

# Local Adaptive Grouped Regularization and its Oracle Properties

Wesley Brooks, Jun Zhu, Zudi Lu

July 9, 2014

## 1 Introduction

Whereas the coefficients in traditional linear regression are scalar constants, the coefficients in a varying coefficients regression (VCR) model are functions - often *smooth* functions - of some effect-modifying variable (Cleveland and Grosse, 1991; Hastie and Tibshirani, 1993).

Current practice for VCR models relies on global model selection to decide which variables should be included in the model, meaning that predictors are identified as relevant or irrelevant over the entire domain  $\mathcal{D}$ . Antoniadis, Gijbels, and Verhasselt (2012) describe a method for globally selecting the

relevant predictors in a VCR model where the coefficient functions are estimated with P-splines. Wang, Li, and Huang (2008) show a method for doing global variable selection in a VCR model where the coefficient functions are estimated by basis expansion.

Local adaptive grouped regularization (LAGR) is developed here as a method to select only the locally relevant predictors at any specific location  $\mathbf{s}$  in the domain  $\mathcal{D}$  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 a univariate effect-modifying variable, and by Sun, Yan, Zhang, and Lu (2014) for two-dimensional effect-modifying variable and autocorrelation among the observed response. These methods minimize the boundary effect (Hastie and Loader, 1993) by estimating the coefficients as local polynomials of odd degree (usually locally linear).

For linear regression models, the adaptive lasso (AL) (Zou, 2006) produces consistent estimates of the coefficients and has been shown to have appealing properties for automating 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. For data where the observed variables fall into mutually exclusive groups that are known in advance, the adaptive group

lasso (Wang and Leng, 2008; Yuan and Lin, 2006) has similar oracle properties to the adaptive lasso while doing selection at the level of groups rather than individual variables. The proposed LAGR method uses the adaptive group lasso for local variable selection and coefficient estimation in a locally linear regression model. We show that LAGR possesses 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 VCR model is described in Section 2; the proposed LAGR technique and its oracle properties are presented in Section 3; in Section 5, the performance of the proposed LAGR technique is evaluated in a simulation study, and in Section 6 the proposed method is applied to the Boston house price dataset. Proofs of the theorems appear in Appendix 6.

## 2 Varying coefficients regression

### 2.1 Model

Consider  $n$  data points, observed at sampling locations  $\mathbf{s}_i = (s_{i,1}, s_{i,2})^T$  for  $i = 1, \dots, n$ , which are distributed in a spatial domain  $\mathcal{D} \subset \mathbb{R}^2$  according to a density  $f(\mathbf{s})$  with  $\mathbf{s} \in \mathcal{D}$ . For  $i = 1, \dots, n$ , let  $y(\mathbf{s}_i)$  and  $\mathbf{x}(\mathbf{s}_i)$  denote,

respectively, the univariate response and the  $(p + 1)$ -variate vector of covariates measured at location  $\mathbf{s}_i$ . At each location  $\mathbf{s}_i$ , assume that the outcome is related to the covariates by a linear regression where the coefficients  $\boldsymbol{\beta}(\mathbf{s}_i)$  are functions in two dimensions and  $\varepsilon(\mathbf{s}_i)$  is random error at location  $\mathbf{s}_i$ . That is,

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

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

$$\boldsymbol{\varepsilon} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2). \quad (2)$$

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, Mei, and Yan, 2008). The augmented local design matrix at location  $\mathbf{s}_i$  is

$$\mathbf{Z}(\mathbf{s}_i) = (\mathbf{X} \quad \mathbf{L}_i \mathbf{X} \quad \mathbf{M}_i \mathbf{X}) \quad (3)$$

where  $\mathbf{X}$  is the unaugmented matrix of covariates,  $\mathbf{L}_i = \text{diag}\{s_{i',1} - s_{i,1}\}$  and

$\mathbf{M}_i = \text{diag}\{s_{i',2} - s_{i,2}\}$  for  $i' = 1, \dots, n$ .

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

$$\boldsymbol{\zeta}(\mathbf{s}_i) = (\boldsymbol{\beta}(\mathbf{s}_i)^T, \nabla_u \boldsymbol{\beta}(\mathbf{s}_i)^T, \nabla_v \boldsymbol{\beta}(\mathbf{s}_i)^T)^T$$

## 2.2 Local Likelihood and Coefficient Estimation

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

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

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 when the coefficient functions are smooth, the coefficients at location  $\mathbf{s}$  can approximate the coefficients within some neighborhood of  $\mathbf{s}$ , with the quality of the approximation declining as the distance from  $\mathbf{s}$  increases.

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

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

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

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

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

Letting  $\mathbf{W}(\mathbf{s}) = \text{diag}\{K_h(\|\mathbf{s} - \mathbf{s}_i\|)\}$  be a diagonal matrix of kernel weights, the local likelihood is maximized by weighted least squares:

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

Thus, we have

$$\tilde{\boldsymbol{\zeta}}(\mathbf{s}) = \{\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}(\mathbf{s})\}^{-1} \mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Y}$$

By Theorem 3 of Sun et al. (2014), for any given  $\mathbf{s}$

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

### 3 Local Variable Selection with LAGR

#### 3.1 The LAGR-Penalized Local Likelihood

Estimating the local coefficients by (??) 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,  $\boldsymbol{\zeta}_j(\mathbf{s}) = (\beta_j(\mathbf{s}), \nabla_u \beta_j(\mathbf{s}), \nabla_v \beta_j(\mathbf{s}))^T$  for  $j = 1, \dots, p$ . 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. The proposed LAGR penalty is an adaptive  $\ell_1$  penalty akin to the adaptive group lasso (Wang and Leng, 2008; Zou, 2006).

More specifically, we consider the penalized local sum of squares at location

$\mathbf{s}$ :

$$\mathcal{J}(\boldsymbol{\zeta}(\mathbf{s})) = \mathcal{S}(\boldsymbol{\zeta}(\mathbf{s})) + \mathcal{P}(\boldsymbol{\zeta}(\mathbf{s}))$$

where  $\mathcal{S}(\boldsymbol{\zeta}(\mathbf{s})) = (1/2) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}^T \mathbf{W}(\mathbf{s}) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}^T$  is the locally weighted sum of squares,  $\mathcal{P}(\boldsymbol{\zeta}(\mathbf{s})) = \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s})\|$  is a local adaptive grouped regularization (LAGR) penalty, and  $\|\cdot\|$  is the  $L_2$ -norm.

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

### 3.2 Oracle Properties

For a local model at location  $\mathbf{s}$ , define the following terms.

