Regularization for Spatial Panel Time Series Using the Adaptive LASSO

Clifford Lam and Pedro CL Souza

Department of Statistics and 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 such a structure is usually assumed before further analysis. We propose to estimate this by penalizing the elements in the spatial weight matrices using the adaptive LASSO proposed by Zou (2006). Non-asymptotic oracle inequalities and the asymptotic sign consistency of the estimators are proved when the dimension of the time series can be larger than the sample size, and they tend to infinity jointly. Asymptotic normality of the LASSO/adaptive LASSO estimator for the model regression parameter is also presented. All the proofs involve non-standard analysis of LASSO/adaptive LASSO estimators, since our model, albeit like a standard regression, always has the response vector as one of the covariates. A block coordinate descent algorithm is introduced, with simulations and a real data analysis carried out to demonstrate the performance of our estimators.

Key words and phrases: Spatial econometrics, adaptive LASSO, sign consistency, asymptotic normality, non-asymptotic oracle inequalities, spatial weight matrices

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 (Elhorst, 2003). Spatial autoregressive models (SAR) can be seen as another formulation of a spatial error model (Lesage and Pace, 2009).

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 (Lesage and Polasek, 2008), to imposing special structures. For example, the contiguity structure has contagious regions having corresponding elements in the spatial weight matrix set to one and zero otherwise (Lesage and Pace, 2009). 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 spatial 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 (2.2). Few researchers attempted to estimate the spatial 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 matrix $\mathbf{W}_1^*$ is estimated as $\widehat{g}(d_{ij})$, where d_{ij} is a distance measure specified by the user. Beenstock and Felsenstein (2012) suggested using a moment estimator for the spatial weight matrix. Bhattacharjee and Jensen-Butler (2013) proposes to estimate the spatial weight matrix by first estimating the error covariance matrix. However, estimating a large error covariance matrix can be inaccurate as the dimension of the panel is large and can be close to the sample size - one of the major characteristics of a large time series panel. 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 spatial 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 spatial 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 spatial 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 a partial solution to the criticism and potential sensitivity towards "arbitrary" specification of these matrices if they themselves can be estimated from data as well. In fact in Lemma 2, we have specified how the error upper bound for the estimation of $\boldsymbol{\beta}^*$ in model (2.2) is related to the error of the estimated/assumed spatial weight matrices. This result sheds some lights on the potential seriousness of wrongly specifying the spatial 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 spatial 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 the supplementary article Lam and Souza (2014).

2 The Model

A commonly used model for describing spatial interaction in a panel of time series is the spatial lag model,

$$\mathbf{y}_t = \rho \mathbf{W} \mathbf{y}_t + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T. \tag{2.1}$$

See equation (19.5) of Anselin et al. (2006) for instance, which is a stacked version of the above. Here, $\mathbf{y}_t$ is an $N \times 1$ vector of response variables, and $\mathbf{X}_t$ is an $N \times K$ matrix of exogenous covariates. The spatial weight matrix $\mathbf{W}$ has elements that express the strength of interaction between location i (row) and j (column). Therefore, $\mathbf{W}$ can be interpreted as the presence and strength of a link between nodes (the observations) in a network representation that matches the spatial weights structure (Anselin et al., 2006). In this paper, such a structure is assumed to be constant across time points $t = 1, \ldots, T$, so that $\mathbf{W}$ remains constant for $t = 1, \ldots, T$. The parameter ρ is called the spatial autoregressive coefficient.

However, to utilize model (2.1), the spatial weight matrix $\mathbf{W}$ has to be specified. As briefly stated in the Introduction, estimation accuracy of model parameters can crucially depend on the correct specification of $\mathbf{W}$. Moreover, Plümper and Neumayer (2010) points out that a common practice of row-standardization in the specification of $\mathbf{W}$ in model (2.1) is in fact problematic, since it alters not only the metric or unit of the spatial lag, but also the relative weight given to the observations.

With all these considerations, we consider a more general form of the spatial lag model,

$$\mathbf{y}_t = \mathbf{W}_1^* \mathbf{y}_t + \mathbf{W}_2^* \mathbf{X}_t \boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{2.2}$$

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 $N \times N$ spatial weight matrices to be estimated, $\mathbf{X}_t$ is an $N \times K$ matrix of centered exogenous variables at time t, $\boldsymbol{\beta}^*$ 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 spatial 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

$$\mathbf{y}_t = \mathbf{\Pi}_1^*\mathbf{W}_2^*\mathbf{X}_t\boldsymbol{\beta}^* + \mathbf{\Pi}_1^*\boldsymbol{\epsilon}_t, \quad \mathbf{\Pi}_1^* = (\mathbf{I}_N - \mathbf{W}_1^*)^{-1}, \tag{2.3}$$

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 spatial 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 spatial 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 (2.2) 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 (Yao and Brockwell, 2006)

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

$$\mathbf{y}_t = \rho\mathbf{W}\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} - \rho\mathbf{W}\mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t = (\mathbf{I}_N + \lambda\mathbf{W}')\mathbf{v}_t$. Model (2.2) 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 $\mathbf{\Sigma}_\epsilon = (\mathbf{I}_N + \lambda\mathbf{W}')\mathrm{var}(\mathbf{v}_t)(\mathbf{I}_N + \lambda(\mathbf{W}')^{\mathrm{T}})$. In the LASSO /adaptive LASSO estimator to be introduced in the next section, we can certainly restrict $\mathbf{W}_2 = \mathbf{I}_N - \mathbf{W}_1$ in order to estimate the spatial weight matrix of such a spatial ARMA error model. This is however not pursued in the present paper. 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 Spatial Weight Matrices

The spatial 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 spatial 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:

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

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 $\mathbf{y}_t = (\mathrm{POP}_{1t}, \ldots, \mathrm{POP}_{Nt}, \mathrm{TAX}_{1t}, \ldots, \mathrm{TAX}_{Nt})^{\mathrm{T}}$ 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

$$\mathbf{X}_t = \begin{pmatrix}
\mathrm{EMP}_{1t} & \mathrm{PUBS}_{1t} & \mathrm{INC}_{1t} & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathrm{EMP}_{Nt} & \mathrm{PUBS}_{Nt} & \mathrm{INC}_{Nt} & 0 & 0 & 0 \\
0 & 0 & 0 & \mathrm{EMP}_{1t} & \mathrm{PUBS}_{1t} & \mathrm{INC}_{1t} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \mathrm{EMP}_{1t} & \mathrm{PUBS}_{Nt} & \mathrm{INC}_{Nt}
\end{pmatrix},$$

