q_{n,j} \wedge d_n \leq \hat{r}\} \right) + d_n k + n_1 \right).$$

As a result, the above procedure to implement TSRGA has the following guarantee.

Corollary 3.6.1. *Suppose $\hat{J}$ and $\{\hat{r}_j : j \in \hat{J}\}$ satisfy the sure-screening property* (3.12) *as $n_1 \to \infty$, and assume (C1)-(C6). If $\max_{1 \leq j \leq p_n} q_{n,j} = O(n_1^\alpha)$, then the above procedure achieves an error of order*

$$\frac{1}{d_n} \sum_{j=1}^{p_n} \|\mathbf{B}_j^* - \hat{\mathbf{B}}_j\|_F^2 = O_p\left(\frac{\mathfrak{s}_n \xi_n^2}{n^2 d_n} \log \frac{n^2 d_n}{\xi_n^2} + \frac{\xi_n^2}{n^2 \delta_n^2} \mathbf{1}\{J_o \neq \emptyset\} \right)$$

with a communication complexity per computing node of order

$$O_p\left(n_1^{\max\{2\alpha,1\}} s_n^2 + (s_n^2 n_1^\alpha d_n + n_1) \log \frac{n^2 d_n}{\xi_n^2} + s_n^{10} \log \frac{n^2 d_n}{\xi_n^2} + d_n s_n^8 \left(\log \frac{n^2 d_n}{\xi_n^2} \right)^2 \right).$$

The proof of Corollary 3.6.1 is an accounting on the communication costs shown above, whose details are relegated to Section 3.8.3. The communication complexity is still free of the ambient dimension p_n, but the dimension of the predictors $\max_{1\leq j\leq p_n} q_{n,j}$ comes into play, which was not a factor in the purely feature-distributed case. The additional communication between segments could inflate the communication complexity compared to the purely feature-distributed case. If $\alpha \leq 0.5$ and $s_n = O(1)$, the communication complexity, up to poly-logarithmic factors, reduces to $O_p(n_1 + n_1^{\alpha} d_n + d_n)$, which is no larger than the purely feature-distributed case $O_p(n_1 + d_n)$ if $d_n = O(n_1^{1-\alpha})$. On the other hand, if $\alpha > 0.5$ and $s_n = O(1)$, the communication complexity becomes $O_p(n_1^{2\alpha} + n_1^{\alpha} d_n)$ (again ignoring poly-logarithmic terms), which is higher than the purely feature-distributed case. These costs are incurred in the greedy search as well as in the determination of $\hat{\lambda}_k$. Finally, we note that the above procedure is sequential, and certain improvements can be achieved with some carefully designed communication protocol. However, methods or algorithms for speeding up convergence or lowering communication of the proposed TSRGA with horizontal partition is left for future research.

3.7 Conclusion

This chapter presented a two-stage relaxed greedy algorithm (TSRGA) for estimating high-dimensional multivariate linear regression models with feature-distributed data. Our main contribution is that the communication complexity of TSRGA is independent of the feature dimension, which is often very large in feature-distributed data. Instead, the complexity depends on the sparsity of the underlying model, making the proposed approach a highly scalable and efficient method for analyzing large data sets. We also briefly discussed applying TSRGA to huge data sets that require both vertical and horizontal partitions.

We would like to point out a possible future extension. In some applications, it is of paramount importance to protect the privacy of each node's data. Thus, modifying TSRGA

so that privacy can be guaranteed for feature-distributed data is an important direction for future research.

3.8 Supplementary details

3.8.1 Second-stage RGA with feature-distributed data

The following algorithm presents the pseudo-code for the implementation of the second-stage RGA with feature-distributed data.

3.8.2 Proofs

This section presents the essential elements of the proofs of our main results. Further technical details are relegated to Section 3.8.3.

The analysis of TSRGA relies on what we call the "noiseless updates," a theoretical device constructed as follows. Initialize $\mathbf{G}^{(0)} = \mathbf{0}$ and $\mathbf{U}^{(0)} = \tilde{\mathbf{Y}}$. For $1 \leq k \leq K_n$, suppose $(\hat{j}_k, \tilde{\mathbf{B}}_{\hat{j}_k})$ is chosen according to (3.3) by the first-stage RGA. The noiseless updates are defined as

$$\mathbf{G}^{(k)} = (1 - \lambda_k)\hat{\mathbf{G}}^{(k-1)} + \lambda_k \mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k}, \tag{3.24}$$

where

$$\lambda_k \in \arg\min_{0 \leq \lambda \leq 1} \|\tilde{\mathbf{Y}} - (1 - \lambda)\hat{\mathbf{G}}^{(k-1)} - \lambda \mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k}\|_F^2. \tag{3.25}$$

Recall that $\tilde{\mathbf{Y}} = \sum_{j=1}^{p_n} \mathbf{X}_j \mathbf{B}_j^*$ is the noise-free part of the response. Thus $\mathbf{G}^{(k)}$ is unattainable in practice. Similarly, we can define the noiseless updates for the second-stage RGA, with $\tilde{\mathbf{B}}_{\hat{j}_k}$ replaced by $\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top$ in (3.24) and (3.25). By definition of the updates, for first- and second-stage RGA,

Algorithm 2: Feature-distributed second-stage RGA

Input: Number of required iterations K_n, $L_n > 0$, pre-selected $\hat{J}$.

Output: Each worker $1 \le c \le M$ has the coefficient matrices $\{\hat{\mathbf{B}}_j : j \in \mathcal{I}_c\}$ to use for prediction.

Initialization: $\hat{\mathbf{B}}_j = \mathbf{0}$, for all j, and $\hat{\mathbf{G}}^{(0)} = \mathbf{0}$

for $k = 1, 2, \ldots, K_n$ **do**

  Workers $c = 1, 2, \ldots, M$ **in parallel do**

    if $k > 1$ **then**

      Receive $(c^*, \hat{\lambda}_{k-1}, \sigma_{\hat{j}_{k-1}}, \mathbf{u}_{\hat{j}_{k-1}}, \mathbf{v}_{\hat{j}_{k-1}})$ from the master.

      $\hat{\mathbf{G}}^{(k-1)} = (1 - \hat{\lambda}_{k-1})\hat{\mathbf{G}}^{(k-2)} + \hat{\lambda}_{k-1}\sigma_{\hat{j}_{k-1}}\mathbf{u}_{\hat{j}_{k-1}}\mathbf{v}_{\hat{j}_{k-1}}^\top$.

      $\hat{\mathbf{B}}_j = (1 - \hat{\lambda}_{k-1})\hat{\mathbf{B}}_j$ for $j \in \mathcal{I}_c \cap \hat{J}$.

      if $c = c^*$ **then**

        $\hat{\mathbf{B}}_{\hat{j}_{k-1}^{(c)}} = \hat{\mathbf{B}}_{\hat{j}_{k-1}} + \hat{\lambda}_{k-1}\hat{\mathbf{\Sigma}}^{-1}_{\hat{j}_{k-1}^{(c)}}\mathbf{U}_{\hat{j}_{k-1}^{(c)}}\hat{\mathbf{S}}^{(c)}_{k-1}\mathbf{V}^\top_{\hat{j}_{k-1}^{(c)}}$

      end

    end

    $\hat{\mathbf{U}}^{(k-1)} = \mathbf{Y} - \hat{\mathbf{G}}^{(k-1)}$

    $(\hat{j}_k^{(c)}, \hat{\mathbf{S}}_k^{(c)}) \in \arg\max_{\substack{j \in \mathcal{I}_c \cap \hat{J} \\ \|\mathbf{S}_k\|_* \le L_n}} |\langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{S}_k\mathbf{V}_j^\top\rangle|$

    $\rho_c = |\langle \hat{\mathbf{U}}^{(k-1)}, \mathbf{X}_{\hat{j}_k^{(c)}}\hat{\mathbf{\Sigma}}^{-1}_{\hat{j}_k^{(c)}}\mathbf{U}_{\hat{j}_k^{(c)}}\hat{\mathbf{S}}_k^{(c)}\mathbf{V}^\top_{\hat{j}_k^{(c)}}\rangle|$

    Find the leading singular value decomposition:

$$\mathbf{X}_{\hat{j}_k^{(c)}}\hat{\mathbf{\Sigma}}^{-1}_{\hat{j}_k^{(c)}}\mathbf{U}_{\hat{j}_k^{(c)}}\hat{\mathbf{S}}_{k^{(c)}}\mathbf{V}^\top_{\hat{j}_k^{(c)}} = \sigma_{\hat{j}_k^{(c)}}\mathbf{u}_{\hat{j}_k^{(c)}}\mathbf{v}^\top_{\hat{j}_k^{(c)}}$$

    Send $(\sigma_{\hat{j}_k^{(c)}}, \mathbf{u}_{\hat{j}_k^{(c)}}, \mathbf{v}_{\hat{j}_k^{(c)}}, \rho_c)$ to the master.

  end

  Master do

    Receives $\{(\sigma_{\hat{j}_k^{(c)}}, \mathbf{u}_{\hat{j}_k^{(c)}}, \mathbf{v}_{\hat{j}_k^{(c)}}, \rho_c) : c = 1, 2, \ldots, M\}$ from the workers.

    $c^* = \arg\max_{1 \le c \le M} \rho_c$

    $\sigma_{\hat{j}_k} = \sigma_{\hat{j}_k^{(c^*)}}, \mathbf{u}_{\hat{j}_k} = \mathbf{u}_{\hat{j}_k^{(c^*)}}, \mathbf{v}_{\hat{j}_k} = \mathbf{v}_{\hat{j}_k^{(c^*)}}$

    $\hat{\mathbf{G}}^{(k)} = (1 - \hat{\lambda}_k)\hat{\mathbf{G}}^{(k-1)} + \hat{\lambda}_k\sigma_{\hat{j}_k}\mathbf{u}_{\hat{j}_k}\mathbf{v}^\top_{\hat{j}_k}$, where $\hat{\lambda}_k$ is determined by

$$\hat{\lambda}_k \in \arg\min_{0 \le \lambda \le 1} \|\mathbf{Y} - (1 - \lambda)\hat{\mathbf{G}}^{(k-1)} - \lambda\sigma_{\hat{j}_k}\mathbf{u}_{\hat{j}_k}\mathbf{v}^\top_{\hat{j}_k}\|_F^2.$$

    Broadcast $(c^*, \hat{\lambda}_k, \sigma_{\hat{j}_k}, \mathbf{u}_{\hat{j}_k}, \mathbf{v}_{\hat{j}_k})$ to all workers.

  end

end

$$\begin{aligned}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 \leq & \|\tilde{\mathbf{Y}} - \mathbf{G}^{(k)}\|_F^2 + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \\ \leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \end{aligned} \tag{3.26}$$

Recursively applying (3.26), we have for any $1 \leq l \leq k$,

$$\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 \leq \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-l)}\|_F^2 + 2\sum_{j=1}^{l} \langle \mathbf{E}, \hat{\mathbf{G}}^{(k-j+1)} - \mathbf{G}^{(k-j+1)} \rangle. \tag{3.27}$$

(3.27) bounds the empirical prediction error at step k by the empirical prediction error at step $k-l$ and a remainder term involving the noise and the noiseless updates up to step l. This will be handy in numerous places throughout the proofs.

Two other useful identities are

$$\max_{\substack{1 \leq j \leq p_n \\ \|\mathbf{B}_j\|_* \leq L_n}} \langle \mathbf{A}, \mathbf{X}_j \mathbf{B}_j \rangle = \sup_{\substack{\mathbf{B}_j \in \mathbb{R}^{q_n, j \times d_n}, j=1,2,\ldots,p_n \\ \sum_j \|\mathbf{B}_j\|_* \leq L_n}} \left\langle \mathbf{A}, \sum_{j=1}^{p_n} \mathbf{X}_j \mathbf{B}_j \right\rangle \tag{3.28}$$

and

$$\max_{\substack{j \in \hat{J} \\ \|\mathbf{S}\|_* \leq L_n}} \langle \mathbf{A}, \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{S} \mathbf{V}_j^\top \rangle = \sup_{\sum_{j \in \hat{J}} \|\mathbf{S}_j\|_* \leq L_n} \left\langle \mathbf{A}, \sum_{j \in \hat{J}} \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{S}_j \mathbf{V}_j^\top \right\rangle, \tag{3.29}$$

where $\mathbf{A} \in \mathbb{R}^{n \times d_n}$ is arbitrary. These identities hold because the maximum of the inner product is attained at the extreme points in the ℓ_1 ball. The proofs are omitted for brevity.

We first prove an auxiliary lemma which guarantees sub-linear convergence of the empirical prediction error, whose proof makes use of the noiseless updates introduced above.

