University of Kent

Kent Academic Repository

Cherodian, Rowan (2023) *Eigenvector Spatial Filtering and Lasso: Theory and Applications*. Doctor of Philosophy (PhD) thesis, University of Kent,.

Downloaded from

https://kar.kent.ac.uk/101045/ The University of Kent's Academic Repository KAR

The version of record is available from

This document version

UNSPECIFIED

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Eigenvector Spatial Filtering and Lasso: Theory and Applications

A Thesis Submitted in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

by

Rowan Cherodian

Supervised by Dr Sylvain Barde and Dr Guy Tchuente

School of Economics

University of Kent

April 2023

©Rowan Cherodian 2023

Declaration

I hereby declare that this thesis is my own work, as are all errors and omissions, with the exception of chapter 3 which is co-authored work with Dr. Sylvain Barde and Dr. Guy Tchuente. The copyright of this thesis rests with the author(s). This thesis consists of 178 pages excluding references and contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree.

Abstract

This thesis focuses on Eigenvector Spatial Filtering (ESF), developed by Griffith (2000, 2003) as a methodology designed to handle general cross-sectional/spatial dependence. ESF uses a subset of eigenvectors from a spatial weights matrix in a linear regression framework to approximate/control for any spatially correlated terms in the underlying data-generating process. Thus, ESF has the key advantage that it does not require the researcher to specify which parts of the model are spatially correlated. This advantage is the main driver behind ESF's recent increasing popularity with applied economists. I extend the theory around ESF and its Lasso (Tibshirani, 1996) implementation for the cases when the structural equation being studied includes and excludes endogenous variables, and demonstrate how the method can be applied in several empirical applications.

This thesis is comprised of four chapters, the first is a literature review, the second and third are in econometrics, and the fourth is in environmental economics. The econometrics chapters propose Moran's i based Lasso procedures for estimating exogenous and endogenous right-hand-side regression parameters when the data is spatially dependent. The last chapter of this thesis uses the procedure proposed in

the second chapter to account for spatial dependence when testing for the presence of the environmental Kuznets curve for forests.

The first chapter provides a review of some common spatial economic models and how they are conventionally estimated, an overview of Kojevnikov et al. (2021) limit theorems for cross-sectionally dependent random variables (used in Chapter 3), and a summary of penalised regressions with a focus on Lasso.

The second chapter, entitled "Moran's i Lasso: for spatially correlated models" provides a theoretical contribution. After an extensive evaluation of existing procedures to select the relevant subset of eigenvectors for ESF, I develop a new selection method called Moran's i Lasso (Mi-Lasso). The procedure uses information about the overall level of spatial dependence present in the underlying data-generating process, contained in the Moran's i, to determine a point estimate for the Lasso tuning parameter. I derive performance bounds and show the necessary conditions for consistent eigenvector selection. The key advantages of the proposed estimator are that it is intuitive and substantially faster than Lasso based on cross-validation or any proposed forward stepwise procedure. Our main simulation results show the proposed selection procedure performs well in finite samples and an application on house prices. Compared to existing selection procedures, I find, Mi-Lasso has one of the smallest biases and mean squared errors across a range of sample sizes and levels of spatial correlation. Additionally, through an evaluation of the properties of the spectral decomposition, I note that ESF can also handle higher-order spatial lags, which is confirmed in a simulation experiment.

The third chapter, entitled "Moran's i 2-Stage Lasso: for spatial models with

endogenous variables" also provides a theoretical contribution, is co-authored work with Dr. Sylvain Barde and Dr. Guy Tchuente. It proposes a new way of estimating a spatial model that includes endogenous variables when the researchers' main concerns are estimating only the direct effect and/or misspecification of the spatial weights matrix and spatial model. The proposed procedure uses Mi-Lasso to select the first and second-stage relevant eigenvectors and then uses the union of selected eigenvectors as controls in a two-stage least squares regression. The procedure is called Moran's i 2-Stage Lasso (Mi-2SL). We show the conditions necessary for consistent and asymptotically normal parameter estimation assuming the support (relevant) set of eigenvectors is known. Our Monte Carlo simulation results also show that Mi-2SL performs well when the spatial weights matrix has a high degree of misspecified links. Our empirical application replicates Cadena and Kovak (2016) instrumental variables estimates using Mi-2SL and shows that Mi-2SL can boost the performance of the first stage.

Finally, the fourth chapter, entitled "The Environmental Kuznets Curve for forests: an application of Mi-Lasso" is an application of Mi-Lasso to a hotly debated question in environmental economics. Does the relationship between a country's economic development (proxied by per capita GDP) and its deforestation rate follow the inverse U-shaped curve postulated by the classic environmental Kuznets curve for forests? I use the Mi-Lasso methodology proposed in chapter 2 to account for spatial dependence of an unknown functional form when testing for the presence of the environmental Kuznets curve for forests. I find evidence of a non-linear relationship, which in some cases is a more complicated predicted inverse U-shaped

curve, the average peak rate of deforestation appears to be falling with time, while the income required for the deforestation rate to start falling is increasing with time. Additionally, I find, that if the spatial dependence is not accounted for, the OLS estimates of income exhibit an absolute upward bias.

Acknowledgements

Firstly, I am indebted to Dr. Guy Tchuente for spotting my interest in the field of econometrics during the micro-econometrics training course in the first year of this degree and for inspiring/encouraging me to make econometrics my field of research. Without Dr. Tchuente's guidance, patience, mentorship, extensive support, and time this thesis would not be possible. Secondly, I am also very grateful to Dr. Sylvain Barde for the comprehensive support and advice he has provided me throughout this thesis. Thirdly, I would also like to express my gratitude to Prof Dr. Tony Thirlwall who inspired and encouraged me to publish my first research paper and start this degree. Fourthly, I am also very grateful to Prof Dr. Iain Fraser for his help and guidance with chapter 4.

I would like to express my deepest admiration for the unwavering support offered to me by the former director of the Economics PhD program at the University of Kent, the late Prof Dr. Miltos Makris. When I decided to make the switch to the field of econometrics he was extremely supportive and generously helped me secure funding for additional training.

I would also like to thank my parents, sister, and countless friends who have

provided me with extensive personal support over the many years I have been working on this thesis. Finally, I would like to thank every person who I have met along this journey such as (this list is not exhaustive) my Ph.D. colleagues, academics within the School of Economics, the Ph.D. students and academics I met while at Aix-Marseille University or any of the conferences I have attended, and the many individuals who have hosted me around the world while I produced this thesis.

Contents

List of Figures

List of Tables

Notation

I adopt the following notation and conventions: let $\boldsymbol{H}$ be some matrix; I then denote the (i,j)th element of $\boldsymbol{H}$ as h_{ij}. Similarly, if $\boldsymbol{b}$ is a vector, then b_i denotes the ith element of $\boldsymbol{b}$. $\boldsymbol{I}$ is an $n \times n$ identity matrix and $tr(\cdot)$ is the trace operator. If $\boldsymbol{H}$ is a square matrix, then $\boldsymbol{H}^{-1}$ denotes the inverse of $\boldsymbol{H}$. If $\boldsymbol{b} \in \mathbb{R}^m$ is a vector I define the ℓ_q norm of $\boldsymbol{b}$ as $||\boldsymbol{b}||_q = (\sum_{m=1}^{M} |b_m|^q)^{1/q}$ for $q \in \mathbb{N}$. $||\boldsymbol{b}||_\infty$ is the largest element in $\boldsymbol{b}$ and $||\boldsymbol{b}||_0$ is the number of non-zero elements in $\boldsymbol{b}$. The support operator is $\text{supp}(\boldsymbol{b}) = \{k \in \{1, \ldots, K\} : b_k \neq 0\}$. Let A be some set then $\grave{A}$ is the complement of A, $\boldsymbol{b}_A$ is a vector with elements $b_k \mathbb{1}\{k \in A\}$ and $\boldsymbol{H}_A$ is a matrix with elements $h_{ij}\mathbb{1}\{j \in A\}$ where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

Acronyms

BSS	Best subset selection
CV	Cross-validation
DGP	Data-generating process
ESF	Eigenvector Spatial Filtering
EKC	Environmental Kuznets Curve (for all environmental indicators)
EKCf	Environmental Kuznets Curve for forests
GMM	Generalised Methods of Moments
GS2SLS	Generalised spatial two stage least squares
GDP	Gross domestic product
i.i.d.	Independent and identically distributed
IV	Instrumental variables
IC	Irrepresentable condition
LLN	Law of large number
Lasso	Least absolute shrinkage and selection operator
LOWESS	Locally Weighted Scatterplot Smoothing (Cleveland and Devlin, 1988)
MESS	Matrix exponential spatial specification
ML	Maximum Likelihood
MSE	Mean squared error
MSPE	Mean squared prediction error
Mi-2SL	Moran's i 2 stage Lasso
Mi-Lasso	Moran's i Lasso
OLS	Ordinary Least Squares
SAR(p)	Pth-order spatial autoregressive model
pLasso	Post Lasso
PCNM	Principal Coordinate of Neighbour Matrices (Borcard and Legendre, 2002)

QML	Quasi-Maximum Likelihood
RE	Restricted eigenvalue condition
SARAR	Spatial autoregressive model with autoregressive error disturbance
SDM	Spatial Durbin model
SDMAR	Spatial Durbin model with autoregressive error disturbance
SEM	Spatial error model
SLX	Spatial lag of X model
SWM	Spatial weights matrix
2SLS	Two-stage least squares

Chapter 1

Literature Review

1.1 What is and why do we need spatial econometrics

Spatial correlation or (weak) cross-sectional dependence means there is stochastic dependence between the cross-sectional observations,[1] thus, the standard (cross-sectional) assumption that the observations are independent is violated. If a classical linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ is estimated by Ordinary Least Squares (OLS), the standard errors and/or parameter estimates will be biased and inconsistent. Cross-sectional dependence can be expressed by following the moment condition (Anselin, 2001)

$$cov[y_i, y_j] = \mathbb{E}[y_i \cdot y_j] - \mathbb{E}[y_i]\mathbb{E}[y_j] \neq 0, \; for \; some \; i \neq j,$$

[1] We will use the term spatial correlation or cross-sectional dependence interchangeably.

where subscripts i and j refer to individual observations, y_i is the value of the random variable of interest. This equation is of interest in a spatial context when configurations of nonzero i, j pairs can be interpreted as spatial interactions, spatial structure, or spatial arrangement of observations.

Spatial models are designed to account for the role of spatial interactions in the variables being studied. The basic principle of these models stems from Tobler's first law of geography "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Spatial spillovers are often observed between various types of units and their characteristics. Some examples include pollution, local tax rates, local public spending, and migration. Other more obscure examples also exist, such as citizens demands of their government may be affected by neighbouring countries' political conditions. In many applications, space is based on distance. However, it can also be based on other factors such as differences in city size, population, net migration, production, or political system (Kelejian and Piras, 2017).[2]

1.1.1 Spatial Data Generating Processes

There are many spatial data-generating processes (DGP) proposed in the literature. The most commonly used spatial economic model is the first-order spatial autoregressive model (SAR(1)), a linear model which includes a spatial lag of the dependent variable is

[2]The following discussion on spatial modeling and estimation is based on Ahrens (2017) and (Kelejian and Piras, 2017).

$$\boldsymbol{y} = \rho_0 \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{v} \tag{1.1}$$

$$\boldsymbol{y} = (\boldsymbol{I} - \rho_0 \boldsymbol{W})^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{v}) \tag{1.2}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)'$ is a $n \times 1$ vector containing the dependent variable, $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)$ is a $n \times k$ matrix of the k covariates (with full column rank), $\boldsymbol{v} = (v_1, \cdots, v_n)'$ is a $n \times 1$ vector of identically and independently distributed (i.i.d.) disturbances with the moment condition $\mathbb{E}[\boldsymbol{v}|\boldsymbol{X}] = 0$ and $\mathbb{E}[\boldsymbol{v}\boldsymbol{v}'|\boldsymbol{X}] = \sigma^2\boldsymbol{I}$, $\boldsymbol{\beta}_0$ is an $k \times 1$ vector of parameters, and ρ_0 is a scalar parameter that describes the degree of spatial correlation in the dependent variables. The $n \times n$ spatial weights matrix (SWM) $\boldsymbol{W}$ describes the spatial or socio-economic relationship between the cross-sectional units, $w_{ij} \neq 0$ implies a meaningful interaction of units j on unit i. In such cases, unit j is often referred to as a neighbour of unit i. These interactions can stem from various sources, such as spillovers, externalities, geographic location, regulations, technology, government policy, or government expenditure. The diagonal w_{ii} is set equal to zero as no unit can affect itself. $\boldsymbol{W}\boldsymbol{y}$ is typically referred to as the spatial lag of $\boldsymbol{y}$.

The reduced form of the SAR model is (1.2) which assumes/requires the matrix $(\boldsymbol{I} - \rho_0 \boldsymbol{W})$ is non-singular and reveals the outcome of the spatial unit i (y_i) is affected directly by their explanatory variables and indirectly by the explanatory variables of all other spatial units. Lee (2002) showed

$$\mathbb{E}[(\boldsymbol{W}\boldsymbol{y})'\boldsymbol{v}] = \sigma^2 tr(\boldsymbol{W}(\boldsymbol{I} - \rho_0\boldsymbol{W})^{-1}).$$

When $\rho \neq 0$, the term $tr(\boldsymbol{W}(\boldsymbol{I} - \rho\boldsymbol{W})^{-1})$ in generally non-zero, in such cases, the spatial lag is endogenous. Thus, if (1.1) is estimated by OLS the estimates will generally be inconsistent, unless the number of spatial units is not fixed (Lee, 2002). The intuition is that y_i can affect y_j and y_j can affect y_i. This phenomenon is commonly referred to as reverse causality or simultaneity.

The second spatial DGP we will consider is the spatial error model (SEM),[3] shown in (1.3) to (1.6),

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{r}, , \tag{1.3}$$

$$\boldsymbol{r} = \delta_0\boldsymbol{W}\boldsymbol{r} + \boldsymbol{v} \tag{1.4}$$

$$\boldsymbol{r} = (\boldsymbol{I} - \delta_0\boldsymbol{W})^{-1}\boldsymbol{v}, \tag{1.5}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + (\boldsymbol{I} - \delta_0\boldsymbol{W})^{-1}\boldsymbol{v}. \tag{1.6}$$

Where the scalar parameter δ_0 describes the degree of spatial correlation in the error term and $(\boldsymbol{I} - \delta_0\boldsymbol{W})$ is assumed non-singular. The spatial effects in this model stem from unobservable shocks across units. If variables in $\boldsymbol{X}$ are weakly exogenous, the parameter vector β_0 can be estimated consistently by OLS.

[3] A less common spatial error process is the spatial moving average process proposed by Haining (1978).

The properties of the error term $\boldsymbol{r}$ are as follows:

$$\begin{aligned}\mathbb{E}[\boldsymbol{r}] &= (\boldsymbol{I} - \delta_0 \boldsymbol{W})^{-1}\, \mathbb{E}[\boldsymbol{v}] = 0 \\ \mathbb{E}[\boldsymbol{r}\boldsymbol{r}'] &= (\boldsymbol{I} - \delta_0 \boldsymbol{W})^{-1}\, \mathbb{E}[\boldsymbol{v}\boldsymbol{v}'](\boldsymbol{I} - \delta_0 \boldsymbol{W})^{-1} \\ &= (\boldsymbol{I} - \delta_0 \boldsymbol{W})^{-1} \sigma_{\boldsymbol{v}}^2 (\boldsymbol{I} - \delta_0 \boldsymbol{W})^{-1} \neq \sigma_{\boldsymbol{v}}^2 \boldsymbol{I}.\end{aligned} \tag{1.7}$$

The error term $\boldsymbol{r}$ is spatially dependent, has a mean of zero, and is heteroskedastic. Making OLS estimates inefficient (Kelejian and Piras, 2017).

The third spatial process we will consider is the spatial lag of $\boldsymbol{X}$ model (SLX) shown in (1.8) (Vega and Elhorst, 2015).

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\psi}_0 + \boldsymbol{v}, \tag{1.8}$$

where the $k \times 1$ parameter vector $\boldsymbol{\psi}_0$ describes the degree of spatial correlation in each of the k exogenous variables. $\boldsymbol{W}\boldsymbol{X}$ are spatial lags of exogenous variables, sometimes referred to as 'Durbin regressors'. The model can include Durbin regressors to capture local spillovers (externalities) in the exogenous variables (Anselin, 2003). In this case, the OLS estimates of (1.8) are consistent (as long as the columns of $\boldsymbol{W}\boldsymbol{X}$ are linearly independent of $\boldsymbol{X}$), however omitting Durbin regressors will yield inconsistent estimates.[4] An empirical example of where the SLX model is estimated is Bloom et al. (2013).

It is common in spatial economic modelling to combine these three spatial pro-

[4]If $\boldsymbol{W}$ is row-normalised, $\boldsymbol{X}$ must not contain an intercept, or the model will suffer from multicollinearity.

cesses. When the SAR model is combined with the SLX model, we will refer to this as the spatial Durbin model (SDM). When the SAR is combined with the SEM, we will refer to this as the spatial autoregressive model with autoregressive error disturbance (SARAR) model. The final, the most general model we consider, combines all three spatial processes and is referred to as the spatial Durbin model with autoregressive error disturbance (SDMAR). All spatial processes we will consider are nested in the SDMAR model and can be recovered by setting the relevant spatial parameter(s) to zero. I assume the weights matrix is the same for the lag of $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{u}$, which does not have to be the case.

There is a close relationship between spatial models and time-series models. For example, the time-series autoregressive model (of order 1) is a special case of the SAR (of order 1) (1.1) models as the time-series autoregressive model can be recovered by setting just the lower sub-diagonal of $\boldsymbol{W}$ is equal to 1 in the SAR, i.e.

$$\boldsymbol{W} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & & \\ 0 & 1 & 0 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \\ 0 & & & 1 & 0 \end{bmatrix}. \tag{1.9}$$

However, unlike in the time-series case, spatial data has no natural ordering, and the elements of the weights matrix are not necessarily binary. It is important to note that spatial models are generally more complicated than time-series models as all of the off-diagonal elements of $\boldsymbol{W}$ can theoretically be non-zero, unlike (1.9).

One noteworthy exception when the SAR can be estimated consistently by OLS is when $\boldsymbol{W}$ is upper triangular, i.e., $w_{ij} = 0 \; \forall \; i \geq j$ as in this case $tr(\boldsymbol{W}(\boldsymbol{I}-\rho_0\boldsymbol{W})^{-1}) = 0$, thus,[5] the triangular structure means the effects are only in one direction, so there is no reverse causality. This case is theoretically interesting but rarely used in spatial applications.

1.1.2 Spatial Weights Matrix

The concept of the $n \times n$ SWM $\boldsymbol{W}$ was introduced by Moran (1948). In most empirical applications, the SWM is specified *a-priori* using observable variables. Some common examples are; if j is a contiguous neighbour of i, the researcher may use $w_{ij} = 1$ and $w_{ij} = 0$ otherwise. Alternatively, the researcher can use the distance between the units d_{ij}, in line with the Tobler first law of geography (Tobler, 1970) d_{ij} is discounted, so closer units get higher values. One typical example is inverse distance $w_{ij} = 1/d_{ij}$; other specifications not related to physical distance may also be used, for example, net migration between units or differences in investment.

Non-distance-related specifications are interesting but possess a key disadvantage w_{ij} may not be bounded, as the difference between two units may be arbitrarily close (Kelejian and Piras, 2017). As a standard assumption in spatial econometrics is the elements of the SWM are bounded, this is a serious issue. Physical distance also has several other advantages; it is plausibly exogenous (other factors are often arguably endogenous to the variables being studied). It also gives a symmetric weights matrix,

[5]When $\boldsymbol{W}$ is upper triangular, $\boldsymbol{I} - \rho_0\boldsymbol{W}$ has ones on the diagonal and is an upper triangular matrix, along with $(\boldsymbol{I} - \rho_0\boldsymbol{W})^{-1}$. Given $\boldsymbol{W}$ and $(\boldsymbol{I} - \rho_0\boldsymbol{W})^{-1}$ are both upper triangular matrices, the product of these two matrices have zeros on the diagonal, hence $tr(\boldsymbol{W}(\boldsymbol{I} - \rho_0\boldsymbol{W})^{-1}) = 0$.

which has beneficial mathematical properties such as real and mutually orthogonal eigenvectors, a feature that ESF takes advantage of.

It is difficult to estimate a SWM if the spatial connections of $\boldsymbol{W}$ are unknown. To illustrate this, take the SAR when $\rho_0 = 1$ and write the model as a system of simultaneous equations;

$$
\begin{array}{rcccccccccccccc}
y_1 & = & 0 & + & w_{12}y_2 & + & w_{13}y_3 & + & \cdots & + & w_{1n}y_n & + & \boldsymbol{x}_1'\boldsymbol{\beta} & + & v_1 \\
y_2 & = & w_{21}y_1 & + & 0 & + & w_{23}y_3 & + & \cdots & + & w_{2n}y_n & + & \boldsymbol{x}_2'\boldsymbol{\beta} & + & v_2 \\
 & \vdots & & & & & & & & & & & & & \\
y_n & = & w_{n1}y_1 & + & w_{n2}y_2 & + & \cdots & + & w_{n(n-1)}y_{(n-1)} & + & 0 & + & \boldsymbol{x}_n'\boldsymbol{\beta} & + & v_n.
\end{array}
$$

If we tried to estimate the w_{ij}'s as coefficients $\forall\ i \neq j$ we would be trying to estimate $(n-1)n + k$ parameters with only n observations. Thus, if the SWM is treated as unknown, the model can not be identified (Bhattacharjee and Jensen-Butler, 2013). This result also holds for other spatial models.

It is also common to row-normalise the SWM before estimations. In such cases, the spatial parameter ρ captures the expected change in the outcome of a unit if all other units' outcomes change by one unit. Row-normalisation is based on the implicit assumption that the strength of interaction effects is constant across units. This assumption may not be valid in some situations. Another possible issue with row-normalisation is there generally does not exists a single re-scaling factor for the spatial autoregressive parameters that will yield a specification equivalent to the specification with the un-normalised weights matrix (Kelejian and Prucha, 2010). Thus, Kelejian and Prucha (2010) suggest normalisation by a scalar factor $c_n = c$

(which clearly can vary with n) i.e. $\rho_0 \boldsymbol{W} = (\rho_0^* c)(\boldsymbol{W}^*/c)$ where $\boldsymbol{W}^*$ is the original SWM. Using a single scalar allows for the recovery of the original autoregressive parameters (Kelejian and Prucha, 2010) and maintains symmetry. Some examples of normalisation scalar factors are dividing by the maximal of the row or column sum or the largest eigenvalue.

Lee (2004) showed that the eigenvalues of the matrices $(\boldsymbol{I} - \rho_0 \boldsymbol{W})$ and $(\boldsymbol{I} - \rho_0 \boldsymbol{W})^{-1}$ should lie within the unit circle. A normalised symmetric SWM ensures this is satisfied. Alternatively, one of the following conditions can also be satisfied (a) the row and column sum of $\boldsymbol{W}$ and $(\boldsymbol{I} - \rho_0 \boldsymbol{W})^{-1}$ should be bound in absolute value as $n \to \infty$ (Kelejian and Prucha, 1998, 1999) or (b) the row and column sum of $\boldsymbol{W}$ must not diverge at a rate equal to or faster than the rate of the sample size n (Lee, 2004). These conditions place a limit on the cross-sectional dependence. They ensure the correlation between two spatial units converges to zero as the distance between them increases to infinity.

1.1.3 Estimating a Spatial Model

An important feature of spatial models is the concept of triangular arrays, the general idea is the sequence of observation of the dependent variable and other right-hand-side variables must be indexed by both order and sample size. Take for example the SDMAR the solution to this model includes the $n \times n$ matrices $(\boldsymbol{I} - \rho_0 \boldsymbol{W})^{-1}$ and $(\boldsymbol{I} - \delta_0 \boldsymbol{W})^{-1}$ which is dependent on the sample size n, this implies that changing the sample size will change the vector $\boldsymbol{y}$, even if the model remains unchanged. This is important as much inference in many spatial models depends on results from large

sample theory. For this reason variables in spatial models are generally indexed by the sample size, we acknowledge this fact but suppress the n indexing for ease of notation and instead remind the reader if a variable or parameter is dependent on n when relevant.

To estimate a spatial model the researcher needs to explicitly specify an SWM and the functional form of the spatial process (which parts of the model are spatially correlated) like those discussed in Section 1.1.1. Once the spatial model has been specified, the model needs to be estimated. If the model includes a spatial lag of the dependent variable then there is endogeneity in the model and it needs to be accounted for. The two most common estimation techniques for such models are maximum likelihood (ML) and Generalised Methods of Moments (GMM).

1.1.3.1 Maximum Likelihood

The first method for estimating a spatial model of the type described in Section 1.1.1 was Maximum Likelihood (ML) by Ord (1975). The spatial ML estimator hinges on the assumption of normally distributed and *i.i.d.* innovations. When the innovations are non-Gaussian, the estimator is called the Quasi-Maximum Likelihood (QML) estimator.

For ML, the log-likelihood is complicated, as the observations are not independent, it is not just a sum of individual observations. Consider the SAR(1) model, the estimates of $\boldsymbol{\beta}$, ρ and σ^2 are obtained by maximising the likelihood function $L(\boldsymbol{\beta}, \rho, \sigma^2 | \boldsymbol{y}, \boldsymbol{X})$. Assuming $\boldsymbol{v} = (\boldsymbol{I} - \rho_0 \boldsymbol{W})\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0$ is *i.i.d.* and normally dis-

tributed, taking logs yields the following log-likelihood function;

$$\ln L(\boldsymbol{\beta}, \rho, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln|\boldsymbol{I} - \rho\boldsymbol{W}| - \frac{\boldsymbol{v}'\boldsymbol{v}}{2\sigma^2},$$

where $|\cdot|$ is the determinant operator. The Log-likelihood contains the Jacobian term $\ln|\boldsymbol{I} - \rho\boldsymbol{W}|$ (the determinant of an $n \times n$ matrix). This is difficult to handle computationally, especially when the sample size is large and often has no closed-form solution (Griffith, 2000). This problem becomes more acute if the model includes a spatial lag of the dependent variable and disturbance term ($\rho \neq 0$ and $\delta \neq 0$) as $\ln|\boldsymbol{I} - \rho\boldsymbol{W}|$ and $\ln|\boldsymbol{I} - \delta\boldsymbol{W}|$ (the determinant of two $n \times n$ matrices) both appear in the log-likelihood function. Several methods have been developed to deal with this computation issue, such as Ord (1975) suggested evaluating the determinant in terms of its characteristic roots

$$\ln|\boldsymbol{I} - \delta\boldsymbol{W}| = \sum_{i=1}^{n}\ln(1 - \delta\lambda_i).$$

Lee (2004) gives the conditions that ensure asymptotic normality and consistency of the QML estimator for the general SDMAR model. Violation of the *i.i.d.* error assumption results in QML being inconsistent. Arraiz et al. (2010) demonstrated through extensive Monte Carlo simulations that QML is inconsistent in the presence of heteroskedasticity.

The problem with this method is computational accuracy (Kelejian and Prucha, 1999). Several other methods which approximate the log Jacobian term have also been proposed; simulations based approximation (Barry and Pace, 1999), character-

istic polynomial (Smirnov and Anselin, 2001), and Chebyshev approximation (Pace and LeSage, 2004).

In response to the computational issues described above for estimating the SAR model LeSage and Pace (2007) proposed the matrix exponential spatial specification (MESS)

$$e^{\alpha \boldsymbol{W}} \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{v}, \tag{1.10}$$

where the scalar α captures the spatial dependence and the matrix exponential $e^{\alpha \boldsymbol{W}} = \sum_{i=0}^{\infty} \frac{\alpha^i}{i!} \boldsymbol{W}^i$. Given the inverse of $e^{\alpha \boldsymbol{W}}$, $e^{-\alpha \boldsymbol{W}}$ always exists (Chiu et al., 1996), the reduced form of the MESS model always exists. Thus no constraints are required on the parameter space of α. The $e^{\alpha \boldsymbol{W}}$ in the MESS replaces the $(\boldsymbol{I} - \rho \boldsymbol{W})$ in the SAR model. The conventional geometric decay pattern of influence over space in the SAR is replaced with an exponential decay pattern in the MESS. As the likelihood function of the MESS does not include the determinant of the Jacobian transformation matrix, estimating MESS by QML is substantially easier than the SAR, especially when higher order spatial lags are included (LeSage and Pace, 2007).[6] Debarsy et al. (2015) find the MESS estimated by QML may also be robust to unknown heteroskedasticity when disturbances are not spatially correlated or follow a similar MESS process.

[6]For higher order SAR models estimated by QML, it is challenging to impose tractable parameter spaces (Elhorst et al., 2012).

1.1.3.2 Generalised Methods of Moments

The generalised method of moments (GMM) based estimator generalised spatial two-stage least squares (GS2SLS) was proposed by Kelejian and Prucha (1998, 1999) and is the most popular estimation procedure economists use to estimate spatial economic models. As highlighted in the previous section, ML is computationally demanding especially when the n is large. Additionally ML requires strong distributional assumptions over the structural errors. Kelejian and Prucha (1998, 1999) developed GS2SLS under the assumption of homoskedastic disturbances for the SARAR model. Kelejian and Prucha (2010); Arraiz et al. (2010) extended the estimation method to allow for heteroskedasticity, arguing that the homoskedasticity assumption is often inappropriate for spatial data. As the size and other characteristics of spatial units are often heterogeneous. GS2SLS can also be easily extended to deal with endogenous regressors (Fingleton and Le Gallo, 2008). The key advantages of GS2SLS over ML are; (a) computationally less demanding; (b) can allow for heteroskedasticity; (c) allows for the inclusion of endogenous regressors; (d) does not require distributional assumptions of the disturbance term.

To demonstrate how the procedure works, I will use the SARAR model.[7]

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r} = \boldsymbol{Z}\boldsymbol{\zeta} + \boldsymbol{r}, \tag{1.11}$$

$$\boldsymbol{r} = \delta \boldsymbol{W}\boldsymbol{r} + \boldsymbol{v}, \tag{1.12}$$

where $\boldsymbol{Z} = [\boldsymbol{X}, \boldsymbol{W}\boldsymbol{y}]$ and $\boldsymbol{\zeta} = (\boldsymbol{\beta}', \rho)'$. Assuming $\mathbb{E}[v_i] = 0$, $\mathbb{E}[v_i^2] = \sigma_i^2 \ \forall \ i$

[7]For the full set of model assumptions, please see Kelejian and Prucha (2010); Arraiz et al. (2010).

and $\mathbb{E}[v_i v_j] = 0 \; \forall \; i \neq j$. GMM-based techniques allow for heteroskedasticity and non-Gaussian errors but still assume the observations are independent. Estimation requires an instrument matrix $\boldsymbol{H}$, which must satisfy two necessary conditions, the instruments (1) are orthogonal to the error term (the exclusion restriction), and (2) must be correlated with the endogenous regressor (the relevance condition). Kelejian and Prucha (1998) assume $|\rho| < 1$ holds so they can use the standard expansion of Debreu and Herstein (1953) to rewrite the spatial operator $(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}$ as the limit of the infinite sum,[8]

$$(\boldsymbol{I} - \rho \boldsymbol{W})^{-1} = \rho^0 \boldsymbol{W}^0 + \rho^1 \boldsymbol{W}^1 + \rho^2 \boldsymbol{W}^2 + \ldots. \tag{1.13}$$

Allowing them to rewrite the conditional expected value of $\boldsymbol{y}$ as

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = (\boldsymbol{I} + \rho^1 \boldsymbol{W}^1 + \rho^2 \boldsymbol{W}^2 + \ldots)\boldsymbol{X}\boldsymbol{\beta}.$$

This suggests

$$\boldsymbol{X}, \; \boldsymbol{W}\boldsymbol{X}, \; \boldsymbol{W}^2\boldsymbol{X}, \; \boldsymbol{W}^3\boldsymbol{X}, \ldots, \; \boldsymbol{W}^l\boldsymbol{X},$$

as valid instruments. An implicate assumption of this procedure is $\beta_i \neq 0$. Identification requires the rank of $\boldsymbol{H}$ to be at least $k + 1$. However, to avoid instrument proliferation, l is typically set to 1 or 2 (Arraiz et al., 2010).

The estimation procedure works as follows. The first step is to estimate $\boldsymbol{\zeta}$ by two-stage-least-squares (2SLS) using $\boldsymbol{H}$ as instruments and ignoring the error structure.

[8] N.B. $\rho^0 W^0 = I$.

The 2SLS estimator of $\boldsymbol{\zeta}$ is

$$\hat{\boldsymbol{\zeta}}_{2sls} = (\boldsymbol{Z}'\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{y},$$

where $\boldsymbol{P}_{\boldsymbol{H}} = \boldsymbol{H}(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'$. $\hat{\boldsymbol{\zeta}}_{2sls}$ is a consistent estimator of $\boldsymbol{\zeta}$, however the estimates will not be efficient unless $\delta = 0$.

If $\delta \neq 0$ the second step is to estimate δ using inefficient GMM, based on the 2SLS residuals $\hat{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\zeta}}_{2sls}$. The GMM estimator uses the following moment conditions (Kelejian and Prucha, 2010)

$$\begin{aligned} n^{-1}\,\mathbb{E}[(\boldsymbol{W}\boldsymbol{v})(\boldsymbol{W}\boldsymbol{v})'] &= n^{-1}tr\Big(\boldsymbol{W}\big[diag_{i=1}^{n}(\mathbb{E}[v_i^2])\big]\boldsymbol{W}'\Big), \\ n^{-1}\,\mathbb{E}[(\boldsymbol{W}\boldsymbol{v})'\boldsymbol{v}] &= 0. \end{aligned} \tag{1.14}$$

If homoskedasticity is assumed $n^{-1}\,\mathbb{E}[\boldsymbol{v}'\boldsymbol{v}] = \sigma^2\boldsymbol{I}$, (1.14) becomes,[9]

$$n^{-1}\,\mathbb{E}[(\boldsymbol{W}\boldsymbol{v})'(\boldsymbol{W}\boldsymbol{v})] = \sigma^2 n^{-1}tr(\boldsymbol{W}\boldsymbol{W}').$$

Substituting $\hat{\boldsymbol{v}} = \hat{\boldsymbol{r}} - \delta\boldsymbol{W}\hat{\boldsymbol{r}}$ for $\boldsymbol{v}$ gives the sample moment condition of the initial and inefficient estimator of δ, denoted $\hat{\delta}$. Consistency of $\hat{\delta}$ under homoskedasticity and heteroskedasticity is proven in (Kelejian and Prucha, 1998) and Kelejian and Prucha (2010).

Step three a Cochrane and Orcutt (1949) type transformation is applied to (1.11) and (1.12), by pre-multiplying (1.11) and (1.12) by $(\boldsymbol{I} - \hat{\delta}\boldsymbol{W})$. 2SLS is then applied

[9]This gives the moment condition from Kelejian and Prucha (1999).

to

$$\boldsymbol{y}_{\hat{\delta}} = \boldsymbol{Z}_{\hat{\delta}}\boldsymbol{\zeta} + \boldsymbol{v},$$

where $\boldsymbol{y}_{\hat{\delta}} = (\boldsymbol{I} - \hat{\delta}\boldsymbol{W})\boldsymbol{y}$ and $\boldsymbol{Z}_{\hat{\delta}} = (\boldsymbol{I} - \hat{\delta}\boldsymbol{W})\boldsymbol{Z}$. The GS2SLS estimator of $\boldsymbol{\zeta}$ is

$$\hat{\hat{\boldsymbol{\zeta}}}_{gs2sls} = \left(\boldsymbol{Z}'_{\hat{\delta}}\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{Z}_{\hat{\delta}}\right)^{-1}\boldsymbol{Z}'_{\hat{\delta}}\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{y}_{\hat{\delta}}.$$

Step four, efficient GMM estimator based on $\hat{\hat{\boldsymbol{u}}} = \boldsymbol{y} - \boldsymbol{Z}\hat{\hat{\boldsymbol{\zeta}}}_{gs2sls}$ is used to estimate δ, the GS2SLS residuals are used for the efficient GMM weighting matrix.

GS2SLS is proven to be asymptotically efficient and has similar small sample properties to ML (Kelejian and Prucha, 2010). Even when an optimal GMM estimator is used, the limiting distribution is the same as the ML estimator (with normal disturbances), and identification of spatial parameters can still be an issue, especially when considering models with higher-order moments (Lee, 2007). Breitung and Wigger (2018) note there is a close relationship between the ML and Kelejian and Prucha (1999) GMM estimator, as they derive a GMM estimator that uses a moment equations based on a second-order approximation of the ML scores and finds their proposed estimator yields similar moment conditions to Kelejian and Prucha (1999) and this explains why both estimators typically perform similarly in applied research.

Other GMM-based estimators have also been used to estimate spatial models. Jin and Lee (2018) extend the MESS model to include endogenous variables and spatial lags of the exogenous variables and propose estimating the model by nonlinear

two-stage least squares (N2SLS).[10] They find when the coefficients of the endogenous variables and spatial lags of the exogenous variables are non-zero (in the true model), N2SLS has the $\sqrt{n}$-rate of convergence and is asymptotically normal. However, when coefficients of the endogenous variables and spatial lags of the exogenous variables are zero, N2SLS becomes irregular as it has a slower than $\sqrt{n}$-rate of convergence and is asymptotically non-normal.

Breitung and Wigger (2018) propose a GMM-based estimator that uses a single quadratic form moment equation (i.e., just identified) to estimate an SEM, SAR, and SDM. As the proposed estimator uses just a single moment condition, no GMM weights matrix is required. The proposed moment condition stems from a first-order approximation of the ML scores. They also find the proposed estimator is robust to heteroskedastic and non-Gaussian errors.

1.1.3.3 Spatial Parameter Identification and Robustness

For the SARAR model, spatial parameter identification of δ_0 and ρ_0 is not possible when the SWM for the error is the same as for the dependent variable and parameter vector $\boldsymbol{\beta}_0 = 0$, due to the likelihood being perfectly symmetric for δ and ρ (Kelejian and Piras, 2017). However if $\boldsymbol{\beta}_0 \neq 0$ then δ_0 and ρ_0 can be identified. A common mistake many researchers make is to assume the identification condition restricts either δ or ρ to be equal to zero. However, this may not be the case, as the correlation patterns could be richer than those implied by either the SEM or SAR models (Kelejian and Piras, 2017).

[10] N2SLS is a GMM-based estimator.

Gibbons and Overman (2012) note that when applied researchers estimate spatial models, they implicitly assume the functional form is known and use data-driven model comparison techniques to select models. They argue that in many cases, such an approach yields, at best, only very weak identification of causal effects. Many of the consistency proofs for the methods described so far require the assumption that the estimated spatial economic model is the true DGP. For example, Lee (2004) shows the consistency of QML under the assumption that that estimated spatial economic model is the true DGP.

Gibbons and Overman (2012) also argue that identification by powers of $\boldsymbol{W}$ typically results in weak identification; the degree of collinearity between these often lags is high as they are simply weighted averages of $\boldsymbol{x}_i$ in some neighbourhood of i. Despite being nested, each of the spatial models discussed so far could imply a very different underlying economic process is at work. Thus, Gibbons and Overman (2012) argues comparison of these models may not always be valid. They suggest that researchers use the experimentalist paradigm instead, where a satisfactory strategy must use theoretical arguments or informal reasoning to justify that the source of exogenous variation can identify the parameter(s) of interest.

Historically, applied researchers have generally focused on specifying the SWM rather than the empirical spatial structure (LeSage and Pace, 2014). A typical robustness check in the applied spatial economics literature is to test how sensitive the estimated parameters are to different SWMs. When estimates are found to be sensitive to the choice of SWM, researchers have attributed this sensitivity to the choice of SWM, however as LeSage and Pace (2014) show, as long as the SWMs are

reasonably well correlated, the results should not be overly sensitive to the choice of SWM. Implying the observed sensitivity is driven by model misspecification rather than the SWM choice. Thus, LeSage and Pace (2014) argue that researchers should focus on the model specification rather than finding the 'ideal' SWM.

1.2 Spatial Filtering

The spatial filtering literature aims to eliminate spatial autocorrelation patterns rather than directly estimating any spatial parameters. Getis and Griffith (2002) outline two non-parametric methods to spatially filter data. The first method proposed by Getis (1990) decomposes a data series into non-spatial and spatial series. The procedure depends on finding an appropriate distance d within which nearby units are spatially correlated and then calculating how much each unit contributes to the spatial dependence in the variable of interest. This method is based on the local spatial statistic Getis G_i statistic (Getis and Ord, 1992).[11] The formula for the Getis G_i statistic is:

$$G_i = \frac{\sum_{j=1}^{n} w_{ij,d} x_j}{\sum_{j=1}^{n} x_j}, \quad \forall i \neq j,$$

where $w_{ij,d} = 1$ if units i and j are within the cutoff distance d and zero otherwise. The filtered observation ($\breve{x}_i$) is give by:

$$\breve{x}_i = x_i \frac{[\sum_{j=1}^{n} w_{ij,d}/(n-1)]}{G_i}. \tag{1.15}$$

[11] A local spatial statistic gives a statistic for each spatial unit $x_{i \in n}$ of a variable $\boldsymbol{x}$.

The intuition behind (1.15) is it compares the observed value of G_i with its expected value, $[\sum_{j=1}^{n} w_{ij,d}/(n-1)]$. If these are the same, i.e., no spatial autocorrelation, the filtered observation is the same as the unfiltered, $\breve{x}_i = x_i$. If $G_i > [\sum_{j=1}^{n} w_{ij,d}/(n-1)]$ the difference between x_i and $\breve{x}_i$ is positive, which indicates spatial autocorrelation among high values of $\boldsymbol{x}$ and vice versa. This methodology has two major disadvantages; it can only be used on non-negative data, and each variable has to be filtered separately pre-regression.

Griffith (2000) proposed a second, more versatile procedure called eigenvector spatial filtering (ESF) based on a spectral decomposition of a transformed SWM matrix, eigenvectors from the decomposition are used to filter or approximately destroy any terms involving the SWM. The second approach has become much more popular with applied researchers as it has several key advantages. It allows for both positive and negative spatial correlation, can be applied to both exogenous and endogenous variables in systems of equations (Tiefelsdorf and Griffith, 2007), and is agnostic to the underlying functional form of the spatial process.

1.3 Cross-sectional dependent limit theorem

An essential concept for deriving limit theorems of cross-sectionally dependent random processes is a triangular array, which takes the form

$$
\begin{array}{l}
x_{1,1} \\
x_{2,1}, x_{2,2} \\
x_{3,1}, x_{3,2}, x_{3,3} \\
\vdots \\
x_{n,1}, x_{n,2}, \cdots, x_{n,n} \\
\vdots
\end{array}
$$

A row-wise *i.i.d.* TA $x_{n,1}, x_{n,2}, \cdots, x_{n,n}$ are mutually independent and have the same distribution. But, $x_{n,1}$ and $x_{n+1,1}$ can have different distribution. These dependent processes are defined on a common probability space $(\Psi, \mathcal{F}, P)$ where Ψ is the sampling space (i.e., a set of all possible outcomes), $\mathcal{F}$ is a set of measurable events and P is a probability function from the sample space to $[0, 1]$.

Now we provide a summary of Kojevnikov et al. (2021) limit theorems for cross-sectionally dependent random variables, where the dependence stems from a network/spatial structure. Kojevnikov et al. (2021) extends Doukhan and Louhichi (1999) notion of ψ-dependence for temporal data to cross-sectional data. Modeling cross-sectional dependence requires assuming a certain metric on $N_n = N = \{1, \ldots, n\}$. Consider an undirected network $G_n = G$ on N where $G = (N, E)$ and $E \subseteq \{\{i, j\} : i, j \in N, i \neq j\}$ denotes the set of links. $d_{i,j}$ is the distance between i and j in G, (i.e., the length of the shortest path between nodes i and j given G) is a metric on the set N.

They define cross-sectional dependence as a stochastic dependence pattern of random variables determined by the distance d in G. Let $N_{i,r} = \{j \in N : d_{i,j} \leq r\}$ (set of the nodes within the distance r from node i) and $N_{i,r}^d = \{j \in N : d_{i,j} = r\}$ (set of the nodes exactly distance r from node i).

Now consider the triangular array of random vectors $\{\boldsymbol{y}_{i,n}\}_{i \in N}$, $\boldsymbol{y}_{i,n} \in \mathbb{R}^v$ (for simplicity denote the $\{\boldsymbol{y}_i\}$), laid on a network G, defined on a common probability space $(\Psi, \mathcal{F}_n, P)$. Further, let $\mathcal{C}_n \subset \mathcal{F}_n$ be a sub-σ-field and assume $\{\boldsymbol{y}_i\}$ is weakly dependent given the sequence $\{\mathcal{C}_n\}_{n \geq 1} = \{\mathcal{C}\}$. Specifically, consider two sets of cross-sectional units (of size $a, b \in \mathbb{N}$) with a distance between each other of at least $r > 0$. Let $\mathcal{P}_{a,b,r}$ denote the collections of all pairs

$$\mathcal{P}_{a,b,r} := \{(A, B) : A, B \subset N, |A| = a,\ |B| = b,\ d_{A,B} \geq r\},$$

with $d_{A,B} = \min_{i \in A, j \in B} d_{i,j}$ and let $\mathcal{L}_{v,a}$ denote a family of real values, bounded Lipschitz functions

$$\mathcal{L}_{v,a} = \{f : \mathbb{R}^{v \times a} \to \mathbb{R} : ||f||_\infty < \infty, \mathrm{Lip}(f) < \infty\},$$

with $\mathrm{Lip}(f)$ denoting the Lipschitz constant of f. The functions in $\mathcal{L}_{v,a}$ are Lipschitz with respect to the distance

$$\mathtt{d}_a(\boldsymbol{Q}, \boldsymbol{K}) = \sum_{l=1}^{a} ||\boldsymbol{q}_l - \boldsymbol{k}_l||_2,$$

where $\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_a)$, $\boldsymbol{K} = (\boldsymbol{k}_1, \ldots, \boldsymbol{k}_a)$ are points in $\mathbb{R}^{v \times a}$.[12] Additionally for each set $A \subset N$ let $y_A = \{\boldsymbol{y}_{i,n} : i \in A\}$. They assume there exists a sequence of σ-fields $\{\mathcal{C}\}$ that for each $n \geq 1$, the adjacency matrix $\boldsymbol{A}$ of graph G is $\mathcal{C}$ measurable. The exact definition of ψ dependence used by Kojevnikov et al. (2021) is, the triangular array $\{\boldsymbol{y}_{i,n} = \boldsymbol{y}_i\}_{i \in N}$, $\boldsymbol{y}_{i,n} \in \mathbb{R}^v$ is called conditionally ψ-dependent given $\{\mathcal{C}_n = \mathcal{C}\}$, if each $n \in \mathbb{N}$ there exists a $\mathcal{C}$-measurable sequence $\mu = \{\mu_{r,n} : r \geq 0\}$, $\mu_{0,n} = 1$ a collection of non-random functions $\psi_{a,b} : \mathcal{L}_{v,a} \times \mathcal{L}_{v,b} \to [0, \infty)$ such that for all $(A, B) \in \mathcal{P}_{a,b,r}$ with $r > 0$ and all $f \in \mathcal{L}_{v,a}$ and $g \in \mathcal{L}_{v,b}$,

$$|\operatorname{Cov}(f(y_A), g(y_B)|\mathcal{C})| \leq \psi_{a,b}(f, g)\mu_r \quad \text{a. s.} \tag{1.16}$$

The sequence $\mu_r = \mu_{r,n}$ is the dependence coefficients of $\{y_i\}$.

The idea is by conditioning on the sub-σ-field $\mathcal{C}$ that generates the cross-sectional dependent of the triangular array $\{\boldsymbol{y}_i\}$, the dependence becomes substantially weaker. Kojevnikov et al. (2021) view $\mathcal{C}$ as a 'common shock' that stems from characteristics or actions of central nodes that affect the other nodes through their many links. They are essentially viewing the observed network structure as a realisation of a stochastic network formation process, and by conditioning on the σ-field that generated this process, they can treat the observed network structure as fixed.

For their limit theorems, the first key assumption they make is that the triangular array $\{\boldsymbol{y}_i\}$ is conditionally ψ-dependent given $\{\mathcal{C}\}$ with the dependence coefficients

[12]The Lipschitz constant for a function $f : \mathbb{R}^{v \times a} \to \mathbb{R}$ is the smallest constant C such that $|f(\boldsymbol{Q}) - f(\boldsymbol{K})| \leq C\mathtt{d}_a(\boldsymbol{Q}, \boldsymbol{K}), \forall\, \boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{v \times a}$.

$\{\mu\}$ satisfying the following conditions. For some constant $C > 0$

$$\psi_{a,b}(f,h) \leq Cab(||f||_\infty + \text{Lip}(f))(||g||_\infty + \text{Lip}(g)) \tag{1.17}$$

and $\sup_{n\geq 1} \max_{r\geq 1} \mu_r < \infty$ a.s.

For their (strong) law of large numbers (LLN), they set $v = 1$ without loss of generality (w.l.o.g) as the LLN is applied element-by-element in the vector case. For their LLN, they require two additional assumptions. The first is the following moment condition, for some $p > 1$

$$\sup_{n\geq 1} \max_{i\in N} (\mathbb{E}[|y_i|^p|\mathcal{C}])^{1/p} < \infty \quad \text{a. s.}$$

and the second is

$$n^{-1} \sum_{r\geq 1} \delta_r^d \mu_r \to_{a.s.} 0, \tag{1.18}$$

where $\delta_r^d = n^{-1} \sum_{i\in N} |N_{i,r}^d|$ is a measure of denseness of the network. Equation (1.18) puts a restriction on the rate of decay of dependence and the denseness of the network with respect to the network distance. An example of where this assumption could fail is if one unit is connected to all other units (the star network is an example of such a structure).

Under these three assumptions Kojevnikov et al. (2021) derive their (strong) LLN (Theorem 3.1)

$$\left|\left| \frac{1}{n} \sum_{i\in N} (y_i - \mathbb{E}[y_i|\,\mathcal{C}]) \right|\right|_1 \to_{a.s.} 0.$$

Their central limit theorem is for a sum of random variables that are conditionally

ψ-dependent. Let $\sigma_n^2 = \sigma^2 = \mathrm{Var}(b|\,\mathcal{C})$ where $b = b_n = \sum_{i\in N} y_i$.

To derive their central limit theorem, they replace the second and third assumptions of the LLN with more restive assumptions. For the moment condition they require $p > 4$ (i.e. at least the 4th moment of y_i is finite) and there exists a positive sequence $m_n = m \to \infty$ such that for $k = 1, 2$

$$\frac{n}{\sigma^{2+k}} \sum_{r=0}^{\infty} c_{r,m;k} \mu_r^{1-\frac{2+k}{p}} \to_{a.s.} 0,$$
$$\frac{n^2 \mu_m^{1-1/p}}{\sigma} \to_{a.s.} 0,$$

where

$$\begin{aligned} c_{r,m;k} =& \inf_{\alpha>1} \left[\Delta_{r,m;k\alpha}\right]^{1/\alpha} \left[\delta^d_{r,\alpha/(1-\alpha)}\right]^{1-1/\alpha}, \\ \Delta_{r,m;k} =& n^{-1} \sum_{i\in N} \max_{j\in N^d_{i,r}} |N_{i,m}/N_{j,r-1}|^k, \\ \delta^d_{r,k} =& n^{-1} \sum_{i\in N} |N^d_{i,r}|^k, \end{aligned}$$

and $p > 4$ is as same as in the moment condition, as $n \to \infty$.

Under these regularity conditions and $\mathbb{E}[y_i|\,\mathcal{C}] = 0$ $a.s.$ they show as $n \to \infty$

$$\sup_{t\in\mathbb{R}} \left| P\left\{ \frac{b}{\sigma} \le t|\,\mathcal{C} \right\} - \Phi(t) \right| \to_{a.s.} 0,$$

where Φ denotes the distribution function of $N(0, 1)$.

Now suppose $\sigma^2/(nv^2) \to_{a.s.} 1$ as $n \to \infty$ where v^2 is a random variable that is $\mathcal{D}$ measurable and $\mathcal{D}$ is a sub-sigma field of $\mathcal{C}_n$ for all $n \ge 1$, then $b/\sqrt{n}$ converges

stably to a mixture of normal random variables.

