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

XXXXXXX[1]

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 (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 by, for example, basis expansion (Wang et al., 2008), or local regression (Wang and Xia, 2009). Since the coefficients vary in a VCR model, it is natural to allow that the set of relevant covariates may vary over the domain.

This paper introduces local adaptive grouped regularization (LAGR) for local variable selection and estimation in 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). These methods mitigate the boundary effect by estimating the coefficients as local polynomials of odd degree (usually locally linear) (Hastie and Loader, 1993).

The adaptive Lasso (AL) is a regularization method that simultaneously selects covariates for a regression model and shrinks the coefficient estimates toward zero (Zou, 2006). The AL 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). The method of LAGR uses an adaptive group Lasso that achieves the oracle properties in the local regression setting, where convergence is slower than the typical $n^{1/2}$ rate.

The remainder of this paper is organized as follows. The kernel-based estimation of a VCR model is described in Section 2. The proposed LAGR technique for varying coefficient linear regression and its oracle properties are presented in Section 3. In Section 4, LAGR is applied to the Boston housing price dataset, followed by conclusions and discussion in Section 5. Technical proofs are given in the Appendix.

2. Varying Coefficient Regression

Throughout the paper, we work with the generalized linear model (GLM) (McCullagh and Nelder, 1989). The response of a GLM can come from any distribution in the exponential family, so the familiar linear regression model is included as a special case.

2.1. Varying coefficient model

The response $Y(s)$, covariates $\boldsymbol{X}(s)$, and coefficients $\boldsymbol{\beta}(s)$ in a VCR model are indexed by the location parameter s. Let d be the dimension of the location parameter (e.g., $d = 1$ for a coefficients that vary with time). Assume n observations of the response and the covariates at locations $s_1, \ldots, s_n$ and let $Y_i = Y(s_i)$, $\boldsymbol{X}_i = \boldsymbol{X}(s_i)$, and $\boldsymbol{\beta}_i = \boldsymbol{\beta}(s_i)$. The generalized linear model (GLM) with varying coefficients is written

$$
\begin{aligned}
E[Y(s)|\boldsymbol{X}(s)] &= \mu(s) \\
\eta(s) = g(\mu(s)) &= \boldsymbol{X}'(s)\boldsymbol{\beta}(s) \\
\text{var}[Y(s)|\boldsymbol{X}(s)] &= \phi V(\mu(s))
\end{aligned}
$$

where $g(\cdot)$, $V(\cdot)$, and ϕ are, respectively, the link function, variance function, and dispersion parameter of the response family. For simplicity of notation, assume the canonical link

function.

2.2. Local polynomial regression

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). Near estimation location s, the coefficient functions are well approximated by Taylor's expansion as locally linear functions of the location parameter

$$\boldsymbol{\beta}(t) = \boldsymbol{\beta}(s) + \nabla\boldsymbol{\beta}(s)(t-s) + o(|t-s|). \tag{1}$$

The locally linear approximation is implemented by augmenting the matrix $\boldsymbol{X}$ with interactions between the covariates and the location parameter. Let the augmented covariates be $\boldsymbol{Z}(s) = (\boldsymbol{X}_i \ \text{diag}\{s - s_i\}_{i=1}^n \cdot \boldsymbol{X}_i)$ and the augmented coefficients be $\boldsymbol{\zeta}(s) = (\boldsymbol{\beta}(s)^T, \ \nabla\boldsymbol{\beta}(s)^T)^T$.

