Eigenvector-Based Semiparametric Filtering of Spatial Autocorrelation in Regression Models

Sebastian Juhl

University of Mannheim

March 17, 2021

Motivation

UNIVERSITY OF MANNHEIM

- spatial misspecification can lead to inefficient or even biased and inconsistent parameter estimates in regression models
- social scientists primarily deal with a spatial misspecification problem 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

Spatial Misspecification in Linear Models

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

Spatial Misspecification in Linear Models

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

Spatial Error DGP

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

Spatial-X DGP

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

Spatial Lag DGP

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

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

Spatial Misspecification in Linear Models

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

Spatial Error DGP

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

Spatial-X DGP

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

Spatial Lag DGP

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

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

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

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

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

Spatial Misspecification in Linear Models

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

Spatial Error DGP

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

Spatial-X DGP

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

Spatial Lag DGP

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

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

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

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

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

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

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

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

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

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

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

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)

Spatial Eigenfunction Analysis

ESF is based on the spectral decomposition of a centered (and symmetric/ symmetrized) connectivity matrix: $\boldsymbol{W}$:

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

- demeaning projector $\boldsymbol{M} = (\boldsymbol{I} - \mathbf{1}\mathbf{1}'/n)$ – also ensures that all eigenvectors are orthogonal and uncorrelated
- $\boldsymbol{E}$ are all n eigenvectors
- $\boldsymbol{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues $\boldsymbol{\lambda}$

Spatial Eigenfunction Analysis

ESF is based on the spectral decomposition of a centered (and symmetric/ symmetrized) connectivity matrix: $\boldsymbol{W}$:

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

- demeaning projector $\boldsymbol{M} = (\boldsymbol{I} - \mathbf{1}\mathbf{1}'/n)$ – also ensures that all eigenvectors are orthogonal and uncorrelated
- $\boldsymbol{E}$ are all n eigenvectors
- $\boldsymbol{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues $\boldsymbol{\lambda}$

Direct relationship to the numerator of the global Moran coefficient:

$$MC(\boldsymbol{x}) = \frac{n}{\mathbf{1}'\boldsymbol{W}\mathbf{1}} \frac{\boldsymbol{x}'\boldsymbol{MWMx}}{\boldsymbol{x}'\boldsymbol{Mx}}$$

Eigenvectors as Synthetic Map Patterns

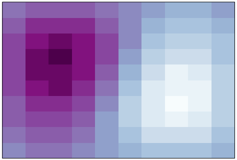
UNIVERSITY OF MANNHEIM

- eigenvectors $\boldsymbol{E}$ depict distinct – and uncorrelated – synthetic map patterns
- corresponding eigenvalues indicate the level of SA

Eigenvectors as Synthetic Map Patterns

- eigenvectors $\boldsymbol{E}$ depict distinct – and uncorrelated – synthetic map patterns
- corresponding eigenvalues indicate the level of SA

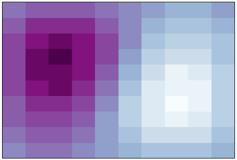
Eigenvector 1
(MC = 1)



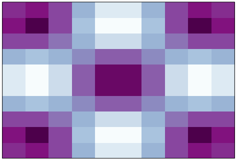
Eigenvectors as Synthetic Map Patterns

- eigenvectors $\boldsymbol{E}$ depict distinct – and uncorrelated – synthetic map patterns
- corresponding eigenvalues indicate the level of SA

Eigenvector 1
(MC = 1)



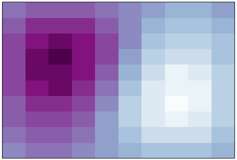
Eigenvector 10
(MC = 0.73)



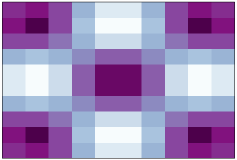
Eigenvectors as Synthetic Map Patterns

- eigenvectors $\boldsymbol{E}$ depict distinct – and uncorrelated – synthetic map patterns
- corresponding eigenvalues indicate the level of SA

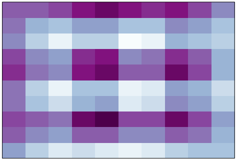
Eigenvector 1
(MC = 1)



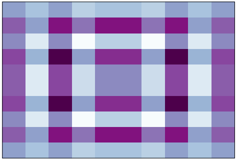
Eigenvector 10
(MC = 0.73)



