Published in final edited form as:
Comput Stat Data Anal. 2015 March 1; 83: 236–250. doi:10.1016/j.csda.2014.10.004.

Domain selection for the varying coefficient model via local polynomial regression

Dehan Kong, **Howard Bondell**[1], and **Yichao Wu**
Department of Statistics, North Carolina State University

Abstract

In this article, we consider the varying coefficient model, which allows the relationship between the predictors and response to vary across the domain of interest, such as time. In applications, it is possible that certain predictors only affect the response in particular regions and not everywhere. This corresponds to identifying the domain where the varying coefficient is nonzero. Towards this goal, local polynomial smoothing and penalized regression are incorporated into one framework. Asymptotic properties of our penalized estimators are provided. Specifically, the estimators enjoy the oracle properties in the sense that they have the same bias and asymptotic variance as the local polynomial estimators as if the sparsity is known as a *priori*. The choice of appropriate bandwidth and computational algorithms are discussed. The proposed method is examined via simulations and a real data example.

Keywords

Bandwidth selection; Oracle properties; Penalized local polynomial fitting; SCAD

1. Introduction

The varying coefficient model (Cleveland, Grosse and Shyu, 1991; Hastie and Tibshirani, 1993) assumes that the covariate effect may vary depending on the value of an underlying variable, such as time. It has been used in a variety of applications, such as longitudinal data analysis, and is given by

$$Y=\mathbf{x}^{\top}\mathbf{a}(U)+\varepsilon, \quad (1)$$

where the predictor vector $\mathbf{x} = (x_1, \ldots, x_p)^{\top}$ represents p features, and correspondingly, $\mathbf{a}(U) = (a_1(U), \ldots, a_p(U))^{\top}$ denotes the effect of different features over the domain of the variable U. Y is the response we are interested in and ε denotes the random error satisfying $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2(U)$.

[1]Corresponding Author. Postal Address: 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203. bondell@stat.ncsu.edu. Phone: +1(919) 515-1914. Fax: +1(919) 515-1169.

The varying coefficient model has been extensively studied. Many methods have been proposed to estimate its parameters. The first group of estimation methods are based on local polynomial smoothing. Examples include, but are not limited to, Fan and Gijbels (1996); Wu, Chiang and Hoover (1998); Hoover, Rice, Wu and Yang (1998); Kauermann and Tutz (1999); Fan and Zhang (2008). The second is polynomial splines-based methods, such as Huang, Wu and Zhou (2002, 2004); Huang and Shen (2004) and references therein. The last group is based on smoothing splines as introduced by Hastie and Tibshirani (1993); Hoover, Rice, Wu and Yang (1998); Chiang, Rice and Wu (2001) and many others. In this paper, we not only consider estimation for the varying coefficient model, but also wish to identify the regions in the domain of U where predictors have an effect and the regions where they may not. This is similar, although different than variable selection, as selection methods attempt to decide whether a variable is active or not while our interest focuses on identifying regions.

For variable selection in a traditional linear model, various shrinkage methods have been developed. They include least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001), adaptive LASSO (Zou, 2006) and excessively others. Although the LASSO penalty gives sparse solutions, it leads to biased estimates for large coefficients due to the linearity of the L1 penalty. To remedy this bias issue, Fan and Li (2001) proposed the SCAD penalty and showed that the SCAD penalized estimator enjoys the oracle property in the sense that not only it can select the correct submodel consistently, but also the asymptotic covariance matrix of the estimator is the same as that of the ordinary least squares estimate as if the true subset model is known as a *priori*. To achieve the goal of variable selection for group variables, Yuan and Lin (2006) developed the group LASSO penalty which penalized coefficients as a group in situations such as a factor in analysis of variance. As with the LASSO, the group LASSO estimators do not enjoy the oracle property. As a remedy, Wang, Chen and Li (2007) proposed the group SCAD penalty, which again selects variables in a group manner.

For the varying coefficient model, some existing works focus on identifying the nonzero coefficient functions, which achieves component selection for the varying coefficient functions. However, each estimated coefficient function is either zero everywhere or nonzero everywhere. For example, Wang et al. (2008) considered the varying coefficient model under the framework of a B-spline basis and used the group SCAD to select the significant coefficient functions. Wang and Xia (2009) combined local constant regression and the group SCAD penalization together to select the components, while Leng (2009) directly applied the component selection and smoothing operator (Lin and Zhang, 2006).

In this paper, we consider a different problem: detecting the nonzero regions for each component of the varying coefficient functions. Specifically, we aim to estimate the nonzero domain of each $a_j(U)$, which corresponds to the regions where the jth component of $\mathbf{x}$ has an effect on Y. To this end, we incorporate local polynomial smoothing together with penalized regression. More specifically, we combine local linear smoothing and group SCAD shrinkage method into one framework, which estimates not only the function coefficients but also their nonzero regions. The proposed method involves two tuning parameters,

namely the bandwidth used in local polynomial smoothing and the shrinkage parameter used in the regularization method. We propose methods to select these two tuning parameters. Our theoretical results show that the resulting estimators have the same asymptotic bias and variance as the original local polynomial regression estimators.

The rest of paper is organized as follows. Section 2 reviews the local polynomial estimation for the varying coefficient model. Section 3 describes our methodology including the penalized estimation method and tuning procedure. Asymptotic properties are presented in Section 4. Simulation examples in Section 5 are used to evaluate the finite-sample performance of the proposed method. In Section 6, we apply our methods to the real data. We conclude with some discussions in Section 7.

2. Local polynomial regression for the varying coefficient model

Suppose we have independent and identically distributed (iid) samples $\{(U_i, \mathbf{x}_i^\top, Y_i)^\top, i=1, \ldots, n\}$ from the population $(U, \mathbf{x}^\top, Y)^\top$ satisfying model (1). As $\mathbf{a}(u)$ is a vector of unspecified functions, a smoothing method must be incorporated for its estimation. In this article, we adopt the local linear smoothing for this varying coefficient model (Fan and Zhang, 1999). For U in a small neighborhood of u, we can approximate the function $a_j(U)$, $1 \le j \le p$, by a linear function

$$a_j(U) \approx a_j(u) + a_j'(u)(U - u).$$

. For a fixed point u, denote $a_j(u)$ and $a'_j(u)$ by a_j and b_j, respectively, and denote their estimates by $\hat{a}_j$ and $\hat{b_j}$, which estimate the function $a_j(\cdot)$ and its derivative at the point u. Note that $(\hat{a}_j, \hat{b_j})$ $(1 \le j \le p)$ can be estimated via local polynomial regression by solving the following optimization problem:

$$\min_{\mathbf{a},\mathbf{b}} \sum_{i=1}^{n} \{Y_i - \mathbf{x}_i^\top \mathbf{a} - \mathbf{x}_i^\top \mathbf{b}(U_i - u)\}^2 (K_h(U_i - u)/K_h(0)), \quad (2)$$

where $\mathbf{a} = (a_1, \ldots, a_p)^\top$ and $\mathbf{b} = (b_1, \ldots, b_p)^\top$, $K_h(t) = K(t/h)/h$, and $K(t)$ is a kernel function. The parameter $h > 0$ is the bandwidth controlling the size of the local neighborhood. It implicitly controls the model complexity. Consequently it is essintial to choose an appropriate smoothing bandwidth in local polynomial regression. We will discuss how to select the bandwidth h in section 2.1.

The kernel function $K(\cdot)$ is a nonnegative symmetric density function satisfying $\int K(t)dt = 1$. There are numerous choices for the kernel function. Examples are Gaussian kernel $(K(t) = \exp(-t^2/2)/\sqrt{2\pi})$ and Epanechnikov kernel $(K(t) = 0.75(1 - t^2)_+)$ among many others. Typically, the estimates are not sensitive to the choice of the kernel function. In this paper, we use the Epanechnikov kernel, which leads to computational efficiency due to its bounded support.

Notice here that our loss function is slightly different from the loss function of the traditional local polynomial regression for the varying coefficient model (Fan and Zhang, 1999). We

have rescaled the original loss function by a term $K_h(0)$. For a fixed h, this change does not affect the estimates. However, this scaling is needed later to correctly balance the loss function and penalty term since $K_h(U_i - u) = K((U_i - h)/h)/h$, we include the term $K_h(0)$ to eliminate the effect of h so that $K_h(U_i - u)/K_h(0) = O(1)$.

Denote $\mathbf{a}_0 = (a_{01}, \ldots, a_{0p})^\top$ and $\mathbf{b}_0 = (b_{01}, \ldots, b_{0p})^\top$ to be the true values of the coefficient functions and their derivatives, respectively, and $\hat{\mathbf{a}} = (\hat{a}_{01}, \ldots, \hat{a}_{0p})^\top$ and $\hat{\mathbf{b}} = (\hat{b}_{01}, \ldots, \hat{b}_{0p})^\top$ as their corresponding local polynomial regression estimates. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ to be a $n \times p$ matrix. Further, denote $\gamma_j = (a_j, b_j)^\top$, $1 \le j \le p$ and $\gamma = (\gamma_1^\top, \ldots, \gamma_p^\top)^\top$ be a $2p$ dimensional vector. Define $\mathbf{U}_u = \mathrm{diag}(U_1 - u, \ldots, U_n - u)$, where $\mathrm{diag}(u_1, \ldots, u_n)$ denotes the matrix with $(u_1, \ldots, u_n)$ on the diagonal and zeros elsewhere. Let $\mathbf{x}_{(j)}$ be the jth column of $\mathbf{X}$ and x_{ij} be the ijth element of $\mathbf{X}$. Denote $\mathbf{\Gamma}_{uj} = (\mathbf{x}_{(j)}, \mathbf{U}_u\mathbf{x}_{(j)})$ for $1 \le j \le p$ and $\mathbf{\Gamma}_u = (\mathbf{\Gamma}_{u1}, \ldots, \mathbf{\Gamma}_{up})$ to be a $n \times 2p$ matrix. Define $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{W}_u = \mathrm{diag}(K_h(U_1 - u)/K_h(0), \ldots, K_h(U_n - u)/K_h(0))$. Using these notations, we can write (2) as

$$\min_{\gamma} (\boldsymbol{Y} - \mathbf{\Gamma}_u \boldsymbol{\gamma})^\top \mathbf{W}_u (\boldsymbol{Y} - \mathbf{\Gamma}_u \boldsymbol{\gamma}), \quad (3)$$

which has the formulation of a weighted least square problem.

2.1. Choosing the bandwidth h

The standard approach to choose the bandwidth is based on the trade-off between bias and variance. The most straightfoward one is the rule of thumb method, see Fan and Gijbels (1996) for details. It is fast in computation. However it highly depends on the asymptotic expansion of the bias and variance, and may not work well in small samples. Moreover, the optimal bandwidth is based on several unknown quantities, for which good estimates may be difficult to obtain. To overcome these deficiencies, we adopt the mean square error (MSE) tuning method. For a detailed review, see Fan and Zhang (1999) and Zhang and Lee (2000). This method uses information provided by a finite sample and hence carries more information about the finite sample. As a result it has the potential of selecting the bandwidth more accurately than other methods such as residual squares criterion (RSC) (Fan and Gijbels, 1995) and cross validation.

