Department of Economics

LASSO-Driven Inference in Time and Space

Victor Chernozhukov
Massachusetts Institute of Technology

Wolfgang K. Härdle
Humboldt-Universität zu Berlin, Singapore Management University, Xiamen University

Chen Huang
University of St. Gallen

Weining Wang[1]
City, University of London

Department of Economics
Discussion Paper Series
No. 18/04

[1] Corresponding author: Weining Wang, Department of Economics, City, University of London, Northampton Square, London EC1V 0HB, UK. Email: weining.wang@city.ac.uk

LASSO-Driven Inference in Time and Space [*]

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

June 10, 2018

Abstract

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

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

1 Introduction

Many applications in economics, finance, and statistics are concerned with a system of ultra high-dimensional objects that communicate within complex dependency channels. Given a complex system involving many factors, one builds a network model by taking a large set of regressions, i.e. regressing every factor in the system on a large subset of other factors. Examples include analysis of financial systemic risk by quantile predictive graphical models with LASSO (Least Absolute Shrinkage and Selection Operator) (Hautsch et al., 2015; Härdle et al., 2016; Belloni et al., 2016) and limit order book network modeling via the penalized vector autoregressive approach (Härdle et al., 2018). Another example is quantifying the spillover effects or

[*] We thank Weibiao Wu, Oliver Linton, Bryan Graham, Manfred Deistler, Hashem Pesaran, Michael Wolf, Valentina Corradi, Zudi Lu, Liangjun Su, Peter Phillips, Frank Windmeijer, Wenyang Zhang and Likai Chen for helpful comments and suggestions. We remain responsible for any errors or omissions.

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

[‡] Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin. Sim Kee Boon Institute for Financial Economics, Singapore Management University. The Wang Yanan Institute for Studies in Economics, Xiamen University.

[§] Faculty of Mathematics and Statistics, University of St. Gallen.

[¶] Department of Economics, City, University of London.

externalities for a social network, especially when the social interactions (or the interconnectedness) is not obvious (Manresa, 2013). In general, a step-by-step LASSO procedure is very helpful for network formation. In pursuing a highly structural approach, one certainly favors a simple set of regressions that allows multiple insights on the econometric structure. Therefore, a sequence of regressions with LASSO is a natural path to take. Especially in cases of reduced forms of simultaneous equation models and structural vector autoregressive (VAR) models, one can attain valuable pre-information on the core structure by running a set of simple regressions with LASSO shrinkage.

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

A second crucial issue is to establish simultaneous inference on parameters. 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 a 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.

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

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

Theoretically, the LASSO tuning parameter selection requires characterizing the asymptotic distribution of the maximum of a high dimensional random vector. Chernozhukov et al. (2013a) develop a Gaussian approximation for the maximum of a sum of high-dimensional random vectors, which is in fact the basic tool for modern high-dimensional estimation. Here it is applied to LASSO inference. Moreover, Chernozhukov et al. (2013b) deliver results for the case of β-mixing processes. Although it is quite common to assume a mixing condition which is at base a concept yielding asymptotic independence, it is not in general easy to verify the condition for a particular process, and some simple linear processes can be excluded from the strong mixing class, Andrews (1984). With an easily accessible dependency concept as in Wu (2005), Zhang and Wu (2017) derive Gaussian approximation results for a wide class of stationary processes. Note that the dependence measure is linked to martingale decompositions and is therefore readily connected with a pool of results on tail probabilities, moment inequalities and central limit theorems of martingale theory. Our results are built on the above-mentioned theoretical works and we extend them substantially to fit into the estimation in a system of regression equations. In particular, our LASSO estimation is with random design for dependent data; therefore, we need to deal with the population implications of the Restricted Eigenvalue (RE) condition. Moreover, we show the interaction between the tail assumption and the dimensionality of the covariates in our theoretical results.

In the meantime, the issue of simultaneous inference is challenging and has motivated a series of research articles. For the case of i.i.d. data, Belloni et al. (2011, 2014), Zhang and Zhang (2014), Javanmard and Montanari (2014), Van de Geer et al. (2014), Neykov et al. (2015), Chernozhukov et al. (2016), Zhu and Bradic (2017), among others, develop confidence intervals of low-dimensional variables in high-dimensional models with various forms of de-biased/orthogonalization methods. Still in the case of i.i.d. data, Belloni et al. (2015b) establish a uniform post-selection inference for the target parameters defined via de-biased Huber's Z-estimators when the dimension of the parameters of interest is potentially larger than the sample size, where they employed the multiplier bootstrap to the estimated residuals. Wild and residual bootstrap-assisted approaches are also studied in Dezeure et al. (2017); Zhang and Cheng (2017) for the case of mean regression. We pick up the line of the inference analysis of Belloni et al. (2015b) and employ it in a temporal and cross-sectional dependence framework, thus making it applicable to a rich class of high-dimensional time series. The core proof strategy is vastly different, as it is well known that the technique for handling the suprema of empirical processes indexed by functional classes with dependent data is not the same as in i.i.d. cases. For instance, the key Bahadur representation in Belloni et al. (2015b) applies maximal inequalities derived in Chernozhukov et al. (2014) for i.i.d. random variables, while we derive the key concentration inequalities based on a martingale approximation method.

The following notations are adopted throughout this paper. For a vector $v = (v_1, \dots, v_p)^\top$, let $|v|_\infty \overset{\text{def}}{=} \max_{1 \leqslant j \leqslant p} |v_j|$ and $|v|_s \overset{\text{def}}{=} (\sum_{j=1}^p |v_j|^s)^{1/s}$, $s \geqslant 1$. For a random variable X, let $\|X\|_q \overset{\text{def}}{=} (\mathsf{E}\,|X|^q)^{1/q}$, $q > 0$. For any function $g : \mathcal{W} \to \mathbb{R}$, $\mathsf{E}_n(g) \overset{\text{def}}{=} n^{-1} \sum_{t=1}^n \{g(\omega_t)\}$ and $G_n(g) \overset{\text{def}}{=} n^{-1/2} \sum_{t=1}^n [g(\omega_t) - \mathsf{E}\{g(\omega_t)\}]$. Given two sequences of positive numbers x_n and y_n, write $x_n \lesssim y_n$ if there exists constant $C > 0$ such that $x_n / y_n \leqslant C$.

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

2 The System Model

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

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

where $X_{j,t} = (X_{jk,t})_{k=1}^{K_j}$, $K_j = \dim(X_{j,t}) \leqslant K$. We allow the dimension K of $X_{j,t}$ and the number of equations, J to be large, potentially larger than n, which interplays 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 (**A Regression Model**)**.** Suppose that we are interested in estimating the predictive model for a response variable U_t:

$$U_t = X_t^\top \gamma^0 + \varepsilon_t, \quad \mathsf{E}\,\varepsilon_t X_t = 0,$$

and also predictive relations between covariates for a strict subset G:

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

where $X_{-k,t} = (X_{\ell,t})_{\ell \neq k} \in \mathbb{R}^{\mathrm{K}-1}$ and $|G|$ is the cardinality of the set G. This is a special SRE model with

$$(Y_{1,t}, X_{1,t}, \varepsilon_{1,t}, \beta_1^0) = (U_t, X_t, \varepsilon_t, \gamma^0),$$

$$(Y_{j,t}, X_{j,t}, \varepsilon_{j,t}, \beta_j^0) = (X_{(j-1),t}, X_{-(j-1),t}, \nu_{(j-1),t}, \delta_{(j-1)}^0), \quad j = 2, \dots, J = (|G| + 1).$$

It can be seen that we only put contemporaneous exogeneity conditions for X_t. It is worth mentioning that this SRE case is closely related to the semiparametric estimation framework studied in Section 2.4 in Belloni et al. (2015b).

Modeling predictive relations between covariates is important for constructing joint confidence intervals for the entire parameter vector $(\gamma_k^0)_{k=1}^K$ in the main regression equation. Indeed, the construction relies on the semi-parametrically efficient point estimators obtained from the empirical analog of the following moment equation:

$$\mathsf{E}[(U_{k,t}^0 - X_{k,t}\gamma_k^0)\nu_{k,t}] = 0, \quad k \in G, \tag{2.1}$$

where $U_{k,t}^0 = U_t - X_{-k,t}^\top \gamma_{-k}^0$ 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 (**Many Regression Models**)**.** Example 1 can be generalized to handle many regression models of the following form:

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

and also predictive relations between covariates:

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

where G should be a strict subset of $\{1, \ldots, K\}$ for the strict exogeneity assumption to hold. This is again a special SRE with

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

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

Here, the understanding of the predictive relations between covariates is important for constructing joint confidence intervals for the entire parameter vector $\{(\gamma_{mk}^0)_{k=1}^K\}_{m=1}^M$. Indeed, the construction relies on the efficient point estimators obtained from the empirical analog of the following orthogonalized moment equation:

$$\mathsf{E}[(U_{mk,t}^0 - X_{k,t}\gamma_{mk}^0)\nu_{k,t}] = 0, \quad k \in G, \quad m = 1, \ldots, M, \tag{2.2}$$

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.

Example 3 (**Simultaneous Equation Systems (SES)**)**.** Suppose there are many regression equations in the following form:

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

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

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

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

$$U_t = \mathbf{B}X_t + \nu_t, \quad \mathsf{E}\,\nu_{m,t}X_t = 0, \quad m = 1, \ldots, M, \tag{2.3}$$

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

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

Example 4 (**Large Vector Autoregression Models**)**.** In the case where the covariates involve lagged variables of the response, SRE can be written as a large vector autoregression model. For example, the VAR(p) model,

$$U_t = \sum_{\ell=1}^{p} \mathbf{B}^\ell U_{t-\ell} + \varepsilon_t, \quad \mathsf{E}\,\varepsilon_{m,t}U_{t-\ell} = 0, \quad m = 1, \ldots, M, \tag{2.4}$$

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

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

2.1 Practical Examples

Example 5 (**Identification Test for Large Structural Vector Autoregression Models**)**.** Denote $U_t = (U_{1,t}, U_{2,t}, \ldots, U_{M,t})^\top$. A large structural VAR can be represented in the following form (without loss of generality, consider only lag one):

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

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

$$U_t = \mathbf{D}U_{t-1} + \nu_t, \tag{2.5}$$

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

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

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

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

In practice, ν_t are replaced by the residuals obtained from a large VAR reduced form regression as in example 4. The estimator of δ_j is denoted as $\widehat{\delta}_j$. It can be obtained by LASSO estimation, which give us a sparse estimator of the jth row of the matrix $\mathbf{A}^{-1}$ up to a scaling factor. Repeating this step for any j, one may formulate estimators for each row and perform simultaneous inference/hypothesis testing on the structural matrix $\mathbf{A}^{-1}$.