2.3. Coefficient estimation via local quasi-likelihood

The quasi-likelihood $Q(\mu, Y)$ is an approximation to the log likelihood. Its derivative the quasi-score function is $q(\mu, Y) = (y-\mu)\{V(\mu)\}^{-1}$, which is a function of the linear predictor η through the link and variance functions. Taylor's approximation (1) is more accurate nearer to s, so a kernel function $K_h(\|s - s_i\|) = h^{-d}K(h^{-1}\|s - s_i\|)$ is used to weight the observations based on their distance from s and the bandwidth parameter h. For instance, the Epanechnikov kernel is:

$$K(x) = (3/4)(1-x^2) \text{ if } x < 1 \text{ and } 0 \text{ otherwise.} \tag{2}$$

Now, letting $\mu_i(s) = g^{-1}\left(\{\boldsymbol{Z}(s)\}_i^T\boldsymbol{\zeta}(s)\right)$ be the mean at s_i approximated via Taylor's expansion of $\boldsymbol{\beta}(t)$ at s, the local quasi-likelihood at $s \in \mathcal{D}$ is:

$$\ell\left(\boldsymbol{\zeta}(s)\right) = \sum_{i=1}^{n} K_h(\|s - s_i\|) Q\left(\mu_i(s), Y_i\right). \tag{3}$$

The local quasi-likelihood (3) is maximized to obtain an estimate $\tilde{\boldsymbol{\zeta}}(s)$ of the local coefficients at s. Since the quasi-likelihood is concave, the maximum is achieved where the local quasi-score is zero:

$$q\left(\tilde{\boldsymbol{\zeta}}(s)\right) = \sum_{i=1}^{n} K_h(\|s - s_i\|)\left(y_i - \tilde{\mu}_i(s)\right)\left\{V\left(\tilde{\mu}_i(s)\right)\right\}^{-1}\boldsymbol{z}_i = \mathbf{0}. \tag{4}$$

The asymptotic distribution of the local coefficients in a varying-coefficient GLM is (Cai et al., 2000):

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

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

So for any given s and under the risk-minimizing bandwidth $h = O\left(n^{-1/(4+d)}\right)$, the estimated local coefficients $\tilde{\boldsymbol{\beta}}(s) = \left(\tilde{\zeta}_1(s), \ldots, \tilde{\zeta}_p(s)\right)^T$ converge in probability at the optimal rate of $O\left(n^{-2/(4+d)}\right)$ and are asymptotically normally distributed. The bias of the local coefficient estimates is proportional to the second derivatives of the true coefficient functions.

3. Local Variable Selection with LAGR

Estimating the local coefficients by (4) has traditionally relied on *a priori* variable selection. Here we introduce a new method of penalized regression to simultaneously select covariates locally and estimate the corresponding local coefficients.

3.1. LAGR Penalized Local Likelihood

Each raw covariate is grouped with its covariate-by-location interactions. Thus, the jth group is $\boldsymbol{\zeta}_{(j)}(s) = \left(\beta_j(s), \ \nabla\beta_j(s)\right)^T$ for $j = 1, \ldots, p$. The proposed penalty is akin to the

adaptive group Lasso (Yuan and Lin, 2006; Wang and Leng, 2008). The method of LAGR entails maximizing the penalized local quasi-likelihood at s:

$$\mathcal{J}\left(\boldsymbol{\zeta}(s)\right) = \ell\left(\boldsymbol{\zeta}(s)\right) - \mathcal{P}\left(\boldsymbol{\zeta}(s)\right), \tag{5}$$

where $\ell\left(\boldsymbol{\zeta}(s)\right)$ is the local quasi-likelihood defined in (3) and $\mathcal{P}\left(\boldsymbol{\zeta}(s)\right) = \sum_{j=1}^{p} \phi_j(s)\|\boldsymbol{\zeta}_{(j)}(s)\|$ is a local adaptive grouped regularization (LAGR) penalty. The LAGR penalty for the jth group of coefficients at s is $\phi_j(s) = \lambda_n\|\tilde{\boldsymbol{\zeta}}_{(j)}(s)\|^{-\gamma}$, where $\lambda_n > 0$ is a local tuning parameter applied to all coefficient groups at s, $\tilde{\boldsymbol{\zeta}}_{(j)}(s)$ is a vector of unpenalized local coefficients for the jth covariate group from (4), and $\gamma > d/2$.

