Econometrics **2015**, *3*, 128-155; doi:10.3390/econometrics3010128

OPEN ACCESS

econometrics

ISSN 2225-1146

www.mdpi.com/journal/econometrics

Article

Two-Step Lasso Estimation of the Spatial Weights Matrix

Achim Ahrens * and Arnab Bhattacharjee

Spatial Economics and Econometrics Centre (SEEC), Heriot-Watt University, Edinburgh,
Scotland EH14 4AS, UK; E-Mail: a.bhattacharjee@hw.ac.uk

* Author to whom correspondence should be addressed; E-Mail: aa1266@hw.ac.uk;
Tel.: +44-0-131-451-3482.

Academic Editor: Kerry Patterson

Received: 31 October 2014 / Accepted: 11 February 2015 / Published: 9 March 2015

Abstract: The vast majority of spatial econometric research relies on the assumption that the spatial network structure is known *a priori*. This study considers a two-step estimation strategy for estimating the $n(n-1)$ interaction effects in a spatial autoregressive panel model where the spatial dimension is potentially large. The identifying assumption is approximate sparsity of the spatial weights matrix. The proposed estimation methodology exploits the Lasso estimator and mimics two-stage least squares (2SLS) to account for endogeneity of the spatial lag. The developed two-step estimator is of more general interest. It may be used in applications where the number of endogenous regressors and the number of instrumental variables is larger than the number of observations. We derive convergence rates for the two-step Lasso estimator. Our Monte Carlo simulation results show that the two-step estimator is consistent and successfully recovers the spatial network structure for reasonable sample size, T.

Keywords: lasso; endogeneity; unkown W; spatial weights matrix

JEL classifications: C23; C33; C52

1. Introduction

This study proposes an estimator, based on the Lasso estimator, for an approximately sparse spatial weights matrix in a high-dimensional setting. The vast majority of spatial econometric research relies on the assumption that the spatial weights matrix, $\mathbf{W}_n$, which measures the strength of interactions between units, is known *a priori*. In applied work, researchers often need to select between standard specifications such as the binary contiguity matrix, inverse distance matrix or other matrices based on some observable notion of distance. The choice of spatial weights has been a focus of criticism of spatial econometric methods, since estimation results highly depend on the researcher's specification of the spatial weights matrix [1–3]. Furthermore, a pre-defined weights matrix does not provide insights into the drivers of socio-economic interactions and general equilibrium effects in a network, but only allows for measuring the general strength of interactions, which is reflected in the size of the spatial autoregressive coefficient.

The shortcomings of employing pre-specified spatial weights are well known. Pinkse *et al.* [4] is one of the first attempts to conduct inferences in a setting where the spatial weights matrix is not known *a priori*. The authors propose a semi-parametric estimator which relies on observable distance measures. Bhattacharjee and Jensen-Butler [5] consider estimation of the spatial weights matrix from the spatial autocovariance matrix in spatial panel models, and show that $\mathbf{W}_n$ is only partially identified. Intuitively, the main issue is that, in contrast to autocovariances, spatial weights relate to the direction and strength of causation between spatial units. Since there are twice as many spatial weights as there are autocovariances, further assumptions are required for identification. Bhattacharjee and Jensen-Butler [5] propose an estimator that provides exact identification under the assumption that the spatial weights matrix is symmetric and n is fixed.[1] Estimation of the spatial weights matrix in a low-dimensional small n panel, under different structural assumptions on the autocovariances or using moment conditions is discussed in [7,8].

The aforementioned literature focuses on a low-dimensional context where typically $n \ll T$. In contrast, Bailey *et al.* [9] consider sparsity of the spatial weights matrix as an alternative identification assumption based on a large T panel setting and the spatial error model. They apply a multiple testing procedure to the matrix of spatial autocorrelation coefficients in order to identify the non-zero interactions, and place weights of $+1$, -1 or zero, depending on whether the autocorrelations are significantly positive, significantly negative or insignificant, respectively. There are also a few previous studies which apply Lasso-type estimators to high-dimensional spatial panel models and assume sparsity.[2] Manresa [13] considers a non-autoregressive panel model with spatially lagged exogenous regressors. Hence, the model does not suffer from simultaneity and the Lasso estimator can be used for dimensionality reduction. Souza [14] and Lam and Souza [15] consider a spatial autoregressive model with additional spatial lags on exogenous

[1] See [6] for a similar approach.

[2] The Lasso has also been applied in the GIS literature, where the focus is on estimation of a spatial model where spatial dependence is a function of geographic distance; see, for example, Huang *et al.* [10] and Wheeler [11]. Likewise, Seya *et al.* [12] assume a known spatial weights matrix and apply the Lasso for spatial filtering. Spatial filtering is different from our approach as filtering treats the spatial weights as nuisance parameters whereas we focus on the recovery of the spatial dependence structure.

regressors. Souza [14] discusses several exclusion restrictions that allow for identification, but require prior knowledge about the network structure. Lam and Souza [15] propose a Lasso-type estimator for spatial weights under the assumption that the error variance decays to zero as T increases, which may be a strong assumption in some applications. By contrast, the method proposed here does not require prior knowledge about the network structure and does not rely on variance decay, but instead exploits exogenous regressors as instruments.

This study explores the estimation of the spatial weights matrix in a panel data setting where T, the number of time periods, is large. The spatial autoregressive or spatial lag model is given by

$$y_{it} = \sum_{j=1}^{n} w_{ij} y_{jt} + \mathbf{x}_{it}' \boldsymbol{\beta}_i + e_{it}, \quad t = 1, \ldots, T; i = 1, \ldots, n \tag{1.1}$$

where y_{it} is the response variable, $\mathbf{x}_{it} = (x_{1,it}, x_{2,it}, \ldots, x_{K,it})'$ is the vector of exogenous regressors and $\boldsymbol{\beta}_i$ is the $K \times 1$ parameter vector with $K \geq 1$. The error term is assumed to be independently distributed, but allowed to be heteroskedastic and non-Gaussian. w_{ij} is the (i, j)th element of the $n \times n$ spatial weights matrix, denoted by $\mathbf{W}_n$, and measures the strength of spill-over effects from unit j to unit i. The spatial weights matrix has zeros on the diagonal, *i*.e., $w_{ii} = 0$ for all i.[3] The first term on the right-hand side is often referred to as the spatial lag, analogous to a temporal lag in time-series models. The spatial autoregressive panel model is a natural extension to cross-sectional spatial autoregressive models as introduced by [16,17]. Spatio-temporal panel models, such as the spatial autoregressive model in (1.1), have recently attracted much attention; see, for example, [18–22].[4]

Estimation of the above model poses two major challenges when $\mathbf{W}_n$ is treated as unknown. First, the model suffers from reverse causality as the response variable appears both on the left and right-hand side of the equation. It is well known that ordinary least squares (OLS) is inconsistent in the presence of endogeneity. Second, the model is not identified unless the number of parameters, $p := n(n-1) + Kn$, is smaller than the number of observations, nT, or further assumptions are made. The identification assumption considered here is sparsity of the weights matrix which requires that each unit is affected by only a limited number of other units. Specifically, the number of units affecting a specific unit i is assumed to be much smaller than T, but we explicitly allow for $p \gg nT$.

The proposed estimation method is a two-step procedure based on the Lasso estimator introduced by Tibshirani [24]. The Lasso is a regularization technique which can, under the sparsity assumption, deal with high-dimensional settings where the number of exogenous regressors is large relative to the number of observations. The ℓ_1-penalization employed by the Lasso sets some of the coefficient estimates to exactly zero, making the Lasso estimator attractive for model selection. The ℓ_1-penalization behaves similarly to the ℓ_0-penalty, as used in the Akaike information criterion and Bayesian information criterion [25,26], but is computationally more attractive due to its convex form. The Lasso is a popular

[3] We implicitly set the spatial autoregressive parameter, which is commonly employed in spatial models, equal to one, since w_{ij} and the spatial autoregressive parameter are not separately identified [5].

[4] See [23] for an overview.

and well-established technique, but its theoretical properties have only recently been better understood. Recent theoretical contributions include [27–35].

Conceptually, identification of a spatial weights matrix requires suitably dealing with the endogeneity inherent in model (1.1). Lam and Souza [15] address this issue by assuming that the error variance asymptotically decays to zero. By contrast, we address endogeneity using instruments. The estimation methodology proceeds in two steps. In the first step, relevant instruments are identified by the Lasso and predictions for $y_{1t}, \ldots, y_{nt}$ are obtained. In the second step, the regression model in (1.1) is estimated, but the spatial lag on the right-hand side is replaced with predictions from the first step. That is, the second-step Lasso selects the neighbours affecting y_{it}. The procedure is conceptually based on two-stage least squares (2SLS), but employs the Lasso for selecting relevant instruments in the first step and for selecting relevant spatial lags in the second step. Figure 1 visualizes the spatial autoregressive model in (1.1) for $n = 2$ and motivates the choice of instruments exploited to identify the spatial weights. In the regression equation with y_{1t} as the dependent variable, we can exploit $\mathbf{x}_{2t}$ as instruments for y_{2t} and vice versa.

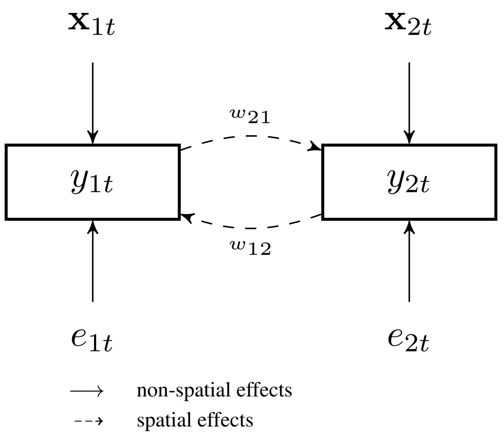

Figure 1. The Spatial Autoregressive Model for $n = 2$.

We also consider the post-Lasso OLS estimator due to Belloni and Chernozhukov [31], which applies ordinary least squares (OLS) to the model selected by the Lasso and aims at reducing the Lasso shrinkage bias. Although the estimation methodology relies on large T asymptotics, Monte Carlo results suggest that the two-step Lasso estimator is able to recover the spatial network structure if T is as small as 50–100. The estimator may be combined with established large T panel estimators such as the Common Correlated Effects estimator [36,37], which controls for strong cross-sectional dependence, and can be extended to dynamic models including temporal lags of the dependent variable as regressors.

Finally, this study is also related to the emerging literature on high-dimensional methods which allow the number of endogenous regressors to be larger than the sample size. The Self-Tuning Instrumental Variable (STIV) due to Gautier and Tsybakov [38] is a generalization of the Dantzig estimator [39] allowing for many endogenous regressors. The focused generalized methods of moments (FGMM) developed in Fan and Liao [40] extends shrinkage GMM estimators as in, e.g., [41] to high-dimensional settings. The two-step Lasso estimator considered in this study is conceptually similar to Lin *et al.* [42] who apply two-step penalized least squares to genetic data. We improve upon Lin *et al.* [42] in that our approach allows for approximate sparsity, non-Gaussian errors and uses the sharper penalty level proposed by Belloni *et al.* [30]. However, our main contribution is to point out that a simple two-step Lasso estimation method can be employed to estimate the spatial weights matrix. The approach does not require any prior knowledge about the network structure, except for the sparsity assumption and a set of exogenous regressors.

The article is organized as follows. In Section 2, we consider a general setting where the number of endogenous regressors and the number of instruments is allowed to be larger than the number of observations. The two-step estimator may be of more general interest for applications with endogeneity in high-dimensions. Section 3 applies the proposed two-step estimator to estimate the spatial autoregressive model in (1.1). In Section 4, we present Monte Carlo results to demonstrate the performance of the two-step Lasso for estimating the spatial weights matrix. Finally, Section 5 concludes.

Notation. The ℓ_q-norm of the vector $\mathbf{a} \in \mathbb{R}^M$ is defined as $\|\mathbf{a}\|_q = (\sum_{m=1}^{M} |a_m|^q)^{1/q}$, $q = 1, 2$. The number of non-zero elements in $\mathbf{a}$ is denoted by $\|\mathbf{a}\|_0$, and $\|\mathbf{a}\|_\infty$ is the largest element in $\mathbf{a}$. We use $((\cdot))$ to denote the typical element of a matrix, e.g., $\mathbf{A} = ((a_{ij}))$. The Frobenius norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F = (\sum_{i,j} |a_{ij}|^2)^{1/2}$. Let $\|\mathbf{A}\|_1$ be the entry-wise ℓ_1 norm, *i.e.*, $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$. The support operator is $\text{supp}(\mathbf{a}) = \{m \in \{1, \ldots, M\} : a_m \neq 0\}$. Let V be a set, then $\bar{V}$ is the complement of V. $\mathbf{a}_V$ is a vector with elements $a_m \mathbf{1}\{m \in V\}$ for $m = 1, \ldots, M$ where $\mathbf{1}$ is the indicator function. The typical element of $\mathbf{A}_V$ is $a_{ij}\mathbf{1}\{j \in V\}$. We use $x_T \lesssim_{\mathrm{P}} z_T$ to denote $x_T = O_P(z_T)$ and $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$.

2. Two-Step Lasso Estimator

In this section, we develop a two-step estimation procedure that allows the number of possibly endogenous regressors as well as the number of instruments to be larger than the sample size. The identifying assumption is approximate sparsity. Section 3 presents the spatial autoregressive model as an application to this setting. The two-step estimator may be of interest in, for example, cross-country growth regressions where the number of regressors is large relative to the number of countries and endogeneity is a potential issue. Furthermore, endogeneity in high dimensions may arise when the aim is to find a sparse linear approximation to a complex non-parametric data-generating process; see earning regressions in [43].