Lemma 3.8.1. *Assume (C1)-(C2) and that* $\sum_{j=1}^{p_n} \|\mathbf{B}_j^*\|_* \leq d_n^{1/2} L$. *RGA has the following*

uniform rate of convergence.

$$\max_{1\le k\le K_n}\frac{(nd_n)^{-1}\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k)}\|_F^2}{k^{-1}}=O_p(1). \tag{3.30}$$

Proof. Let $1\le m\le K_n$ be arbitrary. Note that for any $1\le k\le K_n$,

$$\begin{aligned}
&\langle\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)},\mathbf{X}_{\hat{j}_k}\tilde{\mathbf{B}}_{\hat{j}_k}-\hat{\mathbf{G}}^{(k-1)}\rangle\\
=&\langle\mathbf{Y}-\hat{\mathbf{G}}^{(k-1)},\mathbf{X}_{\hat{j}_k}\tilde{\mathbf{B}}_{\hat{j}_k}-\hat{\mathbf{G}}^{(k-1)}\rangle-\langle\mathbf{E},\mathbf{X}_{\hat{j}_k}\tilde{\mathbf{B}}_{\hat{j}_k}-\hat{\mathbf{G}}^{(k-1)}\rangle\\
\ge&\max_{\substack{1\le j\le p_n\\ \|\mathbf{B}_j\|_*\le L_n}}\{\langle\mathbf{Y}-\hat{\mathbf{G}}^{(k-1)},\mathbf{X}_j\mathbf{B}_j-\hat{\mathbf{G}}^{(k-1)}\rangle\}-2L_n\xi_E\\
\ge&\max_{\substack{1\le j\le p_n\\ \|\mathbf{B}_j\|_*\le L_n}}\{\langle\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)},\mathbf{X}_j\mathbf{B}_j-\hat{\mathbf{G}}^{(k-1)}\rangle\}-4L_n\xi_E.
\end{aligned}\tag{3.31}$$

Put

$$\mathcal{E}_n(m)=\left\{\min_{1\le l\le m}\max_{\substack{1\le j\le p_n\\ \|\mathbf{B}_j\|_*\le L_n}}\langle\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(l-1)},\mathbf{X}_j\mathbf{B}_j-\hat{\mathbf{G}}^{(l-1)}\rangle>\tilde{\tau}d_n^{1/2}\xi_E\right\}, \tag{3.32}$$

for some $\tilde{\tau}>4L_0$. It follows from (3.28) and (3.31) that on $\mathcal{E}_n(m)$, for all $1\le k\le m$,

$$\begin{aligned}
&\langle\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)},\mathbf{X}_{\hat{j}_k}\tilde{\mathbf{B}}_{\hat{j}_k}-\hat{\mathbf{G}}^{(k-1)}\rangle\\
\ge&(1-\frac{4L_0}{\tilde{\tau}})\max_{\substack{1\le j\le p_n\\ \|\mathbf{B}_j\|_*\le L}}\{\langle\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)},\mathbf{X}_j\mathbf{B}_j-\hat{\mathbf{G}}^{(k-1)}\rangle\}\\
\ge&(1-\frac{4L_0}{\tilde{\tau}})\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)}\|_F^2\\
:=&\tau\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)}\|_F^2\\
\ge&0,
\end{aligned}\tag{3.33}$$

where $\tau = 1 - 4L_0/\tilde{\tau}$. This, together with Lemma 3.8.2(iii) in Section 3.8.3, implies

$$\lambda_k = \frac{\langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)} \rangle}{\|\mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}$$

for $1 \leq k \leq m$ on $\mathcal{E}_n(m)$ except for a vanishing event. This, combined with (3.26) and (3.33), yields

$$\begin{aligned}
\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 \leq & \|\tilde{\mathbf{Y}} - \mathbf{G}^{(k)}\|_F^2 + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \\
= & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)} - \lambda_k (\mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)})\|_F^2 + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \\
= & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 - \frac{\langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)} \rangle^2}{\|\mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\|_F^2} + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \\
\leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \left\{ 1 - \frac{\tau^2 \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}{\|\mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\|_F^2} \right\} + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle
\end{aligned} \tag{3.34}$$

for all $1 \leq k \leq m$ on $\mathcal{E}_n(m)$ except for a vanishing event. By (C1), with probability tending to one, $\|\mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \leq 4L_n^2 n\mu$ and $\|\tilde{\mathbf{Y}}\|_F^2 \leq (1 - \epsilon_L)^2 L_n^2 n\mu$. Now by Lemma 3.8.3 and Lemma 3.8.2(ii) in Section 3.8.3, we have

$$\begin{aligned}
\frac{1}{nd_n} \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(m)}\|_F^2 \leq & \frac{4L_0^2 \mu}{1 + m\tau^2} + 2 \sum_{l=1}^{m} \frac{|\langle \mathbf{E}, \hat{\mathbf{G}}^{(l)} - \mathbf{G}^{(l)} \rangle|}{nd_n} \\
= & \frac{4L_0^2 \mu}{1 + m\tau^2} + 2 \sum_{l=1}^{m} |\hat{\lambda}_l - \lambda_l| \frac{|\langle \mathbf{E}, \mathbf{X}_{\hat{j}_l} \tilde{\mathbf{B}}_{\hat{j}_l} - \hat{\mathbf{G}}^{(l-1)} \rangle|}{nd_n} \\
\leq & \frac{4L_0^2 \mu}{1 + m\tau^2} + \frac{8}{1 - \epsilon_L} \frac{m\xi_E^2}{n^2 d_n},
\end{aligned} \tag{3.35}$$

on $\mathcal{E}_n(m)$ except for a vanishing event. Note that by (C2), $m\xi_E^2/(n^2 d_n) \leq m^{-1}(K_n \xi_E/(nd_n^{1/2}))^2 = O_p(m^{-1})$. Furthermore, it is shown in Section 3.8.3 that on $\mathcal{E}_n^c(m)$

except for a vanishing event,

$$\frac{1}{nd_n}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(m)}\|_F^2 \leq \frac{\tilde{\tau}\xi_E}{n\sqrt{d_n}} + \frac{8m\xi_E^2}{(1-\epsilon_L)n^2 d_n}. \tag{3.36}$$

Combining (3.35) and (3.36) yields the desired result. □

Now we are ready to prove the main results.

Proof of Theorem 3.3.1. Since $d_n^{1/2}L \geq \sum_{j=1}^{p_n}\|\mathbf{B}_j^*\|_* \geq \sharp(J_n)\min_{j\in J_n}\sigma_{r_j^*}(\mathbf{B}_j^*)$ and $s_n = o(K_n^2)$, it follows that $\sharp(J_n) = o(K_n)$, and by (C3), with probability tending to one, $\lambda_{\min}(\mathbf{X}(\hat{J}_k \cup J_n)^\top \mathbf{X}(\hat{J}_k \cup J_n)) \geq n\mu^{-1}$, for all $1 \leq k \leq K_n$, where $\hat{J}_k = \{\hat{j}_1, \hat{j}_2, \ldots, \hat{j}_k\}$. Let $\mathcal{G}_n = \{\text{there exists some } j \text{ such that } \operatorname{rank}(\mathbf{B}_j^*) > \operatorname{rank}(\hat{\mathbf{B}}_j^{(\hat{k})})\}$. Then on $\mathcal{G}_n$ except for a vanishing event, it follows from (3.28), (C3), Eckart-Young theorem and (C4) that

$$\begin{aligned}
\min_{1\leq m\leq \hat{k}} \max_{\substack{1\leq j\leq p_n \\ \|\mathbf{B}_j\|_*\leq L}} \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(m)}, \mathbf{X}_j\mathbf{B}_j - \hat{\mathbf{G}}^{(m)}\rangle \geq & \min_{1\leq m\leq \hat{k}} \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(m)}\|_F^2 \\
\geq & n\mu^{-1} \min_{1\leq m\leq \hat{k}} \|\mathbf{B}_j^* - \hat{\mathbf{B}}_j^{(m)}\|_F^2 \\
\geq & n\mu^{-1} \min_{\operatorname{rank}(\mathbf{B})<r_j^*} \|\mathbf{B}_j^* - \mathbf{B}\|_F^2 \\
\geq & \frac{nd_n}{\mu s_n}.
\end{aligned} \tag{3.37}$$

By (3.37), (C2) and (C4), we have $\lim_{n\to\infty}\mathbb{P}\left(\mathcal{G}_n \cap \mathcal{E}_n^c(\hat{k})\right) \leq \lim_{n\to\infty}\mathbb{P}(nd_n^{1/2} \leq \tilde{\tau}\mu s_n\xi_E) = 0$, where $\mathcal{E}_n(\cdot)$ is defined in (3.32). Hence it suffices to show $\lim_{n\to\infty}\mathbb{P}(\mathcal{G}_n \cap \mathcal{E}_n(\hat{k})) = 0$. By (3.37) and the same argument as in (3.34), on $\mathcal{G}_n \cap \mathcal{E}_n(\hat{k})$ except for a vanishing event,

$$\begin{aligned}
\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 \leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \left\{1 - \frac{\tau^2\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}{\|\mathbf{X}_{\hat{j}_k}\tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}\right\} + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)}\rangle \\
\leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \left\{1 - \frac{\tau^2 s_n^{-1}}{4L_0^2\mu^2}\right\} + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)}\rangle,
\end{aligned}$$

and thus

$$nd_n\hat{\sigma}_k^2 \leq \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \left(1 - \frac{\tau^2 s_n^{-1}}{4L_0^2\mu^2}\right) + \|\mathbf{E}\|_F^2 + 2\langle \mathbf{E}, \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}, \rangle$$

for $1 \leq k \leq \hat{k}$. It follows that

$$\begin{aligned}\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2} \leq & \frac{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 + (nd_n)^{-1}\|\mathbf{E}\|_F^2 + 4L_0\xi_E/(nd_n^{1/2})}{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 + (nd_n)^{-1}\|\mathbf{E}\|_F^2 - 4L_0\xi_E/(nd_n^{1/2})} \\ & - \frac{\tau^2 s_n^{-1}}{4L_0^2\mu^2} \frac{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 + (nd_n)^{-1}\|\mathbf{E}\|_F^2 - 4L_0\xi_E/(nd_n^{1/2})} \\ := & A_k - B_k, \end{aligned} \tag{3.38}$$

for $1 \leq k \leq \hat{k}$ on $\mathcal{G}_n \cap \mathcal{E}_n(\hat{k})$ except for a vanishing event. We show in Section 3.8.3 that on $\mathcal{G}_n \cap \mathcal{E}_n(\hat{k})$ except for a vanishing event, for all $1 \leq k \leq \hat{k}$,

$$A_k \leq 1 + \frac{12ML_0\xi_E}{nd_n^{1/2}}, \tag{3.39}$$

and

$$B_k \geq \frac{\tau^2}{4L_0^2\mu^2} s_n^{-1} \frac{1}{1 + \mu M s_n} \left(1 - \frac{4ML_0\xi_E}{nd_n^{1/2}}\right). \tag{3.40}$$

By (3.38)-(3.40), $\max_{1 \leq k \leq \hat{k}} \hat{\sigma}_k^2/\hat{\sigma}_{k-1}^2 \leq 1 - s_n^{-2}C_n$, where

$$C_n = \frac{\tau^2}{4L_0^2\mu^2} \frac{1}{\mu M + s_n^{-1}} \left(1 - \frac{4ML_0\xi_E}{nd_n^{1/2}}\right) - 12ML_0 \frac{s_n^2\xi_E}{nd_n^{1/2}}.$$

By (C2) and (C4), it can be shown that there exists some $v > 0$ such that $C_n \geq v$ with

probability tending to one. Therefore, by the definition of $\hat{k}$,

$$\begin{aligned}\mathbb{P}(\mathcal{G}_n \cap \mathcal{E}_n(\hat{k})) \leq & \mathbb{P}(\hat{k} < K_n, \mathcal{G}_n \cap \mathcal{E}_n(\hat{k})) + \mathbb{P}(\hat{k} = K_n, \mathcal{G}_n \cap \mathcal{E}_n(\hat{k})) \\ \leq & \mathbb{P}(\max_{1 \leq k \leq \hat{k}} \hat{\sigma}_k^2 / \hat{\sigma}_{k-1}^2 \leq 1 - v s_n^{-2}, \hat{k} < K_n) + \mathbb{P}(\hat{k} = K_n, \mathcal{G}_n \cap \mathcal{E}_n(\hat{k})) + o(1) \\ = & \mathbb{P}(\hat{k} = K_n, \mathcal{G}_n \cap \mathcal{E}_n(\hat{k})) + o(1), \end{aligned} \tag{3.41}$$

if $t_n = C s_n^{-2}$ in (3.6) is chosen with $C < v$. In view of (3.41), it remains to show $\mathbb{P}(\hat{k} = K_n, \mathcal{G}_n \cap \mathcal{E}_n(\hat{k})) = o(1)$. Since $s_n = o(K_n)$ by (C4), it follows from (3.37) and Lemma 3.8.1 that

$$\begin{aligned}\mathbb{P}(\hat{k} = K_n, \mathcal{G}_n) \leq & \mathbb{P}\left(\frac{1}{n d_n} \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(K_n)}\|_F^2 \geq \frac{1}{\mu s_n} \right) + o(1) \\ = & \mathbb{P}\left(\frac{(n d_n)^{-1} \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(K_n)}\|_F^2}{K_n^{-1}} \geq \frac{K_n}{\mu s_n} \right) + o(1) \\ = & o(1), \end{aligned}$$

which completes the proof. □

Proof of Lemma 3.3.1. Letting $a_n = \lfloor D s_n^2 \rfloor$ for some arbitrary $D > 0$, we have

$$\begin{aligned}\mathbb{P}(\hat{k} > a_n) \leq & \mathbb{P}\left(\frac{\hat{\sigma}_{a_n}^2}{\hat{\sigma}_{a_n - 1}^2} < 1 - C s_n^2 \right) \\ = & \mathbb{P}\left(C s_n^{-2} < \frac{\hat{\sigma}_{a_n - 1}^2 - \zeta_n^2 - (\hat{\sigma}_{a_n}^2 - \zeta_n^2)}{\zeta_n^2 + \hat{\sigma}_{a_n - 1}^2 - \zeta_n^2} \right) \\ \leq & \mathbb{P}\left(C s_n^{-2} < \frac{\hat{\sigma}_{a_n - 1}^2 - \zeta_n^2}{M^{-1} + \hat{\sigma}_{a_n - 1}^2 - \zeta_n^2} + \frac{4 L_0 \xi_E n^{-1} d_n^{-1/2}}{M^{-1} + \hat{\sigma}_{a_n - 1}^2 - \zeta_n^2} \right) + o(1). \end{aligned} \tag{3.42}$$

Put $A_n = \{\hat{\sigma}^2_{a_n-1} - \zeta_n^2 > 0\}$. Then (3.42) implies

$$\begin{aligned}\mathbb{P}(\hat{k} > a_n, A_n) \leq & \mathbb{P}\left(M^{-1} + \hat{\sigma}^2_{a_n-1} - \zeta_n^2 < \frac{\hat{\sigma}^2_{a_n-1} - \zeta_n^2}{Cs_n^{-2}} + \frac{4L_0 s_n^2 \xi_E}{Cnd_n^{1/2}}, A_n\right) + o(1) \\ \leq & \mathbb{P}\left(M^{-1} < Z_n \frac{s_n^2}{C(a_n-1)} + \frac{4L_0}{C}\frac{s_n^2\xi_E}{nd_n^{1/2}}\right) + o(1),\end{aligned}$$

where

$$Z_n := \max_{1\leq k\leq K_n} \frac{|(nd_n)^{-1}\|\mathbf{Y} - \hat{\mathbf{G}}^{(k)}\|_F^2 - \zeta_n^2|}{k^{-1}}.$$

