Chernozhukov, Victor; Härdle, Wolfgang; Huang, Chen; Wang, Weining

Working Paper

LASSO-driven inference in time and space

cemmap working paper, No. CWP20/19

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Härdle, Wolfgang; Huang, Chen; Wang, Weining (2019) : LASSO-driven inference in time and space, cemmap working paper, No. CWP20/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.2019

This Version is available at:
https://hdl.handle.net/10419/211114

cemmap
centre for microdata methods and practice

LASSO-Driven Inference in Time and Space

Victor Chernozhukov
Wolfgang K. Härdle
Chen Huang
Weining Wang

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP20/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

LASSO-Driven Inference in Time and Space *

Victor Chernozhukov†, Wolfgang K. Härdle‡, Chen Huang§, Weining Wang¶

April 25, 2019

Abstract

We consider the estimation and inference in a system of high-dimensional regression equations allowing for temporal and cross-sectional dependency in covariates and error processes, covering rather general forms of weak dependence. A sequence of regressions with many regressors using LASSO (Least Absolute Shrinkage and Selection Operator) is applied for variable selection purpose, and an overall penalty level is carefully chosen by a block multiplier bootstrap procedure to account for multiplicity of the equations and dependencies in the data. Correspondingly, oracle properties with a jointly selected tuning parameter are derived. We further provide high-quality de-biased simultaneous inference on the many target parameters of the system. We provide bootstrap consistency results of the test procedure, which are based on a general Bahadur representation for the Z-estimators with dependent data. Simulations demonstrate good performance of the proposed inference procedure. Finally, we apply the method to quantify spillover effects of textual sentiment indices in a financial market and to test the connectedness among sectors.

JEL classification: C12, C22, C51, C53
Keywords: LASSO, time series, simultaneous inference, system of equations, Z-estimation, Bahadur representation, martingale decomposition

1 Introduction

Many applications in statistics, economics, finance, biology and psychology are concerned with a system of ultra high-dimensional objects that communicate within complex dependency channels. Given a complex system involving many factors, one builds a network model by taking a large set of regressions, i.e. regressing every factor in the system on a large subset of other factors. Examples include analysis of financial systemic risk by quantile predictive graphical

*We thank Weibiao Wu, Oliver Linton, Bryan Graham, Manfred Deistler, Hashem Pesaran, Michael Wolf, Valentina Corradi, Zudi Lu, Liangjun Su, Peter Phillips, Frank Windmeijer, Wenyang Zhang and Likai Chen for helpful comments and suggestions. We remain responsible for any errors or omissions. Financial support from the Deutsche Forschungsgemeinschaft via IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged.

†Department of Economics and Center for Statistics and Data Science, Massachusetts Institute of Technology.

‡Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin. Sim Kee Boon Institute for Financial Economics, Singapore Management University. The Wang Yanan Institute for Studies in Economics, Xiamen University. Department of Mathematics and Physics Charles University Prague.

§Faculty of Mathematics and Statistics, University of St. Gallen. Corresponding author: chen.huang@unisg.ch

¶Department of Economics, City, University of London. Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin.

models with LASSO (Hautsch et al., 2015; Härdle et al., 2016; Belloni et al., 2016), limit order book network modeling via the penalized vector autoregressive approach (Härdle et al., 2018), analysis of psychology data with temporal and cross- sectional dependencies (Epskamp et al. (2016)). Another example is quantifying the spillover effects or externalities for a social network, especially when the social interactions (or the interconnectedness) is not obvious (Manresa, 2013). Besides, there are numerous applications concerning association network analysis in other fields of applied statistics, see Chapter 7 in Kolaczyk and Csárdi (2014) and Epskamp et al. (2018). In general, a step-by-step LASSO procedure is very helpful for the correlation network formation. In pursuing a highly structural approach, one certainly favors a simple set of regressions that allows multiple insights on the statistical structure of the data. Therefore, a sequence of regressions with LASSO is a natural path to take. Especially in cases of reduced forms of simultaneous equation models and structural vector autoregressive (VAR) models, one can attain valuable pre-information on the core structure by running a set of simple regressions with LASSO shrinkage.

A first important question arising in this framework is how to decide on a unified level of penalty. In this article we advocate an approach to selecting the overall level of the tuning parameter in a system of equations after performing a set of single step regressions with shrinkage. A feasible (block) bootstrap procedure is developed and the consistency of parameter estimation is studied. In addition, we provide a uniform near-oracle bound for the joint estimators. The proposed technique is applicable to ultra-high dimensional systems of regression equations with high-dimensional regressors.

A second crucial issue is to establish simultaneous inference on parameters, which is an important question regarding network topology inference.For example, in a large-scale linear factor pricing model, it is of great interest to check the significance of the intercepts of cross sectional regressions (connected with zero pricing errors), e.g. Pesaran and Yamagata (2017). Our approach is an alternative testing solution compared to the Wald test statistics proposed therein. To achieve the goal of simultaneous inference, we develop a uniform robust post-selection or post-regularization inference procedure for time series data. This method is generated from a uniform Bahadur representation of de-biased instrumental variable estimators. In particular, we need to establish maximal inequalities for empirical processes for a general Huber's Z-estimation. Note that the commonly used technique for independent data, such as the symmetrization technique, is not directly applicable in the dependent data case, see Chapter 11.6 of Kosorok (2008) for a related overview.

Our contribution lies in three aspects. First, we select the penalty level by controlling the aggregated errors in a system of high-dimensional sparse regressions, and we establish the bounds on the estimated coefficients. Furthermore, we show the implication of the restricted eigenvalue (RE) condition at a population level. Secondly, an easily implemented algorithm for effective estimation and inference is proposed. In fact, the offered estimation scheme allows us to make local and global inference on any set of parameters of interest. Thirdly, we run numerical experiments to illustrate good performance of our joint penalty relative to the single equation estimation, and we show the finite sample improvement of our multiplier block bootstrap procedure on the parameter inference. Finally, an application of textual sentiment

spillover effects on the stock returns in a financial market is presented.

In the literature, the fundamental results on achieving near oracle rate for penalized ℓ_1-norm estimators are developed by Bickel et al. (2009). There are many related articles on deriving near-oracle bounds using the ℓ_1-norm penalization function for the i.i.d. case, such as Belloni et al. (2011); Belloni and Chernozhukov (2013). There are also many extensions to the LASSO estimation with dependent data. For example, Kock and Callot (2015) consider the high-dimensional near-oracle inequalities in large vector autoregressive models. However, the majority of the literature imposes a sub-Gaussian assumption on the error distribution; this is rather restrictive and excludes heavy tail distributions. For dependent data, Wu and Wu (2016) discuss the possibility of relaxing the sub-Gaussian assumption by generalizing Nagaev-type inequalities allowing for only moment assumptions. For the case of LASSO the analysis assumes the fixed design, which rules out the most important applications mentioned earlier in the introduction.

Theoretically, the LASSO tuning parameter selection requires characterizing the asymptotic distribution of the maximum of a high dimensional random vector. Chernozhukov et al. (2013a) develop a Gaussian approximation for the maximum of a sum of high-dimensional random vectors, which is in fact the basic tool for modern high-dimensional estimation. Here it is applied to the LASSO inference. Moreover, Chernozhukov et al. (2013b) deliver results for the case of β-mixing processes. Although it is quite common to assume a mixing condition which is at base a concept yielding asymptotic independence, it is not in general easy to verify the condition for a particular process, and some simple linear processes can be excluded from the strong mixing class, Andrews (1984). With an easily accessible dependency concept, Zhang and Wu (2017a) derive Gaussian approximation results for a wide class of stationary processes. Note that the dependence measure is linked to martingale decompositions and is therefore readily connected with a pool of results on tail probabilities, moment inequalities and central limit theorems of martingale theory. Our results are built on the above-mentioned theoretical works and we extend them substantially to fit into the estimation in a system of regression equations. In particular, our LASSO estimation is with random design for dependent data; therefore, we need to deal with the population implications of the Restricted Eigenvalue (RE) condition. Moreover, we show the interaction between the tail assumption and the dimensionality of the covariates in our theoretical results.

In the meantime, the issue of simultaneous inference is challenging and has motivated a series of research articles. For the case of i.i.d. data, Belloni et al. (2011, 2014), Zhang and Zhang (2014), Javanmard and Montanari (2014), van de Geer et al. (2014), Neykov et al. (2015), Chernozhukov et al. (2016), Zhu and Bradic (2018), among others, develop confidence intervals of low-dimensional variables in high-dimensional models with various forms of de-biased/orthogonalization methods. Still in the case of i.i.d. data, Belloni et al. (2015b) establish a uniform post-selection inference for the target parameters defined via de-biased Huber's Z-estimators when the dimension of the parameters of interest is potentially larger than the sample size, where they employ the multiplier bootstrap to the estimated residuals. Wild and residual bootstrap-assisted approaches are also studied in Dezeure et al. (2017); Zhang and Cheng (2017) for the case of mean regression. We pick up the line of the inference analysis of Belloni et al.

(2015b) and employ it in a temporal and cross-sectional dependence framework, thus making it applicable to a rich class of high-dimensional time series. The core proof strategy is different, as it is well known that the technique for handling the suprema of empirical processes indexed by functional classes with dependent data is not the same as in i.i.d. cases. For instance, the key Bahadur representation in Belloni et al. (2015b) applies maximal inequalities derived in Chernozhukov et al. (2014) for i.i.d. random variables, while we derive the key concentration inequalities based on a martingale approximation method.

Our proposed estimation framework is complement to the literature on model selection for Gaussian Graphical model (GGM), see e.g. Yuan and Lin (2007), which has a wide spectrum of applications in statistics. A GGM can be connected with LASSO regression for estimating sparse correlation networks, and therefore is equivalent to our context with a partial correlation network, Meinshausen et al. (2006). In particular, we may find an equation-by-equation relationship to the GGM, and we acknowledge that a similar framework with spatial temporal dependence can be developed. In addition, there is a big literature on social network analysis, which embeds the network information into a dynamic model in advance, see for example Zhu et al. (2017, 2019); Chen et al. (2019); Huang et al. (2016). Relatively, our approach is less structural as we treat the network structure to be unknown and uncover it using LASSO.

The following notations are adopted throughout this paper. For a vector $v = (v_1, \ldots, v_p)^\top$, let $|v|_\infty \stackrel{\text{def}}{=} \max_{1 \leqslant j \leqslant p} |v_j|$ and $|v|_s \stackrel{\text{def}}{=} (\sum_{j=1}^p |v_j|^s)^{1/s}$, $s \geqslant 1$. For a random variable X, let $\|X\|_q \stackrel{\text{def}}{=} (\mathsf{E}\,|X|^q)^{1/q}$, $q > 0$. For any function on a measurable space $g : \mathcal{W} \to \mathbb{R}$, $\mathsf{E}_n(g) \stackrel{\text{def}}{=} n^{-1} \sum_{t=1}^n \{g(\omega_t)\}$ and $G_n(g) \stackrel{\text{def}}{=} n^{-1/2} \sum_{t=1}^n [g(\omega_t) - \mathsf{E}\{g(\omega_t)\}]$. Given two sequences of positive numbers x_n and y_n, write $x_n \lesssim y_n$ if there exists constant $C > 0$ such that $x_n / y_n \leqslant C$. For any finitely discrete measure $\mathcal{Q}$ on a measurable space, let $\mathcal{L}^q(\mathcal{Q})$ denote the space of all measurable functions $f : Z \to \mathbb{R}$ such that $\|f\|_{\mathcal{Q},q} \stackrel{\text{def}}{=} (\mathcal{Q}|f|^q)^{1/q} < \infty$, where $\mathcal{Q}f \stackrel{\text{def}}{=} \int f d\mathcal{Q}$. For a class of measurable functions $\mathcal{F}$, the ϵ-covering number with respect to the $\mathcal{L}^q(\mathcal{Q})$-semimetric is denoted as $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{\mathcal{Q},q})$, and let $\text{ent}(\epsilon, \mathcal{F}) = \log \sup_{\mathcal{Q}} \mathcal{N}(\epsilon \|\bar{F}\|_{\mathcal{Q},q}, \mathcal{F}, \|\cdot\|_{\mathcal{Q},q})$ with $\bar{F} = \sup_{f \in \mathcal{F}} |f|$ (the envelope) denote the uniform entropy number. It should be noted that we suppress the notation of the outer expectation E^* to E and outer probability P^* to P when measurability issues are encountered. Details may be found in the Chapter 1 of Van Der Vaart and Wellner (1996).

The rest of the article is organized as follows. Section 2 shows the system model with a few examples. Section 3 introduces the sparsity method for effective prediction and provides an algorithm for the joint penalty level of LASSO via bootstrap. In Section 4 we propose approaches to implementing individual and simultaneous inference on the coefficients. Main theorems are listed in Section 5. In Section 6 and 7 we deliver the simulation studies and an empirical application on textual sentiment spillover effects. The technical proofs and other details are given in the supplementary materials. The codes to implement the algorithms are publicly accessible via the website www.quantlet.de.

2 The System Model

In this section, we present a general framework which covers many applications in statistics. Consider the system of regression equations (SRE):

$$Y_{j,t} = X_{j,t}^\top \beta_j^0 + \varepsilon_{j,t}, \quad \mathsf{E}\, \varepsilon_{j,t} X_{j,t} = 0, \quad j = 1, ..., J, \quad t = 1, \ldots, n,$$

where $X_{j,t} = (X_{jk,t})_{k=1}^{K_j}$. Without loss of generality, we assume the dimension of the covariates is identical among all equations thereafter, namely $K_j = \dim(X_{j,t}) \equiv K$, for $j = 1, \ldots, J$. We allow the dimension K of $X_{j,t}$ and the number of equations, J to be large, potentially larger than n, which creates an interplay with the tail assumptions on the error processes $\varepsilon_{j,t}$. Both spatial and temporal dependency are allowed and we will obtain results on prediction and inference.

The SRE framework is a system of regression equations, which includes the following important special cases.

Example 1 (**Many Regression Models**)**.** Suppose that we are interested in estimating the predictive models for the response variables $U_{m,t}$:

$$U_{m,t} = X_t^\top \gamma_m^0 + \varepsilon_{m,t}, \quad X_t \in \mathbb{R}^K, \quad \mathsf{E}\, \varepsilon_{m,t} X_t = 0, \quad m = 1, \ldots, M,$$

with auxiliary regressions to model predictive relations between covariates:

$$X_{k,t} = X_{-k,t}^\top \delta_k^0 + \nu_{k,t}, \quad \mathsf{E}\, \nu_{k,t} X_{-k,t} = 0, \quad k = 1, \ldots, K,$$

where $X_{-k,t} = (X_{\ell,t})_{\ell \neq k} \in \mathbb{R}^{K-1}$, and δ_k^0 is defined by the OLS estimator in population, namely $\arg\min_{\delta_k} \frac{1}{n} \sum_{t=1}^n \mathsf{E}(X_{k,t} - X_{-k,t}^\top \delta_k)^2$. This is a special SRE model with

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (U_{j,t}, X_t, \varepsilon_{j,t}, \gamma_j^0), \quad j = 1, \ldots, M,$$

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (X_{(j-M),t}, X_{-(j-M),t}, \nu_{(j-M),t}, \delta_{(j-M)}^0), \quad j = M+1, \ldots, J = M+K.$$

It can be seen that we only put contemporaneous exogeneity conditions for X_t. It is worth mentioning that this SRE case is closely related to the semiparametric estimation framework studied in Section 2.4 in Belloni et al. (2015b). Here, the understanding of the predictive relations between covariates is important for constructing joint confidence intervals for the entire parameter vector $\{(\gamma_{mk}^0)_{k=1}^K\}_{m=1}^M$ in the main regression equations. Indeed, the construction relies on the semi-parametrically efficient point estimators obtained from the empirical analog of the following orthogonalized moment equation:

$$\mathsf{E}[(U_{mk,t}^0 - X_{k,t}\gamma_{mk}^0)\nu_{k,t}] = 0, \quad k = 1, \ldots, K, \quad m = 1, \ldots, M, \tag{2.1}$$

where $U_{mk,t}^0 = U_{m,t} - X_{-k,t}^\top \gamma_{m(-k)}^0$ is the response variable minus the part explained by the covariates other than k. Note that the empirical analog would have all unknown nuisance parameters replaced by the estimators.

Example 2 (**Simultaneous Equation Systems (SES)**)**.** Suppose there are many regression

equations in the following form:

$$U_{m,t} = U_{-m,t}^\top \delta_m^0 + X_t^\top \gamma_m^0 + \varepsilon_{m,t}, \quad m = 1, \ldots, M.$$

Move all the endogenous variables to the left-hand side and rewrite the model in the vector form

$$\mathbf{D} U_t = \mathbf{\Gamma} X_t + \varepsilon_t,$$

which is also called the structural form of the model. Suppose that D is invertible. Then the corresponding reduced form is given by

$$U_t = \mathbf{B} X_t + \nu_t, \quad \mathsf{E}\, \nu_{m,t} X_t = 0, \quad m = 1, \ldots, M, \tag{2.2}$$

with $\mathbf{B} = \mathbf{D}^{-1}\mathbf{\Gamma}$ and $\nu_t = \mathbf{D}^{-1}\varepsilon_t$. In this case the $Y_{j,t}$'s and $X_{j,t}$'s in SRE have no overlapping variables. A high-dimensional SES can be considered as a special case of SRE with

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (U_{j,t}, X_t, \nu_{j,t}, \mathbf{B}_{j\cdot}^\top), \quad j = 1, \ldots, M.$$

Example 3 (**Large Vector Autoregression Models**)**.** In the case where the covariates involve lagged variables of the response, SRE can be written as a large vector autoregression model. For example, the VAR(p) model,

$$U_t = \sum_{\ell=1}^{p} \mathbf{B}^\ell U_{t-\ell} + \varepsilon_t, \quad \mathsf{E}\, \varepsilon_{m,t} U_{t-\ell} = 0, \quad m = 1, \ldots, M, \tag{2.3}$$

where $U_t = (U_{1,t}, U_{2,t}, \ldots, U_{M,t})^\top$, and ε_t is an M-dimensional white noise or innovation process; see e.g. Chapter 2.1 in Lütkepohl (2005). It is a special SRE case again with

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (U_{j,t}, (U_{t-1}^\top, \ldots, U_{t-p}^\top)^\top, \varepsilon_{j,t}, (\mathbf{B}_{j\cdot}^1, \ldots, \mathbf{B}_{j\cdot}^p)^\top), \quad j = 1, \ldots, M.$$

Such dynamics are of interest in biology to understand dynamic gene expression network association using micro array data, see for example Opgen-Rhein and Strimmer (2007); Ramirez et al. (2017); Dimitrakopoulou et al. (2011). It is understood that a crucial feature for many gene networks is their inherent sparsity. The issue of the number of variables involved is potentially larger than the sample size can be addressed by LASSO. Our methodology can help to analyze a gene interaction correlation network in a high dimensional regression scheme. In particular, suppose that each vertex represents a gene j collected at time point t with $U_{j,t}$ as its gene expression and an edge connects two genes if they are correlated.

We refer to Section C.1 in the supplementary materials for more practical examples.

3 Effective Prediction Using Sparsity Method

In this section, we present our model setup and the LASSO estimation algorithm, including the joint penalty selection procedure.

3.1 Sparsity in SRE

The general SRE structure makes it possible to predict $Y_{j,t}$ using $X_{j,t}$ effectively. Note that the dimension of $X_{j,t}$ is large, potentially larger than n. Without loss of generality we assume exact sparsity of β_j^0 throughout the paper:

$$s_j = |\beta_j^0|_0 \leqslant s = o(n), \quad j = 1, \dots J. \tag{3.1}$$

Comment 3.1. It is now well understood that sparsity can be easily extended to approximate sparsity, in which the sorted absolute values of coefficients decrease fast to zero. To be more specific, when β_{jk}^0 is not sparse, we shall define an intermediary optimal value for our true coefficients, i.e. β_{jk}^*. Let $LC_p \stackrel{\text{def}}{=} \min_{|\beta_j|_0 \leqslant p} [\mathsf{E}_n\{X_{j,t}^\top(\beta_j - \beta_j^0)\}^2]^{1/2}$, additionally with proper conditions on the design matrix, the optimal sparsity level is given by $s_j^* = \min_{0 \leqslant p \leqslant (K \wedge n)} LC_p^2 + (\max_{1 \leqslant k \leqslant K} \Psi_{jk}^2)p/n$, where Ψ_{jk}^2 is the long run variance of $\frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_{j,t} X_{jk,t}$. Then the oracle β_{jk}^* is defined to be $\arg \min_{|\beta_j|_0 \leqslant s_j^*} \mathsf{E}_n\{X_{j,t}^\top(\beta_j - \beta_j^0)\}^2$. Thus an additional term involving $LC_{s_j^*}$ will appear in the bound in case of the true signal β_{jk}^0 is not sparse. With approximate sparsity we mean that the true signal is not sparse but nevertheless can be approximated by an exact sparsity set-up well, namely $|\beta_{jk}^0| \leqslant Ak^{-\gamma}$ (ranked in descending order), where $\gamma > 0.5$, and by taking $s_j^* \propto n^{1/(2\gamma)}$ the goal would be achieved.

For this situation one employs an ℓ_1-penalized estimator of β_j^0 of the form:

$$\widehat{\beta}_j = \arg \min_{\beta \in \mathbb{R}^K} \frac{1}{n} \sum_{t=1}^n (Y_{j,t} - X_{j,t}^\top \beta)^2 + \frac{\lambda}{n} \sum_{k=1}^K |\beta_{jk}| \Psi_{jk}, \tag{3.2}$$

where λ is the joint "optimal" penalty level and Ψ_{jk}'s are penalty loadings, which are defined below in (3.3).

A first aim is to obtain performance bounds with respect to the prediction norm:

$$|\widehat{\beta}_j - \beta_j^0|_{j,pr} \stackrel{\text{def}}{=} \left[\frac{1}{n} \sum_{t=1}^n \{X_{j,t}^\top(\widehat{\beta}_j - \beta_j^0)\}^2 \right]^{1/2},$$

where the outside j indicates to use the covariates in the jth equation $X_{j,t}$ in computing the prediction norm, and the Euclidean norm:

$$|\widehat{\beta}_j - \beta_j^0|_2 \stackrel{\text{def}}{=} \{ \sum_{k=1}^K (\widehat{\beta}_{jk} - \beta_{jk}^0)^2 \}^{1/2}.$$

To achieve good performance bounds, we first consider "ideal" choices of the penalty level and the penalty loadings. Let

$$S_{jk} = \frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_{j,t} X_{jk,t},$$

where for a moment we assume to be able to observe $\varepsilon_{j,t} = Y_{j,t} - X_{j,t}^\top \beta_j^0$. In practice one obtains

an approximation by stepwise LASSO. Set

$$\Psi_{jk} \stackrel{\text{def}}{=} \sqrt{\text{avar}(S_{jk})}, \tag{3.3}$$

$$\lambda^0(1-\alpha) \stackrel{\text{def}}{=} (1-\alpha) - \text{quantile of } 2c\sqrt{n} \max_{1\leqslant j\leqslant J, 1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}|, \tag{3.4}$$

where $c > 1$, e.g., $c = 1.1$, and $1-\alpha$ is a confidence level, e.g. $\alpha = 0.1$, where the long run variance is denoted by avar.

Theoretically, we can characterize the rate of $\lambda^0(1-\alpha)$ by the tail probability of S_{jk}, see Theorem 5.1, also via Gaussian Approximation as in corollary 5.4. To calculate $\lambda^0(1-\alpha)$ from data, we can also use a Gaussian approximation based on:

$$Q(1-\alpha) \stackrel{\text{def}}{=} (1-\alpha) - \text{quantile of } 2c\sqrt{n} \max_{1\leqslant j\leqslant J, 1\leqslant k\leqslant K} |Z_{jk}/\Psi_{jk}|,$$

where $\{Z_{jk}\}$ are multivariate Gaussian centered random variables with the same long run covariance structure as $\{S_{jk}\}$. Alternatively, we can employ a multiplier bootstrap procedure to estimate IC empirically to achieve a better finite sample performance; see for example Chernozhukov et al. (2013a). In case of dependent observations over time, it is understood that data cannot be resampled directly as in the the i.i.d. case, as the dependency structure of the underlying processes will be lost. A usual solution to this problem is to consider a block bootstrap procedure, where the data are grouped into blocks, resampled and concatenated. In particular, we will adopt an estimate of IC by a multiplier block bootstrap procedure. The theoretical properties of LASSO and the tuning parameter choices are presented in Section 5.1-5.4.

3.2 Multiplier Bootstrap for the Joint Penalty Level

In this subsection, we introduce an algorithm to approximate the joint penalty level via a block multiplier bootstrap procedure, which is particularly nonoverlapping block bootstrap (NBB). Consider the system of equations with dependent data:

$$Y_{j,t} = X_{j,t}^\top \beta_j^0 + \varepsilon_{j,t}, \quad \mathsf{E}\, \varepsilon_{j,t} X_{j,t} = 0, \quad j = 1, ..., J, \quad t = 1, \ldots, n, \tag{3.5}$$

S1 Run the initial ℓ_1-penalized regression equation by equation, i.e. for the jth equation,

$$\widetilde{\beta}_j = \arg \min_{\beta \in \mathbb{R}^K} \frac{1}{n} \sum_{t=1}^n (Y_{j,t} - X_{j,t}^\top \beta)^2 + \frac{\lambda_j}{n} \sum_{k=1}^{K_j} |\beta_{jk}| \Psi_{jk}, \tag{3.6}$$

where λ_j are the penalty levels and Ψ_{jk} are the penalty loadings. For instance, we can take the X-independence choice using Gaussian approximation (in the heteroscedasticity case): $2c'\sqrt{n}\Phi^{-1}\{1-\alpha'/(2K)\}$ for λ_j, where $\Phi(\cdot)$ denotes the cdf of $\mathrm{N}(0,1)$, $\alpha' = 0.1$, $c' = 0.5$, and choose $\sqrt{\text{lvar}(X_{jk,t}\breve{\varepsilon}_{j,t})}$ for the penalty loadings, where $\breve{\varepsilon}_{j,t}$ are preliminary estimated errors and $\text{lvar}(X_{jk,t}\breve{\varepsilon}_{j,t})$ is an estimate of the long-run variance

$\sum_{\ell=-\infty}^{\infty} \mathsf{E}(X_{jk,t}\breve{\varepsilon}_{j,t}X_{jk,(t-\ell)}\breve{\varepsilon}_{j,(t-\ell)})$, e.g. the Newey-West estimator is given by

$$\sum_{\ell=-p_n}^{p_n} k(\ell/p_n)\,\mathrm{cov}(X_{jk,t}\breve{\varepsilon}_{j,t}, X_{jk,(t-\ell)}\breve{\varepsilon}_{j,(t-\ell)}),$$

with $k(z) = (1-|z|)\mathbf{1}(|z| \leqslant 1)$. We note that the X-independent penalty (using Gaussian approximation) is more conservative, as the correlations among regressors can be adapted in the X-dependent case (using a multiplier bootstrap) with a less aggressive penalty level.

S2 Obtain the residuals for each equation by $\widetilde{\varepsilon}_{j,t} = Y_{j,t} - X_{j,t}^{\top}\widetilde{\beta}_j$, and compute $\Psi_{jk} = \sqrt{\mathrm{lvar}(X_{jk,t}\widetilde{\varepsilon}_{j,t})}$.

S3 Divide $\{\tilde{\varepsilon}_{j,t}\}$ into l_n blocks containing the same number of observations b_n, $n = b_n l_n$, where $b_n, l_n \in \mathbb{Z}$. Then choose $\lambda = 2c\sqrt{n}q^{[B]}_{(1-\alpha)}$, where $q^{[B]}_{(1-\alpha)}$ is the $(1-\alpha)$ quantile of $\max_{1\leqslant j\leqslant J, 1\leqslant k\leqslant K}|Z^{[B]}_{jk}/\Psi_{jk}|$, and $Z^{[B]}_{jk}$ are defined as

$$Z^{[B]}_{jk} = \frac{1}{\sqrt{n}}\sum_{i=1}^{l_n} e_{j,i} \sum_{l=(i-1)b_n+1}^{ib_n} \tilde{\varepsilon}_{j,l}X_{jk,l}, \tag{3.7}$$

$e_{j,i}$ are i.i.d. $\mathrm{N}(0,1)$ random variables independent of the data.

The bootstrap consistency regarding $Z^{[B]}_{jk}$ is proved in Theorem 5.3.

Comment 3.2 (Block bootstrap procedures)**.** (i) Concerning the determination of b_n, we shall report the prediction norm with several block sizes b_n and select the one with the best prediction performance in the simulation study. In addition, if it is the case that n cannot be divided by b_n with no remainder, one can simply take $l_n = \lfloor n/b_n \rfloor$ and drop the remaining observations.

(ii) Other forms of multiplier bootstrap with any random multipliers centered around 0 can also be considered.

(iii) Alternative block bootstrap procedures can be adopted, such as the circular bootstrap and the stationary bootstrap among others; see for example Lahiri et al. (1999) for an overview.

4 Valid Inference on the Coefficients

With a reasonable fitting of LASSO on hand, we can proceed to investigate the issue of simultaneous inference. This section focuses on SRE of Example 2. We allow the covariates in each equation to be different.

The basic idea to facilitate inference is to formulate the estimation in a semi-parametric framework. With partialing out the effect of the nonparametric coefficient(s), we can achieve the desired estimation accuracy of the parametric component of interest. This trick is referred to as "Neyman orthogonalization". Notably, the procedure is equivalent to the well known desparsification procedure in the mean square loss case, which is developed for the inference on the

estimated zero coefficients by LASSO. It thus serves the same purpose of generating a (robust) de-sparsified estimation for LASSO inference.

We list three algorithms to estimate β_{jk}^0. Algorithm 1 is easy to implement and algorithm 2 is tailored to the cases of heavy-tailed distribution of the error term, as Least Absolute Deviation (LAD) regression is well known to be robust against outliers. Algorithm 3 considers a double selection procedure aimed at remedying the bias due to omitted variables by one step selection, while also accounting for the cases of heteroscedastic errors.

Algorithm 1: LS-based algorithm

S1 Consider $Y_{j,t} = X_{jk,t}\beta_{jk}^0 + X_{j(-k),t}^\top \beta_{j(-k)}^0 + \varepsilon_{j,t}$, run (post) LS LASSO procedure (for each j), and keep the quantity $X_{j(-k),t}^\top \widehat{\beta}_{j(-k)}^{[1]}$ for each k.

S2 Run LASSO (for each j,k) by regressing $X_{jk,t} = X_{j(-k),t}^\top \gamma_{j(-k)}^0 + v_{jk,t}$, and keep the residuals as $\widehat{v}_{jk,t} = X_{jk,t} - X_{j(-k),t}^\top \widehat{\gamma}_{j(-k)}$.

S3 Run LS IV regression of $Y_{j,t} - X_{j(-k),t}^\top \widehat{\beta}_{j(-k)}^{[1]}$ on $X_{jk,t}$ using $\widehat{v}_{jk,t}$ as an instrument variable, attaining the final estimator $\widehat{\beta}_{jk}^{[2]}$.

Algorithm 2: LAD-based algorithm

S1 and S2 are the same as Algorithm 1.

S3′ Run LAD IV regression of $Y_{j,t} - X_{j(-k),t}^\top \widehat{\beta}_{j(-k)}^{[1]}$ on $X_{jk,t}$ using $\widehat{v}_{jk,t}$ as an instrument variable, attaining the final estimator $\widehat{\beta}_{jk}^{[2]}$. We refer to Belloni et al. (2015b); Chernozhukov and Hansen (2008) for more details about how to achieve the estimator in this step.

The theoretical properties of the estimators $\widehat{\beta}_{j(-k)}^{[1]}$ and $\widehat{\gamma}_{j(-k)}$ in S1 and S2 are provided in Corollary 5.1 or 5.4 (see Corollary A.1 or A.4 in the supplementary correspondingly if the joint penalty over equations is employed), and Theorem A.4 for post LASSO, respectively. The uniform Bahadur representation and the Central Limit Theorem of the estimator $\widehat{\beta}_{jk}^{[2]}$ in S3 or S3′ are established in Theorem 5.4 and 5.5.

Comment 4.1. Our algorithms follow patterns discussed in Belloni et al. (2015b,a) in the i.i.d. settings. The IV estimator obtained in S3 of Algorithm 1 reduced to the de-biased LASSO estimator (Zhang and Zhang, 2014; van de Geer et al., 2014) and is also first-order equivalent to the double LASSO method in Belloni et al. (2011, 2014). In particular, the estimator under LS IV regression (2-step least square regression) is given by

$$\begin{aligned}\widehat{\beta}_{jk}^{[2]} &= (\widehat{v}_{jk}^\top X_{jk})^{-1}\widehat{v}_{jk}^\top (Y_j - X_{j(-k)}^\top \widehat{\beta}_{j(-k)}^{[1]}) \\ &= (\widehat{v}_{jk}^\top X_{jk})^{-1}\widehat{v}_{jk}^\top Y_j - \sum_{m\neq k} \frac{\widehat{v}_{jk}^\top X_{jm}}{\widehat{v}_{jk}^\top X_{jk}} \widehat{\beta}_{jm}^{[1]}. \end{aligned} \tag{4.1}$$

The second line in (4.1) is exactly the same as the de-biased or de-sparsified LASSO estimator given in Eq. (5) in Zhang and Zhang (2014) or Eq. (5) in van de Geer et al. (2014). As remarked in Belloni et al. (2015b,a), one can alternatively implement an algorithm via double selection as in Belloni et al. (2011, 2014). In particular, heteroscedastic LASSO is employed in S2″ and

the IV regression is replaced by a either LASSO or LAD regression on the target variable and all covariates selected in the first two steps. □

Algorithm 3: Double selection-based algorithm

S1″ Run LS LASSO (for each j) of $Y_{j,t}$ on $X_{j,t}$:

$$\widehat{\beta}_j^{[1]} = \arg\min_{\beta} \frac{1}{n}\sum_{t=1}^{n}(Y_{j,t} - X_{j,t}^{\top}\beta)^2 + \frac{\lambda}{n}|\widehat{\Psi}_j\beta|_1.$$

S2″ Run Heteroscedastic LASSO (for each j,k) of $X_{jk,t}$ on $X_{j(-k),t}$:

$$\widehat{\gamma}_{j(-k)} = \arg\min_{\gamma} \frac{1}{n}\sum_{t=1}^{n}(X_{jk,t} - X_{j(-k),t}^{\top}\gamma)^2 + \frac{\lambda'}{n}|\widehat{\Gamma}_j\gamma|_1,$$

where penalty loadings $\widehat{\Gamma}_j$ can be initialized as $\sqrt{\text{lvar}\{X_{j\ell,t}(X_{jk,t} - \frac{1}{n}\sum_{t=1}^{n} X_{jk,t})\}}$ and then refined by $\sqrt{\text{lvar}(X_{j\ell,t}\widehat{v}_{jk,t})}$, for $\ell \neq k$, and $\widehat{v}_{jk,t} = X_{jk,t} - X_{j(-k),t}^{\top}\widehat{\gamma}_{j(-k)}$ can be obtained by using the initial ones.

S3″ Run LS regression of $Y_{j,t}$ on $X_{jk,t}$ and the covariates selected in S1″ and S2″:

$$\widehat{\beta}_j^{[2]} = \arg\min_{\beta}\{\frac{1}{n}\sum_{t=1}^{n}(Y_{j,t} - X_{j,t}^{\top}\beta)^2 : \text{supp}(\beta_{-k}) \subseteq \text{supp}(\widehat{\beta}_{j(-k)}^{[1]}) \cup \text{supp}(\widehat{\gamma}_{j(-k)})\}.$$

S3‴ Run LAD regression of $Y_{j,t}$ on $X_{jk,t}$ and the covariates selected in S1″ and S2″:

$$\widehat{\beta}_j^{[2]} = \arg\min_{\beta}\{\frac{1}{n}\sum_{t=1}^{n}|Y_{j,t} - X_{j,t}^{\top}\beta| : \text{supp}(\beta_{-k}) \subseteq \text{supp}(\widehat{\beta}_{j(-k)}^{[1]}) \cup \text{supp}(\widehat{\gamma}_{j(-k)})\}.$$

As shown in Belloni et al. (2011) and Belloni et al. (2015a), the double selection approach in S3″ or S3‴ creates an orthogonality condition with respect to the space spanned by the covariates selected by both steps, and thus generates an orthogonal relation to any space spanned by a linear projection of the covariates, e.g. $\widehat{v}_{jk,t}$. Therefore, the inference on the parameters may still be applied as in the framework of Algorithm 1 and 2. Therefore, one may still find the theoretical properties of estimators in S1″, S2″, S3″ (S3‴) in Section 5 according to the links mentioned above.

4.1 Confidence Interval for a Single Coefficient

We discuss an inference framework developed for a single coefficient obtained from the aforementioned algorithms.

Let $\psi_{jk}(Z_{j,t}, \beta_{jk}, h_{jk})$ denote the score function, where $Z_{j,t} = (Y_{j,t}, X_{j,t}^{\top})^{\top}$, $h_{jk}(X_{j(-k),t}) = (X_{j(-k),t}^{\top}\beta_{j(-k)}, X_{j(-k),t}^{\top}\gamma_{j(-k)})^{\top}$. Consider the LAD-based case with $\psi_{jk}(Z_{j,t}, \beta_{jk}, h_{jk}) = \{1/2 - \mathbf{1}(Y_{j,t} \leqslant X_{jk,t}\beta_{jk} + X_{j(-k),t}^{\top}\beta_{j(-k)})\}v_{jk,t}$, define $\omega_{jk} \stackrel{\text{def}}{=} \mathsf{E}\{(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_{jk,t}^0)^2\} = \sum_{\ell=-(n-1)}^{n-1}(1 - \frac{|\ell|}{n})\,\text{cov}(\psi_{jk,t}^0, \psi_{jk,(t-\ell)}^0)$ with $\psi_{jk,t}^0 \stackrel{\text{def}}{=} \psi_{jk}(Z_{j,t}, \beta_{jk}^0, h_{jk}^0)$, and $\phi_{jk} \stackrel{\text{def}}{=} \frac{\partial \mathsf{E}\{\psi_{jk}(Z_{j,t},\beta,h_{jk}^0)\}}{\partial\beta}|_{\beta=\beta_{jk}^0}$.

Suppose we are interested in testing $H_0 : \beta_{jk}^0 = 0$. For this purpose we employ the uniform Bahadur representation (Theorem 5.4) to construct the confidence interval via a multiplier bootstrap procedure. In particular, the distribution of the asymptotically pivotal statistics:

$$T_{jk} = \frac{\sqrt{n}(\widehat{\beta}_{jk}^{[2]} - \beta_{jk}^0)}{\widehat{\sigma}_{jk}}, \tag{4.2}$$

is approximated via its block multiplier bootstrap counterpart:

$$T_{jk}^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{l_n} e_{j,i} \sum_{l=(i-1)b_n+1}^{ib_n} \widehat{\zeta}_{jk,l}, \tag{4.3}$$

where $\widehat{\zeta}_{jk,t}$ are pre-estimators of $\zeta_{jk,t} = -\phi_{jk}^{-1}\sigma_{jk}^{-1}\psi_{jk,t}^0$ such that $\max_{(j,k),(j',k')} |\sum_{i=1}^{l_n} \widehat{\eta}_{j'k',i}\widehat{\eta}_{jk,i} - \sum_{i=1}^{l_n} \eta_{j'k',i}\eta_{jk,i}| = o_{\mathrm{P}}(\{\log(JK)\}^{-2})$, with $\eta_{jk,i} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_{l=(i-1)b_n+1}^{ib_n} \zeta_{jk,l}$ and $\widehat{\eta}_{jk,i} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_{l=(i-1)b_n+1}^{ib_n} \widehat{\zeta}_{jk,l}$, $e_{j,i}$ are independently drawn from $\mathrm{N}(0,1)$, l_n and b_n are the numbers of blocks and block size, respectively.

