Local variable selection and parameter estimation for spatially varying coefficient models

Wesley Brooks

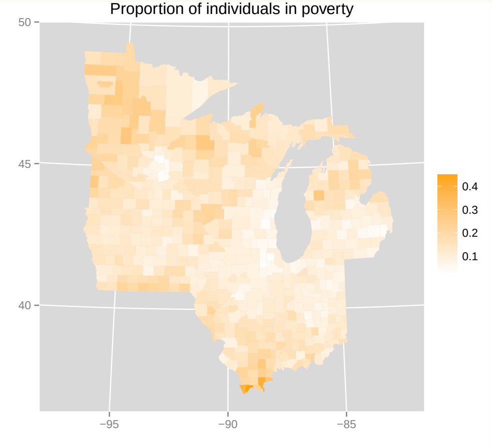
Department of Statistics
University of Wisconsin–Madison

January 16, 2014

Motivation

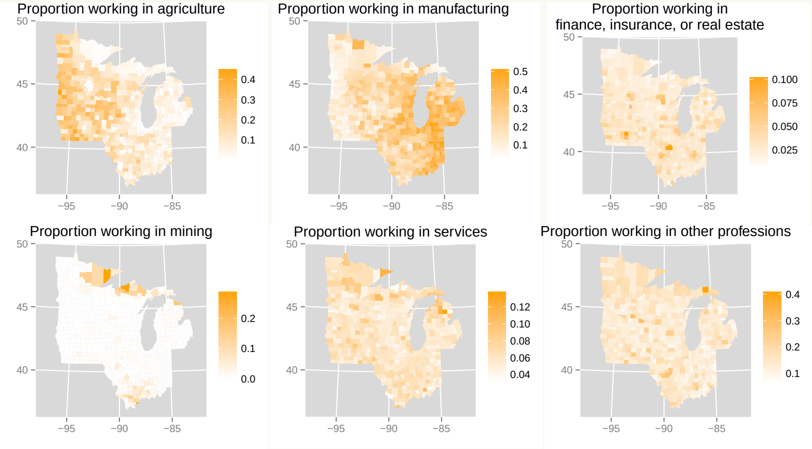
Motivation

Response variable



Motivation

Covariates



Motivation

Scientific questions

- Which of the economic-structure variables is associated with poverty rate?
- What are the sign and magnitude of that association?
- Is poverty rate associated with the same economic-structure variables across the entire region?
- How do the sign and magnitude of the associations vary across the region?

Introduction

Introduction

An overview

- Spatial regression
- Varying coefficient regression
  - Splines
  - Kernels
  - Wavelets
- Model selection via regularization

Introduction

Definitions

- Univariate spatial response process $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$
- Multivariate spatial covariate process $\{\boldsymbol{X}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$
- $n =$ number of observations
- $p =$ number of covariates
- Location (2-dimensional) $\boldsymbol{s}$
- Spatial domain $\mathcal{D}$

Introduction

Spatial linear regression (Cressie, 1993)

- A typical spatial linear regression model

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta} + W(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$$

- $W(\boldsymbol{s})$ is a spatial random effect that accounts for autocorrelation in the response variable
- $\varepsilon(\boldsymbol{s})$ is iid random noise
- The coefficients $\boldsymbol{\beta} = (1, \beta_1, \dots, \beta_p)$ are constant
- Requires *a priori* global variable selection

Introduction

Spatially varying coefficient model (Gelfand *et al.*, 2003)

- A more flexible model: coefficients in a spatial regression model can vary

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$$

- $\{\beta_0(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}, \ldots, \{\beta_p(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ are stationary spatial processes
- Requires *a priori* global variable selection

Introduction

Varying coefficients regression (VCR) (Hastie and Tibshirani, 1993)

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$$

- Assume an effect modifying variable $\boldsymbol{s}$
- Coefficients are functions of $\boldsymbol{s}$

Introduction

Spline-based VCR models (Wood, 2006)

- Splines are a way to parameterize smooth functions
- Estimate the varying coefficients via splines:

$$E\{Y(\boldsymbol{s})\} = \beta_1(\boldsymbol{s})X_1(\boldsymbol{s}) + \cdots + \beta_p(\boldsymbol{s})X_p(\boldsymbol{s})$$

Introduction

Global selection in spline-based VCR models

Regularization methods for global variable selection in VCR models:

- The $\mathcal{L}_2$ norm of a function (e.g. $\int \{f(t)\}^2 dt$) is zero if and only if the function is zero everywhere.
- Use regularization to encourage coefficient functions to be zero
  - SCAD penalty (Wang *et al.*, 2008a)
  - Non-negative garrote penalty (Antoniadis *et al.*, 2012b)

