GW-SELECT

Wesley Brooks

---

1. Introduction

Varying-coefficient regression Hastie and Tibshirani (1993) 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 regression models for spatial data that uses kernel-weighted regression with weights based on the distance between observation locations. The presentation of GWR in Fotheringham et al. (2002) follows the development of local likelihood in **??**.

—Connection between the developments of GWR and nonparametric regression (including kernel smoothing vs local regression). GWR as a kernel-smoother working on the computed coefficients rather than the raw data. Local regression considered better than kernel smoothing (at least on grounds of consistency), adjusting GWR to become a local-regression problem rather than a kernel smoothing problem by focusing on covariate-by-coordinate interactions. Examine the literature of variable selection within local regression models (probably not much, since most local regression is univariate - e.g. LOESS).—

The literature is sparse regarding variable selection in varying-coefficient models. Wheeler (2009) uses the LASSO for local variable selection in GWR models and Fan and Zhang (1999) addresses global variable selection within spline-based varying-coefficient models for response variables that belong to an exponential-family distribution (as in the generalized linear model). Antoniadas et al. (2012) apply the non-negative garrote of Breiman (1995) in local variable selection using P-splines,

and Wang et al. (2008) address variable selection in spline-based models with repeated measurements.

2. Geographically-weighted regression models

2.1. Model

Consider n data observations, made at locations $s_1, \ldots, s_n$. For $i = 1, \ldots, n$, let $y(s_i)$ and $\boldsymbol{x}(s_i)$ be the univariate outcome of interest, and a $(p+1)$-variate vector of covariates measured at location s_i, respectively. At each location s_i, assume that the outcome is related to the covariates by a linear model with coefficients $\boldsymbol{\beta}_i(s_i)$ that may be spatially-varying.

$$y(s_i) = \boldsymbol{x}'(s_i)\boldsymbol{\beta}(s_i) + \epsilon(s_i) \tag{1}$$

Further assume that the error term $\epsilon(s)$ is normally distributed with zero mean and a possibly spatially-varying variance $\sigma^2(s)$

$$\epsilon(s_i) \sim \mathcal{N}\left(0, \sigma^2(s_i)\right) \tag{2}$$

In order to simplify the notation, let subscripts denote the values of data or parameters at the locations where data is observed. Thus, $\boldsymbol{x}(s_i) \equiv \boldsymbol{x}_i \equiv (1, x_{i1}, \ldots, x_{ip})'$, $\boldsymbol{\beta}(s_i) \equiv \boldsymbol{\beta}_i \equiv (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})'$, $y(s_i) \equiv y_i$, and $\sigma^2(s_i) \equiv \sigma_i^2$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)'$. Now equations 1 - 2 can be rewritten

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i + \epsilon_i \tag{3}$$

$$\epsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right) \tag{4}$$

Assume that, given the covariates $\boldsymbol{X}$, observations of the output at different locations are statistically independent of each other. Then the log-likelihood of the observed data is the sum of the

log-likelihood of each individual observation.

$$\ell(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\left[\log\left(2\pi\sigma_i^2\right) + \sigma_i^{-2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_i\right)^2\right] \tag{5}$$

With n observations and $n \times (p+1)$ free parameters, the model is overdetermined. To effectively reduce the number of parameters, assume that the spatially-varying coefficients $\boldsymbol{\beta}(s)$ are *smoothly* varying, and use a kernel smoother to make pointwise estimates of the coefficients. In the setting of spatial data and with a kernel smoother based on the physical distance between observation locations, this method is called geographically-weighted regression (GWR).

2.2. Geographically-weighted regression

Geographically-weighted regression estimates the value of the coefficient surface $\boldsymbol{\beta}(s)$ at each location s_i. Assume for now that there are known weights $w_{ii'}$ based on the distance $\|s_i - s_{i'}\|$ between locations s_i and $s_{i'}$ for all $i, i' \in \{1, \ldots, n\}$.

Coefficient estimation is done by maximizing the local (log-)likelihood at each location (Fotheringham et al., 2002).

$$\ell_i(\boldsymbol{\beta}_i) = -\frac{1}{2}\sum_{i'=1}^{n}\left\{\log(2\pi) + \log\left(\sigma_i^2 w_{ii'}^{-1}\right) + w_{ii'}\sigma_i^{-2}\left(y_{i'} - \boldsymbol{x}_{i'}'\boldsymbol{\beta}_i\right)^2\right\} \tag{6}$$

The first and second derivatives of the local log-likelihood are

$$\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_i} = \sum_{i'=1}^{n}\left[x_{i'j} w_{ii'}\sigma_i^{-2}\left(y_{i'} - \boldsymbol{x}_{i'}'\boldsymbol{\beta}_i\right)\right] \tag{7}$$

$$\left\{\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i'}\right\}_{j,k} = -\sum_{i'=1}^{n}\left\{x_{i'j} x_{i'k} w_{ii'}\sigma_i^{-2}\right\} \tag{8}$$