Let $\widehat{\sigma}_{jk}$ be any consistent estimator of σ_{jk}. Then the confidence interval is given by

$$\mathrm{CI}_{jk}^*(\alpha) : [\widehat{\beta}_{jk}^{[2]} - \widehat{\sigma}_{jk} n^{-1/2} q_{jk}^*(1-\alpha), \widehat{\beta}_{jk}^{[2]} + \widehat{\sigma}_{jk} n^{-1/2} q_{jk}^*(1-\alpha)], \tag{4.4}$$

where $q_{jk}^*(1-\alpha)$ is the $(1-\alpha)$ quantile of the bootstrapped distribution of $|T_{jk}^*|$.

Comment 4.2 (Asymptotic Normality of $\widehat{\beta}_{jk}^{[2]}$)**.** As shown in Corollary 5.5 we have the limit distribution of $\widehat{\beta}_{jk}^{[2]}$:

$$\sigma_{jk}^{-1} n^{1/2} (\widehat{\beta}_{jk}^{[2]} - \beta_{jk}^0) \xrightarrow{\mathcal{L}} \mathrm{N}(0,1), \tag{4.5}$$

where $\sigma_{jk} = (\phi_{jk}^{-2}\omega_{jk})^{1/2}$. Therefore, the two-sided $100(1-\alpha)$ confidence interval by asymptotic normality for β_{jk}^0 is given by

$$\mathrm{CI}_{jk}(\alpha) : [\widehat{\beta}_{jk}^{[2]} - \widehat{\sigma}_{jk} n^{-1/2} \Phi^{-1}(1-\alpha/2), \widehat{\beta}_{jk}^{[2]} + \widehat{\sigma}_{jk} n^{-1/2} \Phi^{-1}(1-\alpha/2)]. \tag{4.6}$$

Comment 4.3 (Residual Multiplier Bootstrap)**.** Alternative bootstrap procedures may be considered as well, e.g. the residual multiplier bootstrap procedure:

$$\widehat{\varepsilon}_{j,t} = Y_{j,t} - X_{j,t}^\top \widehat{\beta}_j^{[1]},$$

then divide $\{\widehat{\varepsilon}_{j,t}\}$ into l_n blocks of size b_n, where $b_n l_n = n$, and for each block $i = 1, \ldots, l_n$,

$$\varepsilon_{j,t}^* = (\widehat{\varepsilon}_{j,t} - \frac{1}{n}\sum_{t=1}^{n} \widehat{\varepsilon}_{j,t}) e_{j,i}, \text{ for } t \in \{(i-1)b_n + 1, \ldots, ib_n\}.$$

Define $Y_{j,t}^* = X_{j,t}^\top \widehat{\beta}_j^{[1]} + \varepsilon_{j,t}^*$ and compute the bootstrap counterpart as

$$T_{jk}^* = \frac{\sqrt{n}(\widehat{\beta}_{jk}^* - \widehat{\beta}_{jk}^{[1]})}{\widehat{\sigma}_{jk}^*},$$

where $\widehat{\beta}^*_{jk}$ and $\widehat{\sigma}^*_{jk}$ are estimated using the bootstrap sample $\{Y^*_{j,t}, X_{j,t}\}$.

4.2 Joint Confidence Region for Simultaneous Inference

We now continue to extend the single coefficient inference to simultaneous inference on a set of coefficients. As shown in the practical examples in Section C.1, it is essential to conduct simultaneous inference on a group of parameters G. In this case, the null hypothesis is: $\mathbf{H}_0 : \beta^0_{jk} = 0$, $\forall (j,k) \in G$, and the alternative $\mathbf{H}_A : \beta^0_{jk} \neq 0$, for some $(j,k) \in G$, where the group G is a set of coefficients with cardinality $|G|$. Suppose for the j-th equation there are p_j target coefficients and the cardinality $|G| = \sum_{j=1}^{J} p_j$. This can be understood as a multiple estimation problem compared to Section 4.1. Without loss of generality, we can rearrange the order of the variables and rewrite the regression equation for each j as (consider the LAD-based model here)

$$Y_{j,t} = \sum_{l=1}^{p_j} X_{jl,t}\beta^0_{jl} + \sum_{l=p_j+1}^{K} X_{jl,t}\beta^0_{jl} + \varepsilon_{j,t}, \quad F_{\varepsilon_j}(0) = 1/2 \tag{4.7}$$

One follows the algorithms to obtain $\widehat{\beta}_{jl} (1 \leqslant l \leqslant p_j)$ for each j. Then the idea of simultaneous inference is very straightforward. We aggregate the statistics T_{jk} in (4.2) by taking the maximum and minimum over the set G. Finally, the component-wise confidence interval is constructed with the quantiles of the bootstrap statistics over all bootstrap samples.

Denote $q^*_G(1-\alpha)$ as the $(1-\alpha)$ quantile of $\max_{(j,k)\in G} |T^*_{jk}|$. A joint confidence region is then:

$$\{\beta \in \mathbb{R}^{|G|} : \max_{(j,k)\in G} T_{jk} \leqslant q^*_G(1-\alpha) \text{ and } \min_{(j,k)\in G} T_{jk} \geqslant -q^*_G(1-\alpha)\}, \tag{4.8}$$

and for each component $(j,k) \in G$, the confidence interval $\widetilde{\mathrm{CI}}^*_{jk}(\alpha)$ is given by $[\widehat{\beta}^{[2]}_{jk} - \widehat{\sigma}_{jk} n^{-1/2} q^*_G(1-\alpha), \widehat{\beta}^{[2]}_{jk} + \widehat{\sigma}_{jk} n^{-1/2} q^*_G(1-\alpha)]$. We show in Corollary 5.7 the consistency of this bootstrap confidence band for simultaneous inference. Note that when there is only one parameter in G for inference, the joint confidence region (4.8) will reduce to the single parameter confidence interval (4.4) as a special case.

5 Main Theorems

In this section, we present the theoretical foundations for the procedures given earlier. In particular, we discuss the properties of the theoretical choices of penalty level and the validity of the other two empirical choices, as well as the theoretical support for the simultaneous inference.

Throughout the whole section, we define $S_{jk} \overset{\text{def}}{=} n^{-1/2} \sum_{t=1}^{n} \varepsilon_{j,t} X_{jk,t}$, $S_{j\cdot} = (S_{jk})_{k=1}^{K}$, and $\Psi_{jk} \overset{\text{def}}{=} \sqrt{\operatorname{avar}(S_{jk})}$, which is the square root of the long-run variance of $X_{jk,t}\varepsilon_{j,t}$, namely $\{\sum_{\ell=-\infty}^{\infty} \mathsf{E}(X_{j,k,t} X_{jk,(t-\ell)} \varepsilon_{j,t} \varepsilon_{j,(t-\ell)})\}^{1/2}$. Recall that for a single equation LASSO, we select the penalty in the following ways:

a) theoretically, for each regression, λ_j is $\lambda^0_j(1-\alpha)$ (IC), i.e. the $(1-\alpha)$ quantile of $2c\sqrt{n} \max_{1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}|$ (note that this penalty takes into account the correlation among regressors and is design adaptive);

b) an empirical choice given a Gaussian approximation result is $Q_j(1-\alpha)$, which is defined to be the $(1-\alpha)$ quantile of $2c\max_{1\leqslant k\leqslant K}\sqrt{n}|Z_{jk}/\Psi_{jk}|$, where Z_{jk}'s are multivariate Gaussian centered random variables with the same long run covariance structure as S_{jk}. Alternatively, a canonical choice disregarding the correlation among regressors can be considered as $\widetilde{Q}_j(1-\alpha) \stackrel{\text{def}}{=} 2c\sqrt{n}\Phi^{-1}\{1-\alpha/(2K)\}$. We shall note that $Q_j(1-\alpha)$ is not feasible but can be estimated by simulations of Gaussian random variable Z_{jk} with estimated long run variance covariance matrix. Typically $\widetilde{Q}_j(1-\alpha)$ is more conservative than $Q_j(1-\alpha)$.

c) another empirical choice of the penalty level is $\Lambda_j(1-\alpha)$ as the $(1-\alpha)$ quantile of $2c\sqrt{n}\max_{1\leqslant k\leqslant K}|Z_{jk}^{[B]}/\widehat{\Psi}_{jk}|$ ($Z_{jk}^{[B]}$'s are defined in (3.7)), and obtainable via the multiplier block bootstrap technique.

5.1 Near Oracle Inequalities under IC

We first provide the near oracle inequalities for the single equation LASSO estimation $\tilde{\beta}_j$ obtained from (3.6) under the ideal choices (IC). For this purpose, a few assumptions and definitions are required.

(A1) For $j=1,\ldots,J, k=1,\ldots,K$, let $X_{jk,t}$ and $\varepsilon_{j,t}$ be stationary processes admitting the following representation forms $X_{jk,t}=g_{jk}(\mathcal{F}_t)=g_{jk}(\ldots,\xi_{t-1},\xi_t)$ and $\varepsilon_{j,t}=h_j(\mathcal{F}_t)=h_j(\ldots,\eta_{t-1},\eta_t)$, where ξ_t,η_t are i.i.d. random elements (innovations or shocks, allowing for overlap, see Comment 5.1) across t, $\mathcal{F}_t=(\ldots,\xi_{t-1},\eta_{t-1},\xi_t,\eta_t)$, $g_{jk}(\cdot)$ and $h_j(\cdot)$ are measurable functions (filters). $\mathsf{E}(X_{jk,t}\varepsilon_{j,t})=0$, for any $j,k\in 1,\cdots,J,1,\cdots,K$.

Definition 5.1. *Let ξ_0 be replaced by an i.i.d. copy of ξ_0^*, and $X_{jk,t}^*=g_{jk}(\ldots,\xi_0^*,\ldots,\xi_{t-1},\xi_t)$. For $q\geqslant 1$, define the functional dependence measure $\delta_{q,j,k,t}\stackrel{\text{def}}{=}\|X_{jk,t}-X_{jk,t}^*\|_q$, which measures the dependency of ξ_0 on $X_{jk,t}$. Also define $\Delta_{m,q,j,k}\stackrel{\text{def}}{=}\sum_{t=m}^{\infty}\delta_{q,j,k,t}$, which measures the cumulative effect of ξ_0 on $X_{jk,t\geqslant m}$. Moreover, we introduce the dependence adjusted norm of $X_{jk,t}$ as $\|X_{jk,\cdot}\|_{q,\varsigma}\stackrel{\text{def}}{=}\sup_{m\geqslant 0}(m+1)^{\varsigma}\Delta_{m,q,j,k}(\varsigma>0)$. Similarly, let η_0 be replaced by an i.i.d. copy of η_0^*, and $\varepsilon_{j,t}^*=h_j(\ldots,\eta_0^*,\ldots,\eta_{t-1},\eta_t)$, we define $\|\varepsilon_{j,\cdot}\|_{q,\varsigma}\stackrel{\text{def}}{=}\sup_{m\geqslant 0}(m+1)^{\varsigma}\sum_{t=m}^{\infty}\|\varepsilon_{j,t}-\varepsilon_{j,t}^*\|_q$ and $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}\stackrel{\text{def}}{=}\sup_{m\geqslant 0}(m+1)^{\varsigma}\sum_{t=m}^{\infty}\|X_{jk,t}\varepsilon_{j,t}-X_{jk,t}^*\varepsilon_{j,t}^*\|_q$.*

It should be noted that (A1) admits a wide class of processes. The largest value of ς which ensures a finite dependence adjusted norm characterizes the dependency structure of the process. The moment-based measure is directly connected with the impulse functions. A few examples for univariate time series Z_t are listed in Appendix C.2 in the supplementary materials.

(A2) Restricted eigenvalue (RE): given $\bar{c}\geqslant 1$, for $\delta\in\mathbb{R}^K$, with probability $1-o(1)$,

$$\kappa_j(\bar{c})\stackrel{\text{def}}{=}\min_{|\delta_{T_j^c}|_1\leqslant\bar{c}|\delta_{T_j}|_1,\,\delta\neq 0}\frac{\sqrt{s_j}|\delta|_{j,pr}}{|\delta_{T_j}|_1}>0,$$

where $T_j\stackrel{\text{def}}{=}\{k:\beta_{jk}^0\neq 0\}$ and $s_j=|T_j|=o(n)$, $\delta_{T_jk}=\delta_k$ if $k\in T_j$, $\delta_{T_jk}=0$ if $k\notin T_j$.

(A3) $\|\varepsilon_{j,\cdot}\|_{q,\varsigma}<\infty$ and $\|X_{jk,\cdot}\|_{q,\varsigma}<\infty$ $(q\geqslant 8)$.

Comment 5.1. We allow for overlap in the elements in ξ_t and η_t, as long as the contemporaneous exogeneity condition $\mathsf{E}(X_{jk,t}\varepsilon_{j,t}) = 0$ is satisfied. For example, consider the VAR(1) model: $Y_t = AY_{t-1} + \varepsilon_t$, with $Y_t, \varepsilon_t \in \mathbb{R}^J$, and suppose that Y_t admits the representation $Y_t = \sum_{l=0}^{\infty} A^l \varepsilon_{t-l}$ with ε_{t-l} as measurable functions of $\xi_{-\infty}, \ldots, \xi_{t-l}$. Thus $X_{jk,t} = g_{jk}(\ldots, \xi_{t-1}) = \sum_{l=0}^{\infty}[A^l]_k \varepsilon_{t-1-l}$, where $[A^l]_k$ is the kth row of the matrix A^l, $k = 1, \ldots, J$. In this case no serial correlation in the innovations ε_t's would be sufficient for $\mathsf{E}(X_{jk,t}\varepsilon_{j,t}) = 0$.

Comment 5.2. We show in Theorem B.1 (see the supplementary materials) that the RE (A2) and RSE (A5) conditions can be implied by assumptions on the corresponding population variance-covariance matrix. This illustrates the feasibility of the RE/RSE assumption.

Lemma 5.1 (Prediction Performance Bound of Single Equation LASSO)**.** *Suppose (A1) and (A2) (with $\bar{c} = \frac{c+1}{c-1}, c > 1$), under the exact sparsity assumption* (3.1) *and given the event $\lambda_j \geqslant 2c\sqrt{n} \max_{1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}|$ and another event which RE holds, then with probability $1 - o(1)$, $\tilde{\beta}_j$ obtained from* (3.6) *satisfy*

$$|\tilde{\beta}_j - \beta_j^0|_{j,pr} \leqslant (1 + 1/c)\frac{\lambda_j\sqrt{s_j}}{n\kappa_j(\bar{c})} \max_{1\leqslant k\leqslant K} \Psi_{jk}. \tag{5.1}$$

In addition, if (A2) (with $2\bar{c}$) holds, then with probability $1 - o(1)$,

$$|\tilde{\beta}_j - \beta_j^0|_1 \leqslant \frac{(1 + 2\bar{c})\sqrt{s_j}}{\kappa_j(2\bar{c})}|\tilde{\beta}_j - \beta_j^0|_{j,pr}. \tag{5.2}$$

Lemma 5.1 follows Theorem 1 of Belloni and Chernozhukov (2013). As the proof is built on inequalities and for the case of dependent data (A1) they remain unchanged, we omit the detailed proof here. To further characterize the rate of IC, we provide a tail probability for $2c\sqrt{n} \max_{1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}|$ under the moment assumption (A3). In particular, the rate depends on the dependence adjusted norm $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$.

Theorem 5.1. *Under (A1) and (A3), we have*

$$\mathrm{P}(2c\sqrt{n} \max_{1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}| \geqslant r) \leqslant C_1\varpi_n n r^{-q} \sum_{k=1}^{K} \frac{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}^q}{\Psi_{jk}^q} + C_2 \sum_{k=1}^{K} \exp\Big(\frac{-C_3 r^2 \Psi_{jk}^2}{n\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}^2}\Big), \tag{5.3}$$

where for $\varsigma > 1/2 - 1/q$ (weak dependence case), $\varpi_n = 1$; for $\varsigma < 1/2 - 1/q$ (strong dependence case), $\varpi_n = n^{q/2-1-\varsigma q}$. C_1, C_2, C_3 are constants depending on q and ς.

Comment 5.3. It can be seen in Theorem 5.1 that the rate of the dependence adjusted norm $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$ plays an important role in the tail probability for $2c\sqrt{n} \max_{1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}|$. Here we discuss the rate under some special cases.

1. **VAR(1)**: Consider the VAR(1) model given by $Y_t = AY_{t-1} + \varepsilon_t$, where $Y_t, \varepsilon_t \in \mathbb{R}^J$, and $\varepsilon_t \sim$ i.i.d. $\mathrm{N}(0, \Sigma)$. In this case $X_{jk,t} = Y_{j,t-1}$ and $K = J$. Suppose there exists a stationary representation of the model as $Y_t = \sum_{l=0}^{\infty} A^l \varepsilon_{t-l}$. Then we have $\|X_{jk,t}\varepsilon_{j,t} - X_{jk,t}^*\varepsilon_{j,t}^*\|_q = \|Y_{j,t-1}\varepsilon_{j,t} - Y_{j,t-1}^*\varepsilon_{j,t}\|_q = \|[A^{t-1}]_j(\varepsilon_0 - \varepsilon_0^*)\varepsilon_{j,t}\|_q \leqslant 2|[A^{t-1}]_j|_1\mu_q^2$, where $\mu_q \overset{\text{def}}{=}$

$\max_j \|\varepsilon_{j,t}\|_q$ and $[A^{t-1}]_j$ is the jth row of the matrix A^{t-1}. Assume $\max_j |[A^t]_j|_1 \leqslant |c|^t$ with $|c| < 1$ (a geometric decay rate). It follows that $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} = \frac{2\mu_q^2}{1-|c|} \sup_{m\geqslant 0}(m+1)^\varsigma \sum_{t=m}^\infty |c|^{t-1} \leqslant (C/|c|) \vee \{C(m^*+1)|c|^{m^*-1}\}$, where $m^* = (-\varsigma/\log|c| - 1) \vee 0$ and $C > 0$ depends on μ_q. Moreover, to justify the geometric decay rate, we consider the example of Network Autoregressive (NAR) model as in Zhu et al. (2017) with $A = \rho W$, where W is a row-normalized adjacency matrix which is pre-specified to indicate the social network connectedness and ρ is the network parameter suggesting the strength of the network effects. In that case, assuming a geometric decay rate $\max_j |[A^t]_j|_1 \leqslant |c|^t$ with $|c| < 1$ again gives similar results.

2. **Spatial MA structure in ε_t**: Consider the model $Y_{j,t} = X_{j,t}^\top \beta_j + \varepsilon_{j,t}$, with $\varepsilon_t = \rho W \varepsilon_t + \eta_t$, where W is a spatial weight matrix, η_t are i.i.d. and have finite qth moments $\mu_q^\eta \stackrel{\text{def}}{=} \max_j \|\eta_{j,t}\|_q$. For simplicity, here we assume $X_{j,t}$ and $\varepsilon_{j,t}$ are independent. Suppose there exists a stationary representation of the error process given by $\varepsilon_t = \sum_{l=0}^\infty \rho^l W^l \eta_{t-l}$. Then we have $\|X_{jk,t}\varepsilon_{j,t} - X_{jk,t}^*\varepsilon_{j,t}^*\|_q \leqslant \|(X_{jk,t} - X_{jk,t}^*)\varepsilon_{j,t}\|_q + \|X_{jk,t}(\varepsilon_{j,t} - \varepsilon_{j,t}^*)\|_q \leqslant \|X_{jk,t} - X_{jk,t}^*\|_q \|\varepsilon_{j,t}\|_q + \|X_{jk,t}\|_q \|[\rho^t W^t]_j(\eta_0 - \eta_0^*)\|_q \leqslant |[(\mathbf{I} - \rho W)^{-1}]_j|_1 \mu_q^\eta \|X_{jk,t} - X_{jk,t}^*\|_q + 2|[\rho^t W^t]_j|_1 \mu_q^\eta \|X_{jk,t}\|_q$. Assume $\max_j |[\rho^t W^t]_j|_1 \leqslant |c|^t$ with $|c| < 1$. It follows that $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \leqslant C_1 \|X_{jk,\cdot}\|_{q,\varsigma} + C_2 \sup_{m\geqslant 0}(m+1)^\varsigma \sum_{t=m}^\infty |c|^t \leqslant C_1 \|X_{jk,\cdot}\|_{q,\varsigma} + C_3(m^*+1)|c|^{m^*-1}$, where $m^* = (-\varsigma/\log|c| - 1) \vee 0$ and $C_1, C_2, C_3 > 0$ depend on μ_q^η and $\|X_{jk,t}\|_q$.

3. **General linear processes**: To study more general spatial and temporal dependency, consider the model $Y_{j,t} = X_{j,t}^\top \beta_j + \varepsilon_{j,t}$, with $\varepsilon_t = \sum_{l=0}^\infty A^l \eta_{t-l}$. Again η_t are i.i.d. and have finite qth moments $\mu_q^\eta \stackrel{\text{def}}{=} \max_j \|\eta_{j,t}\|_q$. If all the A^l are diagonal matrices, there is just temporal dependence, and if $A^l = 0$ for $l \geqslant 1$ there exists only spatial dependence. Let $a_{jk}^t \stackrel{\text{def}}{=} [A^t]_{jk}$ be the element on the jth row and kth column of A^t. Assume $\sum_{t=0}^\infty \sum_k |a_{jk}^t| < \infty$, $X_{j,t}$ and $\varepsilon_{j,t}$ to be independent. We have $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \leqslant C_1 \|X_{jk,\cdot}\|_{q,\varsigma} + C_2 \sup_{m\geqslant 0}(m+1)^\varsigma \sum_{t=m}^\infty \sum_k |a_{jk}^t|$, where $C_1, C_2 > 0$ depend on μ_q^η and $\|X_{jk,t}\|_q$. Moreover, we have $\|\max_{jk}(X_{jk,\cdot}\varepsilon_{j,\cdot})\|_{q,\varsigma} \leqslant \|\max_{jk} X_{jk,\cdot}\|_{q,\varsigma} \|\max_j \varepsilon_{j,\cdot}\|_{q,\varsigma}$, and particularly $\||\varepsilon_t|_\infty\|_q \leqslant \|\max_j \sum_k a_{jk}^t(\eta_{k,0} - \eta_{k,0}^*)\|_q \lesssim q\|\max_k \max_j a_{jk}^t(\eta_{k,0} - \eta_{k,0}^*)\|_q + \sqrt{q\log J}\{\sum_k \max_j (a_{jk}^t)^2(\mu_2^\eta)^2\}^{1/2} \lesssim q\sum_k \max_j |a_{jk}^t| \mu_q^\eta \vee \sqrt{q\log J}\{\sum_k \max_j (a_{jk}^t)^2\}^{1/2}\mu_2^\eta$, where the Rosenthal-Burkholder inequality is applied. Suppose that $\sum_{t=m}^\infty (\sum_k \max_j |a_{jk}^t|) \lesssim J(m \vee 1)^{-c}$, for some constant $c > 0$. If $\varsigma < c$, we have $\|\max_j \varepsilon_{j,\cdot}\|_{q,\varsigma} \leqslant C_3 \sup_{m\geqslant 1}(m+1)^\varsigma (m \vee 1)^{-c} J\sqrt{\log J} \leqslant C_3 \sup_{m\geqslant 1}(m+1)^{\varsigma - c} J\sqrt{\log J}$, where $C_3 > 0$ depends on μ_q^η.

To summarize, if the qth moments are bounded by constant, the dependence adjusted norm $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$ is also bounded in the first two examples where a geometric decay rate on the coefficients is assumed; while in the case of general linear processes, it would depend on the rate of $\sum_{t=0}^\infty \sum_k |a_{jk}^t|$. In particular, suppose $\sum_{t=m}^\infty \sum_k |a_{jk}^t| \lesssim (m \vee 1)^{-c}$ for $c > 0$. If $c > \varsigma$, $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$ is bounded (assume $\|X_{jk,\cdot}\|_{q,\varsigma}$ is bounded).

Under the choice (IC) $\lambda_j^0(1-\alpha)$ is given by the $(1-\alpha)$ quantile of $2c\sqrt{n} \max_{1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}|$, combining the results of Lemma 5.1 and Theorem 5.1 we can get the bounds for $\lambda_j^0(1-\alpha)$ and further obtain the oracle inequalities as in Corollary 5.1.

Corollary 5.1 (Bounds for $\lambda_j^0(1-\alpha)$ and Oracle Inequalities under IC)**.** *Under (A1)-(A3), given $\lambda_j^0(1-\alpha)$ satisfying*

$$\lambda_j^0(1-\alpha) \lesssim \max_{1\leqslant k\leqslant K}\Big\{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}\sqrt{n\log(K/\alpha)} \vee \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}(n\varpi_n K/\alpha)^{1/q}\Big\}, \tag{5.4}$$

and the exact sparsity assumption (3.1)*, then $\tilde{\beta}_j$ obtained from* (3.6) *under IC satisfies*

$$|\tilde{\beta}_j-\beta_j^0|_{j,pr} \lesssim \frac{\sqrt{s_j}}{\kappa_j(\bar{c})}\max_{1\leqslant k\leqslant K}\Psi_{jk}\Big\{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}\frac{\sqrt{\log(K/\alpha)}}{\sqrt{n}} \vee \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}n^{1/q-1}(\varpi_n K/\alpha)^{1/q}\Big\}, \tag{5.5}$$

with probability $1-\alpha-o(1)$, where for $\varsigma > 1/2-1/q$ (weak dependence case), $\varpi_n = 1$; for $\varsigma < 1/2-1/q$ (strong dependence case), $\varpi_n = n^{q/2-1-\varsigma q}$.

Comment 5.4. The Nagaev type of inequality in (5.3) has two terms, namely an exponential term and a polynomial term. It should be noted that if the polynomial term dominates, the above bound does not allow for ultra high dimension of K. Basically, we only allow for a polynomial rate $K = \mathcal{O}(n^{\tilde{c}})$, and the rate of K interplays with the dependence adjusted norm $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$. In particular, to make sure that the estimators are consistent (i.e. the error bounds tend to zero for sufficiently large n), for example, we need $\tilde{c} < q-1-\upsilon q/2-dq$, if there exists q to guarantee $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} = \mathcal{O}(n^d)$ and $0<\upsilon<1$ such that $s_j = \mathcal{O}(n^\upsilon)$.

We now discuss the case of sub-Gaussian tail or sub-exponential tail, which is mostly assumed in the literature.

Comment 5.5. Suppose a stronger exponential moment condition is satisfied,

$$\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,\varsigma} = \sup_{q\geqslant 2} q^{-\nu}\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} < \infty, \tag{5.6}$$

where $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,\varsigma}$ is interpreted as the dependence adjusted sub-exponential ($\nu=2$) or sub-Gaussian ($\nu=1$) norm. Consider the special case of VAR(1). As shown above, we have $\|X_{jk,t}\varepsilon_{j,t}-X^*_{jk,t}\varepsilon^*_{j,t}\|_q \leqslant 2|[A^{t-1}]_j|_1\mu_q^2$. In particular, it is known that $\mu_q \lesssim q$ for sub-exponential variables and $\mu_q \lesssim \sqrt{q}$ for sub-Gaussian variables. Let $\nu=2$ and $\nu=1$ for the two cases respectively, $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,\varsigma} \lesssim (m^*+1)|c|^{m^*-1}$. Then applying the exponential tail bounds as in Lemma B.3 in the supplementary material, we arrive at the following error bounds with probability $1-\alpha-o(1)$,

$$|\tilde{\beta}_j-\beta_j^0|_{j,pr} \lesssim \frac{\sqrt{s_j}}{\kappa_j(\bar{c})}\max_{1\leqslant k\leqslant K}\Psi_{jk}\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,0}\frac{\{\log(K/\alpha)\}^{1/\gamma}}{\sqrt{n}}, \quad \gamma = 2/(2\nu+1), \tag{5.7}$$

as $\lambda_j^0(1-\alpha) \lesssim \sqrt{n}(\log K)^{1/\gamma}\max_{1\leqslant k\leqslant K}\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,0}$. The bound (5.7) works with ultra-high dimensional rate $\exp(n^{r\gamma})$ ($r<1$) of K as only the exponential term shows in the inequality. In particular, suppose $s_j = \mathcal{O}(n^\upsilon)$, and $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,0} = \mathcal{O}(n^d)$, then $r+d+\upsilon/2 < 1/2$ is required to ensure the consistency.

5.2 Gaussian Approximation for Dependent Data

Now we look at the validity of the choice of $Q_j(1-\alpha)$, which relies on a Gaussian approximation theorem. First we define the Kolmogorov distance between any two K-dim random vectors.

Definition 5.2. *Let* $\boldsymbol{X} = (X_1, \cdots, X_K)^\top \in \mathbb{R}^K$, $\boldsymbol{Y} = (Y_1, \cdots, Y_K)^\top \in \mathbb{R}^K$. *The Kolmogorov distance between* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *is defined as*

$$\rho(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{r \geqslant 0} \big| \mathrm{P}(|\boldsymbol{X}|_\infty \geqslant r) - \mathrm{P}(|\boldsymbol{Y}|_\infty \geqslant r) \big|.$$

For each single equation j, aggregate the dependence adjusted norm over $k = 1, \ldots, K$:

$$\| |X_{j,\cdot}|_\infty \|_{q,\varsigma} \stackrel{\text{def}}{=} \sup_{m \geqslant 0} (m+1)^\varsigma \sum_{t=m}^{\infty} \delta_{q,j,t}, \ \delta_{q,j,t} \stackrel{\text{def}}{=} \| |X_{j,t} - X_{j,t}^*|_\infty \|_q, \tag{5.8}$$

where $q \geqslant 1$ and $\varsigma > 0$. Moreover, define the following quantities

$$\begin{aligned} &\Phi_{j,q,\varsigma} \stackrel{\text{def}}{=} 2 \max_{1 \leqslant k \leqslant K} \|X_{jk,\cdot}\|_{q,\varsigma} \|\varepsilon_{j,\cdot}\|_{q,\varsigma}, \ \Gamma_{j,q,\varsigma} \stackrel{\text{def}}{=} 2 \|\varepsilon_{j,\cdot}\|_{q,\varsigma} \Big(\sum_{k=1}^{K} \|X_{jk,\cdot}\|_{q,\varsigma}^{q/2} \Big)^{2/q} \\ &\Theta_{j,q,\varsigma} \stackrel{\text{def}}{=} \Gamma_{j,q,\varsigma} \wedge \{ 2 \| |X_{j,\cdot}|_\infty \|_{q,\varsigma} \|\varepsilon_{j,\cdot}\|_{q,\varsigma} (\log K)^{3/2} \}. \end{aligned} \tag{5.9}$$

It is worth noting that the norm $\| |X_{j,\cdot}|_\infty \|_{q,\varsigma}$ is a kind of aggregated dependence adjusted norm for a vector of processes in comparison to the dependence adjusted norm for a univariate process as in Definition 5.1.

Some additional assumptions are required. Define $L_{1,j} = \{\Phi_{j,4,\varsigma} \Phi_{j,4,0} (\log K)^2\}^{1/\varsigma}$, $W_{1,j} = (\Phi_{j,6,0}^6 + \Phi_{j,8,0}^4)\{\log(Kn)\}^7$, $W_{2,j} = \Phi_{j,4,\varsigma}^2 \{\log(Kn)\}^4$, $W_3 = [n^{-\varsigma} \{\log(Kn)\}^{3/2} \Theta_{j,2q,\varsigma}]^{1/(1/2-\varsigma-1/q)}$, $N_{1,j} = (n/\log K)^{q/2} \Theta_{j,2q,\varsigma}^q$, $N_{2,j} = n(\log K)^{-2} \Phi_{j,4,\varsigma}^{-2}$, $N_{1,j} = \{n^{1/2} (\log K)^{-1/2} \Theta_{j,2q,\varsigma}^{-1}\}^{1/(1/2-\varsigma)}$.

(A4) i) (weak dependency case) Given $\Theta_{j,2q,\varsigma} < \infty$ with $q \geqslant 4$ and $\varsigma > 1/2 - 1/q$, then $\Theta_{j,2q,\varsigma} n^{1/q-1/2} \{\log(Kn)\}^{3/2} \to 0$ and $L_1 \max(W_{1,j}, W_{2,j}) = o(1) \min(N_{1,j}, N_{2,j})$.
ii) (strong dependency case) Given $0 < \varsigma < 1/2 - 1/q$, then $\Theta_{j,2q,\varsigma} (\log K)^{1/2} = o(n^\varsigma)$ and $L_1 \max(W_{1,j}, W_{2,j}, W_{3,j}) = o(1) \min(N_{2,j}, N_{3,j})$.

The assumptions impose mild restrictions on the dependency structure of covariates and error terms. They include a wide class of potential correlation and heterogeneity (including conditional heteroscedasticity), with possible allowance of the lagged dependent variables. Two examples of large VAR and ARCH for high-dimensional time series can be found in Appendix C.2 in the supplementary materials.

Comment 5.6. [Admissible Dimension Rates by the Conditions for Gaussian Approximation] As discussed in Zhang and Wu (2017a), consider the case with $\Theta_{j,2q,\varsigma} = \mathcal{O}(K^{1/q})$ and $\Phi_{j,2q,\varsigma} = \mathcal{O}(1)$, where $\varsigma > 1/2 - 1/q$. Then $\Theta_{j,2q,\varsigma} n^{1/q-1/2} \{\log(Kn)\}^{3/2} \to 0$ becomes $K\{\log(nK)\}^{3q/2} = o(n^{q/2-1})$, which implies that $L_1 \max(W_1, W_2) = o(1) \min(N_1, N_2)$. This means with (A4), the dimension K has to satisfy the condition $K(\log K)^{3q/2} = o(n^{q/2-1})$.

Theorem 5.2 (Gaussian Approximation Results for Dependent Data)**.** *Under (A1) and (A3)-(A4), for each* $j = 1, \ldots, J$ *assume that there exists a constant* $c_j > 0$ *such that* $\min_{1 \leqslant k \leqslant K} \mathrm{avar}(S_{jk}) \geqslant$

c_j, *then we have*

$$\rho(D_j^{-1}S_{j\cdot}, D_j^{-1}Z_j) \to 0, \quad \text{as } n \to \infty, \tag{5.10}$$

where $Z_j \sim \mathrm{N}(0, \Sigma_j)$, Σ_j *is the* $K \times K$ *long-run variance-covariance matrix of* $X_{j,t}\varepsilon_{j,t}$, *and* D_j *is a diagonal matrix with the square root of the diagonal elements of* Σ_j, *namely*

$$\Big\{ \sum_{\ell=-\infty}^{\infty} \mathsf{E}(X_{jk,t}X_{jk,(t-\ell)}\varepsilon_{j,t}\varepsilon_{j,(t-\ell)}) \Big\}^{1/2} = \sqrt{\mathrm{avar}(S_{jk})}, \text{ for } k = 1, \ldots, K.$$

Comment 5.7. The conclusion in Theorem 5.2 can be held with stronger tail assumptions, following Theorem 5.2 in Zhang and Wu (2017a).

Theorem 5.2 justifies the choice of λ_j and $\tilde{Q}_j(1-\alpha)$, which leads to the following corollary:

Corollary 5.2. *Under the conditions of Theorem 5.2, for each* j *we have*

$$\sup_{\alpha\in(0,1)} \big| \mathrm{P}\{ \max_{1\leqslant k\leqslant K} 2c\sqrt{n}|S_{jk}/\Psi_{jk}| \geqslant Q_j(1-\alpha)\} - \alpha \big| \to 0, \quad \text{as } n \to \infty. \tag{5.11}$$

It is worth noting that in practice the variance involved in the Gaussian approximation in 5.2 is not known; we shall discuss how we estimate the variance and also the validity of the Gaussian approximation result with an estimated variance. Given the realization $X_{j,1}\varepsilon_{j,1}, \ldots, X_{j,n}\varepsilon_{j,n}$, we propose to estimate the $K \times K$ long-run variance-covariance matrix Σ_j for $j = 1, \ldots, J$ as follows, given $\mathsf{E}\, X_{j,t}\varepsilon_{j,t} = 0$, and consider:

$$\widehat{\Sigma}_j = \frac{1}{b_n l_n} \sum_{i=1}^{l_n} \Big(\sum_{l=(i-1)b_n+1}^{ib_n} X_{j,l}\varepsilon_{j,l} \Big) \Big(\sum_{l=(i-1)b_n+1}^{ib_n} X_{j,l}\varepsilon_{j,l} \Big)^{\top}. \tag{5.12}$$

Moreover, the following corollary ensures that the Gaussian approximation results still hold if we use the estimate in (5.12).

Corollary 5.3. *Let the conditions of Theorem 5.2 hold, and assume* $\Phi_{j,2q,\varsigma} < \infty$ *with* $q > 4$, $b_n = \mathcal{O}(n^{\eta})$ *for some* $0 < \eta < 1$. *Let* $F_\varsigma = n$, *for* $\varsigma > 1 - 2/q$; $F_\varsigma = l_n b_n^{q/2-\varsigma q/2}$, *for* $1/2 - 2/q < \varsigma < 1 - 2/q$; $F_\varsigma = l_n^{q/4-\varsigma q/2} b_n^{q/2-\varsigma q/2}$, *for* $\varsigma < 1/2 - 2/q$. *Further assume* $n^{-1}\log^2 K \max\{n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2, n^{1/2}b_n^{1/2}\sqrt{\log K}\Phi_{j,8,\varsigma}^2, F_\varsigma^{2/q}\Gamma_{j,2q,\varsigma}^2 K^{2/q}, \Phi_{j,2,0}\Phi_{j,2,\varsigma}v'(b_n)n/\sqrt{\log K}\} = o(1)$, *with* $v'(b_n) = (b_n+1)^{-\varsigma} + 2v_{n,2}/b_n$, $v_{n,2} = \log b_n$ *(resp.* $b_n^{-\varsigma+1}$ *or 1) for* $\varsigma = 1$ *(resp.* $\varsigma < 1$ *or* $\varsigma > 1$*). Then for each* j *we have*

$$\rho(\widehat{D}_j^{-1}S_{j\cdot}, D_j^{-1}Z_j) \to 0, \quad \text{as } n \to \infty, \tag{5.13}$$

where $\widehat{D}_j = \{\mathrm{diag}(\widehat{\Sigma}_j)\}^{1/2}$.

It should be noted that given the Gaussian approximation results in Theorem 5.2, we can have a refined bound for $\lambda_j^0(1-\alpha)$ and also the oracle inequalities under IC.

Corollary 5.4 (Bounds for $\lambda_j^0(1-\alpha)$ and Oracle Inequalities under IC with Gaussian Approximation Results)**.** *Under the conditions of Theorem 5.2 together with (A2), let* $2(\log K)^{-1/2} + \rho(D_j^{-1}S_{j\cdot}, D_j^{-1}Z_j) = o(\alpha)$ *and* $Z_\alpha \geqslant 2\tilde{c}\sqrt{n\log K}$, *where* $\tilde{c}$ *is no less than the* c *in the definition*

of $\lambda_j^0(1-\alpha)$*, then we have* $\lambda_j^0(1-\alpha)$ *satisfying*

$$\lambda_j^0(1-\alpha) \leqslant Z_\alpha, \tag{5.14}$$

and given the exact sparsity assumption (3.1)*, then* $\tilde{\beta}_j$ *obtained from* (3.6) *under IC satisfies*

$$|\tilde{\beta}_j - \beta_j^0|_{j,pr} \lesssim \frac{\sqrt{s_j}}{\kappa_j(\bar{c})} \max_{1\leqslant k\leqslant K} \Psi_{jk} \frac{\sqrt{\log K}}{\sqrt{n}}, \tag{5.15}$$

with probability $1-\alpha-o(1)$.

