Communications in Statistics—Theory and Methods, 37: 2249–2261, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920801931887

Taylor & Francis
Taylor & Francis Group

Penalized Spline Estimation for Varying-Coefficient Models

YIQIANG LU[1], RIQUAN ZHANG[2], AND LIPING ZHU[2]

[1]Institute of Electronic Technology, The PLA Information Engineering University, Zhengzhou, P.R. China
[2]Department of Statistics, East China Normal University, Shanghai, P.R. China

Varying-coefficient models are useful extensions of classical linear models. They arise from multivariate nonparametric regression, nonlinear time series modeling and forecasting, longitudinal data analysis, and others. This article proposes the penalized spline estimation for the varying-coefficient models. Assuming a fixed but potentially large number of knots, the penalized spline estimators are shown to be strong consistency and asymptotic normality. A systematic optimization algorithm for the selection of multiple smoothing parameters is developed. One of the advantages of the penalized spline estimation is that it can accommodate varying degrees of smoothness among coefficient functions due to multiple smoothing parameters being used. Some simulation studies are presented to illustrate the proposed methods.

Keywords Generalized cross validation (GCV); Penalized spline; Smoothing parameter estimation; Varying-coefficient models.

Mathematics Subject Classification 62G05; 62G08; 62G20.

1. Introduction

Here we consider the varying-coefficient models (VCM). Let Y be response variable and t and X be associated covariates. The VCM assume the following structure:

$$Y = X^{\tau}\beta(t) + \epsilon, \tag{1.1}$$

Received October 27, 2006; Accepted October 12, 2007
Address correspondence to Yiqiang Lu, Institute of Electronic Technology, The PLA Information Engineering University, ShangCheng Road, Zhengzhou 450004, P.R. China; E-mail: yiqiang_lu@163.com

where ϵ is independent of (t, X), $E(\epsilon) = 0$, $E(\epsilon^2) = \sigma^2$ and $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))^\tau$ is a p-dimensional vector of unknown coefficient functions. For simplification, we assume that t is univariate. The VCM are useful extension of classical linear models. Model (1.1) permits the interaction between the covariates t and X in such a way that a different level of covariate t is associated with a different linear model. The appeal of this model is that via allowing coefficients $\beta_1, \ldots, \beta_p$ to depend on t, the modeling bias can significantly be reduced and "curse of dimensionality" can be avoided. Another advantage of this model is, similar to the linear model, its interpretability.

The varying-coefficient model arises in many different contexts and has successfully applied to multi-dimensional nonparametric regression, generalized linear models, time series analysis, and longitudinal data analysis. Early application of the VCM in time series contexts can be seen in Haggan and Ozaki (1981). However, nonparametric techniques were not popularized until the work of Chen and Tsay (1993) and Hastie and Tibshirani (1993). For nonlinear time series applications, see Chen and Tsay (1993) and Cai et al. (2000a). The varying-coefficient models have also been popularly used to analyze longitudinal data; This allows one to examine the extent to which covariates affect responses over time. (See Chiang et al., 2001; Hoover et al., 1998; Huang et al., 2002; Wu et al., 1998 among others). Some inference about VCM was studied by Fan and Huang (2005).

There are various methods for fitting the VCM, such as, the kernel method (Wu et al., 1998), local polynomial estimate (Cai et al., 2000a,b), basis function approximation (Huang et al., 2002; Lu and Mao, 2004), smoothing spline estimate (Chiang et al., 2001; Hoover et al., 1998), and so on. While all of these approaches to fitting VCM have demonstrated promise, there are some potential weaknesses. For example, the computation of basis function approximation is easy but the asymptotic normality is difficult to establish. For smoothing spline estimators, the multiple smoothing parameters is difficult to choose. For the kernel and local linear methods, all of the coefficient functions are simply estimated by using the only one bandwidth. The appeal is that the coefficient functions can easily be estimated via a simple local regression. This yield a simple one-step estimation procedure. However, Fan and Zhang (1999) showed that such a one-step method was not optimal when different coefficient functions admit different degree of smoothness and proposed a two-step procedure to accommodate varying degrees of smoothness among coefficient functions. In theory, penalized spline estimates of VCM can also accommodate varying degrees of smoothness among coefficient functions due to multiple smoothing parameters being used. In this article, we proposed that the varying-coefficient models are estimated by penalized spline. Assuming a fixed but potentially large number of knots, it is shown that the penalized spline estimators are strong consistency and asymptotic normality. A systematic optimization algorithm for the selection of multiple smoothing parameters is developed.

The article is organized as follows. Section 2 describes the penalized spline approach to VCM and Sec. 3 gives asymptotic properties of penalized spline estimators. A systematic optimization algorithm for section of smoothing parameters is presented in Sec. 4. In Sec. 5, the methodology is illustrated by some simulated studies. Detailed proofs are in the Appendix.