Introduction

Wavelet methods for VCR models

- Wavelet methods: decompose coefficient function into local frequency components
- Selection of nonzero local frequency components with nonzero coefficients:
  - Bayesian variable selection (Shang, 2011)
  - Lasso (Zhang and Clayton, 2011)
- Sparsity in the local frequency components; not in the local covariates

Geographically weighted regression

Geographically weighted regression

Brundson *et al.* (1998), Fotheringham *et al.* (2002)

- Consider observations at sampling locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$
- $y(\boldsymbol{s}_i) = y_i$ the univariate response at location $\boldsymbol{s}_i$
- $\boldsymbol{x}(\boldsymbol{s}_i) = \boldsymbol{x}_i$ the $(p+1)$-variate vector of covariates at location $\boldsymbol{s}_i$
- Assume $y_i = \boldsymbol{x}_i' \boldsymbol{\beta}_i + \varepsilon_i$ where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$

Geographically weighted regression

Brundson *et al.* (1998), Fotheringham *et al.* (2002)

- The total log likelihood is

$$\ell(\boldsymbol{\beta}) = -(1/2)\left\{n \log\left(2\pi\sigma^2\right) + \sigma^{-2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_i\right)^2\right\}$$

- With n observations and $np + 1$ parameters, the model is not identifiable.
- Idea: to estimate parameters by borrowing strength from nearby observations

Geographically weighted regression

Local regression (Loader, 1999)

Local regression uses a kernel function at each sampling location to weight observations based on their distance from the sampling location.

$$\mathcal{L}_i = \prod_{i'=1}^{n} \left(\mathcal{L}_{i'}\right)^{w_{ii'}}$$

$$\ell_i = \sum_{i'=1}^{n} w_{ii'} \left\{ \log\left(\sigma^2\right) + \sigma^{-2} \left(y_{i'} - \boldsymbol{x}'_{i'}\boldsymbol{\beta}_i\right)^2 \right\}$$

Given the weights, a local model is fit at each sampling location using the local likelihood

Geographically weighted regression

Local likelihood (Loader, 1999)

Weights are calculated via a kernel, e.g. the bisquare kernel:

$$w_{ii'} = \begin{cases} \left\{1 - \left(\phi^{-1}\delta_{ii'}\right)^2\right\}^2 & \text{if } \delta_{ii'} < \phi, \\ 0 & \text{if } \delta_{ii'} \geq \phi \end{cases} \tag{1}$$

where

- ϕ is a bandwidth parameter
- $\delta_{ii'} = \delta(\boldsymbol{s}_i, \boldsymbol{s}_{i'}) = \|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|_2$ is the Euclidean distance between sampling locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$.

Geographically weighted regression

Bandwidth estimation via the AIC_c (Hurvich *et al.*, 1998)

- Smaller bandwidth: less bias, more flexible coefficient surface
- Large bandwidth: less variance, less flexible coefficient surface
- Choose the bandwidth parameter to optimize the bias-variance tradeoff

Geographically weighted regression

Bandwidth estimation via the AIC$_c$ (Hurvich *et al.*, 1998)

- The corrected AIC for bandwidth selection is:

$$\text{AIC}_c = 2n \log \sigma + n \left\{ \frac{n + \nu}{n - 2 - \nu} \right\}$$

  - $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$
  - $\nu = \text{tr}(\boldsymbol{H})$
  - $\boldsymbol{H}_j = \left\{ \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X} \right\}_j$
  - Where subscript j indicates the jth row of the matrix

Geographically weighted regression

Bandwidth estimation via GCV (Wahba, 1990)

- The GCV criterion for bandwidth selection is:

$$\text{GCV} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{(n - \nu)^2}$$

  - $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$
  - $\nu = \text{tr}(\boldsymbol{H})$
  - $\boldsymbol{H}_j = \left\{\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}\right\}_j$
  - Where subscript j indicates the jth row of the matrix

Local variable selection and parameter estimation

Geographically weighted Lasso

Geographically weighted Lasso (Wheeler, 2009)

Within a GWR model, using the Lasso for local variable selection is called the geographically weighted Lasso (GWL).