1.4 Regularisation in Spatial Models

Regularisation methods have recently become popular in the spatial econometric literature. When conventional estimation techinques such as OLS or GMM do not work well, regularisation methods can be used instead.[13] For example, ridge regression is a regularisation method used to address near-perfect multicollinearity (Hastie et al., 2009).

1.4.1 Penalised Regressions

One of the most popular regularisation techniques in the spatial econometric literature is the least absolute shrinkage and selection operator (Tibshirani, 1996, Lasso). Lasso is a *special* case of constrained regressions (1.19) and penalised regressions (1.20). Lasso's popularity stems from the fact that it is capable of both variable/model selection and parameter estimation while also being computationally tractable.

$$\min_{\beta \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{subject to} \quad \boldsymbol{\beta} \in C, \tag{1.19}$$

$$\min_{\beta \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + P(\boldsymbol{\beta}), \tag{1.20}$$

[13] Regularisation is when some regularity conditions are imposed on the model to overcome a problem.

where $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$ is the OLS loss function,[14] C is some set and $P(\cdot)$ is some penalty function. Typically the constraint set C is chosen to be a sublevel set of a norm (or semi-norm) and $P(\cdot)$ to be a multiple of a norm (or semi-norm). Thus, (1.19) and (1.20) can be rewritten as

$$\min_{\beta \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{subject to} \quad \sum_{i=1}^{p} |\beta_i|^q \leq h, \tag{1.21}$$

$$\min_{\beta \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \theta \sum_{i=1}^{p} |\beta_i|^q, \tag{1.22}$$

where h and $\theta \geq 0$ are tuning or regularisation parameters and $q \geq 0$. $q = 1, 2$ corresponds to the ℓ_1 norm and ℓ_2 norm.

It is also possible to use the $\ell_0 = ||\boldsymbol{\beta}||_0 = \sum_{i=1}^{p} \mathbb{1}\{\beta_i \neq 0\}$ norm, which counts the number of non-zero elements.[15] Using the ℓ_0 norm as the penalty is referred to as best subset selection (BSS). BSS aims to find the subset of h regressors that produces the best fit in terms of squared error. For BSS, h is a positive integer and is fixed in advance. The smaller h, the sparser the solution. A sparse solution means that most elements in $\hat{\boldsymbol{\beta}}$ are zero. BSS is a minimization over all subsets of size h, which is an NP-complete problem and thus, computationally infeasible to solve (Foster and George, 1994).

BSS can be approximated in two ways: as a greedy approximation or a convex relaxation. The former leads to forward stepwise regression (iteratively adding vari-

[14] It also is possible to use other loss function, the GMM loss function is a common alternative.

[15] Technically, the ℓ_0 norm is not a norm as it does not satisfy positive homogeneity, $||a\boldsymbol{\beta}||_0 \neq a||\boldsymbol{\beta}||_0 \; \forall \; a \neq 1, 0$.

ables using some selection criterion). However, forward stepwise regression is still difficult to solve as different perturbations will often lead to different results. The latter leads to the Lasso.

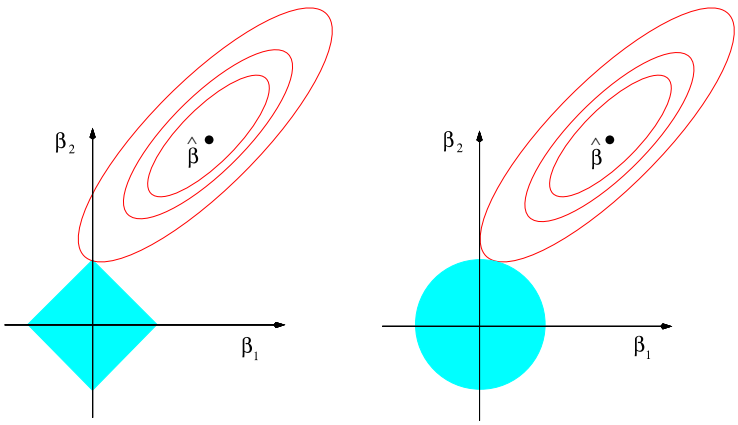
Penalisation based on the ℓ_1 norm ($q = 1$) is called Lasso. Lasso can be considered a convex relaxation BSS of Lasso as $q = 1$ is the smallest value q can take, and the optimisation problem is still convex. As the Lasso shrinks many of the coefficients to 'exactly' zero, the procedure can still be used for variables selection (Hastie et al., 2009). For Lasso, due to Lagrangian duality, there is a data-dependent correspondence between the penalty parameter θ and constraint parameter h.

Another common penalty is the ℓ_2 norm ($q = 2$) and is referred to in the literature as ridge regression (Hoerl and Kennard, 1970). For ridge regressions, there is also a data-dependent correspondence between the penalty parameter θ and constraint parameter h. Ridge regressions shrink the coefficients towards zero, but all the p coefficients will remain non-zero. The ridge regressions optimisation problem is strictly convex (assuming $\theta > 0$). Thus, the ridge regressions solution is always well defined with the closed form

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \theta \boldsymbol{I}_k)^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Ridge regression is a regularisation technique commonly used in the data science literature to correct for near-perfect multicollinearity or when the (re-scaled) Gram matrix $\boldsymbol{X}'\boldsymbol{X}/n$ is unstable.

Figure 1.1: Geometry of Lasso and Ridge Regression

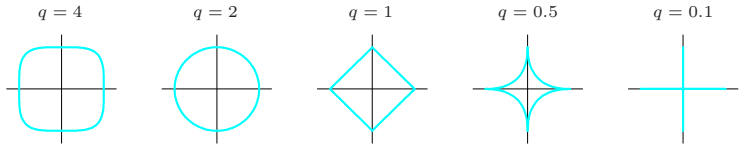


Note: Contours of error and constraint functions for Lasso (left) and ridge regression (right). The solid blue areas are the constraints for $|\beta_1| + |\beta_2| \leq h$ and $\beta_1^2 + \beta_2^2 \leq h$, while red ellipses are the contorts of the residual sum of squares. $\hat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta}$. Based on Figure 3.11 of Hastie et al. (2009).

1.4.1.1 The Geometry of Different Norms and Semi-Norm

To get a better idea of how Lasso shrinks coefficients to 'exactly' zero and ridge toward zero, we present Figure 1.1, which shows the geometry of ridge and Lasso regression for the case when $k = 2$. The ridge constraint is a circle (has no corners) and gives constrained estimates (compared to the OLS estimate) but non-zero coefficients. In comparison, the Lasso constraint is a polytope (in this case, a diamond shape). Thus, the solution hits a vertex (corner) of the constraint and constrains one of the coefficients estimates to exactly zero ($\hat{\beta}_1 = 0$), and the other is non-zero ($\hat{\beta}_2 \neq 0$). For Lasso, when $k > 2$, the diamond becomes a rhomboid and has many faces, flat edges, and corners. This creates substantially more opportunities for es-

Figure 1.2: Constraints of Some of Norms and Semi-Norms



Note: Constraint function of constant value of $\sum_{i=1}^{k} |\beta_i|^q$ for a given q with $k = 2$. Based on Figure 3.12 of Hastie et al. (2009)

timated parameters to be zero (Hastie et al., 2009). This is the intuition as to why Lasso yields a sparse solution, and the ridge solution is not sparse.

Figure 1.2 shows the constraint function for several values of q. This shows when $q \leq 1$, the constraint has corners, so the corresponding estimator will be able to perform model/variable selection. However, as was previously mentioned, the problem is no longer convex when $q < 1$, so it is hard to solve.

1.4.1.2 Lasso for Variable Selection

The critical assumption of Lasso is the sparsity assumption. To understand this consider the following linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v}. \tag{1.23}$$

Additionally, assume that the cardinality of the active set is $s := \sum_{i=1}^{k} \mathbb{1}\{\beta_i \neq 0\}$ (number of non-zero coefficients). A model is typically considered sparse if s is small relative to k. Lasso aims to identify the active set s. Of all the norms (and semi-norms) that could be used for variable selection, Lasso is the most computationally tractable.

Lasso, as a variable selection technique, has the key advantage of handling both low-dimensional and high-dimensional data. The low-dimensional case is defined as the case where the number of parameters k is less than the sample size n ($k < n$). In contrast, the high-dimensional case is defined as where the number of parameters k is greater than the sample size n ($k > n$). When $k > n$, the OLS estimation is infeasible as the matrix $\boldsymbol{X}'\boldsymbol{X}/n$ is rank deficient. However, when k is large relative to n, variable selection is difficult, as hypothesis testing tends to lead to many false positives. Under the assumption the coefficient vector $\boldsymbol{\beta}$ is sparse, Lasso can handle $k >> n$. Tibshirani (1996) argues that sparsity benefits the model interpretation of Lasso.

Large k models are common in economic application (Belloni and Chernozhukov, 2011). Some examples include regional or cross-country growth regressions where the number of covariates is large and the number of regions or countries is small, the 'many-instruments' case in instrument variables models,[16] and for model selection where there are often a large number of competing models.

1.4.1.3 Performance of Lasso

The penalised form of the Lasso problem in vector norm notation is

$$\hat{\boldsymbol{\beta}}_\theta \in \min_{\beta \in \mathbb{R}^k} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \theta||\boldsymbol{\beta}||_1. \tag{1.24}$$

Equation (1.24) defines a family of estimates indexed by the tuning parameter

[16]The definition of 'many-instruments' is when the number of instruments z is large relative to n but with $z < n$. The case where $z > n$ is called 'very-many-instruments'.

$\theta \geq 0$. [17] If $\theta = 0$ the Lasso solution reduces to the OLS solution and $\theta \to \infty$ yields a null vector. More moderate values of θ will result in some parameters being shrunk towards zero and some to exactly zero. Thus, the choice of θ will ultimately determine the estimates $\hat{\boldsymbol{\beta}}_\theta$.

Cross-Validation

The conventional way to select the tuning parameter θ is to choose a value that optimises out-of-sample prediction performance. The expected mean squared prediction error (MSPE) is a common way to measure the predictive performance of a model. The MSPE of (1.23) can be decomposed into (Hastie et al., 2009)

$$\mathbb{E}[(\hat{\boldsymbol{y}} - \boldsymbol{y})^2] = \text{Bias}(\hat{\boldsymbol{y}})^2 + \text{Var}(\hat{\boldsymbol{y}}) + \text{Var}(\boldsymbol{v}),$$

where $\text{Bias}(\hat{\boldsymbol{y}})^2 := (\mathbb{E}[\hat{\boldsymbol{y}}] - \boldsymbol{y})^2$ is squared bias and $\hat{\boldsymbol{y}}$ the estimate of the dependent variable. This shows that as $\text{Var}(\boldsymbol{v})$ is the unavoidable error, the MSPE is a trade-off between the bias and the variance. Under the Gauss-Markov assumptions, it is well known that OLS has the lowest variance of all unbiased linear estimators.[18]

It is important to note that the estimator that minimises the MSPE will not necessarily be unbiased. The Lasso estimator is biased as the ℓ_1 penalty is biased towards zero. However, Lasso can still outperform OLS, at least in terms of MSPE. This can occur because the ℓ_1 penalty reduces the number of covariates, which decreases the variance while increasing the bias (Hastie et al., 2009).

[17] $\theta < 0$ yields an non-unique solution.

[18] the Gauss-Markov assumptions are linearity (implied by (1.23)), exogenous covariates and spherical disturbances ($\mathbb{E}[\boldsymbol{v}\boldsymbol{v}'|\boldsymbol{X}] = \sigma_v^2 \boldsymbol{I}$)

Given that the prediction error is unobservable, the conventional approach uses K-fold cross-validation (CV) to estimate MSPE (Hastie et al., 2009). The general idea is the MSPE is used to assess how one model performs relative to alternatives. The general K-fold CV algorithm to select θ works as follows:

1. Divide the n observations randomly into K approximately equal groups (folds) denoted $F_1, \ldots, F_K$.

2. For $k = 1, \ldots, K$:

   (a) F_k is the validation data set, and the remaining $K - 1$ groups form the training data set.

   (b) For each value of the tuning parameter $\theta \in \{\theta_1, \ldots, \theta_m\}$ estimate $\hat{\boldsymbol{\beta}}_{-k}(\theta)$, where $\hat{\boldsymbol{\beta}}_{-k}$ is an estimate of $\boldsymbol{\beta}$ using the training dataset.

3. For each value of θ compute the average MSPE over all the folds

$$CV(\theta) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in F_k} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{-k}(\theta) \right)^2$$

   where $\boldsymbol{x}_i$ is a vector of explanatory variables associated with observation i.

4. Iterate over the values of θ to find to the θ that minimises $CV(\theta)$.

When $K = n$, the procedure is referred to as Leave-one-out Cross-Validation. It is common in applications to set $K = 5$ or $K = 10$ as the procedure is computationally cumbersome (Hastie et al., 2009).

Theoretical Properties

The theoretical properties of Lasso can be measured in terms of performance regarding parameter estimation and performance in terms of recovering the support of the true parameter vector $\boldsymbol{\beta}_0$. In the following, subscript 0 is used to denote the true vector or value.

Non-asymptotic performance bounds

The predictive performance of Lasso is generally measured by the ℓ_2 prediction norm $||X(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})/\sqrt{n}||_2$. In contrast, $||\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}||_2$ or $||\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}||_1$ measure the loss from the estimation of the parameter vector.

The majority of the literature on (non-asymptotic) performance bounds of Lasso does not consider θ being selected by CV.[19] The bounds are developed based on assumptions about the choice of penalty and conditions on the regressor Gram matrix $\boldsymbol{X}'\boldsymbol{X}/n$, θ is treated as a random variable. For simplicity, I assume the matrix $\boldsymbol{X}$ is non-stochastic.

The bounds are developed from the fundamental inequality (1.25), which stems from the fact that $\hat{\boldsymbol{\beta}}$ minimizes the Lasso objective function.

$$||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2/n + \theta||\hat{\boldsymbol{\beta}}||_1 \leq ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0||_2^2/n + \theta||\boldsymbol{\beta}_0||_1. \tag{1.25}$$

Let $\hat{\boldsymbol{\Delta}} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ and rearranging (1.25) yields

$$||\boldsymbol{X}\hat{\boldsymbol{\Delta}}||_2^2/n \leq 2\boldsymbol{v}'\boldsymbol{X}\hat{\boldsymbol{\Delta}}/n + \theta(||\boldsymbol{\beta}_0||_1 - ||\hat{\boldsymbol{\beta}}||_1). \tag{1.26}$$

[19]The following discussion is based on (Hastie et al., 2015, Chapter 11), (Bühlmann and Van De Geer, 2011, Chapter 6) and Ahrens (2017).

The term $2\boldsymbol{v}'\boldsymbol{X}\hat{\boldsymbol{\Delta}}/n$ is often referred to as the 'empirical process' part of the problem, captures the 'random' part of the problem and by the Hölder inequality with ℓ_∞ and ℓ_1 norms, $|\boldsymbol{v}'\boldsymbol{X}\hat{\boldsymbol{\Delta}}| \leq ||\boldsymbol{v}'\boldsymbol{X}||_\infty ||\hat{\boldsymbol{\Delta}}||_1$. The idea behind the penalty is it should dominate the 'random' part of the problem,

$$2||\boldsymbol{v}'\boldsymbol{X}||_\infty/n \leq \theta, \tag{1.27}$$

this is based on the (arbitrary) assumption $\theta \geq 2\theta_0$ where θ_0 is the true value of θ (Bühlmann and Van De Geer, 2011).

The second right hand side term in (1.26) can be re-written using $||\hat{\boldsymbol{\beta}}_{\grave{\Omega}}||_1 = ||\hat{\boldsymbol{\Delta}}_{\grave{\Omega}}||_1$

$$||\boldsymbol{\beta}_0||_1 - ||\hat{\boldsymbol{\beta}}||_1 \;=\; ||\boldsymbol{\beta}_0||_1 - (||\hat{\boldsymbol{\beta}}_\Omega||_1 + ||\hat{\boldsymbol{\beta}}_{\grave{\Omega}}||_1) \;=\; ||\boldsymbol{\beta}_0||_1 - (||\hat{\boldsymbol{\beta}}_\Omega||_1 + ||\hat{\boldsymbol{\Delta}}_{\grave{\Omega}}||_1),$$

where the $\Omega := \text{supp}(\boldsymbol{\beta}_0)$ is the active set and $\grave{\Omega}$ is the complementary set. Using the reverse triangle inequality $||\hat{\boldsymbol{\beta}}_\Omega||_1 \geq ||\boldsymbol{\beta}_{0,\Omega}||_1 - ||\hat{\boldsymbol{\Delta}}_\Omega||_1$ and given $||\boldsymbol{\beta}_{0,\Omega}||_1 = ||\boldsymbol{\beta}_0||_1$, we have

$$||\boldsymbol{\beta}_0||_1 - ||\hat{\boldsymbol{\beta}}||_1 \;\leq\; ||\boldsymbol{\beta}_0||_1 - (||\hat{\boldsymbol{\beta}}_\Omega||_1 + ||\hat{\boldsymbol{\Delta}}_{\grave{\Omega}}||_1 - ||\hat{\boldsymbol{\Delta}}_\Omega||_1) \;\leq\; ||\hat{\boldsymbol{\Delta}}_\Omega||_1 - ||\hat{\boldsymbol{\Delta}}_{\grave{\Omega}}||_1,$$

Thus, (1.26) reduces to

$$||\boldsymbol{X}\hat{\boldsymbol{\Delta}}||_2^2/n \leq \theta||\hat{\boldsymbol{\Delta}}||_1 + \theta(||\hat{\boldsymbol{\Delta}}_\Omega||_1 - ||\hat{\boldsymbol{\Delta}}_{\hat{\Omega}}||_1) \leq 2\theta||\hat{\boldsymbol{\Delta}}_\Omega||_1 \tag{1.28}$$

$$\leq 2\theta\sqrt{s}||\hat{\boldsymbol{\Delta}}||_2, \tag{1.29}$$

where (1.29) uses the fact that $||\hat{\boldsymbol{\Delta}}_\Omega||_1 \leq \sqrt{s}||\hat{\boldsymbol{\Delta}}||_2$. From here in the low dimensional ($k < n$) setting, when the Gram matrix is well behaved it is relatively simple to derive bounds the prediction norm $||\boldsymbol{X}\hat{\boldsymbol{\Delta}}||_2/\sqrt{n}$ and parameter norm $||\hat{\boldsymbol{\Delta}}||_2$. As in the case the Gram matrix $\boldsymbol{X'X}/n$ is positive definite so the smallest eigenvalue will be positive; that is,

$$\min_{\boldsymbol{\Delta}\in\mathbb{R}^k:\boldsymbol{\Delta}\neq 0} \frac{||\boldsymbol{X}\boldsymbol{\Delta}||_2}{||\boldsymbol{\Delta}||_2} > 0,$$

so we can substitute for either $||\boldsymbol{X}\hat{\boldsymbol{\Delta}}||_2/\sqrt{n}$ or $||\hat{\boldsymbol{\Delta}}||_2$. In contrast, in the high dimensional setting, the Gram matrix will be rank-deficient, and the smallest eigenvalue will be zero. This is because $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X'X}) \leq \min(k, n)$, which can be written as

$$\min_{\boldsymbol{\Delta}\in\mathbb{R}^k:\boldsymbol{\Delta}\neq 0} \frac{||\boldsymbol{X}\boldsymbol{\Delta}||_2}{||\boldsymbol{\Delta}||_2} = 0.$$

In this case, the OLS assumption that the regressor matrix has full column rank has to be weakened. Bickel et al. (2009) proposed the *restricted eigenvalue condition* (RE)

$$\tau_{min} := \min_{\boldsymbol{\Delta} \in \mathcal{D}(c_0, \Omega)} \frac{||\boldsymbol{X}\boldsymbol{\Delta}||_2}{\sqrt{n}||\boldsymbol{\Delta}||_2} > 0, \quad \mathcal{D}(c_0, \Omega) = \{\boldsymbol{\Delta} \in \mathbb{R}^k : ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1 \leq c_0 ||\boldsymbol{\Delta}_{\Omega}||_1\}, \tag{1.30}$$

where $c_0 \geq 1$ is a constant and τ_{min} is the minimum restricted eigenvalue. The minimum is now over the restricted set $\mathcal{D}(c_0, \Omega)$. Assuming $\theta > 0$ and using (1.28) we have

$$0 \leq \theta||\hat{\boldsymbol{\Delta}}||_1 + \theta(||\hat{\boldsymbol{\Delta}}_{\Omega}||_1 - ||\hat{\boldsymbol{\Delta}}_{\grave{\Omega}}||_1)$$
$$||\hat{\boldsymbol{\Delta}}_{\grave{\Omega}}||_1 \leq |\hat{\boldsymbol{\Delta}}||_1 + ||\hat{\boldsymbol{\Delta}}_{\Omega}||_1 \leq 2|\hat{\boldsymbol{\Delta}}||_1.$$

Thus, assuming (1.27) Lasso satisfies $\hat{\boldsymbol{\Delta}} \in \mathcal{D}(c_0, \Omega)$ for $c_0 = 2$.

There are many variants of the RE conditions, for example, the *Compatibility condition* of Bühlmann and Van De Geer (2011), Belloni et al. (2012), and Chetverikov et al. (2021). Bühlmann and Van De Geer (2011) and Hastie et al. (2015) give a comprehensive review of RE and related conditions and show the RE condition holds under fairly general conditions. One sufficient condition is, the appropriate sub-matrices of $n^{-1}\boldsymbol{X}\boldsymbol{X}$ are invertible (Bickel et al., 2009).

Assuming the non-stochastic matrix $\boldsymbol{X}$ satisfies (1.30) and the event (1.27) holds with high probability then τ_{min} can be substituted into (1.29) to derive the following ℓ_2 prediction norm and ℓ_2 parameter norm,

$$||\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)||_2/\sqrt{n} \leq \frac{2\theta\sqrt{s}}{\tau_{min}}, \tag{1.31}$$

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 \leq \frac{2\theta\sqrt{s}}{\tau^2_{min}}. \tag{1.32}$$

Importantly for the above inequalities to hold, the penalty must be chosen so it dominates the term $2||\boldsymbol{v}'\boldsymbol{X}||_\infty/n$ with high probability. Given $\boldsymbol{v}$ is unobservable, the question remains, how to ensure the event (1.27) holds with high probability? The most common approach is to assume $\boldsymbol{v}$ is *i.i.d.* and sub-Gaussian, derive the distribution of $||\boldsymbol{v}'\boldsymbol{X}||_\infty/n$ and then select the penalty parameter accordingly.[20] Belloni et al. (2012) weaken the assumption of Gaussian errors to allow for non-Gaussian and heteroskedastic errors by using Jing et al. (2003) moderate deviation theory for self-normalised sums to derive the smallest penalty that ensures that (1.27) holds as $n \to \infty$.

Chetverikov et al. (2021) develop (non-asymptotic) performance bound of Lasso with θ estimated by CV in the high dimensional setting. Similar to the RE, they impose conditions on the regressor matrix $\boldsymbol{X}$ that ensures that the eigenvalues of the population Gram matrix $\mathbb{E}[\boldsymbol{X}'\boldsymbol{X}]$ are bound away from zero (Chetverikov et al. (2021) Assumption 1) and find under some additional regularity conditions when the conditional distribution of the disturbance term $\boldsymbol{v}$ is Gaussian

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_p\left(\sqrt{\frac{s\log k}{n}} \times \sqrt{log(kn)}\right).$$

[20]For example see Lemma 6.2 in Bühlmann and Van De Geer (2011).

The term $\sqrt{s \log k/n}$ represents the optimal rate of convergence, and the CV Lasso estimator attains this rate up to a small $\sqrt{log(kn)}$ factor.

Model selection performance

It is important to note consistent parameter estimations and model selection consistency are different. Even if the parameter norm $||\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}||_2$ is small it is still feasible the support of $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ differ (Hastie et al., 2015, pg. 301). The definition of model selection consistency is

$$\lim_{n\to\infty} P\big(\operatorname{supp}(\hat{\boldsymbol{\beta}}) = \operatorname{supp}(\boldsymbol{\beta}_0)\big) = 1.$$

A stronger form of selection consistency is sign selection consistency

$$\lim_{n\to\infty} P\big(\operatorname{sign}(\hat{\boldsymbol{\beta}}) = \operatorname{sign}(\hat{\boldsymbol{\beta}}_0\big) = 1,$$

where sign$(\cdot)$ maps positive entry to 1, negative entry to -1, and zero to zero. Zou (2006); Zhao and Yu (2006); Meinshausen and Bühlmann (2006) all show the sufficient and (almost) necessary for Lasso be sign consistent is called the *irrepresentable condition*. Let $\boldsymbol{C}_{\Omega\Omega} = n^{-1}\boldsymbol{X}_{\Omega}'\boldsymbol{X}_{\Omega}$, $\boldsymbol{C}_{\grave{\Omega}\Omega} = n^{-1}\boldsymbol{X}_{\grave{\Omega}}'\boldsymbol{X}_{\Omega}$, $\boldsymbol{C}_{\Omega\grave{\Omega}} = n^{-1}\boldsymbol{X}_{\Omega}'\boldsymbol{X}_{\grave{\Omega}}$, $\boldsymbol{C}_{\grave{\Omega}\grave{\Omega}} = n^{-1}\boldsymbol{X}_{\grave{\Omega}}'\boldsymbol{X}_{\grave{\Omega}}$, where $\boldsymbol{X}_{\Omega}$ is an $n \times s$ matrix with columns corresponding to the active set $\Omega := \operatorname{supp}(\boldsymbol{\beta}_0)$ and s is the cardinality of Ω. $\grave{\Omega}$ is the complementary set and the $n \times (k-s)$ matrix $\boldsymbol{X}_{\grave{\Omega}}$ is defined accordingly. Now the Gram matrix $\boldsymbol{C} = n^{-1}\boldsymbol{X}'\boldsymbol{X}$ can be re-expressed in block-wise form

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{\Omega\Omega} & \boldsymbol{C}_{\Omega\grave{\Omega}} \\ \boldsymbol{C}_{\grave{\Omega}\Omega} & \boldsymbol{C}_{\grave{\Omega}\grave{\Omega}} \end{bmatrix}.$$

The irrepresentable condition (IC) is defined as follows

$$|\boldsymbol{C}_{\grave{\Omega}\Omega}(\boldsymbol{C}_{\Omega\Omega})^{-1}\,\mathrm{sign}(\boldsymbol{\beta}_{\Omega})| < \mathbf{1} - \boldsymbol{\nu}, \tag{1.33}$$

where $\boldsymbol{\nu}$ is a vector of positive constants. The IC can be interpreted as a regularisation constraint of the regression coefficients of $\boldsymbol{X}_{\grave{\Omega}}$ (the irrelevant covariates) on $\boldsymbol{X}_{\Omega}$ (the relevant covariates), the total amount of irrelevant covariates represented by the covariates in the true model must be less than one. Intuitively, the IC states that none of the variables in the inactive set should be highly correlated with the active set, or Lasso will be unable to distinguish between the two sets.

The RE condition discussed above is much weaker than the IC. For the IC (1.33) to hold, it must be the case that $\boldsymbol{C}_{\Omega\Omega}$ is positive definite, implying the condition $s < n$. If $\boldsymbol{C}_{\Omega\Omega}$ is not positive definite, then Lasso can not identify the true model, even if the true support was known.

In classical econometric theory, it is assumed that k is fixed and the Gram matrix $n^{-1}\boldsymbol{X}'\boldsymbol{X} \rightarrow_{a.s.} \boldsymbol{A}$ as $n \rightarrow \infty$, where $\boldsymbol{A}$ is a positive definite matrix. This assumption is reasonable in the low-dimensional setting. However, in the high dimensional setting, this assumption is *not* reasonable. Instead, this is relaxed to $k \rightarrow \infty$ as $n \rightarrow \infty$. For this reason k_n, $\boldsymbol{C}_n$ and $\boldsymbol{\beta}_n$ are indexed by the sample size n.

Knight and Fu (2000) showed, for the fixed k case, the regularity condition under which Lasso is consistent for estimating regression parameters. Leng et al. (2004)

also studied the fixed k with orthogonal design and showed that when prediction accuracy criteria are used to estimate θ, Lasso is generally not consistent for variable selection (it will tend to over-select). Zou (2006) showed for fixed k when the true value of the parameter is zero, there is a positive probability mass, implying Lasso is not consistent for variable selection in this case. Zhao and Yu (2006) allow k to grow faster than n showed that under the IC, Lasso is consistent for variable selection provided k is less than $\exp(n^a)$ for some $0 < a < 1$ and θ grows faster than $\sqrt{n \log(k)}$ when the errors have Gaussian tails. Bühlmann and Van De Geer (2011, Chapter 7) conclude Lasso is consistent for variable selection but only under specific conditions and not in applications where the covariates exhibit strong correlation.

1.4.1.4 Related Lasso Methods

There are many variants of the Lasso estimator. The most common are discussed in this section.

Post-Lasso. As discussed in Section 1.4.1.3 the Lasso estimates are biased towards zero. Post-Lasso is an intuitive method for reducing the bias from the penalisation (assuming Lasso selects the correct set of variables). The Post-Lasso estimator is defined as

$$\min_{\beta \in \mathbb{R}^k} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{subject to} \quad \text{supp}(\boldsymbol{\beta}) = \text{supp}(\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ is the Lasso solution. The Post-Lasso estimator is an OLS regression using the variables selected by Lasso. Belloni and Chernozhukov (2013) show under a condition on the Gram matrix they call the restricted sparse eigenvalue condition,

Post-Lasso convergence rates are at least as good as Lasso rates.

Adaptive Lasso. Zou (2006) argued that the asymptotic setup of Lasso is unfair as θ is the same for every β_j, so instead proposes a version with adaptive weighted called Adaptive Lasso. The adaptive Lasso problem is

$$\hat{\beta}_\theta \in \min_{\beta \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \theta \sum_{j=1}^{k} w_j^{-1} |\beta_j|,$$

where w_j is a constant (weight). Zou studies the fixed k case and proposes the weights can be based on the initial OLS estimates. Huang, Ma and Zhang (2008) examines the case where k grows with n and suggests the weights based on estimates from OLS marginal regressions.

Elastic-net. Proposed by Zou and Hastie (2005) elastic-net uses both the ℓ_1 and ℓ_2 penalty. The elastic net problem is

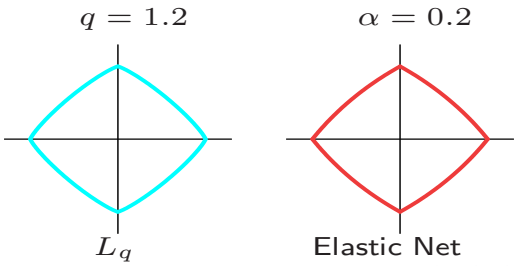
$$\hat{\beta}_\theta \in \min_{\beta \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \theta\big((1-\alpha)||\boldsymbol{\beta}||_1 + \alpha||\boldsymbol{\beta}||_2^2\big),$$

where $\alpha \in [0, 1]$ controls the mix of Lasso and ridge.[21] Zou and Hastie (2005) highlight several disadvantages of Lasso. Firstly, if $k > n$ due to the nature of the convex optimisation problem, it will select at most n variables. Secondly, if $\boldsymbol{X}$ contains groups of variables with high within-group correlation, the Lasso will tend to select only one of the variables from each group. Zou and Hastie (2005) show that elastic nets can outperform Lasso for predictive performance and variable selection.

Figure 1.3 compares the $\ell_{1.2}$ penalty to the Elastic net penalty with $\alpha = 0.2$. The

[21] If $\alpha = 0$ the problem reduces the Lasso and $\alpha = 1$ the problem reduces the ridge regression.

Figure 1.3: Constraints of Elastic Nets vs. Semi-Norms



Note: Constraint function of $\sum_{i=1}^{p} |\beta_i|^q$ with $q = 1.2$ (left) and $(1-\alpha)||\boldsymbol{\beta}||_1 + \alpha||\boldsymbol{\beta}||_2^2$ with $\alpha = 0.2$. Based on Figure 3.13 of Hastie et al. (2009)

key difference is the elastic-net constraint has sharper corners, making it possible to use the elastic-net penalty for variable selection. Using the elastic-net penalty also has computational advantages over the ℓ_q penalty (Hastie et al., 2009). A disadvantage of elastic nets is additional parameters α need to be specified.

Jia and Yu (2010) study the selection consistency of Elastic net in the high dimensional setting. They propose the sufficient condition called the **Elastic Irrepresentable Condition**, which is a weaker condition than the Lasso IC as it allows for a stronger degree of correlation between the covariates.

Dantzig Selector. Introduced by Candes and Tao (2007), the Dantzig Selector problem is:

$$\min ||\boldsymbol{\beta}||_1 \quad \text{subject to} \quad ||\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})||_\infty \leq \theta.$$

Bickel et al. (2009) demonstrate that under a sparse scenario, the Lasso and Dantzig Selector exhibit similar theoretical properties. Gautier and Rose (2011) extend the Dantzig Selector to allow for endogenous variables in the high dimensional

setting. This estimator is referred to as Self-Tuning Instrumental Variables.

Square-root Lasso developed by Belloni et al. (2011, 2014), has the objective function

$$\min ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2 + \theta||\boldsymbol{\beta}||_1.$$

The authors show that the optimal penalty for square-root Lasso does not depend on the noise, giving the estimator a practical advantage over standard Lasso. They motivate this variant of Lasso by citing the problem that much of the work on the performance bounds of Lasso relies on the assumption that the penalty term dominates that term $2||\boldsymbol{v}'\boldsymbol{X}||_\infty/n$. Thus, the optimal penalty depends on the noise level $\sigma_{\boldsymbol{v}}$.

Shrinkage GMM changes the OLS objective function to the GMM objective function (Caner, 2009). Fan and Liao (2014) proposed a Shrinkage GMM method that allows for endogenous variables in the high dimensional setting, and this estimator is referred to as Focused Generalised Methods of Moment.

1.4.1.5 Application of Regularization in Spatial Econometrics

Some examples of Lasso-based estimators used in spatial econometrics are Ahrens and Bhattacharjee (2015), which proposes using a 2-step Lasso estimator to estimate a SWM under the identifying assumption the SWM is sparse. Lam and Souza (2016) use a lasso based procedure to estimate SWMs with block structures. Lam and Souza (2020) use adaptive Lasso to estimate a SWM from a linear combination of candidate SWMs also under the assumption of sparsity. Zhu et al. (2010) propose an

adaptive Lasso-based procedure estimated by ML to combine covariate selection with the SEM, and Cai et al. (2019) propose using Lasso estimated by GMM to also to combine covariate selection with the SEM. Cai and Maiti (2020) extents the results of Cai et al. (2019) and look at the difference in the rate of convergence between Lasso and Post Lasso and find Post Lasso performs at least as well as standard Lasso. Jin and Lee (2018) propose using adaptive group Lasso to estimate the MESS model to fix the problem that when coefficients of the endogenous variables and spatial lags of the exogenous variables are zero in the true model, the N2SLS estimates are irregular. Other regularisation methods have also been used in spatial econometrics. For example, Tchuente (2019) proposes using a regularised two-stage least squares (2SLS) estimator to overcome the problem of weak identification of ρ in the SAR model.

Chapter 2

Moran's I Lasso for spatially correlated models

2.1 Introduction

In conventional spatial economic modeling, the researcher is required to specify (1) a spatial weights matrix (SWM),[1] and (2) which parts of the model are spatially correlated. Historically, applied researchers have generally focused more on specifying the SWM rather than the empirical spatial structure (LeSage and Pace, 2014). A standard robustness check in the applied spatial economic literature tests whether the estimates are sensitive to different SWMs. When estimates are found to be sensitive to the choice of SWM, researchers have attributed this sensitivity to the choice of SWM. However, as LeSage and Pace (2014) shows, as long as the SWMs are

[1] A spatial weights matrix is an $n \times n$ matrix that describes the pair-wise relationship between each of the n cross-sectional units.

reasonably well correlated, the results should not be overly sensitive to the choice of SWM. This implies that the sensitivity many researchers observe is driven by misspecification of the spatial economic model rather than the choice of SWM. Thus LeSage and Pace (2014) argue that researchers should focus on specifying the spatial model rather than finding the 'ideal' SWM.

Eigenvector Spatial Filtering (ESF) developed by Griffith (2000, 2003) which uses a subset of eigenvectors from the SWM as 'controls' to filter or approximate any terms involving the SWM in the underlying model. ESF has recently started receiving substantial attention from applied economic researchers.[2] ESF's main advantage over conventional maximum likelihood (ML) and generalised method of moments (GMM) based techniques is the researcher does not need to explicitly specify which parts of the model are spatially correlated or estimate the corresponding spatial parameters. This feature of ESF is desirable for many applied researchers as they can determine if there is some underlying spatial process via a test for spatial correlation but rarely know the 'true' form of this process. So if the applied researchers' aim is to reduce the bias from cross-sectional dependence in the parameter estimates of the covariates, i.e., estimate just the direct effects, then the fact that ESF is agnostic to the underlying spatial process is beneficial.

The critical challenge for EFS is that the spectral decomposition of the $n \times n$ SWM yields n eigenvectors. If the full set of eigenvectors is included, the resulting model will be a high-dimensional linear model, thus estimation by Ordinary Least

[2]Some examples include Patuelli et al. (2011, 2012); Crespo Cuaresma and Feldkircher (2013); Csereklyei and Stern (2015); Oberdabernig et al. (2018); Battisti and Di Vaio (2008); Grimpe and Patuelli (2011).

Squares (OLS) is infeasible.[3] Griffith (2003) argues that only a subset of eigenvectors is necessary to eliminate the spatial/cross-sectional dependence in the dependent variable. A key question is how many/which eigenvectors are required, I refer to this as the ESF eigenvector selection problem. Several methods for this selection problem have been proposed, such as several forward stepwise iterative greedy algorithms where eigenvectors are iteratively added until some user-specified threshold is reached (Griffith, 2000, 2003; Tiefelsdorf and Griffith, 2007). The forward stepwise iterative/greedy algorithms are simply heuristic approximations to the full ESF selection problem, thus, they are necessarily sub-optimal. Seya et al. (2015) assumes that most of the coefficients associated with the eigenvectors are zero (i.e., sparsity) and proposes using an ℓ_1-penalised regression, i.e., Lasso. Given the Lasso estimates are ultimately determined by the chosen tuning parameter, this turns the eigenvector selection problem into a tuning parameter calibration problem. Seya et al. (2015) propose estimating the tuning parameter using conventional K-fold cross-validation (CV) with prediction accuracy of the loss function, as CV is the conventional way to estimate the tuning parameter. However the theoretical results on CV-Lasso assume the cross-sectional units are independent (Chetverikov et al., 2020) which is hard to justify in the context of ESF as the eigenvectors are derived from a matrix that describes the dependence between the observations. Additionally, the goal of ESF is to eliminate spatial correlation patterns, not prediction accuracy. There is therefor no guarantee that CV prediction accuracy will yield consistent eigenvector selection.

I propose an alternative procedure to solve the ESF Lasso tuning parameter

[3]A high-dimensional model is defined as a model with more parameters to estimate than observations, thus the corresponding Gram matrix will be necessarily rank deficient.

calibration problem using information embedded in the Morans i statistic, called Moran's i Lasso (Mi-Lasso). A Moran's i test (Moran, 1950) on then regression residuals where the spatial correlation is ignored will contain information about the overall level of spatial correlation in the dependent variable. Mi-Lasso uses this information to develop a point estimate for the Lasso tuning parameter. The intuition behind Mi-Lasso works as follows when the level of spatial correlation in the residuals is low, only a small set of eigenvectors will be necessary, so a high level of regularisation is required and vice versa for a high level of residual spatial correlation. Mi-Lasso has several advantages; the method is (1) intuitive, (2) theoretically grounded, and (3) substantially faster than Lasso with K-fold cross-validation (CV) or any of the forward stepwise iterative greedy algorithms suggested in the literature.[4] To evaluate the theoretical properties of the Lasso-based estimator, I use some standard spatial regularity conditions and formalise the implicit ESF assumptions, which imply the terms which include the SWM can be approximated by a subset of eigenvectors to derive non-asymptotic bounds for the coefficients of the eigenvectors, and I also assess the additional conditions required for Mi-Lasso to yield consistent eigenvector selection.

Additionally, I study the case where the underlying model includes an unknown number of higher-order lags of the SWM, as the ESF approximation is expected to perform well in this case. Due to the property that the spectral decomposition of the SWM $\boldsymbol{W}$ gives the same matrix eigenvectors as for higher order power of the

[4]Mi-Lasso only requires estimating a single point on a Lasso path, unlike K-fold cross validation which requires estimating K Lasso paths. The larger K, the more computationally demanding the procedure.

SWM $\boldsymbol{W}^p \ \forall \ p \in \mathbb{Z}^+$. Our simulations confirm that Mi-Lasso performs well for a range of levels of spatial correlation and when the data-generating process includes higher-order lags. Regarding computational time, Mi-Lasso is at least an order of magnitude faster than CV-Lasso in the setup considered for a sample size of 10^4.[5]

Our empirical application uses the Boston Housing Dataset to show that Mi-Lasso also performs well in an empirical application. I find Mi-Lasso selects more than triple the number of eigenvectors compared to existing procedures. However, Mi-Lasso gives a bitter fit of the data in terms of adjusted R^2 and has substantially fewer insignificant eigenvectors than other selection procedure considered. Mi-Lasso is also over 60 times faster than the alternative selection procedures for this application.

The rest of this paper is organised as follows, Section 2.2 describes the underlying model. Section 2.3 discusses the statistical aspects of ESF and looks at existing methods for the ESF eigenvector selection problem. Section 2.4 presents the Mi-Lasso procedure. Section 2.5 derives several theoretical results for our proposed procedure. Section 2.6 provides a Monte Carlo study comparing Mi-Lasso to the main existing selection procedures. Section 2.7 tests the proposed method in an empirical application on house prices. Finally, Section 2.8 offers our concluding remarks.

2.2 Underlying model

Consider the following equation where the endogenous $n \times 1$ vector $\boldsymbol{y}$ is specified as a function of an $n \times k$ matrix of exogenous regressors $\boldsymbol{X}$ and follows some spatial

[5]The forward stepwise procedures are infeasible when the sample size is 10^4.

process such as (2.1) and (2.2). However, the exact form of spatial process (which of the spatial parameters are non-zero) including p is unknown.

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{W}^i \boldsymbol{y} \rho_{i,0} + \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\psi}_0 + \boldsymbol{r}, \tag{2.1}$$

$$\boldsymbol{r} = \delta_0 \boldsymbol{W}\boldsymbol{r} + \boldsymbol{v}, \tag{2.2}$$

where $\boldsymbol{W}$ is an $n \times n$ weights matrix of known constants,[6] $\boldsymbol{\beta}_0$ is the $k \times 1$ parameter vector of interest, $\boldsymbol{\psi}_0$, $\rho_{i,0}$'s and δ_0 describe the degree of spatial correlation in each of the k exogenous variables, the dependent variable and error term. Note simpler spatial models can be recovered by setting the spatial parameters $\rho_{i,0}$'s, δ_0, and/or $\boldsymbol{\psi}_0$ equal to zero, and most spatial models set $p = 1$.

The weights matrix $\boldsymbol{W}$, with typical element w_{ij}, describes the spatial or socio-economic relationship between the cross-sectional units. When $w_{ij} \neq 0$, there is a meaningful interaction of units j on unit i. In such cases, unit j is often referred to as a neighbour of unit i. These interactions can stem from various sources, such as spillovers, externalities, geographic location, regulations, technology, government policy, or government expenditure. I further assume $\min_i \sum_{j=1}^{n} w_{ij} > 0$ with probability 1, $w_{ii} = 0$ by construction and $w_{ij} = w_{ji}$. The variables $\boldsymbol{W}\boldsymbol{r}$, $\boldsymbol{W}\boldsymbol{X}$ and $\boldsymbol{W}^i\boldsymbol{y}$ are typically referred to as first order spatial lags of $\boldsymbol{r}$ and $\boldsymbol{X}$ and ith order spatial lags of $\boldsymbol{y}$.

Let N denote the set of observations $N_n = N = \{1, \ldots, n\}$. All variables are

[6] I allow for the $\boldsymbol{W}$ to be normalised by a scalar factor as it allows for the recovery of the original autoregressive parameters (Kelejian and Prucha, 2010) and maintains symmetry.

normalised as the transformed model is estimated by a Lasso-based procedure. For reasons of generality, I allow the elements of $\boldsymbol{u}_n$, $\boldsymbol{y}_n$, $\boldsymbol{W}_n$ and $\boldsymbol{X}_n$ to be dependent on n- that is to form triangular arrays, however, to simplify the notation I omit the n index. Our analysis is conditioned on realised values, thus, matrices such as $\boldsymbol{X}$ and $\boldsymbol{W}$ are viewed as matrices of constants.

I consider higher-order spatial lags as powers of the weights matrix $\boldsymbol{W}$, more recent papers studying the estimation of higher-order spatial models, have generalised the concept of a higher-order spatial lag to allow for p different weights matrices, thus, replacing $\boldsymbol{W}^i$ with $\boldsymbol{W}_i$ in (2.1). Powers of $\boldsymbol{W}$ are viewed as a special case. Some examples of papers looking at the estimation of higher-order spatial models are Lee and Liu (2010); Badinger and Egger (2013); Gupta and Robinson (2015, 2018); Gupta (2019); Baltagi et al. (2022); Han et al. (2021); Gupta (2018, 2021); Gupta and Qu (2022).

I allow the number of lags p to be unknown. Even if p is known, the estimation of such a model is non-trivial, as shown by Blommestein (1985). When the SWM is binary, powers of the SWM can result in the presence of circular and redundant routes. Proper higher-order spatial lags need to have these circular and redundant routes eliminated. [7]

The reduced form for $\boldsymbol{y}$ is

$$\boldsymbol{y} = \boldsymbol{S}_1^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\psi}_0 + \boldsymbol{S}_2^{-1}\boldsymbol{v}),$$

[7]Both Blommestein and Koper (1992) and Anselin and Smirnov (1996) introduced algorithms to construct proper higher-order spatial lags.

if both $\boldsymbol{S}_1 \equiv (\boldsymbol{I} - \sum_{i=1}^{p} \boldsymbol{W}^i \rho_{i,0})$ and $\boldsymbol{S}_2 \equiv (\boldsymbol{I} - \delta_0 \boldsymbol{W})$ are non-singular.

I now make the following assumptions about (2.1) and (2.2)

Assumption 1.

1. *(a)* $\boldsymbol{W}$ *are stochastic real symmetric* $n \times n$ *matrices with* $w_{ii} = 0$. *(b)* $\boldsymbol{S}_1$ *and* $\boldsymbol{S}_2$ *are non-singular for all* n. *(c) The sequences* $\{\boldsymbol{W}\}$, $\{\boldsymbol{S}_1^{-1}\}$ *and* $\{\boldsymbol{S}_2^{-1}\}$ *are uniformly bounded in both row and column sums.*

2. *(a) The* $n \times k$ *matrices of exogenous variables* $\boldsymbol{X}$ *has full column rank (for a large enough* n*) and (b) all the elements of* $\boldsymbol{X}$ *are non-stochastic and uniformly bound in absolute value for all* n.

3. *The elements of the vector of innovations* $\boldsymbol{v}$ *are identically and independently distributed (i.i.d.) sub-Gaussian triangular arrays with* $\mathbb{E}[\boldsymbol{v}] = 0$ *and* $\mathbb{E}[\boldsymbol{v}\boldsymbol{v}'] = \sigma_v^2 \boldsymbol{I}$ *where* $0 < \sigma_v^2 < \infty$. *Additionally, the innovation's fourth moment is assumed finite.*

Assumption 1.1-1.3 are standard assumptions in the spatial econometrics literature (Kelejian and Prucha, 1998, 1999; Lee, 2004). Assumption 1.1 (a) is required for the spectral decomposition. Assumption 1.1 (b) is necessary to ensure the model is complete, and Assumption 1.1 (c) is necessary to limit the degree of dependence in $\boldsymbol{y}$. Given Assumption 1.1 (a) if $\boldsymbol{W}$ is normalised by the largest eigenvalue then invertibility of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ holds if $\sum_{i=1}^{p} |\rho_{i,0}| < 1$ and $|\delta_0| < 1$. Assumption 1.3 requires the errors to be sub-Gaussian, this assumption allows us to derive a probability for the Lasso tuning parameter dominating the noise of the model. The finite fourth moment is needed for the selection consistency proof.

2.3 Eigenvector Spatial Filtering

Spatial filtering aims at eliminating spatial autocorrelation patterns rather than directly estimating any spatial parameters. This section shows how ESF works, formalises its implicit assumptions, the relationship between the Moran's I and ESF and summaries the existing eigenvector selection procedures.

2.3.1 Spectral Decomposition and Spatial Filtering

I now show how eigenvectors from a spectral decomposition of $\boldsymbol{W}$, can be used to spatially filter the model described in Section 2.2.

Add and subtract $\delta_0 \boldsymbol{W}\boldsymbol{r}$ to (2.1),

$$\begin{aligned}\boldsymbol{y} &= \sum_{i=1}^{p} \boldsymbol{W}^i \boldsymbol{y}\rho_{i,0} + \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\psi}_0 + \delta_0 \boldsymbol{W}\boldsymbol{r} + \boldsymbol{r} - \delta_0 \boldsymbol{W}\boldsymbol{r} \\ &= \sum_{i=1}^{p} \boldsymbol{W}^i \boldsymbol{y}\rho_{i,0} + \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\psi}_0 + \delta_0 \boldsymbol{W}\boldsymbol{r} + [\boldsymbol{I} - \delta_0 \boldsymbol{W}]\boldsymbol{r}.\end{aligned}$$

Then substituting in (2.2) gives

$$\begin{aligned}\boldsymbol{y} &= \sum_{i=1}^{p} \boldsymbol{W}^i \boldsymbol{y}\rho_{i,0} + \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\psi} + \delta_0 \boldsymbol{W}\boldsymbol{r} + \boldsymbol{v} \\ &= \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{W}(\sum_{i=1}^{p} \boldsymbol{W}^{i-1} \boldsymbol{y}\rho_{i,0} + \boldsymbol{X}\boldsymbol{\psi}_0 + \delta_0 \boldsymbol{r}) + \boldsymbol{v}. \qquad (2.3)\end{aligned}$$

As $\boldsymbol{W}$ is real and symmetric matrix (by Assumption 1.1 (a)) the spectral decom-

position of $\boldsymbol{W}$ is given in (2.4)

$$\boldsymbol{W} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}', \tag{2.4}$$

$$\boldsymbol{I} = \boldsymbol{E}'\boldsymbol{E} = \boldsymbol{E}\boldsymbol{E}', \tag{2.5}$$

$$\boldsymbol{E}' = \boldsymbol{E}^{-1}, \tag{2.6}$$

where $\boldsymbol{E}$ is an $n \times n$ matrix of the n eigenvectors $\boldsymbol{e}_{i \in N}$ and $\boldsymbol{\Lambda}$ is a $n \times n$ diagonal matrix of the n eigenvalues $(\lambda_{i \in N})$ from $\boldsymbol{W}$. I list some useful orthogonal properties of the decomposition in (2.5) and (2.6).

It is also important to note that the matrix of eigenvectors $\boldsymbol{E}$ of $\boldsymbol{W}$ is also the matrix of the eigenvectors of $\boldsymbol{W}^i \quad \forall\ i \in \mathbb{Z}^+$. The proof is very simple, recursively multiplying (2.4) by $\boldsymbol{W}$ and substituting in (2.4) and note (2.5)

$$\begin{aligned}
\boldsymbol{W}\boldsymbol{W} = \boldsymbol{W}^2 &= \boldsymbol{W}\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}' = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}'\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}' = \boldsymbol{E}\boldsymbol{\Lambda}^2\boldsymbol{E}' \\
&\vdots \\
\boldsymbol{W}^p &= \boldsymbol{E}\boldsymbol{\Lambda}^p\boldsymbol{E}',
\end{aligned}$$

therefor $\boldsymbol{E}$ is the matrix of eigenvectors of $\boldsymbol{W}^p$.

Another way of writing (2.3) is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r}) + \boldsymbol{v}, \tag{2.7}$$

where $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$ is a linear in parameter function of $\boldsymbol{W}$, $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{r}$.