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

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

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

Example 6 (**Cross-sectional Asset Pricing**)**.** Denote $Y_{j,t}$ as the excess return for asset j and period t. Asset pricing models explain the cross sectional variation in expected returns across assets; see e.g. Cochrane (2009). In particular, the variation of expected cross sectional returns is explained by the exposure to $K-1$ factors $X_{jk,t}$, $k = 1, \ldots, K-1$. One commonly used way to estimate an asset pricing model is to run a system of regression equations:

$$Y_{j,t} = \beta_{j0} + \sum_{k=1}^{K-1} \beta_{jk} X_{jk,t} + \varepsilon_{j,t}, \tag{2.6}$$

where $X_{jk,t}$'s are the factor returns (assumed to be excess returns of zero-cost portfolios). The selection of factors is a critical issue and the SRE framework addresses this issue, in particular when the number of factors K is large. See Feng et al. (2017) for a detailed model-selection exercise on picking asset pricing factors. The factor premiums are $\mathsf{E}(X_{jk,t})$ and the pricing errors are β_{j0}. Usually, asset pricing imposes the restriction that all β_{j0}'s are zero. Our simultaneous inference framework naturally serves the purpose of simultaneously testing the zero pricing errors in a cross sectional regression setup. Namely, we are interested in testing $H_0 : \beta_{j0} = 0, \forall j = 1, \ldots, J$ versus $H_A : \exists\, j$ such that $\beta_{j0} \neq 0$. Our test procedure in Section 4.2 can be directly applied to achieve this goal.

Example 7 (**Network Formation and Spillover Effects**)**.** There is an emerging literature in economics concerning quantifying spillover effects and network formation. One leading example is as in Manresa (2013), which attempts to quantify social returns to research and development (R&D). Here, $U_{j,t}$ is taken to be the log output for firm j and time t. This output is loading on $D_{j,t}$ (capital stock for firm j and period t), and the aggregated spill-overs from the capital stock of other firms $\sum_{i\neq j} w_{ij} D_{i,t}$. The regression equation also controls for other covariates $X_{j,t}$ (e.g., log labor, log capital etc.):

$$U_{j,t} = \beta_j D_{j,t} + \sum_{i\neq j} \omega_{ij} D_{i,t} + \gamma_j^\top X_{j,t} + \varepsilon_{j,t}, \tag{2.7}$$

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

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

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

Comment 2.1. Suppose there is unobserved heterogeneity in $U_{j,t}$, e.g. $U_{j,t} = \alpha_j + \sum_{i \neq j} w_{ij} D_{i,t}$ $+\varepsilon_{j,t}$, where w_{ij} characterizes the spillover of individual i on j, and α_j is the individual fixed effect. For this situation consider the demeaned version to eliminate the individual specific effects and work with the new model: $\widetilde{U}_{j,t} = \sum_{i \neq j} w_{ij} \widetilde{D}_{i,t} + \widetilde{\varepsilon}_{j,t}$, where $\widetilde{U}_{j,t} = U_{j,t} - \frac{1}{n} \sum_{t=1}^n U_{j,t}$, $\widetilde{D}_{i,t} = D_{i,t} - \frac{1}{n} \sum_{t=1}^n D_{i,t}$, $\widetilde{\varepsilon}_{j,t} = \varepsilon_{j,t} - \frac{1}{n} \sum_{t=1}^n \varepsilon_{j,t}$, under the condition that $U_{j,t}$ has no feedback effects on $D_{i,t}$ (for example, $D_{i,t}$ should not be the lagged variable of $U_{j,t}$).

3 Effective Prediction Using Sparsity Method

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

3.1 Sparsity in SRE

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

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

It is now well understood that sparsity can be easily extended to approximate sparsity, in which sorted absolute values of coefficients decrease faster to zero, with an additional bias term in the bound.

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

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

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

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

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

and the Euclidean norm:

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

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

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

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

$$\Psi_{jk} \overset{\text{def}}{=} \sqrt{\operatorname{Var}(S_{jk})}, \tag{3.3}$$

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

where $c > 1$, e.g., $c = 1.1$, and $1-\alpha$ is a confidence level, e.g. $\alpha = 0.01$.

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

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

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

3.2 Multiplier Bootstrap for the Joint Penalty Level

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

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

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

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

where λ_j are the penalty levels and Ψ_{jk} are the penalty loadings. For instance, we can take the X-independence choice using Gaussian approximation (in the heteroscedasticity case): $2c'\sqrt{n}\Phi^{-1}\{1 - \alpha'/(2K_j)\}$ for λ_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 $\sum_{\ell=-\infty}^{\infty} \mathsf{E}(X_{jk,t}\breve{\varepsilon}_{j,t}X_{jk,(t-\ell)}\breve{\varepsilon}_{j,(t-\ell)})$, e.g. the Newey-West estimator is given by

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

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

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

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

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

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

Comment 3.1 (Block bootstrap procedures). (i) Concerning the determination of b_n, we shall report the fitting errors with several block sizes b_n in the simulation study. If it is the case that n cannot be divided by b_n with no remainder, one can take $l_n = \lfloor n/b_n \rfloor$ and drop the remaining observations.

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

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

4 Valid Inference on the Coefficients

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

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

We list three algorithms to estimate a subset of parameters β^0_{jk}, for $(j,k) \in G$. 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 $(j,k) \in G$), and keep the quantity $X^\top_{j(-k),t}\widehat{\beta}^{[1]}_{j(-k)}$.
- S2 Run LASSO (for each $(j,k) \in G$) 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 $Y_{j,t} - X^\top_{j(-k),t}\widehat{\beta}^{[1]}_{j(-k)}$ on $X_{jk,t}$ using $\widehat{v}_{jk,t}$ as an instrument variable, attaining the final estimator $\widehat{\beta}^{[2]}_{jk}$. We refer to Belloni et al. (2015b); Chernozhukov and Hansen (2008) for more details about how to achieve the estimator in this step.

Comment 4.1. Our algorithms follow patterns discussed in Belloni et al. (2013, 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 et al. (2011, 2014). In particular, the estimator under LS IV regression (2-step least square regression) is given by

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

The second line in (4.1) is exactly the same as the de-biased or de-sparsified LASSO estimator given in Eq. (5) in Zhang and Zhang (2014) or Eq. (5) in Van de Geer et al. (2014). As remarked in Belloni et al. (2013, 2015a), one can alternatively implement an algorithm via double selection as in Belloni et al. (2011, 2014). In particular, heteroscedastic LASSO is employed in S2″ and the IV regression is replaced by a either LASSO or LAD regression on the target variable and all covariates selected in the first two steps. □

Algorithm 3: Double selection-based algorithm

S1$''$ Run LS LASSO (for each j) of $Y_{j,t}$ on $X_{j,t}$:

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

S2$''$ Run Heteroscedastic LASSO (for each $(j,k) \in G$) 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 $\sqrt{\operatorname{lvar}\{X_{j\ell,t}(X_{jk,t} - \frac{1}{n}\sum_{t=1}^{n} X_{jk,t})\}}$ and then refined by $\sqrt{\operatorname{lvar}(X_{j\ell,t}\widehat{v}_{jk,t})}$, for $\ell \neq k$, and $\widehat{v}_{jk,t} = X_{jk,t} - X_{j(-k),t}^{\top}\widehat{\gamma}_{j(-k)}$ can be obtained by using the initial ones.

S3$''$ Run LS regression of $Y_{j,t}$ on $X_{jk,t}$ and the covariates selected in S1$''$ and S2$''$:

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

S3$'''$ Run LAD regression of $Y_{j,t}$ on $X_{jk,t}$ and the covariates selected in S1$''$ and S2$''$:

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

As shown in Belloni et al. (2011) and Belloni et al. (2015a), the double selection approach in S3$''$ or S3$'''$ creates an orthogonality condition with respect to the space spanned by the covariates selected by both steps, and thus generates an orthogonal relation to any space spanned by a linear projection of the covariates, e.g. $\widehat{v}_{jk,t}$. Therefore, the inference on the parameters may still be applied as in the framework of Algorithm 1 and 2.

4.1 Confidence Interval for a Single Coefficient

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

Let $\psi_{jk}(Z_{j,t}, \beta_{jk}, h_{jk})$ denote the score function, where $Z_{j,t} = (Y_{j,t}, X_{j,t}^{\top})^{\top}$, $h_{jk}(X_{j(-k),t}) = (X_{j(-k),t}^{\top}\beta_{j(-k)}, X_{j(-k),t}^{\top}\gamma_{j(-k)})^{\top}$. Consider the LAD-based case with $\psi_{jk}(Z_{j,t}, \beta_{jk}, h_{jk}) = \{1/2 - \mathbf{1}(Y_{j,t} \leqslant X_{jk,t}\beta_{jk} + X_{j(-k),t}^{\top}\beta_{j(-k)})\}v_{jk,t}$, define $\omega_{jk} \overset{\text{def}}{=} \mathsf{E}\{(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \psi_{jk,t}^{0})^2\} = \sum_{\ell=-(n-1)}^{n-1}(1 - \frac{|\ell|}{n}) \operatorname{cov}(\psi_{jk,t}^{0}, \psi_{jk,(t-\ell)}^{0})$ with $\psi_{jk,t}^{0} \overset{\text{def}}{=} \psi_{jk}(Z_{j,t}, \beta_{jk}^{0}, h_{jk}^{0})$, and $\phi_{jk} \overset{\text{def}}{=} \frac{\partial \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta, h_{jk}^{0})\}}{\partial \beta}|_{\beta=\beta_{jk}^{0}}$.As shown in Corollary 5.7 we have the limit distribution of $\widehat{\beta}_{jk}^{[2]}$:

$$\sigma_{jk}^{-1} n^{1/2} (\widehat{\beta}_{jk}^{[2]} - \beta_{jk}^{0}) \xrightarrow{\mathcal{L}} \mathrm{N}(0,1), \tag{4.2}$$

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

$$\mathrm{CI}_{jk}(\alpha) : [\widehat{\beta}^{[2]}_{jk} - \widehat{\sigma}_{jk} n^{-1/2}\Phi^{-1}(1-\alpha/2), \widehat{\beta}^{[2]}_{jk} + \widehat{\sigma}_{jk} n^{-1/2}\Phi^{-1}(1-\alpha/2)]. \tag{4.3}$$

Suppose we are interested in testing $H_0 : \beta^0_{jk} = 0$. For this purpose we employ the uniform Bahadur representation (Theorem 5.8) to construct the confidence interval via a bootstrap procedure. In particular, the distribution of the asymptotically pivotal statistics:

$$T_{jk} = \frac{\sqrt{n}(\widehat{\beta}^{[2]}_{jk} - \beta^0_{jk})}{\widehat{\sigma}_{jk}}, \tag{4.4}$$

is approximated via its block multiplier bootstrap counterpart:

$$T^*_{jk} = \frac{1}{\sqrt{n}} \sum_{i=1}^{l_n} e_{j,i} \sum_{l=(i-1)b_n+1}^{ib_n} \widehat{\zeta}_{jk,l}, \tag{4.5}$$

where $\zeta_{jk,t} = -\phi^{-1}_{jk}\sigma^{-1}_{jk}\psi^0_{jk,t}$, $e_{j,i}$ are independently drawn from $\mathrm{N}(0,1)$, l_n and b_n are the numbers of blocks and block size, respectively.

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

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

where $q^*_{jk}(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the bootstrapped distribution of $|T^*_{jk}|$.

