Eigenvector-Based Semiparametric Filtering of Spatial Autocorrelation in Regression Models

Sebastian Juhl

University of Mannheim

March 17, 2021

Motivation

- spatial autocorrelation can lead to inefficient or even biased and inconsistent parameter estimates in regression models
- social scientists primarily address this issue in one of two ways:
  1. ignore it altogether and assume *iid* observations
  2. apply (parametric) spatial regression models

- semiparametric filtering techniques can offer an attractive alternative
  - ease of estimation (standard OLS or ML estimators)
  - straightforward interpretation
  - accounts for spatial autocorrelation at different scales/resolutions
  - easily adaptable to GLMs

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

Spatial Misspecification in Linear Models

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

Spatial Error Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \underbrace{(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\epsilon}}_{\boldsymbol{e}_{SEM}}$$

Spatial-X Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \underbrace{\rho\boldsymbol{W}\boldsymbol{X} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SLX}}$$

Spatial Lag Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \underbrace{\rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SAR}}$$

$$\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

Spatial Misspecification in Linear Models

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

Spatial Error Model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \underbrace{(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\epsilon}}_{\boldsymbol{e}_{SEM}}$$

Spatial-X Model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \underbrace{\rho\boldsymbol{W}\boldsymbol{X} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SLX}}$$

Spatial Lag Model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \underbrace{\rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SAR}}$$

$$\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}(\boldsymbol{X\beta} + \boldsymbol{\epsilon})$$

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

$$\boldsymbol{y} = \boldsymbol{X\beta} \left.\begin{array}{l} +\rho\boldsymbol{W}\boldsymbol{\epsilon} \\ +\rho^2\boldsymbol{W}^2\boldsymbol{\epsilon} \\ +... + \boldsymbol{\epsilon} \end{array}\right\} \boldsymbol{\epsilon}_{SEM}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} \left.\begin{array}{l} +\rho\boldsymbol{W}(\boldsymbol{X\beta} + \boldsymbol{\epsilon}) \\ +\rho^2\boldsymbol{W}^2(\boldsymbol{X\beta} + \boldsymbol{\epsilon}) \\ +... + \boldsymbol{\epsilon} \end{array}\right\} \boldsymbol{\epsilon}_{SAR}$$

Spatial Misspecification in Linear Models

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

Spatial Error Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \underbrace{(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}\boldsymbol{\epsilon}}_{\boldsymbol{e}_{SEM}}$$

Spatial-X Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \underbrace{\rho\boldsymbol{W}\boldsymbol{X} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SLX}}$$

Spatial Lag Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \underbrace{\rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SAR}}$$

$$\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \left.\begin{aligned} &+\rho\boldsymbol{W}\boldsymbol{\epsilon} \\ &+\rho^2\boldsymbol{W}^2\boldsymbol{\epsilon} \\ &+\ldots + \boldsymbol{\epsilon} \end{aligned}\right\} \boldsymbol{\epsilon}_{SEM}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \left.\begin{aligned} &+\rho\boldsymbol{W}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &+\rho^2\boldsymbol{W}^2(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &+\ldots + \boldsymbol{\epsilon} \end{aligned}\right\} \boldsymbol{\epsilon}_{SAR}$$

$\underline{\boldsymbol{e}_{SEM}}$

$$\underbrace{\sum_{r=1}^{\infty} \rho^r \boldsymbol{W}^r \boldsymbol{\epsilon}}_{\text{spatial part}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

$\underline{\boldsymbol{e}_{SLX}}$

$$\underbrace{\rho\boldsymbol{W}\boldsymbol{X}}_{\text{spatial part}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

$\underline{\boldsymbol{e}_{SAR}}$

$$\underbrace{\sum_{r=1}^{\infty} \rho^r \boldsymbol{W}^r (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})}_{\text{spatial part}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}$$

Intuition:

ESF partitions the response variable into i) a systematic trend, ii) a stochastic spatial signal, and iii) *iid* disturbances

Example: SAR DGP

$$\boldsymbol{y}$$

(outcome)

$$\boldsymbol{X}\boldsymbol{\beta} \qquad \sum_{r=1}^{\infty} \rho^r \boldsymbol{W}^r (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) + \epsilon$$

(systematic trend) (autocorrelated part)

$$\sum_{r=1}^{\infty} \rho^r \boldsymbol{W}^r (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \qquad \epsilon$$

(stochastic spatial signal) (*iid* disturbances)

