Detection And Estimation Of Block Structure In Spatial Weight Matrix

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 the use of LASSO penalization to estimate the underlying block structure of the spatial weight matrix of a spatial lag/error model in the absence of exogenous variables. We show that the block structure of the spatial weight matrix is recovered, in the sense that zero blocks are estimated as zeros with high probability. We denote such a property as "zero-block consistent". The method in Lam and Souza (2013) had proven sign-consistency for the elements in the spatial weight matrix only in the presence of exogenous variables and decaying variance of the disturbance, which are not assumed in this paper. The tool developed in this paper can be used as a verification of block structures by applied researchers, or as an exploration tool for estimating unknown block structures. We analyzed the US Senate voting data and correctly identify blocks based on party affiliations. Simulations also show that the method performs well.

Key words and phrases. Spatial weight matrix; LASSO penalization; zero-block consistency; spatial lag/error model; Nagaev-type inequality.

*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

A spatial lag/error model allows the study of spatial dependence among individuals and find applications in a wide array of fields. However, practitioners usually assume a known spatial weight matrix using expert knowledge, or more often just rough proxies like the inverse of "distances" or its arbitrary powers. Unfortunately, estimation accuracy of other parameters in the model depends crucially on the correct specification of the spatial weight matrix (see, for example, Arbia and Fingleton (2008) and Pinkse and Slade (2010)).

With these concerns in mind, there are increasing number of researchers who attempt to estimate the spatial weight matrix together with other important model parameters in a spatial lag/error model. Pinkse et al. (2002) suggested to estimate a nonparametric smooth function for the elements of the spatial weight matrix. 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. These methods can suffer from the need to input an appropriate distance metric, which is still determined by the user, or to estimate a large error covariance matrix, which 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. There are other ad hoc approaches as well, many of which unfortunately lack vigorous analysis of the properties of the resulting estimators. Recently, Lam and Souza (2013) suggested to estimate jointly the spatial weight matrix and other parameters in a spatial lag/error model through the use of adaptive LASSO penalization, which is first developed in Zou (2006) for variable selection problems in standard regression. They provided vigorous analysis of the properties of the resulting estimators, including the spatial weight matrix and other important parameters in the model, and the size of the panel is allowed to be close to or even larger than the sample size.

Motivated by a US Senate voting data set, in this paper, we also focus on estimating the spatial weight matrix. However, our aim is to study the spatial block dependence structure. For the US Senate data, we want to explore if the Republicans and the Democrats form two major blocks based on their Senator's voting records. These blocks may overlap slightly if some senators from different parties have similar voting records. One major use of such a spatial weight matrix goes to social interaction study. See Case (1991) for more details. See also Lee (2002) and Kelejian and Prucha (2002) for the corresponding theoretical treatments. However, these papers assume that the blocks in the spatial weight matrix are known. Moreover, the blocks are assumed not overlapping each other, and constrained to have equal elements inside each block.

We do not assume any prior knowledge of the spatial weight matrix in this paper, other than the fact that it can be formed in blocks, and hence the spatial weight matrix is sparse overall, i.e. with a lot of zero elements. The blocks can be slightly overlapping each other, and we do not know where the blocks are formed. Since there are no obvious exogenous variables, we analyze the US Senate voting data using a spatial lag model in the absence of them, which is essentially a spatial error model. To the best of our knowledge, this has not been done by any researchers previously. We show that accurate within-block estimation is not possible without any exogenous variables. However, we prove that theoretically we are able to detect blocks, or even slightly overlapping ones in the spatial weight matrix, with probability going to 1 as both the sample size and the panel size go to infinity.

The rest of the paper is organized as follows. In section 2, we introduce the spatial lag/error model with blocks in the spatial weight matrix, and proposed a LASSO minimization problem for finding the estimator of the spatial weight matrix. Section 3 presents the concept of zero-block consistency, with probability lower bound of such consistency for the LASSO estimator explicitly given, thus showing that block detection is achieved with high probability. Section 4 relaxed all the previous settings and results to overlapping blocks. Section 5 presents our simulation results as well as the complete analysis of the US Senate voting data. Conclusion is in section 6, and all technical proofs are in section 7.

2 The Model and the LASSO Estimator

We consider a spatial lag model without any exogenous variables,

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

where $\mathbf{y}_t$ is an $N \times 1$ vector of observations at time t, $\boldsymbol{\epsilon}_t$ is a zero mean noise vector of the same size, and $\mathbf{W}^*$ is the spatial weight matrix of size N, with 0 on its main diagonal. We assume that $\left\|\mathbf{W}^*\right\|_\infty \leq \eta < 1$, where $\left\|A\right\|_\infty = \max_i \sum_j |A_{ij}|$ is the L_∞ norm of a matrix A. Model (2.1) is also a spatial error model, since it can be written as

$$\mathbf{y}_t = (\mathbf{I}_N - \mathbf{W}^*)^{-1} \boldsymbol{\epsilon}_t. \tag{2.2}$$

Without loss of generality, we assume the components of $\mathbf{y}_t$ are sorted so that the weight matrix $\mathbf{W}^*$ is block diagonal, with

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1^* & & \\ & \ddots & \\ & & \mathbf{W}_G^* \end{pmatrix}, \quad \boldsymbol{\epsilon}_t = \begin{pmatrix} \boldsymbol{\epsilon}_t^{(1)} \\ \vdots \\ \boldsymbol{\epsilon}_t^{(G)} \end{pmatrix}, \tag{2.3}$$

where G is the number of blocks in $\mathbf{W}^*$. An important assumption for $\{\epsilon_t\}$ is that $\mathrm{cov}(\boldsymbol{\epsilon}_t^{(i)}, \boldsymbol{\epsilon}_t^{(j)}) = \mathbf{0}$ for $i \neq j$. Otherwise, the block structure in $\mathbf{W}^*$ is not identifiable. Detailed assumptions can be found in section 3.1. Relaxation to overlapping blocks is treated in section 4.

For recovering the block structure of the spatial weight matrix in (2.3), if there were exogenous variables, the adaptive LASSO estimator proposed in Lam and Souza (2013) is more than sufficient, since it has been shown that the adaptive LASSO estimator is asymptotically sign consistent for the elements in the spatial weight matrix. In this paper, we complement their results by showing that, even in the absence of exogenous variables, it is still possible to accurately estimate the block structure of the spatial weight matrix Furthermore, the disturbance decay assumption in Lam and Souza (2013) is neither needed nor feasible, or else $\mathbf{y}_t$ would have decaying variance as well.

Before we propose our estimator, we write (2.1) as a linear regression model,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\xi}^* + \boldsymbol{\epsilon}, \tag{2.4}$$

where $\mathbf{y} = \mathrm{vec}\{(\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}\}$, $\boldsymbol{\epsilon} = \mathrm{vec}\{(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^{\mathrm{T}}\}$, $\boldsymbol{\xi}^* = \mathrm{vec}(\mathbf{W}^{*\mathrm{T}})$ and $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}$. The design matrix $\mathbf{Z}$ contains the endogenous variables $\mathbf{y}_t$, and hence least square estimation will be biased. Furthermore, when N is close to T, e.g. $N = T/2$, it has a serious negative effect on the accuracy of the least square estimators since the inverse $(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$ will be ill-conditioned.

Since we assume there is a block structure in $\mathbf{W}^*$, we know that $\boldsymbol{\xi}^*$ is a sparse vector, that is, $\boldsymbol{\xi}^*$ should have a lot of zeros corresponding to the zero blocks in $\mathbf{W}^*$. This motivates us to propose the LASSO penalization on the elements of $\boldsymbol{\xi}$ to obtain

$$\widetilde{\boldsymbol{\xi}} = \min_{\boldsymbol{\xi}} \frac{1}{2T}\left\|\mathbf{y} - \mathbf{Z}\boldsymbol{\xi}\right\|^2 + \gamma_T \left\|\boldsymbol{\xi}\right\|_1, \text{ subj. to } \sum_{j=1}^{N} w_{ij} < 1 \tag{2.5}$$

where $\|\cdot\|_1$ represents the L_1-norm and $\|\cdot\|$ represents the L_2 norm. The row sum condition ensures stability of the system. The rate for the tuning parameter γ_T will be discussed briefly after Theorem 3 in section 3.2.