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

and finally $\boldsymbol{\beta}^* = (\beta_1, \beta_2, 0, 0, \gamma_1, \gamma_2)^{\mathrm{T}}$. Thus both matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are very sparse in this model. Rather than fixing the spatial weight matrices, their sparse estimation 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 (2.3) 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 spatial 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 spatial 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)

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

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 (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 (2.2) 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 $\mathbf{y} = \text{vec}\{(\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\text{T}}\}$, $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\text{T}}$, $\mathbf{X}_{\boldsymbol{\beta}^*} = \mathbf{I}_N \otimes \{(\mathbf{I}_T \otimes \boldsymbol{\beta}^{*\text{T}})(\mathbf{X}_1, \ldots, \mathbf{X}_T)^{\text{T}}\}$, $\boldsymbol{\xi}_j^* = \text{vec}(\mathbf{W}_j^{*\text{T}})$ 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 $\boldsymbol{\xi}^* = (\boldsymbol{\xi}_1^{*\text{T}}, \boldsymbol{\xi}_2^{*\text{T}})^{\text{T}}$ 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. Hence, the theoretical analysis presented in this paper is not typical as in the LASSO under the standard linear regression setting, which is a well-studied topic.

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

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

where $\big\|\cdot\big\|_1$ 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

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

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

Both (3.2) and (3.3) are presented as a version of penalized least squares problem. Note however, that we do not assume the errors in the noise vector $\boldsymbol{\epsilon}_t$ are independent and identically distributed. It can be argued that the generalized least square estimator can improve in practical performance. However, the inverse of the sample covariance matrix of the estimated noise vector $\widehat{\boldsymbol{\epsilon}}_t$ is in general is too noisy in our setting for any real improvement to be seen. And when $N > T$, such an estimator of the inverse does not exist.

4 Properties of LASSO and adaptive LASSO Estimators

An ideal estimator for a spatial 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. We also present the asymptotic normality of the LASSO estimator $\widetilde{\boldsymbol{\beta}}$, which is useful for inference purpose.

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 $a > 0$,

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

where $j = 1, \ldots, NK$, $\ell = 1, \ldots, N$, and $x'_{tj} = f_j(\mathcal{F}'_t)$ with the filtration $\mathcal{F}'_t = (\ldots, \mathbf{e}_{x,-1}, \mathbf{e}'_{x,0}, \mathbf{e}_{x,1}, \ldots, \mathbf{e}_{x,t})$. The vector $\mathbf{e}'_{x,0}$ is 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. For a matrix $\mathbf{A}$, we define $\|\mathbf{A}\|_\infty = \max_i \sum_{j\geq 1} |A_{ij}|$.

A1. The entries in the 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 (2.2) 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 the variables $X_{t,jk}, X_{t,jk}^*, \zeta_{t,j}$ and $\zeta_{t,j}^*$ by the same positive constants D_1, D_2 and q.

A6. Let $\Theta^x_{m,a} = \sum_{t=m}^{\infty} \max_{1\le j\le NK} \theta^x_{t,a,j}$ and $\Theta^\zeta_{m,a} = \sum_{t=m}^{\infty} \max_{1\le j\le N} \theta^\zeta_{t,a,j}$, where $\theta^\zeta_{t,a,j} = \theta^\epsilon_{t,a,j}/\delta_T^{1/2}$. Then we assume $\Theta^x_{m,2w}, \Theta^\zeta_{m,2w} \le Cm^{-\alpha}$ 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 $k \le K$ 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 as demonstrated numerically in section 6. For moderate value of T, if the spatial weight matrices are sparse enough, then a slow decay rate is sufficient, which in practice means that the noise level is only required to be not too large. See also example 2 in section 4.2 for a simple illustration, and a remark therein about estimating the reduced form model (2.3) 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 $\mathbf{W}^*_2\mathbf{X}_t\boldsymbol{\beta}^*$ is then equals to $\sum_{j=1}^d \beta^*_j\mathbf{W}^*_2\mathbf{y}_{t-j} + \mathbf{W}^*_2\mathbf{z}_t\boldsymbol{\beta}^*_2$, where $\boldsymbol{\beta}^* = (\beta^*_1, \ldots, \beta^*_d, \boldsymbol{\beta}^{*T}_2)^T$. 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 $\boldsymbol{\Pi}^*_1$ is defined in (2.3), and B is the backward shift operator. For the inverse operator above to be defined (i.e. the system is stationary), we need

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

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 $\boldsymbol{\Sigma}_{xk}$ and $\boldsymbol{\Sigma}_\zeta$ 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 $\boldsymbol{\Sigma}_{xk}$ have their smallest eigenvalues uniformly bounded away from 0, and the cross covariance between the $\mathrm{cov}(\mathbf{X}_{t,k_1}, \mathbf{X}_{t,k_2})$ is not too strong for all $1 \le k_1 \ne k_2 \le K$.

4.2 Example 2

We demonstrate that the decay assumption A2 is needed in general for estimating the spatial 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 is 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 $\sigma^2_{\epsilon,j} = \mathrm{var}(\epsilon_{t,j})$, $\sigma^2_{X,j} = \mathrm{var}(X_{t,j})$. 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 $\mathrm{cov}(X_{t1}, X_{t2}) = \mathrm{cov}(\epsilon_{t1}, \epsilon_{t2}) = 0$, the bias can be calculated to be converging in probability:

$$\widehat{w}_{12} - w_{12} \xrightarrow{\mathbf{P}} \frac{w_{21}\sigma_{\epsilon,1}^2(1 - w_{12}w_{21})}{w_{21}^2(\sigma_{\epsilon,1}^2 + \sigma_{X,1}^2) + \sigma_{\epsilon,2}^2 + \sigma_{X,2}^2},$$

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

By symmetry of the formulae for the asymptotic biases of $\widehat{w}_{12}$ and $\widehat{w}_{21}$, we can easily see that if $\sigma_{\epsilon,1}^2$ and $\sigma_{\epsilon,2}^2$ 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 $\sigma_{X,1}^2 = \sigma_{X,2}^2 = 0$ (assumption A7 fails), and $\sigma_{\epsilon,1}^2 = \sigma_{\epsilon,2}^2$, we see that the asymptotic bias becomes independent of $\sigma_{\epsilon,j}^2$, 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 $\mathbf{X}_t$, although other exogenous covariates are still needed. See (4.3) in section 4.1 for more details. ■

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

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

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. We are presenting the properties of our estimators over the intersection of these sets. Each of them gives an upper bound for the processes defined inside, which luckily has probability approaching 1 as will be presented in Theorem 1 below.

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). Essentially, this lemma gives the tail probability of a general time series object with dependence defined functionally as in (4.2).

