LAGR and its oracle properties

Wesley Brooks

June 25, 2014

1 Introduction

Here goes the literature review. Essetial elements are:

- nonparametric regression

- varying coefficients regression

- adaptive lasso

- generalized linear models

2 Varying coefficient models

2.1 Model

Consider n data points, observed at sampling locations $\boldsymbol{s}_i = (s_{i,1} \quad s_{i,2})^T$ for $i = 1, \dots, \boldsymbol{s}_n$, 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 location $\boldsymbol{s}_i$, assume that the outcome is related to the covariates by a linear model where the coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i)$ may be spatially-varying and $\varepsilon(\boldsymbol{s}_i)$ is random error at location $\boldsymbol{s}_i$. That is,

$$y(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i). \tag{1}$$

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,

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

2.2 Augment the covariates and the coefficients with location interactions

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

$$\boldsymbol{Z}(\boldsymbol{s}_i) = (\boldsymbol{X} \;\; L_i\boldsymbol{X} \;\; M_i\boldsymbol{X}) \tag{3}$$

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 $Y(\boldsymbol{s}_i) = \left\{\boldsymbol{Z}(\boldsymbol{s}_i)\right\}_i^T \boldsymbol{\zeta}(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i)$, where $\left\{\boldsymbol{Z}(\boldsymbol{s}_i)\right\}_i^T$ 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:

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

2.3 Local likelihood

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

$$\ell\left\{\boldsymbol{\zeta}\right\} = -(1/2) \sum_{i=1}^{n} \left[\log \sigma^2 + \sigma^{-2} \left\{y(\boldsymbol{s}_i) - \boldsymbol{z}'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s}_i)\right\}^2\right]. \quad (4)$$

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 $\boldsymbol{\zeta}(\boldsymbol{s})$:

$$\mathcal{L}\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} = \prod_{i=1}^{n} \left\{\left(2\pi\sigma^2\right)^{-1/2} \exp\left[-(1/2)\sigma^{-2} \left\{y(\boldsymbol{s}_i) - \boldsymbol{z}'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s})\right\}^2\right]\right\}^{K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)}, \quad (5)$$

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

$$K_h(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|) = h^{-2} K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|\right)$$

$$K(x) = \begin{cases} (3/4)(1-x^2) & \text{if } x < 1, \\ 0 & \text{if } x \geq 1. \end{cases} \tag{6}$$

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

$$\ell\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} = -(1/2)\sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)\left[\log\sigma^2 + \sigma^{-2}\left\{y(\boldsymbol{s}_i) - \boldsymbol{z}'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s})\right\}^2\right]. \tag{7}$$

2.4 Estimating the coefficients

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:

$$\mathcal{S}\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} = (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T$$

$$\therefore \tilde{\boldsymbol{\zeta}}(\boldsymbol{s}) = \left\{\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Y} \tag{8}$$

Now Theorem 3 of **?** says that, for any given $\boldsymbol{s}$

$$\sqrt{nh^2 f(\boldsymbol{s})}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{\beta}(\boldsymbol{s}) - (1/2)\kappa_0^{-1}\kappa_2 h^2\left\{\boldsymbol{\beta}_{uu}(\boldsymbol{s}) + \boldsymbol{\beta}_{vv}(\boldsymbol{s})\right\}\right] \xrightarrow{D} N\left(\boldsymbol{0}, \kappa_0^{-2}\nu_0\sigma^2\Psi^{-1}\right)$$

3 Local variable selection with LAGR

Estimating the local coefficients by (8) relies on *a priori* variable selection. The goal of local adaptive grouped regularization (LAGR) is to simultaneously select the locally relevant predictors and estimate the local coefficients. The proposed LAGR penalty is an adaptive ℓ_1 penalty akin to the adaptive group lasso **??**.

3.1 Variable groupings

Each raw covariate in a LAGR model is grouped with its covariate-by-location interactions. 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 $j = 1, \ldots, p$. By the mechanism of the group lasso, variables within the same group are included in or dropped from the model together. The intercept group is left unpenalized.