So the observed Fisher information in the locally weighted sample is

$$\begin{aligned}
\boldsymbol{\mathcal{J}}_i &= \sigma_i^{-2} \begin{pmatrix} \sum_{i'=1}^n w_{ii'} x_{i'1}^2 & \cdots & \sum_{i'=1}^n w_{ii'} x_{i'1} x_{i'p} \\ \vdots & \ddots & \vdots \\ \sum_{i'=1}^n w_{ii'} x_{i'p} x_{i'1} & \cdots & \sum_{i'=1}^n w_{ii'} x_{i'p}^2 \end{pmatrix} && (9) \\
&= \sigma_i^{-2} \sum_{i'=1}^n w_{ii'} \begin{pmatrix} x_{i'1}^2 & \cdots & x_{i'1} x_{i'p} \\ \vdots & \ddots & \vdots \\ x_{i'p} x_{i'1} & \cdots & x_{i'p}^2 \end{pmatrix} && (10) \\
&= \sigma_i^{-2} \sum_{i'=1}^n w_{ii'} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}' && (11)
\end{aligned}$$

The form of the observed Fisher information suggests that the information in the data $\boldsymbol{x}_{i'}$ about the coefficients at location s_i is proportional to the weight $w_{ii'}$

At each location s_i, the ordinary geographically-weighted regression estimator minimizes the objective function:

$$\sum_{i'=1}^n w_{ii'} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i \right)^2 \tag{12}$$

Letting the weight matrix $\boldsymbol{W}_i$ be

$$\boldsymbol{W}_i = \begin{pmatrix} w_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{in} \end{pmatrix} \tag{13}$$

estimation of the ordinary geographically-weighted regression coefficient surface is by weighted least squares:

$$\hat{\boldsymbol{\beta}}_{i,\mathrm{GWR}} = \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{Y} \tag{14}$$

2.3. Smoothing kernel

The bisquare kernel function is used to generate geographic weights based on the distance between observation locations. For estimating the value of the coefficient surface at location s_i, the weight given to the observation at location $s_{i'}$ is

$$w_{ii'} = \begin{cases} \left[1 - \left(\text{bw}^{-1} \left\| s_i - s_{i'} \right\|\right)^2\right]^2 & \text{if } \left\| s_i - s_{i'} \right\| < \text{bw} \\ 0 & \text{if } \left\| s_i - s_{i'} \right\| \geq \text{bw} \end{cases} \tag{15}$$

where bw is the kernel bandwidth.

3. Model selection and shrinkage

In traditional GWR, the model coefficients are calculated by weighted least squares, so model selection must be done *a priori*. In some settings, the Adaptive LASSO (Zou, 2006) has the "oracle" property of asymptotically selecting exactly the correct variables for inclusion in a regression model.

Applying the Adaptive LASSO in the setting of a GWR model requires that a tuning parameter be selected at each location where the coefficients are to be estimated. In Wheeler (2009), the tuning parameter for the LASSO at location s_i is selected to minimize the absolute jackknife prediction error $|y_i - \hat{y}_i^{(i)}|$, which means that the coefficients can only be estimated at the locations where data has been observed. On the other hand, using the local AIC to select the tuning parameter allows coefficients to be estimated at any location where the local likelihood can be calculated. The local AIC is calculated by adding a penalty to the local likelihood, with the sum of the weights around s_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).

3.1. Tuning parameter selection

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

$$\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}'_{i'}\boldsymbol{\beta}_i\right)^2 + \sum_{j=1}^{p} \lambda_{ij}\beta_{ij} \tag{16}$$

Where $\lambda_{ij}, j \in \{1, \ldots, p\}$ are penalties from the Adaptive LASSO (Zou, 2006). Taking the derivatives with respect to β and setting to zero, we see that

$$\hat{\boldsymbol{\beta}}_{i,\mathrm{GWL}} = \left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} - \frac{1}{2}\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{\lambda}_i \tag{17}$$

$$\hat{y}_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{i,\mathrm{GWL}} = \boldsymbol{x}_i\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} - \frac{1}{2}\boldsymbol{x}_i\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{\lambda}_i \tag{18}$$

So unlike in the case of ordinary geographically-weighted regression, the fitted values $\hat{\boldsymbol{Y}}$ are not a linear combination of the observations $\boldsymbol{Y}$. Because GWL is not a linear smoother the AIC and confidence intervals as calculated in Fotheringham et al. (2002) are not accurate for the GWL (Zou, 2006). The local AIC ($\mathrm{AIC}_{\mathrm{loc}}$) is minimized to select the adaptive lasso tuning parameter.

$$\mathrm{AIC}_{\mathrm{loc}} = \sum_{i'=1}^{n} w_{ii'}\hat{\sigma}_i^{-2}\left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_i\right)^2 + 2\mathrm{df}_i \tag{19}$$

Where the estimated local variance $\hat{\sigma}_i^2$ is the variance estimate from the unpenalized local model (i.e. the ordinary GWR model) (Zou et al., 2007).

