*The Annals of Statistics*
2021, Vol. 49, No. 3, 1702–1735
https://doi.org/10.1214/20-AOS2019


# LASSO-DRIVEN INFERENCE IN TIME AND SPACE

BY VICTOR CHERNOZHUKOV[1], WOLFGANG KARL HÄRDLE[2], CHEN HUANG[3] AND WEINING WANG[4]

[1]*Department of Economics and Operations Research Center, Massachusetts Institute of Technology, vchern@mit.edu*
[2]*IRTG1792, Humboldt–Universität zu Berlin, haerdle@hu-berlin.de*
[3]*Department of Economics and Business Economics and CREATES, Aarhus University, chen.huang@econ.au.dk*
[4]*Department of Economics and Related Studies, University of York, weining.wang@york.ac.uk*

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 temporal 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.

**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, that is, regressing every factor in the system on a large subset of other factors. Examples include analysis of financial systemic risk by quantile predictive graphical models with LASSO (Hautsch, Schaumburg and Schienle (2015), Härdle, Wang and Yu (2016), Belloni, Chen and Chernozhukov (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. (2018)). 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). 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 models, one can attain valuable pre-information on the core structure by running a set of simple regressions with LASSO shrinkage.

Received July 2018; revised May 2020.
*MSC2020 subject classifications.* Primary 62M10, 62J99; secondary 62F40.
*Key words and phrases.* LASSO, time series, simultaneous inference, system of equations, $Z$-estimation, Bahadur representation, martingale decomposition.

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 $\ell_1$-norm estimators are developed by Bickel, Ritov and Tsybakov (2009). There are many related articles on deriving near-oracle bounds using the $\ell_1$-norm penalization function for the i.i.d. case, such as Belloni and Chernozhukov (2013), Belloni, Chernozhukov and Hansen (2011). There are also many extensions to the LASSO estimation with dependent data. For example, Basu and Michailidis (2015) study the consistency of the estimator in sparse high-dimensional Gaussian time series models; Kock and Callot (2015) consider the high-dimensional near-oracle inequalities in large vector autoregressive (VAR) models; Lin and Michailidis (2017) look at the regularized estimation and testing for high-dimensional multi-block VAR models. However, the majority of the literature imposes a Gaussian or 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, Chetverikov and Kato (2013) 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, Chetverikov

and Kato (2019) deliver results for the case of $\beta$-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, Chernozhukov and Hansen (2011, 2014), Zhang and Zhang (2014), Javanmard and Montanari (2014), van de Geer et al. (2014), Neykov et al. (2018), Chernozhukov et al. (2018), 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, Chernozhukov and Kato (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, Bühlmann and Zhang (2017), Zhang and Cheng (2017) for the case of mean regression. And more recently, Krampe, Kreiss and Paparoditis (2018) extend the approaches to test large groups of coefficients in sparse VAR models. We pick up the line of the inference analysis of Belloni, Chernozhukov and Kato (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. This allows us to embed the high-dimensional VAR model as a special case. Our 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, Chernozhukov and Kato (2015b) applies maximal inequalities derived in Chernozhukov, Chetverikov and Kato (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 and Bühlmann (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, Härdle and Okhrin (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 $|v|_\infty \stackrel{\text{def}}{=} \max_{1\le j\le p} |v_j|$ and $|v|_s \stackrel{\text{def}}{=} (\sum_{j=1}^p |v_j|^s)^{1/s}$, $s \ge 1$. 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 $a_n$ and $b_n$, write $a_n \lesssim b_n$ if there exists constant $C > 0$ (does not depend on $n$) such that $a_n/b_n \le C$. For a sequence of random variables $x_n$, we use the notation $x_n \lesssim_{\mathrm{P}} b_n$ to denote $x_n = \mathcal{O}_{\mathrm{P}}(b_n)$. 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 $\epsilon$-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 $\mathrm{ent}(\epsilon, \mathcal{F}) = \log\sup_{\mathcal{Q}}\mathcal{N}(\epsilon\|\bar{F}\|_{\mathcal{Q},q}, \mathcal{F}, \|\cdot\|_{\mathcal{Q},q})$ 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 $\mathsf{E}^*$ to $\mathsf{E}$ and outer probability $\mathrm{P}^*$ 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 Sections 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 Material (Chernozhukov et al. (2021)). 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):

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

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 $j = 1,\dots,J$. 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 $U_{m,t}$:

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

with auxiliary regressions to model predictive relations between covariates:

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

where $X_{-k,t} = (X_{\ell,t})_{\ell\neq k} \in \mathbb{R}^{K-1}$, and $\delta_k^0$ 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

$$\begin{aligned}
&(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (U_{j,t}, X_t, \varepsilon_{j,t}, \gamma_j^0), \quad j = 1,\dots,M,\\
&(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (X_{(j-M),t}, X_{-(j-M),t}, \nu_{(j-M),t}, \delta_{(j-M)}^0),\\
&\quad j = M+1,\dots,J = M+K.
\end{aligned}$$

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, Chernozhukov and Kato (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:

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

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

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

with $\mathbf{B} = \mathbf{D}^{-1}\mathbf{\Gamma}$ and $v_t = \mathbf{D}^{-1}\varepsilon_t$. 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

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

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,

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

where $U_t = (U_{1,t}, U_{2,t}, \ldots, U_{M,t})^\top$, and $\varepsilon_t$ is an $M$-dimensional white noise or innovation process; see, for example, 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, Dimitrakopoulou et al. (2011), Opgen-Rhein and Strimmer (2007), Ramirez et al. (2017). 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 Material 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 $\beta_j^0$ throughout the paper:

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

where the $\ell_0$-norm, $|\cdot|_0$, is the number of nonzero components of a vector.

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 $\beta_{jk}^0$ is not sparse, we shall define an intermediary optimal value for our true coefficients, that is, $\beta_{jk}^*$. Let $LC_p \stackrel{\text{def}}{=} \min_{|\beta_j|_0 \le p} [\mathsf{E}_n\{X_{j,t}^\top(\beta_j - \beta_j^0)\}^2]^{1/2}$, additionally with proper conditions on the design matrix, the optimal sparsity level is given by $s_j^* = \min_{0 \le p \le (K \wedge n)} LC_p^2 + (\max_{1 \le k \le K} \Psi_{jk}^2) p/n$, where $\Psi_{jk}^2$ is the long run variance of $\frac{1}{\sqrt{n}} \sum_{t=1}^n \varepsilon_{j,t} X_{jk,t}$. Then the oracle $\beta_{jk}^*$ is defined to be $\arg\min_{|\beta_j|_0 \le s_j^*} \mathsf{E}_n\{X_{j,t}^\top(\beta_j - \beta_j^0)\}^2$. Thus an additional term involving $LC_{s_j^*}$ will appear in the bound in case of the true signal $\beta_{jk}^0$ 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 $|\beta_{jk}^0| \le Ak^{-\gamma}$ (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 $\ell_1$-penalized estimator of $\beta_j^0$ of the form:

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

where $\lambda$ is the joint "optimal" penalty level and $\Psi_{jk}$'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:

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

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

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

To achieve good performance bounds, we first consider "ideal" choices (IC) 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

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

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

where $c > 1$, for example, $c = 1.1$, and $1 - \alpha$ is a confidence level, for example, $\alpha = 0.1$, 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:

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

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, Chetverikov and Kato (2013). 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 Sections 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 non-overlapping block bootstrap (NBB). Consider the system of equations with dependent data:

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

S1 Run the initial $\ell_1$-penalized regression equation by equation, that is, for the $j$th 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 $\lambda_j$ are the penalty levels and $\Psi_{jk}$ 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 $\lambda_j$, where $\Phi(\cdot)$ denotes the cdf of $\mathrm{N}(0,1)$, $\alpha' = 0.1$, $c' = 0.5$, and choose $\sqrt{\mathrm{lvar}(X_{jk,t}\breve{\varepsilon}_{j,t})}$ for the penalty loadings, where $\breve{\varepsilon}_{j,t}$ are preliminary estimated errors and $\mathrm{lvar}(X_{jk,t}\breve{\varepsilon}_{j,t})$ is an estimate of the long-run variance $\sum_{\ell=-\infty}^{\infty} \mathsf{E}(X_{jk,t}\breve{\varepsilon}_{j,t} X_{jk,(t-\ell)}\breve{\varepsilon}_{j,(t-\ell)})$, for example, 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 $k(z) = (1-|z|)\mathbf{1}(|z|\le 1)$. 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 $\max_{1\le j\le J, 1\le k\le K} |Z^{[B]}_{jk}/\Psi_{jk}|$, 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. 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 (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 de-sparsification 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 $\beta^0_{jk}$. 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 $Y_{j,t} = X_{jk,t}\beta^0_{jk} + X^\top_{j(-k),t}\beta^0_{j(-k)} + \varepsilon_{j,t}$, run (post) LS LASSO procedure (for each $j$), and keep the quantity $X^\top_{j(-k),t}\widehat{\beta}^{[1]}_{j(-k)}$ for each $k$.

