Regularization for Spatial Panel Time Series Using the Adaptive LASSO

Clifford Lam*[1] and Pedro CL Souza†[2]

[1]Department of Statistics, London School of Economics and Political Science
[2]Department of Economics, London School of Economics and Political Science

Abstract

This paper proposes a model for estimating the underlying cross-sectional dependence structure of a large panel of time series. Technical difficulties meant past researchers usually assume the dependence structure of the data is known before further analysis. We propose to estimate such a structure by penalizing the elements in the spatial weight matrices, which are essential for specifying dependence structure in our model, using the adaptive LASSO proposed by Zou [2006]. Technical hurdles are overcome with a Nagaev-type inequality for dependent data. Non-asymptotic oracle inequalities, together with the asymptotic sign consistency of the estimators, are presented and proved when the dimension N of the time series can be larger than the sample size T. A block coordinate descent algorithm is introduced for numerical computations. A simulation experiment and a real data analysis are carried out to demonstrate its practical performance.

Key words and phrases. Spatial econometrics; adaptive LASSO; sign consistency; non-asymptotic oracle inequalities; cross-sectional dependence; spatial weight matrices.

*Clifford Lam is lecturer, Department of Statistics, London School of Economics. Email: C.Lam2@lse.ac.uk
†Pedro CL Souza is PhD, Department of Economics, London School of Economics. Email: p.souza@lse.ac.uk

1 Introduction

The study of spatial panel data is of increasing importance in econometrics and many other disciplines. As obtaining large panel of time series data becomes easier, more researchers look into these data as they provide valuable information on spatial-temporal dependence structure. Various models are proposed to study the cross-sectional dependence of variables, including fixed or random effects spatial lag (or spatial autoregressive) and spatial error models (see Elhorst [2003]). Spatial autoregressive models (SAR) can be seen as another formulation of a spatial error model (see Lesage and Pace [2009] for example).

One important feature of these models is the need for the specification of the spatial weight matrix, which is the key in quantifying the spatial lag structure in the panel time series data. Method of specification ranges from using prior expert knowledge (for example see Lesage and Polasek [2008]), to imposing special structures. For example, the contiguity structure has contagious regions having corresponding elements in the weight matrix set to one and zero otherwise (see Lesage and Pace [2009] for more details). The more general "distance metric" has elements corresponding to further away regions smaller than those that are closer together. Exact "distance" specification, however, is not universal. Bavaud [1998] suggested various specifications, including a distance decay model, and their implications and interpretations with theoretical supports. Anselin [2002] has also addressed the issue of weight matrix specification and interpretation.

In this paper, we study a more general form of spatial autoregressive model as detailed in section 2. In the terminology of Anselin [2002], we include both global and local spillover effects, through the terms $\mathbf{W}_1^*\mathbf{y}_t$ and $\mathbf{W}_2^*\mathbf{X}_t\boldsymbol{\beta}^*$ respectively in model (2.1). Few researchers attempted to estimate the weight matrices, including a well known paper by Pinkse et al. [2002]. They estimate a nonparametric smooth function $\widehat{g}(\cdot)$ assuming normality of data, and the (i,j)-th element of the weight matrix $\mathbf{W}_1^*$ is estimated as $\widehat{g}(d_{ij})$, where d_{ij} is a distance measure specified by the user. In our paper, we focus on estimating the spatial weight matrices themselves, which are assumed to be sparse: having a lot of zero entries. There is no need to specify a distance measure for our method as long as the true weight matrices are sparse. We provided non-asymptotic bounds on various estimated quantities on a set with probability approaching 1 asymptotically (see Lemma 2 for example). We demonstrate that sparsity is a common endeavor with a structural equation model in Example 1 in section 3.1.

The aims in estimating the weight matrices are twofold. First, it is not always clear what exactly the spatial dependence structure is for the panel data. Even with expert knowledge of what the spatial matrices should look like, estimating them from data may reveal dependence structures that our assumptions can miss out. Presenting the estimated weight matrix as a network connecting the components of the panel time series provide a visual tool for deeper understanding of cross-sectional dependence structure. Second, as presented previously, there are no universal rules in specifying a spatial weight matrix. We quote a part of the criticism summarized in Arbia and Fingleton [2008], "... arbitrary nature of weight matrix... are not the results obtained conditional on somewhat arbitrary decisions taken about its structure?" Although debate is still on about the sensitivity of results towards the specification of spatial weight matrices, this paper provides partly a solution to the criticism and potential sensitivity towards "arbitrary" specification of these matrices if they themselves can be estimated from the data as well. In fact in Lemma 2, we have specified how the error upper bound for the estimation of $\boldsymbol{\beta}^*$ in model (2.1) is related to the error of the estimated/assumed weight matrices. This result sheds some lights on the potential seriousness of wrongly specifying the weight matrices.

The rest of the paper is organized as follows. In section 2, we introduce the spatial autoregressive model considered, with examples. Section 3 presents the model in a compact form and introduces the minimization problems for obtaining the estimators of the sparse weight matrices. These estimators are analyzed in section 4 using a relatively new concept of time dependence in time series data, with non-asymptotic oracle inequalities and rates of convergence spelt out, as well as asymptotic sign consistency presented. Section 5 discusses the computational issue of our estimators, and presented a block coordinate descent algorithm as a solution. Section 6 presents our extensive simulation results and real data analysis. The paper concludes with section 7, outlining our main contributions and some future research directions. Finally all technical proofs of the theorems in section 4 are presented in section 8.

2 The Model

The spatial autoregressive model we consider is

$$\mathbf{y}_t = \mathbf{W}_1^* \mathbf{y}_t + \mathbf{W}_2^* \mathbf{X}_t \boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{2.1}$$

where $\mathbf{y}_t$ is an $\mathbf{N} \times 1$ vector of dependent time series variables, $\mathbf{W}_j^*$ for $j = 1, 2$ are the $N \times N$ weight matrices to be estimated, $\mathbf{X}_t$ is an $N \times K$ matrix of centered exogenous variables at time t, $\boldsymbol{\beta}^*$ (with the first element fixed at 1 for identifiability) is a vector of K regression parameters for the exogenous variables, and finally $\{\boldsymbol{\epsilon}_t\}$ is an innovation process with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}_\epsilon$, and is independent of $\{\mathbf{X}_t\}$. Both $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are assumed second order stationary. The matrix $\boldsymbol{\Sigma}_\epsilon$ is assumed to have uniformly bounded entries as $N, T \to \infty$. Detailed assumptions A1- A8 can be found in section 4.

The weight matrix $\mathbf{W}_1^*$ has 0 on the main diagonal, and we assume that there exists a constant $\eta < 1$ such that $\left\|\mathbf{W}_1^*\right\|_\infty < \eta < 1$, i.e. $\max_{1 \le i \le N} \sum_{j=1}^{N} |w_{1,ij}^*| < \eta < 1$ uniformly as $N, T \to \infty$, where $w_{1,ij}^*$ is the (i, j)-th element of $\mathbf{W}_1^*$. This regularity condition ensures $\mathbf{y}_t$ has a reduced form

$$\mathbf{y}_t = \boldsymbol{\Pi}_1^* \mathbf{W}_2^* \mathbf{X}_t \boldsymbol{\beta}^* + \boldsymbol{\Pi}_1^* \boldsymbol{\epsilon}_t, \quad \boldsymbol{\Pi}_1^* = (\mathbf{I}_N - \mathbf{W}_1^*)^{-1}, \tag{2.2}$$

with innovations in $\Pi_1^* \boldsymbol{\epsilon}_t$ having finite variances, where $\mathbf{I}_N$ is the identity matrix of size N. See also Corrado and Fingleton [2011] or Kapoor et al. [2007] for a similar row sum regularity condition for the weight matrices in a slightly different spatial model specification. Hence each component y_{tj} is a weighted linear combination of the other components in $\mathbf{y}_t$. If $w_{1,ij}^* \neq 0$, it means that y_{ti} depends on y_{tj} explicitly. An analysis of the links among financial markets is given in section 6 to illustrate the use of such a model.

The weight matrix $\mathbf{W}_2^*$ has 1 on the main diagonal, with the same row sum condition as $\mathbf{W}_1^*$ excluding the diagonal entries. Hence while each component y_{tj} has the same regression coefficients $\boldsymbol{\beta}^*$ for their respective exogenous variables $\mathbf{x}_{t,j}^{\mathrm{T}}$ (the j-th row of $\mathbf{X}_t$), model (2.1) gives flexibility through $\mathbf{W}_2^*$ by allowing each y_{tj} to depend on a linear combination of exogenous variables for other components as well. This is also related to the local spatial spillover effects. For more details please refer to Anselin [2002]. See section 3.1 for an illustrative example with covariates.

Remark 1. The spatial error model with spatial autoregressive-moving average (ARMA) error can

be defined by (see also Yao and Brockwell [2006])

$$\begin{cases} \mathbf{y}_t &= \mathbf{X}_t\boldsymbol{\beta} + \mathbf{u}_t, \\ \mathbf{u}_t &= \rho\mathbf{W}\mathbf{u}_t + (\mathbf{I}_N + \lambda\mathbf{W}')\mathbf{v}_t, \end{cases} \quad \text{implying } \mathbf{y}_t = \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} - \rho\mathbf{W}\mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t = (\mathbf{I}_N + \lambda\mathbf{W}')\mathbf{v}_t$. Model (2.1) entails this spatial ARMA error model, by setting $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, $\mathbf{W}_1^* = \rho\mathbf{W}$, $\mathbf{W}_2^* = \mathbf{I}_N - \rho\mathbf{W}$, and $\boldsymbol{\Sigma}_\epsilon = (\mathbf{I}_N + \lambda\mathbf{W}')\text{var}(\mathbf{v}_t)(\mathbf{I}_N + \lambda(\mathbf{W}')^{\mathrm{T}})$. From assumption A4 in section 4.1, as long as the spatial autocovariance between $x_{t,jk}$ and $x_{t,j'k}$ for $j \neq j'$ decays fast enough as $|j-j'|$ gets larger, the correlation matrix for $\boldsymbol{\epsilon}_t$ can have a general structure, including that of a spatial moving-average structure as above.

3 Sparse Estimation of the Weight Matrices

The weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are assumed to be sparse. We give an example with covariates to illustrate that sparseness of weight matrices is a common endeavor.

3.1 Example 1

Irwin and Geoghegan [2001] considered an example of modeling jointly the population and property tax rate in different counties, assuming that households migration pattern is determined by local tax rate. They gave an example of a very much simplified structural equation model for jointly modeling the two:

$$\begin{aligned} \text{POP}_{it} &= w_1\text{TAX}_{it} + \beta_1\text{EMP}_{it} + \beta_2\text{PUBS}_{it} + \epsilon_{1it}, \\ \text{TAX}_{it} &= w_2\text{POP}_{it} + \gamma_1\text{PUBS}_{it} + \gamma_2\text{INC}_{it} + \epsilon_{2it}, \end{aligned}$$

where POP = total population, TAX = property tax rate, EMP = employment level, PUBS = measure of the quantity and quality of public services, and INC = per capita income of households. The index i represents measurements at county i, while the index t represents period t. If we write $\mathbf{y}_t = (\text{POP}_{1t}, \ldots, \text{POP}_{Nt}, \text{TAX}_{1t}, \ldots, \text{TAX}_{Nt})^{\mathrm{T}}$ where N=number of counties, the model can be written as $\mathbf{y}_t = \mathbf{W}_1^*\mathbf{y}_t + \mathbf{W}_2^*\mathbf{X}_t\boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t$, where

$$\mathbf{X}_t = \begin{pmatrix} \text{EMP}_{1t} & \text{PUBS}_{1t} & \text{INC}_{1t} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{EMP}_{Nt} & \text{PUBS}_{Nt} & \text{INC}_{Nt} & 0 & 0 & 0 \\ 0 & 0 & 0 & \text{EMP}_{1t} & \text{PUBS}_{1t} & \text{INC}_{1t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \text{EMP}_{1t} & \text{PUBS}_{Nt} & \text{INC}_{Nt} \end{pmatrix}, \quad \boldsymbol{\beta}^* = \begin{pmatrix} \beta_1 \\ \beta_2 \\ 0 \\ 0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix},$$

$$\mathbf{W}_1^* = \begin{pmatrix} \mathbf{0} & w_1\mathbf{I}_N \\ w_2\mathbf{I}_N & \mathbf{0} \end{pmatrix}, \quad \mathbf{W}_2^* = \mathbf{I}_{2N}, \quad \boldsymbol{\epsilon}_t = (\epsilon_{11t}, \ldots, \epsilon_{1Nt}, \epsilon_{21t}, \ldots, \epsilon_{2Nt})^{\mathrm{T}}.$$

Thus both the weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are very sparse in this model. Rather than fixing the weight matrices, our sparse estimation of the weight matrices gives flexibility on the network structure between the TAX and POP variables.

For a low dimensional model like this example, a reduced form model can be calculated like that in (2.2) and we can consistently estimate the parameters from the reduced form model. We can then try to recover the parameters $w_1, w_2, \beta_1, \beta_2, \gamma_1$ and γ_2 from the reduced form model parameters. This is also done in Irwin and Geoghegan [2001] for this particular example. However, for higher dimensional model where the weight matrices are our target, the problem can become intractable, and we in general need the decay assumption A2 in section 4.1 for asymptotic sign consistency for all the estimated entries in the weight matrix. See example 2 in section 4.2 as well.

Penalization has become a well-known tool for estimating a sparse vector/matrix over the past two decades. In this paper, we employ the adaptive LASSO developed in Zou [2006] for penalizing the elements in the matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, resulting in the minimization problem (with $\|\cdot\|$ being the usual L_2-norm)

$$\begin{aligned}&\min_{\mathbf{W}_1,\mathbf{W}_2,\boldsymbol{\beta}} \sum_{t=1}^{T} \left\|\mathbf{y}_t - \mathbf{W}_1\mathbf{y}_t - \mathbf{W}_2\mathbf{X}_t\boldsymbol{\beta}\right\|^2 + \gamma_T \sum_{i,j}(v_{1,ij}|w_{1,ij}| + v_{2,ij}|w_{2,ij}|),\\ &\text{subj. to } \sum_{j\neq i}|w_{1,ij}|, \sum_{j\neq i}|w_{2,ij}| < 1,\end{aligned}$$

where γ_T is a tuning parameter with rate given in Theorem 2 in section 4.3, and $v_{r,ij} = 1/|\widetilde{w}_{r,ij}|^k$ for $r = 1, 2$ and some integer $k \geq 1$, with $\widetilde{w}_{r,ij}$ being the solutions of the above minimization problem with all $v_{r,ij}$ set to 1. The $\widetilde{w}_{r,ij}$'s thus represent the LASSO solutions (see e.g. Zhao and Yu [2006]) with constraints. The $v_{r,ij}$ becomes the weight of penalization. The larger the magnitude of $\widetilde{w}_{r,ij}$, the smaller $v_{r,ij}$ becomes, and vice versa. This is a sensible weighting scheme since a larger $\widetilde{w}_{r,ij}$ means $w^*_{r,ij}$ is less likely to be zero, and hence should be penalized less to reduce estimation bias, and vice versa.

The above penalization problem is cumbersome to write and makes presentation and proofs of theorems difficult. Hence we rewrite model (2.1) as a more familiar regression type model:

$$\begin{aligned}\mathbf{y} &= \mathbf{Z}\boldsymbol{\xi}^*_1 + \mathbf{X}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*_2 + \boldsymbol{\epsilon}\\ &= \mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^* + \boldsymbol{\epsilon},\end{aligned} \tag{3.1}$$

where $\mathbf{y} = \text{vec}\{(\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\text{T}}\}$, $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\text{T}}$, $\mathbf{X}_{\boldsymbol{\beta}^*} = \mathbf{I}_N \otimes \{(\mathbf{I}_T \otimes \boldsymbol{\beta}^{*\text{T}})(\mathbf{X}_1, \ldots, \mathbf{X}_T)^{\text{T}}\}$, $\boldsymbol{\xi}^*_j = \text{vec}(\mathbf{W}^{*\text{T}}_j)$ for $j = 1, 2$, and $\boldsymbol{\epsilon} = \text{vec}\{(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^{\text{T}}\}$. Here $\otimes$ represents the Kronecker product, and the vec operator stacks the columns of a matrix into a single vector, starting from the first column. Defining $\mathbf{M}_{\boldsymbol{\beta}^*} = (\mathbf{Z}, \mathbf{X}_{\boldsymbol{\beta}^*})$ as the "design matrix" and $\boldsymbol{\xi}^* = (\boldsymbol{\xi}^{*\text{T}}_1, \boldsymbol{\xi}^{*\text{T}}_2)^{\text{T}}$ as the true "regression parameter", model (3.1) looks like a typical linear model, except that the design matrix $\mathbf{M}_{\boldsymbol{\beta}^*}$ is dependent on $\mathbf{y}$ as well.

With model (3.1), we can find the LASSO solutions by solving

$$\begin{aligned}&(\widetilde{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\xi},\boldsymbol{\beta}} \frac{1}{2T}\left\|\mathbf{y} - \mathbf{M}_{\boldsymbol{\beta}}\boldsymbol{\xi}\right\|^2 + \gamma_T\left\|\boldsymbol{\xi}\right\|_1,\\ &\text{subj. to } \sum_{j\neq i}|w_{1,ij}|, \sum_{j\neq i}|w_{2,ij}| < 1,\end{aligned} \tag{3.2}$$

where $\|\cdot\|_1$ represents the L_1-norm, and the definitions of $\mathbf{M}_{\boldsymbol{\beta}}$ and $\boldsymbol{\xi}$ are parallel to those in model (3.1).

The adaptive LASSO solutions are then

$$
\begin{aligned}
(\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\beta}}) &= \arg\min_{\boldsymbol{\xi},\boldsymbol{\beta}} \frac{1}{2T}\big\|\mathbf{y} - \mathbf{M}_{\boldsymbol{\beta}}\boldsymbol{\xi}\big\|^2 + \gamma_T \mathbf{v}^{\mathrm{T}}|\boldsymbol{\xi}|, \\
&\text{subj. to } \sum_{j\neq i}|w_{1,ij}|, \sum_{j\neq i}|w_{2,ij}| < 1,
\end{aligned}
\tag{3.3}
$$

where $|\boldsymbol{\xi}| = (|\xi_1|, \ldots, |\xi_{2N^2}|)^{\mathrm{T}}$ and $\mathbf{v} = (|\widetilde{\xi}_1|^{-k}, \ldots, |\widetilde{\xi}_{2N^2}|^{-k})^{\mathrm{T}}$. A general block coordinate descent algorithm is introduced in section 5 to carry out the minimization.