The idea behind Griffith's approach is to use the eigenvectors $\boldsymbol{e}_{i \in N}$ as explanatory variables to control or proxy for $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$. This yields the high dimensional ESF reduced form model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{E}\boldsymbol{\gamma}_0 + \boldsymbol{v}, \tag{2.8}$$

where $\boldsymbol{E}\boldsymbol{\gamma}_0$ can be viewed as a linear approximation of $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$. The key problem with (2.8) is it cannot be estimated consistently by OLS, as (2.8) is a high dimensional linear model. Thus, the OLS assumption that the regressor matrix $\boldsymbol{G} = [\boldsymbol{X}, \boldsymbol{E}]$ has full column rank is violated.[8] This implies the Gram matrix $\boldsymbol{G}'\boldsymbol{G}/n$ will be rank-deficient, and its smallest eigenvalue will be zero.

To handle this problem, I make the following assumptions

Assumption 2.

1. $||\boldsymbol{\gamma}_0||_0 = s < n - k$ *where* $s = s_n$ *is the cardinality of the active set* $\Omega := \text{supp}(\boldsymbol{\gamma}_0)$.

2. $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r}) = \boldsymbol{E}\boldsymbol{\gamma}_0 = \boldsymbol{E}_\Omega\boldsymbol{\gamma}_\Omega$ *where* $\boldsymbol{E}_\Omega$ *is an* $n \times s$ *matrix with columns that correspond to* Ω *and* $\boldsymbol{\gamma}_\Omega$ *the corresponding vector of unknown constants.*

Assumption 2.1 is a weak sparsity assumption, and Assumption 2.2 is required for the ESF approximation to be valid. These strong and untestable assumptions formalise the intuition of Griffith (2000, 2003), who argue only a specific subset of eigenvectors ($\boldsymbol{E}_\Omega$) are related to the dependent variable $\boldsymbol{y}$ and will have non-zero

[8] This is because of $\text{rank}(\boldsymbol{G}) = \text{rank}(\boldsymbol{G}'\boldsymbol{G}) \leq \min(n, (n + k))$.

coefficients. These assumptions imply (2.8) can be reduced to the following low-dimensional equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{E}_\Omega\boldsymbol{\gamma}_\Omega + \boldsymbol{v} = \boldsymbol{G}_\Omega\boldsymbol{\Upsilon}_0 + \boldsymbol{v}, \tag{2.9}$$

where $\boldsymbol{\Upsilon}_0 = [\boldsymbol{\beta}_0, \boldsymbol{\gamma}_\Omega]'$ and $\boldsymbol{G}_\Omega = [\boldsymbol{X}, \boldsymbol{E}_\Omega]$.

In principle, (2.9) can be estimated by OLS. However, as $\boldsymbol{E}_\Omega$ is unknown, this is infeasible in practice. Thus, we now have a selection problem.

2.3.2 Relationship between Moran's i and ESF

Griffith (2000) ESF method is based on the Moran's I statistic for spatial autocorrelation (Moran, 1950). The Moran's i is the most popular test for spatial dependence, the test statistic for the Moran's i (m) on the regression residual $\boldsymbol{M_X y} = \hat{\boldsymbol{u}}$ of $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ where $\boldsymbol{M_X} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{X}'$ is:

$$m = \frac{\boldsymbol{y}'\boldsymbol{M_X}\boldsymbol{W}\boldsymbol{M_X}\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{M_X}\boldsymbol{y}} = \frac{\hat{\boldsymbol{u}}'\boldsymbol{W}\hat{\boldsymbol{u}}}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}, \tag{2.10}$$

where $\boldsymbol{W}$ a $n \times n$ real symmetric SWM.[9]

Substituting in (2.4)

$$m = \frac{\hat{\boldsymbol{u}}'\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}\hat{\boldsymbol{u}}}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}$$

De Jong et al. (1984) showed that the maximum and minimum eigenvalues of

[9]The assumption of symmetry of the elements of $\boldsymbol{W}$ is maintained *w.l.o.g.* since $\hat{\boldsymbol{u}}'\boldsymbol{W}\hat{\boldsymbol{u}} = \hat{\boldsymbol{u}}'[(\boldsymbol{W} + \boldsymbol{W}')/2]\hat{\boldsymbol{u}}$ (Kelejian and Prucha, 2001).

$\boldsymbol{M_X W M_X}$ determine the range of m. Tiefelsdorf and Boots (1995) showed that each of the n eigenvalues of this expression represents a distinct m values and all other possible m values are just linear combinations of these n values (Boots and Tiefelsdorf, 2000).

It is important to note that the numerator of m includes $\boldsymbol{E}'\hat{\boldsymbol{u}}$, which given (2.5) is the OLS coefficient estimate from a regression of $\hat{\boldsymbol{u}}$ on $\boldsymbol{E}$.[10] Griffith (2003) argues that each of the n eigenvectors represents mutually orthogonal spatial patterns and only a subset of eigenvector will be relevant to the model, i.e., in a regression framework only a subset of eigenvectors will have non-zero coefficients.

2.3.3 Existing Selection Procedures

A key question is how can $\boldsymbol{E}_\Omega$ be estimated, i.e. how can we select the relevant eigenvectors? One possible solution to this problem is called Best Subset Selection (BSS). The constrained form of the BSS optimisation problem is

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} ||\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{E\gamma}||_2^2 \quad \text{subject to} \quad ||\boldsymbol{\gamma}||_0 \leq h,$$

where the positive integer h is a regularisation parameter that is fixed in advance.

BSS aims to find the h eigenvectors that produce the best fit in terms of squared error. BSS is a discrete optmisation problem over all subsets of size h and can be re-expressed as a subset sum problem which is NP-complete, so is computationally infeasible to solve (Foster and George, 1994).

[10] $(\boldsymbol{E}'\boldsymbol{E})^{-1}\boldsymbol{E}'\hat{\boldsymbol{u}} = \boldsymbol{E}'\hat{\boldsymbol{u}}$

The first type of procedures proposed were forward stepwise iterative greedy algorithms where eigenvectors are iteratively added till some user-specified threshold is reached (Griffith, 2000, 2003; Tiefelsdorf and Griffith, 2007; Murakami and Griffith, 2019), can be viewed as greedy approximations to BSS.

Griffith (2003) proposed a greedy forward stepwise regression procedure, where eigenvectors are iteratively added to

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{2.11}$$

till the spatial correlation in the OLS residual $\hat{\boldsymbol{u}}$, falls below a pre-specified level. Alternatively, selection criteria based on other statistics such as the adjusted-R^2 or some information criterion such as Akaike Information Criterion or Bayesian Information Criterion have also been suggested (Tiefelsdorf and Griffith, 2007; Murakami and Griffith, 2019).

Tiefelsdorf and Griffith (2007) suggested using the standardised Moran's I as the test for spatial correlation. As the test has good power against a wide array of autoregressive models and different residual distributions (Anselin and Rey, 1991). The definition of the standardised Moran's I statistic (z) on the residual $\hat{\boldsymbol{u}}$ is,[11]

$$z = \left(\frac{m - \mathbb{E}[m]}{\sqrt{\mathrm{Var}(m)}} \right) \tag{2.12}$$

with

$$\mathbb{E}[m] = \frac{tr(\boldsymbol{M_X W M_X})}{n - k}$$

[11]Note the matrix $\boldsymbol{X}$ in the orthogonal projection matrix $\boldsymbol{M_X}$ may also include the selected eigenvectors in Tiefelsdorf and Griffith (2007) procedure.

and

$$\text{Var}(m) = \frac{2\Big((n-k)tr\big((\boldsymbol{M_X W M_X})^2\big) - \big[tr(\boldsymbol{M_X W M_X})\big]^2\Big)}{(n-k)^2(n-k-2)}.$$

The greedy search algorithm work by first iterating over the candidate set of eigenvectors $\boldsymbol{E}_c$ for the eigenvector that minimizes z. The selected eigenvector $\boldsymbol{e}_{i \in N}$ is then permanently removed from $\boldsymbol{E}_c$ and permanently added to the design matrix of (2.11), $\hat{\boldsymbol{u}}$ of this updated regression is then tested to check if (2.13) is satisfied,

$$|z| < tol, \tag{2.13}$$

where *tol* is a pre-specified threshold level of z, which they suggest should be dependent on the sample size n.[12] If (2.13) is satisfied the iterations stop, if not the remaining candidate eigenvector set $\boldsymbol{E}_c$ is then iterated over again to find the eigenvector that minimizes z the most and the process continues until (2.13) is satisfied.

Griffith (2003) suggest that only a subset of candidate eigenvectors $\boldsymbol{E}_c \subseteq \boldsymbol{E}$ needs to be considered in the selection process, based on several criteria. First, if $\boldsymbol{y}$ exhibit positive global spatial autocorrelation the candidate set of eigenvectors should be restricted to eigenvectors with associated eigenvalues greater than zero ($\lambda_i > 0$) as positive eigenvalues are associated with at least weak positive spatial autocorrelation and vice versa. Second eigenvectors with small amounts of spatial variation should be excluded, suggesting a minimum threshold eigenvalue of 0.25, which is related to only approximately 5% of the variation attributed to spatial correlation in the dependent variable.

[12] Tiefelsdorf and Griffith (2007) suggest if $n < 50$ then $tol \approx 1.0$ and if $n \approx 500$ then $tol \approx 0.1$.

These forward stepwise procedures through intuitive have several key disadvantages. There is a lot for the user to decide, such as, what statistic or information criterion to use, what threshold cut-off to use, which eigenvectors to include/exclude in the original candidate set, and in which order to add the eigenvectors. These greedy algorithms could also be viewed as ad-hoc/data mining and the estimated models may fall victim to over-fitting. Simply adding the eigenvectors in order of magnitude of their corresponding eigenvalue is another possible strategy, however, as the problem is an approximation to the BSS problem, is it thus also sub-optimal, so there is no guarantee the solution this yields will be correct. Computational time is another serious issue with these sequential methods and the problem becomes more acute when n is large. Computational time can be saved by limiting $\boldsymbol{E}_c$ with one or more of the rules of thumb proposed by Griffith (2003) and Tiefelsdorf and Griffith (2007). However, this is no guarantee these rules will consistently recover $\boldsymbol{E}_\Omega$.

The use of Lasso (Tibshirani, 1996) was proposed by Seya et al. (2015). Lasso can be viewed as a convex approximation of BSS as it swaps the ℓ_0 norm of BSS for the ℓ_1 norm, which is the smallest norm (or semi-norm) that is convex and still adequate for variable selection.[13] As Lasso shrinks many of the coefficients to 'exactly' zero, the procedure can be used for variables selection (Hastie et al., 2009).

Seya et al. (2015) proposed using Lasso to solve the ESF eigenvector selection problem, under the assumption the parameter vector $\boldsymbol{\gamma}_0$ is sparse and the matrix of regressors $\boldsymbol{X}$ has full column rank, so only the $\boldsymbol{\gamma}$ vector is penalised. The penalised

[13]Bridge regressions which uses the ℓ_q with $q \in (0, 1)$ is also adequate for variable selection (Huang, Horowitz and Ma, 2008) however the optimisation problem is non-convex and thus, difficult to solve.

form of Seya et al. (2015) proposed Lasso estimator is

$$[\hat{\boldsymbol{\beta}}_\theta, \hat{\boldsymbol{\gamma}}_\theta] \in \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 + \theta||\boldsymbol{\gamma}||_1\}, \tag{2.14}$$

where $\theta > 0$ is the Lasso regularization or tuning parameter. Equation (2.14) defines a family of estimators indexed by the tuning parameter θ, as the hyperparameter θ will ultimately determine which eigenvectors the Lasso selects.

Seya et al. (2015) proposed estimating $\hat{\theta}$ using prediction accuracy as the criterion. More specifically, they used the k-fold cross-validation (CV) combined with Brent algorithm (Brent, 1973), outlined in Algorithm 1. The Brent algorithm is a root-finding algorithm that allows for the optimisation to be non-convex, the algorithm first tries inverse quadratic interpolation in an attempt to achieve faster convergence which works well if the optimisaition is convex. If it is non-convex and inverse quadratic interpolation fails, (slower) linear interpolation is used instead. CV using the Brent algorithm is the most time-consuming part of the Seya et al. (2015) Lasso procedure.

The theoretical results on CV-Lasso, hinge on the assumption that the cross-sectional units are independent (Chetverikov et al., 2020), which is hard to justify for ESF as the eigenvectors are derived from a matrix that describes the dependence between the observations. CV procedures do exist for cross-sectionally dependent data but they need to be carefully designed, for example see Li et al. (2020).

Some other methods have also been proposed Pace et al. (2013) suggest simply including the first j eigenvectors (sorted by eigenvalue magnitude) where j is simply based on the sample size. Given this fixed rule, Pace et al. (2013) finds the quality

1. Initialize $\hat{\theta} \in (\theta_{min}, \theta_{max})$
2. Divide the n observations randomly into K approximately equal groups (folds) denoted $F_1, \ldots, F_K$.
3. For $k = 1, \ldots, K$:
   (a) F_k is the validation data set and the remaining $K - 1$ groups form the training data set.
   (b) Using $\hat{\theta}$ estimate $\hat{\boldsymbol{\beta}}_{-k}(\hat{\theta})$ and $\hat{\boldsymbol{\gamma}}_{-k}(\hat{\theta})$ where $\hat{\boldsymbol{b}}_{-k}$ is an estimate of $\boldsymbol{b}$ using the training dataset.
4. Calculate
$$CV(\hat{\theta}) = \sum_{k=1}^{K} \sum_{i \in group\ k} \{y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{-k}(\hat{\theta}) - \boldsymbol{e}_i' \hat{\boldsymbol{\gamma}}_{-k}(\hat{\theta})\}$$
   where $\boldsymbol{x}_i$ and $\boldsymbol{e}_i$ are vectors of explanatory variables and eigenvectors associated with observation i.
5. Update $\hat{\theta}$ using Brent (1973) algorithm for minimization without derivatives.
6. Repeat steps 3-5 until convergence $\hat{\theta}_0$

Algorithm 1: Seya et al. (2015) algorithm

of the ESF approximation is sensitive to the underlying spatial processes. The bias in OLS estimates of $\boldsymbol{\beta}$ when the spatial process is ignored will be determined by a combination of, the type of spatial processes, the level of spatial correlation, and the SWM. Chun et al. (2016) argue more eigenvectors are needed when the level of spatial correlation is high compared to when the level of spatial correlation is low, thus, simple rules based on for example sample size may result in a sub-optimal set of eigenvectors being selected.

Chun et al. (2016) through extensive simulations develop equation (2.15) as a rule to select the optimal number of eigenvectors.

$$w = \frac{n_{pos}}{1 + \exp\left[2.1480 - \left(6.1808(m+0.6)^{0.1742}\right)/n_{pos}^{0.1298} + 3.3534/(m+0.6)^{0.1742}\right]}, \tag{2.15}$$

where n_{pos} denotes the number of eigenvectors that exhibit positive spatial correlation (eigenvectors with positive eigenvalues). Equation (2.15) was generated from a limited simulation that assumed the DGP has just spatial autoregressive disturbances, Chun et al. (2016) do not evaluate how their rule performs when the DGP follows some other spatial process.

2.4 Moran's i Lasso

The Lasso estimates are ultimately determined by tuning parameter θ. Supposing $\theta = 0$, the Lasso solution reduces to the OLS solution, whereas when θ is sufficiently large, the penalised parameter vector is shrunk to zero (no eigenvectors selected).

More moderate values of θ will result in some parameters being shrunk towards zero and some to precisely zero. As outlined above ESF has a particular goal, to eliminate any spatial correlation patterns in a linear regression framework. Information about these patterns will be contained in simple regression residuals $\hat{\boldsymbol{u}}$. I propose using this information to determine a point estimate for θ.

It seems reasonable to assume that when the level of spatial correlation in the residuals is low, only a small set of eigenvectors is necessary. Thus, a high level of regularization (value of θ) is required. In contrast, when the level of spatial correlation is high, a large set of eigenvectors will be necessary. Thus, a low level of regularization (value of θ) is required. For the spatial correlation test, I propose using the standardised Moran's I (z), as the test can be used for small samples (Kelejian and Piras, 2017) and has good power against a wide array of autoregressive models and different residual distributions (Anselin and Rey, 1991). As z gives a large value when the correlation is high and small values when the correlation is low, I propose using the inverse of the absolute value of z from the residuals of (2.11) as a point estimate of θ

$$\theta = \frac{1}{Z^a} \tag{2.16}$$

where $Z = |z|$, z is defined in (2.12), and a is a positive constant, $\forall z \neq 0$. I allow $Z_n = Z$ to be dependent on n but suppress the index. This proposed estimator is called Moran i Lasso (Mi-Lasso) and is outlined in Algorithm 2.

As Lasso is a shrinkage estimator, it induces a downward bias on the estimated non-zero coefficients. Post-Lasso (pLasso) uses the Lasso estimator as just a selection procedure (assuming Lasso selects the correct variables), and then OLS is applied to

1. Decompose the SWM to get the candidate set of Eigenvectors $\boldsymbol{E}$.
2. Estimate simple residuals $\hat{\boldsymbol{u}} = \boldsymbol{M}_X \boldsymbol{y}$ where $\boldsymbol{M}_X = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and calculate corresponding the absolute standardised Moran's i of $\hat{\boldsymbol{u}}$ denoted Z
3. Given $a \in \mathbb{N}^+$ estimate

$$[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}] \in \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 + |\frac{1}{Z^a}|||\boldsymbol{\gamma}||_1\} \tag{2.17}$$

Use the Lasso or post-Lasso estimates of (2.17)

Algorithm 2: Mi-Lasso Algorithm

the model selected by Lasso.[14] The Morans' i Post-Lasso (Mi-pLasso) estimator is defined as

$$[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}] = \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 \quad \text{subject to} \quad \text{supp}(\boldsymbol{\gamma}_0) = \text{supp}(\hat{\boldsymbol{\gamma}}_{\frac{1}{Z}}).$$

Another benefit of Post-Lasso is that it provides an easy way to calculate standard errors (assuming the selection is correct).

2.5 Theoretical results

To focus the proceeding analysis on the parameter vector $\boldsymbol{\gamma}$, I use the Frisch-Waugh-Lowell (FWL) partial regression theorem to partial out the $\boldsymbol{X}$ matrix. Tibshirani and Taylor (2011) and Yamada (2017) showed that the FWL theorem could be used in a low-dimensional Lasso setting. Lemma 1 shows that the FWL theorem can also be applied to the high-dimensional case of Mi-Lasso

[14] For formal results on Post-Lasso, see Belloni and Chernozhukov (2013).

Lemma 1. *Consider the following two Lasso regressions:*

$$[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}] = \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 + \frac{1}{Z^a}||\boldsymbol{\gamma}||_1\}, \tag{2.18}$$

$$[\tilde{\boldsymbol{\gamma}}] = \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\gamma||_2^2 + \frac{1}{Z^a}||\boldsymbol{\gamma}||_1\}, \tag{2.19}$$

where $\boldsymbol{X}$ *is an* $n \times k$ *matrix,* $\boldsymbol{E}$ *is an* $n \times n$ *matrix,* $\tilde{\boldsymbol{y}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$, $\tilde{\boldsymbol{E}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{E}$ *with* $\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. *Then if Assumption 1.2 holds* $\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}$

The proof is provided in Section 2.9.1.

I now introduce the following additional notation in the design. Without loss of generality, let $\boldsymbol{C}_{\Omega\Omega} = n^{-1}\tilde{\boldsymbol{E}}'_{\Omega}\tilde{\boldsymbol{E}}_{\Omega}$, $\boldsymbol{C}_{\Omega\grave{\Omega}} = n^{-1}\tilde{\boldsymbol{E}}'_{\Omega}\tilde{\boldsymbol{E}}_{\grave{\Omega}}$, $\boldsymbol{C}_{\grave{\Omega}\Omega} = n^{-1}\tilde{\boldsymbol{E}}'_{\grave{\Omega}}\tilde{\boldsymbol{E}}_{\Omega}$ and $\boldsymbol{C}_{\grave{\Omega}\grave{\Omega}} = n^{-1}\tilde{\boldsymbol{E}}'_{\grave{\Omega}}\tilde{\boldsymbol{E}}_{\grave{\Omega}}$ where $\tilde{\boldsymbol{E}}_{\Omega}$ is an $n \times s$ matrix with columns corresponding to the active set Ω. $\grave{\Omega}$ is the complementary set and the $n \times q$ matrix $\tilde{\boldsymbol{E}}_{\grave{\Omega}}$ is defined accordingly with $q_n = q = s - n$. Now the (re-scaled) Gram matrix $\boldsymbol{C}_n = \boldsymbol{C} = n^{-1}\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}$ expressed in block-wise form

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{\Omega\Omega} & \boldsymbol{C}_{\Omega\grave{\Omega}} \\ \boldsymbol{C}_{\grave{\Omega}\Omega} & \boldsymbol{C}_{\grave{\Omega}\grave{\Omega}} \end{bmatrix}.$$

Similarly let $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_{\Omega}, \boldsymbol{\gamma}_{\grave{\Omega}}]' = [\gamma_1, \ldots, \gamma_s, \gamma_{s+1}, \ldots, \gamma_n]'$.

2.5.1 Non-asymptotic bounds

This section produces performance bounds for the Mi-Lasso estimates of $\boldsymbol{\gamma}$. Given the high-dimensional structure of ESF, the Gram matrix $\boldsymbol{G}'\boldsymbol{G}/n$ is singular. This implies its minimum eigenvalue will be zero. However, as Bickel et al. (2009) showed for Lasso, only the appropriate sub-matrix of the Gram matrix needs to have positive

and finite eigenvalues called the restricted eigenvalue (RE) condition (Assumption 3).

Assumption 3. *Let $\bar{b}$ and t be positive constants and Ω denote the active set. Then the restricted eigenvalue condition holds for $\tilde{\boldsymbol{E}}$, as $n \to \infty$ if*

$$\tau_{min} := \min_{\mathcal{C}(\Omega,\bar{b})} \frac{||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2}{\sqrt{n}||\boldsymbol{\Delta}||_2} \geq t > 0, \tag{2.20}$$

where

$$\mathcal{C}(\Omega,\bar{b}) = \{\boldsymbol{\Delta} \in \mathbb{R}^n : ||\boldsymbol{\Delta}_{\hat{\Omega}}||_1 \leq \bar{b}||\boldsymbol{\Delta}_{\Omega}||_1, \delta \neq 0\} \tag{2.21}$$

and $\boldsymbol{\Delta} = \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$.

Assumption 3 requires that $\boldsymbol{\Delta}$ lies within the restricted set (2.21). As $\boldsymbol{\Delta}$ is the difference between the estimate $\tilde{\boldsymbol{\gamma}}$ and the true parameter $\boldsymbol{\gamma}_0$, the restricted eigenvalue bounds the minimum change in the prediction norm from a deviation $\boldsymbol{\Delta}$ within the restricted set $\mathcal{C}(\Omega,\bar{b})$ relative to the norm of the deviation on the true support $\boldsymbol{\Delta}_{\Omega}$.

By combining Assumptions 1 and 2 with the RE condition, and treating $\boldsymbol{X}$ and $\boldsymbol{E}$ as constants (realisations) I can now establish the ℓ_1 and ℓ_2 parameter norm bounds and the ℓ_2 prediction norm bound for the Mi-Lasso estimates of $\boldsymbol{\gamma}$.

Theorem 1. *Suppose Assumption 1-2 and Assumption 3 holds for $\bar{b} = \frac{b+1}{b-1}$ for some $b \geq 1$ and the regularization parameter satisfies $\frac{1}{Z^a} \geq b2\sqrt{\frac{4\sigma_v^2 \log n}{n}}$ with probability tending to one as $n \to \infty$, then:*

$$||\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 \leq \frac{\left(\frac{1}{b}+1\right)s}{\tau_{min}^2 Z^a n}, \tag{2.22}$$

$$||\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_2 \leq \frac{\left(\frac{1}{b}+1\right)\sqrt{s}}{\tau_{min}^2 Z^a n}, \tag{2.23}$$

$$\frac{1}{\sqrt{n}}||\tilde{\boldsymbol{E}}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)||_2 \leq \frac{\left(\frac{1}{b}+1\right)\sqrt{s}}{\tau_{min} Z^a n}. \tag{2.24}$$

The proof is provided in Section 2.9.1.

The three convergence rates presented in Theorem 1 depend on the number of eigenvectors with non-zero coefficients, the sample size, and Z. They also require that the tuning parameter dominants the noise of the model, by assuming the errors are sub-Gaussian (Assumption 1.3) I prove the probability of this event occurring goes to one as $n \to \infty$ (see proof for further details).

Corollary 1. *If the condition of Theorem 1 are satisfied and* $s/Z^a n = o_p(1)$ *then the bounds* (2.22)-(2.24) *are* $o_p(1)$ *as* $n \to \infty$.

Corollary 1 is satisfied if $Z = O_p(1)$ and s grows at a rate slower than n, which is reasonable as Z is a measure of correlation and is satisfied by Assumption 4.4.

2.5.2 Consistent Eigenvector Selection

This section shows the conditions required for Mi-Lasso to consistently selects the non-zero and zero elements in $\boldsymbol{\gamma}$. Following Zhao and Yu (2006), I say that $\tilde{\boldsymbol{\gamma}} =_s \boldsymbol{\gamma}_0$ if and only if $\text{sign}(\tilde{\boldsymbol{\gamma}}) = \text{sign}(\boldsymbol{\gamma}_0)$ where $\text{sign}(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to zero. I now define selection consistency for Mi-Lasso as

Definition 1. *(Zhao and Yu, 2006) The Mi-Lasso estimates of* $\boldsymbol{\gamma}$ *are selection consistent if*

$$\lim_{n\to\infty} P(\tilde{\boldsymbol{\gamma}} =_s \boldsymbol{\gamma}_0) = 1.$$

The following assumptions are required to prove sign consistency of Mi-Lasso.

Assumption 4. *There exists* $M_1, M_2, M_3 > 0$, $0 \leq c_1 < c_2 \leq 1$ *and a vector of postive constants* $\boldsymbol{\nu}$*, the following holds:*

1.

$$\frac{1}{n}\tilde{\boldsymbol{e}}_i'\tilde{\boldsymbol{e}}_i \leq M_1 \ \ \forall i,$$

2.

$$\boldsymbol{\alpha}'\boldsymbol{C}_{\Omega\Omega}\boldsymbol{\alpha} \geq M_2 \ \ \forall \ ||\boldsymbol{\alpha}||_2^2 = 1,$$

3.

$$n^{\frac{1-c_2}{2}} \min_{i=1,\ldots,s} |\gamma_i| \geq M_3,$$

4.

$$s = O(n^{c_1}),$$

5.

$$|\boldsymbol{C}_{\grave{\Omega}\Omega}(\boldsymbol{C}_{\Omega\Omega})^{-1}\operatorname{sign}(\boldsymbol{\gamma}_\Omega)| \leq \mathbf{1} - \boldsymbol{\nu}.$$

Assumption 4.1 is a normalisation of the transformed eigenvectors. Assumption 4.2 bounds the eigenvalue of the eigenvectors with non-zero coefficients from below, so the inverse of $\boldsymbol{C}_{\Omega\Omega}$ is well behaved. Assumption 4.3 and Assumption 4.4 are important as they ensure convergence in the high dimensional space as $n \to \infty$. Assumption 4.3 ensure there is a difference of size n^{c_2} between the decay rate of $\boldsymbol{\gamma}_\Omega$ and $\sqrt{n}$. Preventing the estimates from being dominated by the disturbance terms, as the disturbance terms aggregate at a rate of $n^{-1/2}$. Assumption 4.4 is a sparsity

assumption that requires the square root of the size of the true model $\sqrt{s}$ to increase at a slower rate than the rate difference, and this prevents the Lasso estimation bias from dominating the model parameters.

Assumption 4.5 (assuming $\boldsymbol{C}_{\Omega\Omega}$ is invertible) shows the Irrepresentable Condition (IC), which is the necessary condition for the consistency of Mi-Lasso selection, the inequality holds element-wise. The IC requires the correlation between the relevant and irrelevant eigenvectors to be zero or weak. In the Mi-Lasso framework, this is likely to be satisfied as the columns of $\boldsymbol{E}$ are mutually orthogonal. However, it is not guaranteed the columns of $\tilde{\boldsymbol{E}}$ will also be mutually orthogonal, as the eigenvectors are projected into the column space of $\boldsymbol{X}$. Unfortunately, in practice, the IC is impossible to verify as we do not know the true parameter vector $\boldsymbol{\gamma}_0$.

The following proposition places a lower bound on the probability of Mi-Lasso picking the true model, which quantitatively relates to the probability of Lasso selecting the correct model. Proposition 1 is a modification of Proposition 1 in Zhao and Yu (2006).

Proposition 1. *Assume Assumption 1, 2 and 4.5 holds for some* $\boldsymbol{\nu} > 0$*, then*

$$\mathbb{P}\left(\tilde{\boldsymbol{\gamma}} =_s \boldsymbol{\gamma}_0\right) \geq \mathbb{P}(A \cap B),$$

for

$$\begin{aligned} A =& \{||(\boldsymbol{C}_{\Omega\Omega})^{-1}\boldsymbol{z}_{\Omega}|| < \sqrt{n}(|\boldsymbol{\gamma}_{\Omega}| - \frac{1}{2Z^a n}||(\boldsymbol{C}_{\Omega\Omega})^{-1}\operatorname{sign}(\boldsymbol{\gamma}_{\Omega})||)\}, \\ B =& \{||\boldsymbol{C}_{\grave{\Omega}\Omega}(\boldsymbol{C}_{\Omega\Omega})^{-1}\boldsymbol{z}_{\Omega} - \boldsymbol{z}_{\grave{\Omega}}|| \leq \frac{1}{2Z^a\sqrt{n}}\boldsymbol{\nu}\}, \end{aligned}$$

where $\boldsymbol{z}_{\Omega} = \frac{1}{\sqrt{n}}\tilde{\boldsymbol{E}}'_{\Omega}\boldsymbol{v}$ *and* $\boldsymbol{z}_{\grave{\Omega}} = \frac{1}{\sqrt{n}}\tilde{\boldsymbol{E}}'_{\grave{\Omega}}\boldsymbol{v}$.

The proof is provided in Section 2.9.1.

Proposition 1 shows that the measure of spatial correlation Z determines the size of the trade-off between events A and B. A higher level of spatial correlation will lead to larger A but smaller B; this makes Mi-Lasso more likely to select irrelevant eigenvectors. In contrast, a larger ν_i has no impact on A but leads to a larger B. So when IC holds with a large ν_i, Mi-Lasso is more likely to select the correct model.

Theorem 2. *Assuming Assumption 1, 2 and 4 hold, and* $c_2 - c_1 = 0.5$. *Given* $s+q = n$ *implies Mi-Lasso is sign consistent for all* $\frac{1}{Z^a}$ *that satisfy* $\frac{1}{Z^a\sqrt{n}} = o_p(n^{\frac{c_2-c_1}{2}}) = o_p(n^{\frac{1}{4}})$ *and* $\frac{1}{n^3 Z^{4a}} \to \infty$, *we have*

$$\mathbb{P}\left(\tilde{\boldsymbol{\gamma}} =_s \boldsymbol{\gamma}_0\right) \geq 1 - O(n^3 Z^{4a}) \to 1 \quad as\ n \to \infty.$$

The proof is provided in Section 2.9.1.

Theorem 2 shows that Mi-Lasso is consistent at selecting the true model if the $4th$ moment of the errors is finite (Assumptions 1.3), Assumptions 1-4 hold and the difference between c_2 and c_1 is 0.5. The greatest difference (between c_2 and c_1) for which Mi-Lasso is consistent is 0.5, smaller differences can also yield consistency, but this would require higher order moments of the errors to be finite. For example, if we assume the $6th$ or $8th$ moment is finite, the difference would need to be 1/3 or 0.25 for Mi-Lasso to be consistent (see proof for further details).

2.6 Monte Carlo Study

To evaluate the finite sample performance of Mi-Lasso and compare it to the main existing selection procedures, I conduct two Monte Carlo exercises where the DGP is,

$$\begin{aligned} \boldsymbol{y} =& \sum_{i=1}^{p} \boldsymbol{W}^i \boldsymbol{y} \rho_i + \beta \boldsymbol{x} + \psi \boldsymbol{W} \boldsymbol{x} + \boldsymbol{v}, \\ \boldsymbol{x} \sim& N(0, \boldsymbol{I}),\ \boldsymbol{v} \sim N(0, \boldsymbol{I}). \end{aligned} \tag{2.25}$$

In both simulations, I set the 'true' parameter value of $\beta = 1$ and $\psi = 0.8$. The elements of $\boldsymbol{W}$, denoted w_{ij}, are independent draws from a Bernoulli distribution with success probability $p = \mu/n$ for some constant $\mu < \infty$, $w_{ii} = 0$ and $w_{ij} = w_{ji}$. By this construction, each unit's expected number of links equals μ, and I set $\mu = 4, 8$. Each SWM is normalised by the maximal of the row (or column) sum, and the eigenvectors are from the normalised SWM. Sample sizes considered are $n = 100, 250, 500$, and I run 1000 replication.

In setup A, I set $p = 1$ so I can evaluate how the method performs with different levels of spatial correlation. In setup B, I set $p = 3$ to check that ESF still performs well in the presence of higher-order spatial lags.

Estimators I compare:

- Mi-Lasso$a$ - Mi-Lasso with $a = 1$ or $a = 2$
- Mi-pLasso$a$ - OLS with the selected eigenvector from Mi-Lasso with $a = 1$ or $a = 2$

- CV-Lasso - Lasso algorithm outlined in Seya et al. (2015)
- CV-pLasso - OLS with the selected eigenvector from CV-Lasso
- FstepZ - forward stepwise algorithm outlined in Tiefelsdorf and Griffith (2007) with a stopping rule $z = 0.1$.

Note an oracle estimator is not possible in the ESF set up, as an oracle estimator requires knowledge of $\text{supp}(\boldsymbol{\gamma}_0)$ which is unknown.

For setup A I consider only positive spatial correlation as this is the most common and set $\rho_1 \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ to see how the selection procedures behave as the level of spatial correlation increases.

Figure 2.1-2.2 shows the bias, MSE, and the number of selected eigenvectors for setup A when $\mu = 4$ and $\mu = 8$. These show the selection behavior of each of the estimators considered is different. CV-Lasso appears to select a similar number of eigenvectors at all the levels of spatial correlation considered. In contrast, FstepZ selects more eigenvectors when the spatial correlation level is lower than high for small sample sizes. Mi-Lasso2 ($a = 2$) behavior is as expected from the intuition of the procedure, selecting a small set of eigenvectors when the level of spatial correlation is low and a large set when the level is high. However, Mi-Lasso1 ($a = 1$) does not select any eigenvectors at any of the levels of spatial correlation is less than 0.7, and very few at higher levels of spatial correlation, implying the level of regularisation is too high.

The Lasso estimators have a smaller bias and MSE than the post-Lasso (pLasso) estimators. When the level of spatial correlation is high Mi-Lasso2 has the best

Figure 2.1: Bias and MSE of β and the number of selected eigenvectors, setup A and $\mu = 4$

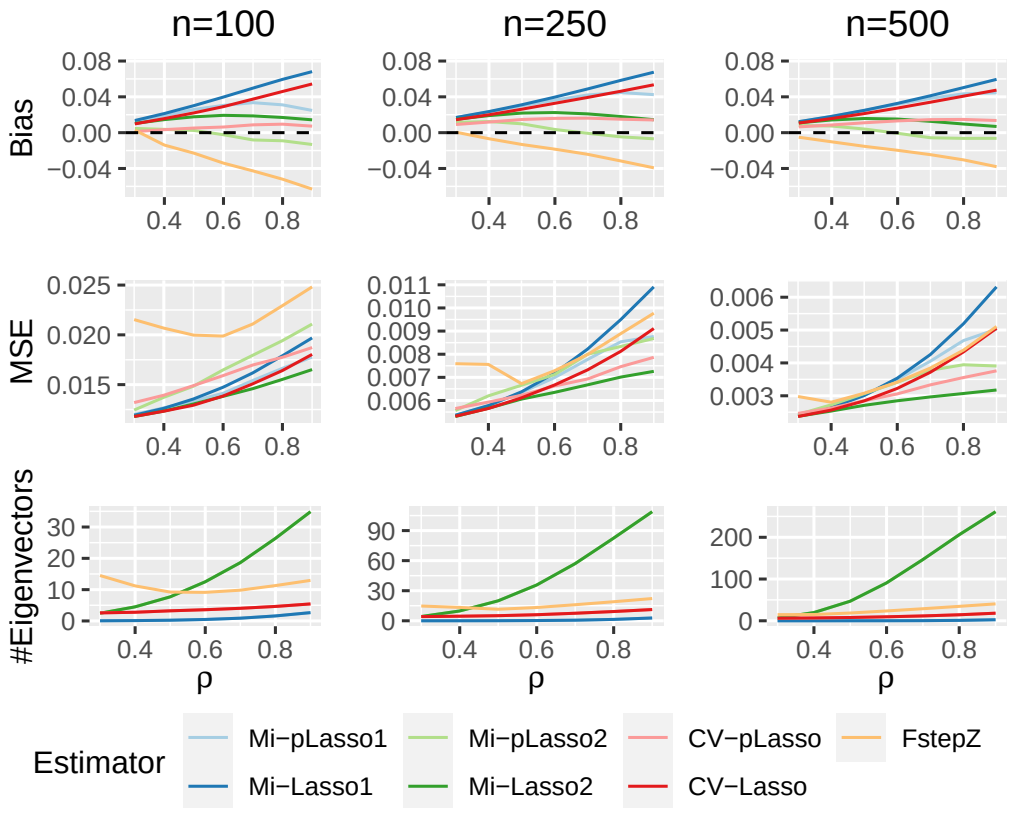
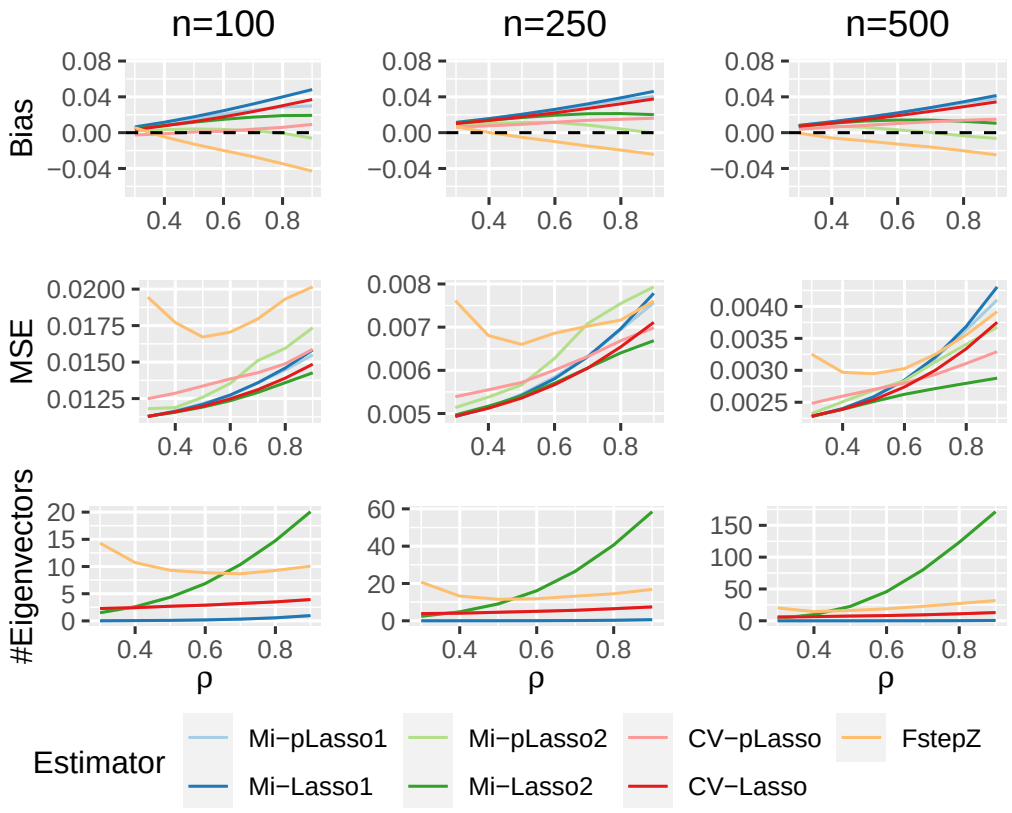


Figure 2.2: Bias and MSE of β and the number of selected eigenvectors, setup A and $\mu = 8$



performance in terms of MSE and performs comparably to the other estimator in terms of bias. Notably, FstepZ has the largest MSE when the sample size is small (100), and also when the level of spatial correlation is low for other sample sizes. FstepZ performance in terms of Bias and MSE improves as the sample size increases and the SWM becomes more sparse. Mi-Lasso1 generally has the worst performance in terms of bias and in terms of MSE for large samples and high spatial correlation, this is likely due to the level of regularisation being too strong. Generally, in terms of bias, the estimators diverge as the level of spatial correlation increases, this is because the bias is determined by an interaction between the level of ρ and the structure of the SWM. Thus, for a given SWM, the larger ρ the larger the bias, so mistakes/variation in selection can have a larger effect.

For setup B, I set $\rho_1 = 0.6$, $\rho_2 = 0.4$ $\rho_3 = 0.5$. Table 2.1 shows the bias, MSE, and the number of selected eigenvectors for setup B. This table confirms that ESF can work well in the presence of higher-order spatial lags. Again I can see here that Mi-Lasso1 selects at most two eigenvectors (in most cases zero) and thus has a larger bias and MSE relative to Mi-Lasso2, which selects the largest set of eigenvectors for a give μ and sample size. In contrast, Mi-Lasso2 always has a smaller bias than CV-Lasso and a comparable MSE. Mi-pLasso2 always has a smaller bias than CV-pLasso and performs comparably in terms of MSE to the other estimators. FstepZ generally performs better in terms of bias and MSE when the SWM is denser ($\mu = 8$) compared to a more sparse SWM ($\mu = 4$).

Table 2.2 shows the computational times of the different estimators used in the simulations. These results show Mi-Lasso2 is the fastest procedure, CV-Lasso is

Table 2.1: Bias, MSE and the number of selected eigenvectors for setup B

n	Estimator	β	No Evecs	β	No Evecs
100	FstepZ	-0.029(0.02)	10	-0.015(0.017)	8
100	CV-Lasso	0.058(0.018)	5	0.037(0.014)	3
100	CV-pLasso	0.023(0.018)	5	0.015(0.015)	3
100	Mi-Lasso1	0.071(0.019)	2	0.047(0.015)	1
100	Mi-pLasso1	0.041(0.017)	2	0.036(0.015)	1
100	Mi-Lasso2	0.032(0.016)	27	0.03(0.014)	12
100	Mi-pLasso2	0.01(0.019)	27	0.014(0.015)	12
250	FstepZ	-0.014(0.008)	16	-0.005(0.007)	13
250	CV-Lasso	0.055(0.009)	9	0.036(0.007)	6
250	CV-pLasso	0.027(0.008)	9	0.023(0.007)	6
250	Mi-Lasso1	0.066(0.01)	1	0.042(0.007)	0
250	Mi-pLasso1	0.053(0.009)	1	0.04(0.007)	0
250	Mi-Lasso2	0.032(0.008)	73	0.031(0.007)	26
250	Mi-pLasso2	0.013(0.008)	73	0.019(0.007)	26
500	FstepZ	-0.014(0.003)	27	-0.007(0.003)	21
500	CV-Lasso	0.047(0.005)	13	0.032(0.003)	9
500	CV-pLasso	0.025(0.004)	13	0.021(0.003)	9
500	Mi-Lasso1	0.055(0.006)	1	0.036(0.004)	0
500	Mi-pLasso1	0.051(0.005)	1	0.036(0.004)	0
500	Mi-Lasso2	0.024(0.003)	164	0.023(0.003)	75
500	Mi-pLasso2	0.008(0.004)	164	0.011(0.003)	75
-	-	$\mu = 4$	-	$\mu = 8$	-

Note: Bias (MSE)

the second fastest, and FstespZ is the slowest procedure for a given sample size. Comparing Mi-Lasso2 to CV-Lasso, I find Mi-Lasso2 is up to 17.70 times faster. The most substantial computational gains are found when the sample size is 500, but even when the sample size is very large (10,000), Mi-Lasso2 is still 4.64 times faster than CV-Lasso. FstepZ is the slowest estimator. However, computational time can be saved by reducing the candidate set of eigenvectors (but all estimators would

Table 2.2: Computation Time and Sample Sizes

Sample Size	Mi-Lasso2	CV-Lasso	FstepZ
250	1 (0.12)	12.08 (1.45)	27.08 (3.25)
500	1 (0.30)	17.70 (5.31)	107.63 (32.29)
1000	1 (2.74)	8.70 (23.85)	321.50 (880.91)
2000	1 (23.11)	9.20 (212.51)	759.12 (17543.24)
10000	1 (3026.28)	4.64 (14045.90)	-

Note: Relative computational time, figures in parenthesis are time in seconds. All procedures exclude eigen-decomposition. The DGP of y is (2.25) with $p = 1$, $\beta = 1$, $\rho_1 = 0.9$ and $\mu = 4$. FstepZ is the forward stepwise algorithm outlined in Tiefelsdorf and Griffith (2007)

benefit in terms of computational time if such reductions are applied).

The key findings from these simulation results in this setup are (1) the Mi-Lasso procedure when $a = 2$ gives a much better performance in terms of both bias and MSE than $a = 1$. (2) the Lasso estimators perform better in terms of MSE than the post-Lasso estimators when the level of spatial correlation is high. (3) the post-Lasso estimators perform better in terms of bias than the Lasso estimators when the level of spatial correlation is high. (4) ESF can handle higher-order spatial lags. (5) Mi-Lasso is substantially faster than both CV-Lasso and FstepZ.

2.7 Empirical Application - Boston Housing Dataset

I now compare the ESF selection procedures using the Boston Housing Dataset, which was first used by Harrison and Rubinfeld (1978) to evaluate the relationship between house prices and demand for clean air. Gilley and Pace (1996) later revisited the dataset when they noted the high spatial correlation in the dataset and proposed

estimating a spatial error model instead. However, as there is no guarantee theirs is the correct specification, and given that the researcher is only concerned with the direct effect (just control for the spatial part of the model), ESF is an appropriate methodology.

Table 2.3: Variables used in Boston housing application

Variable	Description
p	Median values of owner-occupied housing in thousands of U.S. dollars
crim	Per capita crime
zn	Proportion of residential land zoned for lots over 25,000 ft^2 per town
indus	Proportion of non-retail business acres per town
cr	An indicator: 1 if tract borders Charles River; 0 otherwise
nox	Nitric oxide concentration (parts per 10 million) per town
rm	Average number of rooms per dwelling
age	Proportion of owner-occupied units built prior to 1940
dis	Weighted distance to five Boston employment centers
rad	Index of accessibility to radial highways per town
tax	Property-tax rate per $US10,000 per town
ptr	Pupil–teacher ratio per town
black	Percentage of blacks
lsp	Percentage of lower status population

The dataset includes 508 census tracts (spatial units). Table 2.3 describes the variables used in the analysis. The eigenvectors are from a binary SWM where the tracts are connected if they share a border, and SWM is normalised by the maximal of the row (or column) sum.

The following basic model (excluding the eigenvectors) is

$$\begin{aligned} log(p_i) =& \beta_0 + \beta_1 crim_i + \beta_2 zn_i + \beta_3 indus_i + \beta_4 chas_i + \beta_5 nox_i^2 + \beta_6 rm_i + \beta_7 age_i \\ &+ \beta_8 dis_i + \beta_9 rad_i + \beta_{10} tax_i + \beta_{11} ptr_i + \beta_{12} black_i + \beta_{13} lsp_i + \varepsilon_i. \end{aligned}$$

Table 2.4 shows the parameter estimates (excluding eigenvectors) for simple-OLS (ignoring the spatial correlation), Mi-pLasso2, CV-pLasso, and FstepZ.[15]. These results show that some of the OLS estimates are biased by spatial dependence. For example, age had a positive (but insignificant) coefficient when the spatial dependence is ignored, but in the filtering estimates, the coefficient is negative and significant as expected; the coefficient on nox^2, dis, and rm also have a downward bias. Additionally, the filtered estimates also give a substantially better fit of house prices, with Mi-pLasso2 having an adjusted R^2 of 0.978, implying an almost perfect fit of the data. Mi-pLasso2 standard errors are generally the same or smaller than the other estimator.

Table 2.5 shows the computational times, the number of selected eigenvectors, and their significance levels, for the three ESF estimators. There is substantial variation in the number of selected eigenvectors between the procedures. Mi-Lasso2 selected six and three times more eigenvectors than CV-Lasso and FstepZ. However, despite selecting substantially more eigenvectors for Mi-Lasso2, only 0.5 percent of selected eigenvectors are insignificant compared to 28 percent and 19 percent for FstepZ and CV-Lasso. Mi-Lasso2 has more eigenvectors with coefficients significant at the 0.1 percent level than FstepZ or CV-Lasso selected in total, implying these techniques may be under-selecting in this case. Mi-Lasso2 is also over 65 times faster than both FstepZ and CV-Lasso.

[15]Mi-Lasso1 ($a = 1$) estimates are the same as simple-OLS, given no eigenvectors are selected in this case

Table 2.4: Parameter Estimation Results

	Dependent variable:			
	$ln(p)$			
	simple-OLS	FstepZ	CV-pLasso	Mi-pLasso2
	(1)	(2)	(3)	(4)
crim	-0.010^{***}	-0.009^{***}	-0.010^{***}	-0.011^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
zn	0.001^{**}	0.001	0.001^{**}	0.0001
	(0.001)	(0.0005)	(0.0004)	(0.0002)
indus	0.002	-0.0003	0.002	0.004^{***}
	(0.002)	(0.002)	(0.002)	(0.001)
chas	0.104^{***}	0.038	0.065^{**}	0.055^{***}
	(0.034)	(0.030)	(0.026)	(0.016)
nox^2	-0.588^{***}	-0.219^{*}	-0.165^{*}	-0.304^{***}
	(0.114)	(0.125)	(0.091)	(0.051)
rm	0.091^{***}	0.177^{***}	0.209^{***}	0.169^{***}
	(0.017)	(0.015)	(0.014)	(0.008)
age	0.0001	-0.001^{**}	-0.001^{***}	-0.001^{***}
	(0.001)	(0.0004)	(0.0004)	(0.0002)
dis	-0.047^{***}	-0.032^{***}	-0.030^{***}	-0.033^{***}
	(0.008)	(0.007)	(0.006)	(0.004)
rad	0.014^{***}	0.011^{***}	0.012^{***}	0.012^{***}
	(0.003)	(0.002)	(0.002)	(0.001)
tax	-0.001^{***}	-0.0004^{***}	-0.001^{***}	-0.001^{***}
	(0.0002)	(0.0001)	(0.0001)	(0.0001)
ptr	-0.039^{***}	-0.006	-0.023^{***}	-0.036^{***}
	(0.005)	(0.006)	(0.005)	(0.002)
black	-0.003^{***}	-0.005^{***}	-0.005^{***}	-0.005^{***}
	(0.001)	(0.001)	(0.001)	(0.0005)
lsp	-0.029^{***}	-0.020^{***}	-0.018^{***}	-0.022^{***}
	(0.002)	(0.002)	(0.002)	(0.001)
Constant	4.031^{***}	2.655^{***}	2.734^{***}	3.352^{***}
	(0.175)	(0.171)	(0.155)	(0.079)
Adjusted R^2	0.785	0.896	0.893	0.978
Residual Std. Error	0.189 (df = 492)	0.132 (df = 431)	0.134 (df = 461)	0.061 (df = 295)

Note: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in parenthesis. For Mi-Lasso I set $a = 2$

Table 2.5: Computational time and Selected Eigenvectors

	FStepZ	CV-pLasso	Mi-pLasso2
Computational time (seconds)	42.99	39.54	0.64
Number of Eigenvectors	61	31	197
Significant at 0.1% level	16	17	85
Significant at 1% level	7	4	40
Significant at 5% level	13	2	56
Significant at 10% level	8	2	15
Not significant	17	6	1

Note: computational times exclude spectral decomposition. FStepZ uses the 'SpatialFiltering' function from the **R** package 'spdep'.

2.8 Conclusion and Further Work

In this chapter, I have formalised the ESF assumptions and evaluated the existing solutions to the ESF eigenvector selection problem. Our analysis of existing procedures has shown that a dominant selection procedure currently does not exist. The forward-iterative procedure with a user-defined cut-off and eigenvector inclusion criterion can be viewed as ad hoc and are slow, especially as the sample size increases. Seya et al. (2015) proposed using Lasso with prediction accuracy CV to estimate the tuning parameter. However, as ESF aims to estimate $\boldsymbol{\beta}_0$ rather than prediction accuracy, it is unclear if prediction accuracy CV is the best way to estimate the tuning parameter. Additionally, CV-based Lasso procedure is also slow, especially when n is large. The most computationally demanding aspect of the procedure is the CV part.

