Manuscript submitted to *The Econometrics Journal*, pp. 1–19.

Moran's I Lasso for models with spatially correlated data

SYLVAIN BARDE[†], ROWAN CHERODIAN[‡] AND GUY TCHUENTE[*]

[†] *School of Economics, University of Kent, Park Wood Rd, Canterbury, CT2 7FS, UK.*
E-mail: s.barde@kent.ac.uk

[‡] *Population Health, School of Medicine and Population Health, Regent Court (ScHARR), 30 Regent Street, Sheffield, S1 4DA, UK.*
E-mail: r.cherodian@sheffield.ac.uk

[*] *Purdue University, NISER, GLO, Krannert Building, 403 Mitch Daniels Blvd, West Lafayette, IN, 47907-2056, USA.*
E-mail: gtchuent@purdue.edu

Summary

This paper proposes a Lasso-based estimator which uses information embedded in the Moran statistic to develop a selection procedure called Moran's I Lasso (Mi-Lasso) to solve the Eigenvector Spatial Filtering (ESF) eigenvector selection problem. ESF uses a subset of eigenvectors from a spatial weights matrix to efficiently account for any omitted spatially correlated terms in a classical linear regression framework, thus eliminating the need for the researcher to explicitly specify the spatially correlated parts of the model. We proposed the first ESF procedure accounting for post-selection inference. We derive performance bounds and show the necessary conditions for consistent eigenvector selection. The key advantages of the proposed estimator are that it is intuitive, theoretically grounded, able to provide robust inference and substantially faster than Lasso based on cross-validation or any proposed forward stepwise procedure. Our simulation results and an application on house prices demonstrate Mi-Lasso performs well compared to existing procedures in finite samples.

Keywords: *Spectral analysis, cross-sectional dependence, post-selection inference, high-dimensional statistics, Lasso.*

1. INTRODUCTION

In conventional spatial economic modeling, the researcher is required to specify (i) a spatial weights matrix (SWM) describing the pair-wise relationships between the n cross-sectional units and (ii) a spatial model specifying the spatial correlation of the variables. Standard specifications typically include one or more spatial lags of the dependent, exogenous and/or error term (Kelejian and Piras 2017). Historically, applied researchers have generally focused more on specifying the SWM rather than the empirical spatial structure (LeSage and Pace 2014) and a standard robustness check in the applied spatial economic literature is to test sensitivity of the estimates 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 itself. However LeSage and Pace (2014) show that as long as the various SWMs are reasonably well correlated, estimates should not be affected, which implies that the observed sensitivity is driven by misspecification of the spatial economic model rather than the choice of SWM. LeSage and Pace (2014) thus argue that correct specification of the spatial model should receive more attention.

The Eigenvector Spatial Filtering (ESF) approach, introduced by Griffith (2000, 2003), addresses this issue by using a subset of eigenvectors from the SWM as controls, filtering out spatial effects from the model. Its key strength lies in its agnosticism towards the functional form of the underlying spatial process: this offers a distinct advantage over conventional maximum likelihood (ML) and/or generalized method of moments (GMM) based methods, precisely because researchers are not required to explicitly specify the process generating the spatial correlation. The ESF approach assumes instead that spatial parameters are nuisance parameters. This is particularly advantageous when the spatial process is not the primary focus of the estimation, making this feature appealing to applied researchers who seek to obtain the unbiased direct effect of a change in an explanatory variable from data exhibiting cross-sectional dependence.[1] This approach is preferred because establishing the presence of an underlying spatial process through a spatial correlation test is generally easier than determining its exact form.

As a result of this feature, ESF has gained significant attention in applied economics, for examples see Patuelli et al. (2012); Crespo Cuaresma and Feldkircher (2013); Csereklyei and Stern (2015); Oberdabernig et al. (2018); Kourtellos et al. (2020); Sanso-Navarro et al. (2023). In environmental economics, for instance, many outcomes of interest are inherently spatially correlated. For example, when examining greenhouse gas emissions like CO_2, or energy production (Csereklyei and Stern 2015), researchers are concerned with the direct effects of policy interventions or human actions, and spatial parameters can be treated as nuisance parameters. Additionally, in estimating the impact of exogenous shocks at the local level, researchers often leverage spatial variation between regions without explicitly modelling spatial correlation, as seen in studies by Aum et al. (2021); Bargain and Aminjonov (2020).

Despite its appeal, the critical challenge for ESF is that the spectral decomposition of the $n \times n$ SWM yields n eigenvectors and if all are included in the model, it becomes high-dimensional and estimation by Ordinary Least Squares (OLS) is infeasible.[2] Griffith (2003) argues that only a subset of eigenvectors is necessary to eliminate the cross-sectional dependence in the dependent variable. The key question becomes identifying which subset of eigenvectors is required, which we refer to as the ESF eigenvector selection problem. Several solutions to this selection problem have been proposed, such as several stepwise greedy algorithms where eigenvectors are iteratively added until some user-specified threshold is reached (Griffith 2000, 2003; Tiefelsdorf and Griffith 2007). These stepwise greedy algorithms are simple heuristic approximations to the full ESF selection problem, thus, they are necessarily sub-optimal. Under the assumption of sparsity (i.e. most eigenvector coefficients are zero) Seya et al. (2015) propose using an ℓ_1-penalised regression, e.g. Lasso. Given that Lasso estimates are ultimately determined by a 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 as the loss function. However, a first problem is that the existing 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, where the eigenvectors are derived from a matrix that encodes cross-sectional dependence. Additionally, the goal of ESF is to eliminate spatial correla-

[1] Throughout this paper, we will use the terms cross-sectional dependence and spatial dependence interchangeably.

[2] 'High-dimensional' will specifically refer to specifications with more parameters to estimate than observations, leading to a rank-deficient Gram matrix.

tion patterns and provide reliable inference on the parameter(s) of interest, not improve prediction accuracy, and there is no guarantee that running CV with a prediction accuracy loss will yield consistent eigenvector selection. Related to this, the ESF literature provides no results on constructing robust standard errors, and because all the proposed procedures can be viewed as post-model selection estimators, they may suffer from corresponding inference problems (see in particular Leeb and Pötscher, 2008; Belloni et al., 2014; Farrell, 2015), especially if eigenvector selection is inconsistent.

This paper proposes an alternative procedure for choosing the ESF Lasso tuning parameter, called Moran's I Lasso (Mi-Lasso), which uses a transformation of the Moran's I spatial correlation statistic (Moran 1950) as a point estimate for the Lasso tuning parameter. The intuition behind Mi-Lasso is that when the 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. We show that Mi-Lasso has several advantages; the method is (i) intuitive, (ii) theoretically grounded, (iii) able to provide robust inference and (iv) substantially faster than Lasso with K-fold CV or the stepwise iterative greedy algorithms suggested in the literature.