4 Properties of LASSO and adaptive LASSO Estimators

An ideal estimator for a weight matrix is one that recovers the correct locations of zeros and non-zeros in a sparse matrix, along with their correct magnitudes. Corollary 4 and Theorem 5 tell us that under certain conditions such estimators for $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are possible with high probability (as stated in Theorem 1), with explicit rates of convergence given.

In this paper we assume that the processes for the covariates $\{\mathbf{x}_t\} = \{\mathrm{vec}(\mathbf{X}_t)\}$ and for the noise $\{\boldsymbol{\epsilon}_t\}$ are defined by

$$
\mathbf{x}_t = \mathbf{f}(\mathcal{F}_t), \quad \boldsymbol{\epsilon}_t = \mathbf{g}(\mathcal{G}_t), \tag{4.1}
$$

where $\mathbf{f}(\mathcal{F}_t) = (f_1(\mathcal{F}_t), \ldots, f_{NK}(\mathcal{F}_t))^{\mathrm{T}}$ and $\mathbf{g}(\mathcal{G}_t) = (g_1(\mathcal{G}_t), \ldots, g_N(\mathcal{G}_t))^{\mathrm{T}}$ are both vectors of measurable functions defined on the real line. The shift processes $\mathcal{F}_t = (\ldots, \mathbf{e}_{x,t-1}, \mathbf{e}_{x,t})$ and $\mathcal{G}_t = (\ldots, \mathbf{e}_{\epsilon,t-1}, \mathbf{e}_{\epsilon,t})$ are defined by independent and identically distributed (i.i.d.) processes $\{\mathbf{e}_{x,t}\}$ and $\{\mathbf{e}_{\epsilon,t}\}$, and they are independent of each other. Hence $\{\mathbf{x}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are assumed independent. The representation (4.1) is used in Wu [2011] and provides a very general framework for stationary ergodic processes. See Wu [2011] for some examples as well.

For measuring dependence, instead of using traditional measures, like mixing conditions for time series, we use the functional dependence measure introduced in Wu [2005]. This measure lays the framework for applying a Nagaev-type inequality for obtaining the results of our theorems to be presented later. For the time series $\{\mathbf{x}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ in (4.1), define for $a > 0$,

$$
\begin{aligned}
\theta^x_{t,a,j} &= \big\|x_{tj} - x'_{tj}\big\|_a = (E|x_{tj} - x'_{tj}|^a)^{1/a}, \\
\theta^\epsilon_{t,a,\ell} &= \big\|\epsilon_{t\ell} - \epsilon'_{t\ell}\big\|_a = (E|\epsilon_{t\ell} - \epsilon'_{t\ell}|^a)^{1/a},
\end{aligned}
\tag{4.2}
$$

where $j = 1, \ldots, NK$, $\ell = 1, \ldots, N$, and $x'_{tj} = f_j(\mathcal{F}'_t)$, $\mathcal{F}'_t = (\ldots, \mathbf{e}_{x,-1}, \mathbf{e}'_{x,0}, \mathbf{e}_{x,1}, \ldots, \mathbf{e}_{x,t})$, with $\mathbf{e}'_{x,0}$ independent of all other $\mathbf{e}_{x,j}$'s. Hence x'_{tj} is a coupled version of x_{tj} with $\mathbf{e}_{x,0}$ replaced by an i.i.d. copy $\mathbf{e}'_{x,0}$. Finally, we have similar definitions for $\epsilon'_{t\ell}$. Such a definition of "physical" or functional dependence of time series on past "inputs" is used in various papers, for example in Shao [2010] and Zhou [2010].

There are no direct relationships between the usual mixing conditions and this "physical" functional dependence measure. But this measure is easier to handle mathematically and leads to simpler and stronger proofs in our paper, through the Nagaev-type inequality in Lemma 1. Moreover, many well-known processes are not strong mixing, yet can be handled by using the dependence measure (4.2), like the Bernoulli shift process in Andrews [1984].

4.1 Main assumptions and notations

With these definitions in place, we state the main assumptions in the paper. Note that $\|\mathbf{A}\|_\infty = \max_i \sum_{j\geq 1} |A_{ij}|$ for a matrix $\mathbf{A}$.

A1. The entries in the weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are constants as $N, T \to \infty$, on top of the row sum conditions introduced after model (2.1) in section 2.

A2. There exists a constant σ_0^2 such that $\text{var}(\epsilon_{tj}) = \sigma_{\epsilon,j}^2 \leq \delta_T \sigma_0^2$ for all $j = 1, \ldots, N$, with $\delta_T \to 0$ as $T \to \infty$.

A3. Both $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are mean $\mathbf{0}$ second-order stationary, and $\boldsymbol{\epsilon}_t$ is independent of $\mathbf{X}_s$ for each $s \leq t$.

A4. Let $\mathbf{X}_{t,k}$ be the k-th column of $\mathbf{X}_t$, $k = 1, \ldots, K$. Define $\boldsymbol{\zeta}_t = \boldsymbol{\epsilon}_t / \delta_T^{1/2}$. Write $\mathbf{X}_{t,k} = \boldsymbol{\Sigma}_{xk}^{1/2} \mathbf{X}_{t,k}^*$ and $\boldsymbol{\zeta}_t = \boldsymbol{\Sigma}_\zeta^{1/2} \boldsymbol{\zeta}_t^*$, where $\boldsymbol{\Sigma}_{xk}$ and $\boldsymbol{\Sigma}_\zeta$ are covariance matrices for $\mathbf{X}_{t,k}$ and $\boldsymbol{\zeta}_t$ respectively. We assume the elements in $\boldsymbol{\Sigma}_{xk}, \boldsymbol{\Sigma}_\zeta$ are all less than $\sigma_{\max}^2 < \infty$ uniformly as $N, T \to \infty$.

Also, either $\|\boldsymbol{\Sigma}_{xk}^{1/2}\|_\infty \leq S_x < \infty$ uniformly as $N, T \to \infty$, with $\{X_{t,jk}^*\}_{1\leq j\leq N}$ being a martingale difference with respect to the filtration generated by $(X_{t,1k}^*, \ldots, X_{t,jk}^*)$; or, $\|\boldsymbol{\Sigma}_\zeta^{1/2}\|_\infty \leq S_\zeta < \infty$ uniformly as $N, T \to \infty$, with $\{\zeta_{t,j}^*\}_{1\leq j\leq N}$ being a martingale difference with respect to the filtration generated by $(\zeta_{t,1}^*, \ldots, \zeta_{t,j}^*)$.

A5. The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is satisfied for $X_{t,jk}, X_{t,jk}^*, \zeta_{t,j}$ and $\zeta_{t,j}^*$ by the same positive constants D_1, D_2 and q.

A6. Define $\Theta_{m,a}^x = \sum_{t=m}^\infty \max_{1\leq j\leq NK} \theta_{t,a,j}^x$ and $\Theta_{m,a}^\zeta = \sum_{t=m}^\infty \max_{1\leq j\leq N} \theta_{t,a,j}^\zeta$, where $\theta_{t,a,j}^\zeta = \theta_{t,a,j}^\epsilon / \delta_T^{1/2}$. Then we assume $\Theta_{m,2w}^x, \Theta_{m,2w}^\zeta \leq C m^{-\alpha}$ for some $w > 2$, with $\alpha > 0$ and $C > 0$ being constants that can depend on w. These dependence measure assumptions also hold for $\boldsymbol{\zeta}_t^*$ and $\mathbf{X}_{t,k}^*$ for each $k \leq K$ in assumption A4.

A7. Let $\lambda_{\min}(M)$ be the minimum eigenvalue of a square matrix M. Then $\lambda_{\min}(E(\mathbf{x}_t \mathbf{x}_t^{\mathrm{T}})) > u > 0$ uniformly for some constant u as $N, T \to \infty$.

Assumption A1 can be relaxed, so that the weights in $\mathbf{W}_i^*$ can be decaying at a certain rate, at the expense of lengthier proofs. Assumption A2 is needed in general. Otherwise, as demonstrated numerically in section 6, the estimators for the weight matrices will perform badly even if T grows larger and N stays fixed. See example 2 in section 4.2 as well for a simple illustration, and a remark therein about estimating the reduced form model (2.2) instead.

Assumption A3 requires only that $\boldsymbol{\epsilon}_t$ to be independent of $\mathbf{X}_t$, allowing the covariates to be potentially the past values of $\mathbf{y}_t$. If $\mathbf{X}_t = (\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-d}, \mathbf{z}_t)$ where $\mathbf{z}_t$ contains exogenous covariates, the term $\mathbf{W}_2^* \mathbf{X}_t \boldsymbol{\beta}^* = \sum_{j=1}^d \beta_j^* \mathbf{W}_2^* \mathbf{y}_{t-j} + \mathbf{W}_2^* \mathbf{z}_t \boldsymbol{\beta}_2^*$, where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*, \boldsymbol{\beta}_2^{*T})^T$. Hence there is a vector autoregressive part with coefficient matrices $\beta_j \mathbf{W}_2^*$. The reduced form model for $\mathbf{y}_t$ is then

$$\mathbf{y}_t = \left(\mathbf{I}_N - \boldsymbol{\Pi}_1^* \sum_{j=1}^d \beta_j^* \mathbf{W}_2^* B \right)^{-1} \boldsymbol{\Pi}_1^* (\mathbf{W}_2^* \mathbf{z}_t \boldsymbol{\beta}_2^* + \boldsymbol{\epsilon}_t), \tag{4.3}$$

where $\mathbf{\Pi}_1^*$ is defined in (2.2), and B is the backward shift operator. For the inverse operator above to be defined (i.e. the system is stationary), we need

$$\det\left(\mathbf{I}_N - \mathbf{\Pi}_1^* \sum_{j=1}^{d} \beta_j^* \mathbf{W}_2^* z^j\right) \neq 0 \quad \text{for } |z| \leq 1,$$

which impose constraints on $\boldsymbol{\beta}^*$ as well. Allowing past values as covariates extends the applicability of the model, since example 2 in section 4.2 demonstrates that covariates have to be included for sign consistent estimation.

The uniform boundedness assumption in A4 for elements of $\mathbf{\Sigma}_{xk}$ and $\mathbf{\Sigma}_\zeta$ is a direct consequence of the tail assumption in A5. We assume this for notational convenience only. The other half of assumption A4 says that either the cross-correlations between more "distant" components for the k-th covariate $\mathbf{X}_{t,k}$ are getting smaller quick enough, or this happens for the components in the noise $\boldsymbol{\epsilon}_t$. The settings in (4.1) and (4.2) allows us to assume either $\{X_{t,jk}^*\}_j$ or $\{\zeta_{t,j}^*\}_j$ is a martingale difference, which is weaker than assuming that as an independent sequence.

Assumption A5 is a relaxation to normality, allowing sub-gaussian or sub-exponential tails for the concerned random variables. Together with A6, they allow for an application of the Nagaev-type inequality in Lemma 1 for our results. There are many examples of time series where A6 is satisfied. See Chen et al. [2013] for examples in stationary Markov Chains and stationary linear processes. Hence in particular we are allowing the noise series to have weak serial correlation. Finally, assumption A7 is needed for the convergence of $\widetilde{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}^*$. This is a mild condition and is satisfied in particular if all $\mathbf{\Sigma}_{xk}$ have their smallest eigenvalues uniformly bounded away from 0, and the cross covariance between the $\text{cov}(\mathbf{X}_{t,k_1}, \mathbf{X}_{t,k_2})$ is not too strong for all $1 \leq k_1 \neq k_2 \leq K$.

4.2 Example 2

We demonstrate that the decay assumption A2 is needed in general for estimating the weight matrices. In fact this condition is closely related to the conditions of the proximity theorem in Wold [1953] where the variance of the disturbance are small for negligible bias.

Consider $N = 3$, and the model $\mathbf{y}_t = \mathbf{W}\mathbf{y}_t + \mathbf{X}_t\beta + \boldsymbol{\epsilon}_t$, where $\mathbf{X}_t$ is a vector of covariates with mean 0, and denote $\sigma_{\epsilon,j}^2 = \text{var}(\epsilon_{t,j})$, $\sigma_{X,j}^2 = \text{var}(X_{t,j})$. Suppose we know $w_{13} = w_{23} = w_{31} = w_{32} = 0$ and $\beta = 1$, so that essentially the model becomes

$$\begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} = \begin{pmatrix} 0 & w_{12} \\ w_{21} & 0 \end{pmatrix} \begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} + \begin{pmatrix} X_{t1} \\ X_{t2} \end{pmatrix} + \begin{pmatrix} \epsilon_{t1} \\ \epsilon_{t2} \end{pmatrix}, \quad y_{t3} = X_{t3} + \epsilon_{t3}.$$

With $w_{12}, w_{21} < 1$, a simple inversion results in

$$y_{t1} = \frac{w_{12}(\epsilon_{t2} + X_{t2}) + \epsilon_{t1} + X_{t1}}{1 - w_{12}w_{21}}, \quad y_{t2} = \frac{w_{21}(\epsilon_{t1} + X_{t1}) + \epsilon_{t2} + X_{t2}}{1 - w_{12}w_{21}}.$$

The least square estimator for w_{12} is

$$\widehat{w}_{12} = \sum_{t=1}^{T} y_{t2}(y_{t1} - X_{t1}) / \sum_{t=1}^{T} y_{t2}^2 = w_{12} + \sum_{t=1}^{T} y_{t2}\epsilon_{t1} / \sum_{t=1}^{T} y_{t2}^2.$$

Assume proper convergence of all relevant quantities, and that $\mathrm{cov}(X_{t1}, X_{t2}) = \mathrm{cov}(\epsilon_{t1}, \epsilon_{t2}) = 0$, the bias can be calculated to be converging in probability to

$$\widehat{w}_{12} - w_{12} \xrightarrow{\mathrm{P}} \frac{w_{21}\sigma^2_{\epsilon,1}}{1 - w_{12}w_{21}} \Big/ \frac{w^2_{21}(\sigma^2_{\epsilon,1} + \sigma^2_{X,1}) + \sigma^2_{\epsilon,2} + \sigma^2_{X,2}}{(1 - w_{12}w_{21})^2} = \frac{w_{21}\sigma^2_{\epsilon,1}(1 - w_{12}w_{21})}{w^2_{21}(\sigma^2_{\epsilon,1} + \sigma^2_{X,1}) + \sigma^2_{\epsilon,2} + \sigma^2_{X,2}},$$

which is not going to 0 unless either w_{21} or $\sigma^2_{\epsilon,1}$ goes to 0 as $T \to \infty$, since assumption A7 ensures that $\sigma^2_{X,j} > u > 0$ uniformly.

By symmetry of the formulae for the asymptotic biases of $\widehat{w}_{12}$ and $\widehat{w}_{21}$, we can easily see that if $\sigma^2_{\epsilon,1}$ and $\sigma^2_{\epsilon,2}$ are not decaying, these biases can have larger magnitudes then the corresponding weight w_{12} or w_{21}, so that the corresponding estimator cannot be sign consistent even if w_{12} or w_{21} are going to 0 as $T \to \infty$. This demonstrates the necessity of decaying variances for the noise.

If $\sigma^2_{X,1} = \sigma^2_{X,2} = 0$ (assumption A7 fails), and $\sigma^2_{\epsilon,1} = \sigma^2_{\epsilon,2}$, we see that the asymptotic bias becomes independent of $\sigma^2_{\epsilon,j}$, and $\widehat{w}_{12}$ and $\widehat{w}_{21}$ cannot be both sign consistent. Hence it is important that covariates are included in our model. Luckily, assumption A3 allows for past values of $\mathbf{y}_t$ to be our covariates $\mathbf{X}_t$. See (4.3) in section 4.1 for more details.

One final remark is that, for this simple toy example, we may consistently estimate the parameters of the reduced form model like that in (2.2), and recover w_{12} and w_{21} from the estimated reduced form model without assumption A2. But, as explained in example 1, when N is large and a general weight matrix is our target, the problem can become intractable and consistent estimation is then not achievable unless assumption A2 is satisfied. See also section 7 where an instrumental variable approach is mentioned and is still under research to overcome major technical difficulties when used together with LASSO.

We introduce more notations and definitions before presenting our results. Define

$$J = \{j : \xi^*_j \neq 0, \text{and does not correspond to } w^*_{2,ss}, s = 1, \ldots, N\}. \tag{4.4}$$

Hence J is the index set for all truly non-zero weights in $\mathbf{W}^*_1$ and $\mathbf{W}^*_2$ excluding the diagonal entries of $\mathbf{W}^*_2$, which are known to be 1. Define $n = |J|$, $s_1 = \sum_{j\in J} \xi^*_{1,j}$, $s = \sum_{j\in J} \xi^*_j$ and $s_2 = s - s_1$. Denote $\mathbf{v}_S$ a vector $\mathbf{v}$ restricted to those components with index $j \in S$. Let $\lambda_T = cT^{-1/2}\log^{1/2}(T \vee N)$ where c is a large enough constant (see Theorem 1 for the exact value of c), and define the sets

$$\begin{aligned}
\mathcal{A}_1 &= \Big\{ \max_{1\le j,\ell\le N} \max_{1\le k\le K} \Big|\frac{1}{T}\sum_{t=1}^{T} \zeta_{t,j}X_{t,\ell k}\Big| < \lambda_T \Big\}, \\
\mathcal{A}_2 &= \Big\{ \max_{1\le k\le K} \Big|\frac{1}{T}\sum_{j=1}^{N}\sum_{t=1}^{T} \zeta_{t,j}X_{t,jk}\Big| < \lambda_T N^{1/2+1/2w} \Big\}, \\
\mathcal{A}_3 &= \Big\{ \max_{1\le i,j\le N} \Big|\frac{1}{T}\sum_{t=1}^{T} [\zeta_{t,i}\zeta_{t,j} - E(\zeta_{t,i}\zeta_{t,j})]\Big| < \lambda_T \Big\}, \\
\mathcal{A}_4 &= \Big\{ \max_{1\le i,j\le N} \max_{1\le \ell,m\le K} \Big|\frac{1}{T}\sum_{t=1}^{T} X_{t,i\ell}X_{t,jm} - E(X_{t,i\ell}X_{t,jm})\Big| < \lambda_T \Big\}, \\
\mathcal{M} &= \left\{ \max_{1\le t\le T} \max_{1\le j\le N} \max_{1\le k\le K} |X_{t,jk}| < \left(\frac{3\log(T\vee N)}{D_2}\right)^{1/q} \right\},
\end{aligned} \tag{4.5}$$

where w is as defined in assumption A6.