I have proposed an alternative Lasso-based procedure called Morans' i Lasso (Mi-Lasso) that uses information about the level of spatial correlation in the simple regression residuals (ignoring the spatial correlation) to determine a point estimate

for the Lasso tuning parameter instead of using CV. The key benefits of Mi-Lasso are that it is intuitive, theoretically grounded, and substantially faster than Seya et al. (2015) CV Lasso or any forward stepwise procedure proposed and thus, can easily be implemented on large data sets. I have derived performance bounds for the Mi-Lasso estimates of the eigenvectors coefficients and shown the conditions necessary for the estimator to provide consistent eigenvector selection. Our simulation results confirm the estimator performs well in terms of bias and MSE compared to existing selection procedures for a range of levels of spatial correlation and in an empirical application on house prices. Additionally, I have shown using a property of the spectral decomposition and a simulations experiment, that ESF is robust to the presence of an unknown number of higher-order spatial lags in underlying DGP.

A key limitation of the ESF literature is that there are no results on constructing robust standard errors. As all the proposed procedures can be viewed as post-model selection estimators. Thus, all the corresponding estimators suffer from the corresponding post-model selection inference problem (Leeb and Pötscher, 2008). Given the spatial dependence in the model debiasing techniques such as Double Lasso (Belloni et al., 2013) or Partial Lasso (Chernozhukov et al., 2015) does not work well when the covariates being studied are also spatially correlated. A promising avenue of future research in the ESF literature is to extend Mi-Lasso (and other procedures), so standard errors robust to selection mistakes and the spatial dependence in the model can be calculated. Additionally, developing a further understanding of the Mi-Lasso tuning parameter exponent (a) and exploration of alternative variants of Lasso, such as adaptive Lasso, will also be promising avenues for further research.

2.9 Appendix

2.9.1 Proofs

Proof of Lemma 1. Two important point to note is by Assumption 1.2 the $n \times k$ matrix $\boldsymbol{X}$ has full column rank and only the coefficient of the matrix $\boldsymbol{E}$ are being penalized. The objective function in (2.18) is coercive (for minimization) and strictly convex, thus, $[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}]$ is a unique global minimizer. (2.18) is also subdifferentiable, specifically from the Karush-Kuhn-Tucker conditions for Lasso we have

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{E}\hat{\boldsymbol{\gamma}}) = 0 \tag{2.26}$$

$$\boldsymbol{E}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{E}\hat{\boldsymbol{\gamma}}) - \frac{1}{Z^a} s(\hat{\boldsymbol{\gamma}}) = 0 \tag{2.27}$$

where $s(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to $\in [-1, 1]$.

Rearranging (2.26) to make $\hat{\boldsymbol{\beta}}_{\frac{1}{Z}}$ the subject:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{E}\hat{\boldsymbol{\gamma}})$$

Substituting this into (2.27) yields

$$\begin{aligned}
\boldsymbol{E}'(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{E}\hat{\boldsymbol{\gamma}}) - \boldsymbol{E}\hat{\boldsymbol{\gamma}}) - \frac{1}{Z^a} s(\hat{\boldsymbol{\gamma}}) = 0 \\
\boldsymbol{E}'((\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} - (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{E}\hat{\boldsymbol{\gamma}}) - \frac{1}{Z^a} s(\hat{\boldsymbol{\gamma}}) = 0 \\
\boldsymbol{E}'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\hat{\boldsymbol{\gamma}}) - \frac{1}{Z^a} s(\hat{\boldsymbol{\gamma}}) = 0
\end{aligned} \tag{2.28}$$

The left-hand side of (2.28) is a sub-vector of the objective function in (2.19) at

$\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$ and it equals 0, thus, $\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}$, if the minimisation is unique. □

Proof of Theorem 1. By definition

$$\tilde{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma}||_2^2 + \frac{1}{Z^a}||\boldsymbol{\gamma}||_1$$

Denoting $\boldsymbol{\Delta} = \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$, then by the optimality of $\tilde{\boldsymbol{\gamma}}$ and dividing by n

$$\begin{aligned} ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\tilde{\boldsymbol{\gamma}}||_2^2/n + \frac{1}{Z^a n}||\tilde{\boldsymbol{\gamma}}||_1 &\leq ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma}_0||_2^2/n + \frac{1}{Z^a n}||\boldsymbol{\gamma}_0||_1 \\ ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\tilde{\boldsymbol{\gamma}}||_2^2/n - ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma}_0||_2^2/n &\leq \frac{1}{Z^a n}(||\boldsymbol{\gamma}_0||_1 - ||\tilde{\boldsymbol{\gamma}}||_1) \end{aligned} \tag{2.29}$$

Given $\boldsymbol{\Delta}_{\grave{\Omega}} = \tilde{\boldsymbol{\gamma}}_{\grave{\Omega}}$, $\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_\Omega$ and the reverse triangle inequality $||\tilde{\boldsymbol{\gamma}}_\Omega||_1 \geq ||\boldsymbol{\gamma}_\Omega||_1 - ||\boldsymbol{\Delta}_\Omega||_1$, we have

$$\begin{aligned} ||\boldsymbol{\gamma}_0||_1 - ||\tilde{\boldsymbol{\gamma}}||_1 &= ||\boldsymbol{\gamma}_0||_1 - (||\tilde{\boldsymbol{\gamma}}_\Omega||_1 + ||\tilde{\boldsymbol{\gamma}}_{\grave{\Omega}}||_1) = ||\boldsymbol{\gamma}_0||_1 - (||\tilde{\boldsymbol{\gamma}}_\Omega||_1 + ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1) \\ ||\boldsymbol{\gamma}_0||_1 - ||\tilde{\boldsymbol{\gamma}}||_1 &\leq ||\boldsymbol{\gamma}_0||_1 - (||\boldsymbol{\gamma}_\Omega||_1 - ||\boldsymbol{\Delta}_\Omega||_1 + ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1) \leq ||\boldsymbol{\Delta}_\Omega||_1 - ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1 \end{aligned} \tag{2.30}$$

Furthermore

$$\begin{aligned} ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\tilde{\boldsymbol{\gamma}}||_2^2/n - ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma}_0||_2^2/n &= ||\tilde{\boldsymbol{E}}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) - \boldsymbol{v}||_2^2/n - ||\boldsymbol{v}||_2^2/n \\ &= ||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2^2/n - 2\boldsymbol{v}'\tilde{\boldsymbol{E}}\boldsymbol{\Delta}/n \\ &\geq_{(i)} ||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2^2/n - 2||\boldsymbol{v}'\tilde{\boldsymbol{E}}||_\infty/n||\boldsymbol{\Delta}||_1 \\ &\geq_{(ii)} ||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2^2/n - \frac{1}{Z^a bn}||\boldsymbol{\Delta}||_1 \end{aligned} \tag{2.31}$$

(i) uses Hölder inequality with ℓ_∞ and ℓ_1 norms, $2|\boldsymbol{v}'\tilde{\boldsymbol{E}}(\boldsymbol{\Delta})|/n \leq 2||\boldsymbol{v}'\tilde{\boldsymbol{E}}||_\infty/n||\boldsymbol{\Delta}||_1$.

(ii) uses the event

$$T := b2||\boldsymbol{v}'\tilde{\boldsymbol{E}}||_\infty/n \leq \frac{1}{Z^a n} \tag{2.32}$$

where $b \geq 1$ is an arbitrary constant, note this ensures the penalty should dominate the random part of the process.

Combining (2.29), (2.30) and (2.31)

$$\begin{aligned} ||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2^2/n &\leq \frac{1}{bZ^a n}(||\boldsymbol{\Delta}_\Omega||_1 + ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1) + \frac{1}{Z^a n}(||\boldsymbol{\Delta}_\Omega||_1 - ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1) \qquad (2.33) \\ &\leq \left(1 + \frac{1}{b}\right)\frac{1}{Z^a n}||\boldsymbol{\Delta}_\Omega||_1 - \left(1 - \frac{1}{b}\right)\frac{1}{Z^a n}||\boldsymbol{\Delta}_{\grave{\Omega}}||_1 \qquad (2.34) \end{aligned}$$

Given $||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2^2/n > 0$ and using (2.34) we have

$$||\boldsymbol{\Delta}_{\grave{\Omega}}||_1 \leq \bar{b}||\boldsymbol{\Delta}_\Omega||_1$$

where $\bar{b} = (b+1)/(b-1)$, this allow us to use the restricted eigenvalue condition RE($\bar{b}$), substituting in for $||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2^2/n$ in (2.33) gives

$$\begin{aligned} \tau_{min}^2||\boldsymbol{\Delta}||_2^2 &\leq \frac{1}{bZ^a n}||\boldsymbol{\Delta}||_1 + \frac{1}{Z^a n}(||\boldsymbol{\Delta}_\Omega||_1 - ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1) \\ &\leq \left(\frac{1}{b} + 1\right)\frac{1}{Z^a n}||\boldsymbol{\Delta}_\Omega||_1 \\ &\leq \left(\frac{1}{b} + 1\right)\frac{\sqrt{s}}{Z^a n}||\boldsymbol{\Delta}_\Omega||_2 \qquad (2.35) \end{aligned}$$

where the last inequality uses $||\boldsymbol{\Delta}_\Omega||_1 \leq \sqrt{s}||\boldsymbol{\Delta}_\Omega||_2$ which holds by the Cauchy-Schwarz inequality. Which implies the following ℓ_2 parameter bound

$$||\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_2 \leq \frac{\left(\frac{1}{b}+1\right)\sqrt{s}}{\tau_{min}^2 Z^a n}$$

Which is (2.23). Again we can swap the ℓ_2 norm for the ℓ_1 norm and rearrange to gives

$$||\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 \leq \frac{\left(\frac{1}{b}+1\right)s}{\tau_{min}^2 Z^a n}$$

Which is (2.22). Similarly substituting RE($\bar{b}$) in for $||\boldsymbol{\Delta}_\Omega||_2$ in (2.33) and given (2.35) yields

$$||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2^2/n \leq \left(\frac{1}{b}+1\right)\frac{\sqrt{s}}{Z^a \tau n^{3/2}}||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}_\Omega||_2$$

Which implies the following ℓ_2 performance bound

$$\frac{1}{\sqrt{n}}||\tilde{\boldsymbol{E}}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)||_2 \leq \frac{\left(\frac{1}{b}+1\right)\sqrt{s}}{\tau_{min} Z^a n}$$

Which is (2.24).

Now we have obtained (2.22), (2.23) and (2.24) by assuming (2.32). Next we evaluate the probability this event is true i.e. $P(T)$. Let $\frac{1}{Z^a} = t$ and using the definition of $||\cdot||_\infty$ we can rewrite (2.32) as

$$T := \max_{j \in N} 2b|\boldsymbol{v}'\tilde{\boldsymbol{e}}_j| \leq t$$

where $\tilde{\boldsymbol{e}}_j$ is the jth column of $\tilde{\boldsymbol{E}}$.

By a union bound

$$P(\grave{T}) = P(\max_{j \in N} 2b|\boldsymbol{v}'\tilde{\boldsymbol{e}}_j| \geq t) \leq n \max_{j \in N} P(2b|\boldsymbol{v}'\tilde{\boldsymbol{e}}_j| \geq t) \tag{2.36}$$

Given $\boldsymbol{v}$ is sub-Gaussian $(0, \sigma_{\boldsymbol{v}})$ and $\tilde{\boldsymbol{e}}_j$ is a vector of real numbers, thus, $\boldsymbol{v}'\tilde{\boldsymbol{e}}_j$ is also sub-gaussian.

$$P(\grave{T}) \leq n \max_{j \in N} P(2b|\boldsymbol{v}'\tilde{\boldsymbol{e}}_j|/n \geq t) \leq 2n \exp\left(- \frac{n^2 t^2}{2\sigma_{\boldsymbol{v}}^2 \max_{j \in N} ||\tilde{\boldsymbol{e}}_j||_2^2} \right) \leq 2n \exp\left(- \frac{nt^2}{2\sigma_{\boldsymbol{v}}^2} \right)$$

where the final equality holds by assuming $\max_{j \in N} ||\tilde{\boldsymbol{e}}_j||_2 \leq \sqrt{n}$.

let $t = \sqrt{\frac{4\sigma_{\boldsymbol{v}}^2 \log n}{n}}$ we get

$$P(\grave{T}) \leq \frac{2}{n}$$

We now have the following property

$$P(T) = 1 - P(\grave{T}) \geq 1 - \frac{2}{n} \to 1$$

as $n \to \infty$ □

To prove Propostition 1 we state Lemma 2, which is a direct consequence of the Karush-Kuhn-Tucker conditions:

Lemma 2. $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_j, \ldots, \tilde{\gamma}_n)$ *are the Lasso estimates defined by* (2.19) *if and only if*

$$\begin{aligned} \frac{d||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\gamma||_2^2}{d\gamma_j}|_{\gamma_j=\tilde{\gamma}_j} &= \frac{1}{Z^a}\,\mathrm{sign}(\tilde{\gamma}_j) && for\ j : \tilde{\gamma}_j \neq 0 \\ \left|\frac{d||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\gamma||_2^2}{d\gamma_j}\right|_{\gamma_j=\tilde{\gamma}_j} &\leq \frac{1}{Z^a} && for\ j : \tilde{\gamma}_j = 0 \end{aligned}$$

Proof of Propostition 1. Need to show that $A \cap B$ implies $\mathrm{sign}(\tilde{\boldsymbol{\gamma}}_\Omega) = \mathrm{sign}(\boldsymbol{\gamma}_\Omega)$ and $\tilde{\boldsymbol{\gamma}}_{\hat{\Omega}} = 0$

By definition

$$\tilde{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}}[(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma})'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma}) + \frac{1}{Z^a}||\boldsymbol{\gamma}||_1]$$

Let $\boldsymbol{\Delta} = \tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ and define

$$\boldsymbol{d}(\boldsymbol{\Delta}) = \left[(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}))'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}))\right] - (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma})'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma}) + \frac{1}{Z^a}||\boldsymbol{\gamma} + \boldsymbol{\Delta}||_1$$

Then

$$\boldsymbol{\Delta} = \arg\min_{\boldsymbol{\Delta}} \boldsymbol{d}(\boldsymbol{\Delta}) \tag{2.37}$$

Splitting $\boldsymbol{d}(\boldsymbol{\Delta})$ into two parts $\boldsymbol{d}_1(\boldsymbol{\Delta})$ and $\boldsymbol{d}_2(\boldsymbol{\Delta})$. Let

$$\begin{aligned}
\boldsymbol{d}_1(\boldsymbol{\Delta}) &= \big[(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}(\boldsymbol{\gamma}+\boldsymbol{\Delta}))'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}(\boldsymbol{\gamma}+\boldsymbol{\Delta}))\big] - (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma})'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\boldsymbol{\gamma}) \\
&= [(\tilde{\boldsymbol{v}} - \tilde{\boldsymbol{E}}\boldsymbol{\Delta})'(\tilde{\boldsymbol{v}} - \tilde{\boldsymbol{E}}\boldsymbol{\Delta}) - \tilde{\boldsymbol{v}}'\tilde{\boldsymbol{v}}] \\
&= -2\boldsymbol{\Delta}'\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{v}} + \boldsymbol{\Delta}'\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}\boldsymbol{\Delta} \\
&= -2(\sqrt{n}\boldsymbol{\Delta})'\boldsymbol{z} + (\sqrt{n}\boldsymbol{\Delta})'\boldsymbol{C}(\sqrt{n}\boldsymbol{\Delta})
\end{aligned}$$

where $\boldsymbol{z} = \tilde{\boldsymbol{E}}'\tilde{\boldsymbol{v}}/\sqrt{n}$. Differentiate $\boldsymbol{d}_1(\boldsymbol{\Delta})$ w.r.t. $\boldsymbol{\Delta}$

$$\frac{d\boldsymbol{d}_1(\boldsymbol{\Delta})}{d\boldsymbol{\Delta}} = 2\sqrt{n}\big(\boldsymbol{C}(\sqrt{n}\boldsymbol{\Delta}) - \boldsymbol{z}\big) \tag{2.38}$$

Now assuming that $\boldsymbol{\Delta}$ exists such that $\boldsymbol{\Delta}_{\grave{\Omega}} = 0$ and $\boldsymbol{\Delta}_{\Omega}$ is the solution of

$$\boldsymbol{C}_{\Omega\Omega}(\sqrt{n}\boldsymbol{\Delta}_{\Omega}) - \boldsymbol{z}_{\Omega} = -\frac{1}{2Z^a\sqrt{n}}\operatorname{sign}(\boldsymbol{\gamma}_{\Omega}) \tag{2.39}$$

Now, Event A says

$$|(\boldsymbol{C}_{\Omega\Omega})^{-1}\boldsymbol{z}_{\Omega}| < \sqrt{n}(|\boldsymbol{\gamma}_{\Omega}| - \frac{1}{2Z^a n}|(\boldsymbol{C}_{\Omega\Omega})^{-1}\operatorname{sign}(\boldsymbol{\gamma}_{\Omega})|) \tag{2.40}$$

Event B and the IC implies

$$|\boldsymbol{C}_{\grave{\Omega}\Omega}(\boldsymbol{C}_{\Omega\Omega})^{-1}\boldsymbol{z}_{\Omega} - \boldsymbol{z}_{\grave{\Omega}}| \leq \frac{1}{2Z^a\sqrt{n}}(\boldsymbol{1} - |\boldsymbol{C}_{\grave{\Omega}\Omega}(\boldsymbol{C}_{\Omega\Omega})^{-1}\operatorname{sign}(\boldsymbol{\gamma}_{\Omega})|) \tag{2.41}$$

Now, (2.39) and (2.40) implies

$$|\boldsymbol{\Delta}_{\Omega}| < |\boldsymbol{\gamma}_{\Omega}| \tag{2.42}$$

and (2.39) and (2.41) implies

$$-\frac{1}{2Z^a\sqrt{n}}\mathbf{1} \leq \boldsymbol{C}_{\grave{\Omega}\Omega}(\sqrt{n}\boldsymbol{\Delta}_{\Omega}) - \boldsymbol{z}_{\grave{\Omega}} \leq \frac{1}{2Z^a\sqrt{n}}\mathbf{1} \tag{2.43}$$

Thus, by Lemma 2, (2.37), (2.38) and the uniqueness of the Lasso solution, $\text{sign}(\tilde{\boldsymbol{\gamma}}_{\Omega}) = \text{sign}(\boldsymbol{\gamma}_{\Omega})$ and $\tilde{\boldsymbol{\gamma}}_{\grave{\Omega}} = \boldsymbol{\Delta}_{\grave{\Omega}} = 0$.

□

Proof of Theorem 2. This proof works by bounding the tail probability of Proposition 1 using conditions on the disturbance term. By Proposition 1 we have

$$\mathbb{P}\left(\tilde{\boldsymbol{\gamma}} =_s \boldsymbol{\gamma}_0\right) \geq \mathbb{P}(A \cap B)$$

thus,

$$\begin{aligned} 1 - \mathbb{P}(A \cap B) &\leq \mathbb{P}(\grave{A}) + \mathbb{P}(\grave{B}) \\ &\leq \sum_{i=1}^{s} \mathbb{P}\left(|k_i| \geq \sqrt{n}\left(|\gamma_i| - \frac{1}{2Z^a n} b_i\right)\right) + \sum_{i=1}^{q} \mathbb{P}\left(|\xi_i| \geq \frac{1}{2Z^a\sqrt{n}}\nu_i\right) \end{aligned} \tag{2.44}$$

where $\boldsymbol{k} = (k_1, \ldots, k_s)' = \boldsymbol{C}_{\Omega\Omega}^{-1}\boldsymbol{z}_{\Omega}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q)' = \boldsymbol{C}_{\grave{\Omega}\Omega}\boldsymbol{C}_{\Omega\Omega}^{-1}\boldsymbol{z}_{\Omega} - \boldsymbol{z}_{\grave{\Omega}}$ and $\boldsymbol{b} = (b_1, \ldots, b_s)' = \boldsymbol{C}_{\Omega\Omega}^{-1}\,\text{sign}(\boldsymbol{\gamma}_{\Omega})$

Now if we write $\boldsymbol{k} = \boldsymbol{H}_A'\boldsymbol{v}$ where $\boldsymbol{H}_A' = (\boldsymbol{h}_{1,a}, \ldots, \boldsymbol{h}_{s,a})' = \boldsymbol{C}_{\Omega\Omega}^{-1}(n^{-\frac{1}{2}}\tilde{\boldsymbol{E}}_\Omega)$, then

$$\boldsymbol{H}_A'\boldsymbol{H}_A = \boldsymbol{C}_{\Omega\Omega}^{-1}n^{-1}\tilde{\boldsymbol{E}}_\Omega'\tilde{\boldsymbol{E}}_\Omega\boldsymbol{C}_{\Omega\Omega}^{-1} = \boldsymbol{C}_{\Omega\Omega}^{-1}$$

Therefore (using Assumption 4.2) $k_i = \boldsymbol{h}_{i,a}'\tilde{\boldsymbol{v}}$ with

$$||\boldsymbol{h}_{i,a}||_2^2 \leq \frac{1}{M_2} \;\; \forall\, i = 1, \ldots, s. \tag{2.45}$$

Similarly if we write $\boldsymbol{\xi} = \boldsymbol{H}_B'\tilde{\boldsymbol{v}}$ where $\boldsymbol{H}_B' = (\boldsymbol{h}_{1,b}, \ldots, \boldsymbol{h}_{q,b})' = \boldsymbol{C}_{\grave{\Omega}\Omega}\boldsymbol{C}_{\Omega\Omega}^{-1}(n^{-\frac{1}{2}}\tilde{\boldsymbol{E}}_\Omega') - n^{-\frac{1}{2}}\tilde{\boldsymbol{E}}_{\grave{\Omega}}'$, then

$$\begin{aligned}\boldsymbol{H}_B'\boldsymbol{H}_B &= (\boldsymbol{C}_{\grave{\Omega}\Omega}\boldsymbol{C}_{\Omega\Omega}^{-1}(n^{-\frac{1}{2}}\tilde{\boldsymbol{E}}_\Omega') - n^{-\frac{1}{2}}\tilde{\boldsymbol{E}}_{\grave{\Omega}}')(n^{-\frac{1}{2}}\tilde{\boldsymbol{E}}_\Omega\boldsymbol{C}_{\Omega\Omega}^{-1}\boldsymbol{C}_{\grave{\Omega}\Omega} - n^{-\frac{1}{2}}\tilde{\boldsymbol{E}}_{\grave{\Omega}}) \\ &= n^{-1}\tilde{\boldsymbol{E}}_{\grave{\Omega}}'(I - \tilde{\boldsymbol{E}}_\Omega(\tilde{\boldsymbol{E}}_\Omega'\tilde{\boldsymbol{E}}_\Omega)^{-1}\tilde{\boldsymbol{E}}_\Omega')\tilde{\boldsymbol{E}}_{\grave{\Omega}}\end{aligned}$$

Given the eigenvalues of $I - \tilde{\boldsymbol{E}}_\Omega(\tilde{\boldsymbol{E}}_\Omega'\tilde{\boldsymbol{E}}_\Omega)^{-1}\tilde{\boldsymbol{E}}_\Omega'$ are 0 and 1, therefore (using Assumption 4.1) $\xi_i = \boldsymbol{h}_{i,b}'\tilde{\boldsymbol{v}}$ with

$$||\boldsymbol{h}_{i,b}||_2^2 \leq M_1 \;\; \forall\, i = 1, \ldots, s. \tag{2.46}$$

Also note that,

$$\left|\frac{1}{Z^a n}\boldsymbol{b}\right| = \left|\frac{1}{Z^a n}\boldsymbol{C}_{\Omega\Omega}^{-1}\operatorname{sign}(\boldsymbol{\gamma}_\Omega)\right| \leq \frac{1}{Z^a M_2 n}||\operatorname{sign}(\boldsymbol{\gamma}_\Omega)||_2 = \frac{1}{Z^a M_2 n}\sqrt{s} \tag{2.47}$$

Given (2.45), (2.46) and Assumption 1.4 $\mathbb{E}[v_i^4] < \infty$, implies $\mathbb{E}[k_i^4] < \infty$ and $\mathbb{E}[\xi_i^4] < \infty$. In fact for any given constant n-dimensional vector $\boldsymbol{\alpha}$

$$\mathbb{E}(\boldsymbol{\alpha}'\boldsymbol{v})^4 \leq (3)!!||\boldsymbol{\alpha}||_2^2\,\mathbb{E}[v_i^4]$$

For an *i.i.d.* random variable with bound 4^{th} moments, their probability tail is bounded by

$$\mathbb{P}(k_i > t) = O(t^{-4}) \tag{2.48}$$

Rearranging the first summation term of (2.44)

$$\sum_{i=1}^{s} \mathbb{P}\left(|k_i| \geq \sqrt{n}\left(|\gamma_i| - \frac{b_i}{2Z^a n}\right)\right) = \sum_{i=1}^{s} \mathbb{P}\left(|k_i| \geq \frac{b_i}{2Z^a\sqrt{n}}\left(\frac{2nZ^a|\gamma_i|}{b_i} - 1\right)\right)$$

Next, we can use (2.48) to bound the s probabilities:

$$\sum_{i=1}^{s} \mathbb{P}\left(|k_i| \geq \frac{b_i}{2Z^a\sqrt{n}}\left(\frac{2nZ^a|\gamma_i|}{b_i} - 1\right)\right) = \sum_{i=1}^{s} O\left(\left(\frac{b_i}{2Z^a\sqrt{n}}\right)^{-4}\left(\frac{2nZ^a|\gamma_i|}{b_i} - 1\right)^{-4}\right) \tag{2.49}$$

We now need to evaluate the bounds for both terms in (2.49). Starting with the second term. We can use (2.47) and Assumption 4.4 to replace $\frac{nZ^a}{b_i}$ by $\frac{nZ^aM_2}{\sqrt{s}}$ and its associated $O(n^{c_1})$ bound:

$$O\left(\left(\frac{2nZ^a|\gamma_i|}{b_i} - 1\right)^{-4}\right) = O\left(\left(\frac{2nZ^aM_2|\gamma_i|}{\sqrt{s}} - 1\right)^{-4}\right) = O\left(\left(\frac{2nZ^aM_2|\gamma_i|}{O(n^{\frac{c_1}{2}})} - 1\right)^{-4}\right)$$

Next, using Assumption 4.3 to bound $\sqrt{n}|\gamma_i|$ and $\frac{1}{Z^a\sqrt{n}} = o_p(n^{\frac{c_1-c_2}{2}})$, and as all the bounds that contain a power of n cancel out, leaving a term that depends only

on constants M_2 and M_3 and is therefore $O(1)$:

$$O\left(\left(\frac{2nZ^a M_2|\gamma_i|}{O(n^{\frac{c_1}{2}})}-1\right)^{-4}\right)=O\left(\left(\frac{o_p(n^{\frac{c_1-c_2}{2}})2n^{\frac{c_2}{2}}M_3M_2}{O(n^{\frac{c_1}{2}})}-1\right)^{-4}\right)=O_p(1)$$

Note that Assumption 4.3 is used with equality $\forall|\gamma_i|$ rather than as an inequality on $\min|\gamma_i|$. This is because of the negative exponent, -4, as this implies that the highest bound will be obtained for the smallest value of the expression in brackets.

Expression (2.49) now reduces to:

$$\sum_{i=1}^{s}O\left(\left(\frac{b_i}{2Z^a\sqrt{n}}\right)^{-4}\left(\frac{2nZ^a|\gamma_i|}{b_i}-1\right)^{-4}\right)=sO\left(\left(\frac{b_i}{2Z^a\sqrt{n}}\right)^{-4}\right)=sO\left(\frac{16Z^{4a}n^2}{b_i^4}\right)$$

Using (2.47) again to replace $\frac{nZ^a}{b_i}$ by $\frac{nM_2Z^a}{\sqrt{s}}$, integrating s into the bound and ignoring the constants:

$$sO\left(\frac{16Z^{4a}n^2}{b_i^4}\right)=sO\left(\frac{16Z^{4a}n^2M_2^2}{s^2}\right)=O\left(\frac{Z^{4a}n^2}{s}\right)$$

Note that because $s+q=n$ and $s,n>0$, it must be that:

$$\frac{Z^{4a}n^2}{s}<n^3Z^{4a}$$

This is because the left hand side denominator is smaller by a factor s and right hand side larger by a factor n. Therefore:

$$O\left(\frac{Z^{4a}n^2}{s}\right)=o(n^3Z^{4a})$$

Thus,

$$\sum_{i=1}^{s} \mathbb{P}\left(|k_i| \geq \sqrt{n}\left(|\gamma_i| - \frac{b_i}{2Z^a n}\right)\right) = o(n^3 Z^{4a}) \tag{2.50}$$

For the second summation term of (2.44), using (2.48) we have:

$$\sum_{i=1}^{q} \mathbb{P}\left(|\xi_i^n| \geq \frac{1}{2Z^a\sqrt{n}}\nu_i\right) = \sum_{i=1}^{q} O\left(\left(\frac{1}{2Z^a\sqrt{n}}\nu_i\right)^{-4}\right)$$

As ν_i (by Assumption 4) and 2 are both constant they can be ignored

$$\sum_{i=1}^{q} O\left(\left(\frac{1}{2Z^a\sqrt{n}}\nu_i\right)^{-4}\right) = qO\big(Z^{4a}n^2\big)$$

Integrating q into the bound and noting that $s < n$, so if $s = O(n)$ then $q = O(n)$

$$qO\big(Z^{4a}n^2\big) = O(n^3 Z^{4a})$$

thus,

$$\sum_{i=1}^{q} \mathbb{P}\left(|\xi_i^n| \geq \frac{1}{2Z^a\sqrt{n}}\nu_i\right) = O(n^3 Z^{4a}) \tag{2.51}$$

Finally

$$\mathbb{P}(A \cap B) \geq 1 - o(n^3 Z^{4a}) - O(n^3 Z^{4a}) \to 1 \quad as\ n \to \infty \tag{2.52}$$

for $\frac{1}{n^3 Z^{4a}} \to \infty$. □

Chapter 3

Moran's I 2-Stage Lasso for spatial models with endogenous variables.

3.1 Introduction

The main aim of economic modelling is to explain how endogenous variables evolve according to fundamental processes such as productivity, taste, and policy. If the parameter(s) of interest are coefficients of endogenous variables, least squares estimation is invalidated by the endogeneity bias. 'Instrument variables' (IV) have a long history of estimating the parameters of endogenous, dating back to Wright (1928). This research is concerned with the situation where the equation being estimated contains (1) endogenous variables and (2) includes some possibly higher-order spatial process based on some given spatial weights matrix (SWM), and the endogenous variable also includes some possibly higher-order spatial process based on some given

SWM.[1] However, we assume that the exact functional form of the spatial process is *unknown*. There may also be mismeasured links in the SWM, and the researcher is only interested in estimating the direct effect of the right-hand-side variable(s). The spatial parameters are thus considered nuisance parameters. This setup is arguably a realistic situation for many applied researchers as they can test for cross-sectional/spatial dependence using a test such as a Moran's I test (Moran, 1950), but determining the exact form of the spatial process is much more challenging.

Spatial dependence and endogeneity are common in many economic models. Some examples include modeling the relationship between economic growth and energy consumption or pollution, employment and migration, and the effect of policing on crime. Many papers in the econometrics literature have shown how to incorporate endogenous variables into a given spatial model.[2] The Generalised Method of Moment (GMM) based estimation techniques such as Generalised Spatial Two-Stage Least Squares (GS2SLS) are commonly used by applied researchers when estimating a spatial model with an endogenous variable. However, to use any of the proposed GMM-based estimation techniques, the researcher must specify (1) a spatial economic model and (2) define the spatial structure, i.e., the SWM. A misspecified model will yield inconsistent estimates, and this problem will become more acute if the SWM is also misspecified (LeSage and Pace, 2014).

Given the uncertainty of which spatial economic model to use and the spatial parameters are considered nuisance parameters, we propose using Eigenvector Spatial

[1] A spatial weights matrix is an $n \times n$ matrix that describes the pair-wise relationship between each of the n cross-sectional units.

[2] Some recent examples include Hoshino (2018); Jenish (2016); Liu and Lee (2013); Fingleton and Le Gallo (2008).

Filtering (ESF), a methodology developed by Griffith (2000, 2003), as it has the key advantage of being agnostic to the underlying functional form of the spatial process. Instead of explicitly modeling the underlying spatial process, ESF uses a subset of eigenvectors from the SWM as controls in a linear regression framework to approximate the terms involving the SWM instead of estimating any spatial parameters.

Even if we ignore the endogenous variable, we cannot estimate ESF with the complete set of eigenvectors using Ordinary Least Squares (OLS). As the spectral decomposition of SWM yields n eigenvectors, and we have k covariates, the corresponding Gram matrix will be necessarily rank-deficient, as there are more columns/parameters $(n + k)$ than rows/observations (n). Under a sparsity assumption, i.e., only a subset of the eigenvector will have non-zero coefficients, estimation is possible. However, this creates a variable selection problem. To solve this selection problem, we propose using a Lasso-based procedure that uses information contained in the Moran statistic to determine a point estimate for the Lasso tuning parameters. The proposed estimator is called Moran's i two-stage Lasso (Mi-2SL) and is a three-step procedure where the first and second stages are first estimated by a Moran's I based Lasso to extract the relevant eigenvectors. The union of the set of selected eigenvectors is then used as controls in a two-stage least squares (2SLS) regression. The 2SLS part deals with the endogenous variables, and the Moran's I based Lasso deals with the (weak) cross-sectional dependence.[3]

Several studies have used two-stage Lasso procedures in a spatial setting. For example, Peng (2019) estimates a spatial autoregressive model (SAR) by a two-stage

[3]We will use the terms cross-sections dependence and spatial dependence interchangeably.

Lasso procedure to allow heterogeneous peer effects and the identification of the influential individuals in a network. As both stages are high-dimensional, they are both estimated by Lasso. Ahrens (2015) estimate the effect of conflict risk on economic growth using Belloni et al. (2012) two-stage procedure, where Lasso estimates the high-dimensional first stage and the second is a low-dimensional panel SAR model. Additionally, Ahrens and Bhattacharjee (2015); Lam and Souza (2016, 2020) all use two-stage Lasso-based procedures to estimate/select a SWM.

We derive theoretical results on consistent and asymptotically normal parameter estimation under the assumption the support (relevant) set of eigenvectors is known. Even under the assumption of perfect selection, proving the parameter estimates are consistent and asymptotically normal is not trivial. As the eigenvectors are derived from the SWM, a matrix with elements that describe the pair-wise dependence between the observations, so the standard assumption of row-wise independence is difficult to justify. To account for this dependence, we use Kojevnikov et al. (2021) notion of ψ-dependence and their corresponding limit theorems to derive the results on consistency and asymptotic normality.

Our Monty Carlo simulations show the proposed estimator performs well in finite samples. We look at how the proposed procedure performs for various degrees of correlation between the first and second-stage errors and various degrees of mismeasured links in the SWM. Mi-2SL performs best in terms of bias and mean squared errors relative to the other estimators when there is a high degree of mismeasured links. Demonstrating that Mi-2SL and, thus, ESF can perform well in the presence of mismeasured links.

As a motivating application, we apply our methodology to Cadena and Kovak (2016), who analyse the impact of Mexican mobility on local labour markets employment outcomes of natives in the US. They use a standard IV strategy to correct for possible endogeneity when testing if Mexican-born immigrants respond causally to changes in local labour demand. Despite having an explicit spatial dimension in their data they do not account for it in their estimates. A standardised Moran's I test on the first and second-stage residuals indicates significant spatial correlation with a higher spatial correlation level in the first stage than the second for most demographic groups. As we do not know the functional form of the spatial process and have some uncertainty regarding which SWM to use, we re-estimate their model using Mi-2SL to account for the unknown spatial structure. We find using Mi-2SL does not change Cadena and Kovak (2016) overall conclusion. However, it substantially improves the strength of the Bartik instrument in the first stage, which improves the precision (reduces the standard errors) in the second stage.

The rest of the paper is outlined as follows, Section 3.2 presents the underlying model. Section 3.3 proposes the Mi-2SL procedure. Section 3.4 we drive the theoretical properties of 2SLS ESF under perfect selection. Section 3.5 provides Monte Carlo studies to evaluate the finite sample properties of the proposed estimator and how it performs in the presence of mismeasured links. Section 3.6, we apply the proposed procedure to Cadena and Kovak (2016). Finally, Section 3.7 offers our concluding remarks.

3.2 Underlying model

Consider the following equation where the endogenous $n \times 1$ vector $\boldsymbol{y}$ is specified as a function of an $n \times k_1$ matrix of exogenous regressors $\boldsymbol{X}_1$ and an $n \times 1$ endogenous vector $\boldsymbol{x}_2$, as well as follows some spatial process (3.1). However, the exact spatial process (which spatial parameters are non-zero), including p, is unknown.

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_{1,0} + \boldsymbol{x}_2\beta_{2,0} + \sum_{i=1}^{p} \boldsymbol{W}^i\boldsymbol{y}\rho_{i,0} + \boldsymbol{W}\boldsymbol{X}_1\boldsymbol{\psi}_0 + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of innovations,[4], $\boldsymbol{W}$ is a $n \times n$ symmetric weights matrix, $\rho_{i,0}$'s and $\boldsymbol{\psi}_0$ are unknown parameters that represent the degree of spatial correlation in the endogenous variable $\boldsymbol{y}$ and the predetermined exogenous variables $\boldsymbol{X}_1$ with moment conditions $\mathbb{E}[\boldsymbol{X}_1'\boldsymbol{\varepsilon}] = 0$ and $\mathbb{E}[(\boldsymbol{W}\boldsymbol{X}_1, \boldsymbol{X}_1)'\boldsymbol{\varepsilon}] = 0$. The unknown parameter of interest is $\beta_{2,0}$. The regressor $\boldsymbol{x}_2$ is endogenous, in the sense that $\mathbb{E}(\boldsymbol{x}_2'\boldsymbol{\varepsilon}) \neq 0$. The extension to the case where $\boldsymbol{x}_2$ is a matrix is straightforward and omitted for simplicity. The variables $\boldsymbol{W}^i\boldsymbol{y}$ and $\boldsymbol{W}\boldsymbol{X}_1$ are typically referred to as ith and first order spatial lags of $\boldsymbol{y}$ and $\boldsymbol{X}_1$.

Given the uncertainty over the exact functional form of the linear in-parameters spatial process, an alternative representation of (3.1) is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_{1,0} + \boldsymbol{x}_2\beta_{2,0} + f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}_1) + \boldsymbol{\varepsilon},$$

[4]The data generating process of $\boldsymbol{y}$ could also include spatial autoregressive disturbances; however this is excluded from the model for simplicity.

where $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}_1)$ are some index linear function of $\boldsymbol{W}$, $\boldsymbol{y}$ and $\boldsymbol{X}_1$.

We also assume the endogenous variables in $\boldsymbol{x}_2$ also follow some spatial process with an unknown number of lags (l),

$$\boldsymbol{x}_2 = \boldsymbol{X}_1\boldsymbol{\zeta}_{1,0} + \boldsymbol{Z}_2\boldsymbol{\zeta}_{2,0} + \boldsymbol{W}\boldsymbol{X}_1\boldsymbol{\zeta}_{3,0} + \boldsymbol{W}\boldsymbol{Z}_2\boldsymbol{\zeta}_{4,0} + \sum_{i=1}^{l} \boldsymbol{x}_2\boldsymbol{W}^i\zeta_{i,5,0} + \boldsymbol{u}_2, \tag{3.2}$$

where $\boldsymbol{Z}_2$ is an $n \times q$ matrix of instrument variables with $q \geq 1$ and $\mathbb{E}(\boldsymbol{Z}_2'\boldsymbol{\varepsilon}) = 0$ and $\boldsymbol{u}_2$ is a vector of disturbances respectively, with $\mathbb{E}[(\boldsymbol{X}_1, \boldsymbol{Z}_2, \boldsymbol{W}\boldsymbol{X}_1, \boldsymbol{W}\boldsymbol{Z}_2)'\boldsymbol{u}_2] = 0$ and $\mathbb{E}[(\boldsymbol{X}_1, \boldsymbol{Z}_2, \boldsymbol{W}\boldsymbol{X}_1, \boldsymbol{W}\boldsymbol{Z}_2)'\boldsymbol{\varepsilon}] = 0$. The alternative representation of (3.2) is

$$\boldsymbol{x}_2 = \boldsymbol{X}_1\boldsymbol{\zeta}_{1,0} + \boldsymbol{Z}_2\boldsymbol{\zeta}_{2,0} + g(\boldsymbol{W}, \boldsymbol{x}_2, \boldsymbol{X}_1, \boldsymbol{Z}_2) + \boldsymbol{u}_2$$

where $g(\boldsymbol{W}, \boldsymbol{x}_2, \boldsymbol{X}_1, \boldsymbol{Z}_2)$ are some index linear function of $\boldsymbol{W}$, $\boldsymbol{x}_2$, $\boldsymbol{X}_1$ and $\boldsymbol{Z}_2$.

Let $N_n = N = \{1, \ldots, n\}$ be the set of cross-sectional unit indices with $n \in \mathbb{N}$ denoting the number of observations. For reasons of generality, we allow the elements of $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_n$, $\boldsymbol{y} = \boldsymbol{y}_n$, $\boldsymbol{W} = \boldsymbol{W}_n$, $\boldsymbol{Z}_2 = \boldsymbol{Z}_{2,n}$, $\boldsymbol{u}_2 = \boldsymbol{u}_{2,n}$, $\boldsymbol{X}_1 = \boldsymbol{X}_{1,n}$ and $\boldsymbol{x}_2 = \boldsymbol{x}_{2,n}$ to be dependent on n- that is to form triangular arrays. However, to simplify the notation, the n index is omitted. Our analysis is conditioned on realised values. Thus, matrices and vectors such as $\boldsymbol{X}_1$, $\boldsymbol{x}_2$, $\boldsymbol{W}$ and $\boldsymbol{Z}_2$ are viewed as matrices and vectors of constants.

Equation (3.1) contains two sources of endogeneity, $\boldsymbol{x}_2$ because $\mathbb{E}(\boldsymbol{u}_2'\boldsymbol{\varepsilon}) \neq 0$ which in turn causes $\mathbb{E}(\boldsymbol{x}_2'\boldsymbol{\varepsilon}) \neq 0$ and $\boldsymbol{y}$ is also clearly endogenous as it appears on both sides of (3.1). Both sources of endogeneity cause the Ordinary Least Squares (OLS)

estimate of $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{1,0}, \beta_{2,0})'$ to be inconsistent ($\hat{\boldsymbol{\beta}}_{ols} \not\to_p \boldsymbol{\beta}_0$).

Substituting (3.2) into (3.1) gives the reduced form for $\boldsymbol{y}$

$$\boldsymbol{y} = \boldsymbol{S}_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_{1,0} + \boldsymbol{S}_2^{-1}(\boldsymbol{X}_1\boldsymbol{\pi}_{1,0} + \boldsymbol{Z}_2\boldsymbol{\pi}_{2,0} + \boldsymbol{W}\boldsymbol{X}_1\boldsymbol{\pi}_{3,0} + \boldsymbol{W}\boldsymbol{Z}_2\boldsymbol{\pi}_{4,0}) + \boldsymbol{W}\boldsymbol{X}_1\boldsymbol{\psi}_0 + \boldsymbol{d}), \tag{3.3}$$

where $\boldsymbol{\pi}_{1,0} = \beta_{2,0}\boldsymbol{\zeta}_{1,0}$, $\boldsymbol{\pi}_{2,0} = \beta_{2,0}\boldsymbol{\zeta}_{2,0}$, $\boldsymbol{\pi}_{3,0} = \beta_{2,0}\boldsymbol{\zeta}_{3,0}$, $\boldsymbol{\pi}_{4,0} = \beta_{2,0}\boldsymbol{\zeta}_{4,0}$, $\boldsymbol{d} = \boldsymbol{S}_2^{-1}\boldsymbol{u}_2\beta_{2,0} + \boldsymbol{\varepsilon}$ and both $\boldsymbol{S}_1 \equiv (\boldsymbol{I} - \sum_{i=1}^{p} \rho_{i,0}\boldsymbol{W}^i)$, and $\boldsymbol{S}_2 \equiv (\boldsymbol{I} - \sum_{i=1}^{l} \boldsymbol{W}^i\zeta_{i,3,0})$ are non-singular.

3.3 Moran's i 2 Stage Lasso

As discussed above we have two sources of endogeneity in (3.1), $\boldsymbol{x}_2$ and $\boldsymbol{W}^i\boldsymbol{y}$ $\forall i$. Estimation of (3.2) is also not trivial as $\boldsymbol{W}^i\boldsymbol{x}_2$ $\forall i$ is also endogenous. Given we have valid instruments to deal with the endogenous variable $\boldsymbol{x}_2$, the key issue is how to capture the unknown underlying spatial processes in (3.1) and (3.2). Even if the exact underlying spatial process was known, estimation of (3.1) would be feasible but non-trivial. One method would be to first estimate (3.2) by GS2SLS, which was first developed by Kelejian and Prucha (1998) and extended by Drukker et al. (2019) to allow for higher-order spatial lags, which would use higher order spatial lags of the exogenous variables in (3.2) as instruments for $\boldsymbol{W}^i\boldsymbol{x}_2$ $\forall i$. The resulting fitted values can then be used to estimate (3.1). GS2SLS does have the advantage that it can be easily extended to include other right-hand-side endogenous variables. However, the procedure will require the researcher to specify which spatial parameters to estimate,

and given this extra layer of estimation, the standard GS2SLS standard errors would be invalid.

Given the uncertainty regarding the functional form of the spatial part of the underlying model, we propose using eigenvectors $\boldsymbol{E}_n = \boldsymbol{E}$ from a spectral decomposition of $\boldsymbol{W}$ to approximate $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}_1)$ and $g(\boldsymbol{W}, \boldsymbol{x}_2, \boldsymbol{X}_1, \boldsymbol{Z}_2)$ i.e. $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}_1) = \boldsymbol{E}\boldsymbol{\gamma}_{y,0}$ and $g(\boldsymbol{W}, \boldsymbol{x}_2, \boldsymbol{X}_1, \boldsymbol{Z}_2) = \boldsymbol{E}\boldsymbol{\gamma}_{x,0}$ where $\boldsymbol{\gamma}_{y,0}$ and $\boldsymbol{\gamma}_{x,0}$ are vectors of unknown constants. This methodology has the key advantage that it is agnostic to the exact form of $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}_1)$ and $g(\boldsymbol{W}, \boldsymbol{x}_2, \boldsymbol{X}_1, \boldsymbol{Z}_2)$, including the presence of higher-order lags, stemming from the spectral property that the eigenvectors from $\boldsymbol{W}$ and $\boldsymbol{W}^i$ $\forall i \in \mathbb{Z}^+$ are the same. So instead of trying to estimate the system (3.1) and (3.2) the following system could, in principle, be estimated (assuming the approximation is valid)

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\Upsilon}_0 + \boldsymbol{\varepsilon}, \tag{3.4}$$

$$\boldsymbol{x}_2 = \boldsymbol{Z}\boldsymbol{\zeta}_0 + \boldsymbol{u}_2, \tag{3.5}$$

where $\boldsymbol{G} = [\boldsymbol{X}_1, \boldsymbol{x}_2, \boldsymbol{E}]$, $\boldsymbol{\Upsilon}_0 = [\boldsymbol{\beta}'_{1,0}, \beta_{2,0}, \boldsymbol{\gamma}'_{y,0}]'$, $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{Z}_2, \boldsymbol{E}]$ and $\boldsymbol{\zeta}_0 = [\boldsymbol{\zeta}'_{1,0}, \boldsymbol{\zeta}'_{2,0}, \boldsymbol{\gamma}'_{x,0}]'$ with $\mathbb{E}[\boldsymbol{G}'\boldsymbol{\varepsilon}] = 0$ and $\mathbb{E}[\boldsymbol{Z}'\boldsymbol{u}_2] = 0$.

Now the problem with (3.4) and (3.5) is they are both high-dimensional linear regressions, as in each equation the number of parameters is greater than the number of observations. This means both the (re-scaled) Gram matrices $\boldsymbol{G}'\boldsymbol{G}/n$ and $\boldsymbol{Z}'\boldsymbol{Z}/n$ are necessarily rank deficient. Thus, neither (3.4) or (3.5) cannot be estimated by

OLS or (3.4) by 2SLS.

Griffith (2000) argued that only a subset of eigenvectors would be relevant to the data generating process (DGP) of $\boldsymbol{y}$ and $\boldsymbol{x}_2$ will have non-zero coefficients, i.e., the parameter vectors $\boldsymbol{\gamma}_{y,0}$ and $\boldsymbol{\gamma}_{x,0}$ are sparse. The intuition behind this sparsity assumption is each of the n eigenvectors can be viewed as an orthogonal spatial pattern, and only a specific subset of these patterns are relevant to the DGP of $\boldsymbol{y}$ and $\boldsymbol{x}_2$ (Griffith, 2003). Thus, a selection procedure is needed.

Moran's i based Lasso was first proposed in Section 2.4 considered a single equation where all the covariates are exogenous, i.e., (3.1) with $\beta_2 = 0$. They only penalised the $\boldsymbol{\gamma}_y$ coefficients and set the Lasso tuning parameter to $Z^{-a}\ \forall\ z \neq 0$ where a is a positive constant (generally set equal to 1 or 2), Z is the absolute value of the standardised Moran's i (z) of the residual $\hat{\boldsymbol{h}} = \boldsymbol{M_X y}$ where $\boldsymbol{M_X} = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ with,

$$z = \left(\frac{m - \mathbb{E}[m]}{\sqrt{\mathrm{Var}(m)}}\right) \tag{3.6}$$

with

$$\begin{aligned} m &= \frac{\hat{\boldsymbol{h}}'\boldsymbol{W}\hat{\boldsymbol{h}}}{\hat{\boldsymbol{h}}'\hat{\boldsymbol{h}}}, \\ \mathbb{E}[m] &= \frac{tr(\boldsymbol{M_X W M_X})}{n-k}, \\ \mathrm{Var}(m) &= \frac{2\Big((n-k)tr\big((\boldsymbol{M_X W M_X})^2\big) - \big[tr(\boldsymbol{M_X W M_X})\big]^2\Big)}{(n-k)^2(n-k-2)}. \end{aligned}$$

Given that the aim of ESF is to eliminate any spatial correlation patterns, the

intuition behind calibrating the tuning parameter this way is by assuming that when the level of spatial correlation in the residuals is low, only a small set of eigenvectors is necessary, thus a high level of regularization (large tuning parameter) is required. In contrast, when the level of spatial correlation is high, a large set of eigenvectors will be necessary, thus a low level of regularization (small tuning parameter) is required. As z gives a large value when the overall correlation is high and small values when the overall correlation is low, they propose using the absolute value of the standardised Moran's i with a negative exponent as the tuning parameter.[5]

Our proposed procedure is outlined in Algorithm 3 and is called Moran's i 2 stage Lasso (Mi-2SL). The proposed estimator can handle both endogenous covariates and cross-sectional dependence.

The procedure is relatively simple first, do a spectral decomposition of the SWM to get the candidate set of eigenvectors. Then calculate standardised Moran's i on the naive first stage residuals (ignoring the spatial correlation), which is used to estimate (3.7) to get $\hat{\boldsymbol{x}}_2$ using the Lasso (or post-Lasso) estimates and the extracting selected eigenvector $\hat{\boldsymbol{E}}_x$. Subsequently, $\hat{\boldsymbol{x}}_2$ is used instead of $\boldsymbol{x}_2$ to calculate standardised Moran's i for the naive second stage residuals (ignoring the spatial correlation), which is then used to estimate (3.8) to get the selected eigenvector $\hat{\boldsymbol{E}}_y$. Finally estimate β_2 by 2SLS using the union of $\hat{\boldsymbol{E}}_x$ and $\hat{\boldsymbol{E}}_y$ as controls.

[5] A positive tuning parameter is required for the Lasso solution to be unique. Thus, the absolute value is used.