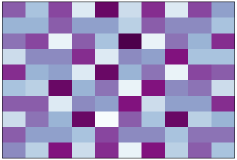
Eigenvector 20
(MC = 0.454)



Eigenvector 44
(MC = 0.066)



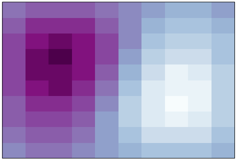
Eigenvector 80
(MC = -0.454)



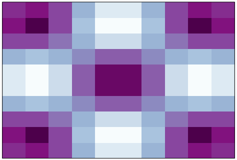
Eigenvectors as Synthetic Map Patterns

- eigenvectors $\boldsymbol{E}$ depict distinct – and uncorrelated – synthetic map patterns
- corresponding eigenvalues indicate the level of SA

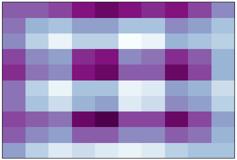
Eigenvector 1
(MC = 1)



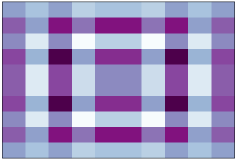
Eigenvector 10
(MC = 0.73)



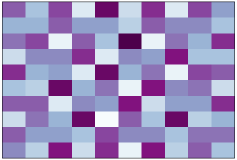
Eigenvector 20
(MC = 0.454)



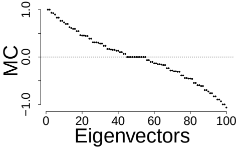
Eigenvector 44
(MC = 0.066)



Eigenvector 80
(MC = -0.454)

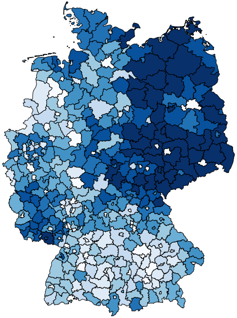


MCs of all
Eigenvectors



- $\boldsymbol{E}$ depict all possible spatial patterns permitted by $\boldsymbol{W}$
- more complex patterns can be obtained by a linear combination of eigenvectors

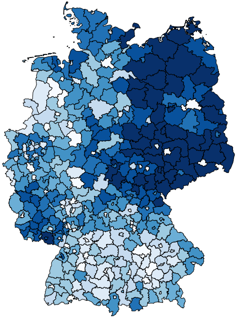
Median Age



MC = 0.404 ($p = 0$)

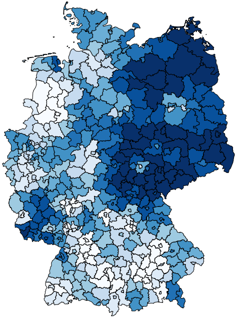
- $\boldsymbol{E}$ depict all possible spatial patterns permitted by $\boldsymbol{W}$
- more complex patterns can be obtained by a linear combination of eigenvectors

Median Age



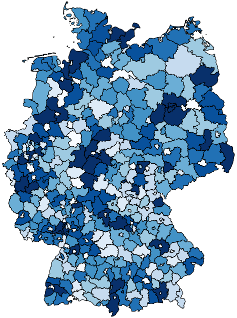
MC = 0.404 (p = 0)

Spatial Filter



MC = 0.404 (p = 0)

Filtered Residuals



MC = -0.089 (p = 0.463)

Eigenvector Selection for the Spatial Filter

However, it is impossible to include all n eigenvectors as regressors

1. Identification of a candidate set $\boldsymbol{E}^C \subset \boldsymbol{E}$ based on
   - sign of SA
   - strength of SA
2. select relevant map patterns $\boldsymbol{E}^* \subset \boldsymbol{E}^C$ using supervised or unsupervised stepwise regression

