London School of Economics and Political Science

Essays on Identification and Estimation of Networks

Pedro Carvalho Loureiro de Souza

Thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy, London, January 2015.

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 50,131 words.

Statement of conjoint work

I can confirm that Chapter 2 and Chapter 3 was jointly co-authored with Professor Clifford Lam and I contributed with 50% of this work.

Statement of use of third party for editorial help

I can confirm that Chapter 1 of this thesis was partially copy-edited for conventions of language, spelling and grammar by American Journal Experts.

Abstract

This thesis consists of three chapters that explore the estimation and identification of networks from observable outcomes and covariates only. This problem is equivalent to estimating the spatial neighbouring matrix from a spatial econometric model. Under three settings, I show how the networks can be recovered entirely from observable non-network data.

In the first chapter, networks are treated as a source of unobserved heterogeneity and dealt with data collected from observing many groups in one period of time. The proposed method estimates the probability that pairs of individuals form connections, which may depend on exogenous factors such as common gender. I derive a maximum likelihood estimator for network effects that is not conditional on network observation, accomplished with recourse to a spatial econometric model with unobserved and stochastic networks. I apply the model to estimate network effects in the context of a program evaluation.

The second chapter assumes the observation of one group over many periods of time and estimates the networks as a collection of pairwise links. We estimate the spatial neighbouring matrix with recourse to the Adaptive Lasso. Non-asymptotic Oracle inequalities, together with the asymptotic sign consistency of the estimators, are presented and proved.

The third chapter shows how the procedure developed in the preceding paper can be used to classify individuals into groups based on similarity of observed behavior. We propose a Lasso estimator that captures the block structure of the spatial neighboring matrix. The main results show that off-diagonal block elements are estimated as zeros with high probability. We correctly identified US Senate's blocks based on party affiliation using only voting data.

Empirical research on social and economic networks has been constrained by the limited availability of data regarding such networks. This collection of papers may therefore provide an useful tool for applied research.

Acknowledgements

I am truly grateful to my supervisors Javier Hidalgo, Steve Pischke and Robin Burgess for their care, support, attention and dedication thoughout my studies at the London School of Economics. I feel I have been very fortunate in having them in my team of supervisors. Their encouragement was pivotal during this journey. A special thanks among faculty go to Oriana Bandiera, Taisuke Otsu, Marcia Schafgans, Tatiana Komarova, Áureo de Paula, Clifford Lam and those who have given incredibly useful comments during the Labour, Development and Econometric Theory seminars.

I am also very grateful to Marcelo de Paiva Abreu, Marcelo Medeiros, Eduardo Loyo, Pedro and Maria Lucia Carvalho for the trust they deposited on me. I would also like to thank for the generous funding of the Royal Economic Society and of the Economics Department.

I also thank my brilliant peers with whom I shared insights, thoughts and a great deal of fun throughout these amazing years in London. To Erika Desseranno, Laura Derksen, Samuel Marden, Jon de Quidt, João Pessoa, Oliver Pardo-y-Reinoso, Thiemo Fetzer, Pedro Franco, Francisco Costa, Oriol Carreras, Fabio Miessi, Daniel Silva, Jason Garred, Markus Riegler, Enrico Mallucci, Munir Squires, Claire Balboni, Dimitri Szerman, many many thanks. I will miss you.

Lastly, I would like to express my gratitude to family and friends who followed me and were an enormous source of strength. I thank you for all your love and care. To Pri for her constant presence, affection and for giving me so much happiness. To mom and dad and their warm and attentive love.

I dedicate this thesis to my grandparents, José, Maria do Carmo, Rubens and Isolda.

Contents

List of Tables

List of Figures

Chapter 1

Estimating Network Effects without Network Data

Abstract. Empirical research on social and economic networks has been constrained by the limited availability of data regarding such networks. This paper develops a method that does not rely on network data to estimate network effects. The proposed method also estimates the probability that pairs of individuals form connections, which may depend on exogenous factors such as common gender. The method may incorporate imperfect network data, such as with self-reported data, with the dual purpose of refining the estimates and testing whether the reported connections positively affect the probability that a link is formed. To achieve those goals, I derive a maximum likelihood estimator for network effects that is not conditional on network observation. Networks are treated as a source of unobserved heterogeneity and dealt with data collected from observing many groups. This is accomplished with recourse to a spatial econometric model with unobserved and stochastic networks. I then apply the model to estimate network effects in the context of a program evaluation. I demonstrate theoretically and empirically that including network effects has important implications for policy assessments.[1]

Keywords: social networks, spillovers, spatial econometrics.

JEL Codes: C21, C49, O12, D85.

[1] I am very grateful to Javier Hidalgo, Steve Pischke, Robin Burgess and Oriana Bandiera for extensive support and feedback on this project. I also thank Clare Balboni, Francisco Costa, Jan-Emmanuel De Neve, Samuel Marden, Priscilla Negreiros, Taisuke Otsu, João Pessoa, Markus Riegler, Munir Squires, Dimitri Szerman, and participants at Econometric Society European Winter Meeting 2014, 61st North American Meeting of the Regional Science Association International 2014, LSE Econometrics, Development, Labour and Summer seminars for extremely helpful comments.

1.1 Introduction

Personal interconnectedness is an important and pervasive feature of human life. Social and economic networks enhance learning in classrooms (Angrist and Lang, 2004; Ammermuller and Pischke, 2009), influence decisions regarding technology adoption (Foster and Rosenzweig, 1995; Conley and Udry, 2010) and serve as mechanisms for informal contractual enforcement (Ambrus et al., 2014). In recent years, the many ways in which social networks affect choices and behavior have been the subject of extensive research (Jackson, 2010). However, incorporating these mechanisms in applied research remains challenging because of the limited availability of network data. Even when networks are able to be observed, these observations are often imperfect, such as when data are self-reported or subject to measurement errors.

This paper develops a method for estimating network effects when network data are either unobserved or imperfectly observed. The method does not rely on network data and derives network effects using only individuals' dependent and explanatory variables data. I specifically propose an estimator that accomplishes three objectives. First, I estimate network spillovers – the difference between expected outcomes when networks are and are not relevant – without network data.[2] Spillovers also capture the extent to which social networks amplify the effect of explanatory variables on outcomes (Miguel and Kremer, 2004). Second, I illuminate structural mechanisms that give rise to network spillovers. I separately identify and estimate Manski's (1993) endogenous effects (the dependence of one's own choices on the choices of others) from exogenous effects (the dependence of one's own choices on the exogenous variables of others), controlling for correlated effects (the similarity of peers in terms of unobservable characteristics).[3] The method also estimates and predicts the probability that pairs of individuals form a connection, which is allowed to depend on exogenous factors such as common gender. Third, I incorporate imperfect network data, such as self-reported network data, with the dual purpose of refining the estimates and providing a test for whether reported connections positively affect the probability that a connection is formed. Rejection of the null demonstrates self-reported network data validity.

To achieve these goals, I propose a spatial econometric model with unobserved and stochastic networks that is coupled with a model for random network formation. I derive a likelihood for the model which is not conditional on network. This likelihood is equivalent to integrating the likelihood conditional on observing the true network with respect to the probability density function

[2] This is also important because OLS estimates are often inconsistent for individual reaction parameters when networks are irrelevant if network spillovers are not included in the regression, and the size of inconsistency depends on the unobserved network.

[3] Endogenous effects are the autoregressive component of a spatial model. Exogenous effects is exogenous component of a spatial model. Correlated effects are captured by fixed effects at the individual level. These are precisely defined with recourse to the model in Section 1.2. The reflection problem is solved if there are asymmetries in the expected network (Kelejian and Prucha (1998), Bramoullé et al. (2009) and De Giorgi et al. (2010) explore similar assumptions when networks are observed) or observation of groups with distinct sizes is available (see also Lee, 2007).

of the stochastic network[4]. Observation of data on individuals' outcomes and explanatory variables in many self-contained groups, such as classrooms in a school, then provides the identifying condition to estimate the model that serves as a substitute for network observation. In essence, networks are treated as a source of unobserved heterogeneity. I allow for time and fixed effects at the individual or group level when panel data are available and when networks are invariant over time.

The estimator for network spillovers is consistent and asymptotically normally distributed under weak identification assumptions because in this case it is not necessary to separately identify endogenous and exogenous effects. In other words, the parameters of the model are identified up to a set and, as I will show, the network spillovers are constant if evaluated at a parameter that belongs to the identified set. Consistency and confidence regions for the structural parameters are provided making use of the set identification framework.[5] To provide point identification for structural parameters of the model, I explore the difference between observed second moments of the data and those implied by the model. I utilize the fact that the presence of social interactions creates dispersion in average outcomes across groups that cannot be explained by independent variables or peer group heterogeneity alone. Such "excess" variance is explored to build an additional moment restriction and to solve a Generalized Method of Moments (GMM) problem which also includes the score conditions implied by the maximizing the likelihood. This completes the requirements for point identification and consistent estimation of the structural parameters of the model.[6]

To illustrate how this method can be applied in practice, I employ the estimator developed herein to investigate treatment effects both on treated and their peers in a setting potentially conducive to spillovers. The randomized intervention of Bandiera et al. (2013)[7] studies the effect on the treated of the provision of livestock and training to low-income households in Bangladesh and finds that the lack of capital and skills is a strong determinant of the occupational choices of the poor. Targeted households begin new livestock-rearing businesses, increase self-employment hours and reduce wage hours. Due to village-level randomization, a large portion of the individuals in the selected villages are treated, which raises the possibility that network effects are important in determining these outcomes, particularly for peers of those who are treated.

Without using network data, I first demonstrate that network spillovers are economically and statistically significant in determining certain outcomes, especially food expenditure and food security. In these cases, spillovers amount to half of the original treatment for both treated households and their peers. Spillovers of occupational choice and livestock are either insignificant or of a small magnitude. To analyze the structural mechanisms that lead to these results, I then

[4] Due to computational reasons, I will focus on an alternative to integrating the likelihood, based on substituting the unknown networks by expected networks

[5] Chernozhukov et al. (2007), Bugni (2010) and Romano and Shaikh (2010).

[6] Graham (2008) uses a similar idea in the context where networks are observed, within the linear-in-means model.

[7] I thank the authors for sharing data.

decompose spillovers into exogenous and endogenous effects. I demonstrate that, regarding occupational choice and assets, a marginal connection to a treated household has an effect in opposite direction to the effect on the treated: an additional connection decreases self-employment hours, increases wage hours and decreases livestock value.[8] On the other hand, a marginal connection to the treated increases food per capita expenditure and food security to a significant extent. These results are consistent with the phenomenon in which peers of treated households partially fill the vacancies left by those who begin new livestock-rearing businesses and suggests a specialization at the village level, where treated households gain comparative advantage in livestock rearing. Estimating the network structure also demonstrates that network densities are fairly low in the majority of cases, suggesting local interactions via personal contacts as opposed to changes in prices in village-level markets. Finally, inclusion of self-reported network data indicates that family links convey meaningful interactions between households, whereas economic (i.e., non-family) links are much less capable of explaining these social dynamics. This result thus reinforces the idea that families are natural *loci* for sharing information and conducting business.

The methods developed in this paper contribute to the spatial econometrics literature, which has to date considered estimation only when networks are observed, non-stochastic and measured without error. The role of randomness in network formation has also received scant attention in spatial models, despite its importance in social networks (Diestel, 2010). The dependence of existing methods on acquiring knowledge of true networks has been stressed as a limitation of the previous literature (Anselin, 2010; Plümper and Neumayer, 2010).[9] Representative papers in the spatial econometrics literature include those by Anselin (1988) and Kelejian and Prucha (1998, 1999, 2001, 2010). Lee (2004, 2007) and Lee et al. (2010) also consider a maximum likelihood estimator. The case in which networks are not observed is explored in Chapters 2 and 3 of the current thesis[10] and Manresa (2013), who consider the estimation of networks when one group is observed for many periods of time and, as a consequence, clearly suit different applications. It is useful to highlight that the latter papers estimate networks as a collection of pairwise links. In contrast, the current paper is concerned with the probability that a link is formed and the role of exogenous factors therein. The identification results reported by Manski (1993), Graham (2008), Bramoullé et al. (2009) and De Giorgi et al. (2010) are also derived under the assumption that networks are observed. In another strand of the literature, stochastic network formation models, such as those described by Holland and Leinhardt (1981), Frank and Strauss (1986) and

[8] The magnitudes of the estimates imply that peers of treated households compensate around 25-30% of the reduction in treated households' wage hours due to exogenous effects. Endogenous effects move in opposite direction reducing the size of the overall spillover effects. Additional details can be found in Section 1.5.

[9] Plümper and Neumayer (2010) show that misspecification of the networks causes serious bias in parameters of the model, which should be a particular concern for the study of social interactions, where these issues frequently appear. Another facet of the same problem emerges in estimation techniques that proposes using peers of peers' exogenous variables as instruments for one's own endogenous variable, such as Kelejian and Prucha (1998, 1999), Bramoullé et al. (2009) and De Giorgi et al. (2010). To the extent that network data suffers from measurement errors, one risks violating relevance or validity assumptions without awareness.

[10] See also Lam and Souza (2013, 2014)

Strauss and Ikeda (1990), also consider the estimation of network structure only when network observations are available.

Beyond its contribution to the spatial econometric literature, this paper provides a method for systematically investigating network effects, with potential applications in many fields, such as peer effects in education (Sacerdote, 2001; Angrist and Lang, 2004; Ammermuller and Pischke, 2009; Bramoullé et al., 2009; De Giorgi et al., 2010), information diffusion and technology adoption (Foster and Rosenzweig, 1995; Bandiera and Rasul, 2006; Conley and Udry, 2010), social networks and labor outcomes (Rees, 1966; Granovetter, 1973; Montgomery (1991); Conley and Topa, 2002; Munshi, 2003; Pellizzari, 2004; Calvó-Armengol and Jackson, 2004) and crime and delinquent behavior (Glaeser et al., 1996; Dell, 2012). In the macroeconomic and trade literature, these methods can be used to study networks as sources of aggregate fluctuations (Acemoglu et al., 2012) and to estimate parameters of gravity equations (Anderson and van Wincoop, 2003). These approaches are particularly relevant when obtaining data on networks is difficult, time-consuming or expensive, which frequently occurs with social network data because reported links are frequently subjective and prone to behavioral biases.

The remainder of the paper is structured as follows. In Section 2, I introduce the model, define network spillovers and illustrate the inconsistencies that arise when networks are not accounted for. In Section 3, I present the estimator for network effects in the absence of network data and explore its asymptotic properties. Section 4 provides a simulation to validate the performance of the estimator in small samples. Section 5 compares the methods in this paper with existing alternatives for estimating spillovers. It also provides an application to treatment spillovers based on the study of Bandiera et al. (2013). Section 6 concludes.

1.2 Model

The model consists of two parts: a model for stochastic network formation and, given a network, a spatial econometric model that connects explanatory variables to outcomes. The former is sufficiently flexible to allow the probability link formation to depend on exogenous characteristics, such as sharing race or gender or the distance between households.[11] This model may also incorporate individual-level characteristics that attract links or, conversely, that make an individual more inclined to form links with others. In this Section, I assume a simple Bernoulli model for network formation; a full account is provided in Appendix 1.B.[12] Given a network, the spatial econometric model has been extensively considered in the literature, such as by Anselin (1988), Lee (2004), Bramoullé et al. (2009), Lee et al. (2010) and De Giorgi et al. (2010); however, in

[11] The model also falls into the Exponential Random Markovian Graphs category. See Holland and Leinhardt (1981), Frank and Strauss (1986) and Strauss and Ikeda (1990).

[12] See also Wasserman and Faust (1994) and Jackson (2010).

contrast to previous papers, I consider the estimation of network effects in the absence of network data.

I assume that data are available for groups $j = 1, \ldots, v$ and individuals $i = 1, \ldots, n_j$. Individuals interact within groups with observed boundaries, but data with respect to networks within groups are not available. For example, information is available on classes that students belong to but information regarding intra-classroom networks is not available; households are known to be located in villages, but the researcher does not have information regarding the pattern of interaction between households.

For each group j, a network is described with a *directed graph* G_j, an unordered collection of ordered pairs of individuals among n_j individuals. This set lists links along with their associated directions: $\{i, k\} \in G_j$ implies individual i affects individual k in group j. For example, if individual 1 affects 2, 2 affects 3 and 3 affects 2, then $G_j = \{\{1, 2\}, \{2, 3\}, \{3, 2\}\}$. As noted by Wasserman and Faust (1994, Ch. 4), Diestel (2010, Ch. 1), Jackson (2010, Ch. 2), Ballobás (2013, Ch. 1) and others, this representation is quite general. For example, Figure 1.1 portrays estimated links between United States senators, as described by Lam and Souza (2014), based on their 2013 voting records. It is also convenient to express the graph with a so-called *neighboring* or *spatial matrix* W_j, of $n_j \times n_j$ dimensions, a representation of G_j with $\{W_j\}_{ik} = 1$ if $\{i, k\} \in G_j$ and $\{W_j\}_{ik} = 0$ otherwise. It is assumed that no individual affects him or herself; thus $\{W_j\}_{ii} = 0$, for all $i \in \{1, \ldots, n_j\}$.[13]

Network formation is random with a probability law, indexed by parameters of interest θ_g. I use a simple model for clarity of explanation only. Suppose a link between individuals is formed with probability δ_1 when the pair shares a characteristic and δ_0 otherwise. To write the probability distribution function, allow $n_j \times n_j$ matrix Q_j to register the commonality of this individual characteristic. If i and k have the same gender, for example, let the elements of the matrix $\{Q_j\}_{ik} = \{Q_j\}_{ki} = 1$ and zero otherwise. Matrix Q_j could also capture if i self-reported a connection with k. In these cases, $P\{\{W_j\}_{ik} = 1 | \{Q_j\}_{ik}\} = \delta_0(1 - \{Q_j\}_{ik}) + \delta_1 \{Q_j\}_{ik}$. The vector of parameters of interest, carried to estimation, is $\theta_g = (\delta_1, \delta_0)'$. Under the assumptions that link formation is homogenous and independent across pairs of individuals, the probability distribution function is[14]

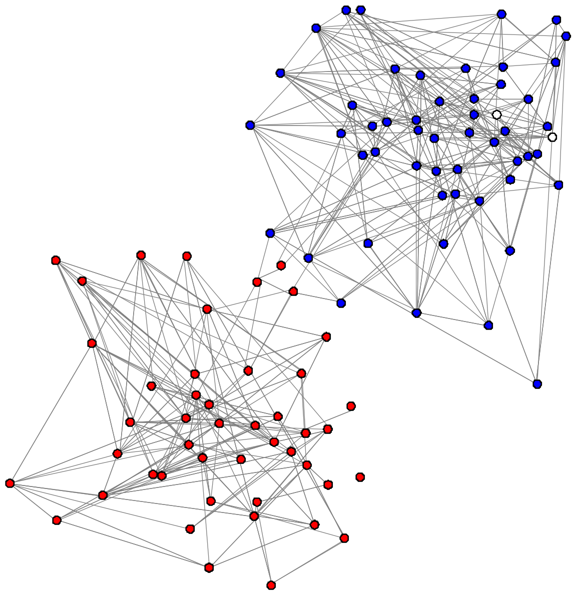
$$\begin{aligned} P\{W_j = w_j | \, Q_j\} \quad &= \quad \prod_{i,k<n_j} (\delta_1^{\{Q_j\}_{ik}} \delta_0^{1-\{Q_j\}_{ik}})^{\{w_j\}_{ik}} \cdot \\ & \quad \cdot ((1-\delta_1)^{\{Q_j\}_{ik}} (1-\delta_0)^{1-\{Q_j\}_{ik}})^{1-\{w_j\}_{ik}}. \qquad (1.1) \end{aligned}$$

Model (1.1) is a simple but arguably truthful representation of situations where differential patterns

[13] G_j and W_j are arrays which depend on the group sizes n_j. In order to keep notation concise, I adopt $G_j \equiv G_{n_j,j}$ and $W_j \equiv W_{n_j,j}$.

[14] This assumption is maintained here only simplicity. In general, link formation may not be independent.

Figure 1.1: Graph example from Lam and Souza (2014).



Note: Red nodes are Senators that belong to the Republican party, blue are Democrats and white are independents.

of associations dominates coalition or strategic behavior, cases in which independence of link formation is violated. A classroom divided along gender or racial lines is possibly an example that satisfies assumption above.

Given a network, it remains to describe a model linking explanatory variables to outcomes. Denote W_j^0 and M_j^0 as two *random* and *unobserved* realizations of a network-generating process, such as the one introduced above. This network is embedded is a spatial econometric model, which incorporates dependence of one's own outcome variable on others' outcome variables and others' exogenous variables. For a particular group $j = 1, \ldots, v$ composed of n_j individuals, the model is given by

$$y_j \quad = \quad \lambda_0 W_j^0 y_j + x_j \beta_{10} + W_j^0 x_j \beta_{20} + v_j \tag{1.2}$$

where y_j is a column vector of dimension $n_j \times 1$, x_j is $n_j \times k$, and v_j is the $n_j \times 1$ disturbance vector. Disturbance term v_j is assumed to follow a structure that allows for spatial dependence, $v_j = \rho_0 M_j^0 v_j + \epsilon_j$, where ϵ_j is $n_j \times 1$, independent and normally distributed with variance σ_0^2. As a particular example, this includes group-level clustering and heteroskedasticity that arises from heterogeneous exposure to disturbances of others.

In Manski's (1993) taxonomy, the term $W_j^0 y_j$ corresponds to the *endogenous effects*, or the dependence of one's own behavior on the behavior of others through link strength scalar parameter λ_0. Parameter β_1, of dimension $k \times 1$, captures the direct effect of one's own exogenous variables on one's own dependent variables. Parameter β_2, of the same dimension, describes the effects of others' exogenous variables on one's own dependent variable. Thus, $W_j^0 x_j$ is denoted as *contextual* or *exogenous effects*. *Correlated effects* are represented by the error $v_j = \rho_0 M_j^0 v_j + \epsilon_j$ and fixed effects, which I describe in Section 1.3.4. This model is similar to the model in Bramoullé et al. (2009) and Lee et al. (2010), among other studies, and is known as the "mixed regressive-spatial autoregressive model" in the spatial econometrics literature (Anselin, 1988). I am then interested in the estimation of usual spatial parameters $\theta_s = (\lambda_0, \beta_{10}', \beta_{20}', \rho_0, \sigma_0^2)'$ and $\theta_g = (\delta_0, \delta_1)$. Hence, the complete set of structural parameters of interest is $\theta = (\theta_s', \theta_g')'$.

Dependence of one's own outcomes on other's outcomes and exogenous variables often means that the overall response to exogenous variation exceeds β_{10}. As a consequence, to the extent that individual network spillovers depend on one's own exogenous variation, estimators for β_{10} that do not account for network spillovers are frequently inconsistent, as I demonstrate immediately below.

Using the series decomposition[15] $(I_{n_j} - \lambda_0 W_j^0)^{-1} = \sum_{s=0}^{\infty} \lambda_0^s (W_j^0)^s$ to obtain the reduced-form model, the expected outcomes are separated into two components: the individual reaction or elasticity with respect to x_j and its effect through the network,

$$\mathbb{E}y_j = x_j\beta_{10} + W_j^0 x_j \beta_{20} + \sum_{s=1}^{\infty} \left(\lambda_0 W_j^0\right)^s \left(x_j\beta_{10} + W_j^0 x_j \beta_{20}\right). \tag{1.3}$$

The term $x_j\beta_{10}$ is understood as the individual-level elasticity with respect to x_j if networks were irrelevant, whereas the second and third terms jointly denote network spillovers, the additional effect on the mean exclusively due to individual interconnectedness:

$$\begin{aligned}\varphi\left(x_j, \theta_0\right) &\equiv W_j^0 x_j \beta_{20} + \sum_{s=1}^{\infty} \left(\lambda_0 W_j^0\right)^s \left(x_j\beta_{10} + W_j^0 x_j \beta_{20}\right) \\ &= \sum_{s=1}^{\infty} \lambda_0^{s-1} \left(W_j^0\right)^s x_j \left(\lambda_0\beta_{10} + \beta_{20}\right). \end{aligned} \tag{1.4}$$

Clearly, if $\lambda_0 = 0$ and $\beta_{20} = 0_{k\times 1}$, or $\delta_1 = \delta_0 = 0$, then $\varphi\left(x_j, \theta_0\right) = 0$. Spillover $\varphi(x_j, \theta_0)$ is a $n_j \times 1$ vector because each individual accrues his or her own spillover.

Separate identification of the individual reaction and network spillovers is relevant in at least two scenarios. Provided that the ultimate goal is to consistently estimate β_{10}, $\varphi\left(x_j, \theta_0\right)$ is a confounding factor. As shown in Subsection 1.2.1, when networks are unaccounted for, consistent estimating β_{10} requires an underlying network structure such that one's own network spillovers are independent of one's own exogenous variation, a condition that breaks down in simple counterexamples.

[15] Conditions for existence of this decomposition are derived in Section 1.3.

Moreover, network spillovers are of interest in their own right, as shown by the plethora of examples in the literature. Glaeser et al. (1996) argue that social interactions explain petty criminal behavior very well, but are also of moderate importance in explaining more serious offenses. Hence, crime prevention policies have indirects effects by reducing of others' proclivity toward criminal activity, and the effect's magnitude then shapes and informs the public policy debate. In another example, Foster and Rosenzweig (1995) reason that farmers' decision to adopt high-yielding seed varieties depends on other farmers' decisions regarding adoption and their accrued profit; consequently, a single farmer's adoption decision multiplies itself by inducing others to adopt also. Finally, note that parameter $\varphi\left(x_j,\theta\right)$ can be explored to optimize treatment effects under a given budget of resources. To the extent that network spillovers are prevalent and positive, often average treatment effects can frequently be maximized by concentrating treatment in fewer groups.

Remark 1. Panel or spatiotemporal models can be naturally introduced from equation (1.2). Index explanatory variables and outcomes at time $t = 1, \dots, T$ and the complete model reads

$$y_{jt} \quad = \quad \lambda_0 W_j^0 y_{jt} + x_{jt}\beta_{10} + W_j^0 x_{jt}\beta_{20} + \alpha_j + \gamma_t + v_{jt} \tag{1.5}$$

where α_j is a vector of $n_j \times 1$ time-invariant coefficients (but allowed to vary at the group or individual levels), which are also denoted, following Manski (1993), as *correlated effects.* The vector γ_t represents time effects. Under the invariance of networks with respect to time, I propose a data transformation that eliminates these nuisance parameters in Subsection 1.3.4. When x_{jt} is a treatment indicator, model (1.5) can be described as a differences-in-differences estimator supplemented with a network component. In the absence of network effects ($\lambda_0 = 0$ and $\beta_{20} = 0_{k\times 1}$), the model is reduced to a standard differences-in-differences. In this context, the terms $\lambda_0 W_j^0$ and $W_j^0 x_{jt}\beta_{20}$ measure the treatment spillovers through the network. □

1.2.1 Inconsistency when Networks are Unaccounted for

Equations (1.3) and (1.4) immediately imply that the aggregate group response to a shock is the sum of one's own variation in the absence of networks (β_{10}) and network spillovers (φ),

$$y_j \quad = \quad x_j\beta_{10} + \varphi(x_j, \theta_0) + \epsilon_j. \tag{1.6}$$

On the one hand, disentangling the two components provides insights into the mechanisms that determine the responses to the shock. In particular, the role of networks is separated from the response in its absence; this construct is useful for example to provide external validity to randomized controlled trials prior to reimplementation in settings in which networks might differ. On the other hand, the omission of $\varphi(x_j, \theta_0)$ biases OLS estimates when one's own spillover is not orthogonal to one's own shock.

Consistency for β_{10} requires that $\mathbb{E}(\varphi(x_j, \theta_0)|x_j) = 0$ for all $i = 1, \ldots, n_j$, the case in which the researcher would be oblivious to network spillovers. At the other extreme, only under perfect correlation between x_j and $\varphi(x_j, \theta_0)$ the OLS estimates are consistent for the sum of β_{10} and full spillovers. In general, however, independence is not generally attained, failing in particular under reciprocated networks or correlation between x_{ij} and x_{kj} for $i \neq k$[16]. In this case, the biasing term $(x_j' x_j)^{-1} x_j' \mathbb{E}(\varphi(x_j, \theta_0)|x_j)$ depends on the network structure, which is unknown; thus, the size and presence of bias are also unknown. I now provide some examples.

Example 1. *(Classrooms and the linear-in-means model).* Manski (1993) proposes the linear-in-means network model in which individuals interact with all others in a given classroom and

$$W_j^0 \;=\; \begin{bmatrix} 0 & \frac{1}{n-1} & \cdots & \frac{1}{n-1} \\ \frac{1}{n-1} & 0 & \cdots & \frac{1}{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n-1} & \frac{1}{n-1} & \cdots & 0 \end{bmatrix} \;=\; \frac{1}{n-1}\iota_n \iota_n' - \frac{1}{n-1} I_n$$

where I_n is the $n \times n$ identity matrix and ι_n is the $n \times 1$ vector of ones. Suppose x_j is a treatment dummy and α is the proportion of the individuals in the group that were treated. The expectation of response conditional on treatment is obtained via the reduced-form model

$$y_j = \left(S_j^0\right)^{-1} x_j \beta_{10} + \left(S_j^0\right)^{-1} W_j^0 x_j \beta_{20} + \left(S_j^0\right)^{-1} \left(R_j^0\right)^{-1} \epsilon_j$$

where $S_j^0 = I_n - \lambda_0 W_j^0$, $R_j^0 = I_n - \rho_0 M_j^0$, $(S_j^0)^{-1} = \frac{n-1}{n-1+\lambda_0} I_n + \frac{\lambda_0}{(n-1+\lambda_0)(1-\lambda_0)} \iota_n \iota_n'$ and $(S_j^0)^{-1} W_j^0 = -\frac{1}{n-1+\lambda_0} I_n + \frac{1+\lambda_0}{(n-1+\lambda_0)(1-\lambda_0)} \iota_n \iota_n'$. The expectation of the outcome of individual i in group j, conditional on not receiving a treatment, is

$$\mathbb{E}\left[y_{ij} |\, x_{ij} = 0\right] \;=\; \alpha n \frac{\lambda_0 \beta_{10} + (1 + \lambda_0)\,\beta_{20}}{(n - 1 + \lambda_0)\,(1 - \lambda_0)}$$

and describes the network spillovers to untreated individuals. Conditioned on receiving a treatment,

$$\mathbb{E}\left[y_{ij} |\, x_{ij} = 1\right] \;=\; \frac{(n-1)\,\beta_{10} - \beta_{20}}{n - 1 + \lambda_0} + \alpha n \frac{\lambda_0 \beta_{10} + (1 + \lambda_0)\,\beta_{20}}{(n - 1 + \lambda_0)\,(1 - \lambda_0)} \tag{1.7}$$

thus, in general, the population difference $\mathbb{E}[y_{ij} |\, x_{ij} = 1] - \mathbb{E}[y_{ij} |\, x_{ij} = 0]$ is approximately β_{10} for a typical classroom size, such as $n = 25$. This result implies that OLS estimates are consistent for β_{10} even if oblivious to network spillovers. □

Example 2. *(Households and local interaction).* Households typically interact with few others,

[16] This type of violation would occur in the case in which individuals who are eligible for a treatment are also more likely to have other eligible individuals in their social networks. Snowballing a treatment is another clear example of violation of the no self-spillover condition $\mathbb{E}(\varphi(x_j, \theta_0)|x_j) = 0$.

and relations are generally reciprocated. For the sake of example, suppose a network consists of isolated subgroups of five households, in which interaction across subgroups is negligible in comparison with interactions within. In this setting, W_j^0 is a block-diagonal matrix with $\frac{n}{5}$ blocks[17], or $W_j^0 = I_{\frac{n}{5}} \otimes (\frac{1}{4}\iota_5\iota_5' - \frac{1}{4}I_5)$. Suppose a proportion α receive a treatment. In contrast to the previous example, the difference $\mathbb{E}[y_{ij}|\, x_{ij} = 1] - \mathbb{E}[y_{ij}|\, x_{ij} = 0]$ is no longer approximately β_{10}, which can be shown by replacing $n = 5$ in equation (1.7). As a consequence, OLS estimates are biased for β_{10} and capture the portion of one's own spillovers that correlate with one's own treatment status. □

Generally, OLS is only consistent for β_{10} in particular network structures. When networks remain unobserved, the implementation of such a strategy depends on hypotheses that rule out feedback mechanisms. In Section 1.3, I provide a method for consistently estimating $\varphi(x_j, \theta)$ under few identifying assumptions that address both motivating elements. The method is based on a maximum likelihood integrated with respect to unobserved networks, resulting in a likelihood that is independent of network observation. In essence, I deal with the networks as unobserved heterogeneity. As will be shown, although the point identification of θ is not obtained without additional assumptions, spillover $\varphi\,(x, \theta)$ is constant within the identified set and thus point-identified. Section (1.3.3) uses additional identifying information to sort through the identified set and reestablish point identification for the structural parameters.

1.3 Estimation of Network Effects

Spatial econometric models dealt with the case of known W_0 and M_0. Under certain conditions, including network observation, Lee (2004) and Lee et al. (2010) show consistency and asymptotic normality of a quasi-maximum likelihood estimator for θ_s. In this scenario, accounting for network effects would not pose a challenge. However, these results are of no use if W_0 and M_0 are unobserved or imperfectly observed, such as when there are measurement errors[18] or data are self-reported. Recently, other papers suggested similar approaches to this problem. Hsieh and Lee (2015)[19] are concerned with a social interactions model in which an observed network is formed endogenously and, for this purpose, propose a bias corrections using a network formation model. In contrast, the current paper does not assume the observation of the network.

In contrast, I deal with networks as a form of unobserved heterogeneity. Networks are randomly formed with certain probability law, homogenous across groups, and observation of many groups is available. More formally, I propose an integrated likelihood approach. The likelihood unconditional on network observation is the integral of the likelihood given a network (from a

[17] For simplicity, assume n is a multiple of 5.

[18] Observation of networks with measurement errors constitute a challenge for methods that are, directly or indirectly, based on network-generated instruments, as validity assumptions are often violated. This is the case of Kelejian and Prucha (1998, 1999), Bramoullé et al. (2009) and others. Also see Plümper and Neumayer (2010).

[19] see also Goldsmith-Pinkham and Imbens (2013)

spatial model soon introduced) with respect to the probability density function for a stochastic network model:

$$\ln \mathcal{L}\left(\theta | y_n, x_n, Q_n\right) \quad = \quad \int \ln \mathcal{L}\left(\theta | y_n, x_n, W_n, M_n\right) dP\left(W_n, M_n | Q_n, x_n, \theta\right) \tag{1.8}$$

where $y_n = (y_1', \ldots, , y_v')'$, $x_n = (x_1', \ldots, x_v')'$, W_n and M_n are a random block matrix with $W_1, \ldots, W_v$ and $M_1, \ldots, M_v$ along the main diagonal. Therefore W_n and M_n have dimension $n \times n$, $n = \sum_{j=1}^{v} n_j$. Likelihood $\ln \mathcal{L}\left(\theta | y_n, x_n, W_n, M_n\right)$ is derived from a spatial model and for simplicity it is assumed independent of Q_n.[20] The probability density function of networks, $P(W_n, M_n | Q_n, x_n, \theta)$, depends on exogenous variables Q_n and x_n and parameters θ. In this way, the probability that peers form a link is affected by individual characteristics Q_n which do not directly affect the mean and exogenous variables x_n. For example, connections may depend on a treatment status dummy[21].

Since there is a finite number of possible graphs, labelled $s = 1, \ldots, g_{nv}$, with $g_{nv} = 2^{\sum_{j=1}^{v} n_j(n_j-1)}$, the full likelihood can be exactly approximated by

$$\ln \mathcal{L}\left(\theta | y_n, x_n, Q_n\right) \quad = \quad \sum_{s=1}^{g_{nv}} \ln \mathcal{L}\left(\theta | y_n, x_n, W_n^s\right) P\left(W_n^s | Q_n, x_n, \theta\right). \tag{1.9}$$

Even for relatively small numbers of n_j and v, g_{nv} is an enormous number. Taking $v = 5$ and $n_j = 10$ for $j = 1, \ldots, v$, the total of number of graphs g_{nv} exceeds 10^{135}. Therefore, evaluation of this integral is computationally costly and burdensome.

I propose a modification that implements a computationally efficient estimator. I substitute W_0 and M_0 for their expected values[22] $W_n^e\left(Q_n, \theta\right) = \int W_n dP\left(W_n | Q_n, x_n, \theta\right)$ and $M_n^e\left(Q_n, \theta\right) = \int M_n dP\left(M_n | Q_n, x_n, \theta\right)$. Estimation of network spillovers and structural parameters is based on the likelihood of the model

$$y_j \quad = \quad \lambda W_j^e\left(Q_j, \theta\right) y_j + x_j \beta_1 + W_j^e\left(Q_j, \theta\right) x_j \beta_2 + v_j^e \tag{1.10}$$

with $v_j^e(Q_j, \theta) = \rho M_j^e(Q_j, \theta) v_j + \epsilon_j$. The term "pseudo-likelihood" is used to distinguish the likelihood of this model from the likelihood of the model with known networks.

Model (1.10) is equivalent to the model if networks were observed in addition to mispecification

[20] This assumption means that characteristics that underpin the network formation do not affect the spatial model directly, but only via the networks.

[21] I rule out endogeneity with respect to outcomes y_n. This is the topic of a future extension to the current paper.

[22] For simplicity of explanation, momentarily assuming W_j^0 and M_j^0 are independent, which does not hold for the rest of the paper.

terms that are close[23] to zero when $\theta = \theta_0$,

$$\begin{aligned} y_j &= \lambda W_j^0 y_j + x_j \beta_{10} + W_j^0 x_j \beta_{20} + \lambda \left\{ W_j^e \left(Q_j, \theta \right) - W_j^0 \right\} y_j \\ &= + \left\{ W_j^e \left(Q_j, \theta \right) - W_j^0 \right\} x_j \beta_{20} + v_j^e. \end{aligned} \tag{1.11}$$

Intuitively, the misspecification terms containing $\{W_j^e(Q_j, \theta) - W_j^0\}$ are of small relevance when a large number of groups is observed. This point is best exemplified if group sizes are constant, condition that is not carried for the remainder of the paper. Under certain conditions, a Law of Large Numbers ensures that $v^{-1} \sum_{j=1}^{v} W_j^0 \xrightarrow{p} W_j^e(Q_j, \theta)$. Averaging the model across groups then implies that misspecification terms are small when $v \longrightarrow \infty$.

The substitution of true networks for expected networks has two consequences. First, the fact that model is inherently misspecified implies that the equality between information matrix and expected hessian does not hold, which will have implications for the expression of the asymptotic variance. Second, the introduction of expected networks implies that pointwise identification of parameters θ is generally not achieved. There are multiple combinations of λ, θ and β_2 such that the model (1.10) is observationally equivalent.

Subsections 1.3.1 to 1.3.3 discuss identification in three scenarios. In Subsection 1.3.1, I show that knowledge of one parameter (I arbitrarily focus the discussion on λ_0) restores identification under the mild additional assumption that there are at least three distinct group sizes. I will show that variation in group sizes allows me to separately identify endogenous and exogenous effects.[24] Knowledge of λ_0 separately identifies the case of a weak connections with high probability (low λ, high δ_0 and δ_1) from the case of strong connections with low probability (high λ, low δ_0 and δ_1). This is then sufficient to fully identify the model.

Subsection 1.3.2 considers the estimation of θ when λ_0 is unknown and no additional information is provided. In this case, the true parameter θ_0 is identified up to a set Θ_0. Importantly, I demonstrate that parameters in the identified set yield network spillovers equal to the spillovers evaluated at the true parameter. That is, for all $\theta \in \Theta_0$, $\varphi(x_j, \theta) = \varphi(x_j, \theta_0)$. Hence, network spillovers are point-identified. I provide the set estimator and confidence regions for the parameters. In the interest of generality, the test for network data validity is also proposed in this context. I adapt the ideas of Chernozhukov et al. (2007), Romano and Shaikh (2010) and Bugni (2010) to provide confidence regions for the structural parameter θ.

The problem with unknown λ_0 can be analogously interpreted as an under-identified Generalized Method of Moments (GMM) problem in which moment conditions are given by the score of

[23] Comparison between likelihood computed with expected network and true networks can be found in Tables 1.F.1 and 1.F.2 in the Appendix.

[24] As also shown by Lee (2007) for the case in which networks are known. Asymmetries in the network, such as those considered by Kelejian and Prucha (1998,1999), Bramoullé et al. (2009) and De Giorgi et al. (2010) could also be used to provide identification. These would in turn require asymmetries in Q_n.

the likelihood. The previous non-identification result manifests itself as the absence of one moment condition relative to the number of parameters. In Subsection 1.3.3, I then make full use of the model to obtain one additional moment condition which restores point identification of θ.

Earlier work on identification of social interactions observed that the presence of social interactions generates dispersion of average group outcomes beyond what can be explained by variance of explanatory variables of peer group heterogeneity alone (Glaeser et al., 1996; Graham, 2008). I implement this idea in the case where networks are unknown. This introduces an additional moment condition: the difference between observed and model-implied across-group outcome variance. As I will show, this restores identification. Consistency and asymptotic normality of the GMM estimator follows. Before proceeding, I formally derive the likelihood.

Define $S_j^e(Q_j,\theta) \equiv I - \lambda W_j^e(Q_j,\theta)$, $S_j^0(\lambda) \equiv I - \lambda W_j^0$, $S_j^0 \equiv S_j^0(\lambda_0)$, $R_j^e(\theta) \equiv I - \rho M_j^e(Q_j,\theta)$, $R_j^0(\rho) \equiv I - \rho M_j^0$, $R_j^0 \equiv R_j^0(\rho_0)$, $Z_j^e(Q_j,\theta_c) = (x_j, W_j^e(Q_j,\theta_c)x_j)$ and the block matrices $W_n^0(Q_n,\theta_c) = \text{diag}(W_1^0(Q_1,\theta_c),\ldots,W_v^0(Q_v,\theta_c))$, $W_n^e(Q_n,\theta_c) = \text{diag}(W_1^e(Q_1,\theta_c),\ldots,$ $W_v^e(Q_v,\theta_c))$, $M_n^e(Q_n,\theta_c) = \text{diag}(M_1^e(Q_1,\theta_c),\ldots,$ $M_v^e(Q_v,\theta_c))$, $S_n^e(Q_n,\theta_c) =$ $\text{diag}(S_1^e(Q_1,\theta_c),\ldots,S_v^e(Q_1,\theta_c))$, and $Z_n^e(Q_n,\theta_c) = (Z_1^{e'}(Q_1,\theta_c),\ldots,Z_v^{e'}(Q_v,\theta_c))'$. Model (1.2) can be denoted $y_n = \lambda_0 W_n^0 y_n + x_n\beta_{10} + W_n^0 x_n \beta_{20} + v_n$, where $v_n = (v_1',\ldots,v_v')'$. The pseudo-likelihood is

$$\begin{aligned}\ln \mathcal{L}_n^e(\theta|\, y,x,Q_n) &= -\frac{n}{2}\ln\left(2\pi\sigma^2\right) + \ln\left|S_n^e(Q_n,\theta)\right| + \ln\left|R_n^e(Q_n,\theta)\right| \\ &\quad -\frac{1}{2\sigma^2}\epsilon_n^{e'}(Q_n,\theta)\,\epsilon_n^e(Q_n,\theta)\end{aligned} \tag{1.12}$$

with $\epsilon_n^e(Q_n,\theta) = R_n^e(Q_n,\theta)\left(S_n^e(Q_n,\theta)y_n - Z_n^e(Q_n,\theta)\beta\right)$ for $\beta = (\beta_1',\beta_2')'$. Parameters β and σ^2 are concentrated out of the likelihood, simplifying derivations and implementation. Denote $\theta_c = \theta \setminus \left\{\beta,\sigma^2\right\}$ the non-concentrated parameters. At each θ_c, the closed-form solutions for the concentrated parameters are

$$\begin{aligned}\hat\beta(Q_n,\theta_c) &= \left(Z_n^{e'}(Q_n,\theta_c)R_n^{e'}(Q_n,\theta_c)R_n^e(Q_n,\theta_c)Z_n^e(Q_n,\theta_c)\right)^{-1}\cdot \\ &\quad \cdot Z_n^{e'}(Q_n,\theta_c)R_n^{e'}(Q_n,\theta_c)R_n^e(Q_n,\theta_c)S_n^e(Q_n,\theta_c)y_n \\ \hat\sigma^2(Q_n,\theta_c) &= \frac{1}{n}(S_n^e(Q_n,\theta_c)y_n - Z_n^e(Q_n,\theta_c)\hat\beta(\theta_c))'R_n^{e'}(Q_n,\theta_c)R_n^e(Q_n,\theta_c)(S_n^e(Q_n,\theta_c)y_n \\ &\quad -Z_n^e(Q_n,\theta_c)\hat\beta(\theta_c)) \\ &= \frac{1}{n}y_n'S_n^{e'}(Q_n,\theta_c)R_n^{e'}(Q_n,\theta_c)P_n^e(Q_n,\theta_c)R_n^e(Q_n,\theta_c)S_n^e(Q_n,\theta_c)y_n\end{aligned}$$

where P_n^e is the projection matrix

$$\begin{aligned}P_n^e(Q_n,\theta_c) &= I_n - R_n^e(Q_n,\theta_c)Z_n^e(Q_n,\theta_c)\left(Z_n^{e'}(Q_n,\theta_c)R_n^{e'}(Q_n,\theta_c)R_n^e(Q_n,\theta_c)Z_n^e(Q_n,\theta_c)\right)^{-1}\cdot \\ &\quad \cdot Z_n^{e'}(Q_n,\theta_c)R_n^{e'}(Q_n,\theta_c)\end{aligned}$$

and $P_n^e \equiv P^e\left(Q_n, \theta_c^0\right)$. The final form for the concentrated pseudo-likelihood brought to maximization is

$$\begin{aligned} \ln \mathcal{L}_n^c\left(\theta_c \middle|\, y_n, x_n, Q_n\right) &= -\frac{n}{2}\left(\ln\left(2\pi\right)+1\right) - \frac{n}{2}\ln\hat{\sigma}^2\left(Q_n, \theta_c\right) + \left|S_n^e\left(Q_n, \theta_c\right)\right| \\ &\quad + \left|R^e\left(Q_n, \theta_c\right)\right|. \end{aligned} \tag{1.13}$$

The final estimator is $\hat{\theta} = (\hat{\theta}_c', \hat{\beta}(\hat{\theta}_c)', \hat{\sigma}^2(\hat{\theta}_c))'$, where $\hat{\theta}_c \equiv \arg\max_{\theta\in\Theta_c} \ln \mathcal{L}_n^c\left(\theta_c \middle|\, y_n, x_n, Q_n\right)$. I now lay formal hypothesis to guarantee asymptotic properties of the estimator.

1.3.1 Pointwise identification of θ when λ_0 is known

In this subsection, I present the basic assumptions for consistent estimation and pointwise identification of the parameters in the model. Identification Assumption 6, required for pointwise identification of θ, holds only if λ_0 is known to the researcher[25]. Assumptions 1-5 are maintained throughout the remaining subsections.

The first assumption defines the true model, properties of the networks and homogeneity of the probability law (P) that generates (unobserved) networks across groups. The zero main diagonal is essentially an identification condition and implies that no individual affects him or herself. The independence of P with respect to β and σ^2 allows me to concentrate these parameters, as described previously, and is taken for simplicity only as results do not depend crucially on it.

Assumption 1. *For each group $j = 1, \ldots, v$, data are generated according to the model*

$$y_j = \lambda_0 W_j^0 y_j + x_j\beta_{10} + W_j^0 x_j \beta_{20} + v_j$$

with $v_j = \rho_0 M_j^0 v_j + \epsilon_j$ and $\epsilon_j \sim \mathbb{N}\left(0, \sigma^2 I\right)$. The elements of x_n and Q_n are uniformly bounded constants. Let $mat_{n_j}\left(\{0,1\}\right)$ denote the space of n_j-by-n_j-by-2 matrices with entries in $\{0,1\}$ and zero main diagonal, let $(\Omega, \mathcal{F}, P)$ be a a probability space with $\mathcal{F}$ as σ-algebra of subsets of Ω and P as probability measure. $\{W_j^0, M_j^0\}$ is particular realization from a random matrix[26]*, a measurable map from $(\Omega, \mathcal{F})$ to $mat_{n_j}\left(\{0,1\}\right)$, with probability distribution function $P\left(W_j, M_j \middle|\, \theta, x_j\right)$ with common functional form across groups. P does not depend on β or σ^2.*

In some applications, it is customary to conduct a row-sum normalization of W_j, the operation consisting of replacing W_j by a W_j^* with $\{W_j^*\}_{ik} = \{W_j\}_{ik}/\sum_{s=1}^{n_j}\{W_j\}_{is}$ (Anselin, 1988, Kelejian and Prucha, 1998, 1999, 2001, 2010, Lee, 2004, 2007, Lee et al., 2010). This implies that all individuals in the group are affected by and affect others to the same extent: row sums of W_j^* add

[25] In fact, Assumption 6 holds in the case where one parameter among λ_0, β_{20} and θ_g^0 is known. For simplicity, I arbitrarily focus the argument on λ_0.