4.3 Main results

We first present a Nagaev-type inequality for a general time series $\{\mathbf{x}_t\}$ under similar settings in (4.1) and (4.2), which is a combination of Theorems 2(ii) and 2(iii) of Liu et al. [2013].

Lemma 1. *For a zero mean time series process $\mathbf{x}_t = \mathbf{f}(\mathcal{F}_t)$ as defined in (4.1) with dependence measure $\theta^x_{t,a,j}$ as defined in (4.2), assume $\Theta^x_{m,w} \le Cm^{-\alpha}$ for some $w > 2$ and constants $C, \alpha > 0$. Then there exists constants C_1, C_2 and C_3 independent of v, T and the index j such that*

$$P\bigg(\bigg|\frac{1}{T}\sum_{t=1}^{T} x_{t,j}\bigg| > v\bigg) \le \frac{C_1 T^{w(\frac{1}{2}-\widetilde{\alpha})}}{(Tv)^w} + C_2 \exp\big(-C_3 T^{\widetilde{\beta}} v^2\big),$$

where $\widetilde{\alpha} = \alpha \wedge (1/2 - 1/w)$, and $\widetilde{\beta} = (3 + 2\widetilde{\alpha}w)/(1+w)$.

Furthermore, assume another zero mean time series process $\{\mathbf{z}_t\}$ (can be the same process $\{\mathbf{x}_t\}$) with both $\Theta^x_{m,2w}, \Theta^z_{m,2w} \le Cm^{-\alpha}$, as in assumption A6. Then provided there is a constant μ such that $\max_j \|x_{tj}\|_{2w}, \max_j \|z_{tj}\|_{2w} \le \mu < \infty$, the above Nagaev-type inequality holds for the product process $\{x_{tj}z_{t\ell} - E(x_{tj}z_{t\ell})\}$.

Remark 2. Note if $\alpha > 1/2 - 1/w$, then $w(1/2 - \widetilde{\alpha}) = \widetilde{\beta} = 1$, simplifying the form of the inequality. Hereafter we assume $\alpha > 1/2 - 1/w$ where w is in assumption A6, and is large enough as specified in Remark 3. We assume this purely for the simplification of all results. For instance, if $\alpha < 1/2 - 1/w$, then we can define $\lambda_T = cT^{-\widetilde{\beta}/2}\log^{1/2}(T \vee N)$ and (more complicated) rates of convergence in different theorems can be derived.

Proof of Lemma 1. The first part is a direct consequence of Theorems 2(ii) and 2(iii) of Liu et al. [2013]. The second part follows from $E(x_{tj}z_{t\ell}) = E(x'_{tj}z'_{t\ell})$, and using the generalized Hölder inequality,

$$\begin{aligned}
\theta^{xz}_{t,w,j\ell} = \|x_{tj}z_{t\ell} - x'_{tj}z'_{t\ell}\|_w &\le \|x_{tj}z_{t\ell} - x_{tj}z'_{t\ell}\|_w + \|x_{tj}z'_{t\ell} - x'_{tj}z'_{t\ell}\|_w \\
&\le \max(\|x_{tj}\|_{2w}, \|z'_{t\ell}\|_{2w})(\theta^x_{t,2w,j} + \theta^z_{t,2w,\ell}) \\
&\le \mu(\theta^x_{t,2w,j} + \theta^z_{t,2w,\ell}),
\end{aligned}$$

so that

$$\Theta^{xz}_{m,w} \le \sum_{t=m}^{\infty} \max_{j,\ell} \mu(\theta^x_{t,2w,j} + \theta^z_{t,2w,\ell}) \le \mu(Cm^{-\alpha} + Cm^{-\alpha}) = 2\mu C m^{-\alpha}.$$

The result follows by applying the first part of Lemma 1. □

With Lemma 1, we can use the union sum inequality to find an explicit probability lower bound for the event $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4$. The proof of the theorem is relegated to the Appendix.

Theorem 1. *Let assumptions A3 - A6 be satisfied. Suppose $\alpha > 1/2 - 1/w$, and suppose for the applications of the Nagaev-type inequality in Lemma 1 for the processes in $\mathcal{A}_1$ to $\mathcal{A}_4$, the constants C_1, C_2 and C_3 are*

the same. Then with $c \geq \sqrt{3/C_3}$ *where* c *is the constant defined in* λ_T*, we have*

$$P(\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4 \cap \mathcal{M}) \geq 1 - 4C_1K^2\left(\frac{C_3}{3}\right)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T \vee N)} - \frac{4C_2K^2N^2 + D_1NTK}{T^3 \vee N^3}.$$

It approaches 1 if we assume further that $N = o(T^{w/4-1/2}\log^{w/4}(T))$*.*

Remark 3. With tail assumptions A5, we can easily show that $\|\zeta_{tj}\|_{2w}, \|x_{tj}\|_{2w} < \infty$ for any $w > 0$ (see the proof of Theorem 1 in the Appendix), and there are many examples with $\Theta^x_{m,2w}, \Theta^\zeta_{m,2w} \leq Cm^{-\alpha}$ where only the constant C is dependent on w (see for example the stationary linear process example 2.2 in Chen et al. [2013]). Therefore we can set w to be large enough so that $N = o(T^{w/4-1/2}\log^{w/4}(T))$ from the beginning, ensuring $P(\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4 \cap \mathcal{M}) \to 1$.

Lemma 2. *Let assumptions A1 to A7 be satisfied. Denote* $\widetilde{\mathbf{W}}_1$ *and* $\widetilde{\mathbf{W}}_2$ *any estimators for* $\mathbf{W}_1^*$ *and* $\mathbf{W}_2^*$ *respectively (not necessarily the LASSO estimators). Define a generic notation* $\mathbf{A}^\otimes = \mathbf{I}_N \otimes \mathbf{A}$ *for a matrix* $\mathbf{A}$*, and denote* $\mathbf{y}^v = (\mathbf{y}_1^\mathrm{T}, \ldots, \mathbf{y}_T^\mathrm{T})^\mathrm{T}$*,* $\mathbf{X} = (\mathbf{X}_1^\mathrm{T}, \ldots, \mathbf{X}_T^\mathrm{T})^\mathrm{T}$*.*

Then on $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4$*, the least square estimator* $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^\mathrm{T}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}\widetilde{\mathbf{W}}_2^\otimes\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}(\mathbf{I}_{TN} - \widetilde{\mathbf{W}}_1^\otimes)\mathbf{y}^v$ *is well-defined, and*

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{a_1(s_2 + N^{\frac{1}{2}+\frac{1}{2w}})\lambda_T\delta_T^{1/2}}{N} + \frac{a_2}{N}\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1,$$

where the constants a_1 *and* a_2 *are defined in Theorem 3.*

The proof is relegated to the Appendix. If we treat $\widetilde{\mathbf{W}}_1$ and $\widetilde{\mathbf{W}}_2$ as some assumed weight matrices, for example distance weight matrices with a particular distance metric, this lemma together with Theorem 1 tells us that with high probability, the error upper bound for estimating $\boldsymbol{\beta}^*$ is related to the error for estimating the weight matrices through $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1$. As long as $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1$ is much less than N, estimation error is related to how sparse the weight matrix $\mathbf{W}_2^*$ (i.e., s_2) is. Otherwise, the error can be large. We provide some simulation results for the estimation of $\boldsymbol{\beta}^*$ in section 6.

We now present an oracle inequality for the error bounds of the LASSO and adaptive LASSO estimators $\widetilde{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\xi}}$ respectively. The proof is presented in the Appendix.

Theorem 2. *Let assumptions A1-A7 be satisfied. Suppose* $\alpha > 1/2 - 1/w$*, and suppose* $\lambda_T = o(\delta_T^{1/2})$*,* $\lambda_T N^{1/w} = O(\delta_T^{1/2})$ *and* $s_2 = O(N^{1/2}\delta_T^{1/4}/\lambda_T^{1/2})$*. Then there is a tuning parameter* γ_T *with* $\gamma_T \asymp \delta_T$ *such that on* $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4$*, the LASSO estimator* $\widetilde{\boldsymbol{\xi}}$ *satisfies*

$$\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1 \leq 4\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J^*\|_1, \quad \textit{so that} \quad \|\widetilde{\boldsymbol{\xi}}_{J^c} - \boldsymbol{\xi}_{J^c}^*\|_1 \leq 3\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J^*\|_1.$$

For $\widehat{\boldsymbol{\xi}}$*, denote* $\xi_{S,\min\,/\,\max} = \min\,/\,\max_{j\in S}\xi_j$ *and* $\widetilde{J}$ *the LASSO estimator for* J *in (4.4). Then*

$$\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1 \leq \frac{4|\widetilde{\xi}_{\widetilde{J},\max}|^k}{|\widetilde{\xi}_{J,\min}|^k}\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J^*\|_1, \quad \textit{so that} \quad \|\widehat{\boldsymbol{\xi}}_{J^c} - \boldsymbol{\xi}_{J^c}^*\|_1 \leq \left(\frac{4|\widetilde{\xi}_{\widetilde{J},\max}|^k}{|\widetilde{\xi}_{J,\min}|^k} - 1\right)\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J^*\|_1.$$

For the exact value of the constant B where $\gamma_T = B\delta_T$, see the proof of the theorem which is relegated to the Appendix. The rate $\lambda_T = o(\delta_T^{1/2})$ implies that the rate of decay for the standard deviation of the noise is slower than λ_T.

The results in Theorem 2 are consistent with the properties of the LASSO estimators under the usual linear regression settings (see (3.2) of Bickel et al. [2009]). With these oracle inequalities, we need to introduce a restricted eigenvalue condition which is similar to condition (3.1) of Bickel et al. [2009]. We however define this condition on a population covariance matrix instead, since our raw design matrix $\mathbf{M}_{\boldsymbol{\beta}^*}$ in (3.1) is always random:

A8. *Restricted eigenvalue condition*: Let $\widehat{\boldsymbol{\Sigma}}^* = T^{-1}\mathbf{M}_{\boldsymbol{\beta}^*}^{\mathrm{T}}\mathbf{M}_{\boldsymbol{\beta}^*}$, and $\boldsymbol{\Sigma} = E(\widehat{\boldsymbol{\Sigma}}^*)$. Define

$$\kappa(r) = \min\left\{\frac{\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}\right\|}{\left\|\boldsymbol{\alpha}_R\right\|}, \frac{\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}\right\|}{\left\|\boldsymbol{\alpha}_{R^c}\right\|} : |R| \leq r, \boldsymbol{\alpha} \in \mathbb{R}^{2N^2}\backslash\{\mathbf{0}\}, \left\|\boldsymbol{\alpha}_{R^c}\right\|_1 \leq c_0\left\|\boldsymbol{\alpha}_R\right\|_1\right\},$$

where $c_0 = \frac{8}{|\boldsymbol{\xi}^*_{J,\min}|^k} - 1$. Then we assume $\kappa(n) > 0$ uniformly as $N, T \to \infty$.

This condition is automatically satisfied if $\boldsymbol{\Sigma}$ has the smallest eigenvalue bounded uniformly away from 0. Similar population restricted eigenvalue condition is also introduced in Zhou et al. [2009] for the analysis of LASSO and adaptive LASSO estimators when the design matrix is formed by i.i.d. rows which are multivariate normally distributed.

Theorem 3. *Let assumption A8 and the assumptions in Theorem 2 be satisfied. Suppose also* $\lambda_T n, \gamma_T n^{1/2} = o(1)$, $(N^{1/2w} + s_2 N^{-1/2})\lambda_T \gamma_T^{-1/2} \log^{1/q}(T \vee N) = o(n^{1/2})$, $n = o(N \log^{-2/q}(T \vee N))$, *where* γ_T *is the same as in Theorem 2. Then on* $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4 \cap \mathcal{M}$, *for large enough* N, T,

$$\left\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\right\| \leq \frac{5\gamma_T n^{1/2}}{\kappa^2(n)}, \quad \left\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\right\| \leq \frac{5\gamma_T n^{1/2}}{\kappa^2(n)|\xi^*_{J,\min}|^k}.$$

Furthermore, for N, T *large enough and suitable constants* a_1 *and* a_2, *on* $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4 \cap \mathcal{M}$,

$$\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 \leq a_1\left(\frac{s_2}{N} + N^{\frac{1}{2w}-\frac{1}{2}}\right)\lambda_T \delta_T^{1/2} + \frac{20 a_2 \gamma_T n}{N\kappa^2(n)},$$
$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 \leq a_1\left(\frac{s_2}{N} + N^{\frac{1}{2w}-\frac{1}{2}}\right)\lambda_T \delta_T^{1/2} + \frac{25 a_2 |\xi^*_{J,\max}|^k \gamma_T n}{N\kappa^2(n)|\xi^*_{J,\min}|^{2k}}.$$

The proof is relegated to the Appendix. Theorems 2 and 3 together implies the following.

Corollary 4. *Under the assumptions of Theorems 2 and 3, for large enough* N, T,

$$\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1 \leq \frac{20\gamma_T n}{\kappa^2(n)}, \quad \left\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1 \leq \frac{25|\xi^*_{J,\max}|^k \gamma_T n}{\kappa^2(n)|\xi^*_{J,\min}|^{2k}}.$$

Corollary 4 says that, in addition to the assumptions in Theorem 3, if $\gamma_T n = o(1)$ also, then all the LASSO and adaptive LASSO estimators from (3.2) and (3.3) converge to their respective true quantities in L_1 norm on the set $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4 \cap \mathcal{M}$, which has probability approaching 1 with explicit probability lower bound shown in Theorem 1. The need for large enough N, T are merely for the simplification of the different error bounds, and can be removed at the expense of more complicated expressions. The proof is omitted.

We conclude this section with the sign consistency theorem for the weight matrices. In the following and hereafter we denote $\mathbf{M}_{AB}$ a matrix $\mathbf{M}$ with rows restricted to the set A and columns to the set B. The proof of the Theorem can be found in the Appendix.

Theorem 5. *Let the assumptions in Theorem 2 and 3 be satisfied. Assume further that* $\lambda_{\min}(\mathbf{\Sigma}_{JJ})$ *is uniformly bounded away from 0, and* $n = o\Big(\gamma_T^{-\frac{2k}{k+1}}\Big)$. *Then on* $\mathcal{A}_1 \cap \cdots \mathcal{A}_4 \cap \mathcal{M}$ *and for large enough* N, T,

$$sign(\widehat{\boldsymbol{\xi}}) = sign(\boldsymbol{\xi}^*).$$

This theorem says that with a suitable rate of decay for the noise variances and the true spatial weight matrices sparse enough, we can correctly estimate the sign (i.e. 0, positive or negative) of every element in the spatial weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ on $\mathcal{A}_1 \cap \cdots \mathcal{A}_4 \cap \mathcal{M}$. Hence asymptotic sign consistency is achieved by Theorem 1. This is very important in recovering the correct sparse pattern for understanding the underlying cross-sectional dependence structure of the panel data.

The rate $n = o\Big(\gamma_T^{-\frac{2k}{k+1}}\Big)$ suggests that the number of non-zero elements allowed in the weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ without violating sign consistency depends on the rate of decay for the variance of the noise. For instance if $\gamma_T \asymp \lambda_T \log^{1/2}(T \vee N)$ and $k = 1$, then $n = o(T^{1/2}\log^{-1}(T \vee N))$. If $k = 2$, $n = o(T^{2/3}\log^{-4/3}(T \vee N))$. Theoretically when k is larger, n can grow larger too. However in practice k cannot be much larger than 1 since all weights $\xi_j \leq 1$, meaning that the weights v_j in the adaptive LASSO problem (3.3) can be too large for the truly non-zero components if k is large, rendering some of them being penalized wrongly to zero. Also, the error bound for $\widehat{\boldsymbol{\xi}}$ in Corollary 4 gets worse as k increases. Hence we set $k \leq 3$ in the theorem.

5 Practical Implementation

In this section, we provide details of the block coordinate descent (BCD) algorithm for carrying out the minimizations for (3.2) and (3.3). We need the BCD algorithm since the objective functions in these problems are not convex in $(\boldsymbol{\xi}, \boldsymbol{\beta})$, although given $\boldsymbol{\beta}$, they are convex in $\boldsymbol{\xi}$ and vice versa.

The BCD algorithm is closely related to the Iterative Coordinate Descent of Fan and Lv [2011], and is also discussed in Friedman et al. [2010] and Dicker et al. [2010]. While it is difficult to establish global convergence of the BCD algorithm without convexity, it is easy to see that for (3.2) and (3.3), each iteration delivers an improvement of the objective functions since given one parameter, the objective functions are convex in the other. From our experience, starting from an appropriate initial value, a minimum will be achieved with good performance in practice. Indeed in the simulation experiments in section 6 (not shown), it is found that the algorithm is robust to a variety of initial values chosen.

We choose blocks to take advantage of intra-block convexity. The parameter $\boldsymbol{\beta}$ forms one block, and for $j = 1, \ldots, N$, $\boldsymbol{\eta}_j^{\mathrm{T}} = (\boldsymbol{\eta}_{1j}^{\mathrm{T}}, \boldsymbol{\eta}_{2j}^{\mathrm{T}})$ = the j-th row of $(\mathbf{W}_1, \mathbf{W}_2)$ form N other blocks. Given the values of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}_{-j} = (\boldsymbol{\eta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\eta}_{j-1}^{\mathrm{T}}, \boldsymbol{\eta}_{j+1}^{\mathrm{T}}, \ldots, \boldsymbol{\eta}_N^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\eta}_j$ is solved by the Least Angle Regression algorithm (LARS) of Efron et al. [2004]. Given $\boldsymbol{\xi}$, $\boldsymbol{\beta}$ is solved by the ordinary least square (OLS) estimator.

The Block Coordinate Descent Algorithm

0. Start with an initial value $\boldsymbol{\xi} = \boldsymbol{\xi}^{(0)}$. This can be obtained by using $\boldsymbol{\beta}^{(0)} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}^v$ (for notations see Lemma 2), and solves (3.2) given $\boldsymbol{\beta}^{(0)}$ using LARS. This gives $\boldsymbol{\xi}^{(0)}$.

