GW-SELECT

Wesley Brooks

1. Introduction

2. Geographically-weighted regression models

2.1. Model

Let univariate $Y(s)$ and $(p+1)$-variate $\boldsymbol{X}(s) = (1, X_1(s), \dots, X_p(s))$ be spatial processes indexed by location s. For every s, let there be local regression coefficients $\boldsymbol{\beta}(s) = (\beta_0(s), \beta_1(s), \dots, \beta_p(s))$ 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 $i = 1, \dots, n$, let $\boldsymbol{x}(s_i)$ and $y(s_i)$ for $i = 1, \dots, n$ be realizations of the processes at location s_i. The likelihood of the observed data is

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

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 $s_1, \dots, s_n$. The observed data are $\{\boldsymbol{X}(s_i), Y(s_i) : i \in 1, \dots, n\}$. The response Y is univariate; $\boldsymbol{X}(s)$ is p-variate: $\boldsymbol{X}(s) = (X_0(s), X_1(s), \dots, X_p(s))^T$, 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 k^{th} 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. Model selection and shrinkage

Coefficient shrinkage and variable selection are accomplished locally via the Lasso (Tibshirani, 1996). At each location s_i, the ordinary geographically-weighted regression estimator minimizes the objective function:

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

The objective minimized by the geographically-weighted also (GWL) is:

$$\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}'_{i'} \boldsymbol{\beta}_i \right)^2 + \sum_{j=1}^{p} \lambda_{ij} \beta_{ij} \tag{2}$$

Where $\lambda_{ij}, j \in \{1, \ldots, p\}$ are penalties from the adaptive lasso (Zou, 2006). The LAR algorithm is a stepwise regression algorithm; the number of steps to include in the model is chosen by minimizing the local AIC, with the sum of the weights arounds_i $\sum_{i'=1}^{n} w_{ii'}$ playing the role of the sample size and the number of nonzero coefficients in $\boldsymbol{\beta}_i$ playing the role of the "degrees of freedom" (df_i) (Zou et al., 2007). The estimated local variance $\hat{\sigma}_i^2$ is the variance estimate of the unpenalized local model (Zou et al., 2007).

Thus:

$$\text{AIC}_{\text{loc}} = \sum_{i'=1}^{n} w_{ii'} \hat{\sigma}_i^{-2} \left(y_{i'} - \boldsymbol{x}'_{i'} \hat{\boldsymbol{\beta}}_i \right)^2 + 2\text{df}_i \tag{3}$$

Because GWL is not a linear smoother (there is no smoothing matrix $\boldsymbol{S}$ such that $b\hat{m}y = \boldsymbol{S}\boldsymbol{y}$) the AIC and confidence intervals as calculated in Fotheringham et al. (2002) are not accurate for the GWL (Zou, 2006).

3.1. Bandwidth selection

The bandwidth is selected to minimize the total AIC. Because of the kernel weights and the application of the lasso, the sample size and degrees of freedom are different for each observation. The total AIC is found by taking the sum over all of the observed data:

$$\mathrm{AIC}_{\mathrm{tot}} = \sum_{i=1}^{n} \left(\hat{\sigma}_i^{-2} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_i \right)^2 + \log \hat{\sigma}_i^2 + 2\mathrm{df}_i \left(\sum_{i'=1}^{n} w_{ii'} \right)^{-1} \right) \tag{4}$$

The bandwidth that minimizes $\mathrm{AIC}_{\mathrm{tot}}$ is found by a line search.

