Local Adaptive Grouped Regularization and its Oracle Properties

Wesley Brooks[a], Jun Zhu[b], Zudi Lu[c]

[a]*Department of Statistics, University of Wisconsin, Madison, WI 53706*

[b]*Department of Statistics and Department of Entomology, University of Wisconsin, Madison, WI 53706*

[c]*School of Mathematical Sciences, The University of Southampton Highfield, Southampton UK*

Abstract

Varying coefficient regression is a flexible technique for modeling data where the coefficients are functions of some effect-modifying parameter, often time or location. While there are a number of methods for variable selection in a varying coefficient regression model, the existing methods all do global selection, which includes or excludes each covariate over the entire domain of the effect-modifying parameter. Presented here is local adaptive grouped regularization, a method of local variable selection in varying coefficient regression. This method selects the variables that are associated with the response at a specific point in the domain of the effect-modifying parameter, and simultaneously estimates the coefficients of those covariates. In particular, the method applies the adaptive group Lasso in a local regression model with locally linear coefficient estimates. Oracle properties of the proposed method are established under local linear regression and local generalized linear regression. The method's finite sample properties are assessed in a simulation study. For illustration, the method is used to identify which covariates are associated with house prices in each census tract of the Boton house price data set.

Keywords: Nonparametric, variable selection

Email addresses: wrbrooks@uwalumni.com (Wesley Brooks), jzhu@stat.wisc.edu (Jun Zhu), Z.Lu@soton.ac.uk (Zudi Lu)

1. Introduction

Whereas the coefficients in traditional linear regression are scalar constants, the coefficients in a varying coefficient regression (VCR) model are functions - often *smooth* functions - of some effect-modifying variable (Cleveland and Grosse, 1991; Hastie and Tibshirani, 1993). Here we treat the case of a VCR model on a spatial domain where the spatial locaton is a two-dimensional effect-modifying parameter. Current practice for VCR models relies on global model selection to decide which variables should be included in the model, meaning that covariates are selected for inclusion or exclusion over the entire spatial domain. Various methods have been developed by using, for example, P-splines (Antoniadas et al., 2012), basis expansion (Wang et al., 2008), and local regression (Wang and Xia, 2009). Since the coefficients vary in a VCR model, in principle there is no reason that the best model must use the same set of covariates everywhere on the domain - that is, some of the coefficients may be zero in part of the domain. New methodology is developed here for guiding the decision of which covariates belong in the VCR model at a specific location, or local variable selection, as the literature on how to do so is currently scarce.

Specifically, local adaptive grouped regularization (LAGR) is developed here as a method of local variable selection at any location in the domain of a VCR model. The method of LAGR applies to VCR models where the coefficients are estimated using locally linear kernel smoothing. Using kernel smoothing for nonparametric regression is described in detail in Fan and Gijbels (1996). The extension to estimating VCR models is made by Fan and Zhang (1999) for a VCR with a univariate effect-modifying variable, and by Sun et al. (2014) for a two-dimensional effect-modifying variable and autocorrelation among the obverved response. These methods minimize the boundary effect (Hastie and Loader, 1993) by estimating the coefficients as local polynomials of odd degree (usually locally linear). In this work, we assume a two dimensional effect modifying parameter but changing its dimensionality affects only the rate of convergence.

For standard linear regression models, the least absolute shrinkage and selection operator

(Lasso) is a penalized regression method that simultaneously selects variables for the regression model and shrinks the coefficient estimates toward zero (Tibshirani, 1996). However, the Lasso can be inconsistent for variable selection and inefficient for coefficient estimation (Zou, 2006). The adaptive Lasso (AL) is a refinement of the Lasso that produces consistent estimates of the coefficients and has been shown to have appealing properties for variable selection, which under suitable conditions include the "oracle" property of asymptotically including exactly the correct set of covariates and estimating their coefficients as well as if the correct covariates were known in advance (Zou, 2006). For data where the obvserved variables fall into mutually exclusive groups that are known in advance, the adaptive group Lasso has similar oracle properties to the adaptive Lasso but does selection on groups rather than individual variables (Yuan and Lin, 2006; Wang and Leng, 2008). The main innovation of the proposed LAGR method is to use the adaptive group Lasso for local variable selection and coefficient estimation in a locally linear regression model, where each group consists of a single covariate and its interactions with location. Further, we extend the method to non-Gaussian responses.

We show that for both LAGR posesses the oracle properties of asymptotically selecting exactly the correct local covariates and estimating their local coefficients as accurately as would be possible if the identity of the nonzero coefficients for the local model were known in advance. The remainder of this document is organized as follows. The kernel-based estimation of a VCR model is described in Section 2. The proposed LAGR technique and its oracle properties are presented in Section 3. In Section 4, the performance of the proposed LAGR technique is evaluated in a simulation study, and in Section 5 the proposed method is applied to the Boston house price dataset. In Section 6, LAGR pis extended to varying coefficient generalized linear regression and the oracle properties for this setting are established. Technical proofs are left to the appendix.

2. Varying Coefficient Regression

2.1. Varying Coefficient Model

Consider n data points, observed at sampling locations $\boldsymbol{s}_i = (s_{i,1},\ s_{i,2})^T$ for $i = 1, \ldots, n$, which are distributed in a domain $\mathcal{D} \subset \mathbb{R}^2$ according to a density f. For $i = 1, \ldots, n$, let $y_i = y(\boldsymbol{s}_i)$ and $\boldsymbol{x}_i = \boldsymbol{x}(\boldsymbol{s}_i)$ 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 regression where the coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i)$ are functions in two dimensions and ε_i is random error at location $\boldsymbol{s}_i$. That is,

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_i(\boldsymbol{s}_i) + \varepsilon_i. \tag{1}$$

Further assume that the error term ε_i is normally distributed with zero mean and variance σ^2, and that ε_i, $i = 1, \ldots, n$ are independent. That is, for $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_n)$.

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 (Fan and Gijbels, 1996). Here, to prepare for the estimation of locally linear coefficients, we augment the local design matrix with covariate-by-location interactions in two dimensions (Wang et al., 2008). Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ be the design matrix of observed covariate values. Then the augmented local design matrix at location $\boldsymbol{s}_i$ is defined to be $\boldsymbol{Z}(\boldsymbol{s}_i) = (\boldsymbol{X}\ \ \boldsymbol{L}_i\boldsymbol{X}\ \ \boldsymbol{M}_i\boldsymbol{X})$, where $\boldsymbol{L}_i = \mathrm{diag}\{s_{i',1} - s_{i,1}\}_{i'=1}^n$ and $\boldsymbol{M}_i = \mathrm{diag}\{s_{i',2} - s_{i,2}\}_{i'=1}^n$.

Let $\boldsymbol{Z}_i = \{\boldsymbol{Z}(\boldsymbol{s}_i)\}_i$ denote the ith row of the matrix $\boldsymbol{Z}(\boldsymbol{s}_i)$ as a column vector. Let $\boldsymbol{\zeta}(\boldsymbol{s}_i) = \left(\boldsymbol{\beta}(\boldsymbol{s}_i)^T,\ \nabla_u \boldsymbol{\beta}(\boldsymbol{s}_i)^T,\ \nabla_v \boldsymbol{\beta}(\boldsymbol{s}_i)^T\right)^T$ denote the vector of local coefficients at location $\boldsymbol{s}_i$, augmented with the local gradients of the coefficient surfaces in the two dimensions, denoted ∇_u and ∇_v. Now we have $Y_i = \boldsymbol{Z}_i^T \boldsymbol{\zeta}_i + \varepsilon_i$.

2.2. Coefficient Estimation via Local Likelihood

Let $\boldsymbol{\zeta} = (\boldsymbol{\zeta}(\boldsymbol{s}_1), \ldots, \boldsymbol{\zeta}(\boldsymbol{s}_n))^T$ denote a matrix of the local coefficients at all observation locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. The total log-likelihood of the observed data is the sum of the log-likelihood of each individual observation:

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

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. When the coefficient functions are smooth, though, the coefficients $\boldsymbol{\zeta}(\boldsymbol{s})$ at location $\boldsymbol{s}$ can be approximated by the coefficients $\boldsymbol{\zeta}(\boldsymbol{t})$, where $\boldsymbol{t}$ is within some neighborhood of $\boldsymbol{s}$. This intuition is formalized by the local log-likelihood at location $\boldsymbol{s} \in \mathcal{D}$:

$$\ell(\boldsymbol{\zeta}(\boldsymbol{s})) = -(1/2) \sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|) \left[\log \sigma^2 + \sigma^{-2} \left\{ y_i - \boldsymbol{z}_i^T \boldsymbol{\zeta}(\boldsymbol{s}) \right\}^2 \right] \tag{3}$$

where h is a bandwidth parameter and the $K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)$ for $i = 1, \ldots, n$ are local weights from a kernel function. For instance, the Epanechnikov kernel is defined as (Samiuddin and el Sayyad, 1990):

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

The local log-likelihood at $\boldsymbol{s}$ is maximized to obtain an estimate $\tilde{\boldsymbol{\zeta}}(\boldsymbol{s})$ of the local coefficients. Let $\boldsymbol{W}(\boldsymbol{s}) = \text{diag}\left\{K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)\right\}_{i'=1}^{n}$ denote a diagonal matrix of kernel weights. The local

likelihood can be maximized by minimizing a locally weighted least squares:

$$\mathcal{S}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\},^T \tag{5}$$

and the minimizer is

$$\tilde{\boldsymbol{\zeta}}(\boldsymbol{s}) = \left\{\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1} \boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Y}. \tag{6}$$

By Theorem 3 of Sun et al. (2014), for any given $\boldsymbol{s}$, the estimated local coefficients $\tilde{\boldsymbol{\beta}}(\boldsymbol{s}) = \left(\tilde{\zeta}_1(\boldsymbol{s})^T, \ldots, \tilde{\zeta}_p(\boldsymbol{s})^T\right)^T$ converge in probability to $\boldsymbol{\beta}(\boldsymbol{s}) + 2^{-1}\kappa_0^{-1}\kappa_2 h^2 \left\{\nabla^2_{uu}\boldsymbol{\beta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}(\boldsymbol{s})\right\}$ are asymptotically normally distributed and the true $\boldsymbol{\beta}(\boldsymbol{s})$ at the optimal rate of $O\left(n^{-1/3}\right)$. The estimated local coefficients are asymptotically unbiased, with finite-sample bias proportional to the second derivatives of the true coefficient functions.

