Oracle properties of local adaptive grouped regularization

Wesley Brooks

---

1. Spatially varying coefficients regression

1.1. Model

Consider n data points, observed at sampling locations $\boldsymbol{s}_i = (s_{i,1} \;\; s_{i,2})^T$ for $i = 1, \ldots, \boldsymbol{s}_n$, which are distributed in a spatial domain $D \subset \mathbb{R}^2$ according to a density $f(\boldsymbol{s})$. For $i = 1, \ldots, n$, let $y(\boldsymbol{s}_i)$ and $\boldsymbol{x}(\boldsymbol{s}_i)$ denote, respectively, the univariate response and the $(p+1)$-variate vector of covariates measured at location $\boldsymbol{s}_i$. At each observation location $\boldsymbol{s}_i$, assume that the outcome is related to the covariates by a generalized linear model (GLM) where the coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i)$ may be spatially varying. That is, the distribution of Y conditional on X is:

$$f_{Y|X}(y|x) = \exp\{y\}$$

$$\mu(\boldsymbol{s}_i) = g^{-1}\{\eta(\boldsymbol{s}_i)\} \tag{1}$$

$$\eta(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta}(\boldsymbol{s}_i) \tag{2}$$

Further assume that the error term $\varepsilon(\boldsymbol{s}_i)$ is normally distributed with zero mean and variance σ^2, and that $\varepsilon(\boldsymbol{s}_i)$, $i = 1, \ldots, n$ are independent. That is,

$$\boldsymbol{\varepsilon} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{3}$$

In the context of nonparametric regression, the boundary-effect bias can be reduced by local polynomial modeling, usually in the form of a locally linear model (**?**). Here, locally linear coefficients are estimated by augmenting the local design matrix with covariate-by-location interactions in two dimensions as proposed by **?**. The augmented local design matrix at location $\boldsymbol{s}_i$ is

$$\boldsymbol{Z}(\boldsymbol{s}_i) = (\boldsymbol{X} \;\; L_i\boldsymbol{X} \;\; M_i\boldsymbol{X}) \tag{4}$$

where $\boldsymbol{X}$ is the unaugmented matrix of covariates, $L_i = \text{diag}\{s_{i'_1} - s_{i_1}\}$ and $M_i = \text{diag}\{s_{i'_2} - s_{i_2}\}$ for $i' = 1, \ldots, n$.

Now we have that $Y(\boldsymbol{s}_i) = \{\boldsymbol{Z}(\boldsymbol{s}_i)\}_i^T \boldsymbol{\zeta}(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i)$, where $\{\boldsymbol{Z}(\boldsymbol{s}_i)\}_i^T$ is the ith row of the matrix $\boldsymbol{Z}(\boldsymbol{s}_i)$ as a row vector, and $\boldsymbol{\zeta}(\boldsymbol{s}_i)$ is the vector of local coefficients at location $\boldsymbol{s}_i$, augmented with the local gradients of the coefficient surfaces in the two spatial dimensions, indicated by ∇_u and ∇_v:

$$\boldsymbol{\zeta}(\boldsymbol{s}_i) = \left(\boldsymbol{\beta}(\boldsymbol{s}_i)^T \;\; \nabla_u\boldsymbol{\beta}(\boldsymbol{s}_i)^T \;\; \nabla_v\boldsymbol{\beta}(\boldsymbol{s}_i)^T\right)^T$$

1.2. Estimation

The total log-likelihood of the observed data is the sum of the log-likelihood of each individual observation:

$$\ell\{\boldsymbol{\zeta}\} = -(1/2)\sum_{i=1}^{n}\left[\log\sigma^2 + \sigma^{-2}\left\{y(\boldsymbol{s}_i) - \boldsymbol{z}'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s}_i)\right\}^2\right]. \tag{5}$$

Since there are a total of $n \times 3(p+1)+1$ parameters for n observations, the model is not identifiable and it is not possible to directly maximize the total likelihood. But since the coefficient functions are smooth, the coefficients at location $\boldsymbol{s}$ can approximate the coefficients within some neighborhood

of $\boldsymbol{s}$, with the quality of the approximation declining as the distance from $\boldsymbol{s}$ increases.

