

Local Adaptive Grouped Regularization and its Oracle Properties for Varying Coefficient Regression

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 mostly do global selection, which includes or excludes each covariate over the entire domain. Presented here is a new local adaptive grouped regularization (LAGR) method for local variable selection in spatially varying coefficient linear and generalized linear regression. LAGR selects the covariates that are associated with the response at any point in space, and simultaneously estimates the coefficients of those covariates by tailoring the adaptive group Lasso toward 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 finite sample properties of LAGR are assessed in a simulation study and for illustration, the Boston housing price data set is analyzed.

Keywords: adaptive Lasso, local generalized linear regression, local linear regression, regularization method, nonparametric, varying coefficient model

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 parameter (Cleveland and Grosse, 1991; Hastie and Tibshirani, 1993). Here we treat the case of a VCR model on a spatial domain where the spatial location 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 any 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 given 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. 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 model with a univariate effect-modifying parameter, and by Sun et al. (2014) for a two-dimensional effect-modifying parameter in a spatial VCR with autocorrelation. These methods mitigate the boundary effect by estimating the coefficients as local polynomials of odd degree (usually locally linear) (Hastie and Loader, 1993). In this work, we focus on a two-dimensional effect-modifying parameter and discuss the effect of different dimensionality on the results.

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

(Lasso) is a penalized regression method that simultaneously selects covariates 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 observed covariates 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 covariates (Yuan and Lin, 2006; Wang and Leng, 2008). The main innovation of the proposed LAGR method is to tailor the adaptive group Lasso toward 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 from varying coefficient linear regression to varying coefficient generalized linear regression for responses that are not necessarily Gaussian. 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 paper is organized as follows. The kernel-based estimation of a VCR model is described in Section 2. The proposed LAGR technique for varying coefficient linear regression and its oracle properties are presented in Section 3. In Section 4, the finite-sample properties of LAGR are evaluated in a simulation study, and in Section 5 LAGR is applied to the Boston housing price dataset. In Section 6, LAGR is extended to varying coefficient generalized linear regression and the oracle properties for this setting are established, followed by conclusions and discussion in section 7. Technical proofs are left to the appendices.

2. Varying Coefficient Regression

2.1. Varying Coefficient Model

Consider n observation locations $\mathbf{s}_i = (s_{i,1}, s_{i,2})^T$ for $i = 1, \dots, n$, which are distributed in a domain $\mathcal{D} \subset \mathbb{R}^2$ according to a density f . For $i = 1, \dots, n$, let $Y_i = Y(\mathbf{s}_i)$ and $\mathbf{X}_i = \mathbf{X}(\mathbf{s}_i)$ denote, respectively, the univariate response and the $(p+1)$ -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 the two dimensions and ε_i is random error at location \mathbf{s}_i . That is,

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta}(\mathbf{s}_i) + \varepsilon_i. \quad (1)$$

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

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 design matrix with covariate-by-location interactions in two dimensions (Wang et al., 2008). Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ be the design matrix of observed covariate values. Then the augmented local design matrix at location \mathbf{s}_i is defined to be $\mathbf{Z}(\mathbf{s}_i) = (\mathbf{X} \quad \mathbf{L}_i \mathbf{X} \quad \mathbf{M}_i \mathbf{X})$, where $\mathbf{L}_i = \text{diag}\{s_{i',1} - s_{i,1}\}_{i'=1}^n$ and $\mathbf{M}_i = \text{diag}\{s_{i',2} - s_{i,2}\}_{i'=1}^n$.

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

2.2. Coefficient Estimation via Local Likelihood

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

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

Since there are a total of $3(p+1)n + 1$ 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 $\boldsymbol{\zeta}(\mathbf{s})$ at location \mathbf{s} can be approximated by the coefficients $\boldsymbol{\zeta}(\mathbf{t})$, where \mathbf{t} is within some neighborhood of \mathbf{s} . This intuition is formalized by the following local log-likelihood at location $\mathbf{s} \in \mathcal{D}$:

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

where h is a bandwidth parameter, $\|\cdot\|$ is the ℓ_2 -norm, and $K_h(\|\mathbf{s} - \mathbf{s}_i\|)$ for $i = 1, \dots, n$ are local weights from a kernel function. For instance, the Epanechnikov kernel is defined as $K_h(\|\mathbf{s}_i - \mathbf{s}_{i'}\|) = h^{-2} K(h^{-1} \|\mathbf{s}_i - \mathbf{s}_{i'}\|)$ where $K(x) = (3/4)(1 - x^2)$ if $x < 1$, and 0 otherwise (Samiuddin and el Sayyad, 1990).

