Detection and estimation of block structure in spatial weight matrix

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

In many economic applications, it is often of interest to categorize, classify or label individuals by groups based on similarity of observed behavior. We propose a method that captures group affiliation or, equivalently, estimates the block structure of a neighboring matrix embedded in a Spatial Econometric model. The main results of the LASSO estimator shows that off-diagonal block elements are estimated as zeros with high probability, property defined as "zero-block consistency". Furthermore, we present and prove zero-block consistency for the estimated spatial weight matrix even under a thin margin of interaction between groups. The tool developed in this paper can be used as a verification of block structure by applied researchers, or as an exploration tool for estimating unknown block structures. We analyzed the US Senate voting data and correctly identified 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 Associate Professor, 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

Classification problems are a common endeavor in Economics and Econometrics research. This is the problem of identifying and assigning individuals to groups based on their observed behavior or common characteristics. This problem can come in many formats. Examples include estimating groups of countries such that their income levels are mutually dependent, industrial inter-linkages and many issues regarding strategic interaction among economic agents. In the nonparametric case, see the classical examples in Ferraty and Vieu (2006). Identification of groups can be used to improve prediction, or can itself be the main purpose of a study.

A spatial weight matrix $\mathbf{W}$ can be used to indicate the existence of groups which are represented as diagonal blocks, producing a block diagonal matrix $\mathbf{W}$. Elements w_{ij} that fall outside blocks are therefore zero, indicating that there is no connection between individuals i and j. The classification into groups can describe, for example, *de facto* political parties operating at a Congress, abstracting from self-denominated labels. Political history is full of examples where parties operate jointly, pressing for a single agenda, thus behaving like a single political entity. Another example is defector policymakers, who effectively operate in a more similar way to political parties other than the one he or she pledged alliance. In both cases, it is useful to have an empirical tool that classifies individuals into groups, independently of labeled political affiliation.

The purpose of this paper is to show the properties of a LASSO-based estimator that uncovers the block structure of an unknown spatial weight matrix when only the outcomes (the response variables) are observed. Estimating the block structure of a spatial weight matrix is also a useful addition to the Spatial Econometrics literature, which usually assumes a known spatial weight matrix using expert knowledge, or more often just rough proxies like the inverse of "distances" or its arbitrary powers.

As shown in Arbia and Fingleton (2008) and Pinkse and Slade (2010), estimation accuracy of other parameters in a spatial lag/error model depends crucially on the correct specification of the spatial weight matrix. With these concerns in mind, there are other attempts in the literature to estimate the spatial weight matrix together with other important 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 theoretical 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 was first developed in Zou (2006) for variable selection problems in standard regression. They provided theoretical 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. However, in their paper, the authors assumed the existence of exogenous covariates, which are not necessarily observed in a setting when the interest lies purely on classifying individuals into groups.

In this paper, our objective is to estimate the block structure of the spatial weight matrix in a spatial lag/error model *in the absence of exogenous covariates* (see model (2.3) and section 2 for details in how we arrive at such a model for estimation). We then propose a LASSO estimator that captures with high probability all the zeros that fall outside blocks of interactions, property defined as "zero-block consistency". We can also estimate the diagonal blocks to be non-zero with probability 1. In section 4, we show zero-block consistency of the LASSO estimator of a spatial weight matrix even when there is a slight overlap between the groups. In other words, there is a small number of "hybrid" individuals.

Motivated by a set of US Senate voting data, in this paper we use the method to explore if the Republicans and the Democrats form two major blocks based on their voting records. We find that along the year of 2013, the method correctly identifies two groups, with Independent Senators behaving mostly as Democrats. The margin of interaction – defined as the Senators with cross-partisan links – is as small as seven Senators, a clear indication of strong polarization in the political chamber. Interestingly, for retrospective years, the degree of interaction was substantially higher, spiking at the last years of the Bush administration.

An interesting computational aspect of a spatial weight matrix with blocks of zeros in the off-diagonal is that we can store it in the computer as a banded matrix which reduces the amount of memory used. This provides another motivation for the development of our estimators in this paper to detect block structure in the spatial weight matrix.

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

One of the most commonly-used model for describing spatial interaction in a panel 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 for example equation (19.5) of Anselin et al. (2006), 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 so-called spatial weight matrix $\mathbf{W}$ has elements that express the strength of interaction between location i (row) and j (column). Therefore, the spatial weight matrix $\mathbf{W}$ 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). Such a structure is assumed to be constant across time points $t = 1, \ldots, T$. The parameter ρ is called the spatial autoregressive coefficient. The spatial lag model (2.1) is typically considered as the specification of the equilibrium outcome of a spatial or social interaction process, in which the value of the dependent variable for one agent is jointly determined with that of the neighboring

agents (Elhorst, 2010). As an example, in the empirical literature on strategic interaction among local governments (Brueckner, 2003), the spatial lag model is theoretically consistent with the situation where taxation and expenditures on public services interact with that in nearby jurisdictions.

To utilize model (2.1), the spatial weight matrix $\mathbf{W}$ has to be specified. Yet, recent researches suggest that the estimation accuracy of the model depends crucially on the correct specification of $\mathbf{W}$. See Arbia and Fingleton (2008) and Pinkse and Slade (2010) for some empirical experiments on this. Moreover, Lemma 2 of Lam and Souza (2013) also shows that if the estimation of $\mathbf{W}$ is not good enough, estimation accuracy of $\boldsymbol{\beta}$ can potentially suffer. Furthermore, 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.

Observing the drawbacks of model (2.1), Lam and Souza (2013) proposes to estimate the spatial weight matrix together with other parameters in the model, using

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

The term $\rho\mathbf{W}$ in model (2.1) is replaced by the spatial weight matrix $\mathbf{W}_1$, to be estimated from the data. The addition of matrix $\mathbf{W}_2$ is a generalization to model (2.1). Model (2.2) allows the spatial weight matrix to be estimated from the data, which overcomes the various drawbacks that are mentioned in the paragraph above when using a spatial lag model. They showed, among various results, that the elements of the spatial weight matrix can be sign-consistently estimated using the adaptive LASSO, i.e. the non-zeros in $\mathbf{W}_1$ and $\mathbf{W}_2$ are estimated with the correct signs, and the zeros in them are estimated as zeros, with probability going to 1.

In this paper, we are motivated to estimate the block structure of a spatial weight matrix. As our primary interest resides is detecting or classifying groups of individuals based on their outcome variables, it is not always the case that exogenous covariates exist or be relevant to a particular empirical question. For example, for the US senators' data, the main objective is to classify them into different *de facto* parties, irrespective of other potential variables that could explain observed behavior. As a consequence, the results in Lam and Souza (2013) cannot be directly applied.