2. Penalized Spline Estimation

Firstly, let's review penalized spline estimation of nonparametric regression. Consider a regression problem $y_i = \eta(x_i) + \epsilon_i,\ i = 1, \ldots, n$ where $x_i \in [0, 1]$ and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. The unknown univariate function $\eta(\cdot)$ can be estimated by a penalized B-spline (Eilers and Marx, 1996). Assume that:

$$\eta(x) = c_1 B_1(x) + \cdots + c_N B_N(x),$$

where $\{B_j(x)\}_{j=1}^N$ are B-spline basis of degree m, $N = m + k + 1$, and k is the number of interior knots. A penalized B-spline estimation of η is via the minimization of a penalized least square score

$$\frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \sum_{i=1}^{N} c_j B_j(x_i)\right\}^2 + \lambda \sum_{j=l+1}^{N} (\Delta^l c_j)^2, \tag{2.1}$$

where Δ^l is the lth difference operator and $\sum_{j=k+1}^{N}(\Delta^l c_j)^2$ is the approximation of $\int_0^1 [\eta^{(l)}(x)]^2 dx$ (Eilers and Marx, 1996). The first term of expression (2.1) measures the goodness of fit and the second term penalizes the roughness of η. The smoothing parameter λ controls the trade-off between the two conflicting goals. It is often to set $m = 3$ and $l = 2$. Thus, the obtained estimate of $\eta(\cdot)$ has the 2th continuous derivative. Define the (i, j)th element of B to be $B_j(x_i)$, D_l to be the matrix representation of the difference operator Δ^l, $Y = (y_1, \ldots, y_n)^\tau$, and $D = D_l^\tau D_l$. Then the penalized B-spline estimation reduces to the minimization of

$$(Y - Bc)^\tau (Y - Bc) + n\lambda c^\tau D c \tag{2.2}$$

with respect to c.

Now, let's apply penalized B-spline approach to varying-coefficient model. Suppose that $\{Y_i, X_i^\tau, t_i\}_{i=1}^n$ is a random sample from model (1.1). Furthermore, assume that $\beta_r(t)$ are lth differentiable of all $t \in [a, b]$ and

$$\beta_r(t_i) = c_{r1} B_{r1}(t_i) + \cdots + c_{rN_r} B_{rN_r}(t_i) = B_r^\tau(t_i) c_r,$$

where $c_r = (c_{r1}, \ldots, c_{rN_r})^\tau$ and $B_r(t_i) = (B_{r1}(t_i), \ldots, B_{rN_r}(t_i))^\tau$ is B-spline basis. We can obtain the penalized spline estimator $\hat{\beta}_r(t)$ of $\beta_r(t)$ by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_{i1} B_1^\tau(t_i) c_1 - \cdots - x_{ip} B_p^\tau(t_i) c_p\right)^2 + \sum_{r=1}^{p} \lambda_r c_r^\tau Q_r c_r, \tag{2.3}$$

where λ_r's are non negative smoothing parameter and Q_r's are defined similarly to D in Eq. (2.2). Let $z_i = (x_{i1} B_1^\tau(t_i), \ldots, x_{ip} B_p^\tau(t_i))^\tau$ and $\theta = (c_1^\tau, \ldots, c_n^\tau)^\tau$. (2.3) can be expressed as:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\tau \theta)^2 + \sum_{r=1}^{p} \lambda_r \theta^\tau S_r \theta, \tag{2.4}$$

where $S_r = \mathrm{diag}(0, \ldots, 0, Q_r, 0, \ldots, 0)$ such that $\theta^\tau S_r \theta = c_r^\tau Q_r c_r$. The S_r's are non negative definite coefficient matrices defining the penalties, each with associated smoothing parameter λ_r.

Although other bases (i.e., truncated power function basis; Yu and Ruppert, 2002) could be used, the modeling methods are the same.

3. Asymptotic Properties

In this section, we present results on the strong consistency and asymptotic normality for penalized B-spline estimators of VCM. There are two types of asymptotics that one could use, increasing number of knots and fixed knots. The first approach does not assume that $\beta_r(t), r = 1, \ldots, p$, are spline functions, so to obtain consistency one must allow the number of knots in the spline model to increase as the sample size increases. The second approach assumes that $\beta_r(t)$ is itself a spline function with fixed k_r knots and lies somewhere in a gray zone between parameter and nonparametric modeling. Like parametric modeling, there is a fixed number of parameter and the parameters can be estimated at $\sqrt{n}$ rates. Like nonparametric modeling, the model is flexible enough to adapt to coefficient functions of unknown form, and regularization by penalty terms is needed to avoid overfitting. Asymptotic results with an increasing number of knots are limited to rates of convergence; at least this is true of all results of which we are aware. Ruppert (2002) showed that asymptotic bias is small and negligible compared to the variance when nonparametric regression is not itself a spline but approximated by a spline function with the fixed knots. Furthermore, fixed-knots asymptotics give a more practical result, convergence to known normal distribution. With asymptotic distribution known, this distribution can be used, for inference. Thus, for the remainder of the article the working assumption is that for some m, $\beta_r(t)$ $(r = 1, \ldots, p)$ are spline function of degree m.

We denote $\lambda = (\lambda_1, \ldots, \lambda_p)$ by $\lambda_n = (\lambda_{n1}, \ldots, \lambda_{np})$ to indicate dependence on sample size. The following assumptions are needed to prove asymptotics.

Assumption 1. The parametric space Θ is compact. The true parameter vector θ_0 is an interior point of Θ.

Assumption 2.

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} z_i z_i^{\tau} = \Omega \quad \text{a.s.}$$

exits and is non singular.

Theorem 3.1. *Under Assumptions* 1 *and* 2, *if the smoothing parameter* $\lambda_{nr} = o(1)$, $r = 1, \ldots, p$, *then a sequence of penalized least estimators minimizing expression* (2.4) *is a strong consistent estimators of* θ_0.