3.2 Write down the LAGR-penalized local likelihood

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

$$\begin{aligned} Q\{\boldsymbol{\zeta}(\boldsymbol{s})\} &= \mathcal{S}\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} + \mathcal{J}\{\boldsymbol{\zeta}(\boldsymbol{s})\} \\ &= (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s}) \left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T + \sum_{j=1}^{p} \phi_j(\boldsymbol{s}) \|\boldsymbol{\zeta}_j(\boldsymbol{s})\| \end{aligned} \tag{9}$$

which is the sum of the weighted sum of squares $\mathcal{S}\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\}$ 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 (8).

3.3 Oracle properties of LAGR

Theorem 1 (Asymptotic normality). *If $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$ and $hn^{-1/2}b_n \xrightarrow{p} \infty$ then*

$$h\sqrt{n}\left[\hat{\boldsymbol{\beta}}_{(a)}(\boldsymbol{s}) - \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) - \frac{\kappa_2 h^2}{2\kappa_0}\{\nabla^2_{uu}\boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}_{(a)}(\boldsymbol{s})\}\right] \xrightarrow{d} N(0, f(\boldsymbol{s})^{-1}\kappa_0^{-2}\nu_0\sigma^2\Psi^{-1})$$

Theorem 2 (Selection consistency). *If $h^{-1}n^{-1/2}a_n \xrightarrow{p} \infty$ and $hn^{-1/2}b_n \xrightarrow{p} \infty$ then $P\left\{\|\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s})\| = 0\right\} \to 0$ if $j \leq p_0$ and $P\left\{\|\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s})\| = 0\right\} \to 1$ if $j > p_0$.*

Remarks Together, Theorem 1 and Theorem 2 indicate that the LAGR estimates have the same asymptotic distribution as a local regression model where the nonzero coefficients are known in advance **?**, and that the LAGR estimates of true zero coefficients go to zero with probability one. Thus, selection and estimation by LAGR has the oracle property.

A note on rates To prove the oracle properties of LAGR, we assumed that $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$ and $hn^{-1/2}b_n \xrightarrow{p} \infty$. Therefore, $h^{-1}n^{-1/2}\lambda_n(\boldsymbol{s}) \to 0$ for $j \leq p_0$ and $hn^{-1/2}\lambda_n(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\|^{-\gamma} \to \infty$ for $j > p_0$.

We require that $\lambda_n(\boldsymbol{s})$ can satisfy both assumptions. Suppose $\lambda_n(\boldsymbol{s}) = n^{\alpha}$, and recall that $h = O(n^{-1/6})$ and $\|\tilde{\zeta}_p(\boldsymbol{s})\| = O(h^{-1}n^{-1/2})$. Then $h^{-1}n^{-1/2}\lambda_n(\boldsymbol{s}) = O(n^{-1/3+\alpha})$ and $hn^{-1/2}\lambda_n(\boldsymbol{s})\|\tilde{\zeta}_p(\boldsymbol{s}\|^{-\gamma} = O(n^{-2/3+\alpha+\gamma/3})$.

So $(2-\gamma)/3 < \alpha < 1/3$, which can only be satisfied for $\gamma > 1$.

4 Extension to GLLMs

4.1 Model

Generalized linear models (GLM) extend the linear model to distributions other than gaussian. The generalized local linear model (GLLM) is an extension of the GLM to varying coefficient models via local regression.

As was the case for local linear regression models, the GLLM coefficients are smooth functions of location, called $\boldsymbol{\beta}(\boldsymbol{s})$. If the response variable y is from an exponential-family distribution then its density is

$$f\left\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right\} = c\left\{y(\boldsymbol{s})\right\} \times \exp\left[\frac{\theta(\boldsymbol{s})y(\boldsymbol{s}) - b\left\{\theta(\boldsymbol{s})\right\}}{a\left\{\phi(\boldsymbol{s})\right\}}\right]$$