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

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

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, Chernozhukov and Kato (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 Material 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 Corollary 5.6.

COMMENT 4.1. Our algorithms follow patterns discussed in Belloni, Chernozhukov and Kato (2015b, 2015a) 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, Chernozhukov and Hansen (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, Chernozhukov and Kato (2015b, 2015a), one can alternatively implement an algorithm via double selection as in Belloni, Chernozhukov and Hansen (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″:

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

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

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

As shown in Belloni, Chernozhukov and Hansen (2011) and Belloni, Chernozhukov and Kato (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, for example, $\widehat{v}_{jk,t}$. Therefore, the inference on the parameters may still be applied as in the framework of Algorithms 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 $\psi_{jk}(Z_{j,t}, \beta_{jk}, h_{jk}) = \{1/2 - \mathbf{1}(Y_{j,t} \le X_{jk,t}\beta_{jk} + X_{j(-k),t}^{\top}\beta_{j(-k)})\}v_{jk,t}$, define $\omega_{jk} \overset{\text{def}}{=} \mathsf{E}\{(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_{jk,t}^{0})^2\} = \sum_{\ell=-(n-1)}^{n-1}(1 - \frac{|\ell|}{n})\operatorname{cov}(\psi_{jk,t}^{0}, \psi_{jk,(t-\ell)}^{0})$ with $\psi_{jk,t}^{0} \overset{\text{def}}{=} \psi_{jk}(Z_{j,t}, \beta_{jk}^{0}, h_{jk}^{0})$, and $\phi_{jk} \overset{\text{def}}{=} \frac{\partial\,\mathsf{E}\{\psi_{jk}(Z_{j,t},\beta,h_{jk}^{0})\}}{\partial\beta}|_{\beta=\beta_{jk}^{0}}$.

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

$$\max_{(j,k),(j',k')}\left|\sum_{i=1}^{l_n}\widehat{\eta}_{j'k',i}\widehat{\eta}_{jk,i} - \sum_{i=1}^{l_n}\eta_{j'k',i}\eta_{jk,i}\right| = o_{\mathrm{P}}\big(\{\log(JK)\}^{-2}\big),$$

with $\eta_{jk,i} \overset{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{l=(i-1)b_n+1}^{ib_n}\zeta_{jk,l}$ and $\widehat{\eta}_{jk,i} \overset{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{l=(i-1)b_n+1}^{ib_n}\widehat{\zeta}_{jk,l}$, $e_{j,i}$ are independently drawn from $\mathrm{N}(0,1)$, $l_n$ and $b_n$ are the numbers of blocks and block size, respectively. More discussion on how one can construct the consistent pre-estimators $\widehat{\zeta}_{jk,t}$ is stated in the Supplementary Material; see Comment B.5.

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

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

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 $\beta_{jk}^0$ 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, for example, 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 $i = 1, \ldots, l_n$,

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

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: $\mathbf{H}_0 : \beta_{jk}^0 = 0$, $\forall (j,k) \in G$, and the alternative $\mathbf{H}_A : \beta_{jk}^0 \neq 0$, 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)

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

One follows the algorithms to obtain $\widehat{\beta}_{jl} (1 \leq l \leq p_j)$ 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

$$\left\{ \beta \in \mathbb{R}^{|G|} : \max_{(j,k)\in G} T_{jk} \leq q_G^*(1-\alpha) \text{ and } \min_{(j,k)\in G} T_{jk} \geq -q_G^*(1-\alpha) \right\}, \tag{4.8}$$

and for each component $(j,k) \in G$, the confidence interval $\widetilde{\mathrm{CI}}_{jk}^*(\alpha)$ is given by $[\widehat{\beta}_{jk}^{[2]} - \widehat{\sigma}_{jk} n^{-1/2} q_G^*(1-\alpha), \widehat{\beta}_{jk}^{[2]} + \widehat{\sigma}_{jk} n^{-1/2} q_G^*(1-\alpha)]$. We show in Corollary 5.8 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 $S_{jk} \stackrel{\text{def}}{=} n^{-1/2}\sum_{t=1}^{n}\varepsilon_{j,t}X_{jk,t}$, $S_{j\cdot} = (S_{jk})_{k=1}^{K}$, and $\Psi_{jk} \stackrel{\text{def}}{=} \sqrt{\text{avar}(S_{jk})}$, which is the square root of the long-run variance of $X_{jk,t}\varepsilon_{j,t}$, namely $\{\sum_{\ell=-\infty}^{\infty}\mathsf{E}(X_{jk,t}X_{jk,(t-\ell)}\varepsilon_{j,t}\varepsilon_{j,(t-\ell)})\}^{1/2}$. Recall that for a single equation LASSO, we select the penalty in the following ways:

(a) theoretically, for each regression, $\lambda_j$ is $\lambda_j^0(1-\alpha)$ (IC), that is, the $(1-\alpha)$ quantile of $2c\sqrt{n}\max_{1\le k\le K}|S_{jk}/\Psi_{jk}|$ (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 $2c\max_{1\le k\le K}\sqrt{n}|Z_{jk}/\Psi_{jk}|$, 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 $2c\sqrt{n}\max_{1\le k\le K}|Z_{jk}^{[B]}/\widehat{\Psi}_{jk}|$ ($Z_{jk}^{[B]}$'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 $j = 1,\dots,J, k = 1,\dots,K$, let $X_{jk,t}$ and $\varepsilon_{j,t}$ be stationary processes admitting the following representation forms $X_{jk,t} = g_{jk}(\mathcal{F}_t) = g_{jk}(\dots,\xi_{t-1},\xi_t)$ and $\varepsilon_{j,t} = h_j(\mathcal{F}_t) = h_j(\dots,\eta_{t-1},\eta_t)$, where $\xi_t,\eta_t$ are i.i.d. random elements (innovations or shocks, allowing for overlap; see Comment 5.1) across $t$, $\mathcal{F}_t = (\dots,\xi_{t-1},\eta_{t-1},\xi_t,\eta_t)$, $g_{jk}(\cdot)$ and $h_j(\cdot)$ are measurable functions (filters). $\mathsf{E}(X_{jk,t}\varepsilon_{j,t}) = 0$, for any $j,k \in 1,\dots,J,1,\dots,K$.

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

It should be noted that (A1) admits a wide class of processes. The largest value of $\varsigma$ 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 Material.

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

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

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

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

COMMENT 5.1. We allow for overlap in the elements in $\xi_t$ and $\eta_t$, 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 $\varepsilon_{t-l}$ 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 $k$th row of the matrix $A^l$, $k = 1, \ldots, J$. In this case no serial correlation in the innovations $\varepsilon_t$'s would be sufficient for $\mathsf{E}(X_{jk,t}\varepsilon_{j,t}) = 0$.