(D.1) Let  $a_n = \max \{\phi_j(\mathbf{s}), j \leq p_0\}$  be the largest penalty applied to a covariate group whose true coefficient norm is nonzero, and  $b_n = \min \{\phi_j(\mathbf{s}), j > p_0\}$  be the smallest penalty applied to a covariate group whose true coefficient norm is zero.

(D.2) Let  $\mathbf{Z}_k(\mathbf{s})$  be the augmented design matrix for covariate group  $k$ , and let  $\mathbf{Z}_{-k}(\mathbf{s})$  be the augmented design matrix for all the data except



covariate group  $k$ . Similarly, let  $\zeta_k(\mathbf{s})$  be the augmented coefficients for covariate group  $k$  and  $\zeta_{-k}(\mathbf{s})$  be the augmented coefficients for all covariate groups except  $k$ .

$$(D.3) \text{ Let } \kappa_0 = \int_{R^2} K(\|\mathbf{s}\|) d\mathbf{s}, \kappa_2 = \int_{R^2} [(1, 0)\mathbf{s}]^2 K(\|\mathbf{s}\|) d\mathbf{s} = \int_{R^2} [(0, 1)\mathbf{s}]^2 K(\|\mathbf{s}\|) d\mathbf{s}, \\ \text{and } \nu_0 = \int_{R^2} K^2(\|\mathbf{s}\|) d\mathbf{s}.$$

Assume the following conditions.

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

(A.2) There are  $p_0 < p$  covariates  $\mathbf{X}_a(\mathbf{s})$  with nonzero local regression coefficients, indicated by  $\beta_a(\mathbf{s})$ . Without loss of generality, assume these are covariates  $1, \dots, p_0$ . The remaining covariates  $\mathbf{X}_b(\mathbf{s})$  have true coefficients equal to zero, indicated by  $\beta_b(\mathbf{s})$ .

(A.3)  $\mathbf{X}(\mathbf{s}_1), \dots, \mathbf{X}(\mathbf{s}_n)$  are *iid* random variables that are independent of  $\varepsilon(\mathbf{s}_1), \dots, \varepsilon(\mathbf{s}_n)$ . Also  $\Psi = E\{\mathbf{X}(\mathbf{s}_1)\mathbf{X}^T(\mathbf{s}_1)\}$  is positive-definite,  $E|\mathbf{X}(\mathbf{s}_1)|^{2q} < \infty$ , and  $E|\varepsilon(\mathbf{s}_1)|^{2q} < \infty$ .

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

(A.5)  $h = h_n = O(n^{-1/6})$

**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*

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

**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  $Pr \left\{ \|\hat{\zeta}_j(\mathbf{s})\| = 0 \right\} \rightarrow 0$  if  $j \leq p_0$  and  $Pr \left\{ \|\hat{\zeta}_j(\mathbf{s})\| = 0 \right\} \rightarrow 1$  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 (Sun et al., 2014), 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 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,  $h^{-1}n^{-1/2}\lambda_n(\mathbf{s}) \rightarrow 0$  for  $j \leq p_0$  and  $hn^{-1/2}\lambda_n(\mathbf{s}) \|\zeta_j(\mathbf{s})\|^{-\gamma} \rightarrow \infty$  for  $j > p_0$ . We require that  $\lambda_n(\mathbf{s})$  satisfy both assumptions. Suppose  $\lambda_n(\mathbf{s}) = n^\alpha$ , and recall that  $h = O(n^{-1/6})$  and  $\|\tilde{\zeta}_p(\mathbf{s})\| = O(h^{-1}n^{-1/2})$ . Then  $h^{-1}n^{-1/2}\lambda_n(\mathbf{s}) = O(n^{-1/3+\alpha})$  and  $hn^{-1/2}\lambda_n(\mathbf{s}) \|\tilde{\zeta}_p(\mathbf{s})\|^{-\gamma} = O(n^{-2/3+\alpha+\gamma/3})$ . 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  $\lambda_n(\mathbf{s})$  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) (Akaike, 1973; Mallows, 1973; Stein, 1981).

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  $\lambda_n(\mathbf{s})$  (Hurvich, Simonoff, and Tsai, 1998). That is, let

$$\begin{aligned}\hat{d}f(\lambda; \mathbf{s}) &= \sum_{j=1}^p I\left(\|\hat{\boldsymbol{\zeta}}(\lambda; \mathbf{s})\| > 0\right) + \sum_{j=1}^p \frac{\|\hat{\boldsymbol{\zeta}}(\lambda; \mathbf{s})\|}{\|\tilde{\boldsymbol{\zeta}}(\mathbf{s})\|} (p_j - 1) \\ \text{AIC}_c(\lambda; \mathbf{s}) &= \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) \sigma^{-2} \left\{ y(\mathbf{s}_i) - \mathbf{z}'(\mathbf{s}_i) \hat{\boldsymbol{\zeta}}(\lambda; \mathbf{s}) \right\}^2 + 2\hat{d}f(\lambda; \mathbf{s}) \\ &\quad + \frac{2\hat{d}f(\lambda; \mathbf{s}) \left\{ \hat{d}f(\lambda; \mathbf{s}) + 1 \right\}}{\sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) - \hat{d}f(\lambda; \mathbf{s}) - 1}\end{aligned}$$

where the local coefficient estimate is written  $\hat{\boldsymbol{\zeta}}(\lambda; \mathbf{s})$  to emphasize that it

depends on the tuning parameter.

## 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  $\beta(\mathbf{s})$ . If the response variable  $y$  is from an exponential-family distribution then its density is

$$f\{y(\mathbf{s})|\mathbf{x}(\mathbf{s}),\theta(\mathbf{s})\} = c\{y(\mathbf{s})\} \times [\exp\{\theta(\mathbf{s})y(\mathbf{s}) - b\{\theta(\mathbf{s})\}\}]$$

where  $\phi$  and  $\theta$  are parameters and

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

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.

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

$$\begin{aligned}
E \{y(\mathbf{s})|\mathbf{x}(\mathbf{s})\} &= g^{-1} \{ \eta(\mathbf{s}) \} \\
\text{Var} \{y(\mathbf{s})|\mathbf{x}(\mathbf{s})\} &= [g' \{ \mu(\mathbf{s}) \}]^{-1} = V \{ \mu(\mathbf{s}) \} \\
\frac{d}{d\mu} g^{-1} \{ \mu(\mathbf{s}) \} &= V \{ \mu(\mathbf{s}) \}
\end{aligned}$$

## 4.2 Local quasi-likelihood

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

$$\ell^* \{\zeta(\mathbf{s})\} = \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) Q \left[ g^{-1} \{ \mathbf{z}'(\mathbf{s}_i) \zeta(\mathbf{s}) \}, Y(\mathbf{s}_i) \right]. \quad (7)$$

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

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

and  $\hat{\mu}(\mathbf{s}_i; \mathbf{s}) = g^{-1} \left\{ \mathbf{z}'(\mathbf{s}_i) \hat{\boldsymbol{\zeta}}(\mathbf{s}) \right\}$  is the mean at location  $\mathbf{s}_i$  estimated using the coefficients  $\hat{\boldsymbol{\zeta}}(\mathbf{s})$  fitted at location  $\mathbf{s}$ . Except for the  $K_h(\|\mathbf{s} - \mathbf{s}_i\|)$  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  $\hat{\boldsymbol{\zeta}}(\mathbf{s})$ .