1. Decompose the SWM to get the candidate set of eigenvectors $\boldsymbol{E}$.

2. Estimate simple first stage residuals $\hat{\boldsymbol{r}} = \boldsymbol{M}_H \boldsymbol{x}_2$ where $\boldsymbol{M}_H = \boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'$ and $\boldsymbol{H} = (\boldsymbol{X}_1, \boldsymbol{Z}_2)$ and calculate corresponding the standardised Moran's I of $\hat{\boldsymbol{r}}$ denoted z_x

3. Estimate

$$[\hat{\boldsymbol{\zeta}}_1, \hat{\boldsymbol{\zeta}}_2, \hat{\boldsymbol{\gamma}}_x] \in \min\{||\boldsymbol{x}_2 - \boldsymbol{X}_1\boldsymbol{\zeta}_1 - \boldsymbol{Z}_2\boldsymbol{\zeta}_2 - \boldsymbol{E}\boldsymbol{\gamma}_x||_2^2 + |z_x|^{-a}||\boldsymbol{\gamma}_x||_1\} \tag{3.7}$$

   Use the Lasso or post-Lasso estimates of (3.7) to get the fitted $\hat{\boldsymbol{x}}_2$ and save the selected set of eigenvector $\hat{\boldsymbol{E}}_x$.

4. Estimate simple second stage residuals $\hat{\boldsymbol{h}} = \boldsymbol{M}_{\hat{\boldsymbol{X}}}\boldsymbol{y}$ where $\boldsymbol{M}_{\hat{\boldsymbol{X}}} = \boldsymbol{I} - \hat{\boldsymbol{X}}(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'$ and $\hat{\boldsymbol{X}} = (\boldsymbol{X}_1, \hat{\boldsymbol{x}}_2)$ and calculate corresponding the standardised Moran's I of $\hat{\boldsymbol{h}}$ denoted z_y

5. Estimate

$$[\hat{\boldsymbol{\beta}}_1, \hat{\beta}_2, \hat{\boldsymbol{\gamma}}_y] \in \min\{||\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \hat{\boldsymbol{x}}_2\beta_2 - \boldsymbol{E}\boldsymbol{\gamma}_y||_2^2 + |z_y|^{-a}||\boldsymbol{\gamma}_y||_1\} \tag{3.8}$$

   and save the selected set of eigenvector $\hat{\boldsymbol{E}}_y$.

6. Estimate β_2 by 2SLS using both $\hat{\boldsymbol{E}}_x$ and $\hat{\boldsymbol{E}}_y$ as controls.

Algorithm 3: Mi-2SL Algorithm

3.4 Theoretical results

We will now derive some theoretical properties of the 2SLS ESF procedure. In scalar notation, the spatial model (3.1) can be rewritten as

$$y_i = \sum_{l=1}^{p}\sum_{j=1}^{n} w_{ij}^{l} y_i \rho_{l,0} + \sum_{j=1}^{k_1} x_{ij,1}\beta_{j,1,0} + x_{i,2}\beta_{2,0} + \sum_{j=1}^{n}\sum_{l=1}^{k_1} w_{ij} x_{il,1}\psi_{l,0} + \varepsilon_i,\ i = 1, \ldots, n.$$

We now make the following assumptions about (3.1) and (3.2).

Assumption 5.

1. *(a)* $\boldsymbol{W}$ *are stochastic real symmetric* $n \times n$ *matrices with* $w_{ii} = 0$. *(b)* $\boldsymbol{S}_1$ *and* $\boldsymbol{S}_2$ *are non-singular for all* n. *The sequences* $\{\boldsymbol{W}\}$, $\{\boldsymbol{S}_1^{-1}\}$ *and* $\{\boldsymbol{S}_2^{-1}\}$ *are uniformly bounded in both row and column sums. (c) The largest eigenvalue of* $\boldsymbol{W}$ *is bound,* $\max_i \lambda_i < \infty$.

2. *(a) The* $n \times q$ *instrument matrices* $\boldsymbol{Z}_2$ *and the* $n \times (k_1 + 1)$ *matrices* $[\boldsymbol{X}_1, \boldsymbol{x}_2]$ *both have full column rank (for a large enough* n*),* $\mathbb{E}[\boldsymbol{X}_1'\boldsymbol{\varepsilon}] = 0$ *and* $\mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{\varepsilon}] = 0$ *and (b) all the elements of* $\boldsymbol{Z}_2$, $\boldsymbol{x}_2$ *and* $\boldsymbol{X}_1$ *are uniformly bound in absolute value.*

3. *The innovations* $\{\varepsilon_i : 1 \leq i \leq n, n \geq 1\}$ *are identically distributed triangular arrays. Further the innovations* $\{\varepsilon_i : 1 \leq i \leq n\}$ *are for each* n *distributed (jointly) independently with* $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ *and* $\mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2 \in (0, \infty)$.

Assumptions 5.1(b)-3 are standard assumptions in the spatial econometrics literature (Kelejian and Prucha, 1998, 1999; Lee, 2004). Assumption 5.1 (a) is required

for the spectral decomposition. Assumption 5.1 (b) is necessary to ensure the model is complete and the assumption that the row and column sums are uniformly bound and to limit the degree of dependence in $\boldsymbol{y}$ and $\boldsymbol{x}_2$. Assumption 5.1 (c) ensure the elements of the eigenvectors have the same dependence coefficient as the elements of the SWM. Assumption 5.2 (a) $\mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{\varepsilon}] = 0$ is necessary for variables in $\boldsymbol{Z}_2$ to be valid instruments and the assumption $q \geq 1$ corresponds to the order condition which says the number of instruments must be greater than or equal to the number of endogenous variables.

We also make the following assumptions about the ESF approximation.

Assumption 6.

1. $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}_1) = \boldsymbol{E}\boldsymbol{\gamma}_{y,0} = \boldsymbol{E}_{\Omega_y}\boldsymbol{\gamma}_{\Omega_y}$ *and* $g(\boldsymbol{W}, \boldsymbol{x}_2, \boldsymbol{X}_1, \boldsymbol{Z}_2) = \boldsymbol{E}\boldsymbol{\gamma}_{x,0} = \boldsymbol{E}_{\Omega_x}\boldsymbol{\gamma}_{\Omega_x}$ *where* $\boldsymbol{E}_{\Omega_y}$ *and* $\boldsymbol{E}_{\Omega_x}$ *are* $n \times s_2$ *and* $n \times s_1$ *matrices with columns that correspond to the active sets* $\Omega_y := \text{supp}(\boldsymbol{\gamma}_{y,0})$ *and* $\Omega_x := \text{supp}(\boldsymbol{\gamma}_{x,0})$*, and* $\boldsymbol{\gamma}_{\Omega_y}$ *and* $\boldsymbol{\gamma}_{\Omega_x}$ *the corresponding vectors of unknown constants.*

2. $|\Omega| = s < n - k_1 - q$ *where* $\Omega = \Omega_y \cup \Omega_x$.

Assumption 6.1 is a strong assumption that says weighted sets of eigenvectors are enough to approximate all the spatially correlated terms in the first and second stage of the model (i.e., the ESF approximation is valid). We view these as approximations of $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}_1)$ and $g(\boldsymbol{W}, \boldsymbol{x}_2, \boldsymbol{X}_1, \boldsymbol{Z}_2)$. Assumption 6.1 is a weak sparsity assumption on the total number of eigenvectors with non-zero coefficients. Unfortunately, verifying these assumptions in practice or even in simulations is impossible.

Under these assumptions, we now have the following low dimensional ESF reduced form system of equations

$$\boldsymbol{y} = \boldsymbol{G}_{\Omega}\boldsymbol{\Upsilon}_{\Omega} + \boldsymbol{\varepsilon}, \tag{3.9}$$

$$\boldsymbol{x}_2 = \boldsymbol{Z}_{\Omega}\boldsymbol{\zeta}_{\Omega} + \boldsymbol{u}_2, \tag{3.10}$$

where $\boldsymbol{G}_{\Omega} = [\boldsymbol{X}_1, \boldsymbol{x}_2, \boldsymbol{E}_{\Omega}]$, $\boldsymbol{Z}_{\Omega} = [\boldsymbol{X}_1, \boldsymbol{Z}_2, \boldsymbol{E}_{\Omega}]$, $\boldsymbol{\Upsilon}_{\Omega} = [\boldsymbol{\beta}_{1,0}, \beta_{2,0}, \boldsymbol{\gamma}_{\Omega}]'$ and $\boldsymbol{\zeta}_{\Omega} = [\boldsymbol{\zeta}_{1,0}, \boldsymbol{\zeta}_{2,0}, \boldsymbol{\zeta}_{3,\Omega}]'$.

We will now derive a consistency proof for estimating $\boldsymbol{\Upsilon}_{\Omega}$ by 2SLS, assuming Ω is known. Even under perfect selection, estimating (3.9)-(3.10) by 2SLS is non-trivial, as the standard weak law of large numbers (LLN) and central limit theorem for triangular arrays used for spatial models requires assuming the row-wise independence. This is not realistic for the elements of $\boldsymbol{E}$, which are constructed from a linear transformation of $\boldsymbol{W}$, a matrix which encapsulates the spatial structure. We thus, need to formalise this dependence and apply appropriate limit theorems.

To model this dependence, we use the notion of ψ-dependence first proposed by Doukhan and Louhichi (1999) for time-series data and adapted by Kojevnikov et al. (2021) to allow for cross-sectional dependence. This allows us to use the limit theorems proposed by Kojevnikov et al. (2021). Roughly speaking, ψ-dependence measures the strength of dependence between two sets of random variables by the covariance of non-linear functions of the random variables.

Suppose we observe some spatial structure which is a function of N, and let

$d_{ij,n} = d_{ij}$ denote the (shortest) distance between ith and jth cross-sectional unit given the spatial structure, the distance d is a metric on the set N. We view the cross-sectional dependence as a stochastic dependence pattern of random variables on the metric d in the spatial structure.

Let $\{w_{ij}, 1 \leq i \leq n, n \geq 1\}$, $j = 1, \ldots, n$, be a triangular arrays of random variables, where $w_{ij} = w_{ij,n}$ denotes the i, jth element of matrix $\boldsymbol{W}$ which is derived from the spatial structure. For any $a \in \mathbb{N}$, we endow $\mathbb{R}^a$ with distance

$$\mathtt{d}_a(\boldsymbol{q}, \boldsymbol{h}) = \sum_{l=1}^{a} |q_l - h_l|,$$

where $\boldsymbol{q} = (q_1, \ldots, q_a)$ and $\boldsymbol{h} = (h_1, \ldots, h_a)$ are points in $\mathbb{R}^a$.

Let $\mathcal{L}_a$ denote a family of real values, bounded Lipschitz functions

$$\mathcal{L}_a = \{f : \mathbb{R}^a \to \mathbb{R} : ||f||_\infty < \infty;\ \mathrm{Lip}(f) < \infty\},$$

where $\mathrm{Lip}(f)$ is the Lipschitz constant of f,[6] and $||f||_\infty = \sup_x |f(x)|$ is the sup-norm of f.

Now consider two sets of cross-sectional units (of size a and $b \in \mathbb{N}$) with a distance between each other of at least $r > 0$. Let $\mathcal{P}_{a,b;r}$ denote the collections of all pairs

$$\mathcal{P}_{a,b;r} = \{(A, B) : A, B \subset N, |A| = a,\ |B| = b,\ d_{A,B} \geq r\}$$

where $d_{A,B} = \min_{i \in A} \min_{j \in B} d_{ij}$ (note $\mathcal{P}_{a,b;r}$, $d_{A,B}$ and d_{ij} are also implicitly indexed

[6]The Lipschitz constant for a function $f : \mathbb{R}^a \to \mathbb{R}$ is the smallest constant C such that $|f(\boldsymbol{q}) - f(\boldsymbol{h})| \leq C\mathtt{d}_a(\boldsymbol{q}, \boldsymbol{h}), \forall\ \boldsymbol{q}, \boldsymbol{h} \in \mathbb{R}^a$.

by n, but we omit the index to simplify the notation). For each set A of positive integers we say $w_A = \{w_{ij} : i \in A\}$.

We take $\{\mathcal{C}_n = \mathcal{C}\}$ be a sequence of given σ-fields, such that for each $n \geq 1$, the spatial weights matrix $\boldsymbol{W}_n = \boldsymbol{W}$ is $\mathcal{C}$-measurable. Definition 2 gives the exact definition of conditional ψ dependence we use.

Definition 2. *(Kojevnikov et al., 2021) The triangular array* $\{w_{ij,n} = w_{ij}, 1 \leq i \leq n, n \geq 1\}$, $j = 1, \ldots, n$ *is called conditionally* ψ*-dependent given* $\{\mathcal{C}_n = \mathcal{C}\}$, *if for each* $n \in \mathbb{N}$ *there exists a* $\mathcal{C}$*-measurable sequence* $\mu_r = \{\mu_r = \mu_{r,n} : r \geq 0\}$, $\mu_0 = 1$, *and a collection of non-random functions* $\psi_{a,b} : \mathcal{L}_a \times \mathcal{L}_b \to [0, \infty)$ *such that for all* $(A, B) \in \mathcal{P}_{a,b;r}$ *with* $r > 0$ *and all* $f \in \mathcal{L}_a$ *and* $g \in \mathcal{L}_b$,

$$|\operatorname{Cov}(f(w_A), g(w_B)|\mathcal{C})| \leq \psi_{a,b}(f, g)\mu_r \quad \text{a. s.} \tag{3.11}$$

The sequence $\{\mu_r\}$ *is the dependence coefficients of* $\{w_{ij}\}$.

We will now explicitly specify the latent spatial formation process. We consider binary connectivity based on physical distance plus some stochastic elements. Specifically, the connection for each pair of spatial units i and j $(i \neq j)$ is randomly realised if and only if

$$w_{ij} = \mathbb{1}\{\phi_{ij} \geq \eta_{ij}\},$$

where the ϕ_{ij}'s and η_{ij}'s are random variables such that $\phi_{ij,n} = \phi_{ij} = \phi_{ji}$, $\eta_{ij} = \eta_{ji}$ and $\{\eta_{ij} \ : \ i < j\}$ is *i.i.d.* and independent of $\phi = (\phi_{ij})_{i<j}$. The random variable ϕ_{ij} which determines the formation probabilities, is assumed to be a function of

observable characteristics $\boldsymbol{l}_{ij,n} = \boldsymbol{l}_{ij}$ (e.g., the physical distance between the spatial units) and unit specific unobservable characteristics $\boldsymbol{t}_{i,n} = \boldsymbol{t}_i$ (i.e. $\phi_{ij} = f(\boldsymbol{t}_i, \boldsymbol{t}_j, \boldsymbol{l}_{ij})$ where $f(\cdot)$ is some function). Thus, the σ-field $\mathcal{C}$ is generated by $\boldsymbol{t}_i$, $\boldsymbol{t}_j$ and $\boldsymbol{l}_{ij}$ for all i and j.

3.4.1 Estimation consistency

In scalar notation (3.9) can be rewritten as

$$y_i = \sum_{j=1}^{k_1} x_{ij,1}\beta_{j,1,0} + x_{i,2}\beta_{2,0} + \sum_{j=1}^{s} e_{ij}\gamma_{j,\Omega} + \varepsilon_i = \sum_{j=1}^{(k_1+1+s)} g_{ij,\Omega}\Upsilon_{j,\Omega} + \varepsilon_i,$$
$$x_{i,2} = \sum_{j=1}^{k_1} x_{ij,1}\zeta_{j,1,0} + \sum_{j=1}^{q} z_{ij,2}\zeta_{j,2,0} + \sum_{j=1}^{s} e_{ij}\zeta_{j,3,\Omega} + u_{i,2} = \sum_{j=1}^{(k_1+q+s)} z_{ij,\Omega}\psi_{j,\Omega} + u_{i,2},$$

for $i = 1, \ldots, n$.

We now state the additional assumptions for consistent estimation of $\beta_{2,0}$ by 2SLS.

Assumption 7.

1. *The triangular array $\{w_{ij}\}$, is conditionally ψ-dependent given $\{\mathcal{C}\}$ with the dependence coefficients $\{\mu_r\}$ satisfying the following condition. For some constant $C > 0$*

$$\psi_{a,b}(f, g) \leq Cab(||f||_\infty + \text{Lip}(f))(||g||_\infty + \text{Lip}(g)) \tag{3.12}$$

2. *for some* $l > 2$ $\sup_{n\geq 1} \max_{i\in N} \left(\mathbb{E}[\sum_{j=1}^{(k_1+1+s)} |g_{ij,\Omega}|^l | \mathcal{C}] \right)^{1/l} < \infty$ *a.s.,*
$\sup_{n\geq 1} \max_{i\in N} \left(\mathbb{E}[|y_i|^l | \mathcal{C}] \right)^{1/l} < \infty$ *a.s. and*
$\sup_{n\geq 1} \max_{i\in N} \left(\mathbb{E}[\sum_{j=1}^{(k_1+q+s)} |z_{ij,\Omega}|^l | \mathcal{C}] \right)^{1/l} < \infty$ *a.s.*

3.
$$n^{-1} \sum_{r=1}^{\infty} \delta_r^d \mu_r \to_{a.s.} 0, \quad n \to \infty \tag{3.13}$$
where $\delta_r^d = n^{-1} \sum_{i\in N} |N_{i,r}^d|$ *and* $N_{i,r}^d = \{j \in N : d_{i,j} = r\}$ *denotes the set of cross-sectional units exactly distance* r *from unit* i*.*

4. $\mathbb{E}[\boldsymbol{Z}_\Omega' \boldsymbol{\varepsilon} | \mathcal{C}] = 0$.

Assumption 7.1 is from Kojevnikov et al. (2021) and the function $\psi_{a,b}$ satisfies Assumption 7.1 if

$$\sup_{n\geq 1} \max_{i\in N} \mathbb{E}[|w_{ij}|^q | \mathcal{C}_n] < \infty \quad \text{a. s.}$$

for some $q > 4$ and $\forall j$. Assumption 7.2 states that all variables have conditional finite second moments, so all are $\mathcal{C}$ measurable. Assumption 7.3 is also from Kojevnikov et al. (2021) and puts a restriction on the denseness of the spatial structure and the rate of decay of dependence with regards to the distance between the spatial units. In the mixing literature, it is common to assume the mixing coefficients can be summed $n^{-1} \sum_{r=1}^{\infty} \mu_r = O_p(1)$ as $n \to \infty$. A sufficient condition for Assumption 7.3, in this case, is if the average number of neighbours at distance r grows slower than the sample size n i.e. $\sup_{r\geq 1} \delta_r^d = o_p(n)$. The basic idea of this assumption is it requires the number of spatial connects at distance r must not grow too fast as r

increases, however, as the precise condition (3.13) includes the dependence coefficient μ_r, this assumption can be relaxed if μ_r decreases appropriately fast with r. This assumption seems reasonable as in the literature on estimating SWMs a sparse spatial structure is often assumed (Ahrens and Bhattacharjee, 2015; Lam and Souza, 2016, 2020). An example of where Assumption 7.3 could fail is if one unit is connected to all other units, such as in the star network. This is because the distance between any two units is never larger than 2.[7] Assumption 7.4 requires the instruments (including $\boldsymbol{E}_\Omega$) are uncorrelated with the error, conditional on $\mathcal{C}$.

It is important to clarify that assumption 7.1, which requires that w_{ij} are ψ-dependent triangular arrays, carries over to the eigenvector elements e_{ik}, which are generated by a linear combination of w_{ij}, λ_k and e_{jk},

[7] $\delta_1^d = 2(n-1)/n$, $\delta_2^d = (n-2)(n-1)/n$ and $\delta_r^d = 0$ for $r \geq 3$

$$\boldsymbol{W} = \sum_{k=1}^{n} \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k'$$

$$\boldsymbol{W}\boldsymbol{e}_k = \lambda_k \boldsymbol{e}_k, \quad \forall k \in N$$

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \begin{bmatrix} e_{1k} \\ e_{2k} \\ \vdots \\ e_{nk} \end{bmatrix} = \lambda_k \begin{bmatrix} e_{1k} \\ e_{2k} \\ \vdots \\ e_{nk} \end{bmatrix}$$

$$\begin{bmatrix} \sum_{j=1}^{n} w_{1j} e_{jk}/\lambda_k \\ \sum_{j=1}^{n} w_{2j} e_{jk}/\lambda_k \\ \vdots \\ \sum_{j=1}^{n} w_{nj} e_{jk}/\lambda_k \end{bmatrix} = \begin{bmatrix} e_{1k} \\ e_{2k} \\ \vdots \\ e_{nk} \end{bmatrix}. \tag{3.14}$$

The second equality holds as the columns of the eigenvectors matrix ($\boldsymbol{e}_k$) are mutually orthonormal,[8] and (3.14) holds if and only if $\lambda_k \neq 0 \ \forall k$ i.e. none of the eigenvalues are zero.

Lemma 3. *Suppose the triangular array* $\{w_{ij}\}$*, with* $w_{ij} \in \mathbb{R}$ *satisfies Assumption 7.1 with dependence coefficient* $\{\mu_r\}$*. For each* $n \geq 1$ *let* $\{\lambda_{k,n} = \lambda_k\}_{k \in N}$*,* $\lambda_k \in \mathbb{R}$ *and* $\{\boldsymbol{e}_{k,n} = \boldsymbol{e}_k\}_{k \in N}$*,* $\boldsymbol{e}_k \in \mathbb{R}^n$ *be a sequence of* $\mathcal{C}$ *measurable random scalars and random vectors with* $\max_{k \in N} |\lambda_k| \leq \infty$ *a.s. and* $||\boldsymbol{e}_k||_2^2 = 1 \ \forall k$*. Then the array* $\{e_{ik}\}$ *defined by* $e_{ik} = \sum_{j=1}^{n} w_{ij} e_{jk}/\lambda_k, i = 1, \ldots, n$ *and* $k = 1, \ldots, n$ *is conditionally* ψ*-dependent*

[8] $\boldsymbol{e}_j'\boldsymbol{e}_j = 1$ and $\boldsymbol{e}_j'\boldsymbol{e}_k = 0 \ \forall j \neq k$

given $\{\mathcal{C}\}$ *with the dependence coefficients* $\{\mu_r\}$,

$$\Big| \operatorname{Cov}\left(f\Big(\sum_{j\in N} w_a e_{jk}/\lambda_k\Big), g\Big(\sum_{j\in N} w_b e_{jk}/\lambda_k\Big) | \, \mathcal{C}\right)\Big| \leq \psi_{a,b}(f_c, g_c)\mu_r \quad \text{a. s.}$$

The proof is provided in Section 3.8.1.

Lemma 3 shows that as long as the largest eigenvalue is bound and the eigenvectors are mutually orthogonal (both of these requirements are satisfied by Assumption 5.1) the eigenvector elements will have the same dependence coefficients $\{\mu_r\}$ as the elements of the SWM.

Theorem 3. *Assuming Assumption 5-7 holds we have*

$$\hat{\boldsymbol{\Upsilon}}_\Omega \to_p \boldsymbol{\Upsilon}_\Omega,$$

where $\hat{\boldsymbol{\Upsilon}}_\Omega$ *is the 2SLS estimate of* $\boldsymbol{\Upsilon}_\Omega$ *from* (3.9)*, as* $n \to \infty$.

The proof is provided in Section 3.8.1.

Theorem 3 shows that under an appropriate mixing condition, some additional regularity conditions and if Ω is known, we could estimate $\boldsymbol{\Upsilon}_\Omega$ consistently by 2SLS. The proof of Theorem 3 uses the weak LLN for triangular arrays, which gives convergence in probability, and the strong LLN for cross-sectionally dependent random variables of Kojevnikov et al. (2021) which gives almost sure convergence, thus, overall gives convergence in probability. An almost sure convergence result could be obtained similarly by using the strong LLN for triangular arrays instead of the weak LLN for triangular arrays.

3.4.2 Asymptotic Distribution

We need some additional assumptions, to derive the asymptotic distribution of the 2SLS estimator for known Ω.

Assumption 8.

1. *for some* $l > 4$ $\sup_{n\geq 1} \max_{i\in N} \left(\mathbb{E}[\sum_{j=1}^{(k_1+1+s)} |g_{ij,\Omega}|^l | \mathcal{C}] \right)^{1/l} < \infty$ *a.s.,*
   $\sup_{n\geq 1} \max_{i\in N} \left(|y_i|^l | \mathcal{C}] \right)^{1/l} < \infty$ *and*
   $\sup_{n\geq 1} \max_{i\in N} \left(\mathbb{E}[\sum_{j=1}^{(k_1+q+s)} |z_{ij,\Omega}|^l | \mathcal{C}] \right)^{1/l} < \infty$

2. *There exists a positive sequence* $m_n = m \to \infty$ *such that for* $k = 1, 2$

$$n\boldsymbol{\Sigma}^{-(2+k)} \sum_{r=0}^{\infty} c_{r,m;k} \mu_r^{1-\frac{2+k}{l}} \to_{a.s.} 0, \tag{3.15}$$

$$n^2 \mu_m^{1-1/l} \boldsymbol{\Sigma}^{-1} \to_{a.s.} 0, \tag{3.16}$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{Z}_\Omega' \boldsymbol{Z}_\Omega | \mathcal{C}] \sigma_\varepsilon^2$, $c_{r,m;k} = \inf_{\alpha>1} \left[\Delta_{r,m;k\alpha}\right]^{1/\alpha} \left[\delta^d_{r,\alpha/(1-\alpha)}\right]^{1-1/\alpha}$,
$\delta^d_{r,k} = n^{-1} \sum_{i\in N} |N^d_{i,r}|^k$, $\Delta_{r,m;k} = n^{-1} \sum_{i\in N} \max_{j\in N^d_{i,r}} |N_{i,m}/N_{j,r-1}|^k$, $N_{i,r} = \{j \in N : d_{i,j} \leq r\}$, $N^d_{i,r} = \{j \in N : d_{i,j} = r\}$ *and* $l > 4$ *is as same as in Assumption 8.1. As* $n \to \infty$.

Assumption 8.1 states that all variables have at least conditional fourth finite moment, so are all $\mathcal{C}$ measurable, which is in line with many spatial and 2SLS models. Assumption 8.2 is from Kojevnikov et al. (2021) and limits the extent of the spatial dependence of the random variables through restrictions on the spatial structure. When the spatial structure is given $c_{r,m;k}$ can be computed, it is composed of

two parts $\Delta_{r,m;k\alpha}$ and $\delta^d_{r,\alpha/(1-\alpha)}$, which capture the denseness of the spatial structure through the average size of neighbourhoods and the average shell size of the neighbourhood. Note that after r goes beyond a certain level $\Delta_{r,m;k}$ tends to decrease fast, as the set $N_{j,r-1}$ becomes large quickly. For (3.16) to be satisfied μ_r (the spatial dependence) needs to decay fast enough as r becomes large, this is because it will become increasingly difficult to find a slowly increasing sequence m to satisfy the condition.

Theorem 4. *Assuming Assumptions 5-8 holds we have*

$$\sqrt{n}(\hat{\boldsymbol{\Upsilon}}_\Omega - \boldsymbol{\Upsilon}_\Omega) \to_d N(0, \left(\mathbb{E}[\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega | \mathcal{C}]\, \mathbb{E}[\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega | \mathcal{C}]^{-1}\, \mathbb{E}[\boldsymbol{Z}'_\Omega \boldsymbol{G}_\Omega | \mathcal{C}]\right)^{-1} \sigma^2_\varepsilon)$$

where $\hat{\boldsymbol{\Upsilon}}_\Omega$ *is the 2SLS estimate of* $\boldsymbol{\Upsilon}_\Omega$ *from* (3.9)*, as* $n \to \infty$*.*

The proof is provided in Section 3.8.1.

Theorem 4 shows that if Ω is known, then under an appropriate mixing condition, restriction on the denseness of the spatial structure, and some additional regularity conditions, the 2SLS estimate of $\boldsymbol{\Upsilon}_\Omega$ and thus, $\beta_{2,0}$ will be asymptotically normal. It is important to note that the rate of convergence will be $n^{-1/2}$.

3.5 Simulation

In this section, we provide simulation evidence to assess the finite sample performance of the Mi-2SL estimator and compare its performance to correctly and incorrectly specified/estimated models. We generate the following system of equations (3.17) -

(3.18) where the equation includes a SAR(4) and the endogenous variable a SAR(1):

$$\boldsymbol{y} = \sum_{i=1}^{4} \boldsymbol{W}^i \boldsymbol{y}\rho_i + \alpha\boldsymbol{\iota} + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}, \tag{3.17}$$

$$\boldsymbol{x}_2 = + \alpha\boldsymbol{\iota} + \zeta_1\boldsymbol{z}_2 + \zeta_2\boldsymbol{x}_1 + \boldsymbol{W}\boldsymbol{x}_2\zeta_3 + \boldsymbol{v}, \tag{3.18}$$

$$(u_i, v_i) \sim N\left(0, \begin{pmatrix} \sigma_u^2 & \sigma_{v,u} \\ \sigma_{v,u} & \sigma_v^2 \end{pmatrix}\right),$$

$$\boldsymbol{z}_2 \sim N(0, \boldsymbol{I}),$$

$$\boldsymbol{x}_1 \sim N(0, \boldsymbol{I}),$$

where u_i and v_i are the ith elements of $\boldsymbol{u}$ and $\boldsymbol{v}$, $\boldsymbol{\iota}$ is a unit vector, $\beta_2 = 1$ is the parameter of interest and $\boldsymbol{W}$ is a symmetric SWM. Additionally we set $\sigma_u^2 = \sigma_v^2 = 1$ $\sigma_{v,u}^2 = 0.3, 0.6, 0.9$, $\zeta_1 = \zeta_2 = \alpha = \beta_1 = \beta_2 = 1$, $\zeta_3 = 0.6$ and $\rho_1 = 0.6$, $\rho_2 = 0.5$, $\rho_3 = 0.4$ and $\rho_4 = 0.3$.

The elements of $\boldsymbol{W}$ denoted w_{ij}, are independent draws from a Bernoulli distribution with success probability $p = \mu/n$ for some constant $\mu < \infty$, $w_{ii} = 0$ and $w_{ij} = w_{ji}$. By this construction, each unit's expected number of links equals μ. We set$\mu = 4, 8$ to see how the estimators perform at different SWM densities. Each SWM is normalised by the largest row sum and the eigenvectors are from the normalised SWM. Sample sizes considered are $n = 100, 250, 500$, and we run 1000 replication.

The models estimated are:

1. Simple OLS, i.e., estimating $\boldsymbol{y} = \alpha\boldsymbol{\iota} + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}$ by OLS (denoted

simpleOLS) ignoring the spatial process and that $\boldsymbol{x}_1$ is endogenous.

2. Simple IV i.e. estimating $\boldsymbol{y} = \alpha\boldsymbol{\iota} + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}$ by IV where $\boldsymbol{z}_1$ is used as instrument for $\boldsymbol{x}_1$ (denoted simpleIV).

3. Spatial Autoregressive (SAR(p)) model but equation for $\boldsymbol{x}_2$ ignores the spatial process i.e. estimating $\boldsymbol{y} = \alpha\boldsymbol{\iota} + \sum_{i=1}^{p}\rho_i\boldsymbol{W}^i\boldsymbol{y} + \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{u}$, by GS2SLS where $\sum_{i=1}^{p+2}\boldsymbol{W}^i\boldsymbol{x}_1$ are being used as an instrument for $\sum_{i=1}^{p}\boldsymbol{W}^i\boldsymbol{y}$ and $\boldsymbol{z}_2$ is used as instrument for $\boldsymbol{x}_2$ (denoted GS2SLSa-SAR(p)).

4. Spatial Autoregressive (SAR(p)) model with $\boldsymbol{x}_2$ estimated as a SAR(1) i.e. estimating $\boldsymbol{y} = \alpha\boldsymbol{\iota}+\sum_{i=1}^{p}\rho_i\boldsymbol{W}^i\boldsymbol{y}+\boldsymbol{x}_1\beta_1+\boldsymbol{x}_2\beta_2+\boldsymbol{u}$, by GS2SLS where $\sum_{i=1}^{p+2}\boldsymbol{W}^i\boldsymbol{x}_1$ are being used as an instrument for $\sum_{i=1}^{p}\boldsymbol{W}^i\boldsymbol{y}$, $\boldsymbol{z}_2$ is used as instrument for $\boldsymbol{x}_2$ and $\boldsymbol{W}\boldsymbol{z}_2$ being used as an instrument for $\boldsymbol{W}\boldsymbol{x}_2$ (denoted GS2SLSb-SAR(p)) i.e. the true model when SAR(4) is estimated.

5. The estimates from the Mi-2SL Algorithm 3 with the first-stage fitted values from Lasso (step 3) and $a = 1$ (denoted Mi-2SLl-a1).

6. The estimates from the Mi-2SL Algorithm 3 with the first-stage fitted values from Lasso (step 3) and $a = 2$ (denoted Mi-2SLl-a2).

Note GS2SLSb-SAR(4)* is the correctly specified and estimated model. The results present in this section are only for β_2. See the tables in Section 3.8.2 for the results for β_1, SAR(2), SAR(3), and Mi-2SL estimates with the first-stage fitted values from post Lasso (Mi-2SLpl).

Table 3.1 and 3.2 shows the bias and MSE of β_2 and the number of selected eigenvectors at the first-stage and second-stages, and the number of eigenvectors used in the IV regression (the union of the first and second stage selected eigenvectors) when $\mu = 4$ and $\mu = 8$ for different levels of error correlation and sample sizes.

Table 3.1: Bias and MSE of β_2 and number of selected eigenvectors, $\mu = 4$

$\sigma^2_{v,u}$	Estimator	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]
		n=100		n=250		n=500	
0.3	GS2SLSa-SAR(1)	0.009(0.016)	-	0.018(0.005)	-	0.017(0.003)	-
0.3	GS2SLSa-SAR(4)	-0.001(0.016)	-	0.002(0.006)	-	0.001(0.003)	-
0.3	GS2SLSb-SAR(1)	-0.001(0.013)	-	0.016(0.005)	-	0.015(0.003)	-
0.3	GS2SLSb-SAR(4)*	-0.008(0.014)	-	0.001(0.005)	-	0.001(0.003)	-
0.3	simpleOLS	0.427(0.212)	-	0.28(0.082)	-	0.296(0.09)	-
0.3	simpleIV	0.055(0.048)	-	0.047(0.01)	-	0.048(0.006)	-
0.3	Mi-2SLl-a1	0.006(0.022)	32[32,0]	0.016(0.007)	27[27,0]	0.014(0.003)	82[82,0]
0.3	Mi-2SLl-a2	0.098(0.088)	81[80,10]	0.068(0.02)	169[164,23]	0.11(0.026)	418[410,78]
0.6	GS2SLSa-SAR(1)	0.01(0.016)	-	0.018(0.005)	-	0.017(0.003)	-
0.6	GS2SLSa-SAR(4)	0.001(0.017)	-	0.003(0.006)	-	0.002(0.003)	-
0.6	GS2SLSb-SAR(1)	0.004(0.013)	-	0.017(0.005)	-	0.016(0.003)	-
0.6	GS2SLSb-SAR(4)*	-0.005(0.014)	-	0.001(0.005)	-	0.001(0.003)	-
0.6	simpleOLS	0.58(0.365)	-	0.431(0.189)	-	0.447(0.202)	-
0.6	simpleIV	0.052(0.048)	-	0.046(0.01)	-	0.047(0.006)	-
0.6	Mi-2SLl-a1	-0.001(0.023)	34[34,0]	0.015(0.007)	31[31,0]	0.012(0.003)	91[91,0]
0.6	Mi-2SLl-a2	0.092(0.056)	81[80,11]	0.07(0.019)	168[165,23]	0.118(0.029)	414[409,78]
0.9	GS2SLSa-SAR(1)	0.012(0.017)	-	0.019(0.005)	-	0.017(0.003)	-
0.9	GS2SLSa-SAR(4)	0.001(0.017)	-	0.004(0.007)	-	0.002(0.003)	-
0.9	GS2SLSb-SAR(1)	0.008(0.013)	-	0.019(0.005)	-	0.016(0.002)	-
0.9	GS2SLSb-SAR(4)*	-0.002(0.014)	-	0.002(0.005)	-	0.001(0.003)	-
0.9	simpleOLS	0.733(0.564)	-	0.582(0.341)	-	0.598(0.36)	-
0.9	simpleIV	0.05(0.049)	-	0.045(0.01)	-	0.047(0.006)	-
0.9	Mi-2SLl-a1	-0.006(0.025)	35[35,0]	0.014(0.007)	36[36,0]	0.012(0.004)	99[99,0]
0.9	Mi-2SLl-a2	0.079(0.043)	80[80,11]	0.067(0.017)	169[167,22]	0.111(0.024)	411[409,78]

Note: Bias (MSE) and GS2SLSb-SAR(4)* is the correctly specified and estimated model.

These tables show the number of selected eigenvectors is always higher in the second-stage than first-stage regardless of the exponent a, which is reasonable given in this simulation setup, the overall level of spatial correlation is higher in the DGP of $\boldsymbol{y}$ (second stage) than $\boldsymbol{x}_2$ (first stage), due to the higher-order lags. The larger

the exponent, the more eigenvectors are selected regardless of error correlation, μ or sample size. For small sample sizes (100) the number of included eigenvector does not change much as the error correlation increases, where as for larger sample sizes (500) when $a = 1$ the number of included eigenvectors increases and when $a = 2$ the number of included eigenvectors decrease, for a given μ. From the number of selected eigenvectors at each stage we can see this decrease is actually being driven by greater overlap between the sets of eigenvectors selected at the first and second stages.

Table 3.2: Bias and MSE of β_2 and number of selected eigenvectors, $\mu = 8$

$\sigma^2_{v,u}$	Estimator	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]
		n=100		n=250		n=500	
0.3	GS2SLSa-SAR(1)	0.004(0.012)	-	0.011(0.005)	-	0.013(0.003)	-
0.3	GS2SLSa-SAR(4)	-0.002(0.015)	-	0.003(0.006)	-	0.002(0.003)	-
0.3	GS2SLSb-SAR(1)	0.003(0.011)	-	0.008(0.005)	-	0.011(0.002)	-
0.3	GS2SLSb-SAR(4)*	-0.009(0.014)	-	0.001(0.005)	-	0.001(0.003)	-
0.3	simpleOLS	0.231(0.061)	-	0.302(0.096)	-	0.261(0.07)	-
0.3	simpleIV	0.016(0.017)	-	0.038(0.012)	-	0.034(0.005)	-
0.3	Mi-2SLl-a1	0.002(0.015)	6[6,0]	0.014(0.007)	31[31,0]	0.014(0.003)	38[38,0]
0.3	Mi-2SLl-a2	0.013(0.022)	34[33,3]	0.067(0.022)	183[180,20]	0.056(0.012)	371[366,40]
0.6	GS2SLSa-SAR(1)	0.004(0.012)	-	0.012(0.005)	-	0.012(0.003)	-
0.6	GS2SLSa-SAR(4)	-0.001(0.015)	-	0.003(0.006)	-	0.002(0.003)	-
0.6	GS2SLSb-SAR(1)	0.006(0.011)	-	0.01(0.005)	-	0.011(0.002)	-
0.6	GS2SLSb-SAR(4)*	-0.007(0.014)	-	0.001(0.005)	-	0.001(0.003)	-
0.6	simpleOLS	0.382(0.153)	-	0.452(0.209)	-	0.412(0.171)	-
0.6	simpleIV	0.013(0.018)	-	0.037(0.012)	-	0.034(0.005)	-
0.6	Mi-2SLl-a1	-0.003(0.017)	7[7,0]	0.012(0.007)	36[36,0]	0.013(0.003)	45[45,0]
0.6	Mi-2SLl-a2	0.01(0.021)	34[33,3]	0.066(0.02)	181[180,20]	0.058(0.012)	368[365,40]
0.9	GS2SLSa-SAR(1)	0.006(0.012)	-	0.012(0.005)	-	0.012(0.002)	-
0.9	GS2SLSa-SAR(4)	0.001(0.016)	-	0.003(0.006)	-	0.002(0.003)	-
0.9	GS2SLSb-SAR(1)	0.01(0.011)	-	0.011(0.005)	-	0.012(0.002)	-
0.9	GS2SLSb-SAR(4)*	-0.003(0.014)	-	0.002(0.005)	-	0.001(0.003)	-
0.9	simpleOLS	0.533(0.29)	-	0.602(0.367)	-	0.562(0.317)	-
0.9	simpleIV	0.01(0.018)	-	0.036(0.012)	-	0.033(0.005)	-
0.9	Mi-2SLl-a1	-0.006(0.018)	7[7,0]	0.011(0.007)	40[40,0]	0.011(0.003)	52[52,0]
0.9	Mi-2SLl-a2	0.007(0.021)	33[33,3]	0.06(0.017)	181[180,19]	0.058(0.011)	367[366,40]

Note: Bias (MSE) and GS2SLSb-SAR(4)* is the correctly specified and estimated model.

Generally, we observe that the bias and MSEs of Mi-2SL are smaller when a stronger penalty is applied, $a = 1$. When $a = 1$ Mi-2SL generally beats the misspecified SAR(1)s, simpleOLS, and simpleIV in terms of bias and simpleOLS and simpleIV in terms of MSE. The bias of $a = 1$ Mi-2SL also fall as the level of error correlation increases, where as the MSE tends to stay the same or increases, for a given sample size and μ. In contrast, when $a = 2$ Mi-2SL can perform worse in terms of both bias and MSE than simpleIV, this is likely due to over-selection in this case. The bias of Mi-2SL tends to falls as the level of error correlation increases, for a given μ and sample size. The biases and MSEs are smaller for denser SWMs, $\mu = 4$ compared to $\mu = 8$, with Mi-2SL having the most substantial improvement, when $a = 2$.

A recent study by Lewbel et al. (2021) showed that GS2SLS is robust to misspecification of links in a binary SWM and will still produce consistent and asymptotically normal estimates when the rate of misspecification grows slower than n^{c-1} $\forall c \leq 0.5$. We now run a Monte Carlo experiment to see how Mi-2SL performs in this situation for the case when $\sigma^2_{v,u}$=0.6. We generate misclassified links using the same set up as Lewbel et al. (2021), let $h_{ij} = w_{ij}t_{ij,1} + (1 - w_{ij})t_{ij,2}$ for all $i \neq j;\ i > j;\ h_{ij} = h_{ji}$; where $t_{ij,1}$ and $t_{ij,2}$ are Bernoulli random variables with success probabilities $1 - \tau_{1,i}$ and $\tau_{2,i}$ respectively. Therefore, $\tau_{i,1}$ is the misclassification probability that $h_{ij} = 0$ when the true $w_{ij} = 1$, and $\tau_{i,2}$ is the misclassification probability that $h_{ij} = 1$ when the true $w_{ij} = 0$. We set $\tau_{i,1} = r_i n^{c-1}$ and $\tau_{i,2} = 100 r_i n^{c-2}$; where $r_i = (\sum_{j=1}^{n} w_{ij}/\mu + |u_i|)/3$. For each i, the probability of misclassification increases in the number of individual i's links $\sum_{j=1}^{n} w_{ij}$, and in the magnitude of i's

Table 3.3: Bias and MSE of β_2 and number of selected eigenvectors, $\mu = 4$ with misspecified links

c	Estimator	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]
		n=100		n=250		n=500	
-	simpleOLS	0.58(0.367)	-	0.432(0.19)	-	0.445(0.2)	-
-	simpleIV	0.071(0.052)	-	0.043(0.009)	-	0.044(0.006)	-
0.1	GS2SLSa-SAR(1)	0.033(0.024)	-	0.018(0.005)	-	0.017(0.003)	-
0.1	GS2SLSa-SAR(4)	0.017(0.028)	-	0.006(0.006)	-	0.003(0.003)	-
0.1	GS2SLSb-SAR(1)	0.027(0.02)	-	0.017(0.005)	-	0.016(0.003)	-
0.1	GS2SLSb-SAR(4)*	0.005(0.022)	-	0.003(0.006)	-	0.001(0.003)	-
0.1	Mi-2SLl-a1	0.014(0.026)	32[32,0]	0.015(0.007)	29[29,0]	0.01(0.004)	86[86,0]
0.1	Mi-2SLl-a2	0.096(0.069)	79[78,8]	0.064(0.018)	164[160,21]	0.108(0.025)	409[404,73]
0.3	GS2SLSa-SAR(1)	0.044(0.03)	-	0.022(0.006)	-	0.02(0.003)	-
0.3	GS2SLSa-SAR(4)	0.032(0.037)	-	0.012(0.007)	-	0.007(0.005)	-
0.3	GS2SLSb-SAR(1)	0.043(0.024)	-	0.022(0.006)	-	0.019(0.003)	-
0.3	GS2SLSb-SAR(4)*	0.013(0.026)	-	0.009(0.006)	-	0.005(0.004)	-
0.3	Mi-2SLl-a1	0.027(0.033)	29[29,0]	0.019(0.008)	25[25,0]	0.015(0.004)	77[77,0]
0.3	Mi-2SLl-a2	0.109(0.082)	74[73,7]	0.053(0.016)	153[151,16]	0.096(0.021)	400[397,59]
0.5	GS2SLSa-SAR(1)	0.058(0.039)	-	0.032(0.007)	-	0.027(0.004)	-
0.5	GS2SLSa-SAR(4)	0.041(0.054)	-	0.026(0.008)	-	0.018(0.004)	-
0.5	GS2SLSb-SAR(1)	0.062(0.034)	-	0.032(0.007)	-	0.027(0.004)	-
0.5	GS2SLSb-SAR(4)*	0.025(0.036)	-	0.022(0.008)	-	0.016(0.004)	-
0.5	Mi-2SLl-a1	0.032(0.041)	22[22,0]	0.029(0.008)	15[15,0]	0.023(0.004)	51[51,0]
0.5	Mi-2SLl-a2	0.073(0.07)	61[60,4]	0.045(0.014)	121[119,8]	0.06(0.013)	371[369,30]
0.7	GS2SLSa-SAR(1)	0.069(0.042)	-	0.042(0.008)	-	0.037(0.005)	-
0.7	GS2SLSa-SAR(4)	0.069(0.069)	-	0.039(0.01)	-	0.034(0.006)	-
0.7	GS2SLSb-SAR(1)	0.075(0.041)	-	0.043(0.008)	-	0.038(0.005)	-
0.7	GS2SLSb-SAR(4)*	0.048(0.054)	-	0.035(0.009)	-	0.032(0.005)	-
0.7	Mi-2SLl-a1	0.052(0.051)	10[10,0]	0.04(0.009)	5[5,0]	0.038(0.005)	14[14,0]
0.7	Mi-2SLl-a2	0.064(0.062)	34[33,1]	0.045(0.012)	57[56,3]	0.042(0.008)	246[246,6]

Note: Bias (MSE) GS2SLSb-SAR(4)* is the correctly specified and estimated model and $\sigma^2_{v,u} = 0.6$. $c = 0$ is the same was when $\sigma^2_{v,u} = 0.6$ in Table 3.1.

unobserved error ($|u_i|$). This construction makes the measurement errors endogenous, correlated with the errors. The eigenvectors are from $\boldsymbol{H}$.

Table 3.4: Bias and MSE of β_2 and number of selected eigenvectors, $\mu = 8$ with misspecified links

c	Estimator	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]	β_2	No Vecs [2nd,1st]
		n=100		n=250		n=500	
-	simpleOLS	0.386(0.157)	-	0.451(0.209)	-	0.409(0.169)	-
-	simpleIV	0.021(0.019)	-	0.034(0.012)	-	0.029(0.005)	-
0.1	GS2SLSa-SAR(1)	0.013(0.014)	-	0.011(0.005)	-	0.011(0.003)	-
0.1	GS2SLSa-SAR(4)	0.01(0.018)	-	0.004(0.006)	-	0.003(0.003)	-
0.1	GS2SLSb-SAR(1)	0.017(0.013)	-	0.008(0.005)	-	0.01(0.003)	-
0.1	GS2SLSb-SAR(4)*	0.003(0.016)	-	0.001(0.006)	-	0.001(0.003)	-
0.1	Mi-2SLl-a1	0.008(0.015)	6[6,0]	0.011(0.007)	34[34,0]	0.01(0.003)	44[44,0]
0.1	Mi-2SLl-a2	0.017(0.021)	31[30,3]	0.058(0.017)	179[177,18]	0.054(0.012)	366[363,40]
0.3	GS2SLSa-SAR(1)	0.017(0.015)	-	0.013(0.006)	-	0.012(0.003)	-
0.3	GS2SLSa-SAR(4)	0.017(0.019)	-	0.008(0.006)	-	0.005(0.003)	-
0.3	GS2SLSb-SAR(1)	0.021(0.014)	-	0.012(0.006)	-	0.011(0.003)	-
0.3	GS2SLSb-SAR(4)*	0.009(0.016)	-	0.005(0.006)	-	0.003(0.003)	-
0.3	Mi-2SLl-a1	0.011(0.016)	5[5,0]	0.011(0.007)	32[32,0]	0.012(0.003)	41[41,0]
0.3	Mi-2SLl-a2	0.019(0.021)	27[26,2]	0.055(0.017)	173[172,16]	0.054(0.011)	360[357,35]
0.5	GS2SLSa-SAR(1)	0.023(0.016)	-	0.02(0.007)	-	0.015(0.003)	-
0.5	GS2SLSa-SAR(4)	0.02(0.022)	-	0.017(0.009)	-	0.011(0.004)	-
0.5	GS2SLSb-SAR(1)	0.027(0.015)	-	0.019(0.007)	-	0.015(0.003)	-
0.5	GS2SLSb-SAR(4)*	0.012(0.019)	-	0.012(0.008)	-	0.009(0.003)	-
0.5	Mi-2SLl-a1	0.013(0.018)	4[4,0]	0.014(0.009)	25[25,0]	0.016(0.004)	30[30,0]
0.5	Mi-2SLl-a2	0.018(0.021)	20[19,2]	0.046(0.019)	157[155,11]	0.041(0.009)	334[333,23]
0.7	GS2SLSa-SAR(1)	0.025(0.017)	-	0.03(0.009)	-	0.022(0.004)	-
0.7	GS2SLSa-SAR(4)	0.028(0.024)	-	0.029(0.011)	-	0.021(0.004)	-
0.7	GS2SLSb-SAR(1)	0.031(0.017)	-	0.031(0.009)	-	0.023(0.004)	-
0.7	GS2SLSb-SAR(4)*	0.02(0.021)	-	0.022(0.011)	-	0.019(0.004)	-
0.7	Mi-2SLl-a1	0.018(0.019)	1[1,0]	0.026(0.011)	12[12,0]	0.025(0.004)	11[11,0]
0.7	Mi-2SLl-a2	0.018(0.022)	9[9,1]	0.035(0.016)	100[99,5]	0.029(0.007)	236[235,8]

Note: Bias (MSE), GS2SLSb-SAR(4)* is the correctly specified and estimated model and $\sigma^2_{v,u} = 0.6$. $c = 0$ is the same was when $\sigma^2_{v,u} = 0.6$ in Table 3.2.

Table 3.3 and Table 3.4 shows the bias and MSE of β_2 and the number of selected eigenvectors at the first-stage and second-stages, and the number of eigenvectors used in the IV regression (i.e., the union of the first and second stage selected eigenvectors) when $\mu = 4$ and $\mu = 8$ for different rates of misspecified links (c) and sample sizes.[9]

These results show that Mi-2SL has good performance in terms of both bias and

[9]See Section 3.8.2 for the full set of simulation results.

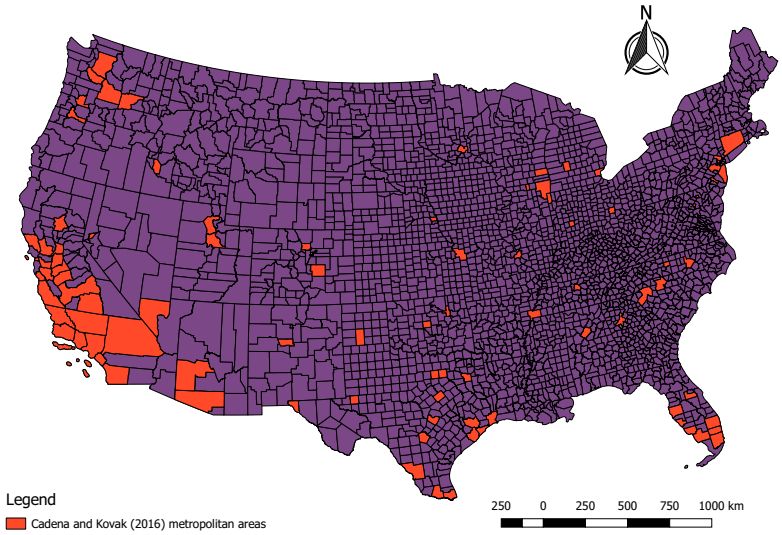
MSE when the SWM contains misspecified links, with its performance improving as c increases relative to the other estimated models. Mi-2SL has the smallest bias of all estimators considered when $c = 0.7$, $a = 1$ $n = 100$ and $\mu = 8$, even beating the correctly specified and estimated model, GS2SLSb-SAR(4) and in the other cases performs similarly to the other estimators for $c = 0.7$. Mi-2SL performs better in terms of both bias and MSE when $a = 1$ than $a = 2$ but the difference between the two falls as c increases, likely due to the reduction in the number of selected/included eigenvectors when $a = 2$. The denser the SWM the smaller the bias and MSE, regardless of the sample size or level of misspecification.