More specifically, we establish the theoretical properties of Mi-Lasso by formalising the assumption, implicit in the ESF literature, that the terms which include the SWM can be represented by a subset of eigenvectors. Under some standard spatial regularity conditions, we then derive non-asymptotic bounds for the coefficients of the eigenvectors and also assess the additional conditions required for Mi-Lasso to yield consistent eigenvector selection. Simulations confirm that Mi-Lasso performs well, both in terms of bias and of coverage, 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. Finally, we examine the practical performance of Mi-Lasso with an empirical application using the Boston Housing Dataset. We find that Mi-Lasso selects more than triple the number of eigenvectors compared to existing ESF procedures, and is more conservative then them in terms of assessing the significance of estimated parameters.

The rest of this paper is organised as follows, Section 2 describes the underlying model. Section 3 discusses the statistical aspects of ESF and looks at existing methods for the ESF eigenvector selection problem. Section 4 presents the Mi-Lasso procedure and derives several theoretical results. Section 5 provides a Monte Carlo study comparing Mi-Lasso to the main existing selection procedures. Section 6 tests the proposed method in an empirical application on house prices. Finally, Section 7 offers our concluding remarks.

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 process:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r}) + \boldsymbol{v}, \tag{2.1}$$

where $\boldsymbol{\beta}_0$ is the $k \times 1$ parameter vector of interest and $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$ is a function that, as explained below, consists of a linear combination of spatial lags of arbitrary power. $\boldsymbol{W}$ is an $n \times n$ SWM of known constants,[3] $\boldsymbol{y}$, $\boldsymbol{X}$ and an $n \times 1$ vector $\boldsymbol{r}$. One example of

[3] We allow for $\boldsymbol{W}$ 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.

such a model is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \sum_{i=1}^{p} \boldsymbol{W}^i \boldsymbol{y}\rho_{i,0} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\psi}_0 + \boldsymbol{r}, \tag{2.2}$$

$$\boldsymbol{r} = \delta_0 \boldsymbol{W}\boldsymbol{r} + \boldsymbol{v}, \tag{2.3}$$

where $\boldsymbol{\psi}_0$, $\rho_{i,0}$'s and δ_0 describe the degree of spatial correlation in each of the k exogenous variables, the dependent variable and error term. Simpler spatial models can be recovered by setting the spatial parameters $\rho_{i,0}$, δ_0, and/or $\boldsymbol{\psi}_0$ equal to zero, and most spatial models set $p = 1$. The reduced form for $\boldsymbol{y}$ of the DGP (2.2) - (2.3) is:

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

assuming that 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.

The SWM $\boldsymbol{W}$, 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. We 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 normalised, as the transformed model is estimated by a Lasso-based procedure. For reasons of generality, we 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 we omit the n index. Our analysis is conditioned on realised values of $\boldsymbol{X}$ and $\boldsymbol{W}$. We consider higher-order spatial lags only as powers of the SWM $\boldsymbol{W}$ and we allow the number of lags p to be unknown.[4] 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.[5]

We now make the following assumptions about variables in Equation (2.1)

Assumption 2.1. (Regularity of DGP)

1 *(a)* $\boldsymbol{W}$ *are stochastic real symmetric* $n \times n$ *matrices with* $w_{ii} = 0$. *(b) The sequence* $\{\boldsymbol{W}\}$ *is uniformly bounded in both row and column sums.*

2 *The* $n \times k$ *matrices of exogenous variables* $\boldsymbol{X}$ *has full column rank (for large enough* n*) and all the elements of* $\boldsymbol{X}$ *are 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.*

[4] 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 (2.2). Powers of $\boldsymbol{W}$ are viewed as a special case.

[5] Algorithms to construct 'proper' higher-order spatial lags have been proposed in the literature, for example see Anselin and Smirnov (1996).

Assumption 2.1.1-2.1.3 are standard assumptions in the spatial econometrics literature (Kelejian and Prucha 1998; Lee 2004). Assumption 2.1.1 (a) is required for the spectral decomposition, and symmetric SWMs are common in spatial applications, where the SWM is often based on connectivity or distance between pairs of units. A non-symmetric SWM is also possible if the corresponding eigenvectors are real (a symmetric SWM guarantees this). Assumption 2.1.1 (b) is necessary to limit the degree of dependence in $\boldsymbol{y}$. Given Assumption 2.1.1 (a) if the true model is (2.2)-(2.3) and $\boldsymbol{W}$ 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 2.1.2 ensures that the Gram matrix $\boldsymbol{X}'\boldsymbol{X}/n$ is invertible. Assumption 2.1.3 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. If the errors are not sub-Gaussian then results presented in Section 4 hold as long as the Lasso tuning parameter dominates the noise of the model. The finite fourth moment is needed for the selection consistency proof.

3. EIGENVECTOR SPATIAL FILTERING

3.1. Spectral Decomposition and Spatial Filtering

We now show how eigenvectors from a spectral decomposition of $\boldsymbol{W}$ can be used to spatially filter the model described in Section 2. Given assumption 2.1.1 (a), the spectral decomposition of the real and symmetric matrix $\boldsymbol{W}$ is given by:

$$\boldsymbol{W} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}', \tag{3.4}$$

where $\boldsymbol{E}$ is an $n \times n$ matrix of n eigenvectors $\boldsymbol{e}_{i \in N}$ and $\boldsymbol{\Lambda}$ is a $n \times n$ diagonal matrix of n eigenvalues $(\lambda_{i \in N})$ from $\boldsymbol{W}$. The intuition behind ESF is to use individual eigenvectors $\boldsymbol{e}_{i \in N}$ as explanatory variables to proxy for $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$, yielding a high dimensional reduced form model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{E}\boldsymbol{\gamma}_0 + \boldsymbol{v}, \tag{3.5}$$

This requires the following assumptions:

Assumption 3.1. (Sparse Spectral Representation)

1 $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r}) = \boldsymbol{E}\boldsymbol{\gamma}_0$ *where* $\boldsymbol{E}$ *is the* $n \times n$ *matrix of eigenvectors of* $\boldsymbol{W}$.

2 $||\boldsymbol{\gamma}_0||_0 = s < n - k$ *where* $s = s_n$ *is the cardinality of the active set* $\Omega := \operatorname{supp}(\boldsymbol{\gamma}_0)$.

Assumption 3.1.1 ensures that $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$ can be controlled for exactly using the eigenvectors $\boldsymbol{E}$. This assumption is satisfied if $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$ is a linear combination of spatial lags, as each additive term in $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r})$ is pre-multiplied by $\boldsymbol{W}$. For example, substituting (2.3) into (2.2) and using the spectral decomposition of $\boldsymbol{W}$ (3.4) gives:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}' \left(\sum_{i=1}^{p} \boldsymbol{W}^{i-1}\boldsymbol{y}\rho_{i,0} + \boldsymbol{X}\boldsymbol{\psi}_0\boldsymbol{r}\delta_0 \right) + \boldsymbol{v} \tag{3.6}$$

Substituting $\boldsymbol{\gamma}_0 := \boldsymbol{\Lambda}\boldsymbol{E}' \left(\sum_{i=1}^{p} \boldsymbol{W}^{i-1}\boldsymbol{y}\rho_{i,0} + \boldsymbol{X}\boldsymbol{\psi}_0\boldsymbol{r}\delta_0 \right)$ recovers (3.5).

