GW-SELECT

Wesley Brooks

1. Introduction

Varying-coefficient regression is a technique used in spatial statistics to model a non-stationary process. Geographically Weighted Regression (GWR) Fotheringham et al. (2002) is a method of fitting varying-coefficient models.

The goal of this project is to develop a method that can answer these questions and work out the conditions under which the answers are reliable. Variable selection is via the Adaptive-LASSO, applied independently to each local model. The LASSO for GWR models was introduced in Wheeler (2009). That paper describes a process for model fitting that is similar to our process in the case of Gaussian data, but does not cover the Generalized Linear Model (GLM) case for non-Gaussian data that nevertheless follows an exponential-family distribution. Justification for the process in Wheeler (2009) is by simulation - the paper does not present theoretical proof of consistency of its estimators nor does it prove an oracle property for its variable selection.

The method in Wheeler (2009) requires cross-validation to select the model's tuning parameter and thus local models can be built only at locations where the output variable has been observed. It will be desirable to build local models that interpolate between the locations where the observations were taken. Using an information criterion like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to select the tuning parameter could solve this problem.

Another option is weighted cross-validation.

2. Geographically-weighted regression models

2.1. Model

Let univariate $Y(s)$ and $(p+1)$-variate $\boldsymbol{X}(s) = (1, X_1(s), \ldots, X_p(s))$ be spatial processes indexed by location s. For every s, let there be local regression coefficients $\boldsymbol{\beta}(s) = (\beta_0(s), \beta_1(s), \ldots, \beta_p(s))$ and local variance $\sigma^2(s)$ such that,

$$Y(s)|\boldsymbol{X}(s) \sim \mathcal{N}\left(\boldsymbol{X}(s)\boldsymbol{\beta}(s), \sigma^2(s)\right)$$

For $i = 1, \ldots, n$, let $\boldsymbol{x}(s_i)$ and $y(s_i)$ for $i = 1, \ldots, n$ be realizations of the processes at location s_i. The likelihood of the observed data is

$$L = \prod_{i=1}^{n}\left(\sqrt{2\pi\sigma^2(s_i)}\exp\left(-\frac{1}{2\sigma^2(s_i)}\left(y(s_i) - \boldsymbol{x}(s_i)'\boldsymbol{\beta}(s_i)\right)^2\right)\right)$$

2.2. Geographically-weighted regression

Data consists of n observations, sampled from spatial processes $\boldsymbol{X}(s)$ (predictor variables) and $Y(s)$ (response variable) at n unique locations $s_1, \ldots, s_n$. The observed data are $\{\boldsymbol{X}(s_i), Y(s_i) : i \in 1, \ldots, n\}$. The response Y is univariate; $\boldsymbol{X}(s)$ is p-variate: $\boldsymbol{X}(s) = (X_0(s), X_1(s), \ldots, X_p(s))^T$, where $X_0(s) \equiv 1 \, \forall s$

The data are compiled into matrices $X_{n\times p} = (\boldsymbol{X}(s_1), \ldots, \boldsymbol{X}(s_n))^T$ and $Y_{n\times 1} = (Y(s_1), \ldots, Y(s_n))^T$

Assume that at each location s, the response $Y(s)$ is related to the predictor variables by a model:

$$Y(s_i) = f\left(\boldsymbol{X}(s_i)\right)$$

The focus here is on the case where $f(\cdot)$ is a linear model model with coefficients $\boldsymbol{\beta}(s) = (\beta_0(s), \beta_1(s), \ldots, \beta_p(s))$:

$$Y(s_i) = \boldsymbol{X}(s_i)^T \boldsymbol{\beta}(s_i) + \epsilon(s_i)$$

$$\epsilon_i \sim \boldsymbol{\mathcal{N}}\left(0, \sigma_i^2\right)$$

Since the response variable is univariate, observed values are denoted, e.g., $y(s_i)$, while the vector of observed predictors at location s_i is denoted $\boldsymbol{x}(s_i)$. The kth element of $\boldsymbol{x}(s_i)$ is $x_k(s_i)$. The Adaptive-LASSO is used for variable selection with a different tuning parameter selected for each local model, so the Adaptive-LASSO tuning parameter at location s_i is denoted $\lambda(s_i)$.

3. Methods

The basic operation of GWR is to build a regression model at each of a set of pre-specified locations. At each model location, all of the observations in the data set are weighted based on their distance from the model location (the weights are uniquely specified by the combination of kernel function and bandwidth). The weighted observations are then used to build a regression model for that location.