We note that the allowed dimension K is still of polynominal rate restricted by (A4).

5.3 Multiplier Block Bootstrap Procedure

In this subsection, we discuss how $\Lambda_j(1-\alpha)$ is attainable via block bootstrap. The data over $t = 1, \ldots, n$ are divided into l_n blocks with the same number of observations b_n, $n = b_n l_n$ (without loss of generality), where $b_n, l_n \in \mathbb{Z}$.

Recall that $\Lambda_j(1-\alpha) = 2c\sqrt{n}q_{j,(1-\alpha)}^{[B]}$, $q_{j,(1-\alpha)}^{[B]}$ is the $(1-\alpha)$ quantile of $\max_{1\leqslant k\leqslant K}|Z_{jk}^{[B]}/\Psi_{jk}|$, where $Z_{jk}^{[B]}$ are defined as

$$Z_{jk}^{[B]} = \frac{1}{\sqrt{n}} \sum_{i=1}^{l_n} e_{j,i} \sum_{l=(i-1)b_n+1}^{ib_n} \varepsilon_{j,l} X_{jk,l}, \tag{5.16}$$

and $e_{j,i}$ are i.i.d. $\mathrm{N}(0,1)$ random variables independent of X and ε.

In fact, the above construction relies on knowing the true residuals $\varepsilon_{j,t}$. In practice, one needs to pre-estimate them using a conservative choice of penalty levels and loadings. The issue of generated errors can be dealt with using a similar argument as in the proof of Corollary 5.3.

Theorem 5.3 (Validity of Multiplier Block Bootstrap Method)**.** *Under (A1) and (A3), and assume* $\Phi_{j,2q,\varsigma} < \infty$ *with* $q > 4$*,* $b_n = \mathcal{O}(n^\eta)$ *for some* $0 < \eta < 1$ *(the detailed rate is calculated in* (B.2) *in the supplementary materials), then we have*

$$\sup_{\alpha\in(0,1)} \big|\, \mathrm{P}\big(\max_{1\leqslant k\leqslant K} |S_{jk}/\Psi_{jk}| \geqslant q_{j,(1-\alpha)}^{[B]} \big) - \alpha \big| \to 0, \text{ as } n \to \infty. \tag{5.17}$$

5.4 Joint Penalty over Equations

Recall that the theoretical choice $\lambda^0(1-\alpha)$ is defined as the $(1-\alpha)$ quantile of $\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} 2c\sqrt{n}|S_{jk}/\Psi_{jk}|$. The empirical choices of the joint penalty level can be:

a) $Q(1-\alpha)$: the $(1-\alpha)$ quantile of $2c \max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} \sqrt{n}|Z_{jk}/\Psi_{jk}|$. In practice, one can take an alternative choice such that $\widetilde{Q}(1-\alpha) \stackrel{\text{def}}{=} 2c\sqrt{n}\Phi^{-1}\{1-\alpha/(2KJ)\}$.

b) $\Lambda(1-\alpha) \stackrel{\text{def}}{=} 2c\sqrt{n}q_{(1-\alpha)}^{[B]}$, where $q_{(1-\alpha)}^{[B]}$ is the $(1-\alpha)$ quantile of $\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} |Z_{jk}^{[B]}/\Psi_{jk}|$.

Section A in the supplementary material provides the main theorems for joint equation estimation. In particular, the dimension along $k = 1, \ldots, K$ and $j = 1, \ldots, J$ will be considered

together by vectorization, resulting in the dimension of KJ. Following the results for the single equation (where j is fixed), we generalize the theorems above to multiple equations case by changing the dimension from K to KJ, see Section A in the Appendix for more details.

5.5 Post-Model Selection Estimation

LASSO estimation is known to be biased especially for large coefficients. Therefore, a post-selection step helps to reduce the bias by running an OLS as a second step on the selected covariates in the first step. In particular, we consider the 2-step OLS post-LASSO estimator:

i) ℓ_1-penalized regression (LASSO selection)

$$\breve{\beta}_j = \arg\min_{\beta \in \mathbb{R}^K} \frac{1}{n}\sum_{t=1}^{n}(Y_{j,t} - X_{j,t}^\top \beta)^2 + \frac{\lambda}{n}\sum_{k=1}^{K}|\beta_{jk}|\Psi_{jk}, \tag{5.18}$$

where λ is the joint penalty level.

ii) We run the post-selection regression (OLS estimation)

$$\widehat{\beta}_j^{[P]} = \arg\min_{\beta \in \mathbb{R}^K}\{\frac{1}{n}\sum_{t=1}^{n}(Y_{j,t} - X_{j,t}^\top \beta)^2 : \beta_k = 0, k \notin \widehat{T}_j\}, \tag{5.19}$$

where $\widehat{T}_j \stackrel{\text{def}}{=} \text{supp}(\breve{\beta}_j) = \{k \in \{1,\ldots,K\} : \breve{\beta}_{jk} \neq 0\}$.

To provide the prediction performance bounds for the OLS post-LASSO estimators, we need the following restricted sparse eigenvalue (RSE) condition:

(A5) Restricted sparse eigenvalue (RSE): given $p < n$, for $\delta \in \mathbb{R}^K$, with probability $1 - o(1)$,

$$\tilde{\kappa}_j(p)^2 \stackrel{\text{def}}{=} \min_{|\delta_{T_j^c}|_0 \leqslant p, \delta \neq 0} \frac{|\delta|_{j,pr}^2}{|\delta|_2^2} > 0, \quad \phi_j(p) \stackrel{\text{def}}{=} \max_{|\delta_{T_j^c}|_0 \leqslant p, \delta \neq 0} \frac{|\delta|_{j,pr}^2}{|\delta|_2^2} > 0.$$

Here p denotes the restriction on the length of the active set of T_j^c. When $T_j = \emptyset$, (A5) is reduced to the standard sparse eigenvalue condition. Moreover, let $\mu_j(p) \stackrel{\text{def}}{=} \frac{\sqrt{\phi_j(p)}}{\tilde{\kappa}_j(p)}$, and denote by $\widehat{p}_j \stackrel{\text{def}}{=} |\widehat{T}_j \backslash T_j|$ the number of components outside $T_j \stackrel{\text{def}}{=} \text{supp}(\beta_j^0) = \{k \in \{1,\ldots,K\} : \beta_{jk}^0 \neq 0\}$ selected by LASSO in the first step.

The performance bounds for the OLS post-LASSO estimator are shown in Theorem A.4 in the supplementary materials.

5.6 Simultaneous Inference

This subsection develops theory corresponding to Section 4. A key Bahadur representation which linearize the estimator for a proper application of the central limit theorem for inference is provided.

Recall that for each $j = 1, \ldots, J$, the following model is considered

$$\begin{aligned} Y_{j,t} &= \sum_{k=1}^{p_j} X_{jk,t}\beta_{jk}^0 + \sum_{k=p_j+1}^{K} X_{jk,t}\beta_{jk}^0 + \varepsilon_{j,t}, \quad \mathsf{E}(\varepsilon_{j,t}X_{j,t}) = 0, \quad F_{\varepsilon_j}(0) = 1/2, \\ X_{jk,t} &= X_{j(-k),t}^\top \gamma_{j(-k)}^0 + v_{jk,t}, \quad \mathsf{E}(v_{jk,t}X_{j(-k),t}) = 0, \quad k = 1, \ldots, p_j, \end{aligned} \tag{5.20}$$

where we define $\gamma_{j(-k)}^0 \stackrel{\text{def}}{=} \arg\min_{\gamma_{j(-k)}} \mathsf{E}(X_{jk,t} - X_{j(-k),t}^\top \gamma_{j(-k)})^2$, and let F_{ε_j} denote the distribution function of $\varepsilon_{j,t}$. In this subsection, we show the validity of the joint confidence region for simultaneous inference on $H_0 : \beta_{jk}^0 = 0, \forall (j,k) \in G$, with $|G| = \sum_{j=1}^J p_j$. In particular, for $j = 1, \ldots, J$, $\beta_{jk}^0 (k = 1, \ldots, p_j)$ are the target parameters. Theoretically, we formulate the estimation as a general Z-estimation problem, with the leading examples as the LAD/LS cases. Nevertheless, it can also include a more general class of loss functions.

For each $(j,k) \in G$, we define the score function as $\psi_{jk}\{Z_{j,t}, \beta_{jk}, h_{jk}(X_{j(-k),t})\}$, where $Z_{j,t} \stackrel{\text{def}}{=} (Y_{j,t}, X_{j,t}^\top)^\top$ and the vector-valued function $h_{jk}(\cdot)$ is a measurable map from $\mathbb{R}^{K-1}$ to $\mathbb{R}^M$ (M is fixed). In particular, in our linear regression case we have $h_{jk}(X_{j(-k),t}) = (X_{j(-k),t}^\top \beta_{j(-k)}, X_{j(-k),t}^\top \gamma_{j(-k)})^\top$, and for the LAD regression $\psi_{jk}\{Z_{j,t}, \beta_{jk}, h_{jk}(X_{j(-k),t})\} = \{1/2 - \mathbf{1}(Y_{j,t} \leqslant X_{jk,t}\beta_{jk} + X_{j(-k),t}^\top \beta_{j(-k)})\}(X_{jk,t} - X_{j(-k),t}^\top \gamma_{j(-k)})$.

Assume that there exists $s = s_n \geqslant 1$ such that $|\beta_{j(-k)}^0|_0 \leqslant s$, $|\gamma_{j(-k)}^0|_0 \leqslant s$, for each $(j,k) \in G$. Moreover, we assume that the nuisance function $h_{jk}^0 = (h_{jk,m}^0)_{m=1}^M$ admits a sparse estimator $\widehat{h}_{jk} = (\widehat{h}_{jk,m})_{m=1}^M$ of the form

$$\widehat{h}_{jk,m}(X_{j(-k),t}) = X_{j(-k),t}^\top \widehat{\theta}_{jk,m}, \quad |\widehat{\theta}_{jk,m}|_0 \leqslant s, \quad m = 1, \ldots, M,$$

where the sparsity level s is small compared to n ($s \ll n$).

The true parameter β_{jk}^0 is identified as a unique solution to the moment condition

$$\mathsf{E}[\psi_{jk}\{Z_{j,t}, \beta_{jk}^0, h_{jk}^0(X_{j(-k),t})\}] = 0. \tag{5.21}$$

However, the object $\arg \operatorname{zero}_{\beta_{jk} \in \widehat{\mathcal{B}}_{jk}} \mathsf{E}_n |[\psi_{jk}\{Z_{j,t}, \beta_{jk}, h_{jk}^0(X_{j(-k),t})\}]|$ does not necessarily exist due to the discontinuity of the function ψ_{jk}. The estimator $\widehat{\beta}_{jk}$ is obtained as a Z-estimator by solving the sample analogue of (5.21)

$$\mathsf{E}_n[\psi_{jk}\{Z_{j,t}, \widehat{\beta}_{jk}, \widehat{h}_{jk}(X_{j(-k),t})\}] \leqslant \inf_{\beta_{jk} \in \widehat{\mathcal{B}}_{jk}} |\mathsf{E}_n[\psi_{jk}\{Z_{j,t}, \beta_{jk}, \widehat{h}_{jk}(X_{j(-k),t})\}]| + o(n^{-1/2} g_n^{-1}),$$

where $g_n \stackrel{\text{def}}{=} \{\log(e|G|)\}^{1/2}$ and $\widehat{\mathcal{B}}_{jk}$ is defined in (C2).

We now lay out the following conditions needed in this section, which are assumed to hold uniformly over $(j,k) \in G$.

(C1) Orthogonality condition:

$$\mathsf{E}\left[\partial_h \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta_{jk}^0, h)|X_{j(-k),t}\}\big|_{h=h_{jk}^0(X_{j(-k),t})} h(X_{j(-k),t})\right] = 0, \tag{5.22}$$

for any $h \in \mathcal{H}_{jk} \cup \{h_{jk}^0\}$, where $\mathcal{H}_{jk}$ is defined in (C5).

(C2) The true parameter β_{jk}^0 satisfies (5.21). Let $\mathcal{B}_{jk}$ be a fixed and closed interval and $\widehat{\mathcal{B}}_{jk}$ be a possibly stochastic interval such that with probability $1-o(1)$, $[\beta_{jk}^0 \pm c_1 r_n] \subset \widehat{\mathcal{B}}_{jk} \subset \mathcal{B}_{jk}$, where $r_n \stackrel{\text{def}}{=} n^{-1/2}(\log a_n)^{1/2} \max_{(j,k)\in G} \|\psi_{jk,\cdot}^0\|_{2,\varsigma} + n^{-1} r_\varsigma (\log a_n)^{3/2} \big\| \max_{(j,k)\in G} |\psi_{jk,\cdot}^0| \big\|_{q,\varsigma}$, $r_n \lesssim \rho_n$ (ρ_n is defined in (C5)), $a_n \stackrel{\text{def}}{=} \max(JK, n, e)$, and $\psi_{jk,t}^0 \stackrel{\text{def}}{=} \psi_{jk}\{Z_{j,t}, \beta_{jk}^0, h_{jk}^0(X_{j(-k),t})\}$. $r_\varsigma = n^{1/q}$ for $\varsigma > 1/2 - 1/q$ and $r_\varsigma = n^{1/2-\varsigma}$ for $\varsigma < 1/2 - 1/q$.

(C3) Properties of the score function: the map $(\beta, h) \mapsto \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h) | X_{j(-k),t}\}$ is twice continuously differentiable, and for every $\vartheta \in \{\beta, h_1, \ldots, h_M\}$, $\mathsf{E}[\sup_{\beta\in\mathcal{B}_{jk}} |\partial_\vartheta \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h_{jk}^0(X_{j(-k),t}) | X_{j(-k),t}\}|^2] \leqslant C_1$; moreover, there exist measurable functions $\ell_1(\cdot), \ell_2(\cdot)$, constants $L_{1n}, L_{2n} \geqslant 1$, $\upsilon > 0$ and a cube $\mathcal{T}_{jk}(X_{j(-k),t}) = \times_{m=1}^M \mathcal{T}_{jk,m}(X_{j(-k),t})$ in $\mathbb{R}^M$ with center $h_{jk}^0(X_{j(-k),t})$ such that for every $\vartheta, \vartheta' \in \{\beta, h_1, \ldots, h_M\}$ we have $\sup_{(\beta,h)\in\mathcal{B}_{jk}\times\mathcal{T}_{jk}(X_{j(-k),t})} |\partial_\vartheta \partial_{\vartheta'} \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h) | X_{j(-k),t}\}| \leqslant \ell_1(X_{j(-k),t})$, $\mathsf{E}\{|\ell_1(X_{j(-k),t})|^4\} \leqslant L_{1n}$, and for every $\beta, \beta' \in \mathcal{B}_{jk}$, $h, h' \in \mathcal{T}_{jk}(X_{j(-k),t})$ we have $\mathsf{E}[\{\psi_{jk}(Z_{j,t}, \beta, h) - \psi_{jk}(Z_{j,t}, \beta', h')\}^2 | X_{j(-k),t}] \leqslant \ell_2(X_{j(-k),t})(|\beta-\beta'|^\upsilon + |h-h'|_2^\upsilon)$, and $\mathsf{E}\{|\ell_2(X_{j(-k),t})|^4\} \leqslant L_{2n}$.

(C4) Identifiability: $2|\mathsf{E}[\psi_{jk}\{Z_{j,t}, \beta, h_{jk}^0(X_{j(-k),t})\}]| \geqslant |\phi_{jk}(\beta - \beta_{jk}^0)| \wedge c_1$ holds for all $\beta \in \mathcal{B}_{jk}$, where $\phi_{jk} \stackrel{\text{def}}{=} \partial_\beta \mathsf{E}[\psi_{jk}\{Z_{j,t}, \beta_{jk}^0, h_{jk}^0(X_{j(-k),t})\}]$ and $|\phi_{jk}| \geqslant c_1$.

(C5) Properties of the nuisance function: with probability $1 - o(1)$, $\widehat{h}_{jk} \in \mathcal{H}_{jk}$, where $\mathcal{H}_{jk} = \times_{m=1}^M \mathcal{H}_{jk,m}$ and each $\mathcal{H}_{jk,m}$ being the class of functions of the form $\tilde{h}_{jk,m}(X_{j(-k),t}) = X_{j(-k),t}^\top \theta_{jk,m}$, $|\theta_{jk,m}|_0 \leqslant s$, $\tilde{h}_{jk,m} \in \mathcal{T}_{jk,m}$. There exists sequence of constants $\rho_n \downarrow 0$ such that $\mathsf{E}[\{\tilde{h}_{jk,m}(X_{j(-k),t}) - h_{jk,m}^0(X_{j(-k),t})\}^2] \lesssim \rho_n^2$.

(C6) The class of functions $\mathcal{F}_{jk} = \{z \mapsto \psi_{jk}\{z, \beta, \tilde{h}(x_{j(-k)})\} : \beta \in \mathcal{B}_{jk}, \tilde{h} \in \mathcal{H}_{jk} \cup \{h_{jk}^0\}\}$ (z is a random vector taking values in a Borel subset of a Euclidean space which contains the vectors $x_{j(-k)}$ as subvectors) is pointwise measurable and has measurable envelope $F_{jk} \geqslant \sup_{f\in\mathcal{F}_{jk}} |f|$, such that $F = \max_{(j,k)\in G} F_{jk}$ satisfies $\mathsf{E}\{F^q(z)\} < \infty$ for some $q \geqslant 4$.

(C7) The second-order moments of scores are bounded away from zero: $\omega_{jk} = \mathsf{E}\{(\frac{1}{\sqrt{n}} \sum_{t=1}^n \psi_{jk,t}^0)^2\} \geqslant c_1$.

(C8) Dimension growth rates: $\rho_{n,\upsilon}(L_{2n} s \log a_n)^{1/2} + n^{-1/2} r_\varsigma (s \log a_n)^{3/2} \|F(z_t)\|_q + \rho_n^2 n^{1/2} = o(g_n^{-1})$. In particular, for the mean regression case $\rho_{n,\upsilon} = \rho_n s$ and $\rho_{n,\upsilon} = \rho_n^{1/2}$ for the median regression case. $n^{-1/2}(s \log a_n)^{1/2} \max_{f\in\mathcal{F}'} \|f(z_t)\|_2 + n^{-1} r_\varsigma (s \log a_n)^{3/2} \|\bar{F}'(z_t)\|_q = \mathcal{O}(\rho_n)$. $\mathcal{F}' = \{z \mapsto \psi_{jk}\{z, \beta, \tilde{h}(x_{j(-k)})\} : (j,k) \in G, \beta \in \mathcal{B}_{jk}, \tilde{h} \in \mathcal{H}_{jk} \cup \{h_{jk}^0\}\}$ with $\bar{F}' = \sup_{f\in\mathcal{F}'} |f|$.

(C9) Let $B_\Phi^h = \max_{m\in\{1,2\}} \Phi_{m,2,\varsigma}^h$, $B_\Omega^h = \max_{m\in\{1,2\}} \Omega_{m,q,\varsigma}^h$, $B_\Phi'^h = \max_{m\in\{1,2\}} \Phi_{m,2,\varsigma}'^h$, and $B_\Omega'^h = \max_{m\in\{1,2\}} \Omega_{m,q,\varsigma}'^h$ (see (B.9), (B.10) and (B.15) in the supplementary for the definitions of $\Phi_{m,2,\varsigma}^h$, $\Omega_{m,q,\varsigma}^h$, $\Phi_{2,\varsigma}^\beta$, $\Omega_{q,\varsigma}^\beta$, $\Phi_{m,2,\varsigma}'^h$, $\Omega_{m,q,\varsigma}'^h$, $\Phi_{2,\varsigma}'^\beta$, $\Omega_{q,\varsigma}'^\beta$). The following restrictions are assumed:

$$s\rho_n (\log a_n)^{1/2} B_\Phi^h + n^{-1/2} r_\varsigma \rho_n s^2 (\log a_n)^{3/2} B_\Omega^h = o(g_n^{-1}),$$

$$\rho_n (s \log a_n)^{1/2} \Phi_{2,\varsigma}^\beta + n^{-1/2} r_\varsigma \rho_n (s \log a_n)^{3/2} \Omega_{q,\varsigma}^\beta = o(g_n^{-1}),$$

$$B_{\Phi}^{'h}\rho_n s^{1/2} = \mathcal{O}(\max_{f\in\mathcal{F}'}\|f(z_t)\|_2),\ B_{\Omega}^{'h}\rho_n s^{1/2} = \mathcal{O}(\|\bar{F}'(z_t)\|_q),$$

$$\Phi_{2,\varsigma}^{'\beta}\rho_n = \mathcal{O}(\max_{f\in\mathcal{F}'}\|f(z_t)\|_2),\ \Omega_{q,\varsigma}^{'\beta}\rho_n = \mathcal{O}(\|\bar{F}'(z_t)\|_q).$$

(C9') Consider the stronger exponential moment condition as in (5.6) and corresponding to (C5), assume that $\mathsf{E}[\{\tilde{h}_{jk,m}(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}^2] \lesssim (\rho_n^e)^2$. Recall the definitions of $\Phi^h_{m,\psi_\nu,0}$, $\Phi^\beta_{\psi_\nu,0}$, $\Phi^{'h}_{m,\psi_\nu,0}$, $\Phi^{'\beta}_{\psi_\nu,0}$ in (B.17) and (B.20) in the supplementary. The following restrictions are assumed:

$$n^{-1/2}(\log a_n)^{1/\gamma}\max_{(j,k)\in G}\|\psi^0_{jk,\cdot}\|_{\psi_\nu,0} \lesssim r_n,$$

$$(s\log a_n)^{1/\gamma}[\rho^e_{n,\upsilon} \vee \rho^e_n\{(s^{1/2}\max_{m\in\{1,2\}}\Phi^h_{m,\psi_\nu,0}) \vee \Phi^\beta_{\psi_\nu,0}\}] = o(g_n^{-1}),$$

$$n^{-1/2}(s\log a_n)^{1/\gamma}\max_{f\in\mathcal{F}'}\|f(z.)\|_{\psi_\nu,0} = \mathcal{O}(\rho^e_n),$$

$$\rho^e_n\{(s^{1/2}\max_{m\in\{1,2\}}\Phi^{'h}_{m,\psi_\nu,0}) \vee \Phi^{'\beta}_{\psi_\nu,0}\} = \mathcal{O}(\max_{f\in\mathcal{F}'}\|f(z.)\|_{\psi_\nu,0}),$$

in particular, for the mean regression case $\rho^e_{n,\upsilon} = \rho^e_n s$ and $\rho^e_{n,\upsilon} = \sqrt{\rho^e_n}$ for the median regression case.

(C10) The density of error $f_{\varepsilon_j}(\cdot)$ is continuously differentiable and both of $f_{\varepsilon_j}(\cdot)$ and $f'_{\varepsilon_j}(\cdot)$ are bounded from the above.

Conditions (C1)-(C4) and (C7) assume mild restrictions on the Z-estimation problems. They include the LAD-based regression (used in Algorithm 2) with nonsmooth score function. Conditions (C2) and (C8) imply that $\max_{(j,k)\in G}\|\psi^0_{jk,\cdot}\|_{2,\varsigma} \lesssim s^{1/2}\max_{f\in\mathcal{F}'}\|f(z_t)\|_2$ and $\|\max_{(j,k)\in G}|\psi^0_{jk,\cdot}|\|_{q,\varsigma} \lesssim s^{3/2}\|\bar{F}'(z_t)\|_q$. In (C5), we suppose that the nuisance parameters have estimators with good sparsity and convergence rate properties. As discussed in previous sections, given the ideal choice of the tuning parameter, the oracle inequalities provided in Corollary 5.1 ensures that our proposed algorithms can produce the estimator of the form $|\widehat{\beta}^{[1]}_{j(-k)} - \beta^0_{j(-k)}|_{j,pr} \lesssim \{\sqrt{s\log(a_n)/n} \vee n^{1/q-1}(\varpi_n a_n)^{1/q}\}\max_{1\leqslant k\leqslant K}\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$, where for $\varsigma > 1/2 - 1/q$ (weak dependence case), $\varpi_n = 1$; for $\varsigma < 1/2 - 1/q$ (strong dependence case), $\varpi_n = n^{q/2-1-\varsigma q}$. The moments of the envelopes are assumed to be finite in (C6).

Comment 5.8. [Discussion on the dimension growth rates] Consider the special case of VAR(1) model. Following the discussion in Comment 5.3, given a geometric decay rate, we have $L_{2n}, B^h_\Phi, B^{'h}_\Phi, \Phi^\beta_{2,\varsigma}, \Phi^{'\beta}_{2,\varsigma}, \max_{f\in\mathcal{F}'}\|f(z_t)\|_2, \max_{(j,k)\in G}\||\psi^0_{jk,\cdot}|\|_{2,\varsigma} \lesssim M_n$, where M_n only depends on the $2q$-th moments of ε_t and ς. Moreover, suppose these quantities are bounded by constant and let $d_n \stackrel{\text{def}}{=} (|G| \vee J)$, we have $B^h_\Omega, B^{'h}_\Omega \lesssim d_n^{1/q}(1 \vee s^{1/2}\rho_n)$, $\Omega^\beta_{q,\varsigma}, \Omega^{'\beta}_{q,\varsigma} \lesssim d_n^{1/q}s^{1/2}\rho_n$ for mean regression case, and $B^h_\Omega, B^{'h}_\Omega \lesssim d_n^{3/(4q)}(1 \vee s^{1/2}\rho_n)$, $\Omega^\beta_{q,\varsigma}, \Omega^{'\beta}_{q,\varsigma} \lesssim d_n^{1/(2q)}s^{1/2}\rho_n$ for the median regression. Moreover, $\|F(z_t)\|_q, \|F'(z_t)\|_q \lesssim d_n^{1/q}(1 \vee \rho_n)$, $\|\max_{(j,k)\in G}|\psi^0_{jk,\cdot}|\|_{q,\varsigma} \lesssim d_n^{1/q}(1 \vee \rho_n)$. The detailed derivation of these rates can be found in the Comment B.3 in the supplementary. Inserting

them into (C8) and (C9) yields

$$n^{-1/2}s^2(\log a_n)^{3/2} + n^{-1}r_\varsigma s^3(\log a_n)^{5/2}d_n^{1/q} + n^{-1/2}r_\varsigma s^{3/2}(\log a_n)^2 d_n^{1/q} = o(1),$$

and

$$n^{-1/4}s^{3/4}(\log a_n)^{5/4} + n^{-1/2}r_\varsigma^{1/2}s^{5/4}(\log a_n)^{7/4}d_n^{3/(8q)} + n^{-1/2}r_\varsigma s^{3/2}(\log a_n)^2 d_n^{3/(4q)} = o(1),$$

for the smooth and non-smooth cases respectively. As a result, we only allow the dimension $(|G| \vee J)$ is of polynomial order with respect to n if q is not tending to infinity. In particular, under the case of $\varsigma > 1/2$ and $q = \infty$, the required rate reduces to $n^{-1/2}s^2(\log a_n)^{3/2} + n^{-1}s^3(\log a_n)^{5/2} + n^{-1/2}s^{3/2}(\log a_n)^2 = o(1)$ or $n^{-1/4}s^{3/4}(\log a_n)^{5/4} + n^{-1/2}s^{5/4}(\log a_n)^{7/4} + n^{-1/2}s^{3/2}(\log a_n)^2 = o(1)$, respectively. In the ideal case where we have weak dependency, the dimension growth rates are slightly slower than the i.i.d. case as in Belloni et al. (2015b) (i.e., $s^2 \log a_n^3 = o(n)$ or $s^3 \log a_n^5 = o(n)$ for the smooth or non-smooth case, respectively), as we apply a different way to bound the dependence adjusted norm in the concentration inequality.

More generally, suppose $\max\{L_{2n}, B_\Phi^h, B_\Phi^{\prime h}, \Phi_{2,\varsigma}^\beta, \Phi_{2,\varsigma}^{\prime\beta}, \max_{f\in\mathcal{F}'}\|f(z_t)\|_2, \max_{(j,k)\in G}\||\psi_{jk,\cdot}^0|\|_{2,\varsigma}\} = \mathcal{O}(n^{k_1})$, and $\max\{B_\Omega^h, B_\Omega^{\prime h}, \Omega_{q,\varsigma}^\beta, \Omega_{q,\varsigma}^{\prime\beta}, \|F(z_t)\|_q, \|F'(z_t)\|_q, \|\max_{(j,k)\in G}|\psi_{jk,\cdot}^0|\|_{q,\varsigma}\} = \mathcal{O}(n^{k_2})$, with $0 \leqslant k_1 \leqslant k_2$, and let $s = \mathcal{O}(n^v)$, $\log a_n = \mathcal{O}(n^r)$. Then (C8) and (C9) imply that

$$r < \max\left\{\frac{1-4v-2k_1}{3}, -\frac{2}{5q} + \frac{2-6v-2k_2}{5}, -\frac{1}{2q} + \frac{1-3v-2k_2}{4}\right\}, \text{ if } \varsigma > 1/2 - 1/q,$$

$$r < \max\left\{\frac{1-4v-2k_1}{3}, \frac{2\varsigma+1-6v-2k_2}{5}, \frac{2\varsigma-3v-2k_2}{4}\right\}, \text{ if } \varsigma < 1/2 - 1/q,$$

and

$$r < \max\left\{\frac{1-3v-4k_1}{5}, -\frac{2}{7q} + \frac{2-5v-2k_2}{7}, -\frac{1}{2q} + \frac{1-3v-2k_2}{4}\right\}, \text{ if } \varsigma > 1/2 - 1/q,$$

$$r < \max\left\{\frac{1-3v-4k_1}{3}, \frac{2\varsigma+1-5v-2k_2}{7}, \frac{2\varsigma-3v-2k_2}{4}\right\}, \text{ if } \varsigma < 1/2 - 1/q,$$

for the smooth and non-smooth cases.

Theorem 5.4. *[Uniform Bahadur Representation] Under conditions (A1)-(A4) and (C1)-(C10), with probability* $1 - o(1)$*, we have*

$$\max_{(j,k)\in G} |n^{1/2}\sigma_{jk}^{-1}(\widehat{\beta}_{jk} - \beta_{jk}^0) + n^{-1/2}\sigma_{jk}^{-1}\phi_{jk}^{-1}\sum_{t=1}^n \psi_{jk,t}^0| = o(g_n^{-1}), \text{ as } n \to \infty, \tag{5.23}$$

where $\sigma_{jk}^2 \stackrel{\text{def}}{=} \phi_{jk}^{-2}\omega_{jk}$, $\omega_{jk} \stackrel{\text{def}}{=} \mathsf{E}(\frac{1}{\sqrt{n}}\sum_{t=1}^n \psi_{jk,t}^0)^2$.

Comment 5.9. The same conclusion as in Thereom 5.4 can be drawn with assuming stronger exponential moment conditions in (5.6) and using (C9') instead of (C6), (C8) and (C9). This is implied by Lemma B.8, B.9 and B.10 in the supplementary material.

We now discuss the rates implication under (C9'). Suppose all the dependence adjusted norms are bounded by constant with an appropriately chosen ν, the restrictions in (C9') would

imply $n^{-1/2}(\log a_n)^{2/\gamma+1/2}s^{2/\gamma+1} = o(1)$ for the case of smooth score, and $n^{-1/4}(\log a_n)^{3/(2\gamma)}s^{3/(2\gamma)+1/2} = o(1)$ for the non-smooth case, where $\gamma = 2/(2\nu+1)$. For example, when $\nu = 1/2, \gamma = 1$ the required rates would be $s^6 \log^5 a_n = o(n)$ and $s^6 \log^8 a_n = o(n)$ for the smooth and non-smooth cases respectively.

The results in Theorem 5.4 imply the asymptotic normality of the proposed estimator by Algorithm 1 and 2 by applying central limit theorems and Gaussian Approximation.

Corollary 5.5. *Under conditions (A1)-(A4) and (C10), for any $(j,k) \in G$ the estimators obtained by Algorithm 1 and 2 satisfy*

$$\sigma_{jk}^{-1} n^{1/2}(\widehat{\beta}_{jk}^{[2]} - \beta_{jk}^0) \xrightarrow{\mathcal{L}} \mathrm{N}(0,1).$$

Theorem 5.5. *[Uniform-Dimensional Central Limit Theorem] Under the same conditions as in Theorem 5.4, assume that $\|\psi_{jk,\cdot}^0\|_{2,\varsigma} < \infty$, we have*

$$\sigma_{jk}^{-1} n^{1/2}(\widehat{\beta}_{jk} - \beta_{jk}^0) \xrightarrow{\mathcal{L}} \mathrm{N}(0,1),$$

uniformly over $(j,k) \in G$.

Consider the vector $\widetilde{\zeta}_t \stackrel{\text{def}}{=} \mathrm{vec}\{(\zeta_{jk,t})_{(j,k)\in G}\}$, $\zeta_{jk,t} \stackrel{\text{def}}{=} -\sigma_{jk}^{-1}\phi_{j,k}^{-1}\psi_{jk,t}^0$, and define the aggregated dependence adjusted norm as follows:

$$\|\widetilde{\zeta}_\cdot\|_{q,\varsigma} \stackrel{\text{def}}{=} \sup_{m\geqslant 0}(m+1)^\varsigma \sum_{t=m}^{\infty} \||\widetilde{\zeta}_t - \widetilde{\zeta}_t^*|_\infty\|_q, \tag{5.24}$$

where $q \geqslant 1$, and $\varsigma > 0$. Moreover, define the following quantities

$$\begin{aligned}
&\Phi_{q,\varsigma}^\zeta \stackrel{\text{def}}{=} \max_{(j,k)\in G} \|\zeta_{jk,\cdot}\|_{q,\varsigma},\ \Gamma_{q,\varsigma}^\zeta \stackrel{\text{def}}{=} \Big(\sum_{(j,k)\in G} \|\zeta_{jk,\cdot}\|_{q,\varsigma}^q\Big)^{1/q}, \\
&\Theta_{q,\varsigma}^\zeta \stackrel{\text{def}}{=} \Gamma_{q,\varsigma}^\zeta \wedge \{\|\widetilde{\zeta}_\cdot\|_{q,\varsigma}(\log|G|)^{3/2}\}.
\end{aligned} \tag{5.25}$$

Define $L_1^\zeta = \{\Phi_{2,\varsigma}^\zeta \Phi_{2,0}^\zeta (\log|G|)^2\}^{1/\varsigma}$, $W_1^\zeta = \{(\Phi_{3,0}^\zeta)^6 + (\Phi_{4,0}^\zeta)^4\}\{\log(|G|n)\}^7$, $W_2^\zeta = (\Phi_{2,\varsigma}^\zeta)^2\{\log(|G|n)\}^4$, $W_3^\zeta = [n^{-\varsigma}\{\log(|G|n)\}^{3/2}\Theta_{q,\varsigma}^\zeta]^{1/(1/2-\varsigma-1/q)}$, $N_1^\zeta = (n/\log|G|)^{q/2}(\Theta_{q,\varsigma}^\zeta)^q$, $N_2^\zeta = n(\log|G|)^{-2}(\Phi_{2,\varsigma}^\zeta)^{-2}$, $N_3^\zeta = \{n^{1/2}(\log|G|)^{-1/2}(\Theta_{q,\varsigma}^\zeta\})^{1/(1/2-\varsigma)}$.

(A6) i) (weak dependency case) Given $\Theta_{q,\varsigma}^\zeta < \infty$ with $q \geqslant 2$ and $\varsigma > 1/2 - 1/q$, then $\Theta_{q,\varsigma}^\zeta n^{1/q-1/2}\{\log(|G|n)\}^{3/2} \to 0$ and $L_1^\zeta \max(W_1^\zeta, W_2^\zeta) = o(1)\min(N_1^\zeta, N_2^\zeta)$.
ii) (strong dependency case) Given $0 < \varsigma < 1/2 - 1/q$, then $\Theta_{q,\varsigma}^\zeta(\log|G|)^{1/2} = o(n^\varsigma)$ and $L_1^\zeta \max(W_1^\zeta, W_2^\zeta, W_3^\zeta) = o(1)\min(N_2^\zeta, N_3^\zeta)$.

Corollary 5.6 (Consistency of the Bootstrap Confidence Interval)**.** *Under (A6) and the same conditions as in Theorem 5.4, for each $(j,k) \in G$ assume that there exists a constant $c > 0$ such that $\min_{(j,k)\in G} \mathrm{avar}\,(n^{-1/2}\sum_{t=1}^n \zeta_{jk,t}) \geqslant c$, with probability $1 - o(1)$, we have*

$$\sup_{\alpha\in(0,1)} |\,\mathrm{P}(\beta_{jk}^0 \in \widetilde{\mathrm{CI}}_{jk}(\alpha),\ \forall (j,k) \in G) - (1-\alpha)| = o(1), \quad \text{as } n \to \infty, \tag{5.26}$$

where $\widetilde{\mathrm{CI}}_{jk}(\alpha) \stackrel{\text{def}}{=} \left[\widehat{\beta}_{jk} \pm \widehat{\sigma}_{jk} n^{-1/2} q(1-\alpha)\right]$, *and* $q(1-\alpha)$ *is the* $(1-\alpha)$ *quantile of the* $\max_{(j,k)\in G} |\mathcal{Z}_{jk}|$, *where* $\mathcal{Z}_{jk}$*'s are the standard normal random variables and* $\widehat{\sigma}_{jk}$ *is a consistent estimator of* σ_{jk}.

Following Theorem 5.4, a joint confidence region and the corresponding confidence interval for each component can be constructed via a block bootstrap method. In particular, the bootstrap statistic are defined by $\frac{1}{\sqrt{n}} \sum_{i=1}^{l_n} e_{j,i} \sum_{l=(i-1)b_n+1}^{ib_n} \widehat{\zeta}_{jk,l}$, where $e_{j,i}$'s are independent and identically distributed draws of standard normal random variables and are independent with respect to the data sample $(Z_{j,t})_{j=1}^{J}$. Recall that $\widehat{\zeta}_{jk,t}$ are pre-estimators with a certain range of accuracy.

Corollary 5.7 (Validity of Multiplier Bootstrap)**.** *Under the same conditions as in Theorem 5.4, assume* $\Phi_{q,\varsigma}^{\zeta} < \infty$ *with* $q > 4$, $b_n = \mathcal{O}(n^{\eta})$ *for some* $0 < \eta < 1$ *(the detailed rate is specified in* (B.27)*), we have*

$$\sup_{\alpha\in(0,1)} |\,\mathrm{P}(\beta_{jk}^0 \in \widetilde{\mathrm{CI}}_{jk}^*(\alpha),\, \forall (j,k) \in G) - (1-\alpha)| = o(1), \quad \text{as } n \to \infty, \tag{5.27}$$

where $\widetilde{\mathrm{CI}}_{jk}^*(\alpha) \stackrel{\text{def}}{=} \left[\widehat{\beta}_{jk} \pm \widehat{\sigma}_{jk} n^{-1/2} q^*(1-\alpha)\right]$, *and* $q^*(1-\alpha)$ *is the* $(1-\alpha)$ *conditional quantile of* $\max_{(j,k)\in G} \frac{1}{\sqrt{n}} |\sum_{i=1}^{l_n} e_{j,i} \sum_{l=(i-1)b_n+1}^{ib_n} \widehat{\zeta}_{jk,l}|$.