For a fixed smoothing bandwidth h, define $MSE(h)$ as

$$MSE(h) = E\{\mathbf{x}^\top \hat{\mathbf{a}}(U) - \mathbf{x}^\top \mathbf{a}(U)\}^2.$$

By direct calculation, we have

$$MSE(h) = E\{\mathbf{B}^\top(U)\mathbf{\Omega}(U)\mathbf{B}(U) + \mathrm{tr}(\mathbf{\Omega}(U)\mathbf{V}(U))\}, \quad (4)$$

where $\mathbf{B}(U) = \mathrm{Bias}(\hat{\mathbf{a}}(U))$, $\mathbf{\Omega}(U) = E(\mathbf{x}\mathbf{x}^\top | U)$ and $\mathbf{V}(U) = \mathrm{Cov}(\hat{\mathbf{a}}(U))$. The estimated $MSE(h)$ is given by

$$\hat{MSE}(h)=n^{-1}\sum_{i=1}^{n}\{\hat{\mathbf{B}}^{\top}(U_i)\hat{\mathbf{\Omega}}(U_i)\hat{\mathbf{B}}(U_i)+\mathrm{tr}(\hat{\mathbf{\Omega}}(U_i)\hat{\mathbf{V}}(U_i))\}.$$

which depends on the estimates $\hat{\mathbf{B}}(U)$, $\hat{\mathbf{\Omega}}(U)$ and $\hat{\mathbf{V}}(U)$ of $\mathbf{B}(U)$, $\mathbf{\Omega}(U)$ and $\mathbf{V}(U)$. A grid search can be applied to select the optimal bandwidth h which minimizes the $\hat{MSE}(h)$.

The estimation of $\mathbf{B}(U)$, $\mathbf{\Omega}(U)$ and $\mathbf{V}(U)$ was discussed in Fan and Zhang (2008), and we give a brief review here. For each given u, we have $\mathbf{\Omega}(u) = E(\mathbf{x}\mathbf{x}^{\top}|u)$, and we can estimate it by a kernel smoother

$$\hat{\mathbf{\Omega}}(u)=\frac{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\top}K_h(U_i-u)}{\sum_{i=1}^{n}K_h(U_i-u)}.$$

Fan and Zhang (2008) summarized the forms of the bias $\hat{\mathbf{B}}(u)$ and variance $\hat{\mathbf{V}}(U)$. Introduce the $p \times 2p$ matrix M, where the $(j, 2j - 1)$ $(1 \le j \le n)$ elements of $\mathbf{M}$ are 1, and the remaining elements are 0. The estimated bias is given by

$$\hat{\mathbf{B}}(u)=\hat{\mathrm{bias}}(\hat{\mathbf{a}}(u))=\mathbf{M}(\mathbf{\Gamma}_u^{\top}\mathbf{W}_u\mathbf{\Gamma}_u)^{-1}\mathbf{\Gamma}_u^{\top}\mathbf{W}_u\hat{\tau},$$

where the ith element of $\hat{\boldsymbol{\tau}}$ is

$$2^{-1}\mathbf{x}_i^{\top}\{\hat{\mathbf{a}}^{(2)}(u)(U_i-u)^2+3^{-1}\hat{\mathbf{a}}^{(3)}(u)(U_i-u)^3\}$$

and $\hat{\mathbf{a}}^{(2)}(u)$ and $\hat{\mathbf{a}}^{(3)}(u)$ denote some estimates of the second and third derivatives of function $\mathbf{a}(\cdot)$ at the point u. These two unknown quantities $\hat{\mathbf{a}}^{(2)}(u)$ and $\hat{\mathbf{a}}^{(3)}(u)$ can be obtained by a local cubic fitting with an appropriate pilot bandwidth h_*.

The estimated variance is given by

$$\hat{\mathbf{V}}(U)=\hat{\mathrm{cov}}(\hat{\mathbf{a}}(u))=\mathbf{M}(\mathbf{\Gamma}_u^{\top}\mathbf{W}_u\mathbf{\Gamma}_u)^{-1}(\mathbf{\Gamma}_u^{\top}\mathbf{W}_u^2\mathbf{\Gamma}_u)(\mathbf{\Gamma}_u^{\top}\mathbf{W}_u\mathbf{\Gamma}_u)^{-1}\mathbf{M}^{\top}\hat{\sigma}^2(u).$$

The estimator $\hat{\sigma^2}(u)$ can be obtained as a by product when we use local cubic fitting with a pilot bandwidth h_*. Denote $\mathbf{\Gamma}_{uj}^*=(\mathbf{x}_{(j)}, \mathbf{U}_u\mathbf{x}_{(j)}, \mathbf{U}_u^2\mathbf{x}_{(j)}, \mathbf{U}_u^3\mathbf{x}_{(j)})$ and $\mathbf{\Gamma}_u^*=(\mathbf{\Gamma}_{u1}^*, \ldots, \mathbf{\Gamma}_{up}^*)$. We have

$$\hat{\sigma}^2(u)=\frac{\boldsymbol{Y}^{\top}\{\mathbf{W}_u^*-\mathbf{W}_u^*\mathbf{\Gamma}_u^*(\mathbf{\Gamma}_u^{*\top}\mathbf{W}_u^*\mathbf{\Gamma}_u^*)^{-1}\mathbf{\Gamma}_u^{*\top}\mathbf{W}_u^*\}\boldsymbol{Y}}{\mathrm{tr}\{\mathbf{W}_u^*-(\mathbf{\Gamma}_u^{*\top}\mathbf{W}_u^*\mathbf{\Gamma}_u^*)^{-1}(\mathbf{\Gamma}_u^{*\top}\mathbf{W}_u^{*2}\mathbf{\Gamma}_u^*)\}},$$

where $\mathbf{W}_u^*$ is $\mathbf{W}_u$ with h replaced by h_*.

The pilot bandwidth h_* is used for a pilot local cubic fitting and Fan and Gijbels (1995) introduced the RSC to select it. However, they only studied the univariate case, which applies to the varying coefficient model with one component. As we are considering the varying coefficient model with several components, their method is not applicable. Instead, we use a five-fold cross validation to select an optimal smoothing bandwidth for the pilot fitting as in Fan and Gijbels (1992) for example. Specifically, we divide the data into five roughly equal parts, denoted as $\{(U_i, \mathbf{x}_i^\top, Y_i^\top)^\top, i \in S(j)\}$ for $j = 1, 2, \cdots, 5$, where $S(j)$ is the set of subject indices corresponding to the jth part. For each j, we treat $\{(U_i, \mathbf{x}_i^\top, Y_i^\top)^\top, i \in S(j)\}$ as the validation data set, and the remaining four parts of data as the training data set. For a candidate bandwidth h and each $i \in S(j)$, we apply a local polynomial fitting to the training data set to estimate $\mathbf{a}(u)$ at $u = U_i$ by solving the minimization problem similar to (2). After we get the estimates $\hat{\mathbf{a}}(U_i)$ for all $i \in S(j)$, we can get the corresponding prediction $\hat{Y}_i = \mathbf{x}_i^\top \hat{\mathbf{a}}(U_i)$. The cross validation error corresponding to a fixed h is defined as

$$CV(h) = \sum_{j=1}^{5} \sum_{i \in S(j)} (Y_i - \hat{Y}_i)^2.$$

We select the pilot bandwidth h_* by minimizing the cross validation error.

3. Penalized local polynomial regression estimation

In practice, it can be of certain interest to detect the nonzero region of each function component of the vector $\mathbf{a}$. To achieve this goal, shrinkage methods can be applied. Notice that $a'_j(u) = 0$ for $u \in [c_1, c_2]$ as long as $a_j(u) = 0$ for $u \in [c_1, c_2]$. Consequently if the function estimates are zero over certain regions, the corresponding derivative estimates should also be zero. Consequently, we treat (a_j, b_j) as a group and do penalization together for each $1 \le j \le p$. To achieve variable selection as well as accurate estimation, a group SCAD penalty (Wang, Chen and Li, 2007) is then added to (2) to get sparse solutions for $\mathbf{a}$ and $\mathbf{b}$.

Recall that we need to solve the minimization problem (3). It can be rewritten as a least square problem with new data $(\mathbf{W}_u^{1/2}\boldsymbol{Y}, \mathbf{W}_u^{1/2}\boldsymbol{\Gamma}_u)$:

$$\min_{\boldsymbol{\gamma}} (\mathbf{W}_u^{1/2}\boldsymbol{Y} - \mathbf{W}_u^{1/2}\boldsymbol{\Gamma}_u\boldsymbol{\gamma})^\top (\mathbf{W}_u^{1/2}\boldsymbol{Y} - \mathbf{W}_u^{1/2}\boldsymbol{\Gamma}_u\boldsymbol{\gamma}).$$

For a traditional linear model, the covariates are typically adjusted to a same scale before adding the penalty. We apply a similar procedure by rescaling each column of the covariate $\mathbf{W}_u^{1/2}\boldsymbol{\Gamma}_u$ to have a same variance. Denote s_j to be the standard deviation of the pseudo covariates $x_{ij}(K_h(U_i - u)/K_h(0))^{1/2}$ $(1 \le i \le n)$ and r_j that of $x_{ij}(U_i - u)(K_h(U_i - u)/K_h(0))^{1/2}$ $(1 \le i \le n)$. In other words, $(s_1, r_1, s_2, r_2, \ldots, s_p, r_p)^\top$ are the standard deviations for each column of the pseudo covariate $\mathbf{W}_u^{1/2}\boldsymbol{\Gamma}_u$. We can standardize the covariates first and apply

penalization thereafter. Yet, a same effect can be achieved by keeping the covariates unchanged but adjusting the penalty correspondingly. This is the procedure that we are going to adopt in this article, as detailed shortly.

In most situations, such a rescaling is only needed for a better finite sample performance. However, for the varying coefficient model, the convergence rates for function estimate, $\hat{a}_j$, and derivative estimate, $\hat{b_j}$, are at different orders. Hence this rescaling is also necessary theoretically. We have shown in Lemma 1 in Appendix A that $s_j = O_P(1)$ and $r_j = O_P(h)$. Based on these results, s_j and r_j can properly adjust the effect of the different rates of convergence of the function and derivative estimates as presented next.

For the local polynomial regression, it is no longer appropriate to use n as the sample size because not all observations contribute equally to the estimation at any given location. In fact, some will contribute nothing if the kernel has a bounded support. Thus motivated, we define the effective sample size as $m=\sum_{i=1}^{n} K_h(U_i-u)/K_h(0)$. The penalized local polynomial regression estimates $(\hat{\mathbf{a}}_\lambda^\top, \hat{\mathbf{b}}_\lambda^\top)^\top$ can be obtained by solving

$$\min_{\mathbf{a},\mathbf{b}} \left[\sum_{i=1}^{n} \{Y_i - \mathbf{x}_i^\top \mathbf{a} - \mathbf{x}_i^\top \mathbf{b}(U_i-u)\}^2 (K_h(U_i-u)/K_h(0)) + m\sum_{j=1}^{p} P_\lambda(\sqrt{s_j^2 a_j^2 + r_j^2 b_j^2}) \right], \quad (5)$$

where the SCAD penalty function $P\lambda(\cdot)$ (Fan and Li, 2001) is symmetric with $P\lambda(0) = 0$ and its first order derivative defined as

$$P_\lambda'(t) = \lambda\{I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\}$$

for $t > 0$ and some constant $a > 2$. In this paper, we use $a = 3.7$ as suggested by Fan and Li (2001).