The proof of Theorem 3.1 is given in Appendix A.

Theorem 3.2. *Under Assumptions* 1 *and* 2, *if the smoothing parameter* $\lambda_{nr} = o(n^{-1/2})$, $r = 1, \ldots, p$, *then a sequence of penalized least estimators* $\hat{\theta}_{n,\lambda_n}$ *minimizing expression* (2.4) *is asymptotically normally distributed. That is*:

$$n^{1/2}(\hat{\theta}_{n,\lambda_n} - \theta_0) \xrightarrow{D} Normal(0, \sigma^2 \Omega^{-1}).$$

The proof of Theorem 3.2 is given in Appendix B.

Remark 3.1. Let $B^{\tau}(t_0) = (B_1(t_0), \ldots, B_{k_r}(t_0))$ be the B-spline base and C_r be coefficient matrix such that $c_r = C_r\theta$. We have:

$$n^{1/2}(\hat{\beta}_r(t_0) - \beta_r(t_0)) \xrightarrow{D} Normal(0, \sigma^2 B^{\tau}(t_0) C_r \Omega^{-1} C_r^{\tau} B(t_0)).$$

Remark 3.2. The asymptotic variance in Theorem 3.2 does not involve λ since λ goes to 0 sufficiently fast as n tends infinity. For finite sample inference, one would expect this asymptotic variance to over-estimated the variance of $\hat{\theta}_n$. Therefore, for purpose of inference, we give the variance of $\hat{\theta}_n$ when λ and n are fixed:

$$\mathrm{var}(\hat{\theta}) = \left(\sum_{i=1}^{n} z_i z_i^{\tau} + \sum_{r=1}^{p} n\lambda_r S_r\right)^{-1} \left(\sum_{i=1}^{n} z_i z_i^{\tau}\right) \left(\sum_{i=1}^{n} z_i z_i^{\tau} + \sum_{r=1}^{p} n\lambda_r S_r\right)^{-1} \sigma^2.$$

4. Selection of Smoothing Parameters

Given a fixed value of the number of knots, the uniform knots are often placed. Ruppert (2002) has a detailed study about the choice of the number of knots. For smooth, either monotonic or unimodal, functions 10–15 knots seem quite adequate and that is what we recommend. Also, more than 20 knots would be needed when the estimating coefficient functions have many local minima and maxima. The advantage of penalized splines is that the problem with knot placement can be partially alleviated by abandoning pure regression splines in favor of penalized regression splines (Parker and Rice, 1985; Wahba, 1980). But in this case, model flexibility is again controlled by a smoothing parameter λ, rather than the basis dimension. Therefore, in this section, we mainly discuss the selection of smoothing parameters.

For convenience, we define $G = \sum_{i=1}^{n} z_i z_i^{\tau}$ and replace $n\lambda_{nr}$ by λ_r. The penalized B-spline estimation reduces to minimize

$$(Y - G\theta)^{\tau}(Y - G\theta) + \sum_{r=1}^{p} \lambda_r \theta^{\tau} S_r \theta. \tag{4.1}$$

Given the smoothing parameters, the minimizer of (4.1) is easily solved, but the smoothing parameters have to be estimated. In principle, this can be achieved by minimizing GCV score

$$V = \frac{n\|Y - AY\|^2}{[\mathrm{tr}(I - A)]^2}, \tag{4.2}$$

with respect to λ_r's. A is the influence or hat matrix of the model, i.e., $\widehat{Y} = AY$, and depends on the smoothing parameters.

The practical difficulty in using GCV lies in the minimization of express (4.2), which has the potential to be prohibitively expensive numerically. A simple grid search for p smoothing parameters using k grid points per parameter would require of the order of nq^2k^p operations (q is the dimension of θ). A useful efficient multiple smoothing parameters estimates based on the exact GCV score was pioneered by Gu and Wahba (1991). Wood (2000) provided an effective means of selecting the degree smoothness for terms in generalized additive model (GAM). Wood (2004)

further developed a stable and efficient multiple smoothing parameters estimation, which allows the fixed penalty term.

Enlightened by these thoughts of the selection of multiple smoothing parameters, we developed an efficient method of selecting the smoothing parameters of penalized spline estimation for VCM. The basic approach is to perform Newton updates of the log smoothing parameters. Working the log smoothing parameters has advantage of ensuring that the smoothing parameter estimates are positive, and that is also justified heuristically by the fact the plots of GCV functions for one-dimensional smoothes appear more susceptible to quadratic approximation than to equivalent plots on the original scale. Hence, derivative of the GCV score with respect to the log smoothing parameters is of primary concern.