Comparing these simulation results to those presented in Section 2.6, we find that adding an endogenous variable that is also spatially correlated changes the behavior of the exponent a. In Section 2.6 $a = 2$ gives the best performance in terms of MSE and bias, whereas when the endogenous variable is added $a = 1$ gives the best performance in terms of MSE and bias. Estimation of the first stage (3.18) is equivalent to the setup A when $\rho_1 = 0.6$ in Section 2.6 (with an additional exogenous variable added), the selection behavior in the first stage in these simulations is similar, when $a = 1$ in most cases no eigenvectors are selected, and when $a = 2$ a larger set is selected. In contrast, the selection behavior in the second stage is systematically selecting more eigenvectors, so in this case, $a = 2$ is likely over-selecting. Thus when the level of spatial correlation in the variable ($\boldsymbol{y}$) being studied is large ($p = 4$) adding an endogenous variable that is also spatially correlated appears to be causing over-selection in the second stage when $a = 2$, a stronger penalty ($a = 1$) is need to improves the performance of the procedure.

3.6 Application on impact of migration on labour markets

This section revisits the empirical application of Cadena and Kovak (2016) empirical work. They find 'that low-skilled Mexican-born immigrants' location choices respond strongly to changes in local labour demand, which helps equalize spatial differences in employment outcomes for low-skilled native workers' using an IV strategy. Cadena and Kovak (2016) starts from the observation that over the Great Recession low-educated Mexican-born male immigrants were more mobile than their native counterparts. Given this observation, they want to test if their location choice is being driven by local labour market conditions and leverage the geographic variation in employment changes during the Great Recession as a natural experiment. They argue that the change in labour market conditions over the Great Recession can be approximately measured by changes in employment as wages are sticky downward. Thus, they look at the effect changes in employment has on population changes for 20 different demographic groups. The demographic groups are split by gender (males and females), education (high school or less and some college or more), and nativity (native-born, foreign-born, Mexican-born, and other foreign-born). The unit of observation is a metropolitan area and they include 95 metropolitan area in their IV analysis. Figure 3.1 shows the 95 metropolitan areas included in the Cadena and Kovak (2016) IV analysis and reveals clear spatial heterogeneity.

Figure 3.1: Metropolitan areas in Cadena and Kovak (2016)



Their system of equations is

$$\Delta pop_i = \beta_0 + \beta_1 \Delta emp_i + \beta_2 mex_i + \beta_3 policy_i + \beta_4 287g_i + u_i, \tag{3.19}$$

$$\Delta emp_i = \psi_0 + \psi_1 \Delta bartik_i + \psi_2 mex_i + \psi_3 policy_i + \psi_4 287g_i + v_i, \tag{3.20}$$

where the index i is for the metropolitan area, Δpop_i is the proportional change in working-age population from 2006–2010, Δemp_i is the proportional change in employment from 2006–2010, mex_i is the share of Mexicans-born population in 2000, $policy_i$ and $287g_i$ are both immigration policy controls, and $\Delta bartik_i$ is the 'Bartik instrument' Bartik (1991), which predicts changes in local labour demand by assuming that in each industry national employment changes are proportionately allocated across cities, based on each cities initial industry composition of employment. For

the reader's convince, Table 3.5 replicates their main IV results, Table 4 in Cadena and Kovak (2016). We have also added the first stage (full) F-statistic, so this can be compared to the partial F-statistic and give further insight into the actual impact of the Bartik instrument in their estimates. Table 3.5 shows the full F-statistic is always smaller than the partial F-statistic, implying that in their specification, the Bartik, which is supposed to give the identification, is not helping the first stage.

Table 3.5: Replication of main IV results (Table 4) in Cadena and Kovak (2016)

	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
Panel A: Men, high school or less					
Change in log of group-specific	0.223	0.007	0.402	0.992	-0.675
employment	(0.166)	(0.09)	(0.409)	(0.468)	(0.278)
First stage F-statistic	11.42	10.94	12.76	8.28	17.11
Partial F-statistic (Bartik)	35.74	36.14	25.31	11.94	45.61
Panel B: Men, some college or more					
Change in log of group-specific	0.27	0.411	-0.237	-0.475	-0.161
employment	(0.157)	(0.192)	(0.264)	(0.387)	(0.329)
First stage F-statistic	7.19	6.74	13.07	14.59	13.02
Partial F-statistic (Bartik)	23.89	21.9	37.76	31.79	36.89
Panel C: Women, high school or less					
Change in log of group-specific	0.145	-0.405	0.272	1.811	-0.979
employment	(0.168)	(0.287)	(0.504)	(0.665)	(0.556)
First stage F-statistic	10.43	8.76	15.21	6.04	22.42
Partial F-statistic (Bartik)	28.59	26.09	26.76	13.74	39.17
Panel D: Women, some college or more					
Change in log of group-specific	-0.066	-0.054	-0.754	0.438	-1.092
employment	(0.378)	(0.42)	(0.716)	(0.919)	(0.738)
First stage F-statistic	1.53	1.56	3.39	7.75	3.49
Partial F-statistic (Bartik)	5.85	5.58	12.97	27.33	13.12

Note: Robust standard errors appear in parentheses. See Cadena and Kovak (2016) Table 4 for further details.

For the SWM, we will use a binary distance-based SWM where $w_{ij} = 1$ if the distance between the metropolitan areas is less than A kilometres and zero otherwise. We consider cutoff distances (A) of 500km, 700km, and 900km. Note 500km is the

smallest distance that ensures every metropolitan area has at least one neighbour.

Table 3.6: standardised Moran's i of first and second stage (Cadena and Kovak, 2016)

SWM cutoff	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
	Panel A: Men, high school or less				
900	12.17***, 4.9***	13.34***, 3.6***	10.38***, 3.91***	11.18***, 3.59***	11.38***, 0.95
700	10.59***, 4.33***	11.96***, 2.86***	8.83***, 3.88***	9.38***, 2.6***	10.25***, 2.28**
500	10.1***, 3.25***	11.36***, 1.04	8.83***, 2.86***	9.7***, 2.53**	9.79***, 2.7***
	Panel B: Men, some college or more				
900	12.85***, 1.13	13.31***, 0.36	11.24***, 3.06***	12.91***, 1.62	12.38***, 2.34**
700	11.31***, 1.39	11.8***, 0.22	10***, 2.01**	11.84***, 0.35	11.17***, 1.3
500	10.43***, 0.37	10.67***, 0.01	9.93***, 2.91***	12.05***, 0.25	10.35***, 1.89*
	Panel C: Women, high school or less				
900	10.08***, 3.4***	13.42***, 3.02***	4.67***, 5.05***	5.81***, 4.34***	9.78***, 0.61
700	9.34***, 2.85***	12.84***, 2.3**	4.13***, 5.01***	3.97***, 4.04***	8.98***, 1.36
500	8.35***, 2.47**	11.41***, 1.71*	4.49***, 3.03***	4.21***, 3.39***	8.79***, 1.53
	Panel D: Women, some college or more				
900	12.79***, 3.95***	12.44***, 2.53**	13***, 6.14***	6.44***, -0.63	14.05***, 4.23***
700	11.27***, 3.53***	10.98***, 2.03**	11.71***, 2.96***	6.83***, -0.19	12.53***, 3.4***
500	9.14***, 2.77***	8.87***, 1.28	9.89***, 2.92***	6.53***, 0.1	10.5***, 3.15***

Note: first stage, second stage. *p<0.1; **p<0.05; ***p<0.01.

Table 3.6 shows the standardised Moran's i for the first and second stages of Cadena and Kovak (2016) IV regressions for the three SWMs considered. This table shows that the standardised Moran's i of the first stage is always significant at the one percent level, and the second stage in most samples is also significant at the ten percent level. In almost all cases, the first-stage has a substantially higher level of spatial correlation than the second-stage. For low-educated Mexican-born male migrants, the standardised Moran's i is significant at the five percent level in both stages for all three SWMs, with a test statistic three times larger in the first than second stage.

Table 3.7 shows the Mi-2SLl (the first stage fitted values from the Lasso estimates,

Table 3.7: Mi-2SLl results of Cadena and Kovak (2016) with SWM cutoff 500km

	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
Panel A: Men, high school or less					
Change in log of group-specific	0.275	0.164	0.328	1.385	-0.6
employment	(0.18)	(0.1)	(0.436)	(0.308)	(0.321)
First stage F-statistic	62.47	48.25	63.73	80.35	435.1
Partial F-statistic (Bartik)	75.9	65.84	39.54	56.82	144.96
Number of vecs [1st,2nd]	8[8,0]	15[15,0]	6[6,0]	14[14,0]	4[4,0]
Panel B: Men, some college or more					
Change in log of group-specific	0.211	0.323	-0.223	0.19	-0.125
employment	(0.175)	(0.243)	(0.18)	(0.679)	(0.251)
First stage F-statistic	21.77	27.04	35.61	200.98	42.07
Partial F-statistic (Bartik)	33.81	46.72	55.71	190.03	63.06
Number of vecs [1st,2nd]	5[5,0]	7[7,0]	4[4,0]	19[19,0]	4[4,0]
Panel C: Women, high school or less					
Change in log of group-specific	0.148	-0.442	0.272	1.811	-0.737
employment	(0.158)	(0.304)	(0.504)	(0.665)	(0.458)
First stage F-statistic	41.51	30.29	15.21	6.04	49.4
Partial F-statistic (Bartik)	64.18	31.98	26.76	13.74	95.59
Number of vecs [1st,2nd]	1[1,0]	6[6,0]	0[0,0]	0[0,0]	1[1,0]
Panel D: Women, some college or more					
Change in log of group-specific	-0.02	-0.037	-0.581	0.061	-0.987
employment	(0.301)	(0.381)	(0.517)	(1.118)	(0.508)
First stage F-statistic	12.32	10.64	22.37	24.51	21.14
Partial F-statistic (Bartik)	9.62	8.17	26.81	36.29	24.32
Number of vecs [1st,2nd]	1[1,0]	1[1,0]	1[1,0]	1[1,0]	2[2,0]

Note: First stage fitted values from Lasso estimates (step 3 in Algorithm 3) with $a = 2$ and the cutoff for the SWM 500km. Robust standard errors appear in parentheses.

step 3 in Algorithm 3) with $a = 2$ and an SWM with a cut-off of 500km.[10] No eigenvectors are selected in any of the second stages due to the low levels of spatial correlation in each of the second stages, so the fitted values from Lasso and post-Lasso yield the same results. The tables of results with SWM with larger cut-off and where the first stage fitted values from post Lasso (Mi-2SLpl) are in Section 3.8.3. The Mi-2SL results do not change the conclusion of Cadena and Kovak (2016) that low-educated Mexican-born migrants respond positively to changes in employment. For low-educated Mexican-born males we find the magnitude of the coefficient increases by approximatly a standard error. A key general impact of the included eigenvectors is a substantial improvement in the first stage F-statistic and partial F-statistic. For example, for low-educated Mexican-born males, the first-stage F-statistic and partial F-statistic increased from 8.28 and 11.94 to 80.35 and 56.82. This improvement in the first-stage estimates leads to an increase in the precision of the second-stage estimates, which can be seen by the reduction in the estimated standard errors on employment change from 0.468 to 0.308. The smaller partial F-statistic than full F-statistic also implies that the Bartik is now having a stronger positive effect, at least in the case of low-educated Mexican-born migrants.

3.7 Conclusion

In conclusion, we have proposed a new two-stage lasso-based procedure to estimate classical regression parameters of endogenous variables in the presence of spatial

[10]When $a = 1$ no eigenvectors are selected, thus the results for this case are equivalent to Table 3.5.

correlation of an unknown functional form called Morans' i 2 stage Lasso (Mi-2SL). Under the assumption that the relevant set of eigenvectors is known, an appropriate mixing condition holds, some further restriction on the spatial structure, and some additional regularity conditions, we show that the ESF 2SLS parameter estimates are consistent and asymptotic normality.

Our simulations result imply the Mi-2SL estimators have a reasonable performance against a range of error corrections and have good performance when the SWM also includes misclassified links, compared to existing selection procedures, when the second stage includes a higher overall level of spatial correlation than the first. Our empirical application, where we replicate the IV results of Cadena and Kovak (2016) where the overall spatial correlation is higher in the first-stage than the second, demonstrates the benefits of using Mi-2SL by improving the first-stage partial F-statistic and full F-statistic and the reduction in second stage standard errors.

Avenues of further research include extending the consistency and asymptotic normality to allow for mistakes in selection (accounting for the fact that Ω need to be estimated), greater theoretical understanding of why Mi-2SL/ESF performs well in the presence of misclassified links and if ESF also performs well if there is misspecification in non-binary SWMs, as well as further theoretical understanding of how the choose the Mi-2SL exponents (a).

3.8 Appendix

3.8.1 Additional Lemmas and proofs main results

The following Lemma shows that as linear transformation (denoted h_c) is a Lipschitz functions, then taking $f_c = f(h_c(\cdot))$ is also a Lipschitz functions, as long as the linear transformation is bound.

Lemma 4. *If $f(r_i)$ is a bounded real Lipschitz function with $f \in \mathcal{L}_a$ and $\{r_i \in \mathbb{R}\}$ are triangular arrays and $h_c(r_i)$ is a bounded real linear Lipschitz function which linearly transforms r_i by the triangular arrays $\{c_i\}$, $c_i \in \mathbb{R}$ with $\max_i |c_i| < \infty$. Then $f_c(r_i) = f\big(h_c(r_i)\big)$ is also a bounded real Lipschitz with $f_c \in f$.*

Proof of Lemma 4. Let the linear function h_c be multiplicative i.e. $h_c(r_i) := c_i r_i$, $\mathbb{R} \to \mathbb{R}$ is Lipschitz iff there exists a constant K_h such that $|h_c(x_1) - h_c(x_2)| = |c_i x_1 - c_i x_2| \leq K_h|(x_1 - x_2)|$ where x_1 and x_2 are points in $\mathbb{R}$. Note by the Cauchy-Schwartz inequality $(c_i(x_1 - x_2))^2 \leq c_i^2(x_1 - x_2)^2$. Thus, $h_c(\cdot)$ is Lipschitz as $K_h = |c_i| < \infty$.

Now as f is a bounded real Lipschitz function $|f(c_i x_1) - f(c_i x_2)| \leq K_f|c_i' x_1 - c_i' x_2|$,

$$|f_c(x_1) - f_c(x_1)| = |f(c_i x_1) - f(c_i x_2)| \leq K_f|c_i(x_1 - x_2)| \tag{3.21}$$

$$\leq K_c|(x_1 - x_2)| \tag{3.22}$$

where (3.22) uses the Cauchy-Schwartz inequality and $K_c := K_f K_h$.

Similarly let the linear function h_c be additive i.e. $h_c(r_i) := c_i + r_i$, $\mathbb{R} \to \mathbb{R}$ is Lipschitz iff there exists a constant K_h such that $|h_c(x_1) - h_c(x_2)| = |(c_i + x_1) - (c_i +$

$x_2)| \leq |(x_1 - x_2)|$, note $K_h = 1$. Thus,

$$|f_c(x_1) - f_c(x_1)| = |f(c_i + x_1) - f(c_i + x_2)| \leq K_f|(c_i + x_1) - (c_i + x_2)| = K_f|x_1 - x_2|$$
$$= K_c|(x_1 - x_2)|$$

where $K_f = K_c K_h = K_c$. □

Proof of Lemma 3. As e_{jk} and λ_k satisfy the linear transformation requirements in Lemma 4, thus, these transformations are $f_c \in f$ and $g_c \in g$, so we have

$$\left| \operatorname{Cov}\left(f\left(\sum_{j\in N} w_a e_{jk}/\lambda_k\right), g\left(\sum_{j\in N} w_b e_{jk}/\lambda_k\right) | \, \mathcal{C}\right)\right| = |\operatorname{Cov}(f_c(w_a), g_c(w_b)|\, \mathcal{C})|$$
$$\leq \psi_{a,b}(f_c, g_c)\mu_r \;\; a.s.$$

The inequality holds due to (3.11). □

Let $\boldsymbol{A}$ be some $n \times s$ matrix with typical elements a_{ij} and column $\boldsymbol{a}_j$, further the columns are orthonormal and $\boldsymbol{B}$ be some $n \times d$ matrix with typical elements b_{ij} and column $\boldsymbol{b}_j$, thus,

$$\boldsymbol{A}'\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n1} & \cdots & a_{ns} \end{bmatrix}' \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n1} & \cdots & a_{ns} \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{k=1}^{n} a_{k1}a_{k1} & \sum_{k=1}^{n} a_{k1}a_{k2} & \cdots & \sum_{k=1}^{n} a_{k1}a_{ks} \\ \sum_{k=1}^{n} a_{k2}a_{k1} & \sum_{k=1}^{n} a_{k2}a_{k2} & \cdots & \sum_{k=1}^{n} a_{k2}a_{ks} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} a_{ks}a_{k1} & \sum_{k=1}^{n} a_{ks}a_{k2} & \cdots & \sum_{k=1}^{n} a_{ks}a_{ks} \end{bmatrix}$$

and

$$\boldsymbol{A}'\boldsymbol{B} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n1} & \cdots & a_{ns} \end{bmatrix}' \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1d} \\ b_{21} & b_{22} & \cdots & b_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nd} \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{k=1}^{n} a_{k1}b_{k1} & \sum_{k=1}^{n} a_{k1}b_{k2} & \cdots & \sum_{k=1}^{n} a_{k1}b_{kd} \\ \sum_{k=1}^{n} a_{k2}b_{k1} & \sum_{k=1}^{n} a_{k2}b_{k2} & \cdots & \sum_{k=1}^{n} a_{k2}b_{kd} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} a_{ks}b_{k1} & \sum_{k=1}^{n} a_{ks}b_{k2} & \cdots & \sum_{k=1}^{n} a_{ks}b_{kd} \end{bmatrix}$$

Lemma 5. *Suppose the triangular arrays* $\{a_{ki}\}$, $a_{ki} \in \mathbb{R}$ *which satisfies Assumption 7.1 with dependence coefficient* $\{\mu_r\}$ *and the columns of* $\boldsymbol{A}$ *are orthonormal (*$||\boldsymbol{a}_i||_2^2 = 1 \ \forall i$*). Then the array* $\{\sum_{k=1}^{n} a_{ki}a_{kj}\}$, $i = 1, \ldots, s$, $j = 1, \ldots, s$ *is conditionally*

ψ/dependent given $\{\mathcal{C}\}$ with the dependence coefficients $\{\mu_r\}$.

$$\Big|\operatorname{Cov}\big(f\Big(\sum_{k\in N} a_a a_{kj}\Big), g\Big(\sum_{k\in N} a_b a_{kj}\Big)|\,\mathcal{C}\,\big)\Big| \le \psi_{a,b}(f_c, g_c)\mu_r \quad \text{a. s.}$$

where $a_a = a_{ka}$ and $a_b = a_{kb}$

Proof of Lemma 5. As $||\boldsymbol{a}_j||_2^2 = 1$ satisfy the linear transformation requirements in Lemma 4, thus, these transformations are $f_c \in f$ and $g_c \in g$, so we have

$$\Big|\operatorname{Cov}\big(f\Big(\sum_{k\in N} a_a a_{kj}\Big), g\Big(\sum_{k\in N} a_b a_{kj}\Big)|\,\mathcal{C}\,\big)\Big| = \big|\operatorname{Cov}(f_c(a_a), g_c(a_b)|\,\mathcal{C})\big| \le \psi_{a,b}(f_c, g_c)\mu_r \;\; a.s.$$

The inequality holds due to (3.11). □

Lemma 6. *Suppose the triangular array $\{a_{ki}\}$, $a_{ki} \in \mathbb{R}$ which satisfies Assumption 7.1 with dependence coefficient $\{\mu_r\}$. For each $n \ge 1$ let $\{b_{kj}\}$, $b_{kj} \in \mathbb{R}$ be a sequence of $\mathcal{C}$ measurable random variables with $\sup_{n\ge1}\max_{i\in N}\left(\mathbb{E}[|b_{kj}|^2|\,\mathcal{C}]\right)^{1/2} < \infty,\ \forall j$. Then the array $\{\sum_{k=1}^n a_{ki}b_{kj}\},\ i = 1,\dots,s,\ j = 1,\dots,d$ is conditionally ψ-dependent given $\{\mathcal{C}\}$ with the dependence coefficients $\{\mu_r\}$.*

$$\Big|\operatorname{Cov}\big(f\Big(\sum_{k\in N} a_a b_{kj}\Big), g\Big(\sum_{k\in N} a_b b_{kj}\Big)|\,\mathcal{C}\,\big)\Big| \le \psi_{a,b}(f_c, g_c)\mu_r \quad \text{a. s.}$$

where $a_a = a_{ka}$ and $a_b = a_{kb}$

Proof of Lemma 6. As $\sup_{n\ge1}\max_{i\in N}\left(\mathbb{E}[|b_{kj}|^2|\,\mathcal{C}]\right)^{1/2} < \infty,\ \forall j$ satisfy the linear transformation requirements in Lemma 4, thus, these transformations are $f_c \in f$ and $g_c \in g$, so we have

$$\Big|\operatorname{Cov}\Big(f\Big(\sum_{k\in N} a_a b_{kj}\Big), g\Big(\sum_{k\in N} a_b b_{kj}\Big)|\,\mathcal{C}\,\Big)\Big| = \big|\operatorname{Cov}(f_c(a_a), g_c(a_b)|\,\mathcal{C})\big| \le \psi_{a,b}(f_c, g_c)\mu_r a.s.$$

The inequality holds due to (3.11). □

Note a special case of Lemma 6 is then $d = 1$, i.e. $\boldsymbol{B}$ is a vector.

Proof of Theorem 3. Starting from the 2SLS solutions of (3.9)

$$\hat{\boldsymbol{\Upsilon}}_\Omega = \big((\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}'_\Omega \boldsymbol{G}_\Omega/n)\big)^{-1}(\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}'_\Omega \boldsymbol{y}/n)$$

Substituting in (3.9) yields

$$\begin{aligned}
\hat{\boldsymbol{\Upsilon}}_\Omega &= \big((\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}'_\Omega \boldsymbol{G}_\Omega/n)\big)^{-1} && (3.23)\\
&\quad \times (\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}'_\Omega(\boldsymbol{G}_\Omega \boldsymbol{\Upsilon}_\Omega + \boldsymbol{\varepsilon})/n)\\
\hat{\boldsymbol{\Upsilon}}_\Omega - \boldsymbol{\Upsilon}_\Omega &= \big((\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}'_\Omega \boldsymbol{G}_\Omega/n)\big)^{-1} && (3.24)\\
&\quad \times (\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}'_\Omega \boldsymbol{\varepsilon}/n) && (3.25)
\end{aligned}$$

Expressing $\boldsymbol{Z}'_\Omega \boldsymbol{Z}_\Omega/n$, $\boldsymbol{G}'_\Omega \boldsymbol{Z}_\Omega/n$, $\boldsymbol{Z}'_\Omega \boldsymbol{G}_\Omega/n$ and $\boldsymbol{Z}'_\Omega \boldsymbol{\varepsilon}/n$ in block-wise form,

$$
\begin{aligned}
\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega/n &= \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1/n & \boldsymbol{X}_1'\boldsymbol{Z}_2/n & \boldsymbol{X}_1'\boldsymbol{E}_\Omega/n \\ \boldsymbol{Z}_2'\boldsymbol{X}_1/n & \boldsymbol{Z}_2'\boldsymbol{Z}_2/n & \boldsymbol{Z}_2'\boldsymbol{E}_\Omega/n \\ \boldsymbol{E}_\Omega'\boldsymbol{X}_1/n & \boldsymbol{E}_\Omega'\boldsymbol{Z}_2/n & \boldsymbol{E}_\Omega'\boldsymbol{E}_\Omega/n \end{bmatrix} \\
\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega/n &= \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1/n & \boldsymbol{X}_1'\boldsymbol{Z}_2/n & \boldsymbol{X}_1'\boldsymbol{E}_\Omega/n \\ \boldsymbol{x}_2'\boldsymbol{X}_1/n & \boldsymbol{x}_2'\boldsymbol{Z}_2/n & \boldsymbol{x}_2'\boldsymbol{E}_\Omega/n \\ \boldsymbol{E}_\Omega'\boldsymbol{X}_1/n & \boldsymbol{E}_\Omega'\boldsymbol{Z}_2/n & \boldsymbol{E}_\Omega'\boldsymbol{E}_\Omega/n \end{bmatrix} \\
\boldsymbol{Z}_\Omega'\boldsymbol{G}_\Omega/n &= \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1/n & \boldsymbol{X}_1'\boldsymbol{x}_2/n & \boldsymbol{X}_1'\boldsymbol{E}_\Omega/n \\ \boldsymbol{Z}_2'\boldsymbol{X}_1/n & \boldsymbol{Z}_2'\boldsymbol{x}_2/n & \boldsymbol{Z}_2'\boldsymbol{E}_\Omega/n \\ \boldsymbol{E}_\Omega'\boldsymbol{X}_1/n & \boldsymbol{E}_\Omega'\boldsymbol{x}_2/n & \boldsymbol{E}_\Omega'\boldsymbol{E}_\Omega/n \end{bmatrix} \\
\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}/n &= \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{\varepsilon}/n \\ \boldsymbol{Z}_2'\boldsymbol{\varepsilon}/n \\ \boldsymbol{E}_\Omega'\boldsymbol{\varepsilon}/n \end{bmatrix}
\end{aligned}
$$

Given the elements of $\boldsymbol{X}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{Z}_2$ are triangular arrays of real number that are bound in absolute value (Assumptions 5.2). Additionally by Assumptions 5.3 and the LLN of triangular arrays we have $\boldsymbol{X}_1'\boldsymbol{X}_1/n \to_p \mathbb{E}[\boldsymbol{X}_1'\boldsymbol{X}_1]$, $\boldsymbol{X}_1'\boldsymbol{Z}_2/n \to_p \mathbb{E}[\boldsymbol{X}_1'\boldsymbol{Z}_2]$, $\boldsymbol{Z}_2'\boldsymbol{X}_1/n \to_p \mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{X}_1]$, $\boldsymbol{Z}_2'\boldsymbol{Z}_2/n \to_p \mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{Z}_2]$, $\boldsymbol{x}_2'\boldsymbol{X}_1/n \to_p \mathbb{E}[\boldsymbol{x}_2'\boldsymbol{X}_1]$, $\boldsymbol{X}_1'\boldsymbol{x}_2/n \to_p \mathbb{E}[\boldsymbol{X}_1'\boldsymbol{x}_2]$, $\boldsymbol{Z}_2'\boldsymbol{x}_2/n \to_p \mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{x}_2]$, $\boldsymbol{x}_2'\boldsymbol{Z}_2/n \to_p \mathbb{E}[\boldsymbol{x}_2'\boldsymbol{Z}_2]$, $\boldsymbol{X}_1'\boldsymbol{\varepsilon}/n \to_p \mathbb{E}[\boldsymbol{X}_1'\boldsymbol{\varepsilon}] = 0$ and $\boldsymbol{Z}_2'\boldsymbol{\varepsilon}/n \to_p \mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{\varepsilon}] = 0$ as $n \to \infty$

By Lemma 3 - 6 terms involving $\boldsymbol{E}_\Omega$ are weakly ψ-dependence with dependence coefficient $\{\mu_r\}$, thus, under Assumptions 7 and the LLN of Kojevnikov et al. (2021) we have

$\boldsymbol{X}_1'\boldsymbol{E}_\Omega/n \to_{a.s.} \mathbb{E}[\boldsymbol{X}_1'\boldsymbol{E}_\Omega|\,\mathcal{C}]$, $\boldsymbol{Z}_2'\boldsymbol{E}_\Omega/n \to_{a.s.} \mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{E}_\Omega|\,\mathcal{C}]$, $\boldsymbol{E}_\Omega'\boldsymbol{X}_1/n \to_{a.s.} \mathbb{E}[\boldsymbol{E}_\Omega'\boldsymbol{X}_1|\,\mathcal{C}]$, $\boldsymbol{E}_\Omega'\boldsymbol{Z}_2/n \to_{a.s.} \mathbb{E}[\boldsymbol{E}_\Omega'\boldsymbol{Z}_2|\,\mathcal{C}]$, $\boldsymbol{x}_2'\boldsymbol{E}_\Omega/n \to_{a.s.} \mathbb{E}[\boldsymbol{x}_2'\boldsymbol{E}_\Omega|\,\mathcal{C}]$, $\boldsymbol{E}_\Omega'\boldsymbol{x}_2/n \to_{a.s.} \mathbb{E}[\boldsymbol{E}_\Omega'\boldsymbol{x}_2|\,\mathcal{C}]$, $\boldsymbol{E}_\Omega'\boldsymbol{E}_\Omega/n \to_{a.s.} \mathbb{E}[\boldsymbol{E}_\Omega'\boldsymbol{E}_\Omega|\,\mathcal{C}]$ and $\boldsymbol{E}_\Omega'\boldsymbol{\varepsilon}/n \to_{a.s.} \mathbb{E}[\boldsymbol{E}_\Omega'\boldsymbol{\varepsilon}|\,\mathcal{C}] = 0$, as $n \to \infty$

Thus, we have $\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega/n \to_p \mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]$, $\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega/n \to_p \mathbb{E}[\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]$ $\boldsymbol{Z}_\Omega'\boldsymbol{G}_\Omega/n \to_p \mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{G}_\Omega|\,\mathcal{C}]$ and $\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}/n \to_p \mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}|\,\mathcal{C}] = 0$.

Finally applying the Continuous Mapping Theorem we have

$$\begin{aligned}\hat{\boldsymbol{\Upsilon}}_\Omega - \boldsymbol{\Upsilon}_\Omega \to_p &\big(\mathbb{E}[\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]\,\mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]^{-1}\,\mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{G}_\Omega|\,\mathcal{C}]\big)^{-1} \\ &\times \mathbb{E}[\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]\,\mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]^{-1}\,\mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}|\,\mathcal{C}] \\ &= 0.\end{aligned}$$

□

Proof of Theorem 4. Multiplying (3.25) by $\sqrt{n}$ gives

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\Upsilon}}_\Omega - \boldsymbol{\Upsilon}_\Omega) = &\big((\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}_\Omega'\boldsymbol{G}_\Omega/n)\big)^{-1} \\ &\times (\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}/\sqrt{n})\end{aligned}$$

Under Assumptions 5-7, the LLN of triangular arrays and the LLN of Kojevnikov et al. (2021) that matrices involving $\boldsymbol{G}_\Omega$ and $\boldsymbol{Z}_\Omega$ are $O_p(1)$ (proved in Theorem 3). We now need to look at the behavior of

$$\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}/\sqrt{n} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{\varepsilon}/\sqrt{n} \\ \boldsymbol{Z}_2'\boldsymbol{\varepsilon}/\sqrt{n} \\ \boldsymbol{E}_\Omega'\boldsymbol{\varepsilon}/\sqrt{n} \end{bmatrix}$$

Given ε_i is a triangular array of identically distributed random variables that is (jointly) independently distributed for each n with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma_\varepsilon^2 < \infty$ (Assumption 5.2), and $\mathbb{E}[\boldsymbol{X}_1'\boldsymbol{X}_1|\,\mathcal{C}]$ and $\mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{Z}_2|\,\mathcal{C}]$ are finite and non-singular (implied by Assumption 5.2 and 7.2). Then the central limit theorem for triangular arrays implies $\boldsymbol{X}_1'\boldsymbol{\varepsilon}/\sqrt{n} \to_d N(0, \sigma_\varepsilon^2\,\mathbb{E}[\boldsymbol{X}_1'\boldsymbol{X}_1|\,\mathcal{C}])$ and $\boldsymbol{Z}_2'\boldsymbol{\varepsilon}/\sqrt{n} \to_d N(0, \sigma_\varepsilon^2\,\mathbb{E}[\boldsymbol{Z}_2'\boldsymbol{Z}_2|\,\mathcal{C}])$.

Lemma 4, 3 and 6 insure that the element of $\boldsymbol{E}_\Omega'\boldsymbol{\varepsilon}$ are ψ-dependent with dependence coefficients μ_r. Additionally given assumptions 5.3, 7.4 and 8 we can apply the CLT of Kojevnikov et al. (2021), $\boldsymbol{E}_\Omega'\boldsymbol{\varepsilon}/\sqrt{n} \to_d N(0, \sigma_\varepsilon^2\,\mathbb{E}[\boldsymbol{E}_\Omega'\boldsymbol{E}_\Omega|\,\mathcal{C}])$.

Thus, we have

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\Upsilon}}_\Omega - \boldsymbol{\Upsilon}_\Omega) &= \big((\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}_\Omega'\boldsymbol{G}_\Omega/n)\big)^{-1} \\
&\quad \times (\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega/n)(\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega/n)^{-1}(\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}/\sqrt{n}) \\
&\to_d N(0, \big(\mathbb{E}[\boldsymbol{G}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]\,\mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{Z}_\Omega|\,\mathcal{C}]^{-1}\,\mathbb{E}[\boldsymbol{Z}_\Omega'\boldsymbol{G}_\Omega|\,\mathcal{C}]\big)^{-1}\sigma_\varepsilon^2)
\end{aligned}$$

Final we show that $\boldsymbol{Z}_\Omega'\boldsymbol{\varepsilon}$ has a finite second moment. By Minkowski's inequality

we have

$$(\mathbb{E}[|\varepsilon_i|^4|\,\mathcal{C}])^{1/4} = (\mathbb{E}[|y_i - \sum_{j=1}^{(k_1+1+s)} g_{ij,\Omega}\Upsilon_{j,\Omega}|^4|\,\mathcal{C}])^{1/4}$$
$$\leq (\mathbb{E}[|y_i|^4|\,\mathcal{C}])^{1/4} + (\mathbb{E}[\sum_{j=1}^{(k_1+1+s)} |g_{ij,\Omega}|^4|\,\mathcal{C}])^{1/4} \sum_{j=1}^{(k_1+1+s)} \Upsilon_{j,\Omega} < \infty$$

under Assumption 8.1, for $i = 1, \ldots, n$. Then by the Cauchy-Schwarz inequality

$$\mathbb{E}[\sum_{j=1}^{(k_1+1+s)} |z_{ij,\Omega}\varepsilon_i|^2|\,\mathcal{C}] \leq \mathbb{E}[\sum_{j=1}^{(k_1+1+s)} |z_{ij,\Omega}|^4|\,\mathcal{C}]\,\mathbb{E}[|\varepsilon_i|^4|\,\mathcal{C}] < \infty$$

under Assumption 8.1, for $i = 1, \ldots, n$. □

3.8.2 Full simulation tables

This section includes the full simulation tables (tables with more misspecified models and versions of Mi-2SL) Tables 3.8-3.11. The simulation setup is as described in Section 3.5

3.8.3 Additional application tables

This section includes the Mi-2SL estimates for larger SWM cutoffs and when the first-stage fitted values from post Lasso estimates (step 3 in Algorithm 3), Tables 3.12-3.16.

Table 3.8: Bias, MSE and number of selected eigenvectors, $\mu = 4$ (full)

$\sigma^2_{v,u}$	Estimator	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]
			n=100			n=250			n=500	
0.3	GS2SLSa-SAR(1)	0.009(0.016)	0.023(0.029)	-	0.018(0.005)	0.012(0.009)	-	0.017(0.003)	0.016(0.005)	-
0.3	GS2SLSa-SAR(2)	-0.017(0.014)	-0.008(0.025)	-	-0.002(0.005)	-0.008(0.009)	-	-0.005(0.002)	-0.005(0.004)	-
0.3	GS2SLSa-SAR(3)	-0.007(0.015)	-0.005(0.025)	-	0.002(0.009)	-0.005(0.01)	-	0.001(0.003)	-0.001(0.005)	-
0.3	GS2SLSa-SAR(4)	-0.001(0.016)	-0.012(0.027)	-	0.002(0.006)	-0.006(0.01)	-	0.001(0.003)	-0.001(0.005)	-
0.3	GS2SLSb-SAR(1)	-0.001(0.013)	0.032(0.025)	-	0.016(0.005)	0.014(0.009)	-	0.015(0.003)	0.017(0.005)	-
0.3	GS2SLSb-SAR(2)	-0.015(0.011)	-0.011(0.023)	-	-0.004(0.005)	-0.009(0.008)	-	-0.005(0.002)	-0.005(0.004)	-
0.3	GS2SLSb-SAR(3)	-0.011(0.013)	-0.011(0.023)	-	0.001(0.005)	-0.007(0.009)	-	-0.001(0.002)	-0.002(0.004)	-
0.3	GS2SLSb-SAR(4)*	-0.008(0.014)	-0.017(0.024)	-	0.001(0.005)	-0.009(0.009)	-	0.001(0.003)	-0.003(0.005)	-
0.3	simpleOLS	0.427(0.212)	-0.313(0.159)	-	0.28(0.082)	-0.206(0.052)	-	0.296(0.09)	-0.218(0.054)	-
0.3	simpleIV	0.055(0.048)	0.065(0.09)	-	0.047(0.01)	0.031(0.015)	-	0.048(0.006)	0.033(0.009)	-
0.3	Mi-2SLl-a1	0.006(0.022)	0.019(0.036)	32[32,0]	0.016(0.007)	0.014(0.012)	27[27,0]	0.014(0.003)	0.01(0.006)	82[82,0]
0.3	Mi-2SLpl-a1	0.011(0.022)	0.014(0.036)	30[30,0]	0.017(0.007)	0.014(0.012)	26[26,0]	0.014(0.003)	0.01(0.006)	79[79,0]
0.3	Mi-2SLl-a2	0.098(0.088)	-0.083(0.179)	81[80,10]	0.068(0.02)	-0.046(0.027)	169[164,23]	0.11(0.026)	-0.086(0.028)	418[410,78]
0.3	Mi-2SLpl-a2	0.095(0.063)	-0.064(0.103)	81[78,10]	0.065(0.015)	-0.042(0.022)	158[146,23]	0.095(0.017)	-0.066(0.018)	411[392,78]
0.6	GS2SLSa-SAR(1)	0.01(0.016)	0.022(0.03)	-	0.018(0.005)	0.011(0.009)	-	0.017(0.003)	0.016(0.005)	-
0.6	GS2SLSa-SAR(2)	-0.016(0.014)	-0.01(0.025)	-	-0.002(0.005)	-0.009(0.009)	-	-0.005(0.002)	-0.005(0.004)	-
0.6	GS2SLSa-SAR(3)	-0.006(0.016)	-0.009(0.025)	-	0.001(0.012)	-0.006(0.01)	-	0.001(0.003)	-0.001(0.005)	-
0.6	GS2SLSa-SAR(4)	0.001(0.017)	-0.017(0.028)	-	0.003(0.006)	-0.007(0.01)	-	0.002(0.003)	-0.002(0.005)	-
0.6	GS2SLSb-SAR(1)	0.004(0.013)	0.028(0.025)	-	0.017(0.005)	0.012(0.009)	-	0.016(0.003)	0.017(0.005)	-
0.6	GS2SLSb-SAR(2)	-0.012(0.011)	-0.015(0.023)	-	-0.003(0.005)	-0.011(0.009)	-	-0.005(0.002)	-0.006(0.004)	-
0.6	GS2SLSb-SAR(3)	-0.008(0.013)	-0.017(0.023)	-	0.002(0.005)	-0.009(0.009)	-	-0.001(0.002)	-0.003(0.005)	-
0.6	GS2SLSb-SAR(4)*	-0.005(0.014)	-0.024(0.025)	-	0.001(0.005)	-0.012(0.009)	-	0.001(0.003)	-0.004(0.005)	-
0.6	simpleOLS	0.58(0.365)	-0.466(0.275)	-	0.431(0.189)	-0.358(0.137)	-	0.447(0.202)	-0.371(0.143)	-
0.6	simpleIV	0.052(0.048)	0.069(0.092)	-	0.046(0.01)	0.032(0.016)	-	0.047(0.006)	0.034(0.01)	-
0.6	Mi-2SLl-a1	-0.001(0.023)	0.026(0.039)	34[34,0]	0.015(0.007)	0.014(0.012)	31[31,0]	0.012(0.003)	0.011(0.006)	91[91,0]
0.6	Mi-2SLpl-a1	0.007(0.023)	0.018(0.038)	32[32,0]	0.016(0.007)	0.013(0.012)	31[31,0]	0.013(0.003)	0.011(0.006)	88[88,0]
0.6	Mi-2SLl-a2	0.092(0.056)	-0.072(0.086)	81[80,11]	0.07(0.019)	-0.047(0.025)	168[165,23]	0.118(0.029)	-0.093(0.029)	414[409,78]
0.6	Mi-2SLpl-a2	0.102(0.051)	-0.072(0.079)	80[78,11]	0.086(0.02)	-0.062(0.024)	154[142,23]	0.135(0.028)	-0.106(0.026)	405[384,78]
0.9	GS2SLSa-SAR(1)	0.012(0.017)	0.021(0.03)	-	0.019(0.005)	0.011(0.009)	-	0.017(0.003)	0.016(0.005)	-
0.9	GS2SLSa-SAR(2)	-0.015(0.013)	-0.012(0.025)	-	-0.001(0.005)	-0.009(0.009)	-	-0.005(0.002)	-0.005(0.004)	-
0.9	GS2SLSa-SAR(3)	-0.005(0.016)	-0.013(0.026)	-	0.001(0.018)	-0.007(0.012)	-	0.001(0.003)	-0.002(0.005)	-
0.9	GS2SLSa-SAR(4)	0.001(0.017)	-0.021(0.029)	-	0.004(0.007)	-0.009(0.01)	-	0.002(0.003)	-0.003(0.005)	-
0.9	GS2SLSb-SAR(1)	0.008(0.013)	0.024(0.026)	-	0.019(0.005)	0.011(0.009)	-	0.016(0.002)	0.016(0.005)	-
0.9	GS2SLSb-SAR(2)	-0.008(0.011)	-0.019(0.023)	-	-0.002(0.005)	-0.012(0.009)	-	-0.005(0.002)	-0.007(0.004)	-
0.9	GS2SLSb-SAR(3)	-0.005(0.013)	-0.024(0.024)	-	0.002(0.005)	-0.012(0.009)	-	-0.001(0.002)	-0.004(0.005)	-
0.9	GS2SLSb-SAR(4)*	-0.002(0.014)	-0.032(0.026)	-	0.002(0.005)	-0.015(0.01)	-	0.001(0.003)	-0.006(0.005)	-
0.9	simpleOLS	0.733(0.564)	-0.619(0.436)	-	0.582(0.341)	-0.51(0.268)	-	0.598(0.36)	-0.524(0.279)	-
0.9	simpleIV	0.05(0.049)	0.073(0.095)	-	0.045(0.01)	0.033(0.016)	-	0.047(0.006)	0.035(0.01)	-
0.9	Mi-2SLl-a1	-0.006(0.025)	0.031(0.042)	35[35,0]	0.014(0.007)	0.014(0.013)	36[36,0]	0.012(0.004)	0.011(0.007)	99[99,0]
0.9	Mi-2SLpl-a1	0.004(0.025)	0.019(0.041)	34[34,0]	0.015(0.007)	0.013(0.013)	35[35,0]	0.013(0.004)	0.01(0.006)	96[96,0]
0.9	Mi-2SLl-a2	0.079(0.043)	-0.056(0.06)	80[80,11]	0.067(0.017)	-0.044(0.021)	169[167,22]	0.111(0.024)	-0.086(0.024)	411[409,78]
0.9	Mi-2SLpl-a2	0.113(0.046)	-0.082(0.06)	80[77,11]	0.096(0.022)	-0.072(0.025)	150[137,22]	0.162(0.036)	-0.133(0.031)	397[376,78]

Note: Bias (MSE) and GS2SLSb-SAR(4)* is the correctly specified and estimated model.

Table 3.9: Bias, MSE and number of selected eigenvectors, $\mu = 8$ (full)

$\sigma^2_{v,u}$	Estimator	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]
			n=100			n=250			n=500	
0.3	GS2SLSa-SAR(1)	0.004(0.012)	0.009(0.023)	-	0.011(0.005)	0.01(0.01)	-	0.013(0.003)	0.011(0.005)	-
0.3	GS2SLSa-SAR(2)	-0.008(0.012)	-0.007(0.026)	-	-0.006(0.006)	-0.009(0.01)	-	-0.005(0.002)	-0.005(0.005)	-
0.3	GS2SLSa-SAR(3)	0.001(0.027)	-0.008(0.029)	-	0.001(0.007)	-0.004(0.009)	-	0.001(0.003)	0.001(0.005)	-
0.3	GS2SLSa-SAR(4)	-0.002(0.015)	-0.01(0.027)	-	0.003(0.006)	-0.006(0.01)	-	0.002(0.003)	-0.001(0.005)	-
0.3	GS2SLSb-SAR(1)	0.003(0.011)	0.011(0.022)	-	0.008(0.005)	0.013(0.009)	-	0.011(0.002)	0.012(0.005)	-
0.3	GS2SLSb-SAR(2)	-0.005(0.011)	-0.008(0.023)	-	-0.007(0.004)	-0.011(0.008)	-	-0.006(0.002)	-0.005(0.005)	-
0.3	GS2SLSb-SAR(3)	-0.008(0.013)	-0.012(0.024)	-	-0.001(0.005)	-0.007(0.009)	-	0.001(0.003)	-0.001(0.005)	-
0.3	GS2SLSb-SAR(4)*	-0.009(0.014)	-0.018(0.025)	-	0.001(0.005)	-0.009(0.009)	-	0.001(0.003)	-0.002(0.005)	-
0.3	simpleOLS	0.231(0.061)	-0.191(0.06)	-	0.302(0.096)	-0.24(0.072)	-	0.261(0.07)	-0.207(0.049)	-
0.3	simpleIV	0.016(0.017)	0.026(0.033)	-	0.038(0.012)	0.027(0.022)	-	0.034(0.005)	0.021(0.008)	-
0.3	Mi-2SLl-a1	0.002(0.015)	0.017(0.029)	6[6,0]	0.014(0.007)	0.01(0.012)	31[31,0]	0.014(0.003)	0.014(0.005)	38[38,0]
0.3	Mi-2SLpl-a1	0.002(0.015)	0.017(0.029)	5[5,0]	0.014(0.007)	0.009(0.012)	31[31,0]	0.014(0.003)	0.014(0.005)	38[38,0]
0.3	Mi-2SLl-a2	0.013(0.022)	0.006(0.04)	34[33,3]	0.067(0.022)	-0.047(0.029)	183[180,20]	0.056(0.012)	-0.035(0.016)	371[366,40]
0.3	Mi-2SLpl-a2	0.01(0.02)	0.007(0.037)	31[29,3]	0.062(0.017)	-0.039(0.024)	176[169,20]	0.056(0.009)	-0.033(0.012)	351[334,40]
0.6	GS2SLSa-SAR(1)	0.004(0.012)	0.008(0.023)	-	0.012(0.005)	0.009(0.01)	-	0.012(0.003)	0.011(0.005)	-
0.6	GS2SLSa-SAR(2)	-0.007(0.012)	-0.008(0.026)	-	-0.006(0.006)	-0.01(0.011)	-	-0.005(0.002)	-0.005(0.005)	-
0.6	GS2SLSa-SAR(3)	-0.001(0.019)	-0.01(0.027)	-	0.001(0.006)	-0.006(0.009)	-	0.001(0.003)	0.001(0.005)	-
0.6	GS2SLSa-SAR(4)	-0.001(0.015)	-0.013(0.027)	-	0.003(0.006)	-0.007(0.011)	-	0.002(0.003)	-0.002(0.005)	-
0.6	GS2SLSb-SAR(1)	0.006(0.011)	0.008(0.022)	-	0.01(0.005)	0.011(0.009)	-	0.011(0.002)	0.012(0.005)	-
0.6	GS2SLSb-SAR(2)	-0.001(0.011)	-0.012(0.024)	-	-0.006(0.004)	-0.012(0.008)	-	-0.005(0.002)	-0.006(0.005)	-
0.6	GS2SLSb-SAR(3)	-0.004(0.013)	-0.018(0.024)	-	-0.001(0.005)	-0.01(0.009)	-	0.001(0.003)	-0.002(0.005)	-
0.6	GS2SLSb-SAR(4)*	-0.007(0.014)	-0.024(0.025)	-	0.001(0.005)	-0.012(0.009)	-	0.001(0.003)	-0.003(0.005)	-
0.6	simpleOLS	0.382(0.153)	-0.341(0.137)	-	0.452(0.209)	-0.392(0.167)	-	0.412(0.171)	-0.359(0.134)	-
0.6	simpleIV	0.013(0.018)	0.03(0.033)	-	0.037(0.012)	0.028(0.022)	-	0.034(0.005)	0.022(0.008)	-
0.6	Mi-2SLl-a1	-0.003(0.017)	0.022(0.031)	7[7,0]	0.012(0.007)	0.01(0.013)	36[36,0]	0.013(0.003)	0.014(0.006)	45[45,0]
0.6	Mi-2SLpl-a1	-0.003(0.016)	0.022(0.031)	6[6,0]	0.013(0.007)	0.009(0.013)	35[35,0]	0.013(0.003)	0.014(0.006)	45[45,0]
0.6	Mi-2SLl-a2	0.01(0.021)	0.007(0.04)	34[33,3]	0.066(0.02)	-0.046(0.027)	181[180,20]	0.058(0.012)	-0.037(0.015)	368[365,40]
0.6	Mi-2SLpl-a2	0.012(0.021)	0.005(0.038)	30[28,3]	0.082(0.02)	-0.061(0.026)	173[165,20]	0.079(0.013)	-0.056(0.015)	342[324,40]
0.9	GS2SLSa-SAR(1)	0.006(0.012)	0.007(0.023)	-	0.012(0.005)	0.008(0.01)	-	0.012(0.002)	0.011(0.005)	-
0.9	GS2SLSa-SAR(2)	-0.006(0.013)	-0.009(0.027)	-	-0.006(0.006)	-0.011(0.011)	-	-0.005(0.002)	-0.005(0.005)	-
0.9	GS2SLSa-SAR(3)	-0.001(0.016)	-0.012(0.026)	-	0.001(0.006)	-0.007(0.01)	-	0.002(0.003)	0.001(0.005)	-
0.9	GS2SLSa-SAR(4)	0.001(0.016)	-0.016(0.027)	-	0.003(0.006)	-0.009(0.011)	-	0.002(0.003)	-0.002(0.005)	-
0.9	GS2SLSb-SAR(1)	0.01(0.011)	0.005(0.022)	-	0.011(0.005)	0.01(0.01)	-	0.012(0.002)	0.011(0.005)	-
0.9	GS2SLSb-SAR(2)	0.002(0.011)	-0.016(0.024)	-	-0.004(0.004)	-0.014(0.009)	-	-0.005(0.002)	-0.006(0.005)	-
0.9	GS2SLSb-SAR(3)	-0.001(0.013)	-0.023(0.024)	-	0.001(0.005)	-0.012(0.009)	-	0.001(0.003)	-0.003(0.005)	-
0.9	GS2SLSb-SAR(4)*	-0.003(0.014)	-0.03(0.026)	-	0.002(0.005)	-0.015(0.009)	-	0.001(0.003)	-0.005(0.005)	-
0.9	simpleOLS	0.533(0.29)	-0.491(0.258)	-	0.602(0.367)	-0.543(0.306)	-	0.562(0.317)	-0.511(0.265)	-
0.9	simpleIV	0.01(0.018)	0.033(0.035)	-	0.036(0.012)	0.03(0.023)	-	0.033(0.005)	0.023(0.008)	-
0.9	Mi-2SLl-a1	-0.006(0.018)	0.025(0.033)	7[7,0]	0.011(0.007)	0.01(0.014)	40[40,0]	0.011(0.003)	0.015(0.006)	52[52,0]
0.9	Mi-2SLpl-a1	-0.006(0.018)	0.025(0.033)	7[7,0]	0.012(0.007)	0.01(0.014)	39[39,0]	0.011(0.003)	0.015(0.006)	51[51,0]
0.9	Mi-2SLl-a2	0.007(0.021)	0.011(0.04)	33[33,3]	0.06(0.017)	-0.041(0.023)	181[180,19]	0.058(0.011)	-0.036(0.013)	367[366,40]
0.9	Mi-2SLpl-a2	0.01(0.02)	0.007(0.038)	29[27,3]	0.094(0.023)	-0.072(0.027)	172[164,19]	0.099(0.019)	-0.075(0.019)	335[317,40]

Note: Bias (MSE) and GS2SLSb-SAR(4)* is the correctly specified and estimated model.