3. Local Variable Selection with LAGR

3.1. LAGR-Penalized Local Likelihood

Estimating the local coefficients by (6) relies on *a priori* variable selection. Here we develop a new method of penalized regression to simultaneously select local covariates and estimate the local coefficients. For this purpose, each raw covariate is grouped with its covariate-by-location interactions. That is, $\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) = (\beta_j(\boldsymbol{s}),\ \nabla_u\beta_j(\boldsymbol{s}),\ \nabla_v\beta_j(\boldsymbol{s}))^T$ for $j = 1, \ldots, p$. The proposed LAGR penalty is an adaptive ℓ_1 penalty akin to the adaptive group Lasso (Yuan and Lin, 2006; Wang and Leng, 2008). By the mechanism of the adaptive group Lasso, variables within the same group are included in or dropped from the model together. The intercept group is left unpenalized.

To select and estimate the local coefficients at location $\boldsymbol{s}$, we minimize a penalized local sum of squares at location $\boldsymbol{s}$:

$$\mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \mathcal{S}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) + \mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right),$$

where $\mathcal{S}\left(\boldsymbol{\zeta}\left(\boldsymbol{s}\right)\right)$ is defined in (5), $\mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \sum_{j=1}^{p} \phi_j(\boldsymbol{s}) \|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|$ is a local adaptive grouped regularization (LAGR) penalty, and $\|\cdot\|$ is the L_2-norm. The LAGR penalty for the jth group of coefficients at location $\boldsymbol{s}$ is $\phi_j(\boldsymbol{s}) = \lambda_n \|\tilde{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\|^{-\gamma}$, where $\lambda_n > 0$ is a local tuning parameter applied to all coefficients at location $\boldsymbol{s}$ and $\tilde{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})$ is a subset of the vector of unpenalized local coefficients from (6).

3.2. Oracle Properties

For a local model at location $\boldsymbol{s}$, we let $a_n = \max\{\phi_j(\boldsymbol{s}), j \leq p_0\}$ be the largest penalty applied to a covariate group whose true coefficient norm is nonzero and $b_n = \min\{\phi_j(\boldsymbol{s}), j > p_0\}$ be the smallest penalty applied to a covariate group whose true coefficient norm is zero. Let $\boldsymbol{Z}_{(k)}(\boldsymbol{s})$ be the augmented design matrix for covariate group k, and let $\boldsymbol{Z}_{(\text{-}k)}(\boldsymbol{s})$ be the augmented design matrix for all the data except covariate group k. Similarly, let $\boldsymbol{\zeta}_{(k)}(\boldsymbol{s})$ be the augmented coefficients for covariate group k and $\boldsymbol{\zeta}_{(\text{-}k)}(\boldsymbol{s})$ be the augmented coefficients for all covariate groups except k. Let $\nabla\zeta_k(\boldsymbol{s}) = (\nabla_u\zeta_k(\boldsymbol{s}), \nabla_v\zeta_k(\boldsymbol{s}))^T$ and $\nabla^2\zeta_j(\boldsymbol{s}) = \begin{pmatrix} \nabla^2_{uu}\zeta_k(\boldsymbol{s}) & \nabla^2_{uv}\zeta_k(\boldsymbol{s}) \\ \nabla^2_{vu}\zeta_k(\boldsymbol{s}) & \nabla^2_{vv}\zeta_k(\boldsymbol{s}) \end{pmatrix}$. Let $\kappa_0 = \int_{R^2} K(\|\boldsymbol{s}\|)ds$, $\kappa_2 = \int_{R^2}[(1,0)\boldsymbol{s}]^2 K(\|\boldsymbol{s}\|)ds = \int_{R^2}[(0,1)\boldsymbol{s}]^2 K(\|\boldsymbol{s}\|)ds$, and $\nu_0 = \int_{R^2} K^2(\|\boldsymbol{s}\|)ds$. Finally, let $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution, respectively, as $n \to \infty$.

Assume the following conditions.

(A.1) The kernel function $K(\cdot)$ is bounded, positive, symmetric, and Lipschitz continuous on $\mathbb{R}$, and has bounded support.

(A.2) There are $p_0 < p$ covariates $\boldsymbol{X}_{(a)}(\boldsymbol{s})$ with nonzero local regression coefficients, denoted $\boldsymbol{\beta}_{(a)}(\boldsymbol{s}) \neq \mathbf{0}$. Without loss of generality, assume these are covariates $1, \ldots, p_0$. The remaining $p - p_0$ covariates $\boldsymbol{X}_{(b)}(\boldsymbol{s})$ have true coefficients equal to zero, denoted $\boldsymbol{\beta}_{(b)}(\boldsymbol{s}) = \mathbf{0}$.

(A.3) $\boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)$ are independent random vectors that are independent of $\varepsilon_1, \ldots, \varepsilon_n$. Also $\Psi(\boldsymbol{s}) = E\left\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{X}^T(\boldsymbol{s})|\boldsymbol{s}\right\}$ is positive-definite and differentiable at location $\boldsymbol{s}$, $E\left|\boldsymbol{X}(\boldsymbol{s})\right|^{2q} < \infty$, and $E\left|\varepsilon(\boldsymbol{s})\right|^{2q} < \infty$ for some $q > 2$.

(A.4) The coefficient functions $\beta_j(\cdot)$, $j = 1, \ldots, p$ have continuous second partial derivatives.

(A.6) The function $f(\boldsymbol{s})$ is differentiable at $\boldsymbol{s}$ and $f(\boldsymbol{s}) > 0$.

(A.6) $E\left\{\left|\boldsymbol{X}(\boldsymbol{s})\right|^3 |\boldsymbol{s}\right\}$ is continuous as location $\boldsymbol{s}$.

(A.7) $E\left\{Y(\boldsymbol{s})^4|\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{s}\right\}$ is bounded as location $\boldsymbol{s}$.

(A.8) $h = O\left(n^{-1/6}\right)$

Under these conditions, we obtain the following:

Theorem 1 (Asymptotic normality)**.** *Under (A.1)-(A.8), if* $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$ *and* $hn^{-1/2}b_n \xrightarrow{p} \infty$ *then*

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

where $\left\{\nabla^2_{uu}\boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}_{(a)}(\boldsymbol{s})\right\} = \left(\nabla^2_{uu}\boldsymbol{\beta}_1(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}_1(\boldsymbol{s}), \ldots, \nabla^2_{uu}\boldsymbol{\beta}_{p_o}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}_{p_o}(\boldsymbol{s})\right)^T$.

Theorem 2 (Selection consistency)**.** *Under (A.1)-(A.8), if* $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$ *and* $hn^{-1/2}b_n \xrightarrow{p} \infty$*, then*

$$P\left\{\|\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\| = \mathbf{0}\right\} \to 0 \text{ if } j \leq p_0, \text{ and } P\left\{\|\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\| = \mathbf{0}\right\} \to 1 \text{ if } j > p_0.$$

Together, Theorem 1 and Theorem 2 indicate that the LAGR estimates have the same asymptotic distribution as a local regression model where the true nonzero coefficients are known in advance (Sun et al., 2014), and that the LAGR estimates of true zero coefficients tend to zero with probability one. Thus, selection and estimation by LAGR has the oracle property. To establish 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, $h^{-1}n^{-1/2}\lambda_n \to 0$ for $j \leq p_0$ and $hn^{-1/2}\lambda_n\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|^{-\gamma} \to \infty$ for $j > p_0$. We require that λ_n satisfy both assumptions. Suppose $\lambda_n = n^{\alpha}$. Since $h = O\left(n^{-1/6}\right)$ and $\|\tilde{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\| = O\left(h^{-1}n^{-1/2}\right)$, it follows that $h^{-1}n^{-1/2}\lambda_n = O\left(n^{-1/3+\alpha}\right)$ and $hn^{-1/2}\lambda_n\|\tilde{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-\gamma} = O\left(n^{-2/3+\alpha+\gamma/3}\right)$. Thus, $(2-\gamma)/3 < \alpha < 1/3$, which can only be satisfied for $\gamma > 1$.

3.3. Tuning Parameter Selection

In practical application, it is necessary to select the LAGR tuning parameter λ_n for each local model. A popular approach in other Lasso-type problems is to select the tuning parameter that maximizes a criterion that approximates the expected log-likelihood of a new, independent data set drawn from the same distribution. This is the framework of Mallows' Cp, Stein's unbiased risk estimate (SURE) and Akaike's information criterion (AIC) (Mallows, 1973; Stein, 1981; Akaike, 1973).