1. At step r, set $\boldsymbol{\beta}^{(r)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_2^{\otimes}(r-1)^{\mathrm{T}}\mathbf{W}_2^{\otimes}(r-1)\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_2^{\otimes}(r-1)^{\mathrm{T}}(\mathbf{I}_{TN} - \mathbf{W}_1^{\otimes}(r-1))\mathbf{y}^v$, where $\mathbf{W}_j^{\otimes}(r) = \mathbf{I}_N \otimes \mathbf{W}_j(r)$, with $\mathbf{W}_1(r), \mathbf{W}_2(r)$ the weight matrices recovered from $\boldsymbol{\xi}^{(r)}$.

2. Using LARS, solve sequentially for $j = 1, \ldots, N$,

$$\boldsymbol{\eta}_j^{(r)} = \arg\min_{\boldsymbol{\eta}_j} \left\|\mathbf{y} - \mathbf{M}_{\boldsymbol{\beta}^{(r)}}\boldsymbol{\eta}\right\|^2 + \lambda\left\|\boldsymbol{\eta}_j\right\|_1, \text{ subj. to } \left\|\boldsymbol{\eta}_{1j}\right\|_1 < 1, \left\|\boldsymbol{\eta}_{2j}\right\|_1 < 2,$$

where $\boldsymbol{\eta} = (\check{\boldsymbol{\eta}}_1^{\mathrm{T}}, \check{\boldsymbol{\eta}}_2^{\mathrm{T}})^{\mathrm{T}}$ with $\check{\boldsymbol{\eta}}_i = (\boldsymbol{\eta}_{i1}^{(r-1)\mathrm{T}}, \ldots, \boldsymbol{\eta}_{i,j-1}^{(r-1)\mathrm{T}}, \boldsymbol{\eta}_{ij}^{\mathrm{T}}, \boldsymbol{\eta}_{i,j+1}^{(r-1)\mathrm{T}}, \ldots, \boldsymbol{\eta}_{iN}^{(r-1)\mathrm{T}})^{\mathrm{T}}$. Then

$$\boldsymbol{\xi}^{(r)} = (\boldsymbol{\eta}_{11}^{(r)\mathrm{T}}, \ldots, \boldsymbol{\eta}_{1N}^{(r)\mathrm{T}}, \boldsymbol{\eta}_{21}^{(r)\mathrm{T}}, \ldots, \boldsymbol{\eta}_{2N}^{(r)\mathrm{T}})^{\mathrm{T}}.$$

3. Iterate steps 1-2 until $\left\|\boldsymbol{\xi}^{(r)} - \boldsymbol{\xi}^{(r-1)}\right\|_1$ is smaller than some pre-set number. The LASSO solution is then $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\xi}}) = (\boldsymbol{\beta}^{(r)}, \boldsymbol{\xi}^{(r)})$.

4. Take $\boldsymbol{\xi}^{(0)} = \widetilde{\boldsymbol{\xi}}$. Repeat steps 1-3 for the adaptive LASSO solutions, where in step 2 the penalty function is modified to $\lambda\mathbf{v}_j^{\mathrm{T}}|\boldsymbol{\eta}_j|$, with the components in $\mathbf{v}_j$ having the form $1/|\widetilde{\xi}_j|^k$.

Cross-validation is performed to select the tuning parameter λ, with prediction error in L_2-norm being the criterion used. In steps, denote T_a a set with consecutive time points. Then the sample with $t \in T_a$ is the test set. We then compute the LASSO solutions $(\widetilde{\boldsymbol{\beta}}_{a,\lambda}, \widetilde{\boldsymbol{\xi}}_{a,\lambda})$ for the sample with $t \in T_a$ on a grid of values of λ. Recovering the weight matrices $\mathbf{W}_{1;a,\lambda}$ and $\widetilde{\mathbf{W}}_{2;a,\lambda}$ from $\widetilde{\boldsymbol{\xi}}_{a,\lambda}$, we then solves

$$\lambda_{CV} = \arg\min_{\lambda} \sum_a \sum_{t \in T_a^c} \left\|\mathbf{y}_t - (\mathbf{I}_N - \widetilde{\mathbf{W}}_{1;a,\lambda})^{-1}\widetilde{\mathbf{W}}_{2;a,\lambda}\mathbf{X}_t\widetilde{\boldsymbol{\beta}}_{a,\lambda}\right\|^2.$$

In practice, we usually set T_a to have 3/5 to 4/5 of the total time points, and define $T_1, \ldots, T_5$ by moving the test data window forward.

6 Numerical Examples

We give detailed simulation results in section 6.1 for our LASSO and adaptive LASSO estimators. A set of stock markets data is analyzed in section 6.2 to visualize the connection among international financial markets.

6.1 Simulation Results

We generate data from model (2.1) and investigate the practical performance of the LASSO and adaptive LASSO estimators.

First, we generate independent Gaussian data from the model as a baseline for studying the performance of the estimators. To this end, we generate the spatial weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ by randomly setting elements in a row of the matrices (except diagonal elements) to be either 0.3 or 0, with an overall sparsity level (i.e. n, the number of non-zero elements) set at a pre-specified level. If the sum of a row excluding any diagonal elements is larger than 1, then we normalize it by 1.1 times the L_1 norm of the row. We set $\boldsymbol{\beta}^* = (1, 0.5)^{\mathrm{T}}$. The covariate matrix $\mathbf{X}_t$ has independent rows $\mathbf{x}_{t,j}^{\mathrm{T}}$ generated by $\mathbf{x}_{t,j} \sim N(\mathbf{0}, (\sigma_{x,ij}))$ where $\sigma_{x,11} = \sigma_{x,22} = 2$ and $\sigma_{x,12} = 0.5$ for each time t. Finally the noise $\boldsymbol{\epsilon}_t$ is a spatially uncorrelated Gaussian white noise with mean $\mathbf{0}$ and variance $\sigma_\epsilon^2 = \frac{\log(T \vee N)}{\sqrt{T}} \Big/ \frac{\log(50)}{\sqrt{50}}$, so that $\sigma_\epsilon^2 = 1$ for the case $N = 25, T = 50$.

We simulate 2 different pairs of $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$, and generate data 100 times according to the scheme above for each pair. Hence in total 200 set of data is generated and analyzed for each particular (N, T) combination. We used $N = 25, 50, 75$ and $T = 50, 100, 200$ to explore the effects of dimension on the performance of our estimators when it can be larger than the sample size T.

Table 1 shows the results of this baseline simulation. From $T = 50$ to 100 the sensitivity (see the table for definition) improved hugely, while specificity remains at a similar level. It is intuitive since the non-zero elements are relatively small, and hence when T is too small they cannot be picked up easily. Bias are mostly negative, meaning that we usually underestimate the non-zero values of the weight matrices. Also it is clear that the performance of the adaptive LASSO is much better than LASSO in general. It is of interest to note that while the L_1 error norm can be large, the L_2 error norm is usually much smaller. These are consistent with the results in Theorem 3, where the L_2 error norm goes to 0 as long as $\gamma_T n^{1/2} = o(1)$, but for the L_1 error norm to go to 0 we need $\gamma_T n = o(1)$ in general.

Table 2 shows the average value of the tuning parameter λ using 5 fold cross-validation. Clearly the true sparsity level is approximately retained at all combinations of N and T.

Table 3 consider two more cases. One is when the covariates include a lagged variable $\mathbf{y}_{t-1}$ on top of X_t. We set $\boldsymbol{\beta}^* = (1, 0.5, 0.15)^{\mathrm{T}}$ which ensures the model for $\mathbf{y}_t$ is stationary. While when $N = 25$ results are similar to the baseline simulations, for $N = 50$ and 75 the performance is getting worse in general. This indicates that while in theory it is fine to include lagged variables, we may need a larger T or a limited N for good performance in practice.

Another case is when the noise exhibits spatial correlations. To this end, we randomly pick the off-diagonal elements in the noise covariance matrix to be 0.3, while keeping it sparse with around 95% elements still 0. The performance is similar to the baseline simulations in general. This is consistent with our theories. In particular this scenario fits assumption A4 (see section 4.1): when there are weak or no spatial correlations in the covariates, then the spatial correlation structure in the noise can be general.

Finally, Table 4 shows some results when some assumptions are violated. The first case is setting the variance of the noise equal to $\sigma_\epsilon^2 = 1$, instead of letting it decay as in the baseline simulations. Clearly the performance is worse in general even when $T = 200$. The results are consistent with Example 2 in section 4.2. The performance when there are no covariates is also shown in the table. The poor performance all round under the absence of covariates is again consistent with Example 2 in section 4.2. Lastly, we simulate the noise using the t_3 distribution rather than normal distribution, violating the tail assumption A5 in section 4.1. While the performance is worse in general, it is still better than when there are no covariates or no variance decay. Hence the method is more robust to fat tails.

6.2 Analysis of stock markets data

Performance of stock markets around the world are well-known to be under mutual influence of each other. More diverse geographic production and globalization deepen this fact. Financial linkages are also well-documented.

To study the dependence structure of worldwide stock markets in more detail, we use model (2.1) to analyze the data. We estimate the spatial weight matrix $\mathbf{W}_1^*$ using the adaptive LASSO estimator. The response variable $\mathbf{y}_t$ is taken as the panel of stock market returns for the 26 biggest world markets. We

		$\kappa = 0.95$						$\kappa = 0.99$					
		$T = 50$		$T = 100$		$T = 200$		$T = 50$		$T = 100$		$T = 200$	
		$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$	Specificity	92.59%	98.86%	96.11%	98.55%	96.29%	98.72%	96.66%	98.79%	97.81%	98.37%	98.45%	98.89%
	Sensitivity	78.76%	30.59%	96.54%	79.49%	99.54%	95.45%	68.05%	47.51%	95.04%	88.67%	99.02%	97.05%
	Bias	-.0335	-.1368	-.0048	-.0338	-.0075	-.0235	-.0622	-.1086	-.0194	-.0262	-.0101	-.0136
	LASSO L_1	7.8323	7.6321	10.7733	10.0291	11.5132	11.7350	2.6372	1.8735	9.3369	8.6764	7.1839	7.0933
	LASSO L_2	0.9919	1.8643	0.8634	1.0301	0.8831	0.9978	0.2447	0.3078	0.6217	0.6680	0.3749	0.4043
	AdaLASSO L_1	7.2676	8.0508	3.4799	3.9148	2.1098	2.0332	1.8196	1.6875	0.9946	1.0002	0.5095	0.5035
	AdaLASSO L_2	1.3345	2.1873	0.4630	0.7927	0.1883	0.2669	0.2834	0.3868	0.1088	0.1398	0.0398	0.0487
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0631		.0362		.0336		.0197		.0169		.0113	
$N = 50$	Specificity	95.22%	98.92%	92.70%	98.81%	97.04%	98.72%	98.48%	99.99%	96.78%	98.89%	97.77%	98.37%
	Sensitivity	76.82%	40.86%	84.82%	35.94%	98.77%	92.24%	83.09%	57.48%	90.80%	74.92%	99.27%	97.40%
	Bias	-.0179	-.0676	-.0557	-.1928	-.0068	-.0429	-.0307	-.0381	-.0624	-.1204	-.0091	-.0142
	LASSO L_1	63.5781	29.9079	24.4255	32.5351	8.0875	9.9059	42.2633	31.3463	32.1011	27.0674	26.1828	25.1055
	LASSO L_2	7.2787	9.2961	2.9375	7.9924	1.6371	1.9855	3.8004	3.5031	0.5061	0.7882	1.2952	1.3716
	AdaLASSO L_1	27.0909	7.8570	3.7208	8.8239	0.7099	1.4649	6.2692	6.2161	5.0422	5.0152	2.5933	2.4494
	AdaLASSO L_2	4.8459	7.8570	3.7208	8.8239	0.7099	1.4649	1.0332	1.4425	0.5519	0.9805	0.1705	0.2057
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0609		.0582		.0511		.0520		.0392		.0180	
$N = 75$	Specificity	95.24%	98.92%	97.03%	99.48%	96.60%	98.33%	98.25%	99.95%	98.84%	99.57%	97.53%	98.35%
	Sensitivity	76.03%	41.21%	85.81%	73.51%	95.59%	90.38%	80.96%	46.06%	95.69%	84.43%	99.51%	96.42%
	Bias	-.0178	-.0673	-.0199	-.0901	-.0088	-.0651	-.0284	-.0400	-.0173	-.0342	-.0042	-.0145
	LASSO L_1	63.6976	56.7362	53.1475	51.5816	49.5190	46.2353	82.5607	53.8498	80.7792	69.5684	31.1932	24.9452
	LASSO L_2	7.3024	9.1839	8.2361	11.4955	2.6296	3.7519	7.6407	6.7404	5.3195	5.4675	1.2661	1.2283
	AdaLASSO L_1	27.4082	29.4429	33.6911	44.6034	19.6262	28.4384	15.9184	15.2392	7.2987	7.4012	5.5895	5.1590
	AdaLASSO L_2	4.9189	7.7206	4.3452	9.8865	1.7318	4.5238	2.6822	3.7875	0.8878	1.3655	0.3218	0.4329
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0591		.0547		.0396		.0718		.0440		.0238	

Table 1: Baseline Simulations. All values are averages over 200 simulations. The parameter κ represents the true sparsity level of the both $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$. Specificity is the percentage of zeros estimated as zeros. Sensitivity is the percentage of non-zeros estimated as non-zeros. LASSO L_1 is the L_1 error norm $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1$ for the LASSO estimator, and AdaLASSO represents the adaptive LASSO. Bias is the sum of error for the estimated non-zero values without taking absolute values.

		$\kappa = 0.95$			$\kappa = 0.99$		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$N = 25$	λ	26.56	31.37	32.31	31.25	22.69	52.31
	Sparsity	96.58%	94.25%	94.20%	99.41%	98.01%	98.76%
$N = 50$	λ	8.21	31.25	31.54	13.13	31.28	43.71
	Sparsity	94.27%	96.36%	95.65%	98.50%	99.04%	98.78%
$N = 75$	λ	8.30	12.54	34.31	10.85	13.61	31.08
	Sparsity	94.28%	94.52%	94.72%	98.13%	98.40%	99.06%

Table 2: Average value of the tuning parameter λ over 200 simulations using five fold cross-validation. The parameter κ represents the true sparsity level of the weight matrix $\mathbf{W}_1^*$ in the simulation.

use daily data available for the whole of 2012 ($T = 252$). See Table 5 for details of the markets and their respective indices.

For the covariates we use the S&P Global 1200 Index and the Dow Jones World Stock Index. By definition, firms that belong to the world index are constituents of the indices of some markets. Hence the exogeneity of the covariates cannot be sustained. Nevertheless, the global variables are included with the purpose of eliminating a global-wide variance that could prevent the identification of $\mathbf{W}_1^*$. Due to the lack of variance in the cross-sectional dimension, $\mathbf{W}_2^*$ is unidentified and hence it is simply set as the identity matrix.

The model is estimated by the adaptive LASSO, with the tuning parameter λ chosen by cross-validation described in section 5. Figure 1 shows the graph of $\widehat{\mathbf{W}}_1$, where a non-zero $\widehat{W}_{1,ij}$ is represented by an edge directed from market i to j. With only 38 directed edges out of $26^2 - 26 = 650$ possible, this is a very sparse graph. It is clear that there is a subgraph dominated by eastern countries, another by western countries, and a third for the United States and Switzerland.

We carry out further study by seeking a connection between the estimated spatial weight matrix $\widehat{\mathbf{W}}_1$ and the number of common opening hours among different markets. Define

$$\text{Common Opening Hours}_{i,j} \quad = \quad \max\left\{\frac{\text{Close Time}_i - \max\left\{\text{Open Time}_i, \text{Open Time}_j\right\}}{\text{Close Time}_i - \text{Open Time}_i}, 0\right\}$$

as the time of market i exposed within a day to market j. The numerator is simply the number of hours of market i subject to the influence from the j-th one, even if the latter has already closed before market i opens. The fraction is therefore the ratio of hours of market i subject to the influence of market j. It is naturally bounded below by zero.

In Figure 2, the elements of $\widehat{\mathbf{W}}_1$ are plotted against the common opening hours. From this figure, it is clear that for markets with less overlapping of opening hours, the estimated elements are zero in $\widehat{\mathbf{W}}_1$. In fact, markets are only affecting each other if they are commonly open for at least half of their opening times.