3 Zero-Block Consistency of the LASSO Estimator

Before presenting the main results of this paper, we introduce the notations to be used for the rest of the paper, and the main technical assumptions.

3.1 Main assumptions and notations

(i) The spatial weight matrix $\mathbf{W}^*$ is block diagonal as in (2.3), with at least one $\mathbf{W}_i^* \neq \mathbf{0}$, and $\left\|\mathbf{W}^*\right\|_\infty \leq \eta < 1$ uniformly as $T, N \to \infty$, where η is a constant. We also assume, uniformly as $T, N \to \infty$,

$$\left\|\mathbf{W}^*\right\|_1 \leq \eta_c,$$

where $\left\|A\right\|_1 = \max_j \sum_i |A_{ij}|$ is the L_1 norm of a matrix A, and η_c is a constant.

(ii) The vector $\boldsymbol{\epsilon}_t$ can be partitioned as in (2.3), with the length of $\boldsymbol{\epsilon}_t^{(j)}$ the same as the size of $\mathbf{W}_j^*$. Furthermore, $E(\boldsymbol{\epsilon}_t) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}_t^{(i)}, \boldsymbol{\epsilon}_t^{(j)}) = \mathbf{0}$ for $i \neq j$. Also, $\text{var}(\epsilon_{tj}) \leq \sigma_\epsilon^2 < \infty$ uniformly as $T, N \to \infty$, where σ_ϵ^2 is a positive constant.

(iii) Define $d_T = \frac{N}{T}$. Then we assume $d_T \to d \in [0, 1)$ as $T, N \to \infty$.

(iv) The series $\{\boldsymbol{\epsilon}_t\}$ is causal, with

$$\boldsymbol{\epsilon}_t = \sum_{i \geq 0} \boldsymbol{\Phi}_i \boldsymbol{\eta}_{t-i}, \quad \boldsymbol{\Phi}_0 = \mathbf{I}_N,$$

where $\boldsymbol{\eta}_t = (\eta_{t1}, \ldots, \eta_{tN})^{\mathrm{T}}$, and the η_{ti}'s are independent and identically distributed random variables with mean 0 and variance σ^2, having finite fourth moments. Furthermore, we assume that uniformly as $N, T \to \infty$,

$$\sum_{i \geq 1} \left\|\boldsymbol{\Phi}_i\right\| \leq \frac{\sigma(1 - \sqrt{d}) - \epsilon - c}{\sigma(1 + \sqrt{d}) + \epsilon},$$

for some constants $\epsilon, c > 0$.

(v) The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is satisfied for η_{ti} and ϵ_{ti} for all integer t and $i = 1, \ldots, N$, for the same constants D_1, D_2 and q.

(vi) There are constants $w > 2$ and $\alpha > \frac{1}{2} - \frac{1}{w}$ such that for all positive integer m,

$$\sum_{i \geq m} \big\|\mathbf{\Phi}_i\big\|_\infty \leq C m^{-\alpha} \Big(\max_{i,j} |J_{ij}|\Big)^{-\frac{1}{2w}},$$

where $C > 0$ is a constant (can depend on w), and J_{ij} =The index set for the non-zero elements of the j-th row of $\mathbf{\Phi}_i$.

Assumption (i) assumes the absolute row sum of $\mathbf{W}^*$ is uniformly less than 1, which is a regularity condition to ensure that the model is stationary and has a reduced form (2.2). Assumption (ii) is an important identifiability condition for the block structure of $\mathbf{W}^*$. Assumptions (iii) and (iv) facilitate the bounding of the minimum eigenvalue of a sample covariance matrix of the observations using random matrix theories. They also make bounding various terms in the proof much easier. Assumption (v) is a relaxation to normality, allowing for sub-gaussian or sub-exponential tailed distributions. Together with assumption (v), assumption (vi) allows us to apply the Nagaev-type inequality in Theorem 1 to determine the tail probability of the mean of the product process $\{\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})\}$. It can actually be relaxed to allow for $0 < \alpha < 1/2 - 1/w$ at the expense of more complicated rate in the Nagaev-type inequality in Theorem 1. See Remark 1 after Theorem 1 for more details on this.

There are more notations and definitions before we move to our main results. Define the set

$$H = \{j : \xi_j^* = 0 \ \text{ and corresponds to the zero blocks in } \ \mathbf{W}^*\}. \tag{3.1}$$

In other words, the set H excludes those zeros within the diagonal blocks $\mathbf{W}_i^*$ for $i = 1, \ldots, G$. Define n =maximum size of $\mathbf{W}_i, i = 1, \ldots, G$. For the rest of the paper, we use the notation $\mathbf{v}_S$ to denote a vector $\mathbf{v}$ restricted to those components with index $j \in S$. Hence, for instance, we have $\boldsymbol{\xi}_H^* = \mathbf{0}$ by definition. Let $\lambda_T = cT^{-1/2}\log^{1/2}(T \vee N)$, where c is a constant (see Corollary 2 for the exact value of c). Finally, define the set

$$A_\epsilon = \Big\{ \max_{1 \leq i,j \leq N} \Big| \frac{1}{T} \sum_{t=1}^{T} [\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})] \Big| < \lambda_T \Big\}. \tag{3.2}$$

3.2 Main results

We first present a theorem and its corollary concerning the probability lower bound of the set defined in (3.2). Then we present the zero-block consistency of the LASSO estimator $\widetilde{\boldsymbol{\xi}}$, which is the main result of the paper.

Theorem 1. *With the causal representation for $\boldsymbol{\epsilon}_t$ in assumption (iv), together with assumptions (v) and (vi), there exists constants C_1, C_2 and C_3 independent of T, v and the indices i, j, such that*

$$P\Big(\Big|\frac{1}{T}\sum_{t=1}^{T}[\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})] > v\Big|\Big) \leq \frac{C_1 T}{(Tv)^w} + C_2 \exp\big(-C_3 T v^2\big).$$

The proof of Theorem 1 is relegated to section 7. This theorem utilizes Lemma 1 of Lam and Souza (2013), where a functional dependence measure for a general time series is presented and discussed. With the causal representation of $\boldsymbol{\epsilon}_t$ and assumptions (v) and (vi), the conditions in Lemma 1 of Lam and Souza (2013) are satisfied, and hence the Nagaev-type inequality there can be invoked.

Remark 1. If $0 < \alpha < 1/2 - 1/w$, then the inequality in Theorem 1 becomes

$$P\Big(\Big|\frac{1}{T}\sum_{t=1}^{T}[\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})] > v\Big|\Big) \leq \frac{C_1 T^{w(1/2-\alpha)}}{(Tv)^w} + C_2 \exp\big(-C_3 T^{\beta} v^2\big),$$

where $\beta = (3+2\alpha w)/(1+w)$. Consequently, we need to redefine $\lambda_T = cT^{-\beta/2}\log^{1/2}(T \vee N)$ and any rates of convergence in the paper needed to be modified. For the sake of clarity we do not present those results in the paper, but just assume $\alpha > 1/2 - 1/w$, as in assumption (vi).

The following corollary is an immediate consequence of Theorem 1.

Corollary 2. *Assume the conditions in Theorem 1. With the same constants C_1,C_2 and C_3 as in Theorem 1, we set the constant c in λ_T such that $c \geq \sqrt{3/C_3}$. Then we have*

$$P(A_\epsilon) \geq 1 - C_1\Big(\frac{C_3}{3}\Big)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T \vee N)} - \frac{C_2 N^2}{T^3 \vee N^3}.$$

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

Proof of Corollary 2. By the union sum inequality, putting $v = \lambda_T$ in the result of Theorem 1,

$$\begin{aligned}
P(A_\epsilon^c) &\leq \sum_{1\leq i,j\leq N} P\Big(\Big|\frac{1}{T}\sum_{t=1}^{T}[\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})]\Big| \geq \lambda_T\Big) \\
&\leq N^2\Big(\frac{C_1 T}{(T\lambda_T)^w} + C_2\exp(-C_3 T\lambda_T^2)\Big) \\
&= \frac{C_1 N^2}{c^w T^{w/2-1}\log^{w/2}(T \vee N)} + C_2 N^2 \exp(-c^2 C_3 \log(T \vee N)) \\
&= \frac{C_1 N^2}{c^w T^{w/2-1}\log^{w/2}(T \vee N)} + \frac{C_2 N^2}{(T \vee N)^{c^2 C_3}} \\
&\leq C_1\Big(\frac{C_3}{3}\Big)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T \vee N)} + \frac{C_2 N^2}{T^3 \vee N^3},
\end{aligned}$$

for $c \geq \sqrt{3/C_3}$. The result follows. □

Remark 2. Assumption (vi) is satisfied, for instance, if $\alpha \geq 1/2$, $|I_{ij}|$ is finite uniformly for all i, j, and

$$\sum_{i\geq m}\|\boldsymbol{\Phi}_i\|_1 \leq Cm^{-\alpha}.$$

If assumption (v) is also satisfied, we can actually set w to be any constant larger than 2, so that the condition $N = o(T^{w/4-1/2}\log^{w/4}(T))$ is satisfied for a large enough constant w. In light of Remark 1, we can allow for $\alpha < 1/2$ as well, with more complicated rate for the lower bound of $P(A_\epsilon)$.