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*

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

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 $\max_j \|x_{tj}\|_{2w}, \max_j \|z_{tj}\|_{2w} \leq \mu < \infty$ *where* μ *is a constant, the above Nagaev-type inequality holds for the product process* $\{x_{tj}z_{t\ell} - E(x_{tj}z_{t\ell})\}$.

The proof of this lemma is in the Supplement.

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. ■

With Lemma 1, we can use the union sum inequality to find an explicit probability lower bound for the event $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4 \cap \mathcal{M}$. All our results below, except for the asymptotic normality for $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ in Theorem 6, are presented on this particular event. The proof of the theorem is in the Supplement.

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 assumption A5, we can show that for any $w > 0$, $\|\zeta_{tj}\|_{2w}, \|x_{tj}\|_{2w} < \infty$ (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). Hence, we can set w to be large enough so that $N = o(T^{w/4-1/2}\log^{w/4}(T))$, 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

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

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

The proof is in the Supplement. If we treat $\widetilde{\mathbf{W}}_1$ and $\widetilde{\mathbf{W}}_2$ as some assumed spatial weight matrices, for example distance 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 spatial weight matrices through $\big\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\big\|_1$. As long as $\big\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\big\|_1$ is much less than N, estimation error is related to how sparse the 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 Supplement.

Theorem 2. *Let assumptions A1-A7 be satisfied. Suppose* $\alpha > 1/2 - 1/w$, *with* $\lambda_T = o(\delta_T^{1/2})$, $\lambda_T N^{1/w} = O(\delta_T^{1/2})$ *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*

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

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*

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

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 Bickel et al. (2009) equation (3.2)). In particular, the error $\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\|_1$, which is a sum of N^2 elements, is bounded by $4\|\widetilde{\boldsymbol{\xi}}_J-\boldsymbol{\xi}^*_J\|_1$, which is a sum of just n elements, with n potentially much smaller than N^2. With these oracle inequalities, we can find upper bounds of $\|\widetilde{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\|_1$ and $\|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}^*\|_1$ with high probability. However, we need to introduce a restricted eigenvalue condition first, which is similar to condition (3.1) of Bickel et al. (2009). Yet, unlike Bickel et al. (2009), we 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 $\widehat{\boldsymbol{\Sigma}}^* = T^{-1}\mathbf{M}^{\mathrm{T}}_{\boldsymbol{\beta}^*}\mathbf{M}_{\boldsymbol{\beta}^*}$, and $\boldsymbol{\Sigma} = E(\widehat{\boldsymbol{\Sigma}}^*)$. Define, for $\boldsymbol{\alpha} \neq \mathbf{0}$,

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

where $c_0 = \frac{8}{|\boldsymbol{\xi}^*_{J,\min}|^k} - 1$. Then 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,

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

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

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq a_1\left(\frac{s_2}{N} + N^{\frac{1}{2w}-\frac{1}{2}}\right)\lambda_T \delta_T^{1/2} + \frac{20 a_2 \gamma_T n}{N\kappa^2(n)},$$

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

The proof is in the Supplement. 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. If $\widehat{\boldsymbol{\xi}}_{J^c} = \mathbf{0}$, that is, the adaptive LASSO estimator estimates all the zeros correctly, then the above bound for $\widehat{\boldsymbol{\xi}}$ means that on average, the error for estimating each non-zero component is of order γ_T. It turns out that $\widehat{\boldsymbol{\xi}}$ has a high probability of being sign consistent, so in particular, we indeed have $\widehat{\boldsymbol{\xi}}_{J^c} = \mathbf{0}$ with high probability. See Theorem 5 below.

The need for large enough N, T in the above theorem is merely for the simplification of the different error bounds, and can be removed at the expense of more complicated expressions. The proof is omitted.

Next, we give the sign consistency for estimating the spatial weight matrices with $\widehat{\boldsymbol{\xi}}$. 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 Supplement.

Theorem 5. *Let the assumptions in Theorem 2 and 3 be satisfied. Assume further that* $\lambda_{\min}(\boldsymbol{\Sigma}_{JJ})$ *is uniformly bounded away from 0, and* $n = o\left(\gamma_T^{-\frac{2k}{k+1}}\right)$*. 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 through 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 spatial 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 $n = o(T^{1/2} \log^{-1}(T \vee N))$.

We conclude this section with the asymptotic normality of $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$. The proof can be found in the Supplement.

Theorem 6. *Let the assumptions in Theorem 5 be satisfied. Define*

$$P_0^x(X_{t,sk}) = E(X_{t,sk}|\mathcal{F}_0) - E(X_{t,sk}|\mathcal{F}_{-1}),$$
$$P_0^\epsilon(\epsilon_{t,j}) = E(\epsilon_{t,j}|\mathcal{G}_0) - E(\epsilon_{t,j}|\mathcal{G}_{-1})$$

where $\mathcal{F}_t$ and $\mathcal{G}_t$ are defined in (4.1). Then assuming further that $\big\|\mathbf{W}_2^*\big\| < \infty$,

$$\max_{1\le s\le N}\max_{1\le k\le K}\sum_{t\ge 0}\delta_T^{1/2}E^{1/2}\{P_0^x(X_{t,sk})^2\} < \infty, \quad \max_{1\le j\le N}\sum_{t\ge 0}E^{1/2}\{P_0^\epsilon(\epsilon_{t,j})^2\} < \infty,$$

and $s_2 = o(N^{\frac{1}{2}+\frac{1}{2w}} \log^{-1/2}(T \vee N))$, $\gamma_T^{1/2} n$, $\gamma_T^{3/2} n^2 = o(N^{\frac{1}{2}+\frac{1}{2w}} T^{-1/2})$, we have for $\boldsymbol{\alpha} \in \mathbb{R}^K$ with $\big\|\boldsymbol{\alpha}\big\| = 1$,

$$T^{1/2} s_0^{-1/2} \boldsymbol{\alpha}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} N(0,1),$$

where $s_0 = \sum_\tau \mathrm{tr}\left\{E(\mathbf{X}_{t+\tau}\mathbf{G}^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{X}_t^{\mathrm{T}})\mathbf{W}_2^*\boldsymbol{\Sigma}_\epsilon(\tau)\mathbf{W}_2^{*\mathrm{T}}\right\}$, with $\boldsymbol{\Sigma}_\epsilon(\tau) = E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^{\mathrm{T}})$, and $\mathbf{G} = E(T^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_2^{*\otimes\mathrm{T}}\mathbf{W}_2^{*\otimes}\mathbf{X})$. The same asymptotic normality holds for $\widehat{\boldsymbol{\beta}}$.