For any point u in the domain of U, we can obtain the estimate of $\mathbf{a}(u)$ using our penalized local polynomial regression method. To detect the nonzero regions for any component of the varying coefficient functions, a set of dense grid is chosen over the whole domain of U, say $(u_1, \ldots, u_N)$. We obtain the estimates of $\mathbf{a}(u)$ on these grid points first. If the estimates of a certain component function are zero on a certain number of consecutive grid points, for instance, say estimates of $a_j(u)$ are zero on grid points $\{u_{l1}, u_{l_1+1}, \ldots, u_{l_2}\}$, we claim the estimate of the function $a_j(u)$ is zero over the domain $[u_{l_1}, u_{l_2}]$.

To present details of our algorithms and regularization parameter selection for the penalization, we next introduce some notations. Denote $\boldsymbol{\beta}_j = (s_j a_j, r_j b_j)^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$, and $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$. Denote $\hat{\beta} = (\hat{\beta}_1^\top, \ldots, \hat{\beta}_p^\top)^\top$ to be the local polynomial estimates of $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^\top, \ldots, \boldsymbol{\beta}_{0p}^\top)^\top$, and $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda 1}^\top, \ldots, \hat{\beta}_{\lambda p}^\top)^\top$ to be the penalized local polynomial estimates when the regularization parameter is λ.

3.1. Algorithms

We discuss how to solve (5) in this subsection. For the SCAD penalization problem, Fan and Li (2001) proposed the local quadratic approximation (LQA) algorithm to optimize the penalized loss function. With LQA, the optimization problem can be solved using a modified Newton-Raphson algorithm. The LQA estimator, however, cannot achieve sparse solutions directly. A thresholding has to be applied to shrink small coefficients to zero. To remedy this issue, Zou and Li (2008) proposed a new approximation method based on local linear approximation (LLA). The advantange of the LLA algorithm is that it inherits the computational efficiency of LASSO and also produces sparse solutions. Denote $\boldsymbol{\beta}_j^{(k)}=(s_j a_j^{(k)}, r_j b_j^{(k)})^\top$. Given the estimate $\{\hat{\beta}_j^{(k)}, j=1,\ldots,p\}$ at the kth iteration, we solve

$$\min_{a,b}\left[(\sum_{i=1}^{n}\{Y_i-\mathbf{x}_i^\top\mathbf{a}-\mathbf{x}_i^\top\mathbf{b}(U_i-u)\}^2(K_h(U_i-u)/K_h(0))+m\sum_{j=1}^{p}w_j^{(k)}\|\boldsymbol{\beta}_j\|\right],$$

to get updated estimate $\{\hat{\beta}_j^{(k+1)}, j=1,\ldots,p\}$, where $w_j^{(k)}=|P_\lambda'(\|\hat{\beta}_j^{(k)}\|)|$ with $\|\cdot\|$ denoting the l_2-norm of the vector. Repeat the iterations until convergence, and the estimate at the convergence is defined as the LLA estimate. The initial value of $\hat{\beta}_j^{(0)}$ can be chosen as the unpenalized local polynomial estimates. Based on our limited numerical experience, one step estimates already perform very competitively and it is not necessary to iterate further. See Zou and Li (2008) for similar discussions. Consequently, the one step estimate is adopted due to its computational efficiency.

3.2. Tuning of the regularization parameter

While tuning the regularization parameter λ, we adopt the Bayesian information criterion (BIC). Let $\hat{\mathbf{a}}_\lambda$ and $\hat{\mathbf{b}_\lambda}$ be the solutions of the optimization problem (5) for a fixed λ. The BIC is given by

$$BIC(\lambda)=m\log\left(\frac{RSS(\lambda)}{m}\right)+(\log m)\times df,$$

where $RSS(\lambda)=\sum_{i=1}^{n}\{Y_i-\mathbf{x}_i^\top\hat{\mathbf{a}}_\lambda-\mathbf{x}_i^\top\hat{\mathbf{b}}_\lambda(U_i-u)\}^2(K_h(U_i-u)/K_h(0))$. The degrees of freedom (df) are given as $\sum_{j=1}^{p}I(\|\hat{\beta}_{\lambda j}\|>0)+\sum_{j=1}^{p}\|\hat{\beta}_{\lambda j}\|(d_j-1)/\|\hat{\beta}_j\|$, where d_j = 2 as we use local linear polynomial regression, see Yuan and Lin (2006).

4. Asymptotic properties

Without loss of generality, we assume that only the first 2s components of β are nonzero. Denote $\boldsymbol{\beta}_N=(\boldsymbol{\beta}_1^\top,\ldots,\boldsymbol{\beta}_s^\top)^\top$ and $\boldsymbol{\beta}_Z=(\boldsymbol{\beta}_{s+1}^\top,\ldots,\boldsymbol{\beta}_p^\top)^\top$. Denote $\boldsymbol{\Gamma}_{huj}$ = $(\mathbf{x}_{(j)}/s_j, \mathbf{U}_u\mathbf{x}_{(j)}/r_j)$ for 1 ≤ j ≤ p and $\boldsymbol{\Gamma}_{hu}$ = $(\boldsymbol{\Gamma}_{hu1}, \ldots, \boldsymbol{\Gamma}_{hup})$. Recall that $m=\sum_{i=1}^{n}K_h(U_i-u)/K_h(0)$, which is the effective sample size. Our objective function can be written as

$$Q(\boldsymbol{\beta})=(\boldsymbol{Y}-\boldsymbol{\Gamma}_{hu}\boldsymbol{\beta})^{\top}\mathbf{W}_u(\boldsymbol{Y}-\boldsymbol{\Gamma}_{hu}\boldsymbol{\beta})+m\sum_{j=1}^{p}P_{\lambda_n}(\|\boldsymbol{\beta}_j\|).$$

Denote $b_n = (nh)^{-1/2}$. We first state the following conditions:

Conditions (A)

(A1) The bandwidth satisfies $nh \to \infty$ and $n^{1/7}h \to 0$.

(A2) $a_n^2 nh \to 0$, where $a_n=\max_{1\le j\le s}P'_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)$.

Condition (A1) is a condition on the bandwidth of the local polynomial regression, which guarantees the bias to dominate the variance while estimating the varying coefficient functions and their derivatives. Condition (A2) is a condition on the penalty function and the strength of the true signals, which can be equivallently written as $a_n = o(b_n)$.

For the SCAD penalty function, we have $\max_{1\le j\le s}|P''_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)| \to 0$ when $n \to \infty$ and $\liminf_{n\to\infty}\liminf_{\theta\to 0^+}P'_{\lambda_n}(\theta)/\lambda_n>0$. Moreover, as $a_n \le \lambda_n$, by condition (A2), we have $\lambda_n^2 nh \to 0$, which indicates $b_n = o(\lambda_n)$. These results will be used in the proof of our paper.

Under Conditions (A), if $\lambda_n \to 0$ as $n \to \infty$, we have the following theorems and corollary:
Theorem 1. *Our penalized local polynomial estimate* $\hat{\boldsymbol{\beta}}_{\lambda}$ *satisfies that* $\|\hat{\boldsymbol{\beta}}_{\lambda}-\boldsymbol{\beta}_0\| = O_P(b_n)$.

Theorem 1 gives the consistency rate for our penalized estimate. By theorem 1 and Lemma 1 in Appendix, we have $\hat{a}_{\lambda j} - a_{j0} = O_P(b_n)$ and $\hat{b_{\lambda j}} - b_{j0} = O_P(b_n/h)$ for any $1 \le j \le p$, which indicates that the rate of consistency for our penalized estimates of the varying coefficient functions is b_n while that of the derivative estimates is b_n/n.

Theorem 2. *With probability tending to 1, for any* $\boldsymbol{\beta}_N$ *satisfying* $\|\boldsymbol{\beta}_N - \boldsymbol{\beta}_{0N}\| = O_P(b_n)$ *and any constant* $C > 0$, *we have*

$$Q(\boldsymbol{\beta}_N^{\top},\mathbf{0})=\min_{\|\boldsymbol{\beta}_Z\|\le Cb_n}Q(\boldsymbol{\beta}_N^{\top},\boldsymbol{\beta}_Z^{\top}).$$

Theorem 2 indicates that we can capture the true zero components with probability going to 1.

Denote $\boldsymbol{\Sigma}_s$ the upper left $2s \times 2s$ submatrix of $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is defined in equation (A.1) in Appendix A. Let $T(\boldsymbol{\beta}_l)$ be a matrix function

$$T(\boldsymbol{\beta}_l)=\frac{\partial\{P'_{\lambda_n}(\|\boldsymbol{\beta}_l\|)\frac{\boldsymbol{\beta}_l}{\|\boldsymbol{\beta}_l\|}\}}{\partial\boldsymbol{\beta}_l^{\top}}=P''_{\lambda_n}(\|\boldsymbol{\beta}_l\|)\frac{\boldsymbol{\beta}_l\boldsymbol{\beta}_l^{\top}}{\|\boldsymbol{\beta}_l\|^2}+P'_{\lambda_n}(\|\boldsymbol{\beta}_l\|)\frac{\boldsymbol{\beta}_l\boldsymbol{\beta}_l^{\top}}{\|\boldsymbol{\beta}_l\|^3}+\frac{P'_{\lambda_n}(\|\boldsymbol{\beta}_l\|)}{\|\boldsymbol{\beta}_l\|}\mathbf{I}_2,$$

where $\boldsymbol{\beta}_l$ is a two dimensional vector and $\mathbf{I}_2$ is the 2×2 identity matrix.

Denote $\mathbf{H} = \mathrm{diag}(T(\boldsymbol{\beta}_{01}), \ldots, T(\boldsymbol{\beta}_{0s}))$, which is a $2s \times 2s$ matrix, and $\mathbf{d}=(P'_{\lambda_n}(\|\boldsymbol{\beta}_{01}\|)\boldsymbol{\beta}_{01}^{\top}/\|\boldsymbol{\beta}_{01}\|, \ldots, P'_{\lambda_n}(\|\boldsymbol{\beta}_{0s}\|)\boldsymbol{\beta}_{0s}^{\top}/\|\boldsymbol{\beta}_{0s}\|)^{\top}$ which is a $2s$ dimensional vector.