While the spectral representation assumption 3.1.1 helps to obtain the reduced form (3.5), this specification is high-dimensional and cannot be estimated consistently by OLS, due to the violation of the assumption that the regressor matrix $\boldsymbol{G} = [\boldsymbol{X}, \boldsymbol{E}]$ has full column rank.[6] This implies a rank-deficient Gram matrix $\boldsymbol{G'G}/n$ with zero-valued eigenvalues.

Assumption 3.1.2 addresses this problem by requiring that the spatial process be sparse, such that enough degrees of freedom are available to estimate the model by OLS. This is a strong assumption, in the sense it implies exact sparsity, requiring enough elements of $\boldsymbol{\gamma}_0$ to be exactly 0. Note that s does not need to be small relative to n, just small enough for OLS estimation (with the relevant set), which is not overly restrictive unless n is very small.

It is important to note that the ESF Literature implicitly assumes a sparse spectral representation, which is formalised by assumption 3.1. Sparsity or approximate sparsity is commonly observed in many spatial processes and as detailed below, an intuitive justification is that each of the eigenvectors as representing a specific dimension of full spatial domain and only a subset of these dimensions/patterns will be related to $\boldsymbol{y}$.

If assumption 3.1 holds then $f(\boldsymbol{W}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{r}) = \boldsymbol{E}\boldsymbol{\gamma}_0 = \boldsymbol{E}_\Omega\boldsymbol{\gamma}_\Omega$ 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. Thus (3.5) can be reduced to the following low-dimensional equation, where $\boldsymbol{\Upsilon}_0 = [\boldsymbol{\beta}_0, \boldsymbol{\gamma}_\Omega]'$ and $\boldsymbol{G}_\Omega = [\boldsymbol{X}, \boldsymbol{E}_\Omega]$.

$$\boldsymbol{y} = \boldsymbol{G}_\Omega \boldsymbol{\Upsilon}_0 + \boldsymbol{v}, \tag{3.7}$$

In principle, (3.7) can then be estimated by OLS. However, as $\boldsymbol{E}_\Omega$ is unknown, this is infeasible in practice. Thus, we now have a selection problem.

3.2. Existing ESF Selection Procedures and the Moran's I

ESF methods differ in the way they solve the problem of identifying $\boldsymbol{E}_\Omega$, the relevant set of eigenvectors. The first type of procedures proposed are forward stepwise greedy algorithms, where eigenvectors are iteratively added until some user-specified threshold is reached (Griffith 2000, 2003; Tiefelsdorf and Griffith 2007). Griffith (2003) proposes iteratively adding eigenvectors in a greedy manner to the base regression:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{3.8}$$

until the spatial correlation in the OLS residual $\hat{\boldsymbol{u}}$ falls below a pre-specified level. Tiefelsdorf and Griffith (2007) suggest using the standardised version of Moran's I statistic for spatial autocorrelation (Moran 1950) as the criterion for the greedy algorithm.

The test statistic for the Moran's I (m) on the regression residual $\boldsymbol{M_X y} = \hat{\boldsymbol{u}}$ of $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ where $\boldsymbol{M_X} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{X}'$ is given by:

$$m = \frac{\boldsymbol{y}'\boldsymbol{M_X}\boldsymbol{W}\boldsymbol{M_X}\boldsymbol{y}}{\boldsymbol{y}'\boldsymbol{M_X}\boldsymbol{y}} = \frac{\hat{\boldsymbol{u}}'\boldsymbol{W}\hat{\boldsymbol{u}}}{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}, \tag{3.9}$$

where $\boldsymbol{W}$ is a $n \times n$ real symmetric SWM.[7] Substituting (3.4) in (3.9):

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

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

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

The standardised version of Moran's I statistic (Z) on the residual $\hat{\boldsymbol{u}}$ is:[8]

$$\begin{aligned} Z =& \left(\frac{m - \mathbb{E}[m]}{\sqrt{\operatorname{Var}(m)}} \right), \qquad (3.10) \\ \mathbb{E}[m] =& \frac{tr(\boldsymbol{M_X W M_X})}{n-k}, \\ \operatorname{Var}(m) =& \frac{2\Big((n-k) tr\big((\boldsymbol{M_X W M_X})^2 \big) - \big[tr(\boldsymbol{M_X W M_X}) \big]^2 \Big)}{(n-k)^2(n-k-2)}. \end{aligned}$$

In practice, the forward stepwise greedy algorithm searches for the eigenvector within the candidate set $\boldsymbol{E}_c$ that minimizes $|Z|$. The selected eigenvector $\boldsymbol{e}_{i \in N}$ is then removed from $\boldsymbol{E}_c$ and added to the design matrix of (3.8), and the residuals $\hat{\boldsymbol{u}}$ of this updated regression are tested to check if $|Z| < \epsilon$, where ϵ is a pre-specified threshold level of Z, which they suggest should be dependent on the sample size n.[9] If the condition is satisfied the iterations stop, if not the algorithm continues searching in the remaining candidate eigenvector set $\boldsymbol{E}_c$, with this iterative process continuing until $|Z| < \epsilon$.

Tiefelsdorf and Griffith (2007) motivate the use of Moran's I based on its power against a wide array of autoregressive models and residual distributions (Anselin and Rey 1991), and the fact it can be used for small samples (Kelejian and Piras 2017). De Jong et al. (1984) show that the range of m is the range of the eigenvalues of $\boldsymbol{M_X W M_X}$, and all possible realisations of m are just linear combinations of these n eigenvalues (Tiefelsdorf and Boots 1995; Boots and Tiefelsdorf 2000), implying that Moran's I can be decomposed into the contribution provided by each eigenvector. This is visible from the fact that the numerator of m includes $\boldsymbol{E}'\hat{\boldsymbol{u}}$, which given the orthogonality of $\boldsymbol{E}$ is the OLS estimate from a regression of $\hat{\boldsymbol{u}}$ on $\boldsymbol{E}$. Griffith (2003) thus argues that the n eigenvectors represents mutually orthogonal spatial patterns, and only a subset $\boldsymbol{E}_c \subseteq \boldsymbol{E}$ will be relevant to the model, i.e., in a regression framework only that subset will have non-zero coefficients. On this basis, Griffith (2003) makes several recommendations. First, if $\boldsymbol{y}$ exhibits positive global spatial autocorrelation then $\boldsymbol{E}_c$ should be restricted to those eigenvectors with positive eigenvalues and thus associated with at least weak positive spatial autocorrelation. Second, eigenvectors with small eigenvalues should be excluded from $\boldsymbol{E}_c$, 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.