Problem (4.1) has the following influence matrix:

$$A = G\left(GG^{\tau} + \sum_{r=1}^{p} \lambda_r S_r\right)^{-1} G^{\tau}.$$

Suppose that the QR decomposition of G follows as $G = QR$. Defining $S = \sum_{r=1}^{p} \lambda_r S_r$ and B as a square root of S such that $B^{\tau}B = S$, a singular value decomposition can be formed as:

$$\begin{bmatrix} R \\ B \end{bmatrix} = UDV^{\tau},$$

where the column of U are column of an orthogonal matrix and V is an orthogonal matrix and D is the diagonal matrix of singular value. Now defining the submatrix U_1 and U_2 of U such that $R = U_1 DV^{\tau}$ and $B = U_2 DV^{\tau}$, we have that:

$$G = QU_1 DV^{\tau} \quad \text{and} \quad G^{\tau}G + S = VD^2V^{\tau}.$$

Turning to the derivatives of GCV, it is convenient to write the influence matrix as

$$A = G\left(G^{\tau}G + \sum_{r=1}^{p} \lambda_r S_r\right)^{-1} G^{\tau} = GH^{-1}G^{\tau},$$

where $H = G^{\tau}G + \sum_{r=1}^{p} \lambda_r S_r$. Defining $\eta_i = \log(\lambda_i)$, we then have the following theorem.

Theorem 4.1. *Let* $\alpha = \|Y - AY\|^2$ *and* $\delta = n - \mathrm{tr}(A)$, *so that* $V = \frac{n\alpha}{\delta^2}$. *Then*

$$\frac{\partial V}{\partial \eta_i} = \frac{n}{\delta^2}\frac{\partial \alpha}{\partial \eta_i} - \frac{2n\alpha}{\delta^3}\frac{\partial \delta}{\partial \eta_i}$$

and

$$\frac{\partial V}{\partial \eta_i \partial \eta_j} = -\frac{2n}{\delta^3}\frac{\partial \delta}{\partial \eta_j}\frac{\partial \alpha}{\partial \eta_i} + \frac{n}{\delta^2}\frac{\partial^2 \alpha}{\partial \eta_i \partial \eta_j} + \frac{6n\alpha}{\delta^4}\frac{\partial \delta}{\partial \eta_j}\frac{\partial \delta}{\partial \eta_i} - \frac{2n}{\delta^3}\frac{\partial \alpha}{\partial \eta_j}\frac{\partial \delta}{\partial \eta_i} - \frac{2n\alpha}{\delta^3}\frac{\partial^2 \delta}{\partial \eta_i \partial \eta_j}.$$

Theorem 4.1 can be easily obtained by the principle of derivatives.

Theorem 4.2. *Let* $Y_1 = U_1^\tau Q^\tau Y$, $M_i = D^{-1}V^\tau S_i V D^{-1}$ *and* $K_i = M_i U_1^\tau U_1$. *Then we have*:

$$\frac{\partial \alpha}{\partial \eta_i} = 2\lambda_i (Y_1^\tau M_i Y_1 - Y_1^\tau K_i Y_1), \tag{4.3}$$

$$\frac{\partial \delta}{\partial \eta_i} = \lambda_i \mathrm{tr}(K_i), \tag{4.4}$$

$$\frac{\partial^2 \alpha}{\partial \eta_i \partial \eta_j} = 2\lambda_i \lambda_j Y_1^\tau (M_i K_j + M_j K_i + K_i M_j - M_i M_j - M_j M_i) Y_1 + \delta_{ij} \frac{\partial \alpha}{\partial \eta_i}, \tag{4.5}$$

$$\frac{\partial^2 \delta}{\partial \eta_i \partial \eta_j} = -[\lambda_i \lambda_j \mathrm{tr}(K_i M_j) + \lambda_i \lambda_j \mathrm{tr}(K_j M_i) - \delta_{ij} \mathrm{tr}(K_i)], \tag{4.6}$$

where $\delta_{ij} = 1$ *if* $i = j$ *and* 0 *otherwise.*

The proof of Theorem 4.2 is given in Appendix C.

From Theorems 4.1 and 4.2, the gradient $g = \left(\frac{\partial}{\partial \eta}\right) V(\eta)$ and the Hessian $K = \left(\frac{\partial^2}{\partial \eta \partial \eta^\tau}\right) V(\eta)$ can be evaluated so that Newton's method can be used to find the optimal λ fairly efficiently.

Now let's briefly describe the algorithm for the minimization of $V(\lambda)$ as function of smoothing parameter λ. The algorithm operates on $\eta_i = \log(\lambda_i)$.

1. Initialization: Set $\eta_- = \eta_0$, $\Delta\eta = 0$, and $V_- = \infty$. Reasonable starting value of λ_i is proportional to $1/\mathrm{tr}(S_i)$ (Wood, 2000).
2. Iteration:
   (a) Set $\eta = \eta_- + \Delta\eta$.
   (b) Evaluate the gradient $g = \left(\frac{\partial}{\partial \eta}\right) V(\lambda)$ and the Hessian $K = \left(\frac{\partial^2}{\partial \eta \partial \eta^\tau}\right) V(\lambda)$ according to Theorems 4.1 and 4.2;
   (c) Calculate the increment $\Delta\eta = -\widetilde{K}^{-1} g$ where $\widetilde{K} = K + \mathrm{diag}(e)$ is positive definite. If K itself is positive definite "enough", e is simple set to 0.
   (d) Check convergence conditions. If conditions fail, set $\eta_- = \eta$, $V_- = V$, go to (a).
3. Compute return value: Return θ and V at converge η.

5. Simulation

Consider the regression model

$$Y = \sin(15\pi u) X_1 + 4u(1-u) X_2 + \epsilon,$$

where u follows a uniform distribution on $[0, 1]$ and X_1 and X_2 are normally distributed with correlation coefficient $2^{-1/2}$. Furthermore, the marginal distributions of X_1 and X_2 are the standard normal, and ϵ, t, and (X_1, X_2) are independent. The random variable ϵ follows a normal distribution with mean zero and standard variance $\sigma = 0.3$. The σ is chosen so that the noise to signal ratio is about 1:5.