6 Simulation Study

In this section, we illustrate the performance of our proposed methodology under different simulation scenarios. The first part concerns the performance of the jointly selected penalty level over equations, and the second part discusses the simultaneous inference.

6.1 Estimation with a Jointly Selected Penalty Level

Consider the system of regression equations:

$$Y_{j,t} = X_t^\top \beta_j^0 + \varepsilon_{j,t},, \quad t = 1, \ldots, n,\ j = 1, \ldots, J, \tag{6.1}$$

where $X_t \in \mathbb{R}^K$. We generate X_t independently from $\mathrm{N}(0, \Sigma)$, where $\Sigma_{k_1,k_2} = \gamma^{|k_1-k_2|}$, $\gamma = 0.5$, $\varepsilon_{j,t} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$. The coefficient vectors β_j are assumed to be sparse. In particular, we divide the indices $\{1, \ldots, K\}$ evenly into blocks with fixed block size 5. $\beta_{jk}^0 = 10$ if k and j belong to the same block and 0 otherwise.

We take $n = 100$, # of bootstrap replications = 5000. We set $J, K = 50, 100$ and 150. The prediction norm $|\widehat{\beta}_j - \beta_j^0|_{j,pr}$ and the Euclidean norm $|\widehat{\beta}_j - \beta_j^0|_2$ ratios are presented in Table 6.1. The ratios measure the relative difference between the results using the penalty level determined from the equation-by-equation case and from the joint equation case (λ_j and λ are selected by the multiplier bootstrap procedure). In particular, a ratio smaller than 1 indicates a better performance of using the jointly selected penalty level.

	$J = K = 50$	$J = K = 100$	$J = K = 150$
	Prediction norm		
Mean	0.9634	0.9474	0.9347
Median	0.9695	0.9516	0.9371
Std.	0.0323	0.0272	0.0254
	Euclidean norm		
Mean	0.9590	0.9429	0.9286
Median	0.9679	0.9468	0.9316
Std.	0.0367	0.0292	0.0286

Table 6.1: Prediction norm and Euclidean norm ratios (overall λ relative to equation-by-equation λ_j's, average over equations). Results (mean, median and standard deviation) are computed over 1000 replications.

It is evident from Table 6.1 that the proposed estimation procedure delivers much better performance in terms of the two measures. In particular, the superiority tends to be more evident (more than 10%) with higher dimension of the covariates and more equations.

Still consider the system of regression equations as in (6.1), but here we generate the data with dependency by following the Appendix D in Zhang and Wu (2017b). In particular, assume the linear process such that $X_t = \sum_{\ell=0}^{\infty} A_\ell \xi_{t-\ell}$, with $A_\ell = (\ell+1)^{-\rho-1} M_\ell$, where M_ℓ are independently drawn from Ginibre matrices, i.e. all the entries of M_ℓ are i.i.d. $\mathrm{N}(0,1)$, and in practice the sum is truncated to $\sum_{\ell=0}^{1000}$. We set ρ to be 1.0 for the weaker dependence and 0.1 for the stronger dependence cases respectively. Let $\xi_{k,t} = e_{k,t}(0.8e_{k,t-1}^2 + 0.2)^{1/2}$ where $e_{k,t}$ are i.i.d. distributed as $t(d)/\sqrt{d/(d-2)}$ and $t(d)$ is the Student's t with degree of freedom d (take $d = 8$ for example). ε_t are generated by following the same fashion independently.

We take $n = 100$, # of bootstrap replications = 5000, $J, K = 50, 100$ and 150. Based on bias-variance tradeoff, several approaches were suggested to determine the optimal choice of b_n for univariate case. Concerning the high-dimensional case, we propose to take the one which gives the lowest prediction norm as the optimal choice. Below we report the average prediction norm $J^{-1}\sum_{j=1}^{J} |\widehat{\beta}_j - \beta_j^0|_{j,pr}$ with several block sizes b_n under different settings and the minimal ones are in bold.

	$\rho = 0.1$ (stronger dependency)			$\rho = 1.0$ (weaker dependency)		
	$J = K = 50$	$J = K = 100$	$J = K = 150$	$J = K = 50$	$J = K = 100$	$J = K = 150$
$b_n = 2$	2.0721	2.9122	3.5932	**2.0165**	2.6270	3.2286
$b_n = 4$	2.0627	2.8924	3.5617	2.0303	**2.6183**	3.2225
$b_n = 6$	2.0487	2.9007	3.5235	2.0834	2.6288	**3.2198**
$b_n = 8$	**2.0388**	2.8841	**3.5073**	2.2149	2.6502	3.2320
$b_n = 10$	2.0521	**2.8836**	3.5268	2.3576	2.7099	3.2975
$b_n = 12$	2.0581	2.9065	3.5687	2.5592	2.8310	3.3895

Table 6.2: The prediction norm (average over equations) using several choices of b_n. Results are computed over 1000 simulations.

From Table 6.2, it is apparent that a larger block size is required for the stronger dependency

case. Moreover, the choice also depends on the dimensionality, which is more evident for relatively weaker dependent data. We note that when $J = K = 50$, $\rho = 1.0$ the ordinary multiplier bootstrap (with $b_n = 1$) produces 2.1003 as the average prediction norm, therefore we suggest $b_n = 2$ for this case.

The prediction norm $|\widehat{\beta}_j - \beta_j^0|_{j,pr}$ and the Euclidean norm $|\widehat{\beta}_j - \beta_j^0|_2$ ratios (using the optimal b_n suggested in Table 6.2 for each case correspondingly) are presented in Table 6.3. Again we report the results with the jointly estimated λ (selected by multiplier block bootstrap) relative to using the single equation λ_j's (selected by the multiplier bootstrap).

	$\rho = 0.1$ (stronger dependency)			$\rho = 1.0$ (weaker dependency)		
	$J = K = 50$	$J = K = 100$	$J = K = 150$	$J = K = 50$	$J = K = 100$	$J = K = 150$
	Prediction norm					
Mean	0.9141	0.8534	0.8250	0.9356	0.8786	0.8326
Median	0.9165	0.8532	0.8255	0.9384	0.8792	0.8330
Std.	0.0436	0.0377	0.0326	0.0380	0.0338	0.0296
	Euclidean norm					
Mean	0.9017	0.8447	0.8114	0.9251	0.8648	0.8154
Median	0.9062	0.8453	0.8135	0.9290	0.8652	0.8157
Std.	0.0515	0.0401	0.0348	0.0453	0.0368	0.0317

Table 6.3: Prediction norm and Euclidean norm ratios (overall λ relative to equation-by-equation λ_j's, average over equations). Results (mean, median and standard deviation) are computed over 1000 replications.

The results show that the coefficient estimation performance measured by both the prediction norm and the Euclidean norm is in favor of the joint penalty level with multiplier block bootstrap approach. The results are robust over different dimension cases with stronger or weaker dependency.

6.2 Simultaneous Inference

In this subsection we consider the following regression model for the purpose of simultaneous inference on the parameters within a system of equations

$$Y_{j,t} = d_{j,t}\alpha_j^0 + X_t^\top \beta_j^0 + \varepsilon_{j,t},\ d_{j,t} = X_t^\top \theta_j^0 + v_{j,t},\ t = 1, \ldots, n,\ j = 1, \ldots, J, \tag{6.2}$$

where $\alpha_j^0 = \alpha^0$ for all j. Also, $\beta_j^0, \theta_j^0 \in \mathbb{R}^K$ are assumed to be sparse. In particular, we divide the indices $1, \ldots, K$ evenly into blocks with a fixed block size 5, β_{jk}^0 and θ_{jk}^0 are independently drawn from Unif$[0, 5]$ and Unif$[0, 0.25]$ respectively, if k and j belong to the same block and 0 otherwise. The way to generate X_t, ε_t and v_t is same as the dependent data setting above.

We consider the sample size $n = 100$. Our goal is to estimate and make inferences on the target variables $d_{j,t}$'s based on the procedure proposed in Section 4. We evaluate and compare the empirical power and size performance of the confidence intervals constructed by the asymptotic distribution theory (4.6), block bootstrap (4.4) and the simultaneous confidence regions via block bootstrap (4.8). The bootstrap statistics are computed based on 5000 replications and we also take the optimal block size according to the numerical comparison conducted above.

Note that the case of $\alpha^0 = 0$ gives the size performance under the null hypothesis, while α^0 uniformly lies in $[0, 2.5]$ and $[0, 5]$ illustrate the power results.

Table 6.4 shows the average rejection rate of $H_0^j : \alpha_j^0 = 0$ over j for individual (or multiple) inference and the rejection rate of $H_0 : \alpha_1^0 = \cdots = \alpha_J^0 = 0$ for simultaneous inference under different settings of J, K and ρ. Multiple testing procedure via step-down method, see e.g. Romano and Wolf (2005); Chernozhukov et al. (2013a), is considered to control the false positives in evaluating the power performance. The rejection rates are computed over 1000 simulation samples.

	$\rho = 0.1$ (stronger dependency)			$\rho = 1.0$ (weaker dependency)		
	$J = K = 50$	$J = K = 100$	$J = K = 150$	$J = K = 50$	$J = K = 100$	$J = K = 150$
			$\alpha^0 = 0$			
Ind. Asym.	0.0166	0.0126	0.0126	0.0242	0.0148	0.0119
Ind. Boot.	0.0303	0.0202	0.0155	0.0224	0.0169	0.0141
Simult. Boot.	0.0260	0.0473	0.0527	0.0520	0.0547	0.0587
			$\alpha^0 \sim \text{Unif}[0, 2.5]$			
Ind. Asym.	0.8714	0.8558	0.8553	0.8763	0.8622	0.8572
Ind. Boot.	0.8746	0.8573	0.8566	0.8761	0.8629	0.8578
Mult. Boot.	0.8413	0.8027	0.8004	0.8438	0.8249	0.8091
			$\alpha^0 \sim \text{Unif}[0, 5]$			
Ind. Asym.	0.9376	0.9247	0.9282	0.9380	0.9319	0.9269
Ind. Boot.	0.9390	0.9254	0.9331	0.9288	0.9325	0.9273
Mult. Boot.	0.9282	0.9070	0.9072	0.9262	0.9182	0.9082

Table 6.4: Average rejection rate of $H_0^j : \alpha_j^0 = 0$ over j for the individual (or multiple) inference and the rejection rate of $H_0 : \alpha_1^0 = \cdots = \alpha_J^0 = 0$ for simultaneous inference under several true α^0 values (given the significance level = 0.05).

It is shown that for individual inference our proposed individual bootstrap approach provides a closer size control to the nominal α and more powerful empirical rejection probabilities compared to constructing the confidence intervals by asymptotic normality in most of the cases. Moreover, the simultaneous inference outperforms the individual inference in size accuracy and in terms of the power performance, the multiple testing is relatively conservative after controlling the false positives. Overall, we observe that the results using bootstrap approach are robust over different dimension settings under either stronger or weaker dependency cases.

7 Empirical Analysis: Textual Sentiment Spillover Effects

Financial markets are driven by information, and this is a well-known phenomenon among investors. More frequent news and availability of sentiment data allows study of the impact of firm-specific investor sentiment on market behavior such as stock returns, volatility and liquidity; see Baker and Wurgler, 2006; Tetlock, 2007, among others. Moreover, powerful statistical tools (e.g. LASSO-type estimators) are being used to model complex relationships among individuals. For example, Audrino and Tetereva (2017) analyze the influence of news on US and European companies by constructing a sparse predictive network via adaptive LASSO and

related testing procedures. In this section the developed technology is applied to study textual sentiment spillover effects across individual stocks. This is different from the "equation-by-equation" analysis in Audrino and Tetereva (2017), since we build up a system of regression equations and implement the estimation and the inference of the network jointly.

7.1 Data Source

The empirical study in this paper is carried out based on the financial news articles published on the NASDAQ community platform from January 2, 2015 to December 29, 2015 (252 trading days). The data were gathered via a self-written web scraper to automate the downloading process. The dataset is available at the Research Data Centre (RDC), Humboldt-Universität zu Berlin. Moreover, unsupervised learning approaches are employed to extract sentiment variables from the articles. Two sentiment dictionaries: the BL option lexicon (Hu and Liu, 2004) and the LM financial sentiment dictionary (Loughran and McDonald, 2011) were used in Zhang et al. (2016). For each article i (published on day t), the average proportion of positive/negative words using BL or LM lexica - $Pos^{\text{BL}}_{j,i,t}$, $Neg^{\text{BL}}_{j,i,t}$, $Pos^{\text{LM}}_{j,i,t}$, $Neg^{\text{LM}}_{j,i,t}$ - are considered as the text sentiment variables. Furthermore, the bullishness indicator for stock j on day t with the related articles $i = 1, \ldots, m$ (based on a particular lexicon) is constructed by following Antweiler and Frank (2004)

$$B_{j,t} = \log[\{1 + m^{-1} \sum_{i=1}^{m} \mathbf{1}(Pos_{j,i,t} > Neg_{j,i,t})\}/\{1 + m^{-1} \sum_{i=1}^{m} \mathbf{1}(Pos_{j,i,t} < Neg_{j,i,t})\}]. \quad (7.1)$$

We refer to Zhang et al. (2016) for more details about the data gathering and processing procedure. 63 individual stocks which are S&P 500 component stocks from 9 Global Industrial Classification Standard (GICS) sectors are considered. They are traded at NSDAQ Stock Exchange or NYSE. The list of the stock symbols and the corresponding company names can be found in Table D.1 in Appendix D in the supplementary materials.

The daily log returns $R_{j,t}$ and log volatilities $\log(\sigma^2_{j,t})$ for the stocks over the same time span are taken as response variables. More precisely, the Garman and Klass (1980) range-based measure to represent the volatility level is employed:

$$\sigma^2_{j,t} = 0.511(u_{j,t} - d_{j,t})^2 - 0.019\{r_{j,t}(u_{j,t} + d_{j,t}) - 2u_{j,t}d_{j,t}\} - 0.383r^2_{j,t}, \quad (7.2)$$

where $u_{j,t} = \log(P^H_{j,t}) - \log(P^O_{j,t}), d_{j,t} = \log(P^L_{j,t}) - \log(P^O_{j,t}), r_{j,t} = \log(P^C_{j,t}) - \log(P^O_{j,t})$, with $P^H_{j,t}, P^L_{j,t},$, $P^O_{j,t}$, and $P^C_{j,t}$ denote the highest, lowest, opening and closing prices, respectively. In addition, the S&P 500 index returns and Chicago Board Options Exchange volatility index (VIX) are included as the state variables. The financial time series data were originally obtained from Datastream, and GICS sector information was found at Compustat.

7.2 Model Setting and Results

We now construct a network model to detect the spillover effects from sentiment variables to financial variables by

$$\begin{aligned} r_{j,t} &= c_j + B_t^\top \beta_j + z_t^\top \gamma_j + r_{j,t-1}\delta_j + \varepsilon_{j,t}, \\ \log \sigma_{j,t}^2 &= c_j + B_t^\top \beta_j + z_t^\top \gamma_j + \log \sigma_{j,t-1}^2 \delta_j + \varepsilon_{j,t}, \end{aligned} \tag{7.3}$$

where $j = 1, \ldots, J$ indicate the stock symbols, $B_t = (B_{1,t}, \ldots, B_{J,t})^\top$ and z_t includes the state variables.

It is of interest to make inferences on the parameters $\beta_j \in \mathbb{R}^J$, $j = 1, \ldots J$. Following the framework introduced in Section 4, an estimation procedure with three steps needs to be implemented.

S1 For each j, run LASSO on (7.3) and keep the estimator $\widehat{\beta}_{j(-j)}^{[1]}$, $\widehat{\gamma}_j^{[1]}$, $\widehat{\delta}_j^{[1]}$ and $\widehat{c}_j^{[1]}$.

S2 For each j, run LASSO on $B_{j,t} = (B_{-j,t}^\top, z_t^\top, r_{j,t-1})^\top \theta_j + v_{j,t}$ to model the dependence among sentiment variables. In particular, we propose to take the joint penalty level obtained via block multiplier bootstrap (discussed in Section 3.2) for this regression system. Keep the residuals as $\widehat{v}_{j,t} = B_{j,t} - (B_{-j,t}^\top, z_t^\top, r_{j,t-1})^\top \widehat{\theta}_j$.

S3 For each (j,k), run IV regression of $r_{j,t} - \widehat{c}_j^{[1]} - B_{-j,t}^\top \widehat{\beta}_{j(-j)}^{[1]} - z_t^\top \widehat{\gamma}_j^{[1]} - r_{j,t-1}\widehat{\delta}_j^{[1]}$ on $B_{k,t}$ using $\widehat{v}_{k,t}$ as an instrument variable. Then we obtain the final estimator $\widehat{\beta}_{jk}^{[2]}$.

If for stock j, the sentiment variable of firm k is selected into the active set after the individual significance test i.e., the null hypothesis $H_0^{jk} : \beta_{jk} = 0$ is rejected under the block multiplier bootstrap procedure (as discussed in Section 6.1 we pre-determine $b_n = 5$ by choosing the one gives the lowest prediction norm in the LASSO estimation in S1 on a grid search), then we put a directional edge from k to j. As a result, we achieve a $0-1$ adjacency matrix describing the dependency network from sentiment variable to financial variable. Note that the diagonal elements in the matrix show the self-effect of stocks.

The graphical network for stock returns and volatility modelled by (7.3) based on BL and LM lexica (from 01/02/15 to 12/29/15) is depicted in Figures 7.1-7.2.

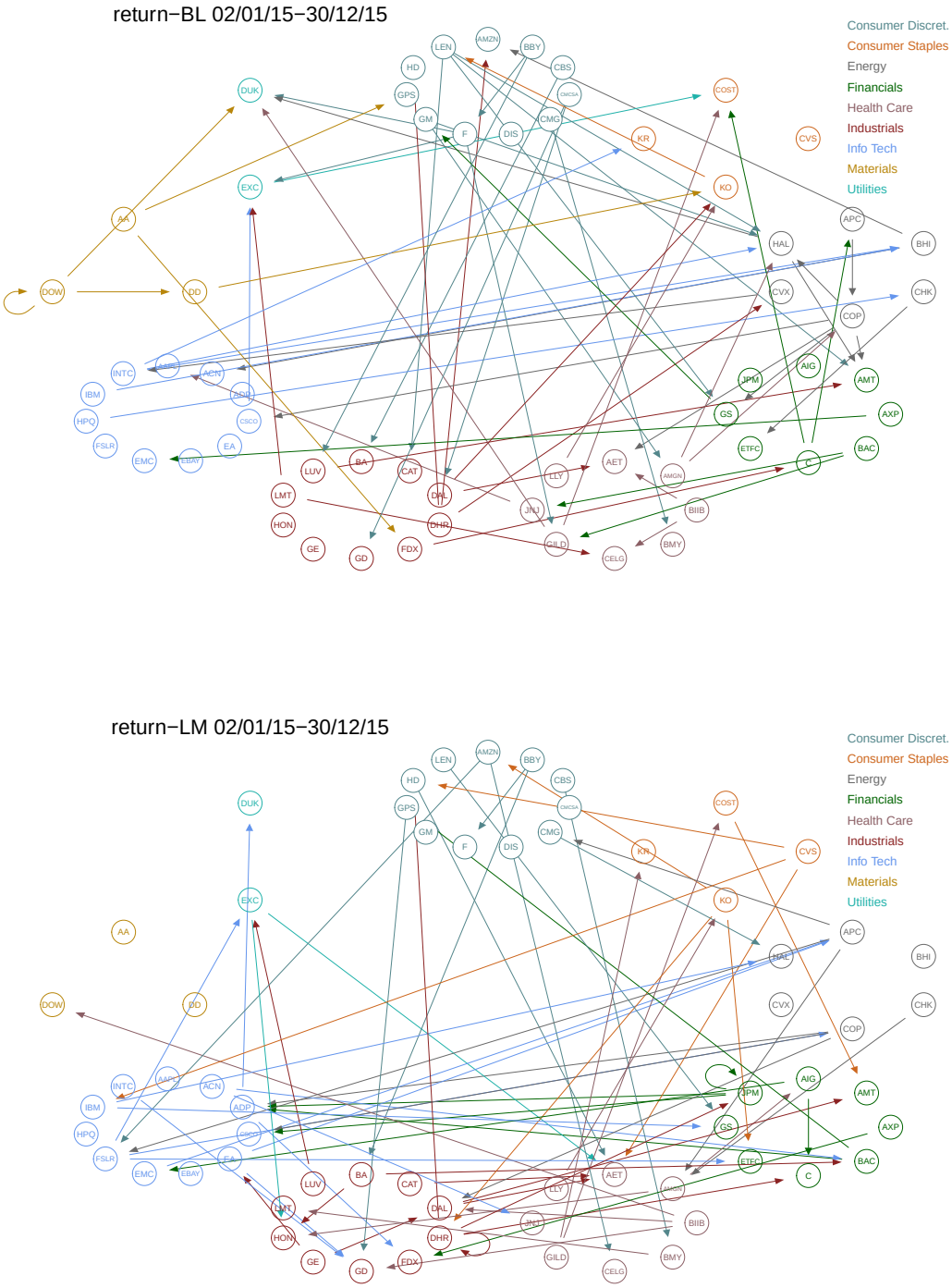

Figure 7.1: The dependency network among individual stocks from sentiment variables to return.

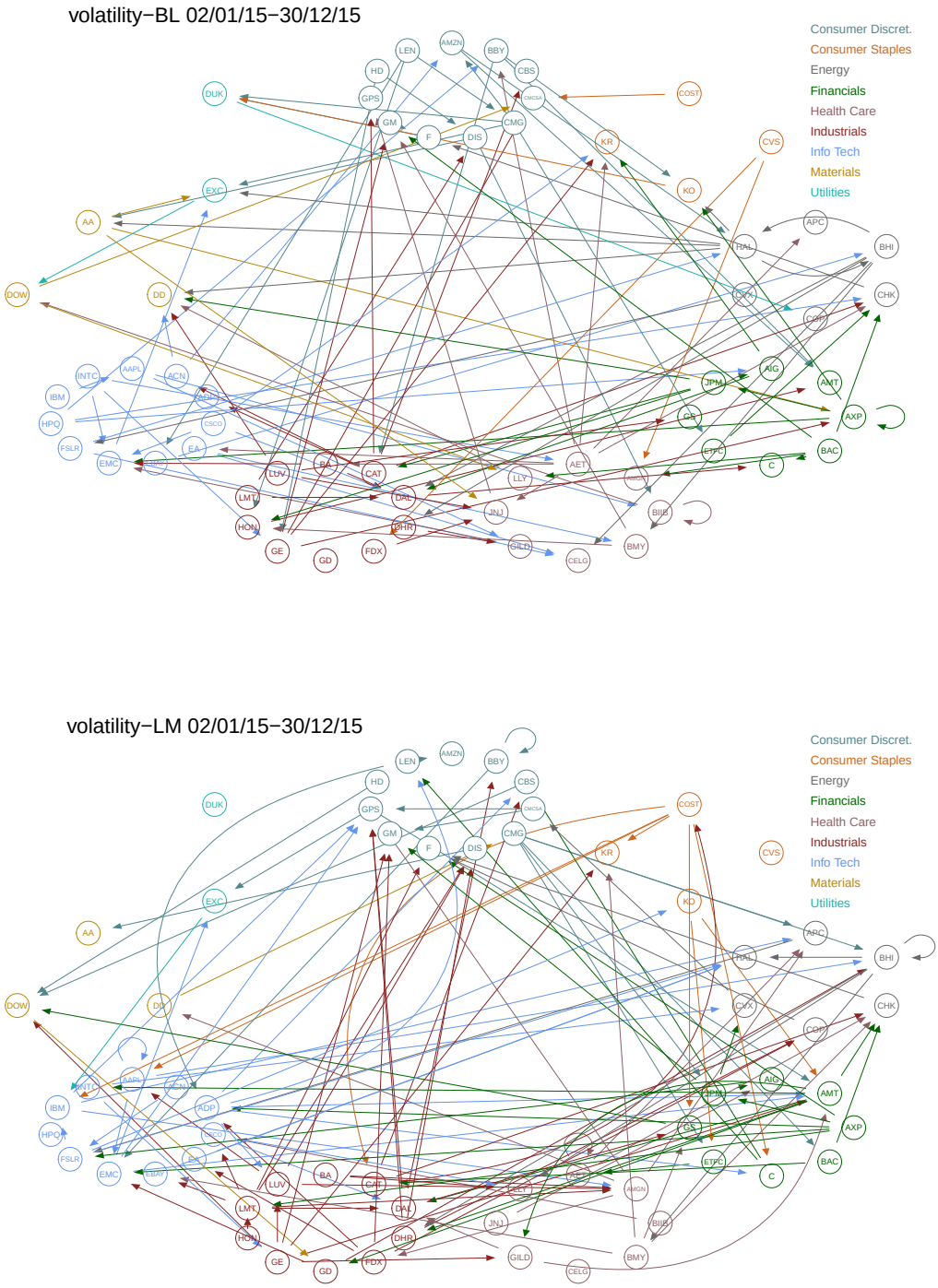

Figure 7.2: The dependency network among individual stocks from sentiment variables to volatility.

Figures 7.1-7.2 depict the dependency networks among individual stocks. Given that the time series of returns and volatility are scaled and centered before implementing the estimation procedure, we find even denser spillover effects in the volatility analysis. This indicates the stock volatility is more sensitive to sentiment than returns. Moreover, the relationships between sectors are also of interest. The simultaneous confidence region constructed via the bootstrap approach introduced in Section 4.2 may help us to detect whether the sentiment information

from one sector has joint influence on the returns of the stocks in another sector. In particular, we look at the null hypothesis: $H_0^{S_1,S_2} : \beta_{jk} = 0, \ \forall j \in S_1, k \in S_2$, where S_1 and S_2 represent two groups of stocks that belong to two sectors, respectively. The conclusion that the sentiment from sector S_2 has a joint effect on the returns or volatility of sector S_1 can be drawn if the null hypothesis is rejected with the simultaneous confidence region (4.8) under the significance level = 0.05.

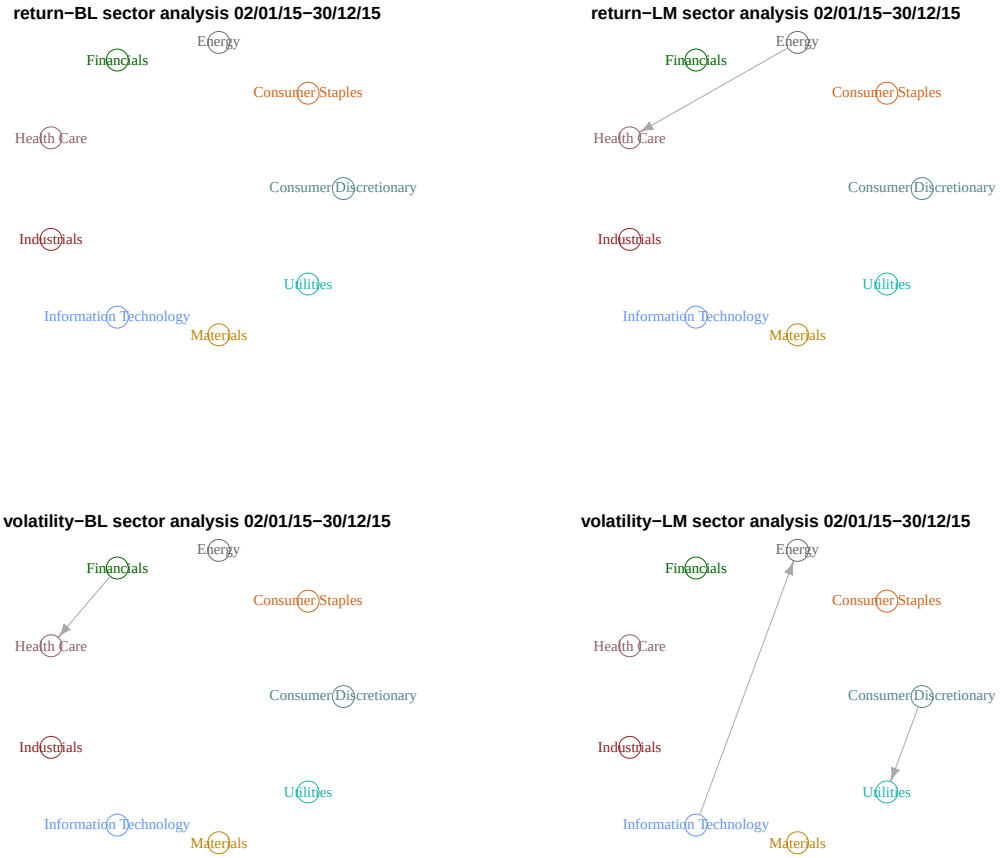

Figure 7.3: The dependency network among sectors from sentiment variables to financial variables.

Figure 7.3 describes the spillover effect network from sentiment to financial variables on the sector levels. In particular, the connections from energy to health care is found to be significant in the analysis of stock returns; while if volatility is focused on then the spillover effects from financials to health care, from information technology to energy, also from consummer discretionary to utilities are detected.

Comment 7.1 (Link to GGM)**.** Another popular way to conduct the network analysis in the literature is the GGM, which is corresponding to the estimation of a high dimensional precision matrix. And under the Gaussian assumption our SRE can be linked to a nodal wise GGM. In particular, one can estimate the coefficients in each equation of SRE by using a sparse Graphical model estimation, for example the LASSO type estimation as in Yuan and Lin (2007), and thus we build the link equation-by- equation.

Consider the VAR(1) as an example, the jth equation in the SRE is given by $Y_{j,t} = \Phi_{j\cdot} Y_{t-1} + \varepsilon_{j,t}$, where Y_t is covariance stationary with $\mathrm{Var}(Y_t) = \Gamma$ (p.d.). Correspondingly, we look at the vector $\widetilde{Y}_{j,t} = (Y_{j,t}, Y_{1,t-1}, \ldots, Y_{J,t-1})^\top$ belonging to an undirected graph (V_j, E_j) with vertex set $(1, \ldots, J+1)$. Suppose $\widetilde{Y}_{j,t} \sim \mathrm{MVN}(0, \Sigma_j)$, $\Sigma_j = \begin{bmatrix} \Gamma_{jj} & \Phi_{j\cdot}\Gamma \\ (\Phi_{j\cdot}\Gamma)^\top & \Gamma \end{bmatrix}$. Define $C_j \stackrel{\text{def}}{=} \Phi_{j\cdot}\Gamma\Phi_{j\cdot}^\top$, then we have the precision matrix as $\Theta_j = \Sigma_j^{-1} = \begin{bmatrix} (\Gamma_{jj} - C_j)^{-1} & -(\Gamma_{jj} - C_j)^{-1}\Phi_{j\cdot} \\ -\Phi_{j\cdot}^\top(\Gamma_{jj} - C_j)^{-1} & \Gamma^{-1} + \Phi_{j\cdot}^\top(\Gamma_{jj} - C_j)^{-1}\Phi_{j\cdot} \end{bmatrix}$. It can be seen that $\Phi_{jk} = 0$ would imply that the $(1, k+1)$th element of Θ_j is zero and vice versa. In addition, a LASSO type estimator proposed in Yuan and Lin (2007) can be obtained by solving

$$\widehat{\Theta}_j = \arg\max_{\Theta} \{-\log\det(\Theta) + \mathrm{trace}(S_j\Theta) + \lambda_j \sum_{\ell k} |\Theta_{\ell k}|\},$$

where $S_j \stackrel{\text{def}}{=} n^{-1}\sum_{t=1}^n \widetilde{Y}_{j,t}\widetilde{Y}_{j,t}^\top$.

In an unreported simulation study we compare the estimation performance between our proposed approach and the nodal wise GGM under the VAR(1) model. The results show that the nodal wise GGM which is approximated to SRE has worse prediction performance than our method, which can be obtained from the authors upon request.

Supplementary Material

A Theorems for Joint Penalty over Equations

Recall that the theoretical choice $\lambda^0(1-\alpha)$ is defined as the $(1-\alpha)$ quantile of $\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} 2c\sqrt{n}|S_{jk}/\Psi_{jk}|$. First, we provide the analogue results of Theorem 5.1 and Corollary 5.1.

Theorem A.1. *Under (A1) and (A3), we have*

$$\begin{aligned}\mathrm{P}(2c\sqrt{n}\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J}|S_{jk}/\Psi_{jk}|\geqslant r) \leqslant & C_1\varpi_n n r^{-q}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}^q}{\Psi_{jk}^q} \\ & + C_2\sum_{j=1}^{J}\sum_{k=1}^{K}\exp\Big(\frac{-C_3 r^2\Psi_{jk}^2}{n\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}^2}\Big), \qquad (A.1)\end{aligned}$$

where for $\varsigma > 1/2-1/q$ (weak dependence case), $\varpi_n = 1$; for $\varsigma < 1/2-1/q$ (strong dependence case), $\varpi_n = n^{q/2-1-\varsigma q}$. C_1, C_2, C_3 are constants depending on q and ς.

Corollary A.1 (Bound for $\lambda^0(1-\alpha)$ and Oracle Inequalities under IC)**.** *Under (A1) and (A3), given $\lambda^0(1-\alpha)$ satisfies*

$$\lambda^0(1-\alpha)\lesssim \max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J}\Big\{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}\sqrt{n\log(KJ/\alpha)}\vee\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}(n\varpi_n KJ/\alpha)^{1/q}\Big\}, \quad (A.2)$$

additionally assume that the RE condition (A2) holds uniformly over equations $j=1,\ldots,J$ with probability $1-o(1)$, and under the exact sparsity assumption (3.1), then $\widehat{\beta}_j$ obtained from (3.2) under IC satisfy

$$|\widehat{\beta}_j-\beta_j^0|_{j,pr}\lesssim C\sqrt{s}\max_{1\leqslant k\leqslant K}\Psi_{jk}\max_{1\leqslant j\leqslant J}\Big\{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}\frac{\sqrt{\log(KJ/\alpha)}}{\sqrt{n}}\vee\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}n^{1/q-1}(\varpi_n KJ/\alpha)^{1/q}\Big\}, \quad (A.3)$$

with probability $1-\alpha-o(1)$, where for $\varsigma > 1/2-1/q$ (weak dependence case), $\varpi_n = 1$; for $\varsigma < 1/2-1/q$ (strong dependence case), $\varpi_n = n^{q/2-1-\varsigma q}$, and the constant C depends on the RE constants.

The other empirical choices of the joint penalty level can be:

a) $Q(1-\alpha)$: the $(1-\alpha)$ quantile of $2c\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J}\sqrt{n}|Z_{jk}/\Psi_{jk}|$. In practice, one can take an alternative choice such that $\widetilde{Q}(1-\alpha)\stackrel{\mathrm{def}}{=}2c\sqrt{n}\Phi^{-1}\{1-\alpha/(2KJ)\}$.

b) $\Lambda(1-\alpha)\stackrel{\mathrm{def}}{=}2c\sqrt{n}q^{[B]}_{(1-\alpha)}$, where $q^{[B]}_{(1-\alpha)}$ is the $(1-\alpha)$ quantile of $\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J}|Z^{[B]}_{jk}/\Psi_{jk}|$.

For a) again we need the Gaussian approximation results for the vectorized process $\widetilde{\mathcal{S}}\stackrel{\mathrm{def}}{=}\mathrm{vec}[\{(S_{jk})_{k=1}^K\}_{j=1}^J]=\frac{1}{\sqrt{n}}\sum_{t=1}^n\widetilde{\mathcal{X}}_t$, where $\widetilde{\mathcal{X}}_t\stackrel{\mathrm{def}}{=}\mathrm{vec}[\{(X_{jk,t}\varepsilon_{j,t})_{k=1}^K\}_{j=1}^J]$ similar to Theorem 5.2 and Corollary 5.2 to justify the choice of λ as $Q(1-\alpha)$.

Let $\mathcal{X}_t \overset{\text{def}}{=} \text{vec}[\{(X_{jk,t})_{k=1}^K\}_{j=1}^J]$. We first aggregate the dependence adjusted norm over $j = 1, \ldots, J$ and $k = 1, \ldots, K$:

$$\||\mathcal{X}_\cdot|_\infty\|_{q,\varsigma} \overset{\text{def}}{=} \sup_{m\geqslant 0}(m+1)^\varsigma \sum_{t=m}^{\infty} \delta_{q,t},\ \delta_{q,t} \overset{\text{def}}{=} \||\mathcal{X}_t - \mathcal{X}_t^*|_\infty\|_q, \tag{A.4}$$

where $q \geqslant 1$, and $\varsigma > 0$. Moreover, define the following quantities

$$\Phi_{q,\varsigma} \overset{\text{def}}{=} 2 \max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} \|X_{jk,\cdot}\|_{q,\varsigma}\|\varepsilon_{j,\cdot}\|_{q,\varsigma},\ \Gamma_{q,\varsigma} \overset{\text{def}}{=} 2\Big(\sum_{j=1}^J \|\varepsilon_{j,\cdot}\|_{q,\varsigma}^{q/2}\Big)^{2/q}\Big(\sum_{k=1}^K\sum_{j=1}^J \|X_{jk,\cdot}\|_{q,\varsigma}^{q/2}\Big)^{2/q}$$

$$\Theta_{q,\varsigma} \overset{\text{def}}{=} \Gamma_{q,\varsigma} \wedge \{\||\mathcal{X}_\cdot|_\infty\|_{q,\varsigma}\|\varepsilon_{j,\cdot}\|_{q,\varsigma}(\log KJ)^{3/2}\}. \tag{A.5}$$

Let $L_1 = [\Phi_{4,\varsigma}\Phi_{4,0}\{\log(KJ)\}^2]^{1/\varsigma}$, $W_1 = (\Phi_{6,0}^6 + \Phi_{8,0}^4)\{\log(KJn)\}^7$, $W_2 = \Phi_{4,\varsigma}^2\{\log(KJn)\}^4$, $W_3 = [n^{-\varsigma}\{\log(KJn)\}^{3/2}\Theta_{2q,\varsigma}]^{1/(1/2-\varsigma-1/q)}$, $N_1 = \{n/\log(KJ)\}^{q/2}\Theta_{2q,\varsigma}^q$, $N_2 = n\{\log(KJ)\}^{-2}\Phi_{4,\varsigma}^{-2}$, $N_3 = [n^{1/2}\{\log(KJ)\}^{-1/2}\Theta_{2q,\varsigma}^{-1}]^{1/(1/2-\varsigma)}$.

(A4') i) (weak dependency case) Given $\Theta_{2q,\varsigma} < \infty$ with $q \geqslant 4$ and $\varsigma > 1/2 - 1/q$, then $\Theta_{2q,\varsigma} n^{1/q-1/2}\{\log(KJn)\}^{3/2} \to 0$ and $L_1 \max(W_1, W_2) = o(1)\min(N_1, N_2)$.
ii) (strong dependency case) Given $0 < \varsigma < 1/2 - 1/q$, then $\Theta_{2q,\varsigma}\{\log(KJ)\}^{1/2} = o(n^\varsigma)$ and $L_1 \max(W_1, W_2, W_3) = o(1)\min(N_2, N_3)$.