COMMENT 5.2. We show in Theorem B.2 (see the Supplementary Material) 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* $\lambda_j \geq 2c\sqrt{n}\max_{1\leq k\leq K}|S_{jk}/\Psi_{jk}|$ *and another event which RE holds*, *then with probability* $1 - o(1)$, $\tilde{\beta}_j$ *obtained from* (3.6) *satisfy*

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

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

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

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 $2c\sqrt{n}\max_{1\leq k\leq K}|S_{jk}/\Psi_{jk}|$ 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*

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

*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* $\varsigma$.

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 $2c\sqrt{n}\max_{1\leq k\leq K}|S_{jk}/\Psi_{jk}|$. Here we discuss the rate under some special cases.

1. *VAR(1)* (Example 3, continued): 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.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 $\|X_{jk,t}\varepsilon_{j,t} - X^*_{jk,t}\varepsilon^*_{j,t}\|_q = \|Y_{j,t-1}\varepsilon_{j,t} - Y^*_{j,t-1}\varepsilon_{j,t}\|_q = \|[A^{t-1}]_j(\varepsilon_0 - \varepsilon_0^*)\varepsilon_{j,t}\|_q \le 2|[A^{t-1}]_j|_1\mu_q^2$, where $\mu_q \stackrel{\text{def}}{=} \max_j \|\varepsilon_{j,t}\|_q$ and $[A^{t-1}]_j$ is the $j$th row of the matrix $A^{t-1}$. Assume $\max_j |[A^t]_j|_1 \le |c|^t$ with $|c| < 1$ (a geometric decay rate). It follows that $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} = \frac{2\mu_q^2}{1-|c|}\sup_{m\ge 0}(m+1)^{\varsigma}\sum_{t=m}^{\infty}|c|^{t-1} \le (C/|c|) \vee \{C(m^*+1)|c|^{m^*-1}\}$, where $m^* = (-\varsigma/\log|c| - 1) \vee 0$ and $C > 0$ depends on $\mu_q$. 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 $\rho$ is the network parameter suggesting the strength of the network effects. In that case, assuming a geometric decay rate $\max_j |[A^t]_j|_1 \le |c|^t$ with $|c| < 1$ again gives similar results.

2. *Spatial autoregressive structure in* $\varepsilon_t$: 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, $\eta_t$ are i.i.d. and have finite $q$th 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 $\|X_{jk,t}\varepsilon_{j,t} - X^*_{jk,t}\varepsilon^*_{j,t}\|_q \le \|(X_{jk,t} - X^*_{jk,t})\varepsilon_{j,t}\|_q + \|X_{jk,t}(\varepsilon_{j,t} - \varepsilon^*_{j,t})\|_q \le \|X_{jk,t} - X^*_{jk,t}\|_q\|\varepsilon_{j,t}\|_q + \|X_{jk,t}\|_q\|[\rho^t W^t]_j(\eta_0 - \eta_0^*)\|_q \le |[(\mathbf{I} - \rho W)^{-1}]_j|_1\mu_q^{\eta}\|X_{jk,t} - X^*_{jk,t}\|_q + 2|[\rho^t W^t]_j|_1\mu_q^{\eta}\|X_{jk,t}\|_q$. Assume $\max_j |[\rho^t W^t]_j|_1 \le |c|^t$ with $|c| < 1$. It follows that $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \le C_1\|X_{jk,\cdot}\|_{q,\varsigma} + C_2\sup_{m\ge 0}(m+1)^{\varsigma}\sum_{t=m}^{\infty}|c|^t \le C_1\|X_{jk,\cdot}\|_{q,\varsigma} + C_3(m^*+1)|c|^{m^*-1}$, where $m^* = (-\varsigma/\log|c| - 1) \vee 0$ and $C_1, C_2, C_3 > 0$ depend on $\mu_q^{\eta}$ 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 $\eta_t$ are i.i.d. and have finite $q$th 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 $l \ge 1$ there exists only spatial dependence. Let $a_{jk}^t \stackrel{\text{def}}{=} [A^t]_{jk}$ be the element on the $j$th row and $k$th 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 $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \le C_1\|X_{jk,\cdot}\|_{q,\varsigma} + C_2\sup_{m\ge 0}(m+1)^{\varsigma}\sum_{t=m}^{\infty}\sum_k |a_{jk}^t|$, where $C_1, C_2 > 0$ depend on $\mu_q^{\eta}$ and $\|X_{jk,t}\|_q$. Moreover, we have $\|\max_{jk}(X_{jk,\cdot}\varepsilon_{j,\cdot})\|_{q,\varsigma} \le \|\max_{jk} X_{jk,\cdot}\|_{q,\varsigma}\|\max_j \varepsilon_{j,\cdot}\|_{q,\varsigma}$, and particularly $\||\varepsilon_t|_{\infty}\|_q \le \|\max_j\sum_k a_{jk}^t(\eta_{k,0} - \eta_{k,0}^*)\|_q \lesssim q\|\max_k\max_j a_{jk}^t(\eta_{k,0} - \eta_{k,0}^*)\|_q + \sqrt{q\log J}\{\sum_k \max_j (a_{jk}^t)^2(\mu_2^{\eta})^2\}^{1/2} \lesssim q\sum_k \max_j |a_{jk}^t|\mu_q^{\eta} \vee \sqrt{q\log J} \times \{\sum_k \max_j (a_{jk}^t)^2\}^{1/2}\mu_2^{\eta}$, 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 $\|\max_j \varepsilon_{j,\cdot}\|_{q,\varsigma} \le C_3\sup_{m\ge 1}(m+1)^{\varsigma}(m \vee 1)^{-c}J\sqrt{\log J} \le C_3\sup_{m\ge 1}(m+1)^{\varsigma-c}J\sqrt{\log J}$, where $C_3 > 0$ depends on $\mu_q^{\eta}$.

To summarize, if the $q$th 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 $2c\sqrt{n}\max_{1\le k\le K}|S_{jk}/\Psi_{jk}|$, 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*

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

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

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

*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 that a stronger exponential moment condition is satisfied,

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

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 $\|X_{jk,t}\varepsilon_{j,t} - X^*_{jk,t}\varepsilon^*_{j,t}\|_q \le 2|[A^{t-1}]_j|_1\mu_q^2$. 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.4 in the Supplementary Material, we arrive at the following error bounds with probability $1-\alpha-o(1)$,

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

as $\lambda_j^0(1-\alpha) \lesssim \sqrt{n}(\log K)^{1/\gamma}\max_{1\le k\le K}\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,0}$. 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.

In the special case with i.i.d. data, the dependence adjusted norm would be $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} \le 2\|X_{jk,t}\varepsilon_{j,t}\|_q$, and $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu,0}$ will be bounded by a constant which is relevant to the corresponding tail assumptions of the moments. Compared to the standard rate for LASSO estimators such as in Theorem 1 of Belloni and Chernozhukov (2013) with independent errors, our results will be the same for the case of Gaussian innovation (i.e. $\nu=0$). Moreover, for time series data, disregarding the dependency adjusted norm term, our convergence rate of prediction norm $\sqrt{s_j\log K/n}$ (given $\nu=0$) is also of the same order as the rate for stable Gaussian processes studied in Basu and Michailidis (2015).

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

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

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

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

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

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

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 $L_{1,j} = \{\Phi_{j,4,\varsigma} \Phi_{j,4,0} (\log K)^2\}^{1/\varsigma}$, $W_{1,j} = (\Phi^6_{j,6,0} + \Phi^4_{j,8,0}) \{\log(Kn)\}^7$, $W_{2,j} = \Phi^2_{j,4,\varsigma} \{\log(Kn)\}^4$, $W_{3,j} = [n^{-\varsigma} \{\log(Kn)\}^{3/2} \times \Theta_{j,2q,\varsigma}]^{1/(1/2-\varsigma-1/q)}$, $N_{1,j} = (n/\log K)^{q/2} \Theta^q_{j,2q,\varsigma}$, $N_{2,j} = n (\log K)^{-2} \Phi^{-2}_{j,4,\varsigma}$, $N_{3,j} = \{n^{1/2} (\log K)^{-1/2} \Theta^{-1}_{j,2q,\varsigma}\}^{1/(1/2-\varsigma)}$.