where ϕ and θ are parameters and

$$\begin{aligned}
E\left\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right\} =& \mu(\boldsymbol{s}) = b'\left\{\theta(\boldsymbol{s})\right\} \\
\theta(\boldsymbol{s}) =& (g \circ b')^{-1}\left\{\eta(\boldsymbol{s})\right\} \\
\eta(\boldsymbol{s}) =& \boldsymbol{x}^T(\boldsymbol{s})\boldsymbol{\beta}(\boldsymbol{s}) = g\left\{\mu(\boldsymbol{s})\right\} \\
\mathrm{Var}\left\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right\} =& b''\left\{\theta(\boldsymbol{s})\right\} a\left\{\phi(\boldsymbol{s})\right\}
\end{aligned}$$

The function $g(\cdot)$ is called the link function. If its inverse $g^{-1}(\cdot) = b'(\cdot)$ then the composition $(g \circ b')(\cdot)$ is the identity function. This particular choice of g

is called the canonical link. We follow the practice of **?** in assuming the use of the canonical link because it is simple and because using an alternative link function would hardly affect the local fit.

Under the canonical link function, the expressions for the mean and variance of the response variable can be simplified to

$$\begin{aligned} E\left\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right\} =& g^{-1}\left\{\eta(\boldsymbol{s})\right\} \\ \mathrm{Var}\left\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right\} =& a\left\{\phi(\boldsymbol{s})\right\}/g'\left\{\mu(\boldsymbol{s})\right\} = V\left\{\mu(\boldsymbol{s})\right\} \times a\left\{\phi(\boldsymbol{s})\right\} \end{aligned}$$

4.2 Local quasi-likelihood

Assuming the canonical link, all that is required is to specify the mean-variance relationship via the variance function, $V\left\{\mu(\boldsymbol{s})\right\}$. Then the GLLM coefficients can be estimated by maximizing the local quasi-likelihood

$$\ell\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} = \sum_{i=1}^{n} K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|) Q\left[g^{-1}\left\{z'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s})\right\}, Y(\boldsymbol{s}_i)\right]. \qquad (10)$$

The local quasi-likelihood generalizes the local log-likelihood that was used to estimate coefficients in the local linear model case. The quasi-likelihood is convex, and is defined in terms of its derivative, the quasi-score function

$$\frac{\partial}{\partial\mu} Q(\mu, y) = \frac{y-\mu}{V(\mu)}.$$

4.3 Estimation

Under these conditions, the local quasi-likelihood is maximized where

$$\ell' \left\{ \hat{\boldsymbol{\zeta}}(\boldsymbol{s}) \right\} = \sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|) \frac{y(\boldsymbol{s}_i) - \hat{\mu}(\boldsymbol{s}_i)}{V\left\{\hat{\mu}(\boldsymbol{s}_i)\right\} g'\left\{\hat{\mu}(\boldsymbol{s}_i)\right\}} \boldsymbol{z}(\boldsymbol{s}_i) = \boldsymbol{0}_{3p}. \tag{11}$$

Except for the $K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)$ term, this is the same as the normal equations for estimating coefficients in a GLM. The method of iteratively reweighted least squares (IRLS) is used to solve for $\hat{\boldsymbol{\zeta}}(\boldsymbol{s})$.

4.4 Distribution of the local coefficients

The asymptotic distribution of the local coefficients in a varying-coefficients GLM with a one-dimensional effect-modifying parameter are given in **?**. For coefficients that vary in two dimensions (e.g. spatial location), the asymptotic distribution under the canonical link is

$$\sqrt{nh^2 f(\boldsymbol{s})} \left[\hat{\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{\beta}(\boldsymbol{s}) - (1/2)\kappa_0^{-1}\kappa_2 h^2 \left\{ \boldsymbol{\beta}_{uu}(\boldsymbol{s}) + \boldsymbol{\beta}_{vv}(\boldsymbol{s}) \right\} \right] \xrightarrow{D} N\left(\boldsymbol{0}, \kappa_0^{-2}\nu_0 V\left\{\mu(\boldsymbol{s})\right\} \Psi^{-1} \right)$$

4.5 LAGR penalty

As in the case of linear models, the LAGR for GLMs is a grouped ℓ_1 regularization method.

