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 $\boldsymbol{s}_i = (s_{i,1},\ s_{i,2})^T$ for $i = 1, \ldots, n$, which are distributed in a domain $\mathcal{D} \subset \mathbb{R}^2$ according to a density f. For $i = 1, \ldots, n$, let $Y_i = Y(\boldsymbol{s}_i)$ and $\boldsymbol{X}_i = \boldsymbol{X}(\boldsymbol{s}_i)$ denote, respectively, the univariate response and the $(p+1)$–vector of covariates measured at location $\boldsymbol{s}_i$. At each location $\boldsymbol{s}_i$, assume that the outcome is related to the covariates by a linear regression where the coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i)$ are functions in the two dimensions and ε_i is random error at location $\boldsymbol{s}_i$. That is,

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

Further assume that the error term ε_i is normally distributed with zero mean and variance σ^2, and that ε_i, $i = 1, \ldots, n$ are independent. That is, for $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ where $\boldsymbol{I}_n$ 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 and the coefficient surfaces with interactions on location in two dimensions (Wang et al., 2008). Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ be the design matrix of observed covariate values. Then the augmented local design matrix at location $\boldsymbol{s}$ is defined to be $\boldsymbol{Z}(\boldsymbol{s}) = (\boldsymbol{X} \ \ \boldsymbol{L}(\boldsymbol{s})\boldsymbol{X} \ \ \boldsymbol{M}(\boldsymbol{s})\boldsymbol{X})$, where $\boldsymbol{L}(\boldsymbol{s}) = \mathrm{diag}\{s_{i',1} - s_1\}_{i'=1}^n$ and $\boldsymbol{M}(\boldsymbol{s}) = \mathrm{diag}\{s_{i',2} - s_2\}_{i'=1}^n$. The vector of augmented local coefficients at location $\boldsymbol{s}$ is defined to be $\boldsymbol{\zeta}(\boldsymbol{s}) = \left(\boldsymbol{\beta}(\boldsymbol{s})^T,\ \nabla_u \boldsymbol{\beta}(\boldsymbol{s})^T,\ \nabla_v \boldsymbol{\beta}(\boldsymbol{s})^T\right)^T$, where $\nabla_u \boldsymbol{\beta}(\boldsymbol{s})$ and $\nabla_v \boldsymbol{\beta}(\boldsymbol{s})$ denote the local gradients of the coefficient surfaces.

2.2. Coefficient Estimation via Local Likelihood

Let $\boldsymbol{\zeta} = (\boldsymbol{\zeta}(\boldsymbol{s}_1), \ldots, \boldsymbol{\zeta}(\boldsymbol{s}_n))^T$ denote a matrix of the local coefficients at all observation locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ and let $\{\boldsymbol{Z}(\boldsymbol{s})\}_i$ denote the ith row of the matrix $\boldsymbol{Z}(\boldsymbol{s})$ as a column vector.

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

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

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

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

where $\boldsymbol{Z}_i = \left\{\boldsymbol{Z}(\boldsymbol{s})\right\}_i$, h is a bandwidth parameter, $\|\cdot\|$ is the ℓ_2-norm, and $K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)$ for $i = 1, \ldots, n$ are local weights from a kernel function. For instance, the Epanechnikov kernel is defined as $K_h(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|) = h^{-2}K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|\right)$ 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 $\tilde{\boldsymbol{\zeta}}(\boldsymbol{s})$ of the local coefficients at $\boldsymbol{s}$. Let $\boldsymbol{W}(\boldsymbol{s}) = \text{diag}\left\{K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)\right\}_{i'=1}^n$ denote a diagonal matrix of kernel weights. The local likelihood (3) can be maximized by minimizing a locally weighted least squares:

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

The minimizer of (4) is

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

By Theorem 3 of Sun et al. (2014), for any given $\boldsymbol{s}$, the estimated local coefficients $\tilde{\boldsymbol{\beta}}(\boldsymbol{s}) = \left(\tilde{\zeta}_1(\boldsymbol{s}), \ldots, \tilde{\zeta}_p(\boldsymbol{s})\right)^T$ converge in probability at the optimal rate of $O\left(n^{-1/3}\right)$ 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, $\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) = (\beta_j(\boldsymbol{s}),\ \nabla_u\beta_j(\boldsymbol{s}),\ \nabla_v\beta_j(\boldsymbol{s}))^T$ for $j = 1, \ldots, p$. The proposed LAGR penalty is 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 $\boldsymbol{s}$, we minimize a penalized local sum of squares at location $\boldsymbol{s}$:

$$\mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \mathcal{S}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) + \mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right), \tag{6}$$

where $\mathcal{S}\left(\boldsymbol{\zeta}\left(\boldsymbol{s}\right)\right)$ is the locally weighted least squares defined in (4), $\mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \sum_{j=1}^{p} \phi_j(\boldsymbol{s})\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|$ is a local adaptive grouped regularization (LAGR) penalty. The LAGR penalty for the jth group of coefficients at location $\boldsymbol{s}$ is $\phi_j(\boldsymbol{s}) = \lambda_n\|\tilde{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\|^{-\gamma}$, where $\lambda_n > 0$ is a local tuning parameter applied to all coefficients at location $\boldsymbol{s}$, $\tilde{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})$ is a the vector of unpenalized local coefficients for the jth 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 $\boldsymbol{\zeta}(\boldsymbol{s})$ will be made available in an `R` package.

3.2. Oracle Properties

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

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 $j = 1, \ldots, p$ have continuous second-order partial derivatives at $\boldsymbol{s}$.

(C.3) The covariates $\boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)$ are random vectors that are independent of $\varepsilon_1, \ldots, \varepsilon_n$. Also $\boldsymbol{\Psi}(\boldsymbol{s}) = E\left\{\boldsymbol{X}(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s})^T|\boldsymbol{s}\right\}$ and $\boldsymbol{\Psi}_{(a)}(\boldsymbol{s}) = E\left\{\boldsymbol{X}_{(a)}(\boldsymbol{s})\boldsymbol{X}_{(a)}(\boldsymbol{s})^T|\boldsymbol{s}\right\}$ are positive-definite and differentiable at location $\boldsymbol{s}$.

(C.4) $E\left\{|\boldsymbol{X}(\boldsymbol{s})|^3 |\boldsymbol{s}\right\}$ and $E\left\{Y(\boldsymbol{s})^4 | \boldsymbol{X}(\boldsymbol{s}), \boldsymbol{s}\right\}$ are continuous at a given location $\boldsymbol{s}$.

(C.5) The observation locations $\{\boldsymbol{s}_i\}$ 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

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

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

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

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 $\boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)$ were assumed to be *iid* in Sun et al. (2014), which is not required here. The existence of of $\boldsymbol{\Psi}(\cdot)$ is needed for the existence of the limiting distribution of $\hat{\boldsymbol{\beta}}(\boldsymbol{s})$; 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 $\{\boldsymbol{s}_i\}$ 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), $h^{-1}n^{-1/2}\lambda_n \to 0$ for all $j \leq p_0(\boldsymbol{s})$ and $hn^{-1/2}\lambda_n\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|^{-\gamma} \to \infty$ for all $j > p_0(\boldsymbol{s})$. We require that λ_n satisfy both assumptions. Suppose $\lambda_n = n^{\alpha}$. Since $h = O\left(n^{-1/6}\right)$ and $\|\tilde{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\| = O\left(h^{-1}n^{-1/2}\right)$, it follows that $h^{-1}n^{-1/2}\lambda_n = O\left(n^{-1/3+\alpha}\right)$ and $hn^{-1/2}\lambda_n\|\tilde{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-\gamma} = O\left(n^{-2/3+\alpha+\gamma/3}\right)$. Thus, $(2-\gamma)/3 < \alpha < 1/3$, which can only be satisfied for $\gamma > 1$.

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

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

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

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

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

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 AIC to select the tuning parameter λ_n. That is, let

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

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

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

4. Simulation Study

4.1. Simulation Setup

A simulation study was conducted to assess the performance of the method described in Sections 2–3. Data were simulated on the domain $[0, 1]^2$, which was divided into a 20×20 grid. Each of $p = 5$ covariates $X_1, \ldots, X_5$ was simulated by a Gaussian random field (GRF) with mean zero, nugget variance 0.2, and exponential covariance $\mathrm{Cov}(X_{ij}, X_{i'j}) = \sigma_x^2 \exp(-\tau_x^{-1}\delta_{ii'})$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0.1$ is the range parameter, and $\delta_{ii'} = \|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|$. Correlation was induced between the covariates by multiplying the design matrix