This motivates us to study the following model:

$$\mathbf{y}_t = \mathbf{W}^*\mathbf{y}_t + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{2.3}$$

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. This model is in fact model (1.6) in LeSage and Pace (2008), with the term ρC there replaced by the spatial weight matrix $\mathbf{W}^*$, to be estimated from data.

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. This ensures that $(\mathbf{I}_N - \mathbf{W}^*)^{-1}$ exists, so that $\mathbf{y}_t = (\mathbf{I}_N - \mathbf{W}^*)^{-1}\boldsymbol{\epsilon}_t$ is stationary. Model (2.3) allows us to study the dependence of one dependent variable on the neighboring ones. In the context of the US senate voting data analysis to be carried out in section 5.3, we are studying the dependence structure of one senator's voting pattern on the other senators, which is captured by the spatial weight matrix $\mathbf{W}^*$. Note that there were other attempts to estimate connectedness in the US Congress in the literature. See, for example,

Fowler (2006).

Since we are interested in studying the block structure of $\mathbf{W}^*$, without loss of generality, we assume the components of $\mathbf{y}_t$ are sorted so that the spatial weight matrix $\mathbf{W}^*$ is block diagonal, with

$$\mathbf{W}^* = \begin{pmatrix} \mathbf{W}_1^* & & \\ & \ddots & \\ & & \mathbf{W}_G^* \end{pmatrix}, \quad \boldsymbol{\epsilon}_t = \begin{pmatrix} \boldsymbol{\epsilon}_t^{(1)} \\ \vdots \\ \boldsymbol{\epsilon}_t^{(G)} \end{pmatrix}, \tag{2.4}$$

where G is the number of blocks in $\mathbf{W}^*$. The blocks will potentially represent the dependence structure of voting patterns of senators from within the Republican, the Democrats, and other parties in the US senate voting data. An important assumption for $\{\epsilon_t\}$ is that $\text{cov}(\boldsymbol{\epsilon}_t^{(i)}, \boldsymbol{\epsilon}_t^{(j)}) = \mathbf{0}$ 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. Such a relaxation is necessary since we expect that even under polarization of political parties, there are few individual senators from different parties sharing similar political views, thus voting similarly on certain issues. Then the corresponding elements in the spatial weight matrix are non-zero, connecting the blocks representing different parties. Hence the blocks in the spatial weight matrix will be slightly overlapping in the end.

As presented in earlier paragraphs, for recovering the block structure of the spatial weight matrix in (2.4), if there were exogenous covariates, 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 covariates, 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. The disturbance decay assumption entails that the maximum variance of the disturbances in $\boldsymbol{\epsilon}_t$ are decaying as the sample size goes to infinity. In view of the block structure of $\mathbf{W}^*$ in (2.4), the matrix $\boldsymbol{\Pi}^* = (\mathbf{I}_N - \mathbf{W}^*)^{-1}$ also has the same block structure, say

$$\boldsymbol{\Pi}^* = \begin{pmatrix} \boldsymbol{\Pi}_1^* & & \\ & \ddots & \\ & & \boldsymbol{\Pi}_G^* \end{pmatrix},$$

with $\boldsymbol{\Pi}_j^*$ having the same size as $\mathbf{W}_j^*$ in (2.4). Hence $\mathbf{y}_t^{(j)} = \boldsymbol{\Pi}_j^* \boldsymbol{\epsilon}_t^{(j)}$, and is uncorrelated with $\boldsymbol{\epsilon}_t^{(i)}$ for $1 \leq i \neq j \leq G$ by the assumption that $\text{cov}(\boldsymbol{\epsilon}_t^{(i)}, \boldsymbol{\epsilon}_t^{(j)}) = \mathbf{0}$ for $i \neq j$. Without a block structure in $\mathbf{W}^*$, a response variable y_{ti} and a disturbance variable ϵ_{tj} cannot be uncorrelated in general. This is the reason why the disturbance decay assumption is not needed in our setting, but is needed in general in Lam and Souza (2013).

Before proposing our estimator, we write (2.3) as a linear regression model,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\xi}^* + \boldsymbol{\epsilon}, \tag{2.5}$$

where $\mathbf{y} = \text{vec}\{(\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\text{T}}\}$, $\boldsymbol{\epsilon} = \text{vec}\{(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^{\text{T}}\}$, $\boldsymbol{\xi}^* = \text{vec}(\mathbf{W}^{*\text{T}})$ and $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\text{T}}$. Here, the operator vec denotes the column by column vectorization of a matrix, while $\otimes$ denotes the kronecker product between two matrices. 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 $\boldsymbol{\xi} = \mathrm{vec}(\mathbf{W}^{\mathrm{T}})$ to obtain

$$\widetilde{\boldsymbol{\xi}} = \min_{\boldsymbol{\xi}} \frac{1}{2T}\big\|\mathbf{y} - \mathbf{Z}\boldsymbol{\xi}\big\|^2 + \gamma_T\big\|\boldsymbol{\xi}\big\|_1, \text{ subj. to } \sum_{j=1}^{N} w_{ij} < 1, \tag{2.6}$$

where $\big\|\mathbf{v}\big\|_1 = \sum_i |v_i|$ represents the L_1-norm of the vector $\mathbf{v}$ and $\big\|\mathbf{v}\big\| = (\sum_i v_i^2)^{1/2}$ represents the L_2 norm, and we denote the elements of $\mathbf{W}$ as w_{ij}. Since $\boldsymbol{\xi}$ is a vector containing all the elements of the spatial weight matrix $\mathbf{W}$, the above penalization problem can be viewed as a least square estimation for the elements of $\mathbf{W}$ (represented as the vector $\boldsymbol{\xi}$) with constraint on the magnitude of $\big\|\boldsymbol{\xi}\big\|_1$ (the absolute sum of all the elements of $\mathbf{W}$). That is, $\widetilde{\boldsymbol{\xi}}$ is the solution to the following problem:

$$\min_{\boldsymbol{\xi}} \frac{1}{2T}\big\|\mathbf{y} - \mathbf{Z}\boldsymbol{\xi}\big\|^2, \text{ subj. to } \big\|\boldsymbol{\xi}\big\|_1 \le c_T \text{ and } \sum_{j=1}^{N} w_{ij} < 1,$$

where c_T is determined by the tuning parameter γ_T. The row sum constraint in (2.6) and the above ensure the stationarity of the estimated model. The rate for the tuning parameter γ_T will be discussed after Theorem 3 in section 3.3.