Theorem A.2. *Under (A1), (A3) and (A4'), for each $k = 1, \ldots, K$, $j = 1, \ldots, J$ assume that there exists a constant $c > 0$ such that $\min_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} \text{avar}(S_{jk}) \geqslant c$, then we have*

$$\rho(D^{-1}\widetilde{\mathcal{S}}, D^{-1}\widetilde{\mathcal{Z}}) \to 0, \quad \text{as } n \to \infty, \tag{A.6}$$

where $\widetilde{\mathcal{Z}} \sim \text{N}(0, \Sigma_{\widetilde{\mathcal{X}}})$, $\Sigma_{\widetilde{\mathcal{X}}}$ is the $JK \times JK$ long-run variance-covariance matrix of $\widetilde{\mathcal{X}}_t$, and D is a diagonal matrix with the square root of the diagonal elements of $\Sigma_{\widetilde{\mathcal{X}}}$, namely

$$\Big\{\sum_{\ell=-\infty}^{\infty} \mathsf{E}(X_{jk,t}X_{jk,(t-\ell)}\varepsilon_{j,t}\varepsilon_{j,(t-\ell)})\Big\}^{1/2} = \sqrt{\text{avar}(S_{jk})},\ \text{for } k = 1, \ldots, K, j = 1, \ldots, J.$$

Corollary A.2. *Under the conditions of Theorem A.2, we have*

$$\sup_{\alpha\in(0,1)} |\,\text{P}\{\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} 2c\sqrt{n}|S_{jk}/\Psi_{jk}| \geqslant Q(1-\alpha)\} - \alpha| \to 0, \quad \text{as } n \to \infty. \tag{A.7}$$

Corollary A.3. *Under the conditions of Theorem A.2, and assume $\Phi_{2q,\varsigma} < \infty$ with $q > 4$, $b_n = \mathcal{O}(n^\eta)$ for some $0 < \eta < 1$. Let $F_\varsigma = n$, for $\varsigma > 1 - 2/q$; $F_\varsigma = l_n b_n^{q/2-\varsigma q/2}$, for $1/2 - 2/q < \varsigma < 1 - 2/q$; $F_\varsigma = l_n^{q/4-\varsigma q/2} b_n^{q/2-\varsigma q/2}$, for $\varsigma < 1/2 - 2/q$. Given $n^{-1}\{\log(KJ)\}^2 \max\{n^{1/2}b_n^{1/2}\Phi_{2q,\varsigma}^2, n^{1/2}b_n^{1/2}\sqrt{\log(KJ)}\Phi_{8,\varsigma}^2, F_\varsigma^{2/q}\Gamma_{2q,\varsigma}^2(KJ)^{2/q}, \Phi_{2,0}\Phi_{2,\varsigma}v'(b_n)n/\sqrt{\log(KJ)}\} = o(1)$, where $v'(b_n) = (b_n+1)^{-\varsigma} + 2v_{n,2}/b_n$, $v_{n,2} = \log b_n$ (resp. $b_n^{-\varsigma+1}$ or 1) for $\varsigma = 1$ (resp. $\varsigma < 1$ or $\varsigma > 1$), then we have*

$$\rho(\widehat{D}^{-1}\widetilde{\mathcal{S}}, D^{-1}\widetilde{\mathcal{Z}}) \to 0, \quad \text{as } n \to \infty, \tag{A.8}$$

where $\widehat{D} = \{\text{diag}(\widehat{\Sigma}_{\widetilde{\mathcal{X}}})\}^{1/2}$, $\widehat{\Sigma}_{\widetilde{\mathcal{X}}} = \frac{1}{b_n l_n}\sum_{i=1}^{l_n}(\sum_{l=(i-1)b_n+1}^{ib_n}\widetilde{\mathcal{X}}_l)(\sum_{l=(i-1)b_n+1}^{ib_n}\widetilde{\mathcal{X}}_l)^\top$.

Similar to Corollary 5.4, we can provide a refined bound for $\lambda^0(1-\alpha)$ and also the oracle inequalities under IC as follows.

Corollary A.4 (Bounds for $\lambda^0(1-\alpha)$ and Oracle Inequalities under IC with Gaussian Approximation Results)**.** *Under the conditions of Theorem A.2, suppose* $2\{\log(KJ)\}^{-1/2}+\rho(D^{-1}\widetilde{\mathcal{S}}, D^{-1}\widetilde{\mathcal{Z}}) = o(\alpha)$ *and let* $Z_\alpha \geqslant 2\tilde{c}\sqrt{n\log(KJ)}$*, where* $\tilde{c}$ *is no less than the* c *in the definition of* $\lambda^0(1-\alpha)$*, then we have* $\lambda^0(1-\alpha)$ *satisfying*

$$\lambda^0(1-\alpha) \leqslant Z_\alpha, \tag{A.9}$$

additionally assume that the RE condition (A2) holds uniformly over equations $j = 1, \ldots, J$ *with probability* $1 - o(1)$*, and given the exact sparsity assumption* (3.1)*, then* $\widehat{\beta}_j$ *obtained from* (3.2) *under IC satisfies*

$$|\widehat{\beta}_j - \beta_j^0|_{j,pr} \lesssim C\sqrt{s} \max_{1\leqslant k\leqslant K} \Psi_{jk} \frac{\sqrt{\log(KJ)}}{\sqrt{n}}, \tag{A.10}$$

with probability $1-\alpha-o(1)$*, where the constant* C *depends on the RE constants.*

Next, we need to show the validity of b). Let $\widetilde{\mathcal{Z}}^{[B]} \stackrel{\text{def}}{=} \text{vec}[\{(Z_{jk}^{[B]})_{k=1}^K\}_{j=1}^J]$ and $\widetilde{\Psi} \stackrel{\text{def}}{=} \text{vec}[\{(\Psi_{jk})_{k=1}^K\}_{j=1}^J]$. Similarly to Theorem 5.3 we have the following results:

Theorem A.3. *Under (A1), (A3), and assume* $\Phi_{2q,\varsigma} < \infty$ *with* $q > 4$*,* $b_n = \mathcal{O}(n^\eta)$ *for some* $0 < \eta < 1$ *(the detailed rate is calculated in* (B.3)*), then*

$$\tilde{\rho}_n \stackrel{\text{def}}{=} \sup_{r\in\mathbb{R}} |\,\mathrm{P}(|\widetilde{\mathcal{Z}}^{[B]}/\widetilde{\Psi}|_\infty \leqslant r|\mathcal{X}., \varepsilon.) - \mathrm{P}(|\widetilde{\mathcal{Z}}/\widetilde{\Psi}|_\infty \leqslant r)| \to 0, \text{ as } n\to\infty, \tag{A.11}$$

and

$$\sup_{\alpha\in(0,1)} |\,\mathrm{P}(|\widetilde{\mathcal{S}}/\widetilde{\Psi}|_\infty \geqslant q_{(1-\alpha)}^{[B]}) - \alpha| \to 0, \text{ as } n\to\infty. \tag{A.12}$$

Lastly, we show the performance bounds for the OLS post-LASSO estimator in the following theorem.

For each $p \leqslant n$, $\widetilde{T}_j \subset \{1, \cdots, K\}$, $|\widetilde{T}_j \backslash T_j| \leqslant p$, we define the class of functions $\mathcal{G}_{\widetilde{T}_j} = \{\varepsilon_{j,t} X_{j,t}^\top \delta/|\delta|_{j,pr}, \text{supp}(\delta) \subseteq \widetilde{T}_j, |\delta|_2 = 1\}$. The covering number of the function class is given by $\sup_{\mathcal{Q}} \mathcal{N}(\epsilon, \mathcal{G}_{\widetilde{T}_j}, \|\cdot\|_{\mathcal{Q},1})$. Also define $\mathcal{F}_{j,p} = \{\mathcal{G}_{\widetilde{T}_j} : \widetilde{T}_j \subset \{1, \cdots, K\}, |\widetilde{T}_j \backslash T_j| \leqslant p\}$. For any $f \in \mathcal{F}_{j,p}$, there exists a set $F_{j,p}$ such that $\min_{f'\in F_{j,p}} \|f - f'\|_{\mathcal{Q},1} \leqslant \epsilon$, and the cardinality of the set is denoted by $|F_{j,p}|$. Consider the vector ϑ_t of length $|F_{j,p}|$, such that for $l = 1, \ldots, |F_{j,p}|$, there is $\vartheta_{l,t} = (f - \mathsf{E}\, f)/\psi_f$ with $\psi_f = \{\text{avar}(G_n(f))\}^{1/2}$, corresponding to each $f \in F_{j,p}$. The aggregated dependence adjusted norm is given by

$$\|\vartheta.\|_{q,\varsigma} \stackrel{\text{def}}{=} \sup_{m\geqslant 0} (m+1)^\varsigma \sum_{t=m}^\infty \||\vartheta_t - \vartheta_t^*|_\infty\|_q, \tag{A.13}$$

where $q \geqslant 1$, and $\varsigma > 0$. Moreover, define the following quantities (for simplicity we drop the

subscripts j, p)

$$\Phi^{\vartheta}_{q,\varsigma} \stackrel{\text{def}}{=} \max_{1\leqslant l\leqslant |F_{j,p}|} \|\vartheta_{l,\cdot}\|_{q,\varsigma},\ \Gamma^{\vartheta}_{q,\varsigma} \stackrel{\text{def}}{=} \Big(\sum_{l=1}^{|F_{j,p}|} \|\vartheta_{l,\cdot}\|^q_{q,\varsigma} \Big)^{1/q},$$
$$\Theta^{\vartheta}_{q,\varsigma} \stackrel{\text{def}}{=} \Gamma^{\vartheta}_{q,\varsigma} \wedge \{\|\vartheta_{\cdot}\|_{q,\varsigma}(\log|F_{j,p}|)^{3/2}\}. \tag{A.14}$$

To evoke the Gaussian approximation on $G_n(f)/\psi_f$, we need to impose the following assumptions additionally. Define $L^{\vartheta}_1 = \{\Phi^{\vartheta}_{2,\varsigma}\Phi^{\vartheta}_{2,0}(\log|F_{j,p}|)^2\}^{1/\varsigma}$, $W^{\vartheta}_1 = \{(\Phi^{\vartheta}_{3,0})^6 + (\Phi^{\vartheta}_{4,0})^4\}\{\log(|F_{j,p}|n)\}^7$, $W^{\vartheta}_2 = (\Phi^{\vartheta}_{2,\varsigma})^2\{\log(|F_{j,p}|n)\}^4$, $W^{\vartheta}_3 = [n^{-\varsigma}\{\log(|F_{j,p}|n)\}^{3/2}\Theta^{\vartheta}_{q,\varsigma}]^{1/(1/2-\varsigma-1/q)}$, $N^{\vartheta}_1 = (n/\log|F_{j,p}|)^{q/2}(\Theta^{\vartheta}_{q,\varsigma})^q$, $N^{\vartheta}_2 = n(\log|F_{j,p}|)^{-2}(\Phi^{\vartheta}_{2,\varsigma})^{-2}$, $N^{\vartheta}_3 = \{n^{1/2}(\log|F_{j,p}|)^{-1/2}(\Theta^{\vartheta}_{q,\varsigma}\})^{1/(1/2-\varsigma)}$.

(A7) i) (weak dependency case) Given $\Theta^{\vartheta}_{q,\varsigma} < \infty$ with $q \geqslant 2$ and $\varsigma > 1/2 - 1/q$, then $\Theta^{\vartheta}_{q,\varsigma} n^{1/q-1/2}\{\log(|F_{j,p}|n)\}^{3/2} \to 0$ and $L_1 \max(W^{\vartheta}_1, W^{\vartheta}_2) = o(1)\min(N^{\vartheta}_1, N^{\vartheta}_2)$.
ii) (strong dependency case) Given $0 < \varsigma < 1/2 - 1/q$, then $\Theta^{\vartheta}_{q,\varsigma}(\log|F_{j,p}|)^{1/2} = o(n^{\varsigma})$ and $L_1 \max(W^{\vartheta}_1, W^{\vartheta}_2, W^{\vartheta}_3) = o(1)\min(N^{\vartheta}_2, N^{\vartheta}_3)$.

Comment A.1. For a random vector $z_t \in R^K$, suppose there exist constants $C, D > 0$, such that $\max_k \mathsf{E}\{\exp(|z_{k,t}/D|^q)\} \leqslant C$. Then by Jensen's inequality it follows that $\||z_t|_\infty\|_q \leqslant D(\log K + \log C)^{1/q}$. In particular, for the case of sub-Gaussian random variables, there exists constant $D > 0$ such that $\mathsf{E}\{\exp(|z_{k,t}/D|^2)\} - 1 \leqslant 1$, which implies $\||z_t|_\infty\|_2 \lesssim D\sqrt{\log K}$.

Similar to the discussion in Remark 5.6, consider the case with $\Theta^{\vartheta}_{q,\varsigma} = \mathcal{O}((\log|F_{j,p}|)^{1/q})$ and $\Phi^{\vartheta}_{q,\varsigma} = \mathcal{O}(1)$, where $\varsigma > 1/2 - 1/q$. Then $\Theta^{\vartheta}_{q,\varsigma} n^{1/q-1/2}\{\log(|F_{j,p}|n)\}^{3/2} \to 0$ becomes $\log|F_{j,p}|\{\log(n|F_{j,p}|)\}^{3q/2} = o(n^{q/2-1})$, which implies that $L^{\vartheta}_1 \max(W^{\vartheta}_1, W^{\vartheta}_2) = o(1)\min(N^{\vartheta}_1, N^{\vartheta}_2)$.

As shown in the proof of Theorem A.4, $|F_{j,p}| \lesssim K^p(6\mu_j(p)\sigma/\epsilon)^{s+p}$ with $\epsilon = \sqrt{p\log K + (p+s)\log(6\mu_j(p)\sigma)}(4\sqrt{n})^{-1}$. This means with (A7), the dimension K has to satisfy the condition $\{p\log K + (s+p)\log(\sqrt{n})\}^{1+3q/2} = o(n^{q/2-1})$, where we consider the case such that $|F_{j,p}|$ is larger than n.

Theorem A.4 (Prediction Performance Bounds for OLS Post-LASSO)**.** *Given (A1), (A3) and (A7), suppose (A2) (with $\bar{c} = \frac{c+1}{c-1}, c > 1$) and (A5) (with $\widehat{p}_j = |\widehat{T}_j \setminus T_j|$) hold uniformly over equations with probability $1 - o(1)$, then under the exact sparsity assumption* (3.1)*, for any $\tau > 0$, there is a constant C_τ independent of n, for all $j = 1, \ldots, J$ we have*

$$|\widehat{\beta}^{[P]}_j - \beta^0_j|_{j,pr} \leqslant C_\tau \max_{1\leqslant k\leqslant K} \Psi_{jk} \sqrt{\frac{p\log K + (p+s)\{\log(6\mu_j(p)\sigma) + \log n/2\}}{n}}$$
$$+ \mathbf{1}(T_j \not\subseteq \widehat{T}_j) C\sqrt{s} \max_{1\leqslant k\leqslant K} \Psi_{jk} \max_{1\leqslant j\leqslant J} \{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma} \frac{\sqrt{\log(KJ/\alpha)}}{\sqrt{n}} \vee \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} n^{1/q-1}(\varpi_n KJ/\alpha)^{1/q}\}, \tag{A.15}$$

with probability $1 - \alpha - \tau - o(1)$, where for $\varsigma > 1/2 - 1/q$ (weak dependence case), $\varpi_n = 1$; for $\varsigma < 1/2 - 1/q$ (strong dependence case), $\varpi_n = n^{q/2-1-\varsigma q}$. $\sigma = \max_j\{\text{avar}(n^{-1/2}\sum_{t=1}^n \varepsilon_{j,t})\}^{1/2}$ and the constant C depends on the RE constants.

In particular, suppose the Gaussian approximation results hold for $\lambda^0(1-\alpha)$, the bound for it can be replaced according to Corollary A.4.

B Detailed Proofs

B.1 Proofs of Single Equation Estimation

Proof of Theorem 5.1. For each $j = 1, \dots J$, $k = 1, \dots, K$, applying Theorem 2 of Wu and Wu (2016) gives

$$\mathrm{P}(\sqrt{n}|S_{jk}| \geqslant x) \leqslant \frac{C_1' \varpi_n n \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}^q}{x^q} + C_2' \exp\big(\frac{-C_3 x^2}{n\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}^2}\big),$$

where for $\varsigma > 1/2 - 1/q$, $\varpi_n = 1$; for $\varsigma < 1/2 - 1/q$, $\varpi_n = n^{q/2-1-\varsigma q}$. C_1', C_2', C_3 are three constants depending on q and ς. It follows that the conclusion holds if we set $x = (2c)^{-1}\Psi_{jk} r$. □

Proof of Theorem 5.2. According to the Minkowski's inequality and Hölder's inequality, we have

$$\begin{aligned}\sum_{t=m}^{\infty} \|X_{jk,t}\varepsilon_{j,t} - X_{jk,t}^*\varepsilon_{j,t}^*\|_q &\leqslant \sum_{t=m}^{\infty} \{\|X_{jk,t}(\varepsilon_{j,t} - \varepsilon_{j,t}^*)\|_q + \|(X_{jk,t} - X_{jk,t}^*)\varepsilon_{j,t}^*\|_q\} \\ &\leqslant \sum_{t=m}^{\infty} \{\|X_{jk,t}\|_{2q}\|\varepsilon_{j,t} - \varepsilon_{j,t}^*\|_{2q} + \|X_{jk,t} - X_{jk,t}^*\|_{2q}\|\varepsilon_{j,t}\|_{2q}\}.\end{aligned}$$

Thus, it is easy to see that

$$\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \leqslant \|X_{jk,\cdot}\|_{2q,0}\|\varepsilon_{j,\cdot}\|_{2q,\varsigma} + \|X_{jk,\cdot}\|_{2q,\varsigma}\|\varepsilon_{j,\cdot}\|_{2q,0} \leqslant 2\|X_{jk,\cdot}\|_{2q,\varsigma}\|\varepsilon_{j,\cdot}\|_{2q,\varsigma}.$$

Consequently, we have the following relationships:

$$\begin{aligned}&\max_{1\leqslant k\leqslant K} \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \leqslant 2 \max_{1\leqslant k\leqslant K} \|X_{jk,\cdot}\|_{2q,\varsigma}\|\varepsilon_{j,\cdot}\|_{2q,\varsigma}, \\ &(\sum_{k=1}^{K} \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}^q)^{1/q} \leqslant 2\|\varepsilon_{j,\cdot}\|_{2q,\varsigma}(\sum_{k=1}^{K} \|X_{jk,\cdot}\|_{2q,\varsigma}^q)^{1/q}, \\ &\|X_{j,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \leqslant 2\|X_{j,\cdot}\|_{2q,\varsigma}\|\varepsilon_{j,\cdot}\|_{2q,\varsigma}.\end{aligned}$$

Therefore, the conditions in Theorem 3.2 of Zhang and Wu (2017a) can be verified for the K-dimensional stationary process $X_{j,t}\varepsilon_{j,t}$. Finally, applying that theorem yields the Gaussian approximation results. □

Proof of Corollary 5.2. It follows directly from the Gaussian approximation results in Theorem 5.2. □

Proof of Corollary 5.3. The proof follows that of Corollary 5.4 in Zhang and Wu (2017a). For $w > 0$, we have

$$\begin{aligned}\rho(\widehat{D}_j^{-1}S_{j\cdot}, D_j^{-1}Z_j) &= \sup_{r\geqslant 0} |\,\mathrm{P}(|\widehat{D}_j^{-1}S_{j\cdot}|_\infty \geqslant r) - \mathrm{P}(|D_j^{-1}Z_j|_\infty \geqslant r)| \\ &\leqslant \rho(D_j^{-1}S_{j\cdot}, D_j^{-1}Z_j) + \sup_{r\geqslant 0} \mathrm{P}(||D_j^{-1}Z_j|_\infty - r| \leqslant w) + \mathrm{P}(|(D_j^{-1} - \widehat{D}_j^{-1})S_{j\cdot}|_\infty \geqslant w) \\ &\lesssim \rho(D_j^{-1}S_{j\cdot}, D_j^{-1}Z_j) + w\sqrt{\log K} + \mathrm{P}(|(D_j^{-1} - \widehat{D}_j^{-1})S_{j\cdot}|_\infty \geqslant w),\end{aligned}$$

where the last line uses the arguments of Theorem 3 in Chernozhukov et al. (2015). Let $V_{n,j} \stackrel{\text{def}}{=} \max_{1\leqslant k\leqslant K}|\Psi_{jk}/\widehat{\Psi}_{jk}-1|$ and $L_{n,j} \stackrel{\text{def}}{=} \max_{1\leqslant k\leqslant K}|\Psi_{jk}^2-\widehat{\Psi}_{jk}^2|$. Then $|(D_j^{-1}-\widehat{D}_j^{-1})S_{j\cdot}|_\infty \leqslant V_{n,j}|D_j^{-1}S_{j\cdot}|_\infty$. As $\min_{1\leqslant k\leqslant K}\widehat{\Psi}_{jk}^2 \geqslant c_j$, let $w=xy$, $0<x<c_j/2$, $y>0$, then

$$\begin{aligned}\mathrm{P}(|(D_j^{-1}-\widehat{D}_j^{-1})S_{j\cdot}|_\infty \geqslant w) &\leqslant \mathrm{P}(V_{n,j}\geqslant 2x/c_j)+\mathrm{P}(|D_j^{-1}S_{j\cdot}|_\infty \geqslant c_jy/2)\\ &\leqslant \mathrm{P}(L_{n,j}\geqslant x)+\rho(D_j^{-1}S_{j\cdot},D_j^{-1}Z_j)+\mathrm{P}(|D_j^{-1}Z_j|_\infty\geqslant c_jy/2).\end{aligned}$$

It follows that

$$\rho(\widehat{D}_j^{-1}S_{j\cdot},D_j^{-1}Z_j)\leqslant \rho(D_j^{-1}S_{j\cdot},D_j^{-1}Z_j)+xy\sqrt{\log K}+\mathrm{P}(L_{n,j}\geqslant x)+\mathrm{P}(|D_j^{-1}Z_j|_\infty\geqslant c_jy/2).$$

In particular, $L_{n,j}\leqslant L_{n,j,1}+L_{n,j,2}$, with $L_{n,j,1}=\max_{1\leqslant k\leqslant K}|\Psi_{jk}^2-\mathsf{E}\,\widehat{\Psi}_{jk}^2|$ and $L_{n,j,2}=\max_{1\leqslant k\leqslant K}|\,\mathsf{E}\,\widehat{\Psi}_{jk}^2-\widehat{\Psi}_{jk}^2|$.

As for $L_{n,j,1}$, applying Theorem 5.1 of Zhang and Wu (2017a), for $u\geqslant n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2$, we have

$$\mathrm{P}(nL_{n,j,1}\geqslant u)\lesssim \frac{F_\varsigma\Gamma_{j,2q,\varsigma}^q}{u^{q/2}}+K\exp\Big(-\frac{C_ju^2}{nb_n\Phi_{j,8,\varsigma}^4}\Big),$$

where the constants C_j depend on η, q, and ς. Then we have $\mathrm{P}(L_{n,j,1}>x)\to 0$, as $n\to\infty$, if we set $x=\frac{\sqrt{\log K}}{n}\max\{n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2, n^{1/2}b_n^{1/2}\sqrt{\log K}\Phi_{j,8,\varsigma}^2, F_\varsigma^{2/q}\Gamma_{j,2q,\varsigma}^2\}$.

For $L_{n,j,2}$, define $v'(b_n)=(b_n+1)^{-\varsigma}+2v_{n,2}/b_n$, $v_{n,2}=\log b_n$ (resp. $b_n^{-\varsigma+1}$ or 1) for $\varsigma=1$ (resp. $\varsigma<1$ or $\varsigma>1$). It can be shown that $L_{n,j,2}\leqslant \Phi_{j,2,0}\Phi_{j,2,\varsigma}v'(b_n)$. Note that $v'(b_n)$ is a special case of $v(b_n)$ in the proof of Theorem 5.3 given $n\to\infty$, and the conclusion follows similarly.

It follows that $\mathrm{P}(L_{n,j}>x)\to 0$, as $n\to\infty$, if we set

$$x=\frac{\sqrt{\log K}}{n}\max\{n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2, n^{1/2}b_n^{1/2}\sqrt{\log K}\Phi_{j,8,\varsigma}^2, F_\varsigma^{2/q}\Gamma_{j,2q,\varsigma}^2, \Phi_{j,2,0}\Phi_{j,2,\varsigma}v'(b_n)n/\sqrt{\log K}\}.$$

Moreover, given Theorem 5.2 and choosing $y=C\sqrt{\log K}$ (the constant $C>0$ is sufficiently large) yields the conclusion. □

Proof of Corollary 5.4. Let $\tilde{\rho}_n \stackrel{\text{def}}{=} \rho(D_j^{-1}S_{j\cdot},D_j^{-1}Z_j)$ and by its definition, we have

$$\begin{aligned}\mathrm{P}(2c\sqrt{n}\max_{1\leqslant k\leqslant K}|S_{jk}/\Psi_{jk}|\leqslant Z_\alpha) &\geqslant \mathrm{P}(2c\sqrt{n}\max_{1\leqslant k\leqslant K}|Z_{jk}/\Psi_{jk}||\leqslant Z_\alpha)-\tilde{\rho}_n\\ &\geqslant 1-\sum_{k=1}^K\mathrm{P}\{|Z_{jk}/\Psi_{jk}|\geqslant Z_\alpha/(2c\sqrt{n})\}-\tilde{\rho}_n\\ &\geqslant 1-\sum_{k=1}^K 2\{Z_\alpha/(2c\sqrt{n})\}^{-1}\exp[-Z_\alpha^2/\{2(2c\sqrt{n})^2\}]-\tilde{\rho}_n\\ &\geqslant 1-2(\log K)^{-1/2}-\tilde{\rho}_n,\end{aligned}$$

where we have applied the union bound, the tail probability of Gaussian random variable and the fact that $Z_\alpha=2\tilde{c}\sqrt{n\log K}\geqslant 2c\sqrt{n\log K}$ ($\tilde{c}$ is no less than the c in the definition of $\lambda_j^0(1-\alpha)$).

It follows that $\lambda_j^0(1-\alpha)\leqslant Z_\alpha$ as $1-\alpha=\mathrm{P}\{2c\sqrt{n}\max_{1\leqslant k\leqslant K}|S_{jk}/\Psi_{jk}|\leqslant\lambda_j^0(1-\alpha)\}\leqslant$

$\mathrm{P}(2c\sqrt{n}\max_{1\leqslant k\leqslant K}|S_{jk}/\Psi_{jk}| \leqslant Z_\alpha)$, given $2(\log K)^{-1/2}+\tilde{\rho}_n = o(\alpha)$ (note that Theorem 5.2 ensures that $\tilde{\rho}_n \to 0$ with a polynomial rate as $n\to\infty$).

□

Proof of Theorem 5.3. Let $S_{jk,i} = \frac{1}{\sqrt{n}}\sum_{l=(i-1)b_n+1}^{ib_n} X_{jk,l}\varepsilon_{j,l}$, we first need to prove that

$$\rho_{n,j} \stackrel{\text{def}}{=} \sup_{r\in\mathbb{R}} \big|\,\mathrm{P}\big\{\max_{1\leqslant k\leqslant K}(Z_{jk}^{[B]}/\Psi_{jk}) \leqslant r | X_{j,\cdot},\varepsilon_{j,\cdot}\big\} - \mathrm{P}\big\{\max_{1\leqslant k\leqslant K}(\widetilde{Z}_{jk}/\Psi_{jk})\leqslant r\big\}\big|$$
$$= \sup_{r\in\mathbb{R}} \big|\,\mathrm{P}\big\{\max_{1\leqslant k\leqslant K}\big(\sum_{i=1}^{l_n} e_{j,i}S_{jk,i}/\Psi_{jk}\big) \leqslant r | X_{j,\cdot},\varepsilon_{j,\cdot}\big\} - \mathrm{P}\big\{\max_{1\leqslant k\leqslant K}(\widetilde{Z}_{jk}/\Psi_{jk})\leqslant r\big\}\big| \to 0, \text{ as } n\to\infty.$$

Given the sample variance covariance matrix $(K\times K)$ $\Sigma_{j,n} = \sum_{\ell=-n}^{n}(1-|\ell|/n)\Gamma_j(\ell)$, where $\Gamma_j(\ell) = \mathsf{E}(X_{j,t}\varepsilon_{j,t}X_{j,t-\ell}^\top\varepsilon_{j,t-\ell})$, let $\widetilde{Z}_j = (\widetilde{Z}_{jk})_{k=1}^K \sim \mathrm{N}(0,\Sigma_{j,n})$. In addition, define $\Sigma_{j,b_n} = \sum_{\ell=-b_n}^{b_n}(1-|\ell|/b_n)\Gamma_j(\ell)$ and $\widehat{\Sigma}_j = \sum_{i=1}^{l_n} S_{j,i}S_{j,i}^\top$, where $S_{j,i} = (S_{jk,i})_{k=1}^K$. Let $\Psi_j = \mathrm{diag}(\Psi_{jk})$, $\delta_j = \delta_{j1}+\delta_{j2}$, with $\delta_{j1} = |\Psi_j^{-1}\widehat{\Sigma}_j\Psi_j^{-1} - \Psi_j^{-1}\Sigma_{j,b_n}\Psi_j^{-1}|_{\max}$ and $\delta_{j2} = |\Psi_j^{-1}\Sigma_{j,b_n}\Psi_j^{-1} - \Psi_j^{-1}\Sigma_{j,n}\Psi_j^{-1}|_{\max}$, where $|\cdot|_{\max}$ is the maximum norm of a matrix. According to Theorem 2 of Chernozhukov et al. (2015), $\rho_{n,j}$ is bounded by $\pi(\delta_{j1})\vee\pi(\delta_{j2})$, with $\pi(\delta_j) \stackrel{\text{def}}{=} C\delta_j^{1/3}\{1\vee a_K^2\vee\log(1/\delta_j)\}^{1/3}(\log K)^{1/3}$, where $a_K = \mathsf{E}(\max_{1\leqslant k\leqslant K} Z_{jk}/\Psi_{jk}) \leqslant \sqrt{2\log K}$.

For the first part,

$$\delta_{j1} = \max_{1\leqslant k_1,k_2\leqslant K}\left|\frac{\sum_{i=1}^{l_n}S_{jk_1,i}S_{jk_2,i}}{\Psi_{jk_1}\Psi_{jk_2}} - \frac{l_n\,\mathsf{E}(S_{jk_1,i}S_{jk_2,i})}{\Psi_{jk_1}\Psi_{jk_2}}\right|$$
$$\leqslant \frac{\max_{1\leqslant k_1,k_2\leqslant K}\big|\sum_{i=1}^{l_n}S_{jk_1,i}S_{jk_2,i} - l_n\,\mathsf{E}(S_{jk_1,i}S_{jk_2,i})\big|}{\min_{1\leqslant k_1,k_2\leqslant K}\Psi_{jk_1}\Psi_{jk_2}}.$$

We need to analyze the tail probability of δ_{j1}. Applying Theorem 5.1 of Zhang and Wu (2017a), for $x\geqslant n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2$, we have

$$\mathrm{P}\left(n\delta_{j1}\geqslant\frac{x}{\min_{1\leqslant k_1,k_2\leqslant K}\Psi_{j_1k_1}\Psi_{j_2k_2}}\right) \lesssim \frac{KF_\varsigma\Gamma_{j,2q,\varsigma}^q}{x^{q/2}} + K^2\exp\left(-\frac{C_jx^2}{nb_n\Phi_{j,8,\varsigma}^4}\right),$$

for all large n, where $F_\varsigma = n$, for $\varsigma > 1-2/q$; $F_\varsigma = l_nb_n^{q/2-\varsigma q/2}$, for $1/2-2/q<\varsigma<1-2/q$; $F_\varsigma = l_n^{q/4-\varsigma q/2}b_n^{q/2-\varsigma q/2}$, for $\varsigma<1/2-2/q$. The constants C_j depend on η, q, and ς. This ensures that when $x = \max\{n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2, n^{1/2}b_n^{1/2}(\log K)^{1/2}\Phi_{j,8,\varsigma}^2c_n, K^{2/q}F_\varsigma^{2/q}\Gamma_{j,2q,\varsigma}^2c_n\}$, $c_n^{-1}=o(1)$, the tail probability tends to 0, as $n\to\infty$.

It follows that $\pi(\delta_{j1})\to 0$ as $n\to\infty$, given $x = o\{n(\log K)^{-2}\}$, which implies the following conditions on b_n:

$$b_n = o\{n(\log K)^{-4}\Phi_{j,2q,\varsigma}^{-4}\wedge n(\log K)^{-5}\Phi_{j,8,\varsigma}^{-4}c_n^{-2}\},\quad F_\varsigma = o\{n^{q/2}(\log K)^{-q}K^{-1}\Gamma_{j,2q,\varsigma}^{-q}c_n^{-q/2}\}.$$

For the second part, by defining $\psi_j \stackrel{\text{def}}{=} \min_{1\leqslant k_1,k_2\leqslant K}\Psi_{jk_1}\Psi_{jk_2}$, we have

$$\delta_{j2} \leqslant \left|\psi_j^{-1}\Big\{\sum_{b_n<|\ell|\leqslant n}(1-|\ell|/n)\Gamma_j(\ell) + \sum_{\ell=-b_n}^{b_n}|\ell|(-1/n+1/b_n)\Gamma_j(\ell)\Big\}\right|_{\max}.$$

Recall that

$$|\Gamma_{j,k_1,k_2}(\ell)| = \Big| \sum_{h=0}^{\infty} \mathsf{E}\{(\mathcal{P}_h(X_{jk,0}\varepsilon_{j0})\mathcal{P}_h(X_{jk_2,\ell}\varepsilon_{j,\ell})\} \Big| \\ \leqslant \sum_{h=0}^{\infty} \|X_{jk_1,h}\varepsilon_{j,h} - X^*_{jk_1,h}\varepsilon^*_{j,h}\|_2 \|X_{jk_2,h+\ell}\varepsilon_{j,h+\ell} - X^*_{jk_2,h+\ell}\varepsilon^*_{j,h+\ell}\|_2,$$

where the operator is given by $\mathcal{P}_h(\cdot) \stackrel{\text{def}}{=} \mathsf{E}(\cdot|\mathcal{F}_h) - \mathsf{E}(\cdot|\mathcal{F}_{h-1})$. It follows that

$$\Big| \sum_{b_n<|\ell|\leqslant n} (1-|\ell|/n)\Gamma_{j,k_1,k_2}(\ell) + \sum_{\ell=-b_n}^{b_n} |\ell|(-1/n+1/b_n)\Gamma_{j,k_1,k_2}(\ell) \Big| \\ \leqslant \Delta_{0,2,j,k_1}\Delta_{b_n+1,2,j,k_2} + \frac{2}{n}\Delta_{0,2,j,k_1}\sum_{\ell=b_n+1}^{n}\Delta_{\ell,2,j,k_2} + 2\frac{n-b_n}{nb_n}\Delta_{0,2,j,k_1}\sum_{\ell=1}^{b_n}\Delta_{\ell,2,j,k_2}, \tag{B.1}$$

where $\Delta_{m,2,j,k} = \sum_{t=m}^{\infty}\|X_{jk,t}\varepsilon_{j,t} - X^*_{jk,t}\varepsilon^*_{j,t}\|_2$. Given the fact that $\Delta_{0,2,j,k} \leqslant \Phi_{j,4,0}$, $\Delta_{\ell,2,j,k} \leqslant \Phi_{j,4,\varsigma}\ell^{-\varsigma}$, (B.1) is bounded by $\Phi_{j,4,0}\Phi_{j,4,\varsigma}\{(b_n+1)^{-\varsigma} + 2n^{-1}\sum_{\ell=b_n+1}^{n}\ell^{-\varsigma} + 2\frac{n-b_n}{nb_n}\sum_{\ell=1}^{b_n}\ell^{-\varsigma}\} = \Phi_{j,4,0}\Phi_{j,4,\varsigma}v(b_n)$ for any k_1, k_2, where $v(b_n)$ is a function with respect to b_n. Note that $v(b_n) \lesssim (b_n+1)^{-\varsigma} + 2v_{n,1}/n + 2(n-b_n)v_{n,2}/(nb_n)$, where $v_{n,1} = \log\{n/(b_n+1)\}$ (resp. $n^{-\varsigma+1}$ or $(b_n+1)^{-\varsigma+1}$) for $\varsigma = 1$ (resp. $\varsigma < 1$ or $\varsigma > 1$), $v_{n,2} = \log b_n$ (resp. $b_n^{-\varsigma+1}$ or 1) for $\varsigma = 1$ (resp. $\varsigma < 1$ or $\varsigma > 1$). Therefore, the bound of δ_{j2} would decrease as b_n increases. In particular, we need to impose an addition assumption such that $\Phi_{j,4,0}\Phi_{j,4,\varsigma}v(b_n) = o\{(\log K)^{-2}\}$ to guarantee $\pi(\delta_{j2}) \to 0$.

The results for the two parts above ensure that $\rho_{n,j} \to 0$ as $n \to \infty$, given $x = o\{n(\log K)^{-2}\}$ and $\Phi_{j,4,0}\Phi_{j,4,\varsigma}v(b_n) = o\{(\log K)^{-2}\}$, which imply the following conditions on b_n:

$$\begin{aligned}
&b_n = o\{n(\log K)^{-4}\Phi_{j,2q,\varsigma}^{-4} \wedge n(\log K)^{-5}\Phi_{j,8,\varsigma}^{-4}c_n^{-2}\},\\
&F_\varsigma = o\{n^{q/2}(\log K)^{-q}K^{-1}\Gamma_{j,2q,\varsigma}^{-q}c_n^{-q/2}\}, \text{ with } c_n^{-1} = o(1).\\
&\Phi_{j,4,0}\Phi_{j,4,\varsigma}\{b_n^{-1} + \log(n/b_n)/n + (n-b_n)\log b_n/(nb_n)\}(\log K)^2 = o(1), \text{ if } \varsigma = 1;\\
&\Phi_{j,4,0}\Phi_{j,4,\varsigma}\{b_n^{-1} + n^{-\varsigma} + (n-b_n)b_n^{-\varsigma+1}/(nb_n)\}(\log K)^2 = o(1), \text{ if } \varsigma < 1;\\
&\Phi_{j,4,0}\Phi_{j,4,\varsigma}\{b_n^{-1} + n^{-1}b_n^{-\varsigma+1} + (n-b_n)/(nb_n)\}(\log K)^2 = o(1), \text{ if } \varsigma > 1.
\end{aligned} \tag{B.2}$$

At last, combining the Gaussian approximation results for S_{jk}/Ψ_{jk} and applying Theorem 3.1 in Chernozhukov et al. (2013a), we have

$$\sup_{\alpha\in(0,1)} \big|\,\mathrm{P}\big(\max_{1\leqslant k\leqslant K}|S_{jk}/\Psi_{jk}| \geqslant q_{j,(1-\alpha)}^{[B]}\big) - \alpha\big| \lesssim \rho_{n,j} + \pi'(z) + \mathrm{P}(\delta_j \geqslant z),$$

where $\pi'(z) = z^{1/3}\{1 \vee \log(K/z)\}^{2/3}$. We need to pick z such that $\pi'(z) + \mathrm{P}(\delta_j \geqslant z) \to 0$ as $n \to \infty$ and it can be obtained by taking $z = r_n^{1/2}/(\log K)$, with
$r_n = n^{-1}\max\{n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2, n^{1/2}b_n^{1/2}(\log K)^{1/2}\Phi_{j,8,\varsigma}^2 c_n, K^{2/q}F_\varsigma^{2/q}\Gamma_{j,2q,\varsigma}^2 c_n, n\Phi_{j,2,0}\Phi_{j,2,\varsigma}v(b_n)\}$, $c_n^{-1} = o(1)$.