[26] In fact, $\{W_j^0, M_j^0\}$ are arrays and full notation should include respective dimensions, $\{W_{n_j,j}^0, M_{n_j,j}^0\}$. This is suppressed for simplicity.

to one. This assumption is avoided here on the basis of anecdotal observation that individuals are generally not homogenous in terms of their connection to others in the group. In classrooms, for example, some students may be more affected by peers than others. I leave networks to be, more simply, a collection of binary numbers.

It is well-known that under row-sum normalization condition, $|\lambda_0| < 1$ suffices for uniform boundedness of W_j^0 and $(S_j^0)^{-1}$, with $S_j^0 \equiv I_{n_j} - \lambda_0 W_j^0$ (Anselin, 1988). In the current setting, I propose the following notion of boundedness: let $\max_i \ |\lambda_0 \sum_{k=1}^{n} \{W_j^0\}_{ik}| \leq 1$, and so no row multiplied by λ_0 in absolute value exceeds one. This includes row-sum normalization as a special case; for constant row sums W_j^0 across rows, $\lambda_0 \sum_{k=1}^{n} \{W_j^0\}_{ik} = \lambda_0^* \sum_{k=1}^{n} \{W_j^{*0}\}_{ik}$ with $\lambda_0^* = \lambda_0 \sum_{s=1}^{n_j} \{W_j\}_{1s}$. In this case, it is clear that letting W_j^0 as a collection of binary numbers and $|\lambda_0|$ closer to zero is only a normalization option. Formally,

Assumption 2. *The sequence of n-by-n realized matrices $\lambda_0 W_n^0$ and $(S_n^0)^{-1}$ and expected matrices $\lambda W_n^e(Q_n, \theta)$ and $(S_n^e(Q_n, \theta))^{-1}$ are uniformly bounded. $W_n^e(Q_n, \theta)$ exists for all $\theta \in \Theta$.*

The next assumption guarantees y_j has an equilibrium and its mean and variance are well defined.

Assumption 3. *S_j^0 is nonsingular, $j = 1, \ldots, n$.*

Asymptotics on v and n_j, without any specific order of divergence, is necessary to guarantee that the misspecification term goes to zero asymptotically and variance terms are consistently estimated in the limit.

Assumption 4. *$n \to \infty$ where $n = \sum_{j=1}^{v} n_j$.*

As a minor technical point, it is only necessary that non-concentrated parameters belong to a compact parameter set Θ_c.

Assumption 5. *The parameter set Θ_c is compact and the true parameter $\theta_c^0 \in \Theta_c^0$.*

Next, I lay out the identifications conditions required for point identification of parameters. The Assumption resembles similar conditions of Bramoullé et al. (2009) and Lee et al. (2010).

Assumption 6. *(Identification). λ_0 is known, network effects do not cancel out ($\beta_{20} \neq \lambda_0 \beta_{10}$), and x_n, $W_n^e(Q_n, \theta_c^0)\, x_n$ and $\left(W_n^e(Q_n, \theta_c^0)\right)^2 x_n$ are linearly independent.*

It is useful to note that variation in group sizes is often sufficient to assure independence between x_n, $W_n^e(Q_n, \theta_c^0)\, x_n$ and $\left(W_n^e(Q_n, \theta_c^0)\right)^2 x_n$. This is also seen in the subgroup model of Lee (2007) where individuals are sorted in many groups. In particular, let the probabilistic model for network formation be the pure Bernoulli, where links are formed with probability δ_0, independent of exogenous characteristic. Then $W_j^e(Q_j,\theta_c^0) = \delta_0(\iota_{n_j}\iota'_{n_j} - I_{n_j})$ and $(W_j^e(Q_j,\theta_c^0))^2 = \delta_0^2(n_j-2)(\iota_{n_j}\iota'_{n_j} + I_{n_j})$. With at least three distinct values of n_j, independence condition in the previous Proposition is guaranteed[27].

Under the conditions introduced above, I present the basic Theorem. Proofs are found in the Appendix 1.D.

Theorem 1. *Under assumptions 1-6, $\hat{\theta}$ is a consistent estimator for θ_0, i.e., $\hat{\theta}\xrightarrow{p}\theta_0$.*

Asymptotic distribution can be obtained from a Taylor expansion around the point $\frac{\partial \ln \mathcal{L}^e\left(\hat{\theta}\middle|y_n,x_n,Q_n\right)}{\partial\theta} = 0$. For a point $\tilde{\theta}$ between $\hat{\theta}$ and θ_0,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) = \left[\frac{1}{n}\frac{\partial^2 \ln \mathcal{L}^e(\tilde{\theta}\Big|\, y_n,x_n,Q_n)}{\partial\theta\partial\theta'}\right]^{-1}\frac{1}{\sqrt{n}}\frac{\partial \ln \mathcal{L}^e\left(\theta_0|\, y_n,x_n,Q_n\right)}{\partial\theta}. \tag{1.14}$$

The variance matrix of the score vector is $\Sigma_n(\lambda_0) \equiv \mathbb{E}[\frac{1}{\sqrt{n}}\frac{\partial \ln \mathcal{L}^e(\theta_0|y_n,x_n,Q_n)}{\partial\theta}\cdot\frac{1}{\sqrt{n}}\frac{\partial \ln \mathcal{L}^e(\theta_0|y_n,x_n,Q_n)}{\partial\theta'}]$. In the limit, $\hat{\theta}\xrightarrow{p}\theta_0$, which implies $\tilde{\theta}\xrightarrow{p}\theta_0$ and so the Hessian matrix converges to $\Omega_n(\lambda_0) = \mathbb{E}[\frac{1}{n}\frac{\partial \ln \mathcal{L}^e(\theta_0|y_n,x_n,Q_n)}{\partial\theta\partial\theta'}]$. As the model is inherently misspecified, the Hessian is not equal to the expected outer product of the gradient. The asymptotic variance-covariance matrix converges instead to the usual sandwich estimator. That is,

Theorem 2. *Under assumptions 1-5, $\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{p}N(0,\Sigma^{-1}(\lambda_0)\Omega(\lambda_0)\Sigma^{-1}(\lambda_0))$, where $\Sigma(\lambda_0) = \lim_{n\to\infty}\Sigma_n(\lambda_0)$ and $\Omega(\lambda_0) = \lim_{n\to\infty}\Omega_n(\lambda_0)$.*

1.3.2 Set identification of θ when λ_0 is unknown

There is one simple way asymptotic independence of the matrices is violated. Any path $\{\lambda_+,\beta_{2+},\theta_c^+\}$ such that $W_n^e\left(Q_n,\theta_c^+\right)x_n\beta_{2+} = W_n^e\left(Q_n,\theta_c^0\right)x_n\beta_{20}$ and $\lambda_+W_n^e\left(Q_n,\theta_c^+\right) = \lambda_0W_n^e\left(Q_n,\theta_c^0\right)$ results in a similar reduced-form, constituting a breakdown of Assumption 6. Parameters are not individually identified, which is compatible with the difficulty in separately identifying a large number of weak connections from a small number of strong connections. I now turn to the problem of estimation and inference on the identified set.

Using assumptions 1-5 only, I employ methods of estimation and inference on set-identified models of Chernozhukov et al. (2007), Romano and Shaikh (2010) and Bugni (2010) to establish desired results. The point of departure from classic asymptotic analysis is the observation that the identified set $\Theta_0 = \{\tilde{\theta}\in\Theta : F_n(\tilde{\theta}) = F_n(\theta_0)\}$, for $F_n\left(\theta\right) = \mathbb{E}\ln\mathcal{L}_n^e\left(\theta\right)$, and the estimated set $\hat{\Theta} = \{\tilde{\theta}\in\Theta : \ln\mathcal{L}_n^e(\tilde{\theta}) = \inf_{\theta\in\Theta}\ln\mathcal{L}_n^e\left(\theta\right)\}$ are not singletons.

[27] That is, if there are three distinct values of n_j, the only conformable vectors c_1, c_2 and c_3 such that $xc_1 + \delta_0(\text{diag}(\iota_{n_1}\iota'_{n_1},\ldots,\iota_{n_j}\iota'_{n_j}) - I_n)xc_2 + (\text{diag}((n_1-2)\iota_{n_1}\iota'_{n_1},\ldots,(n_j-2)\iota_{n_j}\iota'_{n_j}) + I_n)^2xc_3 = 0$ are $c_1 = c_2 = c_3 = 0$.

In the current case, the identified set is of considerable importance because for any $\theta \in \Theta_0$, network spillovers are constant and equal to network spillovers evaluated at the true parameter vector, $\varphi(x_n, \theta_0)$. In order to establish this result, define the subset $\Phi(\theta | y_n, x_n) \subseteq \Theta$ as the parameters such that spillovers are equal to $\varphi(x_n, \theta)$, that is,

$$\begin{aligned} \Phi(\theta | y_n, x_n) &= \{\theta^+ \in \Theta : \lambda_+ W_n^e(Q_n, \theta_c^+) = \lambda W_n^e(Q_n, \theta_c), \\ &\quad W_n^e(Q_n, \theta_c^+) x_n \beta_2^+ = W_n^e(Q_n, \theta_c) x_n \beta_2\}. \end{aligned} \tag{1.15}$$

The next Proposition states that θ_0 belongs to the identified set Θ_0 and that it is fully characterized by the subset of Θ such that spillovers are equal to $\varphi(x_n, \theta_0)$.

Proposition 1. *For any $\theta \in \Phi(\theta^0 | y_n, x_n)$, the network spillovers evaluated at θ are equal to network spillovers evaluated at θ_0, $\varphi(x_n, \theta) = \varphi(x_n, \theta_0)$. Also, this is the identified set, $\Phi(\theta^0 | y_n, x_n) = \Theta_0$.*

The objective then is to produce a sequence of sets such that: (*i*) in the limit, they are consistent estimates of Θ_0, in a sense that the Hausdorff set distance metric[28] d_h converges to zero in probability, and (*ii*) select a set $\hat{\Theta}_\alpha$ such that the coverage probability is asymptotically controlled, that is, $\lim_{n\to\infty} P\{\Theta_0 \subseteq \hat{\Theta}_\alpha\}) = 1 - \alpha$ for $\alpha \in [0, 1]$.

These objectives can be fulfilled with the definition of contour sets of the rescaled likelihood $L_n(\theta | y_n, x_n, Q_n) = -n^{-1}[\ln \mathcal{L}_n^e(\theta | y_n, x_n, Q_n) - \inf_{\theta\in\Theta} \ln \mathcal{L}_n^e(\theta | y_n, x_n, Q_n)]$ and $\hat{\Theta}(c_n) = \{\theta \in \Theta : L_n(\theta | y_n, x_n, Q_n) \le c_n\}$. The next Theorem proves that the estimator $\hat{\Theta} = \hat{\Theta}(0)$ is consistent for Θ_0, i.e, $d_h(\hat{\Theta}, \Theta_0) \xrightarrow{p} 0$. In fact, this result can be obtained if any sequence c_n such that $n^{-1} c_n \xrightarrow{p} 0$ is used to produce an alternative estimator $\hat{\Theta}(c_n)$. For the construction of a set that covers Θ_0 with probability α, it is necessary to select $c_n = \hat{c}_n(\alpha)$ such that $\hat{\Theta}(\hat{c}_n(\alpha))$ possesses the desired property.

Notice the event $\{\Theta_0 \subseteq \hat{\Theta}(c_n)\}$ is equivalent to the event $\{\sup_{\theta\in\Theta_0} L_n(\theta | y_n, x_n, Q_n) \le c_n\}$, and hence, in order to build coverage regions for the identified set Θ_0 with predetermined probability α, it suffices to input a $c_n = \hat{c}_n(\alpha)$ such that $\hat{c}_\alpha$ consistently estimates the α-quantile of the test statistic $\sup_{\theta\in\Theta_0} L_n(\theta | y_n, x_n, Q_n)$. That is, for any set $K \subseteq \Theta$, use

$$\hat{c}_n(\alpha) = \inf \left\{ \tilde{c} : P\left\{ \sup_{\theta \in K} L_n(\theta | y_n, x_n, Q_n) \le \tilde{c} \right\} \ge 1 - \alpha \right\}.$$

Given the probability is not known, I will use a bootstrap algorithm to produce usable estimates of $\hat{c}_n(\alpha)$. For the moment, assume $\hat{c}_n(\alpha)$ is known. The next Theorem shows asymptotic properties

[28] The Hausdorff set distance metric is defined

$$d_h(A, B) = \max \left\{ \sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A) \right\}$$

with $d(b, A) = \inf_{a\in A} \|b - a\|$ and $d_h(A, B) = \infty$ if A or B are empty.

of the estimated contour sets $\hat{\Theta}\left(c_n\right)$ for the various choices of c_n.

Theorem 3. *Let c_n be such that $n^{-1}c_n \xrightarrow{p} 0$. (1) Under Assumptions 1-5, if $\Theta_0 \neq \Theta$ and Θ compact, $\Theta_0 \subseteq \hat{\Theta}\left(c_n\right)$ with probability approaching one, $d_h(\hat{\Theta}\left(c_n\right), \Theta_0) = o_p\left(1\right)$ and $d_h(\hat{\Theta}\left(c_n\right), \Theta_0) = O_p(n^{-\frac{1}{2}})$. (2) For $c = \hat{c}_n\left(\alpha\right)$ consistent estimator of the α-quantile of $\sup_{\theta\in\Theta_0} L_n\left(\theta \middle| y_n, x_n, Q_n\right)$, $\lim_{n\to\infty} P\{\Theta_0 \subseteq \hat{\Theta}\left(\hat{c}_n\left(\alpha\right)\right)\}) = 1-\alpha$. (3) Given Proposition 1, the network spillover is point-identified. (4) Point-identification for β_{10} and σ_0^2 is obtained and $(\hat{\beta}_1, \hat{\sigma}^2) \xrightarrow{p} (\beta_{10}, \sigma_0^2)$.*

Obtaining confidence regions for known functions of the identified set is important at least in two circumstances. First, it provides confidence regions for the network spillovers, i.e., confidence regions for Φ_0, the image of Θ_0 under the known function $\varphi\left(x,\theta\right)$ for given $\theta \in \Theta_0$. Second, I will show it provides a framework for validation of network data, when it is available. I now develop these points.

Following Romano and Shaikh (2010), in general terms, let f be a known function with $f : \Theta \to \Upsilon$, with Υ_0^f being the image of Θ_0 under f, and also let $f^{-1}\left(\upsilon\right) = \{\upsilon \in \Upsilon : f\left(\theta\right) = \upsilon\}$. This suggests a modification of the inferential test statistic in the following way: note $\upsilon \in \Upsilon_0^f$ if, and only if, there exists some $\theta \in f^{-1}\left(\upsilon\right)$ subject to $Q_n\left(\theta\right) = 0$, which in turn implies that $\inf_{\theta\in f^{-1}(\upsilon)} Q_n\left(\theta\right) = 0$. As before, the objective is to construct a set $\hat{\Upsilon}_\alpha$ such that coverage probability is $1-\alpha$, i.e., $\lim_{n\to\infty} P\{\Upsilon_0 \subseteq \hat{\Upsilon}_\alpha\} = 1-\alpha$ and, in analogy to the previous case, this set can be defined by selecting $c_n^f\left(\alpha\right)$ such that the event $\{\Upsilon_0^f \subseteq \hat{\Upsilon}_\alpha\}$ is equivalent to $\{\sup_{\upsilon\in\Upsilon_0^f} \inf_{\theta\in f^{-1}(\upsilon)} L_n\left(\theta\right) \leq c_n^f\left(\alpha\right)\}$.

Again, if the α-quantiles of the test statistic $\sup_{\upsilon\in\Upsilon_0^f} \inf_{\theta\in f^{-1}(\upsilon)} L_n\left(\theta\right)$ were available, coverage region with asymptotically controlled error probability α would be obtained directly. Appendix 1.E details a bootstrap algorithm for obtaining consistent estimates $\hat{c}_n^f\left(\alpha\right)$ of $c_n^f\left(\alpha\right)$. For the moment, I now describe the two important applications of this procedure for the context of inference on the network spillovers and network effects.

Remark 2. *(Confidence region for network spillovers).* The procedure above can be applied directly replacing function f with known function $\varphi\left(x;\theta\right)$. In this case, because $\varphi\left(x_n;\theta\right)$ is a function from Θ to $\mathbb{R}^1$, and given Proposition 1 states the network spillovers are constant in the identified set, the image Υ_0^φ is a scalar in $\mathbb{R}$ and the confidence region is actually a confidence *interval*, a subset of $\mathbb{R}^1$. □

Remark 3. *(Testing for reported network connections).* Introduce reporting of network data with recourse to matrix Q_j, making $\{Q_j\}_{ik} = 1$ if individual i in group j reports a link with individual k in the same group, through which it is believed that i affects k. In this case, a reasonable network model is given by a collection of Bernoulli trials with probability link formation depending on link observed reports, that is, model (1.1) with Q_n as described above. In this setting, structural parameter δ_1 is the the estimated probability *given* observation of link reports, and δ_0 otherwise.

The null hypothesis of interest is $\mathcal{H}_0 : \delta_1 - \delta_0 = 0$, with alternative $\mathcal{H}_A : \delta_1 - \delta_0 \neq 0$. In the setting above, suffices to take $\tilde{f} : \Theta \to \mathbb{R}^1$ as $\tilde{f}(\theta) = \delta_1 - \delta_0$ and build appropriate confidence intervals. □

1.3.3 Pointwise identification when λ_0 is unknown using outcome dispersion

In the previous subsection, I showed that parameters of interest are identified up to a set and network spillovers are constant within the identified set. A theoretically feasible restriction to fully identify the model is to assume λ_0 is known: under certain conditions, Theorem 1 proves consistency. Nevertheless, this assumption is unlikely to be satisfied in practice, as λ_0 is rarely observed. In this Section, I increment the problem with one additional restriction which restores point identification, selecting a parameter in the identified set.

This restriction is derived from matching the observed to the model-implied variance of the group-average outcome. The intuition is straightforward. When social interactions are not present, sufficiently large group sizes implies that group averages should be relatively close to population averages conditional on observables. Introduction of social interactions affects dispersion in the following way. Since individuals mirror the choices of the others, outcomes within a group positively correlate. In other words, a positive shock to the group affects individuals not only through individual decision, but also through peer composition. As a consequence, average of group outcome increases to greater extent than in the counterfactual in which social interactions are irrelevant. A similar reasoning applies to a bad shock. It follows that average outcome across groups are more disperse relative to the case in which social interactions are irrelevant.

It has been observed elsewhere[29] that group outcomes are substantially dispersed across groups even when similar along observable characteristics. This anecdotal observation has been denoted as "excess variance" and used to provide identification when networks are known (Graham, 2008). Other papers have contributed to identification using covariance restrictions in the context of social interactions, such as in the survey paper by Blume et al. (2011, p. 872) and references therein.

Since network formation depends on a model described in Section 1.2, the dispersion across groups provides a restriction that includes link strength, probability of link formation and dependence on exogenous characteristics of the others. The relation is usually non-linear and I will show it is sufficient to provide identification. The main idea is that, accounting for variance originating from explanatory variables and the individual or group heterogeneity, the remaining variance can only be explained by social interactions and pattern of association therein. Define, from the outset, the within and between group variance,

$$V_{W,j}(y_n) \; = \; n_j^{-1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \qquad ; \qquad V_{B,j}(y_n) = (\bar{y}_j - \bar{y})^2$$

[29] Hanushek (1971), Rivkin et al. (2005), Glaeser et al. (1996).

where $\bar{y}_j = n_j^{-1}\sum_{i=1}^{n_j} y_{ij}$ and $\bar{y} = v^{-1}\sum_{j=1}^{v}\bar{y}_j$. It is useful to derive the expectation of these quantities in terms of the variance of outcomes as predicted by the model. Then, $\mathbb{E}V_{W,j}(y) = n_j^{-1}\sum_{i=1}^{n_j}\left[\mathbb{V}(y_j)\right]_{ii}$ and $\mathbb{E}V_{B,j}\left(y\right) = n_j^{-2}\iota_{n_j}'\mathbb{V}(y_j)\iota_{n_j}$. From the reduced-form of model (1.2), the covariance matrix of outcomes for group j is given by[30]

$$\begin{aligned}\mathbb{V}(y_j) &= \mathbb{E}(s_j x_j \beta_{10}\beta_{10}^{'}x_j' s_j' + 2s_j^{*}x_j\beta_{10}\beta_{20}^{'}x_j' s_j' + s_j^{*}x_j\beta_{20}\beta_{20}^{'}x_j' s_j^{*'}) \\ &\quad +\mathbb{E}((S_j^0)^{-1}\epsilon_j\epsilon_j'(S_j^0)^{-1'}) \qquad (1.16)\end{aligned}$$

for $s_j = (S_j^0)^{-1} - \mathbb{E}((S_j^0)^{-1})$ and $s_j^{*} = (S_j^0)^{-1}W_j^0 - \mathbb{E}((S_j^0)^{-1}W_j^0)$. In absence of networks, $s_j = I_{n_j}$ and $s_j^{*} = 0_{n_j\times n_j}$ and, therefore, outcome variance is increased when social interactions are considered. As pointed out above, in applications it is usually the case that the latter is larger than the former in the positive semi-definite sense although the reverse relation is theoretically possible for certain parameters. The distance between variances $V_{B,j}$ and $V_{W,j}$ and their theoretical expected counterparts as implied by the model, $\mathbb{E}V_{B,j}(y_n)$ and $\mathbb{E}V_{W,j}(y_n)$, is used to distinguish between competing parameters that belong to the identified set. Given $V_{B,j}$ and $V_{W,j}$ are observed from data, we only need to generate predictions from the model (1.16). Naturally, this strategy depends on the theoretical calculation of $\mathbb{V}(y_j)$, which are often difficult to evaluate analytically but straightforward to compute. I now introduce one particular example where identification is throughoutly proven only with between-variance of outcomes.

Example 3. *(Bernoulli network model).* In a simple setting where link formation is independent and equal to δ_1, I conduct a Series Expansion and take a first-order approximation. That is, $(S_j^0)^{-1} - \mathbb{E}(S_j^0)^{-1} = \lambda_0(W_j^0 - \mathbb{E}W_j^0) + \cdots$ which is approximately $\lambda_0(W_j^0 - \mathbb{E}W_j^0)$ as remaining terms decay in exponential rates. Using independence of the Bernoulli trials that generate links, equation (1.16) simplifies to

$$\mathbb{V}\left\{y_j\right\} = \text{diag}\left(\mathbb{V}\left\{W_j\right\}\left(\lambda^2\text{diag}\left(x_j^{11}\right) + 2\lambda\text{diag}\left(x_j^{12}\right) + \text{diag}\left(x_j^{22}\right) + \lambda^2\sigma^2\iota_{n_j}\right)\right) + \sigma^2 I_{n_j} \qquad (1.17)$$

where $\mathbb{V}\{W_j^0\}$ is the variance of W_j^0, $x_j^{11} = \text{diag}(x_j\beta_{10}\beta_{10}^{'}x_j')$, $x_j^{12} = \text{diag}(x_j\beta_{10}\beta_{20}^{'}x_j')$ and $x_j^{22} = \text{diag}(x_j\beta_{20}\beta_{20}^{'}x_j')$ extracts the main diagonal of a matrix into a column vector or vice-versa, as appropriate. Off-diagonal terms are zero. In the Bernoulli model without dependence on exogenous characteristics, $\mathbb{V}\{W_j\} = \delta_1(1-\delta_1)\iota_{n_j}\iota_{n_j}'$ and, in this case,

$$\begin{aligned}\mathbb{V}\left\{y_j\right\} &= \text{diag}\left(\delta_1(1-\delta_1)\iota_{n_j}\iota_{n_j}'\left(\lambda^2\text{diag}\left(x_j^{11}\right) + 2\lambda\text{diag}\left(x_j^{12}\right) + \text{diag}\left(x_j^{22}\right) + \lambda^2\sigma^2\iota_{n_j}\right)\right) \\ &\quad +\sigma^2 I_{n_j} \\ &= \delta_1(1-\delta_1)\left(\lambda^2\iota_{n_j}'\text{diag}\left(x_j^{11}\right) + 2\lambda\iota_{n_j}'\text{diag}\left(x_j^{12}\right) + \text{diag}\left(x_j^{22}\right) + n_j\lambda^2\sigma^2\right)I_{n_j} + \sigma^2 I_{n_j}\end{aligned}$$

[30] For the panel data with fixed effects, proceed as described in Subsection 1.3.4. In this Section, for simplicity I assume $\rho_0 = 0$. This is not substantial as all results are maintained in the more general case.

and the between-group variance is

$$V_{B,j} = n_j^{-1}\delta_1\left(1-\delta_1\right)\left(\lambda^2\iota'_{n_j}\text{diag}\left(x_j^{11}\right) + 2\lambda\iota'_{n_j}\text{diag}\left(x_j^{12}\right) + \iota'_{n_j}\text{diag}\left(x_j^{22}\right) + n_j\lambda^2\sigma^2\right) + n_j^{-1}\sigma^2.$$

This provides the additional restriction required for the identification of θ. Formally, the Jacobian of the matrix formed by stacking restrictions, including those originating from reduced-form estimation, has full rank, and then Theorem 6 of Rothenberg (1971, p. 585) is applied. Proofs can be found in Appendix 1.D.8. □

The approach suggests a Genaralized Method of Moments estimator with moment conditions given by $q_{1,j}(y_j, x_j, \theta) = \mathbb{E}V_{B,j}(y_j, x_j, \theta) - V_{B,j}(y_j, x_j, \theta)$ and $q_{2,j}(y_j, x_j, \theta) = \mathbb{E}V_{W,j}(y_j, x_j, \theta) - V_{W,j}(y_j, x_j, \theta)$ minimized on the estimated set $\hat{\Theta}$,

$$\hat{\theta} \quad = \quad \arg\min_{\theta\in\hat{\Theta}}\left(\sum_{j=1}^{v} q_j(y_j, x_j, \theta)\right)'\Omega\left(\sum_{j=1}^{v} q_j(y_j, x_j, \theta)\right)$$

where $q_j(y_j, x_j, \theta) = [q_{1,j}(y_j, x_j, \theta), q_{2,j}(y_j, x_j, \theta)]'$ and 2×2 weight matrix Ω. It is equally possible to estimate the same GMM problem on the unrestricted parameter set Θ and introduce score conditions given by the solution of the pseudo-likelihood and assigning arbitrarily large weights to them. Unfortunately, the expected variances are generally difficult to compute. Even in simple examples, one has to rely on very crude approximations of to obtain the expectation of $(S_j^0)^{-1}$. Next, I outline a general procedure for simulating the moment conditions (Gouriéroux and Monfort, 1997) and prove the desired asymptotic properties, including consistency for $\hat{\theta}$. The final estimator is the solution to

$$\hat{\theta} \quad = \quad \arg\min_{\theta\in\hat{\Theta}}\left(\sum_{j=1}^{v} S^{-1}\sum_{s=1}^{S} q_{s,j}(y_j, x_j, \theta)\right)'\Omega\left(\sum_{j=1}^{v} S^{-1}\sum_{s=1}^{S} q_{s,j}(y_j, x_j, \theta)\right) \tag{1.18}$$

where $q_{s,j}(y_j, x_j, \theta) = [V_{B,j}(y_j, x_j, \theta) - V_{B,j}(\hat{y}_{j,s}, x_j, \theta); V_{W,j}(y_j, x_j, \theta) - V_{B,j}(\hat{y}_{j,s}, x_j, \theta)]$ with $\hat{y}_{j,s} = (S_j^s)^{-1}(x_j\beta_1 + W_j^s x_j\beta_2 + e_j^s)$, $S_j^s = (I_{n_j} - \lambda W_j^s)^{-1}$, W_j^s sampled from the distribution of the network-generating model with parameters θ and ϵ_j^s is sampled from a normal distribution with variance σ^2. If the simulator is unbiased, one can expect that $S^{-1}\sum_{s=1}^{S} q_{s,j}(y_j)\xrightarrow{p}q_j(y_j)$ as $S \longrightarrow \infty$ and asymptotic properties follow. In addition, given $\hat{\Theta}$ is $\sqrt{n}$-consistent for Θ_0 on the Hausdorff metric, one might expect minimizing on the set $\hat{\Theta}$ is asymptotically equivalent to minimizing on the identified set Θ_0.

Theorem 4. *If parameters are identified, (i) estimator (1.18), minimized on the estimated set $\hat{\Theta}$, as defined in Section 1.3.2, is consistent for θ_0, $\hat{\theta}\xrightarrow{p}\theta_0$, and (ii) if $S \to \infty$ sufficiently fast, $\sqrt{n}(\hat{\theta} - \theta_0)\xrightarrow{d}N(0, \Sigma^*)$, where $\Sigma^* = (G'(\Omega^*)^{-1}G)^{-1}$, $G = \mathbb{E}\nabla_\theta q_n(y_n, x_n, \theta_0)$ and $\Omega^* = (\mathbb{E}q_n(y_n, x_n, \theta_0)q_n(y_n, x_n, \theta_0)')^{-1}$ with optimal choice of weight matrix Ω^* and $q_n(y_n, x_n, \theta_0) =$*

$\sum_{j=1}^{v} q_j(y_j, x_j, \theta_0)$.

1.3.4 Fixed and Time Effects

In this subsection, I propose a data transformation to eliminate fixed effects, along with corresponding treatment of the variance-covariance matrix induced by this transformation. This is of considerable importance given that explanatory variables x_j may correlate with unobserved components that vary at the group or individual-level, for example an unobserved "good teacher" shock in a classroom or unobserved peer characteristic that may affect learning.

Bramoullé et al. (2009) and Lee (2007) propose eliminating fixed effects subtracting average of connected peers (local differencing) or average of all individuals in a group in a given time period, regardless of connection status (global differencing). Neither approach is available in the current setting: by definition of the problem in the current paper, networks are unobserved, and hence local differencing is not defined. Yet, global differencing cannot be applied in the absence of row-sum normalization. Group fixed effects with the row-sum normalization condition implies that all individuals are affected to the same degree by network spillovers originating for them. When the row-sum normalization condition is removed, heterogeneity of individual responses to fixed effects through networks implies that no data manipulation possibly removes them in the absence of network observation.

For this purpose, I introduce time dimension and time-difference data in order to remove fixed effects. This approach also has the advantage of allowing for individual fixed effects. Let the spatio-temporal model be, for $t = 1, \ldots, T$,

$$y_{jt} \quad = \quad \lambda W_j y_{jt} + x_{jt}\beta_1 + W_j x_{jt}\beta_2 + \alpha_j + \gamma_t + v_{jt} \tag{1.19}$$

where $v_{jt} = \rho M_j v_{jt} + \epsilon_{jt}$. Here, α_j represents a $n_j \times 1$ vector of individual or group fixed effects, or both. In the classical fixed effects case, α_j is allowed to vary over individuals; the group effect case is when $\alpha_j = \dot{\alpha}_j \iota_{n_j}$, with constant scalar $\dot{\alpha}_j$ throughout individuals in group j and does not vary over time. Notation is left sufficiently general to incorporate both cases. Group effects, in Manski's (1993) terminology, are denominated *correlated effects.*

Define $\dot{y}_{jt} = y_{jt} - \bar{y}_{j\cdot}$, $\bar{y}_{j\cdot} = T^{-1}\sum_{t=1}^{T} y_{jt}$, $\dot{x}_{jt} = x_{jt} - \bar{x}_{j\cdot}$, $\bar{x}_{j\cdot} = T^{-1}\sum_{t=1}^{T} x_{jt}$, $\bar{\gamma}_t = \gamma_t - \dot{\gamma}_\cdot$ and $\bar{\gamma}_\cdot = T^{-1}\sum_{t=1}^{T}\gamma_t$. The transformed model is

$$\dot{y}_{jt} \quad = \quad \lambda W_j \dot{y}_{jt} + \dot{x}_{jt}\beta_1 + W_j \dot{x}_{jt}\beta_2 + \dot{\gamma}_t + \dot{v}_{jt}. \tag{1.20}$$

which is a consequence of (1.19) because the time-differenced $W_j y_{jt}$ is equal to $W_j \dot{y}_{jt}$, and similarly for the $W_j \dot{x}_{jt}\beta$, under the hypothesis of invariance of the network over time. Explicitly, the k-th line of the time-differenced $W_j y_{jt}$ is

$$\sum_{i=1}^{n_j} \{W_j\}_{ki} \{y_{jt}\}_i - T^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n_j} \{W_j\}_{ki} \{y_{jt}\}_i \quad = \quad \sum_{i=1}^{n_j} \{W_j\}_{ki} \left(\{y_{jt}\}_i - \{\bar{y}_{j\cdot}\}_i\right) \qquad (1.21)$$

Letting $\dot{y}_{nT} = (\dot{y}'_{11}, \dots \dot{y}'_{1T}, \dots, \dot{y}'_{v1}, \dots, \dot{y}'_{vT})'$ and $\dot{x}_{nT} = (\dot{x}'_{11}, \dots, \dot{x}'_{1T}, \dots, \dot{x}'_{v1}, \dots, \dot{x}'_{vT})'$, and similarly for $\dot{v}$ and $\dot{\gamma}$, the full model can be rewritten $\dot{y}_{nT} = \lambda W_{nT} \dot{y}_{nT} + \dot{x}_{nT}\beta_1 + W_{nT}\dot{x}_{nT}\beta_2 + \dot{\gamma}_T + \dot{v}_{nT}$, where $W_{nT} = \operatorname{diag}\{I_T \otimes W_1, \dots, I_T \otimes W_v\}$. Remaining matrices are defined in a similar way and carry the subscript nT for clarity. The variance-covariance matrix of $\dot{v}_{nT}$ is $\mathbb{E}\left(\dot{v}_{nT}\dot{v}'_{nT}\right) = \sigma_0^2 (R^0_{nT})^{-1} \dot{\Sigma}_{nT} (R^{0\prime}_{nT})^{-1}$, where $\dot{\Sigma}_{nT} = \sigma_0^2 I_{nT} - \sigma_0^2 T^{-1} \cdot \operatorname{diag}(\iota_T \iota'_T \otimes I_{n_1}, \dots, \iota_T \iota'_T \otimes I_{n_v})$. This more complicated form recognizes the dependence in $\dot{v}_{nT}$ introduced by time-average subtraction. Finally, likelihood (1.12) is adjusted to

$$\begin{aligned} \ln \mathcal{L}^e_{nT}\left(\theta \middle| y_{nT}, x_{nT}, Q_{nT}\right) \quad = \quad & -\frac{nT}{2} \ln\left(2\pi\sigma^2\right) + \ln\left|S^e_{nT}\left(Q_{nT}, \theta\right)\right| + \ln\left|R^e_{nT}\left(Q_{nT}, \theta\right)\right| \\ & -\frac{1}{2\sigma^2} \epsilon^e_{nT}\left(Q_{nT}, \theta\right)' \dot{\Sigma}_{nT} \epsilon^e_{nT}\left(Q_{nT}, \theta\right) \qquad (1.22) \end{aligned}$$

where $\epsilon^e_{nT}\left(Q_{nT}, \theta\right) = R^e_{nT}\left(Q_{nT}, \theta\right) \left(\dot{y}_{nT} - \lambda W^e_{nT}\left(Q_{nT}, \theta\right) \dot{y}_{nT} - \dot{x}_{nT}\beta_1 - W^e_{nT}\left(Q_{nT}, \theta\right) \dot{x}_{nT}\beta_2 - \dot{\gamma}\right) = R^e_{nT}\left(Q_{nT}, \theta\right) \left(S^e_{nT}\left(Q_{nT}, \theta\right) \dot{y}_{nT} - \dot{Z}^e_{nT}\left(Q_{nT}, \theta\right) \tilde{\beta}\right)$ and $\dot{Z}^e_{nT}\left(Q_{nT}, \theta\right)$ now also incorporate time effects: $\dot{Z}^e_{jt}\left(Q_j, \theta\right) = (x_{jt}, W^e_j(Q_j, \theta) x_{jt}, \mathbf{1}\left\{t = 1\right\} \iota_{n_j}, \dots, \mathbf{1}\left\{t = T\right\} \iota_{n_j})$ and $\tilde{\beta} = (\beta', \gamma_1, \dots, \gamma_T)'$. In fact, any variable not subject to exogenous effects can be incorporated by adding columns to $\dot{Z}^e_{jt}(Q_{nT}, \theta)$. The concentrators are now

$$\begin{aligned} \hat{\tilde{\beta}}\left(Q_{nT}, \theta\right) \quad &= \quad (Z^{e\prime}_{nT}\left(Q_{nT}, \theta\right) \ddot{\Sigma}_{nT} Z^e_{nT}\left(Q_{nT}, \theta\right))^{-1} Z^{e\prime}_{nT}\left(Q_{nT}, \theta\right) \ddot{\Sigma}_{nT} S^e_{nT}\left(Q_{nT}, \theta\right) y_{nT} \\ \hat{\tilde{\sigma}}^2\left(Q_{nT}, \theta\right) \quad &= \quad \frac{1}{n} (S^e_{nT}\left(Q_{nT}, \theta\right) y - Z^e_{nT}\left(Q_{nT}, \theta\right) \hat{\tilde{\beta}}) \ddot{\Sigma}_{nT} (S^e_{nT}\left(Q_{nT}, \theta\right) y_{nT} - Z^e_{nT}\left(Q_{nT}, \theta\right) \hat{\tilde{\beta}}) \end{aligned}$$

where $\ddot{\Sigma}_{nT} = R^{e\prime}_{nT}\left(Q_{nT}, \theta\right) \dot{\Sigma}_{nT} R^e_{nT}\left(Q_{nT}, \theta\right)$. Concentrated likelihood (1.12) remains unchanged with $\hat{\sigma}^2(Q_{nT}, \theta)$ substituted for $\hat{\tilde{\sigma}}^2(Q_{nT}, \theta)$. Preceding theorems are applied with obvious modifications.

1.4 Simulations and Implementation

In this Section, I conduct a simulation exercise to demonstrate the small-sample empirical properties of the estimator. MATLAB codes are available upon request[31]. The algorithms are presented in Appendix 1.E.

Four simulations sets are performed: purely cross-sectional model (1.2), under $T = 1$ and absence of fixed effects; the panel (1.5) with $T = 5$ and fixed effects but no time effects; with time effects but no fixed effects; and, finally, with both time and fixed effects. Sample sizes are

[31] STATA codes will soon be available.

$(n = 25, v = 250)$, $(n = 100, v = 250)$, $(n = 25, v = 1000)$ and $(n = 100, v = 1000)$. Simulations with smaller n and v can be found in Appendix 1.F.1. In every case, I allow for heterogeneity in group sizes, by sampling n_j from a standard normal distribution with mean n and standard error 5, rounded to the nearest integer.

True parameters are $\theta_s = (0.0125, 1, 1, 0.04, 0.04, 1)'$ and $\theta_g = (0.75, 0.30)'$. In a row-normalized model and with this combination of parameters, $\lambda = 0.0125$ would roughly correspond to an autoregressive parameter of 0.16 for $n = 25$, 0.32 for $n = 50$ and 0.65 for $n = 75$. The probability of common exogenous characteristic is 50%. That is, $P\left\{\{Q_j\}_{ik} = 1\right\} = 0.5$ and zero otherwise. Finally, x and ϵ are drawn from a normal distribution with mean 0 and variance 1. The simulation is composed of 500 repetitions.[32] The average of the estimated standard errors, following the procedure outlined in 1.3.2, is shown in parentheses, while standard deviations of the point estimates computed across replications is shown in square brackets. Simulations are conducted in the absence of information on λ_0.

Simulated results are largely satisfactory in all cases. Convergence to spatial parameters and those that underpin the randomness in networks, is observed, even with small $n = 25$ and $v = 25$. Moreover, the network spillover is correctly estimated. In Table 1.F.3 of Appendix 1.F.1, I show that OLS estimates would be inconsistent at averages $\hat{\beta}_{OLS} = 1.0670$ for $n = 25$ and $\hat{\beta}_{OLS} = 1.1127$ for $n = 50$. This bias is eliminated with the proposed method. Introduction of time dimension and fixed effects do not change the results, despite the fact that estimates of σ^2 now take into account that cross-section and time variation has been eliminated as the consequence of data transformation (Subsection 1.3.4). For the case without time and fixed effects, estimates of disturbance variance is, in most cases, larger than the true value, but this is expected as it captures the misspecification component due to the fact that the observed model is considered under expected networks – naturally different from the true networks. It is also noteworthy that estimated standard errors are very close in most cases to standard errors of point estimates across iterations, demonstrating good performance of the hypothesis testing procedure.

I also show results on three additional cases in Appendix 1.F.1. Tables 1.F.4 and 1.F.5 shows the performance of the estimator with very low sample sizes. It shows that even with small samples up to $n = 25$ and $v = 50$, estimates are acceptably close to true parameters and confidence intervals are correctly estimated. Then, I introduce across-group connections by randomly assigning value 1 to off-block elements of matrix W_j^0 with probability δ_A. Although not explicitly incorporated in theory, it is shown that a small amount of violation from the isolated-group assumption does not deteriorate empirical performance of the estimator. Performance was good up to $\delta_A = 0.05$ or $\delta_A = 0.075$. Finally I conduct estimation and hypothesis testing when λ_0 is known but misspecified, shown in Table 1.F.7 of Appendix 1.F.1. I assume incorrectly $\lambda = 0.0250$, twice the true value.

[32] Using a MacBook Pro 13", Core i7, Early 2013 specification, the average computing time was <1 minute for $(n = 25, v = 250)$ and around 5 minutes for $(n = 100, v = 1000)$.

As expected, I observe halved $\hat{\delta}_1$ and $\hat{\delta}_0$ and $\hat{\beta}_2$ estimated twice the true parameter. Associated standard errors followed the same expected pattern.

I also implement the multivariate network model described in example 4 of Subsection 1.B, where probability of link formation is described by

$$P\left\{\{W_j\}_{ik} = 1|Q_j\right\} \quad = \quad Q^1_{jik}\delta_1 + Q^0_{jik}\delta_0$$

where Q^1_{jik} is the distance between individuals i and k who belong to group j, and respectively for Q^0_{jik}. Distances are sampled independently from a uniform distribution between -2.5 and 2.5, and probabilities are cut such they do not exceed 1 or fall below 0. True values are $\delta_1 = 0.25$ and $\delta_0 = 0.50$, and remaining parameters remain unchanged from previous setting. Results are shown in Table 1.F.8 of Appendix 1.F.1 and are also satisfactory with convergence to true parameters and standard errors also being observed at small values of n and v. Estimation of λ using second moments is also satisfactory.

Table 1.1: Simulations.

	$T = 1.$				$T = 5$, fixed effects.			
	(1)	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*
n	25	100	25	100	25	100	25	100
v	250	250	1000	1000	250	250	1000	1000
$\hat{\lambda}$	0.0117	0.0122	0.0126	0.0119	0.0119	0.0120	0.0118	0.0119
	(0.002)	(0.006)	(0.002)	(0.001)	(0.004)	(0.001)	(0.001)	(0.000)
	[0.002]	[0.005]	[0.002]	[0.001]	[0.003]	[0.001]	[0.000]	[0.000]
$\hat{\beta}_1$	0.9999	1.0001	0.9998	0.9999	1.0006	1.0000	1.0002	0.9999
	(0.009)	(0.005)	(0.005)	(0.003)	(0.004)	(0.003)	(0.002)	(0.001)
	[0.009]	[0.005]	[0.005]	[0.003]	[0.004]	[0.002]	[0.002]	[0.001]
$\hat{\beta}_2$	0.0461	0.0402	0.0400	0.0401	0.0428	0.0398	0.0408	0.0400
	(0.018)	(0.003)	(0.008)	(0.001)	(0.008)	(0.001)	(0.003)	(0.001)
	[0.018]	[0.003]	[0.006]	[0.002]	[0.007]	[0.001]	[0.003]	[0.001]
$\hat{\delta}_1$	0.7166	0.7497	0.7605	0.7492	0.7247	0.7510	0.7389	0.7499
	(0.164)	(0.024)	(0.097)	(0.012)	(0.085)	(0.012)	(0.031)	(0.004)
	[0.162]	[0.025]	[0.081]	[0.013]	[0.073]	[0.011]	[0.036]	[0.006]
$\hat{\delta}_0$	0.2892	0.2995	0.3015	0.2992	0.2918	0.3010	0.2966	0.3002
	(0.062)	(0.007)	(0.031)	(0.004)	(0.032)	(0.003)	(0.015)	(0.002)
	[0.063]	[0.007]	[0.030]	[0.004]	[0.027]	[0.003]	[0.014]	[0.002]
$\hat{\sigma}^2$	1.0571	1.2199	1.0547	1.2228	0.8421	0.9778	0.8451	0.9774
	(0.018)	(0.003)	(0.008)	(0.002)	(0.008)	(0.001)	(0.003)	(0.001)
	[0.019]	[0.011]	[0.009]	[0.006]	[0.007]	[0.004]	[0.003]	[0.002]
$\varphi(x, \hat{\theta})$	-0.0007	0.0137	0.0007	-0.0099	0.0008	-0.0096	0.0005	0.0020
	(0.023)	(0.092)	(0.008)	(0.048)	(0.009)	(0.050)	(0.006)	(0.020)
	[0.006]	[0.007]	[0.001]	[0.002]	[0.001]	[0.001]	[0.000]	[0.000]

Note: True parameters are $\beta_1 = 1$, $\beta_2 = 0.04$, $\delta_1 = 0.75$, $\delta_0 = 0.30$, $\sigma^2 = 1$ and $\varphi(x, \theta) = 0$. $\lambda = 0.0125$.

Standard error in round brackets. Standard error across iterations in square brackets.

Table 1.2: Simulations.

	$T=5$, time effects.				$T=5$, time and fixed effects.			
	(1)	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*
n	25	100	25	100	25	100	25	100
v	250	250	1000	1000	250	250	1000	1000
$\hat{\lambda}$	0.0121	0.0119	0.0123	0.0118	0.0132	0.0121	0.0120	0.0117
	(0.006)	(0.005)	(0.002)	(0.000)	(0.002)	(0.005)	(0.006)	(0.001)
	[0.005]	[0.005]	[0.001]	[0.000]	[0.002]	[0.005]	[0.005]	[0.001]
$\hat{\beta}_1$	1.0001	0.9996	0.9995	0.9999	1.0005	0.9998	1.0000	1.0002
	(0.004)	(0.002)	(0.002)	(0.001)	(0.005)	(0.002)	(0.002)	(0.001)
	[0.004]	[0.002]	[0.002]	[0.001]	[0.004]	[0.002]	[0.002]	[0.001]
$\hat{\beta}_2$	0.0413	0.0400	0.0396	0.0402	0.0411	0.0399	0.0403	0.0402
	(0.007)	(0.001)	(0.003)	(0.001)	(0.007)	(0.002)	(0.004)	(0.001)
	[0.006]	[0.001]	[0.003]	[0.001]	[0.006]	[0.001]	[0.003]	[0.001]
$\hat{\delta}_1$	0.7387	0.7508	0.7588	0.7486	0.7445	0.7524	0.7479	0.7483
	(0.070)	(0.009)	(0.034)	(0.007)	(0.074)	(0.013)	(0.046)	(0.006)
	[0.071]	[0.011]	[0.036]	[0.006]	[0.071]	[0.011]	[0.047]	[0.006]
$\hat{\delta}_0$	0.2958	0.2992	0.3018	0.3003	0.2980	0.2982	0.2969	0.3008
	(0.028)	(0.004)	(0.014)	(0.002)	(0.029)	(0.005)	(0.017)	(0.002)
	[0.027]	[0.003]	[0.013]	[0.002]	[0.027]	[0.003]	[0.015]	[0.002]
$\hat{\sigma}^2$	0.0114	0.0195	0.0110	0.0603	0.0423	0.0029	0.0002	0.0019
	(0.007)	(0.001)	(0.003)	(0.001)	(0.007)	(0.002)	(0.005)	(0.001)
	[0.014]	[0.008]	[0.007]	[0.001]	[0.012]	[0.003]	[0.007]	[0.001]
$\varphi(x,\hat{\theta})$	0.0005	-0.0050	0.0011	0.0036	0.0012	0.0082	-0.0002	0.0003
	(0.009)	(0.039)	(0.005)	(0.016)	(0.008)	(0.035)	(0.005)	(0.021)
	[0.001]	[0.001]	[0.000]	[0.001]	[0.001]	[0.001]	[0.001]	[0.000]

Note: True parameters are $\beta_1 = 1$, $\beta_2 = 0.04$, $\delta_1 = 0.75$, $\delta_0 = 0.30$, $\sigma^2 = 1$ and $\varphi(x,\theta) = 0$. $\lambda = 0.0125$.

Standard error in round brackets. Standard error across iterations in square brackets.

1.5 Application

Empirical research has led to substantial interest in evaluating the effects of randomized policies on targeted individuals. Much less progress has been made on evaluating the spillovers related to those policies, possibly because of problems associated with observing and defining interactions among people. The method developed in the present paper provides a comprehensive evaluation of programs when networks are unknown or unreliable, and information on a large number of groups is available and network effects are suspected.

The importance of assessing spillovers is further highlighted when a large proportion of individuals are subject to a shock. This effect raises the possibility that spillovers or externalities play a key role in overall program results (Angelucci et al., 2010). As an example of this setting, I analyze the effect of a randomized intervention in which a large proportion of individuals was simultaneously targeted. This example also illustrates that randomization in treatment variables can be used to estimate network effects, as opposed to randomization in the group formation (Sacerdote, 2001).