Since $|(nd_n)^{-1}\|\mathbf{Y} - \hat{\mathbf{G}}^{(k)}\|_F^2 - \zeta_n^2| \leq (nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 + 4L_0\xi_E n^{-1}d_n^{-1/2}$, where $\zeta_n^2 = (nd_n)^{-1}\|\mathbf{E}\|_F^2$, it follows from Lemma 3.8.1 that $Z_n = O_p(1)$. Thus $\limsup_{n\to\infty}\mathbb{P}(\hat{k} > a_n, A_n) \to 0$ as $D \to 0$. On A_n^c, it is not difficult to show that

$$\hat{\sigma}^2_{a_n} - \zeta_n^2 \leq \hat{\sigma}^2_{a_n-1} - \zeta_n^2 \leq 0$$

and

$$\max\left\{\frac{1}{nd_n}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(a_n-1)}\|_F^2, \frac{1}{nd_n}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(a_n)}\|_F^2\right\} \leq \frac{4L_0\xi_E}{nd_n^{1/2}}.$$

It follows that on A_n^c,

$$\begin{aligned}\frac{\hat{\sigma}^2_{a_n}}{\hat{\sigma}^2_{a_n-1}} = & 1 - \frac{\hat{\sigma}^2_{a_n-1} - \hat{\sigma}^2_{a_n}}{\hat{\sigma}^2_{a_n-1}} \\ \geq & 1 - \frac{\hat{\sigma}^2_{a_n-1} - \zeta_n^2 - (\hat{\sigma}^2_{a_n} - \zeta_n^2)}{\zeta_n^2 - 4L_0\xi_E n^{-1}d_n^{-1/2}} \\ \geq & 1 - \frac{1}{\zeta_n^2 - 4L_0\xi_E n^{-1}d_n^{-1/2}}\frac{16L_0\xi_E}{nd_n^{1/2}}.\end{aligned}$$

By (C4), we have

$$\mathbb{P}(\hat{k} > a_n, A_n^c) \leq \mathbb{P}\left(Cs_n^{-2} \leq \frac{1}{\zeta_n^2 - 4L_0\xi_E n^{-1} d_n^{-1/2}} \frac{16L_0\xi_E}{nd_n^{1/2}} \right) = o(1),$$

which completes the proof. □

Before proving Theorem 3.3.2, we introduce the following uniform convergence rate for the second-stage RGA, which is also of independent interest.

Theorem 3.8.1. *Assume the same as Theorem 3.3.1, and additionally (C5) and (C6) hold. The second-stage RGA satisfies*

$$\max_{1 \leq m \leq K_n} \frac{(nd_n)^{-1} \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(m)}\|_F^2}{\left(1 - \frac{\tau^2}{64\mu^5\kappa_n}\right)^m + \frac{(m+\kappa_n)\xi_E^2}{n^2 d_n} + \frac{\xi_E^2}{\delta_n^2 n^2} \mathbf{1}\{J_o \neq \emptyset\}} = O_p(1), \tag{3.43}$$

where $\tau < 1$ is an absolute constant.

Proof. By Theorem 3.3.1, we can assume $\text{rank}(\mathbf{B}_j^*) \leq \hat{r}_j$ holds for all j in the following analysis. Let $1 \leq m \leq K_n$ be arbitrary. Observe that for the second-stage RGA, each $\hat{\mathbf{G}}^{(k)}$, $k = 1, 2, \ldots$, lies in the set

$$\mathcal{C}_L = \left\{ \mathbf{H} = \sum_{j \in \hat{J}} \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{D}_j \mathbf{V}_j^\top : \sum_{j \in \hat{J}} \|\mathbf{D}_j\|_* \leq L_n \right\}. \tag{3.44}$$

By (3.29) and a similar argument as (3.31)-(3.33), we have, for all $1 \leq k \leq m$,

$$\begin{aligned}
&\langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^T - \hat{\mathbf{G}}^{(k-1)} \rangle \\
\geq &\tau \max_{\substack{j \in \hat{J}_{\hat{k}} \\ \|\mathbf{S}\|_* \leq L_n}} \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{S} \mathbf{V}_j^\top - \hat{\mathbf{G}}^{(k-1)} \rangle \\
= &\tau \sup_{\mathbf{H} \in \mathcal{C}_L} \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{H} - \hat{\mathbf{G}}^{(k-1)} \rangle,
\end{aligned} \tag{3.45}$$

where $\tau = 1 - 4\mu L_0/\tilde{\tau}$ and $\tilde{\tau} > 4\mu L_0$ on the event

$$\mathcal{F}_n(m) = \left\{ \min_{1\le k\le m} \max_{\substack{j\in\hat{J}_{\hat{k}} \\ \|\mathbf{S}\|_*\le L_n}} \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_j\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{S}\mathbf{V}_j^\top - \hat{\mathbf{G}}^{(k-1)}\rangle > \tilde{\tau} d_n^{1/2}\xi_E \right\}.$$

Define

$$\mathcal{B} = \left\{ \mathbf{H} = \sum_{j\in\hat{J}_{\hat{k}}} \mathbf{X}_j\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{D}_j\mathbf{V}_j^\top : \|\bar{\mathbf{Y}} - \mathbf{H}\|_F^2 \le \frac{9nd_nL_0^2}{16\mu^3\kappa_n} \right\},$$

where

$$\bar{\mathbf{Y}} = \sum_{j\in\hat{J}_o} \mathbf{X}_j\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{L}_j\mathbf{\Lambda}_j\mathbf{R}_j^\top\mathbf{V}_j^\top + \sum_{j\in\hat{J}-\hat{J}_o} \mathbf{X}_j\mathbf{B}_j^*, \tag{3.46}$$

in which $\hat{J}_o = \{j \in \hat{J} : \hat{r} < \min\{q_{n,j}, d_n\}\}$, $\mathbf{\Lambda}_j$ are defined in (C6), and $\mathbf{L}_j$, $\mathbf{R}_j$ are $\hat{r}\times\bar{r}_j$ matrices such that $\mathbf{L}_j^T\mathbf{L}_j = \mathbf{I}_{\bar{r}_j} = \mathbf{R}_j^T\mathbf{R}_j$ to be specified later (recall that $\hat{r} \ge \bar{r}_j = \text{rank}(\mathbf{X}_j^\top\tilde{\mathbf{Y}})$ because of Theorem 3.3.1). We claim that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{B} \subseteq \mathcal{C}_L) = 1, \tag{3.47}$$

whose proof is relegated to Section 3.8.3. Now put $\mathbf{H}^{(l)} = \hat{\mathbf{G}}^{(l)} + (1+\alpha_l)(\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(l)})$ for $l = 1, 2, \ldots$, where

$$\alpha_l = \frac{3\sqrt{nd_n}L_0}{4\mu^{3/2}\sqrt{\kappa_n}\|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(l)}\|_F} \ge 0.$$

Then (3.47) implies that $\mathbb{P}(\mathbf{H}^{(l)} \in \mathcal{C}_L, l = 1, 2, \ldots) \to 1$. Thus by (3.45),

$$\begin{aligned} &\langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)} \rangle \\ &\geq \tau \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{H}^{(k-1)} - \hat{\mathbf{G}}^{(k-1)} \rangle \end{aligned} \tag{3.48}$$

for all $1 \leq k \leq m$ on $\mathcal{F}_n(m)$ except for a vanishing event. Put $\mathcal{H}_n(m) = \{\|\tilde{\mathbf{Y}} - \bar{\mathbf{Y}}\|_F < 2^{-1} \min_{1 \leq l \leq m} \|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(l-1)}\|_F\}$. On $\mathcal{F}_n(m) \cap \mathcal{H}_n(m)$ except for a vanishing event, (3.48) and Cauchy-Schwarz inequality yield

$$\begin{aligned} &\langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)} \rangle \\ \geq &\tau \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{H}^{(k-1)} - \hat{\mathbf{G}}^{(k-1)} \rangle \\ \geq &\tau (1 + \alpha_{k-1}) \left\{ \|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 - \|\tilde{\mathbf{Y}} - \bar{\mathbf{Y}}\|_F \|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F \right\} \\ \geq &\frac{\tau(1 + \alpha_{k-1})}{2} \|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \geq 0 \end{aligned}$$

for all $1 \leq k \leq m$. Notice that $\|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F \geq (2/3)\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F$ for all $1 \leq k \leq m$

on $\mathcal{H}_n(m)$. Hence, by Lemma 3.8.2(ii), (iii), and a similar argument used in (3.34),

$$\begin{aligned}
\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 \leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 - \frac{\langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k}^\top \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)} \rangle^2}{\|\mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k}^\top \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\|_F^2} \\
& + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \\
\leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 - \frac{\tau^2 \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{H}^{(k-1)} - \hat{\mathbf{G}}^{(k-1)} \rangle^2}{4n\mu L_n^2} + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \\
\leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 - \frac{\tau^2 (1 + \alpha_{k-1})^2}{16n\mu L_n^2} \|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^4 + 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)} \rangle \\
\leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 - \frac{\tau^2 \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}{64\mu^4 \kappa_n} \\
& + 2(\hat{\lambda}_k - \lambda_k) \langle \mathbf{E}, \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k}^\top \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)} \rangle \\
\\
\leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \left(1 - \frac{\tau^2}{64\mu^4 \kappa_n}\right) + \frac{8\mu}{1 - \epsilon_L} \frac{\xi_E^2}{n}
\end{aligned}$$

for all $1 \leq k \leq m$ on $\mathcal{F}_n(m) \cap \mathcal{H}_n(m)$ except for a vanishing event. It follows that, on the same event,

$$\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(m)}\|_F^2 \leq \|\tilde{\mathbf{Y}}\|_F^2 \left(1 - \frac{\tau^2}{64\mu^4 \kappa_n}\right)^m + \frac{8\mu}{1 - \epsilon_L} \frac{m\xi_E^2}{n}. \tag{3.49}$$

By (3.29), on $\mathcal{F}_n^c(m) \cap \mathcal{H}_n(m)$ there exists some $1 \leq k \leq m$ such that

$$\begin{aligned}
\tilde{\tau} d_n^{1/2} \xi_E \geq & \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{H}^{(k-1)} - \hat{\mathbf{G}}^{(k-1)} \rangle \\
\geq & (1 + \alpha_{k-1}) \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)} \rangle \\
\geq & \frac{1}{2} (1 + \alpha_{k-1}) \|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \\
\geq & \frac{3\sqrt{nd_n} L_0}{8\mu^{3/2} \sqrt{\kappa_n}} \|\bar{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F,
\end{aligned}$$

which implies

$$\begin{aligned}\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(m)}\|_F^2 \leq & \|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)}\|_F^2+\frac{8\mu}{1-\epsilon_L}\frac{(m-k)\xi_E^2}{n} \\ \leq & 2\|\tilde{\mathbf{Y}}-\bar{\mathbf{Y}}\|_F^2+2\|\bar{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)}\|_F^2+\frac{8\mu}{1-\epsilon_L}\frac{(m-k)\xi_E^2}{n} \\ \leq & \frac{5}{2}\|\bar{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)}\|_F^2+\frac{8\mu}{1-\epsilon_L}\frac{(m-k)\xi_E^2}{n} \\ \leq & \left(\frac{160\tilde{\tau}^2\mu^3}{9L^2}\kappa_n+\frac{8\mu}{1-\epsilon_L}(m-k)\right)\frac{\xi_E^2}{n}. \end{aligned} \tag{3.50}$$

Next, on $\mathcal{H}_n^c(m)$, there exists some $1 \leq k \leq m$ such that $\|\bar{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)}\|_F^2 \leq 4\|\tilde{\mathbf{Y}}-\bar{\mathbf{Y}}\|_F^2$. By (3.27) and the parallelogram law,

$$\begin{aligned}\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(m)}\|_F^2 \leq & \|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(k-1)}\|_F^2+2\sum_{j=k}^{m}\langle \mathbf{E}, \hat{\mathbf{G}}^{(j)}-\mathbf{G}^{(j)}\rangle \\ \leq & 10\|\tilde{\mathbf{Y}}-\bar{\mathbf{Y}}\|_F^2+\frac{8\mu}{1-\epsilon_L}\frac{(m-k)\xi_E^2}{n} \end{aligned} \tag{3.51}$$

on $\mathcal{H}_n^c(m)$ except for a vanishing event. Finally, note that (3.49)-(3.51) are valid for any choice of $\mathbf{L}_j$ and $\mathbf{R}_j$ so long as $\mathbf{L}_j^\top \mathbf{L}_j = \mathbf{I}_{\bar{r}_j} = \mathbf{R}_j^\top \mathbf{R}_j$, $j \in \hat{J}$. In Section 3.8.3, we show that $\mathbf{L}_j$, $\mathbf{R}_j$, $j \in \hat{J}_o$, can be chosen so that

$$\frac{1}{nd_n}\|\tilde{\mathbf{Y}}-\bar{\mathbf{Y}}\|_F^2 \leq 8\mu L^2\frac{\xi_E^2}{(n\delta_n-\xi_E)^2} = O_p\left(\frac{\xi_E^2}{n^2\delta_n^2}\right). \tag{3.52}$$

Hence, by (3.49)-(3.52), the desired result follows. □

Now we are ready to prove our last main result.