7 Conclusion

In this paper, we developed an adaptive LASSO regularization for the spatial weight matrices in a spatial lag model when the dimension of the panel can be larger than the sample size. An important feature for

		Original Simulations				*Time Dependence*				*Spatial Dependence*			
		$T = 100$		$T = 200$		$T = 100$		$T = 200$		$T = 100$		$T = 200$	
		$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
	Specificity	96.11%	98.55%	96.29%	98.72%	94.91%	98.36%	96.09%	99.65%	97.03%	98.51%	96.82%	98.04%
	Sensitivity	96.54%	79.49%	99.54%	95.45%	93.75%	84.31%	99.99%	93.10%	97.06%	89.73%	98.34%	91.98%
	Bias	-.0048	-.0338	-.0075	-.0235	-.0213	-.0162	-.0273	-.0456	-.0075	-.0222	-.0045	-.0124
$N = 25$	Lasso L_1	10.7733	10.0291	11.5132	11.7350	14.0459	11.5385	11.4053	12.1764	2.8841	2.8482	5.2222	5.2578
	Lasso L_2	0.8634	1.0301	0.8831	0.9978	1.2523	1.2454	1.0141	1.1657	0.3635	0.4177	1.2039	1.2388
	AdaLasso L_1	3.4799	3.9148	2.1098	2.0332	4.4262	3.2545	2.8581	2.6339	1.8718	2.0391	2.1397	2.2883
	AdaLasso L_2	0.4630	0.7927	0.1883	0.2669	.5997	.6189	.3954	.4560	0.3339	2.0391	2.1397	2.2883
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0362		.0336		.0512		.0394		.0044		.0068	
	Specificity	92.70%	98.81%	97.04%	98.72%	93.65%	86.70%	95.78%	98.89%	94.22%	99.48%	97.04%	98.48%
	Sensitivity	84.82%	35.94%	98.77%	92.24%	57.36%	9.26%	95.31%	78.57%	97.39%	80.25%	98.68%	97.21%
	Bias	-.0557	-.1928	-.0068	-.0429	-.0937	-.0937	.0022	-.0865	-.0357	-.1175	-.0077	-.0243
$N = 50$	Lasso L_1	24.4255	32.5351	8.0875	9.9059	38.4113	60.9752	40.2475	37.8140	13.8891	17.8884	8.2285	8.5885
	Lasso L_2	2.9375	7.9924	1.6371	1.9855	6.5169	13.1556	2.7584	3.5580	1.2833	3.2541	0.4916	0.6557
	AdaLasso L_1	3.7208	8.8239	0.7099	1.4649	35.4057	56.1173	14.3274	16.5937	13.8437	19.4374	4.5166	5.2155
	AdaLasso L_2	3.7208	8.8239	0.7099	1.4649	6.4232	12.5275	1.4847	3.4627	1.5294	4.0252	0.4110	0.5682
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0582		.0511		.8182		.1312		.0173		.0153	
	Specificity	97.03%	99.48%	96.60%	98.33%	78.07%	99.99%	78.07%	99.98%	99.23%	99.99%	96.25%	98.72%
	Sensitivity	85.81%	73.51%	95.59%	90.38%	24.37%	3.18%	24.37%	0.00%	85.50%	64.82%	99.25%	98.14%
	Bias	-.0199	-.0901	-.0088	-.0651	-.2519	-.2208	-.2318	-0.2791	.0056	-.0731	-.0113	-.0314
$N = 75$	Lasso L_1	53.1475	51.5816	49.5190	46.2353	74.5943	79.5124	61.7143	79.8384	28.9961	27.5597	24.5726	26.4133
	Lasso L_2	8.2361	11.4955	2.6296	3.7519	15.0406	23.8510	14.7065	23.8531	2.1741	5.6786	1.2291	1.6740
	AdaLasso L_1	33.6911	44.6034	19.6262	28.4384	74.5943	79.5124	61.7143	79.8384	13.1129	21.9017	11.1335	13.3919
	AdaLasso L_2	4.3452	9.8865	1.7318	4.5328	15.0406	23.8510	14.7065	23.8531	4.2185	8.2501	0.6687	1.0429
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0547		.0396		1.3902		1.4853		.0728		.0661	

Table 3: Comparisons to the baseline simulations when the covariates include $\mathbf{y}_{t-1}$ (under the columns "Time Dependence") and when the noise exhibits spatial correlations (under the columns "Spatial Dependence"). Refer to Table 1 for the explanations of different items.

		Original Simulations.				*No Variance Decay*				*No Covariates*		*Fat Tails*			
		$T = 100$		$T = 200$		$T = 100$		$T = 200$		$T = 100$	$T = 200$	$T = 100$		$T = 200$	
		$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_1^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
	Specificity	96.11%	98.55%	96.29%	98.72%	94.88%	98.62%	95.04%	98.33%	58.00%	61.92%	94.61%	98.36%	94.38%	98.01%
	Sensitivity	96.54%	79.49%	99.54%	95.45%	95.04%	54.07%	98.52%	80.59%	42.00%	41.55%	85.99%	46.35%	96.89%	76.98%
	Bias	-.0048	-.0338	-.0075	-.0235	-.0056	-.0392	-.0163	-.0638	-.2456	-.2592	-.0098	-.0284	-.0203	-.0591
$N = 25$	Lasso L_1	10.7733	10.0291	11.5132	11.7350	14.7335	13.1964	16.9506	17.3715	17.4593	18.7124	17.4760	15.1448	17.6558	17.7609
	Lasso L_2	0.8634	1.0301	0.8831	0.9978	1.4060	1.8144	1.5490	1.9357	3.7079	3.2821	2.2824	2.8294	2.1185	2.5928
	AdaLasso L_1	3.4799	3.9148	2.1098	2.0332	5.0450	6.3607	4.4057	3.9808	17.4593	18.7124	7.0409	6.9634	5.4979	5.1720
	AdaLasso L_2	0.4630	0.7927	0.1883	0.2669	0.7370	1.5669	0.5351	-0.7717	3.7079	3.2821	1.4905	2.2062	0.9899	1.3242
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0362		.0336		.0558		.0468		—	—	.0521		.0479	
	Specificity	92.70%	98.81%	97.04%	98.72%	89.45%	99.30%	95.16%	98.66%	54.60%	60.19%	89.97%	99.95%	94.83%	98.45%
	Sensitivity	84.82%	35.94%	98.77%	92.24%	84.95%	29.06%	96.23%	71.64%	45.40%	47.32%	81.05%	17.31%	93.32%	57.62%
	Bias	-.0557	-.1928	-.0068	-.0429	-.0691	-.1999	-.0229	-.0897	-.1809	-.2071	-.0611	-.2055	-.0206	-.0881
$N = 50$	Lasso L_1	24.4255	32.5351	8.0875	9.9059	28.7581	30.9560	53.9934	50.8391	55.5067	53.9241	29.2132	35.6802	57.3559	54.7110
	Lasso L_2	2.9375	7.9924	1.6371	1.9855	3.3456	8.3228	4.3632	5.5186	8.9744	9.7472	3.8837	10.1041	5.9993	7.9254
	AdaLasso L_1	3.7208	8.8239	0.7099	1.4649	28.3113	31.7817	15.5946	20.3610	55.5067	53.9241	28.5963	36.0980	19.6727	27.3955
	AdaLasso L_2	3.7208	8.8239	0.7099	1.4649	4.1735	8.8233	1.7174	4.4298	8.9744	9.7472	4.8213	10.4364	3.2320	7.0662
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0582		.0511		.1168		.0987		—	—	.0729		.0612	
	Specificity	97.03%	99.48%	96.60%	98.33%	96.66%	99.89%	92.08%	97.95%	55.76%	60.56%	96.75%	99.77%	97.43%	99.82%
	Sensitivity	85.81%	73.51%	95.59%	90.38%	76.26%	27.76%	87.36%	48.52%	43.20%	47.94%	69.29%	20.07%	82.30%	79.31%
	Bias	-.0199	-.0901	-.0088	-.0651	-.0148	-.0958	-.0097	-.1137	-.2307	-.2109	-.0013	-.0934	-0.0011	-.0761
$N = 75$	Lasso L_1	53.1475	51.5816	49.5190	46.2353	121.240	107.822	79.0006	63.6286	85.0331	88.6169	127.653	102.507	75.2035	78.0112
	Lasso L_2	8.2361	11.4955	2.6296	3.7519	11.2820	16.3793	5.5820	8.2053	9.8341	8.2868	13.1635	18.5941	7.2351	8.1063
	AdaLasso L_1	33.6911	44.6034	19.6262	28.4384	49.3344	65.8009	39.1014	62.1415	84.9210	84.9809	52.9285	79.5576	29.1066	25.1037
	AdaLasso L_2	4.3452	9.8865	1.7318	4.5328	7.9461	18.4804	4.0581	15.3782	8.3519	7.6390	9.5241	22.7887	5.9502	10.5102
	$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	.0547		.0396		.1366		.1282		—	—	.1579		.1003	

Table 4: Comparisons to the baseline simulations when assumptions are violated. Refer to Table 1 for the explanations of different items.

Country	Code	Index	Country	Code	Index
Argentina	ARG	Merval	Australia	AUL	Dow Jones Australian
Austria	AUT	Viena ATX-5	Brazil	BRZ	Dow Jones Brazil Stock
Canada	CAN	S&P/CDNX Composite	Chile	CHL	Santiago SSE Inter-10
China	CHN	Shanghai SE Composite	Egypt	EGP	SE 100
France	FRA	Paris CAC-40	Germany	GER	CDAX Total Return
Hong Kong	HHK	Hang Seng Composite	India	IDI	NSE-50
Indonesia	IDO	Jakarta SE Liquid 45	Italy	ITA	Milan SE MIB-30
Japan	JPN	Nikkei 500	Mexico	MEX	SE Index (INMX)
New Zealand	NZZ	NZSX-15	Russia	RUS	Russia MICEX Composite
Spain	SPA	Madrid SE IBEX-35	Singapore	SIN	Singapore FTSE All-shares
South Africa	STA	FTSE/JSE Top 40 Tradable Stocks	South Korea	SKK	Korea SE Stock Price
Switzerland	SWZ	Swiss Market	Thailand	THA	Thailand SET General
United Kingdom	UKK	S&P United Kingdom	United States	USA	S&P 500

Table 5: Markets and their respective indices used. Data source: *Global Financial Data*.

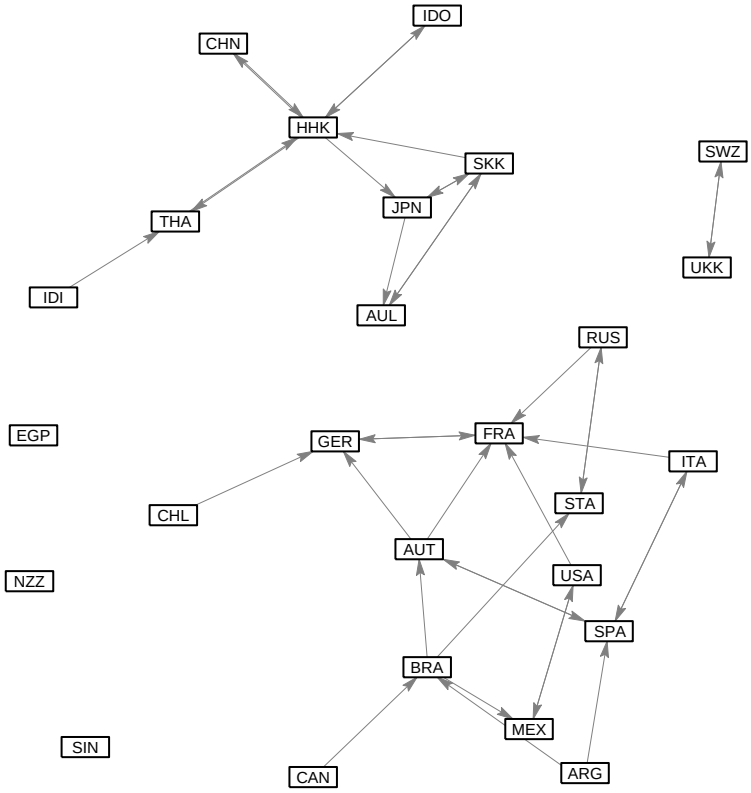

Figure 1: Graph of $\widehat{\mathbf{W}}_1$. An edge directed from market i to j means that $\widehat{W}_{1,ij}$ is non-zero.

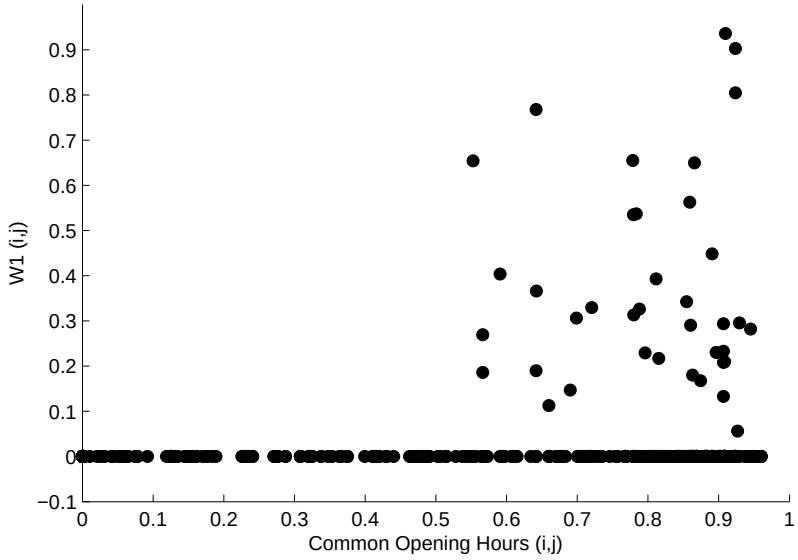

Figure 2: Elements of $\widehat{\mathbf{W}}_1$ plotted against Common Opening Hours.

our LASSO/adaptive LASSO regularized estimation is that unlike many others, our method does not need the specification of the spatial weight matrices or a distance metric for them as in Pinkse et al. [2002]. All parameters in the model are estimated together with the spatial weight matrices, with explicit rates of convergence of various errors stated and proved. In particular, an error upper bound is derived for the regression parameter β^* in our spatial lag model under an arbitrary specification/estimation of the spatial weight matrices, showing that as long as these matrices are specified/estimated with an L_1 error much less than the panel size N, the estimation for $\boldsymbol{\beta}^*$ will be accurate.

The asymptotic sign consistency of the estimated spatial weight matrices is proved as well, showing that we can recover the cross-sectional dependence structure in the spatial weight matrices asymptotically. Another contribution is the development of a practical block coordinate descent algorithm for our method, which is used for the simulation results and a real data analysis.

We argued that covariates are important for our results. Yet there are applications without obvious covariates. Also, the variance of the noise in the panel may not be small enough to satisfy the variance decay assumption in practice. Indeed if enough instruments are available for each covariate, the instrumental variable approach can potentially remove the need for variance decay. There are still major technical hurdles to overcome in this direction. A further study will be to regularize on the reduced form model directly and we impose sparsity on the weight matrices by simple thresholding. This way not even instrumental variables are needed. These are the potential future problems to be tackled.

8 Appendix

Proof of Theorem 1. We first show that, with the tail condition in A5 for a process $\{\mathbf{z}_t\}$, we have for any $w > 0$, $\max_j \left\|z_{tj}\right\|_{2w} \leq \mu_{2w} < \infty$. Hence we can fix a w large enough such that $N = o(T^{w/4-1/2}\log^{w/4}(T))$;

see Remark 3 after Theorem 1. Indeed by the Fubini's Theorem,

$$E|z_{tj}|^{2w} = E\int_0^{|z_{tj}|^{2w}} ds = \int_0^\infty P(|z_{tj}| > s^{1/2w})\, ds \le \int_0^\infty D_1 \exp(-D_2 s^{q/2w})\, ds$$
$$= \frac{4wD_1}{q}\int_0^\infty x^{4w/q-1}e^{-D_2x^2}\, dx = \frac{2wD_1}{qD_2^{2w/q}}\Gamma(2w/q) \text{ [define as } \mu_{2w}^{2w}] < \infty, \tag{8.1}$$

so that $\max_j \left\|z_{tj}\right\|_{2w} \le \mu_{2w} < \infty$ for any $w > 0$. Together with assumption A6, Lemma 1 can then be applied for the processes $\{\zeta_{t,j}X_{t,\ell k}\}$, $\{\zeta_{t,i}\zeta_{t,j} - E(\zeta_{t,i}\zeta_{t,j})\}$ and $\{X_{t,i\ell}X_{t,jm} - E(X_{t,i\ell}X_{t,jm})\}$. Since $\alpha > 1/2 - 1/w$, we have $w(1/2 - \widetilde{\alpha}) = \widetilde{\beta} = 1$ in Lemma 1. The union sum inequality implies

$$P(\mathcal{A}_1^c) \le \sum_{\substack{1\le j,\ell\le N\\ 1\le k\le K}} P\left(\left|T^{-1}\sum_{t=1}^T \zeta_{t,j}X_{t,\ell k}\right| \ge \lambda_T\right) \le N^2K\left(\frac{C_1T}{(T\lambda_T)^w} + C_2\exp(-C_3T\lambda_T^2)\right)$$
$$\le C_1K\left(\frac{C_3}{3}\right)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2KN^2}{T^3\vee N^3}. \tag{8.2}$$

Similarly, we have

$$\begin{aligned} P(\mathcal{A}_3^c) &\le C_1\left(\frac{C_3}{3}\right)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2N^2}{T^3\vee N^3},\\ P(\mathcal{A}_4^c) &\le C_1K^2\left(\frac{C_3}{3}\right)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2K^2N^2}{T^3\vee N^3}. \end{aligned} \tag{8.3}$$

The tail assumption A5 and the union sum inequality imply that

$$P(\mathcal{M}^c) \le NTK\cdot D_1\exp(-3\log(T\vee N)) = \frac{D_1NTK}{T^3\vee N^3}. \tag{8.4}$$

Finally, if we can show that

$$\max_{1\le k\le K}\left\|N^{-\frac{1}{2}-\frac{1}{2w}}\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k}\right\|_{2w} < \infty, \tag{8.5}$$

$$\Theta_{m,2w} = \sum_{t=m}^\infty \max_{1\le k\le K}\left\|N^{-\frac{1}{2}-\frac{1}{2w}}(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k} - \boldsymbol{\zeta}_t'^{\mathrm{T}}\mathbf{X}'_{t,k})\right\|_{2w} \le am^{-\alpha}, \tag{8.6}$$

for some $a > 0$ and all $m \ge 1$, then we can apply Lemma 1 for $\mathcal{A}_2$ to obtain

$$P(\mathcal{A}_2^c) \le \sum_{k=1}^K P\left(\left|T^{-1}\sum_{t=1}^T N^{-\frac{1}{2}-\frac{1}{2w}}\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k}\right| \ge \lambda_T\right)$$
$$\le C_1\left(\frac{C_3}{3}\right)^{w/2}\frac{K}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2K}{T^3\vee N^3}. \tag{8.7}$$

Combining (8.2), (8.3), (8.4) and (8.7), we can then use

$$P(\mathcal{A}_1\cap\ldots\cap\mathcal{A}_4\cap\mathcal{M}) \ge 1 - \sum_{j=1}^4 P(\mathcal{A}_j) - P(\mathcal{M})$$

to yield the conclusion of the Theorem. It remains to show (8.5) and (8.6).