Comment 4.2. Alternative bootstrap procedures may be considered as well, e.g. the residual multiplier bootstrap procedure:

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

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

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

Define $Y^*_{j,t} = X^\top_{j,t}\widehat{\beta}^{[1]}_j + \varepsilon^*_{j,t}$ and compute the bootstrap counterpart as

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

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 2.1, it is essential to conduct simultaneous inference on a group of parameters G. In this case, the null hypothesis is: $\mathbf{H}_0 : \beta_{jk} = 0$, $\forall (j,k) \in G$, and the alternative $\mathbf{H}_A : \beta_{jk} \neq 0$, for some $(j,k) \in G$, where the group G is a set of coefficients with cardinality $|G|$. Suppose for the j-th equation there are p_j target coefficients and the cardinality $|G| = \sum_j^J p_j$. 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)

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

One follows the algorithms to obtain $\widehat{\beta}_{jl}(1 \leqslant l \leqslant p_j)$ for each j. Then the idea of simultaneous inference is very straightforward. We aggregate the statistics T_{jk} in (4.4) 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 $q_G^*(1-\alpha/2)$ as the $(1-\alpha/2)$ quantile of $\max\limits_{(j,k)\in G}|T_{jk}^*|$. A joint confidence region is then:

$$\{\beta \in \mathrm{I\!R}^{|G|} : \max_{(j,k)\in G} T_{jk} \leqslant q_G^*(1-\alpha/2) \text{ and } \min_{(j,k)\in G} T_{jk} \geqslant -q_G^*(1-\alpha/2)\}, \tag{4.8}$$

and for each component $(j,k) \in G$, the confidence interval $\widetilde{\mathrm{CI}}_{jk}^*(\alpha)$ is given by $[\widehat{\beta}_{jk}^{[2]} - \widehat{\sigma}_{jk} n^{-1/2} q_G^*(1-\alpha/2), \widehat{\beta}_{jk}^{[2]} + \widehat{\sigma}_{jk} n^{-1/2} q_G^*(1-\alpha/2)]$. We show in Corollary 5.9 the consistency of this bootstrap confidence band in simultaneous inference.

5 Main Theorems

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

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

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

b) an empirical choice given a Gaussian approximation result is: $Q_j(1-\alpha) \overset{\text{def}}{=} 2c\sqrt{n}\Phi^{-1}\{1-\alpha/(2K_j)\}$;

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

5.1 Near Oracle Inequalities under IC

We first provide the 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 ξ_t, η_t are i.i.d. random elements (innovations or shocks, allowing for overlap, see Comment 5.1) across t, $\mathcal{F}_t = (\dots, \xi_{t-1}, \eta_{t-1}, \xi_t, \eta_t)$, $g_{jk}(\cdot)$ and $h_j(\cdot)$ are measurable functions (filters). $\mathsf{E}(X_{jk,t}\varepsilon_{j,t}) = 0$, for any $j, k \in 1, \cdots, J, 1, \cdots, K$.

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

For more details on this functional dependency measure, see Wu (2005). It should be noted that (A1) admits a wide class of processes. The largest value of ς which ensures a finite dependence adjusted norm characterizes the dependency structure of the process. The moment-based measure is directly connected with the impulse functions. A few examples for univariate time series Z_t are listed in Appendix B in the supplementary materials; for more examples please refer to Wu (2011).

(A2) Restricted eigenvalue (RE): given $\bar{c} \geqslant 1$, for $\beta \in \mathrm{I\!R}^{\mathrm{K_j}}$, with probability $1 - \mathcal{o}(1)$,

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

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

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

Comment 5.1. We allow for overlap in the elements in ξ_t and η_t, as long as the contemporaneous exogeneity condition $\mathsf{E}(X_{jk,t}\varepsilon_{j,t}) = 0$ is satisfied. For example, consider the VAR(1) model: $Y_t = AY_{t-1} + \varepsilon_t$, with $Y_t, \varepsilon_t \in \mathrm{I\!R}^{\mathrm{J}}$, and suppose that Y_t admits the representation $Y_t = \sum_{l=0}^{\infty} A^l \varepsilon_{t-l}$ with ε_{t-l} as measurable functions of $\xi_{-\infty}, \dots, \xi_{t-l}$. Thus $X_{jk,t} = g_{jk}(\dots, \xi_{t-1}) = \sum_{l=0}^{\infty}[A^l]_k \varepsilon_{t-1-l}$, where $[A^l]_k$ is the kth row of the matrix A^l, $k = 1, \dots, J$. In this case no serial correlation in the innovations ε_ts would be sufficient for $\mathsf{E}(X_{jk,t}\varepsilon_{j,t}) = 0$.

Comment 5.2. We show in Theorem A.1 (see the supplementary materials) that the RE (A2) and RSE (A6) conditions can be implied by assumptions on the corresponding population variance-covariance matrix. This illustrates the feasibility of the RE/RSE assumption.

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

obtained from (3.6) *satisfy*

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

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

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

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

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

$$\mathrm{P}(2c\sqrt{n}\max_{1\leqslant k\leqslant K}|S_{jk}/\Psi_{jk}|\geqslant r) \leqslant C_1\varpi_n n r^{-q}\sum_{k=1}^{K}\frac{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}^q}{\Psi_{jk}^q} + C_2\sum_{k=1}^{K}\exp\Big(\frac{-C_3r^2\Psi_{jk}^2}{n\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}^2}\Big), \tag{5.3}$$

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

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

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

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

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

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

with probability $1-\alpha-\mathcal{o}(1)$*, 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 5.3. 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 $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}^q$. 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 $\tilde{c} < q - 1 - \upsilon q/2$, if there exists q as the maximal number to guarantee $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}^q = \mathcal{O}(1)$ 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 5.4. Suppose a stronger exponential moment condition is satisfied,

$$\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu} = \sup_{q\geqslant 2} q^{-\nu}\sum_{t=0}^{\infty}\theta_{q,j,t} = \sup_{q\geqslant 2} q^{-\nu}\sum_{t=0}^{\infty}\|\,\mathsf{E}(X_{jk,t}\varepsilon_{j,t}|\mathcal{F}_0) - \mathsf{E}(X_{jk,t}\varepsilon_{j,t}|\mathcal{F}_{-1})\|_q < \infty, \tag{5.6}$$

where $\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{\psi_\nu}$ is interpreted as the dependence adjusted sub-exponential ($\nu = 1$) or sub-Gaussian ($\nu = 1/2$) norm, and $\theta_{q,j,t}$ denotes the predictive dependence measure. In this case, applying the exponential tail bounds as in Theorem 3 of Wu and Wu (2016), we arrive at the following error bounds with probability $1 - \alpha - o(1)$,

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

This bound (5.7) works with ultra-high dimension rate $\exp(n^{\alpha})$ ($\alpha < 1$) of K as only the exponential term shows in the inequality.

5.2 Gaussian Approximation for Dependent Data

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

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

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

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

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

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

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

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

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

ii) (strong dependency case) Given $0 < \varsigma < 1/2 - 1/q$, then $\Theta_{j,2q,\varsigma}(\log K)^{1/2} = o(n^{\varsigma})$ and $L_1 \max(W_{1,j}, W_{2,j}, W_{3,j}) = o(1)\min(N_{2,j}, N_{3,j})$.

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

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

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

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

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

Theorem 5.2 justifies the choice of λ_j as $Q_j(1-\alpha)$, which leads to the following corollary:

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

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

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 $X_{j,1}\varepsilon_{j,1}, \dots, X_{j,n}\varepsilon_{j,n}$, we propose to estimate the $K \times K$ long-run variance-covariance matrix Σ_j for $j = 1, \dots, J$ as follows, given $\mathsf{E}\, X_{j,t}\varepsilon_{j,t} = 0$, and consider:

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

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

Corollary 5.3. *Let the conditions of Theorem 5.2 hold, and assume $\Phi_{j,2q,\varsigma} < \infty$ with $q > 4$, $b_n = \mathcal{O}(n^{\eta})$ for some $0 < \eta < 1$. Let $F_\varsigma = n$, for $\varsigma > 1 - 2/q$; $F_\varsigma = l_n b_n^{q/2-\varsigma q/2}$, for $1/2 - 2/q < \varsigma < 1 - 2/q$; $F_\varsigma = l_n^{q/4-\varsigma q/2} b_n^{q/2-\varsigma q/2}$, for $\varsigma < 1/2 - 2/q$. Further assume $n^{-1}(\log K)^{3/2} \max\{n^{1/2}b_n^{1/2}\Phi_{j,2q,\varsigma}^2, n^{1/2}b_n^{1/2}\sqrt{\log K}\Phi_{j,8,\varsigma}^2, F_\varsigma^{2/q}\Gamma_{j,2q,\varsigma}^2\} = o(1)$. Then for each j we have*

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

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

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, $n = b_n l_n$ (without loss of generality), where $b_n, l_n \in \mathbb{Z}$.

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

$$Z^{[B]}_{jk} = \frac{1}{\sqrt{n}} \sum_{i=1}^{l_n} e_{j,i} \sum_{l=(i-1)b_n+1}^{ib_n} \varepsilon_{j,l} X_{jk,l}, \tag{5.14}$$

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

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

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

$$\sup_{\alpha \in (0,1)} \big| \mathrm{P}\big(\max_{1 \leqslant k \leqslant K} |S_{jk}/\Psi_{jk}| \leqslant q^{[B]}_{j,(1-\alpha)} \big) - (1-\alpha) \big| \to 0, \text{ as } n \to \infty. \tag{5.15}$$

5.4 Joint Penalty over Equations

In this section we provide results for joint equation estimation. The dimension along $k = 1, \dots, K$ and $j = 1, \dots, J$ will be considered together by vectorization, resulting in the dimension 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.

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

Theorem 5.4. *Under (A1) and (A3), we have*

$$\begin{aligned} \mathrm{P}(2c\sqrt{n} \max_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant J} |S_{jk}/\Psi_{jk}| \geqslant r) \leqslant & C_1 \varpi_n n r^{-q} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|^q_{q,\varsigma}}{\Psi^q_{jk}} \\ & + C_2 \sum_{j=1}^{J} \sum_{k=1}^{K} \exp\Big(\frac{-C_3 r^2 \Psi^2_{jk}}{n \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|^2_{2,\varsigma}} \Big), \end{aligned} \tag{5.16}$$

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

Corollary 5.4 (Bound for $\lambda^0(1-\alpha)$ and Oracle Inequalities under IC)**.** *Under (A1) and (A3), given* $\lambda^0(1-\alpha)$ *satisfies*

$$\lambda^0(1-\alpha) \lesssim \max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J} \Big\{ \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}\sqrt{n\log(KJ/\alpha)} \vee \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}(n\varpi_n KJ/\alpha)^{1/q}\Big\}, \tag{5.17}$$

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

$$|\widehat{\beta}_j - \beta_j^0|_{j,pr} \lesssim C\sqrt{s} \max_{1\leqslant k\leqslant K} \Psi_{jk} \max_{1\leqslant j\leqslant J} \Big\{ \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}\frac{\sqrt{\log(KJ/\alpha)}}{\sqrt{n}} \vee \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma} n^{1/q-1}(\varpi_n KJ/\alpha)^{1/q}\Big\}, \tag{5.18}$$

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

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

a) $Q(1-\alpha) \overset{\text{def}}{=} 2c\sqrt{n}\Phi^{-1}\{1-\alpha/(2KJ)\}$;

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

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

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

$$\||\mathcal{X}_\cdot|_\infty\|_{q,\varsigma} \overset{\text{def}}{=} \sup_{m\geqslant 0}(m+1)^\varsigma \sum_{t=m}^{\infty}\delta_{q,t},\ \delta_{q,t} \overset{\text{def}}{=} \||\mathcal{X}_t - \mathcal{X}_t^*|_\infty\|_q, \tag{5.19}$$

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

$$\begin{aligned}&\Phi_{q,\varsigma} \overset{\text{def}}{=} 2\max_{1\leqslant k\leqslant K, 1\leqslant j\leqslant J}\|X_{jk,\cdot}\|_{q,\varsigma}\|\varepsilon_{j,\cdot}\|_{q,\varsigma},\ \Gamma_{q,\varsigma} \overset{\text{def}}{=} 2\Big(\sum_{j=1}^J\|\varepsilon_{j,\cdot}\|_{q,\varsigma}^{q/2}\Big)^{2/q}\Big(\sum_{k=1}^K\sum_{j=1}^J\|X_{jk,\cdot}\|_{q,\varsigma}^{q/2}\Big)^{2/q}\\ &\Theta_{q,\varsigma} \overset{\text{def}}{=} \Gamma_{q,\varsigma} \wedge \{\||\mathcal{X}_\cdot|_\infty\|_{q,\varsigma}\|\varepsilon_{j,\cdot}\|_{q,\varsigma}(\log KJ)^{3/2}\}.\end{aligned} \tag{5.20}$$

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

(A5) i) (weak dependency case) Given $\Theta_{2q,\varsigma} < \infty$ with $q \geqslant 4$ and $\varsigma > 1/2-1/q$, then $\Theta_{2q,\varsigma}n^{1/q-1/2}\{\log(KJn)\}^{3/2} \to 0$ and $L_1\max(W_1,W_2) = \mathcal{o}(1)\min(N_1,N_2)$.

ii) (strong dependency case) Given $0 < \varsigma < 1/2-1/q$, then $\Theta_{2q,\varsigma}\{\log(KJ)\}^{1/2} = \mathcal{o}(n^\varsigma)$ and $L_1\max(W_1,W_2,W_3) = \mathcal{o}(1)\min(N_2,N_3)$.