These criteria use a so-called covariance penalty to estimate the bias due to using the same data set to select a model and to estimate its parameters (Efron, 2004). We adopt the approximate degrees of freedom for the adaptive group Lasso from Yuan and Lin (2006) and minimize the AICc to select the tuning parameter λ_n (Hurvich et al., 1998). That is, let

$$\hat{df}(\lambda_n; \boldsymbol{s}) = \sum_{j=1}^{p} I\left(\|\hat{\boldsymbol{\zeta}}(\lambda_n; \boldsymbol{s})\| > 0\right) + \sum_{j=1}^{p} \|\hat{\boldsymbol{\zeta}}(\lambda_n; \boldsymbol{s})\| \|\tilde{\boldsymbol{\zeta}}(\boldsymbol{s})\|^{-1}(p_j - 1)$$

$$\begin{aligned}\mathrm{AIC}_c(\lambda_n; \boldsymbol{s}) = &\sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)\sigma^{-2}\left\{y_i - \boldsymbol{z}_i^T\hat{\boldsymbol{\zeta}}(\lambda_n; \boldsymbol{s})\right\}^2 + 2\hat{df}(\lambda_n; \boldsymbol{s}) \\ &+ 2\hat{df}(\lambda_n; \boldsymbol{s})\left\{\hat{df}(\lambda_n; \boldsymbol{s}) + 1\right\}\left\{\sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|) - \hat{df}(\lambda_n; \boldsymbol{s}) - 1\right\}^{-1}\end{aligned}$$

where $I(\cdot)$ is the indicator function and the local coefficient estimate is written $\hat{\boldsymbol{\zeta}}(\lambda_n; \boldsymbol{s})$ to emphasize that it depends on the tuning parameter.

4. Simulation Study

4.1. Simulation Setup

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 $\mathrm{Cov}(X_{ij}, X_{i'j}) = \sigma_x^2 \exp(-\tau_x^{-1}\delta_{ii'})$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0.1$ is the range parameter, and $\delta_{ii'} = \|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|_2$ is the Euclidean distance between locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$.

Correlation was induced between the covariates by multiplying the design matrix $\boldsymbol{X}$ 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 $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon_i$ 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

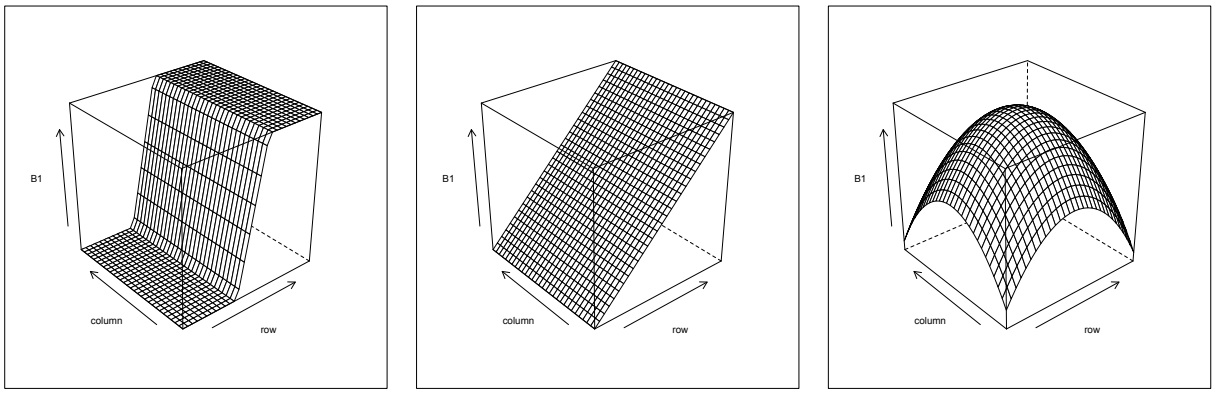

Figure 1: The step function (left), gradient function (middle), and parabola function (right) that were used as the coefficient function $\beta_1(\boldsymbol{s})$ in the VCR model $y_i = x_1(\boldsymbol{s}_i)\beta_1(\boldsymbol{s}_i) + \varepsilon_i$ when generating the data for the simulation study.

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 $\beta_1(\boldsymbol{s})$ (Figure 1). The first is the "step" function $\beta_{\mathbf{step}}(\mathbf{s})$ =1 if $s_x > 0.6$, $5s_x - 2$ if $0.4 < s_x \leq 0.6$, and 0 otherwise. The second is the gradient function, $\beta_{gradient}(\boldsymbol{s}) = s_x$, and the third is the parabola $\beta_{parabola}(\boldsymbol{s}) = 1 - 2\{(s_x - 0.5)^2 + (s_y - 0.5)^2\}$.

In total, three parameters were varied to produce 18 settings, each of which was simulated 100 times. There were the 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 ($\sigma_\varepsilon = 0.5$) and high ($\sigma_\varepsilon = 1$) variance for the random error term.

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 $\hat{\beta}_2(\boldsymbol{s}), \ldots, \hat{\beta}_5(\boldsymbol{s})$ estimated by LAGR were zero. The performance of LAGR was compared to that of a VCR model without variable selection, and to a VCR model with oracle selection. Oracle selection means that exactly the correct set of covariates was used to fit each local model.

Simulation settings			MISE $\hat{\beta}_1$			MISE $\hat{\beta}_2, \ldots, \hat{\beta}_5$	
$\beta_1(\boldsymbol{s})$	ρ	σ_ε	LAGR	VCR	Oracle	LAGR	VCR
step	0	0.5	*0.02*	0.02	**0.01**	**0.00**	0.01
		1.0	*0.03*	0.03	**0.02**	**0.00**	0.02
	0.5	0.5	*0.02*	0.02	**0.01**	**0.00**	0.01
		1.0	*0.03*	0.05	**0.02**	**0.00**	0.03
	0.9	0.5	*0.03*	0.05	**0.01**	**0.00**	0.04
		1.0	*0.12*	0.17	**0.02**	**0.02**	0.15
gradient	0	0.5	0.01	*0.01*	**0.00**	**0.00**	0.00
		1.0	0.03	*0.02*	**0.01**	**0.00**	0.02
	0.5	0.5	0.01	*0.01*	**0.00**	**0.00**	0.01
		1.0	0.04	*0.03*	**0.01**	**0.00**	0.03
	0.9	0.5	*0.03*	0.04	**0.00**	**0.00**	0.04
		1.0	*0.14*	0.14	**0.01**	**0.02**	0.15
parabola	0	0.5	0.01	*0.01*	**0.01**	**0.00**	0.00
		1.0	0.03	*0.02*	**0.02**	**0.00**	0.02
	0.5	0.5	0.01	*0.01*	**0.01**	**0.00**	0.01
		1.0	0.03	*0.03*	**0.02**	**0.00**	0.03
	0.9	0.5	*0.02*	0.04	**0.01**	**0.00**	0.04
		1.0	0.17	*0.14*	**0.02**	**0.03**	0.15

Table 1: For each setting in the simulation study, the mean integrated squared error (MISE) of the coefficient estimates. First, the MISE of $\hat{\beta}_1$ from estimation by local adaptive grouped regularization (LAGR), by locally linear regression without selection (VCR), and by locally linear regression with oracular selection (Oracle). Here, highlighting indicates the **smallest** and *next-smallest* MISE for $\hat{\beta}_1$. Also, the MISE of $\hat{\beta}_2, \ldots, \hat{\beta}_5$ from estimation by LAGR and VCR. Here, highlighting indicates the **smallest** MISE.

4.2. Simulation Results

The MISE of the estimates of $\beta_1(\boldsymbol{s})$ are in Table **??**. Recall that $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$ are exactly zero across the entire domain. Oracle selection will estimate these coefficients perfectly, so we focus on the comparison between estimation by LAGR and by the VCR model with no selection. These results show that for every simulation setting, LAGR estimation is more accurate than the standard VCR model.

From Table 2 we see that LAGR has good ability to identify zero-coefficient covariates. The frequency with which $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$ were dropped from the LAGR models ranged from 0.78 to 0.97. The MISE of the fitted $\hat{y}(\boldsymbol{s})$ is listed in Table 2, where the highlighting is based on which methods estimate an error variance that is closest to the known truth for the simulation. The results are all very similar to each other, indicating that no method was consistently better than the others in this simulation at fitting the model output.

The proposed LAGR method was accurate in selection and estimation, with estimation accuracy for $\beta_1(\boldsymbol{s})$ about equal to that of the VCR model with no selection, and with consistently better accuracy for estimating $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$.

There was minimal difference in the performance of the proposed LAGR method between low ($\sigma_\varepsilon = 0.5$) and high ($\sigma_\varepsilon = 1$) error variance, and between no ($\rho = 0$) and moderate ($\rho = 0.5$) correlation among the covariates. But the selection and estimation accuracy did decline when there was high ($\rho = 0.9$) correlation among the covariates.