I employ data for a large-scale randomized intervention, which provided compelling evidence that occupational choice of the world's poor is determined by a lack of capital and skills (Bandiera et al., 2013). The intervention consisted of the assignment of livestock and skills training, both relevant in terms of the outlay (at approximately USD $140) and duration (training lasted for two years). The authors found significant changes in the occupational choices of the poor, who moved from wage jobs toward self-employment associated with livestock rearing. The program was instituted in 1409 communities, which consisted of clusters of 84 households on average. In each community, households belonging to the bottom quintile of the wealth distribution were identified, and all were eligible for treatment, with certain exceptions. In total, 7953 beneficiaries were surveyed, and all eligible households in the randomly selected communities were treated.

The baseline results comparing the treatment group in selected villages against the treatment group in non-selected villages indicate a dramatic change in the occupational status of targeted households. Four years after treatment, poor women dedicated 92% additional hours to self-employment running their livestock-rearing businesses and moved away from wage hours that were frequently insecure and temporary. This lasting change in occupational status was also associated with higher earnings, higher per capita expenditure, better general wellbeing and higher measures of life satisfaction. After treatment, poor households were classified between near-poor and middle class according to a host of economic indicators.

With recourse to the estimation method developed in this paper, and without network data, I supplement these results with several network-dependent findings. I show that specific program effects are not contained to targeted individuals. Network spillovers affect food expenditure and food security at magnitude around half of the original treatment, but are either insignificant or small determinants of occupational choice and livestock assets. I also shed light on the underlying network structural mechanisms that give rise to these externalities. By separately identifying endogenous and exogenous effects, I am able to estimate the marginal effects of a connection to treated households. I find that the occupational choice of peers of the treated households move in an opposite direction to the treated households: a marginal connection to treated households reduces self-working hours, increases wage hours and decreases livestock value. The magnitudes of the effects are such that exogenous effects counteract 25-30% of the reduction in treated households' wage hours.[33] However, connections to the treated households strongly increase food expenditure and food security. These results are consistent with the interpretation that the treated households gained comparative advantage in livestock rearing, which partially changed the occupational choices of their peers. Overall, network effects are shown to form an integral component of the program evaluation.

There is wide consensus that capital, opportunities, income, information and choices affect the outcomes of peers (Jackson, 2010). In fact, the opportunities of others have been regarded as a form of social capital (Glaeser et al., 2002). In this way, a shock to one's peers can be interpreted in the same fashion as a shock to one's self, and the example described here provides evidence of this mechanism. Now, I turn to a description of the program, followed by the identification strategy and the results.

1.5.1 Program Description

Selection of targeted individuals proceeded in stages. In collaboration with BRAC, a local non-profit organization, the most vulnerable districts were selected based on food-security measures, as described by the World Food Program. Second, BRAC employees selected the poorest communities within each district. Finally, within each community, a combination of a participatory rural appraisal exercise and survey data were used to allocate households to one of five wealth bins. Households belonging to the poorest wealth bins were selected as a potential beneficiary if other eligibility criteria were met, such as not participating as microfinance borrowers and owning no productive assets. Randomization was conducted at the local BRAC branch level, among its 40 offices in Bangladesh, and stratified at the subdistrict level to ensure balance between treated and control groups. Within each subdistrict, one branch was randomly allocated to treatment and another to the control group, and asset transfer was conducted for all selected individuals within the communities covered by the treated BRAC branches. Consequently, a substantial fraction of the community population was treated, raising the possibility that aggregate community-level

[33] This is the ratio between the increase of wage hours due to exogenous effects and the direct effect of reduction of wage hours. These are numbers are averages across all individuals in treated villages, considering the number of treated households in each village and the network parameters which affect the number of expected connections. In this case, endogenous effects counteract exogenous effects which combined produce spillovers of smaller magnitudes. See also Subsection 1.5.2.

effects are substantially larger than the sum of isolated individual treatment effects, including, for example, as a consequence of learning, insurance and informal skills reinforcement from neighbors, who in turn may or may not be in the treatment group themselves. If eligible and selected through the randomization process, households received a transfer of live animals (valued at approximately USD $140) and subsequent skills training for two years that were specifically designed for the chosen asset. Program beneficiaries could select among cows, goats or chickens that added up to the same face value; the large majority chose cows. Participants were required to keep possession of the asset for a minimum of two years, but in practice there were no sanctions in case of noncompliance. All potential beneficiaries of the program and a sample of households across the village wealth distribution were surveyed just before the intervention in 2007 and in two additional waves in 2009 and 2011. The comprehensive survey consisted of household members' sociodemographic characteristics, business assets and activities, land holdings and transfers, financial assets and liabilities, non-business assets, homestead ownership status and improvements, women's empowerment and vulnerability (such as earnings seasonality and food security), and a health module. Network self-reported links were registered when applicable, and data included family outside the household, their business activities, land transfers (through inheritance, mortgage, rent, share, received as dowry or gift, bought or sold), business asset transfers (same possibilities as above), finance links (loans, outstanding lending or transfers) and letting of house ownerships. The questionnaire was applied to all selected and a sample of non-selected households in both treatment and control groups.

1.5.2 Evaluation and Identification Strategies

Treatment effects on the treated could be evaluated comparing the change before and after treatment in the outcomes of selected households who live in a treated village against similar changes in the outcomes of selected households who live in non-treated villages. However, this approach would be unsuitable for estimating the network effects due to two reasons.

First, exclusion of non-treated households in treated and control villages prevents wider evaluation of policy for those groups. Second, as I showed in Subsection 1.2.1, the outcome of the differences-in-differences estimator is unclear when network effects are present because it may or may not capture network spillovers (φ). The extent to which the spillovers are estimated depends on the degree of reciprocation in the network, which is unobserved. When reciprocation is not present or interaction groups are large enough, Example 1 shows that the estimator is consistent for the individual elasticity in the counterfactual in which households are unconnected (β_{10}). On the other hand, separately estimating network-independent β_{10} from network-dependent φ is also important when the researcher desires to evaluate the policy impact in a setting where networks might considerably differ.

To tackle these issues, I consider a triple differences-in-differences with all households in treated and non-treated villages regardless of selection status. Momentarily ignoring network effects, one could specify a double differences-in-differences which would compare changes in outcomes of the selected households before and after treatment against similar changes in outcomes of the non-selected households. However, this strategy would not be sufficient because randomization was conducted at the village level: selection of potential beneficiaries within the villages was determined according to wealth at the baseline. I take two remedial actions. I introduce household fixed effects and I use the control villages to account for different trends in absence of treatment. The third difference eliminates the change before and after treatment in the outcomes of selected households who live in a non-treated village against similar changes in outcomes of non-selected households who also live in non-treated villages.

The final model is then a triple differences-in-differences with household fixed effects. The identification assumption is that trends as observed in the non-treated villages are a good counterfactuals for trends in treated villages. I denote $S_{ij} = 1$ if individual i of village j was selected as a potential beneficiary of the program and $T_{ij} = 1$ if village j was randomly selected for treatment. The model without networks is

$$\begin{aligned} y_{ijt} &= \sum_{s=2}^{3} \beta_{1s} S_{ij} T_{ij} \mathbf{1}\{s=t\} + \sum_{s=2}^{3} \eta_{1s} S_{ij} \mathbf{1}\{s=t\} \\ &\quad + \sum_{s=2}^{3} \eta_{2s} T_{ij} \mathbf{1}\{s=t\} + \gamma_t + \alpha_{ij} + \epsilon_{ijt} \end{aligned} \tag{1.23}$$

where y_{ijt} represents the outcome for individual i in village j at time t, $s = 2$ and 3 are the second and third survey wave (two and four years after treatment, respectively), α_{ij} is a fixed effect at the individual level, γ_t is a full set of time effects, $\mathbf{1}\{\cdot\}$ is an indicator function, and ϵ_{ijt} is the disturbance term, clustered at the village level. The program impact on the treated in the counterfactual in which households are unconnected are β_{12} and β_{13}.

I next introduce network spillovers, which take the form of two additional network-dependent terms attached to equation (1.23). Identification in the network setting follows after identification of the treatment effects on the treated, as introduced above, with added assumptions on variability of group sizes and moment condition based on outcome dispersion, as explained in Section 1.3. The full model in vector notation is

$$\begin{aligned} y_{jt} &= \lambda W_j^0 y_{jt} + \sum_{s=2}^{3} \beta_{1s} ST_j \mathbf{1}\{s=t\} + \sum_{s=2}^{3} W_j^0 ST_j \mathbf{1}\{s=t\} \beta_{2s} + \\ &\quad + \sum_{s=2}^{3} \eta_{1s} S_j \mathbf{1}\{s=t\} + \sum_{s=2}^{3} \eta_{2s} T_j \mathbf{1}\{s=t\} + \gamma_t + \alpha_j + \epsilon_{jt} \end{aligned} \tag{1.24}$$

where W_j^0 is the unobserved household-level network and ST_j is a column vector with the ith line indicating whether individual i was selected and lives in treated village j. Vector $\alpha_j =$

$[\alpha_{1j}, \ldots, \alpha_{n_j j}]$ are household-level fixed effects. The term $\lambda W_j^0 y_{jt}$ represents the endogenous effects – the fact that one's own choice depends on others' choices – and $W_j^0 ST_j \mathbf{1}\{s = t\}\beta_{2s}$ represents exogenous effects, i.e., the dependence of one's own choices on others' treatment status. As explained in Subsection 1.3.4, the correlated effects are captured by the fixed effects and eliminated via the subtraction of time averages. Coefficients β_{22} and β_{23} are interpreted as the marginal effect of treating a peer. Finally, I average network spillovers $\varphi(x_{jt}, \hat{\theta})$ for treated individuals after two and four years (denoted $\hat{\varphi}_{T,2}$ and $\hat{\varphi}_{T,4}$, respectively) and similarly for non-treated individuals (denoted as $\hat{\varphi}_{NT,2}$ and $\hat{\varphi}_{NT,4}$, respectively). It is notable that the overall treatment effect for the treated individuals is the sum of the program effect and spillovers. The construction of the confidence intervals and standard errors is described in Subsection 1.3.2.

Alternative Methods for Estimating Network Effects

There are a variety of methods in the literature to estimate network effects. For example, a possibility in the current setting is to compare non-selected households in treated villages against non-selected households in control villages. Other alternatives explored in the literature introduce variation in the fraction of the population assigned to treatment across groups (Crépon et al., 2012). There are two reasons why the current method improves on these approaches.

The first reason is related to precision of the estimates. Consider two polar cases: general equilibrium effects in which social interactions are intermediated solely by the markets (decrease in the supply of wage hours increases wage in the market) and local interactions (wage jobs left by treated households are occupied through network acquaintances). General equilibrium effects means that all individuals are affected to a small extent by the decisions of others. Networks are dense with weak links. In contrast, local interactions imply strong network spillovers only for those connected to treated households and null for unconnected individuals. The latter case generates large variation in individual outcome which then affects the precision of the estimates.

Second, comparison of non-selected households estimates network spillovers only, which can originate from a combination of endogenous and exogenous effects. In the current setting, for instance, the marginal effect of a connection requires separately identifying endogenous and exogenous effects, which is not possible by comparing non-selected households in treated villages against non-selected households in non-treated villages.

1.5.3 Empirical Results

I consider four sets of outcomes: occupational choice indicators (self-working hours, wage employment hours and specialization in self-employment in Table 1.3), earnings and seasonality (household earnings, in thousands of Bangladeshi Takas, share of income originating from seasonal and regular activities in Table 1.4), livestock assets (number of cows, poultry and livestock value in thousands of Takas in Table 1.5) and per capita expenditures (nonfood and food items and food security in Table 1.6). As an indicator of differential patterns of association, I allow the probability of link formation to depend on the proximity of household identifiers, registered as $Q_{ij} = 1$ and zero otherwise. It has been anecdotally observed that identifiers were allocated while field surveyors followed local streets and roads, and therefore serve as a proxy for geographical distance. This pattern is only a generalization from the purely naive network in which the probability of link formation is constant and independent of any variable.[34]

For each outcome, I show the triple differences-in-differences estimates of the program effects for the treated households ignoring networks, as in equation (1.23). These are shown in odd numbered columns in Tables 1.3-1.6. For example, column 1 of Table 1.3 indicates that treated increased self-working hours in 468.9 and 465.1 hours per year, two and four years after treatment respectively, and these results are significant at the 1% confidence level. Even columns display the results of the triple differences-in-differences augmented with the network module, as in equation (1.24). For example, column 2 of Table 1.3 also indicates treated increased self-working hours in 469.8 and 460.0 hours per year, two and four years after treatment respectively. These numbers are not significantly different from the cases in which networks were ignored in column 1. Therefore, in this particular case, inconsistency due to omission of networks was not a relevant problem.

The following four rows display the results for the network spillovers. Results in this case are not significant at 10% level two years after treatment for treated and nontreated, and point estimates are -6.3 and -3.2 hours per year. However, spillovers are positive and significant four years after treatment at 28.8 and 14.7 self-working hours per year for treated and nontreated respectively, indicating a slight increase in the supply of self-working hours due to spillovers for both types of households. The estimates for the program effect on treated and spillovers, as discussed above, does not depend on separately identifying endogenous and exogenous effects and, hence, do not rely on the presence of group size asymmetries and the moment condition based on outcome dispersion.

Breaking down spillovers in endogenous and exogenous effects then allows me to estimate the marginal effect of the connection to a treated household. These rows are labelled "Link to T". A marginal connection reduces working hours in 24.6 and 17.9 hours per year two and four years after treatment respectively, and are significant at the 1% confidence level. The probabilities of link formation are very high, at 98.3% if individuals live in close vicinity, and 39.6% otherwise indicating that, in this case, network effects operate via general equilibrium. The hypothesis that these numbers are equal is rejected at the 1% level.

[34] Estimation with naive model for probability of link formation is conducted as a robustness in Table 1.F.13 in Appendix 1.F.2. In addition, estimation without fixed effects, time effects and both are also shown to highlight that in their absence network estimates are highly biased.

I present the remaining results in three stages. First, I describe the results for network spillovers for all outcomes. These are followed by the estimates of the network structure and the marginal effect of a connection to a treated household. Finally, I incorporate network data directly into the procedure and demonstrate that the main conclusions remain unchanged. I also show that family self-reported links convey meaningful interaction and mixed results for economic (non-family) links.

Network Spillovers.

As shown in Subsection 1.3.2, it is not necessary to identify the parameters that underpin network formation or those that link explanatory variables to outcomes in a given network, and it is also not necessary to separate endogenous and exogenous effects. It is sufficient that Proposition 1 ensures that spillovers are constant within the identified set.

The current application shows that spillovers on treated and non-treated individuals determined outcomes to a relevant degree. The effect of spillovers was particularly salient in explaining food per capita expenditures. For example, spillovers amounted to 207.0 Takas per year for non-treated individuals after two years, compared with an estimated program effect of 423.9 Takas for treated individuals over the same period. This difference corresponds to a 6.9% increase on top of baseline levels of consumption, or 48.8% of the treatment effect on the treated individuals. The spillover effect is even larger for the treated subpopulation. After two years, spillovers from the treated households to themselves were responsible for an expenditure increase of 380.0 Takas, or 89.6% of the treatment effects. Notably, column 3 of Table 1.3 shows that estimates of treatment effects when networks are not included in the analysis are approximately 40% higher. This difference is attributed to the fact that OLS estimates, as presented in Subsection 1.2.1, may be inconsistent when networks effects are not accounted for.

This result is further confirmed by estimates of food security that are measured by respondents that reported having at least two meals on most days, indicating a positive effect for both the treated and the non-treated groups, across two and four years, ranging from 2.7 percentage points for the non-treated group two years after treatment to 7.1 percentage points for the treated group at the same time. The direct program effect is estimated at 16.9 and 7.6 percentage points (after two years and four years, respectively). Nonfood expenditures are either constant or exhibit a slight increase for the treated group, whereas the non-treated group reduced nonfood consumption after four years. As discussed below, this result can be explained by the reduction in productive assets following the specialization of the peers of the treated group in terms of wage labor.

Spillovers were significant to a small extent in determining self-employment and wage hours, specialization in self-employment, the share of seasonal and regular activities and asset holdings. As discussed above, network spillovers are reduced-form estimates that consist of endogenous effects, or the fact that one's own choice depends on others' choices, and exogenous effects, the fact that one's own choices depend on the treatment of others. Disentangling these structural mechanisms is useful in shedding light on the causes of these results, and this is undertaken in the next Subsection.

Endogenous and Exogenous Effects (or Marginal Value of Connections to the Treated)

I now provide point estimates of structural parameters. Given a network, its full set consists of link strength (λ), one's own response to one's own treatment after two and four years (β_{11} and β_{12}) and exogenous effects (or, in the current setting, the effect of one additional connection to a treated individual, β_{21} and β_{22}). The parameters that capture the network link are the probability of link formation if households are located in close proximity (δ_1), such that $Q_{ij} = 1$ if the difference in household identifiers is less than two[35], and if households are not in close proximity (δ_0). These parameters discriminate between the polar cases in which interactions occur on a localized scale, through personal interconnections and without intermediation of the markets (equivalent to low-density networks, or low δ_0 and δ_1) or through general equilibrium effects in which one's own choices affect all others to a small degree and result in dense networks (high δ_0 and δ_1). As demonstrated in Theorem 4, identification is achieved using the comparison between observed and theoretical across-group dispersion of outcomes as implied by the model. In a social setting, the across-group variation of outcomes cannot be explained by outcome dispersion, peer group heterogeneity or disturbance variance alone. This indicates a moment condition and suggests the use of a GMM criterion that is capable of sorting among structural parameters within the identified set.

In the current application, the estimates show that, whereas treated individuals reduced wage hours (113.5 and 141.9 hours per year, two and four years after treatment, respectively) and increased self-employment hours (469.8 and 460.0 hours per year) associated with livestock rearing, a marginal connection to a treated household had the opposite effect, increasing wage hours (24.6 and 17.9 hours per year for each treated peer) and decreasing self-working hours (13.9 and 13.0 hours per year for each treated peer). Treated individuals specialize in self-employment, and connected peers modestly decrease specialization. Individuals who received treatment left vacancies on wage jobs that were partially filled by individuals located in close geographic proximity[36]. The density of estimated networks is high only for self-employment and wage hours; above 90% for households that live in close proximity and approximately 40% otherwise. The interaction patterns of all other outcomes are much more localized, with densities of approximately 20% or lower in most cases.

[35] Robusteness checks are conducted in Table 1.F.13 of Appendix 1.F.2.

[36] The null hypothesis of no differential association is rejected at the 5% level for all specifications, as shown in Tables (1.3)-(1.6). Given the estimated parameters and the number of treated households in each households, a simple simulation exercise shows that exogenous effects counterbalanced 25-30% of the reduction in wage hours of treated households.

The results demonstrate that treated individuals increased their livestock assets by more than the original treatment. Meanwhile, non-treated individuals reduced their stock of assets. This outcome was not observed for poultry, which is consistent with the low takeover rate of this type of asset. Livestock value followed the same pattern for both groups. Since the treatment also consisted of skills training – specifically targeted for the type of assets provided – and was of long duration (2 years), treated individuals were endowed with a stronger comparative advantage in livestock rearing, whereas connected peers tended to specialize in wage jobs instead.

The final component of the analysis involves the food staples. A marginal connection to a treated peer significantly increases food consumption per capita and food security. In fact, one connection may be responsible for an effect on food expenditures that is equivalent to the direct effect of treatment on the treated individual (443.6 versus 423.9 Takas) and a 9.6 percentage point increase in food security. This finding shows that comovements of occupational choices of the treated and their peers were largely beneficial to all.

Including Network Data

Finally, I make use of network data collected in the survey to reassess the conclusions obtained in their absence. Inclusion of network data serves two primary purposes. First, I show that the main conclusions summarized above remain unchanged (Tables 1.F.9 to 1.F.12 of Appendix 1.F.2). Second, allowing link formation to depend on link reporting enables me to test whether the associated coefficient is significant, which constitutes as a test of

Table 1.3: Occupational Choice.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Self hours.		Wage hours.		Self emp. only.	
	Method	OLS.	Network.	OLS.	Network.	OLS.	Network.
Not function of $\hat{\lambda}$.	Program effect	468.928^{***}	469.774^{***}	-110.799^{***}	-113.531^{***}	0.107^{***}	0.114^{***}
	after 2 years ($\hat{\beta}_{11}$).	(28.62)	(23.20)	(31.07)	(10.61)	(0.02)	(0.01)
	Program effect	465.075^{***}	460.039^{***}	-137.255^{***}	-141.918^{***}	0.112^{***}	0.120^{***}
	after 4 years ($\hat{\beta}_{12}$).	(31.32)	(23.21)	(34.10)	(8.63)	(0.02)	(0.01)
	Spillover on T	—	-6.347	—	26.855^{***}	—	-0.032^{***}
	after 2 years ($\hat{\varphi}_{T,2}$).		(10.55)		(8.45)		(0.01)
	Spillover on T	—	28.847^{***}	—	19.369^{**}	—	-0.025^{***}
	after 4 years ($\hat{\varphi}_{T,4}$).		(9.68)		(8.54)		(0.00)
	Spillover on NT	—	-3.229	—	14.491^{***}	—	-0.018^{***}
	after 2 years ($\hat{\varphi}_{NT,2}$).		(5.37)		(4.55)		(0.00)
	Spillover on NT	—	14.676^{***}	—	10.452^{***}	—	-0.013^{***}
	after 4 years ($\hat{\varphi}_{NT,4}$).		(1.09)		(0.75)		(0.00)
Function of $\hat{\lambda}$.	Link to T	—	-24.604^{***}	—	13.904^{***}	—	-0.050^{***}
	after 2 years ($\hat{\beta}_{21}$).		(2.76)		(2.52)		(0.01)
	Link to T	—	-17.932^{***}	—	13.030^{***}	—	-0.043^{***}
	after 4 years ($\hat{\beta}_{22}$).		(2.76)		(1.59)		(0.01)
	Link probability	—	0.983^{***}	—	0.639^{***}	—	0.192^{***}
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).		(0.03)		(0.03)		(0.01)
	Link probability	—	0.396^{***}	—	0.331^{***}	—	0.106^{***}
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).		(0.01)		(0.01)		(0.00)
	Link strength	—	0.05^{***}	—	0.05^{***}	—	0.15^{***}
	($\hat{\lambda}$).		(0.01)		(0.00)		(0.01)
	p-value $\mathcal{H}_{NV}$.	—	< 0.001	—	< 0.001	—	< 0.001
	Avg treated outcome.	421.8	421.8	646.7	646.7	0.303	0.303
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes: *, ** and *** indicates significance at 10%, 5% and 1% levels. All regressions have household fixed effects. Standard errors clustered at the village level. "Spillover on T" refers to the average $\varphi(x_t, \hat{\theta})$ on the treated only. "Spillovers on NT" refers to equivalent calculation on the non-treated only. "Link to T" refers to the marginal effect of a connection to a treated individual. "Avg treated outcome" refers to the mean outcome of treated at the baseline. "p-value $\mathcal{H}_{NV}$" is the p-value of testing the null hypothesis that household proximity does not affect the probability of link formation. Estimates dependent on the identification strategy for $\hat{\lambda}$ are denoted under the tab "Function of $\hat{\lambda}$". "Self hours" refers to self-working hours per year. "Wage hours" refers to wage working hours per year. "Self emp. only" is a dummy variable if individual is specialized in self-employment.

Table 1.4: Earnings and Seasonality.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Earnings.		Share Seas.		Share Reg.	
	Method	OLS.	Network.	OLS.	Network.	OLS.	Network.
Not function of $\hat{\lambda}$.	Program effect	0.475	0.506^{***}	0.012	-0.028^{***}	0.201^{***}	0.181^{***}
	after 2 years ($\hat{\beta}_{11}$).	(0.46)	(0.12)	(0.02)	(0.01)	(0.02)	(0.01)
	Program effect	2.598^{***}	2.729^{***}	-0.089^{***}	-0.074^{***}	0.191^{***}	0.165^{***}
	after 4 years ($\hat{\beta}_{12}$).	(0.54)	(0.31)	(0.02)	(0.01)	(0.02)	(0.01)
	Spillover on T	—	-0.045	—	-0.051^{***}	—	0.023^{**}
	after 2 years ($\hat{\varphi}_{T,2}$).		(0.10)		(0.02)		(0.01)
	Spillover on T	—	0.008	—	-0.005	—	0.029^{**}
	after 4 years ($\hat{\varphi}_{T,4}$).		(0.11)		(0.02)		(0.01)
	Spillover on NT	—	-0.025	—	-0.023^{***}	—	0.012^{**}
	after 2 years ($\hat{\varphi}_{NT,2}$).		(0.06)		(0.01)		(0.00)
	Spillover on NT	—	0.004	—	-0.002	—	0.015^{***}
	after 4 years ($\hat{\varphi}_{NT,4}$).		(0.09)		(0.01)		(0.00)
Function of $\hat{\lambda}$.	Link to T	—	-0.447	—	-0.010^{***}	—	-0.022^{***}
	after 2 years ($\hat{\beta}_{21}$).		(0.46)		(0.01)		(0.01)
	Link to T	—	-0.326	—	-0.016^{***}	—	-0.015^{**}
	after 4 years ($\hat{\beta}_{22}$).		(0.29)		(0.01)		(0.00)
	Link probability	—	0.075^{***}	—	0.272^{***}	—	0.238^{***}
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).		(0.00)		(0.01)		(0.00)
	Link probability	—	0.023^{***}	—	0.136^{***}	—	0.106^{***}
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).		(0.00)		(0.00)		(0.00)
	Link strength	—	0.50^{***}	—	0.20^{***}	—	0.20^{***}
	($\hat{\lambda}$).		(0.17)		(0.08)		(0.05)
	p-value $\mathcal{H}_{NV}$.	—	< 0.001	—	< 0.001	—	< 0.001
	Avg treated outcome.	4.607	4.607	0.674	0.674	0.478	0.478
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes: Earnings in thousand of Takas per year. "Share Seas." refers to the share of seasonal earnings relative to total earnings. "Share Reg." refers to share of regular earnings, as reported by the respondent, relative to total earnings. See also Table 1.3.

network data validity. I combine network reports into two categories: family and economic (non-family) links. Non-family links include an ensemble of many categories of self-reported links, such as business and labor relationships, financial assets and liabilities and household ownership. The null hypothesis of no network validity was rejected at the 1% level for all specifications regarding occupational choice, earnings and seasonality. The results for livestock holding and expenditures are more nuanced. Whereas for most specifications, the null of no validity was rejected for family links, economic links are much less capable of conveying interactions that influence the outcomes of others. This result suggests that families are natural *loci* that favor asset transactions, particularly when those transactions involve cows, and through which food consumption and expenditures flow.

Table 1.5: Livestock.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Cows.		Poultry.		Livestock Value.	
	Method	OLS.	Network.	OLS.	Network.	OLS.	Network.
Not function of $\hat{\lambda}$.	Program effect	1.119^{***}	1.131^{***}	2.147^{***}	2.120^{***}	10.326^{***}	10.417^{***}
	after 2 years ($\hat{\beta}_{11}$).	(0.04)	(0.03)	(0.42)	(0.50)	(0.56)	(0.39)
	Program effect	1.078^{***}	1.102^{***}	1.294^{**}	1.326^{***}	10.984^{***}	11.175^{***}
	after 4 years ($\hat{\beta}_{12}$).	(0.03)	(0.03)	(0.62)	(0.50)	(0.64)	(0.40)
	Spillover on T	—	-0.033^{***}	—	0.099	—	-0.221^{***}
	after 2 years ($\hat{\varphi}_{T,2}$).		(0.01)		(0.17)		(0.07)
	Spillover on T	—	-0.057^{***}	—	-0.087	—	-0.459^{***}
	after 4 years ($\hat{\varphi}_{T,4}$).		(0.00)		(0.20)		(0.07)
	Spillover on NT	—	-0.020^{***}	—	0.059	—	-0.132^{***}
	after 2 years ($\hat{\varphi}_{NT,2}$).		(0.01)		(0.10)		(0.04)
	Spillover on NT	—	-0.033^{***}	—	-0.052	—	-0.274^{***}
	after 4 years ($\hat{\varphi}_{NT,4}$).		(0.01)		(0.08)		(0.04)
Function of $\hat{\lambda}$.	Link to T	—	-0.996^{***}	—	1.277	—	-10.456^{***}
	after 2 years ($\hat{\beta}_{21}$).		(0.16)		(4.12)		(1.90)
	Link to T	—	-1.285^{***}	—	-2.725	—	-16.464^{***}
	after 4 years ($\hat{\beta}_{22}$).		(0.17)		(4.11)		(2.33)
	Link probability	—	0.024^{***}	—	0.007^{**}	—	0.013^{***}
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).		(0.00)		(0.00)		(0.00)
	Link probability	—	0.012^{***}	—	0.009^{***}	—	0.007^{***}
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).		(0.00)		(0.00)		(0.00)
	Link strength	—	0.50^{***}	—	0.50	—	0.50^{***}
	($\hat{\lambda}$).		(0.03)		(0.38)		(0.16)
	p-value $\mathcal{H}_{NV}$.	—	< 0.001	—	< 0.001	—	< 0.001
	Avg treated outcome.	0.083	0.083	1.79	1.79	0.940	0.940
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes: "Cows" refers to the number of cows held by the household, and similarly for poultry. Livestock value evaluates in thousands of Takas at market value. See also Table 1.3.

Table 1.6: Expenditures.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Nonfood PCE.		Food PCE.		Food Security.	
	Method	OLS.	Network.	OLS.	Network.	OLS.	Network.
Not function of $\hat{\lambda}$.	Program effect	-242.239	-220.509	585.304^{**}	423.929^{***}	0.189^{***}	0.169^{***}
	after 2 years ($\hat{\beta}_{11}$).	(293.34)	(164.53)	(247.19)	(134.22)	(0.03)	(0.01)
	Program effect	175.022	278.277	585.415^{***}	445.063^{***}	0.010^{***}	0.076^{***}
	after 4 years ($\hat{\beta}_{12}$).	(375.16)	(174.72)	(227.38)	(134.27)	(0.03)	(0.01)
	Spillover on T	—	-8.526	—	380.002^{***}	—	0.017^{***}
	after 2 years ($\hat{\varphi}_{T,2}$).		(68.25)		(55.82)		(0.00)
	Spillover on T	—	-171.985^{**}	—	243.172^{***}	—	0.071^{***}
	after 4 years ($\hat{\varphi}_{T,4}$).		(68.15)		(56.88)		(0.02)
	Spillover on NT	—	-5.039	—	206.992^{***}	—	0.027^{***}
	after 2 years ($\hat{\varphi}_{NT,2}$).		(40.34)		(30.14)		(0.00)
	Spillover on NT	—	-101.655^{*}	—	132.459^{***}	—	0.032^{***}
	after 4 years ($\hat{\varphi}_{NT,4}$).		(52.65)		(40.73)		(0.01)
Function of $\hat{\lambda}$.	Link to T	—	-14.185	—	443.619^{***}	—	0.096^{***}
	after 2 years ($\hat{\beta}_{21}$).		(988.46)		(85.36)		(0.01)
	Link to T	—	-2649.43^{***}	—	249.126^{***}	—	0.087^{***}
	after 4 years ($\hat{\beta}_{22}$).		(980.96)		(84.79)		(0.01)
	Link probability	—	0.032^{***}	—	0.132^{***}	—	0.128^{***}
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).		(0.00)		(0.01)		(0.00)
	Link probability	—	0.009^{***}	—	0.080^{***}	—	0.052^{***}
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).		(0.00)		(0.00)		(0.00)
	Link strength	—	0.50^{***}	—	0.20^{**}	—	0.50^{**}
	($\hat{\lambda}$).		(0.14)		(0.11)		(0.21)
	p-value $\mathcal{H}_{NV}$.	—	< 0.001	—	< 0.001	—	< 0.001
	Avg treated outcome.	1054.5	1054.5	2953.7	2953.7	0.457	0.457
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes: "Nonfood PCE" refers to non-food per capita expenditure in thousands of Takas per year, and similarly for food per capita expenditures. Food security is a dummy equal to one if households have at least two meals in most days. Estimates of the program impact on nonfood per capita expenditure on the treated using the triple differences model (column 1) was the only case which does not match well the estimates obtained from the double differences which compares the selected individuals in treated villages against selected in nontreated villages. See Bandiera et al. (2013) and Table 1.3.

1.6 Conclusion

Social and economic networks are useful for understanding many aspects of individual choice, decisions and behavior. Although there has recently been substantial progress on the theoretical underpinnings of network formation, empirical research frequently remains constrained by the availability of network data. The contribution of this paper is then to provide a method for estimating network effects in the absence of network data. The method also estimates the probability that pairs of individuals form a connection based on individual characteristics such as common gender. I also incorporate imperfect network data with the dual purpose of refining the estimates and providing a test for its validity.

The key contribution of the paper was to derive a maximum likelihood estimator that is not conditional on network data. It is obtained by integrating a likelihood conditional on networks which originates from a spatial econometric model with respect to the probability density function of the stochastic network. In this setting, I showed how the observation of outcomes and explanatory variables for many groups such as classrooms serves as a substitute for the network observation. This approach then offers a procedure for estimating network effects using datasets that were previously not suited for this purpose.

Empirical research has led to substantial interest in evaluating the effects of randomized policies on targeted individuals. Much less progress has been made on evaluating the spillovers related to those policies. To illustrate how the method can be applied in practice, I employed the estimator to investigate the impact of a large-scale randomized intervention on the peers of those who were treated. This is the intervention of Bandiera et al. (2013), which consisted of the provision of livestock and skill training to low-income households in Bangladesh.

The proposed estimator met three objectives and yielded useful insights on the wider effects of the policy. The first objective was to provide – in the absence of network data – a consistent and asymptotically normal estimator of network spillovers. In the application, I found that network spillovers were economically and statistically significant in determining some outcomes, especially food per capita expenditure and food security. Network spillovers were responsible for an increase of 206.9 Takas in yearly food per capita expenditure compared with a treatment effect of 423.9 Takas on the treated.[37]

The second objective of the paper was to elucidate the structural mechanisms that gave rise to these spillovers. I derived a method to separately identify endogenous and exogenous effects, controlling for correlated effects, in the absence of network data by using the variability in group sizes. I further solved the problem of separately identifying a few strong links from a large number of weak links by using the "excess" outcome variance that cannot be explained by independent

[37] Respectively an 14% and 7% increase relative to food consumption levels at the baseline.

variables or peer group heterogeneity alone.[38] For this purpose, I reinterpreted the estimator as the solution of a Generalized Method of Moments problem in which moment conditions were given by the score of the likelihood. In this case, the earlier identification difficulty originated from the absence of one moment condition relative to the number of parameters. I then explored the difference between observed second moments of the outcomes and those implied by the model to provide an additional restriction which completes the identification requirements. I am then able to show that the solution of this problem is a consistent and asymptotically normal estimator to the structural parameters of the model.

In the application studied herein, I found that a marginal connection to the treated led to effects in opposite direction to the treatment effect on the treated. Regarding occupational choice and livestock value, one additional connection to a treated household decreased self-employment by 24.6 hours per year, added 13.9 wage hours per year and decreased livestock value by 10.4 thousand Takas. Treated households increased their self-employment hours, decreased their wage hours and increased the value of their livestock. In contrast, regarding food per capita expenditure and food security, a marginal connection to the treated was in the same direction to the treatment effect on the treated, and often of strong magnitudes. A marginal connection to the treated increased food per capita expenditure by 443.6 Takas per year and increased food security by 9.6 percentage points, compared with direct treatment effects of, respectively, 424.0 Takas per year and 16.9 percentage points. With the exception of self-employment and wage hours, I also found that network densities were fairly low, which suggested local interactions through personal contacts rather than through prices and markets. These results are consistent with the interpretation that treated individuals gained comparative advantage in livestock rearing. The randomized policy then generated a village-level occupational specialization in which treated households were employed in rearing the livestock, partially changing the occupational choice and well-being of their peers as measured by food consumption.

The third objective of this paper was to incorporate imperfect network data, such as when data are self-reported, with the dual purpose of refining the estimates and proposing a test for whether reported connections positively affect the estimated connection probability. In the application, I found that reported family links have a greater effect than the reported economic (non-family) links in determining the outcomes of others. The test rejected the null hypothesis that family links do not influence the number of cows but failed to reject the similar influence of economic links. The same holds true for livestock value, indicating that family ties facilitated asset transactions.

The method developed in the present paper contributes to the spatial econometrics literature that has to date considered only models for which networks are accurately known (Anselin (2010) and references therein). Similarly, the literature on the identification of network models addressed

[38] These are similar in essence to the identification ideas in Lee (2007) and Graham (2008), which explore the case in which networks are observed.

a number of techniques only when networks could be observed (Manski (1993), Bramoullé et al. (2009), De Giorgi et al. (2010) and others). This novel method can be applied in many fields, from peer effects (Ammermuller and Pischke, 2009), crime and delinquent behavior (Glaeser et al., 1996) to the estimation of parameters of gravity equations (Anderson and van Wincoop, 2003).

The interest in networks to this date has not been matched with availability of network data, possibly because of problems associated with observing and defining interactions among people. The method developed in the present paper provided a systematic procedure for estimating network effects when networks are unknown or unreliable and information on a large number of groups is available. This ability has shown to be particularly relevant in estimating effects of exogenous variation policy through randomized controlled trials both on treated and their peers. In this way, the paper demonstrated both theoretically and empirically that including network effects may have important implications for policy assessments. Estimating network spillovers and distinguishing among endogenous, exogenous and correlated effects in the absence of network data is certainly a useful empirical tool for future applied research.

Appendix

1.A Summary of Notation.

$\beta = \left(\beta_1', \beta_2'\right)$, $\theta_c = \theta \setminus \left\{\beta, \sigma^2\right\}$, $n = \sum_{j=1}^{v} n_j$.

I_n an identity matrix of dimensions $n \times n$, ι_n is a $n \times 1$ vector of ones.

$y_n = \left(y_1', \ldots, y_j', \ldots, y_v'\right)'$, $y_j = y_{n_j,j} \left(y_{1j}, \ldots, y_{ij}, \ldots, y_{n_j j}\right)'$, $j = 1, \ldots v$, $i = 1, \ldots, n_j$.

$x_n = \left(x_1', \ldots, x_j', \ldots, x_v'\right)'$, $x_j = x_{n_j,j} = \left(x_{1j}', \ldots, x_{ij}', \ldots, x_{n_j j}\right)'$, $j = 1, \ldots v$, $i = 1, \ldots, n_j$.

$\epsilon_n = \left(\epsilon_1', \ldots, \epsilon_j', \ldots, \epsilon_v'\right)'$, $\epsilon_{n_j,j} = \left(\epsilon_{1j}, \ldots, \epsilon_{ij}, \ldots, \epsilon_{n_j j}\right)'$, $j = 1, \ldots v$, $i = 1, \ldots, n_j$.

$y_j = \lambda_0 W_j^0 + x_j \beta_{10} + W_j^0 x_j \beta_{20} + v_j$, $v_j = \rho_0 M_j^0 v_j + \epsilon_j$.

$Z_j^0 = Z_{n_j,j}^0 = \left(x_j, W_j^0 x_j\right)$, $Z_n^0 = \left(Z_1^{0'}, \ldots, Z_v^{0'}\right)'$.

$y_j = \lambda W_j^e\left(Q, \theta_c\right) + x_j \beta_1 + W_j^e\left(Q, \theta_c\right) x_j \beta_2 + \hat{v}_j$.

$Z_j^e\left(Q_j, \theta_c\right) = Z_{n_j,j}^e\left(Q, \theta_c\right) = \left(x_j, W_j^e\left(Q, \theta_c\right) x_j\right)$, $Z_n^e\left(Q, \theta_c\right) = \left(Z_1^{e'}\left(Q_1, \theta_c\right), \ldots, Z_v^{e'}\left(Q_v, \theta_c\right)\right)'$.

$W_n^0 = \operatorname{diag}\left(W_1^0, \ldots, W_v^0\right)$, $W_n^e\left(Q_n, \theta_c\right) = \operatorname{diag}\left(W_1^e\left(Q_1, \theta_c\right), \ldots, W_v^e\left(Q_v, \theta_c\right)\right)$.

$S_j^0\left(\lambda\right) = S_{n_j,j}^0\left(\lambda\right) = I_{n_j} - \lambda W_j^0$, $S_j^0 = S_j^0\left(\lambda_0\right)$, $S_n^0 = \operatorname{diag}\left(S_1^0, \ldots, S_v^0\right)$.

$\left(S_n^0\right)^{-1} = \lambda_0 G_n^0 + I_n$, $G_n^0 = W_n^0\left(S_n^0\right)^{-1}$.

$S_j^e\left(Q_j, \theta\right) = I_{n_j} - \lambda W_j^e\left(Q_j, \theta\right)$, $S_n^e\left(Q_j, \theta_c\right) = \operatorname{diag}\left(S_1^e\left(Q_1, \theta_c\right), \ldots, S_v^e\left(Q_1, \theta_c\right)\right)$.

$\left(S_n^e\left(Q, \theta_c\right)\right)^{-1} = I_n + \lambda G_n^e\left(Q_n, \theta_c\right)$, $G_n^e\left(Q_n, \theta_c\right) \equiv W_n^e\left(Q_n, \theta_c\right)\left(S_n^e\left(Q_n, \theta_c\right)\right)^{-1}$.

$R_j^0\left(\rho\right) = I_{n_j} - \rho M_j^0$, $R_j^0 = R_j^0\left(\rho_0\right)$, $R_n^0 = \operatorname{diag}\left(R_1^0, \ldots, R_v^0\right)$.

$R_j^e\left(\theta\right) = I_{n_j} - \rho M_j^e\left(Q_j, \theta\right)$, $R_n^e\left(Q_n, \theta_c\right) = \operatorname{diag}\left(R_1^e\left(Q_1, \theta_c\right), \ldots, R_v^e\left(Q_v, \theta_c\right)\right)$.

$M_n^0 = \operatorname{diag}\left(M_1^0, \ldots, M_v^0\right)$, $M_n^e\left(Q_n, \theta_c\right) = \operatorname{diag}\left(M_1^e\left(Q_1, \theta_c\right), \ldots, M_v^e\left(Q_v, \theta_c\right)\right)$.

$P_n^e\left(Q_n, \theta_c\right) = I_n - R_n^e\left(Q_n, \theta_c\right) Z_n^e\left(Q_n, \theta_c\right)\left[Z_n^{e'}\left(Q_n, \theta_c\right) R_n^{e'}\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) Z_n^e\left(Q_n, \theta_c\right)\right]^{-1} Z_n^{e'}\left(Q_n, \theta_c\right) R_n^{e'}\left(Q_n, \theta_c\right)$.

$B_n\left(Q_n, \theta_c\right) = \lambda\left(W_n^0 - W_n^e\left(Q, \theta_c\right)\right)\left(I_n + \lambda_0 G_n^0\right)$.

$\mu\left[A\right]$ and $\Sigma\left[A\right]$ denote the expectation and variance-covariance matrix of vector A.

1.B Alternative network models.

I previously described the probability of link formation as dependent on a dummy for sharing exogenous characteristic with independence link formation. I now expand the classes of models in two different directions: I first allow the probability of link formation to depend on a continuous measure, such as distance between households location. Because many modes of social interactions can occur in parallel, it is also important to allow for a multivariate network formation model. In second place, I drop link independence assumption with recourse to the Exponential Random Markovian Graphs (ERMG) family of models, as introduced by Frank and Strauss (1986) and expanded by Wasserman and Pattison (1996). These are presented in form of examples.

Example 4. *(Multivariate network model).* Several forms of relations coexist; arguably, a truthful representation of the probability of link formation will then depend on a number of factors. Allow then Q^I_{ji} as $1 \times k^I$ to be a matrix of individual's i characteristics that underpin probability of link formation and depend exclusively on individual, non-relational, characteristics. For example, this may encompass testing whether males may tend to form more connections than the rest of the population, or personal income may have a relation to social interactions. Let Q^R_{jk} be characteristics of the potential recipient of the link that may generate attraction, of dimension $1 \times k^R$ and, finally, Q^B_{jik} common, shared characteristics, such as belonging to the same gender, or continuous geographic distance between households, with dimension $1 \times k^B$. Coefficients are captured with recourse to θ^I_g, θ^R_g and θ^B_g of compatible dimensions.

$$P\left\{\{W_j\}_{ik} = 1 | Q_j\right\} \quad = \quad Q^I_{ji}\theta^I_g + Q^R_{jk}\theta^R_g + Q^B_{jik}\theta^B_g. \tag{1.25}$$

Because probabilities should stay in the range $[0, 1]$, it is plausible to use, instead, $P\left\{\{W_j\}_{ik} = 1 | Q_j\right\} = \text{logit}(Q^I_{ji}\theta^I_g + Q^R_{jk}\theta^R_g + Q_{jik}\theta^B_g)$ or the equivalent probit version. It is important to note that, even without using the second moments to provide identification, it is still possible to conduct hypothesis testing in the partial identification framework, as long as there is no collinearity among Q^I_{ji}, Q^R_{ji} and Q^B_{jik} for all i, k and j. More specifically, suppose one is interested in whether race commonality affects the probability of link formation. The researcher can then test $\mathcal{H}_0 : \theta^B_g = 0$, with the procedure outlined in Subsection 1.3.2, although it will not be possible to identify the magnitude of the effect unless as a solution to equation (1.18) is provided. □

Example 5. *(ERMG family).* Models of statistic network formation have a long tradition in the literature of estimation of network structure given observations from random graphs generators (Holland and Leinhardt (1981), Frank and Strauss (1986), Strauss and Ikeda (1990) and Snijders (2011)) and are of considerable generality, including the case where link formation are not independent. In particular, Frank and Strauss (1986) proved that, if the graph is such that edges without common nodes are independent conditional on all remaining edges (that is, the graph is Markovian[39]) and homogeneous[40], and all isomorphic graphs have same probability,

$$P\{W_j = w_j\} \quad = \quad \frac{1}{\kappa(\theta_g)} \cdot \exp\left\{\theta^0_g T(w_j) + \sum_{s=1}^{n-1} \theta^s_g S_s(w_j)\right\} \tag{1.26}$$

where $T(w_j) = \sum_{i,k,l} \{w_j\}_{ik} \{w_j\}_{kl} \{w_j\}_{li}$ is the number of triangles, and $S_s(w_j)$ is the number of s-stars in w_j. $\kappa(\theta_g)$ is a normalization constant that depends on parameters $\theta_g = (\theta^0_g, \theta^1_g, \ldots, \theta^{n-1}_g)'$. The Markovian assumption is a relatively mild hypothesis and states that, although dependence between the existence of edges may happen, this cannot be so for edges which do not possess a common node. This formulation is particularly appealing as it provides a probability law for network formation under minimal hypothesis, along with its sufficient statistics. Wasserman and Pattison (1996) expand the class of models to incorporate any set of sufficient statistics $Z(w_j)$, such that

$$P\{W_j = w_j\} \quad = \quad \frac{1}{\kappa(\theta_g)} \cdot \exp\left\{\theta'_g Z(w_j)\right\}. \tag{1.27}$$

[39] Let D be a graph whose nodes are all possible edges of G, that is, all pairs of nodes of G, containing therefore $n!\,(n-1)!$ nodes. If the existence of an edge between $\{a, b\}$ in G depends on the existence of an edge between $\{c, d\}$, conditional on all rest of the graph, then $\{a, b\}$ and $\{c, d\}$ are neighbors in D. The Markovian assumption means, therefore, that all $\{s, t\}$ and $\{u, v\}$ are nonneighbors for different s, t, u and v.

[40] That is, nodes are a priori indistinguishable.

Note that, as a consequence of homogeneity, edges have equal probability of being formed with expected network $W_j^e(\theta_g) = p\iota_{n_j}\iota'_{n_j} - pI_{n_j}$. This is the same expectation as the one obtained in the simple Bernoulli model. □

1.C Score Vector and Hessian Matrix.