- The GWL requires estimating a Lasso tuning parameter for each local model
- Wheeler (2009) estimates the local Lasso tuning parameter at location $\boldsymbol{s}_i$ by minimizing a jacknife criterion: $|y_i - \hat{y}_i^{(i)}|$
- The jacknife criterion can only be calculated where data are observed, making it impossible to use the GWL to impute missing data or to estimate the value of the coefficient surface at new locations
- Also, the Lasso is known to be biased in variable selection and suboptimal for coefficient estimation

Local variable selection and parameter estimation

Geographically weighted adaptive elastic net (GWEN)

- Local variable selection in a GWR model using the adaptive elastic net (AEN) (Zou and Zhang, 2009)

$$
\begin{aligned}
\mathcal{S}(\boldsymbol{\beta}_i) &= -2\ell_i(\boldsymbol{\beta}_i) + \mathcal{J}_2(\boldsymbol{\beta}_i) \\
&= \sum_{i'=1}^{n} w_{ii'} \left\{ \log \sigma_i^2 + \left(\sigma_i^2\right)^{-1} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i\right)^2 \right\} \\
&+ \alpha_i \lambda_i^* \sum_{j=1}^{p} |\beta_{ij}| / \gamma_{ij} \\
&+ (1-\alpha_i) \lambda_i^* \sum_{j=1}^{p} \left(\beta_{ij}/\gamma_{ij}\right)^2
\end{aligned}
$$

Local variable selection and parameter estimation

Geographically weighted adaptive elastic net (GWEN)

- The AEN penalty function is

$$\mathcal{J}_2(\boldsymbol{\beta}_i) = \alpha_i \lambda_i^* \sum_{j=1}^{p} |\beta_{ij}| / \gamma_{ij} + (1 - \alpha_i) \lambda_i^* \sum_{j=1}^{p} (\beta_{ij} / \gamma_{ij})^2$$

- Under suitable conditions, the AEN has an oracle property for selection in linear regression

Local variable selection and parameter estimation

Tuning parameter estimation

To estimate an AEN tuning parameter for each local model, use a local BIC that allows fitting a local model at any location within the spatial domain

$$
\begin{aligned}
\mathsf{BIC}_i &= -2\sum_{i'=1}^{n} \ell_{ii'} + \log\left(\sum_{i'=1}^{n} w_{ii'}\right)\mathsf{df}_i \\
&= \sum_{i'=1}^{n} w_{ii'}\left\{\log(2\pi) + \log\left(\hat{\sigma}_i^2\right) + \hat{\sigma}_i^{-2}\left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i'}\right)^2\right\} \\
&+ \log\left(\sum_{i'=1}^{n} w_{ii'}\right)\mathsf{df}_i
\end{aligned}
$$

Local variable selection and parameter estimation

Bandwidth parameter estimation

- Traditional GWR:
  - $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$
  - So traditional GWR is a linear smoother
  - $\nu = \text{tr}(\boldsymbol{H})$ is the degrees of freedom for the model
- GWEN:
  - $\hat{\boldsymbol{y}} = \boldsymbol{H}^{*}\boldsymbol{y} + \boldsymbol{T}^{*}\boldsymbol{\gamma}$
- GWEN is not a linear smoother
  - There is no projection matrix for GWEN so the degrees of freedom cannot be estimated by the trace of the projection matrix.
- Solution: use GWEN for selection then fit local model for the selected variables via traditional GWR
  - Now df $= \nu = \text{tr}(\boldsymbol{H})$

Local variable selection and parameter estimation

Locally linear coefficient estimation

- GWR, GWEN, GWAL: coefficients locally constant
  - as in Nadaraya-Watson kernel smoother
  - Leads to bias where there is a gradient at the boundary
- Solution: local polynomial modeling
  - First-order polynomial: locally linear coefficients
- Augment with covariate-by-location interactions
  - Two-dimensional
  - Augment with selected covariates only

Simulation study

Simulation study

Simulating covariates

- 30×30 grid on $[0, 1] \times [0, 1]$
- Five covariates $\tilde{X}_1, \ldots, \tilde{X}_5$
- Gaussian random fields:

$$\tilde{X}_j \sim N(0, \mathbf{\Sigma}) \text{ for } j = 1, \ldots, 5$$
$$\{\Sigma\}_{i,i'} = \text{exp}\{-\tau^{-1}\delta_{ii'}\} \text{ for } i, i' = 1, \ldots, n$$

- Colinearity: ρ
  - none ($\rho = 0$)
  - moderate ($\rho = 0.5$)