Consider the case with $\Theta_{2q,\varsigma} = \mathcal{O}((KJ)^{1/q})$ and $\Phi_{2q,\varsigma} = \mathcal{O}(1)$ where $\varsigma > 1/2 - 1/q$. Then $\Theta_{2q,\varsigma} n^{1/q-1/2}\{\log(KJn)\}^{3/2} \to 0$ becomes $KJ\{\log(nKJ)\}^{3q/2} = \mathcal{o}(n^{q/2-1})$, which implies $L_1 \max(W_1, W_2) = \mathcal{o}(1)\min(N_1, N_2)$. This means that to make (A5) hold, the dimension KJ has to satisfy the condition such that $KJ\{\log(KJ)\}^{3q/2} = \mathcal{o}(n^{q/2-1})$.

Theorem 5.5. *Under (A1), (A3) and (A5), for each $k = 1, \dots, K$, $j = 1, \dots, J$ assume that there exists a constant $c > 0$ such that $\min\limits_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant J} \mathrm{Var}(X_{jk,t}\varepsilon_{j,t}) \geqslant c$, then we have*

$$\rho(D^{-1}\widetilde{\mathcal{S}}, D^{-1}\widetilde{\mathcal{Z}}) \to 0, \quad \textit{as } n \to \infty, \tag{5.21}$$

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

$$\{\sum_{\ell=-\infty}^{\infty} \mathsf{E}(X_{jk,t}X_{jk,(t-\ell)}\varepsilon_{j,t}\varepsilon_{j,(t-\ell)})\}^{1/2} = \sqrt{\mathrm{Var}(S_{jk})}, \ \textit{for } k = 1, \dots, K, j = 1, \dots, J.$$

Corollary 5.5. *Under the conditions of Theorem 5.5, we have*

$$\sup_{\alpha \in (0,1)} |\,\mathrm{P}\{\max_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant J} 2c\sqrt{n}|S_{jk}/\Psi_{jk}| \leqslant Q(1-\alpha)\} - (1-\alpha)| \to 0, \quad \textit{as } n \to \infty. \tag{5.22}$$

Corollary 5.6. *Under the conditions of Theorem 5.5, and assume $\Phi_{2q,\varsigma} < \infty$ with $q > 4$, $b_n = \mathcal{O}(n^\eta)$ for some $0 < \eta < 1$. Let $F_\varsigma = n$, for $\varsigma > 1 - 2/q$; $F_\varsigma = l_n b_n^{q/2-\varsigma q/2}$, for $1/2 - 2/q < \varsigma < 1 - 2/q$; $F_\varsigma = l_n^{q/4-\varsigma q/2} b_n^{q/2-\varsigma q/2}$, for $\varsigma < 1/2 - 2/q$. Given*

$$n^{-1}\{\log(KJ)\}^{3/2} \max\{n^{1/2}b_n^{1/2}\Phi^2_{2q,\varsigma}, n^{1/2}b_n^{1/2}\sqrt{\log(KJ)}\Phi^2_{8,\varsigma}, F_\varsigma^{2/q}\Gamma^2_{2q,\varsigma}\} = \mathcal{o}(1),$$

then we have

$$\rho(\widehat{D}^{-1}\widetilde{\mathcal{S}}, D^{-1}\widetilde{\mathcal{Z}}) \to 0, \quad \textit{as } n \to \infty, \tag{5.23}$$

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

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

Theorem 5.6. *Under (A1), (A3), and assume $\Phi_{2q,\varsigma} < \infty$ with $q > 4$, $b_n = \mathcal{O}(n^\eta)$ for some $0 < \eta < 1$ (the detailed rate is calculated in* (A.2)*), then*

$$\tilde{\rho}_n \overset{\mathrm{def}}{=} \sup_{r \in \mathbb{R}} |\,\mathrm{P}(|\widetilde{\mathcal{Z}}^{[B]}/\widetilde{\Psi}|_\infty \leqslant r|\mathcal{X}_\cdot, \varepsilon_\cdot) - \mathrm{P}(|\widetilde{\mathcal{Z}}/\widetilde{\Psi}|_\infty \leqslant r)| \to 0, \textit{ as } n \to \infty, \tag{5.24}$$

and

$$\sup_{\alpha \in (0,1)} |\,\mathrm{P}(|\widetilde{\mathcal{S}}/\widetilde{\Psi}|_\infty \leqslant q^{[B]}_{(1-\alpha)}) - (1-\alpha)| \to 0, \textit{ as } n \to \infty. \tag{5.25}$$

5.5 Post-Model Selection Estimation

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

i) ℓ_1-penalized regression (LASSO selection)

$$\breve{\beta}_j = \arg\min_{\beta\in\mathbb{R}^{\mathrm{K_j}}} \frac{1}{n}\sum_{t=1}^{n}(Y_{j,t} - X_{j,t}^{\top}\beta)^2 + \frac{\lambda}{n}\sum_{k=1}^{K_j}|\beta_{jk}|\Psi_{jk}, \tag{5.26}$$

where λ is the joint penalty level obtained above.

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

$$\widehat{\beta}_j^{[P]} = \arg\min_{\beta\in\mathbb{R}^{\mathrm{K_j}}} \{\frac{1}{n}\sum_{t=1}^{n}(Y_{j,t} - X_{j,t}^{\top}\beta)^2 : \beta_k = 0, k \notin \widehat{T}_j\}, \tag{5.27}$$

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

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

(A6) Restricted sparse eigenvalue (RSE): given $p < n$, for $\beta \in \mathbb{R}^{\mathrm{K}}$, with probability $1 - o(1)$,

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

Here p denotes the restriction on the length of the active set of T_j^c. When $T_j = \emptyset$, (A6) is reduced to the standard sparse eigenvalue condition. Moreover, let $\mu_j(p) \overset{\text{def}}{=} \frac{\sqrt{\phi_j(p)}}{\tilde{\kappa}_j(p)}$, and denote by $\widehat{p}_j \overset{\text{def}}{=} |\widehat{T}_j \setminus T_j|$ the number of components outside $T_j \overset{\text{def}}{=} \operatorname{supp}(\beta_j^0) = \{k \in \{1,\dots,K_j\} : \beta_{jk}^0 \neq 0\}$ selected by LASSO in the first step. The performance bounds for the OLS post-LASSO estimator are shown in the following theorem:

Theorem 5.7 (Prediction Performance Bounds for OLS Post-LASSO)**.** *Given (A1) and (A3), suppose (A2) (with $\bar{c} = \frac{c+1}{c-1}, c > 1$) and (A6) (with $\widehat{p}_j = |\widehat{T}_j \setminus T_j|$) hold uniformly over equations with probability $1 - o(1)$, then under the exact sparsity assumption (3.1), for any $\tau > 0$, there is a constant C_τ independent of n, for all $j = 1,\dots,J$ we have*

$$\begin{aligned}|\widehat{\beta}_j^{[P]} - \beta_j^0|_{j,pr} &\leqslant C_\tau\sigma_j\sqrt{\frac{\widehat{p}_j\log(KJ) + \log(e\mu_j(\widehat{p}_j))}{n}} \\ &+ \mathbf{1}(T_j \nsubseteq \widehat{T}_j)C\sqrt{s}\max_{1\leqslant k\leqslant K}\Psi_{jk}\max_{1\leqslant j\leqslant J}\{\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{2,\varsigma}\frac{\sqrt{\log(KJ/\alpha)}}{\sqrt{n}} \vee \|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}n^{1/q-1}(\varpi_n KJ/\alpha)^{1/q}\},\end{aligned} \tag{5.28}$$

with probability $1 - \alpha - \tau - o(1)$, where for $\varsigma > 1/2 - 1/q$ (weak dependence case), $\varpi_n = 1$; for $\varsigma < 1/2 - 1/q$ (strong dependence case), $\varpi_n = n^{q/2-1-\varsigma q}$. σ_j is the square root of the long-run variance of $\varepsilon_{j,t}$, and the constant C depends on the RE constants.

The proof of Theorem 5.7 is a direct application of Theorem 5 of Belloni and Chernozhukov (2013) by inserting the bound for $\lambda^0(1-\alpha)$ (5.17) provided in Corollary 5.4, and thus is omitted.

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

$$
\begin{aligned}
Y_{j,t} &= \sum_{k=1}^{p_j} X_{jk,t}\beta^0_{jk} + \sum_{k=p_j+1}^{K_j} X_{jk,t}\beta^0_{jk} + \varepsilon_{j,t}, \quad F_{\varepsilon_j}(0) = 1/2, &(5.29)\\
X_{jk,t} &= X^\top_{j(-k),t}\gamma^0_{j(-k)} + v_{jk,t}, \quad \mathsf{E}(v_{jk,t}|X_{j(-k),t}) = 0, \quad k = 1, \dots, p_j. &(5.30)
\end{aligned}
$$

In this subsection, we show the validity of the joint confidence region for simultaneous inference on $H_0 : \beta^0_{jk} = 0, \forall (j,k) \in G$, with $|G| = \sum_{j=1}^J p_j$. In particular, for $j = 1, \dots, J$, β^0_{jk} ($k = 1, \dots, p_j$) 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 function.

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 $Z_{j,t} \overset{\text{def}}{=} (Y_{j,t}, X^\top_{j,t})^\top$ and the vector-valued function $h_{jk}(\cdot)$ is a measurable map from $\mathrm{I\!R}^{\mathrm{K}-1}$ to $\mathrm{I\!R}^{\mathrm{M}}$ (M is fixed). In particular, in our linear regression case we have $h_{jk}(X_{j(-k),t}) = (X^\top_{j(-k),t}\beta_{j(-k)}, X^\top_{j(-k),t}\gamma_{j(-k)})^\top$, and for the LAD regression $\psi_{jk}\{Z_{j,t}, \beta_{jk}, h_{jk}(X_{j(-k),t})\} = \{1/2 - \mathbf{1}(Y_{j,t} \leqslant X_{jk,t}\beta_{jk} + X^\top_{j(-k),t}\beta_{j(-k)})\}(X_{jk,t} - X^\top_{j(-k),t}\gamma_{j(-k)})$.

