GW-SELECT

WESLEY BROOKS

1. Introduction

Varying-coefficient models are a technique of local regression used in geostatistical analysis when one suspects that the effect of some covariate is not constant across the domain of a model. One method of fitting a varying-coefficient model is Geographically Weighted Regression (GWR) [**?**]. In GWR, a local regression model is fit at each point of interest (often, the points of interest are the locations where the data was collected). The technique of fitting the local regression model is to give a weight between zero and one to each observation based on its geographical distance from the point of interest, then do a weighted regression analysis to get the local model.

This paper discusses ongoing work in the area of GWR, focusing on an effort to establish a method of variable selection for GWR models. Variable selection can mean several things:

- What are the predictor variables that have no effect on the measured output, anywhere in the model's domain?
- Which coefficients are constant throughout the model's domain?
- Which coefficients are zero in some regions but non-zero in others?

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 [**?**]. 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 [**?**] 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.

2. Data

2.1. **Observational data.** Observational data comes from four sources:

Poverty data: From the U.S. Census Bureau, this data lists the proportion of families and individuals living in poverty at the county level for the midwestern states of Minnesota, Iowa, Wisconsin, Illinois, Indiana, and Michigan. County populations and other social-economic variables are tracked as well. The goal of analysis with this data set is to estimate the effect that socio-economic variables have on the poverty rate]

Soybean aphid data: Counts of aphids on soybean plats from 2003-2011, including the location of each measurement and an estimate of the crops grown nearby.

Mountain Pine Beetle Data: Collected in Canada's Cascade Mountains, this includes annual measurements from 1972-1986 of the intensity of Pine Beetle infestation at each location, along with other landscape and climatic covariates.

Land-cover data: The land use (in 1905, 1915, and 1986) of a region in northern Wisconsin's Chequamegon-Nicolet National Forest is compiled at the level of quarter sections of the Public Land Survey System (PLSS). Some predictor variables, calculated from land use statistics, are to be used in a model that describes the primary vegetation of each quarter section.

2.2. Simulated data.

Simulation studies are used to validate the method of analysis.

3. Model definitions

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 $X_{n\times p} = (\boldsymbol{X}(s_1), \dots, \boldsymbol{X}(s_n))^T$ and $Y_{n\times 1} = (Y(s_1), \dots, Y(s_n))^T$

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

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

Two cases are considered here: first, $f(\cdot)$ may be a linear model model with coefficients $\boldsymbol{\beta}(s) = (\beta_0(s), \beta_1(s), \dots, \beta_p(s))$:

$$Y(s_i) = \boldsymbol{X}(s_i)^T \boldsymbol{\beta}(s_i) + \epsilon(s_i), \text{ where } \epsilon_i \sim N\left(0, \sigma_i^2\right) \text{ for } i = 1, \dots, n$$

Alternatively, $f(\cdot)$ may be a Generalized Linear Model, with ling function $g(\cdot)$ and again with coefficients $\boldsymbol{\beta}(s) = (\beta_0(s), \beta_1(s), \dots, \beta_p(s))$:

$$\mathbb{E}\left(Y(s_i)\right) = g(\boldsymbol{X}(s_i)^T \boldsymbol{\beta}(s_i))$$

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 $\lambda(s_i)$.

4. 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.

4.1. 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):

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

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.)

4.2. Model fitting.

4.2.1. *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:

$$\{\hat{\boldsymbol{\lambda}},\ \hat{\text{bw}}\} = \underset{\boldsymbol{\lambda},\text{bw}}{\text{argmin}}\,(\text{CV error}) = \underset{\boldsymbol{\lambda},\text{bw}}{\text{argmin}}\left(\sum_{i=1}^{n} |y(s_i) - \tilde{y}\,(s_i, \lambda(s_i), \text{bw})|\right)$$