These forward stepwise procedures, through intuitive, have several key disadvantages. First, a lot of parameters are left to the user's discretion, such as, which statistic or information criterion to use, what threshold ϵ to use, which eigenvectors to include in the initial $\boldsymbol{E}_c$, and in which order to add the eigenvectors. Second, these greedy algorithms could also be at risk of data mining, with estimated models falling victim to over-fitting. Third, all these approaches are heuristics that aim to simplify the original, and infeasible, subset sum problem; therefore the solutions they obtain will be sub-optimal, with no guarantee they are close to the optimal one. Finally, these sequential methods carry a large computational burden, which becomes more acute when n is large. This can be

[8] 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.

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

mitigated by limiting $\boldsymbol{E}_c$ with the rules of thumb mentioned above, but again with no guarantee these rules will consistently recover $\boldsymbol{E}_\Omega$.

This motivates Seya et al. (2015) to propose using Lasso (Tibshirani 1996), which shrinks many of the coefficients to zero, and can thus be used for variable selection. Seya et al. (2015) use Lasso under the assumption that the parameter vector $\boldsymbol{\gamma}_0$ is sparse and the matrix of regressors $\boldsymbol{X}$ has full column rank, so that only the $\boldsymbol{\gamma}$ vector is penalised. The resulting Lasso estimator is:

$$[\hat{\boldsymbol{\beta}}_\theta, \hat{\boldsymbol{\gamma}}_\theta] \in \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 + \theta||\boldsymbol{\gamma}||_1\}, \tag{3.11}$$

where $\theta > 0$ is the Lasso regularization or tuning parameter. Equation (3.11) defines a family of estimators indexed by the tuning parameter θ, a hyperparameter that ultimately determines which eigenvectors the Lasso selects.

Seya et al. (2015) propose using k-fold cross-validation (CV) combined with the Brent algorithm to estimate $\hat{\theta}$, with prediction accuracy as the loss function. 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 optimisation 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. Additionally, because the theoretical results on CV-Lasso hinge on the assumption that the cross-sectional units are independent (Chetverikov et al. 2020), it is hard to justify their validity for ESF, where eigenvectors are derived from a matrix that encodes cross-sectional dependence. 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 of the ESF approximation is sensitive to the underlying spatial processes. Chun et al. (2016) argue that more eigenvectors are needed when the level of spatial correlation is high compared to when it is low, thus simple rules based, for example, on sample size may result in a sub-optimal set of eigenvectors being selected. Additionally, it is also important to note that none of these proposed procedures take into account the post-selection inference problem (Leeb and Pötscher 2008), so are unlikely to provide robust inference.

4. THE MORAN'S *I* LASSO ESTIMATOR

4.1. Moran's I Lasso framework for eigenvector selection

Lasso estimates are ultimately a function of the tuning parameter θ. Supposing $\theta = 0$, the Lasso solution reduces to the OLS solution, whereas with a sufficiently large θ 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, the goal of ESF is to eliminate spatial correlation patterns in a linear regression framework. Information about these patterns will be contained in the regression residuals $\hat{\boldsymbol{u}}$, and we propose using these 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

```
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 Estimate
      $[\hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\gamma}}_L] \in \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 + \frac{1}{Z^2} \cdot ||\boldsymbol{\gamma}||_1\}$   (4.12)
  and save the selected eigenvectors, denoted $\boldsymbol{E}_L$.
4 Let $\hat{\boldsymbol{\beta}}_{pr}$ be the OLS coefficient from a regression of $\bar{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{E}_L\hat{\boldsymbol{\gamma}}_L$ on $\boldsymbol{M}_E\boldsymbol{X}$
  where $\boldsymbol{M}_E := \boldsymbol{I} - \boldsymbol{E}_L(\boldsymbol{E}_L'\boldsymbol{E}_L)^{-1}\boldsymbol{E}_L'$.
5 Calculate a conventional (heteroskedastic) standard error for $\hat{\boldsymbol{\beta}}_{pr}$.
```

Algorithm 1: Mi-Lasso Algorithm

of eigenvectors will be necessary. Thus, a low level of regularization (value of θ) is required. Following Tiefelsdorf and Griffith (2007) we propose using the standardised Moran's I (3.10) to measure the spatial correlation of the residuals due to its previously mentioned properties, in particular its decomposability with respect to the set of eigenvectors $\boldsymbol{E}$ used in the Lasso estimation. As Z takes on large values when the correlation is high and small values when the correlation is low, we propose using the inverse of the square of Z from the residuals of (3.8) as a point estimate of θ,

$$\theta = \frac{1}{Z^2}, \qquad \forall \ Z \neq 0 \tag{4.13}$$

The inverse square ensures that the tuning parameter is always positive regardless of the value of Z, as a positive tuning parameter is required for a unique Lasso solution to exist. This choice is further discussed in section 4.3. The proposed estimator is called Moran's I Lasso (Mi-Lasso) and is outlined in Algorithm 1.

As Lasso is a shrinkage estimator, it induces a downward bias on the estimated non-zero coefficients. One option is to use post-Lasso, where the Lasso estimator is first used as a selection procedure, before OLS is used to estimate the model selected by Lasso. This straightforwardly provides unbiased estimates, assuming Lasso selects the correct eigenvectors. The Moran's I Post-Lasso (Mi-pLasso) estimator is defined as:

$$[\hat{\boldsymbol{\beta}}_{pL}, \hat{\boldsymbol{\gamma}}_{pL}] = \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 \quad \text{subject to} \quad \text{supp}(\boldsymbol{\gamma}_0) = \text{supp}(\hat{\boldsymbol{\gamma}}_L). \tag{4.14}$$

However, in practice there is no guarantee Lasso will yield perfect selection. Wüthrich and Zhu (2023) highlight the often large impact that imperfect selection and the resulting omitted variable bias has on the standard errors of post-Lasso and double post-Lasso (Belloni et al. 2014). Their recommendation, which is to use instead the high-dimensional OLS method of Cattaneo et al. (2018), cannot be used here because by construction the number of regressors in (3.5) is larger than n, ruling out even high-dimensional OLS. Moreover, our estimation procedure, which is Lasso-based, enables us to improve on existing Lasso ESF methods, such as Seya et al. (2015).