Theorem 3 in section 3 shows that the solution $\widetilde{\boldsymbol{\xi}}$ for the LASSO penalization problem in (2.6) is *zero-block consistent* - that is, the zero off-diagonal blocks in $\mathbf{W}^*$ in (2.4) for model (2.3), with corresponding zero patterns in $\boldsymbol{\xi}^* = \mathrm{vec}(\mathbf{W}^{*\mathrm{T}})$, are estimated as zeros in $\widetilde{\boldsymbol{\xi}}$ with probability going to 1. The theorem also says that the diagonal blocks are estimated to be non-zero with probability equal to 1. In the context of the US senate voting data, if the Republican party and the Democrat party are forming two blocks in the spatial weight matrix $\mathbf{W}^*$ because of the political polarity in their voting patterns, the spatial weight matrix $\widetilde{\mathbf{W}}$ recovered from the LASSO estimator $\widetilde{\boldsymbol{\xi}}$ in (2.6) will be able to show such blocks with high probability.

3 Zero-Block Consistency of the LASSO Estimator

Before presenting the main results of this paper, we introduce the notation to be used for the rest of the paper, and the main technical assumptions. The definition of zero-block consistency will also be given in the subsection below.

3.1 Main assumptions and notations

(i) The spatial weight matrix $\mathbf{W}^*$ is block diagonal as in (2.4), with at least one $\mathbf{W}_i^* \neq \mathbf{0}$, and $\big\|\mathbf{W}^*\big\|_\infty \le \eta < 1$ uniformly as $T, N \to \infty$, where η is a constant. We also assume, uniformly as $T, N \to \infty$,

$$\big\|\mathbf{W}^*\big\|_1 \le \eta_c,$$

where $\|A\|_1 = \max_j \sum_i |A_{ij}|$ 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 (2.4), 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 $N, T \to \infty$,

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

for some constants $e, c > 0$.

(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 positive 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,

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

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 $\boldsymbol{\Phi}_i$.

Assumption (i) requires the absolute row sum of $\mathbf{W}^*$ to be uniformly less than 1, which is a regularity condition to ensure that the model is stationary. This row sum condition is in fact less restrictive than the commonly used row-standardization, which forces the absolute sum of each row to be equal to 1 in model (2.1). For stationarity, we need $|\rho| < 1$ in the model, so that in effect each row is forced to sum to ρ in the matrix $\rho\mathbf{W}$. See equation (3.3) in Fischer and Wang (2011) and the descriptions therein to learn more details in row-standardization. On the other hand, the row sum condition in assumption (i) merely needs the absolute sum of each row of $\mathbf{W}^*$ to be less than 1, and each of them can be unequal.

We give a hypothetical trade example to illustrate that the row sum condition is reasonable in practice. It is well known that the income of a country can depend on others, for example through trade linkages. Suppose the partners of country A experience a positive income shock. In the situation described above, it is then expected that country A, as demand for its export rises, will experience some positive spillover from partners' income shock. The row sum condition implies that the overall effect perceived from A's point of view will not be larger than the average shock accrued by its partners, weighted by the elements of W corresponding to row that represents country A. In other words, it is supposed that the income shock in the trade partners is not amplified through linkages, which is reasonable to assume to the extent that A's economy is not overly dependent on the export sector.

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. When $q = 2$, the random variables are sub-gaussian, while they are sub-exponential when $q = 1$. When $0 < q < 1$, the random variables are heavy-tailed. Hence assumption (v) is a significant relaxation to normality. 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 plausible values of c). Finally, define the set

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

For $\mathbf{W}^*$ being block diagonal as in (2.4) and an estimator $\widehat{\mathbf{W}}$, we define the estimator $\widehat{\boldsymbol{\xi}} = \text{vec}(\widehat{\mathbf{W}}^{\text{T}})$ to be *zero-block consistent* for estimating $\mathbf{W}^*$ if

$$P(\widehat{\boldsymbol{\xi}}_{\mathbf{H}} = \mathbf{0}) \to 1, \quad T, N \to \infty. \tag{3.3}$$

In this paper when we say that $T, N \to \infty$ together, we mean they approach infinity jointly rather than N being a function of T or vice versa.

3.2 Why LASSO alone is sufficient

Before presenting our main results, readers who are familiar with LASSO for the classical linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ may wonder : how can LASSO be zero-block consistent in our setting, when for a classical linear model, it is generally selection inconsistent unless the necessary condition given by Theorem 1 of Zou (2006), $|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}| \le 1$, is satisfied?

To answer this question, we first clarify the differences between selection consistency in Zou (2006) and zero-block consistency in our paper. The selection consistency in Zou (2006) concerns with the correct identification of zeros and non-zeros in the true regression parameter $\boldsymbol{\beta}^*$ of a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$. However, zero-block consistency concerns only on the correct identification of zeros which are elements of the zero blocks in the block diagonal spatial weight matrix $\mathbf{W}^*$ in (2.4). For the elements in the diagonal blocks $\mathbf{W}_i^*, i = 1, \ldots, G$ in (2.4), we are not concerned with correct identification of zeros and non-zeros. With this in mind, at the very most we can only draw parallels between the two.

One important parallel is that the necessary and sufficient condition for zero-block consistency in our setting, depicted in equation (7.5) in section 7 (see the proof of Theorem 3 therein to see how we arrive at such necessary and sufficient condition), resembles the necessary condition $|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}| \leq 1$ in Theorem 1 of Zou (2006). Using the notation in equation (7.5) in our paper, the matrix $\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D$ depicts the covariance matrix between the columns of the design matrix $\mathbf{Z}$ of model (2.5) corresponding to the set H defined in (3.1), and the columns of $\mathbf{Z}$ corresponding to the set D defined at the beginning of the proof of Theorem 3. This matrix is parallel to the matrix $\mathbf{C}_{21}$ of Zou (2006). Similarly, the matrix $\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D$ is parallel to the matrix $\mathbf{C}_{11}$. For the necessary and sufficient condition (7.5) to be satisfied, a necessary condition can be derived from (7.5) to be

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

which completely resembles the condition $|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}| \leq 1$ in Theorem 1 of Zou (2006), except that $\mathbf{g}_D$ is a vector containing $1, -1$ and some values with magnitude smaller than 1, whereas $\mathbf{s}$ in Zou (2006) contains only 1 or -1.