Where $\tilde{y}\,(s_i, \lambda(s_i), \text{bw})$ is the predicted value of the observation at location s_i from the local model that was fit by leaving out the observation at location s_i:

$$\hat{\boldsymbol{\beta}}\,(s_i, \lambda(s_i), \text{bw}) = \underset{\boldsymbol{\beta}}{\text{argmin}}\left(\sum_{j \neq i} w_{ij}(\text{bw})\left(y(s_j) - \boldsymbol{x}(s_j)^T\boldsymbol{\beta}\right)^2 + \sum_{l=1}^{p} \lambda_l(s_i)\beta_l\right)$$

$$\lambda_l(s_i) = \frac{\lambda(s_i)}{\beta_{OLS,l}(s_i, \text{bw})}$$

$$\hat{\boldsymbol{\beta}}_{OLS}(s_i, \text{bw}) = \underset{\boldsymbol{\beta}}{\text{argmin}}\left(\sum_{j=1}^{n} w_{ij}(\text{bw})\left(y(s_j) - \boldsymbol{x}(s_j)^T\boldsymbol{\beta}\right)^2\right)$$

$$\tilde{y}\,(s_i, \lambda(s_i), \text{bw}) = \boldsymbol{x}(s_i)^T\hat{\boldsymbol{\beta}}\,(s_i, \lambda(s_i), \text{bw})$$

Finally, the weights are given by:

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

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

4.2.2. *Non-Gaussian data.* A GWR model fit to non-Gaussian data from an exponential-family distribution is a geographically-weighted GLM (GW-GLM). Fitting a GW-GLM proceeds my maximum likelihood, rather than least squares. Again, the objective is to minimize a cross-validation criterion[1] that is calculated by summing the CV error at the location of each observation. Variable selection is again by the Adaptive-LASSO with the adaptive weights coming from the unpenalized GLM. The equations below are for data following a binomial distribution[2].

$$\{\hat{\boldsymbol{\lambda}},\ \hat{\text{bw}}\} = \underset{\boldsymbol{\lambda},\text{bw}}{\text{argmin}}\,(\text{CV error}) = \underset{\boldsymbol{\lambda},\text{bw}}{\text{argmin}}\left(\sum_{i=1}^{n} |y(s_i) - \tilde{y}\,(s_i, \lambda(s_i), \text{bw})|\right)$$

[1] I suspect that I am wrong here - minimizing each local model's CV error is probably overfitting because it is concerned only with the value at the model's central location. Instead, a BIC or GCV criterion that maximizes a regularized likelihood over the (weighted) data used in model-fitting will probably be a better choice. It would also allow selection of the Adaptive-LASSO tuning parameter to proceed at locations other than where the data was originally measured (which is not possible under the approach outlined here).

[2] I've also used Poisson-distributed data (counts of aphids on the soybean plants), and the different distribution changes the likelihood function but not the schematic process of model-fitting.

Where $\tilde{y}\left(s_i, \lambda(s_i), \text{bw}\right)$ is the predicted value of the observation at location s_i from the local model that was fit by leaving out the observation at location s_i:

$$\hat{\boldsymbol{\beta}}\left(s_i, \lambda(s_i), \text{bw}\right) = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(\sum_{j \neq i} w_{ij}(\text{bw}) \left[y(s_j)\boldsymbol{x}(s_j)^T\boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x}(s_j)^T\boldsymbol{\beta}}\right) \right] - \sum_{l=1}^{p} \lambda_l(s_i)\beta_l \right)$$

$\lambda_l(s_i) = \dfrac{\lambda(s_i)}{\beta_{UP,l}(s_i, \text{bw})}$ ($\boldsymbol{\beta}_{UP}(s_i, \text{bw})$ are the coefficients from the un-penalized model at location s_i):

$$\hat{\boldsymbol{\beta}}_{UP}(s_i, \text{bw}) = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(\sum_{j=1}^{n} w_{ij}(\text{bw}) \left[y(s_j)\boldsymbol{x}(s_j)^T\boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x}(s_j)^T\boldsymbol{\beta}}\right) \right] \right)$$

$$\eta\left(s_i, \lambda(s_i), \text{bw}\right) = \boldsymbol{x}(s_i)^T\boldsymbol{\beta}\left(s_i, \lambda(s_i), \text{bw}\right)$$

$$\tilde{y}\left(s_i, \lambda(s_i), \text{bw}\right) = g^{-1}\left(\eta\left(s_i, \hat{\lambda}(s_i), \hat{\text{bw}}\right)\right) = \frac{e^{\eta\left(s_i, \hat{\lambda}(s_i), \hat{\text{bw}}\right)}}{1 + e^{\eta\left(s_i, \hat{\lambda}(s_i), \hat{\text{bw}}\right)}}$$

Where $g^{-1}(\cdot)$ is the inverse link function. Finally, the weights are given by:

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

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