The local log-likelihood (3) is maximized to obtain an estimate $\tilde{\boldsymbol{\zeta}}(\mathbf{s})$ of the local coefficients at \mathbf{s} . Let $\mathbf{W}(\mathbf{s}) = \text{diag} \{K_h(\|\mathbf{s} - \mathbf{s}_i\|)\}_{i=1}^n$ denote a diagonal matrix of kernel weights. The local likelihood (3) can be maximized by minimizing a locally weighted least squares:

$$\mathcal{S}(\boldsymbol{\zeta}(\mathbf{s})) = (1/2) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}^T \mathbf{W}(\mathbf{s}) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}. \quad (4)$$

The minimizer of (4) is

$$\tilde{\boldsymbol{\zeta}}(\mathbf{s}) = \{\mathbf{Z}(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}(\mathbf{s})\}^{-1} \mathbf{Z}(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Y}. \quad (5)$$

By Theorem 3 of Sun et al. (2014), for any given \mathbf{s} , the estimated local coefficients $\tilde{\boldsymbol{\beta}}(\mathbf{s}) = (\tilde{\zeta}_1(\mathbf{s}), \dots, \tilde{\zeta}_p(\mathbf{s}))^T$ converge in probability at the optimal rate of $O(n^{-1/3})$ and are asymptotically normally distributed. The bias of the local coefficient estimates is 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 (5) has traditionally relied on variable selection *a priori*; that is, a set of covariates is pre-determined. Here we develop a new method of penalized regression to simultaneously select covariates locally and estimate the corresponding 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 $j = 1, \dots, p$. The proposed LAGR penalty is akin to the adaptive group Lasso (Yuan and Lin, 2006; Wang and Leng, 2008). By the mechanism of the adaptive group Lasso, covariates 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 \mathbf{s} , we minimize a penalized local sum of squares at location \mathbf{s} :

$$\mathcal{J}(\boldsymbol{\zeta}(\mathbf{s})) = \mathcal{S}(\boldsymbol{\zeta}(\mathbf{s})) + \mathcal{P}(\boldsymbol{\zeta}(\mathbf{s})), \quad (6)$$

where $\mathcal{S}(\boldsymbol{\zeta}(\mathbf{s}))$ is the locally weighted least squares defined in (4), $\mathcal{P}(\boldsymbol{\zeta}(\mathbf{s})) = \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_{(j)}(\mathbf{s})\|$ is a local adaptive grouped regularization (LAGR) penalty. The LAGR penalty for the j th

group of coefficients at location \mathbf{s} is $\phi_j(\mathbf{s}) = \lambda_n \|\tilde{\boldsymbol{\zeta}}_{(j)}(\mathbf{s})\|^{-\gamma}$, where $\lambda_n > 0$ is a local tuning parameter applied to all coefficients at location \mathbf{s} , $\tilde{\boldsymbol{\zeta}}_{(j)}(\mathbf{s})$ is a the vector of unpenalized local coefficients for the j th covariate and its interactions on location from (5), and $\gamma > 1$.

Minimization of (6) is by blockwise coordinate descent, where each block is a covariate group (one raw covariate and its interactions on location). Software for estimating $\boldsymbol{\zeta}(\mathbf{s})$ will be made available in an R package.

3.2. Oracle Properties

At location \mathbf{s} , let there be $p_0 < p$ covariates $\mathbf{X}_{(a)}(\mathbf{s})$ with nonzero local regression coefficients, denoted $\boldsymbol{\beta}_{(a)}(\mathbf{s}) \neq \mathbf{0}$. Without loss of generality, assume the indices of these covariates are $1, \dots, p_0$. The remaining $p - p_0$ covariates $\mathbf{X}_{(b)}(\mathbf{s})$ have coefficients equal to zero, denoted $\boldsymbol{\beta}_{(b)}(\mathbf{s}) = \mathbf{0}$. Denote by $a_n = \max \{\phi_j(\mathbf{s}), j \leq p_0\}$ the largest penalty applied to a covariate group whose true coefficient norm is nonzero, and by $b_n = \min \{\phi_j(\mathbf{s}), j > p_0\}$ the smallest penalty applied to a covariate group whose true coefficient norm is zero. 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 $\boldsymbol{\zeta}_{(k)}(\mathbf{s})$ be the augmented coefficients for covariate group k and $\boldsymbol{\zeta}_{(-k)}(\mathbf{s})$ be the augmented coefficients for all covariate groups except k . Let $\nabla \zeta_j(\mathbf{s}) = (\nabla_u \zeta_j(\mathbf{s}), \nabla_v \zeta_j(\mathbf{s}))^T$ and $\nabla^2 \zeta_j(\mathbf{s}) = \begin{pmatrix} \nabla_{uu}^2 \zeta_j(\mathbf{s}) & \nabla_{uv}^2 \zeta_j(\mathbf{s}) \\ \nabla_{vu}^2 \zeta_j(\mathbf{s}) & \nabla_{vv}^2 \zeta_j(\mathbf{s}) \end{pmatrix}$. Let $\kappa_0 = \int_{\mathbb{R}^2} K(\|\mathbf{s}\|) d\mathbf{s}$, $\kappa_2 = \int_{\mathbb{R}^2} [(1, 0)\mathbf{s}]^2 K(\|\mathbf{s}\|) d\mathbf{s} = \int_{\mathbb{R}^2} [(0, 1)\mathbf{s}]^2 K(\|\mathbf{s}\|) d\mathbf{s}$, and $\nu_0 = \int_{\mathbb{R}^2} K^2(\|\mathbf{s}\|) d\mathbf{s}$. Finally, let \xrightarrow{p} and \xrightarrow{d} denote convergence in probability and distribution, respectively, as $n \rightarrow \infty$.

Assume the following regularity conditions.