It turns out that the probability lower bound in Corollary 2 is the same as the probability lower bound for the LASSO estimator $\widetilde{\boldsymbol{\xi}}$ in (2.5) to be zero-block consistent.

Theorem 3. *Under assumptions (i) to (vi), if $\lambda_T = o(\gamma_T)$ and $n = o(\{\gamma_T/\lambda_T\}^{2/3})$, then for large enough T, N, the LASSO solution $\widetilde{\boldsymbol{\xi}}$ in (2.5) is such that*

$$P(\widetilde{\boldsymbol{\xi}}_H = \mathbf{0}) \geq P(A_\epsilon),$$

which approaches 1 as $T, N \to \infty$ if $N = o(T^{w/4-1/2} \log^{w/4}(T))$. If $\gamma_T \to 0$, then for large enough T, N, $P(\widetilde{\boldsymbol{\xi}}_{H^c} \neq \mathbf{0}) = 1$.

The proof of Theorem 3 is relegated to section 7. In words, this theorems says that a zero-block consistent estimator for the spatial weight matrix exists and is given by the LASSO estimator with probability going to 1. The estimator is also a useful one in detecting block structure of the spatial weight matrix, in the sense that the diagonal blocks are estimated to be non-zero at the same time with probability 1, as long as the tuning parameter γ_T goes to 0.

With $\gamma_T \to 0$, the condition for the maximum block size $n = o(\{\gamma_T/\lambda_T\}^{2/3})$ implies that we need $n = o(T^{1/3} \log^{-1/3}(T \vee N))$. In practice, the method performs well even if the maximum block size is relatively large compared to T; see section 5 for simulation results. In theory, γ_T should be chosen to be small in order to align with $\gamma_T \to 0$. Yet if γ_T is too small, it will not allow for a block with reasonable size. And of course, γ_T cannot be set too large also, or the whole weight matrix is shrunk to zero, which is useless albeit still being zero-block consistent. In our simulations, we balance these concerns by considering

$$\gamma_T = \frac{C}{\sqrt{\log(T \vee N)}},$$

where C is a constant to be chosen, and $C = 0.5$ is recommended in section 5. We choose this form of γ_T since this function indeed goes to 0 as $T, N \to \infty$, has nice computational results in practice, and allows for a larger maximum block size in theory since it goes to 0 slowly. In practice this gives excellent results if detection of block structure is the main concern. See section 5 for an alternative using cross-validated tuning parameter, which geared towards within-block sensitivity more on top of blocks detection.

4 Relaxation for Overlapping Blocks

The spatial weight matrix in (2.3) and the theories presented in section 3 do not include the case where some of the blocks are overlapping. Yet in many practical cases, some or all of the blocks are slightly overlapping despite the non-overlapping majority.

Suppose there are $G \geq 2$ non-overlapping sets $I_1, \ldots, I_G \subset \{1, \ldots, N\}$ such that $w^*_{ij} = 0$ for $i \in I_a$ and $j \in I_b$ with $a \neq b$. Then $I_1, \ldots, I_G$ form G groups for the components of $\mathbf{y}_t$, with $G(G-1)$ corresponding zero blocks in the spatial weight matrix $\mathbf{W}^*$ if we order the components so that those in a set I_j are grouped together. We introduce extra conditions in this section so that the zero-block consistency in Theorem 3 is valid for the estimator of these zero blocks.

Define the set

$$H' = \{j : \xi^*_j = 0 \text{ and corr. to one of the } G(G-1) \text{ zero blocks in } W^*\}. \tag{4.3}$$

This set corresponds to H in (3.1) when the blocks are non-overlapping. Consider two additional assumptions below:

(i)' The spatial weight matrix $\mathbf{W}^*$ is such that, for $i \in I_q$, $q = 1, \ldots, G$, we have uniformly as $T, N \to \infty$,

$$\sum_{j \notin I_q} |\pi^*_{ij}| \leq c_\pi \lambda_T,$$

where c_π is a constant, and π^*_{ij} denotes the (i,j)-th element of $\mathbf{\Pi}^* = (\mathbf{I}_N - \mathbf{W}^*)^{-1}$.

(Rii) Define the set $I' = \{1, \ldots, N\}/\bigcup_{i=1}^G I_i$. The vector $\boldsymbol{\epsilon}_t$ can always be partitioned as

$$\boldsymbol{\epsilon}_t = (\boldsymbol{\epsilon}_{I_1}^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_{I_G}^{\mathrm{T}}, \boldsymbol{\epsilon}_{I'}^{\mathrm{T}})^{\mathrm{T}}.$$

Then we assume $\mathrm{cov}(\boldsymbol{\epsilon}_{I_i}, \boldsymbol{\epsilon}_{I_j}) = \mathbf{0}$ for $i \neq j$, and $\mathrm{cov}(\epsilon_{ti}, \epsilon_{tj}) \leq c_\epsilon \lambda_T$ for $i \in I_q$, $q = 1, \ldots, G$ and $j \in I'$, uniformly as $T, N \to \infty$, where $c_\epsilon > 0$ is a constant. Also, $\mathrm{var}(\epsilon_{ti}) \leq \sigma_\epsilon^2 < \infty$ uniformly as $T, N \to \infty$, where σ_ϵ^2 is a positive constant.

Assumption (i)' is an additional assumption on top of (i) in section 3.1. It says that the matrix $(\mathbf{I}_N - \mathbf{W}^*)^{-1}$ should also have approximately the same block structure as $\mathbf{W}^*$, where the elements corresponding to the zero blocks in $\mathbf{W}^*$ should be close to 0, with order specified. This assumption is likely to be true when the blocks are only slightly overlapping, which is what we are concerned with. Assumption (Rii) is to replace (ii) in section 3.1. It says that the noise series for those components not in any blocks should have only weak correlation with those noise series in blocks. Between blocks, the correlation should still be 0 for identifiability of block structure.

We are now ready to present a version of Theorem 3 for overlapping blocks.

Theorem 4. *Suppose there are overlapping blocks in $\mathbf{W}^*$. Under assumptions (i), (i)', (Rii) and (iii) - (vi), if $\lambda_T = o(\gamma_T)$ and $n = o(\{\gamma_T/\lambda_T\}^{2/3})$, then for large enough T, N, the LASSO solution $\widetilde{\boldsymbol{\xi}}$ in (2.5) is such that*

$$P(\widetilde{\boldsymbol{\xi}}_{H'} = \mathbf{0}) \geq P(A_\epsilon),$$

which approaches 1 as $T, N \to \infty$ if $N = o(T^{w/4-1/2} \log^{w/4}(T))$. If $\gamma_T \to 0$, then for large enough T, N, $P(\widetilde{\boldsymbol{\xi}}_{H'^c} \neq \mathbf{0}) = 1$.

This theorem is in parallel with Theorem 3. Zero-block consistency continues to hold even when there are overlapping blocks in the spatial weight matrix.

5 Practical Implementation

We use the Least Angle Regression algorithm (LARS) of Efron et al. (2004) to implement the minimization in (2.5). A unique solution is guaranteed since the minimization problem in (2.5) is convex. The LARS is very fast since the order of complexity of the algorithm is the same as that for ordinary least squares.

