Local variable selection in varying-coefficients regression models

By

Wesley R. Brooks

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

University of Wisconsin–Madison

2015

Date of final oral examination: August 3, 2015

The dissertation is approved by the following members of the Final Oral Committee:

Professor Jun Zhu, Department of Statistics and Department of Entomology

Professor Murray Clayton, Department of Statistics and Department of Plant Pathology

Professor Katherine Curtis, Department of Community and Environmental Sociology

Professor Ronald Gangnon, Department of Biostatistics and Medical Informatics and Department of Population Health Sciences

Professor Bret Hanlon, Department of Statistics

Acknowledgments

With gratitude, I would like to take this opportunity to acknowledge the mentorship, guidance, and not least the patience of my advisor, Jun Zhu. Since it would be possible to enumerate here all of her contributions, I will only say that without her, this would not have been possible. Murray Clayton was on the committee for my M.S. defense years ago - an experience that was so much fun it inspired me to push to this point and, soon, beyond. It is a pleasure to have him present for this defense as well. Thanks to Ron Gangnon for quickly identifying and critiquing the crux of my research, back in its early stages. His clarity of thought and expression helped to crystallize my work into its current form. I am happy, as well, to acknowledge the work of Katherine Curtis in defining my research problem. Her consistent enthusiasm has been refreshing and helped to reassure me that this work provides value to the scientific community. I would also like to apologize for being surprised to bump into her on the sidewalk. My head was in the clouds that morning. Finally, I wish to wish to acknowledge and thank Bret Hanlon for his friendship and for reminding me to always laugh at the world.

Many thanks also go to the people at the U.S. Geological Survey's Wisconsin Water Science Center, who were unfailingly warm, supportive, and cheerful. Their influence helped keep the struggles of graduate school in perspective. In particular, Mike Fienen and Steve Corsi were invaluable for providing support, both professional and financial, over a period of years.

Contents

Abstract i

1 Introduction **1**

2 Local Adaptive Grouped Regularization and its Oracle Properties for Varying Coefficient Regression **3**

2.1 Introduction 3

2.2 Varying Coefficient Regression 6

2.2.1 Varying Coefficient Model 6

2.2.2 Coefficient Estimation via Local Likelihood 7

2.3 Local Variable Selection with LAGR 8

2.3.1 LAGR Penalized Local Likelihood 8

2.3.2 Oracle Property 9

2.3.3 Tuning Parameter Selection 12

2.4 Simulation Study 13

2.4.1 Simulation Setup 13

2.4.2 Simulation Results 14

2.5 Data Example 17

2.6 Extension to Generalized Linear Regression 19

2.6.1 Local GLM and Local Quasi-likelihood Estimation 19
2.6.2 LAGR Penalized Local Likelihood and Oracle Properties 23
2.7 Conclusions and Discussion 24
2.8 Proofs of Theorems 1–2 26
2.8.1 Proof of Theorem 1 26
2.8.2 Proof of Theorem 2 27
2.9 Proofs of Theorems 3–4 29
2.9.1 Proof of Theorem 3 29
2.9.2 Proof of Theorem 4 31
2.10 Lemmas 32

3 A Weighted Likelihood Bootstrap for Inference in Varying Coefficients Regression 37

3.1 Introduction 37
3.2 Varying Coefficients Regression 42
3.2.1 Model 42
3.2.2 Local Adaptive Grouped Regularization 45
3.3 Weighted Likelihood Bootstrap 46
3.3.1 Classes of Bootstrap Procedures 46
3.3.2 Weighted Likelihood Bootstrap 49
3.3.3 Asymptotic Property 50
3.4 Simulation Study 51
3.4.1 Simulation Setup 51
3.4.2 Simulation Results 53

3.5 Discussion and Conclusions 59

3.6 Technical Appendix 61

4 An Empirical Bayes Procedure for Inference in Local Polynomial Models 75

4.1 Introduction 75

4.2 Varying Coefficients Regression 78

4.2.1 Model 78

4.2.2 Local Adaptive Grouped Regularization 80

4.2.3 Bandwidth Estimation 81

4.2.4 Weighted Likelihood Bootstrap 82

4.2.4.1 Bootstrap Resamples of Local Coefficient Estimates . . 82

4.2.4.2 Bootstrap Resamples of Bandwidth Parameter 83

4.3 A Hierarchical Model for Local Inference 84

4.3.1 Posterior Hyperprior 85

4.3.2 Conditional Posterior 86

4.3.3 Marginal Posterior 87

4.4 Simulation Study 89

4.4.1 Simulation Setup 89

4.4.2 Bandwidth Estimation 90

4.4.3 Posterior Simulation 90

4.4.4 Simulation Results 90

4.5 Conclusions and Discussion 94

5 Conclusion and Future Work **97**

5.1 Future Work 97

Chapter 1

Introduction

Varying coefficient regression is a flexible technique for modeling data where the coefficients are functions of some effect-modifying parameter, often time or location in a certain domain. While there are a number of methods for variable selection in a varying coefficient regression model, the existing methods are mostly for 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.

After the properties of estimation by the method of LAGR are established, the

natural next step for statistical inference for the model parameters. The distribution of LASSO-type estimators (like LAGR) is a complicated mixture of a point mass at zero with a continuous density conditional on the estimate being nonzero. Because the Gaussian approximation is not workable in this case, it is common to use Monte Carlo methods such as the bootstrap to simulate the distribution of the coefficient estimates. A weighted likelihood bootstrap approach is developed for simulating the distribution of coefficients estimated by LAGR. This approach is new and is apparently the first uniformly-convergent bootstrap for the so-called "paired" nonparametric regression, where the locations, covariates, and response are iid samples from a joint distribution.

The methods proposed in this dissertation are kernel smoothing methods for nonparametric regression. Any kernel smoothing method includes a bandwidth parameter, which we estimate by minimizing the Akaike Information Criterion (AIC). Then estimation and inference proceed conditional on the selected bandwidth. It is preferable to make confidence statements about the marginal coefficient estimates, which are not dependent on the nuisance bandwidth parameter. An empirical Bayes approach to marginal inference for the coefficients is proposed. The approach is to assume a parametric distribution for the bandwidth, such as the gamma distribution. The weighted likelihood bootstrap is used to simulate the distribution, and its parameters are estimated from the bootstraps. This estimated distribution for the bandwidth parameter is interpreted as the posterior hyperprior in a mixture distribution for the coefficient estimates. A method is proposed for simulating the mixture distribution by the weighted likelihood bootstrap, and it is shown that the distribution of the bootstraps tends in the limit to the unknown mixture.

Chapter 2

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

2.1 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 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). For spatial coefficient functions

that are expressed by a wavelet decomposition, the wavelet components with nonzero coefficients may be identified via Bayesian variable selection (Shang, 2011) or the Lasso (Zhang and Clayton, 2011). 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, termed local variable selection, as the literature on how to do so is currently scarce. Such new methodology provides for more flexible variable selection in regression models with coefficients that vary in space.

Specifically, local adaptive grouped regularization (LAGR) is developed here as a novel method of local variable selection at a 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). However, none of these authors addressed local variable selection. In this work, we focus on local variable selection with a two-dimensional effect-modifying parameter and discuss the effect of dimensionality on the results.

For standard linear regression models, the least absolute shrinkage and selection operator (Lasso) is a regularization 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 selects groups rather than individual covariates (Yuan and Lin, 2006; Wang and Leng, 2008). An innovation here is to develop an adaptive group Lasso for local variable selection and coefficient estimation in a locally linear regression model, where each group consists of a single covariate and its interactions with the effect-modifying parameter. Further, we consider both varying coefficient linear regression for Gaussian response and varying coefficient generalized linear regression for responses that are not necessarily Gaussian. We show that the proposed LAGR method 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 2.2. The proposed LAGR technique for varying coefficient linear regression and its oracle properties are presented in Section 2.3. In Section 2.4, the finite-sample properties of LAGR are evaluated in a simulation study,

and in Section 2.5 LAGR is applied to the Boston housing price dataset. In Section 2.6, 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 given in the appendices.

2.2 Varying Coefficient Regression

2.2.1 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 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,

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon_i. \tag{2.1}$$

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 with interactions between the covariates and 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 ma-

trix at location $\boldsymbol{s} = (u, v)^T$ is defined to be $\boldsymbol{Z}(\boldsymbol{s}) = (\boldsymbol{X} \;\; \boldsymbol{L}(\boldsymbol{s})\boldsymbol{X} \;\; \boldsymbol{M}(\boldsymbol{s})\boldsymbol{X})$, where $\boldsymbol{L}(\boldsymbol{s}) = \text{diag}\{s_{i',1} - u\}_{i'=1}^n$ and $\boldsymbol{M}(\boldsymbol{s}) = \text{diag}\{s_{i',2} - v\}_{i'=1}^n$. 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.

2.2.2 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:

$$\ell(\boldsymbol{\zeta}) = -(1/2)\sum_{i=1}^{n}\left(\log\sigma^2 + \sigma^{-2}\left[y_i - \{\boldsymbol{z}(\boldsymbol{s}_i)\}_i\boldsymbol{\zeta}(\boldsymbol{s}_i)\right]^2\right). \tag{2.2}$$

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 log-likelihood (2.2). 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 $\boldsymbol{s} \in \mathcal{D}$:

$$\ell(\boldsymbol{\zeta}(\boldsymbol{s})) = -(1/2)\sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)\left[\log\sigma^2 + \sigma^{-2}\left\{y_i - \boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s})\right\}^2\right] \tag{2.3}$$

where $\boldsymbol{Z}_i = \{\boldsymbol{Z}(\boldsymbol{s})\}_i$, 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 (2.3) 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 (2.3) can be maximized by minimizing a locally weighted least squares:

$$\mathcal{S}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}. \tag{2.4}$$

The minimizer of (2.4) is the locally weighted least squares estimate

$$\tilde{\boldsymbol{\zeta}}(\boldsymbol{s}) = \left\{\boldsymbol{Z}(\boldsymbol{s})^T \boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1} \boldsymbol{Z}(\boldsymbol{s})^T \boldsymbol{W}(\boldsymbol{s})\boldsymbol{Y}. \tag{2.5}$$

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.

2.3 Local Variable Selection with LAGR

2.3.1 LAGR Penalized Local Likelihood

Estimating the local coefficients by (2.5) 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, $\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) = \left(\beta_j(\boldsymbol{s}),\ \ \nabla_u\beta_j(\boldsymbol{s}),\ \ \nabla_v\beta_j(\boldsymbol{s})\right)^T$ for $j = 1, \ldots, p$. The proposed 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 $\boldsymbol{s}$:

$$\mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \mathcal{S}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) + \mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right), \tag{2.6}$$

where $\mathcal{S}\left(\boldsymbol{\zeta}\left(\boldsymbol{s}\right)\right)$ is the locally weighted least squares defined in (2.4), $\mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \sum_{j=1}^{p} \phi_j(\boldsymbol{s}) \|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|$ 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 vector of unpenalized local coefficients for the jth covariate and its interactions with location from (2.5), and $\gamma > 1$.

Minimization of (2.6) is by blockwise coordinate descent, where each block is a covariate group (one raw covariate and its interactions with location). A companion software package for estimating $\boldsymbol{\zeta}(\boldsymbol{s})$ will be made available in `R` (R Core Team, 2014).

2.3.2 Oracle Property

At location $\boldsymbol{s}$, let there be $p_0(\boldsymbol{s}) < p$ 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 $a_n = \max\left\{\phi_j(\boldsymbol{s}), j \leq p_0(\boldsymbol{s})\right\}$ the largest penalty applied to a covariate group whose true coefficient norm is nonzero, and by $b_n = \min\left\{\phi_j(\boldsymbol{s}), j > p_0(\boldsymbol{s})\right\}$ 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 $\kappa_0 = \int_{\mathbb{R}^2} K(\|\boldsymbol{s}\|)d\boldsymbol{s}$, $\kappa_2 = \int_{\mathbb{R}^2}[(1,0)\boldsymbol{s}]^2 K(\|\boldsymbol{s}\|)d\boldsymbol{s} = \int_{\mathbb{R}^2}[(0,1)\boldsymbol{s}]^2 K(\|\boldsymbol{s}\|)d\boldsymbol{s}$, 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 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 $hn^{-1/2}b_n \xrightarrow{p} \infty$.

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 $\boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)$ were assumed to be *iid* in Sun et al. (2014), which is not required here. The existence of $\boldsymbol{\Psi}(\cdot)$ 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). 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) implies 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 2.1 (Asymptotic normality)**.** *Under (C.1)–(C.8),*

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

where $\left\{\nabla^2_{uu}\boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}_{(a)}(\boldsymbol{s})\right\} = (\nabla^2_{uu}\boldsymbol{\beta}_1(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}_1(\boldsymbol{s}), \ldots, \nabla^2_{uu}\boldsymbol{\beta}_{p_0}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}_{p_0}(\boldsymbol{s}))^T$.

Theorem 2.2 (Selection consistency)**.** *Under (C.1)–(C.8), for* $j > p_0(\boldsymbol{s})$,

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

Theorem 2.1 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 2.2, the LAGR estimates of true zero coefficients tend to zero with probability one. Together, local variable selection and local coefficient estimation by LAGR have the oracle property. The technical proofs of Theorems 2.1 and 2.2 are given in Appendix A.

2.3.3 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' C_p, 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

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

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

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. Here, $d = 2$. More general dimensionality is discussed in Section 2.7.

2.4 Simulation Study

2.4.1 Simulation Setup

A simulation study was conducted to assess the performance of the method described in Sections 2.2–2.3. 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 $\boldsymbol{X}$ 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^2_ε. 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 $\text{Cov}\left(\beta_j(\boldsymbol{s}_i), \beta_j(\boldsymbol{s}_{i'})\right) = \sigma^2_j \exp\left(-\tau_\beta^{-1}\delta_{ii'}\right)$ where $\tau_\beta = 1$ is the range parameter. The scale of the coefficient surface $\beta_j(\boldsymbol{s})$ was set via the variance σ^2_j, and the values used in the simulations were $\sigma^2_1 = 10, \sigma^2_2 = 1, \sigma^2_3 = 0.1$. These values were chosen so that the covariates X_1, X_2, X_3 would have progressively less influence on the response. The coefficient values $\beta_1(\boldsymbol{s}), \ldots, \beta_3(\boldsymbol{s})$ generated in this way are plotted in Figure 2.1.

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 five data sets consisting of 400 observations each. For each data set, estimates were made of the coefficients for three different sample sizes N: the full 400 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 2.3. 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 $\sum_{i'=1}^n w_{ii'}/n = h$.

2.4.2 Simulation Results

The mean integrated squared error (MISE) of the coefficient surface estimates are in Table 2.1, where the MISE is calculated as $\text{MISE}(\beta_j) = N^{-1}\sum_{i=1}^N \{\hat{\beta}_j(\boldsymbol{s}_i) - \beta_j(\boldsymbol{s}_i)\}^2$.

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	2.16	**2.15**	**0.35**	0.36	**0.18**	0.24	**0.15**	0.29
		1.0	2.19	**2.14**	**0.38**	0.38	**0.17**	0.28	**0.16**	0.35
	0.5	0.5	**2.36**	2.47	0.40	**0.35**	**0.19**	0.27	**0.26**	0.48
		1.0	**2.25**	2.48	0.44	**0.39**	**0.18**	0.34	**0.24**	0.58
	0.9	0.5	**3.00**	4.90	**0.68**	1.16	**0.35**	1.07	**0.70**	2.22
		1.0	**2.77**	5.18	**0.61**	1.35	**0.38**	1.37	**0.53**	2.71
200	0	0.5	1.75	**1.72**	0.20	**0.18**	**0.09**	0.15	**0.03**	0.10
		1.0	1.79	**1.78**	0.27	**0.21**	**0.11**	0.22	**0.05**	0.13
	0.5	0.5	1.80	**1.75**	0.25	**0.22**	**0.12**	0.23	**0.05**	0.15
		1.0	1.84	**1.83**	0.32	**0.28**	**0.18**	0.34	**0.07**	0.21
	0.9	0.5	**2.19**	2.37	**0.43**	0.76	**0.36**	0.98	**0.24**	0.75
		1.0	**2.25**	2.66	**0.52**	1.10	**0.57**	1.48	**0.32**	1.01
400	0	0.5	1.34	**1.33**	0.18	**0.15**	**0.06**	0.06	**0.02**	0.05
		1.0	1.37	**1.35**	0.22	**0.17**	**0.08**	0.08	**0.02**	0.05
	0.5	0.5	1.37	**1.35**	0.20	**0.18**	**0.07**	0.09	**0.03**	0.08
		1.0	1.40	**1.39**	0.25	**0.21**	**0.09**	0.13	**0.03**	0.09
	0.9	0.5	**1.55**	1.66	**0.41**	0.47	**0.16**	0.36	**0.15**	0.40
		1.0	**1.57**	1.84	**0.44**	0.64	**0.17**	0.54	**0.14**	0.46

Table 2.1: 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, averaged over five independent data sets for each simulation setting. 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.

The results in Table 2.1 are averaged over the five independent data sets for each simulation setting. In general, the coefficients estimated by LAGR were more accurate in terms of MISE than those estimated by VCR. Although the frequency with which MISE was smaller under LAGR than under VCR for estimating β_1 and β_2 was 8 of 18 cases each, 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 27% greater than for VCR. The MISE for estimating β_4 with $\rho = 0.9$, $\sigma_\varepsilon = 1.0$, and $n = 100$ setting was 5 times greater for VCR than for LAGR, and 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. 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. In fact, estimation by LAGR was always more accurate than estimation by VCR under high cross-covariate correlation ($\rho = 0.9$).

The frequencies of exact zeros among the estimates of each coefficient for each simulation setting are in Table 2.2. The frequency of exact zeros in the coefficient estimates generally increased as covariates grew less influential. 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 under smaller sample sizes, greater error variance, and greater cross-covariate correlation. Under high cross-covariate correlation, the frequency of exact zero esti-

mates was roughly equal (and in the neighborhood of 75%) for β_2, β_3, and β_4, which indicates that under high cross-covariate correlation, LAGR tended to include only the most influential covariate.

2.5 Data Example

The proposed LAGR estimation method was applied to estimate the coefficients in a VCR model for 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 2.2 and are summarized in Table 2.3. The estimated coefficients of CRIM and LSTAT were everywhere negative or exactly zero, suggesting 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, indicating no discernable effect of the property tax rate on house prices in those tracts. The coefficient of RAD is positive in some areas and

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	0.40	0.67	0.76
		1.0	0.00	0.57	0.72	0.80
	0.5	0.5	0.00	0.47	0.68	0.75
		1.0	0.01	0.65	0.77	0.79
	0.9	0.5	0.00	0.76	0.79	0.78
		1.0	0.02	0.84	0.79	0.76
200	0	0.5	0.00	0.28	0.68	0.83
		1.0	0.00	0.39	0.66	0.84
	0.5	0.5	0.00	0.36	0.66	0.81
		1.0	0.00	0.48	0.69	0.84
	0.9	0.5	0.03	0.67	0.74	0.83
		1.0	0.04	0.71	0.69	0.84
400	0	0.5	0.00	0.18	0.56	0.74
		1.0	0.00	0.31	0.64	0.82
	0.5	0.5	0.00	0.24	0.62	0.73
		1.0	0.00	0.36	0.69	0.80
	0.9	0.5	0.02	0.61	0.77	0.73
		1.0	0.02	0.68	0.75	0.80

Table 2.2: 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 regularization.

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 bottom right panel of Figure 2.2 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.

2.6 Extension to Generalized Linear Regression

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

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

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, $V\left(\mu(\boldsymbol{s})\right)$.

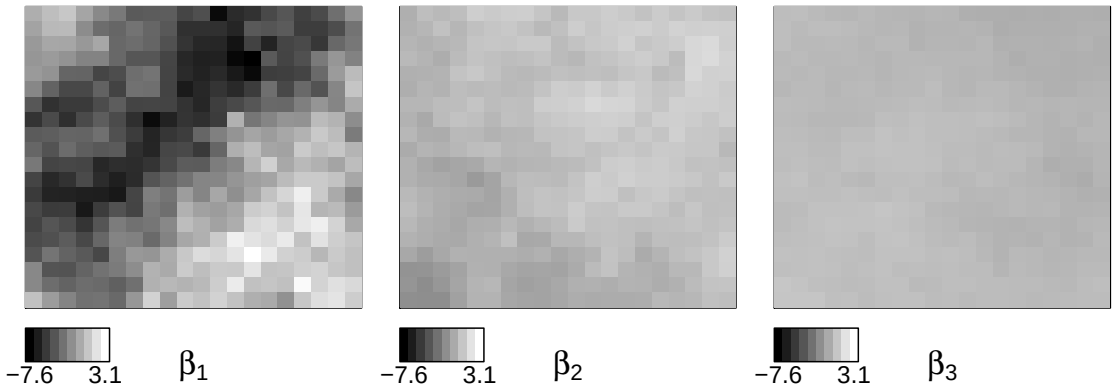

Figure 2.1: Left to right, the values used for coefficients $\beta_1(\boldsymbol{s}), \ldots, \beta_3(\boldsymbol{s})$ in the simulation study.

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 2.3: The mean, standard deviation, and count of zeros among the estimates of the local coefficients in a model for the median house price in 506 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").