The obvious characteristic of this example is that the two coefficient functions admit different degree of smoothness. For the above example, we conducted two simulations with sample size 300 and 500, respectively, each with 100 replications.

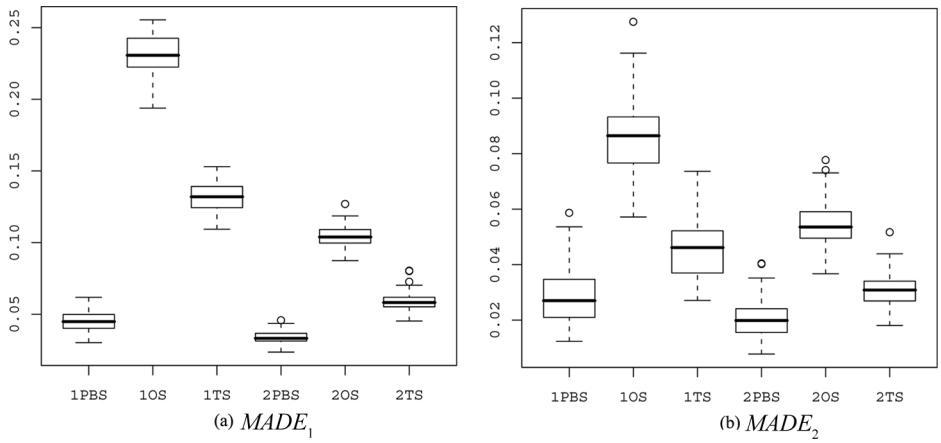

Figure 1. Comparison among one-step, two-step, and penalized estimators based 100 simulations.

With the simulated data, the $\beta_1(u)$ and $\beta_2(u)$ were estimated by penalized B-spline and the smoothing parameter was chosen by the use of the proposed algorithm in Sec. 4. Figure 2 is a typical penalized B-spline estimators based on sample size $n = 500$. The penalized B-spline estimation of $\beta_1(u)$ and $\beta_2(u)$ were also compared with one-step and two-step local linear estimation (Fan and Zhang, 1999). The optimal bandwidths were chosen by the cross-validation method in the one-step and two-step local linear estimation. The performance of the estimators for functional coefficients $\beta_j(u)$ is assessed via mean absolute deviation error

$$MADE_j = n_{grid}^{-1} \sum_{k=1}^{n_{grid}} \{\hat{\beta}_j(u_k) - \beta_j(u_k)\}^2,$$

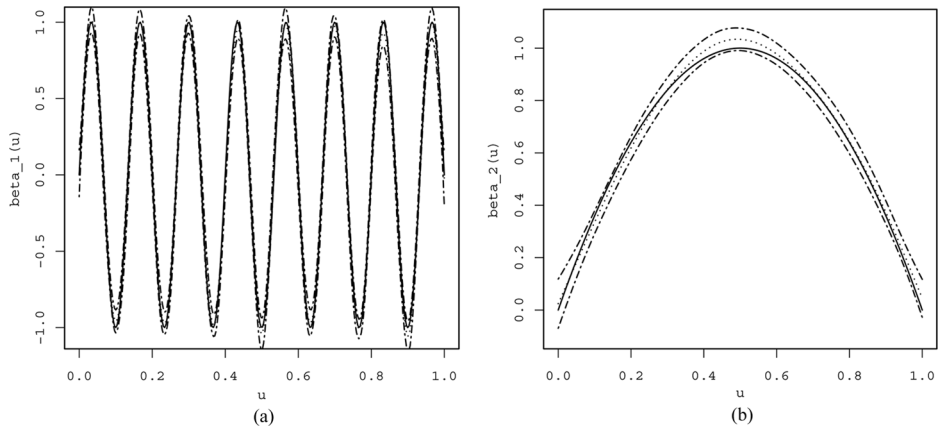

Figure 2. Penalized B-spline estimates of $\beta_1(u)$ and $\beta_2(u)$ based on sample size $n = 500$. Solid curve—true functions; dotted curve—penalized B-spline estimators; dotdash curve—95% pointwise confidence intervals.

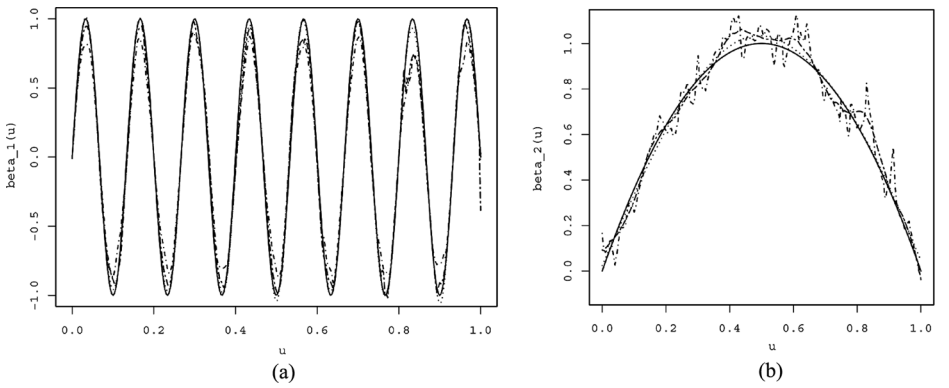

Figure 3. A typical estimates of $\beta_1(u)$ and $\beta_2(u)$ based on sample size $n = 500$. Solid curve—true functions; dotted curve—penalized B-spline estimators; dotdash curve—local linear one-step estimators; twodash curve—two-step estimators.

where $\{u_k, k = 1, \ldots, n_{grid} = 300\}$ are the grid points at which the function $\beta_j(\cdot)$ are estimated.