The likelihood is $\ln \mathcal{L}^e\left(\theta|\, y, x, Q_n\right) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) + \ln\left|S_n^e\left(Q_n,\theta\right)\right| + \ln\left|R_n^e\left(Q_n,\theta\right)\right| - \frac{1}{2\sigma^2}\epsilon_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right)$ where $\epsilon_n^e\left(Q_n,\theta\right) = R_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right)x_n\beta_2\right)$. First-order derivatives are

$$
\begin{array}{rcl}
\frac{\partial \ln \mathcal{L}^e(\theta)}{\partial \lambda} & = & -\mathrm{tr}\left[\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}W_n^e\left(Q_n,\theta\right)\right] + \frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q,\theta\right)R_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial \ln \mathcal{L}^e(\theta)}{\partial \beta_1} & = & -\frac{1}{\sigma^2}x_n' R_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial \ln \mathcal{L}^e(\theta)}{\partial \beta_2} & = & -\frac{1}{\sigma^2}x_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial \ln \mathcal{L}^e(\theta)}{\partial \sigma^2} & = & -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\epsilon_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial \ln \mathcal{L}^e(\theta)}{\partial \rho} & = & -\mathrm{tr}\left[\left(R_n^e\left(Q_n,\theta\right)\right)^{-1}M_n^e\left(Q_n,\theta\right)\right] + \frac{1}{\sigma^2}\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta - W^e\left(Q_n,\theta\right)x_n\beta_2\right)' \cdot \\
 & & M_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial \ln \mathcal{L}^e(\theta)}{\partial \theta_{gi}} & = & -\lambda\mathrm{tr}\left[\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}\nabla_{\theta_{gi}}W_n^e\left(Q_n,\theta\right)\right] - \rho\mathrm{tr}\left[\left(R_n^e\left(Q_n,\theta\right)\right)^{-1}\nabla_{\theta_{gi}}M_n^e\left(Q_n,\theta\right)\right] \\
 & & +\frac{1}{2\sigma^2}\rho\nabla_{\theta_{gi}}M_n^e\left(Q_n,\theta\right)'\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right)x_n\beta_2\right)'\epsilon_n^e\left(Q_n,\theta\right) \\
 & & +\frac{1}{2\sigma^2}R_n^e\left(Q_n,\theta\right)'\nabla_{\theta_{gi}}W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right)'\epsilon^e\left(Q_n,\theta\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \lambda\partial \lambda} & = & -\mathrm{tr}\left[\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}W_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}W_n^e\left(Q_n,\theta\right)\right] \\
 & & -\frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)W_n^e\left(Q_n,\theta\right)y_n \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \lambda\partial \beta_1'} & = & -\frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)x_n \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \lambda\partial \beta_2'} & = & -\frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)W_n^e\left(Q_n,\theta\right)x_n \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \lambda\partial \sigma^2} & = & -\frac{1}{\sigma^4}y_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \lambda\partial \rho} & = & -\frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)M_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q,\theta\right) - \\
 & & \frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)M_n^{e'}\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right)x_n\beta_2\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \lambda\partial \theta_{gi}} & = & -\lambda\mathrm{tr}\left[\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}\nabla_{\theta_{gi}}W_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}W_n^e\left(Q_n,\theta\right)\right] \\
 & & -\mathrm{tr}\left[\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}\nabla_{\theta_{gi}}W_n^e\left(Q_n,\theta\right)\right] \\
 & & +\frac{1}{\sigma^2}y_n'\nabla_{\theta_{gi}}W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_b,\theta\right) - \rho\frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)\nabla_{\theta_{gi}}M_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
 & & -\rho\frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)M_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right)x_n\beta_2\right) \\
 & & -\frac{1}{\sigma^2}y_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)\nabla_{\theta_{gi}}W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_1\partial \beta_1'} & = & \frac{1}{\sigma^2}x_n' R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)x_n \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_1\partial \beta_2'} & = & \frac{1}{\sigma^2}x_n' R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)W_n^e\left(Q_n,\theta\right)x_n \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_1\partial \sigma^2} & = & \frac{1}{\sigma^4}x_n' R_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_1\partial \rho} & = & \frac{1}{\sigma^2}x_n' M_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) - \frac{1}{\sigma^2}x_n' R_n^{e'}\left(Q_n,\theta\right)M_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right)x_n\beta_2\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_1\partial \theta_{gi}} & = & \rho\frac{1}{\sigma^2}x_n'\nabla_{\theta_{gi}}M_n^{e'}\left(Q,\theta\right)\epsilon_n^e\left(Q,\theta\right) \\
 & & +\rho\frac{1}{\sigma^2}x_n R_n^{e'}\left(Q_n,\theta\right)\nabla_{\theta_{gi}}M_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right)x_n\beta_2\right) \\
 & & +\frac{1}{\sigma^2}x_n' R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)\nabla_{\theta_{gi}}W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right). \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_2\partial \beta_2'} & = & \frac{1}{\sigma^2}x_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)W_n^e\left(Q_n,\theta\right)x_n \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_2\partial \sigma^2} & = & \frac{1}{\sigma^4}x_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_2\partial \rho} & = & \rho\frac{1}{\sigma^2}x_n' W_n^{e'}\left(Q_n,\theta\right)M_n^{e'}\left(Q_n,\theta\right)\epsilon_n^e\left(Q_n,\theta\right) + \\
 & & \frac{1}{\sigma^2}x_n' W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)M_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right)x_n\beta_2\right)
\end{array}
$$

$$
\begin{aligned}
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \beta_2 \partial \theta_{gi}} &= -\tfrac{1}{\sigma^2} x_n' \nabla_{\theta_{gi}} W_n^{e'}\left(Q_n,\theta\right) R_n^{e'}\left(Q_n,\theta\right) \epsilon_n^e\left(Q_n,\theta\right) + \rho \tfrac{1}{\sigma^2} x_n' W_n^{e'}\left(Q_n,\theta\right) \nabla_{\theta_{gi}} M_n^{e'}\left(Q_n,\theta\right) \epsilon_n^e\left(Q_n,\theta\right) \\
& \quad \rho \tfrac{1}{\sigma^2} x_n' W_n^{e'}\left(Q_n,\theta\right) R_n^{e'}\left(Q_n,\theta\right) M_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right) \\
& \quad \tfrac{1}{\sigma^2} x_n' W_n^{e'}\left(Q_n,\theta\right) R_n^{e'}\left(Q_n,\theta\right) R_n^e\left(Q_n,\theta\right) \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \sigma^2 \partial \sigma^2} &= \tfrac{n}{2\sigma^4} - \tfrac{1}{\sigma^6} \epsilon_n^{e'}\left(Q_n,\theta\right) \epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \sigma^2 \partial \rho} &= -\tfrac{1}{\sigma^4}\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right) M_n^{e'}\left(Q_n,\theta\right) \epsilon_n^e\left(Q_n,\theta\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \sigma^2 \partial \theta_{gi}} &= \tfrac{1}{\sigma^4} \epsilon_n^{e'}\left(Q_n,\theta\right) R_n^e\left(Q_n,\theta\right) \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right) \\
& \quad -\rho \tfrac{1}{\sigma^4} \epsilon_n^{e'}\left(Q_n,\theta\right) \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \rho \partial \rho} &= \operatorname{tr}\left[\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} M_n^e\left(Q_n,\theta\right)\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} M_n^e\left(Q_n,\theta\right)\right] \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \rho \partial \theta_{gi}} &= \rho \operatorname{tr}\left[\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right)\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} M_n^e\left(Q_n,\theta\right)\right] \\
& \quad -\operatorname{tr}\left[\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right)\right] \\
& \quad -\tfrac{1}{\sigma^2}\left(\lambda y_n + x_n\beta_2\right)' \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)' M_n^{e'}\left(Q_n,\theta\right) \epsilon_n^e\left(Q_n,\theta\right) \\
& \quad +\tfrac{1}{\sigma^2}\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right)' \nabla_{\theta_{gi}} M_n^{e'}\left(Q_n,\theta\right) \epsilon_n^e\left(Q_n,\theta\right) \\
& \quad -\rho \tfrac{1}{\sigma^2}\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right)' M_n^{e'}\left(Q_n,\theta\right) \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right) \cdot \\
& \quad \cdot\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right) \\
& \quad -\lambda \tfrac{1}{\sigma^2}\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right)' M_n^{e'}\left(Q_n,\theta\right) R_n^e\left(Q_n,\theta\right) \\
& \quad \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right) \\
\frac{\partial^2 \ln \mathcal{L}^e(\theta)}{\partial \theta_{gi} \partial \theta_{gk}} &= \lambda^2 \operatorname{tr}\left[\left(S_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gk}} W_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)\right] \\
& \quad -\lambda \operatorname{tr}\left[\left(S_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gi}\theta_{gk}} W_n^e\left(Q_n,\theta\right)\right] \\
& \quad +\rho^2 \operatorname{tr}\left[\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gk}} M_n^e\left(Q_n,\theta\right)\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right)\right] \\
& \quad -\rho \operatorname{tr}\left[\left(R_n^e\left(Q_n,\theta\right)\right)^{-1} \nabla_{\theta_{gi}\theta_{gk}} M_n^e\left(Q_n,\theta\right)\right] \\
& \quad +\rho \tfrac{1}{2\sigma^2} \nabla_{\theta_{gi}\theta_{gk}} M_n^e\left(Q_n,\theta\right)'\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right)' \epsilon_n^e\left(Q_n,\theta\right) \\
& \quad -\rho \tfrac{1}{2\sigma^2} \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right)' \nabla_{\theta_{gk}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right)' \epsilon_n^e\left(Q_n,\theta\right) \\
& \quad -\rho^2 \tfrac{1}{2\sigma^2} \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right)'\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right)' \cdot \\
& \quad \cdot \nabla_{\theta_{gk}} M_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right) \\
& \quad -\rho \tfrac{1}{2\sigma^2} \nabla_{\theta_{gi}} M_n^e\left(Q_n,\theta\right)'\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right)' R_n^e\left(Q_n,\theta\right) \cdot \\
& \quad \cdot \nabla_{\theta_{gk}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right) \\
& \quad -\rho \tfrac{1}{2\sigma^2} \nabla_{\theta_{gk}} M_n^e\left(Q_n,\theta\right)' \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right)' \epsilon_n^e\left(Q_n,\theta\right) \\
& \quad +\tfrac{1}{2\sigma^2} R_n^e\left(Q_n,\theta\right)' \nabla_{\theta_{gi}\theta_{gk}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right)' \epsilon_n^e\left(Q_n,\theta\right) \\
& \quad -\rho \tfrac{1}{2\sigma^2} R_n^e\left(Q_n,\theta\right)' \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right)' \nabla_{\theta_{gk}} M_n^e\left(Q_n,\theta\right) \cdot \\
& \quad \cdot\left(S_n^e\left(Q_n,\theta\right) y_n - x_n\beta_1 - W_n^e\left(Q_n,\theta\right) x_n\beta_2\right) \\
& \quad -\tfrac{1}{2\sigma^2} R_n^e\left(Q_n,\theta\right)' \nabla_{\theta_{gi}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right)' R_n^e\left(Q_n,\theta\right) \nabla_{\theta_{gk}} W_n^e\left(Q_n,\theta\right)\left(\lambda y_n + x_n\beta_2\right)
\end{aligned}
$$

Derivatives $\nabla_{\theta_{gi}} W_j^e\left(Q_j,\theta\right) = \frac{\partial W_j^e\left(Q_j,\theta\right)}{\partial \theta_{gi}}$, $\nabla_{\theta_{gi}\theta_{gk}} W_j^e\left(Q_j,\theta\right) = \frac{\partial^2 W_j^e\left(Q_j,\theta\right)}{\partial \theta_{gi} \partial \theta_{gk}}$ and similarly for derivatives of $M_j^e\left(Q_j,\theta\right)$ and model-dependent and so are omitted here.

1.D Proofs.

1.D.1 Useful Lemmas.

Lemmas without proofs can be found in Kelejian and Prucha (2001), Lee (2004) or Lee et al. (2010).

Lemma 1. *For any $n \times n$ matrix Λ_n with uniformly bounded column sums in absolute value, uniformly bounded $n \times k$ matrix Z_n, and if $u_n \sim N\left(0, \sigma^2 I\right)$ of dimension $n \times 1$, then $\frac{1}{\sqrt{n}} Z_n' \Lambda_n u_n = O_p(1)$.*

Lemma 2. $\mathbb{E}\left(u_n' \Lambda_n u_n\right) = \sigma^2 tr\left(\Lambda_n\right)$ *and* $Var\left(u_n' \Lambda_n u_n\right) = \left(\mu_4 - 3\sigma^4\right) vec_D'\left(\Lambda_n\right) vec_D\left(\Lambda_n\right) + \sigma^4\left[tr\left(\Lambda_n \Lambda_n'\right) + tr\left(\Lambda_n^2\right)\right]$.

Lemma 3. *Define $\Lambda_n^{-1} \equiv (S_n^{0'})^{-1} \tilde{\Lambda}_n^{-1}\left(S_n^0\right)^{-1}$, $(\Lambda_n^e)^{-1} \equiv (S_n^{e'}\left(Q_n, \theta_c^0\right))^{-1} \tilde{\Lambda}_n^{-1}\left(S_n^e\left(Q_n, \theta_c^0\right)\right)^{-1}$ and $\tilde{\Lambda}_n \equiv (S_n^{e'}\left(Q_n, \theta_c\right) R_n^{e'}\left(Q_n, \theta_c\right) P_n^e\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) S_n^e\left(Q_n, \theta_c\right))^{-1}$. Then, for any randomly distributed vector ϵ_n of dimension $n \times 1$ such that $\mathbb{E}\epsilon_i\epsilon_j = 0$ for $i \neq j$ with $\mathbb{E}\epsilon_i^2 < \infty$ and if link formation is independent, $\frac{1}{n}\mathbb{E}(\epsilon_n' \Lambda_n^{-1} \epsilon_n) = \frac{1}{n}\mathbb{E}(\epsilon_n' (\Lambda_n^e)^{-1} \epsilon_n) + o_p(1)$.*

Proof. For simplicity, consider $R_n^e\left(Q_n, \theta_c^0\right) = R_n^0 = I_n$. Proof generalizes immediately otherwise. Then $\frac{1}{n}\mathbb{E}\left\{\epsilon_n'\left[\Lambda_n^{-1} - \left(\mathbb{E}\Lambda_n\right)^{-1}\right]\epsilon_n\right\} = \frac{1}{n}\mathbb{E}\left\{\epsilon_n' \Lambda_n^{-1}\left[\mathbb{E}\Lambda_n - \Lambda_n\right]\left(\mathbb{E}\Lambda_n\right)^{-1}\epsilon_n\right\} = \frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^n \sum_{j=1}^n \epsilon_i\epsilon_j\left[\Lambda_n^{-1}\left(\mathbb{E}\Lambda_n - \Lambda_n\right)\mathbb{E}\Lambda_n^{-1}\right]_{ij}\right\} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{\epsilon_i^2\right\}\mathbb{E}\left[\Lambda_n^{-1}\left(\mathbb{E}\Lambda_n - \Lambda_n\right)\mathbb{E}\Lambda_n^{-1}\right]_{ii}$ as ϵ_i is independent of ϵ_j for $i \neq j$. Because $\mathbb{E}\left[\Lambda_n^{-1}\left(\mathbb{E}\Lambda_n - \Lambda_n\right)\left(\mathbb{E}\Lambda_n\right)^{-1}\right]_{ij} \xrightarrow{p} 0$ and $\mathbb{E}\left\{\epsilon_i^2\right\} < \infty$, then $\frac{1}{n}\mathbb{E}\left\{\epsilon_n'\left[\Lambda_n^{-1} - \left(\mathbb{E}\Lambda_n\right)^{-1}\right]\epsilon_n\right\} = o_p(1)$. Remains to show $\mathbb{E}\Lambda_n = \Lambda_n^e$. By definition, $\Lambda_n = \left(I_n - \lambda_0 W_n^0\right)\tilde{\Lambda}_n(I_n - \lambda_0 W_n^{0'}) = \tilde{\Lambda}_n - \lambda_0 W_n^0 \tilde{\Lambda}_n - \lambda_0 \tilde{\Lambda}_n W_n^{0'} + \lambda_0^2 W_n^0 \tilde{\Lambda}_n W_n^{0'}$. It follows that $\mathbb{E}\Lambda_n = \tilde{\Lambda}_n - \lambda_0 W_n^e\left(Q_n, \theta_c^0\right)\tilde{\Lambda}_n - \lambda_0 \tilde{\Lambda}_n W_n^{e'}\left(Q_n, \theta_c^0\right) + \lambda_0^2 \mathbb{E} W_n^0 \tilde{\Lambda}_n W_n^{0'} = \tilde{\Lambda}_n - \lambda_0 W_n^e\left(Q_n, \theta_c^0\right)\tilde{\Lambda}_n - \lambda_0 \tilde{\Lambda}_n W_n^{e'}\left(Q_n, \theta_c^0\right) + \lambda_0^2 W_n^e\left(Q_n, \theta_c^0\right)\tilde{\Lambda}_n W_n^{e'}\left(Q_n, \theta_c^0\right) = \Lambda_e$ where the second equality holds only if link formation is independent, i.e., if $\mathbb{E}\{W_j^0\}_{ik}\{W_j^0\}_{i'k'} = \mathbb{E}\{W_j^0\}_{ik}\mathbb{E}\{W_j^0\}_{i'k'}$ if either $i \neq i'$ or $k \neq k'$. □

Lemma 4. *Let ϵ_n be a $n \times 1$ stationary, ergodic process with $\mathbb{E}\epsilon_n = 0$. Then $\frac{1}{n}\mathbb{E}(\epsilon_n' \Lambda_n^{-1} \epsilon_n) = \frac{1}{n}\mathbb{E}(\epsilon_n'\left(\Lambda_n^e\right)^{-1}\epsilon_n) + o_p(1)$.*

Proof. Lemma 3 applies with the following modification. Given $\sum_{i=1}^n \sum_{j=1}^n \epsilon_i\epsilon_j\left[\Lambda_n^{-1}\left(\mathbb{E}\Lambda_n - \Lambda_n\right)\mathbb{E}\Lambda_n^{-1}\right]_{ij}$ is a weighted U-statistic, with summable weights, Theorem 3 of Hsing and Wu (2004) is applied to obtain convergence in probability to zero.

□

Lemma 5. $\frac{1}{n}\mathbb{E}\{\beta_0' Z_n^{0'}(S_n^{0'})^{-1} S_n^{e'}\left(Q_n, \theta_c^0\right) R_n^{e'}\left(Q_n, \theta_c^0\right) P_n^e\left(Q_n, \theta_c^0\right) R_n^e\left(Q_n, \theta_c^0\right) S_n^e\left(Q_n, \theta_c^0\right)\left(S_n^0\right)^{-1} Z_n^0 \beta_0\} = o_p(1)$.

Proof. Apply Lemma 4 with minor modifications twice. First, note that $\frac{1}{n}\mathbb{E}\{\beta_0' Z_n^{0'}(S_n^{0'})^{-1} S_n^{e'}\left(Q_n, \theta_c^0\right) R_n^{e'}\left(Q_n, \theta_c^0\right) P_n^e\left(Q_n, \theta_c^0\right) R_n^e\left(Q_n, \theta_c^0\right) S_n^e\left(Q_n, \theta_c^0\right)\left(S_n^0\right)^{-1} Z_n^0 \beta_0\} = \frac{1}{n}\mathbb{E}\{\beta_0' Z_n^{0'} R_n^{e'}\left(Q_n, \theta_c^0\right) P_n^e\left(Q_n, \theta_c^0\right) R_n^e\left(Q_n, \theta_c^0\right) Z_n^0 \beta_0\} + o_p(1)$. Secondly, $\frac{1}{n}\mathbb{E}\{\beta_0' Z_n^{0'} R_n^{e'}\left(Q_n, \theta_c^0\right) P_n^e\left(Q_n, \theta_c^0\right) R_n^e\left(Q_n, \theta_c^0\right) Z_n^0 \beta_0\} = \frac{1}{n}\mathbb{E}\left\{\beta_0' Z_n^{e'}\left(Q_n, \theta_c^0\right) R_n^{e'}\left(Q_n, \theta_c^0\right) P_n^e\left(Q_n, \theta_c^0\right) R_n^e\left(Q_n, \theta_c^0\right) Z_n^e\left(Q_n, \theta_c^0\right)\beta_0\right\} + o_p(1)$. Properties of projection matrix ensures $P_n^e\left(Q_n, \theta_c^0\right) R_n^e\left(Q_n, \theta_c^0\right) Z_n^e\left(Q_n, \theta_c^0\right) = 0$.

□

Lemma 6. $\frac{1}{n}\mathbb{E}\{\epsilon_n'(R_n^{0'})^{-1}(S_n^{0'})^{-1}S_n^{e'}\left(Q_n,\theta_c^0\right)R_n^{e'}\left(Q_n,\theta_c^0\right)P_n^e\left(Q_n,\theta_c^0\right)R_n^e\left(Q_n,\theta_c^0\right)S_n^e\left(Q_n,\theta_c^0\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\} = \sigma_0^2 + o_p\left(1\right)$.

Proof. Direct consequence of Lemma 4 taken with $\theta_c = \theta_c^0$. □

1.D.2 Derivation of pdf of networks.

For the *p1-reciprocity model*, the probability that random matrix W takes a particular value w is

$$\begin{aligned}
P\left(W = w\right) &= \prod_{i<j}\delta_F^{w_{ij}w_{ji}}\prod_{i<j}\delta_A^{w_{ij}\left(1-w_{ji}\right)+\left(1-w_{ij}\right)w_{ji}}\prod_{i<j}\delta_N^{\left(1-w_{ij}\right)\left(1-w_{ji}\right)} \\
&= \exp\left\{\ln\delta_F\sum_{i<j}w_{ij}w_{ji} + \ln\delta_A\sum_{i<j}w_{ij}\left(1-w_{ji}\right)+\left(1-w_{ij}\right)w_{ji} + \ln\delta_N\sum_{i<j}\left(1-w_{ij}\right)\left(1-w_{ji}\right)\right\} \\
&= \frac{1}{\kappa}\exp\left\{\theta_g^1\sum_{i\neq j}w_{ij} + \theta_g^2\sum_{i<j}w_{ij}w_{ji}\right\}
\end{aligned}$$

where $\theta_g^1 = \ln\frac{\delta_A}{\delta_N}$, $\theta_g^2 = \frac{\delta_F\delta_N}{\delta_A^2}$ and $\kappa = \left(\prod_{i<j}\delta_N\right)^{-1}$. Introducing dependence on sharing exogenous characteristics, the pdf is

$$\begin{aligned}
P\left(W = w\middle|\, Q = q\right) &= \prod_{i<j}\left(\delta_{1F}^{q_{ij}}\delta_{0F}^{1-q_{ij}}\right)^{w_{ij}w_{ji}}\prod_{i<j}\left(\delta_{1A}^{q_{ij}}\delta_{0A}^{1-q_{ij}}\right)^{\left(1-w_{ij}\right)w_{ji}+w_{ij}\left(1-w_{ji}\right)}\prod_{i<j}\left(\delta_{1N}^{q_{ij}}\delta_{0N}^{1-q_{ij}}\right)^{\left(1-w_{ij}\right)\left(1-w_{ji}\right)} \\
&= \exp\left\{\ln\left\{\prod_{i<j}\left(\delta_{1F}^{q_{ij}}\delta_{0F}^{1-q_{ij}}\right)^{w_{ij}w_{ji}}\prod_{i<j}\left(\delta_{1A}^{q_{ij}}\delta_{0A}^{1-q_{ij}}\right)^{\left(1-w_{ij}\right)w_{ji}+w_{ij}\left(1-w_{ji}\right)}\cdot\right.\right. \\
&\quad \left.\left.\cdot\prod_{i<j}\left(\delta_{1N}^{q_{ij}}\delta_{0N}^{1-q_{ij}}\right)^{\left(1-w_{ij}\right)\left(1-w_{ji}\right)}\right\}\right\} \\
&= \exp\left\{\sum_{i<j}w_{ij}w_{ji}\left(q_{ij}\ln\delta_{1F} + \left(1-q_{ij}\right)\ln\delta_{0F}\right) + \sum_{i<j}\left(1-w_{ij}\right)w_{ji}\left(q_{ij}\ln\delta_{1A} + \left(1-q_{ij}\right)\ln\delta_{0A}\right)\right. \\
&\quad + \sum_{i<j}w_{ij}\left(1-w_{ji}\right)\left(q_{ij}\ln\delta_{1A} + \left(1-q_{ij}\right)\ln\delta_{0A}\right) \\
&\quad \left. + \sum_{i<j}\left(1-w_{ij}\right)\left(1-w_{ji}\right)\left(q_{ij}\ln\delta_{1N} + \left(1-q_{ij}\right)\ln\delta_{0N}\right)\right\} \\
&= \frac{1}{\kappa}\exp\left\{\theta_g^1\sum_{i\neq j}w_{ij} + \theta_g^2\sum_{i\neq j}w_{ij}q_{ij} + \theta_g^3\sum_{i<j}w_{ij}w_{ji} + \theta_g^4\sum_{i<j}w_{ij}w_{ji}q_{ij}\right\}
\end{aligned}$$

where $\theta_g^1 = \ln\frac{\delta_{0A}}{\delta_{0N}}$, $\theta_g^2 = \ln\frac{\delta_{0N}\delta_{1A}}{\delta_{0A}\delta_{1N}}$, $\theta_g^3 = \ln\frac{\delta_{0F}\delta_{0N}}{\delta_{0A}^2}$, $\theta_g^4 = \ln\frac{\delta_{1F}\delta_{0A}^2\delta_{1N}}{\delta_{0F}\delta_{1A}^2\delta_{0N}}$ and

$$\kappa^{-1} = \exp\left\{\ln\left(\delta_{1N}\delta_{0N}\sum_{i<j}q_{ij}\right)\right\}\prod_{i<j}\delta_{0N}.$$

1.D.3 Lemma.

Lemma 7. *(i) Under Assumption 6, $Z_n^e\left(Q_n,\theta_c^0\right)$ and $G_n^e\left(Q_n,\theta_c^0\right)Z_n^e\left(Q_n,\theta_c^0\right)\beta_0$ are asymptotically independent. (ii) Define*

$$\gamma\left(Q_n,\theta_c\right) \quad = \quad \frac{1}{n}\mathbb{E}\{\beta_0' Z_n^{0'}(S_n^{0'})^{-1}\tilde{P}_n^e\left(Q_n,\theta_c\right)(S_n^0)^{-1}Z_n^0\beta_0\}$$

with $\tilde{P}_n^e\left(Q_n,\theta_c\right) = S_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)$. For every point $\theta_c\in\Theta_c$, the condition $\gamma\left(Q_n,\theta_c\right)>0$ holds.

Proof. (i) Under the assumption, full column rank means that the only solutions for the constants c_1, c_2 and c_3 in the equation $x_nc_1 + W_n^e\left(Q_n,\theta_c^0\right)x_nc_2 + G_n^e\left(Q_n,\theta_c^0\right)x_n\beta_{10}c_3 + G_n^e\left(Q_n,\theta_c^0\right)W_n^e\left(Q_n,\theta_c^0\right)x_n\beta_{20}c_3 = 0$ are $c_1 = c_2 = c_3 = 0$. Under the assumption that $G_n^e\left(Q_n,\theta_c^0\right)\equiv W_n^e\left(Q_n,\theta_c^0\right)\left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1} = \left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}W_n^e\left(Q,\theta_c^0\right)$, i.e., assuming symmetry of $W_n^e\left(Q_n,\theta_c^0\right)$, expression is equal to $x_nc_1 + W_n^e\left(Q_n,\theta_c^0\right)x_nc_2 + \left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}W_n^e\left(Q_n,\theta_c^0\right)x_n\beta_{10}c_3 + \left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\left(W_n^e\left(Q_n,\theta_c^0\right)\right)^2x_n\beta_{20}c_3$, then equivalent to assessing

$$\begin{aligned}
& S_n^e\left(Q_n,\theta_c^0\right)x_nc_1 + S_n^e\left(Q_n,\theta_c^0\right)W_n^e\left(Q_n,\theta_c^0\right)x_nc_2 + W_n^e\left(Q_n,\theta_c^0\right)x_n\beta_{10}c_3 + \left(W_n^e\left(Q_n,\theta_c^0\right)\right)^2x_n\beta_{20}c_3 \\
= \quad & \left(I_n + \lambda W_n^e\left(Q_n,\theta_c^0\right)\right)x_nc_1 + \left(I_n + \lambda W_n^e\left(Q_n,\theta_c^0\right)\right)W_n^e\left(Q_n,\theta_c^0\right)x_nc_2 + W_n^e\left(Q_n,\theta_c^0\right)x_n\beta_{10}c_3 \\
& + \left(W_n^e\left(Q_n,\theta_c^0\right)\right)^2x_n\beta_{20}c_3 \\
= \quad & x_nc_1 + \lambda W_n^e\left(Q_n,\theta_c^0\right)x_nc_1 + W_n^e\left(Q_n,\theta_c^0\right)x_nc_2 + \lambda\left(W_n^e\left(Q_n,\theta_c^0\right)\right)^2x_nc_2 + W_n^e\left(Q_n,\theta_c^0\right)x_n\beta_{10}c_3 \\
& + \left(W_n^e\left(Q_n,\theta_c^0\right)\right)^2x_n\beta_{20}c_3 \\
= \quad & x_nc_1 + W_n^e\left(Q_n,\theta_c^0\right)x_n\left(\lambda c_1 + c_2 + \beta_{10}c_3\right) + \left(W_n^e\left(Q_n,\theta_c^0\right)\right)^2x_n\left(\lambda c_2 + \beta_{20}c_3\right)
\end{aligned}$$

As x_n, $W_n^e\left(Q_n,\theta_c^0\right)x_n$ and $\left(W_n^e\left(Q_n,\theta_c\right)\right)^2x_n$ are linearly independent, $c_1 = 0$, then implying $c_2+\beta_{10}c_3 = 0$ and $\lambda c_2+\beta_{20}c_3 = 0$. Together, $\left(-\lambda\beta_{10}+\beta_{20}\right)c_3 = 0$. Given $\beta_{20}\neq\lambda\beta_{10}$, $c_3 = c_2 = 0$. If $W_n^e(Q_n,\theta_c^0)$ is not symmetric, premultiply the initial expression by $W_n^e(Q_n,\theta_c^0)S_n^e(Q_n,\theta_c^0)(W_n^e(Q_n,\theta_c^0))^{-1} = I_n + \lambda_0W_n^e(Q_n,\theta_c^0)$ and same result follows.

(ii) The reduced-form of the model evaluated at the true vector of parameter θ_0 is

$$y \quad = \quad \left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}Z_n^e\left(Q_n,\theta_c^0\right)\beta_0 + \left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\left(R_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\epsilon_n^e. \tag{1.28}$$

As $\left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1} = I_n+\lambda_0G_n^e\left(Q_n,\theta_c^0\right)$, where $G_n^e\left(Q_n,\theta_c^0\right)\equiv W_n^e\left(Q_n,\theta_c^0\right)\left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}$, the expression above can also be written as

$$y_n \quad = \quad Z_n^e\left(Q_n,\theta_c^0\right)\beta_0 + \lambda_0G_n^e\left(Q_n,\theta_c^0\right)Z_n^e\left(Q_n,\theta_c^0\right)\beta_0 + \left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\left(R_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\epsilon_n^e. \tag{1.29}$$

For separate identification of λ_0 and $\beta_0 = \left(\beta_{10}', \beta_{20}'\right)'$, it is necessary to guarantee that matrices $Z_n^e\left(Q_n, \theta_c^0\right)$

$$G_n^e\left(Q_n, \theta_c^0\right) Z_n^e\left(Q_n, \theta_c^0\right) \beta_0 \quad = \quad W_n^e\left(Q_n, \theta_c^0\right) \left(S_n^e\left(Q_n, \theta_c^0\right)\right)^{-1} Z_n^e\left(Q_n, \theta_c^0\right) \beta_0$$

are not dependent asymptotically. In turn, asymptotic independence of the concerned matrices is a necessary and sufficient condition for $\gamma\left(Q_n, \theta_c\right) > 0$, as I now show. Following Lemma 3, $\gamma\left(Q_n, \theta_c\right)$ is well approximated by $\gamma^e\left(Q_n, \theta_c\right)$, where

$$\gamma^e\left(Q_n, \theta_c\right) \quad = \quad \frac{1}{n}\beta_0' Z_n^{e'}\left(Q_n, \theta_c^0\right) \left(S_n^{e'}\left(Q_n, \theta_c^0\right)\right)^{-1} \tilde{P}_n^e\left(Q_n, \theta_c\right) \left(S_n^e\left(Q_n, \theta_c^0\right)\right)^{-1} Z_n^e\left(Q_n, \theta_c^0\right) \beta_0.$$

Given that $\tilde{P}_n^e\left(Q_n, \theta_c\right) = S_n^{e'}\left(Q_n, \theta_c\right) R_n^{e'}\left(Q_n, \theta_c\right) P_n^e\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) S_n^e\left(Q_n, \theta_c\right)$ is positive definite, then $\gamma\left(Q_n, \theta_c\right) = 0$ if, and only if, $\left(S_n^e\left(Q_n, \theta_c^0\right)\right)^{-1} Z_n^e\left(Q_n, \theta_c^0\right) \beta_0 = 0$, which is equivalent to $Z_n^e\left(Q_n, \theta_c^0\right) \beta_0 + \lambda_0 G_n^e\left(Q_n, \theta_c^0\right) Z_n^e\left(Q_n, \theta_c^0\right) \beta_0 = 0$ using $\left(S_n^e\left(Q_n, \theta_c^0\right)\right)^{-1} = I_n + \lambda_0 G_n^e\left(Q_n, \theta_c^0\right)$ or, essentially, that $Z_n^e\left(Q_n, \theta_c^0\right)$ and $G_n^e\left(Q_n, \theta_c^0\right) Z_n^e\left(Q_n, \theta_c^0\right) \beta_0$ are asymptotically independent. □

1.D.4 Theorem 1.

Proof. (Uniform Convergence). The goal is to show that the concentrated log-likelihood $(n)^{-1}\left[\ln \mathcal{L}_n^c\left(\theta_c\right) - Q_n\left(\theta_c\right)\right]$ converges uniformly to zero on Θ_c, where $F_n\left(\theta_c\right) = \max_{\beta, \sigma^2} \mathbb{E} \ln \mathcal{L}_n^c\left(\theta_c\right)$, that is,

$$\sup_{\theta_c \in \Theta_c} \left| \frac{1}{n} \ln \mathcal{L}_n\left(\theta_c\right) - \frac{1}{n} F_n\left(\theta_c\right) \right| = \sup_{\theta_c \in \Theta_c} \left| \ln \tilde{\sigma}^2\left(\theta_c\right) - \ln \hat{\sigma}^2\left(\theta_c\right) \right| = o_p(1).$$

In first place, misspecification component in $\hat{\sigma}^2\left(Q_n, \theta_c\right)$ is made explicit. Given $S_n^e\left(Q_n, \theta_c\right) = I_n - \lambda W_n^e\left(Q_n, \theta_c\right)$ and $\left(S_n^0\right)^{-1} = \lambda_0 G_n^0 + I_n$ where $G_n^0 = W_n^0\left(S_n^0\right)^{-1}$, then $S_n^e\left(Q_n, \theta_c\right)\left(S_n^0\right)^{-1} = \lambda_0 G_n^0 + I_n - \lambda\lambda_0 W_n^e\left(Q_n, \theta_c\right) G_n^0 - \lambda W_n^e\left(Q_n, \theta_c\right)$. Now $\lambda_0 W_n^e\left(Q_n, \theta_c\right) = \lambda_0 W_n^0 + \lambda_0\left(W_n^e\left(Q_n, \theta_c\right) - W_n^0\right) = I_n - S_n^0 + \lambda_0\left(W_n^e\left(Q_n, \theta_c\right) - W_n^0\right)$ and $S_n^e\left(Q_n, \theta_c\right)\left(S_n^0\right)^{-1} = \left(\lambda_0 - \lambda\right) G_n^0 + I_n + B_n\left(Q_n, \theta_c\right)$ where the misspecification term is defined $B_n\left(Q_n, \theta_c\right) \equiv \lambda\left(W_n^0 - W_n^e\left(Q_n, \theta_c\right)\right) + \lambda\lambda_0\left(W_n^0 - W_n^e\left(Q_n, \theta_c\right)\right) G_n^0 = \lambda\left(W_n^0 - W_n^e\left(Q_n, \theta_c\right)\right)\left(I + \lambda_0 G_n^0\right)$. Therefore, using the reduced-form equation $S_n^e\left(Q_n, \theta_c\right) y_n = S_n^e\left(Q_n, \theta_c\right)\left(S_n^0\right)^{-1} Z_n^0 \beta_0 + S_n^e\left(Q_n, \theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1} \epsilon_n$,

$$\begin{aligned} P_n^e\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) S_n^e\left(Q_n, \theta_c\right) y_n \quad = \quad & P_n^e\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) Z_n^0 \beta_0 + \left(\lambda_0 - \lambda\right) P_n^e\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) G_n^0 Z_n^0 \beta_0 \\ & + P_n^e\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) B_n\left(Q_n, \theta_c\right) Z_n^0 \beta_0 \\ & + P_n^e\left(Q_n, \theta_c\right) R_n^e\left(Q_n, \theta_c\right) S_n^e\left(Q_n, \theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1} \epsilon_n. \end{aligned}$$

Given that $\hat{\sigma}^2\left(Q_n,\theta_c\right)=\frac{1}{n}y_n'S_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)y_n$, $\hat{\sigma}^2\left(Q_n,\theta_c\right)=\sum_{i=1}^{10}K_i\left(Q_n,\theta_g\right)$, where

$$
\begin{aligned}
K_1\left(Q_n,\theta_g\right) &= \frac{1}{n}\left[R_n^e\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]\\
K_2\left(Q_n,\theta_g\right) &= \frac{2}{n}\left(\lambda_0-\lambda\right)\left[R_n^e\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)G_n^0Z_n^0\beta_0\right]\\
K_3\left(Q_n,\theta_g\right) &= \frac{2}{n}\left[R_n^e\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)B_n\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]\\
K_4\left(Q_n,\theta_g\right) &= \frac{2}{n}\left[R_n^e\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]\\
K_5\left(Q_n,\theta_g\right) &= \frac{1}{n}\left(\lambda_0-\lambda\right)^2\left[R_n^e\left(Q_n,\theta_c\right)G_n^0Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)G_n^0Z_n^0\beta_0\right]\\
K_6\left(Q_n,\theta_g\right) &= \frac{2}{n}\left(\lambda_0-\lambda\right)\left[R_n^e\left(Q_n,\theta_c\right)G_n^0Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)B_n\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]\\
K_7\left(Q_n,\theta_g\right) &= \frac{2}{n}\left(\lambda_0-\lambda\right)\left[R_n^e\left(Q_n,\theta_c\right)G_n^0Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]\\
K_8\left(Q_n,\theta_g\right) &= \frac{1}{n}\left[R_n^e\left(Q_n,\theta_c\right)B_n\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)B_n\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]\\
K_9\left(Q_n,\theta_g\right) &= \frac{2}{n}\left[R_n^e\left(Q_n,\theta_c\right)B_n\left(Q_n,\theta_c\right)Z_n^0\beta_0\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]\\
K_{10}\left(Q_n,\theta_g\right) &= \frac{1}{n}\left[R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]'P_n^e\left(Q_n,\theta_c\right)\left[R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]
\end{aligned}
$$

Given Lemma 1, $K_4\left(Q,\theta_g\right)$, $K_7\left(Q,\theta_g\right)$ and $K_9\left(Q,\theta_g\right)$ are $o_p\left(1\right)$. Remains to show the problem in expectation. The concentrators are

$$
\begin{aligned}
\tilde{\beta}\left(Q_n,\theta_c\right) &= \left[Z_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)Z_n^e\left(Q_n,\theta_c\right)\right]^{-1}\cdot\\
&\quad\cdot Z_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\mathbb{E}y_n\\
\tilde{\sigma}^2\left(Q_n,\theta_c\right) &= \frac{1}{n}\mathbb{E}\left\{\left[S_n^e\left(Q_n,\theta_c\right)y_n-Z_n^e\left(Q_n,\theta_c\right)\tilde{\beta}\left(\theta_c\right)\right]'R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)\cdot\right.\\
&\quad\left.\cdot R_n^e\left(Q_n,\theta_c\right)\left[S_n^e\left(Q_n,\theta_c\right)y_n-Z_n^e\left(Q_n,\theta_c\right)\tilde{\beta}\left(\theta_c\right)\right]\right\}.
\end{aligned}
$$

Noticing $P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)Z_n^e\left(Q_n,\theta_c\right)=0$, the expectation

$$
\begin{aligned}
\tilde{\sigma}^2\left(Q_n,\theta_c\right) &= \frac{1}{n}\mathbb{E}\left\{y_n'S_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)y_n\right\}\\
&= \frac{1}{n}\mathbb{E}\left\{\left[\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]'S_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right\}\\
&\quad+\frac{1}{n}\mathbb{E}\left\{\left[\left(S_n^0\right)^{-1}Z_n^0\beta_0\right]'S_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}Z_n^0\beta_0\right\}\\
&= \frac{1}{n}\mathbb{E}\left\{\left[\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]'S_n^{e'}\left(Q_n,\theta_c\right)R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)S_n^e\left(Q_n,\theta_c\right)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon\right\}\\
&\quad+\frac{1}{n}\mathbb{E}\left\{\beta_0'Z_n^{0'}\left[\left(\lambda_0-\lambda\right)G_n^0+I_n+B\left(Q_n,\theta_c\right)\right]'R_n^{e'}\left(Q_n,\theta_c\right)P_n^e\left(Q_n,\theta_c\right)R_n^e\left(Q_n,\theta_c\right)\right.\\
&\quad\left.\left[\left(\lambda_0-\lambda\right)G_n^0+I_n+B_n\left(Q_n,\theta_c\right)\right]Z_n^0\beta_0\right\}
\end{aligned}
$$

and so $\tilde{\sigma}^2(Q,\theta_c) = \sum_{i=1}^{7} \tilde{K}_i(Q,\theta_c)$ with

$$\begin{aligned}
\tilde{K}_1(Q_n,\theta_c) &= \frac{1}{n}\mathbb{E}\left\{\epsilon_n'\left(R_n^{0'}\right)^{-1}\left(S_n^{0'}\right)^{-1} S_n^{e'}(Q_n,\theta_c)\, R_n^{e'}(Q_n,\theta_c)\, P_n^e(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, S_n^e(Q_n,\theta_c)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right\} \\
\tilde{K}_2(Q_n,\theta_c) &= \frac{1}{n}\mathbb{E}\left\{(\lambda_0-\lambda)^2\beta_0' Z_n^{0'} G_n^{0'} R_n^{e'}(Q_n,\theta_c)\, P_n^e(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, G_n^0 Z_n^0\beta_0\right\} \\
\tilde{K}_3(Q_n,\theta_c) &= \frac{2}{n}\mathbb{E}\left\{(\lambda_0-\lambda)\beta_0' Z_n^{0'} G_n^{0'} R_n^{e'}(Q_n,\theta_c)\, P_n^e(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, Z_n^0\beta_0\right\} \\
\tilde{K}_4(Q_n,\theta_c) &= \frac{2}{n}\mathbb{E}\left\{(\lambda_0-\lambda)\beta_0' Z_n^{0'} G_n^{0'} R_n^{e'}(Q_n,\theta_c)\, P_n^e(Q_n,\theta_c)\, R^e(Q_n,\theta_c)\, B(Q_n,\theta_c)\, Z_n^0\beta_0\right\} \\
\tilde{K}_5(Q_n,\theta_c) &= \frac{1}{n}\mathbb{E}\left\{\beta_0' Z_n^{0'} R_n^{e'}(Q_n,\theta_c)\, P_n^e(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, Z_n^0\beta_0\right\} \\
\tilde{K}_6(Q_n,\theta_c) &= \frac{2}{n}\mathbb{E}\left\{\beta_0' Z_n^{0'} R_n^{e'}(Q_n,\theta_c)\, P^e(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, B_n(Q_n,\theta_c)\, Z_n^0\beta_0\right\} \\
\tilde{K}_7(Q_n,\theta_c) &= \frac{1}{n}\mathbb{E}\left\{\beta_0' Z_n^{0'} B_n(Q_n,\theta_c)' R_n^{e'}(Q_n,\theta_c)\, P_n^e(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, B_n(Q_n,\theta_c)\, Z_n^0\beta_0\right\}.
\end{aligned}$$

By Lemma 2, $\tilde{K}_1(Q_n,\theta_c) = K_{10}(Q_n,\theta_c) + o_p(1)$. Also, $\tilde{K}_2(Q_n,\theta_c) = K_5(Q_n,\theta_c) + o_p(1)$, $\tilde{K}_3(Q_n,\theta_c) = K_2(Q_n,\theta_c) + o_p(1)$, $\tilde{K}_4(Q_n,\theta_c) = K_6(Q_n,\theta_c) + o_p(1)$, $\tilde{K}_5(Q_n,\theta_c) = K_1(Q_n,\theta_c) + o_p(1)$, $\tilde{K}_6(Q_n,\theta_c) = K_3(Q_n,\theta_c) + o_p(1)$ and $\tilde{K}_7(Q_n,\theta_c) = K_8(Q_n,\theta_c) + o_p(1)$. As a consequence, $\hat{\sigma}^2(Q_n,\theta_c) - \tilde{\sigma}^2(Q_n,\theta_c) = o_p(1)$ uniformly on θ_c. Convergence is uniform on the parameter space as λ, ρ and θ_c appear as polynomial factors.

(Identification for $\lambda = \lambda_0$). Consider the non-stochastic auxiliary model $y_j = \lambda_0 W_j^e(Q_j,\theta_c^0)\, y_j + x_j\beta_1 + W_j^e(Q_j,\theta_c^0)\, x_j\beta_2 + v_j$ where true neighboring matrices are given by expected network at true parameter values, $W_j^0 = W_j^e(Q_j,\theta_c^0)$ and $M_j^0 = M_j^e(Q_j,\theta_c^0)$. Its likelihood is

$$\ln\mathcal{L}_n^{**}(\theta) = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) + \ln\left|S_n^e(Q_n,\theta)\right| + \ln\left|R_n^e(Q_n,\theta)\right| - \frac{1}{2\sigma^2}\sum_{j=1}^{v}\epsilon_j^{e'}(Q_j,\theta)\,\epsilon_j^e(Q_j,\theta)$$

where $\epsilon_j^e(Q_j,\theta) = R_j^e(Q_j,\theta)\left(S_j^e(Q_j,\theta)\, y_j - x_j\beta_1 - W_j^e(Q_j,\theta)\, x_j\beta_2\right)$. As usual, parameters β and σ^2 can be concentrated out of the likelihood. The concentrators are given by

$$\begin{aligned}
\hat{\beta}^{**}(Q_n,\theta_c) &= \left[Z_n^{e'}(Q_n,\theta_c)\, R_n^{e'}(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, Z_n^e(Q_n,\theta_c)\right]^{-1} Z_n^{e'}(Q_n,\theta_c)\, R_n^{e'}(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, S_n^e(Q_n,\theta_c)\, y_n \\
\hat{\sigma}^{**2}(Q_n,\theta_c) &= \frac{1}{n}\left[S^e(Q_n,\theta_c)\, y_n - Z^e(Q_n,\theta_c)\,\hat{\beta}(\theta_c)\right]' R_n^{e'}(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\left[S_n^e(Q_n,\theta_c)\, y_n - Z^e(Q_n,\theta_c)\,\hat{\beta}(\theta_c)\right] \\
&= \frac{1}{n} y_n' S_n^{e'}(Q_n,\theta_c)\, R_n^{e'}(Q_n,\theta_c)\, P_n^e(Q_n,\theta_c)\, R_n^e(Q_n,\theta_c)\, S_n^e(Q_n,\theta_c)\, y_n
\end{aligned}$$

The final form for the concentrated likelihood is $\ln\mathcal{L}_n^{c**}(\theta_c) = -\frac{n}{2}\left(\ln(2\pi)+1\right) - \frac{n}{2}\ln\hat{\sigma}^2(\theta_c) + \ln\left|S_n^e(Q_n,\theta)\right| + \ln\left|R_n^e(Q_n,\theta)\right|$. The problem in expectation $F_n^{**}(\theta) = \max_{\beta,\sigma^2}\mathbb{E}\ln\mathcal{L}_n^{**}(\theta)$ is $F_n^{**}(\theta) = -\frac{n}{2}\left(\ln(2\pi)+1\right) + \ln\left|S_n^e(Q_n,\theta)\right| + \ln\left|R_n^e(Q_n,\theta)\right| -$

$\frac{n}{2}\tilde{\sigma}^{**2}(\theta)$, where $\tilde{\sigma}^{**2}(Q_n,\theta_c)$ is given by

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\left\{y_n' S_n^{e'}(Q_n,\theta_c) R_n^{e'}(Q_n,\theta_c) P_n^e(Q_n,\theta_c) R_n^e(Q_n,\theta_c) S_n^e(Q_n,\theta_c) y_n\right\} \\
= \; & \frac{1}{n}\mathbb{E}\left\{\epsilon_n' \left(R_n^{e'}\left(Q_n,\theta_c^0\right)\right)^{-1}\left(S_n^{e'}\left(Q_n,\theta_c^0\right)\right)^{-1} S_n^{e'}(Q_n,\theta_c) R_n^{e'}(Q_n,\theta_c) P_n^e \cdot\right. \\
& \left.\cdot (Q_n,\theta_c) R_n^e(Q_n,\theta_c) S_n^e(Q_n,\theta_c)\left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\left(R_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\epsilon_n\right\} \\
& +\frac{1}{n}\mathbb{E}\left\{\beta_0' Z_n^{e'}\left(Q_n,\theta_c^0\right)\left(S_n^{e'}\left(Q_n,\theta_c^0\right)\right)^{-1} S_n^{e'}(Q_n,\theta_c) R_n^{e'}(Q_n,\theta_c) P_n^e(Q_n,\theta_c)\cdot\right. \\
& \left.\cdot R_n^e(Q_n,\theta_c) S_n^e(Q_n,\theta_c)\left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1} Z_n^e\left(Q_n,\theta_c^0\right)\beta_0\right\} \\
= \; & \frac{\sigma^2}{n}\operatorname{tr}\left\{\left(R_n^{e'}\left(Q_n,\theta_c^0\right)\right)^{-1}\left(S_n^{e'}\left(Q_n,\theta_c^0\right)\right)^{-1} S_n^{e'}(Q_n,\theta_c) R_n^{e'}(Q_n,\theta_c) P_n^e(Q_n,\theta_c)\cdot\right. \\
& \left.\cdot R_n^e(Q_n,\theta_c) S_n^e(Q_n,\theta_c)\left(S_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\left(R_n^e\left(Q_n,\theta_c^0\right)\right)^{-1}\right\} \\
& +\frac{(\lambda_0-\lambda)^2}{n}\beta_0' Z_n^{e'}\left(Q_n,\theta_c^0\right) G_n^{e'}\left(Q_n,\theta_c^0\right) R_n^{e'}(Q_n,\theta_c) P_n^e(Q_n,\theta_c) R_n^e(Q_n,\theta_c) G_n^e\left(Q_n,\theta_c^0\right) Z_n^e\left(Q_n,\theta_c^0\right)\beta_0.
\end{aligned}$$

By Jensen's Inequality, $F_n^{**}(\theta) \leq F_n^{**}(\theta_0)$. Identification in the original model follows from

$$\frac{1}{n}F_n(\theta_c) - \frac{1}{n}F_n\left(\theta_c^0\right) \quad = \quad \frac{1}{n}\left[F_n^{**}(\theta_c) - F_n^{**}\left(\theta_c^0\right)\right] + \frac{1}{2}\left[\ln\sigma^{**2}(\theta_c) - \ln\tilde{\sigma}^2(\theta_c) + \ln\tilde{\sigma}^2\left(\theta_c^0\right) - \ln\sigma^{**2}\left(\theta_c^0\right)\right].$$

It is immediate that $\sigma^{**2}\left(\theta_c^0\right) = \sigma_0^2$. Lemmas 5 and 6 imply that $\tilde{\sigma}^2\left(\theta_c^0\right) = \sigma_0^2$. Notice also

$$\begin{aligned}
\tilde{\sigma}^2(\theta_c) \quad = \quad & \frac{1}{n}\cdot\mathbb{E}\left\{\epsilon_n'\left(R_n^{0'}\right)^{-1}\left(S_n^{0'}\right)^{-1} S_n^{e'}(Q_n,\theta_c) R_n^{e'}(Q_n,\theta_c) P_n^e(Q_n,\theta_c) R_n^e(Q_n,\theta_c) S_n^e(Q_n,\theta_c)\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right\} \\
& +\frac{1}{n}\mathbb{E}\left\{\beta_0' Z_n^{0'}\left(S_n^{0'}\right)^{-1} S_n^{e'}(Q_n,\theta_c) R_n^{e'}(Q_n,\theta_c) P_n^e(Q_n,\theta_c) R_n^e(Q_n,\theta_c) S_n^e(Q_n,\theta_c)\left(S_n^0\right)^{-1} Z_n^0\beta_0\right\}.
\end{aligned}$$

Finally, Lemma 3 and Assumption 6 imply $\ln\sigma^{**2}(\theta_c) - \ln\tilde{\sigma}^2(\theta_c) < 0$. This completes the proof. □

1.D.5 Theorem 2.