In order to conduct inference over $\boldsymbol{\beta}$, conditional on the nuisance parameter $\boldsymbol{\gamma}$, we propose an alternative way to calculate standard errors, based on a partial regression framework similar to that of Chernozhukov et al. (2015). The procedure uses the Lasso estimate of $\boldsymbol{\gamma}$ and $\boldsymbol{E}_\Omega$ (the Lasso selected eigenvectors) denoted $\hat{\boldsymbol{\gamma}}_L$ and $\boldsymbol{E}_L$. Specifically, we propose using the estimate of an OLS regression of $\bar{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{E}_L\hat{\gamma}_L$ on $\boldsymbol{M}_E\boldsymbol{X}$ where $\boldsymbol{M}_E := \boldsymbol{I} - \boldsymbol{E}_L(\boldsymbol{E}_L'\boldsymbol{E}_L)^{-1}\boldsymbol{E}_L'$. This corresponds to a partial regression framework, where the Lasso residuals of $\boldsymbol{y}$ on the eigenvectors $\boldsymbol{E}$ are regressed on the residuals of an OLS regression of $\boldsymbol{X}$ on the selected eigenvectors. The difference, namely the lack of selection in the regression of $\boldsymbol{X}$ on $\boldsymbol{E}$, ensures that the resulting partial regression estimator produces the same point estimate as Mi-pLasso (4.14), i.e. $\hat{\boldsymbol{\beta}}_{pr} = \hat{\boldsymbol{\beta}}_{pL}$.[10] The correspondence of the two point estimators means that the theoretical properties derived below carry across, while also ensuring that the procedure provides appropriate coverage, as confirmed in the Section 5 simulations.

In order to focus the theoretical analysis on the parameter vector $\boldsymbol{\gamma}$, we use the FWL partial regression theorem to partial out the $\boldsymbol{X}$ matrix. Yamada (2017) show that the FWL theorem could be used in a low-dimensional Lasso setting. Lemma 4.1 shows that the FWL theorem can also be applied to the high-dimensional case of Mi-Lasso.

Lemma 4.1. *Consider the following two Lasso regressions:*

$$[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}] = \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{E}\boldsymbol{\gamma}||_2^2 + \frac{1}{Z^2}||\boldsymbol{\gamma}||_1\}, \tag{4.15}$$

$$[\tilde{\boldsymbol{\gamma}}] = \min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \{||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{E}}\gamma||_2^2 + \frac{1}{Z^2}||\boldsymbol{\gamma}||_1\}, \tag{4.16}$$

where $\boldsymbol{X}$ is an $n \times k$ matrix, $\boldsymbol{E}$ is an $n \times n$ matrix, $\tilde{\boldsymbol{y}} = \boldsymbol{M_X}\boldsymbol{y}$, $\tilde{\boldsymbol{E}} = \boldsymbol{M_X}\boldsymbol{E}$ with $\boldsymbol{M_X} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Then if Assumptions 2.1 and 3.1 holds $\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}}$.

The proof is provided in the supplementary material.

We 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 complement set and the $n \times q$ matrix $\tilde{\boldsymbol{E}}_{\grave{\Omega}}$ is defined accordingly with $q_n = q = s - n$. The block-wise (re-scaled) Gram matrix $\boldsymbol{C}_n = \boldsymbol{C} = n^{-1}\tilde{\boldsymbol{E}}'\tilde{\boldsymbol{E}}$ is thus:

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

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

4.2. 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 shown by Bickel et al. (2009) for the case of Lasso, the following restricted eigenvalue (RE) condition only requires the appropriate sub-matrix of the Gram matrix to have positive and finite eigenvalues.

Assumption 4.1. (Restricted Eigenvalue) *Let $\bar{b}$ and t be positive constants and Ω*

[10] This due to the idempotent property of $\boldsymbol{M}_E$, combined with the fact that by construction $\boldsymbol{M}_E\boldsymbol{E}_L$ produces a null matrix.

denote the active set. Then the restricted eigenvalue condition holds for $\tilde{\boldsymbol{E}}$*, as* $n \to \infty$ *if we assume:*

$$\tau_{min} := \min_{\mathcal{C}(\Omega,\bar{b})} \frac{||\tilde{\boldsymbol{E}}\boldsymbol{\Delta}||_2}{\sqrt{n}||\boldsymbol{\Delta}||_2} \geq t > 0, \tag{4.17}$$

where

$$\mathcal{C}(\Omega,\bar{b}) = \{\boldsymbol{\Delta} \in \mathbb{R}^n : ||\boldsymbol{\Delta}_{\grave{\Omega}}||_1 \leq \bar{b}||\boldsymbol{\Delta}_{\Omega}||_1, \;\; \Delta \neq 0\} \tag{4.18}$$

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

Assumption 4.1 requires that $\boldsymbol{\Delta}$ lies within the restricted set (4.18). 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 2.1 and 3.1 with the RE condition, and treating $\boldsymbol{X}$ and $\boldsymbol{E}$ as constants (realisations) we 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 4.1. *Suppose Assumption 2.1-3.1 and Assumption 4.1 holds for* $\bar{b} = \frac{b+1}{b-1}$ *for some* $b > 1$ *and the regularization parameter satisfies* $\frac{1}{Z^2} \geq 2b\sqrt{\frac{4\sigma_v^2 \log n}{n}}$ *with probability tending to one as* $n \to \infty$*, then:*

$$||\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 \leq \frac{\left(\frac{1}{b}+1\right)s}{\tau_{min}^2 Z^2 n}, \tag{4.19}$$

$$||\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_2 \leq \frac{\left(\frac{1}{b}+1\right)\sqrt{s}}{\tau_{min}^2 Z^2 n}, \tag{4.20}$$

$$\frac{1}{\sqrt{n}}||\tilde{\boldsymbol{E}}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)||_2 \leq \frac{\left(\frac{1}{b}+1\right)\sqrt{s}}{\tau_{min} Z^2 n}. \tag{4.21}$$

The proof is provided in the supplementary material.

The three convergence rates presented in Theorem 4.1 depend on the number of eigenvectors with non-zero coefficients, the sample size, and Z. They also require that the tuning parameter dominates the noise of the model. By assuming the errors are sub-Gaussian (Assumption 2.1.3) we prove the probability of this event occurring goes to one as $n \to \infty$ (see proof for further details).

Corollary 4.1. *If the condition of Theorem 4.1 are satisfied and* $s/Z^2 n = o_p(1)$ *then the bounds* (4.19)-(4.21) *are* $o_p(1)$ *as* $n \to \infty$.

Corollary 4.1 is satisfied if $Z = O_p(1)$, which is reasonable as Z is a measure of correlation, and s grows at a rate slower than n, which is satisfied by Assumption 4.2.4 below.