For each setting, we obtained 100 estimates of coefficient functions. Figure 1 gives the box plots of $MADE_1$ and $MADE_2$. The x-axis labels in the box plots read as follow: '1PBS' denotes the case for $n = 300$ using penalized B-spline approach; '1OS' denotes the case for $n = 300$ using one-step local linear approach; 1TS' denotes the case for $n = 300$ using two-step local linear approach; '2PBS' denotes the case for $n = 500$ using penalized B-spline approach; and so on. Figure 3 present a typical example of the estimated coefficient functions with the sample $n = 500$.

From Figs. 1–3, we see that, when different coefficient coefficient functions admit different degrees of smoothness, both the penalized B-spline method and two-step method gave fairly good estimates of $\beta_1(u)$ and $\beta_2(u)$. Furthermore, the penalized B-spline method sightly outperforms the two-step method in terms of the observed MADE. But for this case the one-step method could not adequately estimate the coefficient functions. The other simulated examples, in which different coefficient coefficient functions admit different degrees of smoothness, were done. The same results were obtained and omitted for brevity.

6. Appendix

A *Proof of Theorem* 3.1

The penalized least squares $\hat{\theta}$ minimize Eq. (2.4), which can be expanded and written as:

$$Q_{n,\lambda_n}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\tau\theta)^2 + \sum_{r=1}^{p}\lambda_{nr}\theta^\tau S_r\theta$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\tau\theta_0 + z_i^\tau\theta_0 - z_i^\tau\theta)^2 + \sum_{r=1}^{p}\lambda_{nr}\theta^\tau S_r\theta$$

$$= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(z_i^\tau\theta_0 - z_i^\tau\theta) + \frac{1}{n}\sum_{i=1}^{n}(z_i^\tau\theta_0 - z_i^\tau\theta)^2 + \sum_{r=1}^{p}\lambda_{nr}\theta^\tau S_r\theta$$

$$= A_1 + A_2 + A_3 + A_4.$$

The following limits are taken when $n \to \infty$ unless otherwise stated. First, by strong law of large numbers, for almost every ϵ, $A_1 \to \sigma_0^2$. Second, for almost every ϵ, $A_2 \to 0$. Third:

$$A_3 = \frac{1}{n}\sum_{i=1}^{n}(z_i^\tau\theta_0 - z_i^\tau\theta)^2 = (\theta_0-\theta)^\tau\frac{1}{n}\sum_{i=1}^{n}z_iz_i^\tau(\theta_0-\theta) \xrightarrow{a.s} (\theta_0-\theta)^\tau\frac{1}{n}\Omega(\theta_0-\theta) \stackrel{def}{=} Q(\theta).$$

Since Ω is non singular, $Q(\theta)$ has a unique minimum at $\theta = \theta_0$. Finally, since a constant sequence of smoothing parameter $\lambda_{nr} = o(1)$ and $\theta \in \Theta$ compact, we have $A_4 \to 0$. Thus,

$$Q_{n,\lambda_n}(\theta) \to Q(\theta) + \sigma_0^2 \tag{A.1}$$

uniformly for all $\theta \in \Theta$.

Let $\hat{\theta}_{n,\lambda_n}$ be a sequence of the penalized least squares estimators. Let θ' be a limit point of the sequence $\{\hat{\theta}_{n,\lambda_n}\}$, thus there exists a subsequence $\{\hat{\theta}_{n_t,\lambda_{n_t}}\}$ such that it converges to θ'.

Next, we show that:

$$Q_{n_t,\lambda_{n_t}}(\hat{\theta}_{n_t,\lambda_{n_t}}) \to Q(\theta') + \sigma_0^2 \tag{A.2}$$

as $t \to \infty$. We can write:

$$Q_{n_t,\lambda_{n_t}}(\hat{\theta}_{n_t,\lambda_{n_t}}) - Q(\theta') - \sigma_0^2 = [Q_{n_t,\lambda_{n_t}}(\hat{\theta}_{n_t,\lambda_{n_t}}) - Q(\hat{\theta}_{n_t,\lambda_{n_t}}) - \sigma_0^2] + [Q(\hat{\theta}_{n_t,\lambda_{n_t}}) - Q(\theta')].$$

Note that the first term on the right-hand side of the above equation converges to zeros by uniform convergence of $Q_{n,\lambda_n}(\theta)$ to $Q(\theta) + \sigma_0^2$ in Eq. (A.1), using the fact that

$$\left|Q_{n_t,\lambda_{n_t}}(\hat{\theta}_{n_t,\lambda_{n_t}}) - Q(\hat{\theta}_{n_t}) - \sigma_0^2\right| \le \sup_{\theta\in\Theta}\left|Q_{n,\lambda_n}(\theta) - Q(\theta) - \sigma_0^2\right|.$$

And the second term converges to zeros, since $Q(\theta)$ is continuous and $\hat{\theta}_{n_t} \to \theta'$. Then by the triangle inequality, (A.2) holds.