#### 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

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

where  $\Gamma(\mathbf{s}) = E \left[ V \{ \mu(\mathbf{s}) \} X(\mathbf{s}) X(\mathbf{s})^T \right]$ .

#### 4.5 LAGR penalty

As in the case of linear models, the LAGR for GLMs is a grouped  $\ell_1$  regularization method. Now, though, we use a penalized local quasi-likelihood:

$$\begin{aligned}
\mathcal{J}\{\boldsymbol{\zeta}(\mathbf{s})\} &= \ell^*\{\boldsymbol{\zeta}(\mathbf{s})\} + \mathcal{P}\{\boldsymbol{\zeta}(\mathbf{s})\} \\
&= \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) Q\left[g^{-1}\{z'(\mathbf{s}_i)\boldsymbol{\zeta}(\mathbf{s})\}, Y(\mathbf{s}_i)\right] + \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s})\|
\end{aligned} \tag{9}$$

and similarly to the case for gaussian data,  $\phi_j(\mathbf{s}) = \lambda_n(\mathbf{s}) \|\tilde{\boldsymbol{\zeta}}_j(\mathbf{s})\|^{-\gamma}$ , where  $\lambda_n(\mathbf{s}) > 0$  is a the local tuning parameter applied to all coefficients at location  $\mathbf{s}$  and  $\tilde{\boldsymbol{\zeta}}_j(\mathbf{s})$  is the vector of unpenalized local coefficients.

## 4.6 Oracle properties of LAGR in the GLM setting

The oracle properties for LAGR in the GLM setting are similar to those in the gaussian setting:

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

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

**Theorem 4** (Selection consistency). *If  $h^{-1}n^{-1/2}a_n \xrightarrow{p} \infty$  and  $hn^{-1/2}b_n \xrightarrow{p} \infty$  then  $P\left\{\|\hat{\boldsymbol{\zeta}}_j(\mathbf{s})\| = 0\right\} \rightarrow 0$  if  $j \leq p_0$  and  $P\left\{\|\hat{\boldsymbol{\zeta}}_j(\mathbf{s})\| = 0\right\} \rightarrow 1$  if  $j > p_0$ .*



## 5 Simulation Study

### 5.1 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  $30 \times 30$  grid. Each of  $p = 5$  covariates  $X_1, \dots, X_5$  was simulated by a Gaussian random field with mean zero and exponential covariance function  $\text{Cov}(X_{ji}, X_{ji'}) = \sigma_x^2 \exp(-\tau_x^{-1} \delta_{ii'})$  where  $\sigma_x^2 = 1$  is the variance,  $\tau_x = 0.1$  is the range parameter, and  $\delta_{ii'} = \|\mathbf{s}_i - \mathbf{s}_{i'}\|_2$  is the Euclidean distance between locations  $\mathbf{s}$  and  $\mathbf{s}_i$ .

Correlation was induced between the covariates by multiplying the design matrix  $\mathbf{X}$  by  $\mathbf{R}$ , where  $\mathbf{R}$  is the Cholesky decomposition of the covariance matrix  $\mathbf{\Sigma} = \mathbf{R}'\mathbf{R}$ . The covariance matrix  $\mathbf{\Sigma}$  is a  $5 \times 5$  matrix that has ones on the diagonal and  $\rho$  for all off-diagonal entries, where  $\rho$  is the between-covariate correlation.

The simulated response was  $y_i = \mathbf{x}_i' \boldsymbol{\beta}_i + \varepsilon_i$  for  $i = 1, \dots, n$  where  $n = 900$  and the  $\varepsilon_i$ 's were iid Gaussian with mean zero and variance  $\sigma_\varepsilon^2$ . The simulated data included the response  $y$  and five covariates  $x_1, \dots, x_5$ . The true data-generating model uses only  $x_1$ . The variables  $x_2, \dots, x_5$  are included to assess performance in model selection.

Three different functions were used for the coefficient surface  $\beta_1(\mathbf{s})$  (Figure

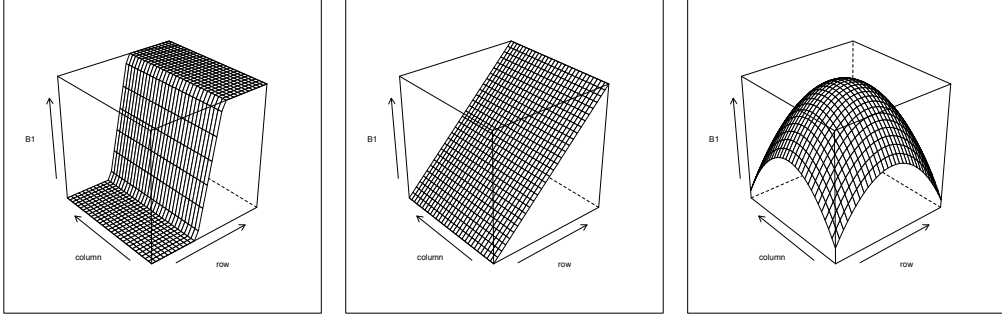


Figure 1: These are, respectively, the step, gradient, and parabola functions that were used for the coefficient function  $\beta_1(\mathbf{s})$  in the VCR model  $y(\mathbf{s}_i) = x_1(\mathbf{s}_i)\beta_1(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i)$  when generating the data for the simulation study.

1). The first is the “step” function:

$$\beta_{step}(\mathbf{s}) = \begin{cases} 1 & \text{if } s_x > 0.6 \\ 5s_x - 2 & \text{if } 0.4 < s_x \leq 0.6 \\ 0 & \text{o.w.} \end{cases}$$

The second is the gradient function,  $\beta_{gradient}(\mathbf{s}) = s_x$ , and the third is the parabola  $\beta_{parabola}(\mathbf{s}) = 1 - 2\{(s_x - 0.5)^2 + (s_y - 0.5)^2\}$ .

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  $\beta_1(\mathbf{s})$ ; 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.