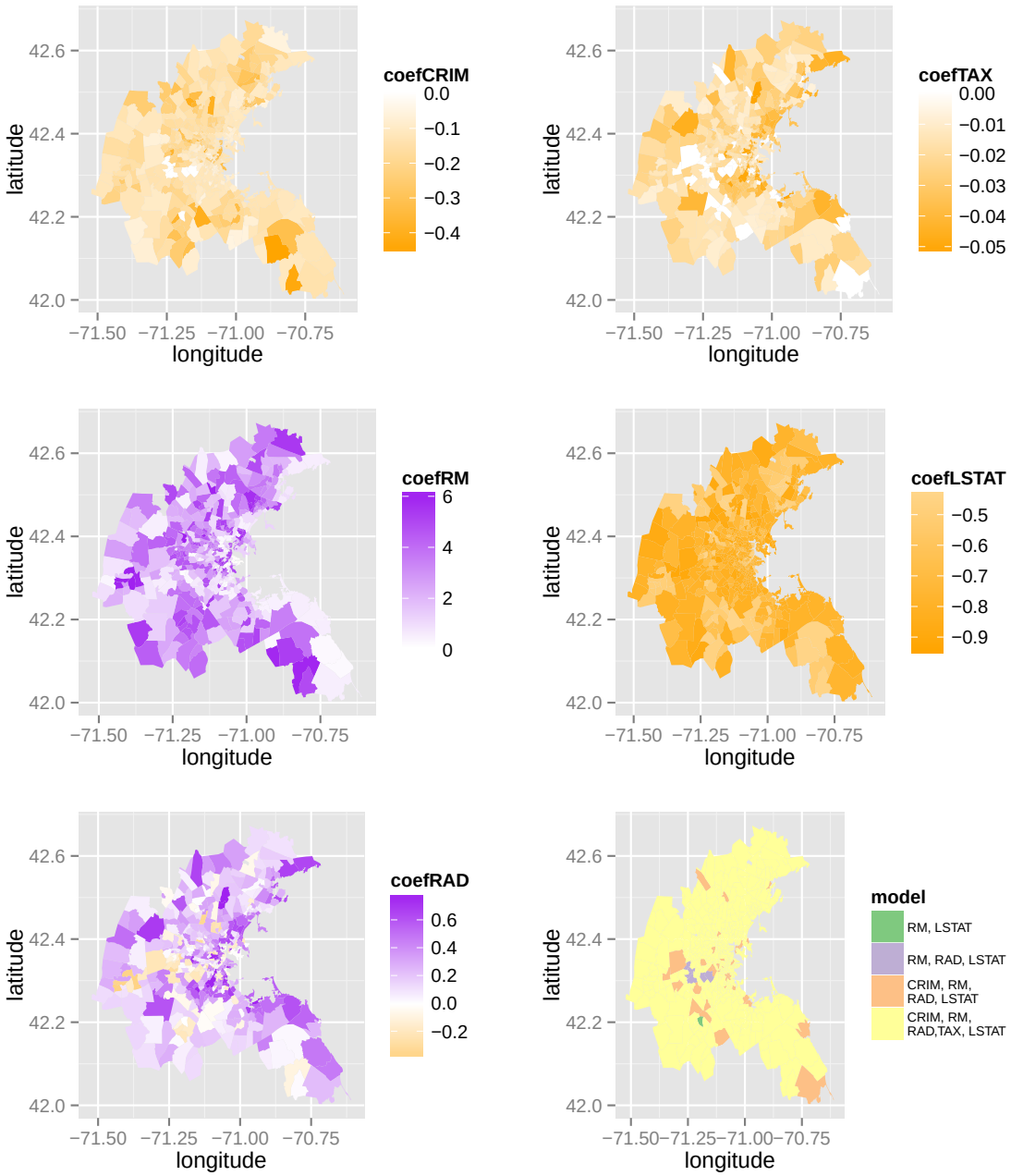

Figure 2.2: A varying coefficient regression model for the median house price in each census tract in Boston in 1970, estimated by local adaptive grouped regularization. In the left column are the estimated coefficients for covariates CRIM (per-capita crime rate), RM (mean number of rooms per house), 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.

Then the local coefficients can be estimated by maximizing the local quasi-likelihood

$$\ell^* \left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \sum_{i=1}^{n} K_h \left(\|\boldsymbol{s} - \boldsymbol{s}_i\|\right) Q \left(g^{-1} \left(\boldsymbol{z}_i^T \boldsymbol{\zeta}(\boldsymbol{s})\right), Y_i\right). \tag{2.7}$$

The local quasi-likelihood (2.7) generalizes the local log-likelihood (2.3) that was used to estimate coefficients in the local linear regression. The local quasi-likelihood (2.7) is concave, 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:

$$(\partial/\partial\boldsymbol{\zeta})\ell^* \left(\hat{\boldsymbol{\zeta}}(\boldsymbol{s})\right) = \sum_{i=1}^{n} K_h \left(\|\boldsymbol{s} - \boldsymbol{s}_i\|\right) \left(y_i - \hat{\mu}(\boldsymbol{s}_i; \boldsymbol{s})\right) \left\{V \left(\hat{\mu}(\boldsymbol{s}_i; \boldsymbol{s})\right)\right\}^{-1} \boldsymbol{z}_i = \mathbf{0}_{3p}, \tag{2.8}$$

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:

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

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

2.6.2 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 negative 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 coefficients. In addition to the definitions and conditions of Section 2.3.2, let

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

and 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 concave and has a unique maximizer.

Theorem 2.1 (Asymptotic normality)**.** *Under (C.1)–(C.10),*

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

Theorem 2.2 (Selection consistency)**.** *Under (C.1)–(C.10), if* $j > p_0(\boldsymbol{s})$*,*

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

By Theorem 2.1, 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 2.1 has been replaced by $\boldsymbol{\Gamma}_{(a)}(\boldsymbol{s})^{-1}$ in Theorem 2.1 because the variance of the response in the GLM case depends on the expectation of the response. Theorem 2.2 gives the same result for the GLM setting as Theorem 2.2 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.

2.7 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 innovation provides a natural and heretofore lacking flexibility to variable selection for varying coefficient regression models, as any covariate may be included in part of and not necessarily the entire domain of interest. This is in contrast to the existing literature

on variable selection for VCR models that focuses on global variable selection, where a covariate is either included in or excluded from the model over its entire domain. 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. Under our refined conditions, we have established the oracle property for the LAGR method.

Here we have 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 may quickly degrade. 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 $\gamma > d/2$.

A possible future direction to take is extension to local regression for spatio-temporal data such that the regression coefficients vary not only in space but also over time. This extension is left to future research.

2.8 Proofs of Theorems 1–2

2.8.1 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

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

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

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

and $\hat{\boldsymbol{u}}_{(b)n} = \boldsymbol{0}$. 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),

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

the result of Theorem 2.1 follows. □

2.8.2 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

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

Thus,

$$
\begin{aligned}
&\left(n^{-1}h^2\right)^{1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} \\
=&\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\left(n^{-1}h^2\right)^{1/2}\left[\boldsymbol{Y}-\boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})-\frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\
+&\left\{n^{-1}\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{(\text{-}p)}(\boldsymbol{s})\right\}\left(nh^2\right)^{1/2}\left[\boldsymbol{\zeta}_{(\text{-}p)}(\boldsymbol{s})+\frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{(\text{-}p)}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}_{(\text{-}p)}(\boldsymbol{s})\right\}-\hat{\boldsymbol{\zeta}}_{(\text{-}p)}(\boldsymbol{s})\right] \\
+&\left\{n^{-1}\boldsymbol{Z}_{(p)}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{(p)}(\boldsymbol{s})\right\}\left(nh^2\right)^{1/2}\left[\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})+\frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})+\nabla^2_{vv}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})\right\}-\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\right].
\end{aligned}
\tag{2.9}
$$

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 2.1 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 (2.9) 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 (2.9) dominates the sum to the right side. Thus, for large enough n, $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\neq\boldsymbol{0}$ cannot maximize $\mathcal{J}\left(\cdot\right)$, and therefore $P\left\{\hat{\boldsymbol{\zeta}}_{(b)}(\boldsymbol{s})=\boldsymbol{0}\right\}\to 1$. □

2.9 Proofs of Theorems 3–4

2.9.1 Proof of Theorem 3

2.1 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, minimizing $H_n'(\boldsymbol{u})$ is equivalent to minimizing $H_n(\boldsymbol{u})$, 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

$$H_n\left(\boldsymbol{u}\right)=-\Omega_n^T\boldsymbol{u}+(1/2)\boldsymbol{u}^T\Delta_n\boldsymbol{u}+\left(\alpha_n^3/6\right)\sum_{i=1}^{n}q_3\left(\boldsymbol{Z}_i^T\tilde{\boldsymbol{\zeta}}_i,Y_i\right)\left[\boldsymbol{Z}_i^T\boldsymbol{u}\right]^3K\left(h^{-1}\|\boldsymbol{s}-\boldsymbol{s}_i\|\right)$$
$$+\sum_{j=1}^{p}\phi_j\left(\boldsymbol{s}\right)\left\{\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})+h^{-1}n^{-1/2}\boldsymbol{u}\|-\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right\}. \tag{2.10}$$

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 (2.10) is $O_p\left(\alpha_n\right)$. The limiting behavior of the last term of (2.10) 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 $n^{-1/2}\phi_j(\boldsymbol{s})\rightarrow n^{-1/2}\lambda_n\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|^{-\gamma}$ 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,

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

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

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

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

$$n\alpha_n^3 E\left(|\xi_1|^3\right) = O\left(\alpha_n\right) \to 0. \tag{2.11}$$

We observe that (2.11) implies that $n\alpha_n^3\left|E\left(\xi_1\right)\right|^3 \to 0$, and since $E\left(\left|\xi_1 - E\xi_1\right|^3\right) < E\left\{\left(|\xi_1| + |E\xi_1|\right)^3\right\} \to 0$, the Lyapunov condition is satisfied. Thus, Ω_n asymptotically follows a Gaussian distribution and the result follows from the quadratic approximation lemma. □

2.9.2 Proof of Theorem 4

2.2

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

$$\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} = \sum_{i=1}^{n} q_1\Big(\boldsymbol{Z}_i^T\hat{\boldsymbol{\zeta}}(\boldsymbol{s}), Y_i\Big)\,\boldsymbol{Z}_{i(p)}K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) \tag{2.12}$$

From Lemma 2, the right hand side of (2.12) is $O_p\left(1\right)$, 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 $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s}) \neq \mathbf{0}$, 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 (2.12) dominates the sum to the right side. Thus, for large enough n, $\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s}) \neq \mathbf{0}$ cannot maximize $\mathcal{J}\left(\cdot\right)$, and therefore $P\left\{\hat{\boldsymbol{\zeta}}_{(b)}(\boldsymbol{s}) = \mathbf{0}\right\} \to 1$. □

2.10 Lemmas

Lemma 1.

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

and

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

$$= \Lambda + o\left(1\right)$$

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

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

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

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

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

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

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

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

And by the definitions of $q_1(\cdot)$ and $q_2(\cdot)$, we have that

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

Now the expectation of Ω_n is

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

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

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

And, using the symmetry of the kernel function,

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

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

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

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

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

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

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

By the symmetry of the kernel function,

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

Thus,

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

□

Lemma 2.

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

$$= -\Delta + o\left(1\right)$$

and

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

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

$$q_2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}), \mu(\boldsymbol{s}_i,\boldsymbol{z}_i)\right) = q_2\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i), \mu(\boldsymbol{s}_i,\boldsymbol{z}_i)\right) + q_3\left(\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i), \mu(\boldsymbol{s}_i,\boldsymbol{z}_i)\right)\left\{\boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}) - \boldsymbol{z}_i^T\boldsymbol{\zeta}(\boldsymbol{s}_i)\right\}$$

$$= -\rho(\boldsymbol{s}_i,\boldsymbol{z}_i) + o\left(1\right).$$

And by the same arguments as before

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

$$E\left(\delta_i|\boldsymbol{s}_i\right) = -\alpha_n^2\begin{pmatrix}1\\ h^{-1}(s_{i,1}-s_1)\\ h^{-1}(s_{i,2}-s_2)\end{pmatrix}\begin{pmatrix}1\\ h^{-1}(s_{i,1}-s_1)\\ h^{-1}(s_{i,2}-s_2)\end{pmatrix}^T\otimes\boldsymbol{\Gamma}(\boldsymbol{s}_i)K\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)$$

$$E\left(\delta_i\right) = -nf\left(\boldsymbol{s}\right)\text{diag}\left\{\kappa_0,\kappa_2,\kappa_2\right\}\otimes\boldsymbol{\Gamma}\left(\boldsymbol{s}\right) + o\left(n^{-1}\right)$$

Thus,

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

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

$$E\left(\delta_i^2|\boldsymbol{Z}_i=\boldsymbol{z}_i,\boldsymbol{s}_i\right) =$$

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

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

Chapter 3

A Weighted Likelihood Bootstrap for Inference in Varying Coefficients Regression

3.1 Introduction

The previous chapter proposed the method of local adaptive grouped regularization (LAGR) for estimating the local coefficients in a spatially varying coefficients regression (VCR) model. The key benefit of LAGR over existing methods for estimation in a VCR model is its ability to estimate the local value of the coefficient functions and to simultaneously estimate which of the model's coefficients are nonzero on only part of the spatial domain, which was called local variable selection. In this chapter, a method is developed for simulating the distribution of the LAGR estimate by a weighted likelihood bootstrap (WLB).

By assumption, coefficients in a VCR model are smooth functions of the location, denoted $\boldsymbol{\beta}(\cdot)$. The method of LAGR employs local polynomial modeling to estimate

the local coefficients pointwise at a particular location. As described in chapter 2, estimation of the values of the coefficient functions at $\boldsymbol{s}$ proceeds by approximating $\boldsymbol{\beta}(\cdot)$ in a neighborhood of $\boldsymbol{s}$ by a first-order Taylor expansion. Then the local coefficient estimate is a weighted mean of the data where the weights decline with increasing distance from $\boldsymbol{s}$ according to a kernel function.

The method of LAGR selects covariates for local regression models and estimates their coefficients via a grouped L_1 penalization scheme akin to the adaptive group LASSO. This type of penalized regression originated with the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). A later refinement, the adaptive LASSO, was shown to have the oracle properties; it asymptotically selects exactly the correct covariates and estimates them as accurately as if their identities were known in advance (Zou, 2006). The group LASSO and adaptive group LASSO are analogous methods that work on predefined groups of covariates, simultaneously selecting all covariates in a group for inclusion in or exclusion from the model (Yuan and Lin, 2006; Wang and Leng, 2008). It was shown in chapter 2 that LAGR possesses the oracle properties.

By the method of LAGR, we are able to estimate the value of the local coefficients, but to proceed with statistical inference within the VCR model it is necessary to derive the distribution of the estimator. LAGR sits at the intersection of local polynomial regression and LASSO-type estimation, both of which are fields where the distribution of estimators is the subject of recent scholarship. In the setting of VCR where the coefficient functions are estimated using the local polynomial model, asymptotic pointwise confidence intervals can be estimated from the asymptotic normality of the local coefficient estimates (Cai et al., 2000). However, this approach relies on

the asymptotic normality of the coefficient estimates and does not account for the L_1 regularization step of LAGR. As a result, the estimated asymptotic distribution is not necessarily the same as the distribution of realized coefficient estimates.

In the paper that proposed the LASSO, Tibshirani (1996) proposed a variance estimate that treats the coefficients estimated as zero as having zero variance. Later, Osborne et al. (2000) developed method that estimates a positive variance for all coefficient estimates by approximating the LASSO as a linear estimator, which is problematic because the nondifferentiability of estimates at zero is the property that enables the LASSO to simultaneously select covariates and estimate their coefficients. A sandwich estimator based on local quadratic approximation, introduced in Fan and Li (2001), can be applied to the variance of nonzero LASSO estimates and was shown in Fan and Peng (2004) to be consistent for the variance of the nonzero coefficients.

An alternative to calculating the distribution of complex or intractable estimators is to simulate the distribution instead. The bootstrap is a method that repeatedly resamples with replacement from an empirical distribution defined by the data (as in the nonparametric bootstrap) or from a parametric distribution estimated from the data (the parametric bootstrap) in order to simulate the distribution of an estimator (Efron, 1979). In particular, the residual bootstrap resamples the residuals of a regression model and adds the resamples back to the fitted values at each iteration (Freedman, 1981). A residual bootstrap approach for simulating the distribution of the coefficient estimates was proposed by Knight and Fu (2000), critiqued by Chatterjee and Lahiri (2010), and modified to demonstrate consistency in Chatterjee and Lahiri (2011), which also showed that the unmodified residual bootstrap is consistent for the distribution of regression coefficients estimated by the adaptive LASSO. The

results of Chatterjee and Lahiri (2011) are shown to apply in the setting of linear regression with constant coefficients and homoskedastic errors, conditions which may be violated in a VCR model.

The residual bootstrap works by resampling from the empirical distribution of residuals from fitting a regression model, then adding the resampled residuals back to the fitted values. This method tends to work when the random error represented by the residuals is homoskedastic and is the only random component in the data (Freedman, 1981). For data where the random error is heteroskedastic, the wild bootstrap is a kind of residual bootstrap where the distribution of each residual is simulated using only its own value, rather than the empirical distribution of the whole set of residuals (Liu, 1988). The wild bootstrap was quickly adapted for application to nonparametric regression (Härdle and Mammen, 1991), and is a standard simulation method for local polynomial regression (Härdle, 1990; Cao-Abad, 1991; Cao-Abad and Gonzàlez-Manteiga, 1993). Data on a spatial domain, though, are often considered to arise at locations that are random according to some density. The appropriate bootstrap procedure for this kind of data is the pairs bootstrap, which Freedman (1981) calls the "correlation model". In this procedure, resampling is not of the residuals but of whole observations from the data, e.g. $(y_i, \boldsymbol{x}_i, \boldsymbol{s}_i)$. Applications of the pairs bootstrap to LASSO-type estimation have focused on evaluating model selection consistency (Bach, 2008; Hall et al., 2009).

Typically, resampling by the pairs bootstrap entails resampling each observation an integer number of times, resulting in a multinomial random vector that describes how often each observation was resampled for a given bootstrap (Freedman, 1981; Mammen, 1993). The so-called Bayesian bootstrap (BB) was introduced by Rubin

(1981), by contrast, generates this vector of resample counts by drawing from a Dirichlet distribution and uses the resample counts as weights in the log-likelihood. Rubin (1981) proposed that the parameters of the Dirichlet random vector could be interpreted as reflecting prior belief, but the BB method is Bayesian in interpretation, not in execution. Whereas Rubin (1981) proposed the method for estimating moments in a nonparametric model for some data, it was modified to estimate the distribution of parameters in a parametric model and renamed to the weighted likelihood bootstrap (WLB) by Newton and Raftery (1994). We adopt and modify this method for estimating the distribution of local coefficient estimates in a nonparametric VCR model. As in Rubin (1981) and Newton and Raftery (1994), the method works by drawing random bootstrap resampling counts to be used as weights in a log-likelihood. Here, though, the weights are sums of independent Dirichlet random vectors, which causes the resampling counts to be clustered around one, and it is this property which allows us to prove a uniform convergence result for nonparametric regression.

For showing the asymptotic properties of estimation by LAGR in the previous chapter, the expectation and variance of some components of the normal equations for local polynomial regression were worked out and then the components were combined through standard results like Slutsky's theorem. In that setting, the components of the normal equations were kernel-weighted sums of independent observations, but in the WLB setting these are sums of the components of Dirichlet random vectors, and so not independent. The moments for sums of this kind of sum were worked out by Provost and Ho Cheong (2000), based on Imhof (1961) for the sum of chi-squared random variables and the representation of a Dirichlet random vector as a vector of independent Gamma random variables, normalized to have unit sum. Such moment

calculations are used extensively to prove our theoretical results here.

The coefficient functions in a VCR model may realize to different values at each location on the domain. As a result, the data observed at distinct locations, however close together, arise from distinct models. The models for nearby locations are related because the parts that vary with location (such as the coefficient values, the variance of random error, and the spatial covariance of the covariates) are varying smoothly. In the previous chapter, proofs of the convergence of the local coefficients estimated by LAGR utilize uniform convergence results of Sun et al. (2014) to the bound order of the absolute difference between a sum of independent but nonidentically distributed random variables and its expectation. Because the WLB has a fixed quantity of weight to distribute among the observations, the elements of the sum are no longer independent in this setting. A new result is proven, following the framework for uniform convergence in local polynomial models that was used in Hansen (2008), which establishes uniform convergence for the sum of dependent and nonidentically distributed random variables of the kind encountered in the WLB.

This chapter concludes with a simulation study that demonstrates an application of the WLB for simulating the distribution of local coefficients in a VCR model as estimated by LAGR.

3.2 Varying Coefficients Regression

3.2.1 Model

Suppose that data are observed at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ in the domain $\mathcal{D}$, a subset of $\mathbb{R}^2$. For the ith observation we have data vector $\{\boldsymbol{s}_i, Y_i, \boldsymbol{X}_i\}$, consisting of location

$\boldsymbol{s}_i$, the univariate response $Y_i = Y(\boldsymbol{s}_i)$ and the p-vector of covariates $\boldsymbol{X}_i = \boldsymbol{X}(\boldsymbol{s}_i)$, which arise from a random field on $\mathcal{D}$ with covariance $\boldsymbol{\Psi}(\boldsymbol{s}) = E\left\{\boldsymbol{X}^T(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s})|\boldsymbol{s}\right\}$ that is continuous with respect to $\boldsymbol{s}$. The response is related to the covariates through a spatially varying coefficient regression (VCR) model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon_i, \tag{3.1}$$

where $\boldsymbol{\beta}(\boldsymbol{s}_i) = (\beta_1(\boldsymbol{s}_i), \ldots, \beta_p(\boldsymbol{s}_i))^T$ is the p-vector of spatially varying coefficients and $\varepsilon_1, \ldots, \varepsilon_n$ are independent random error terms with mean zero and variance σ_i^2 for $i = 1, \ldots, n$. The coefficients in the spatial VCR model (3.1) are functions of the location parameter $\boldsymbol{s}$ and are assumed to all have continuous second derivatives on the domain $\mathcal{D}$.