We use assumption A4 and we assume first that $\left\|\mathbf{\Sigma}_{xk}^{1/2}\right\|_\infty \le S_x < \infty$, and $\{X_{t,jk}^*\}_{1\le j\le N}$ is a martingale difference with respect to the filtration generated by $(X_{t,1k}^*,\ldots,X_{t,jk}^*)$. Assuming the other part of A4 for the noise results in very similar proof and we omit it. Write $\sum_{j=1}^N \zeta_{t,j}X_{t,jk} = \boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k} = \boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2}\mathbf{X}_{t,k}^* = \sum_{j=1}^N(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_jX_{t,jk}^*$, where $\mathbf{X}_{t,k}$,$\mathbf{X}_{t,k}^*$ are the k-th columns of $\mathbf{X}_t$ and $\mathbf{X}_t^*$ respectively. Then by the independence assumption A3,

$$E\big((\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_jX_{t,jk}^*|(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_s, X_{t,sk}^*,\ s\le j-1\big) = E\big((\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_j|(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_s,\ s\le j-1\big) \cdot E(X_{t,jk}^*|X_{t,sk}^*,\ s\le j-1) = 0,$$

since $\{X_{t,jk}^*\}_{1\le j\le N}$ is a martingale difference. Hence $\{(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_jX_{t,jk}^*\}_{1\le j\le N}$ is a martingale difference. By Lemma 2.1 of Li [2003], assumptions A3, A4 and (8.1), we then have

$$\begin{aligned}
E\Big|N^{-\frac{1}{2}-\frac{1}{2w}}\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k}\Big|^{2w} &= E\Big|N^{-\frac{1}{2}-\frac{1}{2w}}\sum_{j=1}^N(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_jX_{t,jk}^*\Big|^{2w}\\
&\le N^{-2}(36w)^{2w}(1+(2w-1)^{-1})^w\sum_{j=1}^N E|(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_jX_{t,jk}^*|^{2w}\\
&= N^{-2}(36w)^{2w}(1+(2w-1)^{-1})^w\sum_{j=1}^N E|(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_j|^{2w}E|X_{t,jk}^*|^{2w}\\
&\le N^{-2}(36w\mu_{2w})^{2w}(1+(2w-1)^{-1})^w\sum_{j=1}^N E\Big|\max_{1\le j\le N}|\zeta_{t,j}|\Big|^{2w}\left\|\mathbf{\Sigma}_{xk}^{1/2}\right\|_\infty^{2w}\\
&\le N^{-2}(36w\mu_{2w}S_x)^{2w}(1+(2w-1)^{-1})^w\sum_{j=1}^N N\max_{1\le j\le N}E|\zeta_{t,j}|^{2w}\\
&\le (36w\mu_{2w}^2S_x)^{2w}(1+(2w-1)^{-1})^w<\infty,
\end{aligned}$$

so that $\max_{1\le k\le K}\left\|N^{-\frac{1}{2}-\frac{1}{2w}}\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k}\right\|_{2w}<\infty$, which is (8.5).

To prove (8.6), observe that

$$\begin{aligned}
\Theta_{m,2w} &\le \sum_{t=m}^\infty\max_{1\le k\le K}N^{-\frac{1}{2}-\frac{1}{2w}}\left[\left\|\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2}(\mathbf{X}_{t,k}^*-\mathbf{X}_{t,k}'^*)\right\|_{2w}+\left\|(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2}-\boldsymbol{\zeta}_t'^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})\mathbf{X}_{t,k}'^*\right\|_{2w}\right],\\
&\le \sum_{t=m}^\infty\max_{1\le k\le K}N^{-\frac{1}{2}-\frac{1}{2w}}\left[\left\|\sum_{j=1}^N(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_j(X_{t,jk}^*-X_{t,jk}'^*)\right\|_{2w}+\left\|\sum_{j=1}^N(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2}-\boldsymbol{\zeta}_t'^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_jX_{t,jk}'^*\right\|_{2w}\right].
\end{aligned}$$

With similar arguments as before, $\{(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_j(X_{t,jk}^*-X_{t,jk}'^*)\}_j$ and $\{(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2}-\boldsymbol{\zeta}_t'^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_jX_{t,jk}'^*\}_j$ can be shown to be martingale differences with respect to the filtration

$$\mathcal{F}_j=\sigma(X_{t,sk}^*,X_{t,sk}'^*,(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_s,(\boldsymbol{\zeta}_t'^{\mathrm{T}}\mathbf{\Sigma}_{xk}^{1/2})_s,\ s\le j).$$

Hence we can use Lemma 2.1 of Li [2003], assumptions A3, A4, A6 and (8.1) to show that

$$\begin{aligned}\big\|N^{-\frac{1}{2}-\frac{1}{2w}}\sum_{j=1}^{N}(\boldsymbol{\zeta}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{xk}^{1/2})_j(X_{t,jk}^{*}-X_{t,jk}^{\prime *})\big\|_{2w} &\le 36w(1+(2w-1)^{-1})^{1/2}\\ &\quad\cdot\left[N^{-2}\sum_{j=1}^{N}E\Big|\max_{1\le j\le N}|\zeta_{t,j}|\Big|^{2w}\big\|\boldsymbol{\Sigma}_{xk}^{1/2}\big\|_{\infty}^{2w}(\theta_{t,2w,jk}^{x^*})^{2w}\right]^{1/2w}\\ &\le 36w\mu_w S_x(1+(2w-1)^{-1})^{1/2}\max_{1\le j\le N}\theta_{t,2w,jk}^{x^*}.\end{aligned}$$

Similarly,

$$\big\|N^{-\frac{1}{2}-\frac{1}{2w}}\sum_{j=1}^{N}(\boldsymbol{\zeta}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{xk}^{1/2}-\boldsymbol{\zeta}_t^{\prime\mathrm{T}}\boldsymbol{\Sigma}_{xk}^{1/2})_jX_{t,jk}^{\prime *}\big\|_{2w}\le 36w\mu_w S_x(1+(2w-1)^{-1})^{1/2}\max_{1\le j\le N}\theta_{t,2w,j}^{\zeta}.$$

Hence combining and using assumption A6, we have

$$\Theta_{m,2w}\le 36w\mu_w S_x(1+(2w-1)^{-1})^{1/2}(\Theta_{m,2w}^{x^*}+\Theta_{m,2w}^{\zeta})\le 72Cw\mu_w S_x(1+(2w-1)^{-1})^{1/2}m^{-\alpha},$$

which is (8.6). The proof is now completed. □

Proof of Lemma 2. Denote $\mathbf{U}=\mathbf{I}_N\otimes T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}$, and

$$\mathbf{V}=\begin{pmatrix}\mathbf{I}_K\otimes\widetilde{\mathbf{w}}_{21}\\ \vdots\\ \mathbf{I}_K\otimes\widetilde{\mathbf{w}}_{2N}\end{pmatrix},\quad\text{where }\widetilde{\mathbf{w}}_{2j}^{\mathrm{T}}\text{ is the }j\text{-th row of }\widetilde{\mathbf{W}}_2.$$

Then $\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes}\mathbf{X}=\mathbf{V}^{\mathrm{T}}\mathbf{U}\mathbf{V}$, and we decompose $\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*=\sum_{j=1}^{5}I_i$, where

$$\begin{aligned}I_1&=-(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}\mathbf{V}^{T}(\mathbf{U}-E(\mathbf{U}))\mathbf{V}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*),\\ I_2&=(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}T^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}(\mathbf{W}_2^{*\otimes}-\widetilde{\mathbf{W}}_2^{\otimes})\mathbf{X}\boldsymbol{\beta}^*,\\ I_3&=(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}T^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}\boldsymbol{\epsilon}^{v},\\ I_4&=(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}T^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}(\mathbf{W}_1^{*\otimes}-\widetilde{\mathbf{W}}_1^{\otimes})(\mathbf{I}_{TN}-\mathbf{W}_1^{*\otimes})^{-1}\mathbf{W}_2^{*\otimes}\mathbf{X}\boldsymbol{\beta}^*,\\ I_5&=(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}T^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}(\mathbf{W}_1^{*\otimes}-\widetilde{\mathbf{W}}_1^{\otimes})(\mathbf{I}_{TN}-\mathbf{W}_1^{*\otimes})^{-1}\boldsymbol{\epsilon}^{v},\end{aligned}$$

where $\boldsymbol{\epsilon}^v$ is defined similar to $\mathbf{y}^v$. Note by assumptions A1 and A7,

$$\big\|(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}\big\|_{\infty}\le\frac{K^{1/2}}{\lambda_{\min}(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})}\le\frac{K^{1/2}}{\lambda_{\min}(E(\mathbf{U}))\lambda_{\min}(\mathbf{V}^{\mathrm{T}}\mathbf{V})}\le\frac{K^{1/2}}{uN}.\tag{8.8}$$

Then on $\mathcal{A}_4$, using (8.8),

$$\begin{aligned}\big\|\mathbf{I}_1\big\|_1&\le K\big\|(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}\big\|_{\infty}\big\|\mathbf{V}^{\mathrm{T}}\big\|_{\infty}\big\|\mathbf{U}-E(\mathbf{U})\big\|_{\max}\big\|\mathbf{V}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\big\|_{\infty}\\ &\le\frac{K^{3/2}}{uN}\cdot 2N\cdot\lambda_T\cdot\big\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1=\frac{2K^{3/2}\lambda_T}{u}\big\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1.\end{aligned}$$

Similarly on $\mathcal{A}_4$, using (8.8) and assumptions A1, A4,

$$\begin{aligned}
\big\|I_2\big\|_1 &\le \frac{K^{1/2}}{uN}\cdot\big\|T^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}(\mathbf{W}_2^{*\otimes}-\widetilde{\mathbf{W}}_2^{\otimes})\mathbf{X}\big\|_\infty\big\|\boldsymbol{\beta}^*\big\|_1\\
&= \frac{K^{1/2}\big\|\boldsymbol{\beta}^*\big\|_1}{uN}\max_{1\le i\le K}\sum_{j=1}^{K}\Big|\sum_{\ell,s=1}^{N}(w_{2,s\ell}^*-\widetilde{w}_{2,s\ell})\sum_{k=1}^{N}\Big(\widetilde{w}_{2,sk}T^{-1}\sum_{t=1}^{T}X_{t,ki}X_{t,\ell j}\Big)\Big|\\
&\le \frac{K^{1/2}\big\|\boldsymbol{\beta}^*\big\|_1}{uN}\cdot 2K(\sigma_{\max}^2+\lambda_T)\big\|\widetilde{\boldsymbol{\xi}}_2-\boldsymbol{\xi}_2^*\big\|_1 = \frac{2K^{3/2}(\sigma_{\max}^2+\lambda_T)\big\|\boldsymbol{\beta}^*\big\|_1}{uN}\big\|\widetilde{\boldsymbol{\xi}}_2-\boldsymbol{\xi}_2^*\big\|_1.
\end{aligned}$$

Similarly on $\mathcal{A}_1$ and $\mathcal{A}_2$, using (8.8) and assumptions A1, A4,

$$\begin{aligned}
\big\|I_3\big\|_1 &\le \frac{K^{1/2}\delta_T^{1/2}}{uN}\cdot\big\|T^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{W}}_2^{\otimes\mathrm{T}}\boldsymbol{\zeta}^v\big\|_1 = \frac{K^{1/2}\delta_T^{1/2}}{uN}\sum_{k=1}^{K}\Big|\sum_{s,\ell=1}^{N}\widetilde{w}_{2,s\ell}T^{-1}\sum_{t=1}^{T}X_{t,sk}\zeta_{t,\ell}\Big|\\
&= \frac{K^{3/2}\delta_T^{1/2}}{uN}\max_{1\le k\le K}\Big|\sum_{s,\ell=1}^{N}(\widetilde{w}_{2,s\ell}-w_{2,s\ell}^*)T^{-1}\sum_{t=1}^{T}X_{t,sk}\zeta_{t,\ell}+\sum_{s,\ell=1}^{N}w_{2,s\ell}^*T^{-1}\sum_{t=1}^{T}X_{t,sk}\zeta_{t,\ell}\Big|\\
&\le \frac{K^{3/2}\delta_T^{1/2}}{uN}(\lambda_T\big\|\widetilde{\boldsymbol{\xi}}_2-\boldsymbol{\xi}_2^*\big\|_1+\lambda_T N^{\frac{1}{2}+\frac{1}{2w}}+\lambda_T s_2).
\end{aligned}$$

Finally, note that the row sum condition in assumption A1 implies

$$\big\|(\mathbf{I}_N-\mathbf{W}_1^*)^{-1}\big\|_\infty \le \sum_{k\ge 0}\big\|\mathbf{W}_1^*\big\|_\infty^k \le \sum_{k\ge 0}\eta^k = (1-\eta)^{-1}. \tag{8.9}$$

Hence using this, (8.8) and assumptions A1,A4, on $\mathcal{A}_1$ and $\mathcal{A}_4$, we have (tedious algebra omitted)

$$\begin{aligned}
\big\|I_4\big\|_1 &\le \frac{4K^{3/2}\big\|\boldsymbol{\beta}^*\big\|_1(\sigma_{\max}^2+\lambda_T)}{(1-\eta)uN}\big\|\widetilde{\boldsymbol{\xi}}_1-\boldsymbol{\xi}_1^*\big\|_1,\\
\big\|I_5\big\|_1 &\le \frac{2K^{3/2}\lambda_T\delta_T^{1/2}}{(1-\eta)uN}\big\|\widetilde{\boldsymbol{\xi}}_1-\boldsymbol{\xi}_1^*\big\|_1.
\end{aligned}$$

Using the expressions for $\big\|I_1\big\|_1$to $\big\|I_5\big\|_1$, rearranging and simplifying, we thus have

$$\begin{aligned}
\big\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1 &\le \frac{K^{3/2}}{u-2K^{3/2}\lambda_T}\left\{\frac{(s_2+N^{\frac{1}{2}+\frac{1}{2w}})\lambda_T\delta_T^{1/2}}{N}+\frac{4\big\|\boldsymbol{\beta}^*\big\|_1(\sigma_{\max}^2+\lambda^T)+2\lambda_T\delta_T^{1/2}}{(1-\eta)N}\big\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\big\|_1\right\}\\
&\le \frac{a_1(s_2+N^{\frac{1}{2}+\frac{1}{2w}})\lambda_T\delta_T^{1/2}}{N}+\frac{a_2}{N}\big\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\big\|_1,
\end{aligned}$$

which is the inequality for $\big\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1$ if we set constants

$$a_1 \ge \frac{K^{3/2}}{u-2K^{3/2}\lambda_T},\quad a_2 \ge \frac{4K^{3/2}\big\|\boldsymbol{\beta}^*\big\|_1(\lambda_T+\sigma_{\max}^2)+2\lambda_T\delta_T^{1/2}K^{3/2}}{(1-\eta)(u-2K^{3/2}\lambda_T)}.\ \square$$

Proof of Theorem 2. For the LASSO estimator $\widetilde{\boldsymbol{\xi}}$, (3.2) implies

$$\frac{1}{2T}\big\|\mathbf{y}-\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}\big\|^2+\gamma_T\big\|\widetilde{\boldsymbol{\xi}}\big\|_1 \le \frac{1}{2T}\big\|\mathbf{y}-\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\big\|^2+\gamma_T\big\|\boldsymbol{\xi}^*\big\|_1,$$

which, using model (3.1), can be rearranged to

$$\frac{1}{2T}\left\|\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^* - \mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}\right\|^2 \leq \frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) + \frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}}_2 - \mathrm{vec}(\mathbf{I}_N)) + \frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{M}_{\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*) + \gamma_T(\left\|\boldsymbol{\xi}^*\right\|_1 - \left\|\widetilde{\boldsymbol{\xi}}\right\|_1). \tag{8.10}$$

On $\mathcal{A}_2$, using $\epsilon_{tj} = \delta_T^{1/2}\zeta_{tj}$,

$$\left|\frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N)\right| = \left|\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{N}\epsilon_{tj}\sum_{k=1}^{K}X_{t,jk}(\widetilde{\beta}_k - \beta_k^*)\right| \leq \lambda_T\delta_T^{1/2}N^{\frac{1}{2}+\frac{1}{2w}}\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1.$$

On $\mathcal{A}_1$, recalling $s_2 = \left\|\boldsymbol{\xi}_2^* - \mathrm{vec}(\mathbf{I}_N)\right\|_1$,

$$\begin{aligned}\left|\frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}}_2 - \mathrm{vec}(\mathbf{I}_N))\right| &\leq \max_{1\leq j\neq\ell\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{tj}\sum_{k=1}^{K}X_{t,\ell k}(\widetilde{\beta}_k - \beta_k^*)\right|\cdot\left\|\widetilde{\boldsymbol{\xi}}_2 - \mathrm{vec}(\mathbf{I}_N)\right\|_1 \\ &\leq \lambda_T\delta_T^{1/2}\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1(s_2 + \left\|\widetilde{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2^*\right\|_1).\end{aligned}$$

Finally,

$$\left|\frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{M}_{\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*)\right| \leq \max_{\substack{1\leq j\neq\ell\leq N\\ 1\leq k\leq K}}\left\{\left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{tj}y_{t\ell}\right|, \left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{tj}X_{t,\ell k}\right|\cdot\left\|\boldsymbol{\beta}^*\right\|_1\right\}\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1.$$

Writing the ℓ-th row of $\boldsymbol{\Pi}_1$ as $\boldsymbol{\pi}_{1,\ell}^{\mathrm{T}}$, using (8.9), we have on $\mathcal{A}_1$ and $\mathcal{A}_3$,

$$\begin{aligned}\left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{tj}y_{t\ell}\right| &\leq \left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{tj}\boldsymbol{\pi}_{1,\ell}^{*\mathrm{T}}\mathbf{W}_2^*\mathbf{X}_t\boldsymbol{\beta}^*\right| + \left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{tj}\boldsymbol{\pi}_{1,\ell}^{*\mathrm{T}}\boldsymbol{\epsilon}_t\right| \\ &\leq \frac{2\delta_T^{1/2}\left\|\boldsymbol{\beta}^*\right\|_1}{1-\eta}\max_{\substack{1\leq\ell\leq N\\ 1\leq k\leq K}}\left|\frac{1}{T}\sum_{t=1}^{T}\zeta_{tj}X_{t,\ell k}\right| + \frac{\delta_T}{1-\eta}\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}[\zeta_{tj}\zeta_{ti} - E(\zeta_{tj}\zeta_{ti})]\right| + \frac{\delta_T\sigma_0^2}{1-\eta} \\ &\leq \frac{2\lambda_T\delta_T^{1/2}\left\|\boldsymbol{\beta}^*\right\|_1 + \lambda_T\delta_T + \delta_T\sigma_0^2}{1-\eta},\end{aligned}$$

where we used assumption A2 that $|E(\zeta_{ti}\zeta_{tj})| \leq \sigma_0^2$. Combining these bounds, on $\mathcal{A}_1$ and $\mathcal{A}_3$,

$$\left|\frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{M}_{\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*)\right| \leq (\lambda_T\delta_T^{1/2}a_T + c_\eta\delta_T)\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1,$$

$$\text{where } c_\eta = \frac{\sigma_0^2}{(1-\eta)},\ a_T = \left\|\boldsymbol{\beta}^*\right\|_1 + \frac{2\left\|\boldsymbol{\beta}^*\right\|_1 + \delta_T^{1/2}}{1-\eta}.$$