The choice of γ_T is a delicate issue. We have discussed the choice $\gamma_T = C/\sqrt{\log(T \vee N)}$ in section 3.2, which is excellent for blocks detection. Even though accurate estimation of within-block elements is not supported by theories, from our simulation results, we can achieve a higher within-block sensitivity by using cross-validation for finding an optimal γ_T, with prediction error as the criterion. Performance of block detection is slightly hindered, but the within-block non-zero elements are estimated more accurately using the cross-validated tuning parameter. Truly zero elements are more often estimated as non-zeros though. See section 5.1 for more details.

To find a tuning parameter by cross-validation, we first split the sample into a test set T_a with consecutive time points, and the rest as the validation set. We compute the LASSO solutions $\widetilde{\boldsymbol{\xi}}_{a,\gamma}$ using the LARS algorithm on a grid of values of γ for the test set. Then we solves

$$\gamma_{CV} = \arg\min_{\gamma} \sum_{a} \sum_{t \in T_a^c} \left\| \mathbf{y}_t - \widetilde{\mathbf{W}}_{a,\gamma} \mathbf{y}_t \right\|^2,$$

where $\widetilde{\mathbf{W}}_{a,\gamma}$ is the spatial weight matrix recovered from $\widetilde{\boldsymbol{\xi}}_{a,\gamma}$. In the simulations to follow, we compare the performance of using these two different tuning parameters.

5.1 Simulation Results

In this paper, we focus on block detection, and there are no theoretical supports for accurate estimation of elements of $\mathbf{W}^*$ in the non-zero diagonal blocks. We measure the performance of block detection using the *across-block specificity*, defined as the proportion of true zeros in the non-diagonal zero blocks estimated as zeros. For the sake of completeness and independent interest, we include other measures as well to gauge the overall performance of estimating $\mathbf{W}^*$. One is the *within-block sensitivity*, defined as the proportion of true non-zeros estimated as non-zeros, and the *within-block specificity*, defined as the proportion of true zeros in the diagonal blocks estimated as zeros. We also use the L_1 error bound $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1/(N(N-1))$ and the L_2 error bound $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|/\sqrt{N(N-1)}$ for comparing the overall estimation performance across different T, N combinations.

We generate the data using the model $\mathbf{y}_t = \mathbf{W}^*\mathbf{y}_t + \boldsymbol{\epsilon}_t$ for a given triplet (T, N, κ), where κ is the sparsity parameter controlling the overall sparsity of $\mathbf{W}^*$. We generate $\mathbf{W}^*$ by randomly selecting the number of blocks between 5 and 10 with uniform probability on their start and end points. Within all blocks, we choose $[(1-\kappa)N(N-1)]$ elements to be non-zeros with value 0.3. It means that a larger κ represents a sparser $\mathbf{W}^*$. Note that a relatively sparse $\mathbf{W}^*$ may have dense blocks, as the sparsity level is defined for the overall matrix $\mathbf{W}^*$. To ensure stationarity, each element w_{ij}^* of $\mathbf{W}^*$ is divided by $1.1 \times \max\left(1, \sum_{j=1}^N w_{ij}^*\right)$. The covariance matrix for $\{\boldsymbol{\epsilon}_t\}$ is defined in the same way, with the same sparsity κ. Hence the within-block pattern of spatial correlation is very general. In each iteration of the simulation, we generate both $\mathbf{W}^*$ and the data in order to ensure that the simulation is carried over a wide range of true models. Thus, the results are not influenced by a particular choice of $\mathbf{W}^*$.

Table 1 shows the simulation results with tuning parameter $\gamma_T = 0.5/\sqrt{\log(T \vee N)}$. It is clear that on average the estimator is zero-block consistent, since the across-block specificity is always close to 99% in all cases. While within-block accuracy is not guaranteed, the within-block specificity and sensitivity are quite good when T is large. The overall sparsity level is close to κ in all cases.

Table 2 shows the simulation results with cross-validated tuning parameter. While zero-block consistency is not as good as when using $\gamma_T = 0.5/\sqrt{\log(T \vee N)}$, it is still satisfactory. Moreover, the within-block

Table 1: Baseline Simulations.

		$\kappa = 0.90$			$\kappa = 0.95$		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$N = 25$	Within-Block Specificity	64.49%	58.32%	49.32%	88.20%	72.57%	71.84%
	Within-Block Sensitivity	76.20%	85.08%	91.55%	46.35%	77.02%	80.99%
	Across-Block Specificity	98.36%	98.86%	99.03%	99.77%	98.46%	99.86%
	L_1	0.0165	0.0139	0.0124	0.0113	0.0111	0.0101
	L_2	0.0771	0.0566	0.0447	0.0635	0.0510	0.0459
	Sparsity	89.17%	88.42%	87.63%	96.39%	92.23%	93.28%
$N = 50$	Within-Block Specificity	69.11%	64.22%	58.20%	86.80%	73.74%	73.57%
	Within-Block Sensitivity	61.13%	73.34%	83.18%	43.28%	75.83%	81.72%
	Across-Block Specificity	99.27%	99.34%	98.76%	99.89%	97.92%	99.59%
	L_1	0.0140	0.0120	0.0109	0.0106	0.0106	0.0094
	L_2	0.1126	0.0823	0.0643	0.1092	0.0859	0.0740
	Sparsity	90.96%	89.41%	87.30%	96.17%	91.12%	92.26%
$N = 75$	Within-Block Specificity	71.61%	67.20%	63.28%	78.76%	77.57%	74.53%
	Within-Block Sensitivity	51.01%	63.37%	74.31%	55.49%	65.95%	79.17%
	Across-Block Specificity	98.31%	98.67%	98.77%	98.33%	99.42%	99.43%
	L_1	0.0120	0.0104	0.0090	0.0102	0.0088	0.0082
	L_2	0.1211	0.0893	0.0675	0.1267	0.1007	0.0829
	Sparsity	91.00%	89.68%	88.39%	92.93%	93.11%	92.06%

Notes: Penalization parameter selected as $\gamma_T = c/\sqrt{\log(T \vee N)}$ with $c = 0.5$. Simulation size is 1,000.

sensitivity is increased in all cases. Basically, cross-validation is being conservative in penalization. Hence truly zero elements are estimated as non-zeros more often, so that both across-block and within-block specificity suffer, while within-block sensitivity increases due to less penalization from the cross-validated tuning parameter. It may worth identifying the zero blocks first using $\gamma_T = c/\sqrt{\log(T \vee N)}$, then using a cross-validated one for within block estimation.

5.2 Analysis of US Senate bill voting

How polarized is the United States Congress? Do congressmen vote exclusively along partisan lines or are there moments when partisanship gives way to consensus? To shed light on these questions, we use model 2.1 to analyze the voting records for the bills enacted and proposed by the United States Senate from 1993 to 2012, period from the first presidency of Bill Clinton to the first four years under Barack Obama. Polarized voting pattern should give at least two blocks in the spatial weight matrix, one corresponding to the Republicans, and another to the Democrats.

We use data compiled by `GovTrack.us`, a web site that freely keeps track of voting record in both houses. Vote is recorded as 1 for "yes", -1 for "no" and 0 for absent for all bills that were proposed in the period under study. To evaluate the evolution of polarization, we estimate the model within windows of each calendar year, representing the first half or second half of a particular meetings of the biannual legislative branch[1]. The composition of the Senate and the number of voting instances can be found in Table 3.

[1] Congresses begin and end at the third day of January in odd-numbered years. Bills voted in the first two days of January of odd years, if any, are discarded.

Table 2: Baseline Simulations with Cross Validation.