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

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

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

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

$$
\mathsf{E}[\psi_{jk}\{Z_{j,t}, \beta^0_{jk}, h^0_{jk}(X_{j(-k),t})\}] = 0. \tag{5.31}
$$

However, the object $\arg \operatorname*{zero}_{\beta_{jk}\in\widehat{\mathcal{B}}_{jk}} \mathsf{E}_n\, |[\psi_{jk}\{Z_{j,t}, \beta_{jk}, h^0_{jk}(X_{j(-k),t})\}]|$ does not necessarily exist due to the discontinuity of the function ψ_{jk}. The estimator $\widehat{\beta}_{jk}$ is obtained as a Z-estimator by solving the sample analogue of (5.31)

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

where $g_n \overset{\text{def}}{=} \{\log(e|G|)\}^{1/2}$ 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:

$$\frac{\partial}{\partial h} \mathsf{E}\{\psi_{jk}(Z_{j,t}, \beta^0_{jk}, h)|X_{j(-k),t}\}|_{h=h^0_{jk}(X_{j(-k),t})} = 0. \tag{5.32}$$

(C2) The true parameter β^0_{jk} satisfies (5.31). Let $\mathcal{B}_{jk}$ be a fixed and closed interval and $\widehat{\mathcal{B}}_{jk}$ be a possibly stochastic interval such that with probability $1 - o(1)$, $[\beta^0_{jk} \pm c_1 r_n] \subset \widehat{\mathcal{B}}_{jk} \subset \mathcal{B}_{jk}$, where $r_n \stackrel{\text{def}}{=} n^{-1/2}(\log a_n)^{1/2} \max\limits_{(j,k)\in G} \|\psi^0_{jk,\cdot}\|_{2,\varsigma} + n^{-1} r_\varsigma (\log a_n)^{3/2} \| \max\limits_{(j,k)\in G} |\psi^0_{jk,\cdot}| \|_{q,\varsigma}$, $a_n \stackrel{\text{def}}{=} \max(JK, n, e)$, and $\psi^0_{jk,t} \stackrel{\text{def}}{=} \psi_{jk}\{Z_{j,t}, \beta^0_{jk}, h^0_{jk}(X_{j(-k),t})\}$. $r_\varsigma = n^{1/q}$ for $\varsigma > 1/2 - 1/q$ and $r_\varsigma = n^{1/2-\varsigma}$ for $\varsigma < 1/2 - 1/q$.

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

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

(C5) Properties of the nuisance function: with probability $1 - o(1)$, $\widehat{h}_{jk} \in \mathcal{H}_{jk}$, where $\mathcal{H}_{jk} = \times_{m=1}^M \mathcal{H}_{jk,m}$ and each $\mathcal{H}_{jk,m}$ being the class of functions of the form $\tilde{h}_{jk,m}(X_{j(-k),t}) = X^\top_{j(-k),t}\theta_{jk,m}$, $|\theta_{jk,m}|_0 \leqslant s$, $\tilde{h}_{jk,m} \in \mathcal{T}_{jk,m}$ and $\mathsf{E}[\{\tilde{h}_{jk,m}(X_{j(-k),t}) - h^0_{jk,m}(X_{j(-k),t})\}^2] \leqslant C_1 n^{-1} s (\log a_n) \max\limits_{1\leqslant k\leqslant K} (\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|^2_{q,\varsigma} \vee \|X_{jk,\cdot}v_{jk,\cdot}\|^2_{q,\varsigma})$.

(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 has measurable envelope $F_{jk} \geqslant \sup\limits_{f\in\mathcal{F}_{jk}} |f|$, such that $F = \max\limits_{(j,k)\in G} F_{jk}$ satisfies $\mathsf{E}\{F^q(z)\} < \infty$ for some $q \geqslant 4$.

(C7) Dimension growth rates: there exist sequences of constants $\rho_n \downarrow 0, \delta_n \downarrow 0$ such that $\rho_n^{\nu/2}(L_{2n} s \log a_n)^{1/2} + n^{-1/2} r_\varsigma (s \log a_n)^2 = o(g_n^{-1})$ and $n^{-1/2}(s \log a_n)^{1/2} + n^{-1} r_\varsigma (s \log a_n)^2 = \mathcal{O}(\rho_n)$.

(C8) The second-order moments of scores are bounded away from zero, i.e.,
$\omega_{jk} = \mathsf{E}([\frac{1}{\sqrt{n}} \sum_{t=1}^n \psi_{jk}\{Z_{j,t}, \beta^0_{jk}, h^0_{jk}(X_{j(-k),t})\}]^2) \geqslant c_1$.

(C9) 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. Recall that $\mathcal{X}_t \stackrel{\text{def}}{=} \mathrm{vec}[\{(X_{jk,t})_{k=1}^K\}_{j=1}^J]$. For $q > 2$, assume the dependence adjusted norm $\||\mathcal{X}_{\cdot}|_\infty\|_{q,\varsigma}$ is finite. The following restrictions are assumed:

$$
\begin{aligned}
&n^{-1/2}(\log a_n)^{1/2}\max_{(j,k)\in G}\|\psi^0_{jk,\cdot}\|_{2,\varsigma} + n^{-1}r_\varsigma(\log a_n)^{3/2}\Big\|\max_{(j,k)\in G}|\psi^0_{jk,\cdot}|\Big\|_{q,\varsigma} = \mathcal{o}(1),\\
&\Big\|\max_{\beta\in\mathcal{B}_{jk},\,(j,k)\in G}|\partial_\beta\,\mathsf{E}\{\psi_{jk}(Z_{j,\cdot},\beta,h^0_{jk})|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}|\Big\|_{q,\varsigma} = \mathcal{O}\{(s\log a_n)^{1/2}\rho_n^{-1}\},\\
&\max_{\beta\in\mathcal{B}_{jk},\,(j,k)\in G}\|\partial_\beta\,\mathsf{E}\{\psi_{jk}(Z_{j,},\beta,h^0_{jk})|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}\|_{2,\varsigma} = \mathcal{O}(\rho_n^{\nu/2-1}L_{2n}^{1/2}),\\
&\max_{1\leqslant k\leqslant K}(\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}\vee\|X_{jk,\cdot}v_{jk,\cdot}\|_{q,\varsigma})\||\mathcal{X}_{\cdot}|_\infty\|_{2q,\varsigma}\Big\|\max_{h\in\mathcal{H}_{jk},\,(j,k)\in G}|\partial_{h_m}\,\mathsf{E}\{\psi_{jk}(Z_{j,\cdot},\beta^0_{jk},h)|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}|\Big\|_{2q,\varsigma}\\
&= \mathcal{O}(n^{1/2}s^{-1/2}),\\
&\max_{1\leqslant k\leqslant K}(\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}\vee\|X_{jk,\cdot}v_{jk,\cdot}\|_{q,\varsigma})\||\mathcal{X}_{\cdot}|_\infty\|_{4,\varsigma}\max_{h\in\mathcal{H}_{jk},\,(j,k)\in G}\|\partial_{h_m}\,\mathsf{E}\{\psi_{jk}(Z_{j,\cdot},\beta^0_{jk},h)|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}\|_{4,\varsigma}\\
&= \mathcal{O}\{\rho_n^{\nu/2}L_{2n}^{1/2}n^{1/2}s^{-1}(\log a_n)^{-1/2}\},\\
&\max_{1\leqslant k\leqslant K}(\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}\vee\|X_{jk,\cdot}v_{jk,\cdot}\|_{q,\varsigma})\|\partial_{h_m}\,\mathsf{E}\{\psi_{jk}(Z_{j,\cdot},\beta^0_{jk},h)|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}\|_{2,\varsigma}\\
&= \mathcal{O}\{n^{1/2}(s\log a_n)^{-1/2}\rho_n\}.
\end{aligned}
$$

In addition, assume the dependence adjusted sub-Gaussian norm $\|\psi^0_{jk,\cdot}\|_{\psi_{1/2}}$ (defined in (5.6)) is finite.

Conditions (C1)-(C4) and (C8) assume mild restrictions on the Z-estimation problems. They include the LAD-based regression (used in Algorithm 2) with nonsmooth score function. 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 and Comment 5.4 ensure that our proposed algorithms can produce the estimator of the form $|\widehat{\beta}^{[1]}_{j(-k)} - \beta^0_{j(-k)}|_{j,pr} \lesssim \sqrt{s\log(a_n)/n}\max_{1\leqslant k\leqslant K}\|X_{jk,\cdot}\varepsilon_{j,\cdot}\|_{q,\varsigma}$ (with probability $1-\mathcal{o}(1)$), under the exponential moment condition in (C9). The moments of the envelopes are assumed to be bounded in (C6). As indicated in Belloni et al. (2015b), it can be accommodated to a bound growing with n by adjusting (C7). For the case with $n^{-1/2}r_\varsigma(s_n\log a_n)^{2/3} = \mathcal{o}(1)$, (C7) implies $n^{-1}s^2(\log a_n)^3 = \mathcal{o}(1)$ if $\nu = 2$, and $n^{-1}s^3(\log a_n)^5 = \mathcal{o}(1)$ if $\nu = 1$. Moreover, different from the i.i.d. case, (C9) imposes additional constraints on the rate of ρ_n regarding the dependence adjusted norm. As for (C9), for the LAD regression case, since the sub-gradient of the score function $\psi_{jk}\{\cdot\}$ is bounded, it is sufficient to impose the sub-Gaussian assumption on $v_{jk,t}$.

Comment 5.5 (Discussion of the case with linear processes and LAD regression)**.** The estimation requires the dependence adjusted norm $\|\partial_{h_m}\,\mathsf{E}\{\psi_{jk}(Z_{j,\cdot},\beta^0_{jk},h)|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}\|_{q,\varsigma}$ and $\|\partial_\beta\,\mathsf{E}\{\psi_{jk}(Z_{j,\cdot},\beta,h^0_{jk})|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}\|_{q,\varsigma}$ to be finite for $q > 2$. Suppose we have the linear process $X_{jk,t} = \sum_{l=0}^\infty a_{jk,l}\xi_{t-l}$ and $\varepsilon_{j,t} = \sum_{l=0}^\infty b_{j,l}\eta_{t-l}$, with i.i.d. innovations ξ_t and η_t, for all $j = 1,\ldots,J, k = 1,\ldots,K$. Thus $X^*_{jk,t} - X_{jk,t} = a_{jk,t}(\xi^*_0 - \xi_0)$ and $\varepsilon^*_t - \varepsilon_t = b_{j,t}(\eta^*_0 - \eta_0)$.