Proof. Jacobian and Hessian matrices are given in Appendix 1.C. The asymptotic distribution can be obtained from a Taylor expansion around the point $\frac{\partial\ln\mathcal{L}^e\left(\hat{\theta}\middle|y_n,x_n,Q_n\right)}{\partial\theta} = 0$. For a point $\tilde{\theta}$ between $\hat{\theta}$ and θ_0,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \quad = \quad \left[-\frac{1}{n}\frac{\partial\ln\mathcal{L}^e(\tilde{\theta}\big|\, y_n,x_n,Q_n)}{\partial\theta\partial\theta'}\right]^{-1}\frac{1}{\sqrt{n}}\frac{\partial\ln\mathcal{L}^e\left(\theta_0\right|y_n,x_n,Q_n)}{\partial\theta}.$$

(Showing $\frac{1}{n}\frac{\partial^2\ln\mathcal{L}^e\left(\tilde{\theta}\middle|y_n,x_n,Q_n\right)}{\partial\theta\partial\theta'} \xrightarrow{p} \frac{1}{n}\frac{\partial^2\ln\mathcal{L}^e\left(\theta_0\middle|y_n,x_n,Q_n\right)}{\partial\theta\partial\theta'}$*).* Convergence is shown explicitly for three terms: $\frac{\partial^2\ln\mathcal{L}^e\left(\tilde{\theta}\right)}{\partial\lambda\partial\beta_1'}$,

$\frac{\partial^2 \ln \mathcal{L}^e(\tilde{\theta})}{\partial\lambda\partial\sigma^2}$ and $\frac{\partial^2 \ln \mathcal{L}^e(\tilde{\theta})}{\partial\lambda^2}$; other terms can be shown with little or no modifications. For

$$\begin{aligned}
\frac{1}{n}\left\{\frac{\partial^2 \ln \mathcal{L}^e(\tilde{\theta})}{\partial\lambda\partial\beta_1'} - \frac{\partial^2 \ln \mathcal{L}^e(\theta_0)}{\partial\lambda\partial\beta_1'}\right\} &= \frac{1}{n\sigma_0^2} y_n' W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) x_n - \frac{1}{n\tilde{\sigma}^2} y_n' W_n^{e'}\left(Q_n,\tilde{\theta}\right) R_n^{e'}\left(Q_n,\tilde{\theta}\right) x_n \\
&= \frac{1}{n}\left[\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}^2}\right] y_n' W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) x_n \\
&\quad + \frac{1}{n\tilde{\sigma}^2} y_n' \left[W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) - W_n^{e'}(Q_n,\tilde{\theta}) R_n^{e'}(Q_n,\tilde{\theta})\right] x_n.
\end{aligned}$$

The argument follows by noticing $W_n^e\left(Q_n,\theta_0\right)$ and $R_n^e\left(Q_n,\theta_0\right)$ are row and column-sum bounded, so $\frac{1}{n} y_n' W_n^{e'}\left(Q_n,\theta_0\right)$ $R_n^{e'}\left(Q_n,\theta_0\right) x_n = O_p\left(1\right)$, while by continuity of the inverse, $\left[\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}^2}\right] = o_p\left(1\right)$. The second term converges in probability as $\frac{1}{n\tilde{\sigma}^2} y_n' [W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) - W_n^{e'}(Q_n,\tilde{\theta}) R^{e'}(Q_n,\tilde{\theta})] x_n = \frac{1}{n\tilde{\sigma}^2} \beta_0' Z_n^{0'} \left(S_n^{0'}\right)^{-1} \left[W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right)\right.$ $\left. - W_n^{e'}(Q_n,\tilde{\theta}) R_n^{e'}(Q_n,\tilde{\theta})\right] x_n + o_p\left(1\right)$. Given that $Z_n^0 = \left[x_n; W_n^0 x_n\right]$, x_n is non stochastic, W_n^0 is row and column-sum bounded, and $\left[W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) - W_n^{e'}(Q_n,\tilde{\theta}) R_n^{e'}(Q_n,\tilde{\theta})\right] = o_p\left(1\right)$, it has been shown that $\frac{1}{n}\left\{\frac{\partial^2 \ln \mathcal{L}^e(\tilde{\theta})}{\partial\lambda\partial\beta_1'}\right.$ $\left. - \frac{\partial^2 \ln \mathcal{L}^e(\theta_0)}{\partial\lambda\partial\beta_1'}\right\} = o_p\left(1\right)$. The next term is

$$\begin{aligned}
\frac{1}{n}\left\{\frac{\partial^2 \ln \mathcal{L}^e(\tilde{\theta})}{\partial\lambda\partial\sigma^2} - \frac{\partial^2 \ln \mathcal{L}^e\left(\theta_0\right)}{\partial\lambda\partial\sigma^2}\right\} &= \frac{1}{n\sigma_0^4} y_n' W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) \epsilon_n^e\left(Q_n,\theta_0\right) - \frac{1}{n\tilde{\sigma}^4} y_n' W_n^{e'}(Q_n,\tilde{\theta}) R_n^{e'}(Q_n,\tilde{\theta}) \epsilon_n^e(Q_n,\tilde{\theta}) \\
&= \frac{1}{n}\left[\frac{1}{\sigma_0^4} - \frac{1}{\tilde{\sigma}^4}\right] y_n' W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) \epsilon_n^e\left(Q_n,\theta_0\right) \\
&\quad + \frac{1}{n\tilde{\sigma}^4} y_n' \left[W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) \epsilon_n^e\left(Q_n,\theta_0\right) - W_n^{e'}(Q_n,\tilde{\theta}) R_n^{e'}(Q_n,\tilde{\theta}) \epsilon_n^e(Q_n,\tilde{\theta})\right] \\
&= \frac{1}{n}\left[\frac{1}{\sigma_0^4} - \frac{1}{\tilde{\sigma}^4}\right] y_n' W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) \epsilon_n^e\left(Q_n,\theta_0\right) \\
&\quad + \frac{1}{n\tilde{\sigma}^4} y_n' \left[W_n^{e'}\left(Q_n,\theta_0\right) R_n^{e'}\left(Q_n,\theta_0\right) - W_n^{e'}(Q_n,\tilde{\theta}) R_n^{e'}(Q_n,\tilde{\theta})\right] \epsilon_n^e\left(Q_n,\theta_0\right) + o_p\left(1\right)
\end{aligned}$$

as $\epsilon_n^e(Q_n,\tilde{\theta}) = R_n^e(Q_n,\tilde{\theta})(S_n^e(Q_n,\tilde{\theta}) y_n - x_n\tilde{\beta}_1 - W_n^e(Q_n,\tilde{\theta}) x_n\tilde{\beta}_2) - R_n^e\left(Q_n,\theta_0\right)\left(S_n^e\left(Q_n,\theta_0\right) y_n - x_n\beta_{10} - W_n^e\left(Q_n,\theta_0\right) x_n\beta_{20}\right)$ $+\epsilon_n^e\left(Q_n,\theta_0\right) = R_n^e(Q_n,\tilde{\theta})([S_n^e(Q_n,\tilde{\theta}) - S_n^e\left(Q_n,\theta_0\right)] y_n - x_n[\tilde{\beta}_1 - \beta_{10}] - W_n^e(Q_n,\tilde{\theta}) x_n\tilde{\beta}_2 + W_n^e\left(Q_n,\theta_0\right) x_n\beta_{20}) +$ $R_n^e(Q_n,\tilde{\theta})[S_n^e(Q_n,\tilde{\theta}) - S_n^e\left(Q_n,\theta_0\right)] y_n - R_n^e(Q_n,\tilde{\theta}) x_n[\tilde{\beta}_1 - \beta_{10}] - R_n^e(Q_n,\tilde{\theta}) W_n^e(Q_n,\tilde{\theta})\ x_n[\beta_{20} - \tilde{\beta}_2] + R_n^e(Q_n,\tilde{\theta})[W_n^e\left(Q_n,\theta_0\right) -$ $W_n^e(Q_n,\tilde{\theta})] x_n\beta_{20}$ $+[R_n^e(Q_n,\tilde{\theta}) - R_n^e\left(Q_n,\theta_0\right)] S_n^e\left(Q_n,\theta_0\right) y_n$ $-[R_n^e(Q_n,\tilde{\theta}) - R_n^e\left(Q_n,\theta_0\right)] x_n\beta_{10} - [R_n^e(Q_n,\tilde{\theta}) - R_n^e\left(Q_n,\theta_0\right)]$ $W_n^e\left(Q_n,\theta_0\right) x_n\beta_{20} + \epsilon_n^e\left(Q_n,\theta_0\right)$, $[S_n^e(Q_n,\tilde{\theta}) - S_n^e\left(Q_n,\theta_0\right)]$, $[W_n^e\left(Q_n,\theta_0\right) - W_n^e(Q_n,\tilde{\theta})]$ and $[R_n^e(Q_n,\tilde{\theta}) - R_n^e\left(Q_n,\theta_0\right)] = o_p\left(1\right)$, and $R_n^e(Q_n,\tilde{\theta})$ is row and column-sum bounded, then $\frac{1}{n}\left\{\frac{\partial^2 \ln \mathcal{L}^e(\tilde{\theta})}{\partial\lambda\partial\sigma^2} - \frac{\partial^2 \ln \mathcal{L}^e(\theta_0)}{\partial\lambda\partial\sigma^2}\right\} = o_p\left(1\right)$.

By Mean Value Theorem, defining $G_j\left(\lambda,\theta_g\right) = \left(S_j^e\left(Q_j,\theta\right)\right)^{-1} W_j^e\left(Q_j,\theta\right)$, $\operatorname{tr}\left\{G_n^2\left(\bar{\lambda},\bar{\theta}_g\right)\right\} = \operatorname{tr}\left\{G_n^2\left(\lambda_0,\theta_g^0\right)\right\} +$ $2\operatorname{tr}\left\{G_n^3\left(\bar{\lambda},\bar{\theta}_g\right)\right\}\left(\bar{\lambda} - \lambda_0\right) + 2\operatorname{tr}\left\{\nabla_{\theta_g} W_n^e\left(\bar{\lambda},\bar{\theta}_g\right) S_n^e\left(\bar{\lambda},\bar{\theta}_g\right)^{-1} G_n\left(\bar{\lambda},\bar{\theta}_g\right)\right\}\left(\bar{\theta}_g - \theta_0\right) + 2\lambda\operatorname{tr}\left\{W_n^e\left(\bar{\lambda},\bar{\theta}_g\right) \nabla_{\theta_g} W_n^e\left(\bar{\lambda},\bar{\theta}_g\right)\right.$

$G_n\left(\bar{\lambda},\bar{\theta}_g\right)\}\left(\bar{\theta}_g-\theta_0\right)$ then

$$\begin{aligned}\frac{1}{n}\left\{\frac{\partial^2\ln\mathcal{L}^e(\tilde{\theta})}{\partial\lambda^2}-\frac{\partial^2\ln\mathcal{L}^e\left(\theta_0\right)}{\partial\lambda^2}\right\} \quad = \quad & 2\mathrm{tr}\left\{G_n^3\left(\bar{\lambda},\bar{\theta}_g\right)\right\}\left(\bar{\lambda}-\lambda_0\right)+2\mathrm{tr}\left\{\nabla_{\theta_g}W_n^e\left(\bar{\lambda},\bar{\theta}_g\right)S_n^e\left(\bar{\lambda},\bar{\theta}_g\right)^{-1}G_n\left(\bar{\lambda},\bar{\theta}_g\right)\right\}\left(\bar{\theta}_g-\theta_0\right)\\ & +2\lambda\mathrm{tr}\left\{W_n^e\left(\bar{\lambda},\bar{\theta}_g\right)\nabla_{\theta_g}W_n^e\left(\bar{\lambda},\bar{\theta}_g\right)G_n\left(\bar{\lambda},\bar{\theta}_g\right)\right\}\left(\bar{\theta}_g-\theta_0\right)\\ & +\left[\frac{1}{\sigma_0^2}-\frac{1}{\tilde{\sigma}^2}\right]\sum_{j=1}^{v}y_j'W_j^{e'}\left(Q_j,\theta_0\right)R_j^{e'}\left(Q_j,\theta_0\right)R_j^e\left(Q_j,\theta_0\right)W_j^e\left(Q_j,\theta_0\right)y_j\\ & -\frac{1}{\tilde{\sigma}^2}\sum_{j=1}^{v}y_j'\left[W_j^{e'}(Q_j,\tilde{\theta})R_j^{e'}(Q_j,\tilde{\theta})R_j^e(Q_j,\tilde{\theta})-W_j^{e'}(Q_j,\theta_0)R_j^{e'}(Q_j,\theta_0)R_j^e(Q_j,\theta_0)\right]\\ & W_j^e(Q_j,\tilde{\theta})y_j.\end{aligned}$$

By similar arguments, as above, $\frac{1}{n}\left\{\frac{\partial^2\ln\mathcal{L}^e(\tilde{\theta})}{\partial\lambda^2}-\frac{\partial^2\ln\mathcal{L}^e(\theta_0)}{\partial\lambda^2}\right\}=o_p\left(1\right)$.

(Showing $\frac{1}{n}\frac{\partial^2\ln\mathcal{L}^e(\,\theta_0|y_n,x_n,Q_n)}{\partial\theta\partial\theta'}\xrightarrow{p}\mathbb{E}\left(\frac{1}{n}\frac{\partial^2\ln\mathcal{L}^e(\,\theta_0|y_n,x_n,Q_n)}{\partial\theta\partial\theta'}\right)$*).* Terms that generically fit into the format $\omega_x\left(\theta\right)=\frac{1}{n}\varphi'\Delta\left(\theta\right)\varphi$, where φ is non-stochastic vector of dimension n and Δ is a stochastic matrix of conformable dimension can be shown to $\mathbb{V}\left\{\omega_x\left(\theta\right)\right\}\xrightarrow{p}0$. For example, $-\frac{\sigma^2}{n}\frac{\partial^2\ln\mathcal{L}^e(\theta_0)}{\partial\lambda\partial\beta_1'}=\frac{1}{n}x_n'R_n^e\left(Q_n,\theta_0\right)W_n^e\left(Q_n,\theta_0\right)y=\frac{1}{n}x_n'R_n^e\left(Q_n,\theta_0\right)$ $W_n^e\left(Q_n,\theta_0\right)\left[\left(S_n^0\right)^{-1}Z_n^0\beta_0+\left(S_n^0\right)^{-1}\left(R_n^0\right)^{-1}\epsilon_n\right]=\frac{1}{n}x_n'R_n^e\left(Q_n,\theta_0\right)\ W_n^e\left(Q_n,\theta_0\right)\left(S_n^0\right)^{-1}x_n\beta_{10}+\frac{1}{n}x_n'R^e\left(Q_n,\theta_0\right)$ $W_n^e\left(Q_n,\theta_0\right)\left(S_n^0\right)^{-1}W_n^0x_n\beta_{20}+o_p\left(1\right)$. Defining $x_n^{(l)}$ as the l-th column of x_n,

$$\omega_{xl}\left(\theta\right)\equiv\frac{1}{\sigma^2 n}x_n^{(l)'}R_n^e\left(Q_n,\theta_0\right)W_n^e\left(Q_n,\theta_0\right)\left(S_n^0\right)^{-1}x_n^{(l)}=\frac{1}{\sigma^2 n}\sum_{i=1}^{n}\sum_{j=1}^{n}x_{n,i}^{(l)}x_{n,j}^{(l)}\left(R_n^e\left(Q_n,\theta^0\right)W_n^e\left(Q_n,\theta_0\right)\left(S_n^0\right)^{-1}\right)_{ij}.$$

If elements of $\Delta\left(\theta\right)$ are approximately independent (taking, for example, $\left(S_n^0\right)^{-1}=I_n+\lambda W_n^0$ as the first-order Series Expansion), then

$$\mathbb{V}\left(\omega_{xl}\right)=\left[\frac{1}{\sigma^2 n}\right]^2\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_{n,i}^{(l)}x_{n,j}^{(l)}\right)^2\mathbb{V}\left\{\left(R_n^e\left(Q_n,\theta^0\right)W_n^e\left(Q_n,\theta_0\right)\left(S_n^0\right)^{-1}\right)_{ij}\right\}$$

Noticing $\mathbb{V}\left(W_n^0\right)$ is a matrix of constants, $R_n^e\left(Q_n,\theta^0\right)W_n^e\left(Q_n,\theta_0\right)$ is column and row-sum bounded, then $\mathbb{V}\left\{\cdot\right\}$ goes to zero and so does $\mathbb{V}\left(\gamma_l\right)$. An equivalent argument goes through if terms in the middle contains matrix of derivatives. Terms that generically fit into $\omega_\epsilon\left(\theta\right)=\frac{1}{n}\epsilon_n'\Delta\left(\theta\right)\epsilon_n$, for example, $-\frac{\sigma^2}{n}\frac{\partial^2\ln\mathcal{L}^e(\theta)}{\partial\lambda\partial\sigma^2}=\frac{1}{\sigma^2 n}y_n'W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)$ $\epsilon_n^e\left(Q_n,\theta\right)=\frac{1}{\sigma^2 n}\epsilon_n'(S_n^{0'})^{-1}W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)\left(\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}y_n\ -Z_n^e\left(Q_n,\theta\right)\beta\right)=\frac{1}{\sigma^2 n}\epsilon_n'(S_n^{0'})^{-1}$ $W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)\left(\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}y_n\right)+o_p\left(1\right)=\frac{1}{\sigma^2 n}\epsilon_n'(S_n^{0'})^{-1}W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)$ $\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}\left(S_n^0\right)^{-1}\epsilon_n+o_p\left(1\right)$ by Lemma 1, and straightforward adaptation of Lemma 3, converges to

$$\mathbb{E}\left\{-\frac{\sigma^2}{n}\frac{\partial^2\ln\mathcal{L}^e\left(\theta\right)}{\partial\lambda\partial\sigma^2}\right\}\quad=\quad\frac{1}{n}\mathrm{tr}\left\{\mathbb{E}\left((S_n^{0'})^{-1}W_n^{e'}\left(Q_n,\theta\right)R_n^{e'}\left(Q_n,\theta\right)R_n^e\left(Q_n,\theta\right)\left(S_n^e\left(Q_n,\theta\right)\right)^{-1}\left(S_n^0\right)^{-1}\right)\right\}.$$

(Asymptotic distribution). Given existence of higher order moments of ϵ_n, the Central Limit Theorem in Kelejian and

Prucha (2001) can be applied to show that $\frac{1}{\sqrt{n}}\frac{\partial \ln \mathcal{L}^e(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \Omega_\theta)$. Given non-singularity of the Hessian matrix as guaranteed by global identification condition in Theorem 1, it follows that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \quad \xrightarrow{d} \quad N\left(0, \Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1}\right).$$

□

1.D.6 Proposition 1.

Proof. (i). Starting from the definition of the social multiplier,

$$\begin{aligned} \varphi\left(x_n; W_n^e\left(Q_n, \theta_c^+\right), \beta_{10}, \beta_+\right) &= \sum_{j=1}^{\infty} \lambda_+^{j-1}\left(W_n^e\left(Q_n, \theta_c^+\right)\right)^j x_n\left(\lambda_+\beta_{10} + \beta_{2+}\right) = \\ \sum_{j=1}^{\infty} \lambda_0\lambda_+^{-1}\lambda_0^{j-1}\left(W_n^e\left(Q_n, \theta_c^0\right)\right)^j x_n\left(\lambda_+\beta_{10} + \beta_{2+}\right) &= \sum_{j=1}^{\infty} \lambda_0^{j-1}\left(W_n^e\left(Q_n, \theta_c^0\right)\right)^j x_n\left(\lambda_0\beta_{10} + \lambda_0\lambda_+^{-1}\beta_{2+}\right) = \\ \sum_{j=1}^{\infty} \lambda_0^{j-1}\left(W_n^e\left(Q_n, \theta_c^0\right)\right)^j x_n\left(\lambda_0\beta_{10} + \beta_{20}\right) &= \varphi\left(x_n; W^e\left(Q_n, \theta_c^0\right), \lambda_0, \beta_{10}, \beta_{20}\right) \end{aligned} \tag{1.30}$$

where the penultimate equality follows by $W_n^e(Q_n, \theta_c^+)x_n\beta_{2+} - W_n^e\left(Q_n, \theta_c^0\right)x_n\beta_{20} = \lambda_0\lambda_+^{-1}W_n^e\left(Q_n, \theta_c^0\right)x_n\beta_{2+}$ $-W_n^e\left(Q_n, \theta_c^0\right)x_n\beta_{20} = W_n^e\left(Q_n, \theta_c^0\right)x_n\ (\lambda_0\lambda_+^{-1}\beta_{2+} - \beta_{20}) = 0$. (ii). Define $\Phi^*\left(\theta | y_n, x_n, Q_n\right) = \{\tilde{\theta} \in \Theta : Q_n(\tilde{\theta}) = Q_n\left(\theta\right)\}$. Sets $\Phi\left(\theta^0 \middle| y_n, x_n\right) = \Phi^*\left(\theta^0 \middle| y_n, x_n\right)$, as I now show. Inclusion $\Phi\left(\theta^0 \middle| y_n, x_n\right) \subseteq \Phi^*\left(\theta^0 \middle| y_n, x_n\right)$ is immediate from the first part. The reverse $\Phi^*\left(\theta^0 \middle| y_n, x_n\right) \subseteq \Phi\left(\theta^0 \middle| y_n, x_n\right)$ follows from a contradiction: suppose there is a θ^* such that $\theta^* \in \Phi\left(\theta^* | y_n, x_n\right)$ and $\theta^* \notin \Phi^*\left(\theta^* | y_n, x_n\right)$. By construction and Jensen's inequality, $Q_n\left(\theta^*\right) < Q_n\left(\theta^0\right)$. Observation of the reduced-form implies $\epsilon_n^e\left(Q_n, \theta_c^*\right) = \epsilon^e\left(Q_n, \theta^0\right)$, $\ln\left|S_n^e\left(Q_n, \theta^*\right)\right| = \ln\left|S_n^e\left(Q_n, \theta^0\right)\right|$ and $\ln\left|R_n^e\left(Q_n, \theta^*\right)\right| = \ln\left|R_n^e\left(Q, \theta^0\right)\right|$, and so $Q_n\left(\theta^*\right) = Q_n\left(\theta^0\right)$, a contradiction. Therefore, given that $\Phi\left(\theta^0 \middle| y_n, x_n\right) = \Phi^*\left(\theta^0 \middle| y_n, x_n\right)$, for any $\theta_c \in \Phi^*\left(\theta^0 \middle| y_n, x_n\right)$, and, by definition, $\Phi^*\left(\theta^0 \middle| y_n, x_n\right) = \Theta_0$, the result is proven. □

1.D.7 Theorem 3.

Proof. For parts (1) and (2), see Theorem 3.2 and Lemma 3.1 of Chernozhukov et al. (2007). By construction, and uniform convergence of Theorem 1 conditions C.1 with $a_n = n$, degeneracy property C.3 and condition C.4 therein are satisfied. Condition C.2 is guaranteed by uniform convergence and boundness of the objective function on a compact set Θ. Parts (3) and (4) are immediate corollaries. □

1.D.8 Example 3.

The full model is $y_j = \lambda_0 W_j^0 y_j + x_j\beta_{10} + W_j^0 x_j \beta_{20} + \epsilon_j$ with reduced form $y_j = (S_j^0)^{-1}x_j\beta_{10} + (S_j^0)^{-1}W_j^0 x_j\beta_{20} + (S_j^0)^{-1}\epsilon_j$. Then

$$y_j - \mathbb{E}y_j \quad = \quad ((S_j^0)^{-1} - \mathbb{E}(S_j^0)^{-1})x_j\beta_{10} + ((S_j^0)^{-1}W_j^0 - \mathbb{E}\{(S_j^0)^{-1}W_j^0\})x_j\beta_{20} + (S_j^0)^{-1}\epsilon_j$$

and $\mathbb{V}y_j = \mathbb{E}((y_j - \mathbb{E}y_j)(y_j - \mathbb{E}y_j)')$ is

$$\begin{aligned}
\mathbb{V}y_j \quad = \quad & \mathbb{E}\left\{((S_j^0)^{-1} - \mathbb{E}(S_j^0)^{-1}x_j\beta_{10}\beta_{10}'x_j'((S_j^0)^{-1} - \mathbb{E}(S_j^{0'})^{-1})\right\} \\
& +2\mathbb{E}\left\{((S_j^0)^{-1} - \mathbb{E}\{(S_j^0)^{-1}\})x_j\beta_{10}\beta_{20}'x_j'((S_j^0)^{-1}W_j^0 - \mathbb{E}\{(S_j^0)^{-1}W_j^0\})'\right\} \\
& +\mathbb{E}\left\{((S_j^0)^{-1}W_j^0 - \mathbb{E}\{(S_j^0)^{-1}W_j^0\})x_j\beta_{20}\beta_{20}'x_j'((S_j^0)^{-1}W_j^0 - \mathbb{E}\{(S_j^0)^{-1}W_j^0\})'\right\} + \mathbb{E}\left\{(S_j^0)^{-1}\epsilon_j\epsilon_j'(S_j^{0'})^{-1}\right\}.
\end{aligned}$$

Denote these terms sequentially as A_j, B_j, C_j and D_j. $A_j = s_j x_j^{11} s_j'$, where $s_j = ((S_j^0)^{-1} - \mathbb{E}\{(S_j^0)^{-1}\})$ and $x_j^{11} = x_j\beta_{10}\beta_{10}'x_j'$. Then

$$A_j \quad = \quad \begin{bmatrix} \sum_{i,k}\mathbb{E}\{s_{1i}s_{1k}\}x_{ik}^{11} & \cdots & \sum_{i,k}\mathbb{E}\{s_{1i}s_{nk}\}x_{ik}^{11} \\ \vdots & \ddots & \vdots \\ \sum_{i,k}\mathbb{E}\{s_{ni}s_{1k}\}x_{ik}^{11} & \cdots & \sum_{i,k}\mathbb{E}\{s_{ni}s_{nk}\}x_{ik}^{11} \end{bmatrix}$$

where s_{ik} denotes the (i,k)th element of s_j, and similarly for x_j^{11}. Matrix s_j can be approximated $s = I + \lambda_0 W_j^0 + \lambda_0^2(W_j^0)^2 + \cdots - (I + \lambda_0\mathbb{E}W_j^0 + \lambda_0^2\mathbb{E}(W_j^0)^2 + \cdots) \approx \lambda_0(W_j^0 - W_j^e(\theta_j^0))$. Hence s_{ik} is dependent of $s_{i'k'}$ if, and only if, $i = i'$ and $k = k'$. Take w_{ik} as the (i,k)th element of W_j^0. This simplifies term A_j to

$$A_j \quad = \quad \lambda^2 \begin{bmatrix} \sum_i \mathbb{V}\{w_{1i}\}x_{ii}^{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_i \mathbb{V}\{w_{ni}\}x_{ii}^{11} \end{bmatrix}$$

which then implies $A_j = \text{diag}\left(\lambda^2\mathbb{V}\{W_j\}\,\text{diag}(x_j^{11})\right)$. Proceeding in a similar fashion, $B_j = s_j x_j^{12} s_j^{*'}$ with $x_j^{12} = x_j\beta_{10}\beta_{20}'x_j'$ and $s_j^* = W_j^0 + \lambda_0(W_j^0)^2 + \lambda_0^2(W_j^0)^3 + \cdots - (W_j^e(\theta_0) + \lambda_0\mathbb{E}(W_j^0)^2 + \lambda_0^2\mathbb{E}(W_j^0)^3 + \cdots) \approx W_j^0 - W_j^e(\theta_0)$

$$B_j \quad = \quad 2\begin{bmatrix} \sum_{i,k}\mathbb{E}\{s_{1i}s_{1k}^*\}x_{ik}^{12} & \cdots & \sum_{i,k}\mathbb{E}\{s_{1i}s_{nk}^*\}x_{ik}^{12} \\ \vdots & \ddots & \vdots \\ \sum_{i,k}\mathbb{E}\{s_{ni}s_{1k}^*\}x_{ik}^{12} & \cdots & \sum_{i,k}\mathbb{E}\{s_{ni}s_{nk}^*\}x_{ik}^{12} \end{bmatrix} \quad = \quad 2\lambda \begin{bmatrix} \sum_i \mathbb{V}\{w_{1i}\}x_{ii}^{12} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_i \mathbb{V}\{w_{ni}\}x_{ii}^{12} \end{bmatrix}$$

and then $B_j = \text{diag}\left(2\lambda \mathbb{V}\left\{W_j\right\}\text{diag}(x_j^{12})\right)$. The second equality uses independence between Bernoulli trials. For C_j,

$$\begin{aligned} C_j &= \begin{bmatrix} \sum_i \mathbb{E}\left\{s_{1i}^{*2}\right\}x_{ii}^{22} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_i \mathbb{E}\left\{s_{ni}^{*2}\right\}x_{ii}^{22} \end{bmatrix} = \begin{bmatrix} \sum_i \mathbb{V}\left\{w_{1i}\right\}x_{ii}^{22} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_i \mathbb{V}\left\{w_{ni}\right\}x_{ii}^{22} \end{bmatrix} \\ &= \text{diag}\left(\mathbb{V}\left\{W_j\right\}\text{diag}\left(x_j^{22}\right)\right) \end{aligned}$$

Lastly,

$$\begin{aligned} D_j &= \begin{bmatrix} \sum_{i,j} \mathbb{E}\left\{s_{1i}s_{1j}\right\}\mathbb{E}\left\{e_{ij}\right\} & \cdots & \sum_{i,j} \mathbb{E}\left\{s_{1i}s_{nj}\right\}\mathbb{E}\left\{e_{ij}\right\} \\ \vdots & \ddots & \vdots \\ \sum_{i,j} \mathbb{E}\left\{s_{ni}s_{1j}\right\}\mathbb{E}\left\{e_{ij}\right\} & \cdots & \sum_{i,j} \mathbb{E}\left\{s_{ni}s_{nj}\right\}\mathbb{E}\left\{e_{ij}\right\} \end{bmatrix} = \begin{bmatrix} \sum_i \mathbb{E}\left\{s_{1i}^2\right\}\sigma^2 & \cdots & \sum_i \mathbb{E}\left\{s_{1i}s_{ni}\right\}\sigma^2 \\ \vdots & \ddots & \vdots \\ \sum_i \mathbb{E}\left\{s_{ni}s_{1i}\right\}\sigma^2 & \cdots & \sum_i \mathbb{E}\left\{s_{ni}^2\right\}\sigma^2 \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} \sum_i \mathbb{E}\left\{s_{1i}^2\right\} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_i \mathbb{E}\left\{s_{ni}^2\right\} \end{bmatrix} = \lambda^2\sigma^2 \begin{bmatrix} \sum_i \mathbb{V}\left\{w_{1i}\right\} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_i \mathbb{V}\left\{w_{ni}\right\} \end{bmatrix} + \sigma^2 I_{n_j} \\ &= \lambda^2\sigma^2\text{diag}\left(\mathbb{V}\left\{W_j\right\}\iota_{n_j}\right) + \sigma^2 I_{n_j}. \end{aligned}$$

The entire expression reads $\mathbb{V}y_j = \text{diag}\left(\mathbb{V}\left\{W_j\right\}(\lambda^2\text{diag}(x_j^{11}) + 2\lambda\text{diag}(x_j^{12}) + \lambda^2\sigma^2\iota_{n_j})\right) + \sigma^2 I_{n_j}$. Using Theorem 6 of Rothenberg (1971, p. 585), suffices that the jacobian of matrix of restrictions has rank equal to the unknown parameters. The identified set can be translated, in this case, as $\delta\lambda = \delta_0\lambda_0$ and $\beta_2\lambda^{-1} = \beta_{20}\lambda_0^{-1}$, where the combination of the parameters in the right hand side is identified from data; parameters β_{10} and σ_0^2 are point-identified. The jacobian then reads

$$\mathcal{J}(\theta) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \delta & 0 & 0 & \lambda & 0 \\ -\beta_2\lambda^{-2} & 0 & \lambda^{-1} & 0 & 0 \\ J_{K1}(\theta) & J_{K2}(\theta) & J_{K3}(\theta) & J_{K4}(\theta) & J_{K5}(\theta) \end{bmatrix}$$

where

$$\begin{aligned}
J_{K1}(\theta) &= 2n_j^{-1}\delta_1(1-\delta_1)\lambda\left(\iota_{n_j}'\mathrm{diag}(x_j^{11})+n_j\sigma^2\right)\\
J_{K2i}(\theta) &= n_j^{-1}\delta_1(1-\delta_1)\left(\lambda^2\iota_{n_j}'\frac{\partial\mathrm{diag}(x_j^{11})}{\partial\beta_{1i}}+2\lambda\iota_{n_j}'\frac{\partial\mathrm{diag}(x_j^{12})}{\partial\beta_{1i}}\right)\\
J_{K3i}(\theta) &= n_j^{-1}\delta_1(1-\delta_1)\left(2\lambda\iota_{n_j}'\frac{\partial\mathrm{diag}(x_j^{12})}{\partial\beta_{2i}}+\iota_{n_j}'\frac{\partial\mathrm{diag}(x_j^{22})}{\partial\beta_{2i}}\right)\\
J_{K4}(\theta) &= n_j^{-1}(1-2\delta_1)\left(\lambda^2\iota_{n_j}'\mathrm{diag}(x_j^{11})+2\lambda\iota_{n_j}'\mathrm{diag}(x_j^{12})+\iota_{n_j}'\mathrm{diag}(x_j^{22})+n_j\lambda^2\sigma^2\right)\\
J_{K5}(\theta) &= \delta_1(1-\delta_1)-n_j\lambda^2+1.
\end{aligned}$$

Identification is guarateed with $\mathrm{rank}(\mathcal{J}(\theta))=K$, where K is the number of parameters in the structural model. Given σ_0^2 is identified, the last equation gives a solution for δ_1 and λ. Linear independence is guaranteed if the only column vector c that satisfies $\mathcal{J}(\theta)c=0$ is $c=0$. For the case of one exogenous covariate, this immediately implies $c_2=c_5=0$. We then have $c_1\delta+c_4\lambda=0$, $-c_1\beta\lambda^{-2}+c_3\lambda^{-1}=0$ and $c_1J_{K1}(\theta)+c_3J_{K3}(\theta)+c_4J_{K4}(\theta)=0$. Substituting out c_1 and c_3 in the third equation, one obtains the condition that $c_4\left[-\lambda\delta^{-1}J_{K1}(\theta)-\lambda\delta^{-1}\beta J_{K3}(\theta)+J_{K4}(\theta)\right]=0$. If $\lambda\neq0$, it is equivalent to $-\lambda\delta^{-1}J_{K1}(\theta)-\lambda\delta^{-1}\beta J_{K3}(\theta)+J_{K4}(\theta)\neq0$ at θ_0. This condition is empirically testable for all $\theta\in\Theta_0$, which is sufficient as $\theta_0\in\Theta_0$.

1.D.9 Theorem 4.

Proof. (Consistency). Because $\hat\Theta$ converges to Θ_0 in the Hausdorff metric, $\hat\Theta\subseteq\Theta_0^\epsilon$ for $\Theta_0^\epsilon=\{\theta\in\Theta:d(\theta,\Theta_0)\leq\epsilon\}$ with $\epsilon=o(1)$ and $\epsilon\geq0$. It follows that

$$\hat\theta = \arg\min_{\theta\in\Theta_0}\left(\sum_{j=1}^{v}S^{-1}\sum_{s=1}^{S}q_{s,j}(y,\theta)\right)'\Omega\left(\sum_{j=1}^{v}S^{-1}\sum_{s=1}^{S}q_{s,j}(y,\theta)\right)+o_p(1)$$

When S and v are going to infinity,

$$v^{-2}\left(\sum_{j=1}^{v}S^{-1}\sum_{s=1}^{S}q_{s,j}(y,\theta)\right)'\Omega\left(\sum_{j=1}^{v}S^{-1}\sum_{s=1}^{S}q_{s,j}(y,\theta)\right)\xrightarrow{a.s.}\left(\mathbb{E}_y^0\mathbb{E}_{W,e}q_{s,j}(y,\theta)\right)'\Omega\left(\mathbb{E}_y^0\mathbb{E}_{W,e}q_{s,j}(y,\theta)\right)$$

where $\mathbb{E}_{W,e}$ is the conditional expectation taken with respect to the distribution of W and e, given y and x and $\mathbb{E}_y^0$ is the expectation with respect to the true distribution of y, given x. Given that $\left(\mathbb{E}_y^0\mathbb{E}_{W,e}q_{s,j}(y,\theta)\right)'\Omega\left(\mathbb{E}_y^0\mathbb{E}_{W,e}q_{s,j}(y,\theta)\right)=\left(\mathbb{E}_y^0q_j(y,\theta)\right)'\Omega\left(\mathbb{E}_y^0q_j(y,\theta)\right)$ and $\mathbb{E}_y^0q_j(y,\theta_0)=0$ only at θ_0, consistency follows.

(Asymptotic normality). In the cases where $S\to\infty$ fast enough, results follow from standard asymptotic theory and Gouriéroux and Monfort (1997, Ch. 2). $\sqrt{n}(\hat\theta-\theta_0)\xrightarrow{d}N(0,\Sigma^*)$, where $\Sigma_n=(G_n'\Omega_nG_n)^{-1}G_n'\Omega_nO_n\Omega_nG_n(G_n'\Omega_nG_n)^{-1}$, $G_n=\mathbb{E}\nabla_\theta q_j(y_n,\theta_0)$, $O_n=\mathbb{E}q_j(y_n,\theta_0)q_j(y_n,\theta_0)'$ and $\Sigma=\lim_{n\to\infty}\Sigma_n$. Optimal weight matrix is $\Omega_n^*=O_n^{-1}$ and, in this

case, $\Sigma_n^* = (G_n' (\Omega_n^*)^{-1} G_n)^{-1}$ and $\Sigma^* = \lim_{n\to\infty} \Sigma_n^*$. When it can be shown that the local maximum is unique, the estimator can also be seen as the solution to

$$\hat{\theta}^\star \quad = \quad \arg\min_{\theta\in\Theta} \left(\sum_{j=1}^{v} S^{-1} \sum_{s=1}^{S} q_{s,j}^\star\left(y,\theta\right) \right)' \Omega^\star \left(\sum_{j=1}^{v} S^{-1} \sum_{s=1}^{S} q_{s,j}^\star\left(y,\theta\right) \right)$$

where $q_{s,j}^\star\left(y,\theta\right) = \left[\nabla_\theta \ln \mathcal{L}^e\left(\theta\right) \quad q_{s,j}\left(y,\theta\right)\right]'$ and $\Omega^\star$ is a weight matrix of conformable dimensions with possibly arbitrary large weights for the first-order conditions, so that the restriction $\theta \in \hat{\Theta}$ is implemented. In the case where $S \to \infty$ fast enough, given identification, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \Sigma^{\star *}\right)$, where $\Sigma_n^\star = (G_n^{\star\prime} \Omega_n^\star G_n^\star)^{-1} G_n^{\star\prime} \Omega_n^\star O_n^\star \Omega_n^\star G_n^\star (G_n^{\star\prime} \Omega_n^\star G_n^\star)^{-1}$, $G^\star = \mathbb{E}\nabla_\theta q_j^\star(y_j,\theta_0)$, $O^\star = \mathbb{E} q_j^\star(y,\theta_0) q_j^\star(y,\theta_0)'$ and $q_j^\star(y,\theta_0) = \lim_{S\to\infty} S^{-1} \sum_{s=1}^{S} q_{s,j}^\star(y,\theta_0)$ and $\Sigma^\star = \lim_{n\to\infty} \Sigma_n^\star$. Using optimal matrix $\Omega_n^{\star *} = \left(O_n^\star\right)^{-1}$, $\Sigma_n^{\star *} = (G_n^{\star\prime} \left(\Omega_n^{\star *}\right)^{-1} G_n^\star)^{-1}$, $\lim_{n\to\infty} \Sigma_n^{\star *}$. □

1.E Algorithms.

1.E.1 Bootstrap for $c_n\left(\alpha\right)$ and $c_n^f\left(\alpha\right)$

In the case of i.i.d. data, Bugni (2010) proposes a bootstrap algorithm correction consistent for $c_n\left(\alpha\right)$ and adaptable to $c_n^f\left(\alpha\right)$. In the current case, spatial dependence or social interactions in groups prevents immediate application of methods described therein. Instead, I propose bootstrapping at the group-level j, while maintaining within-group observations $i = 1, \ldots, n_j$. In this way, dependence of observed data is preserved. Apart from the straightforward modification proposed here, proofs can be found in the aforementioned paper.

Algorithm 1. *(Bugni (2010) bootstrap). In order to produce confidence regions with coverage probability* $1-\alpha$*,* $\alpha \in (0,1)$*, for* Θ_0*, denoted* $\hat{\Theta}_\alpha^B$ *for a bootstrapped sample of arbitrary size* B*, follow the steps:*

Step 1. Estimate the identified set $\hat{\Theta} = \{\theta \in \Theta : L_n\left(\theta \mid y_n, x_n, Q_n\right) = 0\}$.

Step 2. Define the bootstrapped sample $b = 1, \ldots, B$*, sampling* v *groups with replacement from the data and denote bootstrapped sample* $\{y_n^b, x_n^b, Q_n^b\}$*. Compute*

$$\hat{c}_n^b \quad = \quad \sup_{\theta\in\hat{\Theta}} \sqrt{n} \left(L_n(\theta \mid y_n^b, x_n^b, Q_n^b) - L_n\left(\theta \mid y_n, x_n, Q_n\right) \right).$$

Step 3. Let $\hat{c}_n^B\left(\alpha\right)$ *be the* α *quantile of the empirical distribution of* $\{\hat{c}_n^1, \ldots, \hat{c}_n^B\}$*. The* $(1-\alpha)$ *confidence set for the identified set is*

$$\hat{\Theta}_\alpha^B \quad = \quad \left\{ \theta \in \Theta : \sqrt{n} L\left(\theta \mid y_n, x_n, Q_n\right) \leq \hat{c}_n^B\left(1-\alpha\right) \right\}$$

Next, I produce an adaptation of the algorithm to be able to generate confidence regions for the image of the identified set under known function f, hence completing the statistical toolkit necessary for implementation of remarks 2 and 3.

Algorithm 2. *(Adaptation of Bugni (2010) bootstrap for projection under* f*). The modified algorithm to produce confidence regions with probability* $1-\alpha$*,* $\alpha \in (0,1)$*, for the projection of* Θ_0 *under known function* f*,* Υ_0^f*, denoted* $\hat{\Upsilon}_\alpha^B$*, for a bootstrapped sample of arbitrary size* B *is:*

Step 1. Estimate the projection of the identified set $\hat{\Upsilon} = \left\{ \upsilon \in \Upsilon : \inf_{\theta \in f^{-1}(\upsilon)} L_n \left(\theta | \, y_n, x_n, Q_n \right) = 0 \right\}$.

Step 2. Define the bootstrapped sample $b = 1, \ldots, B$, *sampling* v *groups with replacement from the data and denote bootstrapped sample* $\{y_n^b, x_n^b, Q_n^b\}$. *Compute*

$$\hat{c}_n^{f,b} = \sup_{\upsilon \in \hat{\Upsilon}} \inf_{\theta \in f^{-1}(\upsilon)} \sqrt{n} \left(L_n(\theta | \, y_n^b, x_n^b, Q_n^b) - L_n \left(\theta | \, y_n, x_n, Q_n \right) \right).$$

Step 3. Let $\hat{c}_n^{f,B} \left(\alpha \right)$ *be the* α *quantile of the empirical distribution of* $\{\hat{c}_n^{f,1}, \ldots, \hat{c}_n^{f,B}\}$. *The* $(1 - \alpha)$ *confidence set for the projected identified set* Υ_0 *is*

$$\hat{\Upsilon}_\alpha^{f,B} = \left\{ \upsilon \in \Upsilon : \inf_{\theta \in f^{-1}(\upsilon)} \sqrt{n} L \left(\theta | \, y_n, x_n, Q_n \right) \leq \hat{c}_n^{f,B} \left(1 - \alpha \right) \right\}.$$

1.E.2 Main algorithms

Algorithm 3. *If* λ_0 *is known and there are at least three distinct group sizes* n_j, *follow the steps:*

Step 1. Maximize the concentrated pseudo-likelihood

$$\ln \mathcal{L}_n^c \left(\theta_c | \, y_n, x_n, Q_n \right) \quad = \quad -\frac{n}{2} \left(\ln \left(2\pi \right) + 1 \right) - \frac{n}{2} \ln \hat{\sigma}^2 \left(Q_n, \theta_c \right) + \left| S_n^e \left(Q_n, \theta_c \right) \right| + \left| R^e \left(Q_n, \theta_c \right) \right|$$

with respect to θ_g, *where*

$$\hat{\sigma}^2(Q_n, \theta_c) \quad = \quad \frac{1}{n} y_n' S_n^{e'}(Q_n, \theta_c) R_n^{e'}(Q_n, \theta_c) P_n^e(Q_n, \theta_c) R_n^e(Q_n, \theta_c) S_n^e(Q_n, \theta_c) y_n$$

and $P_n^e(Q_n, \theta_c) = I_n - R_n^e(Q_n, \theta_c) Z_n^e(Q_n, \theta_c) (Z_n^{e'}(Q_n, \theta_c) R_n^{e'}(Q_n, \theta_c) R_n^e(Q_n, \theta_c) Z_n^e(Q_n, \theta_c))^{-1} Z_n^{e'}(Q_n, \theta_c) R_n^{e'}(Q_n, \theta_c)$. *Obtain the full solution* $\hat{\theta} = (\hat{\theta}_c', \hat{\beta}(\hat{\theta}_c)', \hat{\sigma}^2(\hat{\theta}_c))'$, *where* $\hat{\theta}_c \equiv \textit{arg}\max_{\theta \in \Theta_c} \ln \mathcal{L}_n^c \left(\theta_c | \, y_n, x_n, Q_n \right)$ *and*

$$\hat{\beta}(\hat{\theta}_c) \quad = \quad (Z_n^{e'}(Q_n, \theta_c) R_n^{e'}(Q_n, \theta_c) R_n^e(Q_n, \theta_c) Z_n^e(Q_n, \theta_c))^{-1} Z_n^{e'}(Q_n, \theta_c) R_n^{e'}(Q_n, \theta_c) R_n^e(Q_n, \theta_c) S_n^e(Q_n, \theta_c) y_n.$$

Calculate and store the expected network $\hat{W}_n^e = W_n^e(Q_n, \hat{\theta})$.