3.1. Kernels

The kernel for geographic weighting could in principle be any function that applies a positive weight to each observation in the data set. This analysis uses the bisquare kernel throughout, with weights

based on distance (without regard for North-South versus East-West distances):

$$
W(\text{dist, bw}) = \begin{cases} (1 - (\frac{\text{dist}}{\text{bw}})^2)^2, & \text{if dist} < \text{bw} \\ 0, & \text{if dist} \geq \text{bw} \end{cases}
$$

The bisquare kernel has the benefit of giving zero weight to observations beyond the bandwidth, so that each local model depends only on the data from a local neighborhood and not at all on observations from outside that neighborhood.

Picking the bandwidth of the kernel is one crucial part of GWR. We currently have three methods to choose the bandwidth:

Global the bandwidth is the same distance for all local models.

k-**nearest-neighbors** for a given proportion q and n total observations, the bandwidth of each local model is set so that the sum of the geographic weights is $q \times n$.

Effective nearest neighbors an adaptive bandwidth selection method, in which the desired sum-of-squared-residuals is pre-specified, and the bandwidth of each local model is adjusted to meet this criterion. The goal is to broaden the bandwidth in areas where the model does not change much, and to narrow it in areas where there is a steep gradient to the coefficient surface. Pearson residuals are used to find the effective nearest neighbors bandwidth for a GW-GLM.

Each method requires a tuning parameter: this is the sum-of-squared-residuals for effective nearest neighbors, q for k-nearest neighbors, and the bandwidth itself in the case of a global bandwidth. The local tuning parameter is selected by minimizing the sum of absolute cross-validation error at the location of interest This is a sum (and therefor not quite the same as leave-one-out cross-validation) because we will sometimes have more than one observation per location (e.g., each of

the six decennial censuses in the `poverty` dataset gives us an observation at each county.)

3.2. Model fitting

3.2.1. Gaussian data

Fitting a GWR model to Gaussian data is done by minimizing the cross-validation criterion, using the Adaptive-LASSO for variable selection, where the adaptive weights are the Ordinary Least Squares (OLS) estimates of the model coefficients:

$$
\begin{aligned}
\hat{\mathrm{bw}} &= \operatorname*{argmin}_{\mathrm{bw}} (\mathrm{AIC}) = \operatorname*{argmin}_{\mathrm{bw}} \left(\sum_{i=1}^{n} \frac{\left(y(s_i) - \hat{y}\left(s_i, \hat{\lambda}(s_i), \mathrm{bw}\right) \right)^2}{\sigma_i^2} + \log \sigma_i^2 + \frac{2 \times \mathrm{df}}{\log\left(\sum_{j=1}^{n} w_{ij}\right)} \right) \\
\hat{\boldsymbol{\lambda}} &= \operatorname*{argmin}_{\boldsymbol{\lambda}} (\mathrm{AIC}) = \operatorname*{argmin}_{\boldsymbol{\lambda}} \left(\sum_{i=1}^{n} \right)
\end{aligned}
$$

Where $\hat{y}\left(s_i, \lambda(s_i), \mathrm{bw}\right)$ is the predicted value of the observation at location s_i:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}\left(s_i, \lambda(s_i), \mathrm{bw}\right) &= \operatorname*{argmin}_{\boldsymbol{\beta}} \left(\sum_{j=1}^{n} w_{ij}(\mathrm{bw}) \left(y(s_j) - \boldsymbol{x}(s_j)^T \boldsymbol{\beta} \right)^2 + \sum_{l=1}^{p} \lambda_l(s_i) \beta_l \right) \\
\lambda_l(s_i) &= \frac{\lambda(s_i)}{\beta_{OLS,l}(s_i, \mathrm{bw})} \\
\hat{\boldsymbol{\beta}}_{OLS}(s_i, \mathrm{bw}) &= \operatorname*{argmin}_{\boldsymbol{\beta}} \left(\sum_{j=1}^{n} w_{ij}(\mathrm{bw}) \left(y(s_j) - \boldsymbol{x}(s_j)^T \boldsymbol{\beta} \right)^2 \right) \\
\hat{y}\left(s_i, \lambda(s_i), \mathrm{bw}\right) &= \boldsymbol{x}(s_i)^T \hat{\boldsymbol{\beta}}\left(s_i, \lambda(s_i), \mathrm{bw}\right)
\end{aligned}
$$