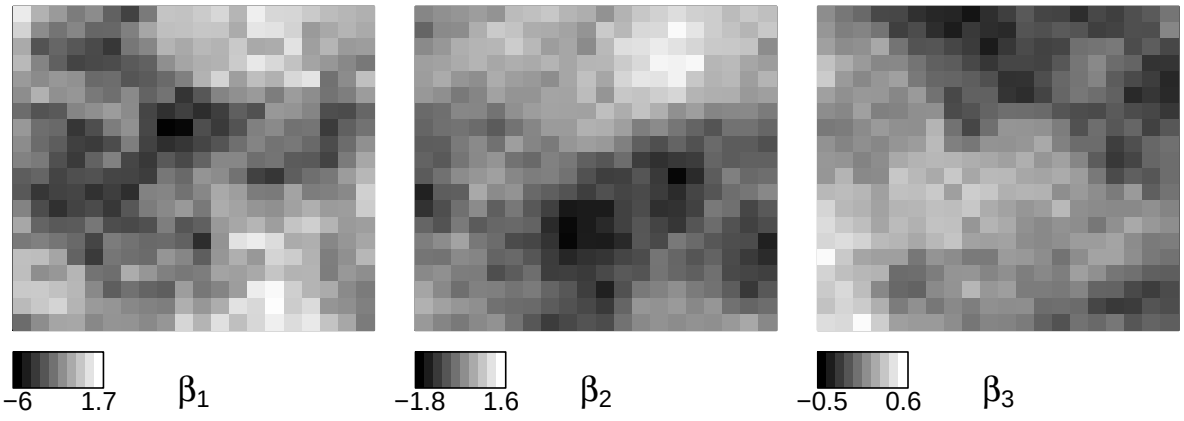

Figure 1: Left to right, the values used for coefficients $\beta_1(\boldsymbol{s}), \ldots, \beta_3(\boldsymbol{s})$ in the simulation study.

$\boldsymbol{X}$ by $\boldsymbol{R}$, where $\boldsymbol{R}$ is the Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{R}^T\boldsymbol{R}$. The covariance matrix $\boldsymbol{\Sigma}$ is a 5×5 matrix that has ones on the diagonal and ρ for all off-diagonal entries, where ρ is the between-covariate correlation.

The simulated response was $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon_i$ for $i = 1, \ldots, n$ where $n = 400$ and the ε_i's were iid Gaussian with mean zero and variance σ_ε^2. The coefficients $\beta_1(\boldsymbol{s}), \ldots, \beta_3(\boldsymbol{s})$ were generated by GRFs, and the fourth coefficient was $\beta_4(\boldsymbol{s}) \equiv 0$. The GRFs for generating $\beta_1(\boldsymbol{s}), \ldots, \beta_3(\boldsymbol{s})$ had mean zero, no nugget variance, and exponential covariance $\mathrm{Cov}\left(\beta_j(\boldsymbol{s}_i), \beta_j(\boldsymbol{s}_{i'})\right) = \sigma_j^2 \exp\left(-\tau_\beta^{-1}\delta_{ii'}\right)$ where $\tau_\beta = 1$ is the range parameter. The scale of the coefficient surface $\beta_j(\boldsymbol{s})$ was set via the variance σ_j^2, and the values used in the simulations were $\boldsymbol{\sigma}^2 = (10, 1, 0.1)$. These values were chosen so that the the covariates $X_1, \ldots, X_4$ would have progressively less influence on the simulated response. The coefficient values $\beta_1(\boldsymbol{s}), \ldots, \beta_3(\boldsymbol{s})$ generated in this way are plotted in Figure 1.

Two parameters were varied to produce six simulation settings. Data were simulated with low ($\rho = 0$), medium ($\rho = 0.5$), or high ($\rho = 0.9$) correlation between the covariates, and with low ($\sigma_\varepsilon = 0.5$) or high ($\sigma_\varepsilon = 1$) variance for the random error term. Each setting was used to generate one data set consisting of 400 observations. For each data set, three estimates were made of the coefficients under three different sample sizes n: the full 400