Table 3.10: Bias, MSE and number of selected eigenvectors, $\mu = 4$ with misspecified links (full)

c	Estimator	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]
			n=100			n=250			n=500	
0.1	GS2SLSa-SAR(1)	0.033(0.024)	0.007(0.045)	-	0.018(0.005)	0.016(0.01)	-	0.017(0.003)	0.017(0.005)	-
0.1	GS2SLSa-SAR(2)	0.003(0.022)	-0.018(0.044)	-	-0.001(0.005)	-0.003(0.01)	-	-0.002(0.003)	-0.003(0.005)	-
0.1	GS2SLSa-SAR(3)	0.017(0.024)	-0.022(0.043)	-	0.004(0.006)	-0.003(0.011)	-	0.002(0.003)	-0.001(0.005)	-
0.1	GS2SLSa-SAR(4)	0.017(0.028)	-0.028(0.042)	-	0.006(0.006)	-0.005(0.011)	-	0.003(0.003)	0.001(0.005)	-
0.1	GS2SLSb-SAR(1)	0.027(0.02)	0.015(0.037)	-	0.017(0.005)	0.016(0.01)	-	0.016(0.003)	0.017(0.005)	-
0.1	GS2SLSb-SAR(2)	0.009(0.016)	-0.024(0.033)	-	0.001(0.005)	-0.005(0.01)	-	-0.003(0.002)	-0.004(0.005)	-
0.1	GS2SLSb-SAR(3)	0.008(0.02)	-0.03(0.035)	-	0.003(0.005)	-0.007(0.01)	-	0.001(0.003)	-0.003(0.005)	-
0.1	GS2SLSb-SAR(4)*	0.005(0.022)	-0.036(0.039)	-	0.003(0.006)	-0.008(0.01)	-	0.001(0.003)	-0.003(0.005)	-
0.1	simpleOLS	0.58(0.367)	-0.476(0.291)	-	0.432(0.19)	-0.357(0.136)	-	0.445(0.2)	-0.368(0.141)	-
0.1	simpleIV	0.071(0.052)	0.041(0.104)	-	0.043(0.009)	0.038(0.016)	-	0.044(0.006)	0.039(0.01)	-
0.1	Mi-2SLl-a1	0.014(0.026)	0.014(0.054)	32[32,0]	0.015(0.007)	0.017(0.013)	29[29,0]	0.01(0.004)	0.013(0.007)	86[86,0]
0.1	Mi-2SLpl-a1	0.02(0.026)	0.009(0.054)	31[31,0]	0.016(0.007)	0.017(0.013)	29[29,0]	0.011(0.004)	0.013(0.007)	83[83,0]
0.1	Mi-2SLl-a2	0.096(0.069)	-0.07(0.1)	79[78,8]	0.064(0.018)	-0.038(0.025)	164[160,21]	0.108(0.025)	-0.083(0.027)	409[404,73]
0.1	Mi-2SLpl-a2	0.107(0.062)	-0.081(0.092)	78[76,8]	0.079(0.018)	-0.049(0.023)	149[137,21]	0.13(0.026)	-0.105(0.027)	398[377,73]
0.3	GS2SLSa-SAR(1)	0.044(0.03)	0.005(0.061)	-	0.022(0.006)	0.018(0.011)	-	0.02(0.003)	0.019(0.006)	-
0.3	GS2SLSa-SAR(2)	0.021(0.024)	-0.015(0.052)	-	0.008(0.006)	0.001(0.011)	-	0.003(0.003)	0.001(0.006)	-
0.3	GS2SLSa-SAR(3)	0.032(0.031)	-0.022(0.053)	-	0.01(0.007)	0.001(0.012)	-	0.007(0.003)	0.002(0.006)	-
0.3	GS2SLSa-SAR(4)	0.032(0.037)	-0.034(0.055)	-	0.012(0.007)	-0.002(0.011)	-	0.007(0.005)	0.002(0.006)	-
0.3	GS2SLSb-SAR(1)	0.043(0.024)	0.01(0.048)	-	0.022(0.006)	0.018(0.011)	-	0.019(0.003)	0.019(0.006)	-
0.3	GS2SLSb-SAR(2)	0.025(0.022)	-0.019(0.048)	-	0.009(0.005)	-0.001(0.01)	-	0.002(0.003)	0.001(0.005)	-
0.3	GS2SLSb-SAR(3)	0.018(0.026)	-0.033(0.046)	-	0.009(0.006)	-0.004(0.011)	-	0.005(0.003)	0.001(0.006)	-
0.3	GS2SLSb-SAR(4)*	0.013(0.026)	-0.042(0.049)	-	0.009(0.006)	-0.006(0.011)	-	0.005(0.004)	0.001(0.006)	-
0.3	simpleOLS	0.58(0.367)	-0.476(0.291)	-	0.432(0.19)	-0.357(0.136)	-	0.445(0.2)	-0.368(0.141)	-
0.3	simpleIV	0.071(0.052)	0.041(0.104)	-	0.043(0.009)	0.038(0.016)	-	0.044(0.006)	0.039(0.01)	-
0.3	Mi-2SLl-a1	0.027(0.033)	0.014(0.075)	29[29,0]	0.019(0.008)	0.02(0.014)	25[25,0]	0.015(0.004)	0.015(0.008)	77[77,0]
0.3	Mi-2SLpl-a1	0.032(0.033)	0.009(0.074)	28[28,0]	0.019(0.008)	0.02(0.014)	25[25,0]	0.015(0.004)	0.014(0.008)	75[75,0]
0.3	Mi-2SLl-a2	0.109(0.082)	-0.072(0.138)	74[73,7]	0.053(0.016)	-0.019(0.024)	153[151,16]	0.096(0.021)	-0.068(0.023)	400[397,59]
0.3	Mi-2SLpl-a2	0.13(0.145)	-0.093(0.283)	73[71,7]	0.065(0.017)	-0.032(0.024)	138[128,16]	0.124(0.025)	-0.093(0.023)	386[368,59]
0.5	GS2SLSa-SAR(1)	0.058(0.039)	0.011(0.073)	-	0.032(0.007)	0.022(0.012)	-	0.027(0.004)	0.023(0.007)	-
0.5	GS2SLSa-SAR(2)	0.045(0.035)	-0.013(0.066)	-	0.022(0.007)	0.012(0.014)	-	0.013(0.003)	0.009(0.006)	-
0.5	GS2SLSa-SAR(3)	0.051(0.044)	-0.028(0.069)	-	0.027(0.019)	0.005(0.026)	-	0.017(0.004)	0.01(0.008)	-
0.5	GS2SLSa-SAR(4)	0.041(0.054)	-0.034(0.075)	-	0.026(0.008)	0.007(0.014)	-	0.018(0.004)	0.01(0.007)	-
0.5	GS2SLSb-SAR(1)	0.062(0.034)	0.011(0.065)	-	0.032(0.007)	0.022(0.012)	-	0.027(0.004)	0.024(0.007)	-
0.5	GS2SLSb-SAR(2)	0.049(0.032)	-0.016(0.062)	-	0.022(0.006)	0.01(0.012)	-	0.013(0.003)	0.009(0.006)	-
0.5	GS2SLSb-SAR(3)	0.035(0.035)	-0.038(0.065)	-	0.021(0.007)	0.005(0.013)	-	0.015(0.004)	0.009(0.007)	-
0.5	GS2SLSb-SAR(4)*	0.025(0.036)	-0.046(0.067)	-	0.022(0.008)	0.003(0.013)	-	0.016(0.004)	0.008(0.006)	-
0.5	simpleOLS	0.58(0.367)	-0.476(0.291)	-	0.432(0.19)	-0.357(0.136)	-	0.445(0.2)	-0.368(0.141)	-
0.5	simpleIV	0.071(0.052)	0.041(0.104)	-	0.043(0.009)	0.038(0.016)	-	0.044(0.006)	0.039(0.01)	-
0.5	Mi-2SLl-a1	0.032(0.041)	0.031(0.089)	22[22,0]	0.029(0.008)	0.03(0.015)	15[15,0]	0.023(0.004)	0.024(0.008)	51[51,0]
0.5	Mi-2SLpl-a1	0.034(0.041)	0.03(0.088)	21[21,0]	0.029(0.008)	0.029(0.015)	15[15,0]	0.023(0.004)	0.023(0.008)	51[51,0]
0.5	Mi-2SLl-a2	0.073(0.07)	-0.008(0.121)	61[60,4]	0.045(0.014)	0.001(0.023)	121[119,8]	0.06(0.013)	-0.018(0.017)	371[369,30]
0.5	Mi-2SLpl-a2	0.088(0.07)	-0.024(0.116)	59[57,4]	0.053(0.014)	-0.006(0.022)	108[102,8]	0.083(0.016)	-0.04(0.017)	350[338,30]
0.7	GS2SLSa-SAR(1)	0.069(0.042)	0.023(0.086)	-	0.042(0.008)	0.026(0.014)	-	0.037(0.005)	0.029(0.008)	-
0.7	GS2SLSa-SAR(2)	0.062(0.042)	-0.002(0.082)	-	0.036(0.008)	0.019(0.014)	-	0.027(0.004)	0.018(0.008)	-
0.7	GS2SLSa-SAR(3)	0.071(0.06)	-0.025(0.082)	-	0.038(0.01)	0.014(0.015)	-	0.032(0.005)	0.021(0.008)	-
0.7	GS2SLSa-SAR(4)	0.069(0.069)	-0.034(0.087)	-	0.039(0.01)	0.012(0.015)	-	0.034(0.006)	0.02(0.008)	-
0.7	GS2SLSb-SAR(1)	0.075(0.041)	0.022(0.082)	-	0.043(0.008)	0.025(0.014)	-	0.038(0.005)	0.029(0.008)	-
0.7	GS2SLSb-SAR(2)	0.065(0.039)	-0.004(0.078)	-	0.036(0.008)	0.017(0.014)	-	0.028(0.004)	0.017(0.008)	-
0.7	GS2SLSb-SAR(3)	0.049(0.048)	-0.032(0.076)	-	0.034(0.009)	0.01(0.014)	-	0.031(0.005)	0.018(0.008)	-
0.7	GS2SLSb-SAR(4)*	0.048(0.054)	-0.043(0.082)	-	0.035(0.009)	0.008(0.015)	-	0.032(0.005)	0.016(0.008)	-
0.7	simpleOLS	0.58(0.367)	-0.476(0.291)	-	0.432(0.19)	-0.357(0.136)	-	0.445(0.2)	-0.368(0.141)	-
0.7	simpleIV	0.071(0.052)	0.041(0.104)	-	0.043(0.009)	0.038(0.016)	-	0.044(0.006)	0.039(0.01)	-
0.7	Mi-2SLl-a1	0.052(0.051)	0.041(0.103)	10[10,0]	0.04(0.009)	0.035(0.016)	5[5,0]	0.038(0.005)	0.031(0.009)	14[14,0]
0.7	Mi-2SLpl-a1	0.052(0.051)	0.04(0.103)	10[10,0]	0.04(0.009)	0.035(0.016)	5[5,0]	0.038(0.005)	0.031(0.009)	14[14,0]
0.7	Mi-2SLl-a2	0.064(0.062)	0.025(0.118)	34[33,1]	0.045(0.012)	0.022(0.021)	57[56,3]	0.042(0.008)	0.02(0.014)	246[246,6]
0.7	Mi-2SLpl-a2	0.065(0.061)	0.023(0.115)	32[31,1]	0.045(0.011)	0.022(0.02)	52[49,3]	0.049(0.009)	0.012(0.014)	231[227,6]

Note: Bias (MSE), GS2SLSb-SAR(4)* is the correctly specified and estimated model and $\sigma^2_{v,u} = 0.6$

Table 3.11: Bias, MSE and number of selected eigenvectors, $\mu = 8$ with misspecified links (full)

c	Estimator	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]	β_2	β_1	No Vecs [2nd,1st]
			n=100			n=250			n=500	
0.1	GS2SLSa-SAR(1)	0.013(0.014)	-0.001(0.026)	-	0.011(0.005)	0.013(0.01)	-	0.011(0.003)	0.012(0.005)	-
0.1	GS2SLSa-SAR(2)	0.003(0.013)	-0.016(0.026)	-	-0.007(0.005)	-0.008(0.009)	-	-0.005(0.003)	-0.005(0.006)	-
0.1	GS2SLSa-SAR(3)	0.008(0.016)	-0.02(0.028)	-	0.002(0.006)	-0.003(0.011)	-	0.002(0.003)	-0.001(0.005)	-
0.1	GS2SLSa-SAR(4)	0.01(0.018)	-0.023(0.029)	-	0.004(0.006)	-0.006(0.01)	-	0.003(0.003)	-0.002(0.005)	-
0.1	GS2SLSb-SAR(1)	0.017(0.013)	-0.003(0.025)	-	0.008(0.005)	0.015(0.01)	-	0.01(0.003)	0.013(0.005)	-
0.1	GS2SLSb-SAR(2)	0.006(0.012)	-0.02(0.025)	-	-0.006(0.005)	-0.008(0.009)	-	-0.005(0.002)	-0.006(0.006)	-
0.1	GS2SLSb-SAR(3)	0.003(0.014)	-0.028(0.026)	-	-0.001(0.006)	-0.007(0.01)	-	0.001(0.003)	-0.003(0.005)	-
0.1	GS2SLSb-SAR(4)*	0.003(0.016)	-0.032(0.027)	-	0.001(0.006)	-0.01(0.01)	-	0.001(0.003)	-0.004(0.005)	-
0.1	simpleOLS	0.386(0.157)	-0.349(0.142)	-	0.451(0.209)	-0.388(0.164)	-	0.409(0.169)	-0.356(0.132)	-
0.1	simpleIV	0.021(0.019)	0.018(0.035)	-	0.034(0.012)	0.035(0.022)	-	0.029(0.005)	0.028(0.009)	-
0.1	Mi-2SLl-a1	0.008(0.015)	0.012(0.03)	6[6,0]	0.011(0.007)	0.016(0.013)	34[34,0]	0.01(0.003)	0.015(0.006)	44[44,0]
0.1	Mi-2SLpl-a1	0.008(0.015)	0.012(0.03)	6[6,0]	0.012(0.007)	0.015(0.013)	33[33,0]	0.011(0.003)	0.014(0.006)	43[43,0]
0.1	Mi-2SLl-a2	0.017(0.021)	-0.003(0.041)	31[30,3]	0.058(0.017)	-0.034(0.024)	179[177,18]	0.054(0.012)	-0.038(0.015)	366[363,40]
0.1	Mi-2SLpl-a2	0.017(0.019)	-0.001(0.039)	28[25,3]	0.077(0.019)	-0.054(0.025)	170[163,18]	0.078(0.014)	-0.06(0.015)	341[323,40]
0.3	GS2SLSa-SAR(1)	0.017(0.015)	-0.002(0.027)	-	0.013(0.006)	0.014(0.011)	-	0.012(0.003)	0.013(0.005)	-
0.3	GS2SLSa-SAR(2)	0.008(0.013)	-0.017(0.029)	-	-0.002(0.006)	-0.004(0.013)	-	-0.003(0.003)	-0.003(0.005)	-
0.3	GS2SLSa-SAR(3)	0.013(0.017)	-0.02(0.029)	-	0.006(0.006)	-0.003(0.012)	-	0.003(0.003)	0.001(0.006)	-
0.3	GS2SLSa-SAR(4)	0.017(0.019)	-0.024(0.031)	-	0.008(0.006)	-0.005(0.011)	-	0.005(0.003)	0.001(0.006)	-
0.3	GS2SLSb-SAR(1)	0.021(0.014)	-0.004(0.026)	-	0.012(0.006)	0.015(0.011)	-	0.011(0.003)	0.014(0.005)	-
0.3	GS2SLSb-SAR(2)	0.011(0.012)	-0.02(0.026)	-	0.001(0.005)	-0.007(0.01)	-	-0.003(0.003)	-0.004(0.005)	-
0.3	GS2SLSb-SAR(3)	0.007(0.015)	-0.028(0.028)	-	0.004(0.006)	-0.007(0.011)	-	0.002(0.003)	-0.001(0.005)	-
0.3	GS2SLSb-SAR(4)*	0.009(0.016)	-0.031(0.029)	-	0.005(0.006)	-0.009(0.011)	-	0.003(0.003)	-0.002(0.006)	-
0.3	simpleOLS	0.386(0.157)	-0.349(0.142)	-	0.451(0.209)	-0.388(0.164)	-	0.409(0.169)	-0.356(0.132)	-
0.3	simpleIV	0.021(0.019)	0.018(0.035)	-	0.034(0.012)	0.035(0.022)	-	0.029(0.005)	0.028(0.009)	-
0.3	Mi-2SLl-a1	0.011(0.016)	0.011(0.032)	5[5,0]	0.011(0.007)	0.016(0.014)	32[32,0]	0.012(0.003)	0.014(0.006)	41[41,0]
0.3	Mi-2SLpl-a1	0.012(0.016)	0.01(0.031)	5[5,0]	0.011(0.007)	0.016(0.014)	32[32,0]	0.012(0.003)	0.014(0.006)	40[40,0]
0.3	Mi-2SLl-a2	0.019(0.021)	-0.003(0.04)	27[26,2]	0.055(0.017)	-0.035(0.024)	173[172,16]	0.054(0.011)	-0.035(0.015)	360[357,35]
0.3	Mi-2SLpl-a2	0.021(0.02)	-0.005(0.038)	24[22,2]	0.072(0.019)	-0.049(0.024)	165[158,16]	0.075(0.013)	-0.054(0.015)	334[317,35]
0.5	GS2SLSa-SAR(1)	0.023(0.016)	-0.002(0.03)	-	0.02(0.007)	0.017(0.013)	-	0.015(0.003)	0.015(0.006)	-
0.5	GS2SLSa-SAR(2)	0.015(0.016)	-0.013(0.031)	-	0.011(0.007)	0.001(0.015)	-	0.004(0.003)	0.003(0.006)	-
0.5	GS2SLSa-SAR(3)	0.019(0.02)	-0.019(0.033)	-	0.014(0.008)	0.002(0.013)	-	0.009(0.003)	0.005(0.006)	-
0.5	GS2SLSa-SAR(4)	0.02(0.022)	-0.024(0.033)	-	0.017(0.009)	0.001(0.014)	-	0.011(0.004)	0.004(0.006)	-
0.5	GS2SLSb-SAR(1)	0.027(0.015)	-0.003(0.029)	-	0.019(0.007)	0.018(0.013)	-	0.015(0.003)	0.016(0.006)	-
0.5	GS2SLSb-SAR(2)	0.019(0.015)	-0.018(0.029)	-	0.011(0.006)	0.001(0.013)	-	0.004(0.003)	0.003(0.006)	-
0.5	GS2SLSb-SAR(3)	0.013(0.017)	-0.029(0.031)	-	0.011(0.007)	-0.002(0.013)	-	0.008(0.003)	0.003(0.006)	-
0.5	GS2SLSb-SAR(4)*	0.012(0.019)	-0.034(0.032)	-	0.012(0.008)	-0.004(0.014)	-	0.009(0.003)	0.002(0.006)	-
0.5	simpleOLS	0.386(0.157)	-0.349(0.142)	-	0.451(0.209)	-0.388(0.164)	-	0.409(0.169)	-0.356(0.132)	-
0.5	simpleIV	0.021(0.019)	0.018(0.035)	-	0.034(0.012)	0.035(0.022)	-	0.029(0.005)	0.028(0.009)	-
0.5	Mi-2SLl-a1	0.013(0.018)	0.013(0.034)	4[4,0]	0.014(0.009)	0.022(0.017)	25[25,0]	0.016(0.004)	0.017(0.007)	30[30,0]
0.5	Mi-2SLpl-a1	0.013(0.018)	0.013(0.034)	3[3,0]	0.015(0.009)	0.021(0.017)	25[25,0]	0.016(0.004)	0.017(0.007)	30[30,0]
0.5	Mi-2SLl-a2	0.018(0.021)	0.004(0.04)	20[19,2]	0.046(0.019)	-0.014(0.029)	157[155,11]	0.041(0.009)	-0.017(0.013)	334[333,23]
0.5	Mi-2SLpl-a2	0.016(0.021)	0.007(0.04)	18[16,2]	0.059(0.019)	-0.028(0.028)	148[143,11]	0.06(0.011)	-0.034(0.013)	310[298,23]
0.7	GS2SLSa-SAR(1)	0.025(0.017)	0.006(0.033)	-	0.03(0.009)	0.021(0.017)	-	0.022(0.004)	0.02(0.007)	-
0.7	GS2SLSa-SAR(2)	0.022(0.017)	-0.006(0.033)	-	0.026(0.009)	0.005(0.026)	-	0.015(0.003)	0.012(0.007)	-
0.7	GS2SLSa-SAR(3)	0.025(0.021)	-0.013(0.035)	-	0.027(0.011)	0.006(0.017)	-	0.02(0.004)	0.012(0.007)	-
0.7	GS2SLSa-SAR(4)	0.028(0.024)	-0.019(0.037)	-	0.029(0.011)	0.004(0.017)	-	0.021(0.004)	0.012(0.007)	-
0.7	GS2SLSb-SAR(1)	0.031(0.017)	0.003(0.032)	-	0.031(0.009)	0.022(0.017)	-	0.023(0.004)	0.02(0.007)	-
0.7	GS2SLSb-SAR(2)	0.026(0.017)	-0.01(0.032)	-	0.026(0.009)	0.008(0.016)	-	0.016(0.003)	0.011(0.007)	-
0.7	GS2SLSb-SAR(3)	0.021(0.019)	-0.022(0.033)	-	0.022(0.01)	0.001(0.016)	-	0.018(0.004)	0.01(0.007)	-
0.7	GS2SLSb-SAR(4)*	0.02(0.021)	-0.027(0.035)	-	0.022(0.011)	-0.001(0.017)	-	0.019(0.004)	0.009(0.007)	-
0.7	simpleOLS	0.386(0.157)	-0.349(0.142)	-	0.451(0.209)	-0.388(0.164)	-	0.409(0.169)	-0.356(0.132)	-
0.7	simpleIV	0.021(0.019)	0.018(0.035)	-	0.034(0.012)	0.035(0.022)	-	0.029(0.005)	0.028(0.009)	-
0.7	Mi-2SLl-a1	0.018(0.019)	0.018(0.036)	1[1,0]	0.026(0.011)	0.028(0.02)	12[12,0]	0.025(0.004)	0.021(0.008)	11[11,0]
0.7	Mi-2SLpl-a1	0.018(0.019)	0.017(0.036)	1[1,0]	0.026(0.011)	0.028(0.02)	12[12,0]	0.025(0.004)	0.021(0.008)	11[11,0]
0.7	Mi-2SLl-a2	0.018(0.022)	0.015(0.042)	9[9,1]	0.035(0.016)	0.012(0.027)	100[99,5]	0.029(0.007)	0.008(0.012)	236[235,8]
0.7	Mi-2SLpl-a2	0.017(0.022)	0.018(0.041)	8[7,1]	0.04(0.017)	0.006(0.026)	94[90,5]	0.035(0.007)	0.002(0.012)	220[215,8]

Note: Bias (MSE), GS2SLSb-SAR(4)* is the correctly specified and estimated model and $\sigma^2_{v,u} = 0.6$

Table 3.12: Mi-2SLl results of Cadena and Kovak (2016) with SWM cutoff 900km

	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
Panel A: Men, high school or less					
Change in log of group-specific	0.312	0.083	0.516	1.558	-0.29
employment	(0.167)	(0.133)	(0.391)	(0.612)	(0.308)
First stage F-statistic	46.88	54.67	50.47	93.47	57.48
Partial F-statistic (Bartik)	52.76	80.16	48.72	23.86	77.39
Number of vecs [1st,2nd]	16[16,0]	19[19,0]	16[16,0]	19[19,0]	10[10,0]
Panel B: Men, some college or more					
Change in log of group-specific	0.338	0.407	0.005	-0.57	0.127
employment	(0.176)	(0.193)	(0.233)	(1.762)	(0.326)
First stage F-statistic	29.32	40.43	29.23	73.26	52.72
Partial F-statistic (Bartik)	62.89	69.16	60.2	38.29	68.66
Number of vecs [1st,2nd]	16[16,0]	19[19,0]	10[10,0]	17[17,0]	13[13,0]
Panel C: Women, high school or less					
Change in log of group-specific	0.083	0.189	0.272	1.811	-0.615
employment	(0.152)	(0.329)	(0.504)	(0.665)	(0.415)
First stage F-statistic	20.63	23.76	15.21	6.04	29.02
Partial F-statistic (Bartik)	49.17	25.6	26.76	13.74	127.88
Number of vecs [1st,2nd]	2[2,0]	19[19,0]	0[0,0]	0[0,0]	3[3,0]
Panel D: Women, some college or more					
Change in log of group-specific	0.231	-0.003	-0.009	-0.033	-0.191
employment	(0.218)	(0.364)	(0.351)	(1.117)	(0.34)
First stage F-statistic	13.1	9.46	21.8	18.8	28.59
Partial F-statistic (Bartik)	27.33	15.79	95.31	33.34	108.93
Number of vecs [1st,2nd]	11[11,0]	10[10,0]	10[10,0]	1[1,0]	15[15,0]

Note: First stage fitted values from Lasso estimates (step 3 in Algorithm 3) with $a = 2$ and the cutoff for the SWM 900km. Robust standard errors appear in parentheses.

Table 3.13: Mi-2SLpl results of Cadena and Kovak (2016) with SWM cutoff 900km

	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
Panel A: Men, high school or less					
Change in log of group-specific	0.312	0.083	0.524	1.661	-0.29
employment	(0.167)	(0.133)	(0.342)	(0.52)	(0.308)
First stage F-statistic	46.88	54.67	44.03	88.95	57.48
Partial F-statistic (Bartik)	52.76	80.16	48.67	22.84	77.39
Number of vecs [1st,2nd]	16[16,0]	19[19,0]	17[16,1]	21[19,2]	10[10,0]
Panel B: Men, some college or more					
Change in log of group-specific	0.338	0.407	0.005	-0.57	0.127
employment	(0.176)	(0.193)	(0.233)	(1.762)	(0.326)
First stage F-statistic	29.32	40.43	29.23	73.26	52.72
Partial F-statistic (Bartik)	62.89	69.16	60.2	38.29	68.66
Number of vecs [1st,2nd]	16[16,0]	19[19,0]	10[10,0]	17[17,0]	13[13,0]
Panel C: Women, high school or less					
Change in log of group-specific	0.083	0.189	0.272	1.811	-0.615
employment	(0.152)	(0.329)	(0.504)	(0.665)	(0.415)
First stage F-statistic	20.63	23.76	15.21	6.04	29.02
Partial F-statistic (Bartik)	49.17	25.6	26.76	13.74	127.88
Number of vecs [1st,2nd]	2[2,0]	19[19,0]	0[0,0]	0[0,0]	3[3,0]
Panel D: Women, some college or more					
Change in log of group-specific	0.231	-0.003	-0.009	-0.033	-0.191
employment	(0.218)	(0.364)	(0.351)	(1.117)	(0.34)
First stage F-statistic	13.1	9.46	21.8	18.8	28.59
Partial F-statistic (Bartik)	27.33	15.79	95.31	33.34	108.93
Number of vecs [1st,2nd]	11[11,0]	10[10,0]	10[10,0]	1[1,0]	15[15,0]

Note: First stage fitted values from post Lasso estimates (step 3 in Algorithm 3) with $a = 2$ and the cutoff for the SWM 900km. Robust standard errors appear in parentheses.

Table 3.14: Mi-2SLl results of Cadena and Kovak (2016) with SWM cutoff 700km

	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
Panel A: Men, high school or less					
Change in log of group-specific	0.231	0.089	0.491	1.544	-0.524
employment	(0.135)	(0.092)	(0.345)	(0.413)	(0.302)
First stage F-statistic	37.17	27.65	45.92	84.3	63.92
Partial F-statistic (Bartik)	66.06	83.1	36.25	42.24	163.15
Number of vecs [1st,2nd]	9[9,0]	13[13,0]	8[8,0]	11[11,0]	4[4,0]
Panel B: Men, some college or more					
Change in log of group-specific	0.473	0.545	-0.091	0.017	0.287
employment	(0.15)	(0.136)	(0.21)	(0.842)	(0.243)
First stage F-statistic	29.93	29.32	23.45	88.71	40.81
Partial F-statistic (Bartik)	88.93	138.98	68.67	69.26	181.3
Number of vecs [1st,2nd]	12[12,0]	13[13,0]	7[7,0]	17[17,0]	11[11,0]
Panel C: Women, high school or less					
Change in log of group-specific	0.221	0.004	0.272	1.811	-0.616
employment	(0.129)	(0.235)	(0.504)	(0.665)	(0.367)
First stage F-statistic	29.66	36.31	15.21	6.04	48.42
Partial F-statistic (Bartik)	72.12	60.39	26.76	13.74	140.81
Number of vecs [1st,2nd]	2[2,0]	8[8,0]	0[0,0]	0[0,0]	2[2,0]
Panel D: Women, some college or more					
Change in log of group-specific	0.267	0.206	-0.176	-0.036	-0.094
employment	(0.218)	(0.288)	(0.327)	(1.134)	(0.342)
First stage F-statistic	34.73	32.15	39.95	17.16	49.31
Partial F-statistic (Bartik)	30.59	27.58	53.89	31.75	66.99
Number of vecs [1st,2nd]	5[5,0]	5[5,0]	4[4,0]	1[1,0]	8[8,0]

Note: First stage fitted values from Lasso estimates (step 3 in Algorithm 3) with $a = 2$ and the cutoff for the SWM 700km. Robust standard errors appear in parentheses.

Table 3.15: Mi-2SLpl results of Cadena and Kovak (2016) with SWM cutoff 700km

	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
Panel A: Men, high school or less					
Change in log of group-specific	0.231	0.089	0.491	1.544	-0.524
employment	(0.135)	(0.092)	(0.345)	(0.413)	(0.302)
First stage F-statistic	37.17	27.65	45.92	84.3	63.92
Partial F-statistic (Bartik)	66.06	83.1	36.25	42.24	163.15
Number of vecs [1st,2nd]	9[9,0]	13[13,0]	8[8,0]	11[11,0]	4[4,0]
Panel B: Men, some college or more					
Change in log of group-specific	0.473	0.545	-0.091	0.017	0.287
employment	(0.15)	(0.136)	(0.21)	(0.842)	(0.243)
First stage F-statistic	29.93	29.32	23.45	88.71	40.81
Partial F-statistic (Bartik)	88.93	138.98	68.67	69.26	181.3
Number of vecs [1st,2nd]	12[12,0]	13[13,0]	7[7,0]	17[17,0]	11[11,0]
Panel C: Women, high school or less					
Change in log of group-specific	0.221	0.004	0.272	1.811	-0.616
employment	(0.129)	(0.235)	(0.504)	(0.665)	(0.367)
First stage F-statistic	29.66	36.31	15.21	6.04	48.42
Partial F-statistic (Bartik)	72.12	60.39	26.76	13.74	140.81
Number of vecs [1st,2nd]	2[2,0]	8[8,0]	0[0,0]	0[0,0]	2[2,0]
Panel D: Women, some college or more					
Change in log of group-specific	0.267	0.206	-0.176	-0.036	-0.094
employment	(0.218)	(0.288)	(0.327)	(1.134)	(0.342)
First stage F-statistic	34.73	32.15	39.95	17.16	49.31
Partial F-statistic (Bartik)	30.59	27.58	53.89	31.75	66.99
Number of vecs [1st,2nd]	5[5,0]	5[5,0]	4[4,0]	1[1,0]	8[8,0]

Note: First stage fitted values from post Lasso estimates (step 3 in Algorithm 3) with $a = 2$ and the cutoff for the SWM 700km. Robust standard errors appear in parentheses.

Table 3.16: Mi-2SLpl results of Cadena and Kovak (2016) with SWM cutoff 500km

	All	Native-born	Foreign-born	Mexican-born	Other foreign-born
Panel A: Men, high school or less					
Change in log of group-specific	0.275	0.164	0.328	1.385	-0.6
employment	(0.18)	(0.1)	(0.436)	(0.308)	(0.321)
First stage F-statistic	62.47	48.25	63.73	80.35	435.1
Partial F-statistic (Bartik)	75.9	65.84	39.54	56.82	144.96
Number of vecs [1st,2nd]	8[8,0]	15[15,0]	6[6,0]	14[14,0]	4[4,0]
Panel B: Men, some college or more					
Change in log of group-specific	0.211	0.323	-0.223	0.19	-0.125
employment	(0.175)	(0.243)	(0.18)	(0.679)	(0.251)
First stage F-statistic	21.77	27.04	35.61	200.98	42.07
Partial F-statistic (Bartik)	33.81	46.72	55.71	190.03	63.06
Number of vecs [1st,2nd]	5[5,0]	7[7,0]	4[4,0]	19[19,0]	4[4,0]
Panel C: Women, high school or less					
Change in log of group-specific	0.148	-0.442	0.272	1.811	-0.737
employment	(0.158)	(0.304)	(0.504)	(0.665)	(0.458)
First stage F-statistic	41.51	30.29	15.21	6.04	49.4
Partial F-statistic (Bartik)	64.18	31.98	26.76	13.74	95.59
Number of vecs [1st,2nd]	1[1,0]	6[6,0]	0[0,0]	0[0,0]	1[1,0]
Panel D: Women, some college or more					
Change in log of group-specific	-0.02	-0.037	-0.581	0.061	-0.987
employment	(0.301)	(0.381)	(0.517)	(1.118)	(0.508)
First stage F-statistic	12.32	10.64	22.37	24.51	21.14
Partial F-statistic (Bartik)	9.62	8.17	26.81	36.29	24.32
Number of vecs [1st,2nd]	1[1,0]	1[1,0]	1[1,0]	1[1,0]	2[2,0]

Note: First stage fitted values from post Lasso estimates (step 3 in Algorithm 3) with $a = 2$ and the cutoff for the SWM 500km. Robust standard errors appear in parentheses.

Chapter 4

The Environmental Kuznets Curve for forest: an application of Mi-Lasso

4.1 Introduction

Understanding the relationship between environmental degradation and economic development is an important question, especially in the context of accelerating climate change as this helps understand if economic development only causes environmental problems, or if it can also be part of their solution. The original hypothesis of Kuznets (1955) was that income inequality increases with development at early stages, up to a point, after this point, inequality decreases. The graphical representation of this hypothesis is the inverted U-shaped curve. One conclusion drawn from

this hypothesis was economic growth would solve the problem of income inequality in the long term. The World-Bank (1993) World Development Report demonstrated empirically that there is also a similar relationship between economic development and environmental degradation. This sparked a large (growing) wave of research on testing for the presence of the environmental Kuznets curve (EKC), i.e., analysing the relationship between economic growth and environmental pollutants such as carbon dioxide, sulphur dioxide, nitrous oxide, etc.

This study focuses on another critical environmental resource/indicator, forests, which is an important resource at both the global and local levels. In the context of the global challenge of climate change, forests play a pivotal role in carbon storage (Seymour and Busch, 2016). At the local level, forest cover plays a central role in biodiversity. Deforestation can arguably be reversed, however, from a biodiversity perspective, all re-forestation is not equal as primary forests tend to have much more biodiversity than plantations.

Deforestation is not a new phenomenon. Historically, human activity has played a crucial role in reshaping the forest landscape for at least 6000 years, with forests being used as a core input for economic development (Williams, 2003, 2008). Between 1982 and 2016, 60% of land use change can be attributed directly to activity and the remaining 40% to indirect human activity, which includes the sources of climate change (Song et al., 2018). Different sources of global forest trends show alternative trends. For example, FAO (2021) concludes that the world is experiencing net forest loss, with a declining rate of deforestation. In contrast, Song et al. (2018) using satellite data found since 1982 that total forest cover has increased, as the decline

in tropical forest cover has been outweighed by the gain in forest cover in boreal, subtropical, and temperate regions. In contrast, economic growth has been relatively sustained over the equivalent period. Caravaggio (2020) argues this is preliminary evidence of the presence of the environmental Kuznets curve for forests (EKCf).

Interestingly over the last 20 years, carbon dioxide emissions from land-use, land-use change, and forestry have shown a slight decrease (Friedlingstein et al., 2021). However, this does not reflect the rise in Brazilian deforestation in recent years (Silva Junior et al., 2021). As many land-use changes are small and hard to detect by remote sensing, these figures need to be taken with caution (Caravaggio, 2020). In the last few decades, there have been several global initiatives related to forests. One example which started in 2005 is the United Nations (UN) program of REDD+ (Reducing emissions from deforestation and forest degradation).

When investigating the EKCf, only a few studies have tried to account for the spatial dimension of the data. Some examples include McPherson and Nieswiadomy (2005); Busa and Waite (2009), and Busa (2013).[1] A standard assumption placed on the cross-sectional units in classical statistical/econometric analysis is independence. If this assumption is violated, i.e., the cross-sectional units are dependent, then the estimates will be biased and/or inconsistent. In cross-country or regional analysis like being conducted for the EKCf, it is hard to justify that the cross-sectional units (countries or regions) are independent. The sample being used is either the whole observed population (all countries/regions) or a specific subset of the observed population. Additionally, deforestation rates are spatially heterogeneous and the

[1] A much larger set of papers have estimated spatial models when looking at other environmental indicators.

geographical distribution of forests (and population/economic activity) is clearly not independent to start with, Spatial modelling is required to account for this.

Classical spatial modelling in economics requires defining two things (1) an $n \times n$ spatial weights matrix (SWM) where the elements define the pair-wise spatial interactions, i.e., the spatial structure, and (2) the spatial economic model like the commonly used first-order spatial autoregressive (SAR(1)) or spatial error model (SEM), i.e., which parts of the structural model are spatially correlated. The problem with parametric structural, spatial models like the SAR(1) is there is no guarantee they are correct specifications. A standard robustness check in the applied spatial economics literature is to test how sensitive the estimated parameters are to different SWMs. When estimates are found to be sensitive to the choice of SWM, researchers have attributed this sensitivity to the choice of SWM; however, as LeSage and Pace (2014) demonstrates, as long as the SWMs are reasonably well correlated, the results should not be overly sensitive to the exact choice of SWM, implying the sensitivity many researchers observe is being driven by the misspecification of the spatial economic model rather than the choice of SWM. Thus, LeSage and Pace (2014) argue that researchers should focus on the spatial model specification rather than finding the 'ideal' SWM.

In the context of the EKCf, I view the spatial parameters in the structural model as nuisance parameters, as I am only interested in the direct effects, i.e., the coefficients on income per capita, and controlling for any possible spatial effects. So rather than trying to specify a structural, spatial model such as the SAR(1) or SEM and then estimating the corresponding spatial parameters, I instead propose using a

methodology called Eigenvector Spatial Filtering (ESF) Griffith (2000, 2003). This approximates the model's spatially correlated parts using a subset of eigenvectors from the SWM as regressors in a linear regression framework. Unlike the conventional approach, where the spatially correlated terms are explicitly specified and the corresponding spatial parameters estimated. ESF has an additional advantage, as demonstrated in simulation in Section 3.5, that the procedure is robust to misspecification of the SWM. To select the relevant subset of eigenvectors, I use the Morans' i Lasso (Mi-Lasso) proposed in Section 2.4. This chapters adds to the EKCf literature by robustly accounting spatial correlation while avoiding misspecification of the spatial parts of the model.

The rest of the paper is outlined as follows, Section 4.2 provides an overview of the theoretical, empirical, and spatial EKCf literature. Section 4.3 describes the data used. Section 4.4 outlines the methodology. Section 4.5 presents our estimated results, and finally, Section 4.6 offers our concluding remarks.

4.2 Literature Review

4.2.1 Theory

Grossman and Krueger (1991) adapted the Kuznets (1955) curve for income inequality to environmental application, which has been widely tested and debated ever since. The World-Bank (1993) Development Report in 1992 inspired further research into the hypothesis by highlighting from observation evidence that after a certain level of economic growth, there is a higher demand for environmental protection,

i.e., reduced environmental degradation. The classic EKC is for all environmental resources. The focus of the EKC literature has been on greenhouse gases as there are long time-series available for this. In contrast, forests have received much less attention.

One of the first papers to propose a theoretical relationship between environmental impact and growth was Ehrlich and Holdren (1971). Specifically, they proposed that the total negative impact of society on the environment (E) follows the relationship $E = PF$ where P is population and F a function on per capita impact, which is determined by factors such as consumption, availability of resources and technology. By assuming technological progress has a natural or weakly beneficial effect, they proposed the adverse effects of using resources and population growth. This sparked a debate on the role of technological progress on the environment, with some arguing that technological progress substantially offsets the negative effect of wealth and population increase. Arrow et al. (1995) argues the EKC relationship can be justified through natural economic progression. A clean agricultural-based economy develops into a dirty industrial resource-depleting manufacturing economy, then into a 'clean' service economy. Suri and Chapman (1998) propose that this process could be accelerated by developed economies exporting polluting processes to less developed countries which have weaker environmental institutions.

In the political economy literature, it is argued that the EKC is a consequence of advanced institutions that directly result from economic development. The idea is that development leads to environmental degradation, which can generally be viewed as externalities. These externalities can only be internalised by advanced institutions,

which requires an advanced level of development. Jones and Manuelli (1995) use an overlapping generations model to show that the inverse-U shaped relationship can be generated through market interaction and the extent of pollution regulation through collective decision-making.

Looking specifically at deforestation (Ehrhardt-Martinez et al., 2002) proposed 'ecological modernisation theory' where economies place more value on the environment and reform themselves to promote environmental goals beyond a certain level of development. This views environmental degradation as short-term disequilibria which are naturally self-correcting. They argue rapid economic transformations cause the deforestation disequilibrium, which is then corrected, i.e., a reduced rate of deforestation or even reforestation, by rural-to-urban population shifts, improved land productivity, and economic evolution.

4.2.2 Empirical

In general empirical investigations into the EKCf estimate the following reduced-form regression:

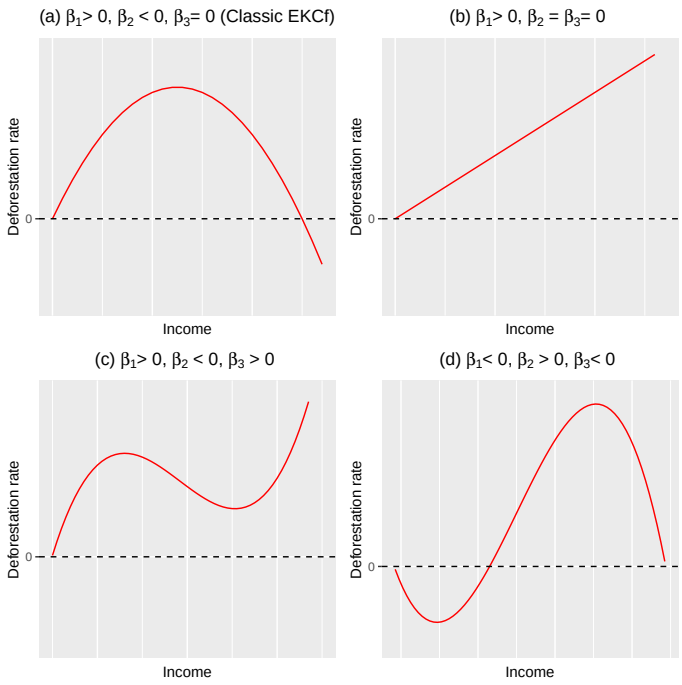
$$deforestation_i = \beta_0 + \beta_1 income_i + \beta_2 income_i^2 + \beta_3 income_i^3 + \sum_{j=1}^{k} \alpha_j \boldsymbol{x}_i + \varepsilon_i,$$

where $\boldsymbol{x}_i$ is a vector of controls.

Figure 4.1 shows expected functional forms for various combinations of coefficient signs. The classic EKCf inverted u-shape (Figure 4.1 (a)) would require a positive

coefficient on income, a negative coefficient on the squared of income, and a zero on the cubic term. This would indicate that the rate of deforestation rises initially with income up to a point and then falls as income rises further. Figure 4.1 (b) a positive linear relationship. Figures 4.1 (c) and (d) show more complicated cubic relationships.

Figure 4.1: Functional Forms for deforestation and income



Empirical studies investigating the EKCf hypothesis of an inverted U-shape have generally not reached a consensus. The first paper to investigate environmental degradation and economic development was Grossman and Krueger (1991) in the context of trade liberalisation. Shafik and Bandyopadhyay (1992) was the first paper to graphicly depict the EKC. They were also the first to use forests as an environmental indicator and other pollutants. They look at 77 countries from 1961 to 1988 and

estimate a panel data model with fixed effects. Each environmental indicator allows for linear, quadratic, and cubic relationships. They find the signs of the estimates do support the hypothesis but are not statistically significant.

The first paper to look solely at EKCf was Cropper and Griffiths (1994), which estimated a panel with fixed effects of 64 tropical countries and found for Latin American and African countries evidence of the EKCf, but not for Asian countries (insignificant results). The author argued higher levels of forest plantations in Asian countries cause this. Their forest data comes from the FAO Production Yearbook. Many other studies have also used this data, e.g., Koop and Tole (1999); Bhattarai and Hammig (2001); Ehrhardt-Martinez et al. (2002); Damette and Delacote (2012). However, the reliability of this source was questioned by Rudel and Roper (1997).

Now looking at more recent country-specific studies. Polomé and Trotignon (2016) who estimate a vector error correction model for the Brazilian Amazon between 1975 and 2014 and find evidence to support the EKCf. However, their less aggregated analysis at the regional level (Araujo et al., 2009) and municipal level (Oliveira and Almeida, 2011) found much weaker evidence. Wang et al. (2019) investigated the EKCf for China between 1970 and 2013 using a provincial-level panel estimated by Generalised Method of Moments and found evidence in support of EKCf. Murshed (2020) investigated the EKCf for Bangladesh using Fully Modified Ordinary Least Squares (Phillips and Hansen, 1990) between 1971 and 2018 and also finds evidence in support of EKCf.

Leblois et al. (2017) use satellite data from Hansen et al. (2013) over the period 2001-2010 for 128 countries, estimating a dynamic panel they find evidence of a linear

relationship only, i.e., evidence against the EKCf. Caravaggio (2020) investigated the EKCf using an unbalanced panel of 114 low-, middle- and high-income countries with a maximum period covered from 1960 to 2015. They estimate an Autoregressive Distributed Lag (p,q) error correct pooled mean group model (Pesaran et al., 1999) by income group to account for the non-stationary in the data and find evidence in support of EKCf.

Choumert et al. (2013) provide an excellent meta-analysis of the EKCf. The authors consider 69 cross-country and country-specific studies between 1992 and 2012 and find the more recent studies with better data, and different econometric techniques tend not to find the hypothesised inverted U-shape. However, Caravaggio (2020), who reviews more recent papers and conducts cross-country analysis, finds evidence supporting the hypothesis.

In conclusion, broadly speaking, the literature has shown mixed results when looking for the presence of the EKCf, with early papers supporting the hypothesis, then until around 2012 refuting the hypothesis, and more recent papers supporting the hypothesis. Thus, the question of the existence of the EKCf is still very much open.

4.2.2.1 Spatial

Most of the research testing the ECKf hypotheses uses conventional statistical techniques, i.e., cross-sectional, panel, or time-series techniques, ignoring possible cross-sectional (spatial) dependence in the data.

Several papers have tried to address this dependence in the EKC literature. For

example, McPherson and Nieswiadomy (2005) estimate a SAR(1), i.e., a mode with a single spatial lag of the dependent variables $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{y}\rho + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ where $\boldsymbol{y}$ is a vector of the dependent variable and $\boldsymbol{X}$ is a matrix of covariates, by Maximum Likelihood (ML). Consistent ML estimation of the SAR(1) requires the underlying model to have normally distributed errors, a strong and arguably unrealistic assumption. Instead of looking at forests, they are looking at threatened bird and mammal species. Maddison (2006) estimate a SAR(1) by robust instrument variables and SEM(1) i.e. a model where the errors $\boldsymbol{u} = \boldsymbol{W}\boldsymbol{u}\lambda + \boldsymbol{v}$ follow a autoregressive process by ML. They use statistical tests to try and determine which spatial model to estimate and a range of different SWMs. Instead of looking at forests, they are looking at various greenhouse gases. Hao et al. (2016) estimate a spatial Durbin model, i.e., a model with a lag of the dependent variable and exogenous variables $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{y}\rho + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{u}$ estimated by ML and look at coal consumption rather than forests. As their model includes a lag of the dependent variable, they can calculate the direct and indirect (spatial) effect. However, they find only the direct effects are significant, despite ρ being significant and negative, which implies negative spatial correlation. Wang and Ye (2017) estimate both a SAR(1) and SEM and look at CO_2. They find the spatial parameters to be significant, but as the spatial spillovers are not of interest, they do not calculate the direct and indirect effects. Chang et al. (2021) estimate a dynamic panel with fixed effects and a spatial lag of the dependent variable estimated by Generalised Method of Moments (Lee and Yu, 2014) and look at air quality instead of forests. The authors find the spatial lag significant but do not calculate the direct and indirect effects.

In contrast, only a few papers have tried to address this dependence in the ECKf literature, Busa and Waite (2009) and Busa (2013) account for the spatial correlation using an alternative spatial filtering approach called principal coordinate of neighbour matrices (PCNM) spatial filtering (Borcard and Legendre, 2002) to account for the spatial correlation in a model agnostic way. PCNM uses principal coordinates (i.e., eigenvectors) from a singular value decomposition of a truncated Euclidean distance matrix. A distance threshold implements the truncation, distances above this threshold are given an arbitrarily large value. They suggest $4(\alpha)$ where α is the threshold. The number of candidate eigenvectors is sensitive threshold (Borcard and Legendre, 2002). Borcard and Legendre (2002) note that negative eigenvalues cannot be used as their coordinates are complex.

Both Busa and Waite (2009) and Busa (2013) only use two eigenvectors without discussing how the eigenvectors were selected or what threshold was used for truncation in their analysis. Busa and Waite (2009) estimate a panel quantile regression model with and without the two eigenvectors. They find the eigenvectors are significant but do not qualitatively change the results. Busa (2013), who also estimate a panel quantile regression model, finds both eigenvectors are significant but do not show/discuss results without the eigenvector. How these eigenvectors are selected and under what conditions consistent selection is achieved in the procedure used is unclear. Additionally, there are few theoretical results for PCNM. Thus, I follow Section 2.4 Mi-Lasso spatial filtering procedure instead as the method works against a wide range of classical spatial economic models, the statistical properties of the estimator have been studied, and the assumptions necessary for consistent

eigenvector selection are known.

4.3 Data

I am working with a panel of 90 countries between 1990–2010. Our forest data is from FRA Global Tables and was first published at 10-year, then 5-year intervals. I thus, have data for 1990, 2000, 2005, and 2010. FRA data has been chosen over FAO data, as FRA data is obtained from officially validated country reports. In contrast, the FAO data uses different definitions of forest and woodlands between developed and developing countries. Thus, FAO is not appropriate for cross-country analysis.

The variable of interest is real Gross Domestic Product per capita (GDPpc) in 2005 dollars (1,000) per capita. Taken from the Pen World Tables for each of the sampling years, which serves as a proxy for economic development. I also include civil liberties (CL) and political freedoms (PR) variation as a proxy for institutions and democracy; obtained from Freedom House, each index is measured on a scale declining scale of 1–7. Population density (popd), forest area (Fora) in 1000's of hectares, land area (Landa) also in 1000's of hectors and a dummy for if the country is developing (Dev).