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

- (C.2) The coefficient functions $\beta_j(\cdot)$ for $j = 1, \dots, p$ have continuous second-order partial derivatives at \mathbf{s} .
- (C.3) The covariates $\mathbf{X}(\mathbf{s}_1), \dots, \mathbf{X}(\mathbf{s}_n)$ are random vectors that are independent of $\varepsilon_1, \dots, \varepsilon_n$. Also $\Psi(\mathbf{s}) = E\{\mathbf{X}(\mathbf{s})\mathbf{X}(\mathbf{s})^T | \mathbf{s}\}$ and $\Psi_{(a)}(\mathbf{s}) = E\{\mathbf{X}_{(a)}(\mathbf{s})\mathbf{X}_{(a)}(\mathbf{s})^T | \mathbf{s}\}$ are positive-definite and differentiable at location \mathbf{s} .
- (C.4) $E\{|\mathbf{X}(\mathbf{s})|^3 | \mathbf{s}\}$ and $E\{Y(\mathbf{s})^4 | \mathbf{X}(\mathbf{s}), \mathbf{s}\}$ are continuous at a given location \mathbf{s} .
- (C.5) The observation locations $\{\mathbf{s}_i\}$ are a sequence of fixed design points on a bounded compact support \mathcal{S} . Further, there exists a positive joint density function $f(\cdot)$ satisfying a Lipschitz condition such that

$$\sup_{\mathbf{s} \in \mathcal{S}} \left| n^{-1} \sum_{i=1}^n [r(\mathbf{s}_i) K_h(\|\mathbf{s}_i - \mathbf{s}\|)] - \int r(\mathbf{t}) K_h(\|\mathbf{t} - \mathbf{s}\|) f(\mathbf{t}) d\mathbf{t} \right| = O(h)$$

where $f(\cdot)$ is bounded away from zero on \mathcal{S} , $r(\cdot)$ is any bounded continuous function, and $K_h(\cdot) = K(\cdot/h)/h^2$.

(C.6) $h = O(n^{-1/6})$.

(C.7) $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$ and $hn^{-1/2}b_n \xrightarrow{p} \infty$.

Conditions (C.1)–(C.4) are common in the literature on nonparametric estimation, for instance see conditions (1)–(3) of Sun et al. (2014) and conditions (5) and (6) of Cai et al. (2000). However, the covariates $\mathbf{X}(\mathbf{s}_1), \dots, \mathbf{X}(\mathbf{s}_n)$ were assumed to be *iid* in Sun et al. (2014), which is not required here. The existence of $\Psi(\cdot)$ is needed for the existence of the limiting distribution of $\hat{\beta}(\mathbf{s})$; its differentiability is used in the Taylor's expansions. Condition (C.4) is used when bounding the remainder term in the Taylor's expansions. Condition (C.5) is the same as condition (4) of Sun et al. (2014), which is more general than strict infill asymptotics, as in the latter case the elements of $\{\mathbf{s}_i\}$ are *iid* as density $f(\cdot)$, and the Lipschitz condition in (C.5) is satisfied with probability one. Under condition (C.6), the

coefficient estimates attain the optimal rate of convergence for bivariate nonparametric regression. Condition (C.7) is needed for establishing the oracle properties, and is a refinement of the condition for the adaptive group lasso (Wang and Leng, 2008). In particular, satisfying (C.7) requires an additional restriction on γ , the unpenalized group norm exponent in the LAGR penalty.

Under condition (C.7), the local penalty tends to zero on covariates with true nonzero coefficients and to infinity on covariates with true zero coefficients. By (C.7), $h^{-1}n^{-1/2}\lambda_n \rightarrow 0$ for all $j \leq p_0$ and $hn^{-1/2}\lambda_n\|\zeta_{(j)}(\mathbf{s})\|^{-\gamma} \rightarrow \infty$ for all $j > p_0$. We require that λ_n satisfy both assumptions. Suppose $\lambda_n = n^\alpha$. Since $h = O(n^{-1/6})$ and $\|\tilde{\zeta}_{(p)}(\mathbf{s})\| = O(h^{-1}n^{-1/2})$, it follows that $h^{-1}n^{-1/2}\lambda_n = O(n^{-1/3+\alpha})$ and $hn^{-1/2}\lambda_n\|\tilde{\zeta}_{(p)}(\mathbf{s})\|^{-\gamma} = O(n^{-2/3+\alpha+\gamma/3})$. Thus, $(2 - \gamma)/3 < \alpha < 1/3$, which can only be satisfied for $\gamma > 1$.

Theorem 1 (Asymptotic normality). *Under (C.1)–(C.8),*

$$\begin{aligned} \{f(\mathbf{s})h^2n\}^{1/2} \left[\hat{\beta}_{(a)}(\mathbf{s}) - \beta_{(a)}(\mathbf{s}) - (2\kappa_0)^{-1}\kappa_2h^2 \left\{ \nabla_{uu}^2\beta_{(a)}(\mathbf{s}) + \nabla_{vv}^2\beta_{(a)}(\mathbf{s}) \right\} \right] \\ \xrightarrow{d} N(0, \kappa_0^{-2}\nu_0\sigma^2\Psi_{(a)}(\mathbf{s})^{-1}), \end{aligned}$$

where $\{\nabla_{uu}^2\beta_{(a)}(\mathbf{s}) + \nabla_{vv}^2\beta_{(a)}(\mathbf{s})\} = (\nabla_{uu}^2\beta_1(\mathbf{s}) + \nabla_{vv}^2\beta_1(\mathbf{s}), \dots, \nabla_{uu}^2\beta_{p_0}(\mathbf{s}) + \nabla_{vv}^2\beta_{p_0}(\mathbf{s}))^T$.