3.2. Oracle Property

At location s, let there be $p_0(s) < p$ covariates $\boldsymbol{X}_{(a)}(s)$ with nonzero local regression coefficients, denoted $\boldsymbol{\beta}_{(a)}(s) \neq \mathbf{0}$. Without loss of generality, assume the indices of these covariates are $1, \ldots, p_0(s)$. The remaining $p - p_0(s)$ covariates $\boldsymbol{X}_{(b)}(s)$ have coefficients equal to zero, denoted $\boldsymbol{\beta}_{(b)}(s) = \mathbf{0}$. Denote by $a_n = \max\left\{\phi_j(s), j \leq p_0(s)\right\}$ the largest penalty applied to a covariate group whose true coefficient norm is nonzero, and by $b_n = \min\left\{\phi_j(s), j > p_0(s)\right\}$ the smallest penalty applied to a covariate group whose true coefficient norm is zero. Let $\boldsymbol{Z}_{(k)}(s)$ be the augmented design matrix for covariate group k, and let $\boldsymbol{Z}_{(\text{-}k)}(s)$ be the augmented design matrix for all the data except group k. Similarly, let $\boldsymbol{\zeta}_{(k)}(s)$ be the augmented coefficients for covariate group k and $\boldsymbol{\zeta}_{(\text{-}k)}(s)$ be the augmented coefficients for all covariate groups except k. Let $\kappa_0 = \int_{\mathbb{R}^2} K(\|s\|)ds$, $\kappa_2 = \int_{\mathbb{R}^2}[(1,0)s]^2 K(\|s\|)ds = \int_{\mathbb{R}^2}[(0,1)s]^2 K(\|s\|)ds$, $\nu_0 = \int_{\mathbb{R}^2} K^2(\|s\|)ds$, and

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

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 s.

(C.3) The covariates $\boldsymbol{X}(s_1), \ldots, \boldsymbol{X}(s_n)$ are random vectors that are independent of $\varepsilon_1, \ldots, \varepsilon_n$. Also, the functions $g'''(s)$, $\nabla\boldsymbol{\Gamma}(s)$, $\nabla\boldsymbol{\Gamma}_{(a)}(s)$, $V(\mu(s, \boldsymbol{z}))$, and $V'(\mu(s, \boldsymbol{z}))$ are continuous at s.

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

(C.5) The observation locations $\{s_i\}$ 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

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

where $f(\cdot)$ is bounded away from zero on $\mathcal{S}$ and $r(\cdot)$ is any bounded continuous function.

(C.6) $h = O\left(n^{-1/(4+d)}\right)$.

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

(C.8) The function $\left(\partial^2/\partial\mu^2\right) Q\left(g^{-1}(\mu), y\right) < 0$ for $\mu \in \mathbb{R}$ and y in the range of the response.

Conditions (C.1)–(C.4) are common in the literature on nonparametric estimation, see conditions (5) and (6) of Cai et al. (2000). The existence of of $\boldsymbol{\Gamma}(\cdot)$ is needed for the existence of the limiting distribution of $\hat{\boldsymbol{\beta}}(s)$; its differentiability is used in the Taylor's expansions. Condition (C.4) is used when bounding the remainder term in the Taylor's expansions. Under condition (C.6), the coefficient estimates attain the optimal rate of convergence for bivariate nonparametric regression. Condition (C.7) is used in establishing the oracle properties.

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), $h^{-d/2}n^{-1/2}\lambda_n \to 0$ for all $j \leq p_0(s)$ and $h^{d/2}n^{-1/2}\lambda_n\|\boldsymbol{\zeta}_{(j)}(s)\|^{-\gamma} \to \infty$ for all $j > p_0(s)$. We require that λ_n satisfy both assumptions. Suppose $\lambda_n = n^{\alpha}$. Since $h = O\left(n^{-1/(4+d)}\right)$ and $\|\tilde{\boldsymbol{\zeta}}_{(p)}(s)\| = O\left(h^{-d/2}n^{-1/2}\right)$, it follows that $h^{-d/2}n^{-1/2}\lambda_n = O\left(n^{\alpha-2/(4+d)}\right)$ and $h^{d/2}n^{-1/2}\lambda_n\|\tilde{\boldsymbol{\zeta}}_{(p)}(s)\|^{-\gamma} = O\left(n^{\{2\gamma-2-d\}/(4+d)+\alpha}\right)$. Thus, $(2+d-2\gamma)/(4+d) < \alpha < 2/(4+d)$, which can only be satisfied for $\gamma > d/2$. Condition (C.8) assures that the local quasi-likelihood is concave and has a unique maximizer.