		$\kappa = 0.90$			$\kappa = 0.95$		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$N = 25$	Within-Block Specificity	52.90%	48.79%	44.93%	73.33%	62.87%	56.89%
	Within-Block Sensitivity	86.02%	90.15%	93.44%	73.05%	86.56%	96.62%
	Across-Block Specificity	93.08%	93.63%	93.57%	93.72%	92.22%	91.34%
	L_1	0.0177	0.0159	0.0153	0.0152	0.0144	0.0136
	L_2	0.0732	0.0577	0.0514	0.0680	0.0558	0.0483
	Sparsity	83.54%	82.93%	81.83%	87.93%	85.26%	83.83%
$N = 50$	Within-Block Specificity	64.76%	60.38%	57.55%	76.43%	71.31%	68.58%
	Within-Block Sensitivity	65.51%	75.78%	84.50%	63.85%	80.63%	90.38%
	Across-Block Specificity	95.93%	96.07%	96.55%	95.82%	94.38%	94.36%
	L_1	0.0151	0.0135	0.0118	0.0127	0.0123	0.0111
	L_2	0.1135	0.0887	0.0674	0.1080	0.0917	0.0745
	Sparsity	87.44%	85.81%	85.03%	90.12%	87.46%	86.80%
$N = 75$	Within-Block Specificity	69.09%	66.53%	63.57%	75.71%	74.57%	71.41%
	Within-Block Sensitivity	54.56%	63.24%	73.56%	60.10%	71.27%	83.60%
	Across-Block Specificity	96.84%	97.18%	97.52%	96.48%	96.70%	96.64%
	L_1	0.0126	0.0109	0.0095	0.0110	0.0098	0.0090
	L_2	0.1243	0.0921	0.0710	0.1298	0.1011	0.0843
	Sparsity	89.46%	88.26%	87.33%	91.03%	90.02%	89.17%

Notes: Tuning parameter γ_T chosen by five-fold Cross-Validation. Simulation size is 250.

Estimation is conducted in absolute disregard of party affiliation, and the tuning parameter γ_T is chosen by cross validation. Hence, specificity can slightly suffer here in light of the simulation results in section 5.1. Thus we are being conservative in drawing the conclusion that the spatial weight matrix has a block structure. The outcome for year 2012 is displayed in Figure 1. The estimated non-zero pairwise links are displayed as a solid line in grey, length of which does not carry any information on its intensity or direction and are purely determined by ease of visualization. The nodes are colored according to party affiliations: Democrats are represented by blue, Republicans by red, and Independents by white.

It is immediately clear from Figure 1 that the Senate behaves as two almost exclusive blocks or groups, defined exclusively along partisan lines, where the Independents behave most similarly to the Democrats. It seems that the two blocks slightly overlap each other, and the results in Theorem 4 can be applied. One Republican forms a block him/herself. Bear in mind that we are using a cross-validated tuning parameter, and hence we are being conservative already in concluding a block structure in the spatial weight matrix.

It is of interest to visualize the number of political collaborations and its evolution throughout the years. To achieve this, we build two measures of cross-partisanship association for a given year. The first is based on the ratio of links with ends on Senators from different parties to the overall number of links. We name this as "Cross-Party Connections". As seen in Figure 2, it is under 4% for all years under study. The second measure is the number of Senators who are the starting points of directed links towards colleagues from different parties, who are generically named "brokers". Both measures represent the number of Senators and links that appear in the frontier and, therefore, could represent collaborative cross-partisan political connections. Both measures show very limited collaboration if compared to the overall legislative activity[2]. It is concluded, therefore, that political affiliations are strong determinants of group identity. It also appears that frontier between the groups and scope for collaborative legislative work is very limited throughout the recent Senates history.

[2] For year 2012, the brokers were identified as Sen. Susan Collins [R, ME], Sen. Lisa Murkowski [R, AK], Sen. Jim DeMint [R, SC], Sen. Claire McCaskill [D, MO], Sen. Scott Brown [R, MA], Sen. Joe Manchin [D, WV], Sen. Rand Paul [R, KY] and Sen. Mike Lee [R, UT]. Other years available upon request.

Figure 1: Visualization of the estimated spatial weight matrix for voting, 2012.

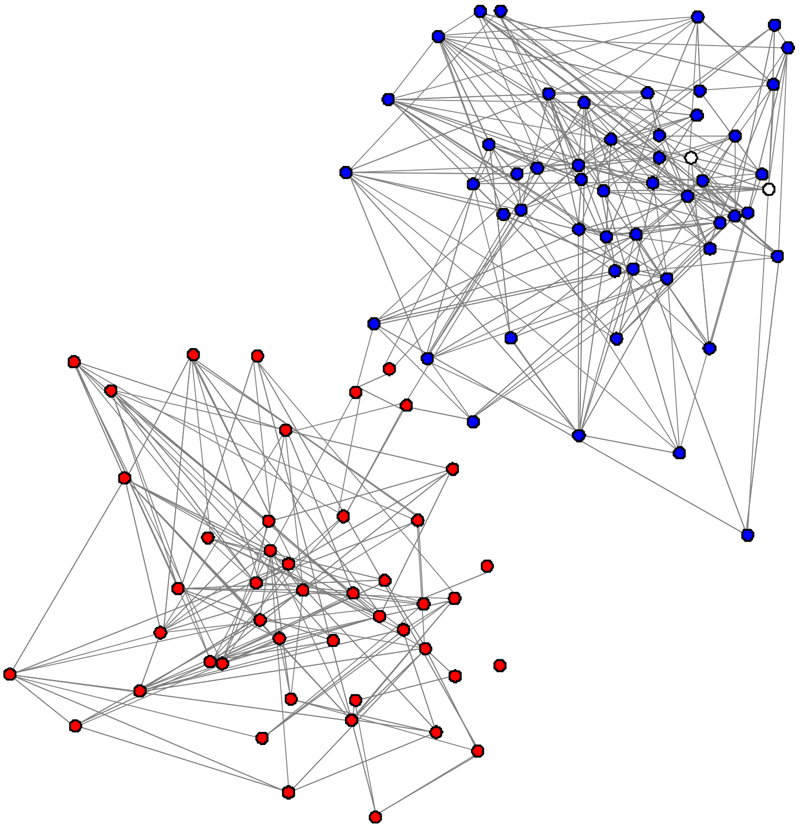

Figure 2: Cross-party collaboration.

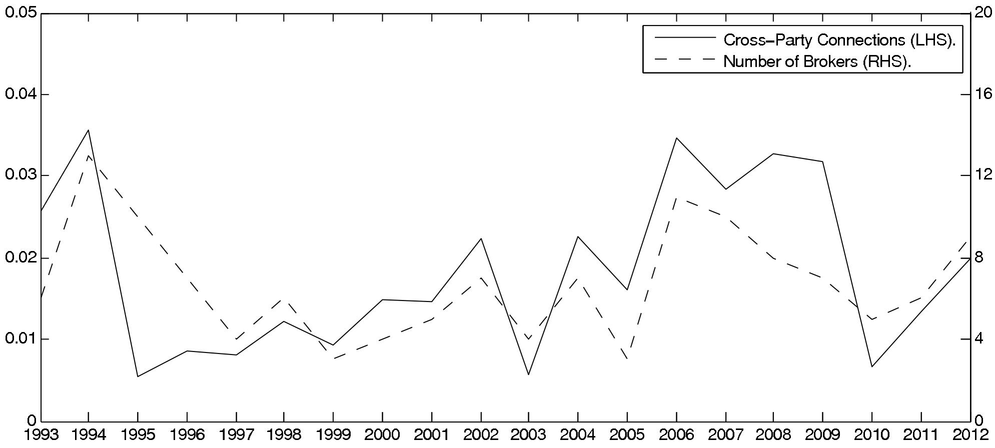

Table 3: Senate Composition.

Year	Congress	Rep	Dem	Ind	Votes
1993	103rd	46	55	0	395
1994					329
1995	104th	53	46	1	613
1996					306
1997	105th	54	45	1	298
1998					314
1999	106th	55	45	1	374
2000					298
2001	107th	49	50	1	380
2002					253
2003	108th	51	48	1	459
2004					216
2005	109th	54	45	1	366
2006					279
2007	110th	49	50	2	442
2008					215
2009	111th	41	61	2	397
2010					299
2011	112th	47	51	2	235
2012					251

6 Conclusion

We developed the LASSO penalization for detecting block structure in a spatial weight matrix, when the size of the panel can be close to the sample size. One distinct feature of our model is the absence of covariates, which is motivated by the US senate voting data example analyzed in this paper. Also, there is no need for the decay of variance of the noise series, like Lam and Souza (2013) does. One contribution of the paper is the derivation of the probability lower bound for the LASSO estimator to be zero-block consistent - a concept that an estimator correctly estimates the non-diagonal zero blocks as zero. We also proved that the diagonal blocks of the estimator are not all zero with probability 1, so that block structure becomes apparent in the estimator. We use the LARS algorithm for practical computation, which is well-established for solving LASSO minimization efficiently, with computational order the same as ordinary least squares iterations. The estimated spatial weight matrix is visualized by a graph with directional edges between components. The absence of edges between two groups of components indicates two blocks. We also allow for the fact that blocks sometimes can overlap slightly, and develop the corresponding theories to show that zero-block consistency still holds in the case of overlapping blocks. The US senate voting data example demonstrates clearly the case of slightly overlapping blocks.