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

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 Material.

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 $L_{1,j} \max(W_{1,j}, W_{2,j}) = o(1) \min(N_{1,j}, N_{2,j})$. 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* $\min_{1 \leq k \leq K} \operatorname{avar}(S_{jk}) \geq c_j$, *then we have*

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

*where* $Z_j \sim \mathrm{N}(0, \Sigma_j)$, $\Sigma_j$ *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* $\Sigma_j$, *namely*

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

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 $\lambda_j$ 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*

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

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 $\Sigma_j$ for $j = 1, \ldots, J$ as follows, given $\mathsf{E}\, X_{j,t}\varepsilon_{j,t} = 0$, and consider

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

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*

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

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

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* $Z_\alpha = 2\tilde{c}\sqrt{n\log K}$, *for* $\tilde{c} \ge \sqrt{2}c$, *where* $c$ *is the one in the definition of* $\lambda_j^0(1-\alpha)$, *then we have* $\lambda_j^0(1-\alpha)$ *satisfying*

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

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

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

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

We note that the allowed dimension $K$ is still of polynomial 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 $t=1,\dots,n$ are divided into $l_n$ blocks with the same number of observations $b_n$, $n=b_nl_n$ (without loss of generality), where $b_n, l_n \in \mathbb{Z}$.

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

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

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

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. We discuss the consistency rate of the bootstrap statistics with generated errors in the Supplementary Material; see Comment B.3 and Theorem B.1.

THEOREM 5.3 (Validity of multiplier block bootstrap method). *Under the conditions of Theorem* 5.2, *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 Material*), *then we have*

$$\sup_{\alpha\in(0,1)} \left|\mathrm{P}\Big(\max_{1\le k\le K}|S_{jk}/\Psi_{jk}| \ge q^{[B]}_{j,(1-\alpha)}\Big) - \alpha\right| \to 0 \quad \text{as } n\to\infty. \tag{5.17}$$

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

(a) $Q(1-\alpha)$: the $(1-\alpha)$ quantile of $2c\max_{1\le k\le K,1\le j\le J}\sqrt{n}|Z_{jk}/\Psi_{jk}|$. 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 $\max_{1\le k\le K,1\le j\le J}|Z^{[B]}_{jk}/\Psi_{jk}|$.

Section A in the Supplementary Material provides the main theorems for joint equation estimation. In particular, the dimension along $k=1,\dots,K$ and $j=1,\dots,J$ 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) $\ell_1$-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 $\lambda$ is the joint penalty level.

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

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

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

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 $1 - o(1)$,

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

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

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 $j = 1,\dots,J$, the following model is considered

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

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 $F_{\varepsilon_j}$ 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 $j = 1,\dots,J$, $\beta_{jk}^0 (k = 1,\dots,p_j)$ 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 $\psi_{jk}\{Z_{j,t}, \beta_{jk}, h_{jk}(X_{j(-k),t})\} = \{1/2 - \mathbf{1}(Y_{j,t} \le X_{jk,t}\beta_{jk} + X_{j(-k),t}^\top \beta_{j(-k)})\}(X_{jk,t} - X_{j(-k),t}^\top \gamma_{j(-k)})$.

Assume that there exists $s = s_n \ge 1$ such that $|\beta_{j(-k)}^0|_0 \le s$, $|\gamma_{j(-k)}^0|_0 \le s$, 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

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

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

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

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

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 $\psi_{jk}$. The estimator $\widehat{\beta}_{jk}$ is obtained as a $Z$-estimator by solving the sample analogue of (5.21)

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

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:

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

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

(C2) The true parameter $\beta_{jk}^0$ 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 $r_n \stackrel{\text{def}}{=} n^{-1/2}\{\log(a_n/\epsilon)\}^{1/2} \max_{(j,k)\in G} \|\psi_{jk,\cdot}^0\|_{2,\varsigma} + n^{-1} r_\varsigma \{\log(a_n/\epsilon)\}^{3/2} \times \|\max_{(j,k)\in G} |\psi_{jk,\cdot}^0|\|_{q,\varsigma}$, $r_n \lesssim \rho_n$ ($\rho_n$ 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 there exists constant $L_n \ge 1$ such that for every $\vartheta \in \{\beta, h_1, \dots, h_M\}$, $\mathsf{E}[\sup_{\beta \in \mathcal{B}_{jk}} |\partial_\vartheta \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h_{jk}^0(X_{j(-k),t}) | X_{j(-k),t}\}|^2] \le L_n$.

Moreover, there exist measurable functions $\ell_1(\cdot), \ell_2(\cdot)$, constants $L_{1n}, L_{2n} \ge 1$, $\upsilon > 0$, and a cube $\mathcal{T}_{jk}(X_{j(-k),t}) = \bigtimes_{m=1}^M \mathcal{T}_{jk,m}(X_{j(-k),t})$ in $\mathbb{R}^M$ with center $h_{jk}^0(X_{j(-k),t})$ such that for every $\vartheta, \vartheta' \in \{\beta, h_1, \dots, h_M\}$ we have $\sup_{(\beta,h) \in \mathcal{B}_{jk} \times \mathcal{T}_{jk}(X_{j(-k),t})} |\partial_\vartheta \partial_{\vartheta'} \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h) | X_{j(-k),t}\}| \le \ell_1(X_{j(-k),t})$, $\mathsf{E}\{|\ell_1(X_{j(-k),t})|^4\} \le L_{1n}$, and for every $\beta, \beta' \in \mathcal{B}_{jk}$, $h, h' \in \mathcal{T}_{jk}(X_{j(-k),t})$ we have $\mathsf{E}[\{\psi_{jk}(Z_{j,t}, \beta, h) - \psi_{jk}(Z_{j,t}, \beta', h')\}^2 | X_{j(-k),t}] \le \ell_2(X_{j(-k),t})(|\beta - \beta'|^\upsilon + |h - h'|_2^\upsilon)$, and $\mathsf{E}\{|\ell_2(X_{j(-k),t})|^4\} \le L_{2n}$.

(C4) Identifiability: $2|\mathsf{E}[\psi_{jk}\{Z_{j,t}, \beta, h_{jk}^0(X_{j(-k),t})\}]| \ge |\phi_{jk}(\beta - \beta_{jk}^0)| \wedge c_1$ 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 $|\phi_{jk}| \ge c_1$.

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

(C6) The class of functions $\mathcal{F}_{jk} = \{z \mapsto \psi_{jk}\{z, \beta, \tilde{h}(x_{j(-k)})\} : \beta \in \mathcal{B}_{jk}, \tilde{h} \in \mathcal{H}_{jk} \cup \{h^0_{jk}\}\}$ ($z$ 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 satisfies the entropy condition $\mathrm{ent}(\epsilon, \mathcal{F}_{jk}) \le Cs\log(a_n/\epsilon)$ for all $0 < \epsilon \le 1$. It also has measurable envelope $F_{jk} \ge \sup_{f\in\mathcal{F}_{jk}} |f|$, such that $F = \max_{(j,k)\in G} F_{jk}$ satisfies $\mathsf{E}\{F^q(z)\} < C$ for some $q \ge 4$.

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

(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. $n^{-1/2}\{s(\log a_n/\epsilon)\}^{1/2}\max_{f\in\mathcal{F}'}\|f(z_t)\|_2 + n^{-1}r_\varsigma\{s(\log a_n/\epsilon)\}^{3/2}\|\bar{F}'(z_t)\|_q = \mathcal{O}(\rho_n)$. $\mathcal{F}' = \{z \mapsto \psi_{jk}\{z, \beta, \tilde{h}(x_{j(-k)})\} : (j,k) \in G, \beta \in \mathcal{B}_{jk}, \tilde{h} \in \mathcal{H}_{jk} \cup \{h^0_{jk}\}\}$ with $\bar{F}' = \sup_{f\in\mathcal{F}'}|f|$.