Under model (2.5), we can use equations (7.8) and (7.12) in section 7 to show that on the set A_ϵ defined in (3.2),

$$\Big|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big)^{-1}\mathbf{g}_D\Big| \leq \Big\|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D\Big\|_\infty \cdot \Big\|\Big(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D\Big)^{-1}\Big\|_\infty \cdot \|\mathbf{g}_D\|_\infty = O(\lambda_T n^{3/2}) = o(1),$$

so that the necessary condition above is satisfied on the set A_ϵ when T, N are large enough, which has $P(A_\epsilon) \to 1$ by Corollary 2. Both equations (7.8) and (7.12) are proved on the basis of the form of the model (2.3) and various assumptions in section 3.1, including the row sum and column sum assumption (i) for the spatial weight matrix $\mathbf{W}^*$ and the causal assumption for the process $\{\boldsymbol{\epsilon}_t\}$ in assumption (iv).

In brief, the special form of our model (2.3) so that $\mathbf{y}_t = \mathbf{\Pi}^*\boldsymbol{\epsilon}_t$, and the assumptions for the spatial weight matrix and the disturbance process, are all reasons for the LASSO estimator in (2.6) to be zero-block consistent.

3.3 Main results

We first present a theorem and its corollary concerning the probability lower bound of the set defined in (3.2), which is the lower bound for the tail probability of the mean of the product process $\{\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})\}$. We show in Theorem 3, the main result of this paper, that this is also the probability lower bound for the LASSO solution $\widetilde{\boldsymbol{\xi}}$ in (2.6) being zero-block consistent. Implications and explanations of our main result will also be discussed after presenting the theorem.

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. *With the same constants C_1,C_2 and C_3, and the same conditions 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

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

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 (2.6) 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 (2.6) 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 $\widetilde{\mathbf{W}}$ for the spatial weight matrix exists and is given by the LASSO estimator $\widetilde{\boldsymbol{\xi}}$ using the relation $\widetilde{\boldsymbol{\xi}} = \text{vec}(\widetilde{\mathbf{W}}^{\text{T}})$, 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. In the context of the US senate voting data analysis in section 5.3, it means that with the number of senators (the dimension N) and the number of voting instances (the number of time points T) large enough, if the voting patterns indeed align with political parties so that the underlying spatial weight matrix is block diagonal as in (2.4) with each block representing a political party, then the probability that the LASSO estimator for the spatial weight matrix has the same block diagonal structure is large. Also, the tuning parameter $\gamma_T \to 0$ means that in practice it has to be small, so that the penalization towards the elements of the spatial weight matrix, through the term $\|\boldsymbol{\xi}\|_1$ in (2.6), cannot be too large. If this is too large, then the whole spatial weight matrix can be estimated as $\mathbf{0}$, which is definitely zero-block consistent, albeit completely useless for our purpose.

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. See section 5 for the introduction of a BIC criterion for choosing γ_T.

4 Relaxation for Overlapping Blocks

The spatial weight matrix in (2.4) 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. As described in the introduction and section 2, this can happen when there are small number of "hybrid" individuals who are interacting with more than one group.

Formally, 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 majority of 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. Note that if the groups are overlapping, then necessarily $\bigcup_{i=1}^G I_i \subset \{1, \ldots, N\}$. 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.

To facilitate understanding of the notation above, we introduce a hypothetical example. For our US senator voting data, suppose there are three major blocks, representing the Republicans, the Democrats and the Independent Senators respectively. However, over a certain period of time, there is one Republican who not only cooperates with some other fellow Republicans, but also with another Democrat and another Independent Senator. Then over this period of time, the voting pattern of this Republican can depend not only on some other fellow Republicans, but also on the Democrat and the Independent Senator with whom he or she is cooperating. Using the notation introduced above, then $G = 3$, but these three senators who are cooperating across parties will not be registered into the sets I_1, I_2 or I_3, since the corresponding elements in the spatial weight matrix $\mathbf{W}^*$ will be non-zero as their voting patterns can depend on each other. Then $I_1 \cup I_2 \cup I_3 \subset \{1, \ldots, N\}$.

Define the set

$$H' = \{j : \xi_j^* = 0 \text{ and corr. to one of the } G(G-1) \text{ zero blocks in } W^*\}. \tag{4.4}$$

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 (2.6) 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 (2.6). A unique solution is guaranteed since the minimization problem in (2.6) is convex. The LARS is very fast since the order of complexity of the algorithm is the same as that for ordinary least squares.

For choosing a suitable γ_T, following Wang et al. (2009), we propose a BIC criterion as below:

$$\mathrm{BIC}(\gamma_T) = \sum_{i=1}^{N} \log\left(T^{-1}\left\|\widetilde{\mathbf{y}}_i - (\mathbf{Z}\widetilde{\boldsymbol{\xi}}_{\gamma_T})_i\right\|^2\right) + |S_{\gamma_T}|\frac{\log(T)}{T}\log(\log(N-1)), \tag{5.1}$$

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 (2.6) with tuning parameter being γ_T. Also, $(\mathbf{Z}\widetilde{\boldsymbol{\xi}}_{\gamma_T})_i$ is the vector with length T which is the portion of the vector $\mathbf{Z}\widetilde{\boldsymbol{\xi}}_{\gamma_T}$ (see (2.5)) 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 in fact the sum of individual BIC criteria for the estimator of the ith row of the spatial weight matrix, with response variable $\widetilde{\mathbf{y}}_i$. We denote γ_{BIC} the tuning parameter that minimizes the BIC criterion in (5.1). This γ_{BIC} will then be used in (2.6) to find the LASSO solution $\widetilde{\boldsymbol{\xi}}$.

5.1 Simulation results

In this paper, we focus on block detection, and there are no theoretical supports for accurate estimation of the 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 between 2 and 4 diagonal blocks as in (2.4), with uniform probability on their start and end points. Models with blocks of fewer than 5 individuals or with within-block sparsity larger than 90% are rejected. The latter condition restricts blocks from being excessively large.

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 $1.1 \times \max\left(1, \sum_{j=1}^{N} w_{ij}^*\right)$. In Table 3, shown in the Appendix, we relax this condition to move close to the non-stationary case. 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 γ_T chosen by minimizing the BIC criteria (5.1) for different values of N and T. The number of replications is 200. 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, and in general gets better as N increases. While within-block accuracy is not guaranteed, the within-block specificity and sensitivity are quite good, even when T is not large. The overall sparsity level is close to κ in most cases. One notable feature is that with N fixed, as T gets larger, the overall sparsity level decreases. This is because as T gets larger, the tuning parameter γ_T selected by the BIC criterion gets smaller, as is evident from Table 1. It means that as T gets larger, BIC does not allow as much penalization to the model. This is because there are many non-zero within-block elements in the main diagonal blocks which can only be detected when T is large enough and γ_T small enough. As T gets larger, it is more beneficial to have a smaller γ_T so that the non-zero parameters are estimated as non-zeros within the diagonal blocks. With a smaller γ_T, the within-block sensitivity certainly increases while the within-block specificity certainly decreases, and hence the overall sparsity decreases. These are exactly what one can observe from Table 1. The choice of tuning parameter when there are many explanatory variables that are highly endogenous like in our case is definitely a future direction for research.