And for the linear regression model, where $h(X_{j(-k),t}) = (X^\top_{j(-k),t}\beta_{j(-k)}, X^\top_{j(-k),t}\gamma_{j(-k)})^\top$, the conditional distribution $F_{\varepsilon_{j,t}|\mathcal{F}_{t-1},X_{j(-k),t}}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}\} = F_{\eta_t}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}-\sum_{l=1}^\infty b_{j,l}\eta_{t-l}\}$. Therefore, we have $|F_{\varepsilon^*_{j,t}|\mathcal{F}^*_{t-1},X_{j(-k),t}}\{(\beta-\beta^0_{jk})X^*_{jk,t}+h_1-h^0_{jk,1}\} - F_{\varepsilon_{j,t}|\mathcal{F}_{t-1},X_{j(-k),t}}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}\}| \lesssim |(\beta-\beta^0_{jk})a_{jk,t}(\xi^*_0-\xi_0)| + |b_{j,t}(\eta^*_0-\eta_0)|$, and similarly, $|f_{\varepsilon^*_{j,t}|\mathcal{F}^*_{t-1},X_{j(-k),t}}\{(\beta-\beta^0_{jk})X^*_{jk,t}+h_1-h^0_{jk,1}\} - f_{\varepsilon_{j,t}|\mathcal{F}_{t-1},X_{j(-k),t}}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}\}| \lesssim |(\beta-\beta^0_{jk})a_{jk,t}(\xi^*_0-\xi_0)| + |b_{j,t}(\eta^*_0-\eta_0)|$.

Consider the LAD regression model, where $\psi_{jk}\{Z_{j,t},\beta,h\} = \{1/2 - \mathbf{1}(Y_{j,t} \leqslant X_{jk,t}\beta + h_1)\}(X_{jk,t}-h_2)$. It is not hard to see that $|\mathsf{E}[\{\psi_{jk}(Z^*_{j,t},\beta,h)-\psi_{jk}(Z_{j,t},\beta,h)\}|\mathcal{F}_{t-1},X_{j(-k),t}]| \lesssim |F_{\varepsilon^*_{j,t}|\mathcal{F}^*_{t-1},X_{j(-k),t}}\{(\beta-\beta^0_{jk})X^*_{jk,t}+h_1-h^0_{jk,1}\} - F_{\varepsilon_{j,t}|\mathcal{F}_{t-1},X_{j(-k),t}}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}\}| + |X^*_{jk,t}-X_{jk,t}| \lesssim |(\beta-\beta^0_{jk})a_{jk,t}(\xi^*_0-\xi_0)|+|b_{j,t}(\eta^*_0-\eta_0)|+|a_{jk,t}(\xi^*_0-\xi_0)|$. Assume $\sum_{t=0}^{\infty}|a_{jk,t}|<\infty$, $\sum_{t=0}^{\infty}|b_{j,t}|<\infty$. It follows that $\|\mathsf{E}\{\psi_{jk}(Z_{j,\cdot},\beta,h)|\mathcal{F}_{\cdot-1},X_{j(-k),\cdot}\}\|_{q,\varsigma}<\infty$. As we can see, the dependence adjusted norm is explicitly linked to the dependency structure of the underlying linear processes.

Moreover, the partial derivatives of $\mathsf{E}\{\psi_{jk}(Z_{j,t},\beta,h)|\mathcal{F}_{t-1},X_{j(-k),t}\}$ are given by: $-\mathsf{E}[f_{\varepsilon_j}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}\}X_{jk,t}(X_{jk,t}-h_2)|\mathcal{F}_{t-1},X_{j(-k),t}]$ (w.r.t. β), $-\mathsf{E}[f_{\varepsilon_j}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}\}(X_{jk,t}-h_2)|\mathcal{F}_{t-1},X_{j(-k),t}]$ (w.r.t. h_1), $-\mathsf{E}[1/2-F_{\varepsilon_j}\{(\beta-\beta^0_{jk})X_{jk,t}+h_1-h^0_{jk,1}\}|\mathcal{F}_{t-1},X_{j(-k),t}]$ (w.r.t. h_2). Then the dependence adjusted norm for the partial derivatives of
$\mathsf{E}\{\psi_{jk}(Z_{j,t},\beta,h)|\mathcal{F}_{t-1},X_{j(-k),t}\}$ are finite and can be verified similarly.

Theorem 5.8 (Uniform Bahadur Representation)**.** *Under conditions (A1), (A3) and (C1)-(C9), with probability* $1-o(1)$, *we have*

$$\max_{(j,k)\in G}|n^{1/2}\sigma^{-1}_{jk}(\widehat{\beta}_{jk}-\beta^0_{jk})+n^{-1/2}\sigma^{-1}_{jk}\phi^{-1}_{jk}\sum_{t=1}^{n}\psi^0_{jk,t}| = o(g_n^{-1}),\ \textit{as } n\to\infty, \tag{5.33}$$

where $\sigma^2_{jk} \overset{\text{def}}{=} \phi^{-2}_{jk}\omega_{jk}$, $\omega_{jk} \overset{\text{def}}{=} \mathsf{E}(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi^0_{jk,t})^2$.

The results in Theorem 5.8 imply the asymptotic normality of the proposed estimator by Algorithm 1 and 2 by applying central limit theorems and Gaussian Approximation.

Corollary 5.7. *Under conditions (A1), (A3), (C8) and (C9), for any* $(j,k)\in G$ *the estimators obtained by Algorithm 1 and 2 satisfy*

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

Theorem 5.9 (Uniform-Dimensional Central Limit Theorem)**.** *Under conditions (A1) and (A3), assume that* $\|\psi^0_{jk,\cdot}\|_{2,\varsigma}<\infty$, *we have*

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

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

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

$$\|\widetilde{\zeta}_{\cdot}\|_{q,\varsigma} \overset{\text{def}}{=} \sup_{m\geqslant 0}(m+1)^{\varsigma}\sum_{t=m}^{\infty}\||\widetilde{\zeta}_t-\widetilde{\zeta}^*_t|_{\infty}\|_q, \tag{5.34}$$

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

$$\begin{aligned}&\Phi^{\zeta}_{q,\varsigma} \overset{\text{def}}{=} \max_{(j,k)\in G}\|\zeta_{jk,\cdot}\|_{q,\varsigma},\ \Gamma^{\zeta}_{q,\varsigma} \overset{\text{def}}{=} \Big(\sum_{(j,k)\in G}\|\zeta_{jk,\cdot}\|^q_{q,\varsigma}\Big)^{1/q},\\ &\Theta^{\zeta}_{q,\varsigma} \overset{\text{def}}{=} \Gamma^{\zeta}_{q,\varsigma}\wedge\{\|\widetilde{\zeta}_{\cdot}\|_{q,\varsigma}(\log|G|)^{3/2}\}.\end{aligned} \tag{5.35}$$

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

(A7) i) (weak dependency case) Given $\Theta_{q,\varsigma} < \infty$ with $q \geqslant 2$ and $\varsigma > 1/2 - 1/q$, then $\Theta_{q,\varsigma} n^{1/q-1/2}\{\log(|G|n)\}^{3/2} \to 0$ and $L_1 \max(W_1^\zeta, W_2^\zeta) = \scriptstyle{\mathcal{O}}(1)\min(N_1^\zeta, N_2^\zeta)$.

ii) (strong dependency case) Given $0 < \varsigma < 1/2 - 1/q$, then $\Theta_{q,\varsigma}(\log|G|)^{1/2} = \scriptstyle{\mathcal{O}}(n^\varsigma)$ and $L_1 \max(W_1^\zeta, W_2^\zeta, W_3^\zeta) = \scriptstyle{\mathcal{O}}(1)\min(N_2^\zeta, N_3^\zeta)$.

Corollary 5.8 (Consistency of the Bootstrap Confidence Interval)**.** *Under (A7) and the same conditions as in Theorem 5.8, for each* $(j,k) \in G$ *assume that there exists a constant* $c > 0$ *such that* $\min_{(j,k)\in G} \mathrm{Var}(\zeta_{jk,t}) \geqslant c$, *with probability* $1 - \scriptstyle{\mathcal{O}}(1)$, *we have*

$$\sup_{\alpha\in(0,1)} |\,\mathrm{P}(\beta_{jk}^0 \in \widetilde{\mathrm{CI}}_{jk}(\alpha),\, \forall(j,k) \in G) - (1-\alpha)| = \scriptstyle{\mathcal{O}}(1), \quad \textit{as } n \to \infty, \tag{5.36}$$

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

Following Theorem 5.8, a joint confidence region and the corresponding confidence interval for each component can be constructed via a block bootstrap method. In particular, the bootstrap statistic are defined by $\frac{1}{\sqrt{n}}\sum_{i=1}^{l_n} e_i \sum_{l=(i-1)b_n+1}^{ib_n} \widehat{\zeta}_{jk,l}$, where e_i's are independent and identically distributed draws of standard normal random variables and are independent with respect to the data sample $(Z_{j,t})_{j=1}^J$.

Corollary 5.9 (Validity of Multiplier Bootstrap)**.** *Under the same conditions as in Theorem 5.8, 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* (A.18)*), we have*

$$\sup_{\alpha\in(0,1)} |\,\mathrm{P}(\beta_{jk}^0 \in \widetilde{\mathrm{CI}}_{jk}^*(\alpha),\, \forall(j,k) \in G) - (1-\alpha)| = \scriptstyle{\mathcal{O}}(1), \quad \textit{as } n \to \infty, \tag{5.37}$$

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

6 Simulation Study

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

6.1 Estimation with a Jointly Selected Penalty Level

Setting 1:

Consider the system of regression equations:

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

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

We take $n = 100$, # of bootstrap replications = 1000. We set $J, K = 50, 100$ and 150. The prediction norm $|\widehat{\beta}_j - \beta^0_j|_{j,pr}$ and the Euclidean norm $|\widehat{\beta}_j - \beta^0_j|_2$ ratios are presented in Table 6.1. The ratios measure the relative difference between the results using the penalty level determined from the equation-by-equation case and from the joint equation case (λ_j and λ are selected by the multiplier block bootstrap procedure introduced in Section 3). In particular, a ratio smaller than 1 indicates a better performance of the jointly selected penalty level.

	$J = K = 50$	$J = K = 100$	$J = K = 150$
	Prediction norm		
Mean	0.8915	0.8381	0.7869
Median	0.9141	0.8698	0.8404
	Euclidean norm		
Mean	0.8962	0.8478	0.7876
Median	0.8922	0.8513	0.8109

Table 6.1: Equation-by-equation to joint equation ratios of prediction norm and Euclidean norm (mean or median over equations). Results are averaged over 1000 simulations.

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

Setting 2:

Consider the Vector Autoregression (VAR) model of order 1:

$$Y_t = \Phi^0 Y_{t-1} + \varepsilon_t, \quad t = 1, \dots, n, \tag{6.2}$$

where $Y_t \in \mathbb{R}^{\mathrm{K}}$. The matrix Φ^0 is set to be a sparse matrix. In particular, it has a block diagonal structure where the blocks are 5×5 matrices with the same parameter ϕ in all blocks. We consider two cases where ϕ equals either 0.05 or 0.15 (in these two cases the largest eigenvalues of Φ^0 are 0.25 and 0.75, respectively). $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \mathbf{I}_K)$.

We take $n = 100$, # of bootstrap replications = 500, $K = 50, 100$ and 150 (in this case $J = K$). We choose $b_n = 4, 10, 20, 25$. The prediction norm $|\widehat{\Phi}_{j\cdot} - \Phi^0_{j\cdot}|_{j,pr}$ (where $\Phi_{j\cdot}$ denotes the jth row of Φ) and the Euclidean norm $|\widehat{\Phi}_{j\cdot} - \Phi^0_{j\cdot}|_2$, $j = 1, \dots, K$ ratios (results with the jointly estimated λ relative to using the single equation λ_j's, which are selected by the multiplier block bootstrap procedure in Section 3) are reported in Table 6.2.