Simulation settings			MISE $\hat{\beta}_1$		MISE $\hat{\beta}_2$		MISE $\hat{\beta}_3$		MISE $\hat{\beta}_4$	
n	ρ	σ_ε	LAGR	VCR	LAGR	VCR	LAGR	VCR	LAGR	VCR
100	0	0.5	**1.43**	1.46	**0.41**	0.43	**0.04**	0.10	**0.05**	0.09
		1.0	**1.41**	1.47	**0.44**	0.55	**0.04**	0.17	**0.03**	0.13
	0.5	0.5	1.47	**1.42**	**0.43**	0.48	**0.04**	0.16	**0.05**	0.13
		1.0	1.48	**1.43**	**0.49**	0.59	**0.05**	0.24	**0.03**	0.19
	0.9	0.5	**1.91**	2.16	**0.76**	1.33	**0.08**	0.73	**0.00**	0.59
		1.0	**1.95**	2.97	**1.08**	1.78	**0.14**	1.08	**0.06**	0.83
200	0	0.5	**1.32**	1.33	0.18	**0.17**	**0.15**	0.24	**0.04**	0.07
		1.0	1.38	**1.35**	0.22	**0.19**	**0.09**	0.26	**0.05**	0.12
	0.5	0.5	**1.41**	1.48	0.21	**0.19**	**0.29**	0.43	**0.06**	0.12
		1.0	**1.51**	1.57	0.26	**0.25**	**0.27**	0.43	**0.08**	0.19
	0.9	0.5	**1.64**	2.02	0.52	**0.51**	**0.85**	1.80	**0.13**	0.49
		1.0	**1.79**	2.40	**0.65**	0.77	**0.92**	1.80	**0.31**	0.82
400	0	0.5	1.23	**1.23**	0.18	**0.17**	**0.06**	0.08	**0.04**	0.05
		1.0	1.28	**1.27**	0.21	**0.19**	**0.07**	0.09	**0.06**	0.10
	0.5	0.5	**1.25**	1.27	**0.21**	0.22	**0.07**	0.13	**0.04**	0.08
		1.0	**1.26**	1.31	0.27	**0.26**	**0.08**	0.15	**0.07**	0.15
	0.9	0.5	**1.43**	1.47	**0.38**	0.56	**0.16**	0.54	**0.07**	0.38
		1.0	**1.40**	1.49	**0.57**	0.77	**0.20**	0.67	**0.22**	0.71

Table 1: For each setting as a combination of sample size n, cross-covariate correlation ρ, and error standard deviation σ_ε, the mean integrated squared error (MISE) of the coefficient estimates. The MISE of $\hat{\beta}_1, \ldots, \hat{\beta}_4$ from estimation by local adaptive grouped regularization (LAGR) is compared to that from estimation by locally linear regression without selection (VCR). **Highlighting** indicates whether LAGR or VCR produced the smaller MISE for each coefficient surface under each simulation setting.

observations, and subsets generated by sampling 100 or 200 unique observations uniformly from the data set. The coefficients were estimated via LAGR and via a VCR model without variable selection as in Section 3. For both estimation methods, the bandwidth parameter was $h = (3/2)n^{-1/6} - 0.36$ with a nearest neighbors type bandwidth, meaning the kernel bandwidth was adjusted at each location $\boldsymbol{s}_i$ to achieve the ratio $\sum_{i'=1}^{n} w_{ii'}/n = h$.

4.2. Simulation Results

The MISE of the coefficient surface estimates are in Table 1. In general, the coefficients estimated by LAGR were more accurate in terms of MISE than those estimated by VCR. The frequency with which MISE was smaller under LAGR than under VCR for estimating

β_1 was 13 of 18 cases, and for β_2 the frequency was was 10 of 18 cases. The improvement by MISE for LAGR over VCR was greater for covariates with smaller influence, with LAGR producing the smaller MISE for β_3 and β_4 in every case. In no case was the MISE for LAGR more than 11% greater than for VCR, while the MISE for estimating β_4 with $\rho = 0.9$, $\sigma_\varepsilon = 0.5$, and $n = 100$ setting was 140 times greater for VCR than for LAGR. That was the exception, as under the other simulation settings the greatest improvement for LAGR over VCR tended to be a $2 - 3$ times reduction in MISE.

With other factors held constant, the MISE for estimating the coefficients tended to be smaller for less influential covariates and under larger sample sizes. On the other hand, the MISE tended to increase with high error variance or increasing correlation between covariates. There was an exception: the MISE for estimating β_3 in simulations where $n = 200$ was surprisingly high under both LAGR and VCR. In terms of MISE, the improvement from estimation by LAGR over VCR was greater for settings with smaller sample sizes, higher correlation between covariates, and greater error variance.