Table 2 introduces slightly overlapping blocks. For any two blocks, their overlapping size is chosen randomly to be $\max(q_1, q_2)$, where q_1 is 5% of the minimum size of the blocks and q_2 is a random integer between 1 and 4. This setting contains the case where $T = 200$ and $N = 75$ with 2 main blocks that are slightly overlapping, which is similar to the situation in the real data analysis in section 5.3, where there are $T = 251$ voting instances and $N = 98$ senators, and two main blocks that are slightly overlapping. Again, the tuning parameter γ_T is chosen such that the BIC criterion in (5.1) is minimized. The results are shown in Table 2. The simulation results show similar pattern as in Table 1: across-block specificity, although shows a slight deterioration, is still around 97% to 99% in most cases. The tuning parameter γ_T selected by the BIC criterion is again decreasing with T, and hence the within-block specificity and the overall sparsity decreases as T increases, but the within-block sensitivity increases, like those in Table 1.

5.2 Simulation results for nonstationary models

In order to see how the stationarity of model (2.3) is important to the practical performance of our method, we show simulation results with adjusted normalization of elements in $\mathbf{W}^*$ in order to move closer to nonstationarity, with results shown in Table 3. We also added results for a nonstationary model in Table 4. They are substantially worse than those in 5.1, which are associated with stationary models.

Table 1: Simulations with non-overlapping blocks.

		$\kappa = 0.90$			$\kappa = 0.95$		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$N = 25$	Within-Block Specificity	80.64%	81.66%	80.20%	96.99%	90.36%	84.31%
		(3.310)	(2.814)	(2.460)	(3.992)	(4.645)	(2.684)
	Within-Block Sensitivity	70.56%	79.44%	89.17%	18.33%	52.22%	87.78%
		(5.832)	(5.566)	(4.578)	(18.829)	(20.268)	(7.566)
	Across-Block Specificity	97.01%	97.60%	97.67%	99.42%	98.70%	98.13%
		(2.035)	(1.857)	(1.819)	(1.139)	(1.738)	(0.718)
	L_1	0.0237	0.0205	0.0215	0.0136	0.0132	0.0124
		(0.002)	(0.001)	(0.003)	(0.001)	(0.001)	(0.000)
	L_2	0.1206	0.0826	0.0769	0.0842	0.0667	0.0511
		(0.014)	(0.006)	(0.011)	(0.006)	(0.005)	(0.005)
	Sparsity	85.94%	83.94%	80.26%	97.75%	93.85%	90.06%
		(2.183)	(2.297)	(3.151)	(2.815)	(3.184)	(1.447)
	γ_{BIC}	0.3500	0.2401	0.1588	0.4979	0.2687	0.1529
		(0.051)	(0.053)	(0.023)	(0.158)	(0.062)	(0.014)
$N = 50$	Within-Block Specificity	77.35%	74.57%	78.75%	89.15%	89.38%	80.27%
		(1.007)	(1.781)	(1.250)	(2.534)	(1.389)	(1.239)
	Within-Block Sensitivity	55.71%	66.02%	75.00%	45.80%	61.86%	87.47%
		(2.846)	(2.374)	(2.796)	(7.885)	(5.029)	(3.129)
	Across-Block Specificity	98.56%	98.94%	98.78%	99.47%	99.42%	98.68%
		(0.501)	(0.347)	(0.361)	(0.282)	(0.325)	(0.408)
	L_1	0.0188	0.0151	0.0139	0.0113	0.0106	0.0112
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	L_2	0.1508	0.1031	0.0782	0.1124	0.0937	0.0875
		(0.007)	(0.004)	(0.002)	(0.005)	(0.004)	(0.004)
	Sparsity	87.46%	87.40%	84.48%	95.03%	93.37%	90.35%
		(0.620)	(0.619)	(0.694)	(1.090)	(0.724)	(0.651)
	γ_{BIC}	0.4807	0.3670	0.1913	0.5048	0.3131	0.1884
		(0.037)	(0.050)	(0.016)	(0.078)	(0.025)	(0.014)
$N = 75$	Within-Block Specificity	82.20%	81.20%	77.47%	89.33%	87.13%	82.46%
		(1.281)	(0.573)	(0.690)	(1.192)	(0.627)	(0.869)
	Within-Block Sensitivity	40.96%	57.24%	68.51%	40.65%	56.74%	81.80%
		(2.620)	(2.863)	(1.274)	(4.172)	(3.329)	(2.437)
	Across-Block Specificity	99.36%	99.45%	99.67%	99.51%	99.63%	99.09%
		(0.324)	(0.316)	(0.179)	(0.168)	(0.248)	(0.349)
	L_1	0.0145	0.0129	0.0116	0.0102	0.0087	0.0091
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	L_2	0.1467	0.1123	0.0867	0.1352	0.0974	0.0919
		(0.007)	(0.005)	(0.003)	(0.005)	(0.004)	(0.004)
	Sparsity	90.75%	88.35%	86.36%	94.71%	93.59%	90.96%
		(0.606)	(0.352)	(0.305)	(0.552)	(0.399)	(0.431)
	γ_{BIC}	0.5591	0.4145	0.2978	0.5690	0.3479	0.2091
		(0.070)	(0.033)	(0.027)	(0.072)	(0.033)	(0.016)

Notes: Standard errors in parenthesis.

Table 2: Simulations with overlapping blocks.