Theorem 2 (Selection consistency). *Under (C.1)–(C.8),*

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

Theorem 1 indicates that the LAGR estimates for true nonzero coefficients have the same asymptotic distribution as a local regression model where the true nonzero coefficients are known in advance. Further, by Theorem 2, the LAGR estimates of true zero coefficients tend to zero with probability one. Together, selection and local coefficient estimation by LAGR has the oracle property. The technical proofs of Theorems 1 and 2 are given in Appendix A.

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

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

where $I(\cdot)$ is the indicator function and the local coefficient estimate is written $\hat{\boldsymbol{\zeta}}(\lambda_n; \mathbf{s})$ 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

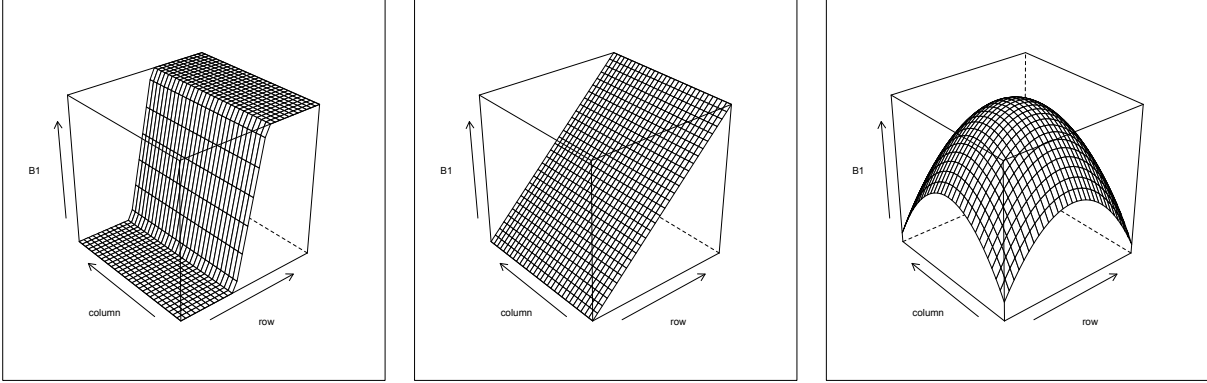


Figure 1: The step function (left), gradient function (middle), and parabola function (right) that were used as the coefficient function $\beta_1(\mathbf{s})$ in the VCR model $y_i = x_1(\mathbf{s}_i)\beta_1(\mathbf{s}_i) + \varepsilon_i$ when generating the data for the simulation study.

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 $\text{Cov}(X_{ij}, X_{i'j}) = \sigma_x^2 \exp(-\tau_x^{-1} \delta_{ii'})$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0.1$ is the range parameter, and $\delta_{ii'} = \|\mathbf{s}_i - \mathbf{s}_{i'}\|$. 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 $\mathbf{\Sigma} = \mathbf{R}^T \mathbf{R}$. The covariance matrix $\mathbf{\Sigma}$ 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 $y_i = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{s}_i) + \varepsilon_i$ for $i = 1, \dots, n$ where $n = 900$ and the ε_i 's were iid Gaussian with mean zero and variance σ_ε^2 . The simulated data included the response y and five covariates x_1, \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 1). Let $\mathbf{s} = (s_u, s_v)^T$. The first is a “step” function $\beta_{\text{step}}(\mathbf{s}) = 1$ if $s_u > 0.6$, $5s_u - 2$ if $0.4 < s_u \leq 0.6$, and 0 otherwise. The second is a gradient function, $\beta_{\text{gradient}}(\mathbf{s}) = s_u$, and the third is a parabola $\beta_{\text{parabola}}(\mathbf{s}) = 1 - 2\{(s_u - 0.5)^2 + (s_v - 0.5)^2\}$.

In total, three parameters were varied to produce 18 settings, each of which was 100 simulated data sets. There were the three functional forms for the coefficient surface $\beta_1(\mathbf{s})$; data were

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

Table 1: For each setting as a combination of coefficient surface $\beta_1(\mathbf{s})$, cross-covariate correlation ρ , and error variance σ^2 , the mean integrated squared error (MISE) of the coefficient estimates. In the middle three columns, 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 $\hat{\beta}_1$. In the last two columns, the MISE of $\hat{\beta}_2, \dots, \hat{\beta}_5$ from estimation by LAGR and VCR. Here, highlighting indicates the **smallest** MISE.

simulated 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 $\hat{\beta}_1(\mathbf{s}), \dots, \hat{\beta}_5(\mathbf{s})$, 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 oracle selection. Oracle selection means that exactly the correct set of covariates was used to fit each local model.

4.2. Simulation Results

The MISE of the coefficient surface estimates are in Table 1. The MISE of true nonzero coefficient $\hat{\beta}_1(\mathbf{s})$ was always the smallest under oracular selection. The difference in MISE of $\hat{\beta}_1(\mathbf{s})$ under LAGR versus VCR estimation was small and each method produced the smaller MISE for nine of 18 settings. 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 of $\beta_2(\mathbf{s}), \dots, \beta_5(\mathbf{s})$ 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 response $\hat{y}(\mathbf{s})$ 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 response.