4.3. Consistent Eigenvector Selection

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

Definition 4.1. *(Zhao and Yu 2006) Mi-Lasso estimates $\tilde{\boldsymbol{\gamma}}$ 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 4.2. (Selection Consistency) *There exists $M_1, M_2, M_3 > 0$, $0 \leq c_1 < c_2 \leq 1$ and a vector of positive constants $\boldsymbol{\nu}$, such that 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 $|\boldsymbol{C}_{\grave{\Omega}\Omega}(\boldsymbol{C}_{\Omega\Omega})^{-1}\,\text{sign}(\boldsymbol{\gamma}_\Omega)| \leq \mathbf{1} - \boldsymbol{\nu}.$

Assumption 4.2.1 is a normalisation of the transformed eigenvectors. Assumption 4.2.2 bounds the eigenvalue of the eigenvectors with non-zero coefficients from below, so the inverse of $\boldsymbol{C}_{\Omega\Omega}$ is well behaved. Assumption 4.2.3 and Assumption 4.2.4 are required as they ensure convergence in the high dimensional space as $n \to \infty$. Assumption 4.2.3 ensures 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, which aggregate at a rate of $n^{-1/2}$. Assumption 4.2.4 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, preventing the Lasso estimation bias from dominating the model parameters. Assumption 4.2.4 is a stronger sparsity assumption than Assumption 3.1.2. Assumption 4.2.5 (assuming $\boldsymbol{C}_{\Omega\Omega}$ is invertible) is 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. The columns of $\tilde{\boldsymbol{E}}$ may not be, however, 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 $\boldsymbol{\gamma}_0$.

Theorem 4.2. *Assuming Assumption 2.1, 3.1 and 4.2 hold, and $c_2 - c_1 = 0.5$. Given $s+q = n$ implies Mi-Lasso is sign consistent for all $\frac{1}{Z^2}$ that satisfy $\frac{1}{Z^2\sqrt{n}} = o_p(n^{\frac{c_2-c_1}{2}}) = o_p(n^{\frac{1}{4}})$ and $\frac{1}{n^3 Z^8} \to \infty$, we have*

$$\mathrm{P}\left(\tilde{\gamma} =_s \gamma_0\right) \geq 1 - O(n^3 Z^8) \to 1 \quad \textit{as } n \to \infty.$$

The proof is provided in the supplementary material.

Theorem 4.2 shows that Mi-Lasso is consistent in selecting the true model if the 4th moment of the errors is finite (Assumptions 2.1.3), Assumptions 2.1-4.2 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 that 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).

Throughout the paper we have specifically used $\theta = Z^{-2}$ as the penalty parameter in the Mi-Lasso method. However, the theoretical results obtained above can hold for a wider range of inverse transformations. Assuming a more general regularization parameter $\theta_a = |Z|^{-a}$, the condition in Theorem 4.1 becomes:

$$\frac{1}{|Z|^a} \geq 2b\sqrt{\frac{4\sigma_{\boldsymbol{v}}^2 \log n}{n}} \tag{4.22}$$

Similarly, the conditions from Theorem 4.2 become:

$$\begin{cases} \dfrac{1}{|Z|^a\sqrt{n}} \;=\; o_p(n^{\frac{c_2-c_1}{2}}) \;=\; o_p(n^{\frac{1}{4}}) \\ \dfrac{1}{n^3|Z|^{4a}} \;\to\; \infty \end{cases} \tag{4.23}$$

Taken together, conditions (4.22) and (4.23) provide the following bounds on the transformation exponent a, where c_n is a sequence converging to 0:

$$\log\left(2b\sqrt{\frac{4\sigma_{\boldsymbol{v}}^2 \log n}{n}}\right) \;\leq\; a\log\left(\frac{1}{|Z|}\right) \;\leq\; c_n\left(\log\left(n^{\frac{c_2-c_1+1}{2}}\right)\right). \tag{4.24}$$

Based on a sensitivity analysis that was run on these bounds for reasonable values of b, $\sigma_{\boldsymbol{v}}^2$ and n, we formulate the following remark, which motivates $a = 2$ as a pragmatic choice.[11]

Remark 4.1. *In presence of significant spatial correlation, $\theta = Z^{-2}$ has a high probability of being amongst the values of $\theta_a = |Z|^{-a}$, with $a \geq 0$, for which the assumptions of Theorem 4.1 and 4.2 are compatible.*

5. SUMMARY OF MONTE CARLO ANALYSIS

In order to evaluate the finite sample performance of Mi-Lasso and compare it to the main existing ESF procedures, we conduct two Monte Carlo exercises with a DGP given by (2.2). The set of estimators compared includes naïve OLS, which regresses $\boldsymbol{y}$ on $\boldsymbol{x}$ ignoring spatial correlation, the CV-Lasso algorithm of Seya et al. (2015), the forward stepwise algorithm of Tiefelsdorf and Griffith (2007), as well as an alternative version of algorithm 1, where post-Lasso is used to obtain the estimate of β in step 4.

In the interest of brevity, Table 1 presents a summary of the results obtained for the case where each spatial unit has an average of $\mu = 4$ connected neighbours. The tables containing the full results of the analysis are provided in the supplementary material. Overall, the findings of the analysis support the theoretical properties discussed in section 4. All ESF estimators perform well in terms of the bias on $\hat{\beta}$, clearly outperforming OLS, and Mi-Lasso provides the same point estimate as a post-Lasso estimation on the selected model, i.e. $\hat{\beta}_{pL} = \hat{\beta}_{pr}$. Instead, the first key difference between the ESF methods lies in the coverage statistics, where Mi-Lasso produces the best coverage, i.e. closest to the optimal values of 0.95 and 0.99. In particular, as discussed in section 4.1, naïvely using post-Lasso on the Lasso-selected eigenvectors leads to poor inference, specifically overly

[11] The results of this sensitivity analysis are available in the supplementary material.