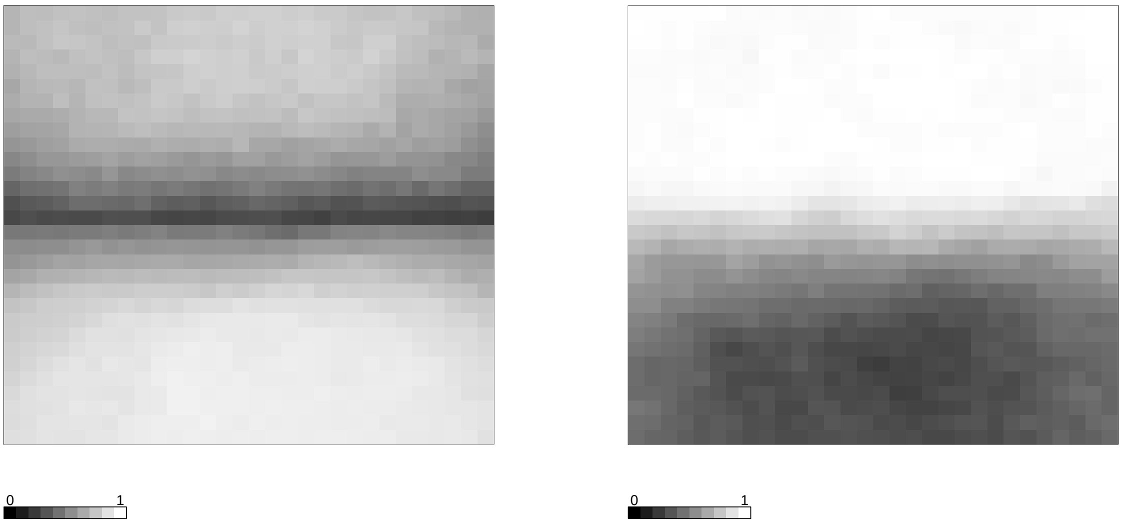
4. Simulation

4.1. Simulation setup

A simulation study was conducted to assess the finite-sample properties of the method described in Sections 2-3. Data was simulated on $[0,1] \times [0,1]$, which was divided into a 30×30 grid. Each of the $p = 5$ covariates was simulated by a Gaussian random field with mean zero and exponential covariance $Cov\left(Z_j(s_i), Z_j(s_{i'})\right) = \sigma^2 \exp\left(-\tau^{-1} \|s_i - s_{i'}\|\right)$ where $\sigma^2 = 1$ is the variance and τ is a range parameter. Correlation was induced between the covariates by multiplying the $\boldsymbol{Z}$ matrix by the Cholesky decomposition of the covariance matrix $\Sigma = \boldsymbol{R}'\boldsymbol{R}$. The covariance matrix is a 5×5 matrix that has ones on the diagonal and ρ for all off-diagonal entries, where ρ is the between-covariate correlation.

The simulated response is $Y(s) = \boldsymbol{X}(s)\boldsymbol{\beta}(s) + \epsilon(s)$, where the coefficient surface used to generate the data is $\boldsymbol{\beta}(s) = (\beta_0(s), \ldots, \beta_5(s)) = (0, \beta_1(s), 0, 0, 0, 0)$ and $\epsilon(s) \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1)$.

In order to evaluate the performance over a range of data conditions, the data was simulated at low and high values of the spatial covariance range, and at low and high values of the between-covariate