□

Comment B.1 (Admissible rate of b_n)**.** Consider the special case with $\Phi_{j,2q,\varsigma} = \mathcal{O}(1)$ and

$\Gamma_{j,2q,\varsigma} = \mathcal{O}(1)$, for $q > 4$. Let $K = \mathcal{O}(n^r)$, $\log K = \mathcal{O}(r \log n) = \mathcal{O}(n^{\bar{r}})$, $c_n = n^s$ with $s > 0$ and assume $1/2 - 2/q < \varsigma < 1 - 2/q$. Then (B.2) implies an admissible rate of $b_n = \mathcal{O}(n^\eta)$ such that $2\bar{r}/\varsigma < \eta < \max\{1 - 5\bar{r} - 2s, (q/2 - q\bar{r} - qs/2 - r - 1)/(q/2 - \varsigma q/2 - 1)\}$.

Comment B.2 (Validity of multiplier block bootstrap under stronger tail assumptions)**.** Note that in case with stronger exponential moment conditions on the underlying processes, we shall change the tail probabilities to bound δ_{j1}.

Let $\Phi_{j,\psi_\nu,\varsigma} = \max_{1 \leqslant k \leqslant p} \sup_{q \geqslant 2} q^{-\nu} \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} < \infty$, then according to Theorem 5.2 of Zhang and Wu (2017a), for all $x > 0$, we have

$$\mathrm{P}(n\delta_{j1} \geqslant \frac{x}{\min_{1 \leqslant k_1,k_2 \leqslant K} \Psi_{j_1k_1}\Psi_{j_2k_2}}) \lesssim K^2 \exp\Big(- \frac{x^\gamma}{4e\gamma(\sqrt{n}b_n\Phi^2_{j,\psi_\nu,0})^\gamma}\Big),$$

where $\gamma = 1/(2\nu + 1)$. This implies that when $x = (\log K\sqrt{n}b_n\Phi^2_{j,\psi_\nu,0})^{1/\gamma}c_n$, with $c_n^{-1} = o(1)$, the tail probability tends to 0, as $n \to \infty$. It follows that $\pi(\delta_{j1}) \to 0$ as $n \to \infty$, given $x = (\log K\sqrt{n}b_n\Phi^2_{j,\psi_\nu,0})^{1/\gamma} = o\{n(\log K)^{-2}\}$. As a result, (B.2) will be replaced by

$$\begin{aligned}
&b_n = o\{n^{\gamma-1/2}(\log K)^{-2\gamma-1}\Phi^{-2}_{j,\psi_\nu,0}c_n^{-\gamma}\}, \text{ with } c_n^{-1} = o(1).\\
&\Phi_{j,4,0}\Phi_{j,4,\varsigma}\{b_n^{-1} + \log(n/b_n)/n + (n - b_n)\log b_n/(nb_n)\}(\log K)^2 = o(1), \text{ if } \varsigma = 1;\\
&\Phi_{j,4,0}\Phi_{j,4,\varsigma}\{b_n^{-1} + n^{-\varsigma} + (n - b_n)b_n^{-\varsigma+1}/(nb_n)\}(\log K)^2 = o(1), \text{ if } \varsigma < 1;\\
&\Phi_{j,4,0}\Phi_{j,4,\varsigma}\{b_n^{-1} + n^{-1}b_n^{-\varsigma+1} + (n - b_n)/(nb_n)\}(\log K)^2 = o(1), \text{ if } \varsigma > 1.
\end{aligned}$$

B.2 Proofs of Joint Equation Estimation

Proof of Theorem A.3. Analogue to the proof of Theorem 5.3, the conclusions are implied by

$$\mathrm{P}\left(n\delta_1 \geqslant \big(\min_{1 \leqslant k_1,k_2 \leqslant K, 1 \leqslant j_1,j_2 \leqslant J} \Psi_{j_1k_1}\Psi_{j_2k_2}\big)^{-1}x\right) \lesssim \frac{JKF_\varsigma\Gamma^q_{2q,\varsigma}}{x^{q/2}} + (JK)^2\exp\Big(- \frac{Cx^2}{nb_n\Phi^4_{8,\varsigma}}\Big),$$

for $x \geqslant n^{1/2}b_n^{1/2}\Phi^2_{2q,\varsigma}$ and all large n, where

$$\delta_1 \stackrel{\text{def}}{=} \max_{1 \leqslant k_1,k_2 \leqslant K, 1 \leqslant j_1,j_2 \leqslant J} \left| \frac{\sum_{i=1}^{l_n} S_{j_1k_1,i}S_{j_2k_2,i}}{\Psi_{j_1k_1}\Psi_{j_2k_2}} - \frac{l_n\,\mathsf{E}(S_{j_1k_1,i}S_{j_2k_2,i})}{\Psi_{j_1k_1}\Psi_{j_2k_2}}\right|.$$

In particular, when $x = \max\{n^{1/2}b_n^{1/2}\Phi^2_{2q,\varsigma}, n^{1/2}b_n^{1/2}\{\log(JK)\}^{1/2}\Phi^2_{8,\varsigma}c_n, (JK)^{2/q}F_\varsigma^{2/q}\Gamma^2_{2q,\varsigma}c_n\}$, $c_n^{-1} = o(1)$, the tail probability tends to 0, as $n \to \infty$.

By similar proof to that of Theorem 5.3, it follows that $\tilde{\rho}_n \to 0$ as $n \to \infty$, given $x = o[n\{\log(KJ)\}^{-2}]$ and $\Phi_{4,0}\Phi_{4,\varsigma}v(b_n) = o\{(\log KJ)^{-2}\}$, which imply the following conditions on

b_n:

$$
\begin{aligned}
&b_n = o[n\{\log(KJ)\}^{-4}\Phi_{2q,\varsigma}^{-4} \wedge n\{\log(KJ)\}^{-5}\Phi_{8,\varsigma}^{-4}c_n^{-2}],\\
&F_\varsigma = o[n^{q/2}\{\log(KJ)\}^{-q}(KJ)^{-1}\Gamma_{2q,\varsigma}^{-q}c_n^{-q/2}], \text{ with } c_n^{-1} = o(1).\\
&\Phi_{4,0}\Phi_{4,\varsigma}\{b_n^{-1} + \log(n/b_n)/n + (n-b_n)\log b_n/(nb_n)\}\{\log(KJ)\}^2 = o(1), \text{ if } \varsigma = 1;\\
&\Phi_{4,0}\Phi_{4,\varsigma}\{b_n^{-1} + n^{-\varsigma} + (n-b_n)b_n^{-\varsigma+1}/(nb_n)\}\{\log(KJ)\}^2 = o(1), \text{ if } \varsigma < 1;\\
&\Phi_{4,0}\Phi_{4,\varsigma}\{b_n^{-1} + n^{-1}b_n^{-\varsigma+1} + (n-b_n)/(nb_n)\}\{\log(KJ)\}^2 = o(1), \text{ if } \varsigma > 1. \qquad \text{(B.3)}
\end{aligned}
$$

Recall that $F_\varsigma = n$, for $\varsigma > 1 - 2/q$; $F_\varsigma = l_n b_n^{q/2-\varsigma q/2}$, for $1/2 - 2/q < \varsigma < 1 - 2/q$; $F_\varsigma = l_n^{q/4-\varsigma q/2} b_n^{q/2-\varsigma q/2}$, for $\varsigma < 1/2 - 2/q$.

The rest of the proof is similar to that of Theorem 5.3 and thus is omitted.

□

Proof of Theorem A.4. For any $\delta, \tilde{\delta} \in \mathbb{R}^{\mathrm{K}}$ in $\mathcal{G}_{\widetilde{T}_j}$, we have

$$
\begin{aligned}
\left|\mathsf{E}_n\left\{\varepsilon_{j,t}\left(\frac{X_{j,t}^\top\delta}{|\delta|_{j,pr}} - \frac{X_{j,t}^\top\tilde{\delta}}{|\tilde{\delta}|_{j,pr}}\right)\right\}\right| &= \left|\mathsf{E}_n\left[\varepsilon_{j,t}\left\{\frac{X_{j,t}^\top(\delta-\tilde{\delta})}{|\delta|_{j,pr}} + \frac{X_{j,t}^\top\tilde{\delta}}{|\delta|_{j,pr}} - \frac{X_{j,t}^\top\tilde{\delta}}{|\tilde{\delta}|_{j,pr}}\right\}\right]\right|\\
&\leqslant \left|\mathsf{E}_n\left[\varepsilon_{j,t}\left\{\frac{X_{j,t}^\top(\delta-\tilde{\delta})}{|\delta|_{j,pr}}\right\}\right]\right| + \left|\mathsf{E}_n\left[\varepsilon_{j,t}\left\{\frac{X_{j,t}^\top\tilde{\delta}}{|\delta|_{j,pr}} - \frac{X_{j,t}^\top\tilde{\delta}}{|\tilde{\delta}|_{j,pr}}\right\}\right]\right|\\
&\leqslant (\mathsf{E}_n\varepsilon_{j,t}^2)^{1/2}\left\{\mathsf{E}_n\left|\frac{X_{j,t}^\top(\delta-\tilde{\delta})}{|\delta|_{j,pr}}\right|^2\right\}^{1/2} + (\mathsf{E}_n\varepsilon_{j,t}^2)^{1/2}\left(\frac{|\tilde{\delta}|_{j,pr} - |\delta|_{j,pr}}{|\delta|_{j,pr}}\right)\\
&\leqslant 2\sigma\mu_j(p)|\delta-\tilde{\delta}|_2.
\end{aligned}
$$

Then by following the proof of Lemma 5 (Step 2) in Belloni and Chernozhukov (2013), we have $\sup_{\mathcal{Q}}\mathcal{N}(\epsilon, \mathcal{G}_{\widetilde{T}_j}, \|\cdot\|_{\mathcal{Q},1}) \lesssim (6\mu_j(p)\sigma/\epsilon)^{s+p}$. And it follows that $|F_{j,p}| \lesssim \binom{K}{p}(6\mu_j(p)\sigma/\epsilon)^{s+p}$.

Moreover, it is not hard to see that $\sup_{f\in\mathcal{F}_{j,p}}|G_n(f)| \leqslant 2\sqrt{n}\epsilon + \sup_{f\in F_{j,p}}|G_n(f)|$. Let $\psi = \max_{f\in F_{j,p}}\psi_f$ (assume ψ is bounded by constant) and applying the Gaussian approximation results on the vector $G_n(f)/\psi_f$ (given (A6)), we have

$$
\begin{aligned}
\mathrm{P}\{\sup_{f\in F_{j,p}}|G_n(f)| \geqslant \kappa_n/2\} &\leqslant \mathrm{P}\{\sup_{f\in F_{j,p}}|G_n(f)/\psi_f| \geqslant \kappa_n/(2\psi)\}\\
&\leqslant 2|F_{j,p}|\{1-\Phi(\kappa_n/(2\psi))\} + d_n\\
&\leqslant 2K^p(6\mu_j(p)\sigma/\epsilon)^{s+p}\exp\{-\kappa_n^2/(8\psi^2)\}\{\kappa_n/(2\psi)\}^{-1} + d_n,
\end{aligned}
$$

as $\binom{K}{p} \leqslant K^p$. Therefore, for $\kappa_n = \psi\sqrt{p\log K + (p+s)\{\log(6\mu_j(p)\sigma) + \log n/2\}}$ and $\epsilon = \sqrt{p\log K + (p+s)\log(6\mu_j(p)\sigma)}(4\sqrt{n})^{-1}$, it follows that $\sup_{f\in\mathcal{F}_{j,p}}|G_n(f)| \leqslant \kappa_n$ (note that $d_n \to 0$ with a polynomial rate as $n\to\infty$).

The rest of the proof is a direct application of Theorem 5 of Belloni and Chernozhukov (2013) by inserting the bound for $\lambda^0(1-\alpha)$ (A.2) provided in Corollary A.1, and thus is omitted. □

B.3 Plausibility of RE and RSE Conditions

Define the s-sparse sphere as $F_\delta = \{\delta : |\delta|_0 \leqslant s, |\delta|_2 = 1\}$. According to Rudelson and Zhou (2012), the ϵ-covering number of F_δ w.r.t. the Euclidean metric is $l = \exp(s\log(3eK/m\epsilon))$, with

$m \geqslant 1$. This is the cardinality of the ϵ-cover set Π_δ of F_δ. Moreover, for any point $\delta \in F_\delta$, let π_δ denote the closest point to δ within Π_δ. Let $\breve{X}_{j,t}^{\pi(\delta)} \stackrel{\text{def}}{=} \{\widetilde{X}_{j,t}^\top \pi(\delta)\}^2 - n^{-1}\pi(\delta)^\top \mathsf{E}\{X_{j,t}X_{j,t}^\top\}\pi(\delta)$, where $\widetilde{X}_j \stackrel{\text{def}}{=} n^{-1/2}X_j$ and $X_j(n \times K)$ is a matrix of $X_{j,t}$. Note that $\breve{X}_{j,t}^{\pi(\delta)}$ is a vector of the cardinality of Π_δ.

Theorem B.1 (Plausibility of RE and RSE)**.** *For any $j = 1, \ldots, J$, suppose the vectors $X_{j,t}$ of length K satisfy*

$$0 < \kappa \leqslant \min_{|\delta|_0 \leqslant s, |\delta|_1 = 1} \delta^\top \mathsf{E}(X_{j,t}X_{j,t}^\top)\delta \leqslant \max_{|\delta|_0 \leqslant s, |\delta|_1 = 1} \delta^\top \mathsf{E}(X_{j,t}X_{j,t}^\top)\delta \leqslant \psi < \infty,$$

where ψ and κ are positive constants. Given $\breve{\Phi}_{2,\varsigma} \stackrel{\text{def}}{=} \max_{\pi(\delta)\in\Pi_\delta} \|\breve{X}_{j,\cdot}^{\pi(\delta)}\|_{2,\varsigma} < \infty$, and for $q > 2$, $\|\max_{\pi(\delta)\in\Pi_\delta} |\breve{X}_{j,\cdot}^{\pi(\delta)}|\|_{q,\varsigma} < \infty$,

$$n^{-1/2}(\log l)^{1/2}\breve{\Phi}_{2,\varsigma} + n^{-1}r_\varsigma(\log l)^{3/2}\|\max_{\pi(\delta)\in\Pi_\delta} |\breve{X}_{j,\cdot}^{\pi(\delta)}|\|_{q,\varsigma} = o(1),$$

where $r_\varsigma = n^{1/q}$ for $\varsigma > 1/2 - 1/q$ and $r_\varsigma = n^{1/2-\varsigma}$ for $\varsigma < 1/2 - 1/q$, then the RE and RSE conditions hold with probability $1 - o(1)$, with $p + s_j \leqslant s$.

Proof of Theorem B.1.

Firstly, we need to check the implication of the population matrix. We know that $\delta^\top X_j^\top X_j\delta/n = |\widetilde{X}_j\delta|_2^2$.Then we have the following inequalities for any point $\delta \in F_\delta$,

$$-|\widetilde{X}_j\{\delta - \pi(\delta)\}|_2 + |\widetilde{X}_j\pi(\delta)|_2 \leqslant |\widetilde{X}_j\delta|_2 \leqslant |\widetilde{X}_j\{\delta - \pi(\delta)\}|_2 + |\widetilde{X}_j\pi(\delta)|_2. \tag{B.4}$$

We first check the right hand side of (B.4). Define $\|\widetilde{X}_j\|_{2,F_\delta} \stackrel{\text{def}}{=} \sup_{\delta\in F_\delta} |\widetilde{X}_j\delta|_2$. As indicated in the proof of Theorem 16 in Rudelson and Zhou (2012), we have $|\widetilde{X}_j\{\delta - \pi(\delta)\}|_2 \leqslant \epsilon\|\widetilde{X}_j\|_{2,F_\delta}$. To bound $\max_{\pi(\delta)\in\Pi_\delta} |\widetilde{X}_j\pi(\delta)|_2$, we invoke the tail probability inequality in Lemma B.2, which gives

$$\mathrm{P}\big(\max_{\pi(\delta)\in\Pi_\delta} |\sum_{t=1}^n \breve{X}_{j,t}^{\pi(\delta)}| \geqslant x\big) = \mathrm{P}\big[\max_{\pi(\delta)\in\Pi_\delta} ||\widetilde{X}_j\pi(\delta)|_2^2 - \pi(\delta)^\top \mathsf{E}\{X_{j,t}X_{j,t}^\top\}\pi(\delta)| \geqslant x\big] \to 0, \text{as } n \to \infty,$$

if $x \gtrsim \sqrt{n\log l}\breve{\Phi}_{2,\varsigma} + r_\varsigma(\log l)^{3/2}\|\max_{\pi(\delta)\in\Pi_\delta} |\breve{X}_{j,\cdot}^{\pi(\delta)}|\|_{q,\varsigma}$.

Therefore, given $\kappa, \psi > 0$, $\kappa - x_n \leqslant |\widetilde{X}_j\pi(\delta)|_2^2 \leqslant x_n + \psi$ holds with probability $1 - o(1)$ for all $\pi(\delta) \in \Pi_\delta$, where $x_n \stackrel{\text{def}}{=} \sqrt{n\log l}\breve{\Phi}_{2,\varsigma} + r_\varsigma(\log l)^{3/2}\|\max_{\pi(\delta)\in\Pi_\delta} |\breve{X}_{j,\cdot}^{\pi(\delta)}|\|_{q,\varsigma}$. In particular, the assumption

$$n^{-1/2}(\log l)^{1/2}\breve{\Phi}_{2,\varsigma} + n^{-1}r_\varsigma(\log l)^{3/2}\|\max_{\pi(\delta)\in\Pi_\delta} |\breve{X}_{j,\cdot}^{\pi(\delta)}|\|_{q,\varsigma} = o(1)$$

ensures that $x_n = o(1)$.

Hence, the right inequality in (B.4) leads to $|\widetilde{X}_j\delta|_2 \leqslant \epsilon\|\widetilde{X}_j\|_{2,F_\delta} + \sqrt{x_n} + \sqrt{\psi}$. Taking the supremum over all $\delta \in F_\delta$ on both sides shows that $\sup_{\delta\in F_\delta} |\widetilde{X}_j\delta|_2 \leqslant (\sqrt{x_n} + \sqrt{\psi})/(1 - \epsilon)$ with probability $1 - o(1)$. Moreover, by the left hand side of (B.4), we have $|\widetilde{X}_j\delta|_2 \geqslant \sqrt{\kappa - x_n} - \epsilon(\sqrt{x_n} + \sqrt{\psi})/(1 - \epsilon)$, with probability $1 - o(1)$.

Collecting the results together, we have shown that for all $\delta \in F_\delta$,

$$\sqrt{\kappa - x_n} - \frac{\epsilon(\sqrt{x_n} + \sqrt{\psi})}{(1-\epsilon)} \leqslant |\widetilde{X}_j \delta|_2 \leqslant \frac{\sqrt{x_n} + \sqrt{\psi}}{(1-\epsilon)}, \tag{B.5}$$

with probability $1 - o(1)$.

Let $c^*(s) = \max_{\delta \in F_\delta} |\widetilde{X}_j \delta|_2$, $c_*(s) = \min_{\delta \in F_\delta} |\widetilde{X}_j \delta|_2$, with properly chosen ϵ, $c^*(s), c_*(s)$ are bounded from above and below, and the desired results follow by the fact $\kappa_j^2(p) \geqslant c_*(s_j + p)$, $\phi_j(p) \leqslant c^*(s_j + p)$, with $s_j + p \leqslant s$. □

B.4 Proofs of Simultaneous Inference

B.4.1 Some Useful Lemmas

Lemma B.1 (Burkholder (1988)). *Let $q > 1$, $q' = \min(q, 2)$. Let $M_n = \sum_{t=1}^n \xi_t$; where $\xi_t \in \mathcal{L}^q$ (i.e., $\|\xi_t\|_q < \infty$) are martingale differences. Then*

$$\|M_n\|_q^{q'} \leqslant K_q^{q'} \sum_{t=1}^n \|\xi_t\|_q^{q'} \quad \textit{where} \quad K_q = \max((q-1)^{-1}, \sqrt{q-1}).$$

Lemma B.2 (Theorem 6.2 of Zhang and Wu (2017a) Tail probabilities for high dimensional partial sums). *For a zero-mean p-dimensional random variable $X_t \in \mathbb{R}^p$, let $S_n = \sum_{t=1}^n X_t$ and assume that $\||X_\cdot|_\infty\|_{q,\varsigma} < \infty$, where $q > 2$ and $\varsigma \geqslant 0$, and $\Phi_{2,\varsigma} = \max_{1 \leqslant j \leqslant p} \|X_{j,\cdot}\|_{2,\varsigma} < \infty$.*
i) If $\varsigma > 1/2 - 1/q$, then for $x \gtrsim \sqrt{n \log p}\Phi_{2,\varsigma} + n^{1/q}(\log p)^{3/2}\||X_\cdot|_\infty\|_{q,\varsigma}$,

$$\mathrm{P}(|S_n|_\infty \geqslant x) \leqslant \frac{C_{q,\varsigma} n (\log p)^{q/2} \||X_\cdot|_\infty\|_{q,\varsigma}^q}{x^q} + C_{q,\varsigma} \exp\left(\frac{-C_{q,\varsigma} x^2}{n\Phi_{2,\varsigma}^2}\right).$$

ii) If $0 < \varsigma < 1/2 - 1/q$, then for $x \gtrsim \sqrt{n \log p}\Phi_{2,\varsigma} + n^{1/2-\varsigma}(\log p)^{3/2}\||X_\cdot|_\infty\|_{q,\varsigma}$,

$$\mathrm{P}(|S_n|_\infty \geqslant x) \leqslant \frac{C_{q,\varsigma} n^{q/2-\varsigma q} (\log p)^{q/2} \||X_\cdot|_\infty\|_{q,\varsigma}^q}{x^q} + C_{q,\varsigma} \exp\left(\frac{-C_{q,\varsigma} x^2}{n\Phi_{2,\varsigma}^2}\right).$$

Lemma B.3 (Tail probabilities for high dimensional partial sums with strong tail assumptions). *For a zero-mean p- dimensional random variable $X_t \in \mathbb{R}^p$, let $S_n = \sum_{t=1}^n X_t$ and assume that $\Phi_{\psi_\nu,\varsigma} = \max_{1 \leqslant j \leqslant p} \sup_{q \geqslant 2} q^{-\nu} \|X_{j,\cdot}\|_{q,\varsigma} < \infty$ for some $\nu \geqslant 0$, and let $\gamma = 2/(1+2\nu)$. Then for all $x > 0$, we have*

$$\mathrm{P}(|S_n|_\infty \geqslant x) \leqslant p \exp\{-C_\gamma x^\gamma / (\sqrt{n}\Phi_{\psi_\nu,0})^\gamma\}.$$

Lemma B.4 (Theorem 1 of El Machkouri et al. (2013)). *Denote $Y_t = f(\mathcal{F}_t)$, where f is some measurable function. Let $S_n = \sum_{t=1}^n Y_t$, and $\delta_{\varsigma,t} = \|Y_t - Y_t^*\|_\varsigma$. If $\mathsf{E}(Y_i) = 0$, $\sum_{t=0}^\infty \delta_{\varsigma,t} < \infty$, some $\varsigma \geqslant 2$, and $\sigma_n^2 \overset{\text{def}}{=} \mathsf{E}(S_n^2) \to \infty$, then*

$$\sigma_n^{-1} S_n \xrightarrow{\mathcal{L}} \mathrm{N}(0,1).$$

Lemma B.5. *Under the same conditions as in Theorem 5.4, let $\tilde{\beta}_{jk}$ be any estimator such that*

$|\tilde{\beta}_{jk} - \beta^0_{jk}| \leqslant C\rho_n$ *with probability* $1 - o(1)$. *Then we have*

$$n^{-1} \max_{(j,k)\in G} \Delta_n \lesssim o(n^{-1/2} g_n^{-1}), \tag{B.6}$$

holds with probability $1 - o(1)$, *where* $\Delta_n \stackrel{\text{def}}{=} n^{1/2} G_n\{\psi_{jk}(Z_{j,t}, \tilde{\beta}_{jk}, \widehat{h}_{jk}) - \psi_{jk}(Z_{j,t}, \beta^0_{jk}, h^0_{jk})\}$.

Proof of Lemma B.5.

As indicated in the proof of Theorem 2 in Belloni et al. (2015b), the entropy $\text{ent}(\epsilon, \widetilde{\mathcal{F}}) \leqslant cs \log(a_n/\epsilon)$ for the function class $\widetilde{\mathcal{F}} = \{z \mapsto \psi_{jk}\{z, \beta, \tilde{h}(x_{j(-k)})\} - \psi_{jk}\{z, \beta^0_{jk}, h^0_{jk}(x_{j(-k)})\} : (j,k) \in G, \beta \in \mathcal{B}_{jk}, |\beta - \beta^0_{jk}| \leqslant C\rho_n, \tilde{h} \in \mathcal{H}_{jk}\}$, which has $2F$ as the envelope (the definition of F is given in (C6)). Therefore, for any $f \in \widetilde{\mathcal{F}}$, there exists a set F_n such that $\min_{f' \in F_n} \|f - f'\|_{\mathcal{Q},2} \leqslant \tilde{\epsilon}$, where $\tilde{\epsilon} \stackrel{\text{def}}{=} \epsilon \|2F\|_{\mathcal{Q},2}$ and the cardinality of the set $|F_n| = (a_n/\epsilon)^{cs}$. Then we have

$$\sup_{f \in \widetilde{\mathcal{F}}} \left| \sum_{t=1}^{n} [f - \pi(f) - \mathsf{E}\{f - \pi(f)\}] \right| \leqslant 2\tilde{\epsilon} n,$$

where $\pi(f) \stackrel{\text{def}}{=} \arg\min_{f' \in \widetilde{\mathcal{F}}} \|f - f'\|_{\mathcal{Q},2}$. Hence, with probability $1 - o(1)$,

$$\begin{aligned}
\max_{(j,k)\in G} \Delta_n &\leqslant n^{1/2} \sup_{f \in \widetilde{\mathcal{F}}} |G_n(f)| \\
&= n \sup_{f \in \widetilde{\mathcal{F}}} \left| [\mathsf{E}_n(f) - \mathsf{E}_n\{\pi(f)\} - \mathsf{E}(f) + \mathsf{E}\{\pi(f)\}] + [\mathsf{E}_n\{\pi(f)\} - \mathsf{E}\{\pi(f)\}] \right| \\
&\leqslant 2n\tilde{\epsilon} + n \max_{f \in F_n} |\mathsf{E}_n(f) - \mathsf{E}(f)| \\
&\leqslant 2n\tilde{\epsilon} + n \max_{f \in F_n} |\mathsf{E}_n(f) - \mathsf{E}_n \mathsf{E}(f|\mathcal{F}_{t-1}, X_{j(-k),t})| + n \max_{f \in F_n} |\mathsf{E}_n \mathsf{E}(f|\mathcal{F}_{t-1}, X_{j(-k),t}) - \mathsf{E}(f)| \\
&= 2n\tilde{\epsilon} + K_n + N_n
\end{aligned} \tag{B.7}$$

Next, we look for the bounds for K_n and N_n, respectively. Note the summands of K_n form martingale differences. Consider the function set F_n, for each $f \in F_n$, let $\varphi_{l,t} \stackrel{\text{def}}{=} f(z_t)$ and $\tilde{\varphi}_{l,t} \stackrel{\text{def}}{=} \varphi_{l,t} - \mathsf{E}(\varphi_{l,t}|\mathcal{F}_{t-1}, X_{j(-k),t})$. Note that φ_t and $\tilde{\varphi}_t$ are vectors of length $|F_n| = (a_n/\epsilon)^{cs}$. For $l = 1, \ldots, |F_n|$, the dependence adjusted norm of $\tilde{\varphi}_{l,t}$ obeys that $\|\tilde{\varphi}_{l,\cdot}\|_{2,\varsigma} \leqslant 2\|\tilde{\varphi}_{l,t}\|_2 \lesssim 8\|\varphi_{l,t}\|_2$. Moreover, we have $\|\varphi_{l,t}\|_2^2 \lesssim L_{2n}\rho^2_{n,\upsilon}$. In particular, for the mean regression case $\rho_{n,\upsilon} = \rho_n s$, while $\rho_{n,\upsilon} = \rho_n^{1/2}$ for the median regression case (by (C5)).

Apply the tail inequality as in Lemma B.2 to the vector $\tilde{\varphi}_t$. As $\max_{1 \leqslant l \leqslant |F_n|} \|\tilde{\varphi}_{l,\cdot}\|_{2,\varsigma} \lesssim \sqrt{L_{2n}}\rho_{n,\upsilon}$ and $\|\max_{1 \leqslant l \leqslant |F_n|} \tilde{\varphi}_{l,\cdot}\|_{q,\varsigma} \lesssim \|4F(z_t)\|_q$ (by (C6)), then we can see that with probability greater than $1 - \mathcal{O}(|F_n|^{-1} + (\log |F_n|)^{-q})$,

$$\begin{aligned}
K_n &\lesssim \sqrt{ns \log(a_n/\epsilon)} \max_{1 \leqslant l \leqslant |F_n|} \|\tilde{\varphi}_{l,\cdot}\|_{2,\varsigma} + r_\varsigma \{s \log(a_n/\epsilon)\}^{3/2} \|\max_{1 \leqslant l \leqslant |F_n|} \tilde{\varphi}_{l,\cdot}\|_{q,\varsigma} \\
&\leqslant \sqrt{nL_{2n} s \log(a_n/\epsilon)} \rho_{n,\upsilon} + r_\varsigma \{s \log(a_n/\epsilon)\}^{3/2} \|8F(z_t)\|_q,
\end{aligned}$$

Hence, we have

$$K_n \lesssim \rho_{Kn}, \tag{B.8}$$

where $\rho_{K_n} \overset{\text{def}}{=} r_{k1} + r_\varsigma r_{k2}$ with $r_{k1} \overset{\text{def}}{=} \sqrt{nL_{2n} s\log(a_n/\epsilon)}\rho_{n,\upsilon}$, $r_{k2} \overset{\text{def}}{=} \{s\log(a_n/\epsilon)\}^{3/2}\|8F(z_t)\|_q$ and $r_\varsigma = n^{1/q}$ for $\varsigma > 1/2 - 1/q$ and $r_\varsigma = n^{1/2-\varsigma}$ for $\varsigma < 1/2 - 1/q$.

Then we handle the term N_n. Again consider the function set F_n, for each $f \in F_n$, let $\breve{\varphi}_{l,t} \overset{\text{def}}{=} \mathsf{E}(\varphi_{l,t}|\mathcal{F}_{t-1}, X_{j(-k),t}) - \mathsf{E}(\varphi_{l,t})$, where $\varphi_{l,t} = f(z_t)$. Then

$$N_n \leqslant \max_{1\leqslant l\leqslant |F_n|} |\sum_{t=1}^{n} \breve{\varphi}_{l,t}|.$$

Moreover, for $l = 1, \ldots, |F_n|$, there is a function g corresponding to each $f \in F_n$ such that $\breve{\varphi}_{l,t} = g(z_t, \beta, \tilde{h})$, where $\beta \in \mathcal{B}_{jk}, |\beta - \beta^0_{jk}| \leqslant C\rho_n, \tilde{h} \in \mathcal{H}_{jk}, (j,k) \in G$. By the mean value theorem and the continuity of the function g, we have

$$\begin{aligned} g(Z_{j,t}, \beta, \tilde{h}) =& \partial_\beta g(Z_{j,t}, \bar{\beta}, \tilde{h})(\beta - \beta^0_{jk}) \\ &+ \sum_{m=1}^{2} \partial_{h_m} g(Z_{j,t}, \beta, \bar{h})\{\tilde{h}_m(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}, \end{aligned}$$

where $(\bar{\beta}, \bar{h}(\cdot))$ is the corresponding point which joins the line segment between $(\beta, \tilde{h}(\cdot))$ and $(\beta^0_{jk}, h^0_{jk}(\cdot))$. Then

$$\begin{aligned} \max_{1\leqslant l\leqslant |F_n|} \sum_{t=1}^{n} \breve{\varphi}_{l,t} =& \max_{\bar{\beta}\in F_n^\beta} \sum_{t=1}^{n} \partial_\beta g(Z_{j,t}, \bar{\beta}, \tilde{h})(\beta - \beta^0_{jk}) \\ &+ \max_{\bar{h}\in F_n^h} \sum_{m=1}^{2} \sum_{t=1}^{n} \partial_{h_m} g(Z_{j,t}, \beta, \bar{h})\{\tilde{h}_m(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}, \end{aligned}$$

where F_n^β and $F_n^{\tilde{h}}$ collect all the points of β and $\tilde{h}$ according to F_n, respectively.

Recall that in our linear model setting, $h^0_{jk}(X_{j(-k),t}) = (X^\top_{j(-k),t}\beta^0_{j(-k)}, X^\top_{j(-k),t}\gamma^0_{j(-k)})^\top = (X^\top_{j(-k),t}\theta^0_{jk,1}, X^\top_{j(-k),t}\theta^0_{jk,2})^\top$, and $\tilde{h}(X_{j(-k),t}) = (X^\top_{j(-k),t}\tilde{\theta}_{jk,1}, X^\top_{j(-k),t}\tilde{\theta}_{jk,2})^\top$, where $\theta^0_{jk,m}$ and $\tilde{\theta}_{jk,m}$ $(m = 1, 2)$ are vectors of length $K-1$. Let $T^0_{jk} \overset{\text{def}}{=} \{1 \leqslant \ell \leqslant K-1 : \theta^0_{jk,1,\ell} \neq 0, \theta^0_{jk,2,\ell} \neq 0\}$, $\widetilde{T}_{jk} \overset{\text{def}}{=} \{1 \leqslant \ell \leqslant K-1 : \tilde{\theta}_{jk,1,\ell} \neq 0, \tilde{\theta}_{jk,2,\ell} \neq 0\}$, and $\breve{X}^{jk}_t \overset{\text{def}}{=} \text{vec}\{(X_{j(-k),t,\ell})_{\ell\in T^0_{jk}\bigcup \widetilde{T}_{jk}}\}$. Now we apply Lemma B.2 on $\sum_{t=1}^n \partial_{h_m} g(Z_{j,t}, \beta, \bar{h})\{\tilde{h}_m(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}$ and $\sum_{t=1}^n \partial_\beta g(Z_{j,t}, \bar{\beta}, \tilde{h})(\beta - \beta^0_{jk})$. To this end, we define the following quantities:

$$\Phi^h_{m,2,\varsigma} \overset{\text{def}}{=} \max_{\bar{h}\in F_n^{\tilde{h}}} \big\||\breve{X}^{jk}_\cdot \partial_{h_m} g(Z_{j,\cdot}, \beta, \bar{h})|_\infty\big\|_{2,\varsigma}, \ \Omega^h_{m,q,\varsigma} \overset{\text{def}}{=} \big\| \max_{\bar{h}\in F_n^{\tilde{h}}} |\breve{X}^{jk}_\cdot \partial_{h_m} g(Z_{j,\cdot}, \beta, \bar{h})|_\infty\big\|_{q,\varsigma}. \quad \text{(B.9)}$$

Let $\chi^m_t \overset{\text{def}}{=} \partial_{h_m} g(Z_{j,t}, \beta, \bar{h})\{\tilde{h}_m(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}$ and define the projector operator $\mathcal{P}_l(\chi^m_t) \overset{\text{def}}{=} \mathsf{E}(\chi^m_t|\mathcal{F}_l) - \mathsf{E}(\chi^m_t|\mathcal{F}_{l-1})$. According to Theorem 1(i) of Wu (2005), it is not hard to see that $\|\chi^m_\cdot\|_{q,\varsigma} \lesssim \sup_{d\geqslant 0}(d+1)^\varsigma \sum_{t=d}^\infty \|\mathcal{P}_0(\chi^m_t)\|_q$, for $m = 1, 2$. Moreover, as $|\tilde{\theta}_{jk,m} - \theta^0_{jk,m}|_1 \lesssim \sqrt{s_j}\rho_n \leqslant \sqrt{s}\rho_n$, we have

$$\begin{aligned} \|\mathcal{P}_0(\chi^m_t)\|_q &\leqslant \big(\mathsf{E}[\mathcal{P}_0\{|\partial_{h_m} g(Z_{j,t}, \beta, \bar{h})\breve{X}^{jk}_t|_\infty\}|\tilde{\theta}_{jk,m} - \theta^0_{jk,m}|_1]^q\big)^{1/q} \\ &\lesssim \sqrt{s}\rho_n\big(\mathsf{E}[\mathcal{P}_0\{|\partial_{h_m} g(Z_{j,t}, \beta, \bar{h})\breve{X}^{jk}_t|_\infty\}]^q\big)^{1/q}. \end{aligned}$$

It follows that $\|\chi_{\cdot}^{m}\|_{q,\varsigma} \lesssim \sqrt{s}\rho_n \| |\breve{X}_{\cdot}^{jk}|_\infty |\partial_{h_m} g(Z_{j,\cdot}, \beta, \bar{h})| \|_{q,\varsigma}$. Then applying the tail probability bounds in Lemma B.2 yields with probability approaching 1,

$$\max_{\bar{h}\in F_n^{\bar{h}}} |\sum_{t=1}^{n} \partial_{h_m} g(Z_{j,t}, \beta, \bar{h})\{\tilde{h}_m(X_{j(-k),t}) - h_{jk,m}^0(X_{j(-k),t})\}| \lesssim r_{N1,m} + r_\varsigma r_{N2,m},$$

where $r_{N1,m} = \sqrt{n}s\rho_n\{\log(a_n/\epsilon)\}^{1/2}\Phi_{m,2,\varsigma}^h$, $r_{N2,m} = s^2\rho_n\{\log(a_n/\epsilon)\}^{3/2}\Omega_{m,q,\varsigma}^h$, and the rates of $\Phi_{m,2,\varsigma}^h$ and $\Omega_{m,q,\varsigma}^h$ are restricted in (C9).

Similarly, by defining

$$\Phi_{2,\varsigma}^{\beta} \stackrel{\text{def}}{=} \max_{\bar{\beta}\in F_n^{\beta}} \|\partial_\beta g(Z_{j,\cdot}, \bar{\beta}, \tilde{h})\|_{2,\varsigma}, \ \Omega_{q,\varsigma}^{\beta} \stackrel{\text{def}}{=} \| \max_{\bar{\beta}\in F_n^{\beta}} |\partial_\beta g(Z_{j,\cdot}, \bar{\beta}, \tilde{h})| \|_{q,\varsigma}, \tag{B.10}$$

we have

$$\max_{\bar{\beta}\in F_n^{\beta}} |\sum_{t=1}^{n} \partial_\beta g(Z_{j,t}, \bar{\beta}, \tilde{h})(\beta - \beta_{jk}^0)| \lesssim r_{N1,0} + r_\varsigma r_{N2,0},$$

where $r_{N1,0} = \rho_n\sqrt{ns\log(a_n/\epsilon)}\Phi_{2,\varsigma}^{\beta}$, $r_{N2,0} = \rho_n\{s\log(a_n/\epsilon)\}^{3/2}\Omega_{q,\varsigma}^{\beta}$. And (C9) constrains the rates of $\Phi_{2,\varsigma}^{\beta}$ and $\Omega_{q,\varsigma}^{\beta}$.

As a result, with probability $1 - o(1)$,

$$N_n \lesssim \rho_{N_n}, \tag{B.11}$$

by letting $\max_{m\in\{0,1,2\}}\{r_{N1,m} + r_\varsigma r_{N2,m}\} = \mathcal{O}(\rho_{N_n})$.