Theorem 3. *Suppose* $\hat{\boldsymbol{\beta}}_N$ *is the local polynomial estimator of the nonzero components, and* $\hat{\boldsymbol{\beta}}_{\lambda N}$ *is the penalized local polynomial estimator for the nonzero components. We have*

$$\hat{\beta}_N - \boldsymbol{\beta}_{0N} = \boldsymbol{\Sigma}_s^{-1}\left(\frac{m}{2n}\mathbf{H} + \boldsymbol{\Sigma}_s\right)\left\{\left(\hat{\beta}_{\lambda N} - \boldsymbol{\beta}_{0N}\right) + \left(\frac{m}{2n}\mathbf{H} + \boldsymbol{\Sigma}_s\right)^{-1}\frac{m}{2n}\mathbf{d}\right\} + op(n^{-1}).$$

Corollary 1. *From theorem 3, we can get*

$$(cov(\hat{\mathbf{a}}_N(u)))^{-1/2}\{\hat{\mathbf{a}}_{\lambda N}(u) - \mathbf{a}_{0N}(u) - bias(\hat{\mathbf{a}}_N(u))\} \xrightarrow{d} N(0, I_s),$$

where $\hat{\mathbf{a}}_{\lambda N}(u)$ *denotes our penalized local polynomial estimates for the nonzero functions.*

This corollary shows that when $n \to \infty$, asymptotically $\hat{\mathbf{a}}_N - \mathbf{a}_{0N}$ and $\hat{\mathbf{a}}_{\lambda N} - \mathbf{a}_{0N}$ have a same distribution (the asymptotic distribution of $\hat{\mathbf{a}}_N - \mathbf{a}_{0N}$ is given in Lemma 3 in the Appendix), which indicates the oracle property. The distribution of $\hat{\mathbf{a}}_{\lambda N} - \mathbf{a}_{0N}$ is asymptotic normal after adjusting bias and variance.

5. Simulation example

Simulation studies are conducted to examine the performance of our penalized local polynomial regression approach and compare it with that of the local polynomial regression method. Specifically, local linear approximation is used for our proposed approach and the unpenalized local polynomial comparison approach.

5.1. Example 1

We first consider a univariate case, where the data are simulated from the model $Y = xa_1(u) + \varepsilon$ with $\varepsilon \sim N(0, 1)$. The true function $a_1(u)$ is defined as

$$a_1(u) = \begin{cases} 50(u-0.3)^3 & \text{if} 0 \le u \le 0.3, \\ 50(u-0.7)^3 & \text{if} 0.7 \le u \le 1, \\ 0 & \text{otherwise}, \end{cases}$$

and it is plotted in panel (a) of Figure 1.

The covariate x_i is generated iid from $N(0, 4)$. The data points U_i are chosen as n equally spaced design points on [0, 1]. The sample sizes are varied to be $n = 100, 200, 500$ in the simulations. We fix 501 equally spaced grid points on [0, 1] and fit the penalized local polynomial regression on each point to estimate $a_1(\cdot)$. We also run local polynomial regression on these grid points to make comparison.

To examine the performance, we run 200 repetitions of Monte Carlo studies. As the zero region for the true function $a_1(u)$ is [0.3, 0.7], we define the correct zero coverage

(correctzero) as the proportion of region in [0.3, 0.7] that is estimated as zero. The mean of correctzero in 100 repetitions will be reported. Moreover, we report the mean square error of the penalized local polynomial regression, which is defined as $\int_0^1(\hat{a}_{1\lambda}(u)-a_1(u))^2du$. The mean square error of the original local polynomial regression is also reported, which is $\int_0^1(\hat{a}_1(u)-a_1(u))^2du$. These MSEs are calculated using numerical integrations based on the estimates on the 501 equally spaced grid points. The MSEs are used to evaluate how well we estimate the nonparametric function. In addition, we generate an independent test data set, which contains $N = 501$ triples ($\tilde{U}_i$, x_i, $\tilde{Y}_i$). The time points $\tilde{U}_i$ are chosen as 501 equally spaced points on [0, 1]. For x_i and $\tilde{Y}_i$ in the test data set, they are randomly generated in the same way as we have generated the training data set. Define the estimation error (EE) for the entire model as

$$EE=\frac{\sum_{i=1}^{N}(\tilde{x}_i a_1(\tilde{U}_i)-\tilde{x}_i\tilde{a}_1(\tilde{U}_i))^2}{N}.$$

The estimation errors of penalized local polynomial regression and local polynomial regression are calculated with $\tilde{a}_1 = \hat{a}_{1\lambda}$ and $\tilde{a}_1 = \hat{a}_1$, respectively. The mean of these estimation errors will be reported, which reflects the prediction error. Performance of two different methods are reported in Table 1 in terms of the averages of MSE, correctzero, and EE over 200 repetitions. The numbers in parentheses are the corresponding standard error.

5.2. Example 2

Next we consider a bivariate case, where the data are simulated from the model $Y = x_1a_1(u) + x_2a_2(u)+\varepsilon$ with $\varepsilon \sim N(0, 1)$. The first function $a_1(u)$ is the same function used in simulation 1. The second component function $a_2(u)$ is defined as

$$a_2(u)=\begin{cases} 100((u-0.5)^2-0.01) & \text{if} 0.4\le u\le 0.6 \\ 0 & \text{otherwise,}\end{cases}$$

a plot of which is given in panel (b) of Figure 1. It can be seen that $a_2(u)$ is not differentiable at point 0.4 and 0.6.

The design point U_i and the noise ε_i are generated in the same way as in Example 1. The bivariate covariate $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ are generate iid from $N(\mathbf{0}, 4\mathbf{I}_2)$. The sample sizes are still set as $n = 100, 200, 500$. For estimation, the penalized local polynomial regression is fitted on 501 equally spaced grid points on [0, 1]. An independent test data set with size $N = 501$ is generated in a similar way. The estimation error of the entire model is defined as

$$EE=\frac{\sum_{i=1}^{N}(\tilde{x}_{i1}a_1(\tilde{U}_i)+\tilde{x}_{i2}a_2(\tilde{U}_i)-\tilde{x}_{i1}\tilde{a}_1(\tilde{U}_i)-\tilde{x}_{i2}\tilde{a}_2(\tilde{U}_i))^2}{N}$$

for any estimate $\tilde{a}_1(\cdot)$ and $\tilde{a}_2(\cdot)$. We run 200 repetitions of Monte Carlo studies and report the average MSE and correctzero for both two functions $a_1(\cdot)$ and $a_2(\cdot)$. The average estimation error for the entire model is also reported. All results are summarized in Table 2.

From these two simulation examples, we can see that our methods perform better than the original local polynomial regression in the sense that it gives smaller estimation error, which indicates better prediction. Meanwhile, we can estimate the zero regions of each component function quite well. When sample size increases, our method can capture the correct zero regions more accurately.

6. Real data application

We apply our method to the Boston housing data, which has been analyzed by various authors, see for instance Harrison and Rubinfeld (1978), Belsley, Kuh and Welsch (1980) and Ibacache-Pulgar, Paula and Cysneiros (2013). The data set is based on the 1970 US census and consists of the median value of owner-occupied homes for 506 census tracks in Boston area. The aim of the study is to find the association between the median house value (MHV) and various predictors. We treat the median house value as the response and consider four predictors: CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), TAX (full-value property-tax rate per $10000), NOX (nitric oxides concentration parts per 10 million). Figure 2 contains the the scatter plot between the outcome variable and each of the four predictors.

We can see from the scatter plot that the outcome variable has a nonlinear relationship with each of the four predictors, which motivates us to construct a more complicated model than the linear regression. Similar as Fan and Huang (2005), we construct a varying-coefficient model by considering an additional variable LSTAT (percentage of lower income status of the population). As the distribution of LSTAT is asymmetric, the square root transformation is employed to make the resulting distribution symmetric as done in Fan and Huang (2005). The histogram of $\sqrt{\mathrm{LSTAT}}$ is plotted in Figure 3. Specifically, we set $U=\sqrt{\mathrm{LSTAT}}$ as in Fan and Huang (2005). In addition, the covariates CRIM, RM, TAX, NOX are denoted as $\mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(5)}$, respectively. We set $\mathbf{x}_{(1)} = \mathbf{1}$ to include the intercept term.

We construct the following model

$$\boldsymbol{Y}=\sum_{j=1}^{5} a_j(U)\mathbf{x}_{(j)}+\varepsilon.$$

When dealing with the data, we center the response first. We also standardize the covariates $\mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(5)}$. Notice that $a_1(\cdot)$ is the intercept function, which denotes the mean coefficient function. We do not penalize this term because generally the mean coefficient function can be any smooth function. We only penalize the function $a_2(\cdot)$, $a_3(\cdot)$, $a_4(\cdot)$, $a_5(\cdot)$. We solve

$$\min_{\mathbf{a},\mathbf{b}}[\sum_{i=1}^{n}\{Y_i-\mathbf{x}_i^{\top}\mathbf{a}-\mathbf{x}_i^{\top}\mathbf{b}(U_i-u)\}^2(K_h(U_i-u)/K_h(0))+m\sum_{j=2}^{5}P_\lambda(\sqrt{s_j^2a_j^2+r_j^2b_j^2})],$$

where $\mathbf{a} = (a_1, a_2, a_3, a_4, a_5)^{\top}$ and $\mathbf{b} = (b_1, b_2, b_3, b_4, b_5)^{\top}$. The Epanechnikov kernel is employed, and the bandwidth ($h = 1.23$) is selected by the MSE tuning method. As we do not penalize the intercept term, we change the definition of degree of freedom used in tuning

λ by $\sum_{j=2}^{5} I(\|\hat{\beta}_{\lambda j}\|>0)+\sum_{j=2}^{5}\|\hat{\beta}_{\lambda j}\|/\|\hat{\beta}_{j}\|$. We use R to implement our method, which is available at ?. The estimated functions using our penalized local polynomial regression ($a_1(\cdot)$, $a_2(\cdot)$, $a_3(\cdot)$, $a_4(\cdot)$, $a_5(\cdot)$) are plotted in panels (a)–(e) of Figure 4.

From Figure 4, we can see that the variable TAX has no effect on the response when U is between 2.6 and 4, and the variable NOX has no effect on the response when U is less than 2.6 or between 4.3 and 4.7.

We have also calculated the prediction error of our penalized local polynomial regression method, and compared the performance with the original local polynomial regression method. Specifically, we randomly pick up 300 samples from the data as the training data and fit using both methods. After that, we use the remaining 206 data points as our test data, and we can get the prediction error given by $\Sigma_{i\in S}(y_i - \hat{y}_i)^2/|S|$, where S denote the indices for the test data set. We repeat the above step for 100 times by choosing different random seeds, and calculate the mean prediction errors for both methods. We have found that the mean of our penalized prediction error (with standard error in the parentheses) 37.7(3.0) is smaller than that of the original local polynomial regression 39.5(3.1), which indicates that we achieve smaller prediction error by using our penalized local polynomial regression method.