Proof of Theorem 3.3.2. Note first that $\mathcal{C}_L$ (defined in (3.44)) is a convex compact set almost surely. Thus we can define $\mathbf{Y}^*$ to be the orthogonal projection of $\mathbf{Y}$ onto $\mathcal{C}_L$. Since

$\hat{\mathbf{G}}^{(m)} \in \mathcal{C}_L$ and $\hat{\sigma}_m^2 \leq \hat{\sigma}_{m_n}^2$ for $m \geq m_n$, it follows that for $m \geq m_n$,

$$\begin{aligned}
\|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m)}\|_F^2 =& \|\mathbf{Y} - \hat{\mathbf{G}}^{(m)}\|_F^2 - \|\mathbf{Y} - \mathbf{Y}^*\|_F^2 + 2\langle \mathbf{Y}^* - \mathbf{Y}, \mathbf{Y}^* - \hat{\mathbf{G}}^{(m)} \rangle \\
\leq& \|\mathbf{Y} - \hat{\mathbf{G}}^{(m_n)}\|_F^2 - \|\mathbf{Y} - \mathbf{Y}^*\|_F^2 \\
=& \|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m_n)}\|_F^2 - 2\langle \tilde{\mathbf{Y}} - \mathbf{Y}^*, \hat{\mathbf{G}}^{(m_n)} - \mathbf{Y}^* \rangle - 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(m_n)} - \mathbf{Y}^* \rangle \\
\leq& 2\|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m_n)}\|_F^2 + \|\mathbf{Y}^* - \tilde{\mathbf{Y}}\|_F^2 - 2\langle \mathbf{E}, \hat{\mathbf{G}}^{(m_n)} - \mathbf{Y}^* \rangle. \qquad (3.53)
\end{aligned}$$

Note that if $\mathbf{H}, \mathbf{G}$ are in $\mathcal{C}_L$ with $\mathbf{H} = \sum_{j \in \hat{J}} \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{S}_j^H \mathbf{V}_j^\top$ and $\mathbf{G} = \sum_{j \in \hat{J}} \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{S}_j^G \mathbf{V}_j^\top$, then by Proposition 1 and (C3) we have

$$\|\mathbf{H} - \mathbf{G}\|_F^2 \geq \frac{n}{\mu^3 \kappa_n} \left\{ \sum_{j \in \hat{J}} \|\mathbf{S}_j^H - \mathbf{S}_j^G\|_* \right\}^2 .$$

Hence

$$|\langle \mathbf{E}, \mathbf{H} - \mathbf{G} \rangle| \leq \mu \xi_E \sum_{j \in \hat{J}} \|\mathbf{S}_j^H - \mathbf{S}_j^G\|_* \leq \xi_E \sqrt{\frac{\mu^5 \kappa_n}{n}} \|\mathbf{H} - \mathbf{G}\|_F. \qquad (3.54)$$

Combining (3.53) and (3.54) yields

$$\|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m)}\|_F^2 \leq 2\|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m_n)}\|_F^2 + \|\mathbf{Y}^* - \tilde{\mathbf{Y}}\|_F^2 + 2\xi_E \sqrt{\frac{\mu^5 \kappa_n}{n}} \|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m)}\|_F.$$

Since $x^2 \leq c + bx$ $(x, b, c \geq 0)$ implies $x \leq (b + \sqrt{b^2 + 4c})/2$, we have

$$\|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m)}\|_F^2 \leq 2\|\mathbf{Y}^* - \tilde{\mathbf{Y}}\|_F^2 + 4\|\mathbf{Y}^* - \hat{\mathbf{G}}^{(m_n)}\|_F^2 + 4\mu^5 \frac{\kappa_n \xi_E^2}{n}. \qquad (3.55)$$

By (3.55) and repeated applications of the parallelogram law, it is straightforward to show

$$\frac{1}{nd_n}\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(m)}\|_F^2 \le \frac{C_1}{nd_n}\left\{\|\tilde{\mathbf{Y}}-\mathbf{Y}^*\|_F^2+\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(m_n)}\|_F^2+\frac{\mu^5\kappa_n\xi_E^2}{n}\right\}$$

for some absolute constant C_1. The right-hand side does not depend on m, so the inequality still holds if we take supremum over $m \ge m_n$ on the left-hand side. Moreover, by (C3) and Theorem 3.3.1, we have

$$\sup_{m\ge m_n}\frac{1}{d_n}\sum_{j=1}^{p_n}\|\mathbf{B}_j^*-\hat{\mathbf{B}}_j^{(m)}\|_F^2 = O_p\left(\frac{1}{nd_n}\left\{\|\tilde{\mathbf{Y}}-\mathbf{Y}^*\|_F^2+\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(m_n)}\|_F^2+\frac{\mu^5\kappa_n\xi_E^2}{n}\right\}\right) \tag{3.56}$$

By Theorem 3.8.1 and the choice of m_n, we have

$$\frac{1}{nd_n}\|\tilde{\mathbf{Y}}-\hat{\mathbf{G}}^{(m_n)}\|_F^2 = O_p\left(\frac{\kappa_n\xi_n^2}{n^2d_n}\log\frac{n^2d_n}{\xi_n^2}+\frac{\xi_n^2}{n^2\delta_n^2}\right). \tag{3.57}$$

By (C6), it is not difficult to show $\bar{\mathbf{Y}}$, defined in (3.46), is in $\mathcal{C}_L$. It follows from the definition of $\mathbf{Y}^*$ that

$$\begin{aligned}\|\tilde{\mathbf{Y}}-\mathbf{Y}^*\|_F^2 =& \|\mathbf{Y}-\mathbf{Y}^*\|_F^2-\|\mathbf{E}\|_F^2-2\langle\mathbf{E},\tilde{\mathbf{Y}}-\mathbf{Y}^*\rangle\\ \le& \|\mathbf{Y}-\bar{\mathbf{Y}}\|_F^2-\|\mathbf{E}\|_F^2-2\langle\mathbf{E},\tilde{\mathbf{Y}}-\mathbf{Y}^*\rangle\\ =& \|\tilde{\mathbf{Y}}-\bar{\mathbf{Y}}\|_F^2+2\langle\mathbf{E},\mathbf{Y}^*-\bar{\mathbf{Y}}\rangle.\end{aligned} \tag{3.58}$$

By (3.54) again,

$$|\langle\mathbf{E},\mathbf{Y}^*-\bar{\mathbf{Y}}\rangle| \le \xi_E\left(\frac{\mu^5\kappa_n}{n}\right)^{1/2}\|\bar{\mathbf{Y}}-\mathbf{Y}^*\|_F. \tag{3.59}$$

Now if $\|\bar{\mathbf{Y}}-\mathbf{Y}^*\|_F \ge 2\|\tilde{\mathbf{Y}}-\mathbf{Y}^*\|_F$, then $\|\bar{\mathbf{Y}}-\mathbf{Y}^*\|_F \le 2\|\bar{\mathbf{Y}}-\tilde{\mathbf{Y}}\|_F$. This, together with

(3.58), (3.59), and (3.52), yields

$$\begin{aligned}\|\tilde{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 \leq & \|\tilde{\mathbf{Y}} - \bar{\mathbf{Y}}\|_F^2 + 4\xi_E \left(\frac{\mu^5 \kappa_n}{n}\right)^{1/2} \|\bar{\mathbf{Y}} - \tilde{\mathbf{Y}}\|_F \\ \leq & 2\|\tilde{\mathbf{Y}} - \bar{\mathbf{Y}}\|_F^2 + 4\mu^5 \frac{\kappa_n \xi_E^2}{n} \\ \leq & 16\mu L_0^2 \frac{n d_n \xi_E^2}{(n\delta_n - \xi_E)^2} + 4\mu^5 \frac{\kappa_n \xi_E^2}{n}. \end{aligned} \tag{3.60}$$

On the other hand, if $\|\bar{\mathbf{Y}} - \mathbf{Y}^*\|_F < 2\|\tilde{\mathbf{Y}} - \mathbf{Y}^*\|_F$, then (3.58) and (3.59) imply

$$\|\tilde{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 \leq \|\tilde{\mathbf{Y}} - \bar{\mathbf{Y}}\|_F^2 + 4\xi_E \left(\frac{\mu^5 \kappa_n}{n}\right)^{1/2} \|\tilde{\mathbf{Y}} - \mathbf{Y}^*\|_F.$$

By a similar argument used to obtain (3.55), this and (3.52) yield

$$\begin{aligned}\|\tilde{\mathbf{Y}} - \mathbf{Y}^*\|_F^2 \leq & 16\mu^5 \frac{\kappa_n \xi_E^2}{n} + 2\|\tilde{\mathbf{Y}} - \bar{\mathbf{Y}}\|_F^2 \\ \leq & 16\mu^5 \frac{\kappa_n \xi_E^2}{n} + 16\mu L_0^2 \frac{n d_n \xi_E^2}{(n\delta_n - \xi_E)^2}. \end{aligned} \tag{3.61}$$

In view of (3.56), (3.57), (3.60), (3.61) and (C5), the deisred result follows. □

3.8.3 Further technical details

In this section, we present some additional auxiliary results along with the proofs of (3.36), (3.39), (3.40), (3.47), (3.52). Some existing results that are useful in our proofs are also stated here for completeness with the references to their proofs in the literature. These results are stated in the forms that are most convenient for our use, which may not be in full generality.

Proposition 1 (Ruhe, 1970)**.** *Let* $\mathbf{A}, \mathbf{B}$ *be matrices with size* $m \times n$ *and* $n \times p$ *respectively.*

Then

$$\sum_{j=1}^{n} \sigma_j^2(\mathbf{A})\sigma_j^2(\mathbf{B}) \geq \|\mathbf{AB}\|_F^2 \geq \sum_{j=1}^{n} \sigma_{n-j+1}^2(\mathbf{A})\sigma_j^2(\mathbf{B}).$$

Remark 3.8.1. One consequence of this inequality we frequently use is $\sigma_1^2(\mathbf{A})\|\mathbf{B}\|_F^2 \geq \|\mathbf{AB}\|_F^2 \geq \sigma_n^2(\mathbf{A})\|\mathbf{B}\|_F^2$. Note also that by transposition the roles of $\mathbf{A}$ and $\mathbf{B}$ can be interchanged on the left- and right-most expressions.

Lemma 3.8.2. *Assume (C1)-(C2) and that* $\sum_{j=1}^{p_n} \|\mathbf{B}_j^*\|_* \leq L$. *Suppose* $L_n = d_n^{1/2} L_0$ *is chosen so that* $L_0 \geq L/(1-\epsilon_L)$ *with* $1 - \epsilon_L \leq 1/(4\mu^2)$. *Then for first- and second-stage RGA, with probability tending to one,*

(i)

$$\inf_{k\geq 1} \frac{1}{nd_n} \|\mathbf{X}_{\hat{j}_k} \tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \geq (1-\epsilon_L)\mu L_0^2 \tag{3.62}$$

$$\inf_{k\geq 1} \frac{1}{nd_n} \|\mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^{\top} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \geq (1-\epsilon_L)\mu L_0^2 \tag{3.63}$$

(ii)

$$\sup_{k\geq 1} |\lambda_k - \hat{\lambda}_k| \leq \frac{2}{(1-\epsilon_L)L_0} \frac{\xi_E}{n\sqrt{d_n}} \tag{3.64}$$

(iii)

$$\max_{1\leq k\leq K_n} \lambda_k \leq 1. \tag{3.65}$$

Proof. We shall prove the results for the second-stage RGA. The corresponding proofs for first-stage RGA follow similarly and thus are omitted. It is also sufficient to prove (i)-(iii) assuming the condition described in (C1) holds almost surely because the event that the condition holds has probability tending to one. It will greatly simplify the exposition

(without repeating that the inequalities holds except on a vanishing event). Note that

$$
\begin{aligned}
&\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)} \rangle \\
=&\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \mathbf{Y} - \hat{\mathbf{G}}^{(k-1)} \rangle - \langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \mathbf{E} \rangle \\
\geq& - |\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \mathbf{E} \rangle| \\
\geq& - \|\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{X}_{\hat{j}_k}^\top \mathbf{E}\|_{op} \|\mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top\|_* \\
\geq& - \mu L_n \xi_E,
\end{aligned}
$$

where the first inequality follows because $\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^T, \mathbf{Y} - \hat{\mathbf{G}}^{(k-1)} \rangle \geq 0$ with probability one and the second inequality follows because the dual norm of the nuclear norm is the operator norm. By Proposition 1, we have

$$
\begin{aligned}
&\|\mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \\
\geq& \|\mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top\|_F^2 - 2\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \hat{\mathbf{G}}^{(k-1)} \rangle \\
\geq& n\mu^{-1} \|\mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top\|_F^2 \\
&+ 2\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)} \rangle - 2\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \tilde{\mathbf{Y}} \rangle \\
\geq& n\mu^{-1} L_n^2 - 2\mu L_n \xi_E - 2\langle \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top, \tilde{\mathbf{Y}} \rangle,
\end{aligned}
$$

where the last inequality follows from the fact that $\hat{\mathbf{S}}_k$ is rank-one with singular value L_n. Thus, by writing $\hat{\mathbf{S}}_k = L_n \mathbf{a}\mathbf{b}^T$ for some unit vectors $\mathbf{a}, \mathbf{b}$, we have $\|\mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^T\|_F^2 =$

$L_n^2\|\mathbf{U}_{\hat{j}_k}\mathbf{a}\mathbf{b}^T\mathbf{V}_{\hat{j}_k}^T\|_F^2 = L_n^2$. Next, observe that

$$\begin{aligned}|\langle \mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top, \tilde{\mathbf{Y}}\rangle| =& \left|\sum_{j=1}^{p_n}\langle \mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top, \mathbf{X}_j\mathbf{B}_j^*\rangle\right| \\ \leq& \sum_{j=1}^{p_n}\|\mathbf{B}_j^*\|_*\|\mathbf{X}_j^\top\mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top\|_{op} \\ \leq& (1-\epsilon_L)L_n^2 n\mu.\end{aligned}$$

Therefore,

$$\begin{aligned}(nd_n)^{-1}\|\mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \geq& \mu^{-1}L_0^2 - 2(1-\epsilon_L)L_0^2\mu - 2\mu L_0\frac{\xi_E}{n\sqrt{d_n}} \\ \geq& 2(1-\epsilon_L)L_0^2\mu - 2\mu L_0\frac{\xi_E}{n\sqrt{d_n}}.\end{aligned}$$

Since $\xi_E = o_p(n\sqrt{d_n})$ by (C2), (3.63) follows.