The structural equation and first-step equations are given by

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}^\star + e_t, \tag{2.1}$$

$$x_{tj} = \mathbf{z}_t' \boldsymbol{\pi}_j^\star + u_{tj}, \qquad j = 1, \ldots, p. \tag{2.2}$$

y_t is the outcome variable and $\mathbf{x}_t$ is a p-dimensional vector of regressors. For notational consistency with Section 3, we use $t = 1, \ldots, T$ to denote distinct units or repeated observations over time. Without loss of generality, we assume that the first $\bar{p}$ regressors are endogenous, *i.e.*, $\mathrm{E}[e_t|x_{tj}] \neq 0$ for $j = 1, \ldots, \bar{p}$ with $\bar{p} \in \{1, \ldots, p\}$. The remaining $p - \bar{p}$ regressors are exogenous. Hence, we allow the set of exogenous regressors to be empty. We assume the existence of $L \geq p$ instruments, $\mathbf{z}_t$, which satisfy the exclusion restriction $\mathrm{E}[e_t|\mathbf{z}_t] = \mathrm{E}[u_{tj}|\mathbf{z}_t] = 0$ for $j = 1, \ldots, p$. If a regressors x_{tj} is exogenous, it serves as an instrument for itself. Hence, $z_{tj} = x_{tj}$ for $j > \bar{p}$. The error terms e_t and u_{tj} are independently distributed, but possibly heteroskedastic and non-Gaussian. The interest lies in obtaining a sparse approximation of $\boldsymbol{\beta}^\star$. While the model in (2.1)–(2.2) assumes that the conditional expectation functions are linear, the framework may be easily generalized to a non-parametric data-generating process as in, for example, Bickel *et al.* [27].

2.1. First-Step Estimation

The aim of the first step is to estimate the conditional expectation function $x^\star_{tj} := \mathrm{E}[x_{tj}|\mathbf{z}_t] = \mathbf{z}'_t\boldsymbol{\pi}^\star_j$ for $j = 1, \ldots, \bar{p}$ where $x^\star_{tj}$ represents the optimal instrument. Note that $x^\star_{tj} = x_{tj}$ if x_{tj} is exogenous, which corresponds to $j = \bar{p} + 1, \ldots, p$. If $L > T$, OLS estimation of the first-step equations in (2.2) is not feasible as the Gram matrix $T^{-1}\mathbf{Z}'\mathbf{Z}$ with $\mathbf{Z} = ((z_{tj}))$ is singular. The Lasso can achieve consistency in a high-dimensional setting where $L > T$ under the assumption of sparsity and further regularity conditions stated below. Exact sparsity requires that the number of nonzero elements in $\boldsymbol{\pi}^\star_j$, *i.e.*, $\|\boldsymbol{\pi}^\star_j\|_0$, is small relative to the sample size. This assumption is too strong in most applications as $\boldsymbol{\pi}^\star_j$ may have many elements that are, although negligible, not exactly zero. Instead, we assume the existence of a sparse parameter vector $\boldsymbol{\pi}^0_j$ that approximates the true parameter vector $\boldsymbol{\pi}^\star_j$. Specifically, as in [30], we assume that for each endogenous regressor j the number of instruments necessary for approximating the conditional expectation function is smaller than the sample size and the associated approximation error $a_{tj}(\mathbf{z}_t) = x^\star_{tj} - \mathbf{z}'_t\boldsymbol{\pi}^0_j$ converges as specified below.[5]

Assumption 2.1. *Consider the model in* (2.2). *There exists a parameter vector* $\boldsymbol{\pi}^0_j$ *for all* $j = 1, \ldots, \bar{p}$ *such that*

$$\mathrm{E}[x_{tj}|\mathbf{z}_t] = \mathbf{z}'_t\boldsymbol{\pi}^0_j + a_{tj}(\mathbf{z}_t), \quad s_1 := \max_{1\le j\le\bar{p}} \|\boldsymbol{\pi}^0_j\|_0 \ll T, \quad A_{s_1} := \max_{1\le j\le\bar{p}} \sqrt{\frac{1}{T}\sum_{t=1}^{T} a^2_{tj}} \lesssim_{\mathrm{P}} \sqrt{\frac{s_1}{T}}$$

The target parameter $\boldsymbol{\pi}^0_j$ can be motivated as the solution to the infeasible oracle program that penalizes the number of non-zero parameters [31]. Under homoskedasticity, we can write the oracle objective function as

$$\min_{\boldsymbol{\pi}_j} \frac{1}{T}\|\mathbf{X}_j - \mathbf{Z}\boldsymbol{\pi}_j\|^2_2 + \frac{\sigma^2}{T}\|\boldsymbol{\pi}_j\|_0$$

where $\mathbf{X}_j$ is the jth column of the matrix $\mathbf{X} = ((x_{tj}))$. The second term represents the noise level and $\sqrt{s_1/T}$ is the convergence rate of the oracle which knows the true model.

The first-step Lasso estimator for endogenous regressor j is defined as

$$\hat{\boldsymbol{\pi}}_j = \arg\min \frac{1}{T}\|\mathbf{X}_j - \mathbf{Z}\boldsymbol{\pi}_j\|^2_2 + \frac{\lambda_1}{T}\left\|\hat{\boldsymbol{\Upsilon}}_{1j}\boldsymbol{\pi}_j\right\|_1$$

The first term is the residual sum of squares and the second term imposes a penalty on the absolute size of the parameters which is increasing in the penalty level λ_1. The Lasso nests OLS with $\lambda_1 = 0$

[5] The subscripts "1" and "2" indicate, where appropriate, that the corresponding terms refer to the first and second step, respectively.

and $\lambda_1 = \infty$ will lead to a null model. $\hat{\boldsymbol{\Upsilon}}_{1j}$ is a diagonal matrix of penalty loadings which account for heteroskedasticity and may be set to the identity matrix under homoskedasticity [30]. The second term imposes a penalty on the absolute size of the coefficients and, thus, shrinks the coefficient estimates towards zero. The Lasso predictions $\hat{\mathbf{X}}_j := \mathbf{Z}\hat{\boldsymbol{\pi}}_j$ replace $\mathbf{X}_j$ in the second step to address endogeneity. For the exogenous regressors, we set $\hat{\mathbf{X}}_j = \mathbf{X}_j$.

The penalty level λ_1 may be selected by cross-validation in order to minimize the prediction error as originally suggested by Tibshirani [24]. Since the primary purpose of our study is not prediction, but recovery of the spatial network structure, we follow an alternative approach that originates from Bickel *et al.* [27]. The penalty level is chosen as the smallest value that, with a high probability, overrules the random part of the data-generating process, which is represented by the score vector $\mathbf{S}_{1j} = -\frac{2}{T}\hat{\boldsymbol{\Upsilon}}_{1j}^{-1}\mathbf{Z}'\mathbf{u}_j$, *i.e.*,

$$\frac{\lambda_1}{T} \geq c \max_{1\leq j\leq \bar{p}} \|\mathbf{S}_{1j}\|_\infty \quad \text{with } c > 1 \tag{2.3}$$

The event in (2.3) plays a crucial role in the derivation of non-asymptotic bounds and convergence rates. Belloni *et al.* [30] show with the use of moderate deviation theory in [44] that as $T \to \infty$, $P(c \max_{1\leq j\leq \bar{p}} \|\mathbf{S}_{1j}\|_\infty > \lambda_1/T) = o_{\mathrm{P}}(1)$ where

$$\lambda_1 = 2c\sqrt{T}\Phi^{-1}(1 - \alpha/(2L\bar{p})) \quad \text{with} \quad \log(1/\alpha) \lesssim \log(\max(L\bar{p}, T)) \tag{2.4}$$

under possibly non-Gaussian and heteroskedastic errors. Note that the term $\bar{p}$ in (2.4) accounts for the number of Lasso regressions in the first step and L is the number of instruments. c is a constant greater than, but close to 1. In applied work, Belloni *et al.* [45] suggest setting $c = 1.1$ and $\alpha = \min(1/T, 0.05)$.

The optimal penalty loadings are given by

$$\boldsymbol{\Upsilon}^0_{1j} = \mathrm{diag}(\gamma_{1j,1}, \ldots, \gamma_{1j,l}, \ldots, \gamma_{1j,L}), \quad \gamma_{1j,l} = \sqrt{\frac{1}{T}\sum_t z_{tl}^2 u_{tj}^2} \tag{2.5}$$

but are infeasible as u_{tj} is unobserved. Under the iterative Algorithm in Appendix A.2, we can construct asymptotically valid penalty loadings, $\hat{\boldsymbol{\Upsilon}}_{1j}$, that are in the probability limit as least as large as the optimal penalty loadings [30].

The properties of the Lasso estimator depend crucially on the Gram matrix $T^{-1}\mathbf{Z}'\mathbf{Z}$. As stated above, OLS is not feasible if $L > T$ as the Gram matrix is singular, which implies that the minimum eigenvalue is zero,

$$\min_{\boldsymbol{\delta}\neq \mathbf{0}} \frac{\|\mathbf{Z}\boldsymbol{\delta}\|_2}{\sqrt{T}\,\|\boldsymbol{\delta}\|_2} = 0$$

Bickel *et al.* [27] introduce the restricted eigenvalue

$$\min_{\|\boldsymbol{\delta}_{\overline{\Omega}_{1j}}\|_1 \leq C\|\boldsymbol{\delta}_{\Omega_{1j}}\|_1, \boldsymbol{\delta}\neq \mathbf{0}} \frac{\|\mathbf{Z}\boldsymbol{\delta}\|_2}{\sqrt{T}\,\left\|\boldsymbol{\delta}_{\Omega_{1j}}\right\|_2}$$

which is defined as the minimum over the restricted set $\|\boldsymbol{\delta}_{\overline{\Omega}_{1j}}\|_1 \leq C\|\boldsymbol{\delta}_{\Omega_{1j}}\|_1$, where $\Omega_{1j} = \mathrm{supp}(\boldsymbol{\pi}_j^0)$ and C is a positive constant. The condition $\|\boldsymbol{\delta}_{\overline{\Omega}_{1j}}\|_1 \leq C\|\boldsymbol{\delta}_{\Omega_{1j}}\|_1$ holds with high probability and, when it does not hold, it is not required to bound the prediction error norm (see Appendix A.1).

Definition 1. *Let C and $\bar{\kappa}$ be positive constants and Ω denote the active set. We say that the restricted eigenvalue condition holds for $\mathbf{M}$, if as $T \to \infty$*

$$\kappa_C(\mathbf{M}) := \min_{\|\boldsymbol{\delta}_{\overline{\Omega}}\|_1 \leq C\|\boldsymbol{\delta}_{\Omega}\|_1, \boldsymbol{\delta} \neq \mathbf{0}} \frac{\sqrt{s}}{\sqrt{T}} \frac{\|\mathbf{M}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}_{\Omega}\|_1} \geq \bar{\kappa} > 0, \quad s := \|\boldsymbol{\delta}\|_0$$

In the above definition of the restricted eigenvalue the ℓ_2-norm in the denominator is replaced with the ℓ_1-norm using the Cauchy-Schwarz inequality, which allows us to relate the ℓ_1-parameter norm to the ℓ_2-prediction norm. The restricted eigenvalue is closely related to the compatibility constant [46]. Bühlmann and Van de Geer [28] provide an extensive overview of related conditions and their relationship. The restricted eigenvalue conditions hold under general conditions; see, e.g., [27,31,47]. One sufficient condition for the restricted eigenvalue is the restricted sparse eigenvalue condition which requires that any appropriate sub-matrix of the Gram matrix has positive and finite eigenvalues [27].

To accommodate heteroskedasticity, we also define the weighted restricted eigenvalue condition [30],

$$\kappa_C^{\omega}(\mathbf{M}) := \min_{\|\boldsymbol{\Upsilon}^0\boldsymbol{\delta}_{\overline{\Omega}}\|_1 \leq C\|\boldsymbol{\Upsilon}^0\boldsymbol{\delta}_{\Omega}\|_1, \boldsymbol{\delta} \neq \mathbf{0}} \frac{\sqrt{s}}{\sqrt{T}} \frac{\|\mathbf{M}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\Upsilon}^0\boldsymbol{\delta}_{\Omega}\|_1}$$

where $\boldsymbol{\Upsilon}^0$ are the optimal penalty loadings as in (2.5). If the restricted eigenvalue condition holds, the weighted restricted eigenvalue is positive as long as the optimal penalty loadings are bounded away from zero and bounded from above, which we maintain in the following.

With respect to the first-step equations in (2.2), we explicitly state the restricted eigenvalue condition as follows:

Assumption 2.2. *The Restricted Eigenvalue Condition holds for* $\mathbf{Z}$*.*

Under Assumption 2.1 and 2.2, using the penalty level as in (2.4), assuming the penalty loadings $\hat{\boldsymbol{\Upsilon}}_{1j}$ are asymptotically valid, then by Theorem 1 in [30], the ℓ_2-prediction error norm of the Lasso estimator has the following rate of convergence

$$\max_{1 \leq j \leq \bar{p}} \frac{1}{\sqrt{T}} \left\|\mathbf{Z}\hat{\boldsymbol{\pi}}_j - \mathbf{Z}\boldsymbol{\pi}_j^{\star}\right\|_2 \lesssim_{\mathrm{P}} \sqrt{\frac{s_1 \log(\max(L\bar{p}, T))}{T}} \tag{2.6}$$

We do not reproduce the proof of Theorem 1 in [30]. However, our main results in Theorem 1 below is a generalization in that we account for the prediction error that arises from the first-step Lasso estimation. The proof is provided in Appendix A.1. The convergence rate in (2.6) is slower than the oracle rate of $\sqrt{s_1/T}$ by a factor of $\sqrt{\log(\max(L\bar{p}, T))}$, which can be interpreted as the cost of not knowing the active set of $\boldsymbol{\pi}_j^0$.