5. 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 based on data from the 1970 U.S. census(Harrison and Rubinfeld, 1978; Gilley and Pace, 1996; Pace and Gilley, 1997). The data are the median price of homes sold in 506 census tracts (MEDV), along

Simulation settings			Zero frequency	MISE $\hat{y}$		
$\beta_1(\boldsymbol{s})$	ρ	σ_ε	$\hat{\beta}_2, \ldots, \hat{\beta}_5$	LAGR	VCR	Oracle
step	0	0.5	0.97	*0.25*	0.26	**0.25**
		1.0	0.96	*1.00*	**1.00**	0.99
	0.5	0.5	0.96	*0.26*	0.26	**0.25**
		1.0	0.92	*0.99*	**1.00**	0.98
	0.9	0.5	0.86	*0.27*	0.30	**0.25**
		1.0	0.85	*1.08*	1.14	**0.98**
gradient	0	0.5	0.96	*0.25*	**0.25**	0.25
		1.0	0.95	**0.99**	*0.99*	0.97
	0.5	0.5	0.94	*0.25*	**0.25**	0.24
		1.0	0.92	*1.00*	**1.00**	0.97
	0.9	0.5	0.80	*0.27*	0.28	**0.24**
		1.0	0.85	*1.09*	1.12	**0.97**
parabola	0	0.5	0.97	*0.25*	**0.25**	0.25
		1.0	0.94	**1.00**	*1.00*	0.98
	0.5	0.5	0.95	**0.25**	0.25	*0.25*
		1.0	0.88	**1.00**	*1.00*	0.97
	0.9	0.5	0.79	*0.26*	0.28	**0.24**
		1.0	0.78	1.13	*1.12*	**0.98**

Table 2: For each setting of the simulation study, the frequency of exact zeroes in the estimates of $\hat{\beta}_2, \ldots, \hat{\beta}_5$ as estimated by local adaptive grouped regularizaton (LAGR). Also, the mean integrated squared error (MISE) for the fitted output of each simulation setting, under variable selection via LAGR, locally linear regression without selection (VCR), and by locally linear regression with oracular selection (Oracle). Highlighting indicates the **closest** and *next-closest* to the actual error variance σ_ε^2 for that simulation setting.

	Mean	SD	Prop. zero
CRIM	-0.07	0.08	0.49
RM	1.92	1.43	0.02
RAD	-0.08	0.13	0.37
TAX	0.00	0.00	1.00
LSTAT	-0.72	0.16	0.01

Table 3: The mean, standard deviation, and proportion of zeros among the local coefficients in a model for the median house price in census tracts in Boston, with coefficients selected and fitted by LAGR.

with the potential covariates 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 $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.

A summary of the local coefficients is in Table 3. It indicates that RM is the only covariate with a positive mean of the local coefficients. The coefficient of the CRIM variable was estimated to be exactly zero at 49% of the locations. The percentage for the RAD variable was 37%.

Estimates of the regression coefficients are plotted in Figure 2. One interesting result is that the TAX variable was nowhere found to be an important predictor of the median house price. Another is that the coefficients of CRIM and LSTAT are everywhere negative or zero (meaning that a greater crime rate or proportion of lower-status individuals is associated with a lower median house price where the effect is discernable) and that of RM is positive (meaning that a greater average number of rooms per house is associated with a greater median house price). The coefficient of RAD is positive in some areas and negative in others. This indicates that there are parts of Boston where access to radial roads is associated with a greater median house price and parts where it is associated with a lesser median house price.

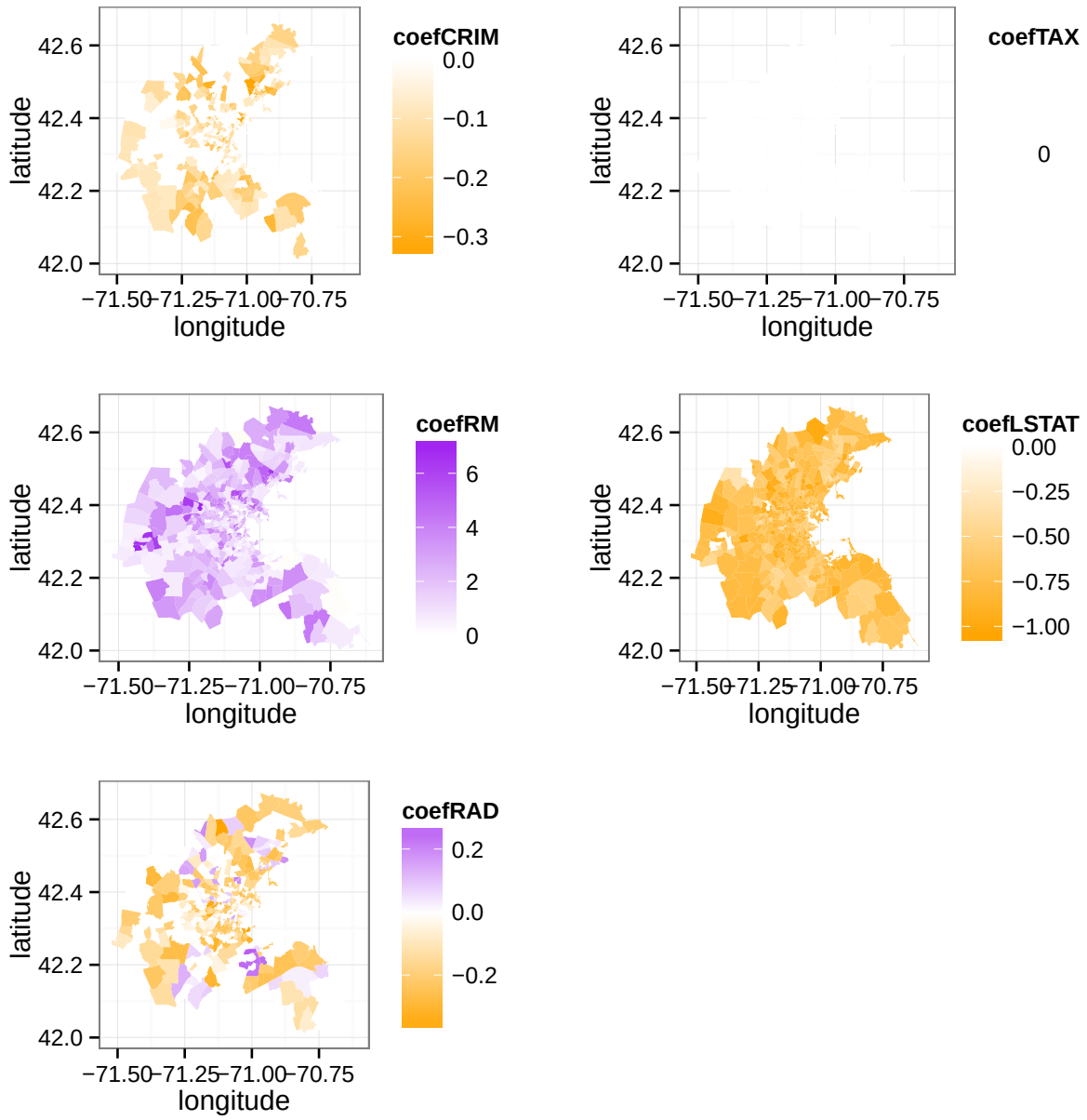

Figure 2: Coefficients for the Boston house price data as estimated by local adaptive grouped regularization.

In their example using the same data, Sun et al. (2014) estimated that the coefficients of RAD annd LSTAT should be constant, at 0.36 and -0.45, respectively. That conclusion differs from our result, which says that the mean local coefficient of RAD is actually negative (-0.08), while our mean fitted local coefficient for LSTAT was more negative than the estimate of Sun et al. (2014).

6. Extension to Generalized Linear Regression

6.1. Model

Generalized linear models (GLMs) extend the linear regression model to a response variable following any distribution in a single-parameter exponential family (McCullagh and Nelder, 1989). As was the case for the local linear regression model, local generalized GLM coefficients are smooth functions of location. If the response variable y is from an exponential-family distribution then its density is

$$f\left(y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s}), \theta(\boldsymbol{s}0) = c\left(y(\boldsymbol{s}0) \times \exp\left[\theta(\boldsymbol{s})y(\boldsymbol{s}) - b\left(\theta(\boldsymbol{s})\right)\right]\right.\right.$$

where ϕ and θ are parameters, $E\left\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right\} = \mu(\boldsymbol{s}) = b'\left(\theta(\boldsymbol{s})\right)$, $\theta(\boldsymbol{s}) = (g \circ b')^{-1}\left(\eta(\boldsymbol{s})\right)$, $\eta(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta}(\boldsymbol{s}) = g\left(\mu(\boldsymbol{s})\right)$, and $\text{Var}\left\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\right\} = b''\left(\theta(\boldsymbol{s})\right)$. 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 g is called the canonical link.