4.6 Oracle properties of LAGR in the GLM setting

5 Simulations

A simulation study was conducted to assess the performance of the method described in Sections 2–3.

Data were simulated on the domain $[0, 1]^2$, which was divided into a 30×30 grid. Each of $p = 5$ covariates $X_1, \ldots, X_5$ was simulated by a Gaussian random field with mean zero and exponential covariance function $\text{Cov}\left(X_{ji}, X_{ji'}\right) = \sigma_x^2 \exp\left(-\tau_x^{-1}\delta_{ii'}\right)$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0.1$ is the range parameter, and $\delta_{ii'}$ is the Euclidean distance $\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|_2$.

Correlation was induced between the covariates by multiplying the matrix $\boldsymbol{X} = (X_1 \cdots X_5)$ by $\boldsymbol{R}$, where $\boldsymbol{R}$ is the Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{R}'\boldsymbol{R}$. The covariance matrix $\boldsymbol{\Sigma}$ 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 was $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i + \varepsilon_i$ for $i = 1, \ldots, n$ where $n = 900$ and the ε_i's were iid Gaussian with mean zero and variance σ_ε^2. The simulated data included the response y and five covariates $x_1, \ldots, x_5$. The true data-generating model uses only x_1. The variables $x_2, \ldots, x_5$ are included to assess performance in model selection.

Three different functions were used for the coefficient surface $\beta_1(\boldsymbol{s})$. They are plotted in Figure 1, and their mathematical forms are listed in (12). The first is a step function, which is equal to zero in 40% of the spatial domain, equal to one in a different 40% of the spatial domain, and increases linearly in the middle

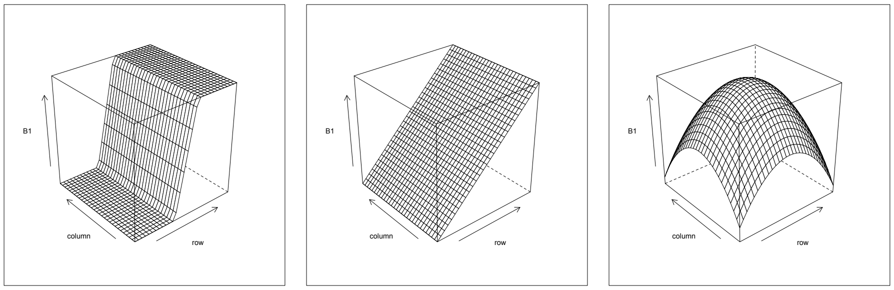

Figure 1: These are, respectively, the step, gradient, and parabola functions that were used for the coefficient function $\beta_1(\boldsymbol{s})$ in the VCR model $y(\boldsymbol{s}_i) = x_1(\boldsymbol{s}_i)\beta_1(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i)$ when generating the data for the simulation study.

20% of the domain. The second is a gradient function, which increases linearly from zero at one end of the domain to one at the other. The final coefficient function is a parabola taking its maximum value of 1 at the center of the domain and falling to zero at each corner of the domain.

$$\begin{aligned}
\beta_{step}(\boldsymbol{s}) &= \begin{cases} 1 & if\ s_x > 0.6 \\ 5s_x - 2 & if\ 0.4 < s_x \leq 0.6 \\ 0 & o.w. \end{cases} \\
\beta_{gradient}(\boldsymbol{s}) &= s_x \\
\beta_{parabola}(\boldsymbol{s}) &= 1 - \frac{(s_x - 0.5)^2 + (s_y - 0.5)^2}{0.5}
\end{aligned} \tag{12}$$

In total, three parameters were varied to produce 18 settings, each of which was simulated 100 times. There were three functional forms for the coefficient surface $\beta_1(\boldsymbol{s})$; data was simulated both with low ($\rho = 0$), medium ($\rho = 0.5$), and high ($\rho = 0.9$) correlation between the covariates; and simulations were made with low ($\sigma_\varepsilon^2 = 0.25$) and high ($\sigma_\varepsilon^2 = 1$) variance for the random error term.