The functions $P_0^x(\cdot)$ and $P_0^\epsilon(\cdot)$ are the predictive dependence measure for $\{X_{t,sk}\}$ and $\{\epsilon_{t,j}\}$ respectively, as defined in Wu (2011). If $X_{t,sk}$ is independent of $\mathcal{F}_0$, then it is independent of $\mathcal{F}_{-1}$ too, so that $P_0^x(X_{t,sk}) = 0$. With this theorem, we can make inference on $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$. We can estimate s_0 by using sample estimates to replace the population ones, $\mathbf{W}_2^*$ by $\widehat{\mathbf{W}}_2$, and $\boldsymbol{\epsilon}_t$ by $\widehat{\boldsymbol{\epsilon}}_t$, the estimated error vector using the adaptive LASSO estimators. We do not pursue the details here.

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. (2013). 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

$$\boldsymbol{\beta}^{(r)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_{2,r-1}^{\otimes\mathrm{T}}\mathbf{W}_{2,r-1}^{\otimes}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_{2,r-1}^{\otimes\mathrm{T}}(\mathbf{I}_{TN} - \mathbf{W}_{1,r-1}^{\otimes})\mathbf{y}^v,$$

   where $\mathbf{W}_{j,r}^{\otimes} = \mathbf{I}_N \otimes \mathbf{W}_{j,r}$, with $\mathbf{W}_{1,r}, \mathbf{W}_{2,r}$ the spatial 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 $1/|\widetilde{\xi}_j|^k$.

We propose a BIC criterion to select the tuning parameter γ_T:

$$\mathrm{BIC}(\gamma_T) = \sum_{i=1}^{N} \log\left(T^{-1}\left\|\widetilde{\mathbf{y}}_i - (\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}_{\gamma_T})_i\right\|^2\right) + |S_{\gamma_T}|\frac{\log(T)}{T}\log(\log(2N-2)), \tag{5.6}$$

where $\mathbf{y} = (\widetilde{\mathbf{y}}_1^{\mathrm{T}}, \ldots, \widetilde{\mathbf{y}}_N^{\mathrm{T}})^{\mathrm{T}}$ with $\widetilde{\mathbf{y}}_i = (y_{i1}, \ldots, y_{iT})^{\mathrm{T}}$. The vector $\widetilde{\boldsymbol{\xi}}_{\gamma_T}$ is the LASSO solution to (3.2) with tuning parameter being γ_T. Also, $(\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}_{\gamma_T})_i$ is the vector with length T which is the portion of the vector $\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}_{\gamma_T}$ (see equation (3.1)) corresponding to $\widetilde{\mathbf{y}}_i$. Finally, the set $S_{\gamma_T} = \{j : (\widetilde{\boldsymbol{\xi}}_{\gamma_T})_j \neq 0\}$, so that $|S_{\gamma_T}|$ counts the number of non-zeros estimated in $\widetilde{\boldsymbol{\xi}}_{\gamma_T}$. This BIC criterion is inspired by the one proposed in Wang et al. (2009), and is in fact the sum of individual BIC criteria for the estimator of the ith row of the two spatial weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$, with response variable $\widetilde{\mathbf{y}}_i$. We denote γ_{BIC} the tuning parameter that minimizes the BIC criterion in (5.6). This γ_{BIC} will then be used in (3.2) to find the LASSO solution $\widetilde{\boldsymbol{\xi}}$. We use the same tuning parameter for the adaptive LASSO estimator in (3.3).

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 (2.2) 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 50 times according to the scheme above for each pair. Hence in total 100 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. In all cases, penalization parameter was chosen via BIC criteria.

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 spatial 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 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, Tables 4 and 5 show 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. 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

It is well-known that worldwide stock markets' performance are dependent on other markets. To study their dependence structure in more detail, we use model (2.2) to analyze markets' returns over 2013. 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 use daily data available for the whole of 2013 ($T = 263$). See Table 6 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 is simply set as the identity matrix. The model is estimated by the adaptive LASSO, with the tuning parameter λ chosen by BIC, as described in section 5.

This setting is also interesting as there is partial knowledge of the intraday linkages: a stock market that ended operations cannot be affect by markets which are yet to open in the same day. Thus the applied example also allow us to explore the robustness of the estimator with respect to not violating this natural impediment. Given the wide geographic dispersion of stock markets, this is set to happen for a relevant number of markets in the data.

To capture this intuition, we define the "common opening hours" index

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

which corresponds to 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 1, the elements of $\widehat{\mathbf{W}}_1$ are plotted against the common opening hours. From this figure, it is clear that for markets with smaller overlap of opening hours, the estimated elements are zero in $\widehat{\mathbf{W}}_1$. In particular, there is no violation of the afore-mentioned restriction and markets are only affecting each other if they are commonly open for at least roughly 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 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. Asymptotic normality of $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are proved, so that making inference on $\boldsymbol{\beta}^*$ is possible. 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 spatial weight matrices by simple thresholding. This way not even instrumental variables are needed. These are the potential future problems to be tackled.

Supplement: Proof of Theorems in the paper

`http://stats.lse.ac.uk/lam/Supp-RSPT.pdf` We present the proofs of Lemma 2, Theorems 1, 2,3 and 5 in the paper.