For (3.64), note first that if the solutions to the line search problems (3.9) and (3.25) (with $\tilde{\mathbf{B}}_{\hat{j}_k}$ replaced by $\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top$) for second-stage RGA are not constrained to be in $[0,1]$, then they are given by

$$\begin{aligned}\hat{\lambda}_{k,uc} =& \frac{\langle \mathbf{Y} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\rangle}{\|\mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\|_F^2}, \\ \lambda_{k,uc} =& \frac{\langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}, \mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\rangle}{\|\mathbf{X}_{\hat{j}_k}\hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1}\mathbf{U}_{\hat{j}_k}\hat{\mathbf{S}}_k\mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\|_F^2}.\end{aligned}$$

Since $\hat{\mathbf{G}}^{(l)}$ can always be expressed as $\hat{\mathbf{G}}^{(l)} = \sum_{j\in\hat{J}}\mathbf{X}_j\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{A}_j\mathbf{V}_j^\top$ with $\sum_{j\in\hat{J}}\|\mathbf{A}_j\|_* \leq$

L_n, it follows that

$$
\begin{aligned}
|\hat{\lambda}_k - \lambda_k| \le |\hat{\lambda}_{k,uc} - \lambda_{k,uc}| =& \frac{|\langle \mathbf{E}, \mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)} \rangle|}{\|\mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\|_F^2} \\
\le& \frac{2L_n \mu \xi_E}{\|\mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^\top - \hat{\mathbf{G}}^{(k-1)}\|_F^2} \\
\le& \frac{2\xi_E}{n d_n^{1/2} (1-\epsilon_L) L_0},
\end{aligned}
$$

with probability tending to one, where the last inequality follows from (3.63).

For (3.65), it suffices to prove that $\lim_{n\to\infty} \mathbb{P}(E_n) = 1$, where $E_n = \big\{\max_{1\le k\le K_n} \lambda_{k,uc} \le 1\big\}$. On E_n^c, there exists some k such that, by Cauchy-Schwarz inequality and (3.27),

$$
\begin{aligned}
\|\mathbf{X}_{\hat{j}_k} \hat{\mathbf{\Sigma}}_{\hat{j}_k}^{-1} \mathbf{U}_{\hat{j}_k} \hat{\mathbf{S}}_k \mathbf{V}_{\hat{j}_k}^T - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \le& \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 \\
\le& \|\tilde{\mathbf{Y}}\|_F^2 + 2\sum_{j=1}^{k-1} \langle \mathbf{E}, \hat{\mathbf{G}}^{(k-j)} - \mathbf{G}^{(k-j)} \rangle \\
=& \|\tilde{\mathbf{Y}}\|_F^2 + 2\sum_{l=1}^{k-1} (\hat{\lambda}_l - \lambda_l) \langle \mathbf{E}, \mathbf{X}_{\hat{j}_l} \hat{\mathbf{\Sigma}}_{\hat{j}_l}^{-1} \mathbf{U}_{\hat{j}_l} \hat{\mathbf{S}}_l \mathbf{V}_{\hat{j}_l}^\top - \hat{\mathbf{G}}^{(l-1)} \rangle \\
\le& \|\tilde{\mathbf{Y}}\|_F^2 + 4K_n L_n \mu \xi_E \max_{1\le l\le k} |\hat{\lambda}_l - \lambda_l|. \qquad (3.66)
\end{aligned}
$$

It is easy to see that

$$
\|\tilde{\mathbf{Y}}\|_F = \left\| \sum_{j=1}^{p_n} \mathbf{X}_j \mathbf{B}_j^* \right\|_F \le (1-\epsilon_L) L_n \sqrt{n\mu}. \qquad (3.67)
$$

Thus, by (3.63), (3.64) and (3.66)-(3.67), we have

$$
\mathbb{P}(E_n^c) \le \mathbb{P}\left((1-\epsilon_L) L_0^2 \mu \{1 - (1-\epsilon_L)\} \le \frac{8\mu}{1-\epsilon_L} \frac{K_n \xi_E^2}{n^2 d_n} \right) + o(1) = o(1),
$$

where the last equality follows from (C2). □

Lemma 3.8.3. *Let $\{a_m\}$ be a nonnegative sequence of reals. If*

$$a_0 \le A, \text{ and } a_m \le a_{m-1}\left(1 - \frac{\xi^2 a_{m-1}}{A}\right) + b_m,$$

for $m = 1, 2, \ldots$, where $b_m \ge 0$ with $b_0 = 0$, then for each m,

$$a_m \le \frac{A}{1 + m\xi^2} + \sum_{k=0}^{m} b_k. \tag{3.68}$$

Proof. We prove by induction. When $m = 0$, (3.68) holds by assumption. Suppose now that (3.68) holds for some $m \ge 1$. Then

$$\begin{aligned}
a_{m+1} \le & a_m\left(1 - \frac{\xi^2 a_m}{A}\right) + b_{m+1} \\
\le & \frac{1}{a_m^{-1} + \xi^2/A} + b_{m+1} \\
\le & \frac{1}{\left(\frac{A}{1+m\xi^2} + \sum_{k=0}^{m} b_k\right)^{-1} + \xi^2/A} + b_{m+1} \\
= & \frac{\frac{A}{1+m\xi^2} + \sum_{k=0}^{m} b_k}{1 + \frac{\xi^2}{A}\left(\frac{A}{1+m\xi^2} + \sum_{k=0}^{m} b_k\right)} + b_{m+1} \\
\le & \frac{A}{1 + (m+1)\xi^2} + \sum_{k=0}^{m+1} b_k,
\end{aligned}$$

where the second inequality follows from $1 - x \le 1/(1 + x)$ for $x \ge 0$. □

Remark 3.8.2. Lemma 3.8.3 is a slight modification of Lemma 3.1 of Temlyakov (2000).

Proof of (3.36). On $\mathcal{E}_n^c(m)$, there exists some $l \le m$ such that

$$\tilde{\tau} d_n^{1/2} \xi_E \ge \max_{\substack{1 \le j \le p_n \\ \|\mathbf{B}_j\|_* \le L_n}} \langle \tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(l-1)}, \mathbf{X}_j \mathbf{B}_j - \hat{\mathbf{G}}^{(l-1)} \rangle \ge \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(l-1)}\|_F^2.$$

By (3.27) and Lemma 3.8.2(ii), it follows that, on $\mathcal{E}_n^c(m)$ except for a vanishing event,

$$\begin{aligned}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(m)}\|_F^2 \leq & \|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(l-1)}\|_F^2 + 2\sum_{k=l}^{m}\langle \mathbf{E}, \hat{\mathbf{G}}^{(k)} - \mathbf{G}^{(k)}\rangle \\ \leq & \tilde{\tau} d_n^{1/2}\xi_E + 2\sum_{k=l}^{m}(\hat{\lambda}_k - \lambda_k)\langle \mathbf{E}, \mathbf{X}_{\hat{j}_k}\tilde{\mathbf{B}}_{\hat{j}_k} - \hat{\mathbf{G}}^{(k-1)}\rangle \\ \leq & \tilde{\tau} d_n^{1/2}\xi_E + \frac{8m\xi_E^2}{n(1-\epsilon_L)},\end{aligned}$$

which is the desired result. □

Proof of (3.39) *and* (3.40). Note first that for any $D > 0$, $(D+x)/(D-x) \leq 1 + 3x/D$ for all $0 \leq x \leq (1-\sqrt{2/3})D$. It is not difficult to see that

$$\begin{aligned}&\mathbb{P}\left\{\frac{4L_0\xi_E}{nd_n^{1/2}} \leq (1-\sqrt{\frac{2}{3}})\left((nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 + (nd_n)^{-1}\|\mathbf{E}\|_F^2\right), 1 \leq k \leq \hat{k}, \mathcal{G}_n\right\} \\ \geq & \mathbb{P}\left\{\frac{4L_0\xi_E}{nd_n^{1/2}} \leq (1-\sqrt{\frac{2}{3}})M^{-1}\right\} - o(1) \\ \to & 1.\end{aligned}$$

Thus, on $\mathcal{G}_n$ except for a vanishing event,

$$\begin{aligned}A_k \leq & 1 + \frac{12L_0\xi_E/(nd_n^{1/2})}{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k)}\|_F^2 + (nd_n)^{-1}\|\mathbf{E}\|_F^2} \\ \leq & 1 + 12ML_0\frac{\xi_E}{nd_n^{1/2}},\end{aligned}$$

for all $1 \leq k \leq \hat{k}$. This proves (3.39). We now turn to (3.40). Since for any positive A and B, $A/(B+x) \geq A(1-x/B)/B$ for all $x \geq 0$, it follows from (3.37) that on $\mathcal{G}_n$ except for a

vanishing event,

$$\begin{aligned} B_k \geq & \frac{\tau^2 s_n^{-1}}{4L_0^2\mu^2} \frac{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2}{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 + (nd_n)^{-1}\|\mathbf{E}\|_F^2} \\ & \times \left(1 - \frac{4L_0\xi_E/(nd_n^{1/2})}{(nd_n)^{-1}\|\tilde{\mathbf{Y}} - \hat{\mathbf{G}}^{(k-1)}\|_F^2 + (nd_n)^{-1}\|\mathbf{E}\|_F^2}\right) \\ \geq & \frac{\tau^2 s_n^{-1}}{4L_0^2\mu^2} \frac{1}{1+\mu M s_n} \left(1 - \frac{4ML_0\xi_E}{nd_n^{1/2}}\right) \end{aligned}$$

for $1 \leq k \leq \hat{k}$, which proves (3.40). □

Proof of (3.47). Let

$$\mathbf{H} = \sum_{j\in\hat{J}} \mathbf{X}_j \hat{\mathbf{\Sigma}}_j^{-1} \mathbf{U}_j \mathbf{D}_j \mathbf{V}_j^\top \in \mathcal{B}.$$

Note that Proposition 1 and (C3) imply

$$\begin{aligned} & \|\bar{\mathbf{Y}} - \mathbf{H}\|_F^2 \\ \geq & n\mu^{-1} \left\{ \sum_{j\in\hat{J}_o} \|\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j(\mathbf{L}_j\mathbf{\Lambda}_j\mathbf{R}_j^\top - \mathbf{D}_j)\mathbf{V}_j^\top\|_F^2 + \sum_{j\in\hat{J}-\hat{J}_o} \|\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{D}_j\mathbf{V}_j^\top - \mathbf{B}_j^*\|_F^2 \right\} \\ \geq & n\mu^{-3} \left\{ \sum_{j\in\hat{J}_o} \|\mathbf{L}_j\mathbf{\Lambda}_j\mathbf{R}_j^\top - \mathbf{D}_j\|_F^2 + \sum_{j\in\hat{J}-\hat{J}_o} \|\mathbf{U}_j^\top\hat{\mathbf{\Sigma}}_j\mathbf{B}_j^*\mathbf{V}_j - \mathbf{D}_j\|_F^2 \right\} \\ \geq & \frac{n}{\mu^3\kappa_n} \left\{ \sum_{j\in\hat{J}_o} \|\mathbf{L}_j\mathbf{\Lambda}_j\mathbf{R}_j^\top - \mathbf{D}_j\|_* + \sum_{j\in\hat{J}-\hat{J}_o} \|\mathbf{U}_j^\top\hat{\mathbf{\Sigma}}_j\mathbf{B}_j^*\mathbf{V}_j - \mathbf{D}_j\|_* \right\}^2. \end{aligned}$$

Since $\mathbf{H} \in \mathcal{B}$, we have

$$\left\{ \sum_{j\in\hat{J}_o} \|\mathbf{L}_j\mathbf{\Lambda}_j\mathbf{R}_j^\top - \mathbf{D}_j\|_* + \sum_{j\in\hat{J}-\hat{J}_o} \|\mathbf{U}_j^\top\hat{\mathbf{\Sigma}}_j\mathbf{B}_j^*\mathbf{V}_j - \mathbf{D}_j\|_* \right\}^2 \leq \frac{9d_nL_0^2}{16} = \frac{9L_n^2}{16}.$$

By the triangle inequality, we have $\sum_{j\in\hat{J}}\|\mathbf{D}_j\|_* \leq 3L_n/4 + \sum_{j\in\hat{J}_o}\|\mathbf{\Lambda}_j\|_* + \sum_{j\in\hat{J}-\hat{J}_o}\|\hat{\mathbf{\Sigma}}_j\mathbf{B}_j^*\|_*$. Because of (C6), and $\hat{J}_o \subset J_o$ (with probability tending to one), $\sum_{j\in\hat{J}_o}\|\mathbf{\Lambda}_j\|_* + \sum_{j\in\hat{J}-\hat{J}_o}\|\hat{\mathbf{\Sigma}}_j\mathbf{B}_j^*\|_* \leq \sum_{j\in\hat{J}}\|\hat{\mathbf{\Sigma}}_j\mathbf{B}_j^*\|_* \leq \mu(1-\epsilon_L)L_n \leq 4^{-1}\mu^{-1}L_n \leq L_n/4$. Hence $\sum_{j\in\hat{J}_{\hat{k}}}\|\mathbf{D}_j\|_* \leq L_n$, which proves $\mathbf{H}\in\mathcal{C}_L$. □

Proposition 2. *Let $\mathbf{A}^*$ be an $m\times n$ matrix and $\mathbf{A}=\mathbf{A}^*+\mathbf{E}$ be its perturbed version. Let $\mathbf{U}_*\mathbf{\Sigma}_*\mathbf{V}_*^\top$ and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be their truncated SVD of rank r_*, respectively. If $\sigma_{r_*}(\mathbf{A}^*) := \sigma_{r_*} > \sigma_{r_*+1}(\mathbf{A}^*) = 0$, and if $\|\mathbf{E}\|_{op} < \sigma_{r_*}$, then*

$$\max\{\mathrm{dist}(\mathbf{U}_*,\mathbf{U}),\mathrm{dist}(\mathbf{V}_*,\mathbf{V})\} \leq \frac{\sqrt{2}\max\{\|\mathbf{E}^\top\mathbf{U}_*\|_{op},\|\mathbf{E}\mathbf{V}_*\|_{op}\}}{\sigma_{r_*}-\|\mathbf{E}\|_{op}},$$

where $\mathrm{dist}(\mathbf{Q},\mathbf{Q}_*) = \min_{\mathbf{R}}\|\mathbf{Q}\mathbf{R}-\mathbf{Q}_*\|_{op}$ for any two orthogonal matrices $\mathbf{Q}$, $\mathbf{Q}^*$ with r columns, where the minimum is taken over all $r\times r$ orthonormal matrices.