(C9) Let $B^h_\Phi = \max_{m\in\{1,2\}}\Phi^h_{m,2,\varsigma}$, $B^h_\Omega = \max_{m\in\{1,2\}}\Omega^h_{m,q,\varsigma}$, $B^{'h}_\Phi = \max_{m\in\{1,2\}}\Phi^{'h}_{m,2,\varsigma}$, and $B^{'h}_\Omega = \max_{m\in\{1,2\}}\Omega^{'h}_{m,q,\varsigma}$ (see (B.10), (B.11) and (B.16) in the Supplementary Material for the definitions of $\Phi^h_{m,2,\varsigma}$, $\Omega^h_{m,q,\varsigma}$, $\Phi^\beta_{2,\varsigma}$, $\Omega^\beta_{q,\varsigma}$, $\Phi^{'h}_{m,2,\varsigma}$, $\Omega^{'h}_{m,q,\varsigma}$, $\Phi^{'\beta}_{2,\varsigma}$, $\Omega^{'\beta}_{q,\varsigma}$). The following restrictions are assumed:

$$
\begin{aligned}
&s\rho_n(\log a_n)^{1/2}B^h_\Phi + n^{-1/2}r_\varsigma\rho_n s^2(\log a_n)^{3/2}B^h_\Omega = o(g_n^{-1}),\\
&\rho_n(s\log a_n)^{1/2}\Phi^\beta_{2,\varsigma} + n^{-1/2}r_\varsigma\rho_n(s\log a_n)^{3/2}\Omega^\beta_{q,\varsigma} = o(g_n^{-1}),\\
&B^{'h}_\Phi\rho_n s^{1/2} = \mathcal{O}\Big(\max_{f\in\mathcal{F}'}\|f(z_t)\|_2\Big), \qquad B^{'h}_\Omega\rho_n s^{1/2} = \mathcal{O}(\|\bar{F}'(z_t)\|_q),\\
&\Phi^{'\beta}_{2,\varsigma}\rho_n = \mathcal{O}\Big(\max_{f\in\mathcal{F}'}\|f(z_t)\|_2\Big), \qquad \Omega^{'\beta}_{q,\varsigma}\rho_n = \mathcal{O}(\|\bar{F}'(z_t)\|_q).
\end{aligned}
$$

(C9′) Consider the stronger exponential moment condition as in (5.6) and corresponding to (C5), assume that $\mathsf{E}[\{\tilde{h}_{jk,m}(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}^2] \lesssim (\rho^e_n)^2$. 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.18) and (B.21) in the Supplementary Material. The following restrictions are assumed:

$$
\begin{aligned}
&n^{-1/2}\{(\log a_n/\epsilon)\}^{1/\gamma}\max_{(j,k)\in G}\|\psi^0_{jk,\cdot}\|_{\psi_\nu,0} \lesssim r_n,\\
&(s\log a_n)^{1/\gamma}\Big[\rho^e_{n,\upsilon} \vee \rho^e_n\Big\{\Big(s^{1/2}\max_{m\in\{1,2\}}\Phi^h_{m,\psi_\nu,0}\Big) \vee \Phi^\beta_{\psi_\nu,0}\Big\}\Big] = o(g_n^{-1}),\\
&n^{-1/2}\{s(\log a_n/\epsilon)\}^{1/\gamma}\max_{f\in\mathcal{F}'}\|f(z.)\|_{\psi_\nu,0} = \mathcal{O}(\rho^e_n),\\
&\rho^e_n\Big\{\Big(s^{1/2}\max_{m\in\{1,2\}}\Phi^{'h}_{m,\psi_\nu,0}\Big) \vee \Phi^{'\beta}_{\psi_\nu,0}\Big\} = \mathcal{O}\Big(\max_{f\in\mathcal{F}'}\|f(z.)\|_{\psi_\nu,0}\Big),
\end{aligned}
$$

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 non-smooth score function. Conditions (C2) and (C8) imply that $_{(j,k)\in G}\max\|\psi^0_{jk,\cdot}\|_{2,\varsigma}\lesssim s^{1/2}\max_{f\in\mathcal{F}'}\|f(z_t)\|_2$ 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 $|\widehat{\beta}^{[1]}_{j(-k)}-\beta^0_{j(-k)}|_{j,\mathrm{pr}}\lesssim_{\mathrm{P}}\{\sqrt{s\log(a_n/\alpha)/n}\vee n^{1/q-1}(\varpi_n a_n/\alpha)^{1/q}\}\max_{1\le k\le K}\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$, 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 (Example 3, continued), given a geometric decay rate, we have $L_{2n}, B^h_\Phi, B^{'h}_\Phi, \Phi^\beta_{2,\varsigma}, \Phi^{'\beta}_{2,\varsigma}, \max_{f\in'}\|f(z_t)\|_2$, $\max_{(j,k)\in G}\|\psi^0_{jk,\cdot}\|_{2,\varsigma}\lesssim M_n$, where $M_n$ only depends on the $2q$-th moments of $\varepsilon_t$ and $\varsigma$. Moreover, suppose these quantities are bounded by constant and let $d_n\overset{\mathrm{def}}{=}(|G|\vee J)$, 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.4 in the Supplementary Material. 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, Chernozhukov and Kato (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 $\max\{L_{2n}, B^h_\Phi, B^{'h}_\Phi, \Phi^\beta_{2,\varsigma}, \Phi^{'\beta}_{2,\varsigma}, \max_{f\in\mathcal{F}'}\|f(z_t)\|_2, \max_{(j,k)\in G}\||\psi^0_{jk,\cdot}|\|_{2,\varsigma}\}=\mathcal{O}(n^{k_1})$, and $\max\{B^h_\Omega, B^{'h}_\Omega, \Omega^\beta_{q,\varsigma}, \Omega^{'\beta}_{q,\varsigma}, \|F(z_t)\|_q, \|F'(z_t)\|_q, \|\max_{(j,k)\in G}|\psi^0_{jk,\cdot}|\|_{q,\varsigma}\}=\mathcal{O}(n^{k_2})$, with $0\le k_1\le k_2$, and let $s=\mathcal{O}(n^v)$, $\log a_n=\mathcal{O}(n^r)$. Then (C8) and (C9) imply that

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

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

and

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

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

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*

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

*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 Theorem 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 Lemmas B.9, B.10 and B.11 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 $v$, 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 $\gamma = 2/(2v+1)$. For example, when $v = 1/2$, $\gamma = 1$ 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 Algorithms 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 Algorithms 1 and 2 satisfy*

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

COROLLARY 5.6 (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:

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

where $q\ge 1$, 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 $q \geq 2$ 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 5.7 (Consistency of the estimated 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* $\min_{(j,k)\in G} \operatorname{avar}(n^{-1/2}\sum_{t=1}^n \zeta_{jk,t}) \geq c$, *with probability* $1 - o(1)$, *we have*

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

*where* $\widetilde{\mathrm{CI}}_{jk}(\alpha) \stackrel{\mathrm{def}}{=} [\widehat{\beta}_{jk} \pm \widehat{\sigma}_{jk} n^{-1/2} q(1-\alpha)]$, *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* $\sigma_{jk}$.

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 statistics 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. More details can be found in Comment B.5 in the Supplementary Material.

COROLLARY 5.8 (Validity of multiplier bootstrap). *Under* (A6) *and 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.28)), *we have*

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

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

COMMENT 5.10 (Admissible rate of $b_n$). Again, consider the special case of VAR(1) with i.i.d. errors (Example 3, continued), with $\Theta_{q,\varsigma}^\zeta = \mathcal{O}(|G|^{1/q})$ and $\Phi_{q,\varsigma}^\zeta = \mathcal{O}(1)$, for $\varsigma > 1$. Then in Corollary 5.8, the restrictions on $b_n$ in (B.28) along with (A6) boil down to a set of simple admissible rates. In particular, letting $\log|G| = \mathcal{O}(n^r)$, we need $2r < \eta < 1 - 5r$ and $|G|(\log|G|)^{3q/2} \vee |G|^2(\log|G|)^q c_n^{q/2} = o(n^{q/2-1})$, where $c_n^{-1} = o(1)$. Note that the rate can be further improved by employing the exponential inequality under stronger tail assumptions.