7. Conclusion and Discussions

In this paper, we propose the domain selection for the varying coefficient model using penalized local polynomial regression. Our method can identify the zero regions for each coefficient function component and perform estimation simultaneously. We further proved that our estimator enjoys the oracle property in the sense that they have the same asymptotic distribution as the local polynomial estimates as if the true sparsity is known. We have evaluated our method using both simulation examples and the Boston housing data. A potential extension of our method is to detect the constant regions for each coefficient function, which can be achieved by penalizing the derivative estimates. But there are some potential issues such as which order polynomial to use and how to adjust the original function estimates if the derivative over a certain region is zero. These problems are beyond the scope of this paper.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Appendix

Appendix A. Lemmas

We will introduce the following Lemmas that would be used in proving the theorems.

Recall that s_j is the standard deviation for the pseudo covariates $x_{ij}(K_h(U_i-u)/K_h(0))^{1/2}$ ($1 \le i \le n$), r_j is the standard deviation for the pseudo covariates $x_{ij}(U_i-u)(K_h(U_i-u)/K_h(0))^{1/2}$ ($1 \le i \le n$), and the effective sample size is defined as $m=\sum_{i=1}^{n} K_h(U_i-u)/K_h(0)$

Lemma 1. *We have* $s_j = O_P(1)$, $r_j = O_P(h)$ *and* $m = O_P(nh)$.

Denote $\mathbf{\Xi}=n^{-1}\mathbf{\Gamma}_u^{\top}\mathbf{W}_u\mathbf{\Gamma}_u$ which is a $2p \times 2p$ matrix. We divide $\mathbf{\Xi}$ into $p \times p$ submatrices.

$$\begin{pmatrix} \mathbf{A}_{11} & \dots & \mathbf{A}_{1p} \\ \dots & \dots & \dots \\ \mathbf{A}_{p1} & \dots & \mathbf{A}_{pp} \end{pmatrix},$$

where the 2×2 matrix $\mathbf{A}_{kl}$ ($1 \le k, l \le p$) is as follows:

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}K_h(U_i-u)/K_h(0) & \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}(U_i-u)K_h(U_i-u)/K_h(0) \\ \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}(U_i-u)K_h(U_i-u)/K_h(0) & \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}(U_i-u)^2K_h(U_i-u)/K_h(0) \end{pmatrix}.$$

We will define the new submatrix $\mathbf{B}_{kl}$ as:

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}K_h(U_i-u)/K_h(0)s_j^2 & \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}(U_i-u)K_h(U_i-u)/K_h(0)s_jr_j \\ \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}(U_i-u)K_h(U_i-u)/K_h(0)s_jr_j & \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{il}(U_i-u)^2K_h(U_i-u)/K_h(0)r_j^2 \end{pmatrix}.$$

Lemma 2. *For any* $1 \le k, l \le p$, *the submatrix* $\mathbf{B}_{kl}$ *is a* 2×2 *matrix with every element on the order of* $O_P(h)$.

Define $\mathbf{\Sigma}$ to be the following matrix

$$\begin{pmatrix} \mathrm{B}_{11} & \dots & \mathrm{B}_{1p} \\ \dots & \dots & \dots \\ \mathrm{B}_{p1} & \dots & \mathrm{B}_{pp} \end{pmatrix}, \quad \text{(A.1)}$$

where every element in $\mathbf{\Sigma}$ is on the order of $O_P(h)$.

The following Lemma is taken directly from Theorem 1 of Fan and Zhang (2008).

Lemma 3. *Under the conditions in* Zhang and Lee (2000), *we have*

$$cov^{-1/2}(\hat{\mathbf{a}}(u))\{\hat{\mathbf{a}}(u) - \mathbf{a}(u) - bias(\hat{\mathbf{a}}(u))\} \xrightarrow{d} N(0, \mathbf{I}_p),$$

with

$$\mathrm{Bias}(\hat{\mathbf{a}}(u))=2^{-1}\mu_2\mathbf{a}''(u)h^2 \ \mathrm{Cov}(\hat{\mathbf{a}}(u))=(nhf(u)E(\mathbf{x}\mathbf{x}^{\top}|U=u))^{-1}\nu_0\sigma^2(u),$$

where $\mu_2 = \int u^2K(u)du$ *and* $\nu_0 = \int K^2(u)du$, $f(u)$ *is the density of* u.

Lemma 4. *From Lemma 3 and* Fan and Gijbels (1996), *we have*

$$\begin{aligned} bias(\hat{\mathbf{a}}(u)) &= O_P(h^2) \\ cov(\hat{\mathbf{a}}(u)) &= O_P((nh)^{-1}) \\ bias(\hat{\mathbf{b}}(u)) &= O_P(h^2) \\ cov(\hat{\mathbf{b}}(u)) &= O_P((nh^3)^{-1}). \end{aligned}$$

Remark: Under condition (A1), We will have $\hat{\mathbf{a}}(u) - \mathbf{a}(u) = O_P(b_n)$ and $\hat{\mathbf{b}}(u) - \mathbf{b}(u) = O_P(b_n/h)$. Consequently, we can get $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_P(b_n)$, where $\hat{\boldsymbol{\beta}}$ is the local polynomial regression estimator and $\boldsymbol{\beta}_0$ is the true value.

Appendix B. Proof of Lemmas

Proof of Lemma 1:

For the following proof, we will denote C as a constant term. Let $f(t)$ be the density function of the random variable U_i. We can see

$$\begin{aligned} &E[x_{ij}(U_i-u)\{K_h(U_i-u)/K_h(0)\}^{1/2}] \\ &=Ch^{1/2}E\{E(x_{ij}(U_i \\ &\quad -u)(K_h(U_i-u))^{1/2}|U_i)\} \\ &=Ch^{1/2}E\{E(x_{ij}|U_i)\int(t-u)(K_h(t-u))^{1/2}f(t)dt\} \\ &=ChE\{E(x_{ij}|U_i)\int \upsilon K(\upsilon)f(u \\ &\quad +h\upsilon)d\upsilon\}=O(h), \end{aligned}$$

and

$$\begin{aligned} &E([x_{ij}(U_i-u)\{K_h(U_i-u)/K_h(0)\}^{1/2}]^2) \\ &=ChE[E\{x_{ij}^2(U_i-u)^2(K_h(U_i-u))|U_i\}] \\ &=ChE\{E(x_{ij}^2|U_i)\int(t-u)^2(K_h(t-u))f(t)dt\} \\ &=CE\{E(x_{ij}^2|U_i)\int h^2\upsilon^2K(\upsilon)f(u \\ &\quad +h\upsilon)d\upsilon\}=O(h^2). \end{aligned}$$

As a result, we can get $Var\{x_{ij}(U_i-u)(K_h(U_i-u)/K_h(0)\}^{1/2}) = O_P(h^2)$. Consequently,

$$r_j=O_P(E[x_{ij}(U_i-u)\{K_h(U_i-u)/K_h(0)\}^{1/2}])+O_P(\sqrt{Var[x_{ij}(U_i-u)\{K_h(U_i-u)/K_h(0)\}^{1/2}]})=O_P(h)$$

Similarly, we will have $s_j = O_P(1)$.

For the effective sample size m, we have

$$E(m)=E(\frac{\sum_{i=1}^{n}K_h(U_i-u)}{K_h(0)})=Cnh\int K_h(t-u)f(t)dt=Cnh\int K(\upsilon)f(u+h\upsilon)d\upsilon=O(nh).$$

Similarly, we have $Var(m) = O(nh)$. As $nh = o(1)$, we can see that

$$m=O_P(E(m))+O_P(\sqrt{Var(m)})=O_P(nh).$$

Proof of Lemma 2:

We can see

$$\sum_{i=1}^{n}x_{ik}x_{il}K_h(U_i-u)=O_P(E(\sum_{i=1}^{n}x_{ik}x_{il}K_h(U_i-u)))+O_P(\sqrt{Var(\sum_{i=1}^{n}x_{ik}x_{il}K_h(U_i-u))}).$$

For the mean, one has

$$\begin{aligned}&E(\sum_{i=1}^{n}x_{ik}x_{il}K_h(U_i\\&\quad-u))\\&=nE(x_{ik}x_{il}K_h(U_i\\&\quad-u))\\&=nE(E(x_{ik}x_{il}K_h(U_i\\&\quad-u))|U_i)=nE(x_{ik}x_{il}\int K_h(t\\&\quad-u)f(t)dt)\\&=nE(x_{ik}x_{il}\int K(\upsilon)f(u\\&\quad+h\upsilon)d\upsilon)=O(n).\end{aligned}$$

For the variance, we can use similar technique and get

$$Var(\sum_{i=1}^{n}x_{ik}x_{il}K_h(U_i-u))=O(n/h).$$

As $s_j = O_P(1)$ and $r_j = O_P(h)$, we will get $n^{-1}\sum_{i=1}^{n}x_{ik}x_{il}K_h(U_i-u)/s_j^2=O_P(h)$. Similarly, we obtain $n^{-1}\sum_{i=1}^{n}x_{ik}x_{il}(U_i-u)K_h(U_i-u)/K_h(0)s_jr_j=O_P(h)$ and $n^{-1}\sum_{i=1}^{n}x_{ik}x_{il}(U_i-u)K_h(U_i-u)/K_h(0)r_j^2=O_P(h)$.

For Lemma 3, see Zhang and Lee (2000) for detailed proof.

For Lemma 4, we can use the similar technique in Fan and Gijbels (1996).

Appendix C. Proof of Theorems and Corollary

Proof of Theorem 1:

We write our tuning parameter λ as λ_n because the tuning parameter depends on the sample size. Let $\alpha_n = b_n + a_n$. Denote $L(\boldsymbol{\beta}) = (\mathbf{Y} - \boldsymbol{\Gamma}_{hu}\boldsymbol{\beta})^\top \mathbf{W}_u(\mathbf{Y} - \boldsymbol{\Gamma}_{hu}\boldsymbol{\beta})$. We want to show that for any $\varepsilon > 0$, there exists a large constant C such that

$$P\{\inf_{\|\mathbf{v}\|=C} Q(\boldsymbol{\beta}_0+\alpha_n\mathbf{v}) > Q(\boldsymbol{\beta}_0)\} \geq 1-\varepsilon, \quad \text{(C.1)}$$

which implies $\hat{\boldsymbol{\beta}_\lambda} - \boldsymbol{\beta}_0 = O_P(\alpha_n)$.