		$\kappa = 0.90$			$\kappa = 0.95$		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$N = 25$	Within-Block Specificity	87.78% (3.983)	74.42% (2.618)	77.56% (2.054)	96.99% (3.448)	89.40% (4.460)	88.46% (1.742)
	Within-Block Sensitivity	50.17% (7.457)	77.04% (4.362)	93.29% (3.142)	18.18% (17.008)	57.14% (19.323)	93.12% (7.471)
	Across-Block Specificity	97.24% (1.476)	94.92% (1.908)	91.32% (2.425)	99.42% (0.848)	98.56% (1.505)	94.86% (1.686)
	L_1	0.0211 (0.001)	0.0253 (0.001)	0.0218 (0.001)	0.0136 (0.000)	0.0131 (0.001)	0.0132 (0.001)
	L_2	0.1032 (0.006)	0.1071 (0.006)	0.0810 (0.006)	0.0846 (0.006)	0.0676 (0.007)	0.0528 (0.004)
	Sparsity	90.47% (2.422)	81.21% (1.594)	79.29% (1.897)	98.03% (2.229)	93.40% (3.010)	88.97% (1.611)
	λ_{BIC}	0.3603 (0.057)	0.2116 (0.030)	0.1411 (0.014)	0.5289 (0.153)	0.2496 (0.047)	0.1588 (0.018)
$N = 50$	Within-Block Specificity	87.79% (0.892)	82.91% (1.494)	77.02% (0.901)	90.51% (2.265)	90.18% (2.380)	87.98% (0.661)
	Within-Block Sensitivity	44.26% (4.556)	61.22% (2.819)	77.42% (1.544)	47.17% (3.450)	53.66% (7.396)	88.45% (2.298)
	Across-Block Specificity	97.61% (0.565)	98.51% (0.818)	97.20% (0.677)	98.88% (0.421)	99.07% (0.318)	98.42% (0.517)
	L_1	0.0199 (0.001)	0.0169 (0.001)	0.0166 (0.000)	0.0110 (0.000)	0.0113 (0.000)	0.0110 (0.000)
	L_2	0.1502 (0.008)	0.1064 (0.004)	0.1006 (0.004)	0.1072 (0.004)	0.1023 (0.003)	0.0834 (0.002)
	Sparsity	87.36% (0.986)	84.70% (1.071)	82.19% (0.522)	94.97% (0.796)	93.64% (1.163)	90.13% (0.323)
	λ_{BIC}	0.4532 (0.072)	0.2909 (0.044)	0.1854 (0.018)	0.4842 (0.054)	0.3131 (0.044)	0.1825 (0.000)
$N = 75$	Within-Block Specificity	80.78% (1.131)	78.59% (0.924)	70.62% (1.067)	92.48% (1.440)	84.60% (0.859)	84.67% (0.897)
	Within-Block Sensitivity	41.47% (1.968)	52.42% (2.573)	71.52% (1.759)	33.05% (5.628)	62.47% (3.444)	78.24% (2.481)
	Across-Block Specificity	98.62% (0.478)	98.70% (0.255)	98.45% (0.291)	99.61% (0.198)	98.83% (0.395)	99.03% (0.361)
	L_1	0.0141 (0.000)	0.0127 (0.000)	0.0112 (0.000)	0.0105 (0.000)	0.0095 (0.000)	0.0097 (0.000)
	L_2	0.1369 (0.005)	0.1140 (0.004)	0.0859 (0.003)	0.1433 (0.005)	0.1118 (0.004)	0.0986 (0.003)
	Sparsity	90.65% (0.581)	89.31% (0.501)	87.01% (0.463)	95.71% (0.837)	92.98% (0.506)	90.60% (0.390)
	λ_{BIC}	0.4904 (0.063)	0.3828 (0.025)	0.2564 (0.024)	0.5821 (0.059)	0.3511 (0.038)	0.2150 (0.010)

Notes: as in Table 1.

In more details, for the first case, we adjust the normalization of elements w_{ij}^* of $\mathbf{W}^*$, which are now divided by $1.05 \times \max\left(0.5, \sum_{j=1}^{N} w_{ij}^*\right)$ (compared to $1.1 \times \max\left(1, \sum_{j=1}^{N} w_{ij}^*\right)$ in baseline simulations). In this way, we ensure that row sum of $\mathbf{W}^*$ is higher than 0.90 in over 60% of the cases for $N = 25$, 70% for $N = 50$ and 95% for $N = 75$. In every case, by design the row-sum is smaller than 1. Apart from this, the simulation setup remains unchanged. As can be seen, in comparison to Table 1, the performance is slightly worse. However, across-block specificity is higher than 95% in all cases. Within-block specificity and sensitivity remains satisfactory and in line with baseline simulations.

Next, we implement a nonstationary case by normalizing the elements w_{ij} by $0.75 \times \max\left(0.01, \sum_{j=1}^{N} w_{ij}^*\right)$. Deterioration in performance can be clearly seen through the worsening of all measures. In particular, the L_1 criterion deteriorated by about 40-50 times and L_2 one around 90-100 times of the values in Table 3.

Table 3: Simulations close to nonstationarity.

		$\kappa = 0.90$			$\kappa = 0.95$		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$N = 25$	Within-Block Specificity	75.51% (2.815)	64.58% (2.996)	73.34% (2.280)	78.66% (2.792)	79.71% (1.760)	83.91% (2.026)
	Within-Block Sensitivity	75.42% (5.327)	81.25% (4.058)	81.67% (4.364)	84.17% (6.107)	88.75% (2.480)	91.25% (3.959)
	Across-Block Specificity	96.36% (1.492)	97.40% (1.374)	99.57% (0.418)	96.96% (0.873)	98.16% (0.741)	98.82% (1.204)
	L_1	0.0269 (0.001)	0.0289 (0.001)	0.0249 (0.001)	0.0237 (0.002)	0.0211 (0.001)	0.0188 (0.001)
	L_2	0.1546 (0.011)	0.1574 (0.011)	0.1319 (0.005)	0.1594 (0.012)	0.1357 (0.006)	0.1151 (0.005)
	Sparsity	84.04% (1.720)	82.31% (1.401)	84.54% (0.999)	87.17% (1.300)	88.83% (0.947)	89.65% (1.390)
	λ_{BIC}	0.3827 (0.056)	0.3004 (0.060)	0.4308 (0.054)	0.2949 (0.031)	0.2718 (0.020)	0.2179 (0.038)
$N = 50$	Within-Block Specificity	73.72% (1.785)	77.22% (1.424)	71.80% (0.995)	86.18% (1.613)	71.69% (1.672)	83.09% (0.996)
	Within-Block Sensitivity	66.63% (1.742)	69.03% (2.404)	84.13% (0.937)	67.68% (3.782)	81.20% (2.797)	88.82% (4.117)
	Across-Block Specificity	98.12% (0.474)	98.35% (0.635)	99.17% (0.118)	97.95% (0.459)	98.64% (0.376)	99.35% (0.398)
	L_1	0.0197 (0.001)	0.0180 (0.001)	0.0161 (0.000)	0.0155 (0.001)	0.0153 (0.000)	0.0133 (0.000)
	L_2	0.1743 (0.008)	0.1396 (0.005)	0.1144 (0.003)	0.1806 (0.007)	0.1725 (0.006)	0.1299 (0.004)
	Sparsity	86.28% (0.380)	84.65% (0.753)	84.46% (0.271)	90.75% (0.750)	90.40% (0.508)	89.94% (0.626)
	λ_{BIC}	0.6407 (0.079)	0.3717 (0.057)	0.3288 (0.023)	0.4343 (0.044)	0.3860 (0.045)	0.2579 (0.052)
$N = 75$	Within-Block Specificity	84.50% (0.569)	78.48% (1.075)	70.77% (1.520)	85.32% (0.972)	77.39% (0.978)	85.06% (0.452)
	Within-Block Sensitivity	39.01% (1.115)	57.57% (1.559)	73.85% (1.417)	58.27% (2.507)	74.91% (1.356)	83.54% (2.005)
	Across-Block Specificity	99.06% (0.337)	99.15% (0.263)	99.43% (0.284)	99.16% (0.417)	98.69% (0.328)	99.12% (0.322)
	L_1	0.0164 (0.000)	0.0132 (0.000)	0.0112 (0.000)	0.0135 (0.000)	0.0108 (0.000)	0.0105 (0.000)
	L_2	0.1745 (0.004)	0.1230 (0.002)	0.0967 (0.003)	0.1967 (0.008)	0.1402 (0.005)	0.1274 (0.005)
	Sparsity	88.64% (0.288)	87.61% (0.443)	87.19% (0.475)	91.34% (0.641)	91.36% (0.332)	90.24% (0.335)
	λ_{BIC}	0.5804 (0.084)	0.4050 (0.079)	0.3199 (0.058)	0.5706 (0.094)	0.3717 (0.040)	0.2357 (0.019)