Simulation settings			MISE $\hat{\beta}_1$			MISE $\hat{\beta}_2, \dots, \hat{\beta}_5$	
$\beta_1(\mathbf{s})$	$\rho$	$\sigma_\varepsilon$	LAGR	VCR	Oracle	LAGR	VCR
step	0	0.5	<i>0.02</i>	0.02	<b>0.01</b>	<b>0.00</b>	0.01
		1.0	<i>0.03</i>	0.03	<b>0.02</b>	<b>0.00</b>	0.02
	0.5	0.5	<i>0.02</i>	0.02	<b>0.01</b>	<b>0.00</b>	0.01
		1.0	<i>0.03</i>	0.05	<b>0.02</b>	<b>0.00</b>	0.03
	0.9	0.5	<i>0.03</i>	0.05	<b>0.01</b>	<b>0.00</b>	0.04
		1.0	<i>0.12</i>	0.17	<b>0.02</b>	<b>0.02</b>	0.15
gradient	0	0.5	0.01	<i>0.01</i>	<b>0.00</b>	<b>0.00</b>	0.00
		1.0	0.03	<i>0.02</i>	<b>0.01</b>	<b>0.00</b>	0.02
	0.5	0.5	0.01	<i>0.01</i>	<b>0.00</b>	<b>0.00</b>	0.01
		1.0	0.04	<i>0.03</i>	<b>0.01</b>	<b>0.00</b>	0.03
	0.9	0.5	<i>0.03</i>	0.04	<b>0.00</b>	<b>0.00</b>	0.04
		1.0	<i>0.14</i>	0.14	<b>0.01</b>	<b>0.02</b>	0.15
parabola	0	0.5	0.01	<i>0.01</i>	<b>0.01</b>	<b>0.00</b>	0.00
		1.0	0.03	<i>0.02</i>	<b>0.02</b>	<b>0.00</b>	0.02
	0.5	0.5	0.01	<i>0.01</i>	<b>0.01</b>	<b>0.00</b>	0.01
		1.0	0.03	<i>0.03</i>	<b>0.02</b>	<b>0.00</b>	0.03
	0.9	0.5	<i>0.02</i>	0.04	<b>0.01</b>	<b>0.00</b>	0.04
		1.0	0.17	<i>0.14</i>	<b>0.02</b>	<b>0.03</b>	0.15

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

The results are presented in terms of the mean integrated squared error (MISE) of the coefficient surface estimates  $\hat{\beta}_1(\mathbf{s}), \dots, \hat{\beta}_5(\mathbf{s})$ , the MISE of the fitted response  $\hat{y}(\mathbf{s})$ , and the frequency with which the coefficient surface estimates  $\hat{\beta}_2(\mathbf{s}), \dots, \hat{\beta}_5(\mathbf{s})$  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 oracular selection. Oracular selection means that exactly the correct set of covariates was used to fit each local model.

## 5.2 Simulation Results

The MISE of the estimates of  $\beta_1(\mathbf{s})$  are in Table ???. Recall that  $\beta_2(\mathbf{s}), \dots, \beta_5(\mathbf{s})$  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  $\beta_2(\mathbf{s}), \dots, \beta_5(\mathbf{s})$  were dropped from the LAGR models ranged from 0.78 to 0.97. The MISE of the fitted  $\hat{y}(\mathbf{s})$  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  $\beta_1(\mathbf{s})$  about equal to that of the VCR model with no selection, and with consistently better accuracy for estimating  $\beta_2(\mathbf{s}), \dots, \beta_5(\mathbf{s})$ .

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

Simulation settings			Zero frequency	MISE		
$\beta_1(\mathbf{s})$	$\rho$	$\sigma_\varepsilon$	$\hat{\beta}_2, \dots, \hat{\beta}_5$	LAGR	$\hat{y}$ VCR	Oracle
step	0	0.5	0.97	<i>0.25</i>	0.26	<b>0.25</b>
		1.0	0.96	<i>1.00</i>	<b>1.00</b>	0.99
	0.5	0.5	0.96	<i>0.26</i>	0.26	<b>0.25</b>
		1.0	0.92	<i>0.99</i>	<b>1.00</b>	0.98
	0.9	0.5	0.86	<i>0.27</i>	0.30	<b>0.25</b>
		1.0	0.85	<i>1.08</i>	1.14	<b>0.98</b>
gradient	0	0.5	0.96	<i>0.25</i>	<b>0.25</b>	0.25
		1.0	0.95	<b>0.99</b>	<i>0.99</i>	0.97
	0.5	0.5	0.94	<i>0.25</i>	<b>0.25</b>	0.24
		1.0	0.92	<i>1.00</i>	<b>1.00</b>	0.97
	0.9	0.5	0.80	<i>0.27</i>	0.28	<b>0.24</b>
		1.0	0.85	<i>1.09</i>	1.12	<b>0.97</b>
parabola	0	0.5	0.97	<i>0.25</i>	<b>0.25</b>	0.25
		1.0	0.94	<b>1.00</b>	<i>1.00</i>	0.98
	0.5	0.5	0.95	<b>0.25</b>	0.25	<i>0.25</i>
		1.0	0.88	<b>1.00</b>	<i>1.00</i>	0.97
	0.9	0.5	0.79	<i>0.26</i>	0.28	<b>0.24</b>
		1.0	0.78	1.13	<i>1.12</i>	<b>0.98</b>

Table 2: 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  $\sigma_\varepsilon^2$  for that setting.

selection and estimation accuracy did decline when there was high ( $\rho = 0.9$ ) correlation among the predictor variables.

## 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 (Gilley and Pace, 1996; Harrison and Rubinfeld, 1978; Pace and Gilley, 1997). The data source 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 covariates. The covariates 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 \$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 predictor variable with a positive mean of the local coefficients. The coefficient of the CRIM variable was estimated to be exactly zero at 49% of

	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.

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 LAGR indicates that the TAX variable was nowhere 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 improved access to radial roads is associated with a greater median house price and parts where it is associated with a lesser median house price.