Finally, the weights are given by:

$$
w_{ij}(\mathrm{bw}) = \begin{cases} \left(1 - \left(\frac{D(s_i, s_j)}{\mathrm{bw}} \right)^2 \right)^2, & \text{if } D(s_i, s_j) < \mathrm{bw} \\ 0, & \text{if } D(s_i, s_j) \geq \mathrm{bw} \end{cases}
$$

Where $D(s_i, s_j)$ is the distance from location of observation i to location of observation j

4. Data analysis

4.1. Simulated data

A simulation study is used to validate the method of analysis. Data was simulated from the following binomial spatial process:

$$Y(s) \sim N\left(\mu(s), \sigma^2\right)$$

$$\mu(s) = \boldsymbol{X}(s)^T \boldsymbol{\beta}(s)$$

$$\sigma^2 = 1$$

$$(X_1(s), X_2(s)) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right)$$

$$(X_3(s), X_4(s)) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right)$$

$$X_5 \sim N(0, 1)$$

$$Z \stackrel{iid}{\sim} N(0, 1)$$

$$\beta_0(s) \equiv 0$$

$$\beta_1(s) = \begin{cases} 2 \text{ if } s_x < 0.5 \\ 0 \text{ otherwise} \end{cases}$$

$$\beta_2(s) = 0$$

$$\beta_3(s) \equiv s_y$$

$$\beta_4(s) \equiv 0$$

$$\beta_5(s) \equiv 0$$

$$\beta_Z(s) \equiv 0$$

Where $s = (s_x, s_y) \in [0, 1] \times [0, 1]$. The variables X_1, X_2, X_3, X_4, and X_5 were generated with autocorrelation. For each of these variables, the spatial covariance is given by $C = \sigma^2 \exp(-\phi^{-1}d)$, where d is the distance between points and ϕ and σ^2 are covariance parameters. Both covariance

parameters were set to 1 during this simulation.

The process is observed on a regular 30×30 grid.

Simulation proceeded by repeating the following process 200 times:

1. Generate the data from the process (given above)

2. Select the optimal bandwidth for a GWR model on the simulated data

3. Generate a model using the selected bandwidth

4. Use the bootstrap to generate the empirical distribution of local coefficients at each observation location

5. At each observation location, note whether each covariate was selected for inclusion

6. At each observation location, note whether each covariate's true coefficient is within the 95% confidence interval as calculated by the bootstrap.

The analysis was carried out using the k-nearest-neighbors bandwidth selection method.

5. References

A.S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically weighted regression: the analysis of spatially varying relationships.* Wiley, 2002. ISBN 9780471496168. URL http://books.google.com/books?id=mPWlG-OG3rQC.

David C Wheeler. Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment and Planning A*, 41:722–742, 2009. doi: 10.1068/a40256.
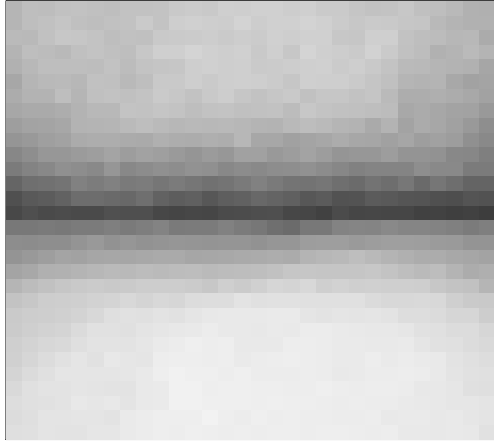
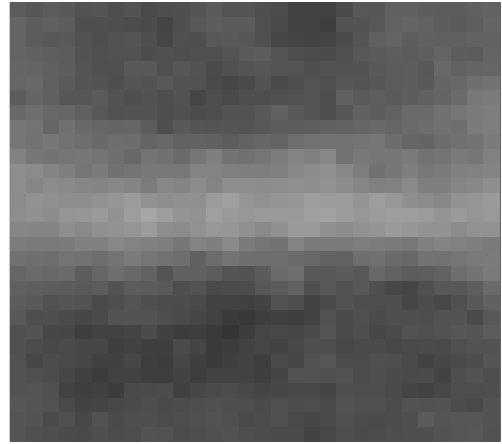(a) Coverage of 95% CI for X_1
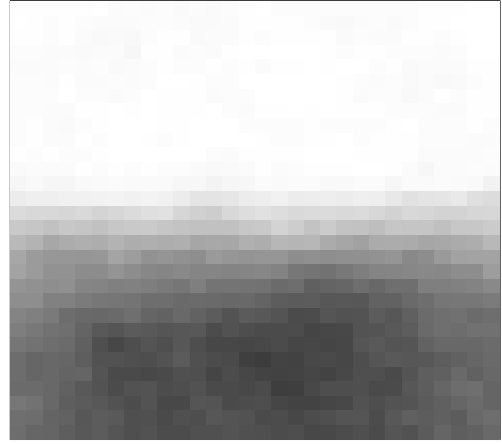
(b) Selection frequency for X_1