Since the vector-valued coefficient function $\boldsymbol{\beta}(\cdot)$ varies smoothly over space, nearby data provide some information about the model at $\boldsymbol{s}$, which is the foundational insight of local regression. The notion of a "nearby" observation is formalized in the kernel function $K_h(\cdot) = h^{-2}K(\cdot/h)$, which assigns a weight to each observation in the data set based on its distance from the location of estimation, $\boldsymbol{s}$, relative to the bandwidth h. Throughout this chapter, we use the Epanechnikov kernel: $K(x) = (3/4)(1 - |x|^2)$ if $|x| < 1$ and $K(x) = 0$ otherwise. Then the local regression estimates of the coefficients at $\boldsymbol{s}$ are found by maximizing the weighted local likelihood

$$\ell_{h,0}(\boldsymbol{s}) = \sum_{i=1}^{n} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) \left\{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}(\boldsymbol{s})\right\}^2.$$

As a kernel smoothing method, local regression is biased for estimating the local coefficient values at $\boldsymbol{s}$, especially if $\boldsymbol{s}$ is near the boundary of the domain (Fan and Gijbels, 1996). In order to reduce the bias, a local polynomial expansion of

the coefficient functions is used. Specifically, we consider the case of a locally linear expansion, which approximates the coefficients as linear functions in a neighborhood of the estimation location $\boldsymbol{s}$.

Let $\boldsymbol{\nabla}\beta_j(\boldsymbol{s}) = \{\nabla_u\beta_j(\boldsymbol{s}), \nabla_v\beta_j(\boldsymbol{s})\}$ denote the vector of directional derivatives of $\beta_j(\cdot)$ at $\boldsymbol{s}$ along the cardinal directions of the location parameter. Then the coefficients at location $\boldsymbol{s}_i$ in a neighborhood of $\boldsymbol{s}$ are approximated by the first order Taylor expansion:

$$\beta_j(\boldsymbol{s}_i) = \beta_j(\boldsymbol{s}) + (\boldsymbol{s}_i - \boldsymbol{s})^T\nabla\beta_j(\boldsymbol{s}) + O(\|\boldsymbol{s}_i - \boldsymbol{s}\|^2) \text{ for } j = 1, \ldots, p,$$

where $\boldsymbol{s} \in \mathcal{D}$ is close to $\boldsymbol{s}_i$, and $\|\cdot\|$ denotes the Euclidean distance. Because the Epanechnikov kernel gives zero weight to observations farther than one bandwidth from the estimation location, the remainder term in the Taylor expansion is $O(h^2)$ for any observation given nonzero weight.

For convenience, denote the value of the jth coefficient and its directional derivatives at $\boldsymbol{s}$ by $\boldsymbol{\zeta}_j(\boldsymbol{s}) = (\beta_j(\boldsymbol{s}), \nabla_u\beta_j(\boldsymbol{s}), \nabla_v\beta_j(\boldsymbol{s}))$. Similarly, the matrix of covariates and their interactions with the two dimensions of the location parameter is denoted $\boldsymbol{Z}_{\boldsymbol{s}} = \{\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{L}_{\boldsymbol{s}}\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{X}(\boldsymbol{s})\}$ where $\boldsymbol{L}_{\boldsymbol{s}}$ and $\boldsymbol{M}_{\boldsymbol{s}}$ are diagonal $n \times n$ matrices with $\{\boldsymbol{L}_{\boldsymbol{s}}\}_{ii} = s_{i1} - s_1$ and $\{\boldsymbol{M}_{\boldsymbol{s}}\}_{ii} = s_{i2} - s_2$.

Now the value of the coefficient function at $\boldsymbol{s}$ is estimated by maximizing the weighted local log-likelihood

$$\ell_{h,1}(\boldsymbol{s}) = \sum_{i=1}^{n} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|)\left[y_i - \{\boldsymbol{z}_s\}_i^T\boldsymbol{\zeta}(\boldsymbol{s})\right]^2, \tag{3.2}$$

which is maximized by the method of weighted least squares

$$\bar{\boldsymbol{\zeta}}(\boldsymbol{s}) = \left(\boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{Z}_{\boldsymbol{s}}\right)^{-1} \boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{y}, \tag{3.3}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{K}_{h,\boldsymbol{s}}$ is the diagonal matrix with $\{\boldsymbol{K}_{h,\boldsymbol{s}}\}_{ii} = K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|)$.

3.2.2 Local Adaptive Grouped Regularization

At $\boldsymbol{s} \in \mathcal{D}$, assume that coefficients $1, \ldots, p_0$ are nonzero and that the remaining coefficients are exactly zero, where $1 \leq p_0 \leq p$. Denote the collection of nonzero local coefficients as $\boldsymbol{\beta}_{(a)} = (\beta_1(\boldsymbol{s}), \ldots, \beta_{p_0}(\boldsymbol{s}))^T$, and that of the zero coefficients as $\boldsymbol{\beta}_{(b)}(\boldsymbol{s}) = (\beta_{p_0+1}(\boldsymbol{s}), \ldots, \beta_p(\boldsymbol{s}))^T$.

The method of LAGR simultaneously estimates the local coefficients and estimates which are equal to zero. Estimation is via penalized local likelihood, adding an adaptive group LASSO penalty to the local likelihood estimation of (3.2). The penalized local likelihood is

$$\ell_h(\boldsymbol{s}) = \sum_{i=1}^{n} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) \left[y_i - \boldsymbol{x}_i^T \{\boldsymbol{\beta}(\boldsymbol{s}) + (\boldsymbol{s}_i - \boldsymbol{s})^T \nabla \boldsymbol{\beta}(\boldsymbol{s})\}\right]^2 + \sum_{j=1}^{p} \phi_j(\boldsymbol{s}) \left\|\boldsymbol{\zeta}_j(\boldsymbol{s})\right\|, \tag{3.4}$$

where $\phi_j(\boldsymbol{s}) = \lambda_n \left\|\bar{\boldsymbol{\zeta}}_j(\boldsymbol{s})\right\|^{-\gamma}$, $\bar{\boldsymbol{\zeta}}_j(\boldsymbol{s})$ is the weighted least squares estimate of $\boldsymbol{\zeta}(\boldsymbol{s})$ defined in (3.3), λ_n is a LASSO tuning parameter, and $\gamma > 1$. In chapter 1 it was shown that, under suitable regularity conditions, the method of LAGR attains a central limit theorem and consistent selection of only the nonzero coefficients:

$$\{nf(\boldsymbol{s})h^d\}^{1/2}\left\{\hat{\boldsymbol{\beta}}_{(a)}(\boldsymbol{s})-\boldsymbol{\beta}_{(a)}(\boldsymbol{s})-(2\kappa_0)^{-1}\kappa_2 h^d\sum_{j=1}^{d}\nabla^2_{jj}\boldsymbol{\beta}_{(a)}(\boldsymbol{s})\right\}\xrightarrow{d}N\left(\mathbf{0},\kappa_0^{-2}\nu_0\sigma^2(\boldsymbol{s})\boldsymbol{\Psi}^{-1}_{(a)}(\boldsymbol{s})\right) \tag{3.5}$$

$$\text{and}\ \ P\left\{\hat{\boldsymbol{\zeta}}_{(b)}=\mathbf{0}\right\}\to 1, \tag{3.6}$$

where $\kappa_0 = \int_{\mathbb{R}^d} K(\|\boldsymbol{s}\|)\mathrm{d}\boldsymbol{s}$, $\kappa_2 = \int_{\mathbb{R}^d} s_1^2 K(\|\boldsymbol{s}\|)\mathrm{d}\boldsymbol{s}$, $\nu_0 = \int_{\mathbb{R}^d} K^2(\|\boldsymbol{s}\|)\mathrm{d}\boldsymbol{s}$, and $\boldsymbol{\Psi}_{(a)}(\boldsymbol{s}) = E\left\{\boldsymbol{X}^T_{(a)}(\boldsymbol{s})\boldsymbol{X}_{(a)}(\boldsymbol{s})|\boldsymbol{s}\right\}$. In (3.5) the local coefficient estimates are biased, with the asymptotic bias of $\hat{\beta}_j(\boldsymbol{s})$ given by $(2\kappa_0)^{-1}\kappa_2 h^d\sum_{j=1}^{d}\nabla^2_{jj}\beta_j(\boldsymbol{s})$, which is proportional to the curvature of $\beta_j(\cdot)$ at $\boldsymbol{s}$.

3.3 Weighted Likelihood Bootstrap

The estimation properties of the method of LAGR are appealing. However, statistical inference requires a model for the variability of the coefficient estimates.

The penalized likelihood is not differentiable when the coefficient is zero, a property that allows the LASSO to simultaneously estimate regression coefficients and shrink some to exactly zero. Because of the nondifferentiability, there is not an analytic formula for the distribution of the parameter estimates. Instead, we develop a bootstrap approach to simulate the distribution of the local coefficients and make inferences about the varying coefficients.

3.3.1 Classes of Bootstrap Procedures

Bootstrap procedures for regression models can be divided into two kinds, which are the so-called fixed-design and paired bootstraps. The fixed-design bootstrap procedures assume that the only random component of the regression model is the error

term $\boldsymbol{\varepsilon}$, with the locations $\mathcal{S}$ and covariates $\boldsymbol{X}$ fixed at their observed values. In the case of linear regression with constant coefficients and homoskedastic error variance, the fitted model can be written as $y_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_i$ for $i = 1, \ldots, n$. The distribution of coefficient estimates $\hat{\boldsymbol{\beta}}$ is simulated via the residual bootstrap by repeatedly fitting the model to bootstrapped responses $y_1^*, \ldots, y_n^*$, where $y_i^* = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \varepsilon_i^*$ and the bootstrap errors $\boldsymbol{\varepsilon}^*$ are sampled with replacement from the residuals $\hat{\boldsymbol{\varepsilon}}$. Chatterjee and Lahiri (2011) showed, for a linear regression model with constant coefficients and homoskedastic errors, that when coefficients are estimated by the adaptive LASSO then the coefficient distribution as simulated by the residual bootstrap converges to the true distribution of the coefficient estimates.

Varying coefficients violate the assumptions of Chatterjee and Lahiri (2011), so the validity of their residual bootstrap method in the VCR setting is unknown. It is likely that the residual bootstrap fails to consistently simulate the coefficient distribution in this case because kernel smoothing methods like LAGR are biased for estimating the local coefficients. Bias in the coefficient estimates affects the residuals through the regression model.

The wild bootstrap is a fixed-design bootstrap procedure for nonparametric regression models with potentially heteroskedastic errors. Under the wild bootstrap, the distribution of each error ε_i is estimated individually from the Studentized residual $\hat{\varepsilon}_i/(1 - h_{ii})$ where $h_{ii} = \partial\hat{y}_i/\partial y_i$. The wild bootstrap does not attempt to simulate the error distribution exactly, but to approximate the distribution from a single observation in a way that consistently estimates at least the first and second moments (Liu, 1988). The wild bootstrap is implemented by sampling v_i^* from a two-point distribution with zero mean - usually either the Rademacher distribu-

tion ($v_i^* = +1$ *w.p.* $1/2$, -1 *w.p.* $1/2$) or Mammen's distribution ($v_i^* = (1 - \sqrt{5})/2$ *w.p.* $(\sqrt{5}+1)/(2\sqrt{5})$, $(1+\sqrt{5})/2$ *w.p.* $(\sqrt{5}-1)/(2\sqrt{5})$). Then the resampled error ε_i^* is calculated from v_i^* and the Studentized residual by $\varepsilon_i^* = v_i^* \varepsilon_i / (1 - h_{ii})$.

In the case of a VCR model, however, it is not clear that any fixed-design bootstrap procedure above is appropriate. The reason is that these resample only the residuals, while our definition of the VCR model assumes that the covariates are realizations of a random process and that the locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ are random according to a density $f(\cdot)$. Since the randomness of the observation locations and the covariates are reflected in the local coefficient estimates, they must also be reflected in the resampling used to simulate the distribution of those estimates. Additionally, the LAGR estimates of the local coefficients in a VCR model are biased, and the biased coefficient estimates affect the residuals. Because of the bias, resampling the residuals may not simulate the correct bootstrap distribution.

After the fixed-design bootstrap, the other kind of bootstrap procedure for regression models is the pairs bootstrap, which was introduced as the so-called correlation model of Freedman (1981). In the setting of ordinary linear regression, this procedure works by resampling the response and the covariates together as the pairs $(Y_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$. In a nonparametric regression setting like the VCR model, the locations are included in the "pairs" as well, leading to resampling the triples $(Y_i, \boldsymbol{X}_i, \boldsymbol{s}_i)$ for $i = 1, \ldots, n$.

While this method does resample from the data-generating process, it poses problems when estimating the coefficients at locations that were resampled zero times because then we have an estimation problem at those locations.

3.3.2 Weighted Likelihood Bootstrap

The weighted likelihood bootstrap (WLB) differs from the standard bootstrap in that it generates continuous weights for each observation, rather than resampling each an integer number of times. Thus, while the NPB draws weights for the observations from an n-parameter Multinomial$(1/n, \ldots, 1/n)$ distribution, the WLB draws the weights from an $(n-1)$-parameter Dirichlet$(1, \ldots, 1)$ distribution. We call the bootstrap procedure the weighted likelihood bootstrap because the distribution of the coefficient estimates is simulated by resampling bootstrap weights for the weighted likelihood.

The procedure to simulate the distribution of the local coefficient estimates is as follows. Let $b = 1, \ldots, B$ index the bootstrap resamples, where B is the total number of resamples. Let $\boldsymbol{D}_1^{*b}, \ldots, \boldsymbol{D}_n^{*b} \overset{iid}{\sim} \text{Dirichlet}(\mathbf{1}_n)$ be n independent n-vectors, each sampled from a uniform Dirichlet distribution. Then their sum, $\boldsymbol{W}^{*b} = \sum_{j=1}^{n} \boldsymbol{D}_j^{*b}$, is the n-vector of likelihood weights for the bth resample.

Then the penalized weighted local log-likelihood for the bth resample is

$$\ell_h^{*b}(\boldsymbol{s}) = \sum_{i=1}^{n} w_i^{*b} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) \left[y_i - x_i^T \left\{\beta(\boldsymbol{s}) + (\boldsymbol{s}_i - \boldsymbol{s})^T \nabla\beta(\boldsymbol{s})\right\}\right]^2 + \lambda_n \sum_{j=1}^{p} \phi_j^{*b}(\boldsymbol{s}) \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|,$$

where $\phi_j^{*b}(\boldsymbol{s}) = \lambda_n \|\bar{\boldsymbol{\zeta}}_j^{*b}(\boldsymbol{s})\|^{-\gamma}$ are adaptive weights for the regularization scheme, λ_n is a sequence of LASSO tuning parameters, and

$$\bar{\boldsymbol{\zeta}}_j^{*b}(\boldsymbol{s}) = \left(\boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{W}^* \boldsymbol{Z}_{\boldsymbol{s}}\right)^{-1} \boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{W}^* \boldsymbol{y}.$$

The WLB coefficients $\hat{\boldsymbol{\beta}}^{*b}(\boldsymbol{s})$ are those that maximize $\ell_h^{*b}(\boldsymbol{s})$ for $b = 1, \ldots, B$. We are now prepared for a statement of the main theorem.

3.3.3 Asymptotic Property

Asymptotics work as follows: let $\boldsymbol{G}_i = \{\boldsymbol{s}, X(\boldsymbol{s}), \varepsilon\}_i = \{\boldsymbol{s}_i, X(\boldsymbol{s}_i), \varepsilon_i\}$ be the collection of random elements from the ith observation in the data set. And let $\mathcal{G}_n = \sigma\langle\{\boldsymbol{G}_i : i \leq n\}\rangle$ be the σ-field generated by the data through the nth observation. Define the conditional expectation $E_{\cdot|G}(\cdots) = E(\cdots|\mathcal{G}_n)$, and the conditional probability $P_{\cdot|G}(\cdots) = P(\cdots|\mathcal{G}_n)$. The limits in this section are taken as $n \to \infty$.

Assume the following regularity conditions:

(C1). The kernel function $K(\cdot)$ is bounded, positive, symmetric, and Lipschitz continuous on $\mathbb{R}$, and has a bounded support.

(C2). The coefficient functions $\beta_j(\cdot)$ for $j = 1, \ldots, p$ have continuous second-order partial derivatives at $\boldsymbol{s}$.

(C3).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 $\boldsymbol{s}$.

(C4). $E\left\{|\boldsymbol{X}(\boldsymbol{s})|^3|\boldsymbol{s}\right\}$ and $E\left\{Y(\boldsymbol{s})^4|\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{s}\right\}$ are continuous at a given location $\boldsymbol{s}$.

(C5). The observation locations $\{\boldsymbol{s}_i\}$ are a sequence of 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

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

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

(C6). $h = O\left(n^{-1/6}\right)$.

(C7). $(nh)^{-1}a_n \xrightarrow{p} 0$ and $(nh)^{-1}b_n \xrightarrow{p} \infty$.

Theorem 3.1. *Let* $H_n = P(\boldsymbol{T}_n \leq x)$ *for* $x \in \mathbb{R}$ *where*

$$\boldsymbol{T}_n = \{h^2 n f(\boldsymbol{s})\}^{1/2} \left[\hat{\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{\beta}(\boldsymbol{s}) - (2\kappa_0)^{-1} \kappa_2 h^2 \{ \sum_{k=1}^{2} \nabla_{kk}^2 \boldsymbol{\beta}(\boldsymbol{s}) \} \right],$$

and let $\hat{H}_n$ *be the conditional cumulative distribution function of the WLB estimate*

$$\boldsymbol{T}_n^* = \{h^2 n^2 f(\boldsymbol{s})\}^{1/2} \left[\hat{\boldsymbol{\beta}}^{*b}(\boldsymbol{s}) - \boldsymbol{\beta}(\boldsymbol{s}) - (2\kappa_0)^{-1} \kappa_2 h^2 \{ \sum_{k=1}^{2} \nabla_{kk}^2 \boldsymbol{\beta}(\boldsymbol{s}) \} \right].$$

Under the regularity conditions, as $n \to \infty$, $\hat{H}_n$ *converges to* H_n, *in the sense that*

$$D(\hat{H}_n, H_n) \xrightarrow{p} 0,$$

where $D(\cdot, \cdot)$ *is the Prokhorov distance metric.*

Proof. The theorem is follows from a series of Lemmas in the technical appendix. □

3.4 Simulation Study

3.4.1 Simulation Setup

A simulation illustrates the WLB inference method. The location parameter for the simulation is two-dimensional. Data were simulated at 196 locations on a uniform 14×14 grid covering the domain $[0, 1] \times [0, 1]$. The model for the simulated data is

$$y_i = \beta_0(s_i) + X_{1i}\beta_1(s_i) + X_{2i}\beta_2(s_i) + \varepsilon_i$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(\mathbf{0}, \sigma^2 I_n)$. The coefficient functions $\beta_0(s)$, $\beta_1(s)$, and $\beta_2(s)$ are plotted in Figure 1. The coefficients X_1 and X_2 were simulated as zero mean Gaussian random fields with exponential covariance model having unit variance, range

parameter 0.1, and nugget variance 0.2. Collinearity was induced in the covariates by multiplying the matrix $(X_1\ X_2)$ by the square root of the matrix having ones on the diagonal and ρ for the off-diagonal entries.

Data were simulated under six settings: low ($\rho = 0.1$), medium ($\rho = 0.5$), or high ($\rho = 0.9$) covariance, crossed with low ($\sigma = 0.5$) or high ($\sigma = 1$) error variance. For each setting, 28 independent replications were simulated. For each of these replications, the coefficient functions were estimated by LAGR and the distribution of these coefficient function estimates was estimated by 100 resamples of the WLB.

The WLB distribution of the coefficient estimates was compared to the asymptotic normal approximation of Cai et al. (2000). The asymptotic distribution of the coefficient estimates as written in (3.5) depends on the local density of observation points, $f(\boldsymbol{s})$ and the local second derivatives of the coefficient functions, $\nabla^2_{jj}\beta_k(\boldsymbol{s})$ for $j = 1, \ldots, d$ and $k = 1, \ldots, p$. These quantities are unknown and so cannot be used for estimation. The local second derivatives of the coefficient functions appear in the asymptotic bias but not the variance; we can avoid having to estimate the second derivatives by ignoring the bias. Meanwhile, some simple calculations show that the expected value of the sum of kernel weights at $\boldsymbol{s}$ is

$$E\sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|) = nEK_h(\|\boldsymbol{s} - \boldsymbol{s}_1\|) = n\kappa_0 f(\boldsymbol{s}) + O_p(h^2) \to n\kappa_0 f(\boldsymbol{s}).$$

Combining this result with (3.5), we have that the local coefficient estimates are approximately distributed as

$$\hat{\boldsymbol{\beta}}(\boldsymbol{s}) \overset{d}{\approx} N\left(\boldsymbol{\beta}(\boldsymbol{s}) + (2\kappa_0)^{-1}\kappa_2 h^d \sum_{j=1}^{d} \nabla^2_{jj}\boldsymbol{\beta}(\boldsymbol{s}), \nu_0\hat{\sigma}_i^2\{h\kappa_0 \sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}(\boldsymbol{s}) K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)\}^{-1}\right).$$

This is the asymptotic distribution to which the bootstrap distribution is compared

in this simulation study. No local variable selection was attempted for the asymptotic approximation.