In their example using the same data, Sun et al. (2014) estimated that the coefficients of RAD and LSTAT should be constant, at 0.36 and  $-0.45$ , respectively. That conclusion differs from our result, which says that the

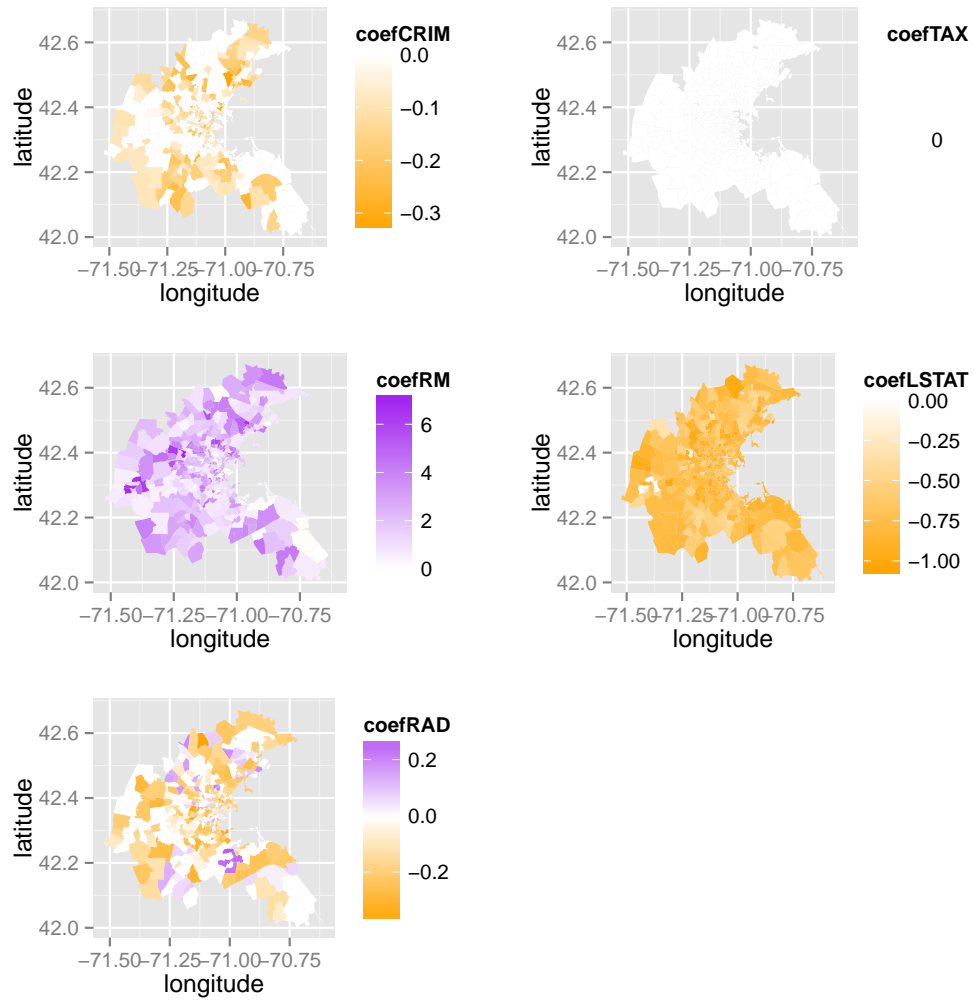


Figure 2: The LAGR estimates of coefficients for the Boston house price data.



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).

## Appendix: Proofs of Theorem 1

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

$$H_n(\mathbf{u}) = (1/2) [\mathbf{Y} - \mathbf{Z}(\mathbf{s}) \{\boldsymbol{\zeta}(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}\}]^T \mathbf{W}(\mathbf{s}) [\mathbf{Y} - \mathbf{Z}(\mathbf{s}) \{\boldsymbol{\zeta}(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}\}] \quad (10)$$

$$+ \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| \quad (11)$$

$$- (1/2) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}^T \mathbf{W}(\mathbf{s}) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\} - \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s})\| \quad (12)$$

$$= (1/2) \mathbf{u}^T \{h^{-2}n^{-1} \mathbf{Z}^T(\mathbf{s}) \mathbf{W}(\mathbf{s}) \mathbf{Z}(\mathbf{s})\} \mathbf{u} \quad (13)$$

$$- \mathbf{u}^T [h^{-1}n^{-1/2} \mathbf{Z}^T(\mathbf{s}) \mathbf{W}(\mathbf{s}) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}] \quad (14)$$

$$+ \sum_{j=1}^p n^{-1/2} \phi_j(\mathbf{s}) n^{1/2} \{\|\boldsymbol{\zeta}_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\boldsymbol{\zeta}_j(\mathbf{s})\|\} \quad (15)$$

Note that the limiting behavior of the last term differs between the cases  $j \leq p_0$  and  $j > p_0$ .

Case  $j \leq p_0$ :

If  $j \leq p_0$ , then  $n^{-1/2}\phi_j(\mathbf{s}) \rightarrow n^{-1/2}\lambda_n(\mathbf{s})\|\zeta_j(\mathbf{s})\|^{-\gamma}$  and  $|\sqrt{n} \{ \|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\zeta_j(\mathbf{s})\| \} | \leq h^{-1}\|\mathbf{u}_j\|$ . Thus,

$$\lim_{n \rightarrow \infty} \phi_j(\mathbf{s}) (\|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\zeta_j(\mathbf{s})\|) \leq h^{-1}n^{-1/2}\phi_j(\mathbf{s})\|\mathbf{u}_j\| \leq h^{-1}n^{-1/2}a_n\|\mathbf{u}_j\| \rightarrow 0$$

Case  $j > p_0$ :

If  $j > p_0$ , then  $\phi_j(\mathbf{s}) (\|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\zeta_j(\mathbf{s})\|) = \phi_j(\mathbf{s})h^{-1}n^{-1/2}\|\mathbf{u}_j\|$ .

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

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

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

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

Thus, the limit of  $H_n(\mathbf{u})$  is the same as the limit of  $H_n^*(\mathbf{u})$  where

$$H_n^*(\mathbf{u}) = (1/2)\mathbf{u}^T \{h^{-2}n^{-1}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}(\mathbf{s})\} \mathbf{u} - \mathbf{u}^T [h^{-1}n^{-1/2}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\zeta(\mathbf{s})\}]$$

if  $\|\mathbf{u}_j\| = 0 \forall j > p_0$ , and  $H_n^*(\mathbf{u}) = \infty$  otherwise. It follows that  $H_n^*(\mathbf{u})$  is convex and its unique minimizer  $\hat{\mathbf{u}}_n$  is found by solving the equation:

$$\underline{0} = \left\{ h^{-2} n^{-1} \mathbf{Z}(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}(\mathbf{s}) \right\} \hat{\mathbf{u}}_n - \left[ h^{-1} n^{-1/2} \mathbf{Z}(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \{ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \boldsymbol{\zeta}(\mathbf{s}) \} \right]. \quad (16)$$