As $P_{\lambda n}(0) = 0$, we will have

$$\begin{aligned}
&Q(\boldsymbol{\beta}_0+\alpha_n\mathbf{v}) \\
&\quad - Q(\boldsymbol{\beta}_0) \geq L(\boldsymbol{\beta}_0 \\
&\quad +\alpha_n\mathbf{v}) \\
&\quad - L(\boldsymbol{\beta}_0) \\
&\quad +m\sum_{j=1}^{s}\{P_{\lambda_n}(\|\boldsymbol{\beta}_{0j} \\
&\quad +\alpha_n\mathbf{v}_j\|) \\
&\quad - P_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)\} \\
&\quad =\alpha_n^2\mathbf{v}^\top\boldsymbol{\Gamma}_{hu}^\top\mathbf{W}_u\boldsymbol{\Gamma}_{hu}\mathbf{v} \\
&\quad - 2\alpha_2\mathbf{v}^\top\boldsymbol{\Gamma}_{hu}^\top\mathbf{W}_u(\boldsymbol{Y} - \boldsymbol{\Gamma}_{hu}\boldsymbol{\beta}_0) \\
&\quad +m\sum_{j=1}^{s}\{P_{\lambda_n}(\|\boldsymbol{\beta}_{0j}+\alpha_n\mathbf{v}_j\|) - P_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)\} \\
&\quad =n\alpha_n^2\mathbf{v}^\top\boldsymbol{\Sigma}\mathbf{v} \\
&\quad - 2\alpha_n\mathbf{v}^\top\boldsymbol{\Gamma}_{hu}^\top\mathbf{W}_u(\boldsymbol{Y} \\
&\quad - \boldsymbol{\Gamma}_{hu}\hat{\beta}) \\
&\quad - 2n\alpha_n\mathbf{v}^\top\boldsymbol{\Sigma}(\hat{\beta} \\
&\quad - \boldsymbol{\beta}_0)+m\sum_{j=1}^{s}\{P_{\lambda_n}(\|\boldsymbol{\beta}_{0j} \\
&\quad +\alpha_n\mathbf{v}_j\|) \\
&\quad - P_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)\} \\
&\quad =A_1+A_2+A_3+A_4.
\end{aligned}$$

By Lemma 2, we have $A_1 = O_P(nh\alpha_n^2)$. As $\hat{\boldsymbol{\beta}}$ is the minimizer of $L(\boldsymbol{\beta})$, we will have $(\partial L(\boldsymbol{\beta})/\partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0$, which indicates $A_2 = 0$. By Lemma 4, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(b_n)$, so we will have $A_3 = O_P(nh\alpha_n^2)$, so by choosing a sufficient large C, A_1 dominates A_3. For the term A_4, we have

$$
\begin{aligned}
A_4 &= m\sum_{j=1}^{s}\{P_{\lambda_n}(\|\boldsymbol{\beta}_{0j}+\alpha_n\mathbf{v}_j\|) - P_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)\} \\
&= m\sum_{j=1}^{s}P'_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)(\|\boldsymbol{\beta}_{0j} \\
&+\alpha_n\mathbf{v}_j\| \\
&\quad - \|\boldsymbol{\beta}_{0j}\|) \\
&+ m\sum_{j=1}^{s}P''_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)(\|\boldsymbol{\beta}_{0j}+\alpha_n\mathbf{v}_j\| - \|\boldsymbol{\beta}_{0j}\|)^2(1 \\
&\quad +o(1)) \le m\sum_{j=1}^{s}P'_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)\alpha_n\|\mathbf{v}_j\| \\
&+ m\sum_{j=1}^{s}\max_{1\le j\le s}|P''_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)|\alpha_n^2\|\mathbf{v}_j\|^2(1 \\
&\quad +o(1)) \le m\sum_{j=1}^{s}a_n\alpha_n\|\boldsymbol{\beta}_{0j}\|\|\mathbf{v}\| \\
&+ ms\alpha_n^2\max_{1\le j\le s}|P''_{\lambda_n}(\|\boldsymbol{\beta}_{0j}\|)|\|\mathbf{v}\|^2(1 \\
&\quad +o(1)) = A_5 + A_6.
\end{aligned}
$$

By Lemma 1, we have $A_5 = O_P(m\alpha_n^2) = O_P(nh\alpha_n^2)$ which is dominated by A_1. We will also have $A_6 = o_P(nh\alpha_n^2)$, which is also dominated by A_1. As a result, (C.1) holds, which indicates $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0 = O_P(\alpha_n)$. Further, under condition (A2), we will have $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0 = OP(b_n)$.

Proof of Theorem 2:

We need to prove

$$P(\|\hat{\beta}_{\lambda j}\| = 0) \to 1 \quad \text{(C.2)}$$

holds for any $s + 1 \le j \le p$. If $\hat{\boldsymbol{\beta}}_{\lambda j} \ne 0$, it should be the solution of the following equation

$$\mathbf{0} = \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\Big|_{\boldsymbol{\beta}=\hat{\beta}_\lambda} = -2\boldsymbol{\Gamma}_{huj}^{\top}\mathbf{W}_u(\boldsymbol{Y} - \sum_{j=1}^{p}\boldsymbol{\Gamma}_{huj}\hat{\beta}_{\lambda j}) + mP'_{\lambda_n}(\|\boldsymbol{\beta}_{\lambda j}\|)\frac{\hat{\beta}_{\lambda j}}{\|\hat{\beta}_{\lambda j}\|} = B_1 + B_2.$$

We can see

$$\begin{aligned} B_1 =& -2\boldsymbol{\Gamma}_{huj}^{\top}\mathbf{W}_u(\boldsymbol{Y} \\ & -\sum_{j=1}^{p}\boldsymbol{\Gamma}_{huj}\hat{\beta}_j) \\ & +2\boldsymbol{\Gamma}_{huj}^{\top}\mathbf{W}_u\sum_{j=1}^{p}(\boldsymbol{\Gamma}_{huj}\hat{\beta}_{\lambda j} \\ & -\boldsymbol{\Gamma}_{huj}\hat{\beta}_j) \\ =& 0+2\boldsymbol{\Gamma}_{huj}^{\top}\mathbf{W}_u\boldsymbol{\Gamma}_u(\hat{\beta}_{\lambda j}-\hat{\beta}_j)=2n\boldsymbol{\Sigma}_j O_P(b_n) \\ =& O_P(nhb_n), \end{aligned}$$

where $\boldsymbol{\Sigma}_j$ is the $(2j-1)$th and $2j$th row of the matrix $\boldsymbol{\Sigma}$. For B_2, under Condition (A2), we have $\|B_2\| = mP'_{\lambda_n}(\|\beta_{\lambda j}\|) = O_P(nh\lambda_n)$. Under Condition (A2), we will have $P(\|B_2\| > \|B_1\|) \to 1$, which indicates with probability tending to one, (C.2) does not hold. Thus $\hat{\boldsymbol{\beta}}_{\lambda j}$ must locate at the place where $Q(\boldsymbol{\beta})$ is not differentiable. Since the only place $Q(\boldsymbol{\beta})$ is not differentiable for $\boldsymbol{\beta}_j$ is the origin, we will have $P(\|\hat{\boldsymbol{\beta}}_{\lambda j}\| = 0) \to 1$ for any $s+1 \le j \le p$.

Proof of Theorem 3:

Let $L(\boldsymbol{\beta}) = (\mathbf{Y} - \boldsymbol{\Gamma}_{hu}\boldsymbol{\beta})^{\top}\mathbf{W}_u(\mathbf{Y} - \boldsymbol{\Gamma}_{hu}\boldsymbol{\beta})$. We will have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\Big|_{\boldsymbol{\beta}=(\hat{\beta}_N^T,\mathbf{0}^T)^T} = 0,$$

for $1 \le j \le s$. Note that $\hat{\boldsymbol{\beta}}_{\lambda N}$ is a consistent estimator,

$$
\begin{aligned}
&\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\Big|_{\boldsymbol{\beta}=(\hat{\beta}_N^\top,\mathbf{0}^\top)^\top} \\
&=\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\Big|_{\boldsymbol{\beta}=(\hat{\beta}_{\lambda N}^\top,\mathbf{0}^\top)^\top} \\
&+mP'_{\lambda_n}(\|\hat{\beta}_{\lambda j}\|)\frac{\hat{\beta}_{\lambda j}}{\|\hat{\beta}_{\lambda j}\|} \\
&=\frac{\partial L(\hat{\beta}_N)}{\partial \boldsymbol{\beta}_j} \\
&+\sum_{l=1}^{s}\{\frac{\partial^2 L(\hat{\beta}_N)}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_l} \\
&\quad +o_P(1)\}(\hat{\beta}_{\lambda l} \\
&-\hat{\beta}_l)+m\{P'_{\lambda_n}(\|\boldsymbol{\beta}_{l0}\|)\frac{\boldsymbol{\beta}_{l0}}{\|\boldsymbol{\beta}_{l0}\|}+(T(\boldsymbol{\beta}_{l0})+o_P(1))(\hat{\beta}_{\lambda l}-\boldsymbol{\beta}_{l0})\}=0+\sum_{l=1}^{s}(2n\boldsymbol{\Sigma}_{jl} \\
&+o_P(1))(\hat{\beta}_{\lambda l} \\
&\quad -\boldsymbol{\beta}_{l0}+\boldsymbol{\beta}_{l0}-\hat{\beta}_l)+m\{P'_{\lambda_n}(\|\boldsymbol{\beta}_{l0}\|)\frac{\boldsymbol{\beta}_{l0}}{\|\boldsymbol{\beta}_{l0}\|} \\
&+(T(\boldsymbol{\beta}_{l0}) \\
&\quad +o_P(1))(\hat{\beta}_{\lambda l}-\boldsymbol{\beta}_{l0})\},
\end{aligned}
$$

where $\boldsymbol{\Sigma}_{jl}$ denotes the (2j −1, 2l−1), (2j −1, 2l), (2j, 2l−1), (2j, 2l) elements of the matrix $\boldsymbol{\Sigma}$.

We will have

$$
(2n\boldsymbol{\Sigma}_s+o_P(1))(\hat{\beta}_N-\boldsymbol{\beta}_{0N})=(m\mathbf{H}+2n\boldsymbol{\Sigma}_s+o_P(1))(\hat{\beta}_{\lambda N}-\boldsymbol{\beta}_{0N})+m\mathbf{d}.
$$

We can write

$$
\begin{aligned}
\hat{\beta}_N-\boldsymbol{\beta}_{0N}&=\boldsymbol{\Sigma}_s^{-1}(\frac{m}{2n}\mathbf{H} \\
&+\boldsymbol{\Sigma}_s)(\hat{\beta}_{\lambda N} \\
&\quad -\boldsymbol{\beta}_{0N})+\frac{m}{2n}\boldsymbol{\Sigma}_s^{-1}\mathbf{d} \\
&+o_P(n^{-1})=\boldsymbol{\Sigma}_s^{-1}(\frac{m}{2n}\mathbf{H} \\
&\quad +\boldsymbol{\Sigma}_s)\{(\hat{\beta}_{\lambda N}-\boldsymbol{\beta}_{0N})+(\frac{m}{2n}\mathbf{H}+\boldsymbol{\Sigma}_s)^{-1}\frac{m}{2n}\mathbf{d}\} \\
&+o_P(n^{-1}).
\end{aligned}
$$