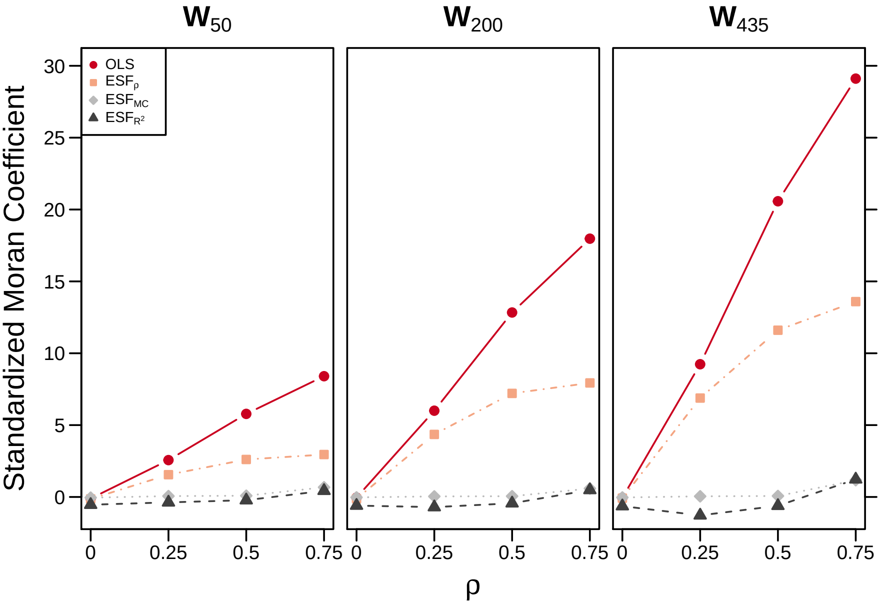
$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X\beta} + \overbrace{\boldsymbol{E\gamma} + \boldsymbol{\epsilon}}^{\boldsymbol{e}_{OLS}} \\ &\approx \boldsymbol{X\beta} + \underbrace{\boldsymbol{E}^*\boldsymbol{\gamma}}_{filter} + \underbrace{\boldsymbol{\epsilon}}_{noise}\end{aligned}$$

- different criteria can be used to select eigenvectors, e.g.:
  - maximization of model fit (e.g., AIC, BIC, R^2)
  - significance level of eigenvectors
  - minimization of residual SA

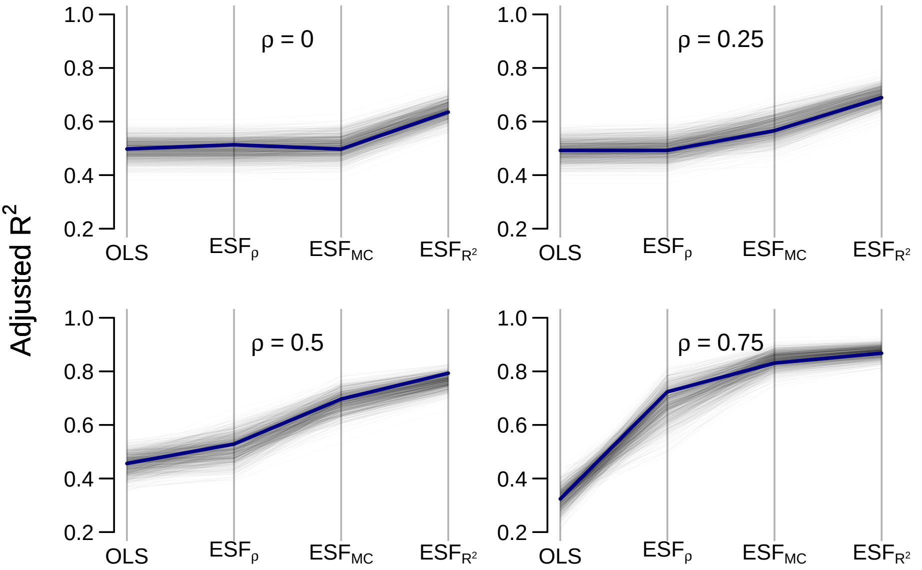
(Preliminary)

Monte Carlo Evidence

Reduction in Residual SA



Improved Model Fit



Conclusion

- ESF offers numerous advantages for political scientists
- however, applicability depends on RQ (as always!)
- things to consider:

<u>Pros</u>

+ 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)

<u>Cons</u>

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

Additional Resources

- `spfilteR` package:
  CRAN: https://CRAN.R-project.org/package=spfilteR
  GitHub: https://github.com/sjuhl/spfilteR
- further projects & working papers
  1. introductory paper on the `spfilteR` package (under review)
  2. project on Moran eigenvector maps and spatial eigenfunction analysis

Feedback & suggestions are highly appreciated!

`sebastian.juhl@gess.uni-mannheim.de`
www.sebastianjuhl.com