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

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

$$\hat{\mathbf{u}}_n \xrightarrow{d} N \left( \frac{\kappa_2 h^2}{2\kappa_0} \{ \nabla_{uu}^2 \boldsymbol{\zeta}_j(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_j(\mathbf{s}) \}, f(\mathbf{s}) \kappa_0^{-2} \nu_0 \sigma^2 \Psi^{-1} \right) \quad (17)$$

the result of Theorem 1 follows.  $\square$

## Appendix: Proof of Theorem 2

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

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

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

$$0 = \mathbf{Z}_p^T(\mathbf{s}) \mathbf{W}(\mathbf{s}) \left\{ \mathbf{Y} - \mathbf{Z}_{-p}(\mathbf{s}) \hat{\zeta}_{-p}(\mathbf{s}) - \mathbf{Z}_p(\mathbf{s}) \hat{\zeta}_p(\mathbf{s}) \right\} - \phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|} \quad (18)$$

$$= \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \zeta(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta(\mathbf{s}) + \nabla_{vv}^2 \zeta(\mathbf{s}) \} \right] \quad (19)$$

$$+ \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{-p}(\mathbf{s}) \left[ \zeta_{-p}(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_{-p}(\mathbf{s}) + \nabla_{vv}^2 \zeta_{-p}(\mathbf{s}) \} - \hat{\zeta}_{-p}(\mathbf{s}) \right] \quad (20)$$

$$+ \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_p(\mathbf{s}) \left[ \zeta_p(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_p(\mathbf{s}) + \nabla_{vv}^2 \zeta_p(\mathbf{s}) \} - \hat{\zeta}_p(\mathbf{s}) \right] \quad (21)$$

$$- \phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|} \quad (22)$$

Thus,

$$\frac{h}{\sqrt{n}} \phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|} = \quad (23)$$

$$\mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \frac{h}{\sqrt{n}} \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \zeta(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta(\mathbf{s}) + \nabla_{vv}^2 \zeta(\mathbf{s}) \} \right] \quad (24)$$

$$+ \left\{ n^{-1} \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{-p}(\mathbf{s}) \right\} h \sqrt{n} \left[ \zeta_{-p}(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_{-p}(\mathbf{s}) + \nabla_{vv}^2 \zeta_{-p}(\mathbf{s}) \} - \hat{\zeta}_{-p}(\mathbf{s}) \right] \\ + \left\{ n^{-1} \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_p(\mathbf{s}) \right\} h \sqrt{n} \left[ \zeta_p(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_p(\mathbf{s}) + \nabla_{vv}^2 \zeta_p(\mathbf{s}) \} - \hat{\zeta}_p(\mathbf{s}) \right] \quad (25)$$

From Lemma 2 of Sun et al. (2014),  $\left\{n^{-1}\mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{-p}(\mathbf{s})\right\} = O_p(1)$   
and  $\left\{n^{-1}\mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_p(\mathbf{s})\right\} = O_p(1)$ .

From Theorem 3 of Sun et al. (2014), we have that  $h\sqrt{n} \left[ \hat{\boldsymbol{\zeta}}_{-p}(\mathbf{s}) - \boldsymbol{\zeta}_{-p}(\mathbf{s}) - \frac{h^2\kappa_2}{2\kappa_0} \{\nabla_{uu}^2 \boldsymbol{\zeta}_{-p}(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_{-p}(\mathbf{s})\} \right] = O_p(1)$  and  $h\sqrt{n} \left[ \hat{\boldsymbol{\zeta}}_p(\mathbf{s}) - \boldsymbol{\zeta}_p(\mathbf{s}) - \frac{h^2\kappa_2}{2\kappa_0} \{\nabla_{uu}^2 \boldsymbol{\zeta}_p(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_p(\mathbf{s})\} \right] = O_p(1)$ .

We showed in the proof of Theorem 1 that

$$h\sqrt{n}\mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \boldsymbol{\zeta}(\mathbf{s}) - \frac{h^2\kappa_2}{2\kappa_0} \{\nabla_{uu}^2 \boldsymbol{\zeta}(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}(\mathbf{s})\} \right] = O_p(1).$$

The right hand side of (23) is  $O_p(1)$ , so for  $\hat{\boldsymbol{\zeta}}_p(\mathbf{s})$  to be a solution, we must have that  $hn^{-1/2}\phi_p(\mathbf{s}) \hat{\boldsymbol{\zeta}}_p(\mathbf{s}) / \|\hat{\boldsymbol{\zeta}}_p(\mathbf{s})\| = O_p(1)$ .

But since by assumption  $\hat{\boldsymbol{\zeta}}_p(\mathbf{s}) \neq \mathbf{0}$ , there must be some  $k \in \{1, 2, 3\}$  such that  $|\hat{\zeta}_{p_k}(\mathbf{s})| = \max\{|\hat{\zeta}_{p_{k'}}(\mathbf{s})| : 1 \leq k' \leq 3\}$ . And for this  $k$ , we have that  $|\hat{\zeta}_{p_k}(\mathbf{s})| / \|\hat{\boldsymbol{\zeta}}_p(\mathbf{s})\| \geq 1/\sqrt{3} > 0$ .

Since  $hn^{-1/2}b_n \rightarrow \infty$ , we have that  $hn^{-1/2}\phi_p(\mathbf{s}) \hat{\boldsymbol{\zeta}}_p(\mathbf{s}) / \|\hat{\boldsymbol{\zeta}}_p(\mathbf{s})\| \geq hb_n/\sqrt{3n} \rightarrow \infty$  and therefore the left hand side of (23) dominates the sum to the right side.

Thus, for large enough  $n$ ,  $\hat{\boldsymbol{\zeta}}_p(\mathbf{s}) \neq \mathbf{0}$  cannot maximize  $\mathcal{J}(\cdot)$ , and therefore  $Pr \left\{ \hat{\boldsymbol{\zeta}}_{(b)}(\mathbf{s}) = \mathbf{0} \right\} \rightarrow 1$ .  $\square$

## Appendix: Lemmas

The next proofs require the following lemmas.