Remark 3.8.3. Proposition 2 is a consequence of the perturbation bounds for singular values (Wedin, 1972). A proof can be found in Chen et al. (2021).

Proof of (3.52). Note first that

$$\begin{aligned}\bar{\mathbf{Y}}-\tilde{\mathbf{Y}} = &\sum_{j\in\hat{J}_o}\mathbf{X}_j\hat{\mathbf{\Sigma}}_j^{-1}(\mathbf{U}_j\mathbf{L}_j-\tilde{\mathbf{U}}_j)\mathbf{\Lambda}_j\tilde{\mathbf{V}}_j^\top \\ &+\sum_{j\in\hat{J}_o}\mathbf{X}_j\hat{\mathbf{\Sigma}}_j^{-1}\mathbf{U}_j\mathbf{L}_j\mathbf{\Lambda}_j(\mathbf{V}_j\mathbf{R}_j-\tilde{\mathbf{V}}_j)^\top.\end{aligned}$$

By triangle inequality,

$$\|\bar{\mathbf{Y}}-\tilde{\mathbf{Y}}\|_F \leq \sqrt{n\mu}\left(\sum_{j\in\hat{J}_o}\|\mathbf{\Lambda}_j\|_F\right)\left\{\max_{j\in\hat{J}_o}\|\mathbf{U}_j\mathbf{L}_j-\tilde{\mathbf{U}}_j\|_{op}+\max_{j\in\hat{J}_o}\|\mathbf{V}_j\mathbf{R}_j-\tilde{\mathbf{V}}_j\|_{op}\right\}. \quad (3.69)$$

Let $\mathbf{U}_{j,\bar{r}_j}$ and $\mathbf{V}_{j,\bar{r}_j}$ be sub-matrices of $\mathbf{U}_j$ and $\mathbf{V}_j$ consisting of column vectors that correspond to the leading $\bar{r}_j$ singular vectors. Write $\mathbf{U}_j=(\mathbf{U}_{j,\bar{r}_j},\mathbf{U}_{j,-\bar{r}_j})$ and $\mathbf{V}_j=(\mathbf{V}_{j,\bar{r}_j},\mathbf{V}_{j,-\bar{r}_j})$.

Since $\mathbf{X}_j^\top \tilde{\mathbf{Y}} = \mathbf{X}_j^\top \mathbf{Y} - \mathbf{X}_j^\top \mathbf{E}$, it follows from Proposition 2 and (C5) that there exist $\bar{r}_j \times \bar{r}_j$ orthonormal matrices $\tilde{\mathbf{L}}_j$ and $\tilde{\mathbf{R}}_j$ such that with probability tending to one,

$$\begin{aligned}\max\left\{\|\mathbf{U}_{j,\bar{r}_j}\tilde{\mathbf{L}}_j - \tilde{\mathbf{U}}_j\|_{op}, \|\mathbf{V}_{j,\bar{r}_j}\tilde{\mathbf{R}}_j - \tilde{\mathbf{V}}_j\|_{op}\right\} \leq & \frac{\sqrt{2}\max\{\|\mathbf{E}^\top \mathbf{X}_j \tilde{\mathbf{U}}_j\|_{op}, \|\mathbf{X}_j^\top \mathbf{E}\tilde{\mathbf{V}}_j\|_{op}\}}{n\delta_n - \|\mathbf{X}_j^\top \mathbf{E}\|_{op}} \\ \leq & \frac{\sqrt{2}\xi_E}{n\delta_n - \xi_E}.\end{aligned}$$

Set $\mathbf{L}_j^\top = (\tilde{\mathbf{L}}_j^\top, \mathbf{0}_{\bar{r}_j \times (\hat{r} - \bar{r}_j)})$ and $\mathbf{R}_j^\top = (\tilde{\mathbf{R}}_j^\top, \mathbf{0}_{\bar{r}_j \times (\hat{r} - \bar{r}_j)})$ for $j \in \hat{J}_o$ in (3.69). Then by (C4) and (C6), it follows that

$$\|\bar{\mathbf{Y}} - \tilde{\mathbf{Y}}\|_F^2 \leq n\mu \left(\sum_{j \in \hat{J}_o} \|\mathbf{\Lambda}_j\|_F\right)^2 \left(\frac{2\sqrt{2}\xi_E}{n\delta_n - \xi_E}\right)^2 \leq 8\mu L^2 n d_n \frac{\xi_E^2}{(n\delta_n - \xi_E)^2}.$$

□

Proof of Corollary 3.6.1. By Lemma 3.3.1, $\sharp(\hat{J}) + \hat{r} = O_p(s_n^2)$. Thus running the first-stage RGA with the just-in-time stopping criterion costs

$$O_p(s_n^2(n_1 + d_n)) \tag{3.70}$$

bytes of communication per computing node. In addition, preparing $\{\hat{\mathbf{\Sigma}}_j^{-1} : j \in \hat{J}\}$ and $(\mathbf{U}_j, \mathbf{V}_j)$ for $j \in \hat{J}$ with $q_{n,j} \wedge d_n > \hat{r}$ costs

$$\begin{aligned}& O_p\left(\sum_{j \in \hat{J}} \{q_{n,j}^2 + (q_{n,j}d_n + \hat{r}(q_{n,j} + d_n))\mathbf{1}\{q_{n,j} \wedge d_n > \hat{r}\}\}\right) \\ =& O_p(n_1^{2\alpha}s_n^2 + n_1^\alpha d_n s_n^2 + s_n^4(n_1^\alpha + d_n)).\end{aligned} \tag{3.71}$$

Since the communication costs per node at the k-th iteration of the second-stage RGA is at

most

$$
\begin{aligned}
&O_p\left(\sum_{j\in\hat{J}}\left(\hat{r}^2\mathbf{1}\{q_{n,j}\wedge d_n>\hat{r}\}+q_{n,j}d_n\mathbf{1}\{q_{n,j}\wedge d_n\le\hat{r}\}\right)+d_nk+n_1\right)\\
=&O_p\left(s_n^6+n_1^\alpha d_n s_n^2+d_nk+n_1\right),
\end{aligned}
$$

running $m_n = O_p(s_n^4 \log(n^2 d_n/\xi_n^2))$ iterations (see Theorem 3.3.2 for the definition of m_n) costs

$$
O_p\left((s_n^6+s_n^2 n_1^\alpha d_n+n_1)s_n^4\log\frac{n^2d_n}{\xi_n^2}+d_ns_n^8\left(\log\frac{n^2d_n}{\xi_n^2}\right)^2\right). \tag{3.72}
$$

Combining (3.70)-(3.72) yields the desired result. □

3.8.4 TSRGA for high-dimensional generalized linear models

In this section, we apply the idea of TSRGA to and propose a modified algorithm for estimating the generalized linear model (GLM). Focusing on the case of a scalar response y_t, the GLM postulates that the probability density function f of y_t (or the probability mass function if y_t is discrete) belongs to the exponential family. In particular,

$$
f(y;\theta)=\exp[y\theta-r(\theta)+h(y)],
$$

and

$$
\mathbb{E}(y_t|x_{t,1},\ldots,x_{t,p_n})=r'\left(\sum_{j=1}^{p_n}\beta_j^*x_{t,j}\right)
$$

where θ is called the natural parameter; r, h are known functions, and r' is the derivative of r, which is also known as the inverse of the link function (see, e.g., Dunn and Smyth,

2018; Han et al., 2023). To maximize the log-likelihood function, scaled as $y\theta - r(\theta)$, one can minimize the following loss function

$$\mathcal{L}_n(\mathbf{X}\boldsymbol{\beta}) = \frac{1}{n}\sum_{t=1}^{n}\left[-y_t\left(\sum_{j=1}^{p_n}\beta_j x_{t,j}\right) + r\left(\sum_{j=1}^{p_n}\beta_j x_{t,j}\right)\right],$$

where $\mathcal{L}_n(\boldsymbol{\tau}) = n^{-1}\sum_{t=1}^{n}(y_t\tau_t - r(\tau_t))$ for $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$.

Interpreting $y_t - r'(\sum_{j=1}^{p_n}\beta_j x_{t,j})$ as the residual, we can implement RGA as follows. First initialize $\hat{\mathbf{G}}^{(0)} = 0$. Then for $k = 1, 2, \ldots, K_n$, find

$$\hat{j}_k \in \arg\max_{1\le j\le p_n}\left|\frac{1}{n}\sum_{t=1}^{n}\left(y_t - r'(\hat{G}_t^{(k-1)})\right)x_{t,j}\right| \tag{3.73}$$

and update

$$\hat{\mathbf{G}}^{(k)} = (1-\hat{\lambda}_k)\hat{\mathbf{G}}^{(k-1)} + \hat{\lambda}_k L s_k \mathbf{z}_{\hat{j}_k}, \tag{3.74}$$

where $\hat{\mathbf{G}}^{(k)} = (\hat{G}_1^{(k)}, \ldots, \hat{G}_n^{(k)})^\top$, $L > 0$ is given, $\mathbf{z}_j = (x_{1,j}, \ldots, x_{n,j})^\top$,

$$s_k = \operatorname{sgn}\left(\frac{1}{n}\sum_{t=1}^{n}\left(y_t - r'(\hat{G}_t^{(k-1)})\right)x_{t,\hat{j}_k}\right),$$

and $\hat{\lambda}_k$ is determined by

$$\hat{\lambda}_k = \arg\min_{\lambda\in[0,1]}\mathcal{L}_n((1-\lambda)\hat{\mathbf{G}}^{(k-1)} + \lambda L s_k \mathbf{z}_{\hat{j}_k}).$$

It is not difficult to see that (3.73) can be easily solved for feature-distributed data and constructing $\hat{\mathbf{G}}^{(k)}$ in each node requires a communication cost of $O(n)$ bytes. The second-stage RGA can be implemented similarly with the set of predictors considered in (3.73) restricted to $\hat{J}$, the set of predictors chosen by the first-stage when the just-in-time criterion

is met. Finally, since $\mathcal{L}_n$ could take negative values, we modify the just-in-time criterion (3.6) as

$$\hat{k} = \min\left\{1 \le k \le K_n : \left|\frac{\mathcal{L}_n(\hat{\mathbf{G}}^{(k)})}{\mathcal{L}_n(\hat{\mathbf{G}}^{(k-1)})} - 1\right| < t_n\right\}. \tag{3.75}$$

In the same spirit as (3.6), (3.75) terminates the first-stage RGA as soon as the improvement in the loss function is below certain threshold, which would save some communication costs and speed up the algorithm.

Next, we examine the performance of this version of TSRGA ((3.73)-(3.75)) using simulations. In the following experiments, the predictors $x_{t,j}$ are generated as in Specification 2. We consider the following two specifications.

Specification 5. (Logit model) The response y_t takes only values in $\{0, 1\}$ and is generated via

$$\mathbb{P}(y_t = y, \theta_t) = \theta_t^y (1-\theta_t)^{1-y}, \quad \theta_t = \frac{1}{1 + \exp(-\sum_{j=1}^{p_n} \beta_j^* x_{t,j})}$$

where $(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*, \beta_5^*) = (-2.4, 1.8, -1.9, 2.8, -2.2)$, $\beta_j^* = 0$ for $j > 5$. For this model, we have $r(\theta) = \log(1 + \exp(\theta))$.

Specification 6. (Poisson model) The response y_t takes values in $\{0, 1, 2, \ldots\}$ and is generated via

$$\mathbb{P}(y_t = y, \theta_t) = \frac{\theta_t^y e^{-\theta_t}}{y!}, \quad \theta_t = \exp(\sum_{j=1}^{p_n} \beta_j^* x_{t,j})$$

and $(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*, \beta_5^*) = (0.15, -0.25, 0.35, -0.45, 0.55)$, $\beta_j^* = 0$ for $j > 5$. For this model, we have $r(\theta) = \exp(\theta)$.

Logit	$n = 800, p = 1200$		$n = 1200, p = 2000$	
	TSRGA	ℓ_1-GLM	TSRGA	ℓ_1-GLM
$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$	0.698	2.185	0.689	2.036
FP	0.018	82.808	0	105.070
FN	0	0	0	0
Accuracy	0.901	0.888	0.901	0.892
Poisson				
$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$	0.135	0.190	0.060	0.144
FP	6.638	25.470	1.830	25.146
FN	0.114	0.008	0.020	0
RMSE	1.324	1.363	1.283	1.329

Table 3.4: Simulation results for estimating high-dimensional GLMs. ℓ_1-GLM is defined in (3.76). The results are based on 500 simulations.

As a benchmark, we compare with the ℓ_1-regularized GLM which solves

$$\min_{\boldsymbol{\beta}} \mathcal{L}_n(\mathbf{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \tag{3.76}$$

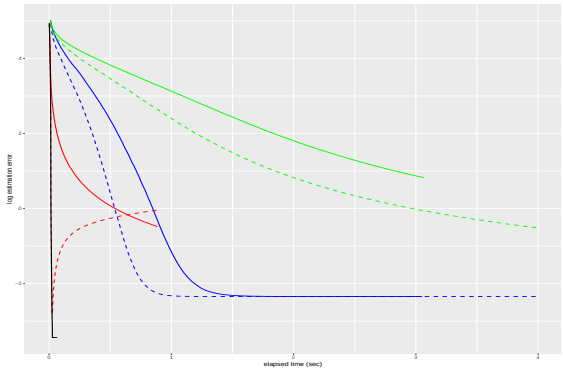
with λ selected by 5-fold cross validation. Table 3.4 reports the parameter estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, the number of irrelevant variables selected (false positives, FP) and the number of relevant variables not selected (false negatives, FN). For the logit model, we additionally report the out-of-sample prediction accuracy on a test set of size 500. For the Poisson model, the out-of-sample prediction error is measured by RMSE. All these figures are averages over 500 independent simulations.