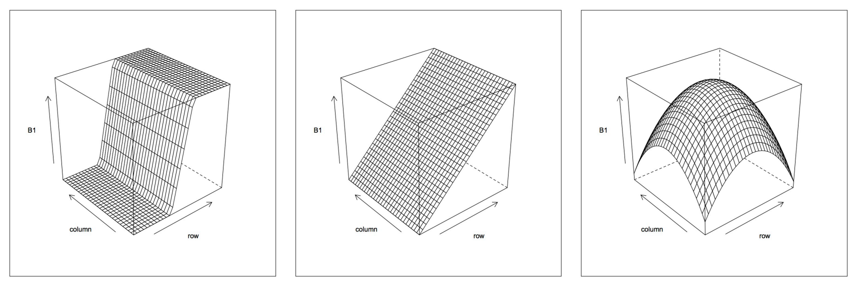
Simulation study

Simulating the response

- $Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta}(\boldsymbol{s}) = \sum_{j=1}^{5} \beta_j(\boldsymbol{s}) X_j(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$
- $\beta_1(\boldsymbol{s})$, the coefficient function for X_1, is nonzero in part of the domain.
- Coefficients for $X_2, \ldots, X_5$ are zero everywhere
- $\varepsilon(\boldsymbol{s}) \sim \ iid \ \ N(0, \sigma^2)$
  - Low noise: $\sigma = 0.5$
  - High noise: $\sigma = 1$

Simulation study

Coefficient functions: step, gradient, and parabola



Simulation study

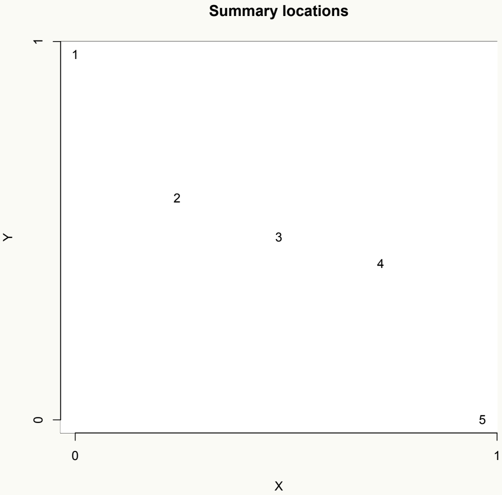
Simulation settings

Each setting simulated 100 times:

Setting	function	ρ	σ^2
1	step	0	0.25
2	step	0	1
3	step	0.5	0.25
4	step	0.5	1
5	gradient	0	0.25
6	gradient	0	1
7	gradient	0.5	0.25
8	gradient	0.5	1
9	parabola	0	0.25
10	parabola	0	1
11	parabola	0.5	0.25
12	parabola	0.5	1

Simulation results

Summary locations



Simulation results

Selection performance

- GWEN selection (60 cases):
  - 21 with no false positives (3 came when $\sigma = 1$, 8 when $\rho = 0.5$)
  - 30 with no false negatives
  - 13 with neither
- GWAL selection (60 cases):
  - 27 with no false positives
  - 26 with no false negatives
  - 17 with neither
- Incerased noise variance led to worse selection performance
- Increased colinearity in the covariates led to worse selection performance
- No consistent difference between GWEN and GWAL

Simulation results

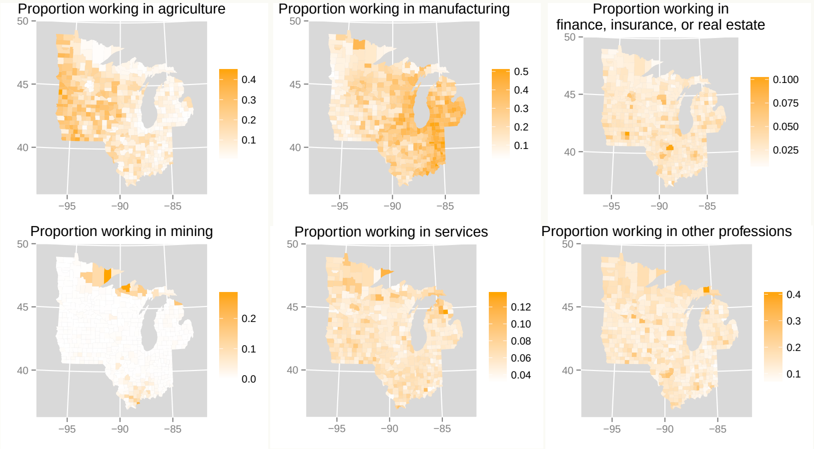
Estimation performance

- Oracular selection
  - best MSE($\hat{\beta}_1$) in 38 of the 60 cases
- Generally small difference between GWR, oracular, GWEN-LLE, and GWAL-LLE
- Incerased noise variance led to worse estimation accuracy
- Increased colinearity in the covariates led to worse estimation accuracy
- Fitting $\hat{y}$: MSE nearest σ^2 split between GWAL-LLE, oracle, and GWR