The frequencies of exact zeros among the estimates of each coefficient for each simulation setting are in Table 2. With one exception, the frequency of exact zeros in the coefficient estimates increased uniformly as covariates grew less influential. The exception is that in the eighteenth simulation the frequencies of exact zeros among the estimates of β_3 and β_4 were 66% and 60%, respectively. In particular, the estimates $\hat{\beta}_1$ were almost never exactly zero, while the estimates $\hat{\beta}_3$ and $\hat{\beta}_4$ were exactly zero more often than not. Exact zero coefficient estimates were generally more frequent for smaller sample sizes and greater cross-covariate correlation. With no cross-covariate correlation, exact zero coefficient estimates were always more frequent under greater error variance, a trend that continued under most but not all settings with moderate or high cross- covariate correlation.

Simulation settings			Zero frequency			
n	ρ	σ_ε	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
100	0	0.5	0.00	0.30	0.54	0.86
		1.0	0.00	0.47	0.62	0.93
	0.5	0.5	0.00	0.50	0.74	0.92
		1.0	0.00	0.56	0.74	0.89
	0.9	0.5	0.01	0.59	0.79	0.96
		1.0	0.01	0.53	0.72	0.78
200	0	0.5	0.00	0.12	0.48	0.76
		1.0	0.00	0.32	0.68	0.84
	0.5	0.5	0.00	0.16	0.44	0.65
		1.0	0.00	0.29	0.63	0.69
	0.9	0.5	0.00	0.50	0.64	0.71
		1.0	0.00	0.43	0.62	0.63
400	0	0.5	0.00	0.05	0.41	0.57
		1.0	0.00	0.18	0.42	0.61
	0.5	0.5	0.00	0.18	0.48	0.71
		1.0	0.00	0.30	0.57	0.74
	0.9	0.5	0.00	0.59	0.64	0.74
		1.0	0.00	0.60	0.66	0.61

Table 2: For each setting as a combination of sample size n, cross-covariate correlation ρ, and error standard deviation σ_ε, the frequency of exact zeroes in the estimates of $\hat{\beta}_1, \ldots, \hat{\beta}_4$ as estimated by local adaptive grouped regularizaton.

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 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 nearest neighbors type bandwidth was set to $h = 0.26$.

The estimates of the local coefficients are plotted in the first five panels of Figure 2 and are summarized in Table 3. The estimated coefficients of CRIM and LSTAT were everywhere negative or exactly zero, meaning that the crime rate and proportion of "lower-status" individuals were associated with a lower median house price. Meanwhile, the coefficient of RM was everywhere estimated to be positive, so the more rooms in the average house was everywhere associated with a higher median house price. The coefficient of TAX was negative in most census tracts, but was estimated to be exactly zero in 50 tracts, meaning that in those tracts there was no discernable effect of the property tax rate on house prices. 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. The sixth panel of Figure 2 shows which covariates were estimated to have a nonzero coefficient in each tract. There were 471 tracts where all LAGR estimated that all the covariates had a nonzero coefficient, 43 tracts where all covariates except for TAX were estimated to have nonzero coefficients, six tracts where the coefficients of CRIM and TAX were estimated to be zero, and one tract where the coefficients of CRIM, RAD, and LSTAT were estimated to be zero.

An interesting result from the data example is the apparent relationship between the property