		$\phi = 0.05$			$\phi = 0.15$		
		$K = 50$	$K = 100$	$K = 150$	$K = 50$	$K = 100$	$K = 150$
		Prediction norm					
$b_n = 4$	Mean	0.8813	0.8464	0.7584	1.0510	1.0381	1.0221
	Median	0.9686	0.9605	0.9501	1.0458	1.0351	1.0231
$b_n = 10$	Mean	0.8858	0.8404	0.7451	1.0812	1.0567	1.0353
	Median	0.9706	0.9636	0.9487	1.0692	1.0506	1.0359
$b_n = 20$	Mean	0.8915	0.8501	0.7471	1.1017	1.0645	1.0465
	Median	0.9702	0.9636	0.9491	1.0874	1.0596	1.0455
$b_n = 25$	Mean	0.8936	0.8491	0.7405	1.1040	1.0715	1.0412
	Median	0.9725	0.9676	0.9486	1.0915	1.0625	1.0395
		Euclidean norm					
$b_n = 4$	Mean	0.8362	0.7947	0.5663	0.9883	0.9630	0.9375
	Median	1	1	1	1.0000	1	1
$b_n = 10$	Mean	0.8411	0.7857	0.5595	0.9773	0.9506	0.9240
	Median	1	1	1	1.0000	1	1
$b_n = 20$	Mean	0.8485	0.7982	0.5721	0.9633	0.9396	0.9192
	Median	1	1	1	0.9996	1	1
$b_n = 25$	Mean	0.8540	0.7977	0.5315	0.9584	0.9376	0.9158
	Median	1	1	1	0.9997	1	1

Table 6.2: Equation-by-equation $\widehat{\lambda}_j$ to the jointly estimated tuning parameter ratios measured by prediction norm and Euclidean norm (mean or median over equations). Results are averaged over 1000 simulations.

This shows that the coefficient estimation performance measured by both the Euclidean norm and the prediction norm is in favor of the joint penalty level approach, especially in a higher dimension case with weaker dependency. More specifically, if there exists stronger dependency in the data, the coefficient estimation with larger block size gives lower errors (in terms of the Euclidean norm).

6.2 Simultaneous Inference

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

$$Y_{j,t} = d_{j,t}\alpha_j^0 + X_t^\top \beta_j^0 + \varepsilon_{j,t},\ d_{j,t} = X_t^\top \theta_j^0 + v_{j,t},\ t = 1, \ldots, n,\ j = 1, \ldots, J, \tag{6.3}$$

where $\alpha_j^0 = \alpha^0$ for all j. Also, $\beta_j^0, \theta_j^0 \in \mathrm{I\!R}^{\mathrm{K}}$ are assumed to be sparse. In particular, we divide the indices $1, \ldots, K$ evenly into blocks with a fixed block size 5, $\beta_{jk}^0 = 0.5/(k - \lfloor \frac{k}{5} \rfloor \times 5), \theta_{jk}^0 = 0.25(k - \lfloor \frac{k}{5} \rfloor \times 5)$ if k and j belong to the same block and 0 otherwise.

Covariates $X_t \in \mathrm{I\!R}^{\mathrm{K}}$ are generated from a VAR(1) process, where the coefficient matrix has a block diagonal structure (e.g., the blocks are 5×5 matrices with all entries in each block equal to 0.1) and the innovations of the VAR process follow $\overset{\text{i.i.d.}}{\sim}$ $\mathrm{N}(0, \mathbf{I}_K)$. For each j, $\varepsilon_{j,t}$ and $v_{j,t}$ are independently drawn from the AR(1) process with the autocorrelation coefficient as 0.5 and i.i.d. $\mathrm{N}(0, 1)$ innovations.

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 power and size performance of the confidence intervals constructed by the asymptotic distribution theory (4.3), block bootstrap (4.6) and the simultaneous confidence regions via the block bootstrap (4.8). The bootstrap statistics are computed based on 500 replications and we take the block size $b_n = 25$ because the numerical study conducted above suggests that larger block size is more favorable in the presence of stronger dependency. To investigate the empirical size and power performances, we generate a sequence of alternatives with either $H_A^j : \alpha_j^0 = a$ (individual inference) or $H_A : \alpha_1^0 = \cdots = \alpha_J^0 = a$ (simultaneous inference), where $a \geqslant 0$ uniformly lies in $[0, 1.5]$. Note that the case of $a = 0$ gives the size performance under the null hypothesis, while $a \geqslant 0$ illustrates the power results.

Figure 6.1 shows the empirical coverage probabilities, namely the average rejection rate of $H_0^j : \alpha_j^0 = 0$ over j for individual inference and the rejection rate of $H_0 : \alpha_1^0 = \cdots = \alpha_J^0 = 0$ for simultaneous inference under different settings of J and K. In particular, the size performances are reported in Table 6.3. The rejection rates are computed over 1000 simulation samples.

	$K = 100, J = 50$	$K = J = 100$	$K = 200, J = 100$	$K = J = 200$
Ind. Asym.	0.000	0.001	0.000	0.000
Ind. Boot.	0.034	0.045	0.038	0.061
Simult. Boot.	0.010	0.010	0.000	0.000

Table 6.3: Size performances with different K, J's, where we take average over j for the individual inference.

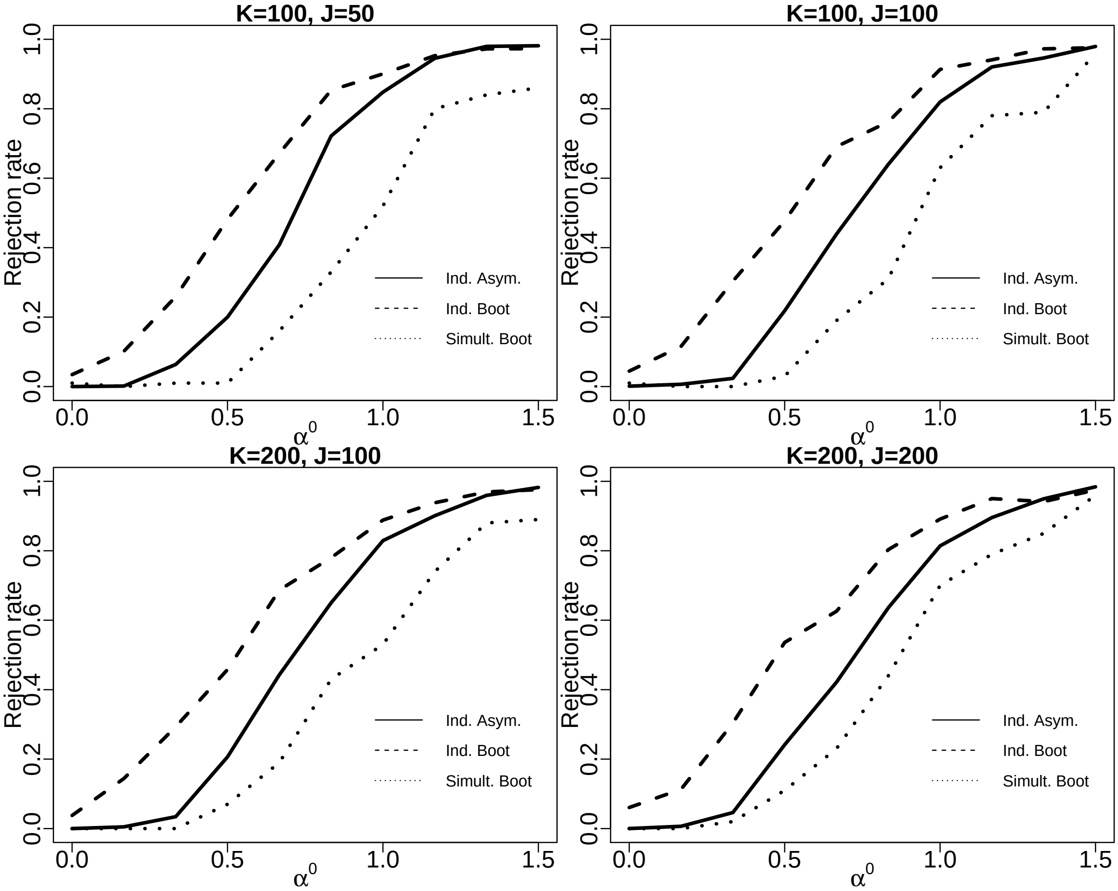


Figure 6.1: Average rejection rate of $H_0^j : \alpha_j^0 = 0$ over j for the individual inference (solid - asymptotics, dashed - bootstrap) and the rejection rate of $H_0 : \alpha_1^0 = \cdots = \alpha_J^0 = 0$ for simultaneous inference (dotted - bootstrap) under several true α^0 values (given the significance level = 0.05).

The results show that the size can be controlled under the significance level 0.05 and the rejection rate converges to 100% as the true α^0 values increase. In particular, for individual inference our proposed individual bootstrap approach provides a much more powerful performance compared to constructing the confidence intervals by asymptotic distributions. We observe that the simultaneous inference is more conservative than the individual inference and the results are robust with increasing dimensions.

7 Empirical Analysis: Textual Sentiment Spillover Effects

Financial markets are driven by information, and this is a well-known phenomenon among investors. More frequent news and availability of sentiment data allows study of the impact of firm-specific investor sentiment on market behavior such as stock returns, volatility and liquidity; see Baker and Wurgler, 2006; Tetlock, 2007, among others. Moreover, powerful statistical tools (e.g. LASSO-type estimators) are being used to model complex relationships among individuals. For example, Audrino and Tetereva (2017) analyze the influence of news on US and European companies by constructing a sparse predictive network via adaptive LASSO and related testing procedures. In this section the developed technology is applied to study textual sentiment spillover effects across individual stocks. This is different from the "equation-by-equation" analysis in Audrino and Tetereva (2017), since we build up a system of regression equations and implement the estimation and the inference of the network jointly.

7.1 Data Source

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

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

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 C.1 in Appendix C in the supplementary materials.

The daily log returns $R_{j,t}$ and log volatilities $\log(\sigma_{j,t}^2)$ 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:

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

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

7.2 Model Setting and Results

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

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

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

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

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

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

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

If for stock j, the sentiment variable of firm k is selected into the active set after the individual significance test i.e., the null hypothesis $H_0^{jk} : \beta_{jk} = 0$ is rejected under the block multiplier bootstrap procedure, block size $b_n = 25$, then we put a directional edge from k to j. As a result, we achieve a $0 - 1$ adjacency matrix describing the dependency network from sentiment variable to financial variable. Note that the diagonal elements in the matrix show the self-effect of stocks.