**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.

TABLE 1

*Prediction norm and Euclidean norm ratios (overall $\lambda$ relative to equation-by-equation $\lambda_j$'s, average over equations). Results (mean, median and standard deviation) are computed over* 1000 *replications*

| | $J = K = 50$ | $J = K = 100$ | $J = K = 150$ |
|---|---|---|---|
| | | 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 |

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 $\beta_j$ 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 $|\widehat{\beta}_j - \beta_j^0|_{j,\mathrm{pr}}$ and the Euclidean norm $|\widehat{\beta}_j - \beta_j^0|_2$ ratios are presented in Table 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 ($\lambda_j$ and $\lambda$ 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.

It is evident from Table 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 $M_\ell$ are independently drawn from Ginibre matrices, that is, all the entries of $M_\ell$ are i.i.d. $\mathrm{N}(0, 1)$, and in practice the sum is truncated to $\sum_{\ell=0}^{1000}$. We set $\rho$ 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). $\varepsilon_t$ 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 trade-off, 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 $J^{-1} \sum_{j=1}^{J} |\widehat{\beta}_j - \beta_j^0|_{j,\mathrm{pr}}$ with several block sizes $b_n$ under different settings and the minimal ones are in bold.

From Table 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

TABLE 2
*The prediction norm (average over equations) using several choices of $b_n$. Results are computed over 1000 simulations*

| | $\rho = 0.1$ (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 |

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 $|\widehat{\beta}_j - \beta_j^0|_{j,\mathrm{pr}}$ and the Euclidean norm $|\widehat{\beta}_j - \beta_j^0|_2$ ratios (using the optimal $b_n$ suggested in Table 2 for each case correspondingly) are presented in Table 3. Again we report the results with the jointly estimated $\lambda$ (selected by the algorithm proposed in Section 3.2 based on multiplier block bootstrap) relative to using the single equation $\lambda_j$'s.

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

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

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, $\beta_{jk}^0$ and $\theta_{jk}^0$ 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 $X_t$, $\varepsilon_t$ and $v_t$ is same as the dependent data setting above.

TABLE 3
*Prediction norm and Euclidean norm ratios (overall $\lambda$ relative to equation-by-equation $\lambda_j$'s, average over equations). Results (mean, median and standard deviation) are computed over 1000 replications*

| | $\rho = 0.1$ (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 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 $\alpha^0$ values (given the significance level = 0.05)*

| | $\rho = 0.1$ (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 |
| | | | $\alpha^0 \sim$ 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 |
| | | | $\alpha^0 \sim$ 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 |

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 $\alpha^0$ uniformly lies in [0, 2.5] and [0, 5] illustrate the power results.

Table 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 $\rho$. Multiple testing procedure via step-down method (see, e.g., Chernozhukov, Chetverikov and Kato (2013), Romano and Wolf (2005)), is considered to control the false positives in evaluating the power performance. The rejection rates are computed over 1000 simulation samples.

It is shown that for individual inference our proposed individual bootstrap approach provides a closer size control to the nominal $\alpha$ 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 (2019) 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 (2019), 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—$\text{Pos}^{\text{BL}}_{j,i,t}$, $\text{Neg}^{\text{BL}}_{j,i,t}$, $\text{Pos}^{\text{LM}}_{j,i,t}$, $\text{Neg}^{\text{LM}}_{j,i,t}$—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)

$$B_{j,t} = \log\left(\left\{1 + m^{-1}\sum_{i=1}^{m}\mathbf{1}(\text{Pos}_{j,i,t} > \text{Neg}_{j,i,t})\right\} \Big/ \left\{1 + m^{-1}\sum_{i=1}^{m}\mathbf{1}(\text{Pos}_{j,i,t} < \text{Neg}_{j,i,t})\right\}\right). \tag{7.1}$$

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 Material.

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:

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

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

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

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 $\widehat{\beta}^{[1]}_{j(-j)}$, $\widehat{\gamma}^{[1]}_j$, $\widehat{\delta}^{[1]}_j$ and $\widehat{c}^{[1]}_j$.

S2 For each $j$, run LASSO on $B_{j,t} = (B^\top_{-j,t}, z_t^\top, r_{j,t-1})^\top \theta_j + v_{j,t}$ 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 $\widehat{v}_{j,t} = B_{j,t} - (B^\top_{-j,t}, z_t^\top, r_{j,t-1})^\top \widehat{\theta}_j$.

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

If for stock $j$, the sentiment variable of firm $k$ is selected into the active set after the individual significance test that is, 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 1–2.

Figures 1–2 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 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 consumer 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 a high-dimensional VAR(1) model as in Example 3, the $j$th 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 $\widetilde{Y}_{j,t} \sim \mathrm{MVN}(0, \Sigma_j)$, $\Sigma_j = \left[\begin{smallmatrix} \Gamma_{jj} & \Phi_{j\cdot}\Gamma \\ (\Phi_{j\cdot}\Gamma)^\top & \Gamma \end{smallmatrix}\right]$. Define $C_j \overset{\text{def}}{=} \Phi_{j\cdot}\Gamma\Phi_{j\cdot}^\top$, then we have the precision matrix as $\Theta_j = \Sigma_j^{-1} = \left[\begin{smallmatrix} (\Gamma_{jj}-C_j)^{-1} & -(\Gamma_{jj}-C_j)^{-1}\Phi_{j\cdot} \\ -\Phi_{j\cdot}^\top(\Gamma_{jj}-C_j)^{-1} & \Gamma^{-1}+\Phi_{j\cdot}^\top(\Gamma_{jj}-C_j)^{-1}\Phi_{j\cdot} \end{smallmatrix}\right]$. It can be seen that $\Phi_{jk} = 0$ would imply that