Analytical Description of ESF I

UNIVERSITY OF MANNHEIM

ESF technique utilizes the spectral decomposition of a centered and symmetrized connectivity matrix: $\boldsymbol{V} = \frac{1}{2}(\boldsymbol{W} + \boldsymbol{W}')$:

$$\boldsymbol{MVM} = \boldsymbol{E\Lambda E}^{-1} = \boldsymbol{E\Lambda E}'$$

- demeaning projector $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{11}'/n)$ ensures that all eigenvectors are orthogonal und uncorrelated
- asymptotically, the eigenfunctions obtained from $\boldsymbol{MVM}$ converge to those of matrix $\boldsymbol{V}$
- $\boldsymbol{E}$ are all n eigenvectors
- $\boldsymbol{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues $\boldsymbol{\lambda}$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \underbrace{\rho\boldsymbol{Wy} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SAR}}$$

$$\Rightarrow \boldsymbol{y} = \boldsymbol{X\beta} + \underbrace{\rho\boldsymbol{E\Lambda E}'\boldsymbol{y} + \boldsymbol{\epsilon}}_{\boldsymbol{e}_{SAR}}$$

Analytical Description of ESF II

$$\boldsymbol{y} = \boldsymbol{X\beta} + \rho \boldsymbol{E\Lambda} \underbrace{\boldsymbol{E'y}}_{\boldsymbol{g}} + \boldsymbol{\epsilon}$$

- $\boldsymbol{g}$ is the OLS estimator of $\boldsymbol{y}$ regressed on $\boldsymbol{E}$

  $(\boldsymbol{E'E})^{-1}\boldsymbol{Ey} \Rightarrow \boldsymbol{Ey}$ since $\boldsymbol{E'E} = \boldsymbol{I}$

Analytical Description of ESF II

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \rho \boldsymbol{E}\boldsymbol{\Lambda} \underbrace{\boldsymbol{E}'\boldsymbol{y}}_{\boldsymbol{g}} + \boldsymbol{\epsilon}$$

- $\boldsymbol{g}$ is the OLS estimator of $\boldsymbol{y}$ regressed on $\boldsymbol{E}$

  $(\boldsymbol{E}'\boldsymbol{E})^{-1}\boldsymbol{E}\boldsymbol{y} \Rightarrow \boldsymbol{E}\boldsymbol{y}$ since $\boldsymbol{E}'\boldsymbol{E} = \boldsymbol{I}$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \rho \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{g} + \boldsymbol{\epsilon}$$

- $\rho\boldsymbol{E}\boldsymbol{\Lambda}$: $\rho\boldsymbol{\Lambda}$ essentially rescales $\boldsymbol{E}$

  $\Rightarrow \boldsymbol{\gamma} = \rho\boldsymbol{\Lambda}\boldsymbol{g}$

Analytical Description of ESF II

$$\boldsymbol{y} = \boldsymbol{X\beta} + \rho \boldsymbol{E\Lambda} \underbrace{\boldsymbol{E'y}}_{\boldsymbol{g}} + \boldsymbol{\epsilon}$$

- $\boldsymbol{g}$ is the OLS estimator of $\boldsymbol{y}$ regressed on $\boldsymbol{E}$

  $(\boldsymbol{E'E})^{-1}\boldsymbol{Ey} \Rightarrow \boldsymbol{Ey}$ since $\boldsymbol{E'E} = \boldsymbol{I}$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \rho \boldsymbol{E\Lambda g} + \boldsymbol{\epsilon}$$

- $\rho\boldsymbol{E\Lambda}$: $\rho\boldsymbol{\Lambda}$ essentially rescales $\boldsymbol{E}$

  $\Rightarrow \boldsymbol{\gamma} = \rho\boldsymbol{\Lambda g}$

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \rho \boldsymbol{E\Lambda E'y} + \boldsymbol{\epsilon} \\ &= \boldsymbol{X\beta} + \boldsymbol{E\gamma} + \boldsymbol{\epsilon} \end{aligned}$$

However, we cannot include all n eigenvectors as regressors!

Constructing the Spatial Filter

- the set of $\boldsymbol{E}$ depict all possible patterns of autocorrelation permitted by the connectivity matrix $\boldsymbol{W}$
- stepwise regression techniques help to select the most relevant patterns $\boldsymbol{E}^*$
- thereby, we can craft a synthetic variable – the spatial filter

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\ &\approx \boldsymbol{X}\boldsymbol{\beta} + \underbrace{\boldsymbol{E}^*\boldsymbol{\gamma}}_{filter} + \underbrace{\boldsymbol{\epsilon}}_{noise} \end{aligned}$$

- different criteria can be used to select eigenvectors, e.g.:
  - model fit
  - significance
  - spatial autocorrelation

Pros and Cons of ESF

Pros

+ ease of model estimation
+ straightforward interpretation of parameters
+ flexibility
  (no need to specify the spatial pattern in each variable)
+ generalizability
  (also applicable to GLMs – with some modifications)

Cons

- "removes" indirect spillovers
- computationally demanding for large N
- over- or undercorrection of spatial autocorrelation possible

(Preliminary)

Monte Carlo Evidence

Reduction in Residual SA