2.2. Second-Step Estimation

Since the second-step is infeasible by OLS if $p > T$, we require, as in the first step, approximate sparsity and assumptions on the Gram matrix.

Assumption 2.3. *Consider the model in* (2.1)*. There exists a parameter vector* $\boldsymbol{\beta}^0$ *such that*

$$\mathrm{E}[y_t|\mathbf{z}_t] = \mathbf{x}_t^{\star\prime}\boldsymbol{\beta}^0 + r_t(\mathbf{z}_t), \quad s_2 := \|\boldsymbol{\beta}^0\|_0 \ll T, \quad R_{s_2} := \sqrt{\frac{1}{T}\sum_{t=1}^{T} r_t^2} \lesssim \sqrt{\frac{s_2}{T}}, \quad \|\boldsymbol{\beta}^0\|_2 \lesssim_{\mathrm{P}} s_2$$

Assumption 2.4. *The Restricted Eigenvalue Condition holds for* $\hat{\mathbf{X}}$*.*

Assumption 2.3 is similar to Assumption 2.1, but assumes $\|\boldsymbol{\beta}^0\|_2 \lesssim s_2$, which allows us to simplify the expression for the convergence rates. Assumption 2.4 could also be written in terms of the optimal instrument matrix $\mathbf{X}^\star$. Specifically, Assumption 2.4 holds if the restricted eigenvalue holds for $\mathbf{X}^\star$ and $\|\hat{\mathbf{X}} - \mathbf{X}^\star\|_\infty$ is small as specified in [28], Corollary 6.8.

For identification of $\boldsymbol{\beta}^0$ we also require, as standard in the IV/GMM literature, that the matrix $\boldsymbol{\Pi}^0 = (\boldsymbol{\pi}_1^0, \dots, \boldsymbol{\pi}_p^0)$ is full column rank.

Assumption 2.5. $\mathrm{rank}(\boldsymbol{\Pi}^0) = p$*.*

The second-step Lasso estimator uses the predictions $\hat{\mathbf{X}}$ as regressors and is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{T}\left\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}\right\|_2^2 + \frac{\lambda_2}{T}\left\|\hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}\right\|_1$$

where the penalty level is set to

$$\lambda_2 = 2c\sqrt{T}\Phi^{-1}(1-\alpha/(2p)) \quad \text{with} \quad \log(1/\alpha) \lesssim \log(\max(L\bar{p}, T)) \tag{2.7}$$

and the penalty loadings are estimated using the algorithm in Appendix A.2.

The crucial difference to the first-step Lasso estimation is that $\mathbf{X}^\star$ is unobservable and we instead use $\hat{\mathbf{X}}$, which is an estimate that in general deviates from the optimal instrument $\mathbf{X}^\star$. For the two-step Lasso estimator, we consider the prediction bound $1/\sqrt{T}\|\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}^\star\boldsymbol{\beta}^\star\|_2$ where predictions obtained using the unknown optimal instrument and the unknown true parameter vector $\boldsymbol{\beta}^\star$ serve as a reference point. Note that, using the triangle inequality,

$$\begin{aligned}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}^\star\boldsymbol{\beta}^\star\right\|_2 &= \frac{1}{\sqrt{T}}\left\|(\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \hat{\mathbf{X}}\boldsymbol{\beta}^0) + (\hat{\mathbf{X}}\boldsymbol{\beta}^0 - \mathbf{X}^\star\boldsymbol{\beta}^0) + (\mathbf{X}^\star\boldsymbol{\beta}^0 - \mathbf{X}^\star\boldsymbol{\beta}^\star)\right\|_2 \\ &\le \frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \hat{\mathbf{X}}\boldsymbol{\beta}^0\right\|_2 + \frac{1}{\sqrt{T}}\left\|\hat{\mathbf{V}}\boldsymbol{\beta}^0\right\|_2 + R_{s_2}\end{aligned}$$

where we define $\hat{\mathbf{V}} = \hat{\mathbf{X}} - \mathbf{X}^\star$ which has the typical element $\hat{v}_{jt} = \mathbf{z}_t'\hat{\boldsymbol{\pi}}_j - \mathbf{z}_t'\boldsymbol{\pi}_j^\star$. The bound for the third term is stated in Assumption 2.3. The convergence rate for the second term follows from prediction norm rate of the first-step Lasso in (2.6). The bound for the first term is derived in Appendix A.1. Combining the three bounds, we have the following result.

Theorem 1. *Consider the model in* (2.1)–(2.2)*. Suppose Assumptions 2.1–2.5 hold. Suppose asymptotically valid penalty loadings are used and the penalty levels* λ_1 *and* λ_2 *are set as in* (2.4) *and* (2.7)*. Then,*

$$\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \mathbf{X}^\star\boldsymbol{\beta}^\star\right\|_2 \lesssim_{\mathrm{P}} s_2^2\sqrt{\frac{s_1\log(\max(L\bar{p}, T))}{T}}$$

Furthermore, if s_1 and s_2, do not depend on T, then

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1 \lesssim_{\mathrm{P}} \sqrt{\frac{\log(\max(L\bar{p}, T))}{T}}$$

The proof is provided in Appendix A.1. As expected, the convergence rates of the ℓ_2-prediction norm depend on the degree of sparsity in the first-step and second-step equation. The second part of the theorem is relevant for the spatial panel model in the next section where the sparsity level (s_1 and s_2) and the dimension of the problem (L and p) depend on the number of units (*i.e.*, n), but not on the time dimension (T).

3. The Spatial Autoregressive Model

This section applies the proposed two-step Lasso procedure to the spatial lag model in (1.1). In Section 3.2, we discuss two extensions to the two-step Lasso estimator; namely, the post-Lasso and thresholded post-Lasso.

3.1. Two-Step Lasso

The structural and reduced form equations can be written as

$$\mathbf{y}_i = \sum_{j=1}^{n} w_{ij}^\star \mathbf{y}_j + \mathbf{X}_i \boldsymbol{\beta}_i^\star + \mathbf{e}_i \tag{3.1}$$

$$\mathbf{y}_j = \sum_{s=1}^{n} \mathbf{X}_s \boldsymbol{\pi}_{j,s}^\star + \mathbf{u}_j, \qquad i, j = 1, \ldots, n \tag{3.2}$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $w_{jj}^\star = 0$, and $\mathbf{X}_i = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')'$ is the $T \times K$ matrix of exogenous regressors. We assume that e_{it} is independently distributed across t, *i.e.*, $E[e_{it}e_{is}] = 0$ for $t \neq s$. The *-superscripts indicate that we interpret the parameters as the true parameter values.

It is evident that the spatial model in (3.1)–(3.2) is an application of the more general model in (2.1)–(2.2). Specifically, the right-hand side regressors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ correspond to endogenous regressors and $\mathbf{X}_i$ corresponds to the exogenous regressors in (2.1). Furthermore, the set of exogenous instruments is given by $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

The choice of instruments is closely related to Kelejian and Prucha [48]. To identify the spatial autoregressive parameter, they suggest the use of first and higher order spatial lags of exogenous regressors as instruments for the endogenous spatial lag. As discussed in the Introduction, we use $\mathbf{X}_j$ as instruments in order to identify $w_{ij}^\star$, which represents the causal impact of $\mathbf{y}_j$ on $\mathbf{y}_i$. Therefore, for identification, we require contemporaneous exogeneity across space:

Assumption 3.1. $\mathrm{E}[e_{it}|\mathbf{x}_{jt}] = 0$ *for all* $i, j = 1, \ldots, n$ *and* $t = 1, \ldots, T$.

In many applications, estimation of (3.1)–(3.2) by 2SLS is not feasible as there appear $n - 1 + K$ and nK regressors on the right-hand side, respectively, which are both potentially larger than T. In order to exploit the Lasso estimator, we require sparseness as in Section 2.

Assumption 3.2. *(a) Consider the model in (3.1). There exists a parameter vector* $\mathbf{w}_i^0 = (w_{i1}^0, \dots, w_{in}^0)$ *for all* $i = 1, \dots, n$ *with* $w_{ii}^0 = 0$ *such that*

$$\mathrm{E}[y_{it}|\mathbf{x}_{1t}, \dots, \mathbf{x}_{nt}] = \sum_{j=1}^{n} w_{ij}^0 y_{jt}^\star + \mathbf{x}_{it}'\boldsymbol{\beta}_i^0 + a_{it}$$

$$s_1 + K \ll T, \qquad A_{s_1} := \max_{1\le i\le n} \sqrt{\frac{1}{T}\sum_{t=1}^{T} a_{it}^2} \lesssim_{\mathrm{P}} \sqrt{\frac{s_1}{T}}$$

where $y_{jt}^\star = \mathrm{E}[y_{jt}|\mathbf{x}_{1t}, \dots, \mathbf{x}_{nt}]$, $s_1 := \max_{1\le i\le n}\|\mathbf{w}_i^0\|_0$ *and* $K = \|\boldsymbol{\beta}_i^0\|_0$

(b) Consider the model in (3.2). There exists a parameter vector $\boldsymbol{\pi}_{i,j}^0$ *for all* $i = 1, \dots, n$ *such that*

$$\mathrm{E}[y_{it}|\mathbf{x}_{1t}, \dots, \mathbf{x}_{nt}] = \sum_{j\neq i} \mathbf{x}_{jt}'\boldsymbol{\pi}_{i,j}^0 + \mathbf{x}_{it}'\boldsymbol{\pi}_{i,i}^0 + r_{it}$$

$$s_2 + K \ll T, \qquad R_{s_2} := \max_{1\le i\le n} \sqrt{\frac{1}{T}\sum_{t=1}^{T} r_{it}^2} \lesssim_{\mathrm{P}} \sqrt{\frac{s_2}{T}}$$

where $s_2 := \max_{1\le i\le n}\sum_{j\neq i}\|\boldsymbol{\pi}_{i,j}^0\|_0$ *and* $K = \|\boldsymbol{\pi}_{i,i}^0\|_0$

To simplify the exposition and without loss of generality, we assume that $\boldsymbol{\beta}_i^0$ does not include any zero elements, implying that all regressors in $\mathbf{x}_{it}$ are relevant determinants of the dependent variable y_{it}. The assumption also guarantees identification as long as $K \geq 1$.

The sparsity assumptions in Assumption 3.2 (a) and Assumption 3.2 (b) are related. To see this, consider the case where $n = 2$. Then, the reduced form equations are given by

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\pi}_{1,1} + \mathbf{X}_2\boldsymbol{\pi}_{1,2} + \mathbf{u}_1$$
$$\mathbf{y}_2 = \mathbf{X}_1\boldsymbol{\pi}_{2,1} + \mathbf{X}_2\boldsymbol{\pi}_{2,2} + \mathbf{u}_2$$

with $\boldsymbol{\pi}_{2,1} = \frac{w_{21}}{1-w_{12}w_{21}}\boldsymbol{\beta}_1$, $\boldsymbol{\pi}_{1,2} = \frac{w_{12}}{1-w_{12}w_{21}}\boldsymbol{\beta}_2$, $\boldsymbol{\pi}_{1,1} = \frac{1}{1-w_{12}w_{21}}\boldsymbol{\beta}_1$ and $\boldsymbol{\pi}_{2,2} = \frac{1}{1-w_{12}w_{21}}\boldsymbol{\beta}_2$ where we assume $|w_{12}| < 1$ and $|w_{21}| < 1$. It becomes evident that, if $\boldsymbol{\pi}_{1,2} = \mathbf{0}$, then $w_{12} = 0$ must hold by assumption given that $\|\boldsymbol{\beta}\|_0 = K$. That is, sparseness of the $\boldsymbol{\pi}_{i,j}$ parameter vectors as specified in Assumption 3.2 (b) implies sparseness of the $\mathbf{W}_n$ matrix in Assumption 3.2 (a) if $n = 2$.

We maintain the following basic assumptions regarding the spatial weights matrix.

Assumption 3.3. *(a) The spatial weights matrix,* $\mathbf{W}_n^0 = ((w_{ij}^0; i, j = 1, \dots, n))$*, is* $n \times n$ *with zeros on the diagonal,* $w_{ii} = 0$*. (b) The spatial weights matrix is time-invariant. (c) The row sums are bounded in absolute value, i.e.,* $\max_i \sum_j |w_{ij}| < 1$.

Assumption 3.3 (a) is standard. Assumption 3.3 (b) is required as the identification strategy exploits variation over time to identify the weights matrix and is standard in the spatial panel econometrics literature; see, e.g., [20]. The assumption corresponds to parameter stability over time in time series.[6]

[6] Whether this assumption is reasonable depends on the application. If there is a regime change at a known date, the model can be estimated for each sub-period separately, assuming that parameter stability holds within in each sub-period and that the time dimension is sufficiently large.