As $\mathrm{P}(K_n + N_n \geqslant x) \leqslant \mathrm{P}(K_n \geqslant x/2) + \mathrm{P}(N_n \geqslant x/2)$ and collecting the results from (B.7), (B.8), and (B.11), we have shown that Δ_n satisfies

$$n^{-1} \max_{(j,k)\in G} \Delta_n \lesssim \rho_{\Delta_n},$$

where $\rho_{\Delta_n} = n^{-1}(\rho_{K_n} + \rho_{N_n}) = o(n^{-1/2}g_n^{-1})$ (given $\tilde{\epsilon}$ is sufficiently small, and using (C8) and (C9)).

□

Comment B.3. [The rates of $\Omega_{m,q,\varsigma}^h$ and $\Omega_{q,\varsigma}^{\beta}$] It is worth discussing the rates of $\Omega_{m,q,\varsigma}^h$ and $\Omega_{q,\varsigma}^{\beta}$ by the definition under some special cases. For example, consider the VAR(1) model as in Comment 5.3 given by $Y_t = AY_{t-1} + \varepsilon_t$, where $Y_t, \varepsilon_t \in \mathbb{R}^J$, and $\varepsilon_t \sim$ i.i.d. $\mathrm{N}(0, \Sigma)$. At first, as shown in the proof of Theorem 5.2, we have

$$\Omega_{m,q,\varsigma}^h = \| \max_{\bar{h}\in F_n^{\bar{h}}} |\breve{X}_{\cdot}^{jk}\partial_{h_m} g(Z_{j,\cdot}, \beta, \bar{h})|_\infty \|_{q,\varsigma} \lesssim \| \max_{(j,k)\in G} |\breve{X}_{\cdot}^{jk}|_\infty \|_{2q,\varsigma} \| \max_{\bar{h}\in F_n^{\bar{h}}} \partial_{h_m} g(Z_{j,\cdot}, \beta, \bar{h}) \|_{2q,\varsigma}.$$

For the first term, it is not hard to see that

$$\| \max_{(j,k)\in G} \{|\breve{X}_t^{jk}|_\infty - |(\breve{X}_t^{jk})^*|_\infty\} \|_{2q} \lesssim |A|_\infty^{t-1} \| |\varepsilon_0|_\infty \|_{2q} \lesssim J^{1/(2q)},$$

where the last inequality is by the union bound, assuming $|A|_\infty < 1$, and the qth moments of

$\varepsilon_{j,0}$ $(\forall j)$ are bounded by a constant μ_q. As for the second term, let $d_n \stackrel{\text{def}}{=} |G| \vee J$. In the mean regression case, for $f \in \widetilde{\mathcal{F}}$, $\mathsf{E}(f(z_t)|\mathcal{F}_{t-1}) = \{X_{jk,t}(\beta^0_{jk} - \beta) + h^0_1 - \tilde{h}_1\}(v_{jk,t} + h^0_2 - \tilde{h}_2)$, it can be seen that

$$
\begin{aligned}
&\| \max_{\bar{h} \in F^{\bar{h}}_n} \{\partial_{h_1} g(Z_{j,t}, \beta, \bar{h}) - \partial_{h_1} g(Z^*_{j,t}, \beta, \bar{h})\} \|_{2q} \\
&\leqslant \| \max_{(j,k)\in G} |v_{jk,t} - v^*_{jk,t}| \|_{2q} + \| \max_{(j,k)\in G} \{|X^\top_{j(-k),t} - (X^\top_{j(-k),t})^*| \max_{\bar{\gamma}_{j(-k)}} |\gamma^0_{j(-k)} - \bar{\gamma}_{j(-k)}|\} \|_{2q} \\
&\lesssim d_n^{1/(2q)} (1 \vee s^{1/2} \rho_n),
\end{aligned}
$$

while in the median regression case, for $f \in \widetilde{\mathcal{F}}$, $\mathsf{E}(f(z_t)|\mathcal{F}_{t-1}) = [\frac{1}{2} - F_{\varepsilon_{j,t}|\mathcal{F}_{t-1}}\{X_{jk,t}(\beta^0_{jk} - \beta) + h^0_1 - \tilde{h}_1\}](v_{jk,t} + h^0_2 - \tilde{h}_2)$,

$$
\begin{aligned}
&\| \max_{\bar{h} \in F^{\bar{h}}_n} \{\partial_{h_1} g(Z_{j,t}, \beta, \bar{h}) - \partial_{h_1} g(Z^*_{j,t}, \beta, \bar{h})\} \|_{2q} \\
&\lesssim \| \max_{(j,k)\in G} |v_{jk,t} - v^*_{jk,t}| \|_{4q} + \| \max_{(j,k)\in G} \{|X^\top_{j(-k),t} - (X^\top_{j(-k),t})^*| \max_{\bar{\gamma}_{j(-k)}} |\gamma^0_{j(-k)} - \bar{\gamma}_{j(-k)}|\} \|_{4q} \\
&\lesssim d_n^{1/(4q)} (1 \vee s^{1/2} \rho_n),
\end{aligned}
$$

where we use the assumption such that the $4q$th moment of the conditional density is bounded. Moreover, we have

$$
\begin{aligned}
&\| \max_{\bar{h} \in F^{\bar{h}}_n} \{\partial_{h_2} g(Z_{j,t}, \beta, \bar{h}) - \partial_{h_2} g(Z^*_{j,t}, \beta, \bar{h})\} \|_{2q} \\
&\leqslant \| \max_{(j,k)\in G} |(X_{j(-k),t} - X^*_{j(-k),t})(\beta^0_{jk} - \beta)| \|_{2q} \\
&\quad + \| \max_{(j,k)\in G} \{|X^\top_{j(-k),t} - (X^\top_{j(-k),t})^*| \max_{\bar{\beta}_{j(-k)}} |\beta^0_{j(-k)} - \bar{\beta}_{j(-k)}|\} \|_{2q} \\
&\lesssim d_n^{1/(2q)} (1 \vee s^{1/2} \rho_n),
\end{aligned}
$$

or $\| \max_{\bar{h} \in F^{\bar{h}}_n} \{\partial_{h_2} g(Z_{j,t}, \beta, \bar{h}) - \partial_{h_2} g(Z^*_{j,t}, \beta, \bar{h})\} \|_{2q} = \mathcal{O}(1)$ for the two cases. Therefore, we are able to conclude that $\Omega^h_{m,q,\varsigma} \lesssim d_n^{1/q}(1 \vee s^{1/2}\rho_n)$ or $\Omega^h_{m,q,\varsigma} \lesssim d_n^{3/(4q)}(1 \vee s^{1/2}\rho_n)$, respectively.

Similarly, it can be shown that $\Omega^\beta_{q,\varsigma} \lesssim d_n^{1/q} s^{1/2} \rho_n$ or $\Omega^\beta_{q,\varsigma} \lesssim d_n^{1/(2q)} s^{1/2} \rho_n$ for the two cases, since

$$
\begin{aligned}
&\| \max_{\bar{\beta} \in F^{\beta}_n} |\partial_\beta g(Z_{j,\cdot}, \bar{\beta}, \tilde{h})| \|_q \\
&\lesssim \| \max_{(j,k)\in G} |X^\top_{j(-k),t} - (X^\top_{j(-k),t})^*| \|_{2q} \| \max_{(j,k)\in G} |\{X^\top_{j(-k),t} - (X^\top_{j(-k),t})^*\}\{\gamma^0_{j(-k)} - \bar{\gamma}_{j(-k)}\}| \|_{2q} \\
&\lesssim d_n^{1/q} s^{1/2} \rho_n,
\end{aligned}
$$

or

$$\begin{aligned}
&\| \max_{\bar{\beta}\in F_n^{\beta}} |\partial_\beta g(Z_{j,\cdot}, \bar{\beta}, \tilde{h})| \|_q \\
&\lesssim \| \max_{(j,k)\in G} |X_{j(-k),t}^\top - (X_{j(-k),t}^\top)^*| \|_{4q} \| \max_{(j,k)\in G} |\{X_{j(-k),t}^\top - (X_{j(-k),t}^\top)^*\}\{\gamma_{j(-k)}^0 - \bar{\gamma}_{j(-k)}\}| \|_{4q} \\
&\lesssim d_n^{1/(2q)} s^{1/2} \rho_n.
\end{aligned}$$

In addition, a similar derivation can show that $\|F(z_t)\|_q \lesssim d_n^{1/q}(1 \vee \rho_n)$ and $\| \max_{(j,k)\in G} |\psi_{jk,\cdot}^0| \|_{q,\varsigma} \lesssim d_n^{1/q}(1 \vee \rho_n)$.

Lemma B.6. *Under the same conditions as in Theorem 5.4, we have with probability* $1 - o(1)$,

$$\max_{(j,k)\in G} |\mathsf{E}_n \psi_{jk}\{Z_{j,t}, \beta_{jk}^0, h_{jk}^0(X_{j(-k),t})\}| \lesssim r_n. \tag{B.12}$$

Proof of Lemma B.6. Consider the class of function $\mathcal{F}_G = \{z \mapsto \psi_{jk}\{z, \beta_{jk}^0, h_{jk}^0(x_{j(-k)})\} : (j,k) \in G\}$, the cardinality of the set is $|G|$. Therefore, the corresponding covering number is given by $\sup_{\mathcal{Q}} \mathcal{N}(\epsilon \|\bar{F}_G\|_{\mathcal{Q},2}, \mathcal{F}_G, \|\cdot\|_{\mathcal{Q},2}) = |G|/\epsilon$, with $\bar{F}_G = \sup_{f\in\mathcal{F}_G} |f|$. Let $\psi_{jk,t}^0 \stackrel{\text{def}}{=} \psi_{jk}\{Z_{j,t}, \beta_{jk}^0, h_{jk}^0(X_{j(-k),t})\}$ and applying the tail probability bounds in Lemma B.2, we have with probability $1 - o(1)$,

$$\max_{(j,k)\in G} |\mathsf{E}_n \psi_{jk,t}^0| \lesssim n^{-1}(r_1 + r_\varsigma r_2) \lesssim r_n, \tag{B.13}$$

where $r_1 = (n \log a_n)^{1/2} \max_{(j,k)\in G} \|\psi_{jk,\cdot}^0\|_{2,\varsigma}$, $r_2 = (\log a_n)^{3/2} \| \max_{(j,k)\in G} |\psi_{jk,\cdot}^0| \|_{q,\varsigma}$, $r_\varsigma = n^{1/q}$ for $\varsigma > 1/2 - 1/q$ and $r_\varsigma = n^{1/2-\varsigma}$ for $\varsigma < 1/2 - 1/q$. □

Lemma B.7. *Under the same conditions as in Theorem 5.4, consider the class of functions* $\mathcal{F}' = \{z \mapsto \psi_{jk}\{z, \beta, \tilde{h}(x_{j(-k)})\} : (j,k) \in G, \beta \in \mathcal{B}_{jk}, \tilde{h} \in \mathcal{H}_{jk} \cup \{h_{jk}^0\}\}$, *we have with probability* $1 - o(1)$,

$$n^{-1/2} \sup_{f\in\mathcal{F}'} |G_n(f)| \lesssim \rho_n. \tag{B.14}$$

Proof of Lemma B.7. The covering number of the function class $\mathcal{F}'$ is given by $\sup_{\mathcal{Q}} \mathcal{N}(\epsilon \|\bar{F}'\|_{\mathcal{Q},2}, \mathcal{F}', \|\cdot\|_{\mathcal{Q},2}) = (a_n/\epsilon)^{cs}$, with $\bar{F}' = \sup_{f\in\mathcal{F}'} |f|$. Also, for any $f \in \mathcal{F}'$, there exists a set F_n' such that $\min_{f'\in F_n'} \|f - f'\|_{\mathcal{Q},2} \leqslant \epsilon \|\bar{F}'\|_{\mathcal{Q},2}$ and the cardinality of the set $|F_n'| = (a_n/\epsilon)^{cs}$.

One can apply the technique we used in the proof of Lemma B.5 to achieve the concentration inequality. Similarly, consider the function set F_n', for each $f \in F_n'$, let $\varphi_{l,t} \stackrel{\text{def}}{=} f(z_t)$ and $\tilde{\varphi}_{l,t} \stackrel{\text{def}}{=} \varphi_{l,t} - \mathsf{E}(\varphi_{l,t}|\mathcal{F}_{t-1}, X_{j(-k),t})$, $l = 1, \ldots, |F_n'|$. We have

$$n \max_{f\in F_n'} |\mathsf{E}_n f - \mathsf{E}_n \mathsf{E}(f|\mathcal{F}_{t-1}, X_{j(-k),t})| \lesssim 4\sqrt{ns \log(a_n/\epsilon)} \max_{f\in\mathcal{F}'} \|f(z_t)\|_2 + r_\varsigma \{s \log(a_n/\epsilon)\}^{3/2} \|4\bar{F}'(z_t)\|_q.$$

For each $f \in F_n'$, there exists a function g such that $g(z_t, \beta, \tilde{h}) = \mathsf{E}\{f(z_t)|\mathcal{F}_{t-1}, X_{j(-k),t}\} - \mathsf{E}\{f(z_t)\}$, where $\beta \in \mathcal{B}_{jk}, \tilde{h} \in \mathcal{H}_{jk} \cup \{h_{jk}^0\}, (j,k) \in G$. As by the mean value theorem and the

continuity of the function g, we have

$$g(Z_{j,t},\beta,\tilde{h}) = \partial_\beta g(Z_{j,t},\bar{\beta},\tilde{h})(\beta-\beta^0_{jk}) + \sum_{m=1}^{2}\partial_{h_m} g(Z_{j,t},\beta,\bar{h})\{\tilde{h}_m(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\},$$

where $(\bar{\beta},\bar{h}(\cdot))$ is the corresponding point which joins the line segment between $(\beta,\tilde{h}(\cdot))$ and $(\beta^0_{jk},h^0_{jk}(\cdot))$. Let $F_n^{\prime\beta}$ and $F_n^{\prime\tilde{h}}$ collect all the points of β and $\tilde{h}$ according to F'_n, and define the following quantities $(m=1,2)$

$$\Phi^{\prime h}_{m,2,\varsigma} \stackrel{\text{def}}{=} \max_{\bar{h}\in F_n^{\prime\tilde{h}}} \big\||\check{X}^{jk}_{\cdot}\partial_{h_m} g(Z_{j,\cdot},\beta,\bar{h})|_\infty\big\|_{2,\varsigma},\ \Omega^{\prime h}_{m,q,\varsigma} \stackrel{\text{def}}{=} \big\|\max_{\bar{h}\in F_n^{\prime\tilde{h}}} |\check{X}^{jk}_{\cdot}\partial_{h_m} g(Z_{j,\cdot},\beta,\bar{h})|_\infty\big\|_{q,\varsigma},$$
$$\Phi^{\prime\beta}_{2,\varsigma} \stackrel{\text{def}}{=} \max_{\bar{\beta}\in F_n^{\prime\beta}} \big\|\partial_\beta g(Z_{j,\cdot},\bar{\beta},\tilde{h})\big\|_{2,\varsigma},\ \Omega^{\prime\beta}_{q,\varsigma} \stackrel{\text{def}}{=} \big\|\max_{\bar{\beta}\in F_n^{\prime\beta}} |\partial_\beta g(Z_{j,\cdot},\bar{\beta},\tilde{h})|\big\|_{q,\varsigma}. \tag{B.15}$$

Then we have with probability approaching 1,

$$\max_{\bar{h}\in F_n^{\prime\tilde{h}}} \big|\sum_{t=1}^{n}\partial_{h_m} g(Z_{j,t},\beta,\bar{h})\{\tilde{h}_m(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}\big| \lesssim r'_{N1,m} + r_\varsigma r'_{N2,m},\ m=1,2,$$
$$\max_{\bar{\beta}\in F_n^{\prime\beta}} \big|\sum_{t=1}^{n}\partial_\beta g(Z_{j,t},\bar{\beta},\tilde{h})(\beta-\beta^0_{jk})\big| \lesssim r'_{N1,0} + r_\varsigma r'_{N2,0},$$

where $r'_{N1,m} = \sqrt{n}s\rho_n\{\log(a_n/\epsilon)\}^{1/2}\Phi^{\prime h}_{m,2,\varsigma}$, $r'_{N2,m} = s^2\rho_n\{\log(a_n/\epsilon)\}^{3/2}\Omega^{\prime h}_{m,q,\varsigma}$, and $r'_{N1,0} = \rho_n\sqrt{ns\log(a_n/\epsilon)}\Phi^{\prime\beta}_{2,\varsigma}$, $r'_{N2,0} = \rho_n\{s\log(a_n/\epsilon)\}^{3/2}\Omega^{\prime\beta}_{q,\varsigma}$. Also (C9) constrains the rates of $\Phi^{\prime h}_{m,2,\varsigma}$, $\Omega^{\prime h}_{m,q,\varsigma}$, $\Phi^{\prime\beta}_{2,\varsigma}$, and $\Omega^{\prime\beta}_{q,\varsigma}$.

The rest of the proof is similar as for Lemma B.5 and thus is omitted. □

Lemma B.8. *Under the same conditions as in Lemma B.5 with (C9') instead of (C6), (C8) and (C9),*

$$n^{-1}\max_{(j,k)\in G}\Delta_n \lesssim o(n^{-1/2}g_n^{-1}), \tag{B.16}$$

holds with probability $1-o(1)$.

Proof of Lemma B.8. We now study the tail probability under stronger tail assumptions. In particular, we need to carry out an analogue proof of Lemma B.5 under (C9').

Specifically, by Lemma B.3, we have $K_n \lesssim n^{1/2}(s\log a_n)^{1/\gamma}\rho^e_{n,\upsilon}$ (in particular, for the mean regression case $\rho^e_{n,\upsilon} = \rho^e_n s$ and $\rho^e_{n,\upsilon} = \sqrt{\rho^e_n}$), and

$$N_n \lesssim n^{1/2}(s\log(a_n/\epsilon))^{1/\gamma}\rho^e_n\{(s^{1/2}\max_{m\in\{1,2\}}\Phi^h_{m,\psi_\nu,0}) \vee \Phi^\beta_{\psi_\nu,0}\},$$
$$\Phi^h_{m,\psi_\nu,0} \stackrel{\text{def}}{=} \max_{\bar{h}\in F_n^{\tilde{h}}}\big\||\check{X}^{jk}_{\cdot}\partial_{h_m} g(Z_{j,\cdot},\beta,\bar{h})|_\infty\big\|_{\psi_\nu,0},\ \Phi^\beta_{\psi_\nu,0} \stackrel{\text{def}}{=} \max_{\bar{\beta}\in F_n^{\beta}}\big\|\partial_\beta g(Z_{j,\cdot},\bar{\beta},\tilde{h})\big\|_{\psi_\nu,0}. \tag{B.17}$$

The rest of the proof is similar as for Lemma B.5 and thus is omitted.

□

Lemma B.9. *Under the same conditions as in Lemma B.6 with (C9') instead of (C6), (C8) and (C9), and assume that* $\max_{(j,k)\in G}\|\psi^0_{jk,\cdot}\|_{\psi_\nu,0}<\infty$, *we have with probability* $1-o(1)$,

$$\max_{(j,k)\in G}|\mathsf{E}_n\psi_{jk}\{Z_{j,t},\beta^0_{jk},h^0_{jk}(X_{j(-k),t})\}|\lesssim n^{-1/2}(\log a_n)^{1/\gamma}\max_{(j,k)\in G}\|\psi^0_{jk,\cdot}\|_{\psi_\nu,0}\lesssim r_n. \tag{B.18}$$

Proof of Lemma B.9. The proof is similar to the proof of Lemma B.6 by replacing the tail probability bounds therein by Lemma B.3. □

Lemma B.10. *Under the same conditions as in Lemma B.7 with (C9') instead of (C6), (C8) and (C9), and assume that* $\max_{f\in\mathcal{F}'}\|f(z_\cdot)\|_{\psi_\nu,0}<\infty$, *we have with probability* $1-o(1)$,

$$n^{-1/2}\sup_{f\in\mathcal{F}'}|G_n(f)|\lesssim\rho^e_n. \tag{B.19}$$

Proof of Lemma B.10. The proof is similar to the proof of Lemma B.7 by replacing the tail probability bounds therein by Lemma B.3. In particular, it can be shown that

$$\begin{aligned}&n^{-1/2}\sup_{f\in\mathcal{F}'}|G_n(f)|\lesssim n^{-1/2}(s\log(a_n/\epsilon))^{1/\gamma}\big[\max_{f\in\mathcal{F}'}\|f(z_\cdot)\|_{\psi_\nu,0}\vee\rho^e_n\{(s^{1/2}\max_{m\in\{1,2\}}\Phi'^h_{m,\psi_\nu,0})\vee\Phi'^\beta_{\psi_\nu,0}\}\big],\\&\Phi'^h_{m,\psi_\nu,0}\stackrel{\text{def}}{=}\max_{\bar h\in F'^h_n}\big\||\breve X^{jk}_\cdot\partial_{h_m}g(Z_{j,\cdot},\beta,\bar h)|_\infty\big\|_{\psi_\nu,0},\ \Phi'^\beta_{\psi_\nu,0}\stackrel{\text{def}}{=}\max_{\bar\beta\in F'^\beta_n}\big\|\partial_\beta g(Z_{j,\cdot},\bar\beta,\tilde h)\big\|_{\psi_\nu,0}.\end{aligned} \tag{B.20}$$

The final conclusion can be achieved by (C9'). □

B.4.2 Proofs of Section 5.6

Proof of Theorem 5.4. The sketch of the proof follows the proof of Theorem 2 in Belloni et al. (2015b).

Step 1: Let $\tilde\beta_{jk}$ be any estimator such that $\max_{(j,k)\in G}|\tilde\beta_{jk}-\beta^0_{jk}|\leqslant C\rho_n$ with probability $1-o(1)$. By rewriting (using the fact that $\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta^0_{jk},h^0_{jk}(X_{j(-k),t})\}]=0$), we have

$$\begin{aligned}\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\tilde\beta_{jk},\widehat h_{jk}(X_{j(-k),t})\}]=&\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta^0_{jk},h^0_{jk}(X_{j(-k),t})\}]\\&+\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,\tilde h(X_{j(-k),t})\}]\big|_{\beta=\tilde\beta_{jk},\tilde h=\widehat h_{jk}}+n^{-1}\Delta_n\end{aligned} \tag{B.21}$$

where $\Delta_n\stackrel{\text{def}}{=}n^{1/2}G_n[\psi_{jk}\{Z_{j,t},\tilde\beta_{jk},\widehat h_{jk}(X_{j(-k),t})\}-\psi_{jk}\{Z_{j,t},\beta^0_{jk},h^0_{jk}(X_{j(-k),t})\}]$.

We first observe that with probability $1-o(1)$, $\max_{(j,k)\in G}\Delta_n\leqslant\sqrt{n}\sup_{f\in\widetilde{\mathcal{F}}}|G_n(f)|$, where $\widetilde{\mathcal{F}}$ is the class of functions defined by $\widetilde{\mathcal{F}}=\{z\mapsto\psi_{jk}\{z,\beta,\tilde h(x_{j(-k)})\}-\psi_{jk}\{z,\beta^0_{jk},h^0_{jk}(x_{j(-k)})\}:(j,k)\in G,\beta\in\mathcal{B}_{jk},|\beta-\beta^0_{jk}|\leqslant C\rho_n,\tilde h\in\mathcal{H}_{jk}\}$. The key to our proof is to achieve a concentration inequality for Δ_n, such that $n^{-1}\max_{(j,k)\in G}\Delta_n\lesssim o(n^{-1/2}g_n^{-1})$ holds with probability $1-o(1)$. This is done in Lemma B.5.

Then we expand the second term in (B.21) by Taylor expansion. Pick any $\beta\in\mathcal{B}_{jk}$ such that $|\beta-\beta^0_{jk}|\leqslant C\rho_n$ and $\tilde h\in\mathcal{H}_{jk}$. For any $(j,k)\in G$, let $(\bar\beta,\bar h(X_{j(-k),t})^\top)^\top$ lie on the line segment between $(\beta,\tilde h(X_{j(-k),t})^\top)^\top$ and $(\beta^0_{jk},h^0_{jk}(X_{j(-k),t})^\top)^\top$. Therefore, we can write

$\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,\tilde{h}(X_{j(-k),t})\}]$ as follows

$$\begin{aligned}
&\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}]+\mathsf{E}(\partial_\beta\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}|X_{j(-k),t}])(\beta-\beta_{jk}^0)\\
&+\sum_{m=1}^{M}\mathsf{E}(\partial_{h_m}\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}|X_{j(-k),t}]\{\tilde{h}_m(X_{j(-k),t})-h_{jk,m}^0(X_{j(-k),t})\})\\
&+\frac{1}{2}\,\mathsf{E}(\partial_\beta^2\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\bar{\beta},\bar{h}(X_{j(-k),t})\}|X_{j(-k),t}])(\beta-\beta_{jk}^0)^2\\
&+\frac{1}{2}\sum_{m,m'=1}^{M}\mathsf{E}(\partial_{h_m}\partial_{h_{m'}}\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\bar{\beta},\bar{h}(X_{j(-k),t})\}|X_{j(-k),t}]\{\tilde{h}_m(X_{j(-k),t})-h_{jk,m}^0(X_{j(-k),t})\}\\
&\qquad\qquad\{\tilde{h}_{m'}(X_{j(-k),t})-h_{jk,m'}^0(X_{j(-k),t})\})\\
&+\frac{1}{2}\sum_{m=1}^{M}\mathsf{E}(\partial_{h_m}\partial_\beta\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\bar{\beta},\bar{h}(X_{j(-k),t})\}|X_{j(-k),t}]\{\tilde{h}_m(X_{j(-k),t})-h_{jk,m}^0(X_{j(-k),t})\})(\beta-\beta_{jk}^0).
\end{aligned}\tag{B.22}$$

It can be seen from the orthogonality condition (5.22) that the third term in (B.22) is zero. By (C3) we have $\mathsf{E}(\partial_\beta\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}|X_{j(-k),t}])=\partial_\beta\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}]$ $=\phi_{jk}$. Moreover, each of the last three terms in (B.22) is $\mathcal{O}(L_{1n}\rho_n^2)=o(n^{-1/2}g_n^{-1})$ (by (C3), (C5) and (C8)). Therefore, we have shown that the second term in (B.21) equals $\phi_{jk}(\tilde{\beta}_{jk}-\beta_{jk}^0)+o(n^{-1/2}g_n^{-1})$, uniformly over $(j,k)\in G$. Then, combining the results in Lemma B.5 gives

$$\begin{aligned}
\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\tilde{\beta}_{jk},\widehat{h}_{jk}(X_{j(-k),t})\}]=&\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}]\\
&+\phi_{jk}(\tilde{\beta}_{jk}-\beta_{jk}^0)+o(n^{-1/2}g_n^{-1}).
\end{aligned}\tag{B.23}$$

Step 2: Next, we need to prove that $\inf_{\beta\in\widehat{\mathcal{B}}_{jk}}|\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta,\widehat{h}_{jk}(X_{j(-k),t})\}]|=o(n^{-1/2}g_n^{-1})$ holds with probability $1-o(1)$. For any $(j,k)\in G$, we focus on any point $\beta_{jk}^*=\beta_{jk}^0-\phi_{jk}^{-1}\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}]$, thus

$$\max_{(j,k)\in G}|\beta_{jk}^*-\beta_{jk}^0|\leqslant C\max_{(j,k)\in G}|\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}]|.$$

By Lemma B.6, we have $|\beta_{jk}^*-\beta_{jk}^0|\lesssim r_n$ uniformly over $(j,k)\in G$. By (C2), $[\beta_{jk}^0\pm c_1r_n]\subset\widehat{\mathcal{B}}_{jk}$ with probability $1-o(1)$, thus β_{jk}^* is contained in $\widehat{\mathcal{B}}_{jk}$ with probability $1-o(1)$. Using the continuity argument as in (B.23) with $\tilde{\beta}_{jk}=\beta_{jk}^*$ and combining the fact that $\phi_{jk}(\beta_{jk}^*-\beta_{jk}^0)=-\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}]$, we have,

$$\begin{aligned}
\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta_{jk}^*,\widehat{h}_{jk}(X_{j(-k),t})\}]&=\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}]+\phi_{jk}(\beta_{jk}^*-\beta_{jk}^0)+o(n^{-1/2}g_n^{-1})\\
&=o(n^{-1/2}g_n^{-1}).
\end{aligned}$$

Therefore,

$$\max_{(j,k)\in G}\inf_{\beta\in\widehat{\mathcal{B}}_{jk}}|\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta,\widehat{h}_{jk}(X_{j(-k),t})\}]|\leqslant|\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta_{jk}^*,\widehat{h}_{jk}(X_{j(-k),t})\}]|=o(n^{-1/2}g_n^{-1}),\tag{B.24}$$

holds with probability $1-o(1)$ uniformly over $(j,k)\in G$.

Step 3: Lastly, it is left to prove that with probability $1-o(1)$, $\max_{(j,k)\in G}|\widehat{\beta}_{jk}-\beta^0_{jk}|\leqslant C\rho_n$, which will lead to the desired Bahadur representation. Consider the class of functions $\mathcal{F}'=\{z\mapsto\psi_{jk}\{z,\beta,\tilde{h}(x_{j(-k)})\}:(j,k)\in G,\beta\in\mathcal{B}_{jk},\tilde{h}\in\mathcal{H}_{jk}\cup\{h^0_{jk}\}\}$. From (B.24) and by the definition of $\widehat{\beta}_{jk}$ we have

$$|\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\widehat{\beta}_{jk},\widehat{h}_{jk}(X_{j(-k),t})\}]|\geqslant|\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,\tilde{h}(X_{j(-k),t})\}]|_{\beta=\widehat{\beta}_{jk},\tilde{h}=\widehat{h}_{jk}}|-n^{-1/2}\sup_{f\in\mathcal{F}'}|G_n(f)|,$$

holds with probability $1-o(1)$ uniformly over $(j,k)\in G$.

Lemma B.7 ensures that $n^{-1/2}\sup_{f\in\mathcal{F}'}|G_n(f)|=\mathcal{O}(\rho_n)$. Furthermore, applying the expansion in (B.22) with $\beta^0_{jk}=\beta$ implies that

$$|\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,\tilde{h}(X_{j(-k),t})\}]-\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,h^0_{jk}(X_{j(-k),t})\}]|\leqslant C(\rho_n+L_{1n}\rho_n^2)=\mathcal{O}(\rho_n).$$

By (C3) along with the fact that $\mathsf{E}[\{\tilde{h}_m(X_{j(-k),t})-h^0_{jk,m}(X_{j(-k),t})\}^2]\leqslant C\rho_n^2$ for all $m=1,\ldots,M$ and any $\tilde{h}=(\tilde{h}_m)_{m=1}^M\in\mathcal{H}_{jk}$, we have with probability $1-o(1)$,

$$|\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,\tilde{h}(X_{j(-k),t})\}]|_{\beta=\widehat{\beta}_{jk},\tilde{h}=\widehat{h}_{jk}}|\geqslant|\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,h^0_{jk}(X_{j(-k),t})\}]|_{\beta=\widehat{\beta}_{jk}}|-\mathcal{O}(\rho_n),\ \text{(B.25)}$$

uniformly over $(j,k)\in G$.

From (B.24) we can see that the left-hand side of (B.25) is $o(n^{-1/2}g_n^{-1})$. Moreover, due to the identification condition (C4), the first term on the right-hand side of (B.25) is bounded from below by $\frac{1}{2}\{|\phi_{jk}(\widehat{\beta}_{jk}-\beta^0_{jk})|\wedge c_1\}$ and this results in $|\widehat{\beta}_{jk}-\beta^0_{jk}|\leqslant o(n^{-1/2}g_n^{-1})+\mathcal{O}(\rho_n)$, with probability $1-o(1)$.

In summary, we have shown that, with probability $1-o(1)$,

$$\begin{aligned}\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\widehat{\beta}_{jk},\widehat{h}_{jk}(X_{j(-k),t})\}]=&\,\mathsf{E}_n[\psi_{jk}\{Z_{j,t},\beta^0_{jk},h^0_{jk}(X_{j(-k),t})\}]\\&+\phi_{jk}(\widehat{\beta}_{jk}-\beta^0_{jk})+o(n^{-2}g_n^{-1}),\end{aligned}\qquad\text{(B.26)}$$

uniformly over $(j,k)\in G$. And with probability $1-o(1)$, the left-hand side is $o(n^{-1/2}g_n^{-1})$ uniformly over $(j,k)\in G$. Lastly, the uniform Bahadur representation can be obtained by solving (B.26) with respect to $(\widehat{\beta}_{jk}-\beta^0_{jk})$.

□

Proof of Corollary 5.5. The proof is an application of Theorem 5.4 with verification of conditions (C1)-(C9).

Here we focus on the estimator by Algorithm 2 as the proof of Algorithm 1 is basically the same. In particular, with the LAD regression case, we have $|G|=1$, $a_n=\max(JK,n)$, $g_n=1$, $M=2$, $h^0_{jk}(X_{j(-k),t})=(X^\top_{j(-k),t}\beta^0_{j(-k)},X^\top_{j(-k),t}\gamma^0_{j(-k)})^\top$, $\psi_{jk}\{Z_{j,t},\beta_{jk},h^0_{jk}(X_{j(-k),t})\}=\{1/2-\mathbf{1}(Y_{j,t}\leqslant X_{jk,t}\beta_{jk}+X^\top_{j(-k),t}\beta^0_{j(-k)})\}(X_{jk,t}-X^\top_{j(-k),t}\gamma^0_{j(-k)})$.

Verification of (C1): Our model setting assumes $F_{\varepsilon_j}(0)=1/2$ and $\mathsf{E}(v_{jk,t}X_{j(-k),t})=0$;

hence we have

$$\begin{aligned}
&\mathsf{E}(\partial_{h_1} \mathsf{E}[\psi_{jk}\{Z_{j,t}, \beta^0_{jk}, h^0_{jk}(X_{j(-k),t})\}|X_{j(-k),t}]h_1(X_{j(-k),t})) \\
&= -\beta^\top_{j(-k)} \mathsf{E}\{f_{\varepsilon_j}(0)v_{jk,t}X_{j(-k),t}\} = 0 \\
&\mathsf{E}(\partial_{h_2} \mathsf{E}[\psi_{jk}\{Z_{j,t}, \beta^0_{jk}, h^0_{jk}(X_{j(-k),t})\}|X_{j(-k),t}]h_2(X_{j(-k),t})) \\
&= -\gamma^\top_{j(-k)} \mathsf{E}[\{1/2 - F_{\varepsilon_j}(0)\}X_{j(-k),t}] = 0
\end{aligned}$$

Verification of (C2): The true parameter β^0_{jk} satisfies (5.21) given $F_{\varepsilon_j}(0) = 1/2$. Moreover, based on the fact that $|\widehat{\beta}^{[1]}_{jk} - \beta^0_{jk}|_{j,pr} \lesssim \sqrt{s(\log a_n)/n}$ (according to Corollary 5.4) and by Remark 2 in Belloni et al. (2015a), with probability $1 - o(1)$, $|\widehat{\beta}^{[2]}_{jk} - \beta^0_{jk}| = o(1/\log n)$, so that for some sufficiently small $c > 0$, $[\beta^0_{jk} \pm c/\log n] \subset \widehat{\mathcal{B}}_{jk} \subset \mathcal{B}_{jk}$, with probability $1 - o(1)$. Then the condition holds.

Verification of (C3): The map

$$\begin{aligned}
(\beta, h) \mapsto &\mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h)|X_{j(-k),t}\} \\
&= \mathsf{E}([1/2 - F_{\varepsilon_j}\{X_{jk,t}(\beta - \beta^0_{jk}) - X^\top_{j(-k),t}\beta^0_{j(-k)} + h_1\}](X_{jk,t} - h_2)|X_{j(-k),t})
\end{aligned}$$

is twice continuously differentiable as f'_{ε_j} is continuous. For every $\vartheta \in \{\beta, h_1, h_2\}$, $\partial_\vartheta \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h)|X_{j(-k),t}\}$ is $-\mathsf{E}[f_{\varepsilon_j}\{X_{jk,t}(\beta - \beta^0_{jk}) - X^\top_{j(-k),t}\beta^0_{j(-k)} + h_1\}X_{jk,t}(X_{jk,t} - h_2)|X_{j(-k),t}]$ (w.r.t. β) or $-\mathsf{E}[f_{\varepsilon_j}\{X_{jk,t}(\beta - \beta^0_{jk}) - X^\top_{j(-k),t}\beta^0_{j(-k)} + h_1\}(X_{jk,t} - h_2)|X_{j(-k),t}]$ (w.r.t. h_1) or $-\mathsf{E}[1/2 - F_{\varepsilon_j}\{X_{jk,t}(\beta - \beta^0_{jk}) - X^\top_{j(-k),t}\beta^0_{j(-k)} + h_1\}|X_{j(-k),t}]$ (w.r.t. h_2). Hence, for every $\beta \in \mathcal{B}_{jk}$,

$$|\partial_\vartheta \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h^0_{jk}(X_{j(-k),t})|X_{j(-k),t}\}| \leqslant C_1 \mathsf{E}(|X_{jk,t}v_{jk,t}|\,|X_{j(-k),t}) \vee C_1 \mathsf{E}(|v_{jk,t}|\,|X_{j(-k),t}) \vee 1.$$

Therefore, the expectation of the square of the right-hand side is bounded. Moreover, let $\mathcal{T}_{jk}(X_{j(-k),t}) = \{\tau \in \mathbb{R}^2 : |\tau_2 - X^\top_{j(-k),t}\beta^0_{j(-k)}| \leqslant c_3\}$, where $c_3 > 0$ is a constant. Then for every $\vartheta, \vartheta' \in \{\beta, h_1, h_2\}$, $\beta \in \mathcal{B}_{jk}$, $h \in \mathcal{T}_{jk}(X_{j(-k),t})$, we have

$$\begin{aligned}
&|\partial_\vartheta \partial_{\vartheta'} \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h)|X_{j(-k),t}\}| \\
&\leqslant C_1[1 \vee \mathsf{E}\{|X^2_{jk,t}(X_{jk,t} - h_2)|\,|X_{j(-k),t}\} \vee \mathsf{E}\{|X_{jk,t}(X_{jk,t} - h_2)|\,|X_{j(-k),t}\} \vee \mathsf{E}(|X_{jk,t}|\,|X_{j(-k),t}) \\
&\qquad \vee \mathsf{E}(|X_{jk,t} - h_2|\,|X_{j(-k),t})].
\end{aligned}$$

In particular,

$$\begin{aligned}
\mathsf{E}\{|X^2_{jk,t}(X_{jk,t} - h_2)|\,|X_{j(-k),t}\} &\leqslant \mathsf{E}\{|(X^\top_{j(-k),t}\gamma^0_{j(-k)} + v_{jk,t})^2(c_3 + |v_{jk,t}|)|\,|X_{j(-k),t}\} \\
&\leqslant 2\,\mathsf{E}\{|\{(X^\top_{j(-k),t}\gamma^0_{j(-k)})^2 + v^2_{jk,t}\}(c_3 + |v_{jk,t}|)|\,|X_{j(-k),t}\} \\
&\leqslant C|X^\top_{j(-k),t}\gamma^0_{j(-k)}|^2.
\end{aligned}$$

And by similar computation we can show that $|\partial_\vartheta \partial_{\vartheta'} \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h)|X_{j(-k),t}\}| \leqslant \ell_1(X_{j(-k),t}) = C'|X^\top_{j(-k),t}\gamma^0_{j(-k)}|^2$, where the constants C, C' dependd on c_3 and C_1. Lastly, for every $\beta, \beta' \in$

$\mathcal{B}_{jk}$, $h, h' \in \mathcal{T}_{jk}(X_{j(-k),t})$ we have

$$\begin{aligned}\mathsf{E}[\{\psi_{jk}(Z_{j,t},\beta,h)-\psi_{jk}(Z_{j,t},\beta',h')\}^2|X_{j(-k),t}] &\leqslant C_1\,\mathsf{E}\{|X_{jk,t}(X_{jk,t}-h_2)^2|\,|X_{j(-k),t}\}|\beta-\beta'| \\ &\quad + C_1\,\mathsf{E}\{(X_{jk,t}-h_2)^2\,|X_{j(-k),t}\}|t_1-t_1'| + (t_2-t_2')^2 \\ &\leqslant C''|X_{j(-k),t}^\top\gamma_{j(-k)}^0|(|\beta-\beta'|+|t_1-t_1'|)+(t_2-t_2')^2 \\ &\leqslant \sqrt{2}(C''|X_{j(-k),t}^\top\gamma_{j(-k)}^0|+2c_3)(|\beta-\beta'|+|t-t'|_2),\end{aligned}$$

where constant C'' depends on c_3 and C_1. Consequently, we have verified the last condition in (C3) by taking $\ell_2(X_{j(-k),t}) = \sqrt{2}(C''|X_{j(-k),t}^\top\gamma_{j(-k)}^0| + 2c_3)$ and $\upsilon = 1$. And given the finite moments conditions on X_t, we have $\mathsf{E}\{|\ell_1(X_{j(-k),t})|^4\} \leqslant L_{1n}$, $\mathsf{E}\{|\ell_2(X_{j(-k),t})|^4\} \leqslant L_{2n}$.