This intuition is formalized by the local likelihood, which is maximized at location $\boldsymbol{s}$ to estimate the local coefficients $\boldsymbol{\zeta}(\boldsymbol{s})$:

$$\mathcal{L}\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} = \prod_{i=1}^{n} \left\{ \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-(1/2)\sigma^{-2}\left\{y(\boldsymbol{s}_i) - \boldsymbol{z}'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s})\right\}^2\right]\right\}^{K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)}, \tag{6}$$

The weights are computed from a kernel function $K_h(\cdot)$ such as the Epanechnikov kernel:

$$\begin{aligned} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|) &= h^{-2} K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|\right) \\ K(x) &= \begin{cases} (3/4)(1-x^2) & \text{if } x < 1, \\ 0 & \text{if } x \geqslant 1. \end{cases} \end{aligned} \tag{7}$$

Thus, the local log-likelihood function is, up to an additive constant:

$$\ell\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} = -(1/2)\sum_{i=1}^{n} K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)\left[\log\sigma^2 + \sigma^{-2}\left\{y(\boldsymbol{s}_i) - \boldsymbol{z}'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s})\right\}^2\right]. \tag{8}$$

Letting $\boldsymbol{W}(\boldsymbol{s})$ be a diagonal weight matrix where $W_{ii}(\boldsymbol{s}) = K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)$, the local likelihood is maximized by weighted least squares:

$$\begin{aligned} \mathcal{S}\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} &= (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \\ \therefore \tilde{\boldsymbol{\zeta}}(\boldsymbol{s}) &= \left\{\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Y} \end{aligned} \tag{9}$$

2. Local variable selection and parameter estimation

2.1. Local variable selection

Local adaptive grouped regularization (LAGR) is explored as a method of local variable selection and coefficient estimation in SVCR models. The proposed LAGR penalty is an adaptive ℓ_1 penalty

akin to the adaptive group lasso (**??**).

Grouped variables are selected together for inclusion in the model. Each group in a LAGR model consists of one covariate and its gradients on the two dimensions of spatial location. That is, $\boldsymbol{\zeta}_j(\boldsymbol{s}) = (\beta_j(\boldsymbol{s}) \quad \nabla_u\beta_j(\boldsymbol{s}) \quad \nabla_v\beta_j(\boldsymbol{s}))^T$ for $j = 1, \ldots, p$.

The objective function for the LAGR at location $\boldsymbol{s}$ is the penalized local sum of squares:

$$\begin{aligned} Q\{\boldsymbol{\zeta}(\boldsymbol{s})\} &= \mathcal{S}\{\boldsymbol{\zeta}(\boldsymbol{s})\} + \mathcal{J}\{\boldsymbol{\zeta}(\boldsymbol{s})\} \\ &= (1/2)\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\}^T \boldsymbol{W}(\boldsymbol{s})\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\}^T + \sum_{j=1}^{p} \phi_j(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\| \end{aligned} \tag{10}$$

which is the sum of the weighted sum of squares $\mathcal{S}\{\boldsymbol{\zeta}(\boldsymbol{s})\}$ and the LAGR penalty $\mathcal{J}\{\boldsymbol{\zeta}(\boldsymbol{s})\}$.

The LAGR penalty for the jth group of coefficients $\boldsymbol{\zeta}_j(\boldsymbol{s})$ at location $\boldsymbol{s}$ is $\phi_j(\boldsymbol{s}) = \lambda_n(\boldsymbol{s})\|\tilde{\boldsymbol{\zeta}}_j(\boldsymbol{s})\|^{-\gamma}$, where $\lambda_n(\boldsymbol{s}) > 0$ is a the local tuning parameter applied to all coefficients at location $\boldsymbol{s}$ and $\tilde{\boldsymbol{\zeta}}_j(\boldsymbol{s})$ is the vector of unpenalized local coefficients from (**??**).

2.2. Computation

2.2.1. Tuning parameter selection

Implementing LAGR requires the selection of local tuning parameters. The criteria commonly used for selecting tuning parameters in lasso-type models are appropriate here, including GCV (**?**), Cp (**?**), AIC (**?**), and BIC (**?**). All of the examples and simulations presented herein used the corrected AIC (AICc) (**?**) for tuning parameter selection:

$$\mathrm{AIC_c}(\boldsymbol{s}) = \hat{\sigma}^{-2}(\boldsymbol{s})\|\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\hat{\boldsymbol{\zeta}}(\boldsymbol{s})\|^2 + 2df + 2\frac{df(df+1)}{\sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|) - df - 1} \tag{11}$$

where df is the degrees of freedom as defined in **?**:

$$df = \sum_{j=1}^{p} I\left\{\|\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s})\|\right\} + 2\sum_{j=1}^{p} \frac{\|\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s})\|}{\|\tilde{\boldsymbol{\zeta}}_j(\boldsymbol{s})\|} \tag{12}$$

3. References