Assumption 3.3 (c) can be interpreted similar to the stationarity condition in time-series econometrics. Assumption (a) and (c) ensure that $\mathbf{I}_n - \mathbf{W}_n^0$ is invertible, where $\mathbf{I}_n$ is the identity matrix of dimension n. Invertibility of $\mathbf{I}_n - \mathbf{W}_n^0$ is required to derive the reduced form equations in (3.2). The assumptions differ from standard assumptions in the spatial econometrics literature in two points; *c.f.*, [48,49]. First, we do not make use of the spatial autoregressive coefficient as the spatial autoregressive coefficient and the spatial weights are not separately identified. Second, we do not apply row-standardization as commonly employed. This is because some of the spatial weights can be negative, a condition negated in a large part of the literature that measures spatial weights by distances or contiguity. Applications that allow for an estimated spatial weights matrix show that negative spatial weights are common in practice; see for example [7–9]. Furthermore, we stress that Assumption 3.3 does not impose any structure on the spatial weights matrix such as symmetry and we allow the interactions effects to be positive and negative.

In order to write the first and second-step Lasso estimator compactly, we introduce some notation. Let $\bar{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $\boldsymbol{\pi}'_j = (\boldsymbol{\pi}'_{j,1}, \ldots, \boldsymbol{\pi}'_{j,n})$ is the corresponding parameter vector. The first-step Lasso estimator solves

$$\min \left\|\mathbf{y}_i - \bar{\mathbf{X}}\boldsymbol{\pi}_i\right\|_2^2 + \lambda_1 \left\|\hat{\boldsymbol{\Upsilon}}_{1,i}\boldsymbol{\pi}_i\right\|_1$$

Let $\hat{\mathbf{y}}_i$ denote the first-step predictions from Lasso estimation and let $\hat{\mathbf{Y}} = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_n)$. Furthermore, define $\mathbf{w}_i = (w_{i1}, \ldots, w_{in})$ with $w_{ii} = 0$, which is the ith row of the spatial weights matrix. This allows us to write the second-step matrix of regressors as $\hat{\mathbf{G}}_i = (\hat{\mathbf{Y}}, \mathbf{X}_i)$ and define the corresponding parameter vector $\boldsymbol{\theta}'_i = (\mathbf{w}_i, \boldsymbol{\beta}'_i)$ The second-step Lasso solves

$$\min \left\|\mathbf{y}_i - \hat{\mathbf{G}}_i\boldsymbol{\theta}_i\right\|_2^2 + \lambda_2 \left\|\hat{\boldsymbol{\Upsilon}}_{2,i}\boldsymbol{\theta}_i\right\|_1$$

We require both $\bar{\mathbf{X}}$ and $\hat{\mathbf{G}}_i$ to be well-behaved as stated in Assumption 3.4.[7]

Assumption 3.4. *The Restricted Eigenvalue Condition holds for $\bar{\mathbf{X}}$ and $\hat{\mathbf{G}}_i$ for all $i = 1, \ldots, n$.*

The penalty levels are set to

$$\lambda_1 = 2c\sqrt{T}\Phi^{-1}(1 - \alpha/(2n^2K)) \tag{3.3}$$

$$\lambda_2 = 2c\sqrt{T}\Phi^{-1}(1 - \alpha/(2n(n-1+K))) \quad \text{with} \quad \log(1/\alpha) \lesssim \log(\max(n^2K, T)) \tag{3.4}$$

Note there are nK and $n - 1 + K$ penalized regressors in the first and second step, respectively, and n Lasso regressions in each step. The penalty loadings are again estimated using Algorithm A.2.

The convergence rates of the two-step Lasso estimator follow from Theorem 1. However, while the general setting in Section 2 allows s_1, s_2 and the number of first and second-step variables to depend on T, we can assume in the spatial panel setting that s_1, s_2 and n are independent of T. Therefore, we obtain the following convergence rates.

[7] To simplify the exposition, the first and second-step Lasso also applies a penalty to $\boldsymbol{\beta}_i$ and $\boldsymbol{\pi}_{i,i}$, although we assume $\|\boldsymbol{\pi}_{i,i}\|_0 = \|\boldsymbol{\beta}_i\|_0 = K$ for identification. For better performance in finite samples, we recommend that the coefficients $\boldsymbol{\beta}_i$ and $\boldsymbol{\pi}_{i,i}$ are not penalized.

Corollary 1. *Consider the model in (3.1)–(3.2). Suppose Assumptions 3.1–3.4 hold. Suppose asymptotically valid penalty loadings are used and the penalty levels are set as in (3.3) and (3.4). Then,*

$$\max_i \frac{1}{\sqrt{T}} \left\| \hat{\mathbf{G}}_i \hat{\boldsymbol{\theta}}_i - \mathbf{G}^\star \boldsymbol{\theta}_i^\star \right\|_2 \lesssim_{\mathrm{P}} \sqrt{\frac{\log(\max(n^2 K, T))}{T}}$$
$$\max_i \left\| \hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^0 \right\|_1 \lesssim_{\mathrm{P}} \sqrt{\frac{\log(\max(n^2 K, T))}{T}}$$

3.2. Post-Lasso and Thresholded Post-Lasso

The shrinkage of the Lasso estimator induces a downward bias which can be addressed by the post-Lasso estimator. The post-Lasso estimator treats the Lasso as a genuine model selector and applies OLS to the set of regressors for which the Lasso coefficient estimate is non-zero. In other words, post-Lasso is OLS applied to the model selected by the Lasso. Formally, the first and second-step post-Lasso estimator of the spatial autoregressive model are defined as

$$\tilde{\boldsymbol{\pi}}_i = \arg\min_{\boldsymbol{\pi}_i} \left\| \mathbf{y}_i - \bar{\mathbf{X}} \boldsymbol{\pi}_i \right\|_2^2 \quad \text{s.t.} \quad \text{supp}(\boldsymbol{\pi}_i) \subseteq \text{supp}(\hat{\boldsymbol{\pi}}_i)$$
$$\tilde{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i} \left\| \mathbf{y}_i - \hat{\mathbf{G}}_i \boldsymbol{\theta}_i \right\|_2^2 \quad \text{s.t.} \quad \text{supp}(\boldsymbol{\theta}_i) \subseteq \text{supp}(\hat{\boldsymbol{\theta}}_i)$$

The thresholded post-Lasso addresses the issue that the Lasso estimator often selects too many variables and that, despite the ℓ_1-penalization, many coefficient estimates are very small, but not exactly zero. The thresholded post-Lasso applies OLS to all spatial lags for which the post-Lasso estimate is larger than a pre-defined threshold τ.[8] While it is in general difficult to select and justify a specific threshold, in the spatial autoregressive model we can use the knowledge that $-1 < w_{ij} < 1$ and assume interaction effects that are smaller than, for example, 0.05 are negligible. For formal results on the post-Lasso and thresholded Lasso, see Belloni and Chernozhukov [31].

4. Monte Carlo Simulation

This Monte Carlo study[9] explores the finite sample performance of the proposed two-step Lasso estimator for estimating the spatial autoregressive model

$$y_{it} = \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} y_{jt} + \eta_i + \mathbf{x}_{it}' \boldsymbol{\beta} + \varepsilon_{it}, \quad t = 1, \ldots, T; i = 1, \ldots, n \tag{4.1}$$

We consider two different spatial weights matrices. Specification 1 is given by

$$w_{ij} = \begin{cases} 1 & \text{if } |j - i| = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i, j = 1, \ldots, n \tag{4.2}$$

[8] The thresholded Lasso estimators considered in [31] apply the threshold to the Lasso estimates whereas we apply the threshold to the post-Lasso estimates.

[9] We are grateful to two anonymous referees who suggested useful extensions to our Monte Carlo simulations.

and specification 2 is given by

$$w_{ij} = \begin{cases} 1 & \text{if } j - i = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i, j = 1, \dots, n \tag{4.3}$$

Subsequently, a row-standardization is applied such that the row sum is equal to $\bar{w}$. The row-standardization ensures that the strength of spill-over effects is constant across i. The strength of spatial interactions is determined by $\bar{w}$, which corresponds to the spatial autoregressive coefficient.

The spatial weights matrix in specification 1 has non-zeros on the sub-diagonal and super-diagonal. Thus, the weights matrix is symmetric and the number of non-zero elements is $2(n - 1)$. In specification 2, only the super-diagonal elements are non-zero, implying $n - 1$ non-zero elements. The structure in (4.3) corresponds to the extreme case where there are only one-way spatial effects. Specification 2 is in our view more challenging than specification 1 as the triangular structure makes it difficult to identify the direction of causal effects. Note that, the spatial weights matrix is in principle identified if the spatial weights matrix is known to be triangular or symmetric. However, the challenge here is to estimate the spatial weights matrix without any prior knowledge. We stress that the estimation strategy does not depend on any particular structure of the spatial weights matrix, but only requires sparsity.

The parameter vector $\boldsymbol{\beta}$ is a K-dimensional vector of ones. Hence, $\boldsymbol{\beta}$ is constant across i, although the estimation method allows for spatial heterogeneity of $\boldsymbol{\beta}$. The exogenous regressors and the spatial fixed effect η_i are drawn from the standard normal distribution, *i.e.*, $x_{k,it} \sim N(0, 1)$ for $k = 1, \dots, K$. The idiosyncratic error is drawn as $\varepsilon_{it} \sim N(0, \sigma^2_{it})$ where

$$\sigma^2_{it} = \frac{(1 + \mathbf{x}'_{it}\boldsymbol{\beta})^2}{\frac{1}{nT}\sum_{i,t}(1 + \mathbf{x}'_{it}\boldsymbol{\beta})^2}$$

which induces conditional heteroskedasticity.

We consider four estimators:[10] (a) The two-step Lasso introduced in Section 3.1. (b) Two-step post-Lasso from Section 3.2. (c) Two-step thresholded post-Lasso with a threshold of $\tau = 0.05$. (d) The oracle estimator. The oracle estimator has full knowledge about which weights are non-zero and applies 2SLS to the true model. The oracle estimator is infeasible as the true model is in general unknown and only serves as a benchmark. The penalty levels are defined as in (3.3)–(3.4) with $c = 1.1$ and $\alpha = \min(1/T, 0.05)$.[11] The penalty loadings are estimated by Algorithm A.2.

We consider a range of different settings. Specifically, $n = \{30, 50, 70\}$, $T = \{50, 100, 500\}$, $K = 1$ and $\bar{w} = \{0.5, 0.7, 0.9\}$. We have also considered $K = 2$ which results in noticeable performance improvements. However, we do not report the results and focus on $K = 1$ which is the minimum requirement for identification. The number of Monte Carlo replications is 1000.

[10] The Lasso estimations were conducted in *R* based on the package *glmnet* by Friedman *et al.* [50]. The code for the two-step Lasso, two-step post-Lasso and thresholded post-Lasso are available on request.

[11] We have also considered, among others, $c = 1.01$ and $\alpha = 0.05/\log(T)$ and did not find significant performance differences.

Table 1. Monte Carlo results: Specification 1. **(a)** Two-step Lasso; **(b)** Two-step post-Lasso; **(c)** Thresholded post-Lasso with $\tau = 0.05$; **(d)** Oracle estimator.

(a)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.70	30	50	19.23	19.01	0.04335	0.04266	0.04335
0.70	30	100	14.62	21.22	0.03812	0.03747	0.03812
0.70	30	500	5.30	24.82	0.02310	0.02305	0.02310
0.70	50	50	13.96	15.13	0.02855	0.02826	0.02855
0.70	50	100	8.86	17.58	0.02653	0.02626	0.02653
0.70	50	500	3.14	22.13	0.01725	0.01718	0.01725
0.70	70	50	11.93	12.35	0.02132	0.02130	0.02132
0.70	70	100	6.09	15.36	0.02071	0.02055	0.02071
0.70	70	500	1.89	20.39	0.01454	0.01446	0.01454
0.90	30	50	2.07	14.99	0.02360	0.02346	0.02360
0.90	30	100	0.97	14.92	0.02002	0.01992	0.02002
0.90	30	500	0.18	18.35	0.01372	0.01365	0.01372
0.90	50	50	1.03	12.69	0.01495	0.01489	0.01495
0.90	50	100	0.29	13.11	0.01354	0.01354	0.01354
0.90	50	500	0.03	15.24	0.00935	0.00931	0.00935
0.90	70	50	0.52	10.37	0.01048	0.01044	0.01048
0.90	70	100	0.15	12.17	0.01058	0.01054	0.01058
0.90	70	500	0.01	13.60	0.00759	0.00758	0.00759

(b)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.70	30	50	8.83	11.55	0.03649	0.03606	0.03649
0.70	30	100	5.64	11.89	0.03090	0.03083	0.03090
0.70	30	500	2.18	12.69	0.02029	0.02022	0.02029
0.70	50	50	4.81	9.38	0.02453	0.02432	0.02453
0.70	50	100	1.90	10.12	0.02187	0.02185	0.02187
0.70	50	500	0.58	10.80	0.01519	0.01519	0.01519
0.70	70	50	3.84	7.65	0.01829	0.01828	0.01829
0.70	70	100	0.87	8.96	0.01746	0.01741	0.01746
0.70	70	500	0.17	9.78	0.01277	0.01273	0.01277
0.90	30	50	0.76	9.83	0.01823	0.01806	0.01823
0.90	30	100	0.25	7.93	0.01267	0.01254	0.01267
0.90	30	500	0.12	7.62	0.00794	0.00787	0.00794
0.90	50	50	0.29	9.27	0.01364	0.01349	0.01364
0.90	50	100	0.04	7.94	0.00947	0.00944	0.00947
0.90	50	500	0.01	6.07	0.00513	0.00512	0.00513
0.90	70	50	0.17	7.26	0.00931	0.00928	0.00931
0.90	70	100	0.02	8.26	0.00854	0.00852	0.00854
0.90	70	500	0.00	5.50	0.00409	0.00408	0.00409