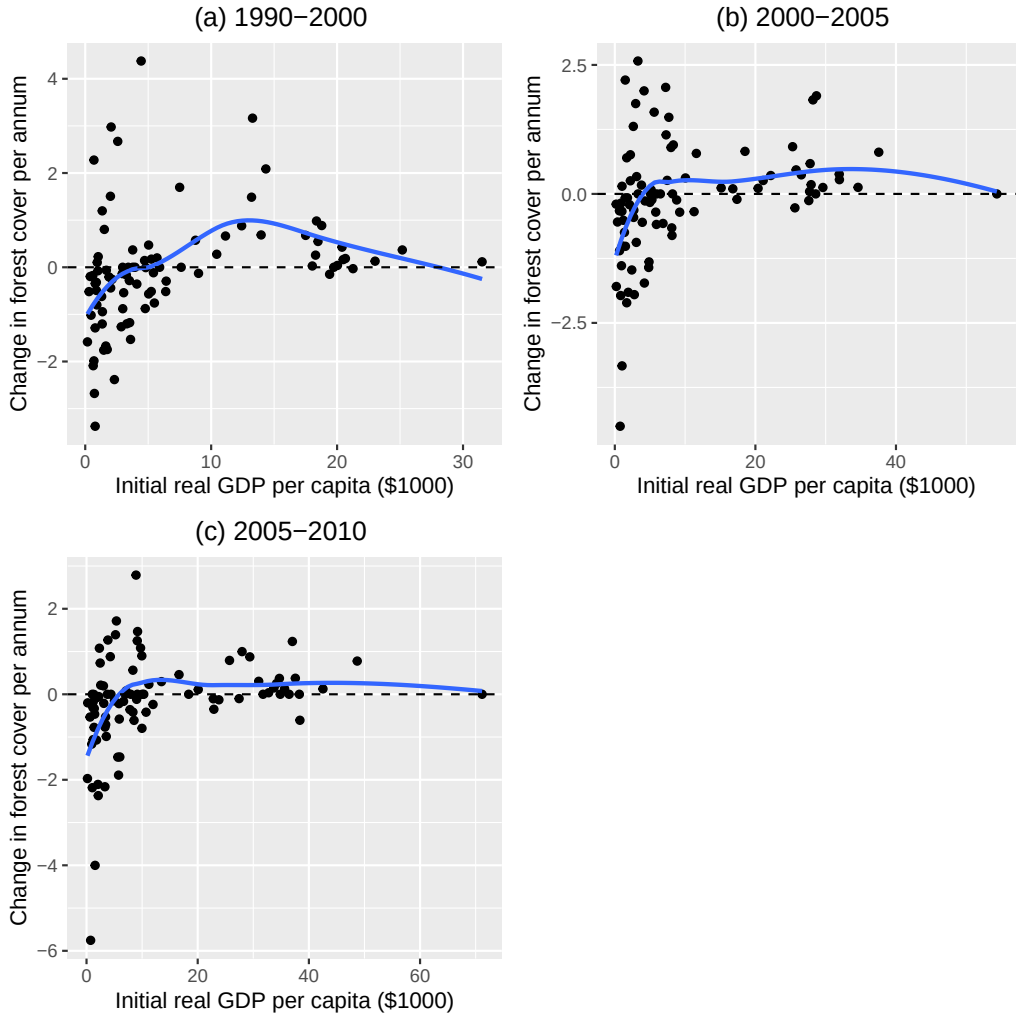
Table 4.1 shows the descriptive statistics for the variables used in the analysis. Interestingly the average yearly deforestation rate in the three periods is negative but very close to zero. GDPpc has become more skewed over the years. Average forest area appears to stay relatively constant over the periods considered.

A visual examination of the raw data is presented in Figure 4.2, which shows

Table 4.1: Descriptive Statistics and definitions

Variable	Definition	Mean	St. Dev.	Min	Max
		1990-2000			
Δf	Change in forest cover pa	−0.066	1.213	−3.374	4.377
y	initial real GDP pc ($1000)	6.885	7.361	0.177	31.515
Fora	Forest area (1000 hectors)	32.976	78.919	0.044	574.839
Landa	Land area (1000 hectors)	97.178	200.734	0.259	942.530
Popd	Population density	1.038	1.268	0.022	8.561
CL	Civil liberties	3.433	1.879	1	7
PR	Political freedoms	3.411	2.167	1	7
Dev	Developing country (dummy)	0.278	0.450	0	1
		2000-2005			
Δf	Change in forest cover pa	−0.074	1.131	−4.503	2.576
y	initial real GDP pc ($1000)	10.214	11.343	0.192	54.354
Fora	Forest area (1000 hectors)	32.251	76.925	0.059	545.943
Landa	Land area (1000 hectors)	97.178	200.734	0.259	942.530
Popd	Population density	1.188	1.448	0.025	10.152
CL	Civil liberties	3.300	1.725	1	7
PR	Political freedoms	3.100	2.120	1	7
PR	Political freedoms	0.278	0.450	0	1
Dev	Developing country (dummy)	0.278	0.450	0	1
		2005-2010			
Δf	Change in forest cover pa	−0.190	1.138	−5.755	2.788
y	initial real GDP pc ($1000)	13.060	14.253	0.169	71.160
Fora	Forest area (1000 hectors)	32.085	76.146	0.067	530.494
Landa	Land area (1000 hectors)	97.178	200.734	0.259	942.530
Popd	Population density	1.257	1.541	0.026	11.073
CL	Civil liberties	2.989	1.764	1	7
PR	Political freedoms	3.044	2.109	1	7
PR	Political freedoms	0.278	0.450	0	1
Dev	Developing country (dummy)	0.278	0.450	0	1

Figure 4.2: Average annual deforestation rate and initial GDP per capita



Note: The blue solid line is a LOWESS (locally weighted scatterplot smoothing) function.

a scatter plot of the average annual rate of deforestation in initial GDPpc for the three periods considered. The curvature of the solid blue line is obtained through a Locally Weighted Scatterplot Smoothing (Cleveland and Devlin, 1988, LOWESS),[2]

[2]LOWESS function splits the data into subsets and then fits a low degree polynomial to each of the subsets using weights least squares.

indicating preliminary evidence of a nonlinear relationship between the deforestation rate and income.

4.4 Methodology

Our underlying cross-country model is:

$$\Delta \boldsymbol{f} = \beta_0 \boldsymbol{\iota} + \beta_1 \boldsymbol{y} + \beta_2 \boldsymbol{y}^2 + \beta_3 \boldsymbol{y}^3 + \boldsymbol{X}\boldsymbol{\zeta} + g(\boldsymbol{W}, \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{y}) + \boldsymbol{u}, \tag{4.1}$$

where $\Delta \boldsymbol{f}$ is an $n \times 1$ vector, $\boldsymbol{\iota}$ is an $n \times 1$ unit vector, $\boldsymbol{X}$ is an $n \times q$ matrix of controls (including a constant) and $g(\boldsymbol{W}, \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{y})$ is some linear in parameter function of the weights matrix $\boldsymbol{W}$ and possibly $\boldsymbol{f}$, $\boldsymbol{X}$ and $\boldsymbol{y}$. For example, $g(\boldsymbol{W}, \Delta \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{y})$ could include not just first-order lags but also higher-order lags (powers). The coefficient of interests are β_1, β_2 and β_3. A cubic term has also been added to allow for a more complicated relationship, the cubic term is the highest order income term that is used in the literature, we thus also include a cubic income term. Due to the availability of the forest data, I consider three periods 1990–2000, 2000–2005, and 2005–2010. GDPpc and the controls are all taken from the starting year, i.e., 1900, 2000, and 2005. This allows us to see the effect that initial income has on the proceeding average deforestation rate.

The SWM $\boldsymbol{W}$ used has a similar setup to Csereklyei and Stern (2015). I use a binary connectivity matrix where pairs of countries get a one if they are neighbours and zero otherwise. Our definition of neighbour is if they share a land border (this includes lakes), and for countries with no land borders, I judge their nearest neigh-

bour. For example, New Zealand's nearest neighbour is Australia. The matrix is symmetric, and the diagonal elements are set equal to zero.[3]

As in our setup, there is substantial uncertainty over which parts of the model exhibit spatial correlation, and as I am not interested in estimating any corresponding spatial parameters, I have generalised the model to include $g(\boldsymbol{W}, \Delta\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{y})$ rather than explicitly specifying which parts of the models are spatially correlated and estimating a classical spatial econometric model like the common SAR(1) or SEM. I instead approximate $g(\boldsymbol{W}, \Delta\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{y})$ using ESF.

4.4.1 Eigenvector Spatial filtering

To account for spatial correlation in this cross-country analysis, I use ESF rather than eigenvectors from a lower triangular distance matrix like the filtering procedure of (Borcard and Legendre, 2002). ESF proposed by Griffith (2000, 2003); Tiefelsdorf and Griffith (2007) takes a more structural approach and uses eigenvectors from a SWM to approximate any spatially correlated omitted variables. The SWM used is assumed to be part of the underlying data-generating process of $\Delta\boldsymbol{f}$. This method is chosen due to the uncertainty over which parts of the model are spatially correlated and the SWM. The idea is to use the eigenvectors $\boldsymbol{E}$ from a spectral decomposition of $\boldsymbol{W}$ as explanatory variables to control or proxy for $g(\boldsymbol{W}, \Delta\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{y})$.[4] This yields the high-dimensional ESF reduced form model:

$$\Delta\boldsymbol{f} = \beta_0\boldsymbol{\iota} + \beta_1\boldsymbol{y} + \beta_2\boldsymbol{y}^2 + \beta_3\boldsymbol{y}^3 + \boldsymbol{X}\boldsymbol{\zeta} + \boldsymbol{E}\boldsymbol{\gamma} + \boldsymbol{u}, \tag{4.2}$$

[3]SWM used available upon request.

[4]A spectral decomposition of $\boldsymbol{W}$ gives n eigenvector.

where $\boldsymbol{E}$ is an $n \times n$ matrix of eigenvectors.

I view $\boldsymbol{E}\boldsymbol{\gamma}$ as a first order approximation of $g(\boldsymbol{W}, \Delta\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{y})$. It is important to note that (4.2) is a high-dimensional linear equation as there are more parameters ($4+k+n$) than observations (n). Thus, estimation by OLS is infeasible. However, Griffith (2000, 2003) argue only specific subset eigenvectors are related to the dependent variable $\Delta\boldsymbol{f}$ and will thus have non-zero coefficients, i.e., the parameter vector $\boldsymbol{\gamma}$ is sparse. OLS estimation is possible if only the subset of relevant eigenvectors is used. However, as the relevant subset is unknown, a selection procedure is required. To solve this selection problem, I use the Lasso-based procedure proposed in Section 2.4. Seya et al. (2015) first proposed selection via Lasso, with the objective function,

$$[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}}, \hat{\boldsymbol{\gamma}}] \in \arg\min\{||\Delta\boldsymbol{f} - \beta_1\boldsymbol{y} - \beta_2\boldsymbol{y}^2 - \beta_3\boldsymbol{y}^3 - \boldsymbol{X}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 + \theta||\boldsymbol{\gamma}||_1\}, \quad (4.3)$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]'$ and θ is the lasso tuning parameter.

It's important to note that the choice of θ determines the number of selected eigenvectors. The selection problem can now be considered a tuning parameter calibration problem. Seya et al. (2015) proposed using k-fold cross-validation prediction accuracy to estimate θ. However, as the aim of ESF is to eliminate patterns of spatial correlation, there is no guarantee the target of prediction accuracy will achieve this. Additionally, when the Lasso tuning parameter is derived by cross-validation, the existing results on theoretical error bounds require the assumptions of random/independent observations (Chetverikov et al., 2021). Given the eigenvectors are derived from a matrix that describes the dependence relationship between

the observations, these bounds are unlikely to hold for ESF. I instead propose using the more intuitive Mi-Lasso, where the tuning parameter is calibrated from the standardised Moran's I (z). Section 2.4 describes the procedure, and Section 2.5 shows the conditions necessary for consistent eigenvector selection and derives finite sample performance bound for the Lasso-based procedure. I estimate the model by post Lasso (OLS estimation with the selected eigenvectors of Mi-Lasso included), and to account for heteroscedasticity robust standard errors are also calculated.

4.5 Results

Table 4.2: Results excluding controls

	Dependent variable:					
	$\Delta \boldsymbol{f}$					
	(1)	(2)	(3)	(4)	(5)	(6)
$\boldsymbol{y}$	0.309***	0.225***	0.182***	0.073*	0.177***	0.058*
	(0.101)	(0.071)	(0.061)	(0.043)	(0.055)	(0.030)
$\boldsymbol{y}^2$	-0.017**	-0.012**	-0.007**	-0.002	-0.006***	-0.001
	(0.008)	(0.005)	(0.003)	(0.002)	(0.002)	(0.001)
$\boldsymbol{y}^3$	0.0003*	0.0002	0.0001**	0.00001	0.00005***	0.00001
	(0.0002)	(0.0001)	(0.00003)	(0.00003)	(0.00002)	(0.00001)
Constant	-0.973***	-0.588***	-0.846***	-0.419**	-1.084***	-0.475***
	(0.285)	(0.222)	(0.285)	(0.169)	(0.316)	(0.135)
Period	1990-2000	1990-2000	2000-2005	2000-2005	2005-2010	2005-2010
Estimator	OLS	Mi-Lasso	OLS	Mi-Lasso	OLS	Mi-Lasso
Adjusted R^2	0.149	0.391	0.129	0.612	0.157	0.784
Partial F-statistic	-	8.15***	-	34.44***	-	16.30***

Note: *p<0.1; **p<0.05; ***p<0.01. Figures in parenthesis are robust standard errors. OLS is a regression with no eigenvectors included and Mi-Lasso is a regression which includes the selected eigenvectors from Mi-Lasso (Section 2.4), see Table 4.5 for the estimates of the eigenvector coefficients. The 'Partial F-statistic' is for an F-test on the included eigenvectors.

Table 4.3: Results including controls

	Dependent variable:					
	$\Delta \boldsymbol{f}$					
	(7)	(8)	(9)	(10)	(11)	(12)
$\boldsymbol{y}$	0.384***	0.259***	0.262***	0.172***	0.212***	0.120***
	(0.096)	(0.089)	(0.071)	(0.058)	(0.061)	(0.042)
$\boldsymbol{y}^2$	-0.025***	-0.018***	-0.009***	-0.005**	-0.006***	-0.003**
	(0.007)	(0.006)	(0.003)	(0.002)	(0.002)	(0.001)
$\boldsymbol{y}^3$	0.0004***	0.0003**	0.0001***	0.00004	0.0001***	0.00002**
	(0.0001)	(0.0001)	(0.00003)	(0.00003)	(0.00002)	(0.00001)
Fora	-0.003*	-0.002	-0.002	-0.004*	-0.001	-0.004**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Landa	0.001	0.001	0.001	0.001	0.0004	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Popd	0.034	0.008	0.039	-0.072	0.010	-0.011
	(0.068)	(0.053)	(0.061)	(0.052)	(0.058)	(0.037)
CL	0.087	-0.061	0.063	0.263**	0.109	0.258
	(0.171)	(0.189)	(0.138)	(0.125)	(0.238)	(0.193)
PR	0.066	0.173	0.159	-0.041	0.009	-0.155
	(0.155)	(0.155)	(0.124)	(0.099)	(0.196)	(0.154)
Dev	0.855	0.869*	0.020	-0.157	-0.136	-0.269
	(0.543)	(0.482)	(0.409)	(0.375)	(0.442)	(0.406)
Constant	-1.813***	-1.129*	-2.029***	-1.465***	-1.698***	-0.924**
	(0.642)	(0.581)	(0.481)	(0.503)	(0.560)	(0.404)
Period	1990-2000	1990-2000	2000-2005	2000-2005	2005-2010	2005-2010
Estimator	OLS	Mi-Lasso	OLS	Mi-Lasso	OLS	Mi-Lasso
Adjusted R^2	0.144	0.348	0.166	0.526	0.118	0.692
Partial F-statistic	-	2.81***	-	1.88**	-	8.78***

Note: *p<0.1; **p<0.05; ***p<0.01. Figures in parenthesis are robust standard errors. OLS is a regression with no eigenvectors included and Mi-Lasso is a regression which includes the selected eigenvectors from Mi-Lasso (Section 2.4), see Table 4.5 for the estimates of the eigenvector coefficients. The 'Partial F-statistic' is for an F-test on the included eigenvectors.

I now present our results where I compare the Mi-Lasso estimated (filtered - including the selected eigenvectors - columns 2, 4, and 6) results with the (unfiltered - columns 1, 3, and 5) OLS results for the three periods considered. Table 4.2 shows

the filtered and unfiltered results for the three periods considered without any controls included. The unfiltered results (columns 1, 3, and 5) indicate a complex cubic relationship between deforestation and income. However, when the selected eigenvectors are included, the magnitude of the coefficients and standard errors shrinks, and the relationship becomes substantially weaker, quadratic (column 2) or linear (columns 4 and 6). The magnitude of the filtered and unfiltered coefficients on income decreases as the periods become more recent. The inclusion of the eigenvectors is further justified by the improvement in the fit of the models (adjusted R^2) in the filtered estimates and given the partial F-test on the Mi-Lasso selected eigenvectors is always significant at the one percent level for all three periods considered.

Table 4.3 presents the unfiltered (OLS - columns 7, 9, and 11) and the filtered (Mi-Lasso - columns 8, 10, and 12) results for the three periods with the controls included. The included controls are generally found to be insignificant. This is reflected by the adjusted R^2 being lower than in the corresponding column in Table 4.2 (except in the case of unfiltered 2000-2005 regressions - column 9/3). The impact of the inclusion of the eigenvectors (filtering) on the income coefficients is similar to Table 4.2 the absolute magnitude of the coefficients as well as their standard errors shrink when the results are filtered, compared to the unfiltered estimates for each of the periods. The main difference between Table 4.3 and 4.2 is the cubic becomes significant in both the filtered and unfiltered estimates, except in the 2000-2005 filtered regression (column 10). Again filtering is further justified by the improvement in the fit of the models (adjusted R^2) in the filtered estimates, and given the partial F-test on the selected eigenvectors is always significant at the five percent level for all three periods

considered.

Table 4.4: Number and significance of selected eigenvectors

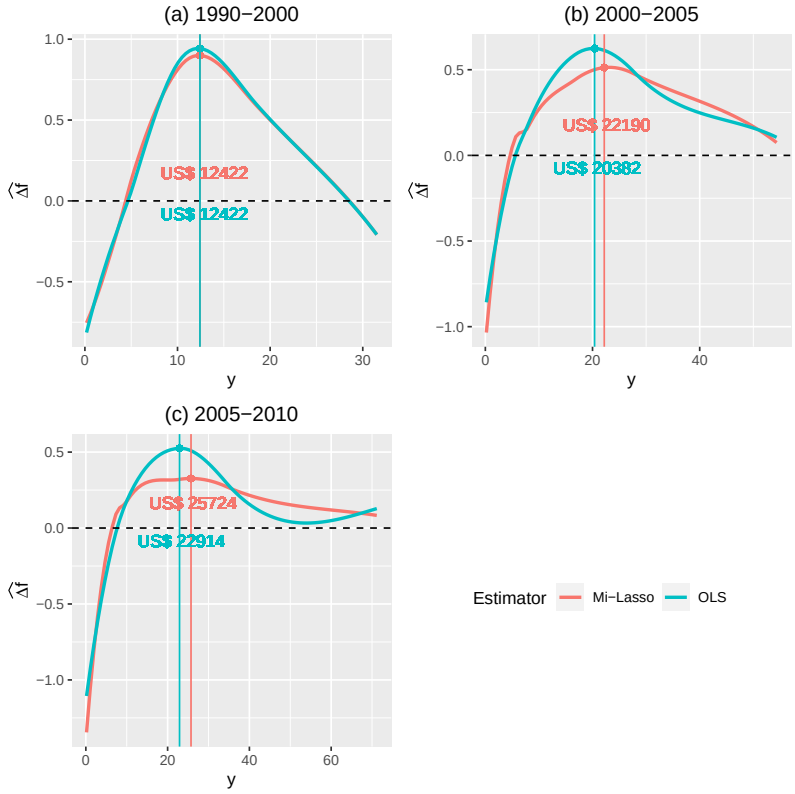
	(2)	(8)	(4)	(10)	(6)	(12)
No of Eigenvectors	4	3	13	8	23	14
Significant at 1% level	3	2	8	4	17	9
Significant at 5% level	0	0	3	4	3	3
Significant at 10% level	1	1	2	0	2	2
Not significant	0	0	0	0	1	0
Period	1990-2000	1990-2000	2000-2005	2000-2005	2005-2010	2005-2010
Controls	excluded	included	excluded	included	excluded	included

Note: Columns (2), (8), (4), (10), (6) and (12) correspond to the equivalent columns in Table 4.2 and 4.3.

Table 4.4 shows the number of selected eigenvectors and their significance. This table shows all the eigenvectors are significant at the 10 percent level, except for the 2005-2010 sample when the controls are excluded (column (6) where 1 of the 23 selected eigenvectors is insignificant). At least half the coefficients on the eigenvectors are significant at the 1 percent level, demonstrating the selected eigenvectors do have explanatory power. Additionally, more eigenvectors are selected when the controls are excluded than included.

Figure 4.3 shows LOWESS functions of the fitted values from the regressions presented in Table 4.3 against initial incomes in the three periods considered with their respective turning points. In the 1990–2000 period (Figure 4.3.a), the turning point for income is the same for the filtered and unfiltered estimates. These estimates look similar to the classic EKC (Figures 4.1 (a)). In contrast, in the 2000–2005 and 2005–2010 periods (Figure 4.3 (b) and Figure 4.3 (c)), the turning point is higher, the peak is flatter, and the second turning point is less clear in the filtered than unfiltered results. Comparing across periods, the per capita income level when the

Figure 4.3: Predicted average annual deforestation rate and initial GDP per capita



Note: $\widehat{\Delta f}$ are the fitted values from the regression results presented in Table 4.3 and the blue and red lines are a LOWESS (locally weighted scatterplot smoothing) function, and the dots show the turning points.

deforestation rate begins to fall is increasing with time, in both the filtered and unfiltered cases. The peak rate of deforestation is also lower in more recent periods. Interestingly comparing the filtered estimates, I find in the 1900–2000 periods after a per capita income of around $12,000 the deforestation rate fall quickly. Whereas, in the 2000–2005 after a per capita income of around $12,000 the deforestation rate is still rising but at a slower rate and by the 2005–2010 periods the deforestation rate remains relatively constant after around $12,000, only falling after about $30,000.

Overall, these results confirm that after accounting for spatial correlation, there is a non-linear relationship between the rate of deforestation and income, however this relationship appears to be changing with time. Like Busa and Waite (2009); Busa (2013) I find the included eigenvectors significant, but unlike Busa and Waite (2009); Busa (2013), I find their inclusion systematically changes the parameter estimates of the covariates.

4.6 Conclusion

In conclusion, I have re-investigated the relationship between a country's per capita GDP and the rate of deforestation for a sample of 90 countries between 1990—2010. To account for the spatial correlation of an unknown functional form and possible misspecification of the SWM, I have used the Moran's I based Lasso (Mi-Lasso) procedure proposed in Section 2.4. I find that if the spatial correlation is ignored, the OLS estimates of income exhibit a systematic absolute upward bias, justifying using Mi-Lasso to estimate the model. In both the filtered (Mi-Lasso) and unfiltered (OLS) estimates, I find evidence of a non-linear relationship that is changing with time. I observe a more complicated relationship in some periods than the originally inverse-U shape postulated by the EKCf. The average peak rate of deforestation appears to be falling with time, while the income level required for the deforestation rate to start falling is increasing with time.

4.7 Appendix

This appendix includes eigenvector parameter estimates for Mi-Lasso results presented in Table 4.2 and 4.3.

Table 4.5: Eigenvector parameter estimates

	Dependent variable:					
	$\Delta \boldsymbol{f}$					
	(2)	(4)	(6)	(8)	(10)	(12)
vec1	-3.643^{***}	-2.182^{***}	-1.699^{***}	-3.106^{***}	-2.738^{***}	-3.242^{***}
	(0.619)	(0.714)	(0.457)	(0.714)	(0.562)	(0.378)
vec2	-3.323^{***}	-2.754^{***}	-2.886^{***}	-3.793^{***}	-3.723^{***}	-5.101^{***}
	(1.094)	(0.586)	(0.323)	(1.229)	(1.002)	(0.780)
vec3	-2.685^{*}	-4.082^{***}	-4.937^{***}	-3.002^{*}	2.178^{**}	1.956^{**}
	(1.601)	(0.734)	(0.635)	(1.772)	(0.921)	(0.775)
vec4	-2.507^{***}	1.981^{**}	1.353^{***}		-2.260^{**}	-1.626^{***}
	(0.863)	(0.835)	(0.458)		(1.012)	(0.424)
vec5		1.953^{***}	2.389^{***}		-2.063^{**}	-1.460^{*}
		(0.738)	(0.601)		(0.985)	(0.811)
vec6		1.963^{***}	-1.147^{**}		1.872^{***}	-1.684^{**}
		(0.475)	(0.547)		(0.595)	(0.777)
vec7		-2.129^{***}	1.176^{**}		-2.238^{**}	-1.634^{**}
		(0.565)	(0.491)		(0.886)	(0.640)
vec8		-1.696^{**}	-1.465^{***}		-2.106^{***}	1.814^{***}
		(0.718)	(0.453)		(0.808)	(0.506)
vec9		-1.489^{**}	-1.320^{***}			1.768^{***}
		(0.687)	(0.373)			(0.628)
vec10		1.607^{*}	0.901			-2.190^{***}
		(0.953)	(0.673)			(0.692)
vec11		1.568^{*}	-1.468^{***}			1.424^{*}
		(0.855)	(0.424)			(0.730)
vec12		-1.821^{***}	-2.040^{***}			1.780^{***}
		(0.664)	(0.609)			(0.610)
vec13		-2.178^{***}	-1.268^{***}			-2.652^{***}
		(0.786)	(0.456)			(0.688)
vec14			1.248^{**}			1.509^{***}
			(0.564)			(0.510)
vec15			1.473^{***}			
			(0.504)			
vec16			-1.998^{***}			
			(0.500)			
vec17			1.546^{***}			
			(0.471)			
vec18			1.496^{*}			
			(0.897)			
vec19			-2.568^{***}			
			(0.617)			
vec20			1.019^{*}			
			(0.520)			
vec21			-1.323^{***}			
			(0.347)			
vec22			1.573^{***}			
			(0.557)			
vec23			-1.108^{***}			
			(0.344)			
Period	1990-2000	2000-2005	2005-2010	1990-2000	2000-2005	2005-2010
Controls	No	No	No	Yes	Yes	Yes

Note: *p<0.1; **p<0.05; ***p<0.01. Figures in parenthesis are robust standard errors. Columns correspond to the equivalent columns in Table 4.2/4.3.

Bibliography

Ahrens, A. (2015), 'Civil conflicts, economic shocks and night-time lights', *Peace Economics, Peace Science and Public Policy* **21**(4), 433–444.

Ahrens, A. (2017), Spatial econometrics and the Lasso estimator: theory and applications, PhD thesis, Heriot-Watt University.

Ahrens, A. and Bhattacharjee, A. (2015), 'Two-Step Lasso Estimation of the Spatial Weights Matrix', *Econometrics* **3**(1), 1–28.

Anselin, L. (2001), Spatial econometrics, *in* B. Baltagi, ed., 'Companion to Econometrics', Basil Blackwell, Oxford, pp. 310–330.

Anselin, L. (2003), 'Spatial externalities, spatial multipliers, and spatial econometrics', *International Regional Science Review* **26**(2), 153–166.

Anselin, L. and Rey, S. (1991), 'Properties of tests for spatial dependence in linear regression models', *Geographical analysis* **23**(2), 112–131.

Anselin, L. and Smirnov, O. (1996), 'Efficient algorithms for constructing proper higher order spatial lag operators*', *Journal of Regional Science* **36**(1), 67–89.

Araujo, C., Bonjean, C. A., Combes, J.-L., Motel, P. C. and Reis, E. J. (2009), 'Property rights and deforestation in the brazilian amazon', *Ecological economics* **68**(8-9), 2461–2468.

Arraiz, I., Drukker, D. M., Kelejian, H. H. and Prucha, I. R. (2010), 'A spatial cliff-ord-type model with heteroskedastic innovations: Small and large sample results*', *Journal of Regional Science* **50**(2), 592–614.

Arrow, K., Bolin, B., Costanza, R., Dasgupta, P., Folke, C., Holling, C. S., Jansson, B.-O., Levin, S., Maler, K.-G., Perrings, C. and Pimentel, D. (1995), 'Economic growth, carrying capacity, and the environment', *Ecological Economics* **15**(2), 91–95.

Badinger, H. and Egger, P. (2013), 'Estimation and testing of higher-order spatial autoregressive panel data error component models', *Journal of Geographical Systems* **15**(4), 453–489.

Baltagi, B. H., Ding, S. and Egger, P. H. (2022), A panel data model with generalized higher-order network effects, *in* 'Essays in Honor of M. Hashem Pesaran: Panel Modeling, Micro Applications, and Econometric Methodology', Emerald Publishing Limited.

Barry, R. P. and Pace, R. K. (1999), 'Monte carlo estimates of the log determinant of large sparse matrices', *Linear Algebra and its Applications* **289**(1), 41 – 54.

Bartik, T. J. (1991), *Who Benefits from State and Local Economic Development*

Policies?, number wbsle *in* 'Books from Upjohn Press', W.E. Upjohn Institute for Employment Research.

Battisti, M. and Di Vaio, G. (2008), 'A spatially filtered mixture of β-convergence regressions for eu regions, 1980–2002', *Empirical Economics* **34**(1), 105–121.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012), 'Sparse models and methods for optimal instruments with an application to eminent domain', *Econometrica* **80**(6), 2369–2429.

Belloni, A. and Chernozhukov, V. (2011), *High Dimensional Sparse Econometric Models: An Introduction*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 121–156.

Belloni, A. and Chernozhukov, V. (2013), 'Least squares after model selection in high-dimensional sparse models', *Bernoulli* **19**(2), 521–547.

Belloni, A., Chernozhukov, V. and Hansen, C. (2013), 'Inference on Treatment Effects after Selection among High-Dimensional Controls†', *The Review of Economic Studies* **81**(2), 608–650.

Belloni, A., Chernozhukov, V. and Wang, L. (2011), 'Square-root lasso: pivotal recovery of sparse signals via conic programming', *Biometrika* **98**(4), 791–806.

Belloni, A., Chernozhukov, V. and Wang, L. (2014), 'Pivotal estimation via square-root lasso in nonparametric regression', *Ann. Statist.* **42**(2), 757–788.

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

Bhattarai, M. and Hammig, M. (2001), 'Institutions and the environmental kuznets curve for deforestation: a crosscountry analysis for latin america, africa and asia', *World development* **29**(6), 995–1010.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009), 'Simultaneous analysis of lasso and dantzig selector', *Ann. Statist.* **37**(4), 1705–1732.

Blommestein, H. J. (1985), 'Elimination of circular routes in spatial dynamic regression equations', *Regional Science and Urban Economics* **15**(1), 121–130.

Blommestein, H. J. and Koper, N. A. M. (1992), 'Recursive algorithms for the elimination of redundant paths in spatial lag operators*', *Journal of Regional Science* **32**(1), 91–111.

Bloom, N., Schankerman, M. and Van Reenen, J. (2013), 'Identifying technology spillovers and product market rivalry', *Econometrica* **81**(4), 1347–1393.

Boots, B. and Tiefelsdorf, M. (2000), 'Global and local spatial autocorrelation in bounded regular tessellations', *Journal of Geographical Systems* **2**(4), 319–348.

Borcard, D. and Legendre, P. (2002), 'All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices', *Ecological modelling* **153**(1-2), 51–68.

Breitung, J. and Wigger, C. (2018), 'Alternative gmm estimators for spatial regression models', *Spatial Economic Analysis* **13**(2), 148–170.

Brent, R. (1973), *Algorithms for minimization without derivatives*, Englewood Cliffs: Prentice-Hall.

Bühlmann, P. and Van De Geer, S. (2011), *Statistics for high-dimensional data*, Springer Series in Statistics, Springer, Heidelberg. Methods, theory and applications.

Busa, J. H. M. (2013), 'Dynamite in the ekc tunnel? inconsistencies in resource stock analysis under the environmental kuznets curve hypothesis', *Ecological Economics* **94**, 116–126.

Busa, J. H. M. and Waite, T. A. (2009), 'Economic prosperity, biodiversity conservation, and the environmental kuznets curve', *Ecological Economics* **68**(7), 2087–2095. Methodological Advancements in the Footprint Analysis.

Cadena, B. C. and Kovak, B. K. (2016), 'Immigrants equilibrate local labor markets: Evidence from the great recession', *American Economic Journal: Applied Economics* **8**(1), 257–90.

Cai, L., Bhattacharjee, A., Calantone, R. and Maiti, T. (2019), 'Variable selection with spatially autoregressive errors: A generalized moments lasso estimator', *Sankhya B: The Indian Journal of Statistics* **81**(1), 146–200.

Cai, L. and Maiti, T. (2020), 'Variable selection and estimation for high-dimensional spatial autoregressive models', *Scandinavian Journal of Statistics* **47**(2), 587–607.

Candes, E. and Tao, T. (2007), 'The dantzig selector: Statistical estimation when p is much larger than n', *Ann. Statist.* **35**(6), 2313–2351.

Caner, M. (2009), 'Lasso-type gmm estimator', *Econometric Theory* **25**(1), 270–290.

Caravaggio, N. (2020), 'A global empirical re-assessment of the environmental kuznets curve for deforestation', *Forest Policy and Economics* **119**, 102282.

Chang, H.-Y., Wang, W. and Yu, J. (2021), 'Revisiting the environmental kuznets curve in china: A spatial dynamic panel data approach', *Energy Economics* **104**, 105600.

Chernozhukov, V., Hansen, C. and Spindler, M. (2015), 'Post-selection and post-regularization inference in linear models with many controls and instruments', *American Economic Review* **105**(5), 486–90.

Chetverikov, D., Liao, Z. and Chernozhukov, V. (2020), 'On cross-validated lasso in high dimensions', *Annals of Statistics* **40**.

Chetverikov, D., Liao, Z. and Chernozhukov, V. (2021), 'On cross-validated Lasso in high dimensions', *The Annals of Statistics* **49**(3), 1300 – 1317.

Chiu, T. Y. M., Leonard, T. and Tsui, K.-W. (1996), 'The matrix-logarithmic covariance model', *Journal of the American Statistical Association* **91**(433), 198–210.

Choumert, J., Motel, P. C. and Dakpo, H. K. (2013), 'Is the environmental kuznets curve for deforestation a threatened theory? a meta-analysis of the literature', *Ecological Economics* **90**, 19–28.

Chun, Y., Griffith, D. A., Lee, M. and Sinha, P. (2016), 'Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters', *Journal of Geographical Systems* **18**(1), 67–85.

Cleveland, W. S. and Devlin, S. J. (1988), 'Locally weighted regression: an approach to regression analysis by local fitting', *Journal of the American statistical association* **83**(403), 596–610.

Cochrane, D. and Orcutt, G. H. (1949), 'Application of least squares regression to relationships containing auto-correlated error terms', *Journal of the American Statistical Association* **44**(245), 32–61.

Crespo Cuaresma, J. and Feldkircher, M. (2013), 'Spatial filtering, model uncertainty and the speed of income convergence in europe', *Journal of Applied Econometrics* **28**(4), 720–741.

Cropper, M. and Griffiths, C. (1994), 'The interaction of population growth and environmental quality', *The American Economic Review* **84**(2), 250–254.

Csereklyei, Z. and Stern, D. I. (2015), 'Global energy use: decoupling or convergence?', *Energy Economics* **51**, 633–641.

Damette, O. and Delacote, P. (2012), 'On the economic factors of deforestation: what can we learn from quantile analysis?', *Economic Modelling* **29**(6), 2427–2434.

De Jong, P., Sprenger, C. and Van Veen, F. (1984), 'On extreme values of moran's i and geary's c', *Geographical Analysis* **16**(1), 17–24.

Debarsy, N., Jin, F. and fei Lee, L. (2015), 'Large sample properties of the matrix exponential spatial specification with an application to fdi', *Journal of Econometrics* **188**(1), 1–21.

Debreu, G. and Herstein, I. N. (1953), 'Nonnegative square matrices', *Econometrica* **21**(4), 597–607.

Doukhan, P. and Louhichi, S. (1999), 'A new weak dependence condition and applications to moment inequalities', *Stochastic Processes and their Applications* **84**(2), 313–342.

Drukker, D. M., Egger, P. H. and Prucha, I. R. (2019), 'Simultaneous equations models with higher-order spatial or social network, interactions'.

Ehrhardt-Martinez, K., Crenshaw, E. M. and Jenkins, J. C. (2002), 'Deforestation and the environmental kuznets curve: A cross-national investigation of intervening mechanisms', *Social Science Quarterly* **83**(1), 226–243.

Ehrlich, P. R. and Holdren, J. P. (1971), 'Impact of population growth', *Science* **171**(3977), 1212–1217.

Elhorst, J., Lacombe, D. J. and Piras, G. (2012), 'On model specification and parameter space definitions in higher order spatial econometric models', *Regional Science and Urban Economics* **42**(1-2), 211–220.

Fan, J. and Liao, Y. (2014), 'Endogeneity in high dimensions', *Ann. Statist.* **42**(3), 872–917.

FAO (2021), Global forest resources assessment 2020, Main report, FAO UN, Rome, Italy.

Fingleton, B. and Le Gallo, J. (2008), 'Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: Finite sample properties*', *Papers in Regional Science* **87**(3), 319–339.

Foster, D. P. and George, E. I. (1994), 'The risk inflation criterion for multiple regression', *Annals of Statistics* **22**(4), 1947–1975.

Friedlingstein, P., Jones, M. W., O'Sullivan, M. et al. (2021), 'Global carbon budget 2021', *Earth System Science Data Discussions* **2021**, 1–191.

Gautier, E. and Rose, C. (2011), 'High-dimensional instrumental variables regression and confidence sets', *arXiv preprint arXiv:1105.2454* .

Getis, A. (1990), 'Screening for spatial dependence in regression analysis', *Papers of the Regional Science Association* **69**(1), 69–81.

Getis, A. and Griffith, D. A. (2002), 'Comparative spatial filtering in regression analysis', *Geographical analysis* **34**(2), 130–140.

Getis, A. and Ord, J. K. (1992), 'The analysis of spatial association by use of distance statistics', *Geographical Analysis* **24**(3), 189–206.

Gibbons, S. and Overman, H. G. (2012), 'Mostly pointless spatial econometrics?*', *Journal of Regional Science* **52**(2), 172–191.

Gilley, O. W. and Pace, R. (1996), 'On the harrison and rubinfeld data', *Journal of Environmental Economics and Management* **31**(3), 403–405.

Griffith, D. A. (2000), 'A linear regression solution to the spatial autocorrelation problem', *Journal of Geographical Systems* **2**(2), 141–156.

Griffith, D. A. (2003), *Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization*, Springer Science & Business Media.

Grimpe, C. and Patuelli, R. (2011), 'Regional knowledge production in nanomaterials: a spatial filtering approach', *The Annals of Regional Science* **46**(3), 519–541.

Grossman, G. M. and Krueger, A. B. (1991), Environmental impacts of a north american free trade agreement, Working Paper 3914, National Bureau of Economic Research.

Gupta, A. (2018), 'Nonparametric specification testing via the trinity of tests', *Journal of Econometrics* **203**(1), 169–185.

Gupta, A. (2019), 'Estimation of spatial autoregressions with stochastic weight matrices', *Econometric Theory* **35**(2), 417–463.

Gupta, A. (2021), 'Efficient closed-form estimation of large spatial autoregressions', *Journal of Econometrics* .

Gupta, A. and Qu, X. (2022), 'Consistent specification testing under spatial dependence', *Econometric Theory* p. 1–42.

Gupta, A. and Robinson, P. M. (2015), 'Inference on higher-order spatial autoregressive models with increasingly many parameters', *Journal of Econometrics* **186**(1), 19–31.

Gupta, A. and Robinson, P. M. (2018), 'Pseudo maximum likelihood estimation of spatial autoregressive models with increasing dimension', *Journal of Econometrics* **202**(1), 92–107.

Haining, R. (1978), 'The moving average model for spatial interaction', *Transactions of the Institute of British Geographers* **3**(2), 202–225.

Han, X., Lee, L.-F. and Xu, X. (2021), 'Large sample properties of bayesian estimation of spatial econometric models', *Econometric Theory* **37**(4), 708–746.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R. et al. (2013), 'High-resolution global maps of 21st-century forest cover change', *science* **342**(6160), 850–853.

Hao, Y., Liu, Y., Weng, J.-H. and Gao, Y. (2016), 'Does the environmental kuznets curve for coal consumption in china exist? new evidence from spatial econometric analysis', *Energy* **114**, 1214–1223.

Harrison, D. and Rubinfeld, D. L. (1978), 'Hedonic housing prices and the demand for clean air', *Journal of Environmental Economics and Management* **5**(1), 81–102.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining, inference and prediction*, 2 edn, Springer.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman & Hall/CRC.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Hoshino, T. (2018), 'Semiparametric spatial autoregressive models with endogenous regressors: With an application to crime data', *Journal of Business & Economic Statistics* **36**(1), 160–172.

Huang, J., Horowitz, J. L. and Ma, S. (2008), 'Asymptotic properties of bridge estimators in sparse high-dimensional regression models', *Annals of Statistics* **36**(2), 587–613.

Huang, J., Ma, S. and Zhang, C.-H. (2008), 'Adaptive lasso for sparse high-dimensional regression models', *Statistica Sinica* pp. 1603–1618.

Jenish, N. (2016), 'Spatial semiparametric model with endogenous regressors', *Econometric Theory* **32**(3), 714–739.

Jia, J. and Yu, B. (2010), 'On model selection consistency of the elastic net when p >> n', *Statistica Sinica* **20**(2), 595–611.

Jin, F. and Lee, L.-F. (2018), 'Irregular n2sls and lasso estimation of the matrix exponential spatial specification model', *Journal of Econometrics* **206**(2), 336 – 358.

Jing, B.-Y., Shao, Q.-M. and Wang, Q. (2003), 'Self-normalized cramér-type large deviations for independent random variables', *Ann. Probab.* **31**(4), 2167–2215.

Jones, L. E. and Manuelli, R. E. (1995), A Positive Model of Growth and Pollution Controls, NBER Working Papers 5205, National Bureau of Economic Research, Inc.

Kelejian, H. H. and Prucha, I. R. (1998), 'A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances', *The Journal of Real Estate Finance and Economics* **17**(1), 99–121.

Kelejian, H. H. and Prucha, I. R. (1999), 'A generalized moments estimator for the autoregressive parameter in a spatial model', *International Economic Review* **40**(2), 509–533.

Kelejian, H. H. and Prucha, I. R. (2001), 'On the asymptotic distribution of the moran i test statistic with applications', *Journal of Econometrics* **104**(2), 219 – 257.

Kelejian, H. H. and Prucha, I. R. (2010), 'Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances', *Journal of Econometrics* **157**(1), 53 – 67. Nonlinear and Nonparametric Methods in Econometrics.

Kelejian, H. and Piras, G. (2017), *Spatial econometrics*, Academic Press.

Knight, K. and Fu, W. (2000), 'Asymptotics for lasso-type estimators', *Annals of Statistics* **28**(5), 1356–1378.

Kojevnikov, D., Marmer, V. and Song, K. (2021), 'Limit theorems for network dependent random variables', *Journal of Econometrics* **222**(2), 882–908.

Koop, G. and Tole, L. (1999), 'Is there an environmental kuznets curve for deforestation?', *Journal of Development economics* **58**(1), 231–244.

Kuznets, S. (1955), 'Economic growth and income inequality', *The American Economic Review* **45**(1), 1–28.

Lam, C. and Souza, P. C. (2020), 'Estimation and selection of spatial weight matrix in a spatial lag model', *Journal of Business & Economic Statistics* **38**(3), 693–710.

Lam, C. and Souza, P. C. L. (2016), 'Detection and estimation of block structure in spatial weight matrix', *Econometric Reviews* **35**(8-10), 1347–1376.

Leblois, A., Damette, O. and Wolfersberger, J. (2017), 'What has driven deforestation in developing countries since the 2000s? evidence from new remote-sensing data', *World Development* **92**, 82–102.

Lee, L.-F. (2002), 'Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models', *Econometric Theory* **18**(2), 252–277.

Lee, L.-F. (2004), 'Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models', *Econometrica* **72**(6), 1899–1925.

Lee, L.-F. (2007), 'Gmm and 2sls estimation of mixed regressive, spatial autoregressive models', *Journal of Econometrics* **137**(2), 489–514.

Lee, L.-F. and Liu, X. (2010), 'Efficient gmm estimation of high order spatial autoregressive models with autoregressive disturbances', *Econometric Theory* **26**(1), 187–230.

Lee, L.-F. and Yu, J. (2014), 'Efficient gmm estimation of spatial dynamic panel data models with fixed effects', *Journal of Econometrics* **180**(2), 174–197.

Leeb, H. and Pötscher, B. M. (2008), 'Can one estimate the unconditional distribution of post-model-selection estimators?', *Econometric Theory* **24**(2), 338–376.

Leng, C., Lin, Y. and Wahba, G. (2004), 'A note on the lasso and related procedures in model selection', *Statistica Sinica* pp. 1273–1284.

LeSage, J. P. and Pace, R. K. (2007), 'A matrix exponential spatial specification', *Journal of Econometrics* **140**(1), 190 – 214. Analysis of spatially dependent data.

LeSage, J. P. and Pace, R. K. (2014), 'The biggest myth in spatial econometrics', *Econometrics* **2**(4), 217–249.

Lewbel, A., Qu, X. and Tang, X. (2021), Social Networks with Mismeasured Links, Boston College Working Papers in Economics 1031, Boston College Department of Economics.

Li, T., Levina, E. and Zhu, J. (2020), 'Network cross-validation by edge sampling', *Biometrika* **107**(2), 257–276.

Liu, X. and Lee, L.-F. (2013), 'Two-stage least squares estimation of spatial autoregressive models with endogenous regressors and many instruments', *Econometric Reviews* **32**(5-6), 734–753.

Maddison, D. (2006), 'Environmental kuznets curves: A spatial econometric approach', *Journal of Environmental Economics and Management* **51**(2), 218–230.

McPherson, M. A. and Nieswiadomy, M. L. (2005), 'Environmental kuznets curve: threatened species and spatial effects', *Ecological Economics* **55**(3), 395–407.

Meinshausen, N. and Bühlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *Annals of Statistics* **34**(3), 1436–1462.

Moran, P. A. P. (1948), 'The interpretation of statistical maps', *Journal of the Royal Statistical Society: Series B (Methodological)* **10**(2), 243–251.

Moran, P. A. P. (1950), 'Notes on Continuous Stochastic Phenomena', *Biometrika* **37**(1-2), 17–23.

Murakami, D. and Griffith, D. A. (2019), 'Eigenvector spatial filtering for large data sets: fixed and random effects approaches', *Geographical Analysis* **51**(1), 23–49.

Murshed, M. (2020), 'Revisiting the deforestation-induced ekc hypothesis: the role of democracy in bangladesh', *GeoJournal* pp. 1–22.

Oberdabernig, D. A., Humer, S. and Crespo Cuaresma, J. (2018), 'Democracy, geography and model uncertainty', *Scottish Journal of Political Economy* **65**(2), 154–185.

Oliveira, R. and Almeida, E. (2011), Deforestation in the brazilian amazonia and spatial heterogeneity: a local environmental kuznets curve approach, *in* '57th Annual North American Meetings of the Regional Science Association International'.

Ord, K. (1975), 'Estimation methods for models of spatial interaction', *Journal of the American Statistical Association* **70**(349), 120–126.

Pace, K. and LeSage, J. (2004), 'Chebyshev approximation of log-determinants of spatial weight matrices', *Computational Statistics & Data Analysis* **45**(2), 179–196.

Pace, R. K., LeSage, J. P. and Zhu, S. (2013), 'Interpretation and computation of estimates from regression models using spatial filtering', *Spatial Economic Analysis* **8**(3), 352–369.

Patuelli, R., Griffith, D. A., Tiefelsdorf, M. and Nijkamp, P. (2011), 'Spatial filtering and eigenvector stability: space-time models for german unemployment data', *International Regional Science Review* **34**(2), 253–280.

Patuelli, R., Schanne, N., Griffith, D. A. and Nijkamp, P. (2012), 'Persistence of regional unemployment: Application of a spatial filtering approach to local labor markets in germany', *Journal of Regional Science* **52**(2), 300–323.

Peng, S. (2019), Heterogeneous Endogenous Effects in Networks, Working paper, arXiv.org.

Pesaran, M. H., Shin, Y. and Smith, R. P. (1999), 'Pooled mean group estimation of dynamic heterogeneous panels', *Journal of the American Statistical Association* **94**(446), 621–634.

Phillips, P. C. and Hansen, B. E. (1990), 'Statistical inference in instrumental variables regression with i (1) processes', *The Review of Economic Studies* **57**(1), 99–125.

Polomé, P. and Trotignon, J. (2016), 'Amazonian deforestation, environmental kuznets curve and deforestation policy: a cointegration approach', *Environmental*

Kuznets Curve and Deforestation Policy: A Cointegration Approach (February 16, 2016) .

Rudel, T. and Roper, J. (1997), 'The paths to rain forest destruction: crossnational patterns of tropical deforestation, 1975–1990', *World Development* **25**(1), 53–65.

Seya, H., Murakami, D., Tsutsumi, M. and Yamagata, Y. (2015), 'Application of lasso to the eigenvector selection problem in eigenvector-based spatial filtering', *Geographical Analysis* **47**(3), 284–299.

Seymour, F. and Busch, J. (2016), *Why forests? Why now?: The science, economics, and politics of tropical forests and climate change*, Brookings Institution Press.

Shafik, N. and Bandyopadhyay, S. (1992), *Economic growth and environmental quality: time-series and cross-country evidence*, Vol. 904, World Bank Publications.

Silva Junior, C. H., Pessoa, A., Carvalho, N. S., Reis, J. B., Anderson, L. O. and Aragao, L. E. (2021), 'The brazilian amazon deforestation rate in 2020 is the greatest of the decade', *Nature ecology & evolution* **5**(2), 144–145.

Smirnov, O. and Anselin, L. (2001), 'Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach', *Computational Statistics & Data Analysis* **35**(3), 301–319.

Song, X.-P., Hansen, M. C., Stehman, S. V., Potapov, P. V., Tyukavina, A., Vermote, E. F. and Townshend, J. R. (2018), 'Global land change from 1982 to 2016', *Nature* **560**(7720), 639–643.

Suri, V. and Chapman, D. (1998), 'Economic growth, trade and energy: implications for the environmental kuznets curve', *Ecological Economics* **25**(2), 195–208.

Tchuente, G. (2019), 'Weak identification and estimation of social interaction models', *arXiv preprint arXiv:1902.06143* .

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Tibshirani, R. J. and Taylor, J. (2011), 'The solution path of the generalized lasso', *Ann. Statist.* **39**(3), 1335–1371.

Tiefelsdorf, M. and Boots, B. (1995), 'The exact distribution of moran's i', *Environment and Planning A: Economy and Space* **27**(6), 985–999.

Tiefelsdorf, M. and Griffith, D. A. (2007), 'Semiparametric filtering of spatial autocorrelation: the eigenvector approach', *Environment and Planning A* **39**(5), 1193–1221.

Tobler, W. R. (1970), 'A computer movie simulating urban growth in the detroit region', *Economic Geography* **46**(sup1), 234–240.

Vega, S. H. and Elhorst, J. P. (2015), 'The slx model', *Journal of Regional Science* **55**(3), 339–363.

Wang, J., Xin, L. and Wang, Y. (2019), 'Economic growth, government policies, and forest transition in china', *Regional Environmental Change* **19**(4), 1023–1033.

Wang, Z. and Ye, X. (2017), 'Re-examining environmental kuznets curve for china's city-level carbon dioxide (co2) emissions', *Spatial Statistics* **21**, 377–389. Regional Economy and Development: A Viewpoint and Application of Spatial Statistics.

Williams, M. (2003), *Deforesting the Earth: From Prehistory to Global Crisis*, University of Chicago Press.

Williams, M. (2008), 'A new look at global forest histories of land clearing', *Annual Review of Environment and Resources* **33**(1), 345–367.

World-Bank (1993), 'The world bank and the environment: The world development report 1992', *Oxford University Press* .

Wright, P. G. (1928), *Tariff on animal and vegetable oils*, Macmillan Company, New York.

Yamada, H. (2017), 'The frisch–waugh–lovell theorem for the lasso and the ridge regression', *Communications in Statistics - Theory and Methods* **46**(21), 10897–10902.

Zhao, P. and Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine learning research* **7**(Nov), 2541–2563.

Zhu, J., Huang, H.-C. and Reyes, P. E. (2010), 'On selection of spatial linear models for lattice data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(3), 389–402.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.