$\beta_1(\boldsymbol{s})$	ρ	σ^2_ε
step	0	0.25
		1
	0.5	0.25
		1
	0.9	0.25
		1
gradient	0	0.25
		1
	0.5	0.25
		1
	0.9	0.25
		1
parabola	0	0.25
		1
	0.5	0.25
		1
	0.9	0.25
		1

Table 1: Listing of the simulation settings used to assess the performance of LAGR models versus oracle selection and no selection.

The simulation settings are enumerated in Table 1.

5.1 Methods for comparison

The performance of LAGR was compared to that of a VCR model without variable selection, and to a VCR model with oracular selection. Oracular selection means that exactly the correct set of covariates was used to fit each local model.

5.2 Results

The results are presented in terms of the mean integrated squared error (MISE) of the coefficient surface estimates $\hat{\beta}_1(\boldsymbol{s}), \ldots, \hat{\beta}_5(\boldsymbol{s})$, the MISE of the fitted response $\hat{y}(\boldsymbol{s})$, and the frequency with which the coefficient surface estimates $\hat{\beta}_1(\boldsymbol{s}), \ldots, \hat{\beta}_5(\boldsymbol{s})$ in the LAGR model were zero.

	LAGR	none	oracle
1	*0.02*	0.02	**0.01**
2	*0.03*	0.03	**0.02**
3	*0.02*	0.02	**0.01**
4	*0.03*	0.05	**0.02**
5	*0.03*	0.05	**0.01**
6	*0.12*	0.17	**0.02**
7	0.01	*0.01*	**0.00**
8	0.03	*0.02*	**0.01**
9	0.01	*0.01*	**0.00**
10	0.04	*0.03*	**0.01**
11	*0.03*	0.04	**0.00**
12	*0.14*	0.14	**0.01**
13	0.01	*0.01*	**0.01**
14	0.03	*0.02*	**0.02**
15	0.01	*0.01*	**0.01**
16	0.03	*0.03*	**0.02**
17	*0.02*	0.04	**0.01**
18	0.17	*0.14*	**0.02**

Table 2: The MISE for the estimates of $\beta_1(\boldsymbol{s})$ in each simulation setting, under variable selection via LAGR, no variable selection, and oracular variable selection. Highlighting indicates the **lowest** and *next-lowest* MISE.

The MISE of the estimates of $\beta_1(\boldsymbol{s})$ are in Table 2.

6 Data example

The proposed LAGR estimation method was used to estimate the coefficients in a VCR model of the effect of some covariates on the price of homes in Boston. The data source is the Boston house price data set of **???**, which is based on the 1970 U.S. census. In the data, we have the median price of homes sold in 506 census tracts (MEDV), along with some potential predictor variables. The predictor variables are CRIM (the per-capita crime rate in the tract), RM (the mean number of rooms for houses sold in the tract), RAD (an index of how accessible the tract is from Boston's radial roads), TAX (the property tax per

	LAGR	none	oracle
1	*0.25*	0.26	**0.25**
2	*1.00*	**1.00**	0.99
3	*0.26*	0.26	**0.25**
4	*0.99*	**1.00**	0.98
5	*0.27*	0.30	**0.25**
6	*1.08*	1.14	**0.98**
7	*0.25*	**0.25**	0.25
8	**0.99**	*0.99*	0.97
9	*0.25*	**0.25**	0.24
10	*1.00*	**1.00**	0.97
11	*0.27*	0.28	**0.24**
12	*1.09*	1.12	**0.97**
13	*0.25*	**0.25**	0.25
14	**1.00**	*1.00*	0.98
15	**0.25**	0.25	*0.25*
16	**1.00**	*1.00*	0.97
17	*0.26*	0.28	**0.24**
18	1.13	*1.12*	**0.98**

Table 3: The MISE for the fitted output in each simulation setting, under variable selection via LAGR, no variable selection, and oracular variable selection. Highlighting indicates the **closest** and *next-closest* to the actual error variance σ_ε^2 for that setting.

	x
1	0.97
2	0.96
3	0.96
4	0.92
5	0.86
6	0.85
7	0.96
8	0.95
9	0.94
10	0.92
11	0.80
12	0.85
13	0.97
14	0.94
15	0.95
16	0.88
17	0.79
18	0.78

Table 4: Proportion of local models under each setting in which the coefficients $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$ are estimated as exactly zero.