Hence utilizing all these bounds, (8.10) becomes

$$\begin{aligned}\frac{1}{2T}\left\|\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}} - \mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\right\|^2 &\leq \lambda_T\delta_T^{1/2}(N^{\frac{1}{2}+\frac{1}{2w}} + s_2 + \left\|\widetilde{\boldsymbol{\xi}}_2 - \boldsymbol{\xi}_2^*\right\|_1)\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 \\ &\quad + (\lambda_T\delta_T^{1/2}a_T + c_\eta\delta_T)\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1 + \gamma_T(\left\|\boldsymbol{\xi}^*\right\|_1 - \left\|\widetilde{\boldsymbol{\xi}}\right\|_1).\end{aligned}$$

Using the result of Lemma 2 on the LASSO estimator $\widetilde{\beta}$, and assuming $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1 > \lambda_T\delta_T^{1/2}$, we have

(tedious algebra omitted)

$$\begin{aligned}\frac{1}{2T}\left\|\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}-\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\right\|^2 &\le a_1\lambda_T\delta_T^{1/2}\left(N^{\frac{1}{2w}}+s_2N^{-\frac{1}{2}}+\lambda_T\delta_T^{1/2}\right)^2\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1\\ &+a_2\lambda_T\delta_T^{1/2}\left(2+N^{\frac{1}{2w}-\frac{1}{2}}+s_2N^{-1}\right)\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1\\ &+(\lambda_Ta_T+c_\eta\delta_T)\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1+\gamma_T(\left\|\boldsymbol{\xi}^*\right\|_1-\left\|\widetilde{\boldsymbol{\xi}}\right\|_1).\end{aligned}$$

Using the rates condition specified in the theorem, the dominant term is $c_\eta\delta_T\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1$, so that there is a constant $D\ge 3a_1+4a_2+c_\eta+a_T$ such that

$$\frac{1}{2T}\left\|\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}-\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\right\|^2\le D\delta_T\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1+\gamma_T(\left\|\boldsymbol{\xi}^*\right\|_1-\left\|\widetilde{\boldsymbol{\xi}}\right\|_1).$$

Setting $\gamma_T=2D\delta_T$, we then have

$$\begin{aligned}D\delta_T\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1&\le\frac{1}{2T}\left\|\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}-\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\right\|^2+D\delta_T\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1\\ &\le 2D\delta_T(\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1+\left\|\boldsymbol{\xi}^*\right\|_1-\left\|\widetilde{\boldsymbol{\xi}}\right\|_1)\\ &=2D\delta_T(\left\|\widetilde{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\right\|_1+\left\|\boldsymbol{\xi}_J^*\right\|_1-\left\|\widetilde{\boldsymbol{\xi}}_J\right\|_1)\\ &\le 4D\delta_T\left\|\widetilde{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\right\|_1.\end{aligned}$$

Hence $\left\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1\le 4\left\|\widetilde{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\right\|_1$, which implies

$$\left\|\widetilde{\boldsymbol{\xi}}_{J^c}-\boldsymbol{\xi}_{J^c}^*\right\|_1\le 3\left\|\widetilde{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\right\|_1.$$

Following exactly the same lines of proof, for the adaptive LASSO estimator $\widehat{\boldsymbol{\xi}}$ we have

$$\frac{1}{2T}\left\|\mathbf{M}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\xi}}-\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\right\|^2\le D\delta_T\left\|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1+\gamma_T\mathbf{v}^{\mathrm{T}}(|\boldsymbol{\xi}^*|-|\widehat{\boldsymbol{\xi}}|).$$

Again set $\gamma_T=2D\delta_T$, then using $2v_j-1\ge v_j$ since $v_j>1$,

$$\begin{aligned}\frac{1}{2T}\left\|\mathbf{M}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\xi}}-\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\right\|^2+2D\delta_T\mathbf{v}^{\mathrm{T}}|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*|-D\delta_T\left\|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1\le 2D\delta_T\mathbf{v}^{\mathrm{T}}(|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*|+|\boldsymbol{\xi}^*|-|\widehat{\boldsymbol{\xi}}|),\ \text{ so}\\ D\delta_T\mathbf{v}^{\mathrm{T}}|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*|\le 4D\delta_T\mathbf{v}_J^{\mathrm{T}}|\widehat{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*|.\end{aligned}$$

It is easy to see that the left hand side is great than $\frac{D\delta_T}{|\widetilde{\xi}_{\widetilde{J},\max}|^k}\left\|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\right\|_1$, while the right hand side is less than $\frac{4D\delta_T}{|\widetilde{\xi}_{J,\min}|^k}\left\|\widehat{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\right\|_1$, where $\widetilde{\xi}_{\widetilde{J},\max}=\max_{j\in\widetilde{J}}\widetilde{\xi}_j$ and $\widetilde{\xi}_{J,\min}=\min_{j\in J}\widetilde{\xi}_j$. The remaining two inequalities for $\widehat{\boldsymbol{\xi}}$ follow immediately. □

Proof of Theorem 3. For $\boldsymbol{\alpha}$ such that $\left\|\boldsymbol{\alpha}_{J^c}\right\|_1\le c_0\left\|\boldsymbol{\alpha}_J\right\|_1$ with $n=|J|$, define $\epsilon=\left\|\widehat{\boldsymbol{\Sigma}}^*-\boldsymbol{\Sigma}\right\|_{\max}$,

$$|\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^*\boldsymbol{\alpha}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}|\le\epsilon\left\|\boldsymbol{\alpha}\right\|_1^2\le\epsilon(1+c_0)^2\left\|\boldsymbol{\alpha}_J\right\|_1^2\le\epsilon n(1+c_0)^2\left\|\boldsymbol{\alpha}_J\right\|^2,$$

so that by assumption A8,

$$\kappa(n)\left\|\boldsymbol{\alpha}_J\right\|\le\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}\right\|\le T^{-1/2}\left\|\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\alpha}\right\|+\epsilon^{1/2}n^{1/2}(1+c_0)\left\|\boldsymbol{\alpha}_J\right\|.\tag{8.11}$$

Put $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*$, so that Theorem 2 implies that $\|\boldsymbol{\alpha}_{J^c}\|_1 \le c_0 \|\boldsymbol{\alpha}_J\|_1$ as $c_0 > 3$. Suppose $\epsilon = O(\lambda_T)$ (to be proved later), and using

$$\frac{1}{2T}\|\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}} - \mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\|^2 \le 4D\delta_T \|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|_1$$

which is an intermediate result from the proof of Theorem 2, we can apply (8.11) to have, on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$,

$$\begin{aligned}
\kappa(n)\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| &\le T^{-1/2}\|\mathbf{M}_{\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*)\| + \epsilon^{1/2}n^{1/2}(1+c_0)\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| \\
&\le T^{-1/2}\|M_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}} - \mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\| + T^{-1/2}\|\mathbf{X}_{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\widetilde{\boldsymbol{\xi}}_2\| + \epsilon^{1/2}n^{1/2}(1+c_0)\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| \\
&\le 2\sqrt{2}D^{1/2}\delta_T^{1/2}\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|_1^{1/2} + T^{-1/2}\left\| 2\|\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*\|_1 \max_{\substack{1\le t\le T \\ 1\le i\le N,\ 1\le k\le K}} |X_{t,ik}|\mathbf{1}_{TN} \right\| \\
&\quad + \epsilon^{1/2}n^{1/2}(1+c_0)\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| \\
&\le 2\sqrt{2}D^{1/2}\delta_T^{1/2}n^{1/4}\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|^{1/2} + h_{1,N,T} + h_{2,N,T}\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1 + h_{3,N,T}\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| \\
&\le 2\gamma_T^{1/2}n^{1/4}\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|^{1/2} + ((1+c_0)n^{1/2}h_{2,N,T} + h_{3,N,T})\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| + h_{1,N,T},
\end{aligned}$$

where $\mathbf{1}_{TN}$ is a vector of ones of size TN, and we used the result in Lemma 2 such that

$$\begin{aligned}
h_{1,N,T} &= 2a_1(3/D_2\log(T\vee N))^{1/q}N^{-1/2}\lambda_T\delta_T^{1/2}(s_2 + N^{\frac{1}{2}+\frac{1}{2w}}), \\
h_{2,N,T} &= 2a_2(3/D_2\log(T\vee N))^{1/q}N^{-1/2}, \quad h_{3,N,T} = \epsilon^{1/2}n^{1/2}(1+c_0).
\end{aligned}$$

With $\epsilon = O(\lambda_T)$ assumed, the explicit rates assumed in Theorem 3 ensure that $h_{1,N,T}, n^{1/2}h_{2,N,T}$ and $h_{3,N,T}$ are all going to 0, with $h_{1,N,T} = o(\gamma_T n^{1/2})$. Hence solving the above quadratic inequality for $\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|^{1/2}$,

$$\begin{aligned}
\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|^{1/2} &\le \frac{\gamma_T^{1/2}n^{1/4} + \left[\gamma_T n^{1/2} + \kappa(n)h_{1,N,T}\right]^{1/2}}{\kappa(n) - (1+c_0)n^{1/2}h_{2,N,T} - h_{3,N,T}}, \text{ so that} \\
\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| &\le \frac{4\gamma_T n^{1/2} + 4\kappa(n)h_{1,N,T}}{(\kappa(n) - (1+c_0)n^{1/2}h_{2,N,T} - h_{3,N,T})^2} \le \frac{5\gamma_T n^{1/2}}{\kappa^2(n)}
\end{aligned}$$

for large enough N, T, which is the inequality for $\widetilde{\boldsymbol{\xi}}$.

To prove the inequality for $\widehat{\boldsymbol{\xi}}$, first note that for large enough N, T,

$$\begin{aligned}
|\widetilde{\boldsymbol{\xi}}_{J,\min}| &\ge |\boldsymbol{\xi}^*_{J,\min}| - |\widetilde{\boldsymbol{\xi}}_{J,\min} - \boldsymbol{\xi}^*_{J,\min}| \ge |\boldsymbol{\xi}^*_{J,\min}| - \|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| \\
&\ge |\boldsymbol{\xi}^*_{J,\min}| - (1-2^{-k})|\boldsymbol{\xi}^*_{J,\min}| = 2^{-k}|\boldsymbol{\xi}^*_{J,\min}|,
\end{aligned}$$

so that $|\widetilde{\boldsymbol{\xi}}_{J,\min}|^k \ge |\boldsymbol{\xi}^*_{J,\min}|^k/2$. Hence using the result in Theorem 2 for $\widehat{\boldsymbol{\xi}}$,

$$\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1 \le \frac{4|\widetilde{\boldsymbol{\xi}}_{\widetilde{J}\max}|^k}{|\boldsymbol{\xi}^*_{J,\min}|^k/2}\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|_1 \le \frac{8}{|\boldsymbol{\xi}^*_{J,\min}|^k}\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|_1 = (1+c_0)\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|_1,$$

so that $\|\widehat{\boldsymbol{\xi}}_{J^c} - \boldsymbol{\xi}^*_{J^c}\|_1 \le c_0 \|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|_1$. Then using an intermediate result

$$\frac{1}{2T}\|\mathbf{M}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\xi}} - \mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*\|^2 \le 4D\delta_T \mathbf{v}_J^{\mathrm{T}}|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J| \le \frac{4D\delta_T}{|\widetilde{\boldsymbol{\xi}}_{J,\min}|^k}\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|_1,$$

which is from the proof of Theorem 2, putting $\boldsymbol{\alpha} = \widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*$ in (8.11), we have on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$,

$$\kappa(n)\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| \le \frac{2\gamma_T^{1/2} n^{1/4}}{|\widetilde{\boldsymbol{\xi}}_{J,\min}|^{k/2}}\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|^{1/2} + ((1+c_0)n^{1/2}h_{2,N,T} + h_{3,N,T})\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| + h_{1,N,T}.$$

Solving for $\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|^{1/2}$ as before and squaring, we obtain

$$\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J\| \le \frac{4\gamma_T n^{1/2}|\widetilde{\boldsymbol{\xi}}_{J,\min}|^{-k} + 4\kappa(n)h_{1,N,T}}{(\kappa(n) - (1+c_0)n^{1/2}h_{2,N,T} - h_{3,N,T})^2} \le \frac{5\gamma_T n^{1/2}}{\kappa^2(n)|\boldsymbol{\xi}^*_{J,\min}|^k}$$

for large enough N, T, which is the inequality for $\widehat{\boldsymbol{\xi}}$. The bounds for $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are obtained by using the results in Lemma 2 and Theorem 2, and substituting the error upper bounds we just proved. It remains to show that $\epsilon = O(\lambda_T)$.

We can easily see that, for $x_{t,j}^{\mathrm{T}}$ the j-th row of $\mathbf{X}_t$,

$$\epsilon = \|\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}\|_{\max} = \max_{1\le i,j\le N}\Bigg\{\Big|T^{-1}\sum_{t=1}^T y_{ti}y_{tj} - E(y_{ti}y_{tj})\Big|, \Big|\boldsymbol{\beta}^{*\mathrm{T}}\Big(T^{-1}\sum_{t=1}^T y_{ti}\mathbf{x}_{t,j} - E(y_{ti}\mathbf{x}_{t,j})\Big)\Big|, \Big|\boldsymbol{\beta}^{*\mathrm{T}}\Big(T^{-1}\sum_{t=1}^T \mathbf{x}_{t,i}\mathbf{x}_{t,j}^{\mathrm{T}} - E(\mathbf{x}_{t,i}\mathbf{x}_{t,j}^{\mathrm{T}})\Big)\boldsymbol{\beta}^*\Big|\Bigg\}.$$

The largest upper bound is given by $\max_{1\le i,j\le N}|T^{-1}\sum_{t=1}^T y_{ti}y_{tj} - E(y_{ti}y_{tj})|$ (details omitted), where using $y_{ti} = \boldsymbol{\pi}_{1,i}^{*\mathrm{T}}\mathbf{W}_2^*\mathbf{X}_t\boldsymbol{\beta}^* + \boldsymbol{\pi}_{1,i}^{*\mathrm{T}}\boldsymbol{\epsilon}_t$ (see (2.2), with $\boldsymbol{\pi}_{1,i}^{*\mathrm{T}}$ the i-th row of $\boldsymbol{\Pi}_1^*$),

$$\begin{aligned}
\Big|T^{-1}\sum_{t=1}^T y_{ti}y_{tj} - E(y_{ti}y_{tj})\Big| &\le \|T^{-1}\sum_{t=1}^T \mathbf{X}_t\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{X}_t^T - E(\mathbf{X}_t\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{X}_t^T)\|_{\max}\cdot\|\mathbf{W}_2^{*\mathrm{T}}\boldsymbol{\pi}^*_{1,i}\|_1^2 \\
&\quad + 2\|T^{-1}\sum_{t=1}^T \mathbf{X}_t\boldsymbol{\beta}^*\boldsymbol{\epsilon}_t^{\mathrm{T}} - E(\mathbf{X}_t\boldsymbol{\beta}^*\boldsymbol{\epsilon}_t^{\mathrm{T}})\|_{\max}\cdot\|\mathbf{W}_2^{*\mathrm{T}}\boldsymbol{\pi}^*_{1,i}\|_1\|\boldsymbol{\pi}_{1,i}\|_1 \\
&\quad + \|T^{-1}\sum_{t=1}^T \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}} - E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}})\|_{\max}\cdot\|\boldsymbol{\pi}_{1,i}\|_1^2 \\
&\le \frac{4\lambda_T\|\boldsymbol{\beta}^*\|_1^2}{(1-\eta)^2} + \frac{4\lambda_T\|\boldsymbol{\beta}^*\|_1}{(1-\eta)^2} + \frac{\lambda_T}{(1-\eta)^2} = \frac{\lambda_T(2\|\boldsymbol{\beta}^*\|_1 + 1)^2}{(1-\eta)^2},
\end{aligned}$$

since it is on $\mathcal{A}_1 \cap \cdots \mathcal{A}_4 \cap \mathcal{M}$. Hence $\epsilon = O(\lambda_T)$. This completes the proof of the theorem. □

Proof of Theorem 5. First, similar to (8.11), we can use assumption A8 for $\|\boldsymbol{\alpha}_{J^c}\|_1 \le c_0\|\boldsymbol{\alpha}_J\|_1$ to arrive at $\kappa(n)\|\boldsymbol{\alpha}_{J^c}\| \le T^{-1/2}\|\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\alpha}\| + \epsilon^{1/2}n^{1/2}(1+c_0)\|\boldsymbol{\alpha}_J\|$. Putting $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*$ and follow the same lines as in the proof of Theorem 3, we can use $\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\| = O(\gamma_T n^{1/2})$ on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$ (by the result of

Theorem 3) to show that, for $j \in J^c$,

$$\widetilde{\xi}_j \le \big\|\widetilde{\boldsymbol{\xi}}_{J^c}\big\| = \big\|\widetilde{\boldsymbol{\xi}}_{J^c} - \boldsymbol{\xi}^*_{J^c}\big\| = O(\gamma_T n^{1/2}). \tag{8.12}$$