(a) Coverage of 95% CI for X_1 (b) Selection frequency for X_1

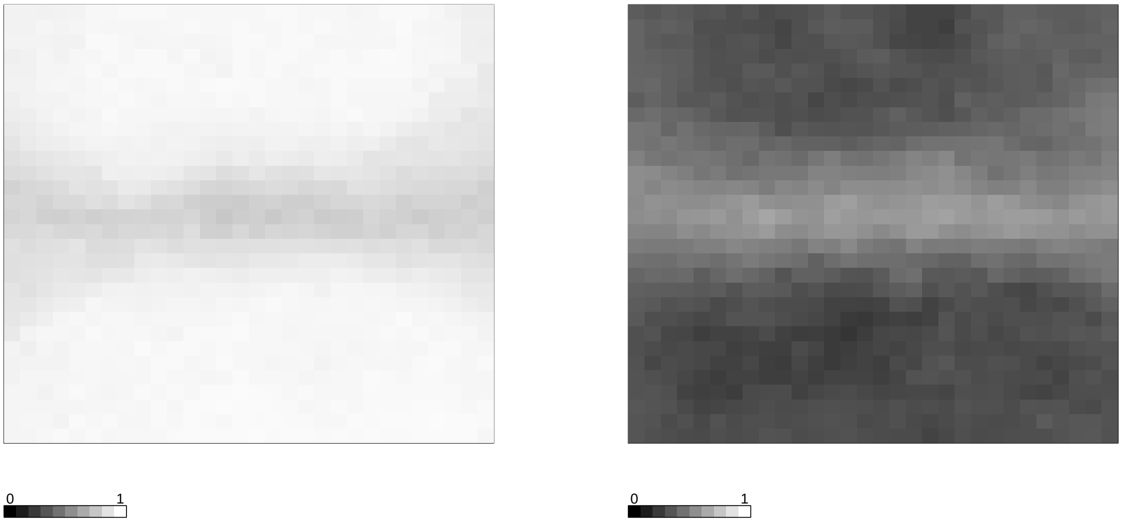
Figure 1: Confidence interval coverage and selection frequency for X_1.

correlation, for two cases of $\beta_1(s)$: the step function $\beta_1(s) = \begin{cases} 0 & \text{if } s_y < 0.5 \\ 1 & \text{o.w.} \end{cases}$ and the constant gradient $\beta_1(s) = 1 - s_y$.

Each case was simulated 100 times.

4.2. Simulation results

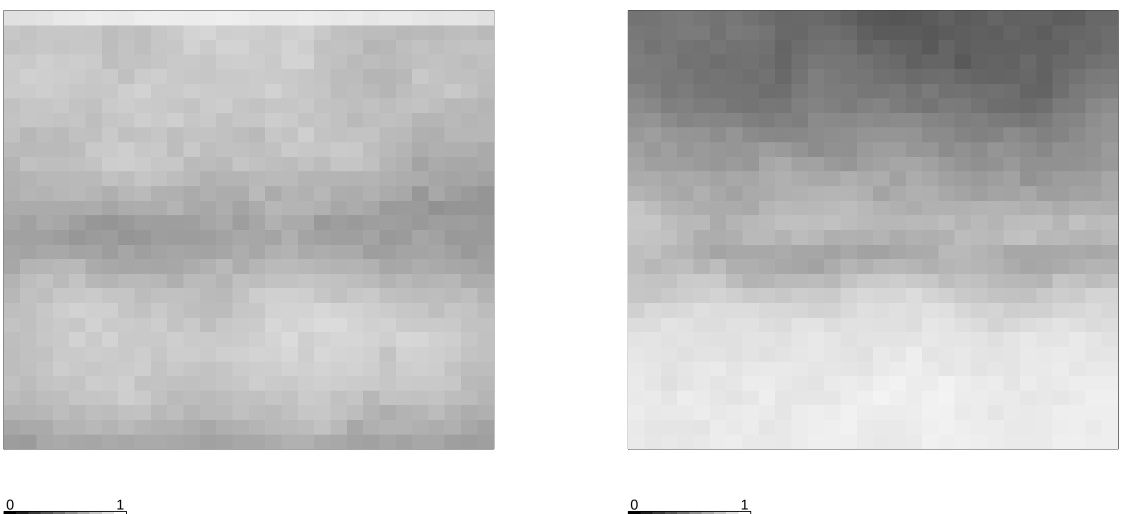
The coverage of the 95% CI and the selection frequency are plotted in the figures.