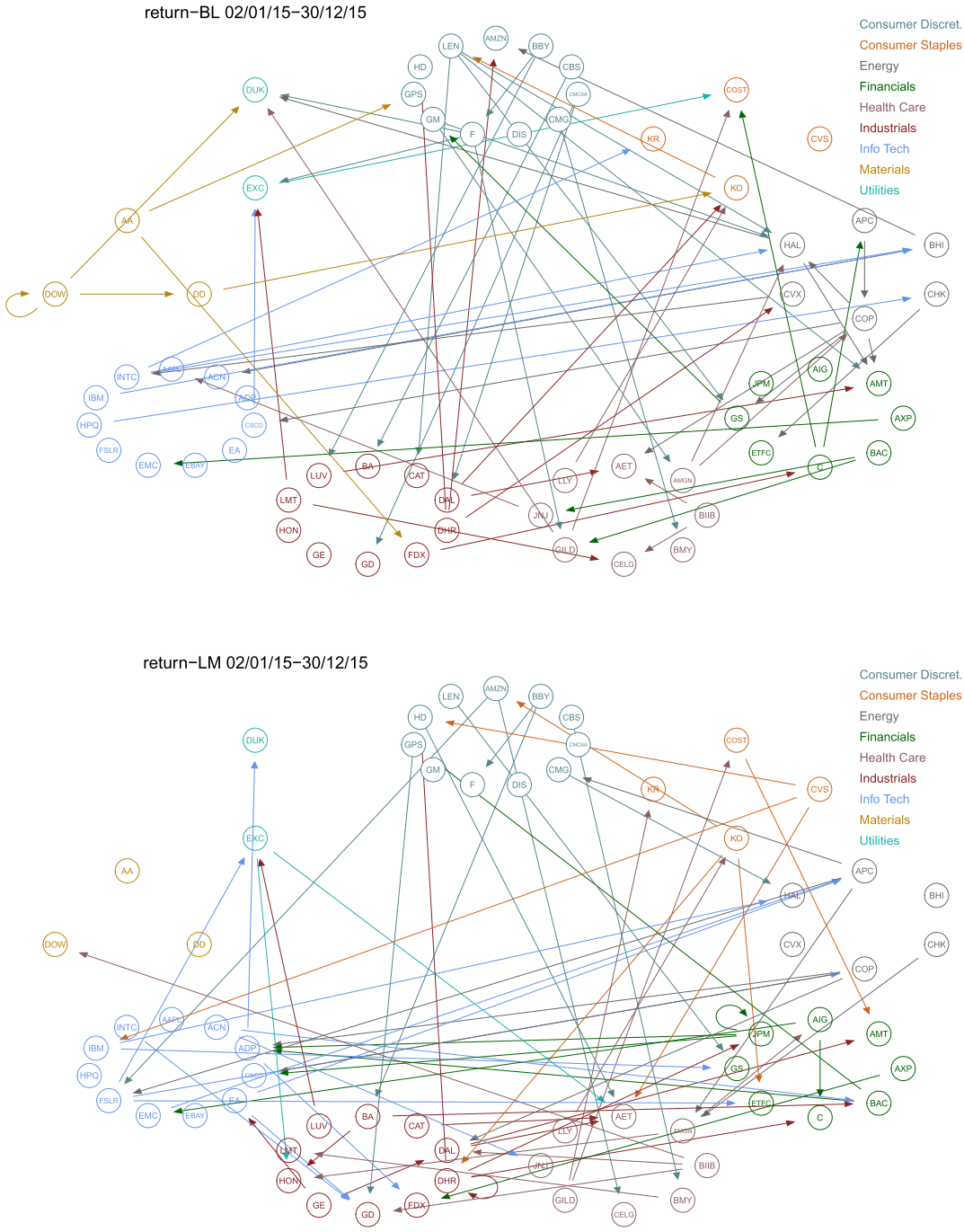


FIG. 1. *The dependency network among individual stocks from sentiment variables to return.*

the $(1, k+1)$th element of $\Theta_j$ is zero and vice versa. In addition, a LASSO type estimator proposed in Yuan and Lin (2007) can be obtained by solving

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

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

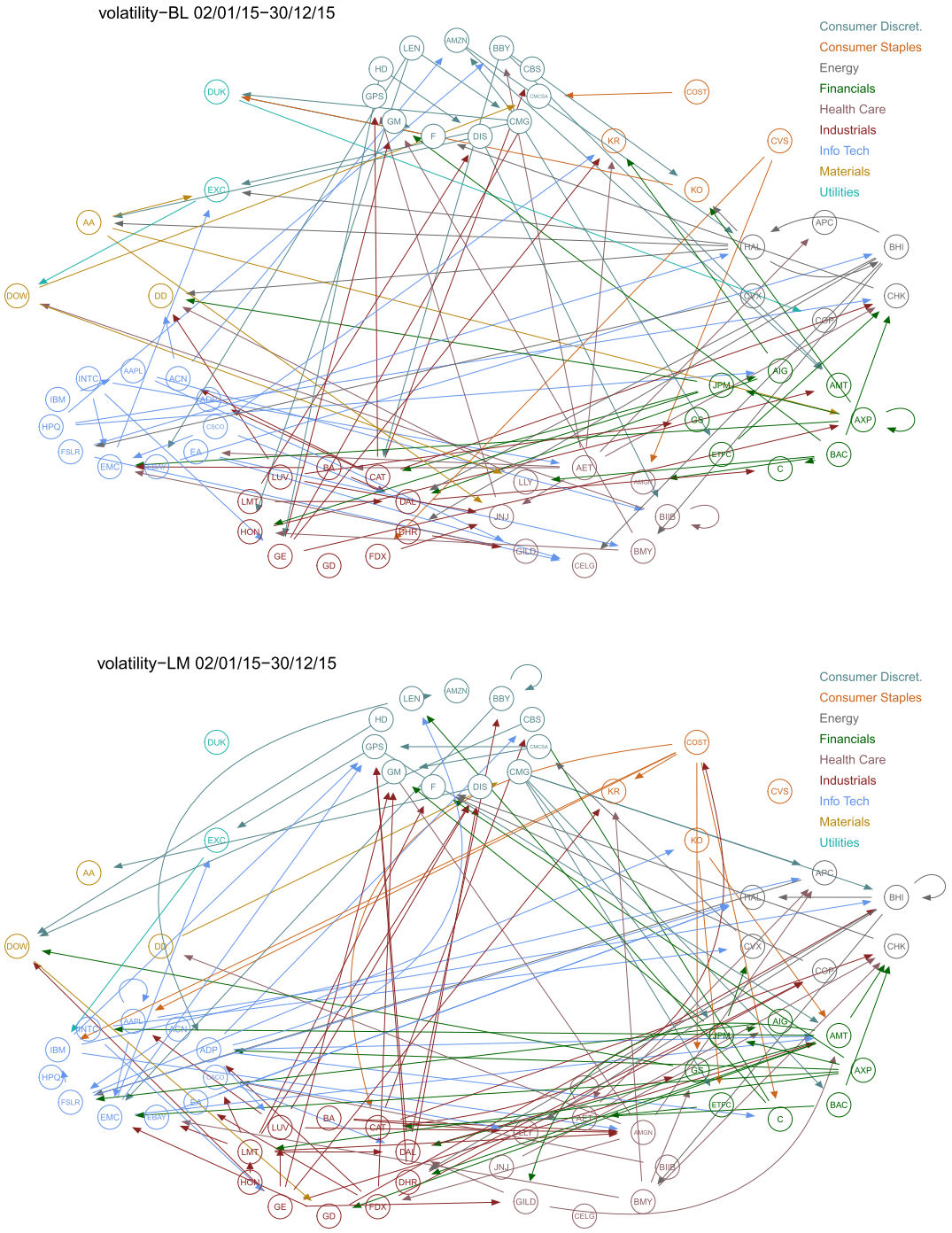


FIG. 2. *The dependency network among individual stocks from sentiment variables to volatility.*

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.

**Acknowledgments.** We thank Oliver Linton, Bryan Graham, Manfred Deistler, Hashem Pesaran, Michael Wolf, Valentina Corradi, Zudi Lu, Liangjun Su, Peter Phillips, Frank Wind-

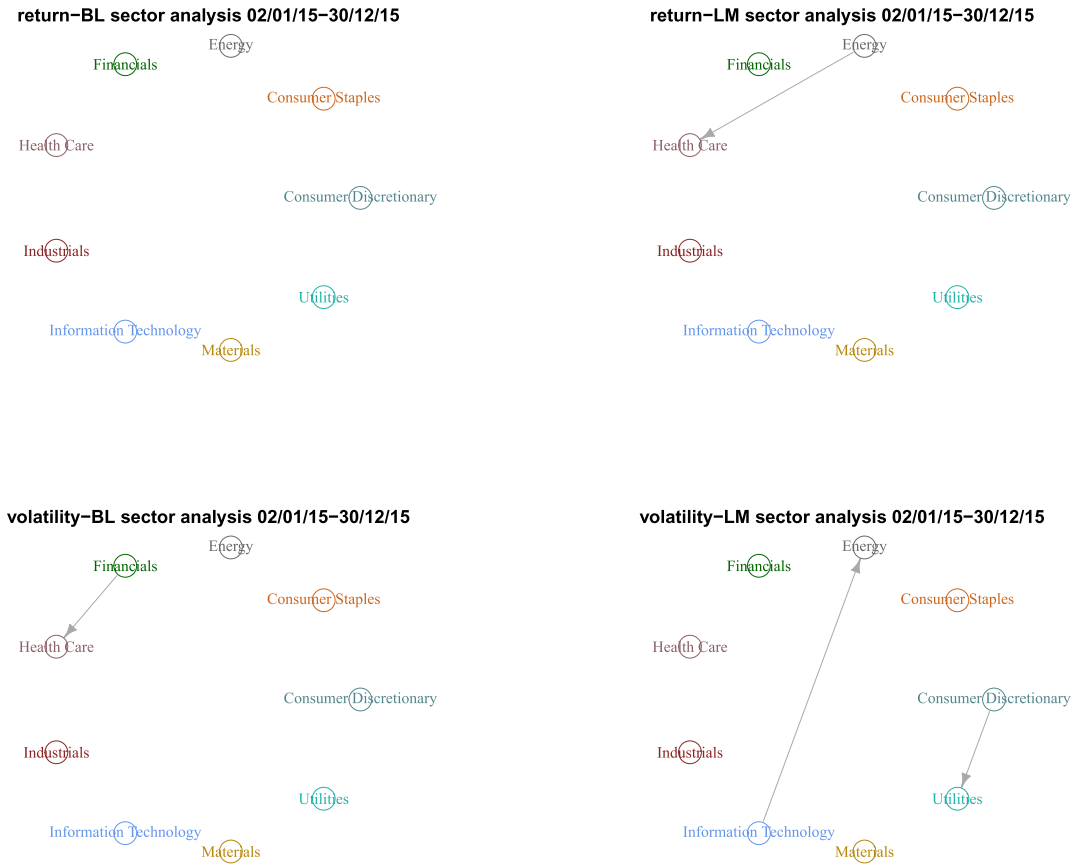


FIG. 3. *The dependency network among sectors from sentiment variables to financial variables.*