Define the set $D = \{j : \xi^*_j \text{ does not corr. to diagonal elements of } \mathbf{W}^*_1, \mathbf{W}^*_2\}$. The KKT condition implies that $\widehat{\boldsymbol{\xi}}$ is a solution to (3.3) if and only if there exists a subgradient

$$\mathbf{g} = \partial(\mathbf{v}^{\mathrm{T}}|\widehat{\boldsymbol{\xi}}|) = \left\{ \mathbf{g} \in \mathbb{R}^{2N^2} : \left\{ \begin{array}{ll} g_i = 0, & i \in D^c; \\ g_i = v_i \mathrm{sign}(\widehat{\xi}_i), & \widehat{\xi}_i \neq 0; \\ |g_i| \le v_i, & \text{otherwise.} \end{array} \right. \right\}$$

such that, differentiating the expression to be minimized in (3.3) with respect to $\boldsymbol{\xi}_D$,

$$T^{-1}\widehat{\mathbf{M}}_D^{\mathrm{T}}\widehat{\mathbf{M}}_D\widehat{\boldsymbol{\xi}}_D - T^{-1}\widehat{\mathbf{M}}^{\mathrm{T}}\mathbf{y} + \gamma_T\mathbf{g}_D + T^{-1}\widehat{\mathbf{M}}_D^{\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}}\mathrm{vec}(\mathbf{I}_N) = \mathbf{0},$$

where we denote $\widehat{\mathbf{M}} = \mathbf{M}_{\widehat{\boldsymbol{\beta}}}$ and $\mathbf{M}^* = \mathbf{M}_{\boldsymbol{\beta}^*}$. Substituting $\mathbf{y} = \mathbf{M}^*_D\boldsymbol{\xi}^*_D + \mathbf{X}_{\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) + \boldsymbol{\epsilon}$,

$$\widehat{\boldsymbol{\Sigma}}_{DD}\widehat{\boldsymbol{\xi}}_D - T^{-1}\widehat{\mathbf{M}}_D^{\mathrm{T}}\mathbf{M}^*_D\boldsymbol{\xi}^*_D + T^{-1}\widehat{\mathbf{M}}_D^{\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) - T^{-1}\widehat{\mathbf{M}}_D^{\mathrm{T}}\boldsymbol{\epsilon} = -\gamma_T\mathbf{g}_D,$$

where $\widehat{\boldsymbol{\Sigma}} = T^{-1}\widehat{\mathbf{M}}^{\mathrm{T}}\widehat{\mathbf{M}}$. For sign consistency of $\widehat{\boldsymbol{\xi}}$, we have $\widehat{\boldsymbol{\xi}}_{J^c\cap D} = \mathbf{0}$ and $\mathrm{sign}(\widehat{\boldsymbol{\xi}}_J) = \mathrm{sign}(\boldsymbol{\xi}^*_J)$. Then it is easy to see that $\widehat{\boldsymbol{\xi}}$ is a sign consistent solution if and only if $\mathrm{sign}(\widehat{\boldsymbol{\xi}}_J) = \mathrm{sign}(\boldsymbol{\xi}^*_J)$ and

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{JJ}\widehat{\boldsymbol{\xi}}_J - T^{-1}\widehat{\mathbf{M}}_J^{\mathrm{T}}\mathbf{M}^*_J\boldsymbol{\xi}^*_J + T^{-1}\widehat{\mathbf{M}}_J^{\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) - T^{-1}\widehat{\mathbf{M}}_J^{\mathrm{T}}\boldsymbol{\epsilon} &= -\gamma_T\mathbf{g}_J; \\ |\widehat{\boldsymbol{\Sigma}}_{J'J}\widehat{\boldsymbol{\xi}}_J - T^{-1}\widehat{\mathbf{M}}_{J'}^{\mathrm{T}}\mathbf{M}^*_J\boldsymbol{\xi}^*_J + T^{-1}\widehat{\mathbf{M}}_{J'}^{\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) - T^{-1}\widehat{\mathbf{M}}_{J'}^{\mathrm{T}}\boldsymbol{\epsilon}| &\le \gamma_T\mathbf{v}_{J'}, \end{aligned}$$

where $J' = J^c \cap D$. Recall from assumption A8 that $\widehat{\boldsymbol{\Sigma}}^* = T^{-1}\mathbf{M}^{*\mathrm{T}}\mathbf{M}^*$ and $\boldsymbol{\Sigma} = E(\widehat{\boldsymbol{\Sigma}}^*)$. Rearranging, these yield

$$\mathrm{sign}(\widehat{\boldsymbol{\xi}}_J) = \mathrm{sign}\big\{\boldsymbol{\xi}^*_J + I_1 + I_2 + I_3 + I_4 + I_5\big\} = \mathrm{sign}(\boldsymbol{\xi}^*_J); \tag{8.13}$$

$$|L_1 + L_2 + L_3 + L_4 + L_5| \le \gamma_T\mathbf{v}_{J'} \tag{8.14}$$

as the necessary and sufficient conditions for $\widehat{\boldsymbol{\xi}}$ to be a sign consistent solution to (3.3), where

$$\begin{aligned} I_1 &= -\boldsymbol{\Sigma}_{JJ}^{-1}\big[T^{-1}(\widehat{\mathbf{M}}_J - \mathbf{M}^*_J)^{\mathrm{T}}(\widehat{\mathbf{M}}_J\widehat{\boldsymbol{\xi}}_J - \mathbf{M}^*_J\boldsymbol{\xi}^*_J)\big], & I_2 &= -\boldsymbol{\Sigma}_{JJ}^{-1}\big[T^{-1}\mathbf{M}^{*\mathrm{T}}_J(\widehat{\mathbf{M}}_J - \mathbf{M}^*_J)\widehat{\boldsymbol{\xi}}_J\big], \\ I_3 &= -\boldsymbol{\Sigma}_{JJ}^{-1}(\widehat{\boldsymbol{\Sigma}}^*_{JJ} - \boldsymbol{\Sigma}_{JJ})(\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J), & I_4 &= -\boldsymbol{\Sigma}_{JJ}^{-1}\big[T^{-1}\widehat{\mathbf{M}}_J^{\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N)\big], \\ I_5 &= \boldsymbol{\Sigma}_{JJ}^{-1}\big[T^{-1}\widehat{\mathbf{M}}_J^{\mathrm{T}}\boldsymbol{\epsilon} - \gamma_T\mathbf{g}_J\big], & D_1 &= T^{-1}(\widehat{\mathbf{M}}_{J'} - \mathbf{M}^*_{J'})^{\mathrm{T}}(\widehat{\mathbf{M}}_J - \mathbf{M}^*_J)\widehat{\boldsymbol{\xi}}_J, \\ D_2 &= T^{-1}(\widehat{\mathbf{M}}_{J'} - \mathbf{M}^*_{J'})^{\mathrm{T}}\mathbf{M}^*_J(\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J), & D_3 &= T^{-1}\mathbf{M}^{*\mathrm{T}}_{J'}(\widehat{\mathbf{M}}_J - \mathbf{M}^*_J)\widehat{\boldsymbol{\xi}}_J, \\ D_4 &= \widehat{\boldsymbol{\Sigma}}^*_{J'J}(\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J), & D_5 &= T^{-1}\widehat{\mathbf{M}}_{J'}^{\mathrm{T}}(\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) - \boldsymbol{\epsilon}). \end{aligned}$$

We first prove that $\big\|\boldsymbol{\Sigma}_{JJ}^{-1}\big\|_\infty \le C$ on $\mathcal{A}\cap\cdots\cap\mathcal{A}_4\cap\mathcal{M}$ for some constant C. To this end, denote $\mathbf{X}^* = \mathbf{X}_{\boldsymbol{\beta}^*}$, and consider the partition $\boldsymbol{\Sigma}_{JJ} = (\mathbf{A}_{ij})_{1\le i,j\le 2}$. Then

$$\mathbf{A}_{11} = E(T^{-1}\mathbf{Z}_J^{\mathrm{T}}\mathbf{Z}_J), \quad \mathbf{A}_{12} = \mathbf{A}_{21}^{\mathrm{T}} = E(T^{-1}\mathbf{Z}_J^{\mathrm{T}}\mathbf{X}^*_J), \quad \mathbf{A}_{22} = E(T^{-1}\mathbf{X}^{*\mathrm{T}}_J\mathbf{X}^*_J).$$

Assumption A1 implies that there are finite number of non-zeros in each row of $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$. Let n_r be the maximum number of non-zeros in a row of $\mathbf{W}_1^*$ or $\mathbf{W}_2^*$. Then n_r is a constant, and each block diagonal $\mathbf{A}_{ij}$ defined above has at most n_r non-zeros in each row. Using the inverse formula of partitioned matrix, we thus have

$$\begin{aligned}
\big\|\mathbf{\Sigma}_{JJ}^{-1}\big\|_\infty &\le \big\|(\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\big\|_\infty + \big\|\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\big\|_\infty \\
&\le n_r\lambda_{\max}\{(\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\} \\
&\quad + n_r\lambda_{\max}(\mathbf{A}_{11}^{-1})\cdot\big\|\mathbf{A}_{12}\big\|_\infty\cdot n_r\lambda_{\max}\{(\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\} \\
&\le n_r\lambda_{\max}(\mathbf{\Sigma}_{JJ}^{-1}) + n_r^3\lambda_{\max}^2(\mathbf{\Sigma}_{JJ}^{-1})\big\|\mathbf{A}_{12}\big\|_{\max} \\
&\le \frac{n_r}{u}+\frac{n_r^3}{u^2}(\sigma_{\max}^2+\lambda_T)(2\big\|\boldsymbol{\beta}^*\big\|_1+1)^2(1-\eta)^{-2}\le C,
\end{aligned}$$

where we use the last part of the proof of Theorem 3 and assumption A4 (details omitted) to arrive at, on $\mathbf{A}_1\cap\cdots\cap\mathcal{A}_4\cap\mathcal{M}$,

$$\big\|\mathbf{A}_{12}\big\|_{\max}\le(\sigma_{\max}^2+\lambda_T)(2\big\|\boldsymbol{\beta}^*\big\|_1+1)^2(1-\eta)^{-2},$$

and the assumption of uniform boundedness, say $\lambda_{\min}(\mathbf{\Sigma}_{JJ})>u>0$ uniformly.

For proving (8.13), it suffices to show that $\big\|I_j\big\|_\infty=o(1)$ since by assumption A1, ξ_j^* is a constant for $j\in J$. Consider

$$\begin{aligned}
\big\|\mathbf{I}_1\big\|_\infty &\le \big\|\mathbf{\Sigma}_{JJ}^{-1}\big\|_\infty\cdot(\big\|T^{-1}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}^{\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}\big\|_\infty\cdot\big\|\widehat{\boldsymbol{\xi}}_{2,J}\big\|_{\max}+\big\|T^{-1}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}^{\mathrm{T}}\mathbf{M}_J^*\big\|_\infty\cdot\big\|\widehat{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\big\|_{\max}) \\
&\le C\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1(\sigma_{\max}^2+\lambda_T)\left\{n_r\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1(1+\big\|\widehat{\boldsymbol{\xi}}_{2,J}-\boldsymbol{\xi}_{2,J}^*\big\|)+\frac{4n_r\big\|\boldsymbol{\beta}^*\big\|_1}{1-\eta}\big\|\widehat{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\big\|\right\} \\
&= O\left(\frac{s_2\lambda_T\gamma_T^{1/2}+\gamma_T n}{N}\cdot\left(\frac{s_2\lambda_T\gamma_T^{1/2}+\gamma_T n}{N}+\gamma_T n^{1/2}\right)\right)=O\left(\frac{\gamma_T^2n^2}{N^2}+\frac{\gamma_T^2n^{3/2}}{N}\right)=o(1),
\end{aligned}$$

where we used the rates assumed in Theorem 2, the last part of the proof of Theorem 3 for the rates of $\big\|T^{-1}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}^{\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}\big\|_\infty$ and $\big\|T^{-1}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}^{\mathrm{T}}\mathbf{M}_J^*\big\|_\infty$ (details omitted, but we also used the fact that these two matrices are of block diagonal structure with at most $2n_r$ non-zero entries in each row), and the results of Theorem 3 for the rates of $\big\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\big\|_1$ and $\big\|\widehat{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\big\|$. We also used $n\le 2n_rN$, so that $\gamma_T n/N\le 2n_r\gamma_T=o(1)$. Similarly, on $\mathcal{A}_1\cap\cdots\mathcal{A}_4\cap\mathcal{M}$,

$$\begin{aligned}
&\big\|I_2\big\|_\infty\le C\big\|T^{-1}\mathbf{M}_J^{*\mathrm{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}\big\|_\infty\big\|\widehat{\boldsymbol{\xi}}_{2,J}\big\|_{\max}=O\left(\frac{2n_rs_2\lambda_T\gamma_T^{1/2}+2n_r\gamma_T n}{N}\right)=O\left(\frac{\gamma_T n}{N}\right)=o(1);\\
&\big\|I_3\big\|_\infty\le C\big\|\widehat{\mathbf{\Sigma}}_{JJ}^*-\mathbf{\Sigma}_{JJ}\big\|_\infty\big\|\widehat{\boldsymbol{\xi}}_J-\boldsymbol{\xi}_J^*\big\|_{\max}=O(2n_r\lambda_T\gamma_T n^{1/2})=o(\lambda_T\gamma_T^{\frac{1}{k+1}})=o(1);\\
&\big\|I_4\big\|_\infty\le C\left\|\begin{pmatrix}T^{-1}\sum_{t=1}^T\mathbf{y}_t(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}\\ T^{-1}\sum_{t=1}^T\mathbf{X}_t\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}\end{pmatrix}\right\|_{\max}=O\left(\frac{s_2\lambda_T\gamma_T^{1/2}+\gamma_T n}{N}\right)=O\left(\frac{\gamma_T n}{N}\right)=o(1);\\
&\big\|I_5\big\|_\infty\le C\left(\big\|T^{-1}\widehat{\mathbf{M}}^{\mathrm{T}}\boldsymbol{\epsilon}\big\|_{\max}+\frac{\gamma_T}{|\widetilde{\boldsymbol{\xi}}_{J,\min}|^k}\right)=O(\gamma_T^{1/2}(\lambda_T+\gamma_T^{1/2})+\gamma_T)=O(\gamma_T)=o(1),
\end{aligned}$$

Hence we have proved (8.13) on $\mathcal{A}_1\cap\cdots\mathcal{A}_4\cap\mathcal{M}$ when N,T are large enough.

For proving (8.14) on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$ when N, T are large enough, it suffices to show by (8.12) that

$$\big\|D_j\big\|_\infty \le \gamma_T / \max_{j \in J^c} |\widetilde{\xi}_j|^k = o\left(\gamma_T/(\gamma_T n^{1/2})^k\right).$$

To show this, consider on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$,

$$\begin{aligned}
\big\|D_1\big\|_\infty &\le \big\|T^{-1}\mathbf{X}^{\mathrm{T}}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J'}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J}\big\|_\infty \big\|\widehat{\boldsymbol{\xi}}_{2,J}\big\|_{\max} \le (\sigma^2_{\max} + \lambda_T) n_r \big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\big\|_1^2 (1 + \big\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\big\|) \\
&= O\left(\frac{\gamma_T^2 n^2}{N^2}\right); \\
\big\|D_2\big\|_\infty &\le \big\|T^{-1}\mathbf{X}^{\mathrm{T}}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,J'}\mathbf{M}^*_J\big\|_\infty \big\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\big\|_{\max} = O\left(\frac{\gamma_T n}{N} \cdot \gamma_T n^{1/2}\right) = O\left(\frac{\gamma_T^2 n^{3/2}}{N}\right); \\
\big\|D_3\big\|_\infty &\le \big\|T^{-1}\mathbf{M}^{*\mathrm{T}}_{J'}\mathbf{X}_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}*,J}\big\|_\infty \big\|\widehat{\boldsymbol{\xi}}_J\big\|_{\max} = O\left(\frac{\gamma_T n}{N}\right); \\
\big\|D_4\big\|_\infty &\le (\big\|\widehat{\boldsymbol{\Sigma}}_{J'J} - \boldsymbol{\Sigma}_{J'J}\big\|_\infty + \big\|\boldsymbol{\Sigma}_{J'J}\big\|_\infty) \big\|\widehat{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\big\|_{\max} = O(\gamma_T n^{1/2}); \\
\big\|D_5\big\|_\infty &\le O\left(\frac{\gamma_T n}{N} + \gamma_T\right).
\end{aligned}$$

The largest order is $\big\|D_4\big\|_\infty = O(\gamma_T n^{1/2})$, which is of smaller order than $\gamma_T/(\gamma_T n^{1/2})^k$ by the assumption $n = o\Big(\gamma_T^{-\frac{2k}{k+1}}\Big)$. This proves (8.14), and completes the proof of the theorem. □

References

Andrews, D. (1984). Nonstrong mixing autoregressive processes. *J. Appl. Probab. 21*(4), 930–934.

Anselin, L. (2002). Under the hood. issues in the specification and interpretation of spatial regression models. *Agric. Econ. 27*(3), 247–267.

Arbia, G. and B. Fingleton (2008). New spatial econometric techniques and applications in regional science. *Papers in Regional Science 87*(3), 311–317.

Bavaud, F. (1998). Models for spatial weights: A systemic look. *Geographical Analysis 30*, 153–171.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist. 37*(4), 1705–1732.

Chen, X., M. Xu, and W. B. Wu (2013). High-dimensional covariance estimation for time series. *Manuscript*.

Corrado, L. and B. Fingleton (2011, January). Where is the economics in spatial econometrics? Working Papers 1101, University of Strathclyde Business School, Department of Economics.

Dicker, L., B. Huang, and X. Lin (2010). Variable selection and estimation with the seamless-l0 penalty. *Working Paper*.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*(2), 407–499.

Elhorst, J. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review 26*(3), 244–268.

Fan, J. and J. Lv (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*, 5467–5484.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Irwin, E. G. and J. Geoghegan (2001). Theory, data, methods: developing spatially explicit economic models of land use change. *Agriculture, Ecosystems and Environment 85*, 7–23.

Kapoor, M., H. H. Kelejian, and I. R. Prucha (2007). Panel data models with spatially correlated error components. *Journal of Econometrics 140*, 97–130.

Lesage, J. and R.-K. Pace (2009). *Introduction to Spatial Econometrics*. New York: CRC Press.

Lesage, J. and W. Polasek (2008). Incorporating transportation network st ructure i n spatial econometric models of commodity flows. *Spatial Economic Analysis 3*(2), 225–245.

Li, Y. (2003). A martingale inequality and large deviations. *Statistics & Probability Letters 62*, 317–321.

Liu, W., H. Xiao, and W. Wu (2013). Probability and moment inequalities under dependence. *Statistica Sinica*. To appear.

Pinkse, J., M. E. Slade, and C. Brett (2002). Spatial price competition: A semiparametric appraoch. *Econometrica 70*(3), 1111–1153.

Shao, X. (2010). Nonstationarity-extended whittle estimation. *Econometric Theory 26*, 1060–1087.

Wold, H. (1953). *Demand Analysis: A Study in Econometrics*. New York: Wiley.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA 102*, 14150–14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *STATISTICS AND ITS INTERFACE 4*, 207–226.

Yao, Q. and P. Brockwell (2006). Gaussian maximum likelihood estimation for arma models ii: Spatial processes. *Bernoulli 12*(3), 403–429.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zhou, S., S. van de Geer, and P. Bühlmann (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. arXiv:0903.2515v1.

Zhou, Z. (2010). Nonparametric inference of quantile curves for nonstationary time series. *Ann. Statist. 38*(4), 2187–2217.

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