(a) Coverage of 95% CI for X_2

(b) Selection frequency for X_2

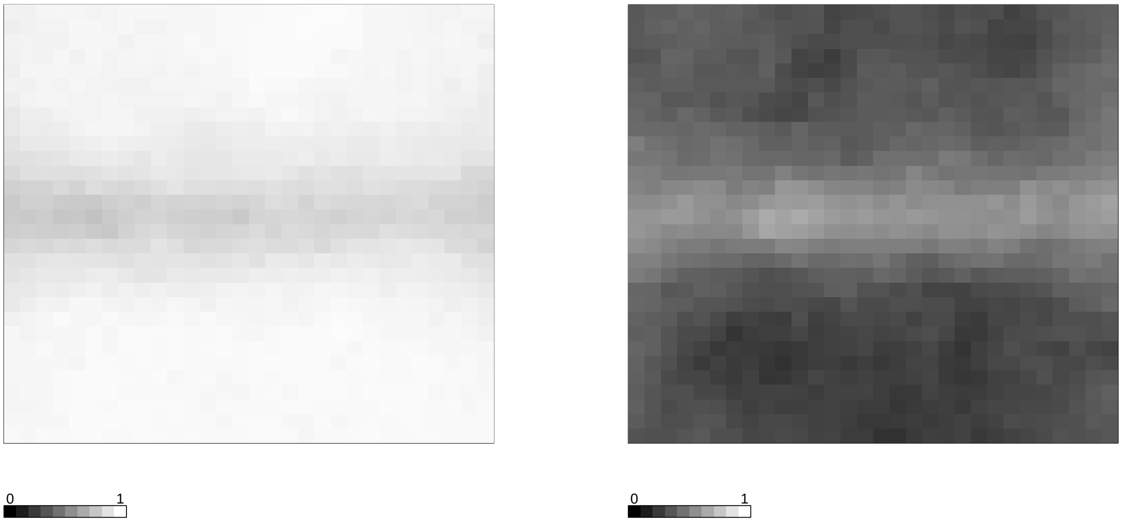
Figure 2: Confidence interval coverage and selection frequency for X_2.