**Lemma 1.**

$$E \left[ \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_1 \left( \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}), Y(\mathbf{s}_i) \right) \{\mathbf{Z}(\mathbf{s}_i)\}_i \right] \rightarrow (1/2) \alpha_n^{-1} h^2 f(\mathbf{s}) \begin{pmatrix} 1 & \nu_1 & \nu_2 \\ \nu_1 & \nu_2 & \nu_3 \\ \nu_2 & \nu_2 & \nu_4 \end{pmatrix} \{1$$

**Lemma 2.**

$$E \left[ \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_2 \left( \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}), Y(\mathbf{s}_i) \right) \{\mathbf{Z}(\mathbf{s}_i)\}_i \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \right] \rightarrow -f(\mathbf{s}) \begin{pmatrix} 1 & \nu_1 & \nu_2 \\ \nu_1 & \nu_2 & \nu_3 \\ \nu_2 & \nu_2 & \nu_4 \end{pmatrix} \otimes$$

**Lemma 3.**

$$E \left[ \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_2 \left( \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}), Y(\mathbf{s}_i) \right) \{\mathbf{Z}(\mathbf{s}_i)\}_i \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \right] \rightarrow -f(\mathbf{s}) \begin{pmatrix} 1 & \nu_1 & \nu_2 \\ \nu_1 & \nu_2 & \nu_3 \\ \nu_2 & \nu_2 & \nu_4 \end{pmatrix} \otimes$$

**Lemma 4.**

$$E \left[ \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_2 \left( \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}), Y(\mathbf{s}_i) \right) \{\mathbf{Z}(\mathbf{s}_i)\}_i \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \right] \rightarrow -f(\mathbf{s}) \begin{pmatrix} 1 & \nu_1 & \nu_2 \\ \nu_1 & \nu_2 & \nu_3 \\ \nu_2 & \nu_2 & \nu_4 \end{pmatrix} \otimes$$

## Appendix: Proof of Theorem 3

*Proof of theorem 3.* Define the  $q$ -functions to be the derivatives of the quasi-likelihood:  $q_j(t, y) = (\partial/\partial t)^j Q(g^{-1}(t), y)$ . Then  $q_1(\eta(\mathbf{s}), \mu(\mathbf{s})) = 0$  and  $q_2(\eta(\mathbf{s}), \mu(\mathbf{s})) = -b''(\eta(\mathbf{s}))$ . Also define  $\bar{\eta}_i()$   $\square$

Define  $H_n(\mathbf{u}) = \mathcal{J}^*(\boldsymbol{\zeta}(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}) - \mathcal{J}^*(\boldsymbol{\zeta}(\mathbf{s}))$ . Then,

$$H_n(\mathbf{u}) = \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) Q\left(g^{-1}\left(\{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}\right), Y(\mathbf{s}_i)\right) \quad (26)$$

$$- \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) Q\left(g^{-1}\left(\{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s})\right), Y(\mathbf{s}_i)\right) \quad (27)$$

$$+ \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}\| - + \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s})\| \quad (28)$$

Define

$$A_1(\mathbf{s}) = \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_1\left(\{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}), Y(\mathbf{s}_i)\right) \{\mathbf{Z}(\mathbf{s}_i)\}_i$$

$$A_2(\mathbf{s}) = \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_2\left(\{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}), Y(\mathbf{s}_i)\right) \{\mathbf{Z}(\mathbf{s}_i)\}_i \{\mathbf{Z}(\mathbf{s}_i)\}_i^T.$$

Then taking the Taylor expansion of  $\mathcal{J}^*(\boldsymbol{\zeta}(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u})$  around  $\boldsymbol{\zeta}(\mathbf{s})$  give us:

$$\mathcal{J}^* (\zeta(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}) = \mathcal{J}^* (\zeta(\mathbf{s})) \quad (29)$$

$$+ A_1(\mathbf{s})^T \mathbf{u} \quad (30)$$

$$+ (1/2)\mathbf{u}^T A_2(\mathbf{s}) \mathbf{u} \quad (31)$$

$$+ (h^{-3}n^{-3/2}/6) \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_3 \left( \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \tilde{\zeta}_i, Y(\mathbf{s}_i) \right) \left[ \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \mathbf{u} \right]^3 \quad (32)$$

$$+ \sum_{j=1}^p \phi_j(\mathbf{s}) \|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}\|$$

where  $\tilde{\zeta}_i$  lies between  $\zeta(\mathbf{s})$  and  $\zeta(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}$ . Thus,

$$H_n(\mathbf{u}) = A_1(\mathbf{s})^T \mathbf{u} \quad (33)$$

$$+ (1/2)\mathbf{u}^T A_2(\mathbf{s}) \mathbf{u} \quad (34)$$

$$+ (h^{-3}n^{-3/2}/6) \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) q_3 \left( \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \tilde{\zeta}_i, Y(\mathbf{s}_i) \right) \left[ \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \mathbf{u} \right]^3 \quad (35)$$

$$+ \sum_{j=1}^p \phi_j(\mathbf{s}) \{ \|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}\| - \|\zeta_j(\mathbf{s})\| \}.$$

Since



$$O \left( E \left| \left( h^{-3} n^{-3/2} / 6 \right) \sum_{i=1}^n K_h (\| \mathbf{s} - \mathbf{s}_i \|) q_3 \left( \{ \mathbf{Z}(\mathbf{s}_i) \}_i^T \tilde{\boldsymbol{\zeta}}_i, Y(\mathbf{s}_i) \right) \left[ \{ \mathbf{Z}(\mathbf{s}_i) \}_i^T \mathbf{u} \right]^3 \right| \right) = O(n^{-1/2} h^{-1}),$$

the third term in (33) is  $O_p(n^{-1/2} h^{-1})$ .

Note that the limiting behavior of the last term of (33) differs between the cases  $j \leq p_0$  and  $j > p_0$ .

**Case  $j \leq p_0$ :** If  $j \leq p_0$ , then  $n^{-1/2} \phi_j(\mathbf{s}) \rightarrow n^{-1/2} \lambda_n(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s})\|^{-\gamma}$  and  $|\sqrt{n} \{ \|\boldsymbol{\zeta}_j(\mathbf{s}) + h^{-1} n^{-1/2} \mathbf{u}_j\| - \|\boldsymbol{\zeta}_j(\mathbf{s})\| \}| \leq h^{-1} \|\mathbf{u}_j\|$ . Thus,

$$\lim_{n \rightarrow \infty} \phi_j(\mathbf{s}) (\|\boldsymbol{\zeta}_j(\mathbf{s}) + h^{-1} n^{-1/2} \mathbf{u}_j\| - \|\boldsymbol{\zeta}_j(\mathbf{s})\|) \leq h^{-1} n^{-1/2} \phi_j(\mathbf{s}) \|\mathbf{u}_j\| \leq h^{-1} n^{-1/2} a_n \|\mathbf{u}_j\| \rightarrow 0$$

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

Since  $h = O(n^{-1/6})$ , if  $h n^{-1/2} b_n \xrightarrow{p} \infty$ , then  $h^{-1} n^{-1/2} b_n \xrightarrow{p} \infty$ .

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

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

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

Thus, the limit of  $H_n(\mathbf{u})$  is the same as the limit of  $H_n^*(\mathbf{u})$  where

$$H_n^*(\mathbf{u}) = A_1(\mathbf{s})^T \mathbf{u} + (1/2) \mathbf{u}^T A_2(\mathbf{s}) \mathbf{u} + o_p(1)$$

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

$$\hat{\mathbf{u}}_n = \{A_2(\mathbf{s})\}^{-1} A_1(\mathbf{s}) \quad (36)$$

by the quadratic approximation lemma (Fan and Gijbels, 1996).