Our proofs utilize results from random matrix theories for bounding extreme eigenvalues of a sample covariance matrix, as well as a Nagaev-type inequality for finding the tail probability of a general time series process. These results can be useful for the theoretical development of other time series researches.

7 Appendix

Proof of Theorem 1. For a random variable z, define the norm $\|z\|_a = [E|z|^a]^{1/a}$. We need to show that there are some constants $\mu, C > 0, w > 2$ and $\alpha > 1/2 - 1/w$ such that

$$\max_{1\leq j\leq N} \left\|\epsilon_{tj}\right\|_{2w} \leq \mu, \tag{7.1}$$

$$\sum_{t=m}^{\infty} \max_{1\leq j\leq N} \left\|\epsilon_{tj} - \epsilon'_{tj}\right\|_{2w} \leq Cm^{-\alpha}, \tag{7.2}$$

where $\boldsymbol{\epsilon}'_t$ has exactly the same causal definition as $\boldsymbol{\epsilon}_t$ as in assumption (iv) with the same values of $\boldsymbol{\Phi}_i$'s and $\boldsymbol{\eta}_j$'s, except for $\boldsymbol{\eta}_0$, which is replaced by an independent and identically distributed copy $\boldsymbol{\eta}'_0$. With (7.1) and (7.2), we can use Lemma 1 of Lam and Souza (2013) for the product process $\{\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})\}$ to complete the proof.

To prove (7.1), by the Fubini's Theorem and assumption (v),

$$\begin{aligned} E|\epsilon_{tj}|^{2w} &= E\int_0^{|\epsilon_{tj}|^{2w}} ds = \int_0^{\infty} P(|\epsilon_{tj}| > s^{1/2w})\, ds \leq \int_0^{\infty} D_1 \exp(-D_2 s^{q/2w})\, ds \\ &= \frac{4wD_1}{q}\int_0^{\infty} x^{4w/q-1}e^{-D_2x^2}\, dx = \frac{2wD_1}{qD_2^{2w/q}}\Gamma(2w/q) = \mu^{2w} < \infty, \end{aligned} \tag{7.3}$$

so that $\max_{1\leq j\leq N}\left\|\epsilon_{tj}\right\|_{2w} \leq \mu < \infty$ for any $w > 0$. This proves (7.1).

To prove (7.2), denote $\boldsymbol{\phi}_{ij}^{\mathrm{T}}$ the j-th row of $\boldsymbol{\Phi}_i$. Then using the causal definition in assumption (iv),

$$|\epsilon_{tj} - \epsilon'_{tj}| = |\boldsymbol{\phi}_{tj}^{\mathrm{T}}(\boldsymbol{\eta}_0 - \boldsymbol{\eta}'_0)| \leq \left\|\boldsymbol{\phi}_{tj}\right\|_1 \max_{i\in J_{tj}} |\eta_{0i} - \eta'_{oi}|,$$

where J_{tj} is the index set of non-zeros in $\boldsymbol{\phi}_{tj}$ as defined in assumption (vi). Hence by assumption (v) on η_{0i} and the calculations in (7.3),

$$\begin{aligned} \left\|\epsilon_{tj} - \epsilon'_{tj}\right\|_{2w} &\leq \left\|\boldsymbol{\phi}_{tj}\right\|_1 \Big[E\big\{\max_{i\in J_{tj}} |\eta_{0i} - \eta'_{0i}|^{2w}\big\}\Big]^{\frac{1}{2w}} \\ &\leq \left\|\boldsymbol{\phi}_{tj}\right\|_1 |J_{tj}|^{\frac{1}{2w}} \max_{i\in J_{tj}} \left\|\eta_{0i} - \eta'_{0i}\right\|_{2w} \\ &\leq \left\|\boldsymbol{\phi}_{tj}\right\|_1 |J_{tj}|^{\frac{1}{2w}} (\max_{i\in J_{tj}} \left\|\eta_{0i}\right\|_{2w} + \max_{i\in J_{tj}} \left\|\eta'_{0i}\right\|_{2w}) \\ &\leq 2\mu\left\|\boldsymbol{\phi}_{tj}\right\|_1 |J_{tj}|^{\frac{1}{2w}}, \end{aligned}$$

so that by assumption (vi), using the same $w > 2$ in the assumption,

$$\begin{aligned} \sum_{t=m}^{\infty} \max_{1\leq j\leq N} \left\|\epsilon_{tj} - \epsilon'_{tj}\right\|_{2w} &\leq 2\mu \sum_{t=m}^{\infty} \max_{1\leq j\leq N} \left\|\boldsymbol{\phi}_{tj}\right\|_1 \max_{1\leq j\leq N} |J_{tj}|^{\frac{1}{2w}} \\ &\leq 2\mu \max_{t,j} |J_{tj}|^{\frac{1}{2w}} \sum_{t=m}^{\infty} \left\|\boldsymbol{\Phi}_t\right\|_{\infty} \\ &\leq 2\mu \max_{t,j} |J_{tj}|^{\frac{1}{2w}} Cm^{-\alpha}\big(\max_{t,j} |J_{tj}|\big)^{-\frac{1}{2w}} \\ &= 2\mu Cm^{-\alpha}, \end{aligned}$$

which is (7.2) since μ, C are constants. This completes the proof of the theorem. □

Proof of Theorem 3. Define the set

$$D = \{j : j \notin H,\ \xi_j^* \text{ does not correspond to the diagonal of } \mathbf{W}^*\},$$

and define $J = D \cup H$. Hence J contains indices for ξ_i not corresponding to the diagonal of $\mathbf{W}^*$.

The KKT condition implies that $\widetilde{\boldsymbol{\xi}}$ is a solution to (2.5) if and only if there exists a subgradient

$$\mathbf{g} = \partial|\widetilde{\boldsymbol{\xi}}| = \left\{ \mathbf{g} \in \mathbb{R}^{2N^2} : \begin{cases} g_i = 0, & i \in J^c; \\ g_i = \text{sign}(\widetilde{\xi}_i), & \widetilde{\xi}_i \neq 0; \\ |g_i| \leq 1, & \text{otherwise.} \end{cases} \right\}$$

such that, differentiating the expression to be minimized in (2.5) with respect to $\boldsymbol{\xi}_J$,

$$\frac{1}{T}\mathbf{Z}_J^{\mathrm{T}}\mathbf{Z}_J - \frac{1}{T}\mathbf{Z}_J^{\mathrm{T}}\mathbf{y} = -\gamma_T \mathbf{g}_J,$$

where the notation $\mathbf{A}_S$ represents the matrix $\mathbf{A}$ restricted to the columns with index $j \in S$. Using $\mathbf{y} = \mathbf{Z}_J\boldsymbol{\xi}_J^* + \boldsymbol{\epsilon}$, the equation above can be written as

$$\frac{1}{T}\mathbf{Z}_J^{\mathrm{T}}\mathbf{Z}_J(\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J^*) - \frac{1}{T}\mathbf{Z}_J^{\mathrm{T}}\boldsymbol{\epsilon} = -\gamma_T \mathbf{g}_J.$$

For $\widetilde{\boldsymbol{\xi}}$ to be zero-block consistent, we need $\widetilde{\boldsymbol{\xi}}_H = \mathbf{0}$, implying $\mathbf{Z}_J(\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J^*) = \mathbf{Z}_D(\widetilde{\boldsymbol{\xi}}_D - \boldsymbol{\xi}_D^*)$. Hence, the KKT condition implies that $\widetilde{\boldsymbol{\xi}}$ is a zero-block consistent solution if and only if

$$\begin{aligned} &\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D(\widetilde{\boldsymbol{\xi}}_D - \boldsymbol{\xi}_D^*) - \frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon} = -\gamma_T \mathbf{g}_H, \\ &\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D(\widetilde{\boldsymbol{\xi}}_D - \boldsymbol{\xi}_D^*) - \frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon} = -\gamma_T \mathbf{g}_D, \end{aligned} \tag{7.4}$$

which can be simplified to

$$\Big|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big)^{-1}\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon} - \gamma_T\mathbf{g}_D\Big) - \frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon}\Big| \leq \gamma_T, \tag{7.5}$$

since $\mathbf{g}_H$ has elements less than or equal to 1.