n	ρ_1	Estimator	Bias	MSE	SD	AASE	CI95	CI99	$\lvert\hat{\Omega}\rvert$
100	0.3	Naïve OLS	0.019	0.012	0.109	0.104	0.934	0.989	-
		Mi-pLasso	0.010	0.013	0.113	0.100	0.918	0.977	3
		Mi-Lasso	0.010	0.013	0.113	0.106	0.938	0.986	3
		Post CV-Lasso	0.005	0.014	0.117	0.098	0.894	0.971	3
		FstepZ	0.003	0.019	0.138	0.104	0.874	0.962	13
100	0.6	Naïve OLS	0.047	0.016	0.116	0.109	0.903	0.981	-
		Mi-pLasso	0.011	0.017	0.130	0.093	0.813	0.904	13
		Mi-Lasso	0.011	0.017	0.130	0.116	0.914	0.985	13
		Post CV-Lasso	0.010	0.016	0.126	0.100	0.870	0.959	4
		FstepZ	-0.026	0.019	0.134	0.100	0.872	0.945	9
100	0.9	Naïve OLS	0.089	0.024	0.127	0.117	0.861	0.959	-
		Mi-pLasso	0.005	0.022	0.148	0.079	0.702	0.820	36
		Mi-Lasso	0.005	0.022	0.148	0.143	0.929	0.979	36
		Post CV-Lasso	0.009	0.019	0.136	0.101	0.839	0.943	5
		FstepZ	-0.056	0.023	0.141	0.098	0.807	0.913	12
250	0.3	Naïve OLS	0.016	0.004	0.063	0.065	0.951	0.991	-
		Mi-pLasso	0.012	0.005	0.066	0.064	0.937	0.984	3
		Mi-Lasso	0.012	0.005	0.066	0.065	0.949	0.990	3
		Post CV-Lasso	0.010	0.005	0.067	0.062	0.937	0.983	4
		FstepZ	0.006	0.007	0.085	0.064	0.894	0.963	19
250	0.6	Naïve OLS	0.036	0.006	0.066	0.067	0.921	0.985	-
		Mi-pLasso	0.013	0.006	0.077	0.059	0.841	0.934	25
		Mi-Lasso	0.013	0.006	0.077	0.070	0.924	0.986	25
		Post CV-Lasso	0.017	0.005	0.069	0.063	0.919	0.977	5
		FstepZ	-0.012	0.006	0.075	0.062	0.904	0.966	13
250	0.9	Naïve OLS	0.063	0.009	0.070	0.070	0.852	0.952	-
		Mi-pLasso	0.005	0.007	0.084	0.047	0.699	0.829	82
		Mi-Lasso	0.005	0.007	0.084	0.082	0.939	0.985	82
		Post CV-Lasso	0.019	0.006	0.076	0.064	0.883	0.960	9
		FstepZ	-0.029	0.007	0.080	0.062	0.847	0.945	19
500	0.3	Naïve OLS	0.012	0.002	0.048	0.046	0.930	0.976	-
		Mi-pLasso	0.008	0.003	0.049	0.045	0.917	0.967	5
		Mi-Lasso	0.008	0.003	0.049	0.046	0.924	0.974	5
		Post CV-Lasso	0.005	0.003	0.050	0.044	0.915	0.970	6
		FstepZ	-0.003	0.003	0.059	0.045	0.896	0.955	18
500	0.6	Naïve OLS	0.033	0.004	0.050	0.047	0.881	0.964	-
		Mi-pLasso	0.002	0.003	0.059	0.038	0.792	0.893	78
		Mi-Lasso	0.002	0.003	0.059	0.051	0.917	0.971	78
		Post CV-Lasso	0.014	0.003	0.054	0.045	0.889	0.964	10
		FstepZ	-0.017	0.003	0.055	0.044	0.875	0.957	21
500	0.9	Naïve OLS	0.059	0.006	0.053	0.050	0.762	0.910	-
		Mi-pLasso	-0.006	0.004	0.065	0.027	0.589	0.703	243
		Mi-Lasso	-0.006	0.004	0.065	0.064	0.938	0.978	243
		Post CV-Lasso	0.013	0.004	0.059	0.045	0.860	0.954	17
		FstepZ	-0.036	0.005	0.058	0.044	0.786	0.901	37

Note: MSE is mean squared error, SD is the standard deviation of $\hat{\beta}$, AASE is the average asymptotic standard error, CI95/99 is the coverage of the 95/99 % confidence intervals.

Table 1: Bias, MSE, SD, AASE, and 95/99 %coverage and selected eigenvectors for μ=4

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
$chas$	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

Table 2: Variables used in Boston housing application

tight confidence intervals and poor coverage. A second difference lies in the computational requirements of the methods, where Mi-Lasso outperforms the other ESF methods by at least an order of magnitude, as it does not require iterative methods to either select the eigenvectors or the Lasso penalisation parameter.

6. EMPIRICAL APPLICATION - BOSTON HOUSING DATASET

We now compare the ESF selection procedures on the Boston Housing Dataset, first used by Harrison and Rubinfeld (1978) to evaluate the relationship between house prices and demand for clean air, and later reveal by Gilley and Pace (1996) upon noting the high spatial correlation in the dataset. This provides an illustration of the kind of applied setting described in the introduction: the parameter of interest is the effect of nitric oxide concentration, but the data is spatially correlated. Given that there is no guarantee that Gilley and Pace (1996) provide the correct specification, and given that the researcher might only concerned with the direct effect of the variable, ESF is an appropriate methodology, allowing to simply control for the spatial effects.[12]

The dataset includes 508 census tracts (spatial units). Table 2 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 the SWM is normalised by the maximal of the row (or column) sum. The following basic model (excluding any eigenvectors) is:

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

Table 3 shows a sub-set of the covariate parameter estimates for naïve OLS (ignoring the spatial correlation), Mi-Lasso, post CV-Lasso, and FstepZ. These results show that the OLS estimates are biased by spatial dependence. For example, the coefficient on nox^2, dis, and rm all exhibit a downward bias in the simple-OLS case. It is important to note that the Mi-Lasso standard errors are generally larger than the other filtered estimates

[12] The data is from the R package 'spdep' (Bivand and Wong 2018).

	Dependent variable: ln(p)			
	Naïve OLS	FstepZ	Post CV-Lasso	Mi-Lasso
	(1)	(2)	(3)	(4)
zn	0.001***	0.001*	0.001***	0.0002
	(0.0004)	(0.0004)	(0.0004)	(0.001)
indus	0.002	-0.0003	0.003	0.005
	(0.002)	(0.002)	(0.002)	(0.003)
chas	0.104***	0.038	0.065***	0.059
	(0.038)	(0.038)	(0.025)	(0.056)
nox^2	-0.588***	-0.219*	-0.126	-0.220
	(0.124)	(0.125)	(0.096)	(0.177)
rm	0.091***	0.177***	0.221***	0.194***
	(0.028)	(0.032)	(0.016)	(0.019)
dis	-0.047***	-0.032***	-0.029***	-0.028**
	(0.008)	(0.007)	(0.006)	(0.011)
black	-0.003***	-0.005***	-0.005***	-0.005*
	(0.001)	(0.001)	(0.001)	(0.002)
lsp	-0.029***	-0.020***	-0.017***	-0.020***
	(0.004)	(0.003)	(0.002)	(0.004)
Adj. R^2	0.785	0.896	0.901	-
Resid. S.E.	0.189	0.132	0.129	-
d.f.	492	431	449	-
No. Eigenvectors	-	61	43	197

Note: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in parenthesis. Full results of covariate estimates can be found in the on-line supplement. The adjusted R^2, residual standard error and the degrees of freedom are omitted for Mi-Lasso as they are not comparable to the other estimator, due to the partial regression step of Algorithm 1.

Table 3: Sub-set of Parameter Estimation Results

and in some cases this changes the significance, for example zoned land (*zn*), adjancency to the Charles river (*chas*) and nitic oxide concentration (*nox*) - the parameter of interest in Harrison and Rubinfeld (1978) - are insignificant for Mi-Lasso and significant for simple-OLS and other filtered procedures. Combined with the findings of the Monte-Carlo analysis regarding coverage, this suggests that the significance obtained with existing ESF might in fact be spurious.