Notes: as in Table 1.

Table 4: Simulations for the nonstationary case.

		$\kappa = 0.90$			$\kappa = 0.95$		
		$T = 50$	$T = 100$	$T = 200$	$T = 50$	$T = 100$	$T = 200$
$N = 25$	Within-Block Specificity	85.32% (0.424)	94.26% (0.479)	88.49% (0.424)	86.88% (3.377)	91.02% (1.752)	91.57% (0.632)
	Within-Block Sensitivity	1.04% (1.240)	4.17% (1.543)	6.67% (0.000)	12.92% (3.753)	19.58% (1.179)	6.67% (0.000)
	Across-Block Specificity	91.85% (0.427)	91.96% (0.108)	91.97% (0.085)	91.76% (3.551)	92.50% (0.403)	92.93% (0.127)
	L_1	0.8141 (0.001)	0.7508 (0.041)	0.7207 (0.000)	0.4677 (0.029)	0.4994 (0.016)	0.5441 (0.001)
	L_2	193.1319 (0.125)	197.9038 (11.178)	163.4174 (0.004)	119.2568 (8.197)	182.1524 (14.187)	186.6742 (0.017)
	Sparsity	96.71% (0.305)	97.40% (0.235)	96.90% (0.124)	92.29% (3.229)	96.25% (0.321)	96.73% (0.251)
	λ_{BIC}	0.6665 (0.000)	0.6143 (0.000)	0.5727 (0.000)	0.3414 (0.248)	0.6238 (0.018)	0.5727 (0.000)
$N = 50$	Within-Block Specificity	91.25% (2.287)	97.35% (0.485)	91.20% (0.509)	94.42% (0.300)	86.49% (0.465)	99.25% (0.072)
	Within-Block Sensitivity	4.54% (1.724)	1.38% (0.304)	9.59% (0.654)	3.96% (0.287)	15.35% (0.678)	2.44% (0.000)
	Across-Block Specificity	92.97% (0.059)	92.99% (0.022)	92.93% (0.051)	92.78% (0.103)	92.01% (0.212)	92.57% (0.000)
	L_1	0.4106 (0.000)	0.4016 (0.000)	0.4021 (0.001)	0.3697 (0.002)	0.4951 (0.011)	0.3512 (0.000)
	L_2	96.3161 (7.643)	109.9296 (0.031)	139.6243 (1.246)	180.1242 (1.095)	743.8054 (63.704)	190.3584 (0.000)
	Sparsity	98.71% (0.213)	99.20% (0.129)	96.93% (0.092)	98.09% (0.078)	95.31% (0.212)	99.66% (0.021)
	λ_{BIC}	0.6665 (0.000)	0.6143 (0.000)	0.5727 (0.000)	0.6665 (0.000)	0.6286 (0.020)	0.5727 (0.000)
$N = 75$	Within-Block Specificity	93.02% (0.610)	95.53% (0.209)	94.70% (0.084)	94.75% (0.241)	95.15% (0.175)	91.49% (0.179)
	Within-Block Sensitivity	4.68% (0.319)	5.23% (0.409)	3.76% (0.311)	0.40% (0.127)	3.15% (0.167)	4.68% (0.471)
	Across-Block Specificity	92.67% (0.012)	92.80% (0.052)	92.11% (0.067)	92.83% (0.097)	91.97% (0.038)	92.89% (0.180)
	L_1	0.2733 (0.000)	0.2775 (0.001)	0.2414 (0.000)	0.2628 (0.000)	0.2612 (0.000)	0.7549 (0.087)
	L_2	65.1182 (0.050)	478.4065 (14.791)	51.7448 (0.018)	148.1981 (0.235)	146.0697 (0.147)	14041.1627 (4394.414)
	Sparsity	98.82% (0.065)	97.96% (0.082)	96.35% (0.059)	98.45% (0.080)	98.46% (0.069)	96.90% (0.131)
	λ_{BIC}	0.6345 (0.000)	0.6143 (0.000)	0.5727 (0.000)	0.6394 (0.014)	0.6143 (0.000)	0.5949 (0.018)

Notes: as in Table 1.

5.3 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 2.3 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 5.

Estimation is conducted in absolute disregard of party affiliation, and the tuning parameter γ_T is

[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.

chosen such that minimizes BIC criterion in (5.1). The outcome for year 2012, which involves $T = 251$ voting instances and $N = 98$ senators, 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 3% 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. 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.