We now show that, on the set A_ϵ as defined in (3.2), (7.5) is true for large enough T, N, thus completing the proof of zero-block consistency of $\widetilde{\boldsymbol{\xi}}$. To this end, there are four terms we need to bound. Define $I_1, \ldots, I_G \subset \{1, \ldots, N\}$ to be the index sets for the G groups of components as in (2.3). Then, consider on the set A_ϵ,

$$\begin{aligned} \Big\|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon}\Big\|_{\max} &= \max_{i\in I_q, j\notin I_q}\Big|\frac{1}{T}\sum_{t=1}^{T} y_{ti}\epsilon_{tj}\Big| = \max_{i\in I_q, j\notin I_q}\Big|\sum_{s\in I_q}\pi_{is}^*\Big(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}\Big)\Big| \\ &\leq \lambda_T \max_{1\leq i\leq N}\sum_{s=1}^{N}|\pi_{is}^*| \leq \frac{\lambda_T}{1-\eta}, \end{aligned} \tag{7.6}$$

where we used (2.2) and $y_{ti} = \sum_{j\in I_q}\pi_{ij}^*\epsilon_{tj}$ for $i \in I_q$ for some q, with π_{ij}^* being the (i,j)-th element of $\boldsymbol{\Pi}^* = (\mathbf{I}_N - \mathbf{W}^*)^{-1}$. The last line follows from assumption (ii) that $\text{cov}(\epsilon_{ti}, \epsilon_{tj}) = 0$ if i and j correspond to different groups, so that on A_ϵ, $|T^{-1}\sum_{t=1}^T \epsilon_{ts}\epsilon_{tj}| \leq \lambda_T$. We also used assumption (i) to arrive at

$$\max_{1\leq i\leq N}\sum_{s=1}^N |\pi^*_{is}| = \big\|\mathbf{\Pi}^*\big\|_\infty \leq \big\|\mathbf{I}_N\big\|_\infty + \sum_{k\geq 1}\big\|\mathbf{W}^*\big\|_\infty^k \leq 1 + \sum_{k\geq 1}\eta^k = \frac{1}{1-\eta}.$$

A potentially larger term is, by similar calculations on A_ϵ,

$$\Big\|\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon}\Big\|_{\max} = \max_{i\in I_q, j\in I_{q'}}\left|\sum_{s\in I_q}\pi^*_{is}\Big(\frac{1}{T}\sum_{t=1}^T \epsilon_{ts}\epsilon_{tj}\Big)\right| \leq \frac{\sigma_\epsilon^2+\lambda_T}{1-\eta}, \tag{7.7}$$

where we used assumption (ii) that $\mathrm{var}(\epsilon_{tj}) \leq \sigma_\epsilon^2$. We also have, on A_ϵ,

$$\Big\|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D\Big\|_\infty \leq n \max_{i\in I_q, j\notin I_q}\left|\frac{1}{T}\sum_{t=1}^T y_{ti}y_{tj}\right| = n \max_{\substack{i\in I_q, j\in I_{q'}\\ q\neq q'}}\left|\sum_{s\in I_q,\ell\in I_{q'}}\pi^*_{is}\pi^*_{j\ell}\Big(\frac{1}{T}\sum_{t=1}^T\epsilon_{ts}\epsilon_{t\ell}\Big)\right| \leq \frac{\lambda_T n}{(1-\eta)^2}. \tag{7.8}$$

Finally, let $\sigma_{\max}(\mathbf{A}) = \lambda_{\max}^{1/2}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ denotes the maximum singular value of the matrix $\mathbf{A}$, and $\sigma_{\min}(\mathbf{A})$ the smallest one. Then

$$\begin{aligned}\Big\|\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big)^{-1}\Big\|_\infty &\leq n^{1/2}\lambda_{\min}^{-1}\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big) \leq n^{1/2}\lambda_{\min}^{-1}\Big(\frac{1}{T}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\Big) = n^{1/2}\lambda_{\min}^{-1}\Big(\frac{1}{T}\sum_{t=1}^T\mathbf{y}_t\mathbf{y}_t^{\mathrm{T}}\Big)\\ &= n^{1/2}\lambda_{\min}^{-1}\bigg(\mathbf{\Pi}^*\Big(\frac{1}{T}\sum_{t=1}^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}}\Big)\mathbf{\Pi}^{*\mathrm{T}}\bigg) \leq n^{1/2}\sigma_{\min}^{-2}(\mathbf{\Pi}^*)\lambda_{\min}^{-1}\Big(\frac{1}{T}\sum_{t=1}^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}}\Big).\end{aligned} \tag{7.9}$$

To bound (7.9), we have

$$\sigma_{\min}^{-2}(\mathbf{\Pi}^*) = \sigma_{\max}^2(\mathbf{I}_N - \mathbf{W}^*) \leq (1+\sigma_{\max}(\mathbf{W}^*))^2 \leq \big(1+\big\|\mathbf{W}^*\big\|_1^{1/2}\big\|\mathbf{W}^*\big\|_\infty^{1/2}\big)^2 \leq (1+\eta^{1/2}\eta_c^{1/2})^2, \tag{7.10}$$

where we used assumption (i) for bounding $\big\|\mathbf{W}^*\big\|_1$ and $\big\|\mathbf{W}^*\big\|_\infty$.

Also, the conditions assumed in assumption (iv) for the η_{ti}'s ensure that Theorem 5.11 on the extreme eigenvalues of a sample covariance matrix in Bai and Silverstein (2010) can be applied. Hence, for each integer $i \geq 0$, we have

$$\lim_{T\to\infty}\lambda_{\min}\Big(\frac{1}{T}\sum_{t=1}^T\boldsymbol{\eta}_{t-i}\boldsymbol{\eta}_{t-i}^{\mathrm{T}}\Big) = \sigma^2(1-\sqrt{d})^2, \quad \lim_{T\to\infty}\lambda_{\max}\Big(\frac{1}{T}\sum_{t=1}^T\boldsymbol{\eta}_{t-i}\boldsymbol{\eta}_{t-i}^{\mathrm{T}}\Big) = \sigma^2(1+\sqrt{d})^2$$

almost surely, where d is specified in assumption (iii). For each i, let U_i be the almost sure set such that the above limits hold. Then on the almost sure set $U = \bigcap_{i\geq 0} U_i$, the above limits hold for all integers $i \geq 0$. Hence on U, for large enough T, N, we have

$$\lambda_{\min}^{1/2}\Big(\frac{1}{T}\sum_{t=1}^T\boldsymbol{\eta}_t\boldsymbol{\eta}_t^{\mathrm{T}}\Big) \geq \sigma(1-\sqrt{d}) - \epsilon, \quad \lambda_{\max}^{1/2}\Big(\frac{1}{T}\sum_{t=1}^T\boldsymbol{\eta}_t\boldsymbol{\eta}_t^{\mathrm{T}}\Big) \leq \sigma(1+\sqrt{d}) + \epsilon,$$

where the constant ϵ is as in assumption (iv). Therefore, on U, for large enough T, N, we have

$$\begin{aligned}
\lambda_{\min}\Big(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}}\Big) &= \sigma_{\min}^2\Big(T^{-1/2}\sum_{i\geq 0}\boldsymbol{\Phi}_i(\boldsymbol{\eta}_{1-i},\ldots,\boldsymbol{\eta}_{T-i})\Big)\\
&\geq \left\{\sigma_{\min}\big(T^{-1/2}(\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_T)\big) - \sum_{i\geq 1}\sigma_{\max}\big(\boldsymbol{\Phi}_i T^{-1/2}(\boldsymbol{\eta}_{1-i},\ldots,\boldsymbol{\eta}_{T-i})\big)\right\}^2\\
&\geq \left\{\lambda_{\min}^{1/2}\Big(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\eta}_t\boldsymbol{\eta}_t^{\mathrm{T}}\Big) - \sum_{i\geq 1}\big\|\boldsymbol{\Phi}_i\big\|\lambda_{\max}^{1/2}\Big(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\eta}_{t-i}\boldsymbol{\eta}_{t-i}^{\mathrm{T}}\Big)\right\}^2\\
&\geq \left\{\sigma(1-\sqrt{d}) - \epsilon - (\sigma(1+\sqrt{d})+\epsilon)\sum_{i\geq 1}\big\|\boldsymbol{\Phi}_i\big\|\right\}^2 \geq c^2, \qquad (7.11)
\end{aligned}$$

where $c > 0$ is a constant as in assumption (iv). Combining (7.10) and (7.11), on U and for large enough T, N, (7.9) becomes

$$\Big\|\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big)^{-1}\Big\|_{\infty} \leq \frac{n^{1/2}(1+\eta^{1/2}\eta_c^{1/2})^2}{c^2}. \qquad (7.12)$$