7. CONCLUSION AND FURTHER WORK

The main aim of this paper is to improve on existing solutions to the ESF eigenvector selection problem. Various methods currently exist, with none dominating clearly in practice. Both the forward-iterative procedures with a user-defined cut-off (Tiefelsdorf and Griffith 2007) and the CV-Lasso method (Seya et al. 2015) are relatively slow, especially as sample size increases. Furthermore, these methods can be seen as ad-hoc, as there is a lack of theoretical results in the ESF literature. In the case of CV-Lasso, for instance, the methodology aims to maximise prediction accuracy, while the goal of ESF is to reduce bias on the parameter of interest β.

This paper starts by formalising the assumptions that underpin ESF, and on the basis of these proposes an alternative Lasso-based procedure called Moran's I Lasso (Mi-Lasso). Rather than use CV to obtain the Lasso penalisation parameter θ, this method uses information about the level of spatial correlation in the naïve regression residuals, measured using Moran's I, to determine a point estimate for θ. In a second step, once the eigenvectors have been selected, it uses a partial regression framework to obtain reliable inference on β. The key benefits of the method are that it is intuitive, theoretically grounded, provides good coverage of the sampling distribution of $\hat{\beta}$, and is substantially faster than stepwise procedures or the CV Lasso of Seya et al. (2015). We have derived performance bounds for the Mi-Lasso estimates of the eigenvector coefficients and shown the conditions necessary for the estimator to provide consistent eigenvector selection. Our simulation results confirm that the estimator performs well in terms of bias, MSE and coverage compared to existing selection procedures for a range of levels of spatial correlation and in an empirical application on house prices.

While this paper provides an important first step in formalising and improving ESF methods, several limitations remain, providing direction for future work. First, it relies on strong sparsity and spectral assumptions for the reduced form to hold, in particular that the SWM used in the estimation is the same as the one in the DGP. This is unrealistic in practice, especially when considering that empirical SWMs will almost never correspond to the exact connectivity of the (unknown) DGP, and therefore warrants a relaxation to approximate sparsity and spectral assumptions. Second, while the simulations carried out suggest that the partial regression framework used in the final step performs well, it falls short of a formal normality proof for the distribution of the resulting estimator, which calls for further investigation in this direction.

8. ACKNOWLEDGEMENTS

We would like to thank Dennis Kristensen (co-editor) and the referees for helpful comments that led to a improved version of the paper. We are also thankful to Hans-Martin Krolzig, Abhimanyu Gupta, Jorge Mateu, Pedro CL Souza, Nicolas Debarsy and Maria Kyriacou for useful discussions and suggestions.

REFERENCES

Anselin, L. and S. Rey (1991). Properties of tests for spatial dependence in linear regression models. *Geographical analysis 23*(2), 112–131.

Anselin, L. and O. Smirnov (1996). Efficient algorithms for constructing proper higher order spatial lag operators. *Journal of Regional Science 36*(1), 67–89.

Aum, S., S. Y. Lee, and Y. Shin (2021). Covid-19 doesn't need lockdowns to destroy jobs: The effect of local outbreaks in korea. *Labour Economics 70*, 101993.

Bargain, O. and U. Aminjonov (2020). Trust and compliance to public health policies in times of covid-19. *Journal of Public Economics 192*, 104316.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies 81*(2), 608–650.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics 37*(4), 1705–1732.

Bivand, R. and D. W. S. Wong (2018). Comparing implementations of global and local indicators of spatial association. *TEST 27*(3), 716–748.

Blommestein, H. J. (1985). Elimination of circular routes in spatial dynamic regression equations. *Regional Science and Urban Economics 15*(1), 121–130.

Boots, B. and M. Tiefelsdorf (2000). Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems 2*(4), 319–348.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association 113*(523), 1350–1361.

Chernozhukov, V., C. Hansen, and M. Spindler (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review 105*(5), 486–90.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2020). On cross-validated lasso in high dimensions. *Annals of Statistics 40*.

Chun, Y., D. A. Griffith, M. Lee, and P. Sinha (2016). Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. *Journal of Geographical Systems 18*(1), 67–85.

Crespo Cuaresma, J. and M. Feldkircher (2013). Spatial filtering, model uncertainty and the speed of income convergence in europe. *Journal of Applied Econometrics 28*(4), 720–741.

Csereklyei, Z. and D. I. Stern (2015). Global energy use: decoupling or convergence? *Energy Economics 51*, 633–641.

De Jong, P., C. Sprenger, and F. Van Veen (1984). On extreme values of moran's i and geary's c. *Geographical Analysis 16*(1), 17–24.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

Gilley, O. W. and R. Pace (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.

Harrison, D. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management 5*(1), 81–102.

Kelejian, H. and G. Piras (2017). *Spatial econometrics*. Academic Press.

Kelejian, H. H. and I. R. Prucha (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 I. R. Prucha (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics 157*(1), 53 – 67.

Kourtellos, A., A. Lenkoski, and K. Petrou (2020). Measuring the strength of the theories of government size. *Empirical Economics 59*, 2185–2222.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica 72*(6), 1899–1925.

Leeb, H. and B. M. Pötscher (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory 24*(2), 338–376.

LeSage, J. P. and R. K. Pace (2014). The biggest myth in spatial econometrics. *Econometrics 2*(4), 217–249.

Li, T., E. Levina, and J. Zhu (2020). Network cross-validation by edge sampling. *Biometrika 107*(2), 257–276.

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika 37*(1-2), 17–23.

Oberdabernig, D. A., S. Humer, and J. Crespo Cuaresma (2018). Democracy, geography and model uncertainty. *Scottish Journal of Political Economy 65*(2), 154–185.

Pace, R. K., J. P. LeSage, and S. Zhu (2013). Interpretation and computation of estimates from regression models using spatial filtering. *Spatial Economic Analysis 8*(3), 352–369.

Patuelli, R., N. Schanne, D. A. Griffith, and P. Nijkamp (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.

Sanso-Navarro, M., F. Sanz Gracia, and M. Vera-Cabello (2023). Terrorism determinants, model uncertainty and space in colombia. *Defence and Peace Economics 34*(1), 92–111.

Seya, H., D. Murakami, M. Tsutsumi, and Y. Yamagata (2015). Application of lasso to the eigenvector selection problem in eigenvector-based spatial filtering. *Geographical Analysis 47*(3), 284–299.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Tiefelsdorf, M. and B. Boots (1995). The exact distribution of moran's i. *Environment and Planning A: Economy and Space 27*(6), 985–999.

Tiefelsdorf, M. and D. A. Griffith (2007). Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning A 39*(5), 1193–1221.

Wüthrich, K. and Y. Zhu (2023). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *Review of Economics and Statistics 105*(4), 982–997.

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 B. Yu (2006). On model selection consistency of lasso. *Journal of Machine learning research 7*(Nov), 2541–2563.