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

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

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

Theorem 2 (Selection consistency)**.** *Under (C.1)–(C.8), if $j > p_0(s)$,*

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

By Theorem 1, the LAGR estimates achieve the same asymptotic distribution as if the nonzero coefficients were known in advance. By Theorem 2, the true zero coefficients are dropped from the model with probability tending to one. Thus, the oracle properties for LAGR in the GLM setting are established. The technical proofs are given in the Appendix.

4. Data Example

The proposed 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 (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 $h = 0.26$ and the local tuning parameters were selected by the AIC.

The estimates of the local coefficients are plotted in the first five panels of Figure 1. 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 negative in others, indicating that the association of RAD with house prices is a local phenomenon. The bottom right panel of Figure 1 shows which covariates were estimated to have a nonzero coefficient in each tract. There were 471 tracts where 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.

5. 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.

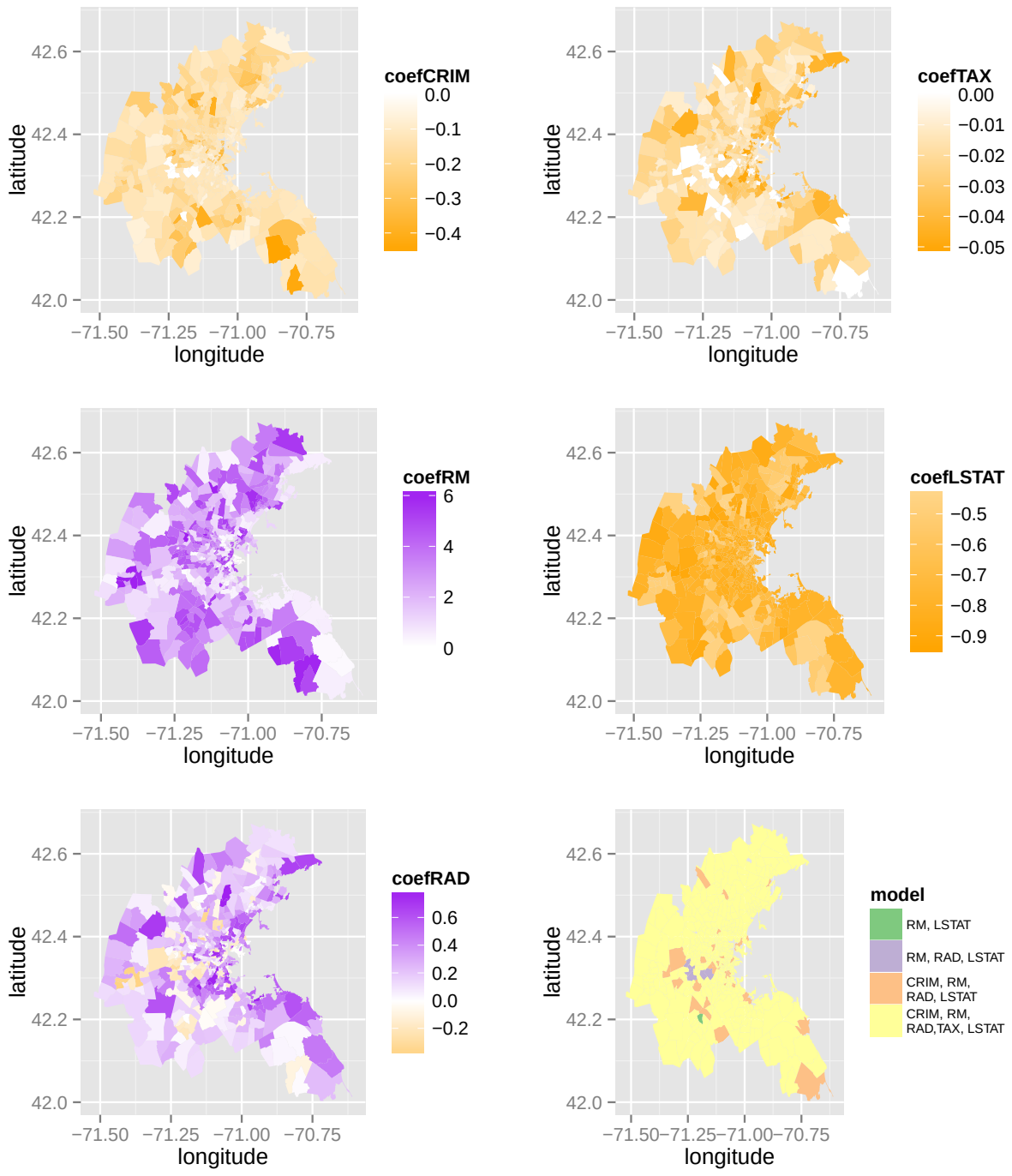

Figure 1: A varying coefficient regression model for the median house price in each census tract in Boston in 1970, estimated by local adaptive grouped regularization. In the left column are the estimated coefficients for covariates CRIM (per-capita crime rate), RM (mean number of rooms per house), and 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.

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.

References

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.

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

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

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

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.

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

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

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.

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

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

Appendix A. Proofs of Theorems 1–2

The results apply for a location parameter of arbitrary dimension, but the intended application was spatial analysis so the proofs proceed with $d = 2$.

Proof of Theorem 1

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 $q_1\left(\eta\left(s, \boldsymbol{z}\right), \mu\left(s, \boldsymbol{z}\right)\right) = \mathbf{0}$, and $q_2\left(\eta\left(s, \boldsymbol{z}\right), \mu\left(s, \boldsymbol{z}\right)\right) = -\rho\left(s, \boldsymbol{z}\right)$. Let

