Published in final edited form as:

Comput Stat Data Anal. 2006 February 10; 50(3): 675–687.

Estimation in Partially Linear Models and Numerical Comparisons¹

Hua Liang

Department of Biostatistics, St. Jude Children's Research Hospital, 332 North Lauderdale St.,
Memphis, TN 38105-2794

Hua Liang: hua.liang@stjude.org

Abstract

Partially linear models with local kernel regression are popular non-parametric techniques. However, bandwidth selection in the models is a puzzling topic that has been addressed in literature with the use of undersmoothing and regular smoothing. In an attempt to address the strategy of bandwidth selection, we review profile-kernel based and backfitting methods for partially linear models, and justify why undersmoothing is necessary for backfitting method and why the “optimal” bandwidth works out for profile-kernel based method. We suggest a general computation strategy for estimating nonparametric functions. We also employ the penalized spline method for partially linear models and conduct intensive simulation experiments to explore the numerical performance of the penalized spline method, profile and backfitting methods. A real example is analyzed with the three methods.

Key words and Phrases

Local linear; Penalized spline; Bandwidth selection; Undersmooth; Profile-kernel based; Linear mixed-effects model

1 Introduction

Partially linear models are, in form, special cases of the additive regression models (Hastie and Tibshirani, 1990; Stone, 1985), which assume that the relation between the response variables and the covariates can be represented as

$$Y_i = X_i^T \beta + g(T_i) + \varepsilon_i, \quad (1)$$

where X_i are d -vector covariates, T_i are scalar covariates, the function $g(\cdot)$ is unknown, and the model's errors ε_i are independent, with a conditional mean zero given the covariates. The models allow easier interpretation of the effect of each variables and may be preferable to a completely nonparametric model since the well-known reason “*curse of dimensionality*”. On the other hand, partially linear models are more flexible than the standard linear model because they combine both parametric and nonparametric components when it is believed that the response variable Y depends on variable X in a linear way but is nonlinearly related to other independent variable T . Another appealing feature of partially linear models is that computation is remarkably easier for partially linear models than for additive models, in which iterative

¹This research was partially supported by the American Lebanese Syrian Associated Charities, and a NIH grant R01 AI 62247-01.

approaches such as a backfitting algorithm (Hastie and Tibishirani, 1990) or marginal integration (Linton and Nielsen, 1995) are necessary. In partially linear models, a closed form of the estimator of β can be explicitly described out even with the backfitting algorithm.

Partially linear models have been paid a great attention in the past decade, since Engle et al. (1986) proposed these models and used them to analyze the relation between electricity usage and average daily temperature. They have been widely studied in the literature. See, for example, the work of Cuzick (1992), Carroll et al. (1997) Severini and Staniswalis (1994), Gao and Anh (1999), and Li (2000), among others. The models have been applied in economics (Schmalensee and Stoker, 1999), biometrics (Zeger and Diggle, 1994), and environmental science (Prada-Sánchez et al., 2000). More recently, Lin and Carroll (2001) considered semiparametric partially generalized linear models for clustered data.

Up to now, several methods have been proposed to consider partially linear models. Engle et al. (1986), Heckman (1986), and Rice (1986) used the smoothing spline technique. Robinson (1988) constructed a feasible least-squares estimator of β by estimating the nonparametric component by a Nadaraya-Waston kernel estimator. Speckman (1988) introduced the idea of a profile least-squares method. Hamilton and Truong (1997) developed the local linear method. Härdle, Liang and Gao (2000) systematically summarized the results obtained to date for partially linear models.

Kernel regression, including local constant (Speckman, 1988) and local linear techniques (Hamilton and Truong, 1997; Opsomer and Ruppert, 1999) has also been used to study the partially linear models. A remarkable characteristic of the kernel-based methods is that under-smoothing has been taken in order to get the root- n estimator of β sometimes (Green et al., 1985; Opsomer and Ruppert, 1999), and this restriction may be canceled (Speckman, 1988; Severini and Staniswalis, 1994). These can be confusing, and it may not be clear for users that which strategy of bandwidth selection is appropriate. We present here a way of clarifying the essential differences by reviewing the profile-kernel based estimator, briefly mentioning backfitting estimator, and analyzing the reasons for applying undersmoothing and regular smoothing. We also suggest an alternative estimator of the nonparametric component and develop the penalized spline with a linear mixed-effects (LME) framework in partially linear models.

The paper is organized as follows. Section 2 reviews the profile-kernel based and backfitting estimators. Empirical-bias bandwidth selection (EBBS) is mentioned. Section 3 develops the penalized spline for partially linear models. Numerical performance of these methods is intensively explored in Section 4. Section 5 investigates a real example to illustrate the approaches.

2 Kernel-based Profile and Backfitting Methods

Note that $E(Y|T) = \{E(X|T)\}^T \beta + g(T)$. It follows that $Y - E(Y|T) = \{X - E(X|T)\}^T \beta + \varepsilon$. An intuitive estimator of β may be defined as the least squares estimator after appropriately estimating $E(Y|T)$ and $E(X|T)$.

In what follows, for any random variable (vector) ζ , let $\hat{E}(\zeta|T)$ be a kernel regression estimator of $E(\zeta|T)$, and let $\tilde{\zeta} = \zeta - E(\zeta|T)$, $\Sigma_{X|T} = \text{cov}\{X - E(X|T)\}$. For example, $\tilde{X}_i = X_i - E(X_i|T_i)$. Denote $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{X}$, $\mathbf{g}$, and $\mathbf{T}$ similarly. Let $m_x(t) = E(X|T = t)$, $m_y(t) = E(Y|T = t)$, and

$$\Psi(m_x, m_y, \beta, Y, X, T) = \{X - m_x(T)\} \left[Y - m_y(T) - \{X - m_x(T)\}^T \beta \right]. \quad (2)$$

The profile-kernel based estimator $\hat{\beta}_P$ solves the estimating equations

$$0 = \sum_{i=1}^n \Psi(\hat{m}_x, \hat{m}_y, \beta, Y_i, X_i, T_i),$$

where $\hat{m}_x$ and $\hat{m}_y$ are two kernel estimators of m_x and m_y . There are a host of alternative methods for estimating m_x and m_y , including higher degree local polynomial kernel methods, kernel methods with varying bandwidths, smoothing and regression splines, etc. Specman (1