Data example: poverty rate in the upper midwest

Data example: poverty rate in the upper midwest

Revisiting the motivating example



Data example: poverty rate in the upper midwest

Data description

- Response: logit-transformed poverty rate in the Upper Midwest states of the U.S.
  - Minnesota, Iowa, Wisconsin, Illinois, Indiana, Michigan
- Covariates: employment structure (raw proportion employed in:)
  - agriculture
  - finance, insurance, and real estate
  - manufacturing
  - mining
  - services
  - other professions
- Data source: U.S. Census Bureau's decennial census of 1970

Data example: poverty rate in the upper midwest

Data description

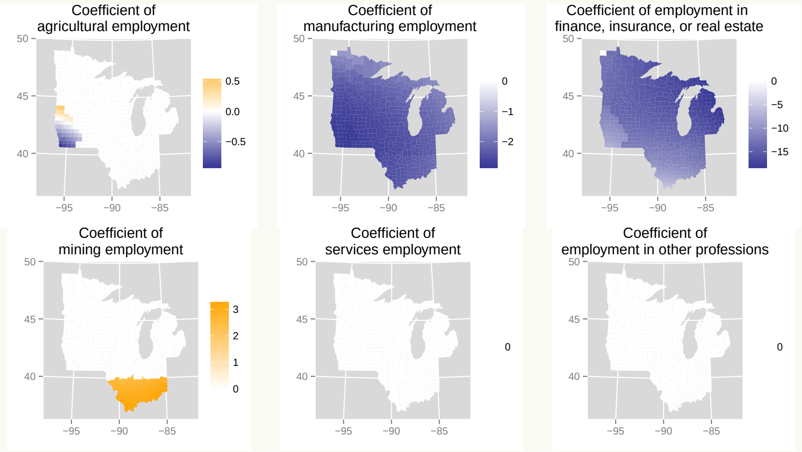
- Data aggregated to the county level
  - counties are areal units
- county centroid treated as sampling location

Data example: poverty rate in the upper midwest

Results from traditional GWR

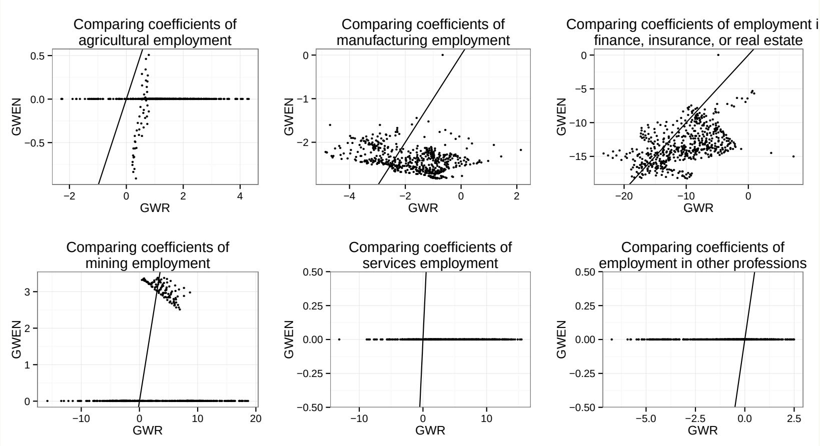
Data example: poverty rate in the upper midwest

Results from GWEN



Data example: poverty rate in the upper midwest

Comparing the coefficients from GWR and the GWEN



Data example: poverty rate in the upper midwest

Results from GWEN-LLE

- Relatively constant compared to GWR
- Services, "other professions" do not affect the poverty rate
- Manufacturing: negative coefficient everywhere
- Finance, insurance, and real estate negative coefficient everywhere
  - Largest magnitude (min: -20, next-largest: -3)
  - GWR comparable to GWEN-LLE
- Manufacturing: negative coefficient everywhere
  - GWR: coefficient greater than zero near Chicago and in NW Minnesota
- Agriculture: nonzero in western Iowa
  - North-south gradient to coefficient
  - ranges positive to negative
- Mining: nonzero in parts south
  - Associated with increased poverty rate
  - Comparable to GWR within far southern range

Future work

Future work

- Apply the GWEN to models for non-Gaussian response variable
- Incorporate spatial autocorrelation in the model
- PalEON project: modeling and mapping tree biomass in the upper midwest

Acknowledgements