The results show that for both the logit and Poisson models, TSRGA yields parsimonious and accurate coefficient estimates, with comparable out-of-sample prediction accuracy to the ℓ_1-GLM defined by (3.76). In particular, the low FP and FN values of TSRGA may be due to its variable selection properties. Though we expect the general conclusions about TSRGA in this chapter, such as the sure-screening property, to hold under the GLM framework, the rigorous mathematical treatment is left for future work.

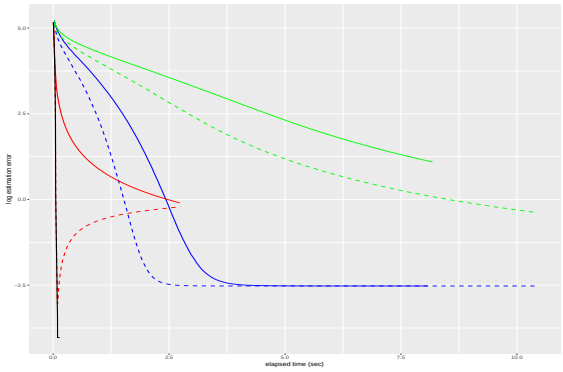
3.8.5 Complementary simulation results

In this section, we present some additional simulation results regarding Specifications 1 and 2. Figures 3.5 and 3.6 plot the parameter estimation error, as in Figures 3.1 and 3.2, against the elapsed time. Clearly, TSRGA converges within the least amount of time. In particular, its second-stage only takes a very short amount of time, thanks to the dimension reduction after the just-in-time stopping criterion. Other methods behave similarly as those in Figures 3.1 and 3.2, as their implementation cost scales directly with the number of iterations.

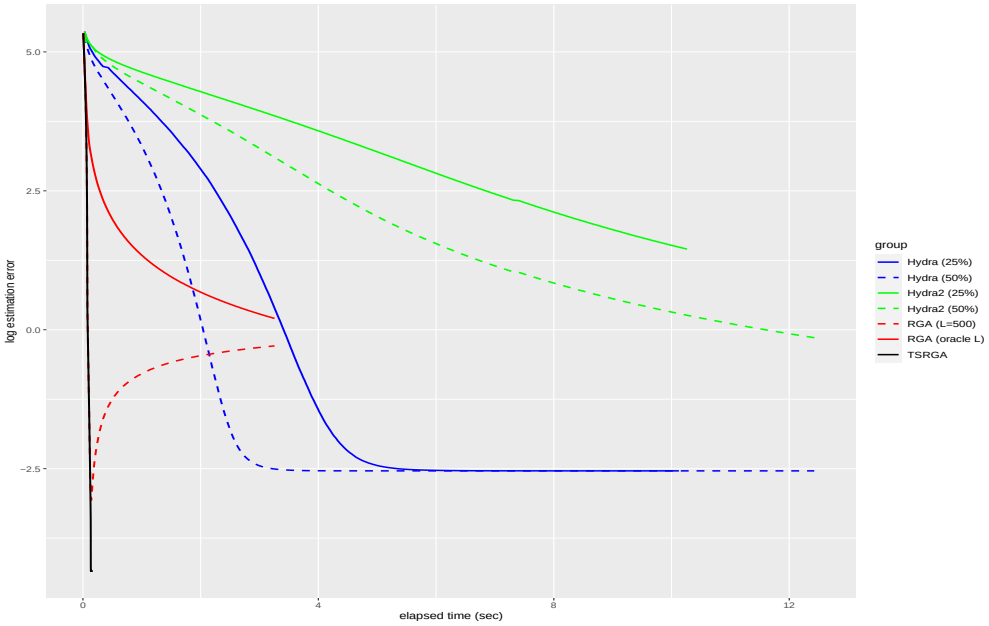
Figures 3.7 and 3.8 plot the out-of-sample prediction error (measures by the root mean square prediction error on an independent test sample) of the methods under Specifications 1 and 2. For Specification 1, the final prediction accuracy of TSRGA, cross-validated Lasso, and Hydra are similar. However, for Specification 2, TSRGA clearly is the most desirable prediction tool among the methods under consideration.



(a) $n = 800$, $p_n = 1200$

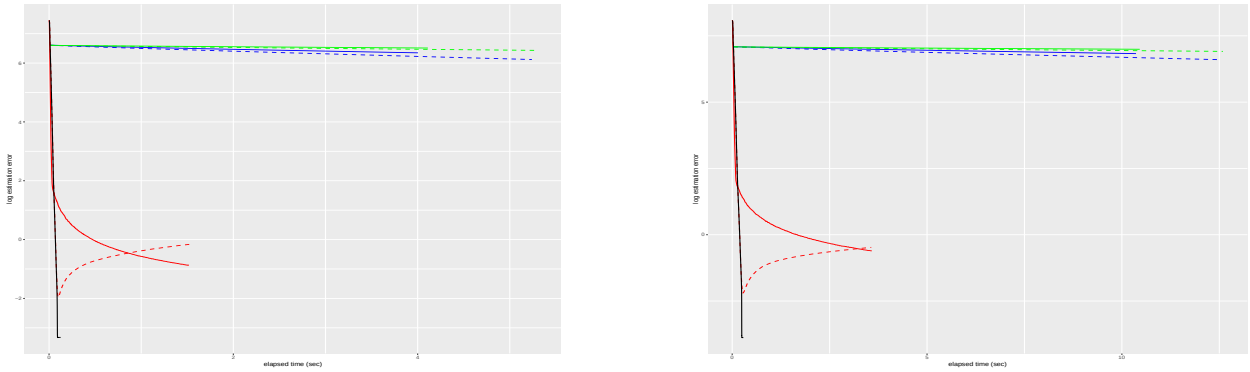


(b) $n = 1200$, $p_n = 2000$



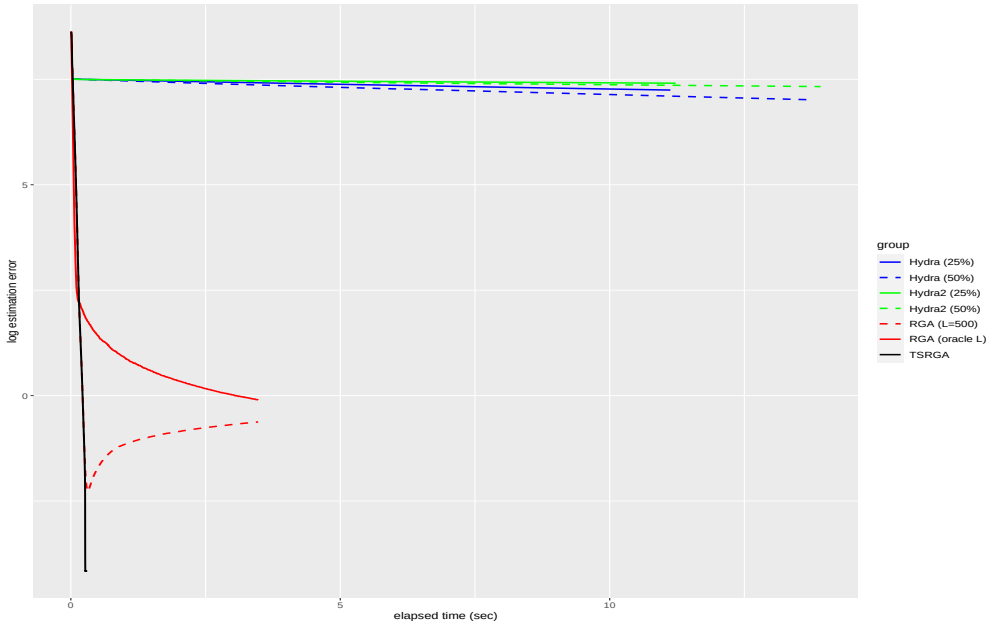
(c) $n = 1500$, $p_n = 3000$

Figure 3.5: Logarithm of parameter estimation errors of various methods against the elapsed time under Specification 1, where n is the sample size and p_n is the dimension of predictors. The results are based on 100 simulations.



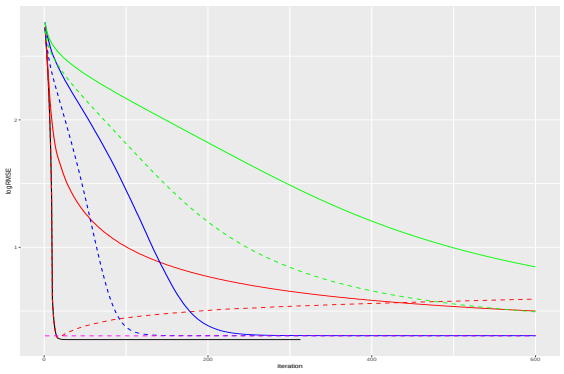
(a) $n = 800$, $p_n = 1200$

(b) $n = 1200$, $p_n = 2000$

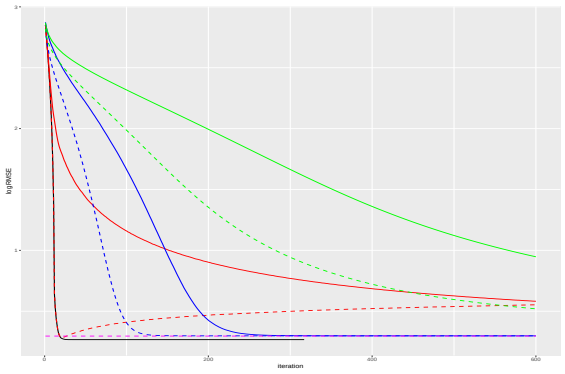


(c) $n = 1500$, $p_n = 3000$

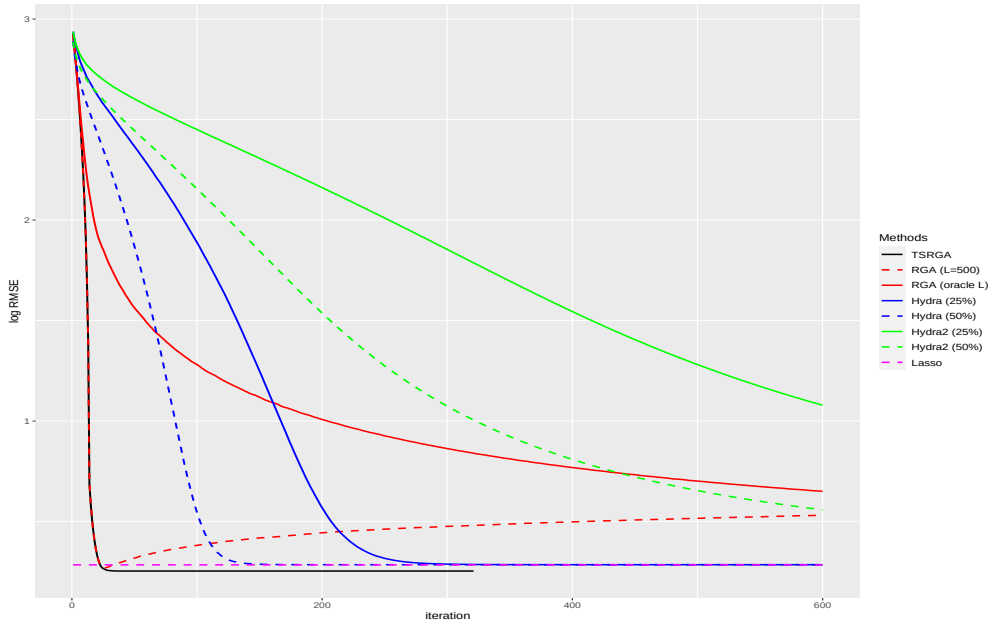
Figure 3.6: Logarithm of parameter estimation errors of various methods against the elapsed time under Specification 2, where n is the sample size and p_n is the dimension of predictors. The results are based on 100 simulations.



(a) $n = 800$, $p_n = 1200$

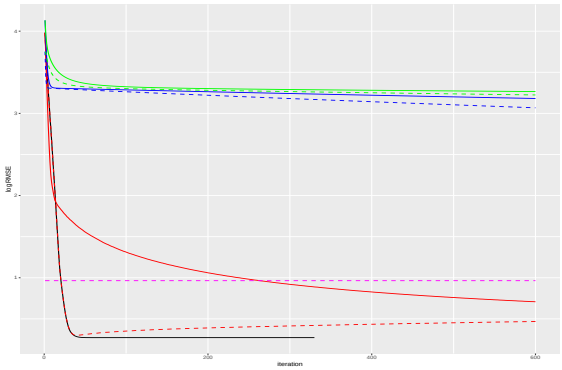


(b) $n = 1200$, $p_n = 2000$

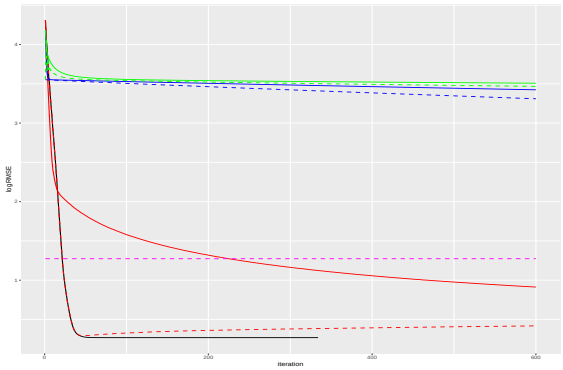


(c) $n = 1500$, $p_n = 3000$

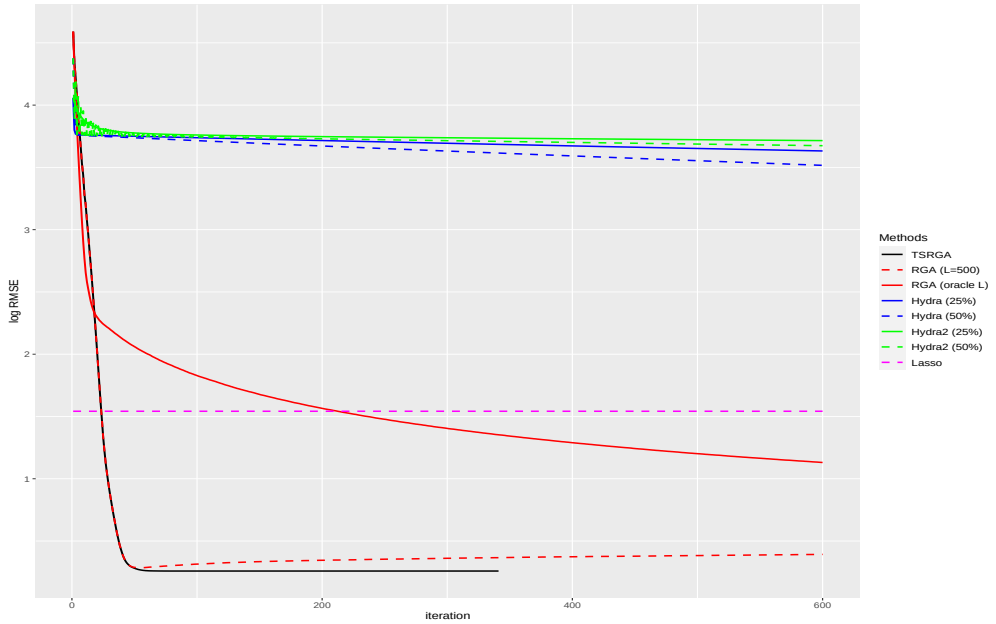
Figure 3.7: Logarithm of out-of-sample prediction errors of various methods under Specification 1, where n is the sample size and p_n is the dimension of predictors. The results are based on 100 simulations.