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

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

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

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

Define

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

and

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

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

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

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

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

the third term in (A.1) is $O_p\left(\alpha_n\right)$. The limiting behavior of the last term of (A.1) differs between the cases $j \le p_0(s)$ and $j > p_0(s)$. *Case $j \le p_0(s)$:* If $j \le p_0(s)$, then $n^{-1/2}\phi_j(s) \to$

$n^{-1/2}\lambda_n\|\boldsymbol{\zeta}_{(j)}(s)\|^{-\gamma}$ and $|\sqrt{n}\left\{\|\boldsymbol{\zeta}_{(j)}(s)+\alpha_n\boldsymbol{u}_{(j)}\|-\|\boldsymbol{\zeta}_{(j)}(s)\|\right\}| \leq h^{-1}\|\boldsymbol{u}_{(j)}\|$. Thus,

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

Case $j > p_0(s)$: If $j > p_0(s)$, then $\phi_j(s)\left(\|\boldsymbol{\zeta}_{(j)}(s)+\alpha_n\boldsymbol{u}_{(j)}\|-\|\boldsymbol{\zeta}_{(j)}(s)\|\right) = \phi_j(s)\alpha_n\|\boldsymbol{u}_{(j)}\|$. 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

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

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

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

if $\|\boldsymbol{u}_j\| = 0\ \forall j > p_0(s)$, 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). 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

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

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 $\left\{\boldsymbol{d}^T Var(\Omega_n)\boldsymbol{d}\right\}^{-1/2}\left\{\boldsymbol{d}^T\Omega_n - \boldsymbol{d}^T E\Omega_n\right\}$. By boundedness of $K(\cdot)$,

linearity of $q_1\left(\boldsymbol{Z}_i^T\boldsymbol{\zeta}(s), Y_i\right)$ in Y_i, and conditions (C.6) and (C.8), we have that

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

We observe that (A.2) 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. □

Proof of Theorem 2

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

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

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

Appendix B. Lemmas

Lemma 1.

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

and

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

Lemma 2.

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

and

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

The proofs are omitted.