(a) Coverage of 95% CI for X_3

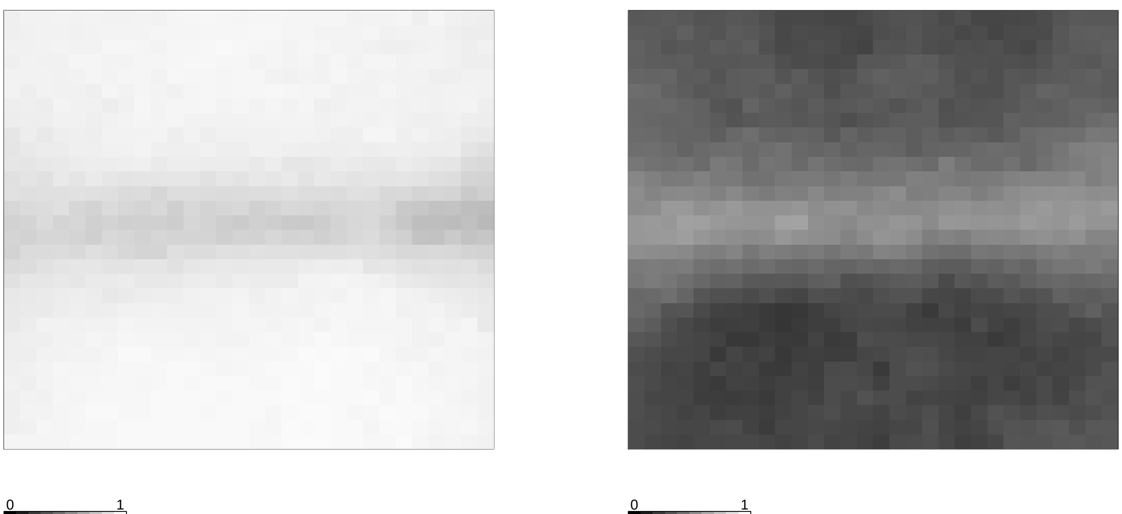
(b) Selection frequency for X_3

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



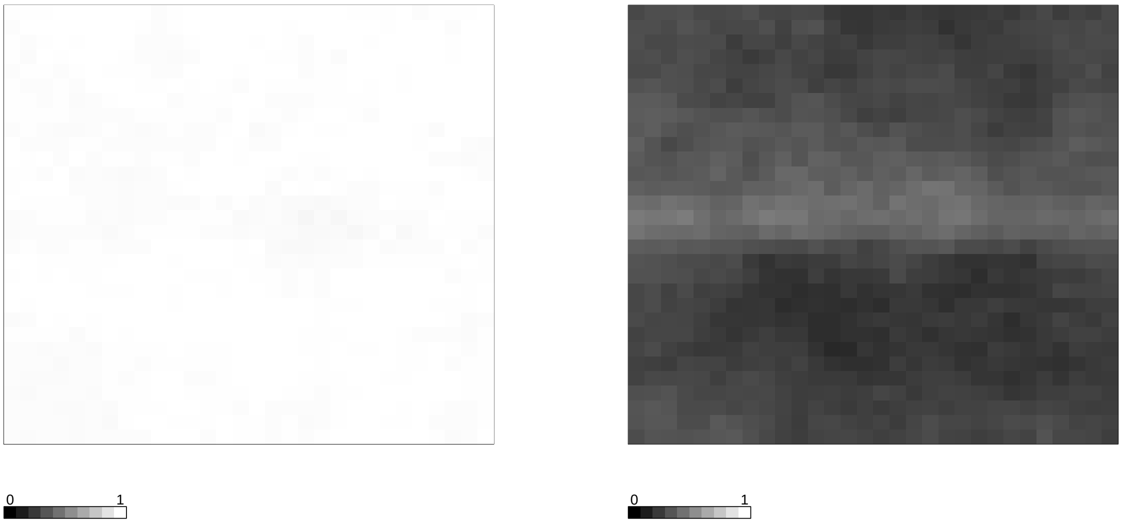
(a) Coverage of 95% CI for X_4 (b) Selection frequency for X_4

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

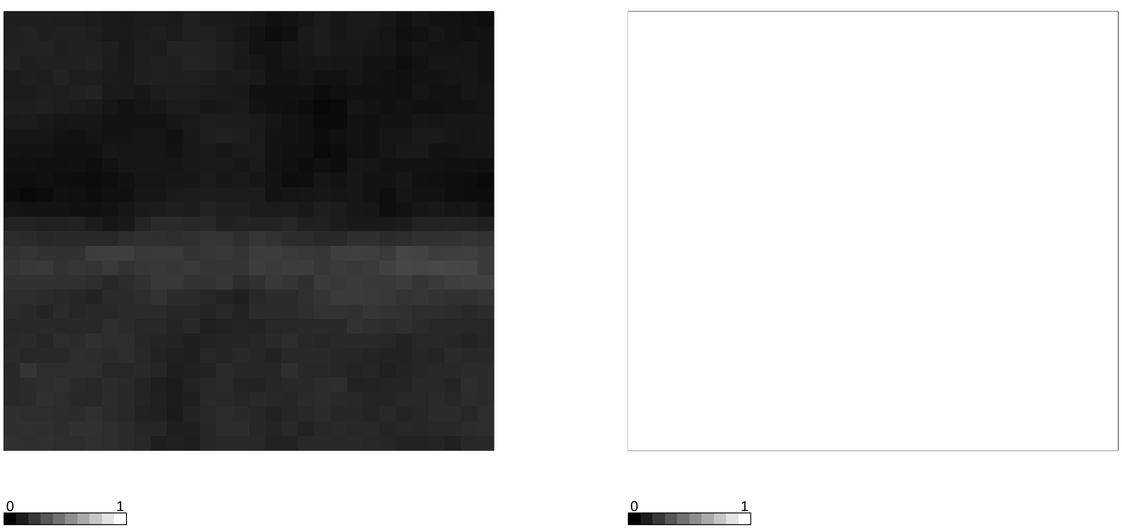
Figure 5: Confidence interval coverage and selection frequency for X_5.



(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.

5. References

A.S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, 2002.

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

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5):2173–2192, 2007.