Overall, the proposed LAGR method was accurate in variable selection and coefficient surface 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 little 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 applied 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

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

Table 2: For each setting as a combination of coefficient surface $\beta_1(\mathbf{s})$, cross-covariate correlation ρ , and error variance σ^2 , the frequency of exact zeroes in the estimates of $\hat{\beta}_2, \dots, \hat{\beta}_5$ as estimated by local adaptive grouped regularization (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.

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 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”). With the Epanechnikov kernel, 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.

A summary of the estimates of the local coefficients is in Table 3. It indicates that RM is the only covariate with a positive mean local coefficient estimate. The percentage of census tracts where the local coefficient was estimated to be exactly zero was 1% for LSTAT, 2% for RM, 37% for RAD, 49% for CRIM, and 100% for TAX.

The estimates of the local coefficients are plotted in Figure 2. One interesting result is that the TAX variable was nowhere found to be associated with 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 discernible) 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.

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 mean estimated 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).

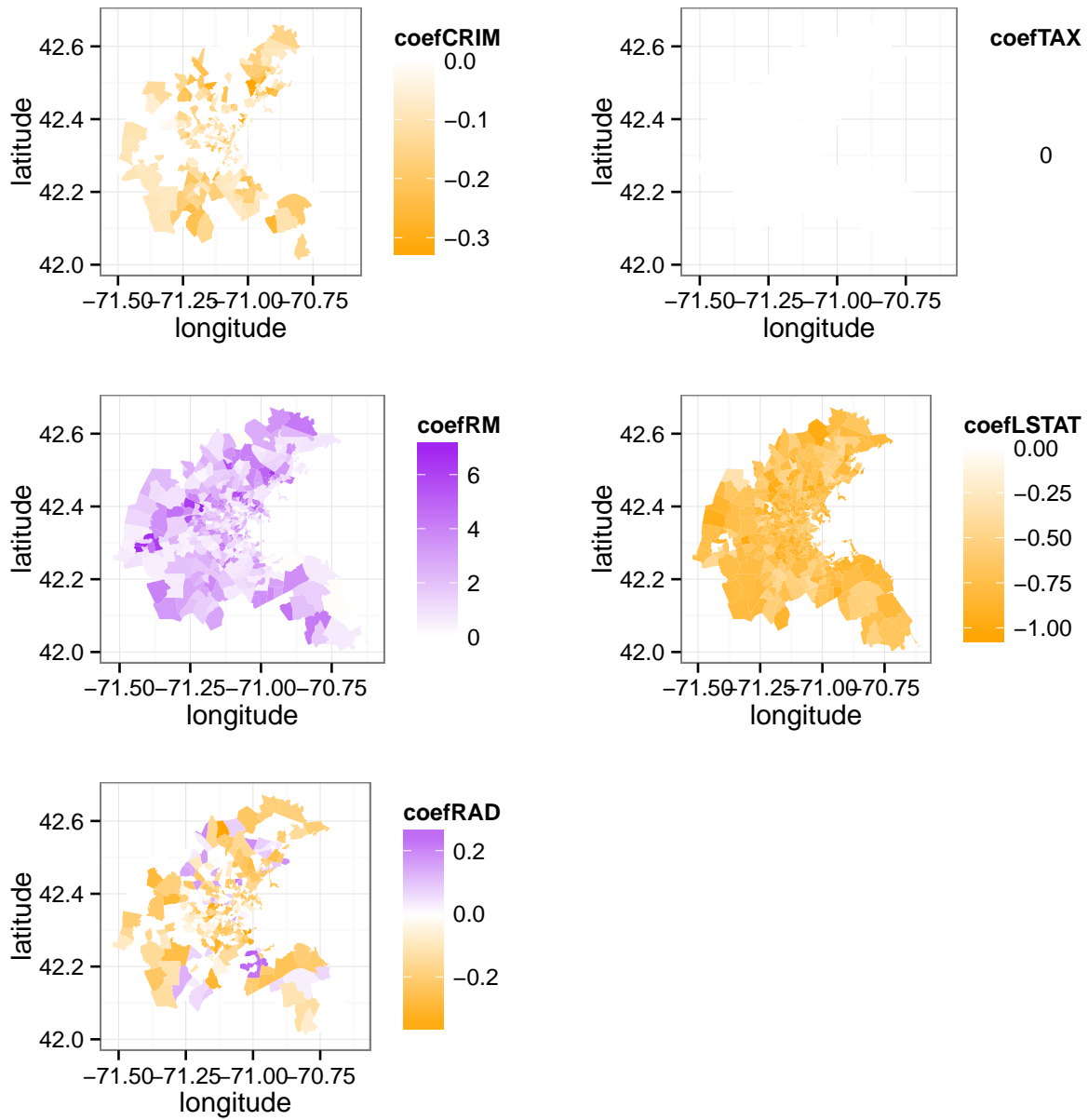


Figure 2: Coefficients for the Boston house price data as estimated by local adaptive grouped regularization.

	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 estimates of the local coefficients in a model for the median house price in census tracts in Boston, with coefficients selected and fitted by local adaptive grouped regularization. The covariates are CRIM (per capita crime rate in the census tract), RM (average number of rooms per home sold in the census tract), RAD (an index of the tract’s access to radial roads), TAX (property tax per USD10,000 of property value), and LSTAT (percentage of the tract’s residents who are considered “lower status”).