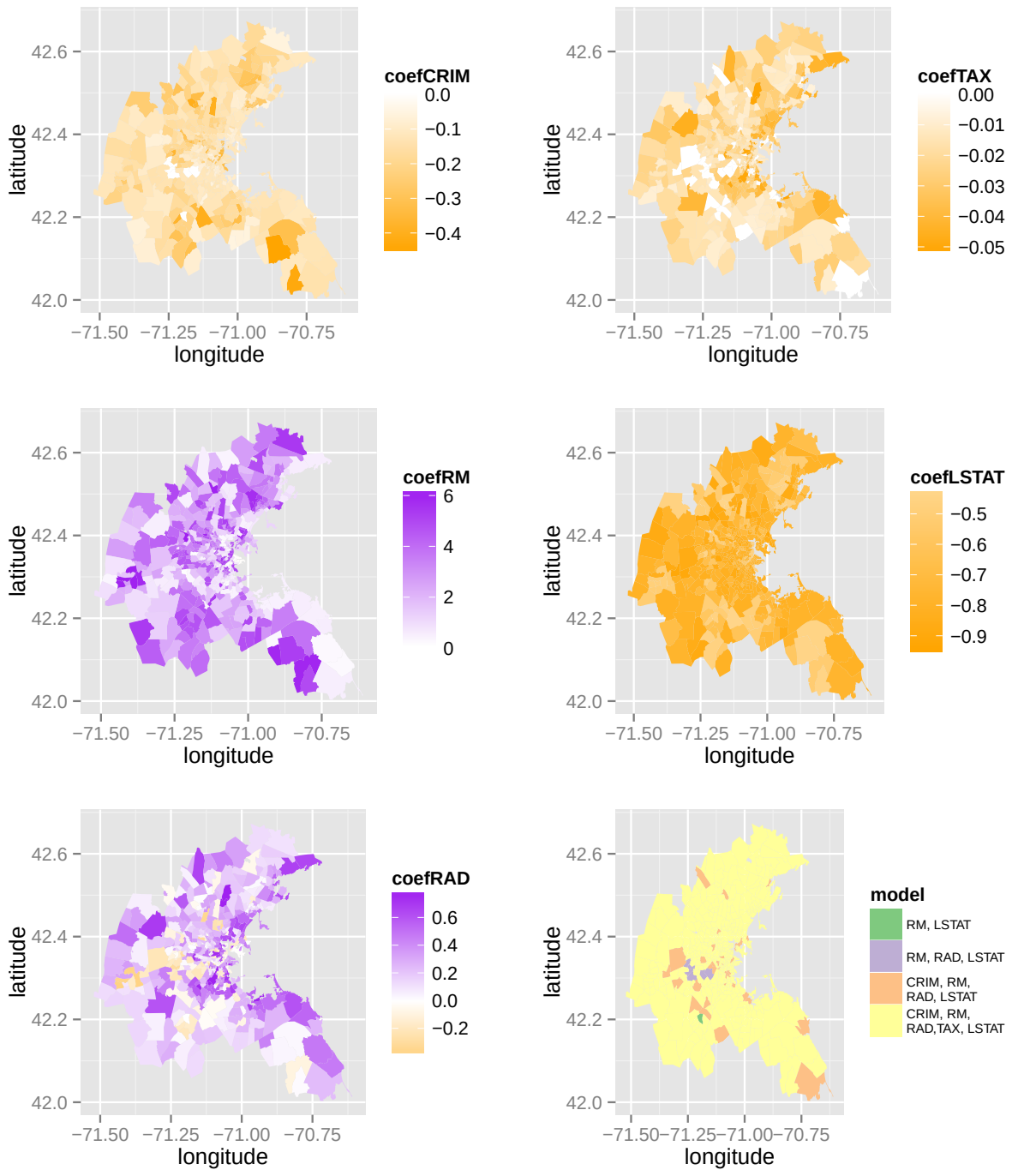

Figure 2: A varying coefficient regression model for the median house price in each census tract in Boston in 1970, estimated by local adaptive grouped regularization. In the left column are the estimated coefficients for covariates CRIM (per-capita crime rate), RM (mean number of rooms per house), and RAD (an index of access to radial roads. In the right column are the estimated coefficients for covariates TAX (property tax per $10,000) and LSTAT (proportion of residents who are "lower status"), and a map indicating which covariates were estimated to have nonzero coefficients in each census tract.

Covariate	Mean	Standard dev.	Zero coef. count
CRIM	-0.15	0.07	7
RM	2.56	1.68	0
RAD	0.21	0.25	1
TAX	-0.02	0.01	50
LSTAT	-0.73	0.13	0

Table 3: The mean, standard deviation, and count 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").

tax rate and the coefficient of the TAX variable. Of the 506 Boston census tracts, there were 50 where the estimated coefficient of the TAX variable was zero. Four of those 50 tracts (8%) were locations where the property tax rate was at least \$666 per \$10000 (which encompasses the two highest property tax rates). In total there were 137 tracts where the property tax rate was at least \$666 per \$10000, which was 27% of all tracts. This suggests that the coefficient of the TAX variable was less likely to be zero in tracts with the highest property tax rates than in other tracts. A rank sum test was used to test the null hypothesis that tracts where the coefficient of the TAX variable was estimated to be zero occurred independently of the property tax rate. The alternative that was considered was that the tracts with an estimated zero TAX coefficient were clustered among the tracts with a lower property tax rate. Of 1000 uniform samples of size 50 from the ranks of the property tax rates, only four had a smaller sum than that observed for the tracts where the TAX coefficient was estimated to be zero. This is compelling evidence that the property tax rate was more likely to have no apparent effect on the median house price in tracts where the property tax was lower.