Now since $\hat{\theta}_{n_t,\lambda_{n_t}}$ is the penalized least squares estimates, that is, minimizer of $Q_{n_t,\lambda_{n_t}}(\cdot)$,

$$Q_{n_t,\lambda_{n_t}}(\hat{\theta}_{n_t,\lambda_{n_t}}) \le Q_{n_t,\lambda_{n_t}}(\theta_0)$$

by letting $t \to \infty$, we have that the left-hand side of the inequality converges to $Q(\theta') + \sigma_0^2$ by (A.2) and the right-hand side of the inequality converges to $Q(\theta_0) + \sigma_0^2 = \sigma_0^2$. Thus,

$$Q(\theta') + \sigma_0^2 \le \sigma_0^2,$$

$Q(\theta') = 0$ follows. Since $Q(\theta)$ has a unique zero at θ_0, the limit point θ' must be θ_0. Since this result holds almost every ϵ, $\hat{\theta}_{n,\lambda_n} \to \theta_0$ almost surely, that is, the penalized least squares estimator $\hat{\theta}_{n,\lambda_n}$ is strongly consistent.

B *Proof of Asymptotic Normality of Theorem* 3.2

The estimators $\hat{\theta}_{n,\lambda_n}$ minimize

$$Q_{n,\lambda_n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\tau\theta)^2 + \sum_{r=1}^{p}\lambda_{nr}\theta^\tau S_r\theta.$$

For consistent estimators $\hat{\theta} = \hat{\theta}_{n,\lambda_n}$, we have:

$$0 = \left.\frac{\partial Q_{n,\lambda_n}}{\partial\theta}\right|_{\hat{\theta}} = -\frac{1}{n}\sum_{i=1}^{n}(y_i - z_i^\tau\theta_0)z_i + \sum_{r=1}^{p}\lambda_{nr}S_r\theta_0 + \left(\frac{1}{n}\sum_{i=1}^{n}z_iz_i^\tau + \sum_{r=1}^{p}\lambda_{nr}S_r\right)(\hat{\theta} - \theta_0).$$

Consequently,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left(\frac{1}{n}\sum_{i=1}^{n}z_iz_i^\tau + \sum_{r=1}^{p}\lambda_{nr}S_r\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_iz_i - \sum_{r=1}^{p}\sqrt{n}\lambda_{nr}S_r\theta_0\right). \tag{A.3}$$

Next, we show that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_iz_i + \sum_{r=1}^{p}\sqrt{n}\lambda_{nr}S_r\theta_0 \xrightarrow{D} N(0, \Omega\sigma_0^2). \tag{A.4}$$

The summand of the first term on the right-hand side is the weighted average of an i.i.d error sequence. Therefore, by the central limit theorem, the first term converges in distribution to $N(0, \sigma_0^2\Omega)$ under Assumption 2; the second term on the right-hand side goes to zeros since $\lambda_{nr} = o(n^{-1/2})$. Thus, (A.4) is established.

Since $\lambda_{nr} = o(n^{-1/2})$ and Assumption 2, we have:

$$\frac{1}{n}\sum_{i=1}^{n}z_iz_i^\tau + \sum_{r=1}^{p}\lambda_{nr}S_r \xrightarrow{P} \Omega. \tag{A.5}$$

From (A.4) and (A.5), applying Slutsky's lemma to expression (A.3), Theorem 3.2 follows.

C *The Proof of Theorem* 4.2

Let $H = G^\tau G + \sum_{r=1}^{p}\lambda_rS_r$ and $\eta_r = \log(\lambda_r)$. From the definition of U_1 and U_2, we have:

$$G = QU_1DV^\tau \tag{A.6}$$

and

$$H = VDU_1^\tau Q^\tau QU_1DV^\tau + VDU_2^\tau U_2DV^\tau = VD^2V^\tau. \tag{A.7}$$

Hence,

$$\frac{\partial H^{-1}}{\partial \eta_i} = -H^{-1}\frac{\partial H}{\partial \eta_i}H^{-1} = -\lambda_i VD^{-2}V^\tau S_i VD^{-2}V^\tau \tag{A.8}$$

and

$$\frac{\partial A}{\partial \eta_i} = G\frac{\partial H^{-1}}{\partial \eta_i}G^\tau = -\lambda_i QU_1 D^{-1}V^\tau S_i VD^{-1}U_1^\tau Q^\tau. \tag{A.9}$$

For the second derivatives, we have:

$$\frac{\partial^2 H^{-1}}{\partial \eta_i \partial \eta_j} = H^{-1}\frac{\partial H}{\partial \eta_j}H^{-1}\frac{\partial H}{\partial \eta_i}H^{-1} - H^{-1}\frac{\partial^2 H}{\partial \eta_i \partial \eta_j}H^{-1} + H^{-1}\frac{\partial H}{\partial \eta_i}H^{-1}\frac{\partial H}{\partial \eta_j}H^{-1}.$$