	$T = 50$		$T = 100$		$T = 200$	
	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$						
Specificity	97.02% (0.011)	98.22% (0.008)	96.74% (0.010)	98.20% (0.008)	96.64% (0.011)	98.36% (0.008)
Sensitivity	78.09% (0.083)	55.38% (0.103)	95.70% (0.042)	86.76% (0.065)	99.35% (0.014)	96.19% (0.032)
Bias	−0.0660 (0.024)	−0.1105 (0.031)	−0.0391 (0.015)	−0.0738 (0.017)	−0.0220 (0.009)	−0.0394 (0.011)
LASSO L_1	18.8344 (2.178)	18.2203 (2.407)	18.0305 (1.780)	18.8540 (2.066)	15.9489 (1.702)	16.8550 (1.810)
LASSO L_2	5.5494 (1.011)	7.2172 (1.046)	3.4079 (0.650)	4.1905 (0.759)	2.2531 (0.481)	2.3123 (0.401)
AdaLASSO L_1	2.1840 (0.368)	2.5987 (0.452)	1.7145 (0.276)	2.0779 (0.357)	1.3482 (0.221)	1.5522 (0.243)
AdaLASSO L_2	1.0609 (0.241)	1.7634 (0.315)	0.4505 (0.140)	0.7627 (0.203)	0.2067 (0.075)	0.2858 (0.096)
Sparsity	0.9349 (0.014)	0.9349 (0.014)	0.9233 (0.012)	0.9233 (0.012)	0.9202 (0.013)	0.9202 (0.013)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.0857 (0.0327)		0.0173 (0.0121)		0.0073 (0.0056)	
$N = 50$						
Specificity	95.70% (0.007)	98.38% (0.005)	96.20% (0.007)	98.35% (0.005)	96.60% (0.006)	98.47% (0.005)
Sensitivity	74.35% (0.045)	42.23% (0.050)	92.54% (0.029)	81.18% (0.043)	98.32% (0.013)	96.15% (0.018)
Bias	−0.0448 (0.011)	−0.0972 (0.016)	−0.0336 (0.006)	−0.0799 (0.011)	−0.0215 (0.004)	−0.0412 (0.006)
LASSO L_1	66.7238 (3.839)	61.6638 (4.002)	64.5299 (4.564)	66.7991 (5.325)	59.4202 (4.480)	63.2523 (4.785)
LASSO L_2	25.2673 (2.012)	31.8719 (2.073)	15.3925 (1.655)	18.7294 (1.509)	9.0062 (1.159)	9.4297 (1.044)
AdaLASSO L_1	7.8904 (0.652)	10.1448 (0.803)	5.8510 (0.637)	7.4972 (0.809)	4.6307 (0.496)	5.5847 (0.585)
AdaLASSO L_2	4.5845 (0.478)	8.2501 (0.604)	1.9969 (0.313)	3.7515 (0.479)	0.8043 (0.178)	1.1878 (0.254)
Sparsity	0.9240 (0.007)	0.9240 (0.007)	0.9182 (0.007)	0.9182 (0.007)	0.9201 (0.007)	0.9201 (0.007)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.0257 (0.0233)		0.0283 (0.0167)		0.0309 (0.0098)	
$N = 75$						
Specificity	95.30% (0.005)	98.90% (0.003)	96.35% (0.004)	98.88% (0.003)	97.16% (0.003)	98.98% (0.003)
Sensitivity	59.54% (0.034)	26.88% (0.033)	85.53% (0.026)	76.25% (0.033)	95.42% (0.013)	96.04% (0.014)
Bias	−0.0224 (0.009)	−0.0973 (0.015)	−0.0277 (0.005)	−0.0911 (0.007)	−0.0196 (0.003)	−0.0483 (0.005)
LASSO L_1	131.4265 (4.475)	111.8097 (5.187)	120.7178 (6.361)	120.3575 (7.167)	113.0090 (6.296)	120.1324 (7.211)
LASSO L_2	65.0015 (3.615)	75.7601 (2.881)	35.5961 (2.409)	46.1954 (2.351)	19.7648 (1.537)	21.3982 (1.777)
AdaLASSO L_1	15.7064 (0.752)	21.8860 (1.054)	10.1854 (0.777)	13.9803 (1.011)	7.8193 (0.627)	9.9623 (0.832)
AdaLASSO L_2	11.7311 (0.867)	20.8000 (0.836)	4.5032 (0.440)	9.9502 (0.795)	1.7424 (0.241)	2.8229 (0.454)
Sparsity	0.9262 (0.005)	0.9262 (0.005)	0.9239 (0.004)	0.9239 (0.004)	0.9262 (0.004)	0.9262 (0.004)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.0274 (0.0200)		0.0343 (0.0170)		0.0348 (0.0101)	

Table 1: Baseline Simulations. All values are averages over 100 simulations. Penalization is chosen via BIC criteria. 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. True sparsity level of the both $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ is $\kappa = 0.95$. Bracketed values are standard deviations.

	$T = 50$		$T = 100$		$T = 200$	
	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$
$N = 25$						
Specificity	99.72% (0.003)	99.86% (0.002)	99.54% (0.003)	99.72% (0.003)	99.47% (0.003)	99.79% (0.002)
Sensitivity	63.14% (0.231)	40.52% (0.214)	94.28% (0.094)	84.98% (0.146)	99.56% (0.032)	97.43% (0.065)
Bias	−0.1189 (0.052)	−0.1305 (0.064)	−0.0678 (0.030)	−0.0881 (0.033)	−0.0328 (0.017)	−0.0424 (0.022)
LASSO L_1	10.7485 (1.581)	8.9524 (1.681)	10.9814 (1.475)	10.5877 (1.653)	9.3435 (1.230)	9.5100 (1.252)
LASSO L_2	1.2140 (0.381)	1.5384 (0.409)	0.7173 (0.245)	0.9196 (0.297)	0.3881 (0.168)	0.4341 (0.186)
AdaLASSO L_1	0.9456 (0.192)	0.9303 (0.209)	0.7516 (0.153)	0.8293 (0.197)	0.5439 (0.112)	0.6016 (0.122)
AdaLASSO L_2	0.2712 (0.107)	0.4082 (0.124)	0.1017 (0.055)	0.1731 (0.086)	0.0334 (0.028)	0.0484 (0.042)
Sparsity	0.9912 (0.004)	0.9912 (0.004)	0.9865 (0.004)	0.9865 (0.004)	0.9856 (0.005)	0.9856 (0.005)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.0232 (0.0114)		0.0100 (0.0076)		0.0084 (0.0061)	
$N = 50$						
Specificity	99.75% (0.002)	99.88% (0.001)	99.57% (0.002)	99.76% (0.001)	99.54% (0.002)	99.82% (0.001)
Sensitivity	61.80% (0.127)	34.74% (0.125)	93.04% (0.056)	84.03% (0.067)	99.14% (0.021)	96.79% (0.035)
Bias	−0.1128 (0.030)	−0.1375 (0.036)	−0.0695 (0.016)	−0.0936 (0.017)	−0.0338 (0.008)	−0.0441 (0.010)
LASSO L_1	33.6100 (4.133)	24.9089 (3.769)	39.8098 (3.169)	36.4643 (3.391)	34.8279 (2.949)	35.1975 (3.435)
LASSO L_2	4.9503 (0.884)	6.4540 (0.983)	2.9538 (0.551)	3.8265 (0.690)	1.6247 (0.299)	1.6463 (0.359)
AdaLASSO L_1	2.9374 (0.453)	2.7847 (0.460)	2.6580 (0.301)	2.8045 (0.350)	1.9250 (0.244)	2.1301 (0.305)
AdaLASSO L_2	1.1193 (0.253)	1.7546 (0.316)	0.4390 (0.134)	0.7485 (0.183)	0.1423 (0.058)	0.1972 (0.090)
Sparsity	0.9915 (0.002)	0.9915 (0.002)	0.9868 (0.002)	0.9868 (0.002)	0.9854 (0.002)	0.9854 (0.002)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.0248 (0.0164)		0.0132 (0.0097)		0.0087 (0.0062)	
$N = 75$						
Specificity	99.79% (0.001)	99.91% (0.001)	99.54% (0.001)	99.74% (0.001)	99.56% (0.001)	99.84% (0.001)
Sensitivity	52.25% (0.140)	24.37% (0.098)	93.66% (0.034)	83.38% (0.056)	99.17% (0.012)	97.46% (0.023)
Bias	−0.1228 (0.023)	−0.1466 (0.030)	−0.0669 (0.009)	−0.0935 (0.013)	−0.0326 (0.005)	−0.0450 (0.008)
LASSO L_1	59.9314 (9.852)	39.7276 (7.405)	80.7885 (4.056)	71.3078 (4.727)	74.6762 (4.159)	74.8206 (5.213)
LASSO L_2	12.1496 (1.295)	15.1889 (1.247)	6.7000 (0.852)	8.3786 (1.078)	3.5099 (0.480)	3.5854 (0.601)
AdaLASSO L_1	5.4167 (0.949)	5.1533 (0.755)	5.3577 (0.391)	5.5670 (0.474)	3.9939 (0.347)	4.4054 (0.441)
AdaLASSO L_2	2.8895 (0.505)	4.2755 (0.446)	0.9576 (0.186)	1.6567 (0.295)	0.2951 (0.092)	0.4148 (0.146)
Sparsity	0.9927 (0.003)	0.9927 (0.003)	0.9861 (0.002)	0.9861 (0.002)	0.9854 (0.001)	0.9854 (0.001)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.0466 (0.0186)		0.0183 (0.0133)		0.0100 (0.0067)	