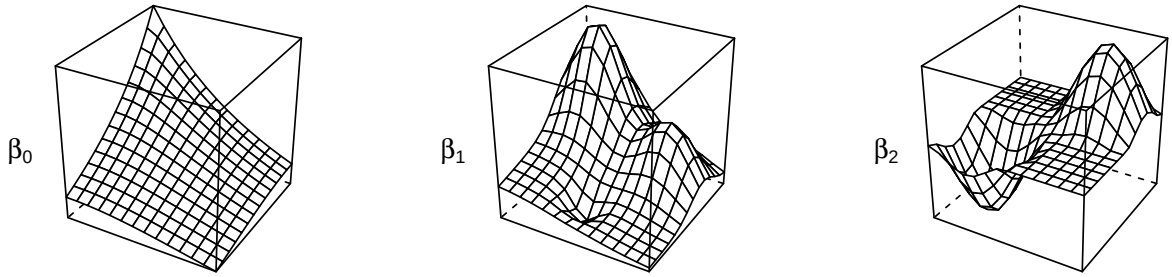

Figure 3.1: The coefficient functions used in the simulation model.

3.4.2 Simulation Results

The mean coverage frequency of nominal 90% confidence intervals (CIs)for the coefficients $\beta_0(\cdot), \beta_1(\cdot)$, and $\beta_2(\cdot)$ are summarized in Tables 3.1 and 3.2, which respectively describe the coverage of WLB CIs and asymptotic CIs. The reported frequencies are the means of coverage frequency for each coefficient over the domain $[0,1] \times [0,1]$. Coverage is reported for the true coefficient functions and for smoothed versions of the truth which represent the estimable coefficient functions.

Coverage of both the true coefficients and the smoothed coefficients by the asymptotic CI was usually greater than by the WLB CI. In general, both achieved mean coverage lower than the nominal 90%. The 90% CIs covered the smoothed coefficients with much greater frequency than they covered the true coefficients.

Coverage of the intercept function $\beta_0(\cdot)$ by both kinds of CI appears to be identical between the true and smoothed versions, probably because the intercept function had minimal curvature. There is a marked difference in the frequency with which nom-

		coverage frequency					
Parameters		true coefficients			smoothed coefficients		
σ	ρ	β_0	β_1	β_2	β_0	β_1	β_2
0.50	0.10	0.77	0.47	0.58	0.77	0.76	0.64
0.50	0.50	0.75	0.48	0.61	0.75	0.72	0.61
0.50	0.90	0.75	0.55	0.70	0.75	0.66	0.54
1.00	0.10	0.82	0.53	0.64	0.82	0.75	0.60
1.00	0.50	0.81	0.51	0.68	0.81	0.68	0.57
1.00	0.90	0.81	0.49	0.64	0.81	0.56	0.41

Table 3.1: Coverage frequencies of nominal 90% weighted likelihood bootstrap confidence intervals (CIs) for the coefficient estimates, averaged over the modeling domain, under the six simulation settings. Columns 3-5 list the frequency that the nominal 90% CI covered the true local value of the coefficient function, while columns 6-8 list the frequency that the nominal 90% CI covered the smoothed local value of the coefficient function.

		coverage frequency					
Parameters		true coefficients			smoothed coefficients		
σ	ρ	β_0	β_1	β_2	β_0	β_1	β_2
0.50	0.10	0.86	0.60	0.62	0.86	0.87	0.86
0.50	0.50	0.85	0.64	0.69	0.85	0.88	0.87
0.50	0.90	0.86	0.78	0.79	0.86	0.87	0.87
1.00	0.10	0.87	0.69	0.69	0.87	0.87	0.87
1.00	0.50	0.88	0.72	0.75	0.88	0.88	0.88
1.00	0.90	0.86	0.80	0.81	0.86	0.86	0.86

Table 3.2: Coverage frequencies of nominal 90% asymptotic confidence intervals (CIs) for the coefficient estimates, averaged over the modeling domain, under the six simulation settings. Columns 3-5 list the frequency that the nominal 90% asymptotic CI covered the true local value of the coefficient function, while columns 6-8 list the frequency that the nominal 90% CI covered the smoothed local value of the coefficient function.

inal 90% confidence intervals for the other coefficient functions cover the smoothed coefficients, compared to the frequency with which they cover the true coefficients. Coverage of the smoothed coefficients is uniformly more frequent and comes close to the nominal 90% coverage (ranging between 0.41 and 0.76 for $\beta_1(\cdot)$ and $\beta_2(\cdot)$). Coverage of the true $\beta_1(\cdot)$ and $\beta_2(\cdot)$ ranged between 0.47 and 0.7.

The reason for the difference appears to be that the local coefficient estimates are biased in the direction of curvature of the coefficient functions. As is apparent from Figure 1, the intercept function β_0 is mostly a gradient from the top left to the bottom right of the domain. The frequency with which the nominal 90% CI covers the true β_0 and the smoothed $\beta_0(\cdot)$ are shown in the second and third panel of Figure 2, respectively. Any difference between the panels is hard to identify. On the other hand, see Figure 3 for the local coverage frequency for the nominal 90% CI for $\beta_1(\cdot)$ and Figure 4 for the local coverage frequency for the nominal 90% CI for $\beta_2(\cdot)$. The nominal 90% CI badly undercovers the true coefficient functions in some areas of the domain, which are colored orange in the figures. These areas are exactly where the true coefficient functions (plotted in the first panels of Figures 2, 3, and 4) are curving. The smoothed versions of these coefficient surfaces are covered at approximately the nominal frequency in the same areas.

It appears from the simulation that the confidence intervals constructed by the WLB are for a smoothed version of the truth, where the smoothing is an artifact of the kernel smoothing we used to estimate the model.

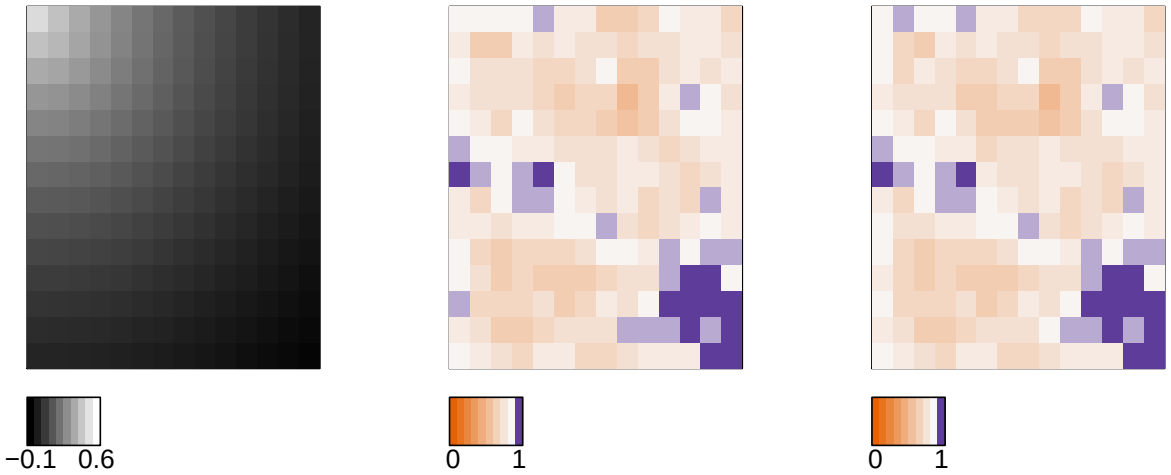

Figure 3.2: Intercept function $\beta_0(\cdot)$ (left panel) and the coverage of nominal 90% confidence intervals (CIs) at 196 locations on the domain $[0,1] \times [0,1]$. The middle panel illustrates the frequency with which the CIs cover the true local intercept, and the right panel illustrates the frequency with which the CIs cover the local value of the smoothed intercept function. Locations of undercoverage are colored orange and locations of overcoverage are colored purple.

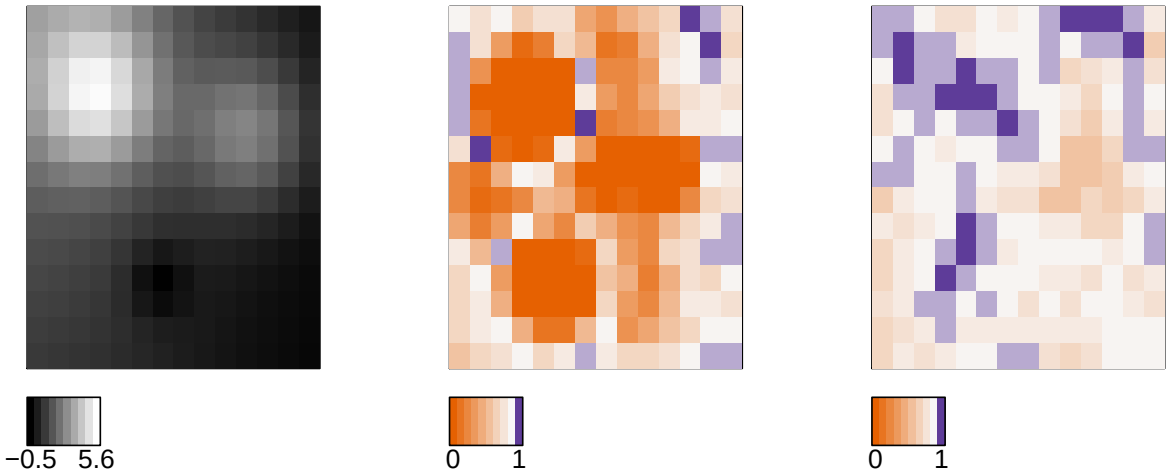

Figure 3.3: Coefficient function $\beta_1(\cdot)$ (left panel) and the coverage of nominal 90% confidence intervals (CIs) at 196 locations on the domain $[0,1] \times [0,1]$. The middle panel illustrates the frequency with which the CIs cover the true local coefficient, and the right panel illustrates the frequency with which the CIs cover the local value of the smoothed coefficient function. Locations of undercoverage are colored orange and locations of overcoverage are colored purple.

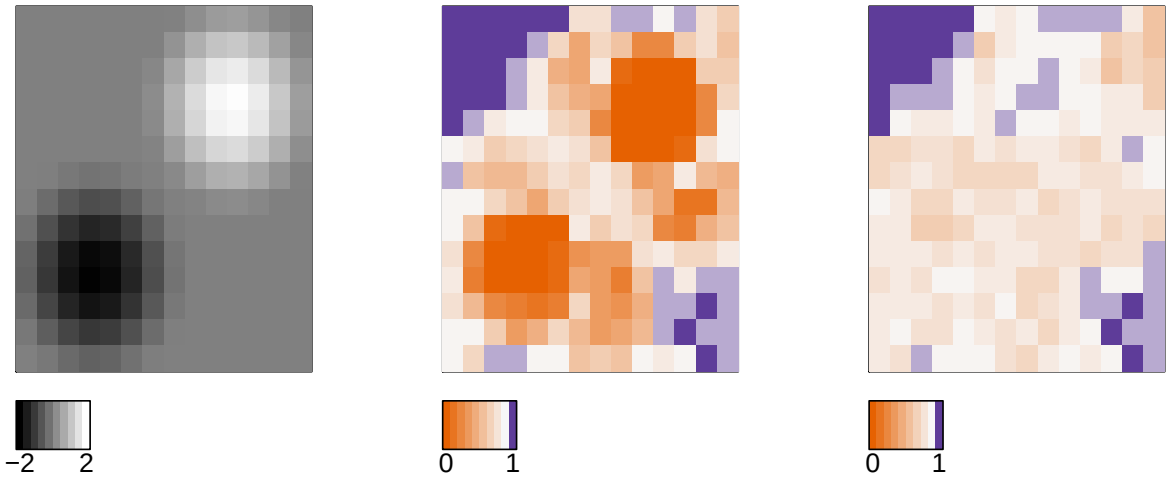

Figure 3.4: Coefficient function $\beta_2(\cdot)$ (left panel) and the coverage of nominal 90% confidence intervals (CIs) at 196 locations on the domain $[0,1] \times [0,1]$. The middle panel illustrates the frequency with which the CIs cover the true local coefficient, and the right panel illustrates the frequency with which the CIs cover the local value of the smoothed coefficient function. Locations of undercoverage are colored orange and locations of overcoverage are colored purple.

3.5 Discussion and Conclusions

Inference for the coefficients in a VCR model can be divided into two categories: either the observation locations are fixed or they are random. The WLB assumes that the observation locations are random, according to the density $f(\boldsymbol{s})$. Under this assumption, the WLB assigns weights to the observations $\{\boldsymbol{s}_i, \boldsymbol{X}_i, Y_i\}$ for $i = 1, \ldots, n$. The WLB resampling scheme thus implicitly simulates the distribution of the locations, the distribution of the covariates $\boldsymbol{X}$, and the distribution of the random errors $\boldsymbol{\varepsilon}$. If the sampling locations are considered fixed, a residual bootstrap method (such as the wild bootstrap or parametric residual bootstrap) may be required.

The rate at which the WLB resamples converge is faster than that of the original LAGR estimate. Specifically, the error of the raw LAGR estimate is $O(h^{-1}n^{-1/2}) = O(n^{-1/3})$, while the error of the WLB coefficient estimates is $O(h^{-1}n^{-1}) = O(n^{-5/6})$. As a result, the WLB estimates must be adjusted in order to be interpreted as simulating the distribution of the coefficients as estimated by LAGR. The adjustment is done by subtracting the mean from the bootstrap estimates, multiplying the difference by $\sqrt{n}$, and then adding the mean back.

Convergence is faster for the WLB because of how the weights are sampled. The sum of n independent Dirichlet($\mathbf{1}_n$) random variables exhibits a central tendency, which means that the weights are clustered near one. This is a consequence of the Central Limit Theorem. On the other hand, the NPB samples from a multinomial distribution with uniform weights, and the BB studied by Rubin (1981) and Newton and Raftery (1994) samples its weights via a single realization of a Dirichlet($\mathbf{1}_n$) distribution, which is a multivariate generalization of the uniform distribution. In both

cases, the mean and variance of the weights tend to one and therefore the mean and variance of the bootstrap resamples tend to the corresponding moments of the target distribution. Meanwhile, the variance of the weights in the WLB tends to $1/n$, which is why the variance of the bootstrap resamples is too small by a factor of n.

The crucial advantage of the WLB is that it converges uniformly on the domain of the location parameter. Because the coefficients vary over the domain, each observation is the product of a unique model that applies only at the location where that observation was made. If the observations had all arisen from the same model, then the familiar law of large numbers and central limit theorem would suffice to show that the coefficient estimates converge to their expectation at a particular rate. Given observations that are generated by a process that varies smoothly over the domain, individual observations can each be approximated via Taylor's expansion. However, that is merely an approximation, and entails an unknown amount of error. Uniform convergence means establishing a supremum for error when approximating the model at $\boldsymbol{s}$ by the weighted mean of nearby approximations, for any $\boldsymbol{s}$. It also means that the supremum goes to zero as $n \to \infty$.

Alternatives like the NPB and the BB are not known to possess the property of uniform convergence, and it is unlikely that they could. Heuristically, convergence is usually established by showing that the probability of encountering any observation that deviates from the expectation by enough to have a noticeable impact on the mean grows more slowly than the sample size. Then any deviations that do occur are eventually washed out in the course of $n \to \infty$. However, when the bootstrap weights are sampled from a uniform distribution (be it discrete as in the NPB, or continuous as in the BB) on $[0, n]$, the limiting behavior can change. Now it is straightforward to show

by the Borel-Cantelli lemma that, for any sample size, repeatedly bootstrapping the weights will inevitable generate bootstraps where one observation carries an arbitrary proportion of the total weight (25%, say), which grows with the sample size. Even as $n \to \infty$, that observation's contribution is not negligible in the mean, which prevents the error supremum from converging to zero.

This heuristic argument also shows why the solution for bootstrapping in a non-parametric regression setting must be for the bootstrap weights to cluster around one. When the weights cluster, the likelihood of any extreme bootstrap weights gets smaller as the sample size increases. The theorem shows that the WLB achieves a balanced condition where the probability of extreme weights declines quickly enough to achieve uniform convergence but not so fast that the bootstrap distribution degenerates to a point mass at the unbootstrapped LAGR estimate. Thus, the convergence of the WLB resamples is orderly, which allows us to interpret those resamples as arising from the approximate distribution of the LAGR estimate.

3.6 Technical Appendix

The following result, due to Provost and Ho Cheong (2000), will be used several times.

Lemma 3.1. *Moments of a linear combination of components of a Dirichlet random vector Let* $\boldsymbol{D} = (D_1, \ldots, D_m)^T$ *be a random vector having an* m*-parameter Dirichlet*$(\mathbf{1}_m)$ *distribution. Further, suppose* $\boldsymbol{a} = (a_1, \ldots, a_m)$ *is a fixed* m*-vector and denote the linear combination of the elements of* $\boldsymbol{a}$ *and* $\boldsymbol{D}$ *by* $S = \sum_{i=1}^{m} a_i D_i$*. Then the* k*th integer moment of* S *is*

$$E(S^k) = \sum_{k_1=0}^{K} \sum_{k_2=0}^{k-k_1} \cdots \sum_{k_{m-1}=0}^{k-k_1-\cdots-k_{m-2}} \frac{k!\Gamma(m)}{\Gamma(k+m)} \prod_{j=1}^{m} \delta_j^{k_j},$$

with $k_m = k - \sum_{i=1}^{m-1} k_i$.

We begin by proving that the kernel-weighted sum of Dirichlet-weighted observations converges uniformly on the domain $\mathcal{D}$. The result differs from Lemma 1 of Sun et al. (2014) in that the Dirichlet weights of the WLB induce dependence in the observations, which violates the assumptions of that Lemma. Our proof of the result repeatedly uses the fact that the Dirichlet weights of the WLB are independent of the observations. The expectation here is with respect to the observations and the Dirichlet weights; for now, the observation locations are considered fixed. The approach taken here is to: truncate the observations and bound the error for doing so, then define a finite set of grid points so that the domain is covered by neighborhoods around the grid points, then find the error supremum by taking the maximum error over the neighborhoods, and finally bounding the remainder by the Bernstein inequality (Hansen, 2008; Sun et al., 2014).

Lemma 3.2. *Uniform convergence*

Let $\{Y_i\}$ *be a sequence of independent random variables,* $\boldsymbol{D}_i \overset{iid}{\sim} \text{Dirichlet}(\mathbf{1}_n)$ *for* $i = 1, \ldots, n$, *and* $\boldsymbol{W}$ *be a random* n*-vector such that* $W_i = \sum_{j=1}^{n} D_{ij}$. *Let* $\{\boldsymbol{s}_i\} \in \mathbb{R}^d$ *be nonrandom location vectors. suppose that for some* $q > 2$, $\max_i E|Y_i|^q < \infty$. *Under regularity conditions C1 and C5, if* $n^{1-2/q}h^d / \log^2 n \to \infty$ *and* $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) < \infty$ *for any* $\boldsymbol{s} \in \mathcal{D}$, *then*

$$\sup_{\boldsymbol{s}\in\mathcal{D}}\left|n^{-1}\sum_{i=1}^{n}\left[K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)W_iY_i-E\{K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)W_iY_i\}|\boldsymbol{G}_n\right]\right|=O_p\left(\left\{n^{-1}h^{-d}\log n\right\}^{1/2}\right)$$

Proof. As noted, we begin by bounding the error that arises when we truncate the observations. Define $\tau_n = n^{1/(q-2)}(\log n)^{1/2}$ and $R_n(\boldsymbol{s}) = n^{-1}\sum_{i=1}^{n} K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)\{W_iY_i - W_iY_iI(|W_iY_i| \leq$ Thus, $R(\boldsymbol{s}) = n^{-1}\sum_{i=1}^{n} K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)W_iY_iI(|W_iY_i|>\tau_n)$. One useful result is that for large enough n, the truncation error vanishes. To show this, we bound the expected frequency of truncation using Chebyshev's inequality: $P(|W_iY_i|>\tau_n) \leq \tau_n^{-q}E|W_iY_i|^q \leq \tau_n^{-q}\Gamma(q+1)E|Y_i|^q$. For $q>2$, $\sum_{n=1}^{\infty}\tau_n^{-q} \leq \sum_{n=1}^{\infty} n^{-1}\log^{-1} n < \infty$, and by assumption, $E|Y_i|^q<\infty$. Thus, by the Borel-Cantelli lemma, there is an N such that for all $n>N$, $|Y_n|<\tau_n$. Since the Y_i are also independent, we have that for $n>N$, $|Y_i|<\tau_n$ for all $i\geq 1$. Thus, for large enough n, $R(\boldsymbol{s})=0$.