$10,000 of property value), and LSTAT (the percentage of the tract's residents who are considered "lower status").

The bandwidth parameter was set to 0.2 for a nearest neighbors-type bandwidth, meaning that the sum of kernel weights for each local model was 20% of the total number of observations. The kernel used was the Epanechnikov kernel.

6.1 Results

Estimates of the regression coefficients are plotted in Figure 2.

One interesting result is that the

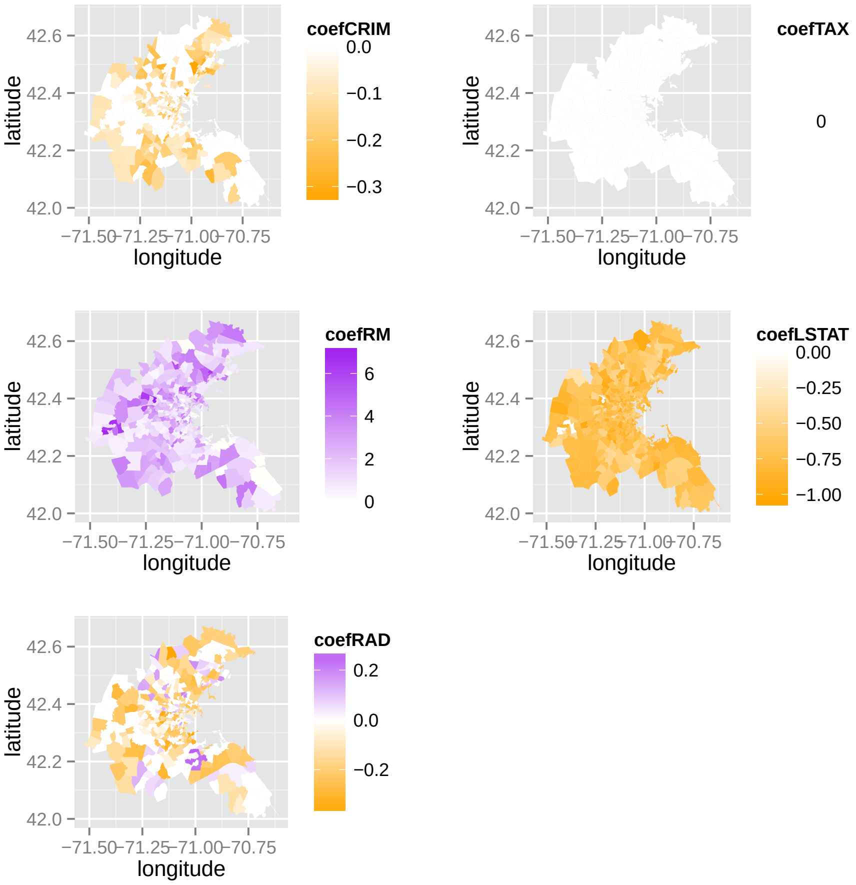

Figure 2: The coefficients for the boston house price data as estimated by LAGR.

A Proofs of theorems

Proof of theorem 1. □

Define $V_4^{(n)}(\boldsymbol{u})$ to be the

$$
\begin{aligned}
V_4^{(n)}(\boldsymbol{u}) &= Q\left\{\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right\} - Q\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} \\
&= (1/2)\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\left\{\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right\}\right]^T \boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\left\{\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right\}\right] \\
&\qquad + \sum_{j=1}^{p} \phi_j(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| \\
&\qquad - (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\} - \sum_{j=1}^{p} \phi_j(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\| \\
&= (1/2)\boldsymbol{u}^T\left\{h^{-2}n^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}\boldsymbol{u} - \boldsymbol{u}^T\left[h^{-1}n^{-1/2}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\
&\qquad + \sum_{j=1}^{p} n^{-1/2}\phi_j(\boldsymbol{s})n^{1/2}\left\{\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| - \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|\right\} \qquad (13)
\end{aligned}
$$