6. Extension to Generalized Linear Regression

6.1. Local GLM and Local Quasi-likelihood Estimation

Generalized linear models (GLMs) extend the linear regression model to a response variable following any distribution in the exponential family (McCullagh and Nelder, 1989). As is the case for the local linear regression model, we now consider local GLM coefficients as smooth functions of location (Cai et al., 2000). Suppose the response variable Y is from an exponential family distribution with $E\{y(\mathbf{s})|\mathbf{x}(\mathbf{s})\} = \mu(\mathbf{s}) = b'(\theta(\mathbf{s}))$, $\theta(\mathbf{s}) = (g \circ b')^{-1}(\eta(\mathbf{s}))$, $\eta(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T \boldsymbol{\beta}(\mathbf{s}) = g(\mu(\mathbf{s}))$, $\text{Var}\{y(\mathbf{s})|\mathbf{x}(\mathbf{s})\} = b''(\theta(\mathbf{s}))$, and link function $g(\cdot)$. Then the probability density is

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

If $g^{-1}(\cdot) = b'(\cdot)$, then the composition $(g \circ b')(\cdot)$ is the identity function. This particular g is called the canonical link. 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 local coefficients can be estimated by maximizing the local quasi-likelihood

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

The local quasi-likelihood (7) generalizes the local log-likelihood (3) that was used to estimate coefficients in the local linear regression. The local quasi-likelihood (7) is convex, and is defined in terms of its derivative, the local 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 function to zero:

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

where $\hat{\mu}(\mathbf{s}_i; \mathbf{s}) = g^{-1}(\mathbf{z}_i^T \hat{\boldsymbol{\zeta}}(\mathbf{s}))$ is the mean at location \mathbf{s}_i evaluated at the estimated coefficients $\hat{\boldsymbol{\zeta}}(\mathbf{s})$ at location \mathbf{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 the two dimensions, the arguments in the proof of Theorem 1 of Cai et al. (2000) can be extended to show that the distribution of the estimated local coefficients is:

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

6.2. LAGR Penalized Local Likelihood and Oracle Properties

Whereas the method of LAGR for local linear regression uses a penalized local likelihood, LAGR for GLMs uses a penalized local quasi-likelihood:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\zeta}(\mathbf{s})) &= \ell^*(\boldsymbol{\zeta}(\mathbf{s})) + \mathcal{P}(\boldsymbol{\zeta}(\mathbf{s})) \\ &= \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) Q(g^{-1}(\mathbf{z}_i^T \boldsymbol{\zeta}(\mathbf{s})), Y_i) + \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_{(j)}(\mathbf{s})\|. \end{aligned}$$

Further, let $\phi_j(\mathbf{s}) = \lambda_n \|\tilde{\boldsymbol{\zeta}}_{(j)}(\mathbf{s})\|^{-\gamma}$, where $\lambda_n > 0$ 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. The following are additional to the definitions and conditions of Section 3.2. Define $\rho(\mathbf{s}, \mathbf{z}) = [g_1(\mu(\mathbf{s}, \mathbf{z}))]^2 \text{Var}\{Y(\mathbf{s}) | \mathbf{X}(\mathbf{s}), \mathbf{s}\}$, 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, $\rho(\mathbf{s}, \mathbf{z}) = V(\mu(\mathbf{s}, \mathbf{z}))$. Let $\boldsymbol{\Gamma}(\mathbf{s}) = E\{\rho(\mathbf{s}, \mathbf{X}(\mathbf{s})) \mathbf{X}(\mathbf{s}) \mathbf{X}(\mathbf{s})^T | \mathbf{s}, \mathbf{Z}(\mathbf{s}) = \mathbf{z}\}$ and

$$\boldsymbol{\Gamma}_{(a)}(\mathbf{s}) = E\{\rho(\mathbf{s}, \mathbf{X}_{(a)}(\mathbf{s})) \mathbf{X}_{(a)}(\mathbf{s}) \mathbf{X}_{(a)}(\mathbf{s})^T | \mathbf{s}, \mathbf{Z}(\mathbf{s}) = \mathbf{z}\}.$$

Assume the following regularity conditions:

(C.8) The functions $g'''(\mathbf{s})$, $\nabla \boldsymbol{\Gamma}(\mathbf{s})$, $\nabla \boldsymbol{\Gamma}_{(a)}(\mathbf{s})$, $V(\mu(\mathbf{s}, \mathbf{z}))$, and $V'(\mu(\mathbf{s}, \mathbf{z}))$ are continuous at \mathbf{s} .

(C.9) The function $(\partial^2/\partial\mu^2) Q(g^{-1}(\mu), y) < 0$ for $\mu \in \mathbb{R}$ and y in the range of the response.

These additional conditions are not uncommon in the nonparametric regression literature (see, e.g., conditions (1) and (2) of Cai et al. (2000)). Condition (C.8) is needed for the Taylor's expansion of the local quasi-likelihood. Condition (C.9) assures that the local quasi-likelihood is convex and has a unique minimizer.