Next, we bound $|ER(\boldsymbol{s})|$ in a way that applies for all n. $|ER(\boldsymbol{s})| = \left|E_{\boldsymbol{W}}E_{\cdot|\boldsymbol{W}}R(\boldsymbol{s})\right| \leq E_{\boldsymbol{W}}\left[n^{-1}\sum_{i=1}^{n} W_iE_{\cdot|\boldsymbol{W}}\left\{|K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)||Y_i|I(|Y_i|>\tau_n/W_i)\right\}\right]$. Since $\tau_n/W_iI(|Y_i|>\tau_n/W_i) \leq |Y_i|$, we have $I(|Y_i|>\tau_n/W_i) \leq |W_iY_i|^{q-1}/\tau_n^{q-1}$. Thus,

$$\begin{aligned}
|ER(\boldsymbol{s})| &\leq \tau_n^{1-q}E_{\boldsymbol{W}}\left[n^{-1}\sum_{i=1}^{n} W_i^qE_{\cdot|\boldsymbol{W}}\left\{|K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)|\,|Y_i|^q\right\}\right]\\
&\leq \tau_n^{1-q}E_{\boldsymbol{W}}\left[\left(\sum_{i=1}^{n} W_i^q\right)E_{\cdot|\boldsymbol{W}}\left\{n^{-1}\sum_{j=1}^{n}|K_h(\|\boldsymbol{s}_j-\boldsymbol{s}\|)|\,|Y_j|^q\right\}\right]\\
&\leq \tau_n^{1-q}n\Gamma(q+1)\int_{\mathbb{R}^d}|K_h(\|\boldsymbol{s}-\boldsymbol{t}\|)|\,E\left(|Y_0|^q|\boldsymbol{s}_0=\boldsymbol{t}\right)f(\boldsymbol{t})\mathrm{d}\boldsymbol{t}\\
&\leq \tau_n^{1-q}n\Gamma(q+1)c_1\int_{\mathbb{R}^d}|K(\|\boldsymbol{u}\|)|\,\mathrm{d}\boldsymbol{u}\\
&\leq \tau_n^{1-q}n\Gamma(q+1)c_1c_2 = O(\tau_n^{1-q}) = O(r_n). \qquad (3.7)
\end{aligned}$$

It follows that with probability one, $\sup_{\boldsymbol{s}\in\mathcal{D}}|R(\boldsymbol{s})-ER(\boldsymbol{s})| = O(r_n)$.

Further, since $\mathcal{D}$ is compact, there is a positive constant c_3 such that $\mathcal{D}\subseteq\{\boldsymbol{s}:$

$\|\boldsymbol{s}\| \leq c_3\}$. We intend to define a regular grid of points such that the domain $\mathcal{D}$ is covered by neighborhoods of the points. Define the neighborhood of each grid point $\boldsymbol{s}_j$ by $N_j = \{\boldsymbol{s} : \|\boldsymbol{s}_j - \boldsymbol{s}\| \leq hr_n\}$. Then $\mathcal{D}$ can be covered by $N \leq c_3/(hr_n)^2$ such neighborhoods and the supremum on $\mathcal{D}$ can be replaced by the maximum over the neighborhoods.

The kernel has compact support, so there is a finite positive constant L such that $\|\boldsymbol{s}\| > L \Rightarrow K(\|\boldsymbol{s}\|) = 0$. Because the kernel is Lipschitz continuous, there is a finite positive constant c_5 such that for all $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d$, $|K(\|\boldsymbol{s}\|) - K(\|\boldsymbol{t}\|)| \leq c_5\|\boldsymbol{s} - \boldsymbol{t}\|$. Let $K^\dagger(\|\boldsymbol{s}\|) = c_5 I(\|\boldsymbol{s}\| \leq 2L)$ and $K_h^\dagger(\|\boldsymbol{s}\|) = h^{-d}K^\dagger\left(h^{-1}\|\boldsymbol{s}\|\right)$. Then $h^{-1}\|\boldsymbol{s} - \boldsymbol{s}_j\| \leq r_n$ for all $\boldsymbol{s} \in N_j$ and, for all $i \geq 1$,

$$\left|K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) - K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\right)\right| \leq r_n K^\dagger\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\right). \tag{3.8}$$

Define $R_1(\boldsymbol{s}) = n^{-1}\sum_{i=1}^n K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|)W_iY_iI(|W_iY_i| \leq \tau_n)$ and $R_1^\dagger(\boldsymbol{s}) = n^{-1}\sum_{i=1}^n K_h^\dagger(\|\boldsymbol{s}_i - \boldsymbol{s}\|)|W_iY_i|I(|W_iY_i| \leq \tau_n)$. By an argument much like (3.7), $E|R_h^\dagger(\boldsymbol{s})| \leq c_6$ for some positive constant c_6. Then by (3.8),

$$\begin{aligned}
\sup_{\boldsymbol{s}\in N_j}|R_1(\boldsymbol{s}) - ER_1(\boldsymbol{s})| &\leq |R_1(\boldsymbol{s}_j) - ER_1(\boldsymbol{s}_j)| + r_n\left(\left|R_1^\dagger(\boldsymbol{s}_j)\right| + E\left|R_1^\dagger(\boldsymbol{s}_j)\right|\right) \\
&\leq |R_1(\boldsymbol{s}_j) - ER_1(\boldsymbol{s}_j)| + r_n\left|R_1^\dagger(\boldsymbol{s}_j) - ER_1^\dagger(\boldsymbol{s}_j)\right| + 2r_nE|R_1^\dagger(\boldsymbol{s}_j)| \\
&\leq |R_1(\boldsymbol{s}_j) - ER_1(\boldsymbol{s}_j)| + \left|R_1^\dagger(\boldsymbol{s}_j) - ER_1^\dagger(\boldsymbol{s}_j)\right| + 2r_nc_6.
\end{aligned}$$

Thus, for large enough n, we can bound the error supremum by

$$\begin{aligned} P\left(\sup_{\boldsymbol{s}\in\mathcal{D}}|R_1(\boldsymbol{s})-ER_1(\boldsymbol{s})|>4c_6r_n\right) \leq & N\max_{1\leq j\leq N} P\left(\sup_{\boldsymbol{s}\in N_j}|R_1(\boldsymbol{s})-ER_1(\boldsymbol{s})|>4c_6r_n\right) \\ \leq & N\max_{1\leq j\leq N} P\left(|R_1(\boldsymbol{s}_j)-ER_1(\boldsymbol{s}_j)|>c_6r_n\right) \\ & + N\max_{1\leq j\leq N} P\left(\left|R_1^\dagger(\boldsymbol{s}_j)-ER_1^\dagger(\boldsymbol{s}_j)\right|>c_6r_n\right). \end{aligned} \tag{3.9}$$

It follows that the probabilities can be bounded by applying Bernstein's inequality. Since $\boldsymbol{W}$ is the sum of n independent Dirichlet random vectors, $R(\boldsymbol{s})$ and $R^\dagger(\boldsymbol{s})$ can be written as sums of independent random variables, each of which is a linear combination of the elements of a Dirichlet random vector. Applying Bernstein's inequality requires bounding each lienar combination and calculating its second moment.

We have seen that for large enough n, $|Y_i|<\tau_n$ with probability one, so assume that $|Y_i|<\tau_n$. Since $K(\|\cdot\|)$ is Lipschitz continuous and has compact support, we have that $K\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)\leq c_7$ for some positive constant c_7. We have that

$$\begin{aligned} |R_1(\boldsymbol{s})-ER_1(\boldsymbol{s})| &= n^{-1}\left|\sum_{i=1}^n Y_iW_iK_h\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)-E\left\{\sum_{i=1}^n Y_iW_iK_h\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)\right\}\right| \\ &= n^{-1}\left|\sum_{i=1}^n\sum_{j=1}^n Y_iD_{ji}K_h\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)-E\left\{\sum_{i=1}^n\sum_{j=1}^n Y_iD_{ji}K_h\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)\right\}\right| \\ &= n^{-1}\left|\sum_{j=1}^n\sum_{i=1}^n\left[Y_iD_{ji}K_h\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)-n^{-1}E\left\{Y_iK_h\left(h^{-1}\|\boldsymbol{s}_i-\boldsymbol{s}\|\right)\right\}\right]\right| \\ &= n^{-1}\left|h^{-2}\sum_{j=1}^n V_j(\boldsymbol{s})\right| \end{aligned}$$

where $V_j(\boldsymbol{s})$ for $j = 1, \ldots, n$ are independent. In addition, we have

$$\begin{aligned}
\text{var } V_j(\boldsymbol{s}) &= E\left(\sum_{i=1}^{n}\left[Y_i D_{ji} K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) - E\left\{Y_i D_{ji} K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right)\right\}\right]\right)^2 \\
&= E_{\boldsymbol{D}} E\left\{\left(\sum_{i=1}^{n} D_{ji}\left[Y_i K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) - E\left\{Y_i K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right)\right\}\right]\right)^2 |\boldsymbol{D}_j\right\} \\
&\leq E\left(\sum_{i=1}^{n} D_{ji}^2\right) E\left(\sum_{i=1}^{n}\left[Y_i K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) - E\left\{Y_i K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right)\right\}\right]\right)^2.
\end{aligned}$$

Since $K^2(\cdot)$ satisfies regularity condition (1), by regularity condition (5), we have

$$\begin{aligned}
&E\left(\sum_{i=1}^{n}\left[Y_i K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right) - E\left\{Y_i K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right)\right\}\right]\right)^2 \\
&\quad = E_{\boldsymbol{Y}} E\left\{\left(\sum_{i=1}^{n} K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}\|\right)\left[Y_i - E\left\{Y_i\right\}\right]\right)^2 |\sigma\langle \boldsymbol{Y}\rangle\right\} \\
&\quad \leq cnh^2.
\end{aligned}$$

Also, because $\sum_{i=1}^{n} D_{ji} = 1$, we have

$$\begin{aligned}
E\left(\sum_{i=1}^{n} D_{ji}\right)^2 &= E\sum_{i=1}^{n} D_{ji}^2 + nED_{j1}\sum_{k\neq 1} D_{jk} \\
&= E\sum_{i=1}^{n} D_{ji}^2 + (n-1)/(n+1) \\
&= 1.
\end{aligned}$$

Thus $E\sum_{i=1}^{n} D_{ji}^2 = 2/(n+1)$, and $\text{var}\left(V_j(\boldsymbol{s})\right) \leq 2cnh^2/(n+1) \leq 2cnh^2$.

Since we have the quantity $|R_1(\boldsymbol{s}) - ER_1(\boldsymbol{s})|$ expressed as the sum of n independent random variables, each of which is bounded by $\pm 2\tau_n c_7$ and with variance

bounded by $2ch^2$. Thus, by Bernstein's inequality,

$$\begin{aligned} P\left(|R_1(\boldsymbol{s}) - ER_1(\boldsymbol{s})| > c_6 r_n\right) &= P\left(\left|\sum_{j=1}^{n} V_j(\boldsymbol{s})\right| > c_6 r_n n h^2\right) \\ &\leq P\left(\sum_{j=1}^{n} |V_j(\boldsymbol{s})| > c_6 r_n n h^2\right) \\ &\leq 2\exp\left\{-(c_6 r_n n h^2)^2 / \left(2\sum_{j=1}^{n} \text{var } V_j(\boldsymbol{s}) + 4/3 c_6 c_7 \tau_n r_n n h^2\right)^{-2}\right\} \\ &\leq 2\exp\left\{-c_6^2 \log n/(2c + 4c_7)\right\} \leq 2n^{-c_6} \end{aligned} \tag{3.10}$$

By an analogous argument, $\left|R_1^\dagger(\boldsymbol{s}) - ER_1^\dagger(\boldsymbol{s})\right| = n^{-1}\left|h^{-2}\sum_{j=1}^{n} V_j^\dagger(\boldsymbol{s})\right|$ where $V_j^\dagger(\boldsymbol{s})$ for $j = 1, \ldots, n$ are independent and $\left|V_j^\dagger(\boldsymbol{s})\right| \leq 2\tau_n c_8$ for a positive constant c_8. Thus,

$$P\left(\left|R_1^\dagger(\boldsymbol{s}) - ER_1^\dagger(\boldsymbol{s})\right| > c_6 r_n\right) \leq 2n^{-c_6}. \tag{3.11}$$

The proof is completed by substituting (3.10) and (3.11) into (3.9):

$$P\left(\sup_{\boldsymbol{s}\in\mathcal{D}} |R_1(\boldsymbol{s}) - ER_1(\boldsymbol{s})| > 4c_6 r_n\right) \leq c(hr_n)^{-2} n^{-c_6} \to 0.$$

□

Lemma 3.3. *Second moment of the normal equations Under regularity conditions (C1) and (C3)-(C6),*

$$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^T\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^*\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{H}^{-1} = \begin{pmatrix} \kappa_0 f(\boldsymbol{s})\boldsymbol{\Psi}(\boldsymbol{s}) & \boldsymbol{0}_{p\times dp} \\ \boldsymbol{0}_{dp\times p} & \kappa_2 f(\boldsymbol{s})\boldsymbol{\Psi}(\boldsymbol{s}) \otimes \boldsymbol{I}_d \end{pmatrix} + O_p\left(c_n \boldsymbol{1}_{(d+1)p}\boldsymbol{1}_{(d+1)p}^T\right)$$

where $c_n = h + \left\{n^{-1}h^{-d}\log n\right\}^{1/2}$.

Proof. Using the definitions of $\boldsymbol{H}$, $\boldsymbol{Z}_{\boldsymbol{s}}$, $\boldsymbol{K}_{h,\boldsymbol{s}}$, and $\boldsymbol{W}$,

$$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^T\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^*\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{H}^{-1} =$$

$$\begin{pmatrix} n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T w_i^* K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) & n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T \otimes (\frac{\boldsymbol{s}_i - \boldsymbol{s}}{h})^T w_i^* K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) \\ n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T \otimes (\frac{\boldsymbol{s}_i - \boldsymbol{s}}{h}) w_i^* K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) & n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T \otimes (\frac{\boldsymbol{s}_i - \boldsymbol{s}}{h})(\frac{\boldsymbol{s}_i - \boldsymbol{s}}{h})^T w_i^* K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|) \end{pmatrix}.$$

By Lemma 1, the expectation is

$$n^{-1}E\left\{\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{H}^{-1}|\mathcal{G}_{n}\right\} =$$
$$\begin{pmatrix} n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{i}\boldsymbol{X}_{i}^{T}K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|) & n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{i}\boldsymbol{X}_{i}^{T}\otimes(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h})^{T}K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|) \\ n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{i}\boldsymbol{X}_{i}^{T}\otimes(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h})K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|) & n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_{i}\boldsymbol{X}_{i}^{T}\otimes(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h})(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h})^{T}K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|) \end{pmatrix}.$$

As $n\to\infty$, due to Lipschitz continuity of $f(\cdot)$, symmetry of the kernel, and regularity condition (C5), the expectation tends to

$$n^{-1}E\left\{\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{H}^{-1}|\mathcal{G}_{n}\right\}\xrightarrow{p}$$
$$\begin{pmatrix} \kappa_{0}f(\boldsymbol{s})\boldsymbol{\Psi}(\boldsymbol{s}) & \boldsymbol{0}_{p\times(dp)} \\ \boldsymbol{0}_{(dp)\times p} & \kappa_{2}f(\boldsymbol{s})\boldsymbol{\Psi}(\boldsymbol{s})\otimes\boldsymbol{I}_{d} \end{pmatrix}+O(h\boldsymbol{1}_{(d+1)p}\boldsymbol{1}_{(d+1)p}^{T}).$$

The difference between the mean and its expectation can be bounded by applying Lemma 1:

$$n^{-1}\left[\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{H}^{-1}-E\left\{\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{H}^{-1}\right\}\right]=O_{p}\left(\{(nh^{d})^{-1}\log n\}^{1/2}\right),$$

which completes the proof of Lemma 3. □

Lemma 3.4. *Bias of the normal equations Under regularity conditions (C1) - (C6),*

$$\boldsymbol{\beta}(\boldsymbol{s})-(\boldsymbol{I}_{p},\boldsymbol{0}_{p\times dp})(\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{Z}_{\boldsymbol{s}})^{-1}\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{m}=-2^{-1}\kappa_{0}^{-1}\kappa_{2}h^{2}\sum_{k=1}^{d}\nabla_{kk}^{2}\boldsymbol{\beta}(\boldsymbol{s})+o_{p}(h^{2}\boldsymbol{1}_{p})$$

Proof. To begin, note that

$$\boldsymbol{\beta}(\boldsymbol{s})-(\boldsymbol{I}_{p},\boldsymbol{0}_{p\times dp})(\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{Z}_{\boldsymbol{s}})^{-1}\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{m}=$$
$$(\boldsymbol{I}_{p},\boldsymbol{0}_{p\times 2p})\boldsymbol{H}^{-1}(n^{-1}\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{H}^{-1})^{-1}n^{-1}\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\left\{\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{\zeta}(\boldsymbol{s})-\boldsymbol{m}\right\} \tag{3.12}$$

and by a standard argument,

$$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\left\{\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{\zeta}(\boldsymbol{s})-\boldsymbol{m}\right\}= \tag{3.13}$$
$$-2^{-1}h^{2}\sum_{j=1}^{p}\begin{Bmatrix} n^{-1}\sum_{i=1}^{n}w_{i}^{*}\boldsymbol{x}_{i}x_{ij}\left(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h}\right)^{T}\nabla^{2}\boldsymbol{\beta}_{j}(\boldsymbol{s}_{i}^{\dagger})\left(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h}\right)K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|) \\ n^{-1}\sum_{i=1}^{n}w_{i}^{*}\boldsymbol{x}_{i}\otimes\left(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h}\right)x_{ij}\left(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h}\right)^{T}\nabla^{2}\boldsymbol{\beta}_{j}(\boldsymbol{s}_{i}^{\dagger})\left(\frac{\boldsymbol{s}_{i}-\boldsymbol{s}}{h}\right)K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|) \end{Bmatrix}$$

where $\boldsymbol{s}_i^\dagger = \boldsymbol{s} + \theta_i^\dagger(\boldsymbol{s}_i - \boldsymbol{s})$ for some $\theta_i^\dagger \in [0,1]$ is a point on the line between $\boldsymbol{s}$ and $\boldsymbol{s}_i$.

We now have that, by Lemma 1, the expectations over the observed data are

$$-(2n)^{-1}h^2\sum_{j=1}^{p}E\left[\sum_{i=1}^{n}w_i^*\boldsymbol{x}_i x_{ij}\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)^T\nabla^2\boldsymbol{\beta}_j(\boldsymbol{s}_i^\dagger)\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)|\mathcal{G}_n\right] =$$
$$-(2n)^{-1}h^2\sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{x}_i x_{ij}\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)^T\nabla^2\boldsymbol{\beta}_j(\boldsymbol{s}_i^\dagger)\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|),$$

and

$$-(2n)^{-1}h^2\sum_{j=1}^{p}E\left[\sum_{i=1}^{n}w_i^*\boldsymbol{x}_i\otimes\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)x_{ij}\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)^T\nabla^2\boldsymbol{\beta}_j(\boldsymbol{s}_i^\dagger)\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)|\mathcal{G}_n\right] =$$
$$-(2n)^{-1}h^2\sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{x}_i\otimes\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)x_{ij}\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)^T\nabla^2\boldsymbol{\beta}_j(\boldsymbol{s}_i^\dagger)\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)$$

.