meijer, Wenyang Zhang and Likai Chen for helpful comments and suggestions. We also thank the Editor and the two anonymous referees for their valuable comments. We remain responsible for any errors or omissions.

Chen Huang is the corresponding author.

**Funding.** Financial support from the Deutsche Forschungsgemeinschaft via IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged.

## SUPPLEMENTARY MATERIAL

**Supplement to "LASSO-driven inference in time and space"** (DOI: 10.1214/20-AOS2019SUPP; .pdf). The supplemental file contains all proofs and technical details.

## REFERENCES

ANDREWS, D. W. K. (1984). Nonstrong mixing autoregressive processes. *J. Appl. Probab.* **21** 930–934. MR0766830 https://doi.org/10.2307/3213710

ANTWEILER, W. and FRANK, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *J. Finance* **59** 1259–1294.

AUDRINO, F. and TETEREVA, A. (2019). Sentiment spillover effects for us and European companies. *J. Bank. Financ.* **106** 542–567.

BAKER, M. and WURGLER, J. (2006). Investor sentiment and the cross-section of stock returns. *J. Finance* **61** 1645–1680.

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BELLONI, A., CHEN, M. and CHERNOZHUKOV, V. (2016). Quantile graphical models: Prediction and conditional independence with applications to financial risk management. Preprint. Available at arXiv:1607.00286.

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163 https://doi.org/10.3150/11-BEJ410

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2011). Inference for high-dimensional sparse econometric models. Preprint. Available at arXiv:1201.0220.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

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** 77–94. MR3335097 https://doi.org/10.1093/biomet/asu056

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

CHEN, C. Y.-H., HÄRDLE, W. K. and OKHRIN, Y. (2019). Tail event driven networks of SIFIs. *J. Econometrics* **208** 282–298. MR3906972 https://doi.org/10.1016/j.jeconom.2018.09.016

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461 https://doi.org/10.1214/14-AOS1230

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2019). Inference on causal and structural parameters using many moment inequalities. *Rev. Econ. Stud.* **86** 1867–1900. MR4009488 https://doi.org/10.1093/restud/rdy065

CHERNOZHUKOV, V. and HANSEN, C. (2008). Instrumental variable quantile regression: A robust inference approach. *J. Econometrics* **142** 379–398. MR2408741 https://doi.org/10.1016/j.jeconom.2007.06.005

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CHERNOZHUKOV, V., KARL HÄRDLE, W., HUANG, C. and WANG, W. (2021). Supplement to "LASSO-driven inference in time and space." https://doi.org/10.1214/20-AOS2019SUPP.

DEZEURE, R., BÜHLMANN, P. and ZHANG, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST* **26** 685–719. MR3713586 https://doi.org/10.1007/s11749-017-0554-2

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. *J. Clin. Bioinformat.* **1** 27. https://doi.org/10.1186/2043-9113-1-27

EPSKAMP, S., WALDORP, L. J., MÕTTUS, R. and BORSBOOM, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivar. Behav. Res.* **53** 453–480. https://doi.org/10.1080/00273171.2018.1454823

GARMAN, M. B. and KLASS, M. J. (1980). On the estimation of security price volatilities from historical data. *J. Bus.* **53** 67–78.

HÄRDLE, W. K., WANG, W. and YU, L. (2016). TENET: Tail-Event driven NETwork risk. *J. Econometrics* **192** 499–513. MR3488092 https://doi.org/10.1016/j.jeconom.2016.02.013

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.

HAUTSCH, N., SCHAUMBURG, J. and SCHIENLE, M. (2015). Financial network systemic risk contributions. *Review of Finance* **19** 685–738.

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

HUANG, D., YIN, J., SHI, T. and WANG, H. (2016). A statistical model for social network labeling. *J. Bus. Econom. Statist.* **34** 368–374. MR3523781 https://doi.org/10.1080/07350015.2015.1039014

JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inf. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629

KOCK, A. B. and CALLOT, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econometrics* **186** 325–344. MR3343790 https://doi.org/10.1016/j.jeconom.2015.02.013

KOLACZYK, E. D. and CSÁRDI, G. (2014). *Statistical Analysis of Network Data with R. Use R!* Springer, New York. MR3288852 https://doi.org/10.1007/978-1-4939-0983-4

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. Springer, New York. MR2724368 https://doi.org/10.1007/978-0-387-74978-5

KRAMPE, J. KREISS, J.-P. and PAPARODITIS, E. (2018). Bootstrap based inference for sparse high-dimensional time series models. Preprint. Available at arXiv:1806.11083.

LAHIRI, S. N. (1999). Theoretical comparisons of block bootstrap methods. *Ann. Statist.* **27** 386–404. MR1701117 https://doi.org/10.1214/aos/1018031117

LIN, J. and MICHAILIDIS, G. (2017). Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *J. Mach. Learn. Res.* **18** Paper No. 117, 49. MR3725456 https://doi.org/10.1631/jzus.a1500279

LOUGHRAN, T. and MCDONALD, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* **66** 35–65.

LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368 https://doi.org/10.1007/978-3-540-27752-1

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

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

NEYKOV, M., NING, Y., LIU, J. S. and LIU, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statist. Sci.* **33** 427–443. MR3843384 https://doi.org/10.1214/18-STS661

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 Syst. Biol.* **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** 416–429.

ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821 https://doi.org/10.1198/016214504000000539

TETLOCK, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *J. Finance* **62** 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. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

WU, W.-B. and WU, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* **10** 352–379. MR3466186 https://doi.org/10.1214/16-EJS1108

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018

ZHANG, X. and CHENG, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112** 757–768. MR3671768 https://doi.org/10.1080/01621459.2016.1166114

ZHANG, D. and WU, W. B. (2017a). Gaussian approximation for high dimensional time series. *Ann. Statist.* **45** 1895–1919. MR3718156 https://doi.org/10.1214/16-AOS1512

ZHANG, D. and WU, W. B. (2017b). Supplement material for "Gaussian approximation for high dimensional time series. Available at *Ann. Statist.* online. https://doi.org/10.1214/16-AOS1512SUPP

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHANG, J. L., HÄRDLE, W. K., CHEN, C. Y. and BOMMES, E. (2016). Distillation of news flow into analysis of stock reactions. *J. Bus. Econom. Statist.* **34** 547–563. MR3547995 https://doi.org/10.1080/07350015.2015.1110525

ZHU, Y. and BRADIC, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* **113** 1583–1600. MR3902231 https://doi.org/10.1080/01621459.2017.1356319

ZHU, X., PAN, R., LI, G., LIU, Y. and WANG, H. (2017). Network vector autoregression. *Ann. Statist.* **45** 1096–1123. MR3662449 https://doi.org/10.1214/16-AOS1476

ZHU, X., WANG, W., WANG, H. and HÄRDLE, W. K. (2019). Network quantile autoregression. *J. Econometrics* **212** 345–358. MR3994021 https://doi.org/10.1016/j.jeconom.2019.04.034