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 $E\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\} = \mu(\boldsymbol{s}) = b'(\theta(\boldsymbol{s}))$, $\theta(\boldsymbol{s}) = (g \circ b')^{-1}(\eta(\boldsymbol{s}))$, $\eta(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta}(\boldsymbol{s}) = g(\mu(\boldsymbol{s}))$, $\text{Var}\{y(\boldsymbol{s})|\boldsymbol{x}(\boldsymbol{s})\} = b''(\theta(\boldsymbol{s}))$, and link function $g(\cdot)$. Then the probability density is

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

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, $V(\mu(\boldsymbol{s}))$. Then the local coefficients can be estimated by maximizing the local quasi-likelihood

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

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:

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

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

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

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:

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

Further, let $\phi_j(\boldsymbol{s}) = \lambda_n\|\tilde{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})\|^{-\gamma}$, where $\lambda_n > 0$ is a the local tuning parameter applied to all coefficients at location $\boldsymbol{s}$ and $\tilde{\boldsymbol{\zeta}}_{(j)}(\boldsymbol{s})$ is the vector of unpenalized local coeffi-

cients. The following are additional to the definitions and conditions of Section 3.2. Define $\rho(\boldsymbol{s},\boldsymbol{z}) = \left[g_1\left(\mu(\boldsymbol{s},\boldsymbol{z})\right)\right]^2 Var\left\{Y(\boldsymbol{s})|\boldsymbol{X}(\boldsymbol{s}),\boldsymbol{s}\right\}$, where $g_1(\cdot) = g_0'(\cdot)/g'(\cdot)$, and $g_0(\cdot)$ is the canonical link function. So when the canonical link is used, $\rho(\boldsymbol{s},\boldsymbol{z}) = V\left(\mu(\boldsymbol{s},\boldsymbol{z})\right)$. Let $\boldsymbol{\Gamma}\left(\boldsymbol{s}\right) = E\left\{\rho\left(\boldsymbol{s},\boldsymbol{X}(\boldsymbol{s})\right)\boldsymbol{X}(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s})^T|\boldsymbol{s},\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{z}\right\}$ and

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

Assume the following regularity conditions:

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

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

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),*

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

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

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

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 $\sigma^2 \boldsymbol{\Psi}_{(a)}(\boldsymbol{s})^{-1}$ in the variance term of Theorem 1 has been replaced by $\boldsymbol{\Gamma}_{(a)}(\boldsymbol{s})^{-1}$ 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 $H_n(\boldsymbol{u}) = \mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right) - \mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right)$ and $\alpha_n = h^{-1}n^{-1/2}$. Then, we have

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

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

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

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

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

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

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

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

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

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

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

the result of Theorem 1 follows. □

Proof of Theorem 2

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

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

Thus,

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

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

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

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

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

and

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

We showed in the proof of Theorem 1 that

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

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

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

Appendix B. Proofs of Theorems 3–4

Proof of Theorem 3

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

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

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

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

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

Define

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

and

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

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

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

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

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

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

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

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

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

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

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

if $\|\boldsymbol{u}_j\|=0\ \forall j>p_0(\boldsymbol{s})$, and $H_n^*(\boldsymbol{u})=\infty$ otherwise. It follows that $H_n^*(\boldsymbol{u})$ is convex and has a unique minimizer, called $\hat{\boldsymbol{u}}_n$. Let $\hat{\boldsymbol{u}}_{(a)n}$ $\Delta_{(a)}$ and $\Omega_{(a)n}$ be, respectively, the parts of $\boldsymbol{u}_n$, Δ,

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

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

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

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

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

$$n\alpha_n^3 E\left(|\xi_1|^3\right) = O\left(\alpha_n\right) \to 0. \tag{B.2}$$

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

Proof of Theorem 4

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

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

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