Step 2. (C.I. of structural parameters). Calculate the asymptotic variance given by Theorem 1. The full expressions of the Jacobian and Hessian are given in Appendix 1.C or can be numerically approximated.

Step 3. (Network spillovers). Network spillovers are calculated as

$$\varphi(x_n, \hat{\theta}) \quad = \quad (I - \lambda_0 \hat{W}_n^e)^{-1} (x_n \hat{\beta}_1 + \hat{W}_n x_j \hat{\beta}_2) - x_n \hat{\beta}_1.$$

Confidence intervals follow from a simple Delta Method, $\sqrt{n^*}(\varphi(x_n, \hat{\theta}) - \varphi(x_n, \theta_0)) \xrightarrow{d} N(0, \nabla\varphi(x_n, \theta_0) \Sigma^{-1}(\lambda_0) \Omega(\lambda_0) \Sigma^{-1}(\lambda_0) \nabla\varphi(x_n, \theta_0))$.

Step 4. (Network data validity). When network data are available, a Delta Method also is employed to provide confidence intervals for the null hypothesis $\mathcal{H}_0 : \delta_1 - \delta_0 = 0$.

Algorithm 4. *The following algorithm generalizes for the case in which* λ_0 *is unknown. If there are at least three distinct group sizes* n_j, *follow the steps:*

Step 1. Select a candidate λ_0.

Step 2. Maximize the concentrated pseudo-likelihood

$$\ln \mathcal{L}_n^c \left(\theta_c | \, y_n, x_n, Q_n \right) \quad = \quad -\frac{n}{2} \left(\ln \left(2\pi \right) + 1 \right) - \frac{n}{2} \ln \hat{\sigma}^2 \left(Q_n, \theta_c \right) + \left| S_n^e \left(Q_n, \theta_c \right) \right| + \left| R^e \left(Q_n, \theta_c \right) \right|$$

with respect to θ_g *and obtain the set of solution* $\hat{\theta} = (\hat{\theta}_c', \hat{\beta}(\hat{\theta}_c)', \hat{\sigma}^2(\hat{\theta}_c))'$ *such that* $\hat{\theta}_c \equiv arg\max_{\theta \in \Theta_c} \ln \mathcal{L}_n^c \left(\theta_c | \, y_n, x_n, Q_n \right)$*. Denote this set* $\hat{\Theta}$*. Full expressions for the concentrated parameters* $\hat{\beta}(\hat{\theta}_c)$ *and* $\hat{\sigma}^2(\hat{\theta}_c)$ *are given in Step 1 of Algorithm 3.*

Step 3. Check if probability of peers forming link is in the $[0,1]$ *range. Otherwise, go back to Step 1 and adjust* λ_0 *accordingly.*

Step 4. (C.I. of structural parameters). Obtain confidence regions for θ_g *following the bootstrap Algorithm 1.*

Step 5. (Network spillovers). Take any point $\hat{\theta}^*$ *in the identified* $\hat{\Theta}$*. Network spillovers are calculated as*

$$\varphi(x_n, \hat{\theta}) \quad = \quad (I - \lambda_0 \hat{W}_n^e)^{-1} (x_n \hat{\beta}_1 + \hat{W}_n x_j \hat{\beta}_2) - x_n \hat{\beta}_1 .$$

Confidence intervals are calculated following Algorithm 2.

Step 6. (Network data validity). When network data are available, Algorithm 2 is reemployed to provide confidence intervals for the null hypothesis $\mathcal{H}_0 : \delta_1 - \delta_0 = 0$*.*

Step 7. (Identifying λ*). Solve the GMM problem*

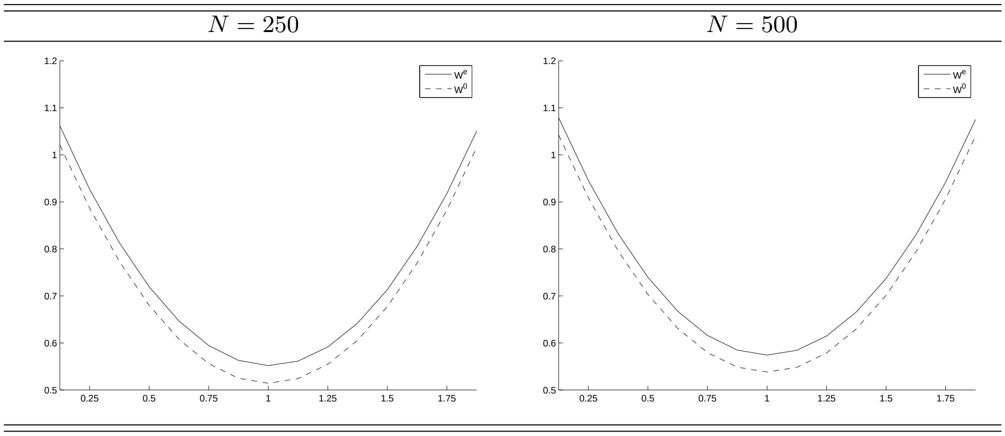
$$\hat{\theta} \quad = \quad \underset{\theta \in \hat{\Theta}}{arg\,min} \left(\sum_{j=1}^{v} S^{-1} \sum_{s=1}^{S} q_{s,j}(y_j, x_j, \theta) \right)' \Omega \left(\sum_{j=1}^{v} S^{-1} \sum_{s=1}^{S} q_{s,j}(y_j, x_j, \theta) \right)$$

where $q_{s,j}(y_j, x_j, \theta) = [V_{B,j}(y_j, x_j, \theta) - V_{B,j}(\hat{y}_j, x_j, \theta); V_{W,j}(y_j, x_j, \theta) - V_{W,j}(\hat{y}_j, x_j, \theta)]'$ *with* $\hat{y}_{j,s} = (S_j^s)^{-1}(x_j \beta_1 + W_j^s x_j \beta_2 + e_j^s)$ *and* $S^s = (I_{n_j} - \lambda W_j^s)^{-1}$*.* W_j^s *is sampled from the distribution of the network-generating model and* e_j^s *is sampled from a normal distribution with variance* σ^2*. Confidence intervals are given in Theorem 4.*

1.F Additional figures and tables.

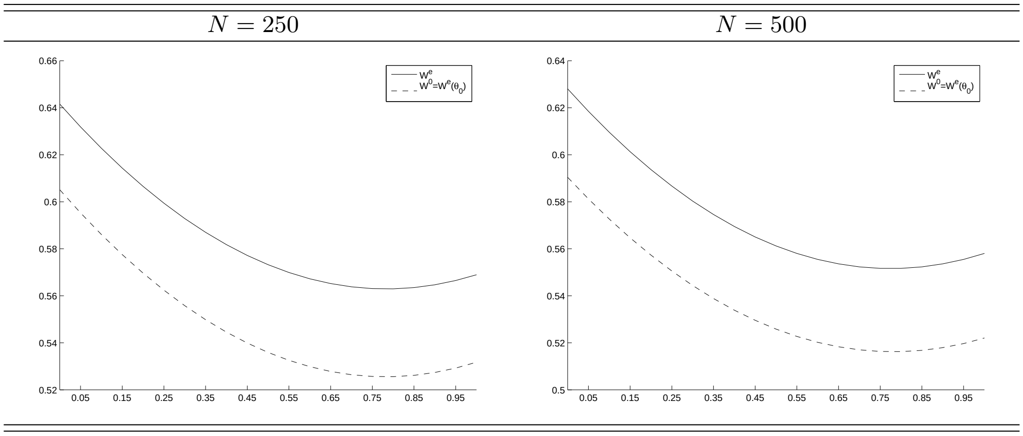
1.F.1 Estimator and simulations.

Table 1.F.1: Likelihood as a function of β_1.



Note: Rescaled additive inverse of likelihood as a function of β_1, with all other parameters at the true value. True $\beta_{10} = 1$. Solid line represents likelihood computed with expected network $W^e = W^e(Q, \theta_0)$, and dashed with real network W^0. True networks are realizations from the stochastic generating process.

Table 1.F.2: Likelihood as a function of δ_1.



Note: Rescaled additive inverse of likelihood as a function of δ_1, with all other parameters at the true value. True $\delta_{10} = 0.75$. Solid line represents likelihood computed with expected network $W^e = W^e(Q, \theta_0)$ and underlying networks are realization from the stochastic generating process. Dashed line $W^0 = W^e(\theta_0)$ is the likelihood where true network is equal to expected network.

Table 1.F.3: Simulations: bias in $\hat{\beta}_{OLS}$.

	$T = 1$.				$T = 5$, fixed effects.			
	(1)	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*
n	25	100	25	100	25	100	25	100
v	250	250	1000	1000	250	250	1000	1000
$\hat{\beta}_1$	1.0670	1.1127	1.0670	1.1120	1.0667	1.1116	1.0669	1.1118
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

	$T = 5$, time effects.				$T = 5$, time and fixed effects.			
	(1)	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*
n	25	100	25	100	25	100	25	100
v	250	250	1000	1000	250	250	1000	1000
$\hat{\beta}_1$	1.0665	1.1110	1.0670	1.1120	1.0660	1.1110	1.0671	1.1120
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

Note: True parameters is $\beta_1 = 1$.

Table 1.F.4: Simulations: baseline model with small n and v.

	$T = 1.$				$T = 5$, fixed effects.			
	(1)	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*
n	15	25	15	25	15	25	15	25
v	25	25	50	50	25	25	50	50
$\hat{\lambda}$	0.0270	0.0219	0.0213	0.0178	0.0199	0.0172	0.0168	0.0141
	[0.022]	[0.019]	[0.011]	[0.014]	[0.015]	[0.011]	[0.011]	[0.005]
$\hat{\beta}_1$	1.0023	0.9972	1.0035	1.0002	1.0023	1.0003	0.9997	1.0003
	(0.036)	(0.030)	(0.025)	(0.021)	(0.017)	(0.014)	(0.013)	(0.010)
	[0.036]	[0.029]	[0.025]	[0.021]	[0.016]	[0.013]	[0.011]	[0.009]
$\hat{\beta}_2$	0.0631	0.0175	0.0654	0.0190	0.1263	0.0501	0.0718	0.0469
	(0.904)	(0.651)	(0.405)	(0.527)	(0.648)	(0.071)	(0.120)	(0.020)
	[0.818]	[0.933]	[0.582]	[0.627]	[0.793]	[0.055]	[0.145]	[0.018]
$\hat{\delta}_1$	0.3637	0.5942	0.6637	0.6264	0.6971	0.7605	0.6823	0.7201
	(1.032)	(0.558)	(0.607)	(0.359)	(0.451)	(0.247)	(0.306)	(0.175)
	[0.804]	[0.498]	[0.556]	[0.374]	[0.360]	[0.228]	[0.258]	[0.163]
$\hat{\delta}_0$	0.0606	0.2101	0.2458	0.2235	0.2551	0.2956	0.2635	0.2901
	(0.539)	(0.234)	(0.276)	(0.139)	(0.161)	(0.103)	(0.114)	(0.070)
	[0.394]	[0.218]	[0.247]	[0.143]	[0.145]	[0.086]	[0.103]	[0.062]
$\hat{\sigma}^2$	1.0248	1.0448	1.0204	1.0475	0.8229	0.8436	0.8286	0.8455
	(1.424)	(0.598)	(0.461)	(0.449)	(0.641)	(0.068)	(0.126)	(0.021)
	[0.071]	[0.059]	[0.050]	[0.042]	[0.026]	[0.021]	[0.018]	[0.015]
$\varphi(x, \hat{\theta})$	0.0062	0.0013	-0.0022	0.0006	0.0008	0.0005	-0.0023	0.0004
	(0.051)	(0.067)	(0.036)	(0.046)	(0.020)	(0.030)	(0.015)	(0.018)
	[0.167]	[0.147]	[0.067]	[0.069]	[0.033]	[0.013]	[0.010]	[0.006]

Note: True parameters are $\beta_1 = 1$, $\beta_2 = 0.04$, $\delta_1 = 0.75$, $\delta_0 = 0.30$, $\sigma^2 = 1$ and $\varphi(x, \theta) = 0$.

Table 1.F.5: Simulations: baseline model with small n and v.

	$T=5$, time effects.				$T=5$, time and fixed effects.			
	(1)	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*
n	15	25	15	25	15	25	15	25
v	25	25	50	50	25	25	50	50
$\hat{\lambda}$	0.0178	0.0141	0.0156	0.0137	0.0197	0.0164	0.0166	0.0142
	[0.014]	[0.007]	[0.001]	[0.004]	[0.017]	[0.013]	[0.012]	[0.006]
$\hat{\beta}_1$	1.0025	1.0015	1.0003	1.0000	0.9997	0.9995	1.0004	1.0000
	(0.016)	(0.014)	(0.011)	(0.009)	(0.017)	(0.015)	(0.012)	(0.010)
	[0.016]	[0.013]	[0.011]	[0.009]	[0.016]	[0.013]	[0.011]	[0.009]
$\hat{\beta}_2$	0.0021	0.0664	0.0632	0.0444	0.0501	0.0808	0.0225	0.0508
	(0.396)	(0.241)	(0.054)	(0.013)	(0.318)	(0.178)	(0.525)	(0.025)
	[0.379]	[0.237]	[0.056]	[0.016]	[0.304]	[0.238]	[0.625]	[0.021]
$\hat{\delta}_1$	0.5643	0.6124	0.6611	0.7255	0.5651	0.6160	0.6345	0.6956
	(0.406)	(0.241)	(0.255)	(0.137)	(0.434)	(0.268)	(0.285)	(0.183)
	[0.360]	[0.237]	[0.259]	[0.161]	[0.353]	[0.224]	[0.253]	[0.162]
$\hat{\delta}_0$	0.2225	0.2479	0.2648	0.2883	0.2102	0.2480	0.2537	0.2732
	(0.164)	(0.091)	(0.106)	(0.052)	(0.178)	(0.108)	(0.110)	(0.070)
	[0.147]	[0.092]	[0.103]	[0.061]	[0.144]	[0.087]	[0.103]	[0.061]
$\hat{\sigma}^2$	0.0139	0.0082	0.0123	0.0041	0.0049	0.0110	0.0372	0.0138
	(0.380)	(0.032)	(0.055)	(0.013)	(0.052)	(0.171)	(0.054)	(0.025)
	[0.053]	[0.018]	[0.037]	[0.031]	[0.023]	[0.027]	[0.027]	[0.021]
$\varphi(x,\hat{\theta})$	-0.0023	-0.0082	-0.0000	0.0011	-0.0004	-0.0006	-0.0000	-0.0008
	(0.022)	(0.032)	(0.016)	(0.021)	(0.022)	(0.028)	(0.015)	(0.019)
	[0.034]	[0.018]	[0.010]	[0.006]	[0.046]	[0.016]	[0.015]	[0.006]

Note: True parameters are $\beta_1 = 1$, $\beta_2 = 0.04$, $\delta_1 = 0.75$, $\delta_0 = 0.30$, $\sigma^2 = 1$ and $\varphi(x, \theta) = 0$.

Table 1.F.6: Simulations: baseline model with across-group connections.

$T = 1$

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(6)*
n	100	100	100	100	100	100	100
v	250	250	250	250	250	250	250
δ_A	0.00	0.01	0.025	0.05	0.075	0.10	0.25
$\hat{\lambda}$	0.0120	0.0123	0.0132	0.0132	0.0140	0.0151	0.0278
	[0.001]	[0.002]	[0.003]	[0.002]	[0.002]	[0.006]	[0.001]
$\hat{\beta}_1$	0.9999	1.0000	1.0000	0.9999	0.9995	0.9992	0.9958
	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)
	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]	[0.005]
$\hat{\beta}_2$	0.0403	0.0396	0.0399	0.0379	0.0355	0.0313	0.0124
	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)	(0.005)	(0.002)
	[0.007]	[0.006]	[0.006]	[0.006]	[0.005]	[0.005]	[0.002]
$\hat{\delta}_1$	0.7569	0.7578	0.7605	0.7857	0.8243	0.8855	1.0000
	(0.081)	(0.074)	(0.081)	(0.079)	(0.086)	(0.076)	(0.071)
	[0.080]	[0.080]	[0.080]	[0.081]	[0.079]	[0.079]	[0.069]
$\hat{\delta}_0$	0.3010	0.3036	0.3035	0.3131	0.3264	0.3484	0.5397
	(0.028)	(0.029)	(0.032)	(0.030)	(0.032)	(0.031)	(0.036)
	[0.031]	[0.030]	[0.030]	[0.030]	[0.030]	[0.030]	[0.035]
$\hat{\sigma}^2$	1.0558	1.0582	1.0629	1.0696	1.0778	1.0849	1.1255
	(0.007)	(0.006)	(0.007)	(0.006)	(0.006)	(0.005)	(0.002)
	[0.009]	[0.009]	[0.010]	[0.010]	[0.010]	[0.010]	[0.010]
$\varphi(x, \hat{\theta})$	0.0008	-0.0010	-0.0015	0.0014	-0.0000	-0.0011	-0.0001
	(0.010)	(0.010)	(0.011)	(0.011)	(0.009)	(0.010)	(0.010)
	[0.002]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]

Note: True parameters are $\beta_1 = 1$, $\beta_2 = 0.04$, $\delta_1 = 0.75$, $\delta_0 = 0.30$, $\sigma^2 = 1$ and $\varphi(x, \theta) = 0$.

Table 1.F.7: Simulations: Bernoulli model under misspecified λ_{ref}.

	$T = 1.$		$T = 5$, FE.		$T = 5$, TE.		$T = 5$, FE and TE.	
	(1)	*(2)*	*(1)*	*(2)*	*(1)*	*(2)*	*(1)*	*(2)*
n	25	100	25	100	25	100	25	100
v	250	250	250	250	250	250	250	250
$\hat{\lambda}$	0.1002	0.0094	0.0092	0.0094	0.0099	0.0105	0.0977	0.0113
	[0.002]	[0.000]	[0.006]	[0.009]	[0.001]	[0.000]	[0.006]	[100.00%]
$\hat{\beta}_1$	1.0016	1.0009	0.9996	1.0008	1.0000	1.0002	1.0005	0.9999
	(0.008)	(0.005)	(0.004)	(0.002)	(0.004)	(0.002)	(0.004)	(0.002)
	[0.009]	[0.005]	[0.004]	[0.002]	[0.004]	[0.002]	[0.004]	[0.002]
$\hat{\beta}_2$	0.0892	0.0817	0.0878	0.0793	0.0826	0.0801	0.0825	0.0813
	(0.037)	(0.007)	(0.019)	(0.003)	(0.010)	(0.002)	(0.011)	(0.003)
	[0.033]	[0.007]	[0.014]	[0.003]	[0.012]	[0.003]	[0.012]	[0.003]
$\hat{\delta}_1$	0.3766	0.3722	0.3606	0.3764	0.3702	0.3745	0.3700	0.3729
	(0.095)	(0.014)	(0.050)	(0.005)	(0.033)	(0.005)	(0.032)	(0.006)
	[0.082]	[0.013]	[0.037]	[0.006]	[0.036]	[0.006]	[0.036]	[0.006]
$\hat{\delta}_0$	0.1427	0.1496	0.1449	0.1499	0.1475	0.1501	0.1488	0.1498
	(0.038)	(0.005)	(0.018)	(0.002)	(0.012)	(0.001)	(0.013)	(0.002)
	[0.029]	[0.004]	[0.014]	[0.002]	[0.013]	[0.002]	[0.013]	[0.002]
$\hat{\sigma}^2$	1.0550	1.2193	0.8458	0.9767	0.0678	0.0495	0.0376	0.0312
	(0.037)	(0.007)	(0.019)	(0.003)	(0.010)	(0.002)	(0.011)	(0.003)
	[0.019]	[0.011]	[0.007]	[0.004]	[0.014]	[0.009]	[0.012]	[0.003]
$\varphi(x, \hat{\theta})$	-0.0032	-0.0095	0.0029	-0.0037	-0.0026	-0.0061	-0.0013	-0.0007
	(0.018)	(0.085)	(0.019)	(0.030)	(0.009)	(0.053)	(0.011)	(0.027)
	[0.006]	[0.006]	[0.007]	[0.001]	[0.001]	[0.009]	[0.001]	[0.001]

Note: True parameters are $\beta_1 = 1$, $\beta_2 = 0.04$, $\delta_1 = 0.75$, $\delta_0 = 0.30$, $\sigma^2 = 1$ and $\varphi(x, \theta) = 0$.

Table 1.F.8: Simulations: multivariate network model.

	$T = 1.$		$T = 5$, FE.		$T = 5$, TE.		$T = 5$, FE and TE.	
	(1)	*(2)*	*(1)*	*(2)*	*(1)*	*(2)*	*(1)*	*(2)*
n	25	100	25	100	25	100	25	100
v	250	250	250	250	250	250	250	250
$\hat{\lambda}$	0.0132	0.0124	0.0122	0.0125	0.0124	0.0125	0.0128	0.0125
	[0.002]	[0.000]	[0.001]	[0.000]	[0.001]	[0.000]	[0.001]	[0.000]
$\hat{\beta}_1$	0.9979	0.9997	0.9999	1.0001	0.9999	1.0000	1.0000	0.9999
	(0.009)	(0.005)	(0.004)	(0.002)	(0.004)	(0.002)	(0.004)	(0.002)
	[0.009]	[0.005]	[0.004]	[0.002]	[0.004]	[0.002]	[0.004]	[0.002]
$\hat{\beta}_2$	0.0432	0.0400	0.0398	0.0398	0.0394	0.0399	0.0403	0.0401
	(0.018)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
	[0.014]	[0.001]	[0.002]	[0.001]	[0.002]	[0.001]	[0.002]	[0.001]
$\hat{\delta}_1$	0.2395	0.2500	0.2509	0.2515	0.2574	0.2509	0.2452	0.2503
	(0.048)	(0.009)	(0.022)	(0.005)	(0.017)	(0.004)	(0.020)	(0.005)
	[0.016]	[0.007]	[0.011]	[0.005]	[0.015]	[0.071]	[0.010]	[0.004]
$\hat{\delta}_0$	0.4949	0.5013	0.5029	0.5021	0.5077	0.5010	0.4969	0.4990
	(0.077)	(0.011)	(0.025)	(0.006)	(0.024)	(0.005)	(0.025)	(0.008)
	[0.018]	[0.010]	[0.011]	[0.007]	[0.015]	[0.068]	[0.012]	[0.005]
$\hat{\sigma}^2$	1.0193	1.0697	0.8159	0.8546	0.1854	0.1956	0.1062	0.2845
	(0.016)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)
	[0.018]	[0.010]	[0.007]	[0.003]	[0.013]	[0.007]	[0.005]	[0.002]
$\varphi(x, \hat{\theta})$	0.0088	0.0027	0.0045	0.0009	0.0053	0.0009	0.0037	0.0021
	(0.007)	(0.002)	(0.003)	(0.008)	(0.004)	(0.009)	(0.003)	(0.010)
	[0.001]	[0.001]	[0.000]	[0.000]	[0.000]	[0.001]	[0.000]	[0.000]

Note: True parameters are $\beta_1 = 1$, $\beta_2 = 0.04$, $\delta_1 = 0.25$, $\delta_0 = 0.50$, $\sigma^2 = 1$ and $\varphi\,(x, \theta) = 0$.

1.F.2 Application.

Table 1.F.9: Occupational Choice with Network Data.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Self hours.		Wage hours.		Self emp. only.	
	Method	Network.	Network.	Network.	Network.	Network.	Network.
		Family.	Economic.	Family.	Economic.	Family.	Economic.
Not function of $\hat{\lambda}$.	Program effect	473.219***	473.581***	−113.002***	−113.146***	0.113***	0.114***
	after 2 years ($\hat{\beta}_{11}$).	(12.99)	(13.89)	(8.33)	(8.33)	(0.01)	(0.01)
	Program effect	464.069***	463.441***	−142.755***	−143.009***	0.120***	0.121***
	after 4 years ($\hat{\beta}_{12}$).	(13.07)	(5.10)	(8.53)	(8.25)	(0.01)	(0.01)
	Spillover on T	−20.438***	−23.097***	24.394***	26.933***	−0.029***	−0.034***
	after 2 years ($\hat{\varphi}_{T,2}$).	(7.01)	(6.95)	(8.50)	(9.21)	(0.01)	(0.01)
	Spillover on T	17.396***	14.734**	19.805**	22.105**	−0.023***	−0.027**
	after 4 years ($\hat{\varphi}_{T,4}$).	(6.41)	(7.04)	(8.37)	(10.30)	(0.00)	(0.01)
	Spillover on NT	−9.771***	−11.346***	12.692***	14.259***	−0.015***	−0.018***
	after 2 years ($\hat{\varphi}_{NT,2}$).	(3.35)	(3.42)	(4.41)	(4.87)	(0.00)	(0.00)
	Spillover on NT	8.317**	7.237***	10.304**	11.703	−0.012***	−0.014***
	after 4 years ($\hat{\varphi}_{NT,4}$).	(3.28)	(1.88)	(5.21)	(13.28)	(0.01)	(0.00)
Function of $\hat{\lambda}$.	Link to T	−40.247***	−27.635***	12.794***	13.663***	−0.045***	−0.051***
	after 2 years ($\hat{\beta}_{21}$).	(1.99)	(1.42)	(2.48)	(2.72)	(0.01)	(0.01)
	Link to T	−30.758***	−20.648***	12.938***	13.721***	−0.040***	−0.045***
	after 4 years ($\hat{\beta}_{22}$).	(1.53)	(1.77)	(1.57)	(2.73)	(0.01)	(0.01)
	Link probability	0.776***	0.759***	0.985***	0.726***	0.336***	0.196***
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).	(0.05)	(0.05)	(0.08)	(0.05)	(0.03)	(0.02)
	Link probability	0.317***	0.464***	0.364***	0.362***	0.115***	0.116***
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
	$\hat{\lambda}$	0.075	0.05	0.05	0.05	0.15	0.15
	p-value $\mathcal{H}_{NV}$.	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	Avg treated outcome.	421.8	421.8	646.7	646.7	0.303	646.7
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes as in Table 1.3.

Table 1.F.10: Earnings and Seasonality with Network Data.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Earnings.		Share Seas.		Share Reg.	
	Method	Network.	Network.	Network.	Network.	Network.	Network.
		Family.	Economic.	Family.	Economic.	Family.	Economic.
Not function of $\hat{\lambda}$.	Program effect	0.562***	0.556***	−0.029***	−0.029***	0.181***	0.182***
	after 2 years ($\hat{\beta}_{11}$).	(0.207)	(0.148)	(0.01)	(0.01)	(0.01)	(0.01)
	Program effect	2.726***	2.806***	−0.075***	−0.075***	0.166***	0.166***
	after 4 years ($\hat{\beta}_{12}$).	(0.196)	(0.108)	(0.01)	(0.01)	(0.01)	(0.01)
	Spillover on T	−0.258**	−0.187*	−0.063***	−0.062***	0.037***	0.030***
	after 2 years ($\hat{\varphi}_{T,2}$).	(0.116)	(0.113)	(0.02)	(0.02)	(0.01)	(0.01)
	Spillover on T	−0.098	−0.188*	−0.016*	−0.016	0.044***	0.035***
	after 4 years ($\hat{\varphi}_{T,4}$).	(0.117)	(0.112)	(0.01)	(0.02)	(0.02)	(0.01)
	Spillover on NT	−0.133**	−0.102*	−0.026***	−0.026***	0.017***	0.014***
	after 2 years ($\hat{\varphi}_{NT,2}$).	(0.060)	(0.062)	(0.01)	(0.01)	(0.01)	(0.01)
	Spillover on NT	−0.051	−0.103	−0.002	−0.007	0.020**	0.017**
	after 4 years ($\hat{\varphi}_{NT,4}$).	(0.057)	(0.78)	(0.01)	(0.01)	(0.01)	(0.00)
Function of $\hat{\lambda}$.	Link to T	−0.236	−0.245	−0.023***	−0.017***	−0.051***	−0.053***
	after 2 years ($\hat{\beta}_{21}$).	(0.456)	(0.478)	(0.02)	(0.00)	(0.01)	(0.01)
	Link to T	−0.740	−0.375	−0.019***	−0.014***	−0.037***	−0.039***
	after 4 years ($\hat{\beta}_{22}$).	(0.541)	(0.596)	(0.01)	(0.00)	(0.01)	(0.01)
	Link probability	0.155***	0.064***	0.234***	0.236***	0.100***	0.077***
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).	(0.01)	(0.00)	(0.02)	(0.01)	(0.01)	(0.00)
	Link probability	0.030***	0.030***	0.152***	0.203***	0.054***	0.051***
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
	$\hat{\lambda}$	0.50	0.50	0.20	0.15	0.50	0.50
	p-value $\mathcal{H}_{NV}$.	< 0.001	< 0.001	< 0.001	0.022	< 0.001	< 0.001
	Avg treated outcome.	4.607	4.607	0.674	0.674	0.478	0.478
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes as in Table 1.3.

Table 1.F.11: Livestock with Network Data.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Cows.		Poultry.		Livestock Value.	
	Method	Network. Family.	Network. Economic.	Network. Family.	Network. Economic.	Network. Family.	Network. Economic.
Not function of $\hat{\lambda}$.	Program effect	1.132^{***}	1.132^{***}	2.116^{***}	2.117^{***}	10.412^{***}	10.420^{***}
	after 2 years ($\hat{\beta}_{11}$).	(0.03)	(0.03)	(0.50)	(0.50)	(365.41)	(0.45)
	Program effect	1.103^{***}	1.101^{***}	1.296^{***}	1.330^{***}	11.175^{***}	11.173^{***}
	after 4 years ($\hat{\beta}_{12}$).	(0.03)	(0.03)	(0.50)	(0.50)	(459.21)	(0.44)
	Spillover on T	-0.032^{***}	-0.033^{***}	0.039	0.107	-0.184^{***}	-0.230^{***}
	after 2 years ($\hat{\varphi}_{T,2}$).	(0.01)	(0.01)	(0.11)	(0.18)	(0.07)	(0.06)
	Spillover on T	-0.055^{***}	-0.055^{***}	0.029	-0.095	-0.407^{***}	-0.456^{***}
	after 4 years ($\hat{\varphi}_{T,4}$).	(0.02)	(0.02)	(0.12)	(0.21)	(0.11)	(0.06)
	Spillover on NT	-0.018^{***}	-0.020^{***}	0.014	0.064	-0.106^{***}	-0.137^{***}
	after 2 years ($\hat{\varphi}_{NT,2}$).	(0.01)	(0.01)	(0.06)	(0.11)	(0.04)	(0.04)
	Spillover on NT	-0.031^{***}	-0.032^{***}	0.011	-0.056	-0.234^{***}	-0.272^{***}
	after 4 years ($\hat{\varphi}_{NT,4}$).	(0.01)	(0.01)	(0.10)	(0.08)	(0.08)	(0.03)
Function of $\hat{\lambda}$.	Link to T	-0.965^{***}	-0.996^{***}	9.169	1.495	-9.251^{***}	-10.634^{***}
	after 2 years ($\hat{\beta}_{21}$).	(0.15)	(0.15)	(19.65)	(4.22)	(2.64)	(1.22)
	Link to T	-1.227^{***}	-1.256^{***}	6.975	-2.914	-14.504^{***}	-16.332^{***}
	after 4 years ($\hat{\beta}_{22}$).	(0.16)	(0.16)	(21.05)	(4.21)	(2.30)	(2.07)
	Link probability	0.039^{***}	0.019^{***}	0.020^{**}	0.008	0.029^{***}	0.010^{**}
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Link probability	0.014^{***}	0.014^{***}	0.011	0.008^{***}	0.008^{***}	0.008^{***}
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
	$\widehat{\lambda}$	0.50	0.50	0.50	0.50	0.50	0.50
	p-value $\mathcal{H}_{NV}$.	0.003	0.300	0.045	1.000	0.024	0.764
	Avg treated outcome.	0.083	0.083	1.79	1.79	0.940	0.940
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes as in Table 1.3.

Table 1.F.12: Expenditures with Network Data.

		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Outcome	Nonfood PCE.		Food PCE.		Food Security.	
	Method	Network. Family.	Network. Economic.	Network. Family.	Network. Economic.	Network. Family.	Network. Economic.
Not function of $\hat{\lambda}$.	Program effect	-208.803	-208.049	421.741***	424.602***	0.169***	0.169***
	after 2 years ($\hat{\beta}_{11}$).	(160.98)	(160.05)	(133.67)	(133.61)	(0.01)	(0.01)
	Program effect	280.309*	279.158	444.980***	447.736***	0.075***	0.076***
	after 4 years ($\hat{\beta}_{12}$).	(145.11)	(178.65)	(133.66)	(133.61)	(0.01)	(0.01)
	Spillover on T	-29.966	-32.452	401.713***	387.106***	0.028***	0.083***
	after 2 years ($\hat{\varphi}_{T,2}$).	(70.13)	(69.80)	(56.88)	(56.47)	(0.01)	(0.03)
	Spillover on T	-161.955**	-161.161**	253.726***	242.561***	0.080**	0.163***
	after 4 years ($\hat{\varphi}_{T,4}$).	(71.28)	(69.72)	(59.58)	(55.82)	(0.03)	(0.05)
	Spillover on NT	-17.507	-19.103	215.298***	208.075***	0.012***	0.033***
	after 2 years ($\hat{\varphi}_{NT,2}$).	(40.98)	(41.09)	(30.18)	(30.97)	(0.00)	(0.01)
	Spillover on NT	-94.620***	-94.869**	135.984***	130.380***	0.033***	0.065***
	after 4 years ($\hat{\varphi}_{NT,4}$).	(26.64)	(39.08)	(51.07)	(29.85)	(0.00)	(0.02)
Function of $\hat{\lambda}$.	Link to T	-311.329	-349.080	343.343***	438.309***	0.102***	0.123***
	after 2 years ($\hat{\beta}_{21}$).	(966.77)	(968.78)	(62.93)	(83.73)	(0.01)	(0.01)
	Link to T	-2386.991**	-2389.737**	190.068***	238.308***	0.088***	0.113***
	after 4 years ($\hat{\beta}_{22}$).	(959.21)	(962.22)	(62.48)	(83.19)	(0.01)	(0.01)
	Link probability	0.020**	0.014**	0.158***	0.132***	0.184***	0.092***
	if $Q_{ij} = 1$ ($\hat{\delta}_1$).	(0.01)	(0.01)	(0.03)	(0.01)	(0.00)	(0.00)
	Link probability	0.013***	0.013***	0.118***	0.087***	0.059***	0.065***
	if $Q_{ij} = 0$ ($\hat{\delta}_0$).	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	$\widehat{\lambda}$	0.50	0.50	0.15	0.20	0.50	0.50
	p-value $\mathcal{H}_{NV}$.	0.389	0.835	0.002	0.159	< 0.001	< 0.001
	Avg treated outcome.	1054.5	1054.5	2953.7	2953.7	0.457	0.457
	Individuals (n).	23029	23029	23029	23029	23029	23029
	Villages (v).	1409	1409	1409	1409	1409	1409
	Survey waves (T).	3	3	3	3	3	3

Notes as in Table 1.3.

Table 1.F.13: Occupational Choice, Bernoulli model.

		(1)	*(2)*	*(3)*
	Outcome	Self hours.	Wage hours.	Self emp. only.
	Method	Network.	Network.	Network.
Not function of $\hat{\lambda}$.	Program effect	474.153***	−112.859***	0.114***
	after 2 years ($\hat{\beta}_{11}$).	(14.55)	(8.34)	(0.01)
	Program effect	464.304***	−143.481***	0.121***
	after 4 years ($\hat{\beta}_{12}$).	(9.50)	(8.47)	(0.01)
	Spillover on T	−26.577***	25.865***	−0.033***
	after 2 years ($\hat{\varphi}_{T,2}$).	(7.92)	(6.55)	(0.01)
	Spillover on T	13.148	22.082***	−0.027***
	after 4 years ($\hat{\varphi}_{T,4}$).	(9.59)	(7.06)	(0.01)
	Spillover on NT	−12.862**	13.714***	−0.018***
	after 2 years ($\hat{\varphi}_{NT,2}$).	(6.56)	(3.77)	(0.00)
	Spillover on NT	6.363	11.708***	−0.015***
	after 4 years ($\hat{\varphi}_{NT,4}$).	(4.58)	(1.97)	(0.00)
Function of $\hat{\lambda}$.	Link to T	−27.891***	13.355***	−0.050***
	after 2 years ($\hat{\beta}_{21}$).	(1.38)	(2.50)	(0.01)
	Link to T	−12.862***	13.758***	−0.045***
	after 4 years ($\hat{\beta}_{22}$).	(1.63)	(1.59)	(0.01)
	Link probability	0.492***	0.380***	0.120***
	($\hat{\delta}_1$).	(0.03)	(0.03)	(0.00)
	$\hat{\lambda}$	0.05	0.05	0.15
	Avg treated outcome.	421.8	646.7	646.7
	Individuals (n).	23029	23029	23029
	Villages (v).	1409	1409	1409
	Survey waves (T).	3	3	3

Notes as in Table 1.3.

Chapter 2

Regularization for Spatial Panel Time Series using the Adaptive Lasso

Abstract. This paper proposes a model for estimating the underlying cross-sectional dependence structure of a large panel of time series. We propose to estimate this by penalizing the elements in the spatial weight matrices using the adaptive LASSO proposed by Zou (2006). Non-asymptotic oracle inequalities and the asymptotic sign consistency of the estimators are proved when the cross-sectional dimension (N) can be larger than the time dimension (T). A block coordinate descent algorithm is introduced, with simulations and a real data analysis carried out.[1]

Keywords: Spatial econometrics; adaptive Lasso; sign consistency; non-asymptotic oracle inequalities; spatial weight matrices.

JEL classification: C31, C33.

[1] Paper coauthored with Clifford Lam, London School of Economics, Department of Statistics.

2.1 Introduction

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

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

In this paper, we study a more general form of spatial autoregressive model as detailed in section 2.2. In the terminology of Anselin (2002), we include both global and local spillover effects, through the terms $\mathbf{W}_1^*\mathbf{y}_t$ and $\mathbf{W}_2^*\mathbf{X}_t\boldsymbol{\beta}^*$ respectively in model (2.2). Few researchers attempted to estimate the spatial weight matrices, including a well known paper by Pinkse et al. (2002). They estimate a nonparametric smooth function $\widehat{g}(\cdot)$ assuming normality of data, and the (i, j)-th element of the matrix $\mathbf{W}_1^*$ is estimated as $\widehat{g}(d_{ij})$, where d_{ij} is a distance measure specified by the user. Beenstock and Felsenstein (2012) suggested using a moment estimator for the spatial weight matrix. Bhattacharjee and Jensen-Butler (2013) proposes to estimate the spatial weight matrix by first estimating the error covariance matrix. However, estimating a large error covariance matrix can be inaccurate as the dimension N of the panel is large and can be close to the sample size T - one of the major characteristics of a large time series panel. Recently, Ahrens and Bhattacharjee (2014) proposes to estimate the spatial weight matrix in a spatial autoregressive model with exogenous instruments by using a two-step LASSO estimation but deal with a restricted version of our model.

In our paper, we focus on estimating the spatial weight matrices themselves, which are assumed to be sparse: having a lot of zero entries. There is no need to specify a distance measure for our method as long as the true spatial weight matrices are sparse. We provided non-asymptotic bounds on various estimated quantities on a set with probability approaching 1 asymptotically (see Lemma

2 for example). We demonstrate that sparsity is a common endeavor with a structural equation model in Example 1 in section 2.3.1.

The aims in estimating the spatial weight matrices are twofold. First, it is not always clear what exactly the spatial dependence structure is for the panel data. Even with expert knowledge of what the spatial matrices should look like, estimating them from data may reveal dependence structures that our assumptions can miss out. Presenting the estimated spatial weight matrix as a network connecting the components of the panel time series provides a visual tool for deeper understanding of cross-sectional dependence structure. Second, as presented previously, there are no universal rules in specifying a spatial weight matrix. We quote a part of the criticism summarized in Arbia and Fingleton (2008), "... arbitrary nature of weight matrix... are not the results obtained conditional on somewhat arbitrary decisions taken about its structure?" Although debate is still on about the sensitivity of results towards the specification of spatial weight matrices, this paper provides a partial solution to the criticism and potential sensitivity towards "arbitrary" specification of these matrices if they themselves can be estimated from data as well. In fact in Lemma 2, we have specified how the error upper bound for the estimation of $\boldsymbol{\beta}^*$ in model (2.2) is related to the error of the estimated/assumed spatial weight matrices. This result sheds some lights on the potential seriousness of wrongly specifying the spatial weight matrices.

The rest of the paper is organized as follows. In section 2.2, we introduce the spatial autoregressive model considered, with examples. Section 2.3 presents the model in a compact form and introduces the minimization problems for obtaining the estimators of the sparse spatial weight matrices. These estimators are analyzed in section 2.4 using a relatively new concept of time dependence in time series data, with non-asymptotic oracle inequalities and rates of convergence spelt out, as well as asymptotic sign consistency presented. Section 2.5 discusses the computational issue of our estimators, and present a block coordinate descent algorithm as a solution. Section 2.6 presents our extensive simulation results and real data analysis. The paper concludes with section 2.7, outlining our main contributions and some future research directions. Finally all technical proofs of the theorems in section 2.4 are presented in section 2.A.

2.2 The Model

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

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

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

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

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

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

where $\mathbf{y}_t$ is an $\mathbf{N} \times 1$ vector of dependent time series variables, $\mathbf{W}_j^*$ for $j = 1, 2$ are the $N \times N$ spatial weight matrices to be estimated, $\mathbf{X}_t$ is an $N \times K$ matrix of centered exogenous variables at time t, $\boldsymbol{\beta}^*$ is a vector of K regression parameters for the exogenous variables, and finally $\{\boldsymbol{\epsilon}_t\}$ is an innovation process with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}_\epsilon$, and is independent of $\{\mathbf{X}_t\}$. Both $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are assumed second order stationary. The matrix $\boldsymbol{\Sigma}_\epsilon$ is assumed to have uniformly bounded entries as $N, T \to \infty$. Detailed assumptions A1- A8 can be found in section 2.4.

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

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

with innovations in $\Pi_1^* \boldsymbol{\epsilon}_t$ having finite variances, where $\mathbf{I}_N$ is the identity matrix of size N. See also Corrado and Fingleton (2011) or Kapoor et al. (2007) for a similar row sum regularity condition for the spatial weight matrices in a slightly different spatial model specification. Hence each component y_{tj} is a weighted linear combination of the other components in $\mathbf{y}_t$. If $w_{1,ij}^* \neq 0$, it means that y_{ti} depends on y_{tj} explicitly. An analysis of the links among financial markets is given in section 2.6 to illustrate the use of such a model.

The spatial weight matrix $\mathbf{W}_2^*$ has 1 on the main diagonal, with the same row sum condition as $\mathbf{W}_1^*$ excluding the diagonal entries. Hence while each component y_{tj} has the same regression coefficients $\boldsymbol{\beta}^*$ for their respective exogenous variables $\mathbf{x}_{t,j}^{\mathrm{T}}$ (the j-th row of $\mathbf{X}_t$), model (2.2) gives flexibility through $\mathbf{W}_2^*$ by allowing each y_{tj} to depend on a linear combination of exogenous variables for other components as well. This is also related to the local spatial spillover effects.

For more details please refer to Anselin (2002). See section 2.3.1 for an illustrative example with covariates.

Remark 1. The spatial error model with spatial autoregressive-moving average (ARMA) error can be defined by (see also Yao and Brockwell, 2006)

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

where $\boldsymbol{\epsilon}_t = (\mathbf{I}_N + \lambda\mathbf{W}')\mathbf{v}_t$. Model (2.2) entails this spatial ARMA error model, by setting $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, $\mathbf{W}_1^* = \rho\mathbf{W}$, $\mathbf{W}_2^* = \mathbf{I}_N - \rho\mathbf{W}$, and $\boldsymbol{\Sigma}_\epsilon = (\mathbf{I}_N + \lambda\mathbf{W}')\mathrm{var}(\mathbf{v}_t)(\mathbf{I}_N + \lambda(\mathbf{W}')^{\mathrm{T}})$. From assumption A4 in section 2.4.1, as long as the spatial autocovariance between $x_{t,jk}$ and $x_{t,j'k}$ for $j \neq j'$ decays fast enough as $|j - j'|$ gets larger, the correlation matrix for $\boldsymbol{\epsilon}_t$ can have a general structure, including that of a spatial moving-average structure as above.

2.3 Sparse Estimation of the Spatial Weight Matrices

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

2.3.1 Example 1

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

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

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

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

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

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

For a low dimensional model like this example, a reduced form model can be calculated like that in (2.3) and we can consistently estimate the parameters from the reduced form model. We can then try to recover the parameters $w_1, w_2, \beta_1, \beta_2, \gamma_1$ and γ_2 from the reduced form model parameters. This is also done in Irwin and Geoghegan (2001) for this particular example. However, for higher dimensional model where the spatial weight matrices are our target, the problem can become intractable, and we in general need the decay assumption A2 in section 2.4.1 for asymptotic sign consistency for all the estimated entries in the spatial weight matrix. See example 2 in section 2.4.2 as well.

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

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

where γ_T is a tuning parameter with rate given in Theorem 2 in section 2.4.3, and $v_{r,ij} = 1/|\tilde{w}_{r,ij}|^k$ for $r = 1, 2$ and some integer $k \geq 1$, with $\tilde{w}_{r,ij}$ being the solutions of the above minimization problem with all $v_{r,ij}$ set to 1. The $\tilde{w}_{r,ij}$'s thus represent the LASSO solutions (e.g. Zhao and Yu, 2006) with constraints. The $v_{r,ij}$ becomes the weight of penalization. The larger the magnitude of $\tilde{w}_{r,ij}$, the smaller $v_{r,ij}$ becomes, and vice versa. This is a sensible weighting scheme since a larger

$\tilde{w}_{r,ij}$ means $w^*_{r,ij}$ is less likely to be zero, and hence should be penalized less to reduce estimation bias, and vice versa.

The above penalization problem is cumbersome to write and makes presentation and proofs of theorems difficult. Hence we rewrite model (2.2) as a more familiar regression type model:

$$\begin{aligned}\mathbf{y} &= \mathbf{Z}\boldsymbol{\xi}^*_1 + \mathbf{X}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^*_2 + \boldsymbol{\epsilon}\\ &= \mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^* + \boldsymbol{\epsilon},\end{aligned} \tag{2.4}$$

where $\mathbf{y} = \mathrm{vec}\{(\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}\}$, $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}$, $\mathbf{X}_{\boldsymbol{\beta}^*} = \mathbf{I}_N \otimes \{(\mathbf{I}_T \otimes \boldsymbol{\beta}^{*\mathrm{T}})(\mathbf{X}_1, \ldots, \mathbf{X}_T)^{\mathrm{T}}\}$, $\boldsymbol{\xi}^*_j = \mathrm{vec}(\mathbf{W}^{*\mathrm{T}}_j)$ for $j = 1, 2$, and $\boldsymbol{\epsilon} = \mathrm{vec}\{(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^{\mathrm{T}}\}$. Here $\otimes$ represents the Kronecker product, and the vec operator stacks the columns of a matrix into a single vector, starting from the first column. Defining $\mathbf{M}_{\boldsymbol{\beta}^*} = (\mathbf{Z}, \mathbf{X}_{\boldsymbol{\beta}^*})$ as the "design matrix" and $\boldsymbol{\xi}^* = (\boldsymbol{\xi}^{*\mathrm{T}}_1, \boldsymbol{\xi}^{*\mathrm{T}}_2)^{\mathrm{T}}$ as the true "regression parameter", model (2.4) looks like a typical linear model, except that the design matrix $\mathbf{M}_{\boldsymbol{\beta}^*}$ is dependent on $\mathbf{y}$ as well.

With model (2.4), we can find the LASSO solutions by solving

$$\begin{aligned}(\widetilde{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\beta}}) &= \arg\min_{\boldsymbol{\xi},\boldsymbol{\beta}} \frac{1}{2T}\|\mathbf{y} - \mathbf{M}_{\boldsymbol{\beta}}\boldsymbol{\xi}\|^2 + \gamma_T\|\boldsymbol{\xi}\|_1,\\ \text{subj. to } &\sum_{j\neq i}|w_{1,ij}|, \sum_{j\neq i}|w_{2,ij}| < 1,\end{aligned} \tag{2.5}$$

where $\|\cdot\|_1$ represents the L_1-norm, and the definitions of $\mathbf{M}_{\boldsymbol{\beta}}$ and $\boldsymbol{\xi}$ are parallel to those in model (2.4). The adaptive LASSO solutions are then

$$\begin{aligned}(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\beta}}) &= \arg\min_{\boldsymbol{\xi},\boldsymbol{\beta}} \frac{1}{2T}\|\mathbf{y} - \mathbf{M}_{\boldsymbol{\beta}}\boldsymbol{\xi}\|^2 + \gamma_T\mathbf{v}^{\mathrm{T}}|\boldsymbol{\xi}|,\\ \text{subj. to } &\sum_{j\neq i}|w_{1,ij}|, \sum_{j\neq i}|w_{2,ij}| < 1,\end{aligned} \tag{2.6}$$

where $|\boldsymbol{\xi}| = (|\xi_1|, \ldots, |\xi_{2N^2}|)^{\mathrm{T}}$ and $\mathbf{v} = (|\widetilde{\xi}_1|^{-k}, \ldots, |\widetilde{\xi}_{2N^2}|^{-k})^{\mathrm{T}}$. A general block coordinate descent algorithm is introduced in section 2.5 to carry out the minimization.