Verification of (C4): For any $\beta \in \mathcal{B}_{jk}$, there exists β' between β_{jk}^0 and β such that

$$\begin{aligned}\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,h_{jk}^0(X_{j(-k),t})\}] =& \partial_\beta\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}](\beta-\beta_{jk}^0) \\ &+\frac{1}{2}\partial_\beta^2\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta',h_{jk}^0(X_{j(-k),t})\}](\beta-\beta_{jk}^0)^2.\end{aligned}$$

Let $\phi_{jk} = \partial_\beta\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}] \geqslant c_1^2$. Since $\partial_\beta^2\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta',h_{jk}^0(X_{j(-k),t})\}] \leqslant C_1\,\mathsf{E}\,|X_{jk,t}^2 v_{jk,t}| \leqslant C_2$, we have

$$2|\,\mathsf{E}[\psi_{jk}\{Z_{j,t},\beta,h_{jk}^0(X_{j(-k),t})\}]| \geqslant 2\phi_{jk}|\beta-\beta_{jk}^0| - C_2(\beta-\beta_{jk}^0)^2 \geqslant \phi_{jk}|\beta-\beta_{jk}^0|,$$

whenever $|\beta-\beta_{jk}^0| \leqslant c_1^2/C_2$.

Verification of (C5): According to Corollary 5.4, with probability $1-o(1)$ we have

$$\|\widehat{\beta}_{j(-k)}^{[1]}-\beta_{j(-k)}^0\|_{j,pr} \lesssim \sqrt{s(\log a_n)/n}, \quad \|\widehat{\gamma}_{j(-k)}-\gamma_{j(-k)}^0\|_{j,pr} \lesssim \sqrt{s(\log a_n)/n},$$

which means the algorithms can provide an estimator of the nuisance function with good sparsity and rate properties given IC λ. Thus, by Lemma 7 in Belloni et al. (2015a), we have (C5) holds.

Verification of (C6): We refer to the proof of Theorem 1 in Belloni et al. (2015a).

Verification of (C7): Recall that $\psi_{jk,t}^0 = \{1/2 - \mathbf{1}(\varepsilon_{j,t} \leqslant 0)\}v_{jk,t}$. Hence, $\mathsf{E}(\frac{1}{\sqrt{n}}\sum_{t=1}^n \psi_{jk,t}^0)^2 = \sum_{\ell=-(n-1)}^{n-1}(1-|\ell|/n)\,\mathsf{E}(\psi_{jk,t}^0\psi_{jk,t-\ell}^0) \geqslant \frac{1}{4}\sum_{\ell=-(n-1)}^{n-1}(1-|\ell|/n)\,\mathsf{E}(v_{jk,t}v_{jk,t-\ell}) \geqslant c_1/4$.

Verification of (C8) and (C9): See Comment 5.8 where we discuss the admissible dimension rates either under the special case of VAR(1) with geometric decay rate (which gives bounded dependence adjusted norm) or more generally with finite dependence adjusted norm in polynomial rates.

Verification of (C9'): See Comment 5.9 and the discussion can be generalized to the case of finite dependence adjusted norm in polynomial rates easily.

□

Lemma B.11. *Let* $\psi_{jk,t}^0 \stackrel{\text{def}}{=} \psi_{jk}\{Z_{j,t},\beta_{jk}^0,h_{jk}^0(X_{j(-k),t})\}$, $T_n^{jk} \stackrel{\text{def}}{=} \sigma_{jk}^{-1}\phi_{jk}^{-1}\sum_{t=1}^n \psi_{jk,t}^0$, *and assume that* $\|\psi_{jk,\cdot}^0\|_{2,\varsigma} < \infty$. *Then*

$$\|T_n^{jk}\|_2 = \mathcal{O}(\sqrt{n}\|\psi_{jk,\cdot}^0\|_{2,\varsigma}), \text{ and } n^{-1/2}T_n^{jk} \xrightarrow{\mathcal{L}} \mathrm{N}(0,1)$$

Proof of Lemma B.11. Define the projector operator $\mathcal{P}_l(X_t) \stackrel{\text{def}}{=} \mathsf{E}(X_t|\mathcal{F}_l) - \mathsf{E}(X_t|\mathcal{F}_{l-1})$. Note that the projection operator is directly linked to the dependence adjusted norm for $X_{jk,t} = g_{jk}(\mathcal{F}_t) = g_{jk}(\ldots, \xi_{t-1}, \xi_t)$, and $\|\mathcal{P}_0(X_{jk,t})\|_2 \leqslant \|g_{jk}(\mathcal{F}_t) - g_{jk}(\mathcal{F}_t^*)\|_2 \leqslant 2\|\mathcal{P}_0(X_{jk,t})\|_2$ (by Theorem 1(i) in Wu, 2005).

Let $J_{l,n}^{jk} \stackrel{\text{def}}{=} \sigma_{jk}^{-1}\phi_{jk}^{-1}\sum_{t=1}^n \mathcal{P}_{t-l}(\psi_{jk,t}^0)$, and it is not hard to see that $T_n^{jk} = \sum_{l=0}^\infty J_{l,n}^{jk}$. As $\sigma_{jk}^{-1}\phi_{jk}^{-1}\mathcal{P}_{t-l}(\psi_{jk,t}^0)$'s form the martingale differences over t, according to Lemma B.1 we can apply the Burkholder Inequality and get $\|J_{l,n}^{jk}\|_2^2 \leqslant (\sigma_{jk}\phi_{jk})^{-2}\sum_{t=1}^n \|\mathcal{P}_{t-l}(\psi_{jk,t}^0)\|_2^2 \lesssim n(\delta_{j,k,l}^\psi)^2$, where $\delta_{j,k,l}^\psi \stackrel{\text{def}}{=} \|\psi_{jk,l}^0 - (\psi_{jk,l}^0)^*\|_2$. Thus, $\|T_n^{jk}\|_2 \lesssim \sqrt{n}\sum_{l=0}^\infty \delta_{j,k,l}^\psi \leqslant \sqrt{n}\|\psi_{jk,\cdot}^0\|_{2,\varsigma} = \mathcal{O}(\sqrt{n}\|\psi_{jk,\cdot}^0\|_{2,\varsigma})$. Then the conclusion that $n^{-1/2}T_n^{jk} \xrightarrow{\mathcal{L}} \mathrm{N}(0,1)$ follows from Lemma B.4 in light of the fact that $\mathsf{E}\,\psi_{jk,t}^0 = 0$ and $\|\psi_{jk,\cdot}^0\|_{2,\varsigma} < \infty$. □

Proof of Theorem 5.5. The proof follows directly from Lemma B.11. □

Proof of Corollary 5.6. We apply the high-dimensional central limit theorem (Theorem 3.2 in Zhang and Wu (2017a)) to the vector $\widetilde{\Im} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{t=1}^n \widetilde{\zeta}_t$ and $\widetilde{\mathcal{Z}} \stackrel{\text{def}}{=} \mathrm{vec}[\{(\mathcal{Z}_{jk})_{k=1}^K\}_{j=1}^J]$ is the corresponding standard Gaussian random vector, with the same correlation structure. Then we have $\rho(D^{-1}\widetilde{\Im}, D^{-1}\widetilde{\mathcal{Z}}) \to 0$, as $n \to \infty$, where D is a diagonal matrix with the square root of the diagonal elements of the long-run variance-covariance matrix of $\widetilde{\zeta}_t$, namely $\{\sum_{\ell=-\infty}^{\ell=\infty}\mathsf{E}(\zeta_{jk,t}\zeta_{jk,(t-\ell)})\}^{1/2}$, for $k = 1, \ldots, K$, $j = 1, \ldots, J$. The rest of the proof is similar to Corollary 5.2 and thus is omitted. □

Proof of Corollary 5.7. The proof is similar to that of Theorem 5.3 and Theorem A.3; therefore, we omit the detailed proof here. In particular, the following conditions on b_n are required:

$$
\begin{aligned}
&b_n = o\{n(\log|G|)^{-4}(\Phi_{q,\varsigma}^\zeta)^{-4} \wedge n(\log|G|)^{-5}(\Phi_{4,\varsigma}^\zeta)^{-4}c_n^{-2}\},\\
&F_\varsigma = o\{n^{q/2}(\log|G|)^{-q}|G|^{-1}(\Gamma_{q,\varsigma}^\zeta)^{-q}c_n^{-q/2}\}, \text{ with } c_n^{-1} = o(1).\\
&\Phi_{2,0}^\zeta\Phi_{2,\varsigma}^\zeta\{b_n^{-1} + \log(n/b_n)/n + (n-b_n)\log b_n/(nb_n)\}(\log|G|)^2 = o(1), \text{ if } \varsigma = 1;\\
&\Phi_{2,0}^\zeta\Phi_{2,\varsigma}^\zeta\{b_n^{-1} + n^{-\varsigma} + (n-b_n)b_n^{-\varsigma+1}/(nb_n)\}(\log|G|)^2 = o(1), \text{ if } \varsigma < 1;\\
&\Phi_{2,0}^\zeta\Phi_{2,\varsigma}^\zeta\{b_n^{-1} + n^{-1}b_n^{-\varsigma+1} + (n-b_n)/(nb_n)\}(\log|G|)^2 = o(1), \text{ if } \varsigma > 1. \qquad \text{(B.27)}
\end{aligned}
$$

where $F_\varsigma = n$, for $\varsigma > 1 - 2/q$; $F_\varsigma = l_n b_n^{q/2-\varsigma q/2}$, for $1/2 - 2/q < \varsigma < 1 - 2/q$; $F_\varsigma = l_n^{q/4-\varsigma q/2}b_n^{q/2-\varsigma q/2}$, for $\varsigma < 1/2 - 2/q$. □

C Supplementary Examples

C.1 Practical Examples of SRE

Example 4 (Identification Test for Large Structural Vector Autoregression Models). Denote $U_t = (U_{1,t}, U_{2,t}, \ldots, U_{M,t})^\top$. A large structural VAR can be represented in the following form (without loss of generality, consider only lag one):

$$\mathbf{A}U_t = \mathbf{B}U_{t-1} + \varepsilon_t,$$

where $\mathbf{A}$(invertible) and $\mathbf{B}$ are $M \times M$ matrices. The structural shocks ε_t satisfy $\mathsf{E}(\varepsilon_t) = 0$ and $\text{Var}(\varepsilon_t) = \mathbf{I}_M$. The corresponding reduced form is given by

$$U_t = \mathbf{D}U_{t-1} + \nu_t, \tag{C.1}$$

with $\mathbf{D} = \mathbf{A}^{-1}\mathbf{B}$ and $\nu_t = \mathbf{A}^{-1}\varepsilon_t$, where ν_t is denoted as the reduced form VAR shocks. Suppose ν_t spans the space of ε_t. The crucial question is the identification of $\mathbf{A}$. Typically, the covariance matrix of the reduced form shock ν_t is estimated with $M(M+1)/2$ restrictions, which are smaller than the M^2 restrictions needed to pin down ε_t. Adopting the identification approach proposed by Stock and Watson (2012), we may use external instruments that are correlated with the shock of interest and are uncorrelated with other shocks. Without loss of generality, suppose the structural shock of interest is $\varepsilon_{j,t}$. Then we can define $z_{j,t}$ as an external instrument for the jth structural shock satisfying

$$\begin{aligned} \mathsf{E}(\varepsilon_{j,t} z_{j,t}) &\neq 0, \\ \mathsf{E}(\varepsilon_{j',t} z_{j,t}) &= 0, \quad \text{for } j' \neq j. \end{aligned}$$

Thus, we propose to regress $z_{j,t}$ on ν_t:

$$z_{j,t} = \nu_t^\top \delta_j + e_{j,t}.$$

In practice, ν_t are replaced by the residuals obtained from a large VAR reduced form regression as in Example 3. The estimator of δ_j is denoted as $\widehat{\delta}_j$. It can be obtained by LASSO estimation, which give us a sparse estimator of the jth row of the matrix $\mathbf{A}^{-1}$ up to a scaling factor. Repeating this step for any j, one may formulate estimators for each row and perform simultaneous inference/hypothesis testing on the structural matrix $\mathbf{A}^{-1}$.

In summary, this is also a special case of SRE with

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (U_{j,t}, U_{-j,t-1}, \nu_t, \mathbf{D}_{j\cdot}^\top), \quad j = 1, \ldots, M,$$

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (z_{(j-M),t}, \nu_t, e_{(j-M),t}, \delta_{(j-M)}), \quad j = M+1, \ldots, 2M.$$

Example 5 (Cross-sectional Asset Pricing). Denote $Y_{j,t}$ as the excess return for asset j and period t. Asset pricing models explain the cross sectional variation in expected returns across assets; see e.g. Cochrane (2009). In particular, the variation of expected cross sectional returns is explained by the exposure to $K-1$ factors $X_{jk,t}$, $k = 1, \ldots, K-1$. One commonly used way to estimate an asset pricing model is to run a system of regression equations:

$$Y_{j,t} = \beta_{j0} + \sum_{k=1}^{K-1} \beta_{jk} X_{jk,t} + \varepsilon_{j,t}, \tag{C.2}$$

where $X_{jk,t}$'s are the factor returns (assumed to be excess returns of zero-cost portfolios). The selection of factors is a critical issue and the SRE framework addresses this issue, in particular when the number of factors K is large. See Feng et al. (2017) for a detailed model-selection exercise on picking asset pricing factors. The factor premiums are $\mathsf{E}(X_{jk,t})$ and the

pricing errors are β_{j0}. Usually, asset pricing imposes the restriction that all β_{j0}'s are zero. Our simultaneous inference framework naturally serves the purpose of simultaneously testing the zero pricing errors in a cross sectional regression setup. Namely, we are interested in testing $H_0 : \beta_{j0} = 0, \forall j = 1, \ldots, J$ versus $H_A : \exists\, j$ such that $\beta_{j0} \neq 0$. Our test procedure in Section 4.2 can be directly applied to achieve this goal.

Example 6 (**Network Formation and Spillover Effects**)**.** There is an emerging literature in economics concerning quantifying spillover effects and network formation. One leading example is as in Manresa (2013), which attempts to quantify social returns to research and development (R&D). Here, $U_{j,t}$ is taken to be the log output for firm j and time t. This output is loading on $D_{j,t}$ (capital stock for firm j and period t), and the aggregated spill-overs from the capital stock of other firms $\sum_{i\neq j} w_{ij} D_{i,t}$. The regression equation also controls for other covariates $X_{j,t}$ (e.g., log labor, log capital etc.):

$$U_{j,t} = \beta_j D_{j,t} + \sum_{i\neq j} \omega_{ij} D_{i,t} + \gamma_j^\top X_{j,t} + \varepsilon_{j,t}, \tag{C.3}$$

where ω_{ij} is referred to as the spillover effects of the R&D development of firm i on firm j. This again is contained in the SRE with

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (U_{j,t}, (D_{j,t}, D_{-j,t}^\top, X_{j,t}^\top)^\top, \varepsilon_{j,t}, (\beta_j, \omega_{(-j)j}^\top, \gamma_j^\top)^\top), \quad j = 1, \ldots, J.$$

Our simultaneous inference procedure (Section 4.2) can be applied to check the significance of the spillover effects for any set of parameters of interest. As an analogy, the presented framework displays a general class of network models, where $U_{j,t}$ is taken to be the nodal response, and $D_{i,t}$ are the nodal covariates. Global or local inference on the network parameters ω_{ij} is the subject of research. Section 7 is devoted to inference on the spillover effects of a textual sentiment index.

Comment C.1. Suppose there is unobserved heterogeneity in $U_{j,t}$, e.g. $U_{j,t} = \alpha_j + \sum_{i\neq j} w_{ij} D_{i,t} + \varepsilon_{j,t}$, where w_{ij} characterizes the spillover of individual i on j, and α_j is the individual fixed effect. For this situation consider the demeaned version to eliminate the individual specific effects and work with the new model: $\widetilde{U}_{j,t} = \sum_{i\neq j} w_{ij} \widetilde{D}_{i,t} + \widetilde{\varepsilon}_{j,t}$, where $\widetilde{U}_{j,t} = U_{j,t} - \frac{1}{n}\sum_{t=1}^n U_{j,t}$, $\widetilde{D}_{i,t} = D_{i,t} - \frac{1}{n}\sum_{t=1}^n D_{i,t}$, $\widetilde{\varepsilon}_{j,t} = \varepsilon_{j,t} - \frac{1}{n}\sum_{t=1}^n \varepsilon_{j,t}$, under the condition that $U_{j,t}$ has no feedback effects on $D_{i,t}$ (for example, $D_{i,t}$ should not be the lagged variable of $U_{j,t}$).

C.2 Examples of the Dependence Measure

1. **AR(1)**: Y_t follows $Y_t = aY_{t-1} + \varepsilon_t$, with $|a| < 1$, $\varepsilon_t \sim$ i.i.d.$(0, \sigma^2)$. Therefore, the MA representation is given by $Y_t = \sum_{l=0}^\infty a^l \varepsilon_{t-l}$ and $Y_t^* = \sum_{l=0}^\infty a^l \varepsilon_{t-l} + a^t \varepsilon_0^* - a^t \varepsilon_0$. $\|Y_t - Y_t^*\|_q = |a|^t \|\varepsilon_0 - \varepsilon_0^*\|_q$, $\Delta_{m,q} \lesssim |a|^m$, $\|Y_\cdot\|_{q,\varsigma} \lesssim \sup_{m\geqslant 0}(m+1)^\varsigma |a|^m < \infty$.

2. **ARCH(1)**: An ARCH (Autoregressive conditionally heteroscedastic) model is given by $Z_t = \sigma_t \varepsilon_t$, $\sigma_t^2 = w + \alpha^2 Z_{t-1}^2$, with $w > 0$, ε_t are i.i.d. shocks and $\mathrm{Var}(Z_t) = \sigma^2 < \infty$. Thus, it is not hard to see that $Z_t^2 = w\sum_{l=0}^\infty \alpha^{2l} \prod_{k=0}^l \varepsilon_{t-k}^2$. Rewrite the model as $Z_t = R(Z_{t-1}, \varepsilon_t) = \sqrt{(w + \alpha^2 Z_{t-1}^2)}\varepsilon_t$. According to Wu and Shao (2004), we have

the Lipschitz constant involved in the Lyapunov type condition ensuring the forward iteration contraction $\sup_{x\neq x'} \frac{|R(x,\varepsilon_0)-R(x',\varepsilon_0)|}{|x-x'|} \leqslant |\alpha\varepsilon_0|$. Let $\mu \stackrel{\text{def}}{=} \mathsf{E}\,|\alpha\varepsilon_0| < 1$ and assume $|\alpha\varepsilon_0| + |R(t_0,\varepsilon_0)|$ has finite qth moment. Then the process Z_t has stationary solutions. Moreover, $\|Z_t - Z_t^*\|_q \leqslant |\mu|^t \|\varepsilon_0 - \varepsilon_0^*\|_q$, and thus $\Delta_{m,q} \lesssim |\mu|^m$. Given $|\mu| < 1$, then we have $\|Z_\cdot\|_{q,\varsigma} \lesssim \sup_{m\geqslant 0}(m+1)^\varsigma |\mu|^m < \infty$.

3. **TAR** (Threshold autoregressive model): $Y_t = \theta_1 Y_{t-1}\mathbf{1}\{Y_{t-1} < \tau\} + \theta_2 Y_{t-1}\mathbf{1}\{Y_{t-1} \geqslant \tau\} + \varepsilon_t$, where θ_1 and θ_2 are two parameters and ε_t are i.i.d. shocks. If $\theta \stackrel{\text{def}}{=} \max\{|\theta_1|, |\theta_2|\} < 1$ and ε_t has a finite α-th order moment, then the TAR model admits a stationary solution with $\|Y_\cdot\|_{q,\varsigma} \lesssim \sup_{m\geqslant 0}(m+1)^\varsigma \theta^m < \infty$.

4. **VAR** (Vector autoregressive model): Without loss of generality we focus on VAR(1) given by $Y_t = AY_{t-1} + \varepsilon_t$, where $Y_t, \varepsilon_t \in \mathbb{R}^J$, and $\varepsilon_t \sim$ i.i.d. $\mathrm{N}(0,\Sigma)$. If the spectral radius of $A^\top A$, $\rho(A^\top A) < 1$, then $\lim_{m\to\infty} \|A\|^m \to 0$, where $\|\cdot\|$ denotes the spectral norm of a matrix. Rewrite the model as $Y_t = \sum_{l=0}^\infty A^l \varepsilon_{t-l}$. The existence of a stationary solution can be checked by Kolmogorov's three series theorem. For each equation j, $Y_{j,t} - Y_{j,t}^* = [A^t]_j(\varepsilon_0 - \varepsilon_0^*)$, where $[A^t]_j$ is the jth row of the matrix A^t. $(\mathsf{E}(|Y_{j,t} - Y_{j,t}^*|^q))^{1/q} \leqslant |[A^t]_j|_1 \| |\varepsilon_0 - \varepsilon_0^*|_\infty \|_q$. It follows that $(\mathsf{E}(|Y_{j,t} - Y_{j,t}^*|^q))^{1/q} \leqslant 2|[A^t]_j|_1 \mu_q$, where $\mu_q \stackrel{\text{def}}{=} \max_{1\leqslant j\leqslant J} \|\varepsilon_{j,0}\|_q$. Suppose $\max_{1\leqslant j\leqslant J} |[A^t]_j|_1 \leqslant |\alpha|^t$ $(|\alpha| < 1)$. Then we have $\max_{1\leqslant j\leqslant J} \|Y_{j,\cdot}\|_{q,\varsigma} \lesssim \mu_q$, $(\sum_{j=1}^J \|Y_{j,\cdot}\|_{q,\varsigma}^q)^{1/q} \lesssim J^{1/q}\mu_q$, and $\| |Y_{j,\cdot}|_\infty \|_{q,\varsigma} \lesssim (J)^{1/q}$ by union bounds.

5. **High-dimensional ARCH**: Consider $Y_t \in \mathbb{R}^J$, a high-dimensional ARCH(1) model follows for example the general specification from Bollerslev et al. (1988) and Hansen and Rahbek (1998): $Z_t = H_t^{1/2}\varepsilon_t$, and $\mathsf{E}(Z_t Z_t^\top | \mathcal{F}_{t-1}) = H_t$, with $\varepsilon_t \sim$ i.i.d. $\mathrm{N}(0, \mathbf{I}_J)$. The specification of the conditional covariance matrix $H_t = \Omega + AZ_{t-1}Z_{t-1}^\top A^\top$, where Ω is positive definite and A is a $J \times J$ matrix. Studying the stationarity condition of the process is not trivial. Define $h_t \stackrel{\text{def}}{=} \mathrm{vech}(H_t)$, the selection matrix D_J $(J^2 \times J(J+1)/2)$ gives $\mathrm{vec}(H_t) = D_J h_t$ and its generalized inverse matrix D_J^+ such that $D_J^+ D_J = \mathbf{I}_{J(J+1)/2}$. The vech notation of the iterations follows $h_t = \mathrm{vech}(\Omega) + D_J^+(A\otimes A)D_J \mathrm{vech}(Y_{t-1}Y_{t-1}^\top)$. Define $\tilde{A} \stackrel{\text{def}}{=} D_J^+(A\otimes A)D_J$, $w \stackrel{\text{def}}{=} \mathrm{vech}(\Omega)$. For simplicity, we look at the process h_t, with the state space representation $h_t = w + G(h_{t-1},\varepsilon_{t-1}) = F(h_{t-1},\varepsilon_{t-1}) = w + \tilde{A}\mathrm{vech}(\{\mathrm{vech}^{-1}(h_{t-1})\}^{1/2}\varepsilon_{t-1}\varepsilon_{t-1}^\top\{\mathrm{vech}^{-1}(h_{t-1})\}^{-1/2})$. The partial derivative matrix is $\Delta_t = \Delta(h_t,\varepsilon_t) = \partial h_{t+1}/\partial h_t^\top = \tilde{A}D_J^+(H_t^{1/2}\varepsilon_t\varepsilon_t^\top H_t^{-1/2} \otimes I_J)D_J$, and $\mathsf{E}\,\Delta_t = \tilde{A}$. Therefore, the spectral radius of $AA^\top$, $\rho(AA^\top) < 1$ ensures a stationary solution to the process h_t. Moreover, by solving the state space iteration recursively, we have $\mathsf{E}\,|h_t - h_t^*|_1 \leqslant 2\,\mathsf{E}\,|\mathcal{P}_0(h_t)|_1 \leqslant |\tilde{A}^t\{\mathrm{vech}(\Sigma) + w\} + \tilde{A}^{t+1}\mathrm{vech}(\Sigma)|_1 \lesssim \{\mathrm{tr}(AA^\top)\}^t$, where the projector operator $\mathcal{P}_l(h_t) \stackrel{\text{def}}{=} \mathsf{E}(h_t|\mathcal{F}_l) - \mathsf{E}(h_t|\mathcal{F}_{l-1})$ and $\Sigma = \mathsf{E}\,H_t = \sum_{i=0}^\infty A^i\Omega(A^i)^\top$. Assume that $\{\mathrm{tr}(AA^\top)\}^t < |c|^t$, with $|c| < 1$, we have $\sum_{j=1}^{J(J+1)/2} \|h_{j,\cdot}\|_{1,\varsigma} \lesssim J(J+1)/2$.

   According to Hafner and Preminger (2009), the iteration formulae are given by $h_t = \varpi(\bar{h}_{t-1}^\star,\varepsilon_{t-1}) + \sum_{l=1}^{m-1}\Pi_{k=1}^l \Delta(\bar{h}_{t-k}^\star,\varepsilon_{t-k})\varpi(\bar{h}_{t-l-1}^\star,\varepsilon_{t-l-1}) + \Pi_{k=1}^m\Delta(\bar{h}_{t-k}^\star,\varepsilon_{t-k})h_{t-m}$, where $\varpi(h,\varepsilon) = w + G(h^\star,\varepsilon) - \Delta(h,\varepsilon)h^\star$, $h^\star$ is the contraction state, and $\bar{h}_{t-k}^\star$'s lie on the line segment between $h^\star$ and h_{t-k}. For ease of derivation, we assume a strong assumption such that $\mathsf{E}\sup_{h_m} \|\Delta(h_m,\varepsilon_m)\|^q < s < 1$ for all $m \geqslant 1$ and $q \geqslant 2$, where $\|\cdot\|$ denotes

the spectral norm of a matrix. Let $h^m = \{(h_1^\top, \ldots, h_m^\top)^\top : |h_t|_2 = 1, t = 1, \ldots, m\}$, it follows $\mathsf{E} \sup_{h^m} \|\Pi_{k=1}^m \Delta(h_{m-k+1}, \varepsilon_{m-k+1})\|^q \leqslant \Pi_{k=1}^m \mathsf{E} \sup_{h_{m-k+1}} \|\Delta(h_{m-k+1}, \varepsilon_{m-k+1})\|^q \leqslant s^m$. Hence, $\max_{1 \leqslant j \leqslant J(J+1)/2} \|h_{j,\cdot}\|_{q,\varsigma} \leqslant C$, $\| |h_\cdot|_\infty \|_{q,\varsigma} \lesssim \| |h_t|_\infty \|_q \lesssim \{J(J+1)/2\}^{1/q}$, and $(\sum_{j=1}^{J(J+1)/2} \|h_{j,\cdot}\|_{q,\varsigma}^q)^{1/q} \lesssim \{J(J+1)/2\}^{1/q}$.

D Additional Details for Empirical Analysis

Consumer Discretionary (11)		**Financials (8)**		GD	General Dynamics Corporation
AMZN	Amazon.com, Inc.	AIG	American International Group, Inc.	GE	General Electric Company
BBY	Best Buy Co. Inc.	AMT	American Tower Corporation (REIT)	HON	Honeywell International Inc.
CBS	CBS Corporation	AXP	American Express Company	LMT	Lockheed Martin Corporation
CMCSA	Comcast Corporation	BAC	Bank of America Corporation	LUV	Southwest Airlines Company
CMG	Chipotle Mexican Grill, Inc.	C	Citigroup Inc.	**Information Technology (11)**	
DIS	Walt Disney Company (The)	ETFC	E*TRADE Financial Corporation	AAPL	Apple Inc.
F	Ford Motor Company	GS	Genpact Limited	ACN	Accenture plc
GM	General Motors Company	JPM	J P Morgan Chase & Co	ADP	Automatic Data Processing, Inc.
GPS	Gap, Inc. (The)	**Health Care (8)**		CSCO	Cisco Systems, Inc.
HD	Home Depot, Inc. (The)	AET	Aetna Inc.	EA	Electronic Arts Inc.
LEN	Lennar Corporation	AMGN	Amgen Inc.	EBAY	eBay Inc.
Consumer Staples (4)		BIIB	Biogen Inc.	EMC	EMC Corporation
COST	Costco Wholesale Corporation	BMY	Bristol-Myers Squibb Company	FSLR	First Solar, Inc.
CVS	CVS Health Corporation	CELG	Celgene Corporation	HPQ	HP Inc.
KO	Coca-Cola Company (The)	GILD	Gilead Sciences, Inc.	IBM	International Business Machines Corporation
KR	Kroger Company (The)	JNJ	Johnson& Johnson	INTC	Intel Corporation
Energy (6)		LLY	Eli Lilly and Company	**Materials (3)**	
APC	Anadarko Petroleum Corporation	**Industrials (10)**		AA	Alcoa Corporation
BHI	Black Hills Corp.	BA	Boeing Company (The)	DD	EI du Pont de Nemours & Co
CHK	Chesapeake Energy Corporation	CAT	Caterpillar, Inc.	DOW	Dow Chemical
COP	ConocoPhillips	DAL	Delta Air Lines, Inc.	**Utilities (2)**	
CVX	Chevron Corporation	DHR	Danaher Corporation	DUK	Duke Energy Corp.
HAL	Halliburton Company	FDX	FedEx Corporation	EXC	Exelon Corporation

Table D.1: The list of the stock symbols and the corresponding company names grouped by industries.

References

Andrews, D. W. (1984). Non-strong mixing autoregressive processes, *Journal of Applied Probability* **21**(4): 930–934.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards, *The Journal of Finance* **59**(3): 1259–1294.

Audrino, F. and Tetereva, A. (2017). Sentiment spillover effects for US and European companies, *SSRN preprint SSRN:2957581* .

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns, *The Journal of Finance* **61**(4): 1645–1680.

Belloni, A., Chen, M. and Chernozhukov, V. (2016). Quantile graphical models: Prediction and conditional independence with applications to financial risk management, *arXiv preprint arXiv:1607.00286* .

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models, *Bernoulli* **19**(2): 521–547.

Belloni, A., Chernozhukov, V. and Hansen, C. (2011). Inference for high-dimensional sparse econometric models, *arXiv preprint arXiv:1201.0220* .

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls, *The Review of Economic Studies* **81**(2): 608–650.

Belloni, A., Chernozhukov, V. and Kato, K. (2015a). Supplement material for "Uniform post selection inference for least absolute deviation regression and other Z-estimation problems", Available at *Biometrika* online.

Belloni, A., Chernozhukov, V. and Kato, K. (2015b). Uniform post selection inference for least absolute deviation regression and other Z-estimation problems, *Biometrika* **102**(1): 77–94.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector, *The Annals of Statistics* **37**(4): 1705–1732.

Bollerslev, T., Engle, R. F. and Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances, *Journal of political Economy* **96**(1): 116–131.

Burkholder, D. L. (1988). Sharp inequalities for martingales and stochastic integrals, *Astérisque* (157-158): 75–94.

Chen, C. Y.-H., Härdle, W. K. and Okhrin, Y. (2019). Tail event driven networks of SIFIs, *Journal of Econometrics* **208**(1): 282–298.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2016). Double machine learning for treatment and causal parameters, *arXiv preprint arXiv:1608.00060* .

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors, *The Annals of Statistics* **41**(6): 2786–2819.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013b). Testing many moment inequalities, *arXiv preprint arXiv:1312.7614* .

Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors, *Probability Theory and Related Fields* **162**(1-2): 47–70.

Chernozhukov, V., Chetverikov, D., Kato, K. et al. (2014). Gaussian approximation of suprema of empirical processes, *The Annals of Statistics* **42**(4): 1564–1597.

Chernozhukov, V. and Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach, *Journal of Econometrics* **142**(1): 379–398.

Cochrane, J. H. (2009). *Asset Pricing: (Revised Edition)*, Princeton university press.

Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap, *Test* **26**(4): 685–719.

Dimitrakopoulou, K., Tsimpouris, C., Papadopoulos, G., Pommerenke, C., Wilk, E., Sgarbas, K. N., Schughart, K. and Bezerianos, A. (2011). Dynamic gene network reconstruction from gene expression data in mice after influenza a (h1n1) infection, *Journal of clinical bioinformatics* **1**(1): 27.

El Machkouri, M., Volný, D. and Wu, W. B. (2013). A central limit theorem for stationary random fields, *Stochastic Processes and their Applications* **123**(1): 1–14.

Epskamp, S., Waldorp, L. J., Mõttus, R. and Borsboom, D. (2016). Discovering psychological dynamics: The gaussian graphical model in cross-sectional and time-series data, *arXiv preprint arXiv* **1609**.

Epskamp, S., Waldorp, L. J., Mõttus, R. and Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data, *Multivariate Behavioral Research* **53**(4): 453–480.

Feng, G., Giglio, S. and Xiu, D. (2017). Taming the factor zoo, *Chicago booth research paper no. 17-04*, The University of Chicago Booth School of Business.

Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data, *The Journal of Business* **53**(1): 67–78.

Hafner, C. M. and Preminger, A. (2009). On asymptotic theory for multivariate garch models, *Journal of Multivariate Analysis* **100**(9): 2044–2054.

Hansen, E. and Rahbek, A. (1998). Stationarity and asymptotics of multivariate ARCH time series with an application to robustness of cointegration analysis, *Preprint. University of Copenhagen* .

Härdle, W. K. Chen, S., Liang, C. and Schienle, M. (2018). Time-varying limit order book networks, *IRTG 1792 Discussion Paper 2018-016*, IRTG 1792, Humboldt Universität zu Berlin, Germany.

Härdle, W. K., Wang, W. and Yu, L. (2016). TENET: Tail-Event driven NETwork risk, *Journal of Econometrics* **192**(2): 499–513.

Hautsch, N., Schaumburg, J. and Schienle, M. (2015). Financial network systemic risk contributions, *Review of Finance* **19**(2): 685–738.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews, *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177.

Huang, D., Yin, J., Shi, T. and Wang, H. (2016). A statistical model for social network labeling, *Journal of Business & Economic Statistics* **34**(3): 368–374.

Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory, *IEEE Transactions on Information Theory* **60**(10): 6522–6554.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions, *Journal of Econometrics* **186**(2): 325–344.

Kolaczyk, E. D. and Csárdi, G. (2014). *Statistical analysis of network data with R*, Vol. 65, Springer.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.*, Springer.

Lahiri, S. N. et al. (1999). Theoretical comparisons of block bootstrap methods, *The Annals of Statistics* **27**(1): 386–404.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance* **66**(1): 35–65.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*, Springer Science & Business Media.

Manresa, E. (2013). Estimating the structure of social interactions using panel data, *Unpublished manuscript*, CEMFI, Madrid.

Meinshausen, N., Bühlmann, P. et al. (2006). High-dimensional graphs and variable selection with the lasso, *The annals of statistics* **34**(3): 1436–1462.

Neykov, M., Ning, Y., Liu, J. S. and Liu, H. (2015). A unified theory of confidence regions and testing for high dimensional estimating equations, *arXiv preprint arXiv:1510.08986* .

Opgen-Rhein, R. and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, *BMC systems biology* **1**(1): 37.

Pesaran, M. H. and Yamagata, T. (2017). Testing for alpha in linear factor pricing models with a large number of securities, *USC-INET Research Paper No. 17-13*, USC Dornsife Institute for New Economic Thinking.

Ramirez, R. N., El-Ali, N. C., Mager, M. A., Wyman, D., Conesa, A. and Mortazavi, A. (2017). Dynamic gene regulatory networks of human myeloid differentiation, *Cell systems* **4**(4): 416–429.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing, *Journal of the American Statistical Association* **100**(469): 94–108.

Rudelson, M. and Zhou, S. (2012). Reconstruction from anisotropic random measurements, *Proceedings of the 25th Annual Conference on Learning Theory*, Vol. 23, pp. 10.1–10.28.

Stock, J. H. and Watson, M. W. (2012). Disentangling the channels of the 2007-2009 recession, *Brookings panel on economic activity*, The Brookings Institution.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market, *The Journal of Finance* **62**(3): 1139–1168.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models, *The Annals of Statistics* **42**(3): 1166–1202.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence, *Weak convergence and empirical processes*, Springer, pp. 16–28.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, National Acad Sciences, pp. 14150–14154.

Wu, W. B. and Shao, X. (2004). Limit theorems for iterated random functions, *Journal of Applied Probability* **41**(2): 425–436.

Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors, *Electronic Journal of Statistics* **10**(1): 352–379.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model, *Biometrika* **94**(1): 19–35.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1): 217–242.

Zhang, D. and Wu, W. B. (2017a). Gaussian approximation for high dimensional time series, *The Annals of Statistics* **45**(5): 1895–1919.

Zhang, D. and Wu, W. B. (2017b). Supplement material for "Gaussian approximation for high dimensional time series", Available at the *Annals of Statistics* online, DOI: 10.1214/16-AOS1512SUPP.

Zhang, J. L., Härdle, W. K., Chen, C. Y. and Bommes, E. (2016). Distillation of news flow into analysis of stock reactions, *Journal of Business & Economic Statistics* **34**(4): 547–563.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models, *Journal of the American Statistical Association* **112**(518): 757–768.

Zhu, X., Pan, R., Li, G., Liu, Y., Wang, H. et al. (2017). Network vector autoregression, *The Annals of Statistics* **45**(3): 1096–1123.

Zhu, X., Wang, W., Wang, H. and Härdle, W. K. (2019). Network quantile autoregression, *Journal of Econometrics* **In Press**.

Zhu, Y. and Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models, *Journal of the American Statistical Association* **113**(524): 1583–1600.