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

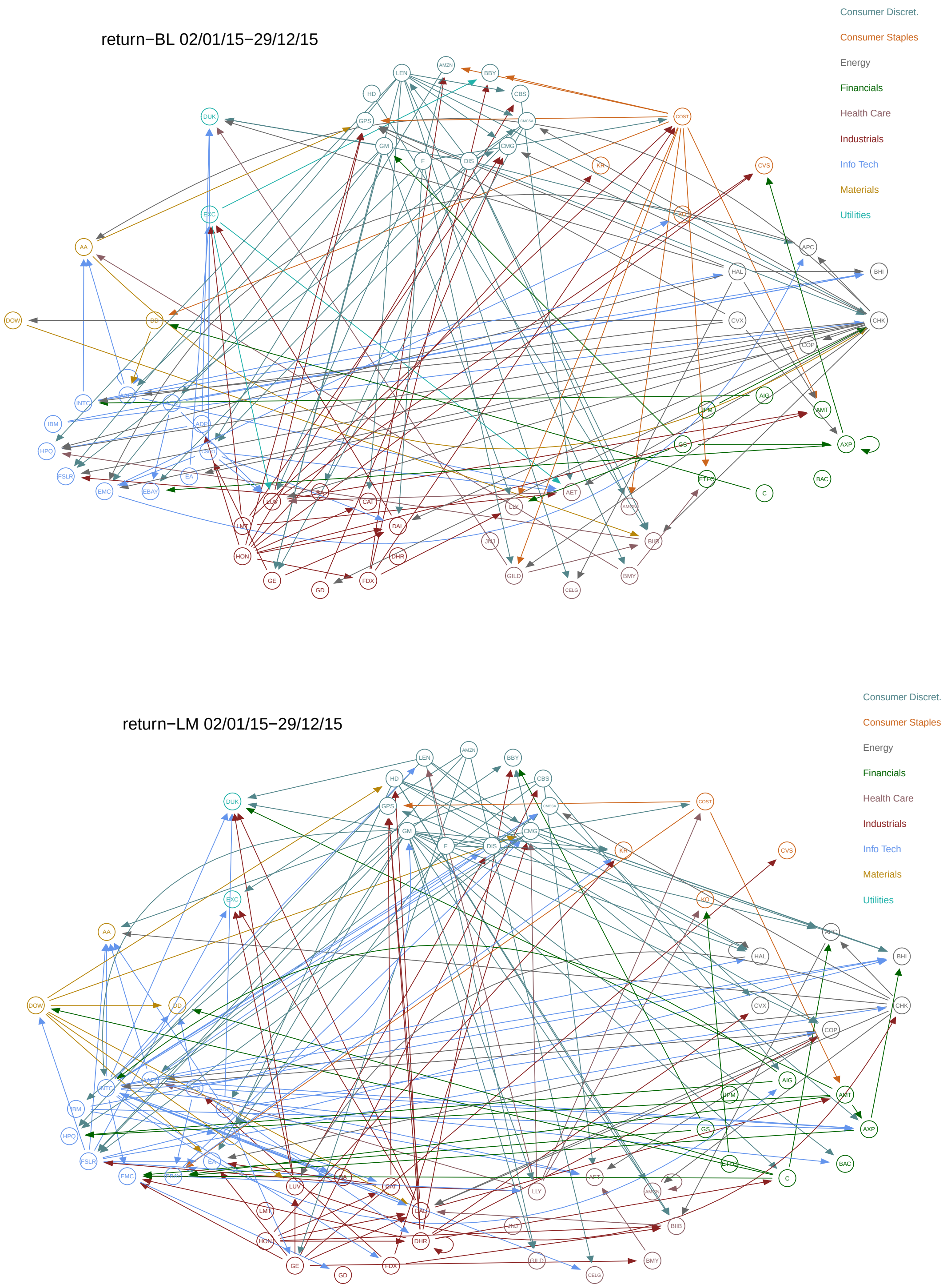


Figure 7.1: The dependency network among individual stocks from sentiment variables to return.

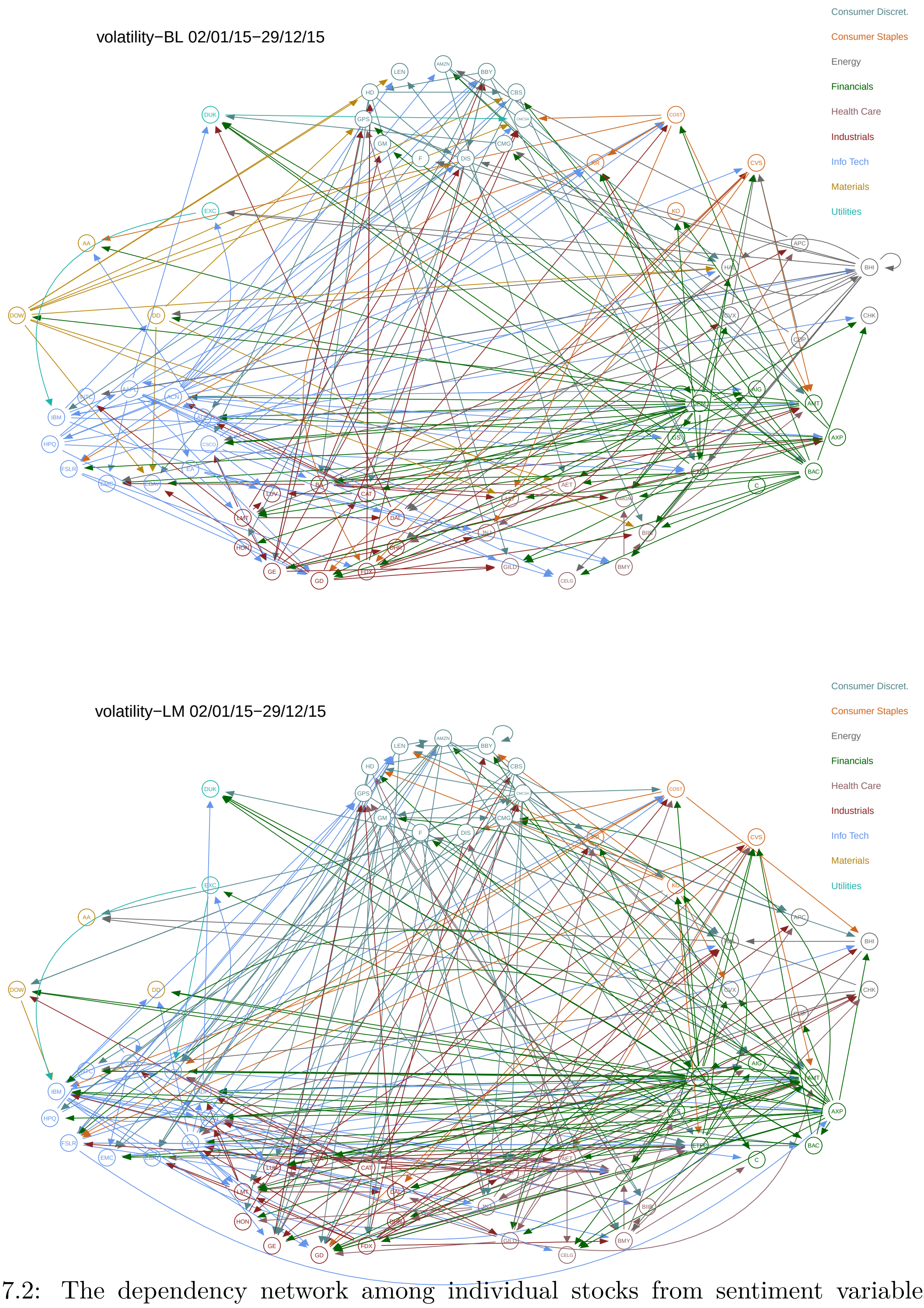
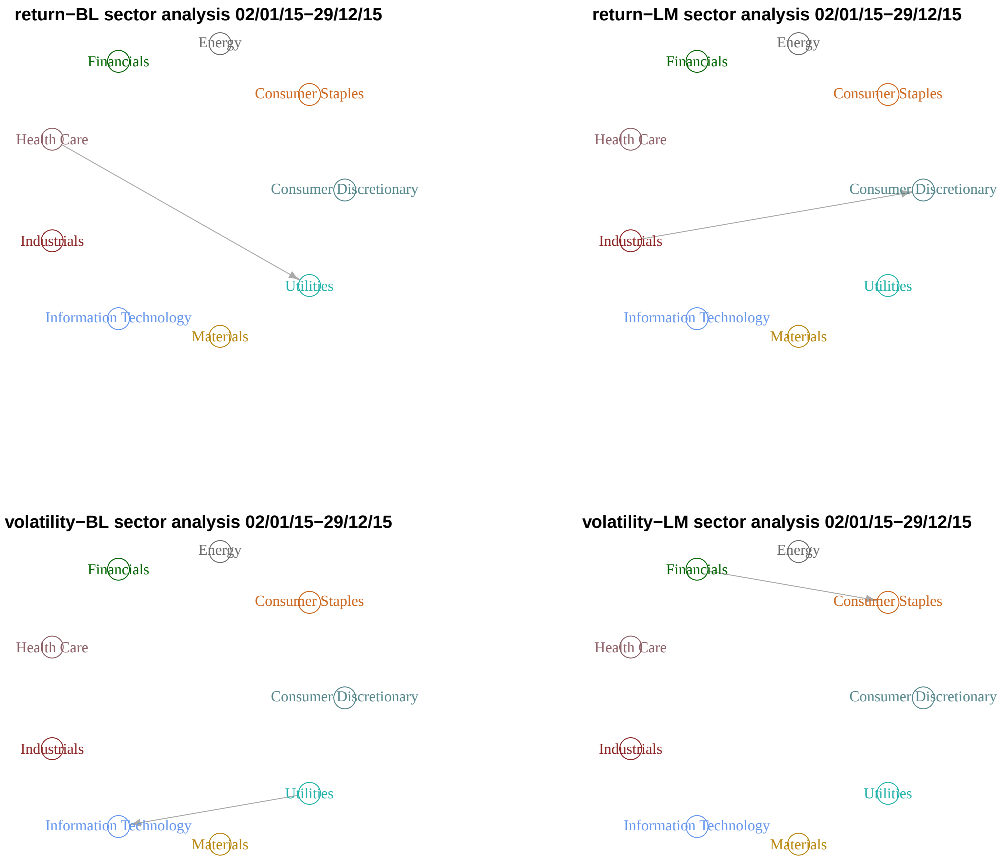


Figure 7.2: The dependency network among individual stocks from sentiment variables to volatility.

Figures 7.1-7.2 depict the dependency networks among individual stocks. Given that the time series of returns and volatility are scaled and centered before implementing the estimation procedure, we find even denser spillover effects in the volatility analysis. This indicates the stock volatility is more sensitive to sentiment than returns. Moreover, the relationships between sectors are also of interest. The simultaneous confidence region constructed via the bootstrap approach introduced in Section 4.2 may help us to detect whether the sentiment information from one sector has joint influence on the returns of the stocks in another sector. In particular, we look at the null hypothesis: $H_0^{S_1,S_2} : \beta_{jk} = 0,\ \forall j \in S_1,\ k \in S_2$, where S_1 and S_2 represent two groups of stocks that belong to two sectors, respectively. The conclusion that the sentiment from sector S_2 has a joint effect on the returns or volatility of sector S_1 can be drawn if the null hypothesis is rejected with the simultaneous confidence region (4.8) under the significance level $= 0.05$.


return−BL sector analysis 02/01/15−29/12/15

return−LM sector analysis 02/01/15−29/12/15

volatility−BL sector analysis 02/01/15−29/12/15

volatility−LM sector analysis 02/01/15−29/12/15

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

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

References

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

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

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

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

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

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

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

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

Belloni, A., Chernozhukov, V. and Kato, K. (2013). Uniform post selection inference for lad regression and other z-estimation problems, *arXiv preprint arXiv:1304.0282* .

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

Wu, W. B. (2011). Asymptotic theory for stationary processes, *Statistics and its Interface* **4**(2): 207–226.

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

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

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

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

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

Zhu, Y. and Bradic, J. (2017). Linear hypothesis testing in dense high-dimensional linear models, *Journal of the American Statistical Association* (accepted, doi:10.1080/01621459.2017.1356319).