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

$$\hat{\mathbf{u}}_n \xrightarrow{d} N \left( \frac{\kappa_2 h^2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_j(\mathbf{s}) + \nabla_{vv}^2 \zeta_j(\mathbf{s}) \}, f(\mathbf{s}) \kappa_0^{-2} \nu_0 \sigma^2 \Psi^{-1} \right) \quad (37)$$

the result of Theorem 1 follows.

*Proof of Theorem 2.* We showed in Theorem 1 that  $\hat{\zeta}_j(\mathbf{s}) \xrightarrow{p} \zeta_j(\mathbf{s}) + \frac{\kappa_2 h^2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_j(\mathbf{s}) + \nabla_{vv}^2 \zeta_j(\mathbf{s}) \}$ , so to complete the proof of selection consistency, it only remains to show that  $Pr \left\{ \hat{\zeta}_j(\mathbf{s}) = \underline{0} \right\} \rightarrow 1$  if  $j > p_0$ .  $\square$

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

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

$$0 = \mathbf{Z}_p^T(\mathbf{s}) \mathbf{W}(\mathbf{s}) \left\{ \mathbf{Y} - \mathbf{Z}_{-p}(\mathbf{s}) \hat{\zeta}_{-p}(\mathbf{s}) - \mathbf{Z}_p(\mathbf{s}) \hat{\zeta}_p(\mathbf{s}) \right\} - \phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|} \quad (38)$$

$$= \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \zeta(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta(\mathbf{s}) + \nabla_{vv}^2 \zeta(\mathbf{s}) \} \right] \quad (39)$$

$$+ \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{-p}(\mathbf{s}) \left[ \zeta_{-p}(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_{-p}(\mathbf{s}) + \nabla_{vv}^2 \zeta_{-p}(\mathbf{s}) \} - \hat{\zeta}_{-p}(\mathbf{s}) \right] \quad (40)$$

$$+ \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_p(\mathbf{s}) \left[ \zeta_p(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_p(\mathbf{s}) + \nabla_{vv}^2 \zeta_p(\mathbf{s}) \} - \hat{\zeta}_p(\mathbf{s}) \right] \quad (41)$$

$$- \phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|} \quad (42)$$

Thus,

$$\frac{h}{\sqrt{n}} \phi_p(\mathbf{s}) \frac{\hat{\boldsymbol{\zeta}}_p(\mathbf{s})}{\|\hat{\boldsymbol{\zeta}}_p(\mathbf{s})\|} = \quad (43)$$

$$\mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \frac{h}{\sqrt{n}} \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \boldsymbol{\zeta}(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \boldsymbol{\zeta}(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}(\mathbf{s}) \} \right] \quad (44)$$

$$\begin{aligned} &+ \left\{ n^{-1} \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{-p}(\mathbf{s}) \right\} h\sqrt{n} \left[ \boldsymbol{\zeta}_{-p}(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \boldsymbol{\zeta}_{-p}(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_{-p}(\mathbf{s}) \} \right] \\ &+ \left\{ n^{-1} \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_p(\mathbf{s}) \right\} h\sqrt{n} \left[ \boldsymbol{\zeta}_p(\mathbf{s}) + \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \boldsymbol{\zeta}_p(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_p(\mathbf{s}) \} \right] - \end{aligned} \quad (45)$$

From Lemma 2 of Sun et al. (2014),  $\left\{ n^{-1} \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{-p}(\mathbf{s}) \right\} = O_p(1)$  and  $\left\{ n^{-1} \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_p(\mathbf{s}) \right\} = O_p(1)$ .

From Theorem 3 of Sun et al. (2014), we have that  $h\sqrt{n} \left[ \hat{\boldsymbol{\zeta}}_{-p}(\mathbf{s}) - \boldsymbol{\zeta}_{-p}(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \boldsymbol{\zeta}_{-p}(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_{-p}(\mathbf{s}) \} \right] = O_p(1)$  and  $h\sqrt{n} \left[ \hat{\boldsymbol{\zeta}}_p(\mathbf{s}) - \boldsymbol{\zeta}_p(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \boldsymbol{\zeta}_p(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_p(\mathbf{s}) \} \right] = O_p(1)$ .

We showed in the proof of Theorem 1 that

$$h\sqrt{n} \mathbf{Z}_p(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \boldsymbol{\zeta}(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \boldsymbol{\zeta}(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}(\mathbf{s}) \} \right] = O_p(1).$$

The right hand side of (43) is  $O_p(1)$ , so for  $\hat{\boldsymbol{\zeta}}_p(\mathbf{s})$  to be a solution, we must have that  $hn^{-1/2} \phi_p(\mathbf{s}) \hat{\boldsymbol{\zeta}}_p(\mathbf{s}) / \|\hat{\boldsymbol{\zeta}}_p(\mathbf{s})\| = O_p(1)$ .

But since by assumption  $\hat{\boldsymbol{\zeta}}_p(\mathbf{s}) \neq \mathbf{0}$ , there must be some  $k \in \{1, 2, 3\}$  such that  $|\hat{\zeta}_{p_k}(\mathbf{s})| = \max\{|\hat{\zeta}_{p_{k'}}(\mathbf{s})| : 1 \leq k' \leq 3\}$ . And for this  $k$ , we have that  $|\hat{\zeta}_{p_k}(\mathbf{s})|/\|\hat{\boldsymbol{\zeta}}_p(\mathbf{s})\| \geq 1/\sqrt{3} > 0$ .

Since  $hn^{-1/2}b_n \rightarrow \infty$ , we have that  $hn^{-1/2}\phi_p(\mathbf{s})\hat{\boldsymbol{\zeta}}_p(\mathbf{s})/\|\hat{\boldsymbol{\zeta}}_p(\mathbf{s})\| \geq hb_n/\sqrt{3n} \rightarrow \infty$  and therefore the left hand side of (43) dominates the sum to the right side. Thus, for large enough  $n$ ,  $\hat{\boldsymbol{\zeta}}_p(\mathbf{s}) \neq \mathbf{0}$  cannot maximize  $\mathcal{J}(\cdot)$ , and therefore  $Pr\left\{\hat{\boldsymbol{\zeta}}_{(b)}(\mathbf{s}) = \mathbf{0}\right\} \rightarrow 1$ .

## 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., N. E. Heckman, and M. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* 90, 141–150.
- 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 infor-

- mation 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*, 661–675.
- 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 nonparametric kernel 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*.
- Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis* 52, 5277–5286.
- 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.