Consequently,

$$\begin{aligned}\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} &= G\frac{\partial^2 H^{-1}}{\partial \eta_i \partial \eta_j}G^\tau \\ &= \lambda_i\lambda_j QU_1 DV^\tau VD^{-2}V^\tau S_j VD^{-2}V^\tau S_i VD^{-2}V^\tau VDU_1^\tau Q^\tau \\ &\quad + \lambda_i\lambda_j QU_1 DV^\tau VD^{-2}V^\tau S_i VD^{-2}V^\tau S_j VD^{-2}V^\tau VDU_1^\tau Q^\tau \\ &\quad - \delta_{ij}\lambda_i QU_1 DV^\tau VD^{-2}V^\tau S_i VD^{-2}V^\tau VDU_1^\tau Q^\tau \\ &= \lambda_i\lambda_j QU_1 D^{-1}V^\tau (S_i VD^{-2}V^\tau S_j + S_j VD^{-2}V^\tau S_i)VD^{-1}U_1^\tau Q^\tau + \delta_{ij}\frac{\partial A}{\partial \eta_i},\end{aligned} \tag{A.10}$$

where $\delta_{ij} = 1$ if $i = j$ and zero otherwise.

By $\alpha = \|Y - AY\|^2$, we have:

$$\frac{\partial \alpha}{\partial \eta_i} = 2Y^\tau \frac{\partial A}{\partial \eta_i}AY - 2Y^\tau \frac{\partial A}{\partial \eta_i}Y$$

and

$$\frac{\partial^2 \alpha}{\partial \eta_i \partial \eta_j} = 2Y^\tau \frac{\partial^2 A}{\partial \eta_i \partial \eta_j}AY + 2Y^\tau \frac{\partial A}{\partial \eta_i}\frac{\partial A}{\partial \eta_j}Y - 2Y^\tau \frac{\partial^2 A}{\partial \eta_i \partial \eta_j}Y.$$

By $\delta = n - \mathrm{tr}(A)$, we have:

$$\frac{\partial \delta}{\partial \eta_i} = -\mathrm{tr}\left(\frac{\partial A}{\partial \eta_i}\right) \quad \text{and} \quad \frac{\partial^2 \delta}{\partial \eta_i \partial \eta_j} = -\mathrm{tr}\left(\frac{\partial^2 A}{\partial \eta_i \partial \eta_j}\right),$$

From (A.9) and (A.10), Eqs. (4.3)–(4.6) can be obtained by some rather tedious manipulation.

Acknowledgments

Lu's research was supported by a grant from the National Natural Science Foundation of China (No. 10501053). Zhu's research was supported by a NSF

grant from National Natural Science Foundation of China (No. 10701035), and ChenGuang project of Shanghai Education Development Foundation (No. 2007CG33). The authors thank the editor and referees for their constructive and detailed suggestions that led to significant improvement in the presentation of the article.

References

Cai, Z., Fan, J., Yao, Q. (2000a). Functional-coefficient regression models for nonlinear times series. *J. Amer. Statist. Assoc.* 95:941–956.

Cai, Z., Fan, J., Li, R. (2000b). Efficient estimation and inference for varying-coefficient models. *J. Amer. Statist. Assoc.* 95:888–902.

Chen, R., Tsay, R. S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* 88:298–308.

Chiang, C.-T., Rice, J. A., Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measure dependent variables. *J. Amer. Statist. Assoc.* 96:605–619.

Eilers, P. H. C., Marx, B. D. (1996). Flexible smoothing with B-spline and penalties. *Statist. Sci.* 11:89–121.

Fan, J., Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* 27:1491–1518.

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

Gu, C., Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Netwon method. *SIAM J. Sci. Statist. Comp.* 12:383–398.

Haggan, V., Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive times series model. *Biometrika* 68:189–196.

Hastie, T., Tibshirani, R. J. (1993). Varying-coefficient model. *J. Roy. Statist. Soc. Ser. B.* 55:757–796.

Hoover, D. R., Rice, J. A., Wu, C. O., Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*. 85:809–822.

Huang, J. Z., Wu, C. O., Zhou, L. (2002). Varying-coefficient models and basis functions approximations for the analysis of repeated measurements. *Biometrika* 89:111–128.

Lu, Y., Mao, S. (2004). Local asymptotics for B-spline estimators of the varying-coefficient model. *Commun. Statist. Theor. Meth.* 33:1119–1138.

Parker, R. L., Rice, J. A. (1985). Discussion on 'Some aspects of the spline smoothing approach to nonparametric regression curve fitting' (by B. W. Silverman). *J. Roy. Statist. Soc. B* 47:40–42.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Computat. Graph. Statist.* 11:735–757.

Wahba, G. (1980). Spline bases, regularization, and generalized cross validation for solving approximation problems with large quanties of noisy data. In: Cheney, W., ed. *Approximation Theory III*. New York: Academic Press, pp. 905–912.

Wood, S. N. (2000). Modelling ang smoothing parameter estimation with multiple quadratic penalties. *J. Roy. Statist. Soc. B* 62:413–428.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* 99:673–686.

Wu, C. O., Chiang, C., Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* 93:1388–1403.

Yu, Y., Ruppert, D. (2002). Penalized spline estimation for partially linear single index models. *J. Amer. Statist. Assoc.* 97:1042–1054.

Copyright of Communications in Statistics: Theory & Methods is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Copyright of Communications in Statistics: Theory & Methods is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.