3.2. Bandwidth selection

The bandwidth is selected to minimize the total AIC ($\mathrm{AIC}_{\mathrm{tot}}$). Because of the kernel weights and the application of the lasso, the sample size and degrees of freedom are different at each location. The total AIC is found by taking the sum over all of the observed data:

$$\mathrm{AIC}_{\mathrm{tot}} = \sum_{i=1}^{n}\left\{\hat{\sigma}_i^{-2}\left(y_i - \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_i\right)^2 + \log\hat{\sigma}_i^2 + 2\mathrm{df}_i\left(\sum_{i'=1}^{n} w_{ii'}\right)^{-1}\right\} \tag{20}$$

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

3.3. Confidence interval estimation

Confidence intervals for the coefficient estimates can be calculated either by the bootstrap (Efron and Tibshirani, 1986) or by using the variables selected by the Adaptive LASSO in a weighted least squares model. To compute coefficient confidence intervals via the bootstrap, the observations with non-zero geographic weights are resampled uniformly with replacement for each of n_B bootstrap replicates. For each bootstrap replicate, the GWL is used to estimate regression coefficients. The local likelihood of the bootstrap replicates may be different from that of the original sample, so the adaptive lasso tuning parameter may differ for each bootstrap replicate. Since the GWL is applied independently to each bootstrap replicate, the variables selected by GWL may be different for each replicate. The, e.g., 95% confidence interval for each regression coefficient is then the (2.5, 97.5) percentiles of the coefficient estimates from the bootstrap replicates.

Unshrunk coefficient estimates are found by using the GWL at each location for variable selection only and then estimating the coefficients for the selected variables by GWR. An unshrunk bootstrap confidence interval is found by estimating the unshrunk coefficients for each of the n_B bootstrap replicates and then calculating the percentiles as above.

A third way to estimate the coefficient confidence intervals is to use the GWL for variable selection only and then to use GWR to calculate a confidence interval based on the assumption of an independent, identically distributed, Gaussian error structure. In this case, the standard error of the regression coefficients is

$$\hat{\mathrm{se}}_{\beta_i} = \left(\tilde{\boldsymbol{X}}_i' \boldsymbol{W}_i \tilde{\boldsymbol{X}}_i\right)^{-1} \tilde{\boldsymbol{X}}_i' \boldsymbol{W}_i \boldsymbol{Y} \tag{21}$$

where $\tilde{\boldsymbol{X}}_i$ is the model matrix including only those variables that are selected by GWL at location i.

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 $y_i = \boldsymbol{x}_i\boldsymbol{\beta}_i + \epsilon_i$ for $i = 1, \ldots, 900$. The simulated data included the output y and five covariates $x_1, \ldots, x_5$. The true data-generating model used only x_1, so $x_2, \ldots, x_5$ are included to test the variable-selection properties of GWL. The coefficient surface of β_1 is described by the "step" function:

$$\beta_1(s) = \begin{cases} 0 & \text{if } s_y < 0.4 \\ 5(s_y - 0.4) & \text{if } 0.4 \leq s_y < 0.6 \\ 1 & \text{o.w.} \end{cases} \tag{22}$$

.

In order to evaluate the performance of GWL under a range of conditions, the data was simulated under 18 different settings for each type of β_1 (Table **??**): high (0.1) and low (0.03) levels of the

autoregression range parameter τ for the Gaussian random fields used to generate the covariates $\boldsymbol{X}_1(s), \ldots, \boldsymbol{X}_5(s)$; three levels (0, 0.5, 0.8) of between-covariate correlation ρ; and three levels (0, 0.03, 0.1) of the autoregression range parameter τ for the Gaussian random field used to generate the error term $\epsilon(s)$. Each case was simulated 100 times.

4.2. Simulation results

Results of the simulation experiment were summarized to asses the consistency in selection and estimation, as well as the coverage properties of the confidence intervals.

5. References

Antoniadas, A., I. Gijbels, and A. Verhasselt (2012). Variable selection in varying-coefficient models using p-splines. *Journal of Computational and Graphical Statistics 21*(3), 638–661.

Breiman, L. (1995). Better subset wregression using the nonnegative garrote. *Technometrics 51*, 373–384.

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science 1*(1), 54–75.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics 27*(5), 1491–1518.

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

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological) 55*(4), pp. 757–796.

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Wheeler, D. C. (2009). Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment and Planning A 41*, 722–742.

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

Zou, H., T. Hastie, and R. Tibshirani (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics 35*(5), 2173–2192.