(c)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.70	30	50	10.04	6.36	0.02561	0.02523	0.02561
0.70	30	100	6.40	6.44	0.02165	0.02145	0.02165
0.70	30	500	2.36	6.60	0.01368	0.01360	0.01368
0.70	50	50	5.71	4.85	0.01621	0.01609	0.01621
0.70	50	100	2.28	5.21	0.01434	0.01430	0.01434
0.70	50	500	0.65	5.36	0.00968	0.00969	0.00968
0.70	70	50	4.59	3.84	0.01200	0.01197	0.01200
0.70	70	100	1.13	4.48	0.01114	0.01110	0.01114
0.70	70	500	0.20	4.69	0.00791	0.00790	0.00791
0.90	30	50	1.57	5.00	0.01316	0.01298	0.01316
0.90	30	100	0.42	3.54	0.00923	0.00915	0.00923
0.90	30	500	0.13	2.35	0.00528	0.00525	0.00528
0.90	50	50	1.06	4.30	0.00877	0.00871	0.00877
0.90	50	100	0.12	3.25	0.00620	0.00615	0.00620
0.90	50	500	0.01	1.54	0.00310	0.00308	0.00310
0.90	70	50	0.56	3.14	0.00589	0.00584	0.00589
0.90	70	100	0.09	3.24	0.00513	0.00510	0.00513
0.90	70	500	0.00	1.21	0.00230	0.00229	0.00230

(d)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.70	30	50	–	–	0.02742	0.02444	0.02742
0.70	30	100	–	–	0.01915	0.01779	0.01915
0.70	30	500	–	–	0.00762	0.00753	0.00762
0.70	50	50	–	–	0.01533	0.01462	0.01533
0.70	50	100	–	–	0.01135	0.01057	0.01135
0.70	50	500	–	–	0.00455	0.00453	0.00455
0.70	70	50	–	–	0.01114	0.01055	0.01114
0.70	70	100	–	–	0.00816	0.00771	0.00816
0.70	70	500	–	–	0.00325	0.00323	0.00325
0.90	30	50	–	–	0.02576	0.02358	0.02576
0.90	30	100	–	–	0.02039	0.01887	0.02039
0.90	30	500	–	–	0.01250	0.01237	0.01250
0.90	50	50	–	–	0.01562	0.01418	0.01562
0.90	50	100	–	–	0.01216	0.01132	0.01216
0.90	50	500	–	–	0.00751	0.00744	0.00751
0.90	70	50	–	–	0.01087	0.01020	0.01087
0.90	70	100	–	–	0.00860	0.00816	0.00860
0.90	70	500	–	–	0.00533	0.00532	0.00533

"False neg." denotes false negative rate in %. "False pos." denotes false positive rate in %. RMSE denotes root-mean-square error. The bias is defined in (4.4). The false negative and false positive rate is 0% for the oracle estimator by construction. The oracle estimator is infeasible in practice and serves only as a reference point. Number of replications is 1000. The number of exogenous regressors is $K = 1$. See description in the main text.

Tables 1 and 2 report the following statistics to assess the performance of the estimators. "False negative" is the average percentage of non-zero elements falsely identified as being zero. "False positive" is the average percentage of zero elements falsely identified as non-zero. Furthermore, let $\widehat{\mathbf{W}}_{(i)}$ be the estimate of the spatial weights matrix from the ith Monte Carlo iteration. The bias is defined as

$$\widehat{\text{bias}}_{(i)} = \frac{1}{n(n-1)} \left\| \widehat{\mathbf{W}}_{(i)} - \mathbf{W}_n \right\|_1 \tag{4.4}$$

where $\|\cdot\|_1$ denotes the entry-wise ℓ_1-norm. Average and median bias across iterations are reported, as well as the root-mean-square error (RMSE). Note that the false negative and false positive rate are 0% for the oracle estimator by construction. We do not report the bias for the estimation of $\boldsymbol{\beta}$, since estimation of $\boldsymbol{\beta}$ is a standard problem.

Table 2. Monte Carlo results: Specification 2. **(a)** Two-step Lasso; **(b)** Two-step post-Lasso; **(c)** Thresholded post-Lasso with $\tau = 0.05$; **(d)** Oracle estimator.

(a)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.50	30	50	41.60	19.36	0.05258	0.04915	0.05258
0.50	30	100	32.82	20.11	0.04119	0.03952	0.04119
0.50	30	500	7.80	18.61	0.01904	0.01877	0.01904
0.50	50	50	39.98	15.89	0.03820	0.03512	0.03820
0.50	50	100	30.54	16.98	0.03055	0.02976	0.03055
0.50	50	500	7.26	17.07	0.01497	0.01483	0.01497
0.50	70	50	40.38	12.76	0.02885	0.02730	0.02885
0.50	70	100	28.60	14.98	0.02477	0.02447	0.02477
0.50	70	500	6.90	15.95	0.01290	0.01284	0.01290

(b)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.50	30	50	36.30	13.00	0.05679	0.05388	0.05679
0.50	30	100	25.70	13.08	0.04681	0.04572	0.04681
0.50	30	500	4.46	13.10	0.02784	0.02779	0.02784
0.50	50	50	32.20	10.46	0.03943	0.03689	0.03943
0.50	50	100	20.85	10.80	0.03353	0.03326	0.03353
0.50	50	500	3.22	11.66	0.02276	0.02274	0.02276
0.50	70	50	32.13	8.41	0.02944	0.02814	0.02944
0.50	70	100	17.51	9.37	0.02653	0.02642	0.02653
0.50	70	500	2.32	10.59	0.01958	0.01956	0.01958

(c)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.50	30	50	37.53	7.81	0.03843	0.03722	0.03843
0.50	30	100	26.45	8.03	0.03247	0.03218	0.03247
0.50	30	500	4.54	7.72	0.01811	0.01806	0.01811
0.50	50	50	33.53	6.00	0.02573	0.02498	0.02573
0.50	50	100	21.68	6.36	0.02257	0.02252	0.02257
0.50	50	500	3.31	6.72	0.01441	0.01438	0.01441
0.50	70	50	33.43	4.72	0.01946	0.01899	0.01946
0.50	70	100	18.29	5.40	0.01762	0.01758	0.01762
0.50	70	500	2.40	5.99	0.01224	0.01223	0.01224

(d)

$\bar{w}$	N	T	False	False		bias	
			neg.	pos.	mean	median	RMSE
0.50	30	50	–	–	0.01335	0.01186	0.01335
0.50	30	100	–	–	0.00934	0.00861	0.00934
0.50	30	500	–	–	0.00351	0.00348	0.00351
0.50	50	50	–	–	0.00794	0.00724	0.00794
0.50	50	100	–	–	0.00559	0.00524	0.00559
0.50	50	500	–	–	0.00210	0.00208	0.00210
0.50	70	50	–	–	0.00561	0.00512	0.00561
0.50	70	100	–	–	0.00388	0.00371	0.00388
0.50	70	500	–	–	0.00149	0.00149	0.00149

See notes in Table 1.

4.1. Specification 1

The first specification in (4.2) defines a sparse, symmetric matrix. Across all n and $\bar{w}$, the performance of the two-step Lasso improves in terms of false negative rate and bias as T increases. For example, if $n = 70$, $T = 50$ and $\bar{w} = 0.7$, in which case the model cannot be estimated by 2SLS, on average more than 88.0% of the non-zero spatial weights are identified by the Lasso. When $\bar{w} = 0.9$, this rate increases to 99.4%. However, the false positive rate of the Lasso estimator is high at approximately 10%–25% and remains high as T increases. This is in line with the known phenomenon that the Lasso estimator often selects too many variables; see, e.g., [28].

The two-step post-Lasso estimator shows substantial performance improvement over the two-step Lasso. The bias is smaller across all T and n, suggesting that post-Lasso OLS estimation successfully addresses the shrinkage bias arising from ℓ_1-penalization. Moreover, the two-step post-Lasso also dominates the two-step Lasso in terms of false negative and false positive rate. This is consistent with [30,31], who show that the post-Lasso often performs as least as good as the Lasso. However, the false positive rate is still relatively high at 5%–13% and does not seem to decrease with T. The thresholded post-Lasso, which sets post-Lasso estimates below 0.05 equal to zero, improves upon the post-Lasso in that it shows a lower false positive rate. While we do not recommend $\tau = 0.05$ as a general threshold, the thresholded post-Lasso reveals that many ‘falsely positive’ post-Lasso estimates are close to zero, but not exactly zero, which explains the high false positive rate. As expected, the oracle estimator which knows the true model, exhibits the lowest bias across all n and T.

Notice that both false negative as well as false positive rate decrease with n. The decrease in the false positive rate is because the number of zero weights increases with n as a proportion of the total number of off-diagonal elements in $\mathbf{W}_n$. The same situation holds in many real spatial applications where the number of neighbors of a region are bounded. In turn such boundedness is a necessity for spatial stationarity; see Assumption 3.3 and the spatial granularity condition in [37]. Note that with large n, standard least squares methods would not work because of high-dimensionality, which underlines the important advantage of the Lasso-based methods proposed in this article.

Figure 2 shows how often each w_{ij} is identified as being non-zero by the estimators for $n = T = 50$. It can be seen that the two-step procedures successfully recover the spatial structure in (4.2). Note that weights to the left of the sub-diagonal and to the right of the super-diagonal (*i.e.*, $w_{13}, w_{24}, w_{31}, \ldots,$ *etc.*) are falsely selected slightly more often relative to other weights. This is likely due to indirect effects, resulting in spatial spillage. For example, w_{13} is selected slightly more often relative to other zero elements as y_{3t} affects y_{1t} through y_{2t}.

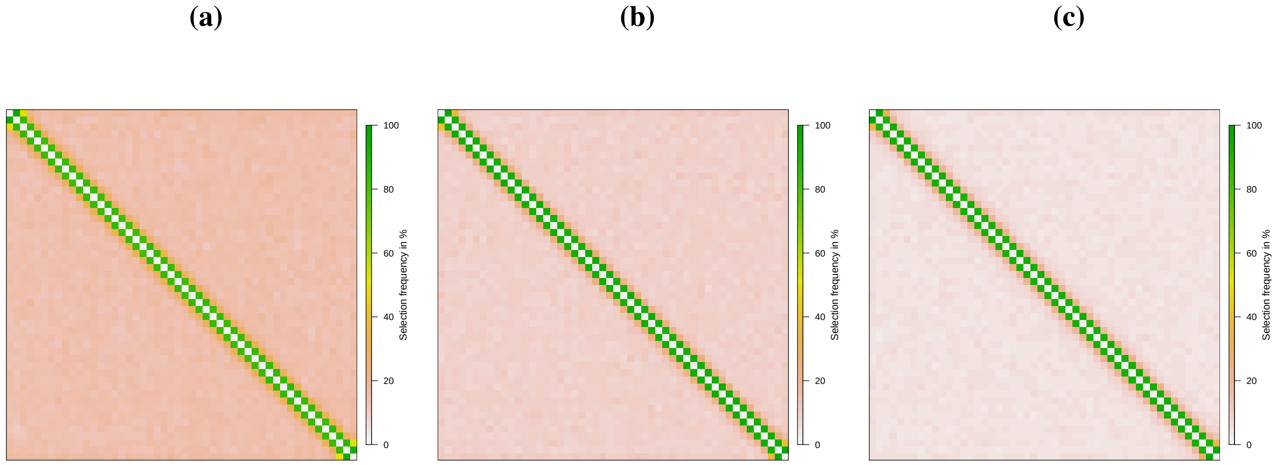

Figure 2. Recovery of Spatial Weights Matrix ($N = 50$, $T = 50$): Specification 1. **(a)** Two-step Lasso; **(b)** Two-step post-Lasso; **(c)** Two-step post-Lasso with $\tau = 0.05$.

4.2. Specification 2

As expected, the performance under specification 2 is not as satisfactory as for specification 1. Table 2 shows that false negative rate and bias decrease in T for all three Lasso-based estimators. As in specification 1, the two-step post-Lasso outperforms the two-step Lasso in terms of the false negative rate. The thresholded Lasso mainly differs from the two-step post-Lasso in that the false positive rate is lower. Figures 3 and 4 show the selection frequency for $n = T = 50$ and $n = 50, T = 500$. For $T = 50$, it can clearly be seen that the elements in the sub-diagonal are selected more often relative to other non-zero elements, stressing the difficulty of identifying the direction of the effects in small samples. This problem reduces with T and is negligible for $T = 500$, see Figure 4.