Table 2: Baseline Simulations. All values are averages over 100 simulations. Penalization is chosen via BIC criteria. 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. True sparsity level of the both $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ is $\kappa = 0.99$. Bracketed values are standard deviations.

	Time Dependence				*Spatial Dependence*			
	$T = 100$		$T = 200$		$T = 100$		$T = 200$	
	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$								
Specificity	96.35% (0.01)	98.04% (0.01)	96.46% (0.01)	98.34% (0.01)	96.73% (0.01)	98.23% (0.01)	96.54% (0.01)	98.12% (0.01)
Sensitivity	94.04% (0.05)	84.39% (0.07)	99.44% (0.01)	93.86% (0.07)	94.71% (0.05)	88.03% (0.06)	99.22% (0.02)	96.00% (0.03)
Bias	−0.04 (0.01)	−0.07 (0.02)	−0.02 (0.01)	−0.05 (0.02)	−0.05 (0.01)	−0.08 (0.02)	−0.02 (0.01)	−0.04 (0.01)
LASSO L_1	18.470 (2.02)	19.625 (1.90)	16.053 (1.78)	16.971 (1.90)	18.121 (1.61)	18.720 (1.67)	16.304 (1.69)	17.494 (1.95)
LASSO L_2	3.721 (0.74)	4.561 (0.79)	2.412 (0.61)	2.706 (0.75)	3.586 (0.58)	4.127 (0.73)	2.351 (0.51)	2.465 (0.50)
AdaLASSO L_1	1.792 (0.33)	2.195 (0.36)	1.349 (0.21)	1.564 (0.26)	1.733 (0.26)	2.043 (0.29)	1.404 (0.23)	1.638 (0.29)
AdaLASSO L_2	0.518 (0.15)	0.855 (0.21)	0.249 (0.13)	0.379 (0.22)	0.481 (0.12)	0.740 (0.19)	0.229 (0.08)	0.314 (0.11)
Sparsity	0.922 (0.01)	0.922 (0.01)	0.920 (0.01)	0.920 (0.01)	0.923 (0.01)	0.923 (0.01)	0.919 (0.01)	0.919 (0.01)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.018 (0.009)		0.010 (0.006)		0.023 (0.013)		0.007 (0.006)	
$N = 50$								
Specificity	95.39% (0.01)	97.96% (0.01)	95.96% (0.01)	98.37% (0.00)	96.15% (0.01)	98.37% (0.01)	96.64% (0.01)	98.51% (0.00)
Sensitivity	91.07% (0.03)	67.93% (0.13)	98.33% (0.01)	86.47% (0.07)	93.58% (0.02)	81.74% (0.03)	98.65% (0.01)	95.69% (0.02)
Bias	−0.04 (0.01)	−0.10 (0.02)	−0.02 (0.00)	−0.07 (0.02)	−0.03 (0.01)	−0.08 (0.01)	−0.02 (0.00)	−0.04 (0.01)
LASSO L_1	73.099 (7.87)	83.666 (11.43)	67.840 (4.70)	80.425 (6.82)	64.892 (4.55)	66.489 (5.64)	59.234 (4.16)	62.960 (4.95)
LASSO L_2	18.115 (3.08)	23.277 (4.09)	11.342 (2.16)	14.537 (2.99)	15.429 (1.49)	18.314 (1.19)	9.142 (0.97)	9.431 (0.97)
AdaLASSO L_1	7.282 (1.26)	10.651 (2.31)	5.622 (0.62)	8.320 (1.24)	5.878 (0.64)	7.395 (0.87)	4.611 (0.47)	5.605 (0.62)
AdaLASSO L_2	2.451 (0.55)	5.180 (1.35)	1.134 (0.31)	2.578 (0.86)	1.993 (0.28)	3.632 (0.35)	0.818 (0.14)	1.239 (0.22)
Sparsity	0.911 (0.01)	0.911 (0.01)	0.914 (0.01)	0.914 (0.01)	0.917 (0.01)	0.917 (0.01)	0.919 (0.01)	0.919 (0.01)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.035 (0.021)		0.027 (0.013)		0.031 (0.018)		0.035 (0.010)	
$N = 75$								
Specificity	92.43% (0.01)	94.97% (0.02)	87.74% (0.01)	90.68% (0.02)	96.44% (0.00)	98.90% (0.00)	97.20% (0.00)	98.99% (0.00)
Sensitivity	70.69% (0.03)	17.31% (0.03)	88.79% (0.02)	25.72% (0.04)	84.79% (0.03)	75.73% (0.03)	95.25% (0.01)	96.17% (0.01)
Bias	−0.03 (0.01)	−0.19 (0.02)	−0.03 (0.00)	−0.20 (0.02)	−0.03 (0.00)	−0.09 (0.01)	−0.02 (0.00)	−0.05 (0.00)
LASSO L_1	209.546 (5.33)	258.505 (5.43)	268.400 (5.61)	308.294 (6.47)	119.755 (6.98)	118.370 (8.39)	112.565 (6.04)	118.902 (7.55)
LASSO L_2	66.675 (5.21)	102.104 (11.58)	71.681 (6.04)	114.267 (15.06)	35.621 (2.69)	45.683 (2.58)	19.816 (1.45)	21.457 (1.23)
AdaLASSO L_1	27.859 (1.20)	43.580 (1.53)	32.307 (1.09)	46.462 (1.71)	10.024 (0.81)	13.732 (1.11)	7.798 (0.55)	9.865 (0.86)
AdaLASSO L_2	10.405 (1.19)	26.359 (2.43)	9.084 (1.07)	26.784 (3.08)	4.512 (0.48)	9.867 (0.80)	1.759 (0.20)	2.824 (0.29)
Sparsity	0.894 (0.01)	0.894 (0.01)	0.841 (0.01)	0.841 (0.01)	0.925 (0.00)	0.925 (0.00)	0.927 (0.00)	0.927 (0.00)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.101 (0.030)		0.105 (0.025)		0.034 (0.021)		0.034 (0.010)	

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.