Note the different limiting behavior of the third term between the cases $j \le p_0$ and $j > p_0$:

Case $j \le p_0$ If $j \le p_0$ then $n^{-1/2}\phi_j(\boldsymbol{s}) \to n^{-1/2}\lambda_n(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\|^{-\gamma}$ and $|\sqrt{n}\left\{\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| - \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|\right\}|$ $h^{-1}\|\boldsymbol{u}_j\|$ so

$$\lim_{n\to\infty} \phi_j(\boldsymbol{s})\left(\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| - \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|\right) \le h^{-1}n^{-1/2}\phi_j(\boldsymbol{s})\|\boldsymbol{u}_j\| \le h^{-1}n^{-1/2}a_n\|\boldsymbol{u}_j\| \to 0$$

Case $j > p_0$ If $j > p_0$ then $\phi_j(\boldsymbol{s})\left(\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| - \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|\right) = \phi_j(\boldsymbol{s})h^{-1}n^{-1/2}\|\boldsymbol{u}_j\|$.

And note that $h = O(n^{-1/6})$ so that if $hn^{-1/2}b_n \xrightarrow{p} \infty$ then $h^{-1}n^{-1/2}b_n \xrightarrow{p} \infty$.

Now, if $\|\boldsymbol{u}_j\| \neq 0$ then

$$h^{-1}n^{-1/2}\phi_j(\boldsymbol{s})\|\boldsymbol{u}_j\| \geq h^{-1}n^{-1/2}b_n\|\boldsymbol{u}_j\| \to \infty$$

. On the other hand, if $\|\boldsymbol{u}_j\| = 0$ then $h^{-1}n^{-1/2}\phi_j(\boldsymbol{s})\|\boldsymbol{u}_j\| = 0$.

Thus, the limit of $V_4^{(n)}(\boldsymbol{u})$ is the same as the limit of $V_4^{*(n)}(\boldsymbol{u})$ where

$$V_4^{*(n)}(\boldsymbol{u}) = \begin{cases} (1/2)\boldsymbol{u}^T\left\{h^{-2}n^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}\boldsymbol{u} - \boldsymbol{u}^T\left[h^{-1}n^{-1/2}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] & \text{if } \|\boldsymbol{u}_j\| = 0 \; \forall j > p \\ \infty & \text{otherwise} \end{cases}$$

From which it is clear that $V_4^{*(n)}(\boldsymbol{u})$ is convex and its unique minimizer is $\hat{\boldsymbol{u}}^{(n)}$:

$$\begin{aligned} 0 &= \left\{h^{-2}n^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}\hat{\boldsymbol{u}}^{(n)} - \left[h^{-1}n^{-1/2}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\ \therefore \hat{\boldsymbol{u}}^{(n)} &= \left\{n^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1}\left[hn^{-1/2}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \end{aligned} \tag{14}$$

By the epiconvergence results of **?** and **?**, the minimizer of the limiting function is the limit of the minimizers $\hat{\boldsymbol{u}}^{(n)}$. And since, by Lemma 2 of **?**,

$$\hat{\boldsymbol{u}}^{(n)} \xrightarrow{d} N\left(\frac{\kappa_2 h^2}{2\kappa_0}\{\nabla^2_{uu}\boldsymbol{\zeta}_j(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_j(\boldsymbol{s})\}, f(\boldsymbol{s})\kappa_0^{-2}\nu_0\sigma^2\Psi^{-1}\right) \tag{15}$$

the result is proven.