Hence combining the bounds (7.6), (7.7), (7.8) and (7.12), on $A_\epsilon \cap U$, for large enough T, N, we have

$$\begin{aligned}
&\Big|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big)^{-1}\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon} - \gamma_T\mathbf{g}_D\Big) - \frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon}\Big|\\
&\leq \Big\|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D\Big\|_{\infty}\Big\|\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big)^{-1}\Big\|_{\infty}\Big\|\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon} - \gamma_T\mathbf{g}_D\Big\|_{\max} + \Big\|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon}\Big\|_{\max}\\
&\leq \frac{\lambda_T n^{3/2}(1+\eta^{1/2}\eta_c^{1/2})^2}{(1-\eta)^2c^2}\left(\frac{\sigma_\epsilon^2+\lambda_T}{1-\eta}+\gamma_T\right) + \frac{\lambda_T}{1-\eta}\\
&= O(\lambda_T n^{3/2}) = o(\gamma_T),
\end{aligned}$$

by the assumption $n = o(\{\gamma_T/\lambda_T\}^{2/3})$. Hence on $A_\epsilon \cap U$, (7.5) is satisfied for large enough T, N, so that $\widetilde{\boldsymbol{\xi}}$ is zero-block consistent, i.e. $\widetilde{\boldsymbol{\xi}}_H = \mathbf{0}$. It is clear then for large enough T, N, $A_\epsilon \cap U \subseteq \{\widetilde{\boldsymbol{\xi}}_H = \mathbf{0}\}$, and hence

$$P(\widetilde{\boldsymbol{\xi}}_H = \mathbf{0}) \geq P(A_\epsilon \cap U) = P(A_\epsilon),$$

since U is an almost sure set. The part where $P(A_\epsilon) \to 1$ if $N = o(T^{w/4-1/2}\log^{w/4}(T))$ is given by the results of Corollary 2. This completes the proof of the first half of Theorem 3.

For the second half, suppose $\widetilde{\boldsymbol{\xi}}_D = \mathbf{0}$. Then using (7.4), we have

$$\mathbf{g}_D = \frac{1}{\gamma_T}\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon} + \frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\boldsymbol{\xi}_D^*\Big) = \frac{1}{\gamma_T}\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{y}\Big).$$

One of the element of $\mathbf{g}_D$ is, for some j, with T, N large enough and on U,

$$\frac{1}{\gamma_T}\Big(\frac{1}{T}\sum_{t=1}^{T}y_{tj}^2\Big) = \frac{1}{\gamma_T}\Big(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\pi}_j^{*\mathrm{T}}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^T\boldsymbol{\pi}_j^*\Big) \geq \frac{\big\|\boldsymbol{\pi}_j^*\big\|^2}{\gamma_T}\lambda_{\min}\Big(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^T\Big) \geq \frac{c^2}{\gamma_T},$$

where $\boldsymbol{\pi}_j^{\mathrm{T}}$ is the j-th row of $\boldsymbol{\Pi}^*$, with $\big\|\boldsymbol{\pi}_j^*\big\| > 1$, and we used (7.11). Since $\gamma_T \to 0$, we have just proved that this particular element goes to infinity as $T, N \to \infty$, which is a contradiction since all elements in

$\mathbf{g}_D$ are less than or equal to 1 in magnitude. Hence we must have $\widetilde{\boldsymbol{\xi}}_D \neq \mathbf{0}$ for large enough T, N. This completes the proof of the theorem. □

Proof of Theorem 4. Define the set

$$D' = \{j : j \notin H', \xi_j \text{ does not correspond to the diagonal of } \mathbf{W}^*\}.$$

Then the proof of this theorem is almost exactly the same as that for Theorem 3 by replacing D with D' and H with H'. The only differences are the bounds in (7.6) and (7.8). Consider, on A_ϵ,

$$\begin{aligned}
\Big\|\frac{1}{T}\mathbf{Z}_{H'}^{\mathrm{T}}\boldsymbol{\epsilon}\Big\|_{\max} &= \max_{i\in I_q, j\notin I_q}\Big|\frac{1}{T}\sum_{t=1}^{T} y_{ti}\epsilon_{tj}\Big| = \max_{i\in I_q, j\notin I_q}\Big|\sum_{s\in I_q}\pi^*_{is}\Big(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}\Big) + \sum_{s\notin I_q}\pi^*_{is}\Big(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}\Big)\Big| \\
&\leq \max_{s\in I_q, j\notin I_q}\Big|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}\Big|\|\boldsymbol{\Pi}^*\|_\infty + \max_{s\notin I_q, j\notin I_q}\Big|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}\Big|\max_{i\in I_q}\sum_{s\notin I_q}|\pi^*_{is}| \\
&\leq \frac{\lambda_T + c_\epsilon\lambda_T}{1-\eta} + (\sigma_\epsilon^2 + \lambda_T)c_\pi\lambda_T = O(\lambda_T),
\end{aligned} \tag{7.13}$$

where we used assumption (Rii) that $\mathrm{cov}(\epsilon_{ts}, \epsilon_{tj}) \leq c_\epsilon\lambda_T$ when $s \in I_q$ for some q and $j \notin I_\ell$ for any ℓ, and assumption (i)' that $\sum_{j\notin I_q}|\pi^*_{ij}| \leq c_\pi\lambda_T$ for $i \in I_q$. Also, on A_ϵ,

$$\begin{aligned}
\Big\|\frac{1}{T}\mathbf{Z}_{H'}^{\mathrm{T}}\mathbf{Z}_{D'}\Big\|_\infty &\leq n \max_{i\in I_q, j\notin I_q}\Big|\sum_{s\in I_q}\pi^*_{js}\Big(\frac{1}{T}\sum_{t=1}^{T}y_{ti}\epsilon_{ts}\Big) + \sum_{s\notin I_q}\pi^*_{js}\Big(\frac{1}{T}\sum_{t=1}^{T}y_{ti}\epsilon_{ts}\Big)\Big| \\
&\leq n\left(\frac{\sigma_\epsilon^2 + \lambda_T}{1-\eta}\right)c_\pi\lambda_T + n\lambda_T\left(\frac{1+c_\epsilon}{1-\eta} + c_\pi(\sigma_\epsilon^2 + \lambda_T)\right)\frac{1}{1-\eta} = O(\lambda_T n),
\end{aligned} \tag{7.14}$$

where we used (7.13) in the last line. The rates in (7.13) and (7.14) are the same as (7.6) and (7.8) respectively, and hence the results in Theorem 3 follows. □

References

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

Bai, Z. and J. Silverstein (2010). *Spectral Analysis of Large Dimensional Random Matrices* (2 ed.). New York: Springer Series in Statistics.

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

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.

Case, A. C. (1991). Spatial patterns in household demand. *Econometrica 59*, 953–965.

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

Kelejian, H. H. and I. R. Prucha (2002, November). 2sls and ols in a spatial autoregressive model with equal spatial weights. *Regional Science and Urban Economics 32*(6), 691–707.

Lam, C. and P. C. L. Souza (2013). Regularization for spatial panel time series using the adaptive lasso. Manuscript.

Lee, L.-F. (2002, April). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory 18*(02), 252–277.

Pinkse, J. and M. E. Slade (2010). The future of spatial econometrics. *Journal of Regional Science 50*(1), 103–117.

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

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