	No Variance Decay				*Fat Tails*			
	$T = 100$		$T = 200$		$T = 100$		$T = 200$	
	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$								
Specificity	96.24% (0.01)	97.87% (0.01)	95.62% (0.01)	97.42% (0.01)	93.44% (0.02)	95.73% (0.02)	91.81% (0.02)	94.55% (0.01)
Sensitivity	95.06% (0.04)	84.64% (0.06)	99.11% (0.02)	95.98% (0.03)	88.76% (0.07)	58.61% (0.10)	98.10% (0.03)	84.91% (0.08)
Bias	−0.04 (0.02)	−0.08 (0.02)	−0.03 (0.01)	−0.05 (0.01)	−0.05 (0.02)	−0.11 (0.05)	−0.04 (0.01)	−0.09 (0.02)
LASSO L_1	19.660 (2.20)	20.683 (2.49)	18.489 (1.55)	20.274 (1.72)	27.561 (3.60)	30.570 (4.47)	25.583 (1.92)	30.797 (2.68)
LASSO L_2	3.907 (0.70)	4.520 (0.69)	2.865 (0.48)	3.023 (0.42)	7.595 (2.21)	9.432 (2.07)	6.636 (1.15)	6.913 (1.23)
AdaLASSO L_1	1.941 (0.36)	2.409 (0.44)	1.654 (0.20)	2.021 (0.24)	4.381 (2.53)	6.091 (2.76)	3.111 (0.57)	4.520 (0.80)
AdaLASSO L_2	0.530 (0.15)	0.844 (0.19)	0.291 (0.08)	0.383 (0.10)	1.970 (2.34)	2.822 (2.15)	1.090 (0.40)	1.445 (0.47)
Sparsity	0.918 (0.01)	0.918 (0.01)	0.914 (0.01)	0.914 (0.01)	0.895 (0.02)	0.895 (0.02)	0.871 (0.02)	0.871 (0.02)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.018 (0.012)		0.010 (0.008)		0.041 (0.030)		0.024 (0.017)	
$N = 50$								
Specificity	95.70% (0.01)	98.15% (0.01)	95.69% (0.01)	97.92% (0.00)	93.57% (0.01)	96.93% (0.01)	92.59% (0.01)	95.99% (0.01)
Sensitivity	92.41% (0.02)	77.73% (0.04)	98.22% (0.01)	93.97% (0.03)	80.41% (0.04)	50.10% (0.06)	94.01% (0.03)	82.04% (0.04)
Bias	−0.04 (0.01)	−0.08 (0.01)	−0.03 (0.00)	−0.05 (0.01)	−0.05 (0.01)	−0.10 (0.02)	−0.04 (0.01)	−0.08 (0.01)
LASSO L_1	70.005 (4.59)	71.366 (4.80)	69.186 (3.93)	74.982 (5.18)	97.516 (6.32)	101.219 (8.91)	94.701 (5.16)	110.642 (7.47)
LASSO L_2	17.093 (1.75)	20.165 (1.59)	11.422 (1.38)	12.697 (1.17)	28.677 (3.55)	35.093 (4.23)	23.099 (2.81)	27.307 (3.30)
AdaLASSO L_1	6.660 (0.63)	8.430 (0.78)	5.722 (0.46)	7.338 (0.72)	12.871 (2.35)	18.654 (3.87)	10.689 (1.64)	15.928 (2.58)
AdaLASSO L_2	2.281 (0.33)	4.163 (0.46)	1.072 (0.21)	1.776 (0.30)	5.379 (1.72)	9.608 (2.53)	3.608 (1.22)	6.023 (1.81)
Sparsity	0.913 (0.01)	0.913 (0.01)	0.913 (0.01)	0.913 (0.01)	0.898 (0.01)	0.898 (0.01)	0.885 (0.01)	0.885 (0.01)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.030 (0.017)		0.035 (0.013)		0.048 (0.031)		0.054 (0.019)	
$N = 75$								
Specificity	95.99% (0.00)	98.77% (0.00)	96.49% (0.00)	98.73% (0.00)	94.16% (0.01)	98.04% (0.00)	94.03% (0.01)	97.33% (0.01)
Sensitivity	83.29% (0.03)	70.74% (0.03)	94.63% (0.02)	93.24% (0.02)	71.87% (0.03)	38.61% (0.03)	88.41% (0.02)	73.94% (0.04)
Bias	−0.03 (0.00)	−0.10 (0.01)	−0.02 (0.00)	−0.07 (0.01)	−0.03 (0.01)	−0.10 (0.01)	−0.04 (0.00)	−0.10 (0.01)
LASSO L_1	129.473 (6.81)	127.115 (9.64)	129.971 (6.92)	139.475 (8.09)	182.960 (8.85)	172.011 (11.09)	184.988 (9.50)	209.753 (14.13)
LASSO L_2	39.202 (2.47)	50.770 (2.52)	24.565 (1.72)	28.827 (2.39)	60.011 (4.52)	78.056 (5.12)	45.624 (4.51)	60.051 (5.58)
AdaLASSO L_1	11.351 (0.87)	15.647 (1.44)	9.717 (0.72)	12.943 (1.00)	21.764 (2.44)	31.231 (3.85)	18.852 (2.80)	29.517 (4.90)
AdaLASSO L_2	5.131 (0.46)	11.422 (0.75)	2.347 (0.24)	4.459 (0.65)	10.260 (1.91)	21.597 (3.09)	6.576 (1.91)	13.822 (2.99)
Sparsity	0.921 (0.00)	0.921 (0.00)	0.921 (0.00)	0.921 (0.00)	0.909 (0.01)	0.909 (0.01)	0.901 (0.01)	0.901 (0.01)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	0.037 (0.016)		0.040 (0.011)		0.065 (0.024)		0.064 (0.014)	