Proof of theorem 2. We showed in Theorem 1 that $\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s}) \xrightarrow{p} \boldsymbol{\zeta}_j(\boldsymbol{s}) + \frac{\kappa_2 h^2}{2\kappa_0}\{\nabla^2_{uu}\boldsymbol{\zeta}_j(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_j(\boldsymbol{s})\}$, so to complete the proof of selection consistency, it only remains to show that $P\left\{\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s}) = 0\right\} \to 1$ if $j > p_0$. □

The proof is by contradiction. Without loss of generality we consider only the case $j = p$.

Assume $\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| \neq 0$. Then $Q\{\boldsymbol{\zeta}(\boldsymbol{s})\}$ is differentiable w.r.t. $\boldsymbol{\zeta}_p(\boldsymbol{s})$ and is minimized where

$$\begin{aligned}
0 &= \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}_{-p}(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{-p}(\boldsymbol{s}) - \boldsymbol{Z}_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\right\} - \phi_p(\boldsymbol{s})\frac{\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})}{\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\|} \\
&= \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\
&\quad + \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{-p}(\boldsymbol{s})\left[\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_{-p}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_{-p}(\boldsymbol{s})\right] \\
&\quad + \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_p(\boldsymbol{s})\left[\boldsymbol{\zeta}_p(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_p(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_p(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\right] \\
&\quad - \phi_p(\boldsymbol{s})\frac{\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})}{\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\|}
\end{aligned} \tag{16}$$

So

$$\begin{aligned}
\frac{h}{\sqrt{n}}\phi_p(\boldsymbol{s})\frac{\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})}{\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\|} &= \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\frac{h}{\sqrt{n}}\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\
&\quad + \left\{n^{-1}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{-p}(\boldsymbol{s})\right\}h\sqrt{n}\left[\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_{-p}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_{-p}(\boldsymbol{s})\right] \\
&\quad + \left\{n^{-1}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_p(\boldsymbol{s})\right\}h\sqrt{n}\left[\boldsymbol{\zeta}_p(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_p(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_p(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\right]
\end{aligned} \tag{17}$$

From Lemma 2 of **?**, $\left\{n^{-1}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{-p}(\boldsymbol{s})\right\} = O_p(1)$ and $\left\{n^{-1}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_p(\boldsymbol{s})\right\} =$

$O_p(1)$.

From Theorem 3 of **?**, we have that $h\sqrt{n}\left[\hat{\boldsymbol{\zeta}}_{-p}(\boldsymbol{s}) - \boldsymbol{\zeta}_{-p}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\zeta_{-p}(\boldsymbol{s}) + \nabla^2_{vv}\zeta_{-p}(\boldsymbol{s})\right\}\right] = O_p(1)$ and $h\sqrt{n}\left[\hat{\boldsymbol{\zeta}}_{p}(\boldsymbol{s}) - \boldsymbol{\zeta}_{p}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\zeta_{p}(\boldsymbol{s}) + \nabla^2_{vv}\zeta_{p}(\boldsymbol{s})\right\}\right] = O_p(1)$.

So the second and third terms of the sum in (17) are $O_p(1)$.

We showed in the proof of 1 that $h\sqrt{n}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] = O_p(1)$.

The three terms of the sum to the right of the equals sign in (17) are $O_p(1)$, so for $\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})$ to be a solution, we must have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})/\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| = O_p(1)$.

But since by assumption $\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s}) \neq 0$, there must be some $k \in \{1, \dots, 3\}$ such that $|\hat{\zeta}_{p_k}(\boldsymbol{s})| = \max\{|\hat{\zeta}_{p_{k'}}(\boldsymbol{s})| : 1 \leq k' \leq 3\}$. And for this k, we have that $|\hat{\zeta}_{p_k}(\boldsymbol{s})|/\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| \geq 1/\sqrt{3} > 0$.

Now since $hn^{-1/2}b_n \to \infty$, we have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})/\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| \geq hb_n/\sqrt{3n} \to \infty$ and therefore the term to the left of the equals sign dominates the sum to the right of the equals sign in (17). So for large enough n, $\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s}) \neq 0$ cannot maximize Q.

So $P\left\{\hat{\boldsymbol{\zeta}}_{(b)}(\boldsymbol{s}) = 0\right\} \to 1$.

References