2.4 Properties of LASSO and adaptive LASSO Estimators

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

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

$$\mathbf{x}_t = \mathbf{f}(\mathcal{F}_t), \quad \boldsymbol{\epsilon}_t = \mathbf{g}(\mathcal{G}_t), \tag{2.7}$$

where $\mathbf{f}(\mathcal{F}_t) = (f_1(\mathcal{F}_t), \ldots, f_{NK}(\mathcal{F}_t))^{\mathrm{T}}$ and $\mathbf{g}(\mathcal{G}_t) = (g_1(\mathcal{G}_t), \ldots, g_N(\mathcal{G}_t))^{\mathrm{T}}$ are both vectors of measurable functions defined on the real line. The shift processes $\mathcal{F}_t = (\ldots, \mathbf{e}_{x,t-1}, \mathbf{e}_{x,t})$ and $\mathcal{G}_t = (\ldots, \mathbf{e}_{\epsilon,t-1}, \mathbf{e}_{\epsilon,t})$ are defined by independent and identically distributed (i.i.d.) processes $\{\mathbf{e}_{x,t}\}$ and $\{\mathbf{e}_{\epsilon,t}\}$, and they are independent of each other. Hence $\{\mathbf{x}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are assumed independent. The representation (2.7) is used in Wu (2011) and provides a very general framework for stationary ergodic processes. See Wu (2011) for some examples as well.

For measuring dependence, instead of using traditional measures, like mixing conditions for time series, we use the functional dependence measure introduced in Wu (2005). This measure lays the framework for applying a Nagaev-type inequality for obtaining the results of our theorems to be presented later. For the time series $\{\mathbf{x}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ in (2.7), define for $a > 0$,

$$\begin{aligned} \theta^x_{t,a,j} &= \|x_{tj} - x'_{tj}\|_a = (E|x_{tj} - x'_{tj}|^a)^{1/a}, \\ \theta^\epsilon_{t,a,\ell} &= \|\epsilon_{t\ell} - \epsilon'_{t\ell}\|_a = (E|\epsilon_{t\ell} - \epsilon'_{t\ell}|^a)^{1/a}, \end{aligned} \tag{2.8}$$

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

There are no direct relationships between the usual mixing conditions and this "physical" functional dependence measure. But this measure is easier to handle mathematically and leads to simpler and stronger proofs in our paper, through the Nagaev-type inequality in Lemma 1. Moreover, many well-known processes are not strong mixing, yet can be handled by using the dependence measure (2.8), like the Bernoulli shift process in Andrews (1984).

2.4.1 Main assumptions and notations

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

A1. The entries in the matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are constants as $N, T \to \infty$, on top of the row sum conditions introduced after model (2.2) in section 2.2.

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

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

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

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

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

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

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

Assumption A1 can be relaxed, so that the weights in $\mathbf{W}_i^*$ can be decaying at a certain rate, at the expense of lengthier proofs. Assumption A2 is needed as demonstrated numerically in section 2.6. For moderate value of T, if the spatial weight matrices are sparse enough, then a slow decay rate is sufficient, which in practice means that the noise level is required to be not too large. Intuitively, low noise limits the correlation between spatial lags of $\mathbf{y}_t$ and the disturbance term, hence limiting a potential source of inconsistency that arises due to the simultaneous nature of the model. See also example 2 in section 2.4.2 for a simple illustration, and a remark therein about estimating the reduced form model (2.3) instead.

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

$$\mathbf{y}_t = \Big(\mathbf{I}_N - \boldsymbol{\Pi}_1^* \sum_{j=1}^d \beta_j^* \mathbf{W}_2^* B\Big)^{-1} \boldsymbol{\Pi}_1^* (\mathbf{W}_2^* \mathbf{z}_t \boldsymbol{\beta}_2^* + \boldsymbol{\epsilon}_t), \tag{2.9}$$

where $\mathbf{\Pi}_1^*$ is defined in (2.3), and B is the backward shift operator. For the inverse operator above to be defined (i.e. the system is stationary), we need

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

which impose constraints on $\boldsymbol{\beta}^*$ as well. Allowing past values as covariates extends the applicability of the model, since example 2 in section 2.4.2 demonstrates that covariates have to be included for sign consistent estimation.

The uniform boundedness assumption in A4 for elements of $\mathbf{\Sigma}_{xk}$ and $\mathbf{\Sigma}_\zeta$ is a direct consequence of the tail assumption in A5. We assume this for notational convenience only. The other half of assumption A4 says that either the cross-correlations between more "distant" components for the k-th covariate $\mathbf{X}_{t,k}$ are getting smaller quick enough, or this happens for the components in the noise $\boldsymbol{\epsilon}_t$. The settings in (2.7) and (2.8) allows us to assume either $\{X_{t,jk}^*\}_j$ or $\{\zeta_{t,j}^*\}_j$ is a martingale difference, which is weaker than assuming that as an independent sequence.

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

2.4.2 Example 2

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

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

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

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

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

The least square estimator for w_{12} is

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

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

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

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

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

If $\sigma_{X,1}^2 = \sigma_{X,2}^2 = 0$ (assumption A7 fails), and $\sigma_{\epsilon,1}^2 = \sigma_{\epsilon,2}^2$, we see that the asymptotic bias becomes independent of $\sigma_{\epsilon,j}^2$, and $\hat{w}_{12}$ and $\hat{w}_{21}$ cannot be both sign consistent. Hence it is important that covariates are included in our model. Luckily, assumption A3 allows for past values of $\mathbf{y}_t$ to be our covariates $\mathbf{X}_t$, although other exogenous covariates are still needed. See (2.9) in section 2.4.1 for more details.

One final remark is that, for this simple toy example, we may consistently estimate the parameters of the reduced form model like that in (2.3), and recover w_{12} and w_{21} from the estimated reduced form model without assumption A2. But, as explained in example 1, when N is large and a general spatial weight matrix is our target, the problem can become intractable and consistent estimation is then not achievable unless assumption A2 is satisfied. See also section 2.7 where an instrumental variable approach is mentioned and is still under research to overcome major technical difficulties when used together with LASSO.

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

$$J = \{j : \xi_j^* \neq 0, \text{and does not correspond to } w_{2,ss}^*, s = 1, \ldots, N\}. \tag{2.10}$$

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

$$\begin{aligned}
\mathcal{A}_1 &= \{\max_{1\le j,\ell\le N}\max_{1\le k\le K}|\frac{1}{T}\sum_{t=1}^{T}\zeta_{t,j}X_{t,\ell k}| < \lambda_T\},\\
\mathcal{A}_2 &= \{\max_{1\le k\le K}|\frac{1}{T}\sum_{j=1}^{N}\sum_{t=1}^{T}\zeta_{t,j}X_{t,jk}| < \lambda_T N^{1/2+1/2w}\},\\
\mathcal{A}_3 &= \{\max_{1\le i,j\le N}|\frac{1}{T}\sum_{t=1}^{T}[\zeta_{t,i}\zeta_{t,j} - E(\zeta_{t,i}\zeta_{t,j})]| < \lambda_T\}, \qquad (2.11)\\
\mathcal{A}_4 &= \{\max_{1\le i,j\le N}\max_{1\le \ell,m\le K}|\frac{1}{T}\sum_{t=1}^{T}X_{t,i\ell}X_{t,jm} - E(X_{t,i\ell}X_{t,jm})| < \lambda_T\},\\
\mathcal{M} &= \left\{\max_{1\le t\le T}\max_{1\le j\le N}\max_{1\le k\le K}|X_{t,jk}| < \left(\frac{3\log(T\vee N)}{D_2}\right)^{1/q}\right\},
\end{aligned}$$

where w is as defined in assumption A6.

2.4.3 Main results

We first present a Nagaev-type inequality for a general time series $\{\mathbf{x}_t\}$ under similar settings in (2.7) and (2.8), which is a combination of Theorems 2(ii) and 2(iii) of Liu et al. (2013).

Lemma 1. *For a zero mean time series process $\mathbf{x}_t = \mathbf{f}(\mathcal{F}_t)$ as defined in (2.7) with dependence measure $\theta^x_{t,a,j}$ as defined in (2.8), assume $\Theta^x_{m,w} \le Cm^{-\alpha}$ for some $w > 2$ and constants $C, \alpha > 0$. Then there exists constants C_1, C_2 and C_3 independent of v, T and the index j such that*

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

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

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

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

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

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

so that

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

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

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

Theorem 1. *Let assumptions A3 - A6 be satisfied. Suppose $\alpha > 1/2 - 1/w$, and suppose for the applications of the Nagaev-type inequality in Lemma 1 for the processes in $\mathcal{A}_1$ to $\mathcal{A}_4$, the constants C_1, C_2 and C_3 are the same. Then with $c \geq \sqrt{3/C_3}$ where c is the constant defined in λ_T, we have*

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

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

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

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

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

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

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

The proof is relegated to the Appendix. If we treat $\widetilde{\mathbf{W}}_1$ and $\widetilde{\mathbf{W}}_2$ as some assumed spatial weight matrices, for example distance matrices with a particular distance metric, this lemma together with Theorem 1 tells us that with high probability, the error upper bound for estimating $\boldsymbol{\beta}^*$ is related to the error for estimating the spatial weight matrices through $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1$. As long as $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1$ is much less than N, estimation error is related to how sparse the matrix $\mathbf{W}_2^*$ (i.e., s_2) is. Otherwise, the error can be large. We provide some simulation results for the estimation of $\boldsymbol{\beta}^*$ in section 2.6.

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

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

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

For $\widehat{\boldsymbol{\xi}}$, denote $\xi_{S,\min/\max} = \min/\max_{j\in S}\xi_j$ and $\widetilde{J}$ the LASSO estimator for J in (2.10). Then

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

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

The results in Theorem 2 are consistent with the properties of the LASSO estimators under the usual linear regression settings (see (3.2) of Bickel et al., 2009). With these oracle inequalities, we need to introduce a restricted eigenvalue condition which is similar to condition (3.1) of Bickel et al. (2009). We however define this condition on a population covariance matrix instead, since our raw design matrix $\mathbf{M}_{\boldsymbol{\beta}^*}$ in (2.4) is always random:

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

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

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

This condition is automatically satisfied if $\boldsymbol{\Sigma}$ has the smallest eigenvalue bounded uniformly away from 0. Similar population restricted eigenvalue condition is also introduced in Zhou et al. (2009)

for the analysis of LASSO and adaptive LASSO estimators when the design matrix is formed by i.i.d. rows which are multivariate normally distributed.

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

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

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

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

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

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

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

Corollary 4 says that, in addition to the assumptions in Theorem 3, if $\gamma_T n = o(1)$ also, then all the LASSO and adaptive LASSO estimators from (2.5) and (2.6) converge to their respective true quantities in L_1 norm on the set $\mathcal{A}_1 \cap \ldots \cap \mathcal{A}_4 \cap \mathcal{M}$, which has probability approaching 1 with explicit probability lower bound shown in Theorem 1. The need for large enough N, T are merely for the simplification of the different error bounds, and can be removed at the expense of more complicated expressions. The proof is omitted.

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

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

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

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

The rate $n = o(\gamma_T^{-\frac{2k}{k+1}})$ suggests that the number of non-zero elements allowed in the spatial weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ without violating sign consistency depends on the rate of decay for the variance of the noise. For instance if $\gamma_T \asymp \lambda_T \log^{1/2}(T \vee N)$ and $k = 1$, then $n = o(T^{1/2} \log^{-1}(T \vee N))$.

2.5 Practical Implementation

In this section, we provide details of the block coordinate descent (BCD) algorithm for carrying out the minimizations for (2.5) and (2.6). We need the BCD algorithm since the objective functions in these problems are not convex in $(\boldsymbol{\xi}, \boldsymbol{\beta})$, although given $\boldsymbol{\beta}$, they are convex in $\boldsymbol{\xi}$ and vice versa.

The BCD algorithm is closely related to the Iterative Coordinate Descent of Fan and Lv (2011), and is also discussed in Friedman et al. (2010) and Dicker et al. (2010). While it is difficult to establish global convergence of the BCD algorithm without convexity, it is easy to see that for (2.5) and (2.6), each iteration delivers an improvement of the objective functions since given one parameter, the objective functions are convex in the other. From our experience, starting from an appropriate initial value, a minimum will be achieved with good performance in practice. Indeed in the simulation experiments in section 2.6 (not shown), it is found that the algorithm is robust to a variety of initial values chosen.

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

<u>The Block Coordinate Descent Algorithm</u>

0. Start with an initial value $\boldsymbol{\xi} = \boldsymbol{\xi}^{(0)}$. This can be obtained by using $\boldsymbol{\beta}^{(0)} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}^v$ (for notations see Lemma 2), and solves (2.5) given $\boldsymbol{\beta}^{(0)}$ using LARS. This gives $\boldsymbol{\xi}^{(0)}$.

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

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

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

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

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

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

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

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

$$\mathrm{BIC}(\gamma_T) = \sum_{i=1}^{N} \log\left(T^{-1}\|\widetilde{\mathbf{y}}_i - (\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}_{\gamma_T})_i\|^2\right) + |S_{\gamma_T}|\frac{\log(T)}{T}\log(\log(2N-2)), \tag{2.12}$$

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 (2.5) with tuning parameter being γ_T. Also, $(\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}_{\gamma_T})_i$ is the vector with length T which is the portion of the vector $\mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}_{\gamma_T}$ (see equation (2.4)) 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 two spatial weight matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$, with response variable $\widetilde{\mathbf{y}}_i$. We denote γ_{BIC} the tuning parameter that minimizes the BIC criterion in (2.12). This γ_{BIC} will then be used in (2.5) to find the LASSO solution $\widetilde{\boldsymbol{\xi}}$. We use the same tuning parameter for the adaptive LASSO estimator in (2.6).

2.6 Numerical Examples

We give detailed simulation results in section 2.6.1 for our LASSO and adaptive LASSO estimators. A set of stock markets data is analyzed in section 2.6.2 to visualize the connection among international financial markets.

2.6.1 Simulation Results

We generate data from model (2.2) and investigate the practical performance of the LASSO and adaptive LASSO estimators.

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

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

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

Table 2.B.3 consider two more cases. One is when the covariates include a lagged variable $\mathbf{y}_{t-1}$ on top of X_t. We set $\boldsymbol{\beta}^* = (1, 0.5, 0.15)^{\mathrm{T}}$ which ensures the model for $\mathbf{y}_t$ is stationary. While when $N = 25$ results are similar to the baseline simulations, for $N = 50$ and 75 the performance is getting worse in general. This indicates that while in theory it is fine to include lagged variables, we may need a larger T or a limited N for good performance in practice.

Another case is when the noise exhibits spatial correlations. To this end, we randomly pick the off-diagonal elements in the noise covariance matrix to be 0.3, while keeping it sparse with around 95% elements still 0. The performance is similar to the baseline simulations in general. This is consistent with our theories. In particular this scenario fits assumption A4 (see section 2.4.1): when there are weak or no spatial correlations in the covariates, then the spatial correlation structure in the noise can be general.

Finally, Tables 2.B.4 and 2.B.5 show some results when some assumptions are violated. The first case is setting the variance of the noise equal to $\sigma_\epsilon^2 = 1$, instead of letting it decay as in the baseline simulations. Clearly the performance is worse in general even when $T = 200$. The results are consistent with Example 2 in section 2.4.2. The performance when there are no covariates is also shown. The poor performance all round under the absence of covariates is again consistent with Example 2 in section 2.4.2. Lastly, we simulate the noise using the t_3 distribution rather than normal distribution, violating the tail assumption A5 in section 2.4.1. While the performance is worse in general, it is still better than when there are no covariates or no variance decay. Hence the method is more robust to fat tails.

2.6.2 Analysis of stock markets data

It is well-known that worldwide stock markets' performance are dependent on other markets. To study their dependence structure in more detail, we use model (2.2) to analyze markets' returns over 2013. We estimate the spatial weight matrix $\mathbf{W}_1^*$ using the adaptive LASSO estimator. The response variable $\mathbf{y}_t$ is taken as the panel of stock market returns for the 26 biggest world markets. We use daily data available for the whole of 2013 ($T = 263$). See Table 2.C.1 for details of the markets and their respective indices.

For the covariates we use the S&P Global 1200 Index and the Dow Jones World Stock Index. By definition, firms that belong to the world index are constituents of the indices of some markets. Hence the exogeneity of the covariates cannot be sustained. Nevertheless, the global variables are included with the purpose of eliminating a global-wide variance that could prevent the identification of $\mathbf{W}_1^*$. Due to the lack of variance in the cross-sectional dimension, $\mathbf{W}_2^*$ is unidentified and is simply set as the identity matrix. The model is estimated by the adaptive LASSO, with the tuning parameter λ chosen by BIC, as described in section 2.5.

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

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

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

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

In Figure 2.C.1, the elements of $\widehat{\mathbf{W}}_1$ are plotted against the common opening hours. From this figure, it is clear that for markets with smaller overlap of opening hours, the estimated elements are zero in $\widehat{\mathbf{W}}_1$. In particular, there is no violation of the afore-mentioned restriction and markets are only affecting each other if they are commonly open for at least roughly half of their opening times.

2.7 Conclusion

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

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

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

Appendix

2.A Proofs

Proof of Theorem 1. We first show that, with the tail condition in A5 for a process $\{\mathbf{z}_t\}$, we have for any $w > 0$, $\max_j \|z_{tj}\|_{2w} \le \mu_{2w} < \infty$. Hence we can fix a w large enough such that $N = o(T^{w/4-1/2} \log^{w/4}(T))$; see Remark 3 after Theorem 1. Indeed by the Fubini's Theorem,

$$\begin{aligned} E|z_{tj}|^{2w} &= E\int_0^{|z_{tj}|^{2w}} ds = \int_0^\infty P(|z_{tj}|> s^{1/2w})\, ds \le \int_0^\infty D_1 \exp(-D_2 s^{q/2w})\, ds \\ &= \frac{4wD_1}{q}\int_0^\infty x^{4w/q-1} e^{-D_2x^2}\, dx = \frac{2wD_1}{qD_2^{2w/q}}\Gamma(2w/q) \text{ [define as } \mu_{2w}^{2w}] < \infty, \end{aligned} \tag{2.13}$$

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

$$\begin{aligned} P(\mathcal{A}_1^c) &\le \sum_{\substack{1\le j,\ell\le N \\ 1\le k\le K}} P\Big(\Big|T^{-1}\sum_{t=1}^T \zeta_{t,j}X_{t,\ell k}\Big| \ge \lambda_T\Big) \le N^2K\Big(\frac{C_1T}{(T\lambda_T)^w} + C_2\exp(-C_3T\lambda_T^2)\Big) \\ &\le C_1K\left(\frac{C_3}{3}\right)^{w/2} \frac{N^2}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2KN^2}{T^3\vee N^3}. \end{aligned} \tag{2.14}$$

Similarly, we have

$$\begin{aligned} P(\mathcal{A}_3^c) &\le C_1\left(\frac{C_3}{3}\right)^{w/2} \frac{N^2}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2N^2}{T^3\vee N^3}, \\ P(\mathcal{A}_4^c) &\le C_1K^2\left(\frac{C_3}{3}\right)^{w/2} \frac{N^2}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2K^2N^2}{T^3\vee N^3}. \end{aligned} \tag{2.15}$$

The tail assumption A5 and the union sum inequality imply that

$$P(\mathcal{M}^c) \le NTK\cdot D_1\exp(-3\log(T\vee N)) = \frac{D_1NTK}{T^3\vee N^3}. \tag{2.16}$$

Finally, if we can show that

$$\max_{1\le k\le K} \|N^{-\frac{1}{2}-\frac{1}{2w}}\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k}\|_{2w} < \infty, \tag{2.17}$$

$$\Theta_{m,2w} = \sum_{t=m}^{\infty} \max_{1\le k\le K} \|N^{-\frac{1}{2}-\frac{1}{2w}}(\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k} - \boldsymbol{\zeta}_t^{'\mathrm{T}}\mathbf{X}'_{t,k})\|_{2w} \le am^{-\alpha}, \tag{2.18}$$

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

$$\begin{aligned} P(\mathcal{A}_2^c) &\le \sum_{k=1}^{K} P\left(\left|T^{-1}\sum_{t=1}^{T} N^{-\frac{1}{2}-\frac{1}{2w}}\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k}\right| \ge \lambda_T\right) \\ &\le C_1\left(\frac{C_3}{3}\right)^{w/2}\frac{K}{T^{w/2-1}\log^{w/2}(T\vee N)} + \frac{C_2K}{T^3\vee N^3}. \end{aligned} \tag{2.19}$$

Combining (2.14), (2.15), (2.16) and (2.19), we can then use

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

to yield the conclusion of the Theorem. It remains to show (2.17) and (2.18).

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

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

since $\{X_{t,jk}^*\}_{1\le j\le N}$ is a martingale difference. Hence $\{(\boldsymbol{\zeta}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{xk}^{1/2})_jX_{t,jk}^*\}_{1\le j\le N}$ is a martingale difference. By Lemma 2.1 of Li (2003), assumptions A3, A4 and (2.13), we then have

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

so that $\max_{1\leq k\leq K}\|N^{-\frac{1}{2}-\frac{1}{2w}}\boldsymbol{\zeta}_t^{\mathrm{T}}\mathbf{X}_{t,k}\|_{2w}<\infty$, which is (2.17).

To prove (2.18), observe that

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

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

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

Hence we can use Lemma 2.1 of Li (2003), assumptions A3, A4, A6 and (2.13) to show that

$$\begin{aligned}
\|N^{-\frac{1}{2}-\frac{1}{2w}}\sum_{j=1}^{N}(\boldsymbol{\zeta}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{xk}^{1/2})_j(X_{t,jk}^*-X_{t,jk}'^*)\|_{2w} &\leq 36w(1+(2w-1)^{-1})^{1/2} \\
&\quad\cdot\left[N^{-2}\sum_{j=1}^{N}E|\max_{1\leq j\leq N}|\zeta_{t,j}||^{2w}\|\boldsymbol{\Sigma}_{xk}^{1/2}\|_\infty^{2w}(\theta_{t,2w,jk}^{x^*})^{2w}\right]^{1/2w} \\
&\leq 36w\mu_wS_x(1+(2w-1)^{-1})^{1/2}\max_{1\leq j\leq N}\theta_{t,2w,jk}^{x^*}.
\end{aligned}$$

Similarly,

$$\|N^{-\frac{1}{2}-\frac{1}{2w}}\sum_{j=1}^{N}(\boldsymbol{\zeta}_t^{\mathrm{T}}\boldsymbol{\Sigma}_{xk}^{1/2}-\boldsymbol{\zeta}_t'^{\mathrm{T}}\boldsymbol{\Sigma}_{xk}^{1/2})_j X_{t,jk}'^{*}\|_{2w}\le 36w\mu_w S_x(1+(2w-1)^{-1})^{1/2}\max_{1\le j\le N}\theta_{t,2w,j}^{\zeta}.$$

Hence combining and using assumption A6, we have

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

which is (2.18). The proof is now completed. □

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

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

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

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

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

$$\|(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})^{-1}\|_\infty\le\frac{K^{1/2}}{\lambda_{\min}(\mathbf{V}^{\mathrm{T}}E(\mathbf{U})\mathbf{V})}\le\frac{K^{1/2}}{\lambda_{\min}(E(\mathbf{U}))\lambda_{\min}(\mathbf{V}^{\mathrm{T}}\mathbf{V})}\le\frac{K^{1/2}}{uN}. \tag{2.20}$$

Then on $\mathcal{A}_4$, using (2.20),

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

Similarly on $\mathcal{A}_4$, using (2.20) and assumptions A1, A4,

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

Similarly on $\mathcal{A}_1$ and $\mathcal{A}_2$, using (2.20) and assumptions A1, A4,

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

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

$$\|(\mathbf{I}_N - \mathbf{W}_1^*)^{-1}\|_\infty \leq \sum_{k\geq 0} \|\mathbf{W}_1^*\|_\infty^k \leq \sum_{k\geq 0} \eta^k = (1-\eta)^{-1}. \tag{2.21}$$

Hence using this, (2.20) and assumptions A1,A4, on $\mathcal{A}_1$ and $\mathcal{A}_4$, we have (tedious algebra omitted)

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

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

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

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

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

Proof of Theorem 2. For the LASSO estimator $\widetilde{\boldsymbol{\xi}}$, (2.5) implies

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

which, using model (2.4), can be rearranged to

$$\begin{aligned}\frac{1}{2T}\|\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\xi}^* - \mathbf{M}_{\widetilde{\boldsymbol{\beta}}}\widetilde{\boldsymbol{\xi}}\|^2 \leq{}& \frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) + \frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{X}_{\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}}_2 - \mathrm{vec}(\mathbf{I}_N))\\ &+ \frac{1}{T}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{M}_{\boldsymbol{\beta}^*}(\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*) + \gamma_T(\|\boldsymbol{\xi}^*\|_1 - \|\widetilde{\boldsymbol{\xi}}\|_1). \qquad (2.22)\end{aligned}$$

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

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

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

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

Finally,

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

Writing the ℓ-th row of $\boldsymbol{\Pi}_1$ as $\boldsymbol{\pi}_{1,\ell}^{\mathrm{T}}$, using (2.21), we have on $\mathcal{A}_1$ and $\mathcal{A}_3$,

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

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

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

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

Hence utilizing all these bounds, (2.22) becomes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

so that by assumption A8,

$$\kappa(n)\|\boldsymbol{\alpha}_J\| \leq \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}\| \leq T^{-1/2}\|\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\alpha}\| + \epsilon^{1/2}n^{1/2}(1+c_0)\|\boldsymbol{\alpha}_J\|. \tag{2.23}$$

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

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

which is an intermediate result from the proof of Theorem 2, we can apply (2.23) to have, on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$,

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

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

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

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

$$\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}^*_J\|^{1/2} \leq \frac{\gamma_T^{1/2} n^{1/4} + \big[\gamma_T n^{1/2} + \kappa(n) h_{1,N,T}\big]^{1/2}}{\kappa(n) - (1+c_0)n^{1/2}h_{2,N,T} - h_{3,N,T}}, \text{ so that}$$

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

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

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

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

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

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

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

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

which is from the proof of Theorem 2, putting $\boldsymbol{\alpha} = \widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*$ in (2.23), we have on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$,

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

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

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

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

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

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

The largest upper bound is given by $\max_{1\le i,j\le N}|T^{-1}\sum_{t=1}^{T} y_{ti}y_{tj} - E(y_{ti}y_{tj})|$ (details omitted), where using $y_{ti} = \boldsymbol{\pi}_{1,i}^{*\mathrm{T}}\mathbf{W}_2^*\mathbf{X}_t\boldsymbol{\beta}^* + \boldsymbol{\pi}_{1,i}^{*\mathrm{T}}\boldsymbol{\epsilon}_t$ (see (2.3), with $\boldsymbol{\pi}_{1,i}^{*\mathrm{T}}$ the i-th row of $\boldsymbol{\Pi}_1^*$),

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

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

Proof of Theorem 5. First, similar to (2.23), we can use assumption A8 for $\|\boldsymbol{\alpha}_{J^c}\|_1 \le c_0\|\boldsymbol{\alpha}_J\|_1$ to arrive at $\kappa(n)\|\boldsymbol{\alpha}_{J^c}\| \le T^{-1/2}\|\mathbf{M}_{\boldsymbol{\beta}^*}\boldsymbol{\alpha}\| + \epsilon^{1/2}n^{1/2}(1+c_0)\|\boldsymbol{\alpha}_J\|$. Putting $\boldsymbol{\alpha} = \widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*$ and follow the same lines as in the proof of Theorem 3, we can use $\|\widetilde{\boldsymbol{\xi}}_J - \boldsymbol{\xi}_J^*\| = O(\gamma_T n^{1/2})$ on $\mathcal{A}_1\cap\cdots\cap\mathcal{A}_4\cap\mathcal{M}$ (by the result of Theorem 3) to show that, for $j\in J^c$,

$$\widetilde{\xi}_j \le \|\widetilde{\boldsymbol{\xi}}_{J^c}\| = \|\widetilde{\boldsymbol{\xi}}_{J^c} - \boldsymbol{\xi}_{J^c}^*\| = O(\gamma_T n^{1/2}). \tag{2.24}$$

Define the set $D = \{j : \xi_j^*$ does not corr. to diagonal elements of $\mathbf{W}_1^*, \mathbf{W}_2^*\}$. The KKT condition implies that $\widehat{\boldsymbol{\xi}}$ is a solution to (2.6) if and only if there exists a subgradient

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

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

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

where we denote $\widehat{\mathbf{M}} = \mathbf{M}_{\widehat{\boldsymbol{\beta}}}$ and $\mathbf{M}^* = \mathbf{M}_{\boldsymbol{\beta}^*}$, and we use A_S to denote the matrix A with columns restricted to the index set S. Substituting $\mathbf{y} = \mathbf{M}^*_D\boldsymbol{\xi}^*_D + \mathbf{X}_{\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) + \boldsymbol{\epsilon}$,

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

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

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

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

$$\mathrm{sign}(\widehat{\boldsymbol{\xi}}_J) = \mathrm{sign}\{\boldsymbol{\xi}^*_J + I_1 + I_2 + I_3 + I_4 + I_5\} = \mathrm{sign}(\boldsymbol{\xi}^*_J); \tag{2.25}$$

$$|D_1 + D_2 + D_3 + D_4 + D_5| \le \gamma_T\mathbf{v}_{J'} \tag{2.26}$$

as the necessary and sufficient conditions for $\widehat{\boldsymbol{\xi}}$ to be a sign consistent solution to (2.6), where

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

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

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

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

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

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

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

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

For proving (2.25), it suffices to show that $\|I_j\|_\infty = o(1)$ since by assumption A1, ξ_j^* is a constant for $j \in J$. Consider

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

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

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

Hence we have proved (2.25) on $\mathcal{A}_1 \cap \cdots \mathcal{A}_4 \cap \mathcal{M}$ when N, T are large enough.

For proving (2.26) on $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_4 \cap \mathcal{M}$ when N, T are large enough, it suffices to show by (2.24) that

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

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

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

The largest order is $\|D_4\|_\infty = O(\gamma_T n^{1/2})$, which is of smaller order than $\gamma_T/(\gamma_T n^{1/2})^k$ by the assumption $n = o(\gamma_T^{-\frac{2k}{k+1}})$. This proves (2.26), and completes the proof of the theorem. □

2.B Simulations

Table 2.B.1: Baseline Simulations. All values are averages over 100 simulations. Penalization is chosen via BIC criteria. Specificity is the percentage of zeros estimated as zeros. Sensitivity is the percentage of non-zeros estimated as non-zeros. LASSO L_1 is the L_1 error norm $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1$ for the LASSO estimator, and AdaLASSO represents the adaptive LASSO. Bias is the sum of error for the estimated non-zero values without taking absolute values. Standard errors in parenthesis. True sparsity level of the both $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ is $\kappa = 0.95$.

		$T = 50$		$T = 100$		$T = 200$	
		$\mathbf{W}_1^*$	$\mathbf{W}_2^*$	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$
$N = 25$	Specificity	97.02% (0.011)	98.22% (0.008)	96.74% (0.010)	98.20% (0.008)	96.64% (0.011)	98.36% (0.008)
	Sensitivity	78.09% (0.083)	55.38% (0.103)	95.70% (0.042)	86.76% (0.065)	99.35% (0.014)	96.19% (0.032)
	Bias	−0.0660 (0.024)	−0.1105 (0.031)	−0.0391 (0.015)	−0.0738 (0.017)	−0.0220 (0.009)	−0.0394 (0.011)
	LASSO L_1	18.8344 (2.178)	18.2203 (2.407)	18.0305 (1.780)	18.8540 (2.066)	15.9489 (1.702)	16.8550 (1.810)
	LASSO L_2	5.5494 (1.011)	7.2172 (1.046)	3.4079 (0.650)	4.1905 (0.759)	2.2531 (0.481)	2.3123 (0.401)
	AdaLASSO L_1	2.1840 (0.368)	2.5987 (0.452)	1.7145 (0.276)	2.0779 (0.357)	1.3482 (0.221)	1.5522 (0.243)
	AdaLASSO L_2	1.0609 (0.241)	1.7634 (0.315)	0.4505 (0.140)	0.7627 (0.203)	0.2067 (0.075)	0.2858 (0.096)
	Sparsity	0.9349 (0.014)	0.9349 (0.014)	0.9233 (0.012)	0.9233 (0.012)	0.9202 (0.013)	0.9202 (0.013)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0857 (0.0327)		0.0173 (0.0121)		0.0073 (0.0056)	
$N = 50$	Specificity	95.70% (0.007)	98.38% (0.005)	96.20% (0.007)	98.35% (0.005)	96.60% (0.006)	98.47% (0.005)
	Sensitivity	74.35% (0.045)	42.23% (0.050)	92.54% (0.029)	81.18% (0.043)	98.32% (0.013)	96.15% (0.018)
	Bias	−0.0448 (0.011)	−0.0972 (0.016)	−0.0336 (0.006)	−0.0799 (0.011)	−0.0215 (0.004)	−0.0412 (0.006)
	LASSO L_1	66.7238 (3.839)	61.6638 (4.002)	64.5299 (4.564)	66.7991 (5.325)	59.4202 (4.480)	63.2523 (4.785)
	LASSO L_2	25.2673 (2.012)	31.8719 (2.073)	15.3925 (1.655)	18.7294 (1.509)	9.0062 (1.159)	9.4297 (1.044)
	AdaLASSO L_1	7.8904 (0.652)	10.1448 (0.803)	5.8510 (0.637)	7.4972 (0.809)	4.6307 (0.496)	5.5847 (0.585)
	AdaLASSO L_2	4.5845 (0.478)	8.2501 (0.604)	1.9969 (0.313)	3.7515 (0.479)	0.8043 (0.178)	1.1878 (0.254)
	Sparsity	0.9240 (0.007)	0.9240 (0.007)	0.9182 (0.007)	0.9182 (0.007)	0.9201 (0.007)	0.9201 (0.007)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0257 (0.0233)		0.0283 (0.0167)		0.0309 (0.0098)	
$N = 75$	Specificity	95.30% (0.005)	98.90% (0.003)	96.35% (0.004)	98.88% (0.003)	97.16% (0.003)	98.98% (0.003)
	Sensitivity	59.54% (0.034)	26.88% (0.033)	85.53% (0.026)	76.25% (0.033)	95.42% (0.013)	96.04% (0.014)
	Bias	−0.0224 (0.009)	−0.0973 (0.015)	−0.0277 (0.005)	−0.0911 (0.007)	−0.0196 (0.003)	−0.0483 (0.005)
	LASSO L_1	131.4265 (4.475)	111.8097 (5.187)	120.7178 (6.361)	120.3575 (7.167)	113.0090 (6.296)	120.1324 (7.211)
	LASSO L_2	65.0015 (3.615)	75.7601 (2.881)	35.5961 (2.409)	46.1954 (2.351)	19.7648 (1.537)	21.3982 (1.777)
	AdaLASSO L_1	15.7064 (0.752)	21.8860 (1.054)	10.1854 (0.777)	13.9803 (1.011)	7.8193 (0.627)	9.9623 (0.832)
	AdaLASSO L_2	11.7311 (0.867)	20.8000 (0.836)	4.5032 (0.440)	9.9502 (0.795)	1.7424 (0.241)	2.8229 (0.454)
	Sparsity	0.9262 (0.005)	0.9262 (0.005)	0.9239 (0.004)	0.9239 (0.004)	0.9262 (0.004)	0.9262 (0.004)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0274 (0.0200)		0.0343 (0.0170)		0.0348 (0.0101)	

Table 2.B.2: Baseline Simulations. All values are averages over 100 simulations. Penalization is chosen via BIC criteria. Specificity is the percentage of zeros estimated as zeros. Sensitivity is the percentage of non-zeros estimated as non-zeros. LASSO L_1 is the L_1 error norm $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1$ for the LASSO estimator, and AdaLASSO represents the adaptive LASSO. Bias is the sum of error for the estimated non-zero values without taking absolute values. Standard errors in parenthesis. True sparsity level of the both $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ is $\kappa = 0.99$.

		$T = 50$		$T = 100$		$T = 200$	
		$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$	Specificity	99.72% (0.003)	99.86% (0.002)	99.54% (0.003)	99.72% (0.003)	99.47% (0.003)	99.79% (0.002)
	Sensitivity	63.14% (0.231)	40.52% (0.214)	94.28% (0.094)	84.98% (0.146)	99.56% (0.032)	97.43% (0.065)
	Bias	−0.1189 (0.052)	−0.1305 (0.064)	−0.0678 (0.030)	−0.0881 (0.033)	−0.0328 (0.017)	−0.0424 (0.022)
	LASSO L_1	10.7485 (1.581)	8.9524 (1.681)	10.9814 (1.475)	10.5877 (1.653)	9.3435 (1.230)	9.5100 (1.252)
	LASSO L_2	1.2140 (0.381)	1.5384 (0.409)	0.7173 (0.245)	0.9196 (0.297)	0.3881 (0.168)	0.4341 (0.186)
	AdaLASSO L_1	0.9456 (0.192)	0.9303 (0.209)	0.7516 (0.153)	0.8293 (0.197)	0.5439 (0.112)	0.6016 (0.122)
	AdaLASSO L_2	0.2712 (0.107)	0.4082 (0.124)	0.1017 (0.055)	0.1731 (0.086)	0.0334 (0.028)	0.0484 (0.042)
	Sparsity	0.9912 (0.004)	0.9912 (0.004)	0.9865 (0.004)	0.9865 (0.004)	0.9856 (0.005)	0.9856 (0.005)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0232 (0.0114)		0.0100 (0.0076)		0.0084 (0.0061)	
$N = 50$	Specificity	99.75% (0.002)	99.88% (0.001)	99.57% (0.002)	99.76% (0.001)	99.54% (0.002)	99.82% (0.001)
	Sensitivity	61.80% (0.127)	34.74% (0.125)	93.04% (0.056)	84.03% (0.067)	99.14% (0.021)	96.79% (0.035)
	Bias	−0.1128 (0.030)	−0.1375 (0.036)	−0.0695 (0.016)	−0.0936 (0.017)	−0.0338 (0.008)	−0.0441 (0.010)
	LASSO L_1	33.6100 (4.133)	24.9089 (3.769)	39.8098 (3.169)	36.4643 (3.391)	34.8279 (2.949)	35.1975 (3.435)
	LASSO L_2	4.9503 (0.884)	6.4540 (0.983)	2.9538 (0.551)	3.8265 (0.690)	1.6247 (0.299)	1.6463 (0.359)
	AdaLASSO L_1	2.9374 (0.453)	2.7847 (0.460)	2.6580 (0.301)	2.8045 (0.350)	1.9250 (0.244)	2.1301 (0.305)
	AdaLASSO L_2	1.1193 (0.253)	1.7546 (0.316)	0.4390 (0.134)	0.7485 (0.183)	0.1423 (0.058)	0.1972 (0.090)
	Sparsity	0.9915 (0.002)	0.9915 (0.002)	0.9868 (0.002)	0.9868 (0.002)	0.9854 (0.002)	0.9854 (0.002)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0248 (0.0164)		0.0132 (0.0097)		0.0087 (0.0062)	
$N = 75$	Specificity	99.79% (0.001)	99.91% (0.001)	99.54% (0.001)	99.74% (0.001)	99.56% (0.001)	99.84% (0.001)
	Sensitivity	52.25% (0.140)	24.37% (0.098)	93.66% (0.034)	83.38% (0.056)	99.17% (0.012)	97.46% (0.023)
	Bias	−0.1228 (0.023)	−0.1466 (0.030)	−0.0669 (0.009)	−0.0935 (0.013)	−0.0326 (0.005)	−0.0450 (0.008)
	LASSO L_1	59.9314 (9.852)	39.7276 (7.405)	80.7885 (4.056)	71.3078 (4.727)	74.6762 (4.159)	74.8206 (5.213)
	LASSO L_2	12.1496 (1.295)	15.1889 (1.247)	6.7000 (0.852)	8.3786 (1.078)	3.5099 (0.480)	3.5854 (0.601)
	AdaLASSO L_1	5.4167 (0.949)	5.1533 (0.755)	5.3577 (0.391)	5.5670 (0.474)	3.9939 (0.347)	4.4054 (0.441)
	AdaLASSO L_2	2.8895 (0.505)	4.2755 (0.446)	0.9576 (0.186)	1.6567 (0.295)	0.2951 (0.092)	0.4148 (0.146)
	Sparsity	0.9927 (0.003)	0.9927 (0.003)	0.9861 (0.002)	0.9861 (0.002)	0.9854 (0.001)	0.9854 (0.001)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0466 (0.0186)		0.0183 (0.0133)		0.0100 (0.0067)	

Table 2.B.3: Comparisons to the baseline simulations when the covariates include $\mathbf{y}_{t-1}$ (under the columns "Time Dependence") and when the noise exhibits spatial correlations (under the columns "Spatial Dependence"). Refer to Table 2.B.1 for the explanations of different items.

		Time Dependence				*Spatial Dependence*			
		$T = 100$		$T = 200$		$T = 100$		$T = 200$	
		$\mathbf{W}_1^*$	$\mathbf{W}_2^*$	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$	$\mathbf{W}_1^*$	$\mathbf{W}_2^*$
$N = 25$	Specificity	96.35% (0.013)	98.04% (0.007)	96.46% (0.010)	98.34% (0.008)	96.73% (0.009)	98.23% (0.007)	96.54% (0.011)	98.12% (0.009)
	Sensitivity	94.04% (0.046)	84.39% (0.067)	99.44% (0.013)	93.86% (0.070)	94.71% (0.047)	88.03% (0.059)	99.22% (0.018)	96.00% (0.033)
	Bias	−0.0361 (0.014)	−0.0686 (0.018)	−0.0195 (0.011)	−0.0477 (0.015)	−0.0464 (0.013)	−0.0761 (0.019)	−0.0235 (0.011)	−0.0395 (0.012)
	LASSO L_1	18.4701 (2.025)	19.6248 (1.896)	16.0530 (1.783)	16.9714 (1.900)	18.1214 (1.609)	18.7202 (1.666)	16.3036 (1.686)	17.4943 (1.950)
	LASSO L_2	3.7210 (0.737)	4.5611 (0.795)	2.4119 (0.612)	2.7056 (0.749)	3.5858 (0.577)	4.1266 (0.729)	2.3510 (0.509)	2.4653 (0.500)
	AdaLASSO L_1	1.7919 (0.325)	2.1952 (0.361)	1.3490 (0.211)	1.5641 (0.264)	1.7328 (0.261)	2.0427 (0.294)	1.4041 (0.233)	1.6380 (0.288)
	AdaLASSO L_2	0.5181 (0.152)	0.8554 (0.213)	0.2486 (0.126)	0.3794 (0.218)	0.4805 (0.124)	0.7400 (0.188)	0.2291 (0.084)	0.3144 (0.107)
	Sparsity	0.9216 (0.012)	0.9216 (0.012)	0.9199 (0.011)	0.9199 (0.011)	0.9225 (0.010)	0.9225 (0.010)	0.9194 (0.011)	0.9194 (0.011)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0184 (0.0094)		0.0101 (0.0060)		0.0230 (0.0127)		0.0070 (0.0058)	
$N = 50$	Specificity	95.39% (0.010)	97.96% (0.006)	95.96% (0.007)	98.37% (0.005)	96.15% (0.007)	98.37% (0.006)	96.64% (0.005)	98.51% (0.005)
	Sensitivity	91.07% (0.029)	67.93% (0.125)	98.33% (0.012)	86.47% (0.065)	93.58% (0.024)	81.74% (0.030)	98.65% (0.012)	95.69% (0.018)
	Bias	−0.0380 (0.008)	−0.0964 (0.021)	−0.0244 (0.005)	−0.0725 (0.017)	−0.0339 (0.006)	−0.0774 (0.009)	−0.0206 (0.004)	−0.0421 (0.006)
	LASSO L_1	73.0988 (7.870)	83.6656 (11.432)	67.8401 (4.699)	80.4248 (6.820)	64.8915 (4.553)	66.4893 (5.642)	59.2335 (4.163)	62.9595 (4.951)
	LASSO L_2	18.1152 (3.076)	23.2770 (4.087)	11.3419 (2.160)	14.5371 (2.991)	15.4289 (1.492)	18.3139 (1.185)	9.1422 (0.970)	9.4313 (0.972)
	AdaLASSO L_1	7.2820 (1.261)	10.6510 (2.312)	5.6221 (0.620)	8.3195 (1.241)	5.8783 (0.641)	7.3947 (0.870)	4.6112 (0.471)	5.6047 (0.618)
	AdaLASSO L_2	2.4515 (0.550)	5.1805 (1.352)	1.1335 (0.313)	2.5783 (0.865)	1.9934 (0.281)	3.6315 (0.351)	0.8176 (0.137)	1.2385 (0.225)
	Sparsity	0.9111 (0.011)	0.9111 (0.011)	0.9140 (0.008)	0.9140 (0.008)	0.9171 (0.007)	0.9171 (0.007)	0.9189 (0.007)	0.9189 (0.007)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.0351 (0.0211)		0.0274 (0.0126)		0.0306 (0.0180)		0.0349 (0.0096)	
$N = 75$	Specificity	92.43% (0.006)	94.97% (0.018)	87.74% (0.009)	90.68% (0.024)	96.44% (0.005)	98.90% (0.003)	97.20% (0.003)	98.99% (0.003)
	Sensitivity	70.69% (0.026)	17.31% (0.032)	88.79% (0.023)	25.72% (0.039)	84.79% (0.025)	75.73% (0.034)	95.25% (0.014)	96.17% (0.010)
	Bias	−0.0335 (0.007)	−0.1890 (0.019)	−0.0299 (0.005)	−0.2028 (0.019)	−0.0260 (0.004)	−0.0920 (0.010)	−0.0196 (0.002)	−0.0489 (0.005)
	LASSO L_1	209.5463 (5.333)	258.5049 (5.432)	268.4002 (5.614)	308.2939 (6.467)	119.7554 (6.982)	118.3699 (8.395)	112.5645 (6.037)	118.9015 (7.546)
	LASSO L_2	66.6747 (5.213)	102.1036 (11.582)	71.6813 (6.040)	114.2675 (15.065)	35.6206 (2.686)	45.6831 (2.580)	19.8162 (1.450)	21.4568 (1.229)
	AdaLASSO L_1	27.8593 (1.200)	43.5799 (1.526)	32.3066 (1.089)	46.4623 (1.705)	10.0237 (0.807)	13.7319 (1.111)	7.7980 (0.553)	9.8648 (0.861)
	AdaLASSO L_2	10.4046 (1.195)	26.3588 (2.434)	9.0844 (1.069)	26.7842 (3.080)	4.5115 (0.475)	9.8667 (0.797)	1.7587 (0.204)	2.8241 (0.292)
	Sparsity	0.8942 (0.006)	0.8942 (0.006)	0.8409 (0.008)	0.8409 (0.008)	0.9248 (0.005)	0.9248 (0.005)	0.9266 (0.003)	0.9266 (0.003)
	$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$	0.1006 (0.0300)		0.1054 (0.0250)		0.0343 (0.0207)		0.0340 (0.0097)	

Table 2.B.4: Comparisons to the baseline simulations when assumptions are violated. Refer to Table 2.B.1 for the explanations of different items.