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

	$T = 100$		$T = 200$	
	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$				
Specificity	96.00% (0.021)	– (–)	93.27% (0.017)	– (–)
Sensitivity	59.22% (0.198)	– (–)	90.63% (0.075)	– (–)
Bias	−0.1089 (0.040)	– (–)	−0.0928 (0.024)	– (–)
LASSO L_1	7.6912 (0.751)	– (–)	7.5505 (0.661)	– (–)
LASSO L_2	7.6912 (0.751)	– (–)	7.5505 (0.661)	– (–)
AdaLASSO L_1	1.5748 (0.228)	– (–)	1.2060 (0.136)	– (–)
AdaLASSO L_2	1.5748 (0.228)	– (–)	1.2060 (0.136)	– (–)
Sparsity	0.9324 (0.029)	– (–)	0.8907 (0.018)	– (–)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	– (–)		– (–)	
$N = 50$				
Specificity	95.76% (0.007)	– (–)	94.40% (0.008)	– (–)
Sensitivity	63.47% (0.070)	– (–)	86.84% (0.039)	– (–)
Bias	−0.0825 (0.015)	– (–)	−0.0804 (0.015)	– (–)
LASSO L_1	27.1855 (1.406)	– (–)	26.3433 (1.840)	– (–)
LASSO L_2	27.1855 (1.406)	– (–)	26.3433 (1.840)	– (–)
AdaLASSO L_1	4.8163 (0.366)	– (–)	3.9884 (0.346)	– (–)
AdaLASSO L_2	4.8163 (0.366)	– (–)	3.9884 (0.346)	– (–)
Sparsity	0.9279 (0.009)	– (–)	0.9032 (0.008)	– (–)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	– (–)		– (–)	
$N = 75$				
Specificity	95.46% (0.007)	– (–)	94.58% (0.006)	– (–)
Sensitivity	57.03% (0.063)	– (–)	76.97% (0.043)	– (–)
Bias	−0.0685 (0.012)	– (–)	−0.0684 (0.012)	– (–)
LASSO L_1	55.0692 (3.474)	– (–)	51.4000 (2.648)	– (–)
LASSO L_2	55.0692 (3.474)	– (–)	51.4000 (2.648)	– (–)
AdaLASSO L_1	8.5933 (0.714)	– (–)	6.9086 (0.544)	– (–)
AdaLASSO L_2	8.5933 (0.714)	– (–)	6.9086 (0.544)	– (–)
Sparsity	0.9283 (0.009)	– (–)	0.9099 (0.006)	– (–)
$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1$	– (–)		– (–)	

Table 5: Simulations without covariates. Comparisons to the baseline simulations when assumptions are violated. Refer to Table 1 for the explanations of different items.

Country	**Index**	**Country**	**Index**
Argentina	Merval	Australia	Dow Jones Australian
Austria	Viena ATX-5	Brazil	Dow Jones Brazil Stock
Canada	S&P/CDNX Comp.	Chile	Santiago SSE Inter-10
China	Shanghai SE Comp.	Egypt	SE 100
France	Paris CAC-40	Germany	CDAX Total Return
Hong Kong	Hang Seng Comp.	India	NSE-50
Indonesia	Jakarta SE Liquid 45	Italy	Milan SE MIB-30
Japan	Nikkei 500	Mexico	SE Index (INMX)
New Zealand	NZSX-15	Russia	Russia MICEX Comp.
Spain	Madrid SE IBEX-35	Singapore	Singapore FTSE All-sh.
South Africa	FTSE/JSE Top 40	South Korea	Korea SE Stock Price
Switzerland	Swiss Market	Thailand	Thailand SET General
United Kingdom	S&P UK	United States	S&P 500

Table 6: Markets and their respective indices used. Data source: *Global Financial Data*.

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.

Anselin, L., J. Le Gallo, and H. Jayet (2006). *Spatial panel econometrics. In: Matyas L, Sevestre P. (eds) The econometrics of panel data, fundamentals and recent developments in theory and practice* (3 ed.). Kluwer, Dordrecht.

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.

Beenstock, M. and D. Felsenstein (2012). Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geographical Analysis 44*(4), 386–397.

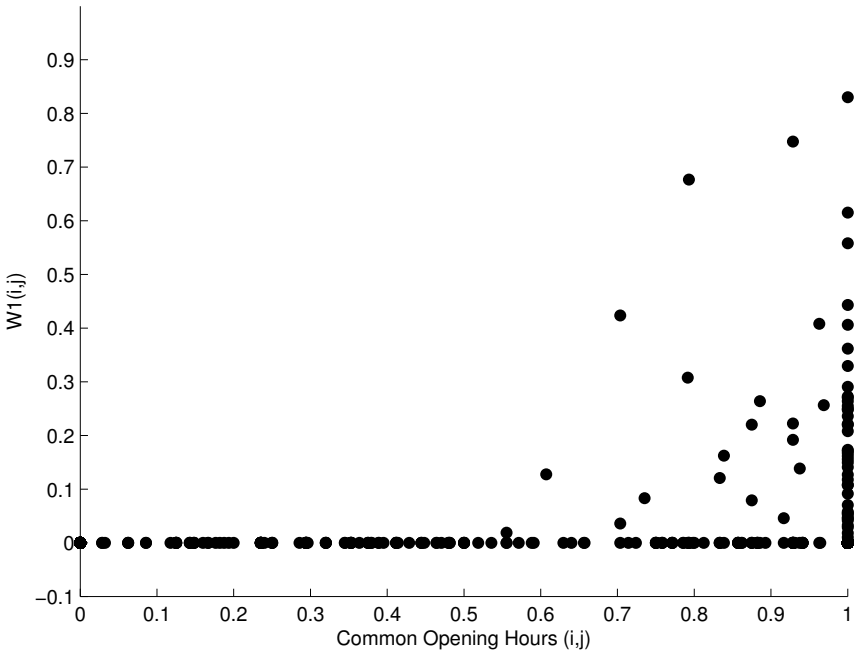

Figure 1: Elements of $\widehat{\mathbf{W}}_1$ plotted against Common Opening Hours.

Bhattacharjee, A. and C. Jensen-Butler (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics 43*(4), 617 – 634.

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, 12). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics 41*(6), 2994–3021.

Corrado, L. and B. Fingleton (2011, Jan). 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 (2013). Variable selection and estimation with the seamless-l0 penalty. *Statistica Sinica 23*, 929–962.

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.

Lam, C. and P. C. L. Souza (2014). Detection and estimation of block structure in spatial weight matrix. Manuscript.

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.

Liu, W., H. Xiao, and W. Wu (2013). Probability and moment inequalities under dependence. *Statistica Sinica 23*, 1257–1272.

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

Plümper, T. and E. Neumayer (2010). Model specification in the analysis of spatial dependence. *European Journal of Political Research 49*(3), 418–442.

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

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 671–683.

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 0*, 1–20.

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.

Department of Statistics, London School of Economics and Political Science

E-mail: c.lam2@lse.ac.uk

Department of Statistics, London School of Economics and Political Science

E-mail: p.souza@lse.ac.uk