6.2. Coefficient Estimation via 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 local coefficients can be estimated by maximizing the local quasi-likelihood

$$\ell^*\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \sum_{i=1}^{n} K_h\left(\|\boldsymbol{s}-\boldsymbol{s}_i\|\right) Q\left(g^{-1}\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s})\right), Y_i\right), \tag{7}$$

where $\boldsymbol{Z}(\boldsymbol{s})$ and $\boldsymbol{\zeta}(\boldsymbol{s})$ are defined in (**??**) and (**??**). 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 $(\partial/\partial\mu)\,Q(\mu, y) = (y-\mu)\{V(\mu)\}^{-1}$. The local quasi-likelihood is maximized by setting the local quasi-score equation to zero:

$$(\partial/\partial\boldsymbol{\zeta})\ell^*\left(\hat{\boldsymbol{\zeta}}(\boldsymbol{s})\right) = \sum_{i=1}^{n} K_h\left(\|\boldsymbol{s}-\boldsymbol{s}_i\|\right)\left(y_i - \hat{\mu}(\boldsymbol{s}_i;\boldsymbol{s})\right)\left\{V\left(\hat{\mu}(\boldsymbol{s}_i;\boldsymbol{s})\right)\right\}^{-1}\boldsymbol{z}_i = \mathbf{0}_{3p}, \tag{8}$$

where $\hat{\mu}(\boldsymbol{s}_i;\boldsymbol{s}) = g^{-1}\left(\boldsymbol{z}_i^T\hat{\boldsymbol{\zeta}}(\boldsymbol{s})\right)$ is the mean at location $\boldsymbol{s}_i$ estimated using the coefficients $\hat{\boldsymbol{\zeta}}(\boldsymbol{s})$ fitted at location $\boldsymbol{s}$. The asymptotic distribution of the local coefficients in a varying-coefficient GLM with a one-dimensional effect-modifying parameter are given in Cai et al. (2000). For coefficients that vary in two dimensions, it follows from Lemmas 1 and 2 that the distribution of the estimated local coefficients is:

$$\left\{nh^2 f(\boldsymbol{s})\right\}^{1/2}\left[\tilde{\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{\beta}(\boldsymbol{s}) - (1/2)\kappa_0^{-1}\kappa_2 h^2\left\{\nabla_{uu}^2\boldsymbol{\beta}(\boldsymbol{s}) + \nabla_{vv}^2\boldsymbol{\beta}(\boldsymbol{s})\right\}\right] \xrightarrow{D} N\left(\mathbf{0}, \kappa_0^{-2}\nu_0\Gamma^{-1}(\boldsymbol{s})\right).$$

6.3. LAGR Penalized Local Likelihood

As in the case of linear models, the LAGR for GLMs is a grouped ℓ_1 regularization method. Now, though, we use a penalized local quasi-likelihood:

$$\mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \ell^*\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) + \mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) \tag{9}$$

$$= \sum_{i=1}^{n} K_h\left(\|\boldsymbol{s} - \boldsymbol{s}_i\|\right) Q\left(g^{-1}\left(\boldsymbol{z}_i^T \boldsymbol{\zeta}(\boldsymbol{s})\right), Y_i\right) + \sum_{j=1}^{p} \phi_j(\boldsymbol{s}) \|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|. \tag{10}$$

As in the case of Gaussian data, let $\phi_j(\boldsymbol{s}) = \lambda_n \|\tilde{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\|^{-\gamma}$, where $\lambda_n > 0$ 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.

6.4. Oracle properties of LAGR in the GLM setting

The following are additional to the definitions and assumptions of Section 3.2:

Define $\rho(\boldsymbol{s}, \boldsymbol{z}) = \left[g_1\left(\mu(\boldsymbol{s}, \boldsymbol{z})\right)\right]^2 Var\left\{Y(\boldsymbol{s}) | \boldsymbol{X}(\boldsymbol{s}), \boldsymbol{s}\right\}$, where $g_1(\cdot) = g_0'(\cdot)/g'(\cdot)$, and $g_0(\cdot)$ is the canonical link function. So when the canonical link is used, $\rho(\boldsymbol{s}, \boldsymbol{z}) = V\left(\mu(\boldsymbol{s}, \boldsymbol{z})\right)$. Let $\Gamma\left(\boldsymbol{s}\right) = E\left\{\rho\left(\boldsymbol{s}, \boldsymbol{X}(\boldsymbol{s})\right) \boldsymbol{X}(\boldsymbol{s}) \boldsymbol{X}(\boldsymbol{s})^T | \boldsymbol{s}, \boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{z}\right\}$.

Assume the following conditions.

(A.9) The functions $g'''\left(\boldsymbol{s}\right)$, $\nabla\Gamma\left(\boldsymbol{s}\right)$, $V\left(\mu\left(\boldsymbol{s}, \boldsymbol{z}\right)\right)$, and $V'\left(\mu\left(\boldsymbol{s}, \boldsymbol{z}\right)\right)$ are continuous at $\boldsymbol{s}$.

(A.10) The function $\left(\partial^2/\partial\mu^2\right) Q\left(g^{-1}\left(\mu\right), y\right) < 0$ for $\mu \in \mathbb{R}$ and y in the range of the response.

Theorem 3 (Asymptotic normality)**.** *Under (A.1)-(A.10), if* $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$ *and* $hn^{-1/2}b_n \xrightarrow{p} \infty$ *then*

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

Theorem 4 (Selection consistency)**.** *Under (A.1)-(A.10), if* $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$ *and* $hn^{-1/2}b_n \xrightarrow{p} \infty$ *then*

$$P\left\{\|\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\| = \mathbf{0}\right\} \to 0 \text{ if } j \leq p_0, \text{ and } P\left\{\|\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\| = \mathbf{0}\right\} \to 1 \text{ if } j > p_0.$$

7. References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki (Eds.), *2nd International Symposium on Information Theory*, pp. 267–281.

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

Cai, Z., J. Fan, and R. Li (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association 95*, 888–902.

Cleveland, W. and E. Grosse (1991). Local regression models. In J. Chambers and T. Hastie (Eds.), *Statistical models in S*. Wadsworth and Brooks/Cole.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association 99*, 619–632.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

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

Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *Annals of Statistics 22*, 1993–2010.

Gilley, O. and R. K. Pace (1996). On the harrison and rubinfeld data. *Journal of Environmental Economics and Management 31*, 403–405.

Harrison, D. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management 5*, 81–102.

Hastie, T. and C. Loader (1993). Local regression: automatic kernel carpentry. *Statistical Science 8*, 120–143.

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

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society Series B 60*, 271–293.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Mallows, C. (1973). Some comments on cp. *Technometrics 15*, 661–675.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, Second Edition*. Taylor and Francis.

Pace, R. K. and O. Gilley (1997). Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics 14*, 333–340.

Samiuddin, M. and G. M. el Sayyad (1990). On nonparametrickernel density estimates. *Biometrika 77*, 865–874.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics 9*, 1135–1151.

Sun, Y., H. Yan, W. Zhang, and Z. Lu (2014). A semiparametric spatial dynamic model. *Annals of Statistics 42*, 700–727.

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

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis 52*, 5277–5286.

Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*, 747–757.

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*, 1556–1569.

Wang, N., C.-L. Mei, and X.-D. Yan (2008). Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. *Environment and Planning A 40*, 986–1005.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B 68*, 49–67.

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

Appendix: Proof of Theorem 1

Proof. Let $H_n(\boldsymbol{u}) = \mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right) - \mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right)$ and $\alpha_n = h^{-1}n^{-1/2}$. Then, we have

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

The limiting behavior of the last term differs between the cases $j \le p_0$ and $j > p_0$. *Case* $j \le p_0$: If $j \le p_0$, then $n^{-1/2}\phi_j(\boldsymbol{s}) \to n^{-1/2}\lambda_n\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|^{-\gamma}$ and $|n^{1/2}\left\{\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) + \alpha_n\boldsymbol{u}_{(j)}\| - \|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right\}| \le h^{-1}\|\boldsymbol{u}_{(j)}\|$. Thus,

$$\phi_j(\boldsymbol{s})\left(\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) + \alpha_n\boldsymbol{u}_{(j)}\| - \|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right) \le \alpha_n\phi_j(\boldsymbol{s})\|\boldsymbol{u}_{(j)}\| \le \alpha_n a_n\|\boldsymbol{u}_{(j)}\| \to 0$$

Case $j > p_0$: If $j > p_0$, then $\phi_j(\boldsymbol{s})\left(\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) + \alpha_n\boldsymbol{u}_{(j)}\| - \|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right) = \phi_j(\boldsymbol{s})\alpha_n\|\boldsymbol{u}_{(j)}\|$. Since $h = O(n^{-1/6})$, if $hn^{-1/2}b_n \xrightarrow{p} \infty$, then $\alpha_n b_n \xrightarrow{p} \infty$. Thus, if $\|\boldsymbol{u}_{(j)}\| \ne 0$, then

$$\alpha_n\phi_j(\boldsymbol{s})\|\boldsymbol{u}_{(j)}\| \ge \alpha_n b_n\|\boldsymbol{u}_{(j)}\| \to \infty.$$

On the other hand, if $\|\boldsymbol{u}_{(j)}\| = 0$, then $\alpha_n\phi_j(\boldsymbol{s})\|\boldsymbol{u}_{(j)}\| = 0$. Thus, the limit of $H_n(\boldsymbol{u})$ is the same as the limit of $H_n^*(\boldsymbol{u})$ where $H_n^*(\boldsymbol{u}) = \infty$ if $\|\boldsymbol{u}_{(j)}\| \ne 0$ for some $j > p_0$, and

$$H_n^*(\boldsymbol{u}) = (1/2)\alpha_n^2\boldsymbol{u}^T\left\{\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}\boldsymbol{u} - \alpha_n\boldsymbol{u}^T\left[\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right]$$

otherwise. It follows that $H_n^*(\boldsymbol{u})$ is convex and its unique minimizer is

$$\hat{\boldsymbol{u}}_n = \left\{n^{-1}\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1}\left[hn^{1/2}\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y}-\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right].$$