Proof of Corollary 1

From Theorem 3, we can get

$$\begin{aligned} cov^{-1/2}(\hat{\beta}_N)\{\hat{\beta}_N &- \boldsymbol{\beta}_{0N} - bias(\hat{\beta}_N)\} = cov^{-1/2}(\hat{\beta}_N)\{\boldsymbol{\Sigma}_s^{-1}(\frac{m}{2n}\mathbf{H} \\ &+\boldsymbol{\Sigma}_s)(\hat{\beta}_{\lambda N} \\ &- \boldsymbol{\beta}_{0N}) + (\frac{m}{2n}\mathbf{H}+\boldsymbol{\Sigma}_s)^{-1}\frac{m}{2n}\mathbf{d} \\ &+o_P(n^{-1}) \\ &- bias(\hat{\beta}_N)\} \\ =& cov^{-1/2}(\hat{\beta}_N)\boldsymbol{\Sigma}_s^{-1}(\frac{m}{2n}\mathbf{H} \\ &+\boldsymbol{\Sigma}_s)(\hat{\beta}_{\lambda N} \\ &- \boldsymbol{\beta}_{0N}) + cov^{-1/2}(\hat{\beta}_N)\boldsymbol{\Sigma}_s^{-1}\frac{m}{2n}\mathbf{d} \\ &- cov^{-1/2}(\hat{\beta}_N)bias(\hat{\beta}_N) \\ &- cov^{-1/2}(\hat{\beta}_N)o_P(n^{-1}) \\ =& T_1+T_2+T_3+T_4. \end{aligned}$$

Under Condition (A2), as $\boldsymbol{\Sigma}_s = O_P(h)$ and $m\mathbf{H}/2n = o(h)$, the first term T_1 is approximately $cov^{-1/2}(\hat{\boldsymbol{\beta}_N})(\hat{\boldsymbol{\beta}_{\lambda N}} - \boldsymbol{\beta}_{0N})$. We can easily see that the second term T_2 is negligible compared to the first term. Similarly, as $b_n = o(1/n)$, the fourth term T_4 is negligible compared to the first term. As a result, we can get

$$cov^{-1/2}(\hat{\beta}_N)\{\hat{\beta}_{\lambda N} - \boldsymbol{\beta}_{0N} - bias(\hat{\beta}_N)\} = cov^{-1/2}(\hat{\beta}_N)\{\hat{\beta}_N - \boldsymbol{\beta}_{0N} - bias(\hat{\beta}_N)\} + o_P(1). \quad \text{(C.3)}$$

From Lemma 3, we can get

$$cov^{-1/2}(\hat{\mathbf{a}}_N(u))\{\hat{\mathbf{a}}_N(u) - \mathbf{a}_{0N}(\mathrm{u}) - bias(\hat{\mathbf{a}}_N(u))\} \xrightarrow{d} N(0, I_s).$$

where $\mathbf{a}_N(u)$ denotes the components that are nonzero. As a result, we can get from (C.3) that

$$cov^{-1/2}(\hat{\mathbf{a}}_N(u))\{\hat{\mathbf{a}}_{\lambda N}(u) - \mathbf{a}_{0N}(\mathrm{u}) - bias(\hat{\mathbf{a}}_N(u))\} \xrightarrow{d} N(0, I_s),$$

where $\hat{\mathbf{a}}_{\lambda N}(u)$ denotes our penalized local polynomial estimates for the function value. This shows the oracle property.

References

Belsley, DA.; Kuh, E.; Welsch, RE. Regression diagnostics: identifying influential data and sources of collinearity. New York-Chichester-Brisbane: John Wiley & Sons; 1980. Wiley Series in Probability and Mathematical Statistics.

Chiang CT, Rice JA, Wu CO. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. Journal of the American Statistical Association. 2001; 96:605–619.

Cleveland, W.; Grosse, E.; Shyu, W. Local regression models. In: Chambers, J.; Hastie, T., editors. Statistical Models. S. London: Chapman & Hall; 1991.

Fan J, Gijbels I. Variable bandwidth and local linear regression smoothers. The Annals of Statistics. 1992; 20:2008–2036.

Fan J, Gijbels I. Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. Journal of the Royal Statistical Society, Series B. 1995; 57:371–394.

Fan, J.; Gijbels, I. Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66. 1 ed.. Chapman & Hall: Chapman & Hall/CRC Monographs on Statistics & Applied Probability; 1996.

Fan J, Huang T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli. 2005; 11:1031–1057.

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

Fan J, Zhang W. Statistical estimation in varying coefficient models. The Annals of Statistics. 1999; 27:1491–1518.

Fan J, Zhang W. Statistical methods with varying coefficient models. Statistics and its Interface. 2008; 1:179–195. [PubMed: 18978950]

Harrison D, Rubinfeld DL. Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management. 1978; 5:81–102.

Hastie T, Tibshirani R. Varying-coefficient models. Journal of the Royal Statistical Society. Series B. 1993; 55:757–796.

Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998; 85:809–822.

Huang JZ, Shen H. Functional coefficient regression models for non-linear time series: a polynomial spline approach. Scandinavian Journal of Statistics. 2004; 31:515–534.

Huang JZ, Wu CO, Zhou L. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika. 2002; 89:111–128.

Huang JZ, Wu CO, Zhou L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statistica Sinica. 2004; 14:763–788.

Ibacache-Pulgar G, Paula GA, Cysneiros FJA. Semiparametric additive models under symmetric distributions. TEST. 2013; 22:103–121.

Kauermann G, Tutz G. On model diagnostics using varying coefficient models. Biometrika. 1999; 86:119–128.

Leng C. A simple approach for varying-coefficient model selection. Journal of Statistical Planning and Inference. 2009; 139:2138–2146.

Lin Y, Zhang HH. Component selection and smoothing in multivariate nonparametric regression. The Annals of Statistics. 2006; 34:2272–2297.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B. 1996; 58:267–288.

Wang H, Xia Y. Shrinkage estimation of the varying coefficient model. Journal of the American Statistical Association. 2009; 104:747–757.

Wang L, Chen G, Li H. Group scad regression analysis for microarray time course gene expression data. Bioinformatics. 2007; 23:1486–1494. [PubMed: 17463025]

Wang L, Li H, Huang J. Variable selection for nonparametric varying-coefficient models for analysis of repeated measurements. Journal of American Statistical Association. 2008; 103:1556–1569.

Wu CO, Chiang CT, Hoover DR. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. Journal of the American Statistical Association. 1998; 93:1388–1402.

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

Zhang W, Lee SY. Variable bandwidth selection in varying-coefficient models. J. Multivariate Anal. 2000; 74:116–134.

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics. 2008; 36:1509–1533. [PubMed: 19823597]

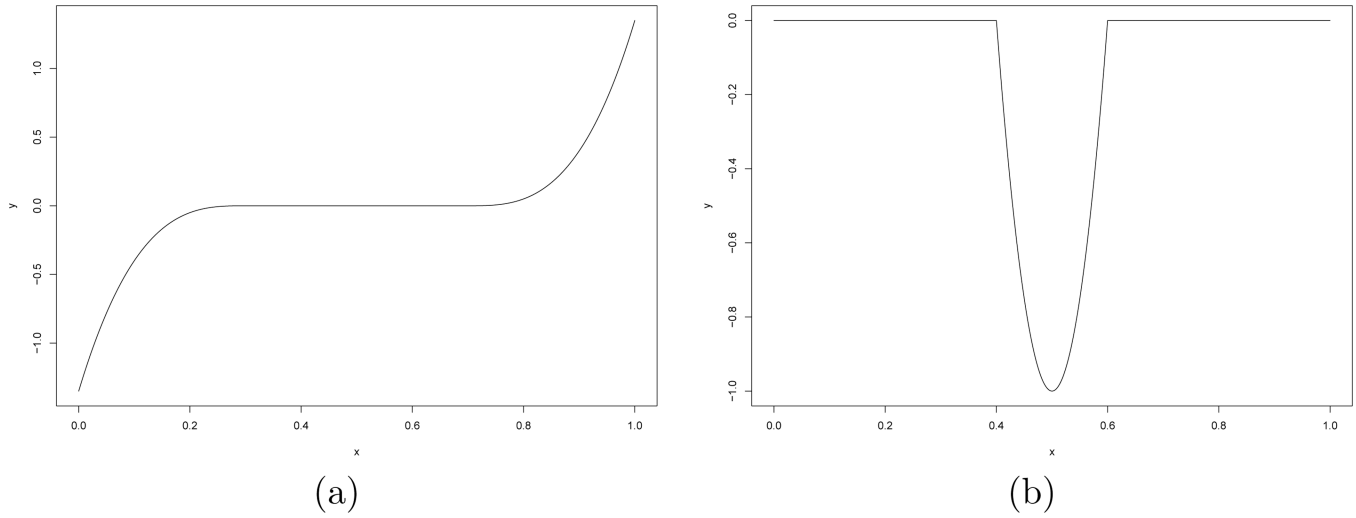

Figure 1.
Plots of $a_1(u)$ (left) and $a_2(u)$ (right) for simulation examples.

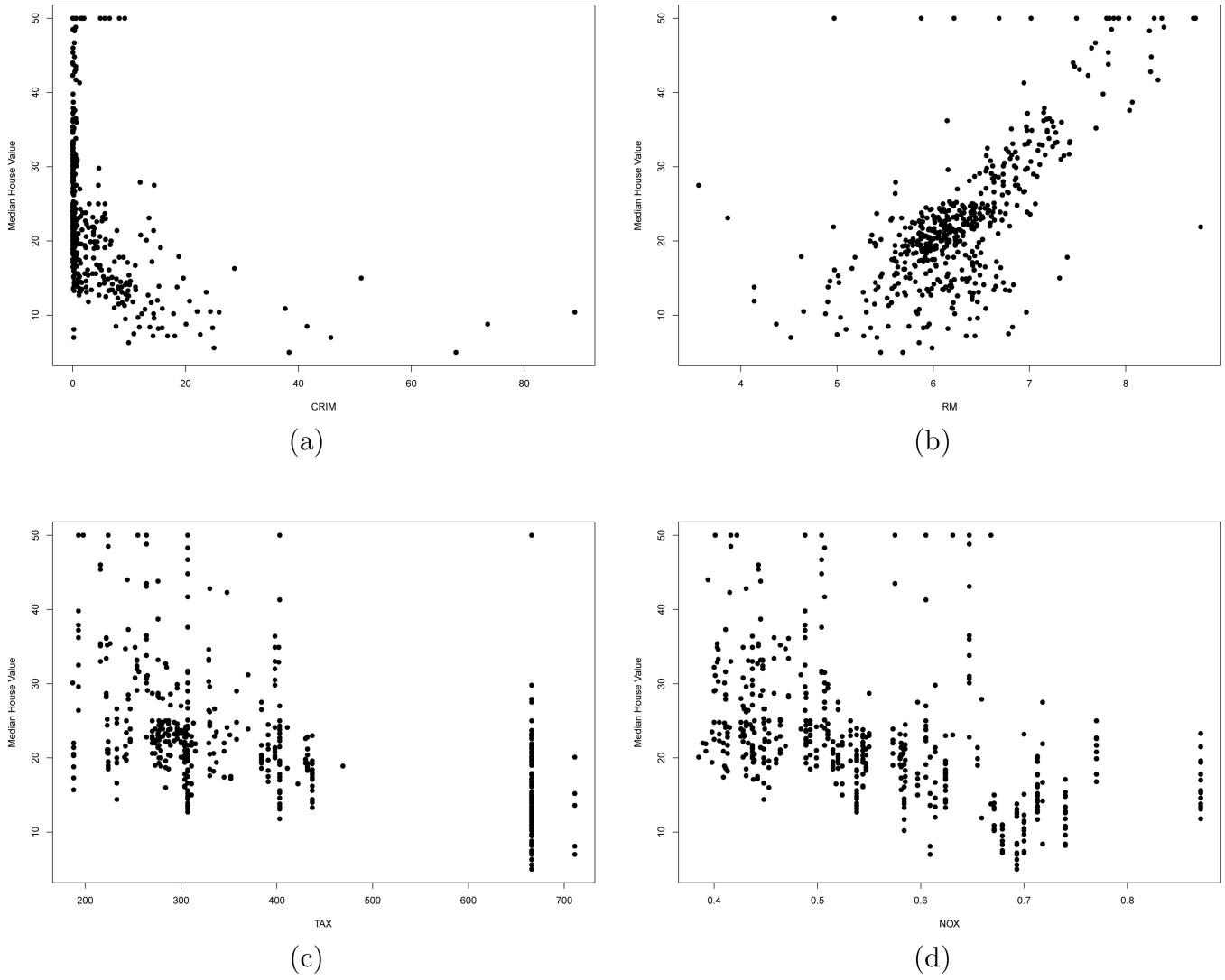

Figure 2.
Scatter plots: Panel (a): MHV versus CRIM, Panel (a): MHV versus RM, Panel (a): MHV versus TAX, Panel (a): MHV versus NOX.