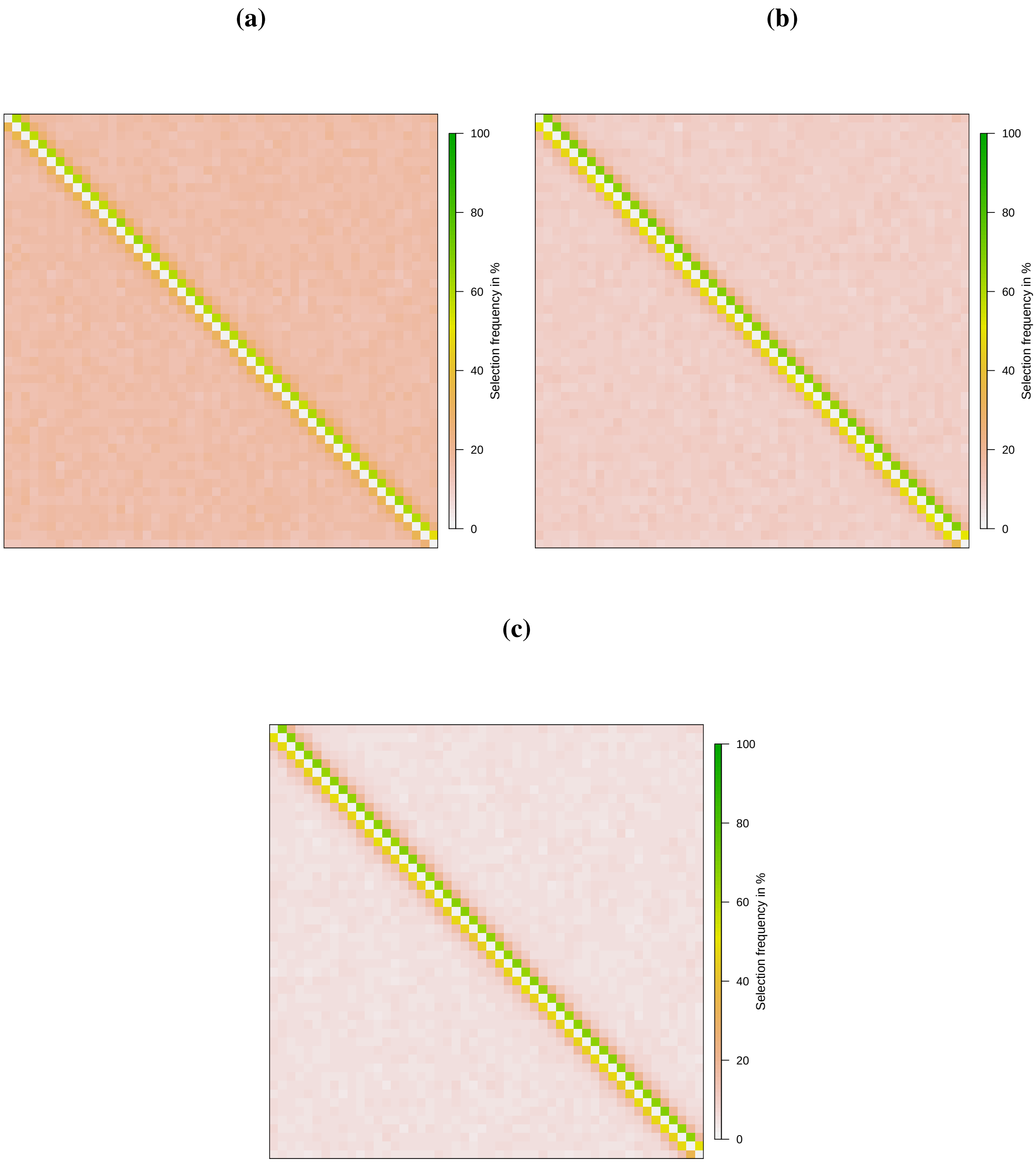

Figure 3. Recovery of Spatial Weights Matrix ($N = 50$, $T = 50$): Specification 2. **(a)** Two-step Lasso; **(b)** Two-step post-Lasso; **(c)** Two-step post-Lasso with $\tau = 0.05$.

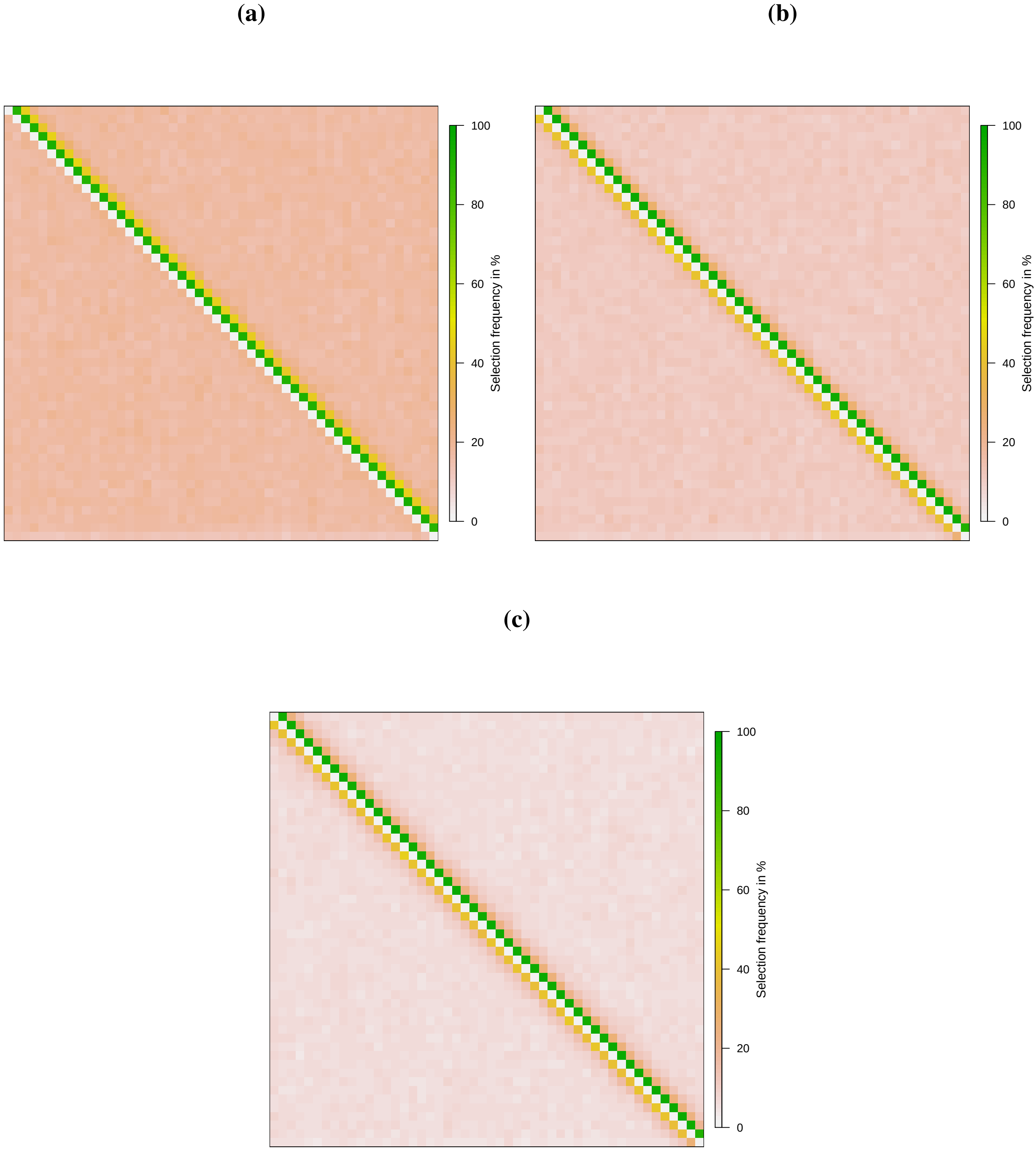

Figure 4. Recovery of Spatial Weights Matrix ($N = 50$, $T = 500$): Specification 2. **(a)** Two-step Lasso; **(b)** Two-step post-Lasso; **(c)** Two-step post-Lasso with $\tau = 0.05$.

Overall, the two-step Lasso performs well in recovering the network structure, even for the more challenging specification 2. However, we observe that the two-step Lasso selects too many spatial lags in small samples, although the performance improves substantially with T in terms of bias and false negative rate. The two-step post-Lasso outperforms the two-step Lasso in terms of bias and selection performance.

5. Conclusions

The identification of interaction effects is crucial for the understanding of how individuals, firms and regions interact. However, to date there is still a lack of methods that allow the estimation of interaction effects, particularly when the spatial dimension is large. Thus, most applied spatial econometric research uses *ad hoc* specifications to incorporate interaction effects. The lack of estimation strategies may also explain why interaction effects in socio-economic processes are often ignored.

We propose a two-step procedure based on the Lasso estimator that accounts for reverse causality and allows estimating interaction effects between units in a spatial autoregressive panel model without requiring any prior knowledge about the network structure. The identifying assumption is sparsity. The two-step estimator can be implemented based on fast algorithms available for the Lasso estimator; e.g., [50]. The estimation methodology is attractive for applied research as the Lasso estimator also serves as a model selector and, hence, is relatively robust to misspecification.

We have derived convergence rates for a general two-step Lasso estimator which allows for the number of endogenous regressors and the number of instruments to be larger than the sample size. We then applied the two-step estimator to the spatial autoregressive panel model. Monte Carlo results confirm that the estimation method recovers the structure of the spatial weights matrix, even if T is as small as 50–100. However, our Monte Carlo results show a tendency for over-selection of spatial weights. The two-step post-Lasso estimator, which in each step applies OLS to the model selected by the Lasso, outperforms the two-step Lasso in terms of bias, false positive and false negative rate.

The use of the two-step Lasso raises several issues shared with other Lasso-type estimators. Controlling uncertainty and conducting inference in the Lasso is challenging and remains an area of ongoing research. Recent contributions include a Lasso significance tests due to Lockhart *et al.* [51] and the sample splitting approaches proposed by Wasserman and Roeder [52] and Meinshausen *et al.* [53] which allow for controlling the false discovery rate. Earlier seminal work on the asymptotic distribution of shrinkage estimators include Fan and Li [54] and Knight and Fu [55]. The former introduces the SCAD penalty. In addition, the choice of an optimal penalty level is an important issue. Penalized estimators typically select the penalty level oriented towards optimizing predictive performance, which may not be appropriate if the purpose is structure recovery. The optimal penalty used here is not based on cross-validation or other model selection criteria commonly employed and is therefore not directly subject to this criticism. Specifically, we follow Bickel *et al.* [27] and Belloni *et al.* [30] in choosing the smallest penalty level that dominates the noise of the problem. Our Monte Carlo results show that the proposed method works quite well in the structure discovery context.

This work suggests several lines of future research. First, given that the two-step post-Lasso outperforms the two-step Lasso, formal results for the two-step post-Lasso are required. Second, the methodology can be extended to the square-root Lasso and square-root post-Lasso. The main advantage of the square-root Lasso is that the optimal penalty level does not depend on the unknown error variance [56,57]. Hence, further performance improvements seem possible. Third, instead of relying on a two-step Lasso estimation method, an alternative estimation strategy may be based on the recent work by Fan and Liao [40] or Gautier and Tsybakov [38] who allow for endogeneity in high dimensions.

These one-step procedures potentially facilitate accounting for uncertainty in model selection and estimation. These ideas are retained for future work.

Acknowledgments

We thank Tapabrata Maiti and Mark Schaffer for helpful suggestions. The comments and constructive criticism by two anonymous referees helped us revise and improve the paper substantially. Their contributions are gratefully acknowledged. Achim Ahrens gratefully acknowledges the support of an ESRC postgraduate research scholarship.

Author Contributions

Achim Ahrens is the main author of the manuscript. Arnab Bhattacharjee contributed in formulation of the problem, writing and revision of the manuscript in advisory capacity.

A. Appendix

A.1. Proof of Theorem 1

Setting. First, we summarize the setting and introduce some notation. We can write the model in (2.1)–(2.2) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \mathbf{e}, \quad \mathbf{X} = \mathbf{Z}\boldsymbol{\Pi}^\star + \mathbf{U}$$

Thus, the reduced form equation for $\mathbf{y}$ is given by

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\Pi}^\star\boldsymbol{\beta}^\star + \mathbf{U}\boldsymbol{\beta}^\star + \mathbf{e}$$

In Assumption 2.3, we assume approximate sparsity.

$$\mathbf{y} = \mathbf{X}^\star\boldsymbol{\beta}^0 + \mathbf{X}^\star(\boldsymbol{\beta}^\star - \boldsymbol{\beta}^0) + \boldsymbol{\varepsilon} = \mathbf{X}^\star\boldsymbol{\beta}^0 + \mathbf{r} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}^\star = \mathbf{Z}\boldsymbol{\Pi}^\star$, $\boldsymbol{\varepsilon} = \mathbf{U}\boldsymbol{\beta}^\star + \mathbf{e}$, $\mathbf{r} = \mathbf{X}^\star(\boldsymbol{\beta}^\star - \boldsymbol{\beta}^0)$ with $1/\sqrt{T}\left\|\mathbf{r}\right\|_2 = R_{s_2}$ and $\boldsymbol{\beta}^0$ is the target parameter vector. As $\mathbf{X}^\star$ is unknown, we use $\hat{\mathbf{X}} = \mathbf{X}^\star + \hat{\mathbf{V}}$ in the second step.

$$\mathbf{y} = (\hat{\mathbf{X}} - \hat{\mathbf{V}})\boldsymbol{\beta}^0 + \mathbf{r} + \boldsymbol{\varepsilon} = \hat{\mathbf{X}}\boldsymbol{\beta}^0 - \hat{\mathbf{V}}\boldsymbol{\beta}^0 + \mathbf{r} + \boldsymbol{\varepsilon} = \hat{\mathbf{X}}\boldsymbol{\beta}^0 + \mathbf{r} + \mathbf{m} + \boldsymbol{\varepsilon}$$

where $\mathbf{m} := -\hat{\mathbf{V}}\boldsymbol{\beta}^0$ is the matrix of prediction errors from the first step weighted by the target parameter vector. Recall, the second-step Lasso estimator solves

$$\min \frac{1}{T}\left\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}\right\|_2^2 + \frac{\lambda_2}{T}\left\|\hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}\right\|_2$$

Let $Q(\boldsymbol{\beta}) = \frac{1}{T}\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2$ and $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$. Furthermore, define the active set $\Omega_2 = \text{supp}(\boldsymbol{\beta}^0)$ and $|\Omega_2| = s_2$.