(a) $n = 800$, $p_n = 1200$



(b) $n = 1200$, $p_n = 2000$



(c) $n = 1500$, $p_n = 3000$

Figure 3.8: Logarithm of out-of-sample prediction errors of various methods under Specification 2, where n is the sample size and p_n is the dimension of predictors. The results are based on 100 simulations.

REFERENCES

Haitham A. Al-Zoubi. The long swings in the spot exchange rates and the complex unit roots hypothesis. *Journal of International Financial Markets, Institutions and Money*, 18 (3):236–244, 2008. ISSN 1042–4431.

Haitham A. Al-Zoubi, Jennifer A. O'Sullivan, and Abdulaziz M. Alwathnani. Business cycles, financial cycles and capital structure. *Annals of Finance*, 14(1):105–123, 2018.

Glen Baxter. An asymptotic result for the finite predictor. *Mathematica Scandinavica*, 10: 137–144, 1962.

Aurélien Bellet, Yingyu Liang, Alireza Bagheri Garakani, Maria-Florina Balcan, and Fei Sha. A distributed frank-wolfe algorithm for communication-efficient sparse learning. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 478–486, 2015.

A. Bertrand and M. Moonen. Distributed adaptive node-specific signal estimation in fully connected sensor networks—part i: Sequential node updating. *IEEE Transactions on Signal Processing*, 58(10):5277–5291, 2010.

A. Bertrand and M. Moonen. Distributed canonical correlation analysis in wireless sensor networks with application to distributed blind source separation. *IEEE Transactions on Signal Processing*, 63(18):4800–4813, 2015.

Alexander Bertrand and Marc Moonen. Distributed adaptive estimation of covariance matrix eigenvectors in wireless sensor networks with application to distributed pca. *Signal Processing*, 104:120–135, 2014.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Herman J. Bierens. Complex unit roots and business cycles: Are they real? *Econometric Theory*, 17(5):962–983, 2001.

Patrick Billingsley. *Convergence of Probability Measures*. Wiley, 1999.

Tim Bollerslev, Robert F. Engle, and Jeffrey M. Wooldridge. A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1):116–131, 1988.

D.R. Brillinger. *Time Series: Data Analysis and Theory*. Holt, Rinehart, and Winston, New York, 1975.

Peter Bühlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34 (2):559–583, 2006.

Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

Leland Bybee, Bryan T Kelly, Asaf Manela, and Dacheng Xiu. Business news and business cycles. Working Paper 29344, National Bureau of Economic Research, October 2021.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

N. H. Chan and C. Z. Wei. Limiting distributions of least squares estimates of unstable autoregressive processes. *The Annals of Statistics*, 16(1):367–401, 03 1988.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 09 2013.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. *Foundations and Trends®in Machine Learning*, 14(5):566–806, 2021.

A. Chudik, G. Kapetanios, and M. Hashem Pesaran. A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica*, 86(4):1479–1512, 2018.

Lisandro Dalcín and Yao-Lung L. Fang. mpi4py: Status update after 12 years of development. *Computing in Science and Engineering*, 23(4):47–54, 2021.

Lisandro Dalcín, Rodrigo Paz, and Mario Storti. Mpi for python. *Journal of Parallel and Distributed Computing*, 65(9):1108–1115, 2005.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

Tomás del Barrio Castro, Gianluca Cubadda, and Denise R. Osborn. On cointegration for processes integrated at different frequencies. *Journal of Time Series Analysis*, 43(3): 412–435, 2022.

Lijun Ding, Yingjie Fei, Qiantong Xu, and Chengrun Yang. Spectral frank-wolfe algorithm: Strict complementarity and linear convergence. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 2535–2544, 13–18 Jul 2020.

Lijun Ding, Jicong Fan, and Madeleine Udell. k fw: A frank-wolfe style algorithm with stronger subproblem oracles. *arXiv preprint arXiv:2006.16142*, 2021.

P.K. Dunn and G.K. Smyth. *Generalized Linear Models With Examples in R*. Springer Texts in Statistics. Springer New York, 2018.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

Yingying Fan, Gareth M. James, and Peter Radchenko. Functional additive regression. *The Annals of Statistics*, 43(5):2296–2325, 2015.

João Ricardo Faria, Juan Carlos Cuestas, and Luis A. Gil-Alana. Unemployment and entrepreneurship: A cyclical relation? *Economics Letters*, 105(3):318–320, 2009.

Olivier Fercoq, Zheng Qu, Peter Richtárik, and Martin Takáč. Fast distributed coordinate descent for non-strongly convex losses. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2014.

David F Findley and Ching-Zong Wei. Moment bounds for deriving time series CLT's and model selection procedures. *Statistica Sinica*, pages 453–480, 1993.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

Priyank Gandhi, Tim Loughran, and Bill McDonald. Using annual report sentiment as a proxy for financial distress in u.s. banks. *Journal of Behavioral Finance*, 20(4):424–436, 2019.

Zhaoxing Gao and Ruey S. Tsay. Divide-and-conquer: A distributed hierarchical factor approach to modeling large-scale time series data. *Journal of the American Statistical Association*, 118(544):2698–2711, 2023.

Dan Garber. Revisiting frank-wolfe for polytopes: Strict complementarity and sparsity. In *Advances in Neural Information Processing Systems*, volume 33, pages 18883–18893, 2020.

Luis A. Gil-Alana. Time series modeling of sunspot numbers using long-range cyclical dependence. *Solar Physics*, 257:371–381, 2009.

Luis A. Gil-Alana and Rangan Gupta. Persistence and cycles in historical oil price data. *Energy Economics*, 45:511–516, 2014.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Yuefeng Han and Ruey S. Tsay. High-dimensional linear regression for dependent data with applications to nowcasting. *Statistica Sinica*, 30(4):1797–1827, 2020.

Yuefeng Han, Ruey S. Tsay, and Wei Biao Wu. High dimensional generalized linear models for temporal dependent data. *Bernoulli*, 29(1):105–131, 2023.

Kathleen Weiss Hanley and Gerard Hoberg. Dynamic interpretation of emerging risks in the financial sector. *The Review of Financial Studies*, 32(12):4543–4603, 02 2019.

Christina Heinze, Brian McWilliams, and Nicolai Meinshausen. Dual-loco: Distributing statistical estimation using random projections. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 875–883, Cadiz, Spain, 2016.

Inge S Helland. Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics*, pages 79–94, 1982.

Yaochen Hu, Di Niu, Jianming Yang, and Shengping Zhou. Fdml: A collaborative machine learning framework for distributed features. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2232–2240, 2019.

Hsueh-Han Huang, Ngai Hang Chan, Kun Chen, and Ching-Kang Ing. Consistent order selection for arfima processes. *The Annals of Statistics*, 2022.

Shuo-Chieh Huang, Ching-Kang Ing, and Ruey S. Tsay. Asymptotic properties of nonstationary arx models with conditional heteroscedasticity. *working paper*, 2023.

C. K. Ing. A note on mean-squared prediction errors of the least squares predictors in random walk models. *Journal of Time Series Analysis*, 22:711–724, 2001.

Ching Kang Ing. Multistep prediction in autoregressive processes. *Econometric Theory*, 19: 254–279, 4 2003.

Ching-Kang Ing. Model selection for high-dimensional linear regression with dependent observations. *The Annals of Statistics*, 48(4):1959–1980, 2020.

Ching-Kang Ing and Tze Leung Lai. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, pages 1473–1513, 2011.

Ching Kang Ing and Chiao Yi Yang. Predictor selection for positive autoregressive processes. *Journal of the American Statistical Association*, 109:243–253, 2014.

Ching Kang Ing, Chor Yiu Sin, and Shu Hui Yu. Prediction errors in nonstationary autoregressions of infinite order. *Econometric Theory*, 26:774–803, 6 2010.

Ching-Kang Ing, Chor-Yiu Sin, and Shu-Hui Yu. Model selection for integrated autoregressive processes of infinite order. *Journal of Multivariate Analysis*, 106:57–71, 4 2012. ISSN 0047259X.

Ching-Kang Ing, Hai-Tang Chiou, and Meihui Guo. Estimation of inverse autocovariance matrices for long memory processes. *Bernoulli*, 22(3):1301–1330, 2016.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. *Proceedings of the 30th International Conference on Machine Learning*, 28(1):427–435, 2013.

Martin Jaggi and Simon Lacoste-Julien. On the global linear convergence of frank-wolfe optimization variants. *Advances in Neural Information Processing Systems*, 28, 2015.

Narasimhan Jegadeesh and Di Wu. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729, 2013.

Anders Bredahl Kock. Consistent and conservative model selection with the adative lasso in stationary and nonstationary autoregressions. *Econometric Theory*, 32(1):243–259, 2016.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, 2009.

T. L. Lai and C. Z. Wei. Asymptotic properties of projections with applications to stochastic regression problems. *Journal of Multivariate Analysis*, 12(3):346–370, 1982.

Qi Lei, Jiacheng Zhuo, Constantine Caramanis, Inderjit S Dhillon, and Alexandros G Dimakis. Primal-dual block generalized frank-wolfe. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gen Li, Xiaokang Liu, and Kun Chen. Integrative multi-view regression: Bridging group-sparse and low-rank models. *Biometrics*, 75(2):593–602, 2019.

Shiqing Ling and W. K. Li. Limiting distributions of maximum likelihood estimators for unstable autoregressive moving-average time series with general autoregressive heteroscedastic errors. *The Annals of Statistics*, 26(1):84–125, 02 1998.

Shiqing Ling and Michael McAleer. Stationarity and the existence of moments of a family of garch processes. *Journal of Econometrics*, 106(1):109–117, 2002. ISSN 0304-4076.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 13–18, Punta Cana, Dominican Republic, 2021.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4): 2164–2204, 2011.

Federico Maddanu and Tommaso Proietti. Modelling persistent cycles in solar activity. *Solar Physics*, 297(13), 2022.

Michael W. McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business and Economic Statistics*, 34(4):574–589, 2016.

Marcelo C. Medeiros and Eduardo F. Mendes. ℓ_1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191 (1):255–271, 2016.

Alan L. Montgomery, Victor Zarnowitz, Ruey S. Tsay, and George C. Tiao. Forecasting the u.s. unemployment rate. *Journal of the American Statistical Association*, 93(442):478–493, 1998.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17 (1):1842–1879, 2016.

Tommaso Proietti and Federico Maddanu. Modelling cycles in climate series: The fractional sinusoidal waveform process. *Journal of Econometrics*, 2024.

Gregory C. Reinsel, Raja P. Velu, and Kun Chen. *Multivariate Reduced-Rank Regression.* Springer New York, NY, 2022.

Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25, 2016.

Axel Ruhe. Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, 10:343–354, 1970.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

V. N. Temlyakov. Weak greedy algorithms. *Advances in Computational Mathematics*, 12(2): 213–227, 2000.

V. N. Temlyakov. Greedy approximation in convex optimization. *Constructive Approximation*, 41(2):269–296, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

R.S. Tsay. *Analysis of Financial Time Series.* John Wiley: Hoboken, NJ., 3 edition, 2010.

Ruey S. Tsay. Order selection in nonstationary autoregressive models. *The Annals of Statistics*, 12(4):1425–1433, 1984.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

Jialei Wang, Jason D. Lee, Mehrdad Mahdavi, Mladen Kolar, and Nathan Srebro. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *Electronic Journal of Statistics*, 11(2):4896–4944, 2017.

Xiangyu Wang, David Dunson, and Chenlei Leng. Decorrelated feature space partitioning for distributed sparse regression. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 802–810, 2016.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

C. Z. Wei. Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *The Annals of Statistics*, 15(4):1667–1682, 1987.

C. Z. Wei. On predictive least squares principles. *The Annals of Statistics*, 20(1):1–42, 1992.

Jiyan Yang, Michael W. Mahoney, Michael A. Saunders, and Yuekai Sun. Feature-distributed sparse regression: A screen-and-clean approach. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2711–2719, 2016.

Hsiang-Yuan Yeh, Yu-Ching Yeh, and Da-Bai Shen. Word vector models approach to text regression of financial risk prediction. *Symmetry*, 12(1), 2020.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006.

Wenjie Zheng, Aurélien Bellet, and Patrick Gallinari. A distributed frank-wolfe framework for learning low-rank matrices with the trace norm. *Machine Learning*, 107(8):1457–1475, 2018.

Zemin Zheng, Yingying Fan, and Jinchi Lv. High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):627–649, 2014.

Jiacheng Zhuo, Qi Lei, Alex Dimakis, and Constantine Caramanis. Communication-efficient asynchronous stochastic frank-wolfe over nuclear-norm balls. In *International Conference on Artificial Intelligence and Statistics*, pages 1464–1474, 2020.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.