By epiconvergence, the minimizer of the limiting function is the limit of the minimizers $\hat{\boldsymbol{u}}_n$ (Geyer, 1994; Knight and Fu, 2000). Since, by Lemma 2 of Sun et al. (2014),

$$\hat{\boldsymbol{u}}_n - \left(2\alpha_n f(\boldsymbol{s})^{1/2}\kappa_0\right)^{-1}\kappa_2 h^2\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\} \xrightarrow{d} N\left(0, \alpha_n^{-2}f(\boldsymbol{s})^{-1}\kappa_0^{-2}\nu_0\sigma^2\Psi(\boldsymbol{s})^{-1}\right)$$

the result of Theorem 1 follows. □

Appendix: Proof of Theorem 2

Proof. We showed in Theorem 1 that $\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s}) \xrightarrow{p} \boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+(2\kappa_0)^{-1}\kappa_2 h^2\nabla^2_{uu}\left\{\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\right\}$, so to complete the proof of selection consistency, it only remains to show that $P\left\{\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})=\boldsymbol{0}\right\}\to 1$ 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 $\mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right)$ is differentiable w.r.t. $\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})$ and is minimized where

$$\begin{aligned}
\boldsymbol{0} =& \boldsymbol{Z}^T_{(p)}(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y}-\boldsymbol{Z}_{(-p)}\left(\boldsymbol{s}\right)\hat{\boldsymbol{\zeta}}_{(-p)}(\boldsymbol{s})-\boldsymbol{Z}_{(p)}\left(\boldsymbol{s}\right)\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\right\}-\phi_{(p)}(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1}\\
=& \boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y}-\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})-(2\kappa_0)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right]\\
&+\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{(-p)}(\boldsymbol{s})\left[\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s})+(2\kappa_0)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s})\right\}-\hat{\boldsymbol{\zeta}}_{(-p)}(\boldsymbol{s})\right]\\
&+\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{(p)}(\boldsymbol{s})\left[\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})+(2\kappa_0)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})\right\}-\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\right]\\
&-\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1}
\end{aligned}$$

Thus,

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

From Lemma 2 of Sun et al. (2014),

$$O_p\left(n^{-1}\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\,\boldsymbol{Z}_{(\text{-}p)}(\boldsymbol{s})\right) = O_p\left(n^{-1}\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{(p)}(\boldsymbol{s})\right) = O_p\left(1\right).$$

From Theorem 3 of Sun et al. (2014), we have that

$$\left(nh^2\right)^{1/2}\left[\hat{\boldsymbol{\zeta}}_{(\text{-}p)}(\boldsymbol{s}) - \boldsymbol{\zeta}_{(\text{-}p)}(\boldsymbol{s}) - \left(2\kappa_0\right)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\zeta_{(\text{-}p)}(\boldsymbol{s}) + \nabla^2_{vv}\zeta_{(\text{-}p)}(\boldsymbol{s})\right\}\right] = O_p\left(1\right)$$

and

$$\left(nh^2\right)^{1/2}\left[\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s}) - \boldsymbol{\zeta}_{(p)}(\boldsymbol{s}) - \left(2\kappa_0\right)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})\right\}\right] = O_p\left(1\right).$$

We showed in the proof of Theorem 1 that

$$\left(nh^2\right)^{1/2}\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s}) - \left(2\kappa_0\right)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] = O_p\left(1\right).$$

The right hand side of (.1) is $O_p(1)$, so for $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})$ to be a solution, we must have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} = O_p\left(1\right)$. But since by assumption $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s}) \neq \boldsymbol{0}$, there must be some $k \in \{1,2,3\}$ such that $|\hat{\zeta}_{(p)_k}(\boldsymbol{s})| = \max\{|\hat{\zeta}_{(p)_m}(\boldsymbol{s})| : 1 \leq m \leq 3\}$. And for this k, we have that $|\hat{\zeta}_{(p)_k}(\boldsymbol{s})|\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} \geq 3^{-1/2} > 0$. Since $hn^{-1/2}b_n \to \infty$, we have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} \geq hb_n\left(3n\right)^{-1/2} \to \infty$ and therefore the left hand side of (.1)

dominates the sum to the right side. Thus, for large enough n, $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s}) \neq \mathbf{0}$ cannot maximize $\mathcal{J}(\cdot)$, and therefore $P\left\{\hat{\boldsymbol{\zeta}}_{(b)}(\boldsymbol{s}) = \mathbf{0}\right\} \to 1$. □

Appendix: Lemmas

The next proofs require the following lemmas. First, let $\boldsymbol{z} \in \mathbb{R}^{3p}$. Define the q-functions to be the derivatives of the quasi-likelihood: $q_j(t,y) = (\partial/\partial t)^j Q\left(g^{-1}(t), y\right)$. Then $q_1\left(\eta\left(\boldsymbol{s},\boldsymbol{z}\right), \mu\left(\boldsymbol{s},\boldsymbol{z}\right)\right) = \mathbf{0}$, and $q_2\left(\eta\left(\boldsymbol{s},\boldsymbol{z}\right), \mu\left(\boldsymbol{s},\boldsymbol{z}\right)\right) = -\rho\left(\boldsymbol{s},\boldsymbol{z}\right)$. Let $\tilde{\boldsymbol{\beta}}_i'' = \Big[(\boldsymbol{s}_i - \boldsymbol{s})^T \left\{\nabla^2\beta_1(\boldsymbol{s})\right\}(\boldsymbol{s}_i - \boldsymbol{s}), \ldots, (\boldsymbol{s}_i - \boldsymbol{s})^T \left\{\nabla^2\beta_p(\boldsymbol{s})\right\}($ be the p-vector of quadratic forms of location interactions on the second derivatives of the coefficient functions.

Lemma 1.

$$E\left[\sum_{i=1}^{n} q_1\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)\boldsymbol{Z}_i K_h\left(\|\boldsymbol{s}-\boldsymbol{s}_i\|\right)\right] = \begin{pmatrix} 2^{-1}n^{1/2}h^3 f(\boldsymbol{s})\kappa_2\Gamma(\boldsymbol{s})\left(\nabla^2_{uu}\boldsymbol{\beta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}(\boldsymbol{s})\right)^T \\ \mathbf{0}_{2p} \end{pmatrix} + o_p\left(h^2\mathbf{1}_{3p}\right)$$

and

$$\begin{aligned} Var\left[\sum_{i=1}^{n} q_1\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)\boldsymbol{Z}_i K_h\left(\|\boldsymbol{s}-\boldsymbol{s}_i\|\right)\right] &= f(\boldsymbol{s})\mathrm{diag}\left\{\nu_0,\nu_2,\nu_2\right\}\otimes\Gamma(\boldsymbol{s}) + o(1) \\ &= \Lambda + o(1) \end{aligned}$$

Lemma 2.

$$\begin{aligned} E\left[\sum_{i=1}^{n} q_2\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)\boldsymbol{Z}_i\boldsymbol{Z}_i^T K_h\left(\|\boldsymbol{s}-\boldsymbol{s}_i\|\right)\right] &= -f(\boldsymbol{s})\mathrm{diag}\left\{\kappa_0,\kappa_2,\kappa_2\right\}\otimes\Gamma(\boldsymbol{s}) + o(1) \\ &= -\Delta + o(1) \end{aligned}$$

and

$$Var\left\{\left(\sum_{i=1}^{n} q_2\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)\boldsymbol{Z}_i\boldsymbol{Z}_i^T K_h\left(\|\boldsymbol{s}-\boldsymbol{s}_i\|\right)\right)_{ij}\right\} = O\left(n^{-1}h^{-2}\right)$$

Appendix: Proof of Theorem 3

Proof. Let $H_n'(\boldsymbol{u}) = \mathcal{J}^*\left(\boldsymbol{\zeta}(\boldsymbol{s})+\alpha_n\boldsymbol{u}\right) - \mathcal{J}^*\left(\boldsymbol{\zeta}(\boldsymbol{s})\right)$ and $\alpha_n = h^{-1}n^{-1/2}$. Then, maximixing $H_n'(\boldsymbol{u})$ is equivalent to maximizing $H_n(\boldsymbol{u})$, where

$$\begin{aligned}
H_n(\boldsymbol{u}) =& n^{-1}\sum_{i=1}^{n} Q\left(g^{-1}\left(\boldsymbol{Z}_i^T\left\{\boldsymbol{\zeta}(\boldsymbol{s})+\alpha_n\boldsymbol{u}\right\}\right), Y_i\right) K\left(h^{-1}\|\boldsymbol{s}-\boldsymbol{s}_i\|\right) \\
&- n^{-1}\sum_{i=1}^{n} Q\left(g^{-1}\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s})\right), Y_i\right) K\left(h^{-1}\|\boldsymbol{s}-\boldsymbol{s}_i\|\right) \\
&+ n^{-1}\sum_{j=1}^{p}\phi_j\left(\boldsymbol{s}\right)\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+\alpha_n\boldsymbol{u}\| - \sum_{j=1}^{p}\phi_j\left(\boldsymbol{s}\right)\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|
\end{aligned}$$

Define

$$\begin{aligned}
\Omega_n =& \alpha_n\sum_{i=1}^{n} q_1\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)\boldsymbol{Z}_i K\left(h^{-1}\|\boldsymbol{s}-\boldsymbol{s}_i\|\right) \\
=& \alpha_n\sum_{i=1}^{n}\omega_i
\end{aligned}$$

and

$$\begin{aligned}
\Delta_n =& \alpha_n^2\sum_{i=1}^{n} q_2\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)\boldsymbol{Z}_i\boldsymbol{Z}_i^T K\left(h^{-1}\|\boldsymbol{s}-\boldsymbol{s}_i\|\right) \\
=& \alpha_n^2\sum_{i=1}^{n}\delta_i
\end{aligned}$$