Thus, as $n\to\infty$, due to continuity of the second derivatives of $\boldsymbol{\beta}(\cdot)$, symmetry of the kernel function, Lipschitz continuity of $f(\cdot)$, and regularity condition (C5), the expectations of the components of (3.13) tend to:

$$-(2n)^{-1}h^2\sum_{j=1}^{p}E\left[\sum_{i=1}^{n}w_i^*\boldsymbol{x}_i x_{ij}\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)^T\nabla^2\boldsymbol{\beta}_j(\boldsymbol{s}_i^\dagger)\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)|\mathcal{G}_n\right]\xrightarrow{p}$$
$$-2^{-1}h^2\kappa_2 f(\boldsymbol{s})\boldsymbol{\Psi}(\boldsymbol{s})\left\{\sum_{l=1}^{d}\nabla_{ll}^2\beta_1(\boldsymbol{s}),\ldots,\sum_{l=1}^{d}\nabla_{ll}^2\beta_p(\boldsymbol{s})\right\}^T+o(h^2\mathbf{1}_p)$$

and (due especially to the symmetry of the kernel function),

$$-(2n)^{-1}h^2\sum_{j=1}^{p}E\left[\sum_{i=1}^{n}\boldsymbol{x}_i\otimes\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)x_{ij}\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)^T\nabla^2\boldsymbol{\beta}_j(\boldsymbol{s}_i^\dagger)\left(\frac{\boldsymbol{s}_i-\boldsymbol{s}}{h}\right)K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)|\mathcal{G}_n\right]$$
$$\xrightarrow{p}o(h^2\mathbf{1}_{dp}).$$

Using Lemma 2 to bound the difference between (3.13) and its expectation, we

have that

$$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\left\{\boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{\zeta}(\boldsymbol{s})-\boldsymbol{m}\right\}\xrightarrow{p}$$
$$\begin{pmatrix} -2^{-1}h^{2}\kappa_{2}f(\boldsymbol{s})\Psi(\boldsymbol{s})\left\{\sum_{l=1}^{d}\nabla_{ll}^{2}\beta_{1}(\boldsymbol{s}),\ldots,\sum_{l=1}^{d}\nabla_{ll}^{2}\beta_{p}(\boldsymbol{s})\right\}^{T} \\ \mathbf{0}_{dp} \end{pmatrix}+o_{p}(h^{2}\mathbf{1}_{(d+1)p}). \tag{3.14}$$

The statement of Lemma 4 now follows by Lemma 3 and Slutsky's theorem. □

Lemma 3.5. *Adaptive weights Under regularity conditions (C1) - (C6),*

$$\{n^{2}h^{d}f(\boldsymbol{s})\}^{1/2}\left(\bar{\boldsymbol{\beta}}^{*}(\boldsymbol{s})-\boldsymbol{\beta}(\boldsymbol{s})-2^{-1}\kappa_{0}^{-1}\kappa_{2}h^{2}\sum_{k=1}^{d}\nabla_{kk}^{2}\boldsymbol{\beta}(\boldsymbol{s})\right)$$
$$\xrightarrow{d}N\left(\mathbf{0},\kappa_{0}^{-2}\nu_{0}\sigma^{2}\boldsymbol{\Psi}^{-1}(\boldsymbol{s})\right)$$

Proof. First, note that

$$\begin{aligned}\{n^{2}h^{d}f(\boldsymbol{s})\}^{1/2}\left(\bar{\boldsymbol{\beta}}^{*}(\boldsymbol{s})-\boldsymbol{\beta}(\boldsymbol{s})\right)&=\{n^{2}h^{d}f(\boldsymbol{s})\}^{1/2}(\boldsymbol{I}_{p},\mathbf{0}_{p\times dp})(\boldsymbol{Z}_{s}^{T}\boldsymbol{K}_{h,s}\boldsymbol{W}^{*}\boldsymbol{Z}_{s})^{-1}\boldsymbol{Z}_{s}^{T}\boldsymbol{K}_{h,s}\boldsymbol{W}^{*}\boldsymbol{\varepsilon}\\&\quad+\{n^{2}h^{d}f(\boldsymbol{s})\}^{1/2}(\boldsymbol{I}_{p},\mathbf{0}_{p\times dp})\{(\boldsymbol{Z}_{s}^{T}\boldsymbol{K}_{h,s}\boldsymbol{W}^{*}\boldsymbol{Z}_{s})^{-1}\boldsymbol{Z}_{s}^{T}\boldsymbol{K}_{h,s}\boldsymbol{W}^{*}\boldsymbol{m}-\boldsymbol{\beta}(\boldsymbol{s})\}\\&=\boldsymbol{J}_{n1}+\boldsymbol{J}_{n2}.\end{aligned}$$

Since $\boldsymbol{D}_{1}^{*},\ldots,\boldsymbol{D}_{n}^{*}$ are independent,

$$\begin{aligned}\text{var}\left(\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{W}^{*}\boldsymbol{\varepsilon}\right)&=\sum_{i=1}^{n}\text{var}\ \left(\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{D}_{i}^{*}\boldsymbol{\varepsilon}\right)\\&=n\text{var}\left(\boldsymbol{H}^{-1}\boldsymbol{Z}_{\boldsymbol{s}}^{T}\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{D}_{1}^{*}\boldsymbol{\varepsilon}\right).\end{aligned}$$

By applying Lemma 1, we get

$$E\{\boldsymbol{H}^{-1}\boldsymbol{Z}_{s}^{T}\boldsymbol{K}_{h,s}\boldsymbol{D}_{i}^{*}\boldsymbol{\varepsilon}|\mathcal{G}_{n}\}=n^{-1}\sum_{i=1}^{n}\{\boldsymbol{Z}_{s}\}_{i}\varepsilon_{i}K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|)$$

and

$$\begin{aligned}&E\{\boldsymbol{H}^{-1}\boldsymbol{Z}_{s}^{T}\boldsymbol{K}_{h,s}\boldsymbol{D}_{1}^{*}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{T}\boldsymbol{D}_{1}^{*}\boldsymbol{K}_{h,s}\boldsymbol{Z}_{s}\boldsymbol{H}^{-1}|\mathcal{G}_{n}\}\\&\quad=2/\{n(n+1)\}\sum_{i=1}^{n}\sum_{j=1}^{i}\{\boldsymbol{Z}_{s}\}_{i}\{\boldsymbol{Z}_{s}\}_{j}^{T}\varepsilon_{i}\varepsilon_{j}K_{h}(\|\boldsymbol{s}_{i}-\boldsymbol{s}\|)K_{h}(\|\boldsymbol{s}_{j}-\boldsymbol{s}\|).\end{aligned}$$

Thus,

$$h^d f(\boldsymbol{s})\text{var}\{\boldsymbol{H}^{-1}\boldsymbol{Z}_s^T\boldsymbol{K}_{h,s}\boldsymbol{W}^*\boldsymbol{\varepsilon}|\mathcal{G}_n\} = \frac{n-1}{n(n+1)}h^d f(\boldsymbol{s})\sum_{i=1}^{n}\{\boldsymbol{Z}_s\}_i\{\boldsymbol{Z}_s\}_i^T\varepsilon_i^2 K_h^2(\|\boldsymbol{s}_i-\boldsymbol{s}\|)$$
$$+\frac{2n}{n(n+1)}h^d f(\boldsymbol{s})\sum_{i=2}^{n}\sum_{j=1}^{i-1}\{\boldsymbol{Z}_s\}_i\{\boldsymbol{Z}_s\}_j^T\varepsilon_i\varepsilon_j K_h(\|\boldsymbol{s}_i-\boldsymbol{s}\|)K_h(\|\boldsymbol{s}_j-\boldsymbol{s}\|)$$
$$\xrightarrow{p}\sigma^2(\boldsymbol{s})f^2(\boldsymbol{s})\begin{pmatrix}\nu_0\boldsymbol{\Psi}(\boldsymbol{s}) & \boldsymbol{0}_{p\times dp}\\ \boldsymbol{0}_{dp\times p} & \nu_2\boldsymbol{\Psi}(\boldsymbol{s})\otimes\boldsymbol{I}_2\end{pmatrix}.$$

By the central limit theorem, Lemma 3.3, and Slutsky's theorem,

$$\boldsymbol{J}_{n1}\xrightarrow{d}N\left(\boldsymbol{0},\nu_0\kappa_0^{-2}\sigma^2\boldsymbol{\Psi}^{-1}(\boldsymbol{s})\right).$$

By Lemma 4, we have that

$$\boldsymbol{J}_{n2}=\{n^2h^d f(\boldsymbol{s})\}^{1/2}-2^{-1}\kappa_0^{-1}\kappa_2 h^2\sum_{k=1}^{d}\nabla_{kk}^2\boldsymbol{\beta}(\boldsymbol{s})+o_p(1\boldsymbol{1}_p),$$

which completes the proof of Lemma 5. □

Lemma 3.6. *Bootstrap distribution of adaptive LASSO estimator Under regularity conditions (C1) - (C7),*

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

where $\left\{\nabla_{uu}^2\boldsymbol{\beta}_{(a)}(\boldsymbol{s})+\nabla_{vv}^2\boldsymbol{\beta}_{(a)}(\boldsymbol{s})\right\}=(\nabla_{uu}^2\boldsymbol{\beta}_1(\boldsymbol{s})+\nabla_{vv}^2\boldsymbol{\beta}_1(\boldsymbol{s}),\ldots,\nabla_{uu}^2\boldsymbol{\beta}_{p_0}(\boldsymbol{s})+\nabla_{vv}^2\boldsymbol{\beta}_{p_0}(\boldsymbol{s}))^T$.

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

have

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

The limiting behavior of the last term differs between the cases $j \leq p_0(\boldsymbol{s})$ and $j > p_0(\boldsymbol{s})$. *Case* $j \leq p_0(\boldsymbol{s})$: If $j \leq p_0(\boldsymbol{s})$, then $\phi_j^*(\boldsymbol{s}) \leq a_n$ and $\left|\left\{\|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s}) + \alpha_n \boldsymbol{u}_{(j)}\| - \|\boldsymbol{\zeta}_{(j)}(\boldsymbol{s})\|\right\}\right| \leq \alpha_n \|\boldsymbol{u}_{(j)}\|$. Thus,

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

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 by assumption $\alpha_n b_n \xrightarrow{p} \infty$, then if $\|\boldsymbol{u}_{(j)}\| \neq 0$, it follows that

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

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

$$
H_n^\dagger(\boldsymbol{u}) = (1/2)\alpha_n^2 \boldsymbol{u}^T \left\{\boldsymbol{Z_s}^T \boldsymbol{W}^* \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{Z_s}\right\} \boldsymbol{u} - \alpha_n \boldsymbol{u}^T \left[\boldsymbol{Z_s}^T \boldsymbol{W}^* \boldsymbol{K}_{h,\boldsymbol{s}} \left\{\boldsymbol{Y} - \boldsymbol{Z_s}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right]
$$

otherwise. It follows that $H_n^\dagger(\boldsymbol{u})$ 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 $\hat{\boldsymbol{u}}_n^*$ corresponding to the true

nonzero coefficients and true zero coefficients. Then

$$\hat{\boldsymbol{u}}^*_{(a)n} = \left\{n^{-1}\boldsymbol{Z}^T_{\boldsymbol{s}(a)}\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{Z}_{\boldsymbol{s}(a)}\right\}^{-1}\left[h\boldsymbol{Z}^T_{\boldsymbol{s}(a)}\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\left\{\boldsymbol{Y} - \boldsymbol{Z}_{\boldsymbol{s}(a)}\boldsymbol{\zeta}_{(a)}(\boldsymbol{s})\right\}\right]$$

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 Lemmas 3 and 4,

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

the statement of Lemma 6 follows. □

Lemma 3.7. *Consistency of bootstrap for covariate selection by adaptive LASSO*

Under regularity conditions (C1) - (C7), for $j > p_0(\boldsymbol{s})$,

$$P\left\{\hat{\boldsymbol{\zeta}}^*_{(j)}(\boldsymbol{s}) = \mathbf{0}|\mathcal{G}_n\right\} \to 1.$$

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

$$\begin{aligned}
\mathbf{0} &= \boldsymbol{Z}^T_{\boldsymbol{s}(p)}\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\left\{\boldsymbol{Y} - \boldsymbol{Z}_{\boldsymbol{s}(-p)}\hat{\boldsymbol{\zeta}}^*_{(-p)}(\boldsymbol{s}) - \boldsymbol{Z}_{\boldsymbol{s}(p)}\hat{\boldsymbol{\zeta}}^*_{(p)}(\boldsymbol{s})\right\} - \phi^*_{(p)}(\boldsymbol{s})\hat{\boldsymbol{\zeta}}^*_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}^*_{(p)}(\boldsymbol{s})\|^{-1} \\
&= \boldsymbol{Z}^T_{\boldsymbol{s}(p)}\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\left[\boldsymbol{Y} - \boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{\zeta}(\boldsymbol{s}) - (2\kappa_0)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\
&+ \boldsymbol{Z}^T_{\boldsymbol{s}(p)}\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{Z}_{\boldsymbol{s}(-p)}\left[\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s}) + (2\kappa_0)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}^*_{(-p)}(\boldsymbol{s})\right] \\
&+ \boldsymbol{Z}^T_{\boldsymbol{s}(p)}\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{Z}_{\boldsymbol{s}(p)}\left[\boldsymbol{\zeta}_{(p)}(\boldsymbol{s}) + (2\kappa_0)^{-1}h^2\kappa_2\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}^*_{(p)}(\boldsymbol{s})\right] \\
&- \phi^*_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}^*_{(p)}(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}^*_{(p)}(\boldsymbol{s})\|^{-1}.
\end{aligned}$$

Thus,

$$
\begin{aligned}
&h\phi_p^*(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}(\boldsymbol{s})\|^{-1} \\
&\quad = h\boldsymbol{Z}_{\boldsymbol{s}(p)}^T\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\left[\boldsymbol{Y} - \boldsymbol{Z}_{\boldsymbol{s}}\boldsymbol{\zeta}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla_{uu}^2\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla_{vv}^2\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\
&\qquad + nh\left\{n^{-1}\boldsymbol{Z}_{\boldsymbol{s}(p)}^T\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{Z}_{\boldsymbol{s}(-p)}\right\}\left[\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla_{uu}^2\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s}) + \nabla_{vv}^2\boldsymbol{\zeta}_{(-p)}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_{(-p)}^*(\boldsymbol{s})\right] \\
&\qquad + nh\left\{n^{-1}\boldsymbol{Z}_{\boldsymbol{s}(p)}^T\boldsymbol{W}^*\boldsymbol{K}_{h,\boldsymbol{s}}\boldsymbol{Z}_{\boldsymbol{s}(p)}\right\}\left[\boldsymbol{\zeta}_{(p)}(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla_{uu}^2\boldsymbol{\zeta}_{(p)}(\boldsymbol{s}) + \nabla_{vv}^2\boldsymbol{\zeta}_{(p)}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s})\right].
\end{aligned}
\tag{3.15}
$$

By Lemma 3,

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

From Lemma 5, we have that

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

and

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

We showed in the proof of Lemma 6 that

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

The right hand side of (3.15) is $O_p(1)$, so for $\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s})$ to be a solution, we must have that $h\phi_p^*(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s})\|^{-1} = O_p(1)$. But since by assumption $\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s}) \neq \boldsymbol{0}$, there must be some $k \in \{1, \ldots, d+1\}$ such that $|\hat{\zeta}_{(p)_k}^*(\boldsymbol{s})| = \max\{|\hat{\zeta}_{(p)_m}^*(\boldsymbol{s})| : 1 \leq m \leq d+1\}$. 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 by assumption $(nh)^{-1}b_n \to \infty$, we have that $h\phi_p^*(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s})\|\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s})\|^{-1} \geq hb_n3^{-1/2} \to \infty$ and therefore the left hand side of (3.15) dominates the sum to the right side. Thus, for large enough n, $\hat{\boldsymbol{\zeta}}_{(p)}^*(\boldsymbol{s}) \neq \boldsymbol{0}$ cannot maximize $\mathcal{J}^*(\cdot)$, and therefore $P\left\{\hat{\boldsymbol{\zeta}}_{(b)}^*(\boldsymbol{s}) = \boldsymbol{0}|\mathcal{G}_n\right\} \to 1$. □

Chapter 4

An Empirical Bayes Procedure for Inference in Local Polynomial Models

4.1 Introduction

Introduced in chapter 2, local adaptive grouped regularization (LAGR) is a method for estimating varying coefficient regression (VCR) models. The method of LAGR utilizes local polynomial regression, a form of kernel smoothing (Fan and Gijbels, 1996). Kernel smoothing requires estimation of a bandwidth parameter $\hat{h}$, and then estimation and inference proceed conditional on $\hat{h}$. However, there is uncertainty in the estimation of $\hat{h}$, which is not reflected when estimates are conditional on $\hat{h}$. Furthermore, the bandwidth is a nuisance parameter, and it is preferable to report results marginal to h rather than conditional on $\hat{h}$.

In this chapter, an empirical Bayes method is proposed to simulate the marginal distribution of the local coefficient estimates. Empirical Bayes methods are those which use the data to estimate a prior distribution. Here, we use a weighted like-

lihood bootstrap (WLB) to estimate the hyperparameters in the distribution of the bandwidth parameter, which is known as a parametric empirical Bayes (PEB) approach. At this level, the bootstrap resamples h^* are interpreted as samples drawn from the posterior distribution for h, which is called the posterior hyperprior. The resamples are not considered independent because they are related by the bootstrapping weights. In cases where the bootstrap is used to estimate a parameter after model selection, the nonparametric delta method of Efron (1981) can reduce the uncertainty of the estimate by smoothing out discontinuities that are due to the model selection (Efron, 2014). Since the method of LAGR was used to select which coefficient functions are locally nonzero (called local variable selection), we employ the nonparametric delta method for an improved estimate of the variability in h.

The particular WLB used here to generate the bootstrap resamples was introduced in chapter 3. Its weights are generated by a sum of independent Dirichlet random vectors, which induces a central tendency in the bootstrap weights. As a result, the weights are not uniformly distributed. Although uniformity is usually a desirable property of bootstrap resampling, it was shown in the previous chapter that the nonuniform weights are necessary for the bootstrap estimates of the local coefficients to converge uniformly on the spatial domain. A side effect of the nonuniform bootstrap weights is that the local coefficient distribution as estimated by the WLB has smaller variance than the true distribution of the non-bootstrapped local coefficient estimator. It is therefore necessary to correct the variance of the WLB distribution for the local coefficient estimates.

A form of the WLB was introduced by Rubin (1981) as the Bayesian Bootstrap (BB). This approach was built on the observation that the standard nonparametric

bootstrap samples from a uniform multinomial distribution where each observation is added to the log-likelihood once for each time it is resampled (Efron, 1979). Then Rubin (1981) replaced the discrete resample counts by uniform continuous weights from a Dirichlet distribution, with the Bayesian part of the BB name coming from the interpretation of parameters of the Dirichlet distribution as hyperparameters that express a nonparametric prior likelihood for the observed data. The WLB name comes from the paper of Newton and Raftery (1994), which used the BB approach in a fully parametric model to generate samples of a model parameter from the likelihood. The WLB samples of the parameter values are then resampled according to a prior distribution by a sampling-importance resampling (SIR) algorithm (Rubin, 1988; Smith and Gelfand, 1992). This process generates samples that are simulation-consistent for the posterior distribution of the parameter (Newton and Raftery, 1994).

The method we develop here uses the WLB both to estimate the prior distribution and then to estimate the marginal posterior. This is an empirical Bayes approach in that the data is used to estimate the hyperprior distribution. The approach follows Laird and Louis (1987) in using the bootstrap to simulate the hyperprior and the marginal posterior both. Unlike the foregoing, however, the method developed here is applied in a nonparametric regression setting. One consequence is that the bandwidth parameter here cannot be estimated by maximum likelihood, and thus we propose to estimate it by AIC instead. Additionally, we develop an approach for adjusting the WLB samples in order to correct their variance. Because the method of LAGR incorporates data-based model selection, estimates such as the bandwidth that minimizes AIC may be discontinuous near the boundaries of model-selection regions (Efron, 2014). Bootstrap aggregation, or bagging, is the process of averaging an esti-

mate over several bootstrap resamples in order to smooth out discontinuities caused by model selection (Breiman, 1996). The expectation and variance of the posterior hyperprior are estimated by bagging the WLB resamples of the bandwidth, using the nonparametric delta method of citetEfron-2014 to estimate the variance. In particular, we assume a gamma model for the distribution of the bandwidth and implement a variance adjustment to correct for WLB underdispersion by multiplying the estimated variance by $n^{1/2}$. Hyperparameters for the gamma distribution of bandwidth are then estimated by the method of moments.

The remainder of the chapter is organized as follows. Section 2 briefly describes varying coefficients regression, local covariate selection and estimation by local adaptive grouped regularization, and the weighted likelihood bootstrap for simulating the distribution of local coefficient estimates. Section 3 describes a hierarchical model for the local coefficient estimates and proposes a method for estimation and inference in the hierarchical framework. In section 4, the proposed method is demonstrated via a simulation study.

4.2 Varying Coefficients Regression

4.2.1 Model

The varying coefficients regression model is the same setting as was used in chapters 2 and 3 of this dissertation.

Consider a response variable Y that is related to the p covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$ on the two-dimensional domain $\mathcal{D} \subset \mathbb{R}^2$ by a linear equation with coefficients $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \ldots, \beta_p(\cdot))^T$. Assume that the coefficients are smoothly varying functions of

location. The model is written

$$Y(\boldsymbol{s}) = \boldsymbol{X}^T(\boldsymbol{s})\boldsymbol{\beta}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \tag{4.1}$$

where $\boldsymbol{s} \in \mathcal{D}$ and $\boldsymbol{\varepsilon}$ is a vector of independent random errors such that $\varepsilon_i \sim N(0, \sigma_i^2)$. Suppose data were observed at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in \mathcal{D}$ and write $(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}_i) = (Y(\boldsymbol{s}_i), \boldsymbol{X}(\boldsymbol{s}_i), \boldsymbol{\beta}(\boldsymbol{s}_i))$ for $i = i, \ldots, n$.