		No Variance Decay				*Fat Tails*			
		$T = 100$		$T = 200$		$T = 100$		$T = 200$	
		$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$	Specificity	96.24% (0.013)	97.87% (0.010)	95.62% (0.012)	97.42% (0.010)	93.44% (0.021)	95.73% (0.016)	91.81% (0.016)	94.55% (0.014)
	Sensitivity	95.06% (0.044)	84.64% (0.065)	99.11% (0.016)	95.98% (0.028)	88.76% (0.066)	58.61% (0.099)	98.10% (0.026)	84.91% (0.081)
	Bias	−0.0422 (0.015)	−0.0765 (0.020)	−0.0255 (0.011)	−0.0459 (0.010)	−0.0486 (0.021)	−0.1109 (0.052)	−0.0382 (0.015)	−0.0871 (0.019)
	LASSO L_1	19.6605 (2.195)	20.6833 (2.493)	18.4888 (1.549)	20.2737 (1.717)	27.5608 (3.601)	30.5705 (4.472)	25.5832 (1.923)	30.7968 (2.678)
	LASSO L_2	3.9067 (0.699)	4.5202 (0.688)	2.8648 (0.483)	3.0230 (0.419)	7.5948 (2.215)	9.4319 (2.066)	6.6359 (1.153)	6.9125 (1.234)
	AdaLASSO L_1	1.9409 (0.358)	2.4089 (0.442)	1.6541 (0.198)	2.0208 (0.243)	4.3815 (2.534)	6.0911 (2.763)	3.1113 (0.566)	4.5200 (0.798)
	AdaLASSO L_2	0.5298 (0.145)	0.8437 (0.195)	0.2911 (0.082)	0.3834 (0.096)	1.9702 (2.338)	2.8222 (2.147)	1.0896 (0.403)	1.4446 (0.468)
	Sparsity	0.9176 (0.013)	0.9176 (0.013)	0.9139 (0.012)	0.9139 (0.012)	0.8948 (0.021)	0.8948 (0.021)	0.8707 (0.018)	0.8707 (0.018)
	$\lVert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rVert_1$	0.0181 (0.0115)		0.0103 (0.0076)		0.0411 (0.0297)		0.0237 (0.0170)	
$N = 50$	Specificity	95.70% (0.006)	98.15% (0.005)	95.69% (0.007)	97.92% (0.005)	93.57% (0.009)	96.93% (0.008)	92.59% (0.008)	95.99% (0.007)
	Sensitivity	92.41% (0.022)	77.73% (0.044)	98.22% (0.013)	93.97% (0.025)	80.41% (0.037)	50.10% (0.062)	94.01% (0.026)	82.04% (0.042)
	Bias	−0.0367 (0.006)	−0.0844 (0.010)	−0.0270 (0.005)	−0.0520 (0.007)	−0.0464 (0.009)	−0.0999 (0.018)	−0.0432 (0.007)	−0.0849 (0.013)
	LASSO L_1	70.0053 (4.594)	71.3662 (4.805)	69.1858 (3.930)	74.9824 (5.178)	97.5160 (6.324)	101.2189 (8.907)	94.7012 (5.161)	110.6415 (7.469)
	LASSO L_2	17.0935 (1.748)	20.1647 (1.592)	11.4220 (1.382)	12.6971 (1.173)	28.6774 (3.548)	35.0926 (4.227)	23.0987 (2.806)	27.3068 (3.296)
	AdaLASSO L_1	6.6601 (0.633)	8.4296 (0.782)	5.7215 (0.465)	7.3379 (0.723)	12.8711 (2.350)	18.6539 (3.868)	10.6893 (1.644)	15.9282 (2.578)
	AdaLASSO L_2	2.2809 (0.327)	4.1629 (0.457)	1.0718 (0.213)	1.7761 (0.298)	5.3795 (1.724)	9.6082 (2.528)	3.6077 (1.219)	6.0230 (1.814)
	Sparsity	0.9129 (0.008)	0.9129 (0.008)	0.9126 (0.007)	0.9126 (0.007)	0.8984 (0.009)	0.8984 (0.009)	0.8850 (0.008)	0.8850 (0.008)
	$\lVert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rVert_1$	0.0303 (0.0173)		0.0352 (0.0133)		0.0482 (0.0305)		0.0538 (0.0191)	
$N = 75$	Specificity	95.99% (0.005)	98.77% (0.004)	96.49% (0.004)	98.73% (0.003)	94.16% (0.006)	98.04% (0.004)	94.03% (0.006)	97.33% (0.006)
	Sensitivity	83.29% (0.027)	70.74% (0.033)	94.63% (0.017)	93.24% (0.019)	71.87% (0.030)	38.61% (0.035)	88.41% (0.023)	73.94% (0.039)
	Bias	−0.0286 (0.005)	−0.0970 (0.008)	−0.0249 (0.003)	−0.0652 (0.006)	−0.0326 (0.005)	−0.1019 (0.013)	−0.0386 (0.004)	−0.1006 (0.010)
	LASSO L_1	129.4730 (6.808)	127.1148 (9.643)	129.9715 (6.918)	139.4748 (8.089)	182.9601 (8.852)	172.0112 (11.093)	184.9879 (9.496)	209.7532 (14.129)
	LASSO L_2	39.2023 (2.474)	50.7696 (2.523)	24.5651 (1.722)	28.8266 (2.391)	60.0113 (4.521)	78.0560 (5.120)	45.6243 (4.509)	60.0513 (5.580)
	AdaLASSO L_1	11.3512 (0.869)	15.6470 (1.436)	9.7166 (0.724)	12.9432 (0.999)	21.7642 (2.443)	31.2314 (3.847)	18.8524 (2.802)	29.5169 (4.897)
	AdaLASSO L_2	5.1309 (0.462)	11.4221 (0.747)	2.3472 (0.241)	4.4590 (0.650)	10.2597 (1.910)	21.5966 (3.088)	6.5760 (1.906)	13.8220 (2.990)
	Sparsity	0.9212 (0.004)	0.9212 (0.004)	0.9211 (0.005)	0.9211 (0.005)	0.9093 (0.006)	0.9093 (0.006)	0.9008 (0.006)	0.9008 (0.006)
	$\lVert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rVert_1$	0.0367 (0.0160)		0.0402 (0.0108)		0.0655 (0.0241)		0.0644 (0.0138)	

Table 2.B.5: Simulations without covariates. Comparisons to the baseline simulations when assumptions are violated. Refer to Table 2.B.1 for the explanations of different items.

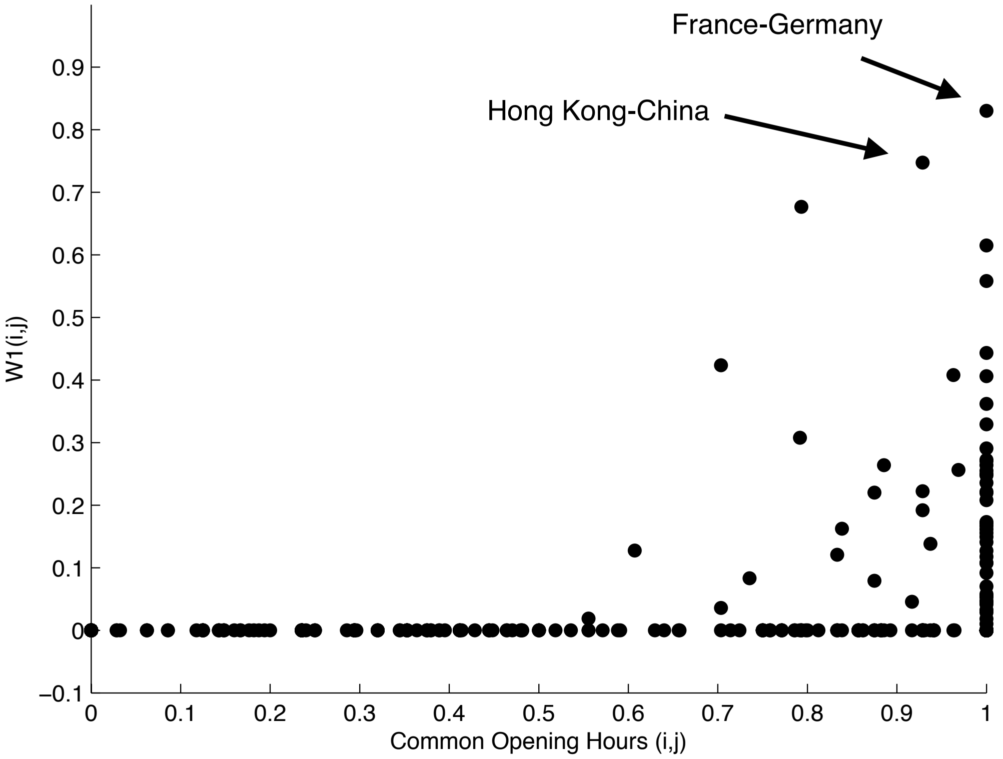
		$T = 100$		$T = 200$	
		$\mathbf{W_1^*}$	$\mathbf{W_2^*}$	$\mathbf{W_1^*}$	$\mathbf{W_2^*}$
$N = 25$	Specificity	96.00% (0.021)	– (–)	93.27% (0.017)	– (–)
	Sensitivity	59.22% (0.198)	– (–)	90.63% (0.075)	– (–)
	Bias	-0.1089 (0.040)	– (–)	-0.0928 (0.024)	– (–)
	LASSO L_1	7.6912 (0.751)	– (–)	7.5505 (0.661)	– (–)
	LASSO L_2	7.6912 (0.751)	– (–)	7.5505 (0.661)	– (–)
	AdaLASSO L_1	1.5748 (0.228)	– (–)	1.2060 (0.136)	– (–)
	AdaLASSO L_2	1.5748 (0.228)	– (–)	1.2060 (0.136)	– (–)
	Sparsity	0.9324 (0.029)	– (–)	0.8907 (0.018)	– (–)
	$\lVert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rVert_1$	– (–)		– (–)	
$N = 50$	Specificity	95.76% (0.007)	– (–)	94.40% (0.008)	– (–)
	Sensitivity	63.47% (0.070)	– (–)	86.84% (0.039)	– (–)
	Bias	-0.0825 (0.015)	– (–)	-0.0804 (0.015)	– (–)
	LASSO L_1	27.1855 (1.406)	– (–)	26.3433 (1.840)	– (–)
	LASSO L_2	27.1855 (1.406)	– (–)	26.3433 (1.840)	– (–)
	AdaLASSO L_1	4.8163 (0.366)	– (–)	3.9884 (0.346)	– (–)
	AdaLASSO L_2	4.8163 (0.366)	– (–)	3.9884 (0.346)	– (–)
	Sparsity	0.9279 (0.009)	– (–)	0.9032 (0.008)	– (–)
	$\lVert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rVert_1$	– (–)		– (–)	
$N = 75$	Specificity	95.46% (0.007)	– (–)	94.58% (0.006)	– (–)
	Sensitivity	57.03% (0.063)	– (–)	76.97% (0.043)	– (–)
	Bias	-0.0685 (0.012)	– (–)	-0.0684 (0.012)	– (–)
	LASSO L_1	55.0692 (3.474)	– (–)	51.4000 (2.648)	– (–)
	LASSO L_2	55.0692 (3.474)	– (–)	51.4000 (2.648)	– (–)
	AdaLASSO L_1	8.5933 (0.714)	– (–)	6.9086 (0.544)	– (–)
	AdaLASSO L_2	8.5933 (0.714)	– (–)	6.9086 (0.544)	– (–)
	Sparsity	0.9283 (0.009)	– (–)	0.9099 (0.006)	– (–)
	$\lVert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rVert_1$	– (–)		– (–)	

2.C Application

Table 2.C.1: Markets and their respective indices used. Data source: *Global Financial Data.*

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

Figure 2.C.1: Elements of $\widehat{\mathbf{W}}_1$ plotted against Common Opening Hours.



Chapter 3

Detection and Estimation of Block Structure in Spatial Weight Matrix

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.[1]

Keywords: Spatial weight matrix; LASSO penalization; zero-block consistency; spatial lag/error model; Nagaev-type inequality.

JEL classification: C31, C33.

[1] Paper coauthored with Clifford Lam, London School of Economics, Department of Statistics.

3.1 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 (3.3) and section 3.2 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 3.4, 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 3.2, 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 3.3 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 3.4 relaxes all the previous settings and results to overlapping blocks. Section 3.5 presents our simulation results as well as the complete analysis of the US Senate voting data. Conclusion is in section 3.6, and all technical proofs are in section 3.A.

3.2 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,

$$\mathbf{y}_t = \rho \mathbf{W} \mathbf{y}_t + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T. \tag{3.1}$$

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 (3.1) 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 (3.1), 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 (3.1) 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 (3.1), Lam and Souza (2013) proposes to estimate the spatial weight matrix together with other parameters in the model, using

$$\mathbf{y}_t = \mathbf{W}_1 \mathbf{y}_t + \mathbf{W}_2 \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T. \tag{3.2}$$

The term $\rho\mathbf{W}$ in model (3.1) 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 (3.1). Model (3.2) 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 are 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:

$$\mathbf{y}_t = \mathbf{W}^*\mathbf{y}_t + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{3.3}$$

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 $\|\mathbf{W}^*\|_\infty \leq \eta < 1$, where $\|A\|_\infty = \max_i \sum_j |A_{ij}|$ 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 (3.3) 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 3.5.3, 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

$$\mathbf{W}^* = \begin{pmatrix} \mathbf{W}_1^* & & \\ & \ddots & \\ & & \mathbf{W}_G^* \end{pmatrix}, \quad \boldsymbol{\epsilon}_t = \begin{pmatrix} \boldsymbol{\epsilon}_t^{(1)} \\ \vdots \\ \boldsymbol{\epsilon}_t^{(G)} \end{pmatrix}, \tag{3.4}$$

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 $\mathrm{cov}(\boldsymbol{\epsilon}_t^{(i)}, \boldsymbol{\epsilon}_t^{(j)}) = \mathbf{0}$ for $i \neq j$. Otherwise, the block structure in $\mathbf{W}^*$ is not identifiable. Detailed assumptions can be found in section 3.3.1. Relaxation to overlapping blocks is treated in section 3.4. 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 (3.4), 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 (3.4), the matrix $\mathbf{\Pi}^* = (\mathbf{I}_N - \mathbf{W}^*)^{-1}$ also has the same block structure, say

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

with $\mathbf{\Pi}_j^*$ having the same size as $\mathbf{W}_j^*$ in (3.4). Hence $\mathbf{y}_t^{(j)} = \mathbf{\Pi}_j^* \boldsymbol{\epsilon}_t^{(j)}$, and is uncorrelated with $\boldsymbol{\epsilon}_t^{(i)}$ for $1 \leq i \neq j \leq G$ by the assumption that $\mathrm{cov}(\boldsymbol{\epsilon}_t^{(i)}, \boldsymbol{\epsilon}_t^{(j)}) = \mathbf{0}$ 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 (3.3) as a linear regression model,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\xi}^* + \boldsymbol{\epsilon}, \tag{3.5}$$

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 $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}$. 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

$$\widetilde{\boldsymbol{\xi}} = \min_{\boldsymbol{\xi}} \frac{1}{2T}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\xi}\|^2 + \gamma_T \|\boldsymbol{\xi}\|_1, \text{ subj. to } \sum_{j=1}^{N} w_{ij} < 1, \tag{3.6}$$

where $\|\mathbf{v}\|_1 = \sum_i |v_i|$ represents the L_1-norm of the vector $\mathbf{v}$ and $\|\mathbf{v}\| = (\sum_i v_i^2)^{1/2}$ 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 $\|\boldsymbol{\xi}\|_1$ (the absolute sum of all the elements of $\mathbf{W}$). That is, $\widetilde{\boldsymbol{\xi}}$ is the solution to the following problem:

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

where c_T is determined by the tuning parameter γ_T. The row sum constraint in (3.6) and the above ensure the stationarity of the estimated model. The rate for the tuning parameter γ_T will be discussed after Theorem 8 in section 3.3.3.

Theorem 8 in section 3.3 shows that the solution $\widetilde{\boldsymbol{\xi}}$ for the LASSO penalization problem in (3.6) is *zero-block consistent* - that is, the zero off-diagonal blocks in $\mathbf{W}^*$ in (3.4) for model (3.3), with corresponding zero patterns in $\boldsymbol{\xi}^* = \text{vec}(\mathbf{W}^{*\text{T}})$, 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 (3.6) will be able to show such blocks with high probability.

3.3 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.

3.3.1 Main assumptions and notations

(i) The spatial weight matrix $\mathbf{W}^*$ is block diagonal as in (3.4), with at least one $\mathbf{W}_i^* \neq \mathbf{0}$, and $\|\mathbf{W}^*\|_\infty \le \eta < 1$ uniformly as $T, N \to \infty$, where η is a constant. We also assume, uniformly as $T, N \to \infty$,

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

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 (3.4), 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, $\text{var}(\epsilon_{tj}) \le \sigma_\epsilon^2 < \infty$ 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,

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

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 (3.1). 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 6 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 6. See Remark 1 after Theorem 6 for more details on this.

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

$$H = \{j : \xi_j^* = 0 \text{ and corresponds to the zero blocks in } \mathbf{W}^*\}. \tag{3.7}$$

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 7 for the plausible values of c). Finally, define the set

$$A_\epsilon = \{\max_{1\le i,j\le N} |\frac{1}{T}\sum_{t=1}^{T}[\epsilon_{ti}\epsilon_{tj} - E(\epsilon_{ti}\epsilon_{tj})]| < \lambda_T\}. \tag{3.8}$$

For $\mathbf{W}^*$ being block diagonal as in (3.4) and an estimator $\widehat{\mathbf{W}}$, we define the estimator $\widehat{\boldsymbol{\xi}} = \text{vec}(\widehat{\mathbf{W}}^{\mathrm{T}})$ to be *zero-block consistent* for estimating $\mathbf{W}^*$ if

$$P(\widehat{\boldsymbol{\xi}}_{\mathbf{H}} = \mathbf{0}) \to 1, \quad T, N \to \infty. \tag{3.9}$$

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.

3.3.2 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), $|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}|\leq 1$, 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 (3.4). For the elements in the diagonal blocks $\mathbf{W}_i^*, i = 1, \dots, G$ in (3.4), 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 (3.5) in section 3.A (see the proof of Theorem 8 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 (3.5) 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 (3.5) corresponding to the set H defined in (3.7), and the columns of $\mathbf{Z}$ corresponding to the set D defined at the beginning of the proof of Theorem 8. 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 (3.5) to be satisfied, a necessary condition can be derived from (3.5) to be

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

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 (3.5), we can use equations (3.8) and (3.12) in section 3.A to show that on the set A_ϵ defined in (3.8),

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

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 7. Both equations (3.8) and (3.12) are proved on the basis of the form of the model (3.3) and various assumptions in section 3.3.1, 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 (3.3) so that $\mathbf{y}_t = \boldsymbol{\Pi}^*\boldsymbol{\epsilon}_t$, and the assumptions for the spatial weight matrix and the disturbance process, are all reasons for the LASSO estimator in (3.6) to be zero-block consistent.

3.3.3 Main results

We first present a theorem and its corollary concerning the probability lower bound of the set defined in (3.8), 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 8, the main result of this paper, that this is also the probability lower bound for the LASSO solution $\widetilde{\boldsymbol{\xi}}$ in (3.6) being zero-block consistent. Implications and explanations of our main result will be discussed after presenting the theorem.

Theorem 6. *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*

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

The proof of Theorem 6 is relegated to section 3.A. 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 6 becomes

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

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 6.

Corollary 7. *With the same constants C_1,C_2 and C_3, and the same conditions as in Theorem 6, we set the constant c in λ_T such that $c \geq \sqrt{3/C_3}$. Then we have*

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

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 7. By the union sum inequality, putting $v = \lambda_T$ in the result of Theorem 6,

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

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

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

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 7 is the same as the probability lower bound for the LASSO estimator $\widetilde{\boldsymbol{\xi}}$ in (3.6) to be zero-block consistent.

Theorem 8. *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 (3.6) 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 8 is relegated to section 3.A. 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 3.5.3, 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 (3.4) 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 (3.6), 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 3.5 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 3.5 for the introduction of a BIC criterion for choosing γ_T.

3.4 Relaxation for Overlapping Blocks

The spatial weight matrix in (3.4) and the theories presented in section 3.3 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 3.2, 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 8 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

$$H' = \{j : \xi_j^* = 0 \ \text{ and corr. to one of the } G(G-1) \text{ zero blocks in } W^*\}. \tag{3.10}$$

This set corresponds to H in (3.7) 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 3.3.1. 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 3.3.1. 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 8 for overlapping blocks.

Theorem 9. *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 (3.6) 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 8. Zero-block consistency continues to hold even when there are overlapping blocks in the spatial weight matrix.

3.5 Practical Implementation

We use the Least Angle Regression algorithm (LARS) of Bradley Efron and Tibshirani (2004) to implement the minimization in (3.6). A unique solution is guaranteed since the minimization problem in (3.6) 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:

$$\mathrm{BIC}(\gamma_T) = \sum_{i=1}^{N} \log\left(T^{-1}\|\widetilde{\mathbf{y}}_i - (\mathbf{Z}\widetilde{\boldsymbol{\xi}}_{\gamma_T})_i\|^2\right) + |S_{\gamma_T}|\frac{\log(T)}{T}\log(\log(N-1)), \tag{3.11}$$

where $\mathbf{y} = (\widetilde{\mathbf{y}}_1^{\mathrm{T}}, \ldots, \widetilde{\mathbf{y}}_N^{\mathrm{T}})^{\mathrm{T}}$ with $\widetilde{\mathbf{y}}_i = (y_{i1}, \ldots, y_{iT})^{\mathrm{T}}$. The vector $\widetilde{\boldsymbol{\xi}}_{\gamma_T}$ is the LASSO solution to (3.6) 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 (3.5)) 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 (3.11). This γ_{BIC} will then be used in (3.6) to find the LASSO solution $\widetilde{\boldsymbol{\xi}}$.

3.5.1 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 $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|_1/(N(N-1))$ and the L_2 error bound $\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\|/\sqrt{N(N-1)}$ 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 (3.4), 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 3.3, 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 3.1 shows the simulation results with tuning parameter γ_T chosen by minimizing the BIC criteria (3.11) 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 3.1. 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 3.1. 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 3.2 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 3.5.3, 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 (3.11) is minimized. The results are shown in Table 3.2. The simulation results show similar pattern as in Table 3.1: 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

Table 3.1: 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% (3.310)	81.66% (2.814)	80.20% (2.460)	96.99% (3.992)	90.36% (4.645)	84.31% (2.684)
	Within-Block Sensitivity	70.56% (5.832)	79.44% (5.566)	89.17% (4.578)	18.33% (18.829)	52.22% (20.268)	87.78% (7.566)
	Across-Block Specificity	97.01% (2.035)	97.60% (1.857)	97.67% (1.819)	99.42% (1.139)	98.70% (1.738)	98.13% (0.718)
	L_1	0.0237 (0.002)	0.0205 (0.001)	0.0215 (0.003)	0.0136 (0.001)	0.0132 (0.001)	0.0124 (0.000)
	L_2	0.1206 (0.014)	0.0826 (0.006)	0.0769 (0.011)	0.0842 (0.006)	0.0667 (0.005)	0.0511 (0.005)
	Sparsity	85.94% (2.183)	83.94% (2.297)	80.26% (3.151)	97.75% (2.815)	93.85% (3.184)	90.06% (1.447)
	γ_{BIC}	0.3500 (0.051)	0.2401 (0.053)	0.1588 (0.023)	0.4979 (0.158)	0.2687 (0.062)	0.1529 (0.014)
$N = 50$	Within-Block Specificity	77.35% (1.007)	74.57% (1.781)	78.75% (1.250)	89.15% (2.534)	89.38% (1.389)	80.27% (1.239)
	Within-Block Sensitivity	55.71% (2.846)	66.02% (2.374)	75.00% (2.796)	45.80% (7.885)	61.86% (5.029)	87.47% (3.129)
	Across-Block Specificity	98.56% (0.501)	98.94% (0.347)	98.78% (0.361)	99.47% (0.282)	99.42% (0.325)	98.68% (0.408)
	L_1	0.0188 (0.000)	0.0151 (0.000)	0.0139 (0.000)	0.0113 (0.000)	0.0106 (0.000)	0.0112 (0.000)
	L_2	0.1508 (0.007)	0.1031 (0.004)	0.0782 (0.002)	0.1124 (0.005)	0.0937 (0.004)	0.0875 (0.004)
	Sparsity	87.46% (0.620)	87.40% (0.619)	84.48% (0.694)	95.03% (1.090)	93.37% (0.724)	90.35% (0.651)
	γ_{BIC}	0.4807 (0.037)	0.3670 (0.050)	0.1913 (0.016)	0.5048 (0.078)	0.3131 (0.025)	0.1884 (0.014)
$N = 75$	Within-Block Specificity	82.20% (1.281)	81.20% (0.573)	77.47% (0.690)	89.33% (1.192)	87.13% (0.627)	82.46% (0.869)
	Within-Block Sensitivity	40.96% (2.620)	57.24% (2.863)	68.51% (1.274)	40.65% (4.172)	56.74% (3.329)	81.80% (2.437)
	Across-Block Specificity	99.36% (0.324)	99.45% (0.316)	99.67% (0.179)	99.51% (0.168)	99.63% (0.248)	99.09% (0.349)
	L_1	0.0145 (0.000)	0.0129 (0.000)	0.0116 (0.000)	0.0102 (0.000)	0.0087 (0.000)	0.0091 (0.000)
	L_2	0.1467 (0.007)	0.1123 (0.005)	0.0867 (0.003)	0.1352 (0.005)	0.0974 (0.004)	0.0919 (0.004)
	Sparsity	90.75% (0.606)	88.35% (0.352)	86.36% (0.305)	94.71% (0.552)	93.59% (0.399)	90.96% (0.431)
	γ_{BIC}	0.5591 (0.070)	0.4145 (0.033)	0.2978 (0.027)	0.5690 (0.072)	0.3479 (0.033)	0.2091 (0.016)

Notes: Standard errors in parenthesis.

as T increases, but the within-block sensitivity increases, like those in Table 3.1.

3.5.2 Simulation results for nonstationary models

In order to see how the stationarity of model (3.3) 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 3.3. We also added results for a

Table 3.2: 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 3.1.

nonstationary model in Table 3.4. They are substantially worse than those in Subsection 3.5.1, which are associated with stationary models.

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 3.1, 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 3.3.

3.5.3 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 3.3 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[2]. The composition of the Senate and the number of voting instances can be found in Table 3.5.

Estimation is conducted in absolute disregard of party affiliation, and the tuning parameter γ_T is chosen such that minimizes BIC criterion in (3.11). The outcome for year 2012, which involves

[2] 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 3.3: 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 3.1.

$T = 251$ voting instances and $N = 98$ senators, is displayed in Figure 3.1. 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 3.1 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 9 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.

Table 3.4: 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 3.1.

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 3.2, 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.

Figure 3.1: Visualization of the estimated spatial weight matrix for voting, 2012.

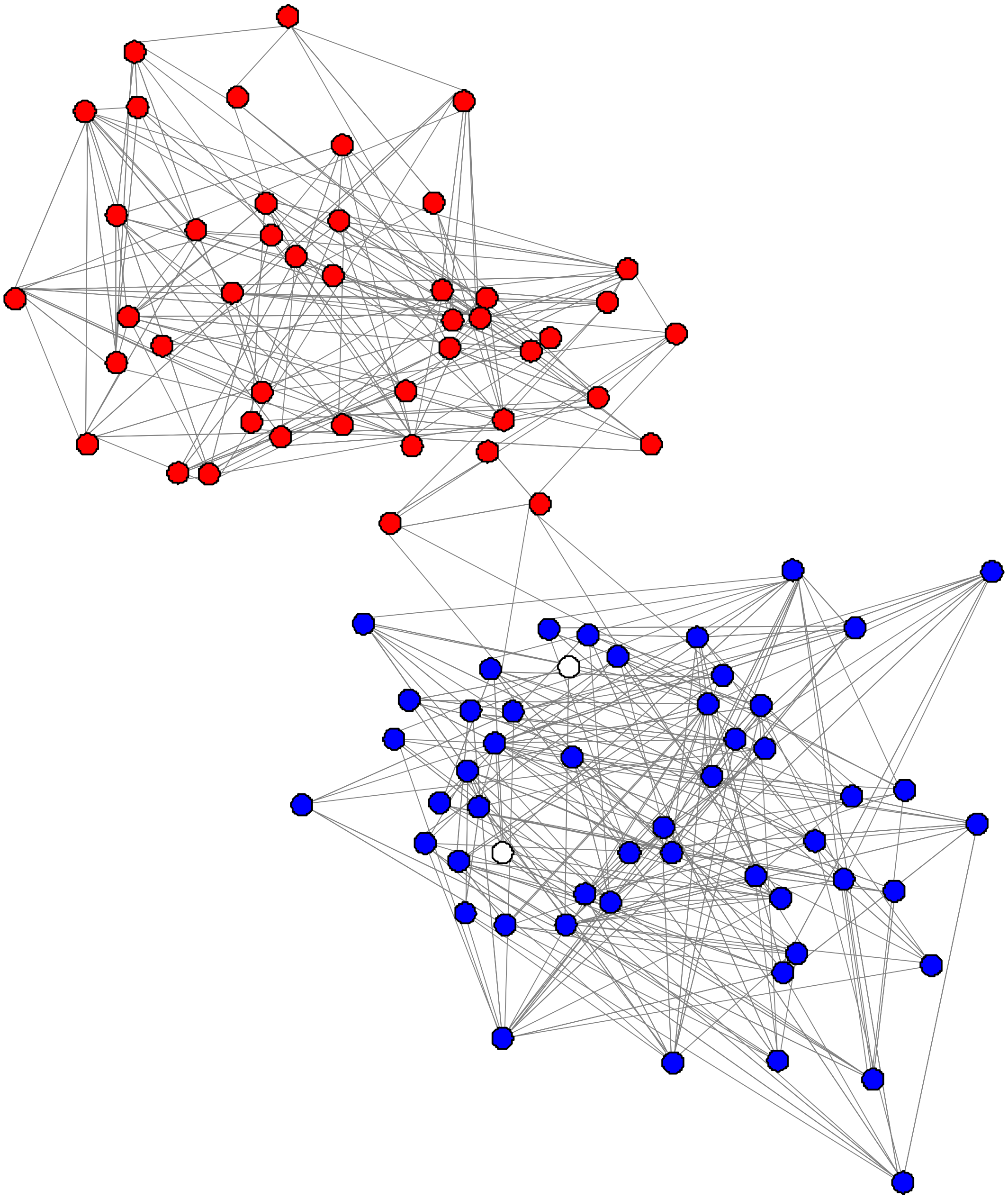


Figure 3.2: Cross-party collaboration.

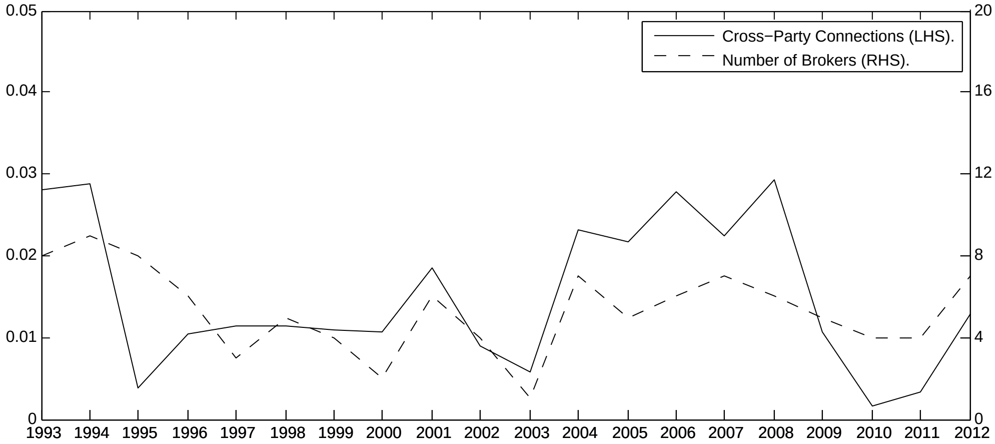


Table 3.5: Senate Composition.

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

3.6 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 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.

Appendix

3.A Proofs

Proof of Theorem 6. 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

$$\max_{1\leq j\leq N} \|\epsilon_{tj}\|_{2w} \leq \mu, \tag{3.1}$$

$$\sum_{t=m}^{\infty} \max_{1\leq j\leq N} \|\epsilon_{tj} - \epsilon'_{tj}\|_{2w} \leq Cm^{-\alpha}, \tag{3.2}$$

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 (3.1) and (3.2), 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 (3.1), by the Fubini's Theorem and assumption (v),

$$\begin{aligned} E|\epsilon_{tj}|^{2w} &= E\int_0^{|\epsilon_{tj}|^{2w}} ds = \int_0^{\infty} P(|\epsilon_{tj}|> s^{1/2w})\, ds \leq \int_0^{\infty} D_1 \exp(-D_2 s^{q/2w})\, ds \\ &= \frac{4wD_1}{q}\int_0^{\infty} x^{4w/q-1} e^{-D_2x^2}\, dx = \frac{2wD_1}{qD_2^{2w/q}}\Gamma(2w/q) = \mu^{2w} < \infty, \end{aligned} \tag{3.3}$$

so that $\max_{1\leq j\leq N} \|\epsilon_{tj}\|_{2w} \leq \mu < \infty$ for any $w > 0$. This proves (3.1).

To prove (3.2), denote $\boldsymbol{\phi}_{ij}^{\mathrm{T}}$ the j-th row of $\boldsymbol{\Phi}_i$. Then using the causal definition in assumption (iv),

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

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 (3.3),

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

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

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

which is (3.2) since μ, C are constants. This completes the proof of the theorem. □

Proof of Theorem 8. 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 (3.6) 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 (3.6) 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

$$\begin{aligned}&\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D(\widetilde{\boldsymbol{\xi}}_D - \boldsymbol{\xi}_D^*) - \frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon} = -\gamma_T\mathbf{g}_H,\\ &\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D(\widetilde{\boldsymbol{\xi}}_D - \boldsymbol{\xi}_D^*) - \frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon} = -\gamma_T\mathbf{g}_D,\end{aligned} \tag{3.4}$$

which can be simplified to

$$|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D)^{-1}(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon} - \gamma_T\mathbf{g}_D) - \frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon}| \leq \gamma_T, \tag{3.5}$$

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

We now show that, on the set A_ϵ as defined in (3.8), (3.5) 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 (3.4). Then, consider on the set A_ϵ,

$$\begin{aligned}\|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\boldsymbol{\epsilon}\|_{\max} = \max_{i\in I_q, j\notin I_q}\left|\frac{1}{T}\sum_{t=1}^T y_{ti}\epsilon_{tj}\right| &= \max_{i\in I_q, j\notin I_q}\left|\sum_{s\in I_q}\pi_{is}^*\Big(\frac{1}{T}\sum_{t=1}^T \epsilon_{ts}\epsilon_{tj}\Big)\right|\\ &\leq \lambda_T \max_{1\leq i\leq N}\sum_{s=1}^N|\pi_{is}^*|\leq \frac{\lambda_T}{1-\eta},\end{aligned} \tag{3.6}$$

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 (3.3) 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 $\mathrm{cov}(\epsilon_{ti}, \epsilon_{tj}) = 0$ 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

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

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

$$\|\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\boldsymbol{\epsilon}\|_{\max} = \max_{i\in I_q, j\in I_{q'}}\left|\sum_{s\in I_q}\pi_{is}^*\Big(\frac{1}{T}\sum_{t=1}^T \epsilon_{ts}\epsilon_{tj}\Big)\right| \leq \frac{\sigma_\epsilon^2 + \lambda_T}{1-\eta}, \tag{3.7}$$

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

$$\|\frac{1}{T}\mathbf{Z}_H^{\mathrm{T}}\mathbf{Z}_D\|_\infty \leq n \max_{i \in I_q, j \notin I_q} \left| \frac{1}{T}\sum_{t=1}^T y_{ti}y_{tj} \right| = n \max_{i \in l_q, j \in l_{q'}} \left| \sum_{s \in I_q, \ell \in I_{q'}} \pi_{is}^* \pi_{j\ell}^* (\frac{1}{T}\sum_{t=1}^T \epsilon_{ts}\epsilon_{t\ell}) \right| \leq \frac{\lambda_T n}{(1-\eta)^2}. \tag{3.8}$$

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

$$\begin{aligned} \|(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D)^{-1}\|_\infty &\leq n^{1/2}\lambda_{\min}^{-1}(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D) \leq n^{1/2}\lambda_{\min}^{-1}(\frac{1}{T}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}) = n^{1/2}\lambda_{\min}^{-1}(\frac{1}{T}\sum_{t=1}^T \mathbf{y}_t\mathbf{y}_t^{\mathrm{T}}) \\ &= n^{1/2}\lambda_{\min}^{-1}\Big(\mathbf{\Pi}^*(\frac{1}{T}\sum_{t=1}^T \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}})\mathbf{\Pi}^{*\mathrm{T}}\Big) \leq n^{1/2}\sigma_{\min}^{-2}(\mathbf{\Pi}^*)\lambda_{\min}^{-1}(\frac{1}{T}\sum_{t=1}^T \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}}). \end{aligned} \tag{3.9}$$

To bound (3.9), we have

$$\sigma_{\min}^{-2}(\mathbf{\Pi}^*) = \sigma_{\max}^2(\mathbf{I}_N - \mathbf{W}^*) \leq (1 + \sigma_{\max}(\mathbf{W}^*))^2 \leq (1 + \|\mathbf{W}^*\|_1^{1/2}\|\mathbf{W}^*\|_\infty^{1/2})^2 \leq (1 + \eta^{1/2}\eta_c^{1/2})^2, \tag{3.10}$$

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 $i \geq 0$, we have

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

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

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

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

$$\begin{aligned}\lambda_{\min}(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}}) &= \sigma_{\min}^2(T^{-1/2}\sum_{i\geq 0}\boldsymbol{\Phi}_i(\boldsymbol{\eta}_{1-i},\ldots,\boldsymbol{\eta}_{T-i}))\\ &\geq \left\{\sigma_{\min}(T^{-1/2}(\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_T)) - \sum_{i\geq 1}\sigma_{\max}(\boldsymbol{\Phi}_i T^{-1/2}(\boldsymbol{\eta}_{1-i},\ldots,\boldsymbol{\eta}_{T-i}))\right\}^2\\ &\geq \left\{\lambda_{\min}^{1/2}(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\eta}_t\boldsymbol{\eta}_t^{\mathrm{T}}) - \sum_{i\geq 1}\|\boldsymbol{\Phi}_i\|\lambda_{\max}^{1/2}(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\eta}_{t-i}\boldsymbol{\eta}_{t-i}^{\mathrm{T}})\right\}^2\\ &\geq \left\{\sigma(1-\sqrt{d}) - e - (\sigma(1+\sqrt{d})+e)\sum_{i\geq 1}\|\boldsymbol{\Phi}_i\|\right\}^2 \geq c^2, \qquad (3.11)\end{aligned}$$

where $c > 0$ is a constant as in assumption (iv). Combining (3.10) and (3.11), on U and for large enough T, N, (3.9) becomes

$$\|(\frac{1}{T}\mathbf{Z}_D^{\mathrm{T}}\mathbf{Z}_D)^{-1}\|_\infty \leq \frac{n^{1/2}(1+\eta^{1/2}\eta_c^{1/2})^2}{c^2}. \qquad (3.12)$$

Hence combining the bounds (3.6), (3.7), (3.8) and (3.12), on $A_\epsilon \cap U$, for large enough T, N, we have

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

by the assumption $n = o(\{\gamma_T/\lambda_T\}^{2/3})$. Hence on $A_\epsilon \cap U$, (3.5) 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 7. This completes the proof of the first half of Theorem 8.

For the second half, suppose $\widetilde{\boldsymbol{\xi}}_D = \mathbf{0}$. Then using (3.4), we have

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

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

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

where $\boldsymbol{\pi}_j^{\mathrm{T}}$ is the j-th row of $\mathbf{\Pi}^*$, with $\|\boldsymbol{\pi}_j^*\| > 1$, and we used (3.11). 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 9. 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 8 by replacing D with D' and H with H'. The only differences are the bounds in (3.6) and (3.8). Consider, on A_ϵ,

$$\begin{aligned}
\|\frac{1}{T}\mathbf{Z}_{H'}^{\mathrm{T}}\boldsymbol{\epsilon}\|_{\max} &= \max_{i\in I_q, j\notin I_q}\left|\frac{1}{T}\sum_{t=1}^{T} y_{ti}\epsilon_{tj}\right| = \max_{i\in I_q, j\notin I_q}\left|\sum_{s\in I_q}\pi_{is}^*(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}) + \sum_{s\notin I_q}\pi_{is}^*(\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj})\right| \\
&\leq \max_{s\in I_q, j\notin I_q}|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}|\|\mathbf{\Pi}^*\|_\infty + \max_{s\notin I_q, j\notin I_q}|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{ts}\epsilon_{tj}|\max_{i\in I_q}\sum_{s\notin I_q}|\pi_{is}^*| \\
&\leq \frac{\lambda_T + c_\epsilon\lambda_T}{1-\eta} + (\sigma_\epsilon^2 + \lambda_T)c_\pi\lambda_T = O(\lambda_T),
\end{aligned} \tag{3.13}$$

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,

$$\begin{aligned}
\|\frac{1}{T}\mathbf{Z}_{H'}^{\mathrm{T}}\mathbf{Z}_{D'}\|_\infty &\leq n\max_{i\in I_q, j\notin I_q}\left|\sum_{s\in I_q}\pi_{js}^*(\frac{1}{T}\sum_{t=1}^{T}y_{ti}\epsilon_{ts}) + \sum_{s\notin I_q}\pi_{js}^*(\frac{1}{T}\sum_{t=1}^{T}y_{ti}\epsilon_{ts})\right| \\
&\leq n\left(\frac{\sigma_\epsilon^2 + \lambda_T}{1-\eta}\right)c_\pi\lambda_T + n\lambda_T\left(\frac{1+c_\epsilon}{1-\eta} + c_\pi(\sigma_\epsilon^2 + \lambda_T)\right)\frac{1}{1-\eta} = O(\lambda_T n),
\end{aligned} \tag{3.14}$$

where we used (3.13) in the last line. The rates in (3.13) and (3.14) are the same as (3.6) and (3.8) respectively, and hence the results in Theorem 8 follows. □

Bibliography

D. Acemoglu, V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016, September 2012.

A. Ahrens and A. Bhattacharjee. Two-step lasso estimation of the spatial weights matrix. 2014. Manuscript.

A. Ambrus, A. G. Chandrasekar, and M. Elliott. Social investments, informal risk sharing and inequality. Unpublished Memo., September 2014.

A. Ammermuller and J.-S. Pischke. Peer effects in european primary schools: Evidence from pirls. *Journal of Labor Economics*, 27(3):315–348, 2009.

J. Anderson and E. van Wincoop. Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192, March 2003.

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

J. D. Angrist and K. Lang. Does school integration generate peer effects? evidence from boston's metco program. *American Economic Review*, 94(5):1613–1634, December 2004.

L. Anselin. *Spatial Econometrics: Methods and Models.* Springer, August 1988.

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

L. Anselin. Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1):3–25, 2010.

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

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

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

B. Ballobás. *Modern Graph Theory.* Springer, corrected 2nd edition edition, 2013.

O. Bandiera and I. Rasul. Social networks and technology adoption in northern mozambique. *The Economic Journal*, 116:869–902, October 2006.

O. Bandiera, R. Burgess, N. Das, S. Gulesci, I. Rasul, and M. Sulaiman. Can basic entrepreneurship transform the economic lives of the poor. *STICERD Working Paper Series*, (EOPP 43), April 2013.

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

M. Beenstock and D. Felsenstein. Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geographical Analysis*, 44(4):386–397, 2012. ISSN 1538-4632. doi: 10.1111/j.1538-4632.2012.00851.x. URL `http://dx.doi.org/10.1111/j.1538-4632.2012.00851.x`.

A. Bhattacharjee and C. Jensen-Butler. Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617 – 634, 2013. ISSN 0166-0462. doi: http://dx.doi.org/10.1016/j.regsciurbeco.2013.03.005. URL `http://www.sciencedirect.com/science/article/pii/S0166046213000288`.

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

L. Blume, W. A. Brock, S. N. Durlauf, and Y. Ioannides. Identification of social interactions. In J. Behabib, A. Bisin, and M. O. Jackson, editors, *Handbook of Social Economics*, volume 1B. Noth-Holland, 2011.

I. J. Bradley Efron, Trevor Hastie and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–451, April 2004.

Y. Bramoullé, H. Djebbari, and B. Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009.

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

F. Bugni. Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, 78(2):735–753, March 2010.

A. Calvó-Armengol and M. O. Jackson. The effects of social networks on employment and inequality. *American Economic Review*, 94(3):426–454, June 2004.

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

V. Chernozhukov, H. Hong, and E. Tamer. Estimation and confidence regions for parameter sets in econometric models. *Econometrican*, 75(5):1243–1284, September 2007.

T. G. Conley and G. Topa. Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics*, 17:303–327, 2002.

T. G. Conley and C. R. Udry. Learning about a new technology: Pineapple in ghana. *American Economic Review*, 100(1):35–69, 2010.

L. Corrado and B. Fingleton. Where is the economics in spatial econometrics? Working Papers 1101, University of Strathclyde Business School, Department of Economics, Jan. 2011. URL `http://ideas.repec.org/p/str/wpaper/1101.html`.

B. Crépon, E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora. Do labor market policies have displacement effects? evidence from a clustered randomized experiment. Technical Report 18597, NBER Working Paper, 2012.

G. De Giorgi, M. Pellizzari, and S. Redaelli. Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics*, 2(2):241–275, April 2010.

M. Dell. Trafficking networks and the mexican drug war. Unpublished Memo., December 2012.

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

R. Diestel. *Graph Theory*. Graduate Texts in Mathematics. Springer, 2010.

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

J. Elhorst. Spatial panel data models. In M. M. Fischer and A. Getis, editors, *Handbook of Applied Spatial Analysis*, pages 377–407. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-03646-0. doi: 10.1007/978-3-642-03647-7_19. URL `http://dx.doi.org/10.1007/978-3-642-03647-7_19`.

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

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

M. M. Fischer and J. Wang. *Spatial Data Analysis: Models, Methods and Techniques (SpringerBriefs in Regional Science)*. Springer, 1st edition. edition, Sept. 2011. ISBN 3642217192. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/3642217192`.

A. D. Foster and M. R. Rosenzweig. Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, 103(6):1176–1209, December 1995.

J. Fowler. Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 71(1): 456–487, 2006. doi: 10.1093/pan/mpl002.

O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81 (395):832–842, 1986.

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

E. Glaeser, B. Sacerdote, and J. A. Scheinkman. Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548, May 1996.

E. Glaeser, D. Laibson, and B. Sacerdote. An economic approach to social capital. *The Economic Journal*, 112(483):F437–F458, November 2002.

P. Goldsmith-Pinkham and G. W. Imbens. Social Networks and the Identification of Peer Effects. *Journal of Business & Economic Statistics*, 31(3):253–264, July 2013.

C. Gouriéroux and A. Monfort. *Simulation-Based Econometric Methods.* Core Lectures. Oxford University Press, 1997.

B. S. Graham. Identifying social interactions through conditional variance restrictions. *Econometrica*, 76(3):645–660, May 2008.

M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, May 1973.

E. Hanushek. Teacher characteristics and gains in student achievement: estimation using micro data. *American Economic Review*, 61(2), May 1971.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, March 1981.

C.-S. Hsieh and L.-f. Lee. A Social Interactions Model with Endogenous Friendship Formation and Selectivity. *Journal of Applied Econometrics*, pages n/a–n/a, Jan. 2015.

T. Hsing and W. B. Wu. On weighted u-statistics for stationary processes. *The Annals of Probability*, 32(2):1600–1631, 2004.

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

M. O. Jackson. *Social and Economic Networks*. Princeton University Press, November 2010.

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

H. H. Kelejian and I. R. Prucha. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, 17(1):99–121, 1998.

H. H. Kelejian and I. R. Prucha. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40:509–533, 1999.

H. H. Kelejian and I. R. Prucha. On the asymptotic distribution of the moran i test statistic with applications. *Journal of Econometrics*, 104:219–57, 2001.

H. H. Kelejian and I. R. Prucha. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157:53–67, 2010.

C. Lam and P. C. Souza. Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews*, August 2014.

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

L.-F. Lee. Asumptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):25, November 2004.

L.-F. Lee. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics*, 140(2):333–374, 2007.

L.-F. Lee and J. Yu. Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, 154:165–185, 2010.

L.-F. Lee, X. Liu, and X. Lin. Specification and estimation of social interaction models with network structures. *Econometrics Journal*, 13:145–176, 2010.

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

J. LeSage and R. K. Pace. *Introduction to Spatial Econometrics*. CRC Press, New York, 2009.

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

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

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

C. F. Manski. Identification of endogenous social effects: the reflection problem. *The Review of Economic Studies*, 60(3):531–542, July 1993.

C. F. Manski. Economic analysis of social interactions. *The Journal of Economic Perspectives*, 14 (3):115–136, Summer 2000.

E. Miguel and M. Kremer. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217, January 2004.

J. D. Montgomery. Social networks and labor-market outcomes: toward an economic analysis. *American Economic Review*, 81(5):1408–1418, December 1991.

K. Munshi. Networks in the modern economy: Mexican migrants in the u.s. labor market. *Quarterly Journal of Economics*, 118(2):549–599, May 2003.

W. Newey and D. McFadden. Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

M. Pellizzari. Do friends and relatives really help in getting a good job? Technical Report 623, Centre for Economic Performance Working Paper, March 2004.

J. Pinkse and M. E. Slade. The future of spatial econometrics. *Journal of Regional Science*, 50(1): 103–117, 2010. URL `http://ideas.repec.org/a/bla/jregsc/v50y2010i1p103-117.html`.

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

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

A. Rees. Information networks in labor markets. *American Economic Review*, 56(1/2):559–566, March 1966.

S. G. Rivkin, E. Hanushek, and J. F. Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, March 2005.

J. R. Romano and A. M. Shaikh. Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211, January 2010.

T. J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

B. Sacerdote. Peer effects with random assingment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704, May 2001.

X. Shao. Nonstationarity-extended whittle estimation. *Econometric Theory*, 26:1060–1087, 2010. ISSN 1469-4360. doi: 10.1017/S0266466609990466. URL http://dx.doi.org/10.1017/S0266466609990466.

T. A. B. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37:131–53, 2011.

D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00693.x. URL http://dx.doi.org/10.1111/j.1467-9868.2008.00693.x.

S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

S. Wasserman and P. Pattison. Logit models and logistic regression for social networks: An introduction to markov graphs and p*. *Psychometrika*, 61(3):401–425, September 1996.

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

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

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

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

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

S. Zhou. Thresholded lasso for high-dimensional variable selection and statistical estimation. Technical Report 1002.1583v2, arXiv.org Collection, 2010.

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

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006. URL http://ideas.repec.org/a/bes/jnlasa/v101y2006p1418-1429.html.