Then it follows from the Taylor expansion of $\mathcal{J}^*\left(\boldsymbol{\zeta}(\boldsymbol{s})+\alpha_n\boldsymbol{u}\right)$ around $\boldsymbol{\zeta}(\boldsymbol{s})$ that

$$\begin{aligned} H_n\left(\boldsymbol{u}\right) =& \Omega_n^T\boldsymbol{u} \\ &+ (1/2)\boldsymbol{u}^T\Delta_n\boldsymbol{u} \\ &+ \left(\alpha_n^3/6\right)\sum_{i=1}^{n} q_3\left(\boldsymbol{Z}_i^T\tilde{\boldsymbol{\zeta}}_i, Y_i\right)\left[\boldsymbol{Z}_i^T\boldsymbol{u}\right]^3 K\left(h^{-1}\|\boldsymbol{s}-\boldsymbol{s}_i\|\right) \\ &+ \sum_{j=1}^{p}\phi_j\left(\boldsymbol{s}\right)\left\{\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+h^{-1}n^{-1/2}\boldsymbol{u}\|-\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right\}. \end{aligned} \tag{.2}$$

where $\tilde{\boldsymbol{\zeta}}_i$ lies between $\boldsymbol{\zeta}(\boldsymbol{s})$ and $\boldsymbol{\zeta}(\boldsymbol{s})+\alpha_n\boldsymbol{u}$. Since $q_3\left(\boldsymbol{Z}_i^T\tilde{\boldsymbol{\zeta}}_i, Y_i\right)$ is linear in Y_i, $K\left(\cdot\right)$ is bounded, and, by condition (A.6),

$$\left(\alpha_n^3/6\right)E\left|\sum_{i=1}^{n} q_3\left(\boldsymbol{Z}_i^T\tilde{\boldsymbol{\zeta}}_i, Y_i\right)\left[\boldsymbol{Z}_i^T\boldsymbol{u}\right]^3 K\left(h^{-1}\|\boldsymbol{s}-\boldsymbol{s}_i\|\right)\right| = O\left(\alpha_n\right),$$

the third term in (.2) is $O_p\left(\alpha_n\right)$. The limiting behavior of the last term of (.2) differs between the cases $j\leq p_0$ and $j>p_0$. *Case $j\leq p_0$:* If $j\leq p_0$, then $n^{-1/2}\phi_j(\boldsymbol{s})\to n^{-1/2}\lambda_n\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|^{-\gamma}$ and $|\sqrt{n}\left\{\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+\alpha_n\boldsymbol{u}_{(j)}\|-\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right\}|\leq h^{-1}\|\boldsymbol{u}_{(j)}\|$. Thus,

$$\lim_{n\to\infty}\phi_j(\boldsymbol{s})\left(\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+\alpha_n\boldsymbol{u}_{(j)}\|-\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right)\leq\alpha_n\phi_j(\boldsymbol{s})\|\boldsymbol{u}_{(j)}\|\leq\alpha_n a_n\|\boldsymbol{u}_{(j)}\|\to 0$$

Case $j>p_0$: If $j>p_0$, then $\phi_j(\boldsymbol{s})\left(\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+\alpha_n\boldsymbol{u}_{(j)}\|-\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right)=\phi_j(\boldsymbol{s})\alpha_n\|\boldsymbol{u}_{(j)}\|$. Since $h=O(n^{-1/6})$, if $hn^{-1/2}b_n\xrightarrow{p}\infty$, then $\alpha_n b_n\xrightarrow{p}\infty$. Now, if $\|\boldsymbol{u}_{(j)}\|\neq 0$, then

$$\alpha_n\phi_j(\boldsymbol{s})\|\boldsymbol{u}_{(j)}\|\geq\alpha_n b_n\|\boldsymbol{u}_{(j)}\|\to\infty.$$

On the other hand, if $\|\boldsymbol{u}_{(j)}\|=0$, then $\alpha_n\phi_j(\boldsymbol{s})\|\boldsymbol{u}_{(j)}\|=0$. By Lemma 2, $\Delta_n=\Delta+O_p\left(\alpha_n\right)$, so the limit of $H_n(\boldsymbol{u})$ is the same as the limit of $H_n^*(\boldsymbol{u})$ where

$$H_n^*(\boldsymbol{u})=\Omega_n^T\boldsymbol{u}+(1/2)\boldsymbol{u}^T\Delta\boldsymbol{u}+o_p\left(1\right)$$

if $\|\boldsymbol{u}_j\| = 0\ \forall j > p_0$, and $H_n^*(\boldsymbol{u}) = \infty$ otherwise. It follows that $H_n^*(\boldsymbol{u})$ is convex and its unique minimizer is

$$\hat{\boldsymbol{u}}_n = \Delta^{-1}\Omega_n + o_p(1)$$

by the quadratic approximation lemma (Fan and Gijbels, 1996). And by epiconvergence, the minimizer of the limiting function is the limit of the minimizers $\hat{\boldsymbol{u}}_n$ (Geyer, 1994; Knight and Fu, 2000). Since Δ is a constant, the normality of $\hat{\boldsymbol{u}}_n$ follows from the normality of Ω_n, which is establised via the Cramér-Wold device. Let $\boldsymbol{d} \in \mathbb{R}^{3p}$ be a unit vector, and let

$$\xi_i = q_1\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)\boldsymbol{d}^T\boldsymbol{Z}_i K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right).$$

Then $\boldsymbol{d}^T\Omega_n = \alpha_n\sum_{i=1}^n \xi_i$. We establish the normality of $\boldsymbol{d}^T\Omega_n$ by checking the Lyapunov condition of the sequence $\left\{\boldsymbol{d}^T Var\left(\Omega_n\right)\boldsymbol{d}\right\}^{-1/2}\left\{\boldsymbol{d}^T\Omega_n - \boldsymbol{d}^T E\Omega_n\right\}$. By boundedness of $K(\cdot)$, linearity of $q_1\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right)$ in Y_i, and assumptions (A.6), (A.7), and (A.9), we have that

$$n\alpha_n^3 E\left(|\xi_1|^3\right) = O\left(\alpha_n\right) \to 0. \tag{.3}$$

We observe that (.3) implies that $n\alpha_n^3\left|E\left(\xi_1\right)\right|^3 \to 0$, and since $E\left(|\xi_1 - E\xi_1|^3\right) < E\left\{\left(|\xi_1| + |E\xi_1|\right)^3\right\} \to 0$, the Lyapunov condition is satisfied. Thus, Ω_n asymptotically follows a Gaussian distribution and the result follows from the quadratic approximation lemma. □

Appendix: Proof of Theorem 4

Proof. We showed in Theorem 3 that $\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s}) \xrightarrow{p} \boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) + (2\kappa_0)^{-1}\kappa_2 h^2\{\nabla_{uu}^2\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) + \nabla_{vv}^2\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\}$, so to complete the proof of selection consistency, it only remains to show that $P\left\{\hat{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s}) = \boldsymbol{0}\right\} \to 1$ 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 $\mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right)$ is differentiable w.r.t. $\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})$ and is

minimized where

$$\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} = \sum_{i=1}^{n} q_1\left(\boldsymbol{Z}_i^T\hat{\boldsymbol{\zeta}}(\boldsymbol{s}), Y_i\right)\boldsymbol{Z}_{i(p)}K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) \tag{.4}$$

From Lemma 1, the right hand side of (.4) is $O_p(1)$, so for $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})$ to be a solution, we must have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} = O_p(1)$. But since by assumption $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s}) \neq \mathbf{0}$, there must be some $k \in \{1,2,3\}$ such that $|\hat{\zeta}_{(p)_k}(\boldsymbol{s})| = \max\{|\hat{\zeta}_{(p)_m}(\boldsymbol{s})| : 1 \leq m \leq 3\}$. And for this k, we have that $|\hat{\zeta}_{(p)_k}(\boldsymbol{s})|\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} \geq 3^{-1/2} > 0$. Since $hn^{-1/2}b_n \to \infty$, we have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} \geq hb_n(3n)^{-1/2} \to \infty$ and therefore the left hand side of (.4) dominates the sum to the right side. Thus, for large enough n, $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s}) \neq \mathbf{0}$ cannot maximize $\mathcal{J}(\cdot)$, and therefore $P\left\{\hat{\boldsymbol{\zeta}}_{(b)}(\boldsymbol{s}) = \mathbf{0}\right\} \to 1$. □

Appendix: Proof of Lemma 1

Proof. **Expectation**: For $j = 1, \ldots, p$, by a Taylor expansion of $\beta_j(\boldsymbol{s}_i)$ around $\boldsymbol{s}$,

$$\beta_j(\boldsymbol{s}_i) = \beta_j(\boldsymbol{s}) + \nabla\beta_j(\boldsymbol{s})(\boldsymbol{s}_i - \boldsymbol{s}) + (\boldsymbol{s}_i - \boldsymbol{s})^T\left\{\nabla^2\beta_j(\boldsymbol{s})\right\}(\boldsymbol{s}_i - \boldsymbol{s}) + o\left(h^2\right)$$

and thus, for $\boldsymbol{x} \in \mathbb{R}^p$,

$$\boldsymbol{x}_i^T\boldsymbol{\beta}(\boldsymbol{s}_i) = \sum_{j=1}^{p} x_{ij}\left[\beta_j(\boldsymbol{s}) + \nabla\beta_j(\boldsymbol{s})^T(\boldsymbol{s}_i - \boldsymbol{s}) + \tilde{\beta}_{ij}''\right] + o\left(h^2\right).$$

Letting $\boldsymbol{z}_i^T = \left\{(1, s_{i,1} - s_1, s_{i,2} - s_2) \otimes \boldsymbol{x}_i^T\right\}$ and $\boldsymbol{\zeta}(\boldsymbol{s}) = \left(\boldsymbol{\beta}(\boldsymbol{s})^T, \nabla_u\boldsymbol{\beta}(\boldsymbol{s})^T, \nabla_v\boldsymbol{\beta}(\boldsymbol{s})^T\right)^T$, we have that

$$\begin{aligned}\boldsymbol{x}_i^T\boldsymbol{\beta}(\boldsymbol{s}_i) - \boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}) =& \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}_i'' + o\left(h^2\right)\\ =& O_p\left(h^2\right).\end{aligned}$$