Theorem 3 (Asymptotic normality). *Under (C.1)–(C.10),*

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

Theorem 4 (Selection consistency). *Under (C.1)–(C.10),*

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

By Theorem 3, the LAGR estimates achieve the same asymptotic distribution as if the nonzero coefficients were known in advance. The difference between the Gaussian and GLM cases is that $\sigma^2 \boldsymbol{\Psi}_{(a)}(\mathbf{s})^{-1}$ in the variance term of Theorem 1 has been replaced by $\boldsymbol{\Gamma}_{(a)}(\mathbf{s})^{-1}$ in Theorem 3 because the variance of the response in the GLM case depends on the expectation of the response. Theorem 4 gives the same result for the GLM setting as Theorem 2 does for the Gaussian setting: the true zero coefficients are dropped from the model with probability tending to one. Thus, the oracle properties for the GLM setting are established. The technical proofs are given in Appendix B and the necessary lemmas are provided in the online supplementary materials.

7. Conclusions and Discussion

We have developed a new method of LAGR and shown its oracle properties for local variable selection and coefficient estimation in VCR models. This is in contrast to the existing literature on variable selection for VCR models that focuses on global variable selection. Further, the method of LAGR extends the adaptive group lasso. In particular, the previous literature on the adaptive group lasso is insufficient for local selection in a VCR model because the local weights are functions of the kernel $K(\cdot)$ and the bandwidth h . As a result, the local observation weights change with sample size and the coefficient estimates converge

at a slower rate than in the traditional adaptive group lasso. Thus, the conditions for oracle properties of the adaptive group lasso must be refined for the LAGR method.

Here we considered the case of two-dimensional effect-modifying parameter. Similar results can be obtained when the effect-modifying parameter has dimension other than two, but in higher dimensions the so-called “curse of dimensionality” means that the estimation accuracy quickly degrades. Since the optimal rate of convergence for nonparametric regression is achieved when $h = O\left(n^{-1/\{4+d\}}\right)$ where d is the dimension of the effect-modifying parameter, it follows that to attain the oracle properties, the exponent in the adaptive weights for LAGR estimation must satisfy $\gamma > d/2$.

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 C_p . *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 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* 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 A. Proofs of Theorems 1–2

Proof of Theorem 1

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

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

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 $n^{-1/2} \phi_j(\mathbf{s}) \rightarrow n^{-1/2} \lambda_n \|\boldsymbol{\zeta}_j(\mathbf{s})\|^{-\gamma}$ and $|n^{1/2} \{\|\boldsymbol{\zeta}_j(\mathbf{s}) + \alpha_n \mathbf{u}_j\| - \|\boldsymbol{\zeta}_j(\mathbf{s})\|\}| \leq h^{-1} \|\mathbf{u}_j\|$. Thus,

$$\phi_j(\mathbf{s}) (\|\boldsymbol{\zeta}_j(\mathbf{s}) + \alpha_n \mathbf{u}_j\| - \|\boldsymbol{\zeta}_j(\mathbf{s})\|) \leq \alpha_n \phi_j(\mathbf{s}) \|\mathbf{u}_j\| \leq \alpha_n a_n \|\mathbf{u}_j\| \rightarrow 0.$$

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

$$\alpha_n \phi_j(\mathbf{s}) \|\mathbf{u}_j\| \geq \alpha_n b_n \|\mathbf{u}_j\| \rightarrow \infty.$$

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

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

otherwise. It follows that $H_n^*(\mathbf{u})$ is convex and has a unique minimizer, called $\hat{\mathbf{u}}_n$. Let $\hat{\mathbf{u}}_{(a)n}$ and $\hat{\mathbf{u}}_{(b)n}$ be, respectively, the subvectors of \mathbf{u}_n corresponding to the true nonzero coefficients and true zero coefficients. Then

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

and $\hat{\mathbf{u}}_{(b)n} = \mathbf{0}$. By epiconvergence, the minimizer of the limiting function is the limit of the minimizers $\hat{\mathbf{u}}_n$ (Geyer, 1994; Knight and Fu, 2000). Since, by Lemma 2 of Sun et al. (2014),

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

the result of Theorem 1 follows. \square

Proof of Theorem 2

Proof. The proof is by contradiction. Without loss of generality we consider only the p th covariate group. Assume $\|\hat{\boldsymbol{\zeta}}_{(p)}(\mathbf{s})\| \neq 0$. Then $\mathcal{J}(\boldsymbol{\zeta}(\mathbf{s}))$ is differentiable w.r.t. $\boldsymbol{\zeta}_{(p)}(\mathbf{s})$ and is minimized where

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

Thus,

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

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

$$O_p(n^{-1} \mathbf{Z}_{(p)}(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{(-p)}(\mathbf{s})) = O_p(n^{-1} \mathbf{Z}_{(p)}(\mathbf{s})^T \mathbf{W}(\mathbf{s}) \mathbf{Z}_{(p)}(\mathbf{s})) = O_p(1).$$

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

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

and

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

We showed in the proof of Theorem 1 that

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

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

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

Appendix B. Proofs of Theorems 3–4

Proof of Theorem 3

The next proofs require the lemmas in the web-based supplemental material. First, let $\mathbf{z} \in \mathbb{R}^{3p}$. Define the q -functions to be the derivatives of the quasi-likelihood: $q_j(t, y) = (\partial/\partial t)^j Q(g^{-1}(t), y)$. Then $q_1(\eta(\mathbf{s}, \mathbf{z}), \mu(\mathbf{s}, \mathbf{z})) = \mathbf{0}$, and $q_2(\eta(\mathbf{s}, \mathbf{z}), \mu(\mathbf{s}, \mathbf{z})) = -\rho(\mathbf{s}, \mathbf{z})$. Let

$$\tilde{\beta}_i'' = \left[(\mathbf{s}_i - \mathbf{s})^T \{ \nabla^2 \beta_1(\mathbf{s}) \} (\mathbf{s}_i - \mathbf{s}), \dots, (\mathbf{s}_i - \mathbf{s})^T \{ \nabla^2 \beta_p(\mathbf{s}) \} (\mathbf{s}_i - \mathbf{s}) \right]^T$$

be the p -vector of quadratic forms of location interactions on the second derivatives of the coefficient functions.