The general approach in the following steps is based on Belloni *et al.* [30] and Bickel *et al.* [27], but accounts for the prediction error from the first step, $1/\sqrt{T}\left\|\mathbf{m}\right\|_2$.

Non-asymptotic ℓ_2-prediction norm bound. In this step, we bound $1/\sqrt{T}\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\|_2$ and treat $1/\sqrt{T}\left\|\mathbf{m}\right\|_2$ as given. The convergence rate of $1/\sqrt{T}\left\|\mathbf{m}\right\|_2$ will be derived in the next step.

By optimality of the Lasso estimate $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}) + \frac{\lambda_2}{T}\left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\beta}}\right\|_1 &\leq Q(\boldsymbol{\beta}^0) + \frac{\lambda_2}{T}\left\|\hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}^0\right\|_1 \\ Q(\hat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^0) &\leq \frac{\lambda_2}{T}\left(\left\|\hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}^0\right\|_1 - \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\beta}}\right\|_1\right) \end{aligned} \tag{A.1}$$

where

$$\begin{aligned} \left\|\hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}^0\right\|_1 - \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\beta}}\right\|_1 &= \left\|\hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}^0_{\Omega_2}\right\|_1 - \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\beta}}_{\Omega_2}\right\|_1 - \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\beta}}_{\bar{\Omega}_2}\right\|_1 \\ &\leq \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\Omega_2}\right\|_1 - \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\bar{\Omega}_2}\right\|_1 \end{aligned} \tag{A.2}$$

using $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^0_{\Omega_2}$, $\hat{\boldsymbol{\beta}}_{\bar{\Omega}_2} = \hat{\boldsymbol{\delta}}_{\bar{\Omega}_2}$ and $\left\|\hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}^0_{\Omega_2}\right\|_1 - \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\beta}}_{\Omega_2}\right\|_1 \leq \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\beta}}_{\Omega_2} - \hat{\boldsymbol{\Upsilon}}_2\boldsymbol{\beta}^0_{\Omega_2}\right\|_1$ by reverse triangle inequality. Futhermore,

$$TQ(\hat{\boldsymbol{\beta}}) - TQ(\boldsymbol{\beta}^0) = \left\|\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}\right\|_2^2 - \left\|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}^0\right\|_2^2 = -2\mathbf{y}'\hat{\mathbf{X}}\hat{\boldsymbol{\delta}} + \hat{\boldsymbol{\beta}}'\hat{\mathbf{X}}'\hat{\mathbf{X}}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{0\prime}\mathbf{X}'\hat{\mathbf{X}}\boldsymbol{\beta}^0$$

Substracting $\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2^2$ from both sides gives

$$\begin{aligned} Q(\hat{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}^0) - \frac{1}{T}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2^2 &= -\frac{2}{T}\boldsymbol{\varepsilon}'\hat{\mathbf{X}}\hat{\boldsymbol{\delta}} - \frac{2}{T}\mathbf{r}'\hat{\mathbf{X}}\hat{\boldsymbol{\delta}} - \frac{2}{T}\mathbf{m}'\hat{\mathbf{X}}\hat{\boldsymbol{\delta}} && \text{(A.3)} \\ &= -\frac{2}{T}\boldsymbol{\varepsilon}'\hat{\mathbf{X}}(\hat{\boldsymbol{\Upsilon}}_2)^{-1}\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}} - \frac{2}{T}\mathbf{r}'\hat{\mathbf{X}}\hat{\boldsymbol{\delta}} - \frac{2}{T}\mathbf{m}'\hat{\mathbf{X}}\hat{\boldsymbol{\delta}} \\ &=_{(i)} -\boldsymbol{S}_2'\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}} - 2R_{s_2}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 - \frac{2}{T}\|\mathbf{m}\|_2\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 \\ &\geq_{(ii)} -\|\boldsymbol{S}_2\|_\infty\left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}\right\|_1 - 2R_{s_2}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 - \frac{2}{T}\|\mathbf{m}\|_2\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 \\ &\geq_{(iii)} -\frac{\lambda_2}{cT}\left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}\right\|_1 - 2R_{s_2}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 - \frac{2}{T}\|\mathbf{m}\|_2\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 && \text{(A.4)} \end{aligned}$$

where (i) uses the Cauchy-Schwarz inequality and the definitions $R_{s_2} = \frac{1}{\sqrt{T}}\|\mathbf{r}\|_2$ and $\boldsymbol{S}_2 := \frac{2}{T}(\hat{\boldsymbol{\Upsilon}}_2)^{-1}\hat{\mathbf{X}}'\boldsymbol{\varepsilon}$. (ii) uses the Hölder inequality. (iii) uses $\lambda_2 \geq cT\|\boldsymbol{S}_2\|_\infty$ which holds as $T \to \infty$. Note that, by substituting for $\|\boldsymbol{S}_2\|_\infty$, we have eliminated the random component. Combining (A.1), (A.2) and (A.4) yields

$$\begin{aligned} \frac{1}{T}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2^2 &\leq 2R_{s_2}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 + \frac{2}{T}\|\mathbf{m}\|_2\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 + \frac{\lambda_2}{cT}\left(\left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\Omega_2}\right\|_1 + \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\bar{\Omega}_2}\right\|_1\right) \\ &\quad + \frac{\lambda_2}{T}\left(\left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\Omega_2}\right\|_1 - \left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\bar{\Omega}_2}\right\|_1\right) \\ &\leq 2R_{s_2}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 + \frac{2}{T}\|\mathbf{m}\|_2\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 + \left(1 + \frac{1}{c}\right)\frac{\lambda_2}{T}\left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\Omega_2}\right\|_1 \\ &\quad - \left(1 - \frac{1}{c}\right)\frac{\lambda_2}{T}\left\|\hat{\boldsymbol{\Upsilon}}_2\hat{\boldsymbol{\delta}}_{\bar{\Omega}_2}\right\|_1 \\ &\leq \left(2R_{s_2} + \frac{2}{\sqrt{T}}\|\mathbf{m}\|_2\right)\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 + \left(1 + \frac{1}{c}\right)\frac{\lambda_2}{T}u\left\|\boldsymbol{\Upsilon}^0_2\hat{\boldsymbol{\delta}}_{\Omega_2}\right\|_1 && \text{(A.5)} \\ &\quad - \left(1 - \frac{1}{c}\right)\frac{\lambda_2}{T}l\left\|\boldsymbol{\Upsilon}^0_2\hat{\boldsymbol{\delta}}_{\bar{\Omega}_2}\right\|_1 \end{aligned}$$

with $0 < l \leq 1 \leq u$. The last step assumes that $\hat{\boldsymbol{\Upsilon}}_2$ is asymptotically valid. Specifically, there are two constants u and l such that $l\boldsymbol{\Upsilon}_2^0 \leq \hat{\boldsymbol{\Upsilon}}_2 \leq u\boldsymbol{\Upsilon}_2^0$ where $l \to_{\mathrm{P}} 1$ and $u \to_{\mathrm{P}} \bar{u}$ with $\bar{u} \geq 1$ [30].

We distinguish between two cases. Case A: If $1/\sqrt{T}\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\|_2 \leq 2R_{s_2} + 2/\sqrt{T}\|\mathbf{m}\|_2$, the bound is established by assumption. Case B: If $1/\sqrt{T}\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\|_2 > 2R_{s_2} + 2/\sqrt{T}\|\mathbf{m}\|_2$, the above equation yields

$$\left\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\bar{\Omega}_2}\right\|_1 \leq c_0\left\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\Omega_2}\right\|_1$$

where $c_0 = u(c+1)/(l(c-1))$ which allows us to invoke the weighted restricted eigenvalue condition,

$$\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_2 \leq 2R_{s_2} + \frac{2}{\sqrt{T}}\|\mathbf{m}\|_2 + \left(1 + \frac{1}{c}\right)u\frac{\lambda_2}{T}\frac{\sqrt{s_2}}{\kappa_{c_0}^{\omega}(\hat{\mathbf{X}})} \tag{A.6}$$

This establishes the non-asymptotic ℓ_2-prediction norm bound, but takes the prediction error $1/\sqrt{T}\,\|\mathbf{m}\|_2$ from the first step as given. Note that if $\mathbf{m} = \mathbf{0}$, we arrive at the bound in Lemma 6 in Belloni *et al.* [30].

Convergence rate of $1/\sqrt{T}\|\mathbf{m}\|_2$. In this step, we derive the convergence rate for $1/\sqrt{T}\|\mathbf{m}\|_2$.

$$\begin{aligned}\|\mathbf{m}\|_2 = \left\|\hat{\mathbf{V}}\boldsymbol{\beta}^0\right\|_2 = \left\|\hat{\mathbf{V}}_{\Omega_2}\boldsymbol{\beta}^0_{\Omega_2}\right\|_2 \leq \left\|\hat{\mathbf{V}}_{\Omega_2}\right\|_F\|\boldsymbol{\beta}^0\|_2 &= \|\boldsymbol{\beta}^0\|_2\left(\sum_{j\in\Omega_2}\sum_i \hat{v}_{ij}^2\right)^{1/2}\\ \leq \|\boldsymbol{\beta}^0\|_2\sum_{j\in\Omega_2}\left(\sum_i \hat{v}_{ij}^2\right)^{1/2} &= \|\boldsymbol{\beta}^0\|_2\sum_{j\in\Omega_2}\left\|\hat{\mathbf{V}}_j\right\|_2\\ \leq \|\boldsymbol{\beta}^0\|_2\, s_2 \max_j\left\|\hat{\mathbf{V}}_j\right\|_2 &\end{aligned} \tag{A.7}$$

where $\hat{\mathbf{V}}_j = \mathbf{Z}\hat{\boldsymbol{\pi}}_j - \mathbf{Z}\boldsymbol{\pi}_j^\star$. By Theorem 1 in Belloni *et al.* [30],

$$\max_j \frac{1}{\sqrt{T}}\left\|\hat{\mathbf{V}}_j\right\|_2 \lesssim_{\mathrm{P}} \sqrt{\frac{s_1\log(\max(L\bar{p}, T))}{T}} \tag{A.8}$$

Substituting (A.7) into (A.8) and assuming $\|\boldsymbol{\beta}^0\|_2 \lesssim s_2$,

$$\frac{1}{\sqrt{T}}\|\mathbf{m}\|_2 = \frac{1}{\sqrt{T}}\left\|\hat{\mathbf{V}}\boldsymbol{\beta}^0\right\|_2 \lesssim_{\mathrm{P}} s_2^2\sqrt{\frac{s_1\log(\max(L\bar{p}, T))}{T}}$$

Convergence rate of ℓ_2-prediction norm bound. The non-asymptotic ℓ_2-prediction bound and the convergence rate for $1/\sqrt{T}\,\|\mathbf{m}\|_2$ allows us to derive the ℓ_2-prediction norm convergence rate. Note that $\lambda_2 \lesssim \sqrt{T\log(L\bar{p}/\alpha)}$ and $R_{s_2} \lesssim_{\mathrm{P}} \sqrt{s_2/T}$ by assumption. By (A.6) and substituting the convergence rate of $1/\sqrt{T}\|\hat{\mathbf{V}}\boldsymbol{\beta}^0\|_2$,

$$\begin{aligned}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\|_2 &\lesssim_{\mathrm{P}} \sqrt{\frac{s_2}{T}} + s_2^2\sqrt{\frac{s_1\log(\max(L\bar{p}, T))}{T}} + \sqrt{\frac{s_2\log(\max(L\bar{p}, T))}{T}}\\ &\lesssim_{\mathrm{P}} s_2^2\sqrt{\frac{s_1\log(\max(L\bar{p}, T))}{T}}\end{aligned}$$

However, we want to bound the deviations from $\hat{\mathbf{X}}\hat{\boldsymbol{\beta}}$ to $\mathbf{X}^\star\boldsymbol{\beta}^\star$. Hence, we apply the triangle inequality

$$\begin{aligned}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\beta}}-\mathbf{X}^\star\boldsymbol{\beta}^\star\right\|_2 &= \left\|(\hat{\mathbf{X}}\hat{\boldsymbol{\beta}}-\hat{\mathbf{X}}\boldsymbol{\beta}^0)+(\hat{\mathbf{X}}\boldsymbol{\beta}^0-\mathbf{X}^\star\boldsymbol{\beta}^0)+(\mathbf{X}^\star\boldsymbol{\beta}^0-\mathbf{X}^\star\boldsymbol{\beta}^\star)\right\|_2\\ &\le \frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\beta}}-\hat{\mathbf{X}}\boldsymbol{\beta}^0\right\|_2+\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{V}}\boldsymbol{\beta}^0\right\|_2+R_{s_2}\\ &\lesssim_{\mathrm{P}} s_2^2\sqrt{\frac{s_1\log(\max(L\bar{p},T))}{T}}\end{aligned}$$

Non-asymptotic ℓ_1-parameter norm bound. Again, we distinguish between two cases. Case A: $\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\overline{\Omega}_2}\|_1 \le 2c_0\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\Omega_2}\|_1$. Then, we can use the definition of the weighted restricted eigenvalue

$$\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 \le (1+2c_0)\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\Omega_2}\|_1 \le (1+2c_0)\frac{\sqrt{s_2}}{\kappa_{2c_0}^{\omega}(\hat{\mathbf{X}})\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2 \tag{A.9}$$