By a Taylor expansion around $\boldsymbol{x}^T\boldsymbol{\beta}(\boldsymbol{s}_i)$, then,

$$\begin{aligned} q_1\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), \mu(\boldsymbol{s}_i, \boldsymbol{z}_i)\right) =& q_1\left(\boldsymbol{x}_i^T\boldsymbol{\beta}(\boldsymbol{s}_i), \mu(\boldsymbol{s}_i, \boldsymbol{z})\right) \\ & - q_2\left(\boldsymbol{x}_i^T\boldsymbol{\beta}(\boldsymbol{s}_i), \mu(\boldsymbol{s}_i, \boldsymbol{z})\right)\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}_i'' \\ & + o\left(h^2\right). \end{aligned}$$

And by (D.A.2)(a) and (D.A.2)(b), we have that

$$q_1\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), \mu(\boldsymbol{s}_i, \boldsymbol{z}_i)\right) = \rho(\boldsymbol{s}_i, \boldsymbol{z}_i)\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}_i'' + o\left(h^2\right).$$

Now the expectation of Ω_n is

$$\begin{aligned} nE\left(\omega_i|\boldsymbol{Z}_i = \boldsymbol{z}_i, \boldsymbol{s}_i\right) &= (1/2)\,\alpha_n\boldsymbol{z}_i q_1\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), \mu(\boldsymbol{s}_i, \boldsymbol{z}_i)\right) K\left(h^{-1}\|\boldsymbol{s} - \boldsymbol{s}_i\|\right) \\ &= (1/2)\,\alpha_n h^2\boldsymbol{z}_i\left\{h^{-2}\rho(\boldsymbol{s}_i, \boldsymbol{z}_i)\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}_i'' + o\left(\mathbf{1}_{3p}\right)\right\} K\left(h^{-1}\|\boldsymbol{s} - \boldsymbol{s}_i\|\right). \end{aligned}$$

To facilitate a change of variables, we observe that $h^{-2}\tilde{\beta}_j'' = \left(\frac{\boldsymbol{s}_i - \boldsymbol{s}}{h}\right)^T\left\{\nabla^2\beta_j(\boldsymbol{s})\right\}\left(\frac{\boldsymbol{s}_i - \boldsymbol{s}}{h}\right)$. Thus,

$$E\left(\omega_i|\boldsymbol{s}_i\right) = (1/2)\,\alpha_n h^2\left[\begin{pmatrix} 1 \\ h^{-1}(s_{i,1} - s_1) \\ h^{-1}(s_{i,2} - s_2) \end{pmatrix} \otimes \left\{\Gamma(\boldsymbol{s}_i)h^{-2}\tilde{\boldsymbol{\beta}}_i''\right\} + o\left(\mathbf{1}_{3p}\right)\right] K\left(h^{-1}\|\boldsymbol{s} - \boldsymbol{s}_i\|\right).$$

And, using the symmetry of the kernel function,

$$E\left(\omega_i\right) = (1/2)\alpha_n h^4 f(\boldsymbol{s})\begin{pmatrix} \kappa_2 \\ h\kappa_3 \\ h\kappa_3 \end{pmatrix} \otimes \left[\Gamma(\boldsymbol{s})\left\{\nabla_{uu}^2\boldsymbol{\beta}(\boldsymbol{s}) + \nabla_{vv}^2\boldsymbol{\beta}(\boldsymbol{s})\right\}\right] + o\left(h^2\mathbf{1}_{3p}\right)$$

where $\{\nabla^2_{uu}\boldsymbol{\beta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}(\boldsymbol{s})\} = (\nabla^2_{uu}\beta_1(\boldsymbol{s}) + \nabla^2_{vv}\beta_1(\boldsymbol{s}), \ldots, \nabla^2_{uu}\beta_p(\boldsymbol{s}) + \nabla^2_{vv}\beta_p(\boldsymbol{s}))^T$. Thus,

$$E(\Omega_n) = \begin{pmatrix} \alpha_n^{-1} 2^{-1} h^2 \kappa_2 f(\boldsymbol{s}) \Gamma(\boldsymbol{s}) \left(\nabla^2_{uu}\boldsymbol{\beta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}(\boldsymbol{s})\right)^T \\ \mathbf{0}_{2p} \end{pmatrix} + o_p\left(h^2 \mathbf{1}_{3p}\right)$$

Variance: By the previous result, $E(\Omega_n) = O(h^2)$. Thus, $var(\Omega_n) \to E(\Omega_n^2)$, and since the observations are independent, $E(\Omega_n^2) = \sum_{i=1}^n E(\omega_i^2)$. And, by Taylor expansion around $\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i)$,

$$\begin{aligned} q_1^2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), Y_i\right) =& q_1^2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i), Y_i\right) \\ & - q_1\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i), Y_i\right) q_2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i), Y_i\right) \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}_i'' \\ & + o\left(h^2\right). \end{aligned}$$

Since $q_1(\cdot,\cdot)$ is the quasi-score function, it follows that

$$E\left(\omega_i^2 | \boldsymbol{Z}_i = \boldsymbol{z}_i, \boldsymbol{s}_i\right) = \alpha_n^2 \rho(\boldsymbol{s}_i, \boldsymbol{z}_i) \boldsymbol{z}_i \boldsymbol{z}_i^T K\left(h^{-1}\|\boldsymbol{s} - \boldsymbol{s}_i\|\right) + o\left(h^2\right).$$

By the symmetry of the kernel function,

$$E\left(\omega_i^2\right) = n^{-1} f(\boldsymbol{s}) \mathrm{diag}\{\nu_0, \nu_2, \nu_2\} \otimes \Gamma(\boldsymbol{s}) + o(1).$$

Thus,

$$Var(\Omega_n) = f(\boldsymbol{s}) \mathrm{diag}\{\nu_0, \nu_2, \nu_2\} \otimes \Gamma(\boldsymbol{s}) + o(1).$$

□

Appendix: Proof of Lemma 2

Proof. **Expectation**: The approach is similar to the proof of Lemma 1. By the Taylor expansion of $q_2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), \mu\left(\boldsymbol{s}_i, \boldsymbol{z}_i\right)\right)$ around $\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i)$:

$$\begin{aligned} q_2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), \mu(\boldsymbol{s}_i, \boldsymbol{z}_i)\right) =& q_2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i), \mu(\boldsymbol{s}_i, \boldsymbol{z}_i)\right) + q_3\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i), \mu(\boldsymbol{s}_i, \boldsymbol{z}_i)\right)\left\{\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}) - \boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i)\right\} \\ =& -\rho(\boldsymbol{s}_i, \boldsymbol{z}_i) + o\left(1\right). \end{aligned}$$

And by the same arguments as before

$$\begin{aligned} E\left(\delta_i|\boldsymbol{Z}_i = \boldsymbol{z}_i, \boldsymbol{s}_i\right) =& -\alpha_n^2\rho(\boldsymbol{s}_i, \boldsymbol{z}_i)\boldsymbol{z}_i\boldsymbol{z}_i^T K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) \\ E\left(\delta_i|\boldsymbol{s}_i\right) =& -\alpha_n^2\begin{pmatrix} 1 \\ h^{-1}(s_{i,1} - s_1) \\ h^{-1}(s_{i,2} - s_2) \end{pmatrix}\begin{pmatrix} 1 \\ h^{-1}(s_{i,1} - s_1) \\ h^{-1}(s_{i,2} - s_2) \end{pmatrix}^T \otimes \Gamma(\boldsymbol{s}_i) K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) \\ E\left(\delta_i\right) =& -nf\left(\boldsymbol{s}\right)\mathrm{diag}\left\{\kappa_0, \kappa_2, \kappa_2\right\} \otimes \Gamma\left(\boldsymbol{s}\right) + o\left(n^{-1}\right) \end{aligned}$$

Thus,

$$E\left(\Delta_n\right) = -f(\boldsymbol{s})\mathrm{diag}\left\{\kappa_0, \kappa_2, \kappa_2\right\} \otimes \Gamma(\boldsymbol{s}) + o\left(1\right)$$

Variance: From the previous result, it follows that $\left\{E\left(\delta_i\right)\right\}^2 = O\left(n^{-2}\right)$. By the definition of δ_i,

$$\begin{aligned} &E\left(\delta_i^2|\boldsymbol{Z}_i = \boldsymbol{z}_i, \boldsymbol{s}_i\right) = \\ &\alpha_n^4\boldsymbol{z}_i^T\boldsymbol{z}_i q_2^2(\boldsymbol{s}_i, \boldsymbol{z}_i)\begin{pmatrix} 1 \\ h^{-1}(s_{i,1} - s_1) \\ h^{-1}(s_{i,2} - s_2) \end{pmatrix}\begin{pmatrix} 1 \\ h^{-1}(s_{i,1} - s_1) \\ h^{-1}(s_{i,2} - s_2) \end{pmatrix}^T \boldsymbol{z}_i\boldsymbol{z}_i^T K^2\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) + o\left(1\right) \end{aligned}$$

And it follows that $E\left(\delta_i^2\right) = O\left(n^{-1}\alpha_n^2\right)$, and $Var\left(\Delta_n\right) = O\left(\alpha_n^2\right)$. □