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

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

Define

$$\Omega_n = \alpha_n \sum_{i=1}^n q_1(\mathbf{Z}_i^T \zeta(\mathbf{s}), Y_i) \mathbf{Z}_i K(h^{-1} \|\mathbf{s} - \mathbf{s}_i\|) = \alpha_n \sum_{i=1}^n \omega_i$$

and

$$\Delta_n = \alpha_n^2 \sum_{i=1}^n q_2(\mathbf{Z}_i^T \zeta(\mathbf{s}), Y_i) \mathbf{Z}_i \mathbf{Z}_i^T K(h^{-1} \|\mathbf{s} - \mathbf{s}_i\|) = \alpha_n^2 \sum_{i=1}^n \delta_i.$$

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

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

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

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

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

$$\lim_{n \rightarrow \infty} \phi_j(\mathbf{s}) (\|\zeta_{(j)}(\mathbf{s}) + \alpha_n \mathbf{u}_{(j)}\| - \|\zeta_{(j)}(\mathbf{s})\|) \leq \alpha_n \phi_j(\mathbf{s}) \|\mathbf{u}_{(j)}\| \leq \alpha_n a_n \|\mathbf{u}_{(j)}\| \rightarrow 0$$

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

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

On the other hand, if $\|\mathbf{u}_{(j)}\| = 0$, then $\alpha_n \phi_j(\mathbf{s}) \|\mathbf{u}_{(j)}\| = 0$. By Lemma 1, $\Delta_n = \Delta + O_p(\alpha_n)$, so the limit of $H_n(\mathbf{u})$ is the same as the limit of $H_n^*(\mathbf{u})$ where

$$H_n^*(\mathbf{u}) = \Omega_{(a)n}^T \mathbf{u}_{(a)} + (1/2) \mathbf{u}_{(a)}^T \Delta_{(a)} \mathbf{u}_{(a)} + o_p(1)$$

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 has a unique minimizer, called $\hat{\mathbf{u}}_n$. Let $\hat{\mathbf{u}}_{(a)n}$, $\Delta_{(a)}$ and $\Omega_{(a)n}$ be, respectively, the parts of \mathbf{u}_n , Δ ,

and Ω_n corresponding to the true nonzero coefficients, and let $\hat{\mathbf{u}}_{(b)n}$ be the subvector of $\hat{\mathbf{u}}_n$ corresponding to the true zero coefficients. Then

$$\hat{\mathbf{u}}_{(a)n} = \Delta_{(a)}^{-1} \Omega_{(a)n} + o_p(1) \text{ and } \hat{\mathbf{u}}_{(b)n} = \mathbf{0}$$

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

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

Then $\mathbf{d}^T \Omega_n = \alpha_n \sum_{i=1}^n \xi_i$. We establish the normality of $\mathbf{d}^T \Omega_n$ by checking the Lyapunov condition of the sequence $\left\{ \mathbf{d}^T \text{Var}(\Omega_n) \mathbf{d} \right\}^{-1/2} \left\{ \mathbf{d}^T \Omega_n - \mathbf{d}^T E \Omega_n \right\}$. By boundedness of $K(\cdot)$, linearity of $q_1(\mathbf{Z}_i^T \boldsymbol{\zeta}(\mathbf{s}), Y_i)$ in Y_i , and conditions (C.6) and (C.8), we have that

$$n \alpha_n^3 E(|\xi_1|^3) = O(\alpha_n) \rightarrow 0. \quad (\text{B.2})$$

We observe that (B.2) implies that $n \alpha_n^3 |E(\xi_1)|^3 \rightarrow 0$, and since $E(|\xi_1 - E\xi_1|^3) < E\{(|\xi_1| + |E\xi_1|)^3\} \rightarrow 0$, the Lyapunov condition is satisfied. Thus, Ω_n asymptotically follows a Gaussian distribution and the result follows from the quadratic approximation lemma. \square

Proof of Theorem 4

Proof. The proof is by contradiction. Without loss of generality we consider only the p th covariate group. Assume $\|\hat{\boldsymbol{\zeta}}_{(p)}(\mathbf{s})\| \neq 0$. Then $\mathcal{J}(\boldsymbol{\zeta}(\mathbf{s}))$ is differentiable w.r.t. $\boldsymbol{\zeta}_{(p)}(\mathbf{s})$ and is minimized where

$$\phi_p(\mathbf{s}) \hat{\boldsymbol{\zeta}}_{(p)}(\mathbf{s}) \|\hat{\boldsymbol{\zeta}}_{(p)}(\mathbf{s})\|^{-1} = \sum_{i=1}^n q_1 \left(\mathbf{Z}_i^T \hat{\boldsymbol{\zeta}}(\mathbf{s}), Y_i \right) \mathbf{Z}_{i(p)} K \left(h^{-1} \|\mathbf{s}_i - \mathbf{s}\| \right) \quad (\text{B.3})$$

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