Case B: If $\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\overline{\Omega}_2}\|_1 > 2c_0\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\Omega_2}\|_1$, then by (A.5) $2R_{s_2}+2/\sqrt{T}\|\mathbf{m}\|_2 \ge \frac{1}{\sqrt{T}}\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\|_2$ must hold. Also, from (A.5)

$$\begin{aligned}\left\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\overline{\Omega}_2}\right\|_1 &\le \left(2R_{s_2}+\frac{2}{\sqrt{T}}\|\mathbf{m}\|_2-\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2\right)\frac{T}{\lambda_2}\frac{1}{\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2\left(\frac{c}{l(c-1)}\right)+c_0\left\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\Omega_2}\right\|_1\\ &\le \left(2R_{s_2}+\frac{2}{\sqrt{T}}\|\mathbf{m}\|_2\right)^2\frac{T}{\lambda_2}\left(\frac{c}{l(c-1)}\right)+\frac{1}{2}\left\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\overline{\Omega}_2}\right\|_1\\ &\le 2\left(2R_{s_2}+\frac{2}{\sqrt{T}}\|\mathbf{m}\|_2\right)^2\frac{T}{\lambda_2}\left(\frac{c}{l(c-1)}\right)\end{aligned}$$

where the second step uses $\max_{x\ge 0} x(2a-x)\le a^2$. In addition, by Case B assumption,

$$\begin{aligned}\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\Omega_2}\|_1 &< \frac{1}{2c_0}\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\overline{\Omega}_2}\|_1\\ \|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 &< \left(1+\frac{1}{2c_0}\right)\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}_{\overline{\Omega}_2}\|_1\\ \|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 &< \left(1+\frac{1}{2c_0}\right)2\left(2R_{s_2}+\frac{2}{\sqrt{T}}\|\mathbf{m}\|_2\right)^2\frac{T}{\lambda_2}\left(\frac{c}{l(c-1)}\right)\end{aligned} \tag{A.10}$$

Combining (A.9) and (A.10),

$$\begin{aligned}\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 &\le (1+2c_0)\frac{\sqrt{s_2}}{\kappa^{\omega}\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2+\left(1+\frac{1}{2c_0}\right)2\left(2R_{s_2}+\frac{2}{\sqrt{T}}\|\mathbf{m}\|_2\right)^2\frac{T}{\lambda_2}\left(\frac{c}{l(c-1)}\right)\\ \|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 &\le 3c_0\frac{\sqrt{s_2}}{\kappa^{\omega}\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2+3\frac{c_0T}{\lambda_2}\left(2R_{s_2}+\frac{2}{\sqrt{T}}\|\mathbf{m}\|_2\right)^2\\ \|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 &\le 3c_0\frac{\sqrt{s_2}}{\kappa^{\omega}\sqrt{T}}\left\|\hat{\mathbf{X}}\hat{\boldsymbol{\delta}}\right\|_2+3\frac{c_0T}{\lambda_2}\left(4R_{s_2}^2+\frac{4}{T}\|\mathbf{m}\|_2^2+8R_{s_2}\frac{1}{\sqrt{T}}\|\mathbf{m}\|_2\right)\end{aligned}$$

where we use that $c/(l(c-1))\le c_0$ and $1+1/(2c_0)\le 3/2$.

ℓ_1-parameter norm convergence rate. In the last step, we derive the ℓ_1-convergence rates. We assume, as stated in the Theorem, that s_1 and s_2 do not depend on T. This assumption may be strong in general,

but is reasonable in the spatial autoregressive panel model where s_1 and s_2 are determined by n and not by T.

$$\begin{aligned}\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 &\lesssim_{\mathrm{P}} \sqrt{s_1 s_2}s_2^2\sqrt{\frac{\log(\max(L\bar{p},T))}{T}} + \frac{s_2}{\sqrt{T\log\left(\max(L\bar{p},T)\right)}} + s_2^4 s_1\sqrt{\frac{\log(\max(L\bar{p},T))}{T}} \\ &\quad + s_2^2\sqrt{\frac{s_1 s_2}{T}} \\ \|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\|_1 &\lesssim_{\mathrm{P}} \sqrt{\frac{\log(\max(L\bar{p},T))}{T}}.\end{aligned}$$

Lastly,

$$\begin{aligned}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1 = \left\|(\boldsymbol{\Upsilon}_2^0)^{-1}\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\right\|_1 &\leq \|(\boldsymbol{\Upsilon}_2^0)^{-1}\|_\infty \left\|\boldsymbol{\Upsilon}_2^0\hat{\boldsymbol{\delta}}\right\|_1 \\ &\lesssim_{\mathrm{P}} \sqrt{\frac{\log(\max(L\bar{p},T))}{T}}\end{aligned}$$

A.2. Algorithm for Estimating Penalty Loadings

The algorithm is reproduced from Algorithm A.1 in Belloni *et al.* [30].

Algorithm 2. *Consider the model* $\mathrm{E}[y_t|\mathbf{x}_t] = \mathbf{x}_t'\boldsymbol{\beta}^0$ *for* $t = 1, \ldots, T$ *where* $\mathbf{x}_t$ *is a* p*-dimensional vector and* $\boldsymbol{\beta}^0$ *is the target value. The initial and refined penalty loadings are given by*

$$\text{initial:}\quad \hat{\gamma}_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_{tj}^2(y_t - \bar{y})^2} \qquad\qquad \text{refined:}\quad \hat{\gamma}_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_{tj}^2\hat{e}_t}$$

where $\bar{y} = T^{-1}\sum y_t$. *Specify the number of iterations* K. *Proceed as follows: (1) Obtain the Lasso or post-Lasso estimate* $\hat{\boldsymbol{\beta}}$ *using the initial penalty loadings and the optimal penalty level* λ. *(2) Obtain the Lasso or post-Lasso residuals* $\hat{e}_t = y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}$ *and update the Lasso or post-Lasso estimate* $\hat{\boldsymbol{\beta}}$ *using the refined penalty loadings. (3) Repeat the second step* K *times.*

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Arbia, G.; Fingleton, B. New spatial econometric techniques and applications in regional science. *Papers Reg. Sci.* **2008**, *87*, 311–317.
2. Harris, R.; Moffat, J.; Kravtsova, V. In search of 'W'. *Spat. Econ. Anal.* **2011**, *6*, 249–270.
3. Corrado, L.; Fingleton, B. Where is the economics in spatial econometrics? *J. Reg. Sci.* **2012**, *52*, 210–239.
4. Pinkse, J.; Slade, M.E.; Brett, C. Spatial price competition: A semiparametric approach. *Econometrica* **2002**, *70*, 1111–1153.

5. Bhattacharjee, A.; Jensen-Butler, C. Estimation of the spatial weights matrix under structural constraints. *Reg. Sci. Urban Econ.* **2013**, *43*, 617–634.
6. Beenstock, M.; Felsenstein, D. Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geogr. Anal.* **2012**, *44*, 386–397.
7. Bhattacharjee, A.; Holly, S. Structural interactions in spatial panels. *Empir. Econ.* **2011**, *40*, 69–94.
8. Bhattacharjee, A.; Holly, S. Understanding interactions in social networks and committees. *Spat. Econ. Anal.* **2013**, *8*, 23–53.
9. Bailey, N.; Holly, S.; Pesaran, M.H. A two stage approach to spatiotemporal analysis with strong and weak cross-sectional dependence. *J. Appl. Econome.* 2014, in press.
10. Huang, H.C.; Hsu, N.J.; Theobald, D.M.; Breidt, F.J. Spatial Lasso with applications to GIS model selection. *J. Comput. Graph. Statist.* **2010**, *19*, 963–983.
11. Wheeler, D.C. Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso. *Environ. Plan. A* **2009**, *41*, 722–742.
12. Seya, H.; Murakami, D.; Tsutsumi, M.; Yamagata, Y. Application of Lasso to the eigenvector selection problem in eigenvector-based spatial filtering. *Geogr. Anal.* **2014**, doi:10.1111/gean.12054.
13. Manresa, E. CEMFI: Madrid, Spain. Estimating the structure of social interactions using panel data. 2014, Unpublished work.
14. Souza, P.C. Department of Statistics, London School of Economics and Political Science, London, UK. Estimating networks: Lasso for spatial weights. 2012, Unpublished work.
15. Lam, C.; Souza, P.C. Department of Statistics, London School of Economics and Political Science, London, UK. Regularization for spatial panel time series using adaptive Lasso. 2013, Unpublished work.
16. Cliff, A.D.; Ord, J.K. *Spatial Autocorrelation: Monographs in Spatial and Environmental Systems Analysis*; Pion Ltd.: London, UK, 1973.
17. Anselin, L. *Spatial Econometrics: Methods and Models*; Springer: New York, NY, USA, 1988.
18. Kapoor, M.; Kelejian, H.H.; Prucha, I.R. Panel data models with spatially correlated error components. *J. Econom.* **2007**, *140*, 97–130.
19. Lee, L.F.; Yu, J. Some recent developments in spatial panel data models. *Reg. Sci. Urban Econ.* **2010**, *40*, 255–271.
20. Lee, L.F.; Yu, J. Estimation of spatial autoregressive panel data models with fixed effects. *J. Econom.* **2010**, *154*, 165–185.
21. Lee, L.F.; Yu, J. Efficient GMM estimation of spatial dynamic panel data models with fixed effects. *J. Econom.* **2014**, *180*, 174–197.
22. Mutl, J.; Pfaffermayr, M. The Hausman test in a Cliff and Ord panel model. *Econom. J.* **2011**, *14*, 48–76.
23. Elhorst, J. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*; Springer: New York, NY, USA, 2014.
24. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* **1996**, *58*, 267–288.

25. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.
26. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
27. Bickel, P.J.; Ritov, Y.; Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732.
28. Bühlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data*; Springer: New York, NY, USA, 2011.
29. Zhao, P.; Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2563.
30. Belloni, A.; Chen, D.; Chernozhukov, V.; Hansen, C. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **2012**, *80*, 2369–2429.
31. Belloni, A.; Chernozhukov, V. Least squares after model selection in high-dimensional sparse models. *Bernoulli* **2013**, *19*, 521–547.
32. Bunea, F.; Tsybakov, A.; Wegkamp, M. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **2007**, *1*, 169–194.
33. Wainwright, M.J. Sharp thresholds for high-dimensional and noisy sparsity recovery using L1-constrained quadratic programming. *IEEE Trans. Inf. Theor.* **2009**, *55*, 2183–2202.
34. Van de Geer, S. High-dimensional generalized linear models and the Lasso. *Ann. Stat.* **2008**, *36*, 614–645.
35. Meinshausen, N.; Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Stat.* **2009**, *37*, 246–270.
36. Pesaran, M.H. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **2006**, *74*, 967–1012.
37. Pesaran, M.H.; Tosetti, E. Large panels with common factors and spatial correlation. *J. Econom.* **2011**, *161*, 182–202.
38. Gautier, E.; Tsybakov, A.B. Toulouse School of Economics, Toulouse, France. High-dimensional instrumental variables regression and confidence sets. 2014, Unpublished work.
39. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2313–2351.
40. Fan, J.; Liao, Y. Endogeneity in high dimensions. *Ann. Stat.* **2014**, *42*, 872–917.
41. Caner, M. Lasso-type Gmm estimator. *Econom. Theor.* **2009**, *25*, 270–290.
42. Lin, W.; Feng, R.; Li, H. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *J. Am. Stat. Assoc.* **2014**, doi:10.1080/01621459.2014.908125.
43. Belloni, A.; Chernozhukov, V. High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation SE - 3*; Alquier, P., Gautier, E., Stoltz, G., Eds.; Springer: Berlin/ Heidelberg, Germany, 2011; pp. 121–156.
44. Jing, B.Y.; Shao, Q.M.; Wang, Q. Self-normalized Cramér-type large deviations for independent random variables. *Ann. Probab.* **2003**, *31*, 2167–2215.
45. Belloni, A.; Chernozhukov, V.; Hansen, C. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **2014**, *81*, 608–650.

46. Van de Geer, S.; Bühlmann, P. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **2009**, *3*, 1360–1392.
47. Raskutti, G.; Wainwright, M.J.; Yu, B. Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **2010**, *11*, 2241–2259.
48. Kelejian, H.H.; Prucha, I.R. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *J. Real Estate Financ. Econ.* **1998**, *17*, 99–121.
49. Kelejian, H.H.; Prucha, I.R. A generalized moments estimator for the autoregressive parameter in a spatial model. *Int. Econ. Rev.* **1999**, *40*, 509–533.
50. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
51. Lockhart, R.; Taylor, J.; Tibshirani, R.J.; Tibshirani, R. A significance test for the Lasso. *Ann. Stat.* **2014**, *42*, 413–468.
52. Wasserman, L.; Roeder, K. High-dimensional variable selection. *Ann. Statist.* **2009**, *37*, 2178–2201.
53. Meinshausen, N.; Meier, L.; Bühlmann, P. p-Values for high-dimensional regression. *J. Am. Statist. Assoc.* **2009**, *104*, 1671–1681.
54. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360.
55. Knight, K.; Fu, W. Asymptotics for Lasso-type estimators. *Ann. Stat.* **2000**, *28*, 1356–1378.
56. Belloni, A.; Chernozhukov, V.; Wang, L. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **2011**, *98*, 791–806.
57. Belloni, A.; Chernozhukov, V.; Wang, L. Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Stat.* **2014**, *42*, 757–788.