Table 5: 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

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

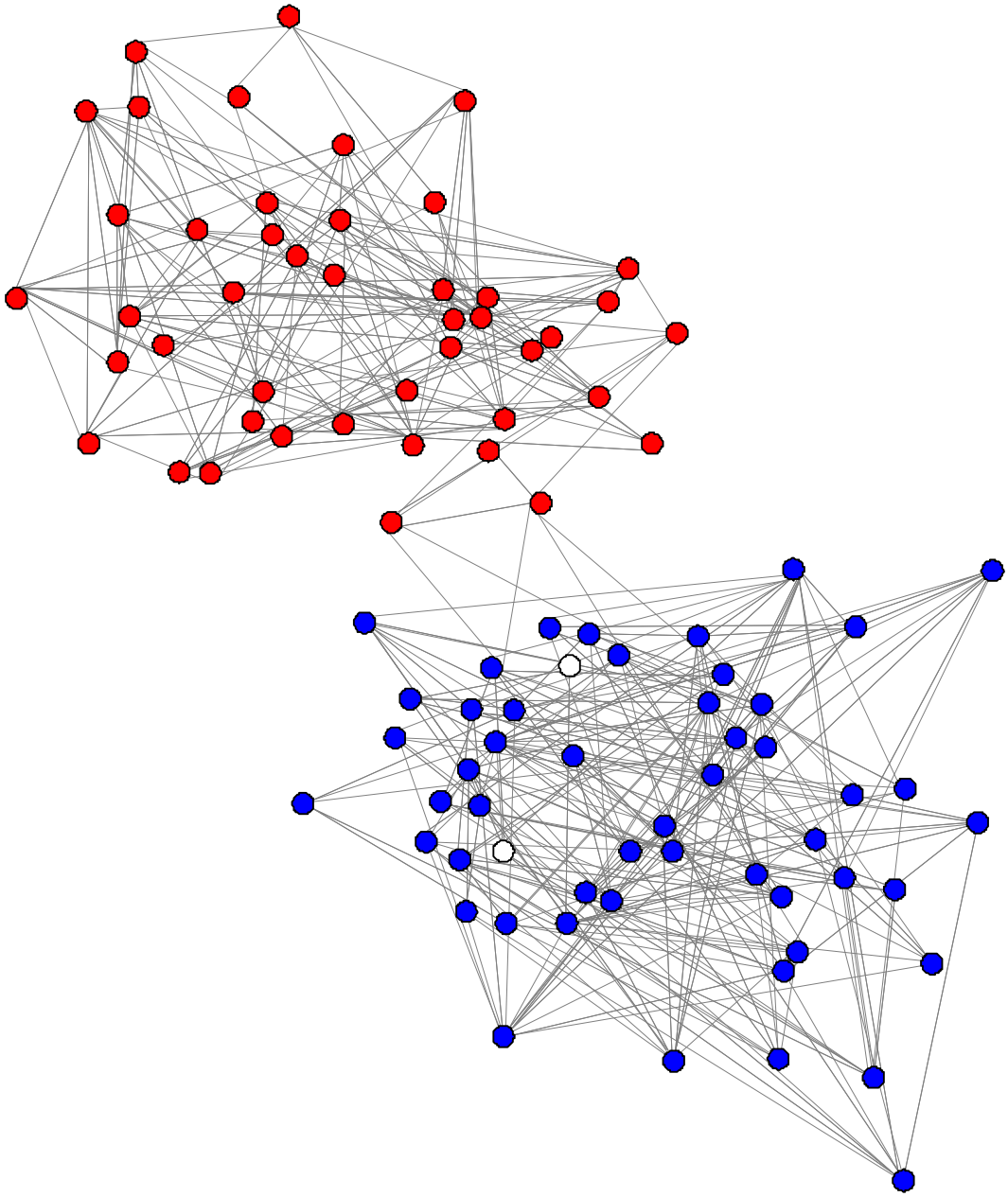
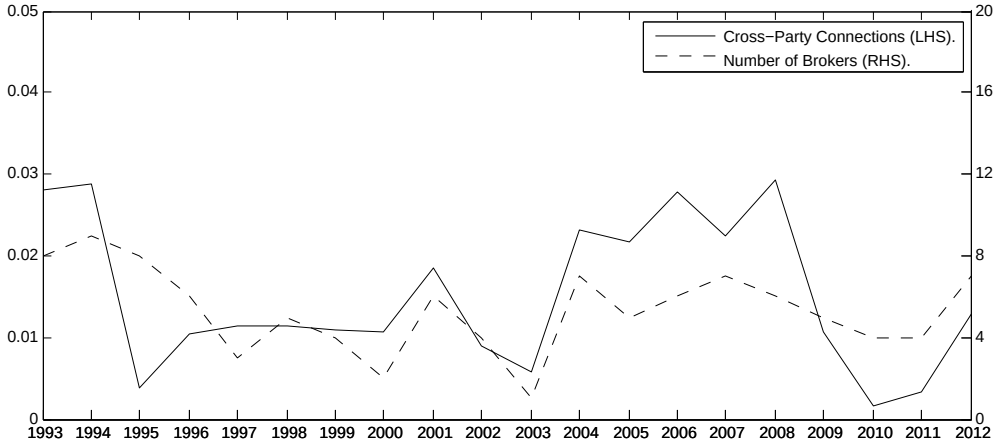


Figure 2: Cross-party collaboration.



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 slightly overlapping blocks. The US senate voting data example demonstrates clearly such a case.

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

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

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

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),

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

so that $\max_{1\le j\le N} \big\|\epsilon_{tj}\big\|_{2w} \le \mu < \infty$ 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),

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

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),

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

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

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

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 (2.6) if and only if there exists a subgradient

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

such that, differentiating the expression to be minimized in (2.6) with respect to $\boldsymbol{\xi}_J$,

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

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 (2.4). Then, consider on the set A_ϵ,

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

where we used the reduced form $\mathbf{y}_t = \boldsymbol{\Pi}^*\boldsymbol{\epsilon}_t = (\mathbf{I}_N - \mathbf{W}^*)^{-1}\boldsymbol{\epsilon}_t$ of model (2.3) 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

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

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

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

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

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

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

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

To bound (7.9), we have

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

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

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 $i \ge 0$, 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 $U = \bigcap_{i\ge 0}U_i$, the above limits hold for all integers $i \ge 0$. Hence on U, for large enough T, N, we have

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

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

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

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,

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

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,

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

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

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.

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.

Brueckner, J. (2003). Strategic interaction among local governments: An overview of empirical studies. *International Regional Science Review 26*(2), 175–188.

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

Elhorst, J. (2010). Spatial panel data models. In M. M. Fischer and A. Getis (Eds.), *Handbook of Applied Spatial Analysis*, pp. 377–407. Springer Berlin Heidelberg.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Berlin: Springer-Verlang.

Fischer, M. M. and J. Wang (2011, September). *Spatial Data Analysis: Models, Methods and Techniques (SpringerBriefs in Regional Science)* (1st Edition. ed.). Springer.

Fowler, J. (2006). Connecting the congress: A study of cosponsorship networks. *Political Analysis 71*(1), 456–487.

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

LeSage, J. and R. K. Pace (2008). *Introduction to Spatial Econometrics*. Chapman and Hall.

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.

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

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.

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