By saying that the coefficient functions are smoothly varying, it is meant that they have continuous second derivatives. The method of maximum local likelihood is used to estimate the values of the coefficients at $\boldsymbol{s} \in \mathcal{D}$. For $\boldsymbol{s}_i$ in a neighborhood of $\boldsymbol{s}$, the values of the coefficients at $\boldsymbol{s}_i$ are approximated by a first order Taylor expansion:

$$\beta_j(\boldsymbol{s}_i) = \beta_j(\boldsymbol{s}) + (\boldsymbol{s}_i - \boldsymbol{s})^T \nabla \beta_j(\boldsymbol{s}) + O(\|\boldsymbol{s}_i - \boldsymbol{s}\|^2).$$

Observations nearest to $\boldsymbol{s}$ contain the most information about the values of $\boldsymbol{\beta}(\boldsymbol{s})$ and $\nabla\boldsymbol{\beta}(\boldsymbol{s})$. This intuition is formalized by the likelihood weights $K_h(\|\boldsymbol{s}_i - \boldsymbol{s}\|)$ for $i = 1, \ldots, n$, which arise from the kernel function $K_h(\cdot) = h^{-2}K(\cdot/h)$ and bandwidth parameter h. An example is the Epanechnikov kernel: $K(t) = 3/4(1 - t^2)$ if $|t| < 1$ and $K(t) = 0$ otherwise. Some notation will allow the likelihood to be written more simply. Denote the concatenation of the jth covariate with its directional derivatives by $\boldsymbol{\zeta}_j(\boldsymbol{s}) = (\beta_j(\boldsymbol{s}), \nabla_u\beta_j(\boldsymbol{s}), \nabla_v\beta_j(\boldsymbol{s}))^T$. Let $\boldsymbol{Z}_{\boldsymbol{s}} = (\boldsymbol{X}\ \boldsymbol{L}_{\boldsymbol{s}}\boldsymbol{X}\ \boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{X})$ where $\boldsymbol{L}_{\boldsymbol{s}}$ and $\boldsymbol{M}_{\boldsymbol{s}}$ are diagonal matrix with $\{\boldsymbol{L}_{\boldsymbol{s}}\}_{ii} = s_{i1} - s_1$ and $\{\boldsymbol{M}_{\boldsymbol{s}}\}_{ii} = s_{i2} - s_2$ for $i = 1, \ldots, n$, and denote the ith row of $\boldsymbol{Z}_{\boldsymbol{s}}$ as a column vector by $\{\boldsymbol{Z}_{\boldsymbol{s}}\}_i$. Then the local log-likelihood is

$$\ell_0(\boldsymbol{s}; h) = \sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|) \left[y_i - \{\boldsymbol{z}_{\boldsymbol{s}}\}_i^T \boldsymbol{\zeta}(\boldsymbol{s}) \right]. \tag{4.2}$$

Let $\boldsymbol{K}_{h,\boldsymbol{s}}$ be the diagonal matrix where $\{\boldsymbol{K}_{h,\boldsymbol{s}}\}_{ii} = K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)$. Then the local coefficient estimates that maximize the local likelihood (4.2), can be calculated by weighted least squares:

$$\bar{\boldsymbol{\zeta}}(\boldsymbol{s}) = \left(\boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{Z}_{\boldsymbol{s}} \right)^{-1} \boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{y}. \tag{4.3}$$

4.2.2 Local Adaptive Grouped Regularization

As the coefficients are smoothly varying functions of location, a coefficient may in general be equal to zero on some portion of the domain. At $\boldsymbol{s} \in \mathcal{D}$, assume that p_0 out of the p coefficient functions are nonzero and assume that these have indices $1, \ldots, p_0$ so that $\beta_j \neq 0$ if $j \leq p_0$ and $\beta_j(\boldsymbol{s}) = 0$ if $j > p_0$. Chapter 2 introduced the method of local adaptive grouped regularization (LAGR), a method of simultaneously estimating which coefficients are locally nonzero and estimating the local values of the nonzero coefficients. The local adaptive grouped regularization (LAGR) estimator at $\boldsymbol{s}$ minimizes the penalized weighted likelihood criterion

$$\ell(\boldsymbol{s}; h) \sum_{i=1}^{n} K_h(\|s_i - s\|) \left[y_i - \{\boldsymbol{z}_{\boldsymbol{s}}\}_i^T \boldsymbol{\zeta}(\boldsymbol{s}) \right]^2 + \sum_{j=1}^{p} \phi_j(\boldsymbol{s}; h_b) \, \|\boldsymbol{\zeta}_j(\boldsymbol{s})\| \tag{4.4}$$

where $\phi_j(\boldsymbol{s}; h_b) = \lambda_n \|\bar{\boldsymbol{\zeta}}_j(\boldsymbol{s}; h_b)\|^{-\gamma}$ for $j = 1, \ldots, p$, $\gamma > 1$, and $\bar{\boldsymbol{\zeta}}_j(\boldsymbol{s}; h_b)$ is the vector of unpenalized weighted likelihood estimates of the jth coefficient function and its first derivatives from (4.2). The maximizer of (4.4) is denoted by $\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s}; h_b)$.

4.2.3 Bandwidth Estimation

Bandwidth h is a global parameter in the VCR model, meaning that it appears in every local likelihood. The method of maximum likelihood is not suitable for estimating the bandwidth because changing the bandwidth changes the degrees of freedom of the model. As a result, estimating the parameters under two different bandwidths in a sense produces two different models, rather than two evaluations at different parameter values. For estimation of the bandwidth h, the adopted framework is that of minimizing the Kullback-Leibler (KL) divergence $I(f, g; h)$ between truth $f(\cdot)$ and the fitted model $g(\cdot; h)$, where the KL divergence is (Kullback and Leibler, 1951)

$$I(f, g; h) = \int \{\log f(x) - \log g(x; h)\} f(x) \mathrm{d}u.$$

The KL divergence generalizes the method of maximum likelihood to the context of model selection, but $I(f, g; h)$ depends on the unknown truth f, and therefore is not practically computable. An estimate of $E\{I(f, g; h)\}$ that can be computed is the Akaike Information Criterion (AIC). Minimizing the AIC generates in an estimate of the model that maximizes the expected density of a theoretical future data set where the covariates and observation locations are identical to the observed data (Akaike, 1973, 1974; Burnham and Anderson, 2002).

The estimated bandwidth is $\hat{h}$ that minimizes the AIC, defined as

$$\mathrm{AIC}(h) = -2\hat{\ell}(h) + 2\hat{\mathrm{df}}(h) \tag{4.5}$$

where

$$\hat{\ell}(h) = \sum_{i=1}^{n} \left\{ y_i - \{\boldsymbol{z}_s\}_i^T \hat{\boldsymbol{\zeta}}(\boldsymbol{s}_i; h) \right\}^2$$

is the maximized total log likelihood and

$$\hat{\mathrm{df}}(h) = \sum_{i=1}^{n} \left[\hat{\mathrm{df}}(\boldsymbol{s}_i; h) \left\{ \sum_{j=1}^{n} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|) \right\}^{-1} \right]$$

is the total degrees of freedom used in estimating the model with bandwidth h. The degrees of freedom for the local fit at $\boldsymbol{s}_i$, $\hat{\mathrm{df}}(\boldsymbol{s}_i; h)$, is the number of nonzero elements in the vector $\hat{\boldsymbol{\zeta}}(\boldsymbol{s}_i; h)$.

4.2.4 Weighted Likelihood Bootstrap

4.2.4.1 Bootstrap Resamples of Local Coefficient Estimates

The distribution of model parameters is simulated by the weighted likelihood bootstrap (WLB), which is described in detail in chapter 3 of this dissertation. The procedure is briefly reviewed here.

Each bootstrap resample is generated by the method of maximum weighted likelihood, as in (4.4), with the kernel likelihood weights multiplied by additional bootstrap weights $\boldsymbol{w}^*$. The vector of bootstrap weights is given by

$$\begin{aligned} w_i^* &= \sum_{k=1}^{n} d_k^* \quad \text{for } i = 1, \ldots, n, \quad \text{where} \\ D_k^* &\overset{iid}{\sim} \text{Dirichlet}(\mathbf{1}_n) \quad \text{for } k = 1, \ldots, n. \end{aligned} \tag{4.6}$$

For a particular bandwidth h, the bootstrap resample $\hat{\boldsymbol{\zeta}}^*(\boldsymbol{s}; h)$ is generated by maximizing the weighted likelihood

$$\ell^*(\boldsymbol{s}; h, \boldsymbol{w}^*) = \sum_{i=1}^{n} w_i^* K_h(\|s_i - s\|) \left[y_i - \{\boldsymbol{z_s}\}_i^T \boldsymbol{\zeta}(\boldsymbol{s})\right]^2 + \sum_{j=1}^{p} \phi_j^*(\boldsymbol{s}; h) \|\boldsymbol{\zeta}_j(\boldsymbol{s})\| \quad (4.7)$$

where $\phi_j^*(\boldsymbol{s}; h) = \lambda_n \|\bar{\boldsymbol{\zeta}}_j^*(\boldsymbol{s}; h)\|^{-\gamma}$, $\bar{\boldsymbol{\zeta}}^*(\boldsymbol{s}; h) = \left(\boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{W}^* \boldsymbol{Z}_{\boldsymbol{s}}\right)^{-1} \boldsymbol{Z}_{\boldsymbol{s}}^T \boldsymbol{K}_{h,\boldsymbol{s}} \boldsymbol{W}^* \boldsymbol{y}$ and $\boldsymbol{W}^{*b}$ is the diagonal matrix having $\{\boldsymbol{W}^{*b}\}_{ii} = w_i^*$.

It was shown in chapter 3 that the distribution of the WLB resamples for the true nonzero coefficients $\boldsymbol{\beta}_{(a)}$ is converging to

$$\{f(\boldsymbol{s}) n^2 h^2\}^{-1/2} \left[\hat{\boldsymbol{\beta}}_{(a)}^*(\boldsymbol{s}) - \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) - (2\kappa_0)^{-1} \kappa_2 h^2 \{\nabla_{11}^2 \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla_{22}^2 \boldsymbol{\beta}_{(a)}(\boldsymbol{s})\}\right] \xrightarrow{d} N\left(0, \kappa_0^{-2} \nu_0 \sigma^2(\boldsymbol{s}) \boldsymbol{\Psi}_{(a)}^{-1}(\boldsymbol{s})\right),$$

which is the same limiting distribution as the original estimates by LAGR, but as a faster rate.

Henceforth, the distribution of the WLB resamples $\hat{\boldsymbol{\zeta}}^*(\boldsymbol{s}; h)$ is written as the posterior distribution $\pi^*\{\boldsymbol{\zeta}(\boldsymbol{s}) | \boldsymbol{y}, h\}$. Implicit in this notation is the assumption of a uniform prior distribution on the local coefficients, but if there is nonuniform prior information about the distribution of the local coefficients, it can be incorporated via sampling-importance resampling (SIR) of the WLB resamples (Smith and Gelfand, 1992).

4.2.4.2 Bootstrap Resamples of Bandwidth Parameter

The WLB may also used to simulate the distribution of the bandwidth parameter that minimizes the AIC. Each WLB resample h^* is the bandwidth that minimizes a bootstrap AIC, where the bootstrap likelihood weights $\boldsymbol{w}^*$ are generated according to (4.6), and the bootstrap AIC, written as AIC*, is defined as

$$\mathrm{AIC}^*(h, \boldsymbol{w}^*) = -2\hat{\ell}^*(h, \boldsymbol{w}^*) + 2\hat{\mathrm{df}}^*(h, \boldsymbol{w}^*) \tag{4.8}$$

where

$$\hat{\ell}^*(h, \boldsymbol{w}^*) = \sum_{i=1}^{n} w_i^* \left\{ y_i - \{\boldsymbol{z}_s\}_i^T \hat{\boldsymbol{\zeta}}^*(\boldsymbol{s}_i; h) \right\}^2$$

is the maximized total log likelihood and

$$\hat{\mathrm{df}}^*(h, \boldsymbol{w}^*) = \sum_{i=1}^{n} \left[w_i^* \hat{\mathrm{df}}^*(\boldsymbol{s}_i; h) \left\{ \sum_{j=1}^{n} w_j^* K_h(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|) \right\}^{-1} \right]$$

is the total degrees of freedom used in estimating the model with bandwidth h and the degrees of freedom for the local fit at $\boldsymbol{s}_i$, $\hat{\mathrm{df}}^*(\boldsymbol{s}_i; h)$, is the number of nonzero elements in the vector $\hat{\boldsymbol{\zeta}}^*(\boldsymbol{s}_i; h)$.

4.3 A Hierarchical Model for Local Inference

The distribution of the local coefficient estimates as simulated by the WLB is conditional on the bandwidth parameter, so we write it as $\pi^*\{\boldsymbol{\beta}(\boldsymbol{s})|\boldsymbol{y}, h\}$. Unconditional confidence statements about the local coefficients must come from the marginal posterior distribution $\pi\{\boldsymbol{\beta}(\boldsymbol{s})|\boldsymbol{y}\}$. Bandwidth h is therefore a nuisance parameter that must be dealt with when making inferences about $\boldsymbol{\beta}(\boldsymbol{s})$. Estimation of $\hat{h}$ by AIC entails maximizing the likelihood over all parameters in the model. Thus, minimizing AIC jointly estimates the parameters $\{\boldsymbol{\beta}(\boldsymbol{s}), h\}$, which suggests the strategy of basing inference about h or $\boldsymbol{\beta}(\boldsymbol{s})$ on draws from a joint bootstrap distribution $\pi^*\{\boldsymbol{\beta}(\boldsymbol{s}), h\}$. However, samples from this distribution are expensive because the AIC is minimized by profiling over the possible range of bandwidths, which entails estimating the model several times over.

4.3.1 Posterior Hyperprior

A less expensive alternative is to frame the marginal posterior $\pi\{\boldsymbol{\beta}(\boldsymbol{s})|\boldsymbol{y}\}$ as a mixture of conditional distributions $\pi\{\boldsymbol{\beta}(\boldsymbol{s})|h, \boldsymbol{y}\}$, where the mixture weights are given by the posterior hyperprior $\pi(h|\boldsymbol{y})$ and the estimated posterior hyperprior $\hat{\pi}(h|\boldsymbol{y})$ is based on a relatively small number S of samples from the WLB bandwidth distribution $\pi^*(h|\boldsymbol{y})$. The marginal posterior distribution is then obtained by integrating the bandwidth out of the model. That is,

$$\pi^*\{\boldsymbol{\beta}(\boldsymbol{s})|\boldsymbol{y}\} = \int \pi^*\{\boldsymbol{\beta}(\boldsymbol{s})|h, \boldsymbol{y}\}\hat{\pi}(h|\boldsymbol{y})\mathrm{d}h. \tag{4.9}$$

Typically, the estimated posterior hyperprior $\hat{\pi}(h|\boldsymbol{y})$ might be calculated from the WLB resamples via a kernel density estimate, which is a nonparametric empirical Bayes procedure (Newton and Raftery, 1994). But since convergence of the WLB resamples is faster than that of the likelihood (by the results of chapter 3), an adjustment is required to the variance of the resamples, and it is easiest to apply this adjustment to a parametric distribution. Thus, we assume a particular parametric form for the bandwidth distribution. Since the bandwidth is a positive, continuous value with no upper bound, we adopt the gamma as a plausible distribution for the bandwidth.

Suppose that the posterior hyperprior $\pi(h|\boldsymbol{y})$ is a $\Gamma(a, b)$ distribution. Then, the mean and variance of the WLB resamples $h^{*1}, \ldots, h^{*S}$ are those of a $\Gamma(na, nb)$ distribution, which accounts for the faster convergence of the WLB resamples without affecting the expectation or the parametric form of the distribution. That is, if $H^* \sim \text{Gamma}(na, nb)$ and $H \sim \text{Gamma}(a, b)$, then $EH = EH^*$ and $\text{var}(H) = n \times \text{var}(H^*)$.

As the estimation of h^* by minimizing the $\mathrm{AIC}^*(h)$ involves model selection, the nonparametric delta method of Efron (2014) is used to estimate the variance of the distribution of H^*, rather than naively calculating the variance of the WLB resamples. By the method of moments, the parameter estimates are $\hat{a} = (h_\cdot^*)^2/(n\tilde{\sigma}_h^2)$ and $\hat{b} = h_\cdot^*/(n\tilde{\sigma}_h^2)$, where

$$
\begin{aligned}
h_\cdot^* &= S^{-1}\sum_{i=1}^{S} h_i^* \\
\tilde{\sigma}_h^2 &= \sum_{i=1}^{n} \hat{\mathrm{cov}}_i^2 \\
\hat{\mathrm{cov}}_i &= \sum_{j=1}^{S}(w_{ij} - w_{\cdot j})(h_j^* - h_\cdot^*)/S \\
w_{\cdot j} &= n^{-1}\sum_{i=1}^{n} w_{ij}.
\end{aligned}
$$

The result is that the estimated posterior hyperprior $\hat{\pi}(h|\boldsymbol{y}) = \Gamma(\hat{a}, \hat{b})$. This is an empirical Bayes procedure in that we have estimated a posterior hyperprior distribution from the likelihood, without an explicit prior distribution (Laird and Louis, 1987).

4.3.2 Conditional Posterior

Our strategy for sampling from the posterior distribution is to draw B bandwidth samples $h_1, \ldots, h_B$ from the distribution $\hat{\pi}(h|\boldsymbol{y})$ and use each to sample from the conditional posterior $\pi(\boldsymbol{\zeta}|h_b, \boldsymbol{y})$, which is simulated by the WLB resamples as $\pi^*(\boldsymbol{\zeta}|h_b, \boldsymbol{y})$ for $b = 1, \ldots, B$. A result of chapter 3 is that the WLB resamples have variance smaller than that of the target distribution by a factor of n, so an adjustment is required to simulate $\pi(\boldsymbol{\zeta}|h_b, \boldsymbol{y})$ by $\pi^*(\boldsymbol{\zeta}|h_b, \boldsymbol{y})$. We may expect, naively, to apply the

adjustment after all the samples have been drawn. However, because the bandwidth h appears in the expression for the bias of the local coefficient estimates, the naive approach ends up overestimating the variance of the local coefficients by inflating the variability of the bias.

The approach for adjusting the WLB resamples to the proper variance without altering the bias is to draw $N > 1$ resamples from each conditional distribution $\pi^*\{\boldsymbol{\zeta}(\boldsymbol{s})|h_b, \boldsymbol{y}\}$, subtract their mean, multiply the difference by $[n\{(2n-1)/(2n)\}\{N/(N+1)\}]^{1/2}$, and then add back the mean. That is, the adjusted WLB resamples $\boldsymbol{\zeta}^\dagger(\boldsymbol{s}; h_b)$ have the same expectation and variance as $\hat{\boldsymbol{\zeta}}(\boldsymbol{s}; h_b)$ where

$$\boldsymbol{\zeta}^\dagger(\boldsymbol{s}; h_b) = \boldsymbol{\zeta}^*_{\cdot}(\boldsymbol{s}; h_b) + [n\{(2n-1)/(2n)\}\{N/(N+1)\}]^{1/2} \{\boldsymbol{\zeta}^*(\boldsymbol{s}; h_b) - \boldsymbol{\zeta}^*_{\cdot}(\boldsymbol{s}; h_b)\}$$

$$\boldsymbol{\zeta}^*_{\cdot}(\boldsymbol{s}; h_b) = N^{-1}\boldsymbol{\zeta}^*(\boldsymbol{s}; h_b)$$

4.3.3 Marginal Posterior

The procedure for sampling from the simulated marginal posterior distribution $\pi\{\boldsymbol{\zeta}(\boldsymbol{s})|\boldsymbol{y}\}$ by drawing B bandwidth samples $h_1, \ldots, h_B$ from the estimated posterior hyperprior $\hat{\pi}(h|\boldsymbol{y})$. Then for each h_b $(b = 1, \ldots, B)$, N samples $\boldsymbol{\zeta}^{\dagger 1}(\boldsymbol{s}; h_b), \ldots, \boldsymbol{\zeta}^{\dagger N}(\boldsymbol{s}; h_b)$ are drawn from the simulated conditional posterior distribution $\pi^*\{\boldsymbol{\zeta}(\boldsymbol{s})|h, \boldsymbol{y}\}$. This procedure produces the $N \times B$ samples

$$\boldsymbol{Z}^\dagger(\boldsymbol{s}) = \left\{\boldsymbol{\zeta}^{\dagger 1}(\boldsymbol{s}; h_1), \ldots, \boldsymbol{\zeta}^{\dagger N}(\boldsymbol{s}; h_B)\right\},$$

which are interpreted as a simulation of the marginal posterior distribution.

For a given bandwidth, the local coefficient estimates by the WLB are converging to some Gaussian distribution. And the process of integrating the bandwidth out of the

hierarchy is to form a linear combination of the bandwidth-conditional distributions, so the limit of the marginal posterior distribution is also Gaussian. We need only find its expectation and variance, which are straightforward to calculate by conditional expectation.

First, the expectation is

$$\begin{aligned} E\ \boldsymbol{Z}^{\dagger}(\boldsymbol{s}) &= E\left[E\{\boldsymbol{\zeta}^{\dagger}(\boldsymbol{s};h)|h\}\right] \\ &= E\left[\boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + (2\kappa_0)^{-1}\kappa_2 h^2\{\nabla^2_{11}\boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla^2_{22}\boldsymbol{\beta}_{(a)}(\boldsymbol{s})\}\right] \\ &= \boldsymbol{\beta}_{(a)}(\boldsymbol{s}) - (2\kappa_0)^{-1}\kappa_2\hat{a}(\hat{a}+1)\hat{b}^{-2}\{\nabla^2_{11}\boldsymbol{\beta}_{(a)}(\boldsymbol{s}) + \nabla^2_{22}\boldsymbol{\beta}_{(a)}(\boldsymbol{s})\}. \end{aligned} \tag{4.10}$$

The variance is

$$\begin{aligned} \text{var}\ \boldsymbol{Z}^{\dagger}(\boldsymbol{s}) =& E\left[\text{var}\{\boldsymbol{\zeta}^{\dagger}(\boldsymbol{s};h)|h\}\right] + \text{var}\left[E\{\boldsymbol{\zeta}^{\dagger}(\boldsymbol{s};h)|h\}\right] \\ =& E\left[\{\kappa_0^2 nh^2 f(\boldsymbol{s})\}^{-1}\nu_0\sigma^2\Psi(\boldsymbol{s})\right] + \text{var}\left[\boldsymbol{\beta}(\boldsymbol{s}) + (2\kappa_0)^{-1}h^2\{\nabla^2_{uu}\boldsymbol{\beta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}(\boldsymbol{s})\}\right] \\ =& \{\kappa_0^2 nf(\boldsymbol{s})\}^{-1}\hat{b}^2(\hat{a}-2)^{-1}(\hat{a}-1)^{-1}\nu_0\sigma^2\Psi(\boldsymbol{s}) \\ &+ (2\kappa_0)^{-1}\hat{a}(4\hat{a}+6)(\hat{a}+1)\hat{b}^{-4}\{\nabla^2_{uu}\boldsymbol{\beta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\beta}(\boldsymbol{s})\}. \end{aligned} \tag{4.11}$$

These are the same expectation and variance as for the marginal posterior distribution of the non-bootstrap local coefficient estimates. So the empirical Bayes procedure simulates the marginal posterior distribution.