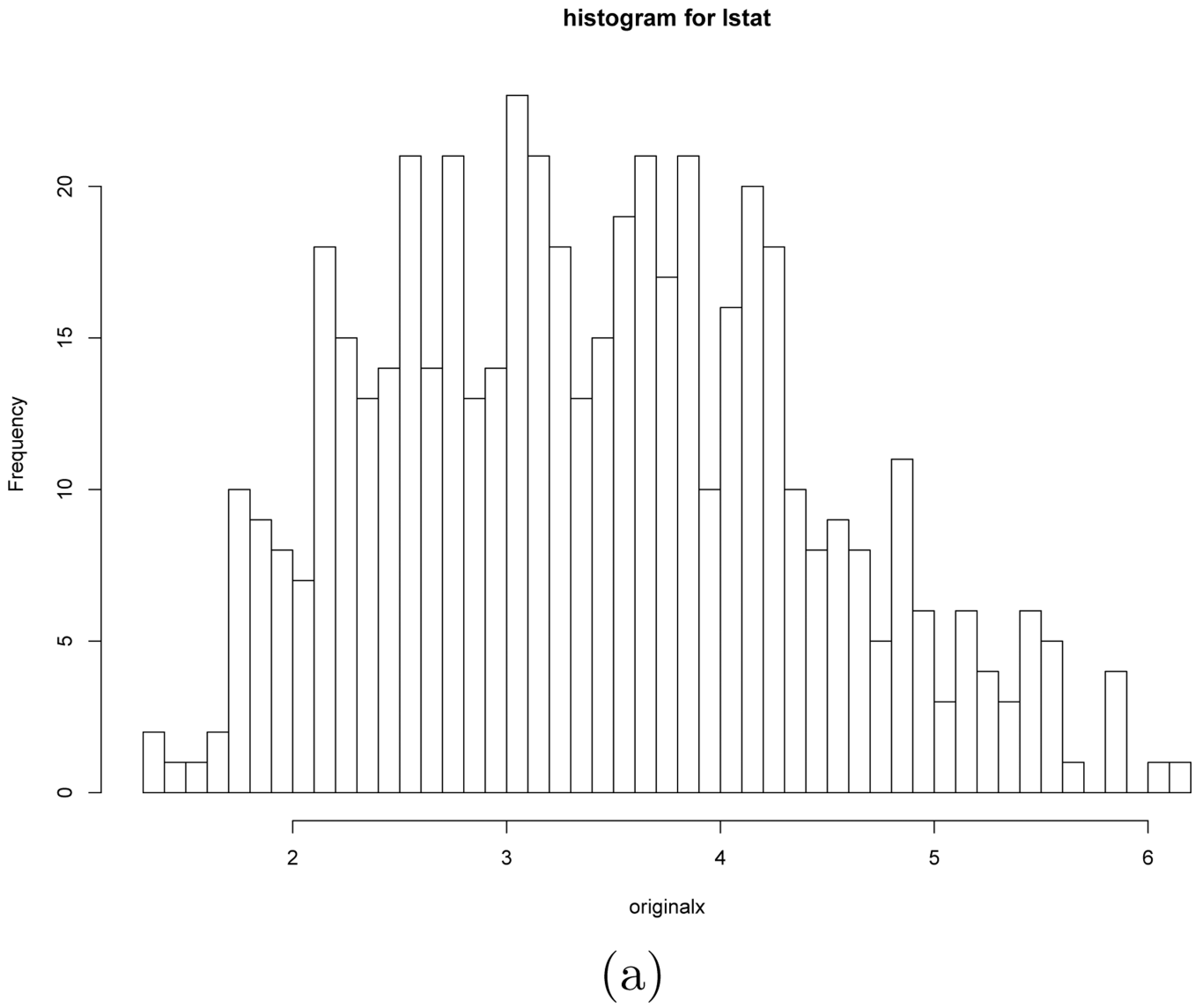

Figure 3.
Histogram for $\sqrt{\mathrm{LSTAT}}$.

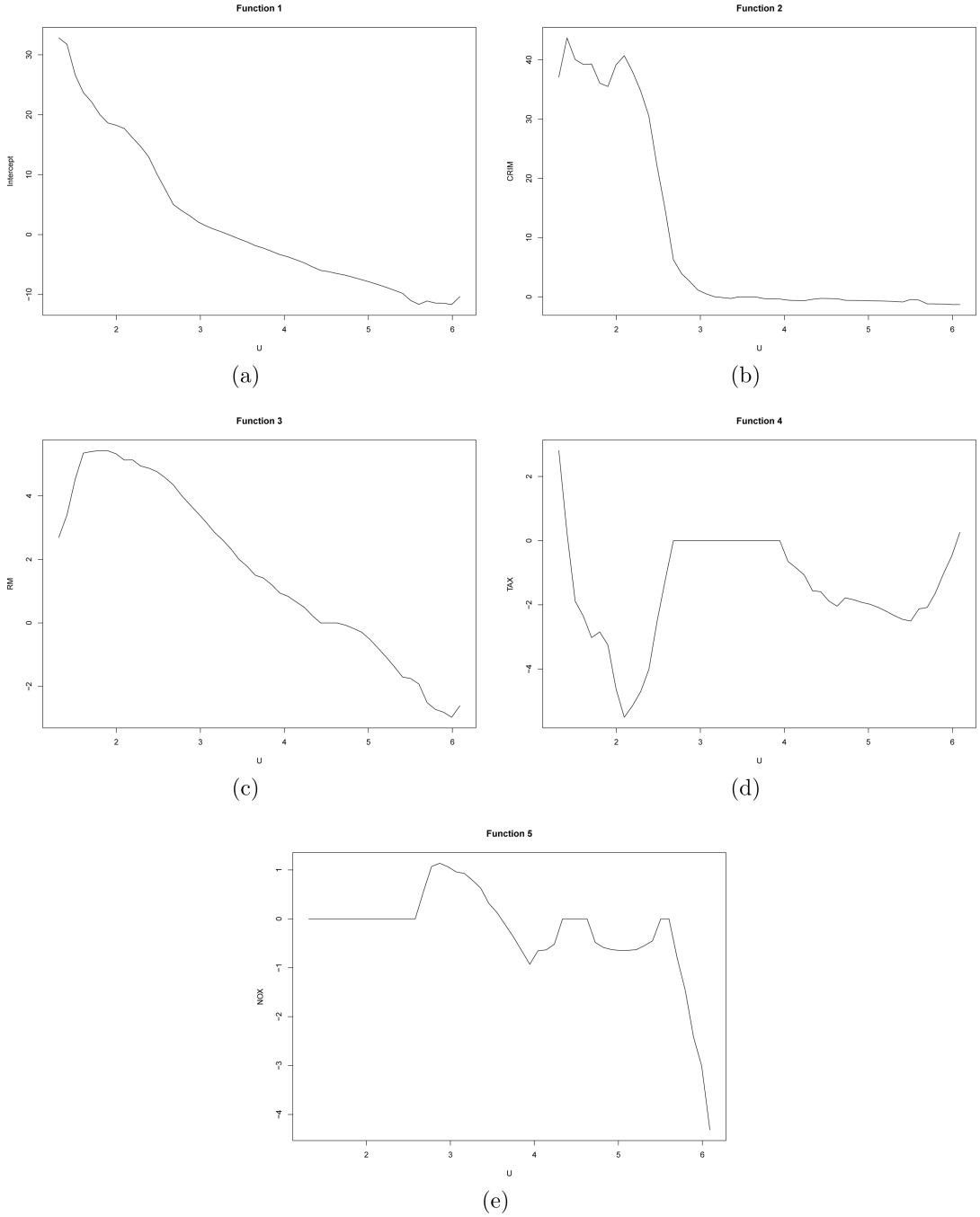

Figure 4.

The penalized local polynomial estimates for the coefficient functions. Panels (a) (b) (c) (d) and (e) correspond to the estimates of the coefficient functions corresponding to the intercept, the variables CRIM, RM, TAX and NOX, respectively.

Table 1

Simulation results for the univariate case using penalized local polynomial regression and original local polynomial regression (denoted as Penalized and Original, respectively) when sample size varies.

Sample size	Method	MSE	correctzero	EE
$n = 100$	Penalized	0.0229(0.0015)	0.9895 (0.003)	0.077 (0.005)
	Original	0.0220(0.0009)	–	0.084 (0.003)
$n = 200$	Penalized	0.0099(0.0008)	0.9903 (0.004)	0.034 (0.002)
	Original	0.0114(0.0004)	–	0.044 (0.002)
$n = 500$	Penalized	0.0035(0.0001)	0.9964 (0.002)	0.013 (0.0004)
	Original	0.0048(0.0002)	–	0.019 (0.0006)

Table 2

Simulation results for the bivariate case using penalized local polynomial regression and original local polynomial regression when sample size varies.

Sample size	Method	MSE (function 1)	correctzero(function 1)	MSE(function 2)	correctzero(function 2)	EE
n = 100	Penalized	0.0300(0.0019)	0.9506 (0.009)	0.0556(0.0019)	0.7511 (0.009)	0.257 (0.009)
	Original	0.0253(0.0010)	–	0.0510(0.0012)	–	0.268 (0.006)
n = 200	Penalized	0.0144(0.0008)	0.9516 (0.007)	0.0222(0.0009)	0.8555 (0.006)	0.113 (0.004)
	Original	0.0126(0.0005)	–	0.0239(0.0007)	–	0.133 (0.003)
n = 500	Penalized	0.0051(0.0002)	0.9793 (0.005)	0.0105(0.0003)	0.8887 (0.003)	0.048 (0.001)
	Original	0.0051(0.0001)	–	0.0116(0.0002)	–	0.060 (0.001)