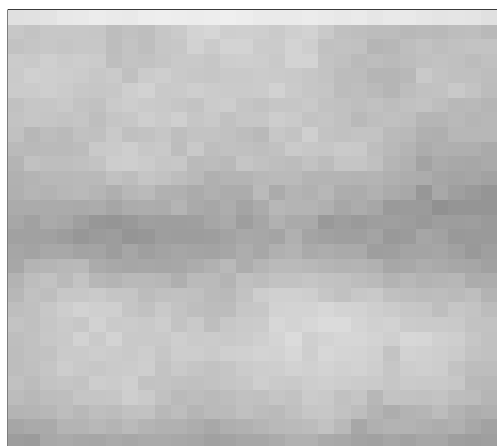Figure 1: Confidence interval coverage and selection frequency for X_1.
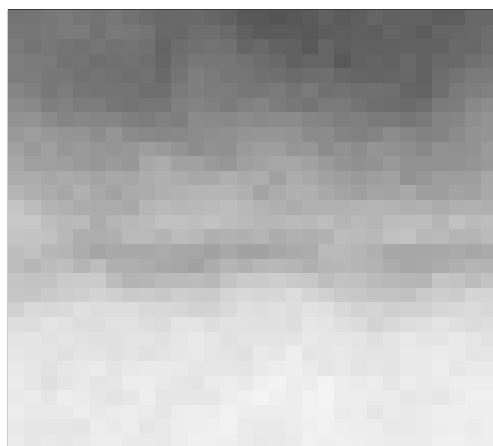

(a) Coverage of 95% CI for X_2

(b) Selection frequency for X_2

Figure 2: Confidence interval coverage and selection frequency for X_2.
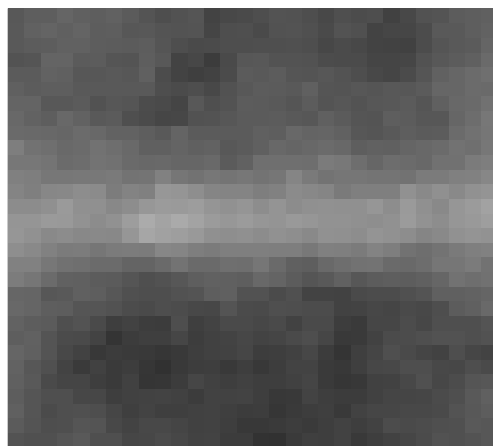
(a) Coverage of 95% CI for X_3

(b) Selection frequency for X_3

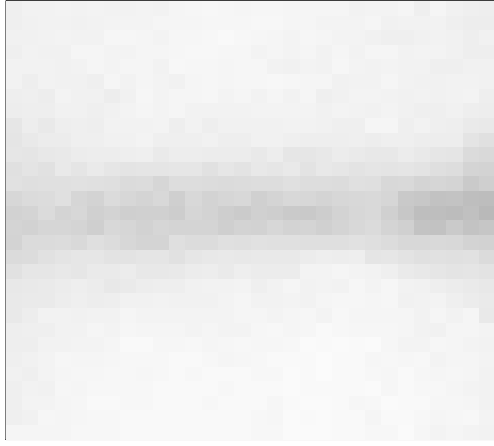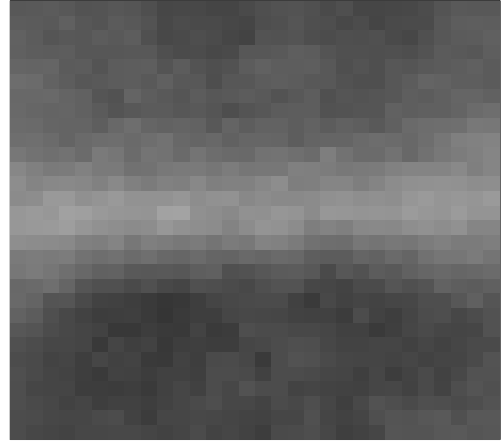Figure 3: Confidence interval coverage and selection frequency for X_3.

(a) Coverage of 95% CI for X_4

(b) Selection frequency for X_4

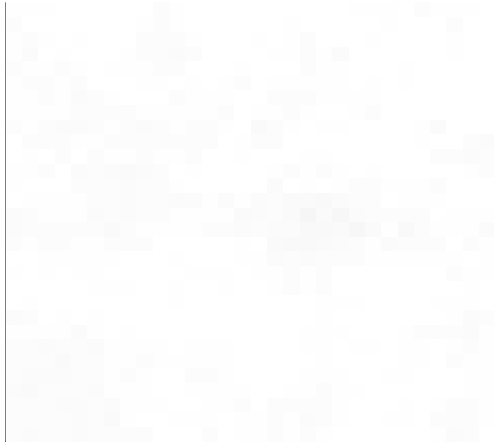Figure 4: Confidence interval coverage and selection frequency for X_4.

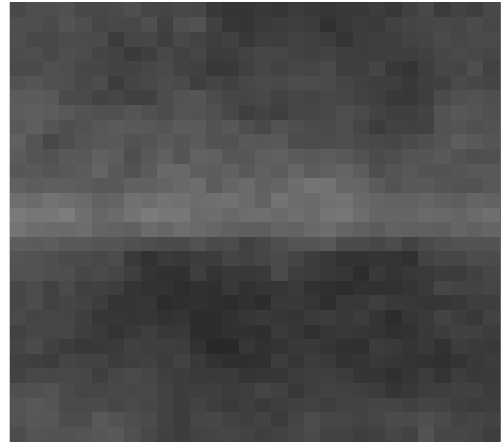(a) Coverage of 95% CI for X_5 (b) Selection frequency for X_5

Figure 5: Confidence interval coverage and selection frequency for X_5.

(a) Coverage of 95% CI for Z (b) Selection frequency for Z

Figure 6: Confidence interval coverage and selection frequency for Z.

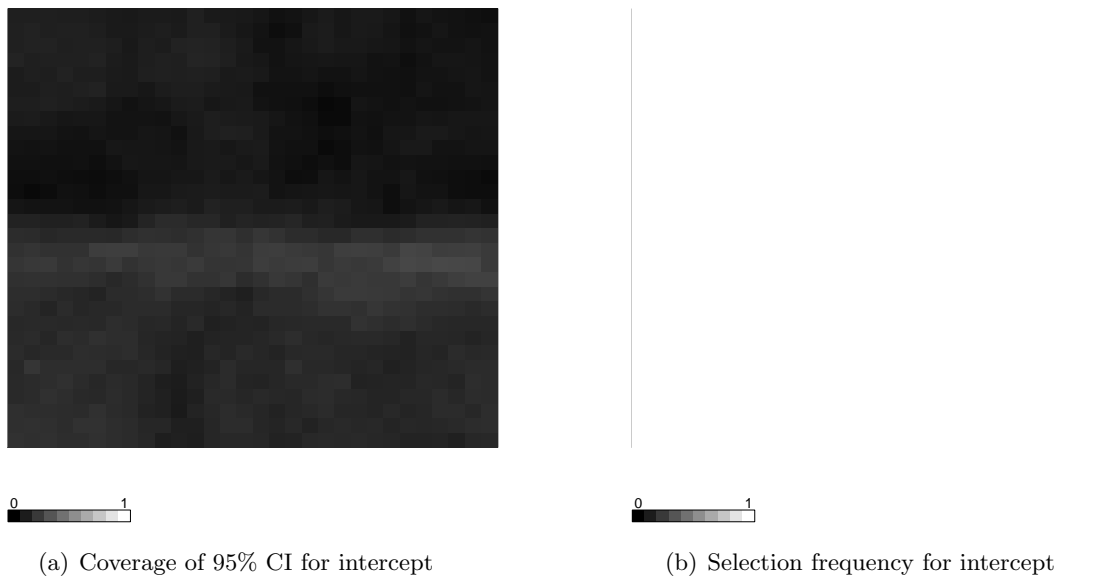(a) Coverage of 95% CI for intercept   (b) Selection frequency for intercept

Figure 7: Confidence interval coverage and selection frequency for intercept.