4.4 Simulation Study

4.4.1 Simulation Setup

A simulation illustrates the empirical Bayes procedure. The location parameter for the simulation is two-dimensional. Data were simulated at 196 locations on a uniform 14×14 grid covering the domain $[0,1] \times [0,1]$. The model for the simulated data is

$$y_i = \beta_0(s_i) + X_{1i}\beta_1(s_i) + X_{2i}\beta_2(s_i) + \varepsilon_i$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ and $\sigma = 0.5$. The coefficient functions $\beta_0(s)$, $\beta_1(s)$, and $\beta_2(s)$ are plotted in Figure 1. The coefficients X_1 and X_2 were simulated as zero mean Gaussian random fields with exponential covariance model having unit variance, range parameter 0.1, and nugget variance 0.2. Collinearity was induced in the covariates by multiplying the matrix $(X_1\ X_2)$ by the square root of the matrix having ones on the diagonal and $\rho = 0.1$ for the off-diagonal entries. The simulation was repeated for a total of 67 runs. In each run, 200 resamples were drawn from the estimated marginal posterior distribution of each local coefficient.

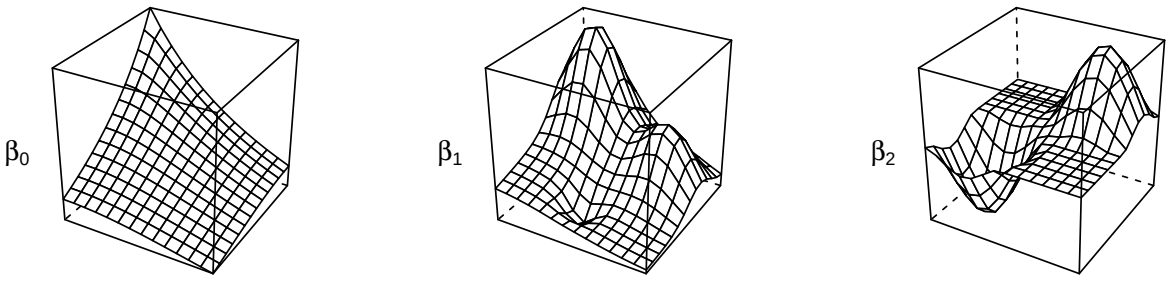

Figure 4.1: The coefficient functions used in the simulation model.

4.4.2 Bandwidth Estimation

For each run of the simulation, the WLB was used to generate $S = 10$ from the posterior hyperprior distribution for bandwidth h. Following the method of section 3.1, these resamples are assumed to be from a $\Gamma(na, nb)$ distribution. Each resample from the posterior hyperprior was estimated by sampling a WLB weight vector $\boldsymbol{w}^*$ and minimizing the $\text{AIC}^*(h, \boldsymbol{w}^*)$ with respect to h. An example AIC profile is plotted in Figure 3, in which the minimum was achieved at $\hat{h} = 0.315$. By the method of section 3.1, the resamples $h_1^*, \ldots, h_{10}^*$ were used to estimate $\hat{a}$ and $\hat{b}$, the parameters of the posterior hyperprior gamma distribution. For one run of the simulation, the WLB samples and the estimated $\hat{\pi}(h|\boldsymbol{y})$ are plotted in Figure 4.

4.4.3 Posterior Simulation

For each $b = 1, \ldots, B$, a bandwidth parameter h_b^* was sampled from the posterior hyperprior $\hat{\pi}(h|\boldsymbol{y})$, and $N = 2$ weight vectors $\boldsymbol{w}_1^{*b}$ and $\boldsymbol{w}_2^{*b}$ were sampled according to (6). Using the sampled bandwidth and likelihood weights, the local coefficients were estimated at each of the $n = 196$ grid points by minimizing (7). Then the variance adjustment method of section 3.2 was applied to the local coefficients.

4.4.4 Simulation Results

A quantile-based 90% confidence interval (CI) was generated for each local coefficient at each location in each run of the simulation. The endpoints of the CIs were the 5th and 95th quantiles of marginal posterior resamples. The results of the simulation are reported in terms of the frequency with which these 90% CIs cover the smoothed

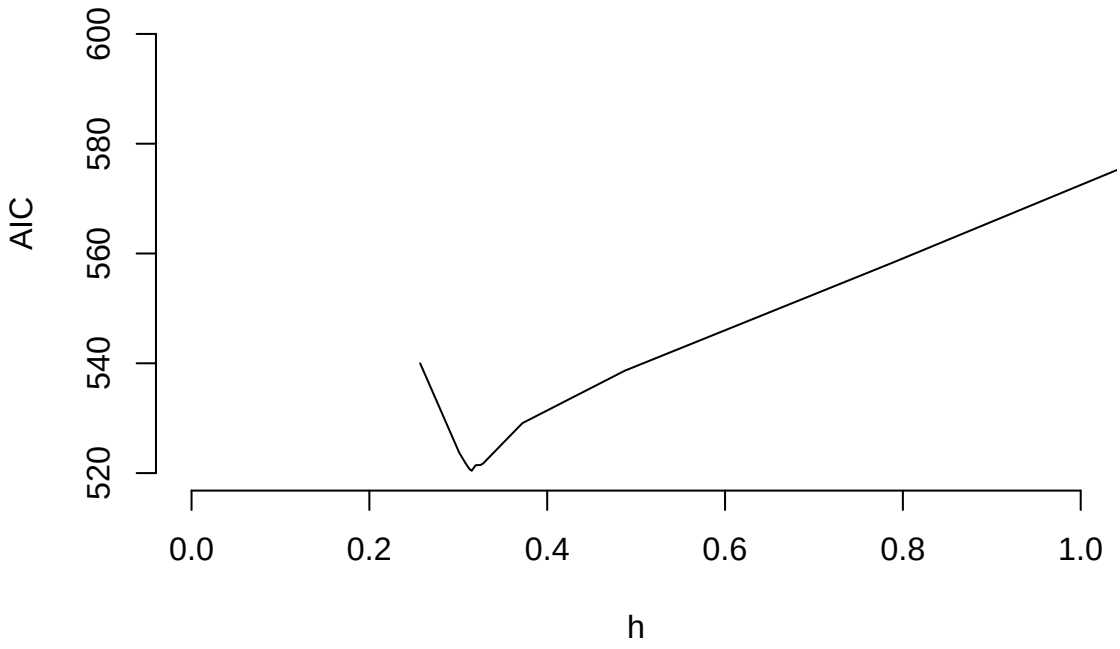

Figure 4.2: Trace of the AIC values calculated while estimating the bandwidth parameter, h, for the simulated data. The parameter estimate $\hat{h}$ is the value of h that minimizes the AIC. For the simulated data, $\hat{h} = 0.315$.

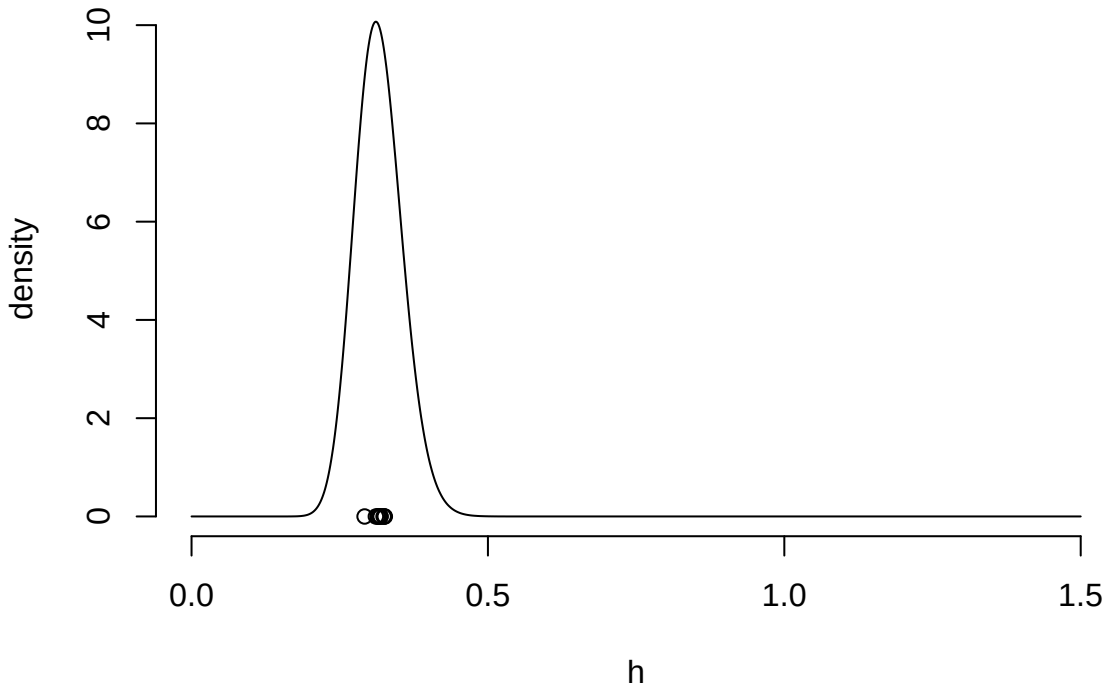

Figure 4.3: Density of the estimated posterior hyperprior distribution, $\hat{\pi}(h|\boldsymbol{y})$, for the simulated data. Also plotted (as circles) are the weighted likelihood bootstrap samples that were used to estimate the parameters of the distribution. The density is that of a $\Gamma(\hat{a}, \hat{b})$ distribution where $\hat{a} = (h_{\cdot}^*)^2/(n\tilde{\sigma}_h^2)$, $\hat{b} = h_{\cdot}^*/(n\tilde{\sigma}_h^2)$, $h_{\cdot}^* = S^{-1}\sum_{j=1}^{S} h_j^*$, and $\tilde{\sigma}_h^2$ is calculated by the nonparametric delta method as described in the manuscript. This is the distribution from which the bandwidth samples are drawn for the hierarchical mixture model.

true coefficients, as in chapter 3. In each run of the simulation, the mean of the posterior hyperrprior bandwidth distribution was used to smooth the true coefficients for the purpose of calculating CI coverage.

Figures 4, 5, and 6 summarize the local values of the coefficients estimates. Figure 4 is two panels: on the left is the actual spatially-varying intercept used to generate the data, and on the right is the coverage frequency of the 90% CIs based on local marginal posterior resamples. The cells of the right panel are all tinted orange, indicating undercoverage, because the nominal 90% CIs in fact achieved a mean coverage frequency of 0.66, across the entire spatial domain. The story is similar for Figures 5 and 6, the left-hand panels of which show the true spatially-varying coefficients $\beta_1(\cdot)$ and $\beta_2(\cdot)$. The middle panels show the coverage frequency of the local 90% CIs. For both $\beta_1(\cdot)$ and $\beta_2(\cdot)$ the CIs mostly undercovered their nominal confidence level, achieving a mean coverage frequency of 0.70 across the spatial domain for $\beta_1(\cdot)$ and for $\beta_2(\cdot)$. The rightmost panels of Figures 5 and 6 plot the frequency of exact zeroes among the WLB resamples from the marginal posteriors of $\beta_1(\cdot)$ and $\beta_2(\cdot)$, with orange indicating less than 50% exact zeroes and purple indicating greater than 50% exact zeroes.

There is no apparent pattern in the plots of coverage frequency for the local CIs for the intercept nor for $\beta_1(\cdot)$. Figure 6 seems to indicate that coverage of $\beta_2(\cdot)$ by the local CIs was greatest in the corners where the true coefficient was exactly zero, and near the "peak" and "trough" of the two bulges in $\beta_2(\cdot)$. Additionally, more than half of the resamples from the marginal posterior were less than zero in the corners where that coefficient's true value is exactly zero. There is a small region in the bottom right of the domain where $\beta_1(\cdot)$ is near zero and a large proportion of the resamples from

the marginal posterior were exactly zero. Undercoverage was a problem for the CIs in this simulation study, where mean coverage was less than was achieved under the same settings but without the empirical Bayes framework in chapter 3. It is encouraging that there is less spatial pattern to the coverage frequency in this study than was observed in chapter 3.

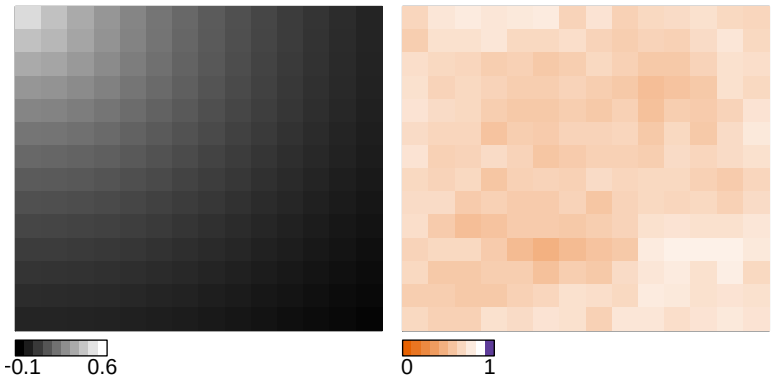

Figure 4.4: Intercept function $\beta_0(\cdot)$ (left panel) and the coverage of nominal 90% confidence intervals (CIs) at 196 locations on the domain $[0,1] \times [0,1]$. The middle panel illustrates the frequency with which the CIs cover the true local intercept, and the right panel illustrates the frequency with which the CIs cover the local value of the smoothed intercept function. Locations of undercoverage are colored orange and locations of overcoverage are colored purple.

4.5 Conclusions and Discussion

In this chapter, an empirical Bayes procedure is proposed for sampling from a posterior distribution of the local coefficients in a local regression model, where inference is to be marginal to the bandwidth parameter. The methods described require interpreting resamples from the WLB as draws from a posterior distribution. The same data are used to estimate the distributions of the nuisance bandwidth param-

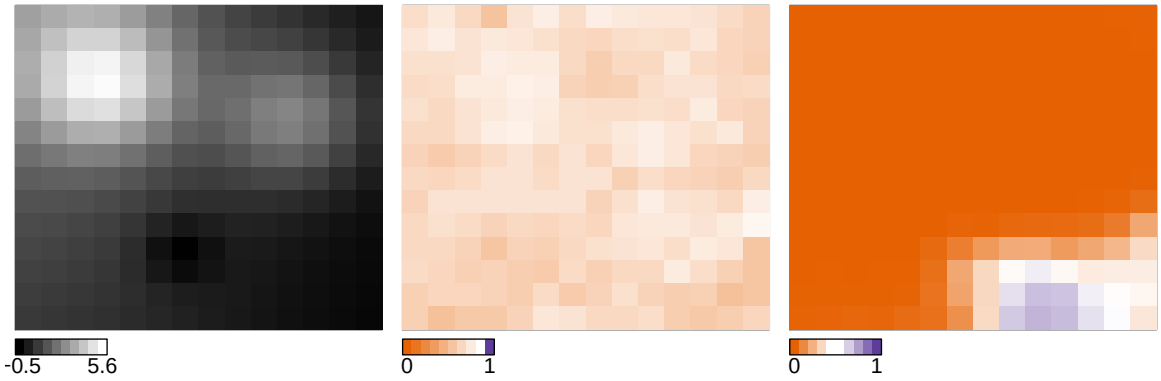

Figure 4.5: Intercept function $\beta_1(\cdot)$ (left panel) and the coverage of nominal 90% confidence intervals (CIs) at 196 locations on the domain $[0,1] \times [0,1]$. The middle panel illustrates the frequency with which the CIs cover the smoothed local coefficient. Locations of undercoverage are colored orange and locations of overcoverage are colored purple. The right panel shows how frequently the local coefficient was estimated as exactly zero.

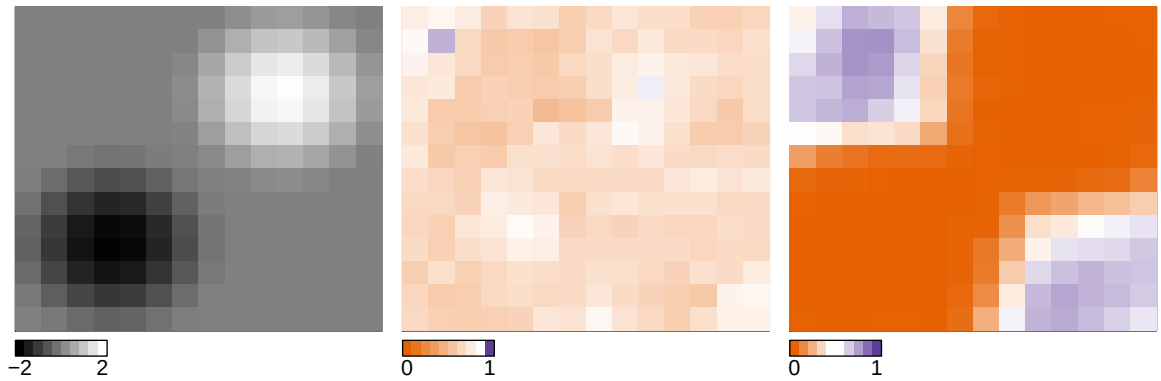

Figure 4.6: Intercept function $\beta_2(\cdot)$ (left panel) and the coverage of nominal 90% confidence intervals (CIs) at 196 locations on the domain $[0,1] \times [0,1]$. The middle panel illustrates the frequency with which the CIs cover the smoothed local coefficient. Locations of undercoverage are colored orange and locations of overcoverage are colored purple. The right panel shows how frequently the local coefficient was estimated as exactly zero.

eter and of the local coefficients, which are the parameters of interest. This process involves a parametric assumption about the distribution of the bandwidth parameter and two variance adjustments due to the way that likelihood weights are sampled for the WLB. A simulation study was used to illustrate the procedure, but the confidence intervals calculated during the simulation study covered the true coefficients with a frequency well below the nominal 90%. It remains unclear whether the undercoverage is due to random chance, the modest sample size, particular features of the simulation setup, or needed refinements to the methodology.

Chapter 5

Conclusion and Future Work

5.1 Future Work

Together, the chapters of this dissertation describe a framework for estimation and inference in a varying-coefficient regression (VCR) model where coefficients may be exactly zero on part of the domain. In chapter 2, the method of local adaptive grouped regularization (LAGR) was proposed for estimation of the local coefficients. The method of LAGR was shown to work in a generalized linear model context. Chapters 3 and 4 proposed a weighted likelihood bootstrap (WLB) for inference, and an empirical Bayes (EB) procedure for simulating the marginal posterior distribution of the local coefficient estimates. These chapters were developed in the context of regression models with a Gaussian response. The extension of the WLB and EB procedures to the setting of generalized linear models is a natural next step. From the experience of developing the current methodology, it seems that this will be a simple generalization.

An R package to estimate local coefficients in a VCR model has been developed

and is available for download from `github.com/wrbrooks/lagr`. While the code is complete and functional, it is a prototype that was developed for the purpose of running the simulations and examples in this dissertation. Once the last bugs are fixed and the documentation is finalized, a manuscript will be forthcoming that describes the software and its use.

A more prospective avenue of future research is to describe LAGR and the WLB in the framework of traditional spatial models, where nearby data are related through a covariance model. The cases considered in this dissertation assume that, conditional on the covariates, all the data are independent. It seems likely, though, that for some kinds of spatial error process, covariance among the random errors may be adequately modeled by the spatially-varying intercept. We anticipate that this will be a future area of research.

Bibliography

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

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.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, Halsinki, Finland.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Burnham, K. P. and D. R. Anderson (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer New York.

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.

Cao-Abad, R. (1991). Rates of convergence for the wild bootstrap in nonparametric regression. *Annals of Statistics 19*(4), 2226–2231.

Cao-Abad, R. and W. Gonzàlez-Manteiga (1993). Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics 2*, 378–388.

Chatterjee, A. and S. Lahiri (2011). Bootstraping lasso estimators. *Journal of the American Statistical Association 106*(494), 608–625.

Chatterjee, A. and S. N. Lahiri (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American mathematical society 138*(12), 4497–4509.

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. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics 7*(1), 1–26.

Efron, B. (1981). Nonparametric estimates of standard error: the jacknife, the bootstrap, and other methods. *Biometrika 68*(3), 589–599.

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

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association 109*(507), 991–1007.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics 32*(3), 928–961.

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

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics 9*(6), 1218–1228.

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.

Hall, P., E. R. Lee, and B. U. Park (2009). Bootstrap-based penalty choice for the LASSO, acheiving oracle performance. *Statistica Sinica 19*, 449–471.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*, 726–748.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Boston MA.

Härdle, W. and E. Mammen (1991). *Bootstrap methods in nonparametric regression.* Springer New York.

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.

Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika 48*(3/4), 419–426.

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

Kullback, S. and R. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*, 79–86.

Laird, N. M. and T. A. Louis (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association 82*(399), 739–750.

Liu, R. Y. (1988). Bootstrap procedures under some non-I.I.D. models. *Annals of Statistics 16*(4), 1696–1708.

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

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics 21*(1), 255–285.

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

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B 56*(1), 3–48.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics 9*(2), 319–337.

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.

Provost, S. B. and Y. Ho Cheong (2000). On the distribution of linear combinations of the components of a Dirichlet random vector. *Canadian Journal of Statistics 28*(2), 417–425.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics 9*, 130–134.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*. Oxford University Press.

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

Shang, Z. (2011). *Bayesian variable selection: theory and applications.* (Doctoral dissertation) Retrieved from ProQuest Dissertations and Theses. (Accession order number No. 3489080).

Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: a sampling-resampling perspective. *The American Statistician 46*(2), 84–88.

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.

Zhang, J. and M. Clayton (2011). Functional concurrent linear regression models for images. *Journal of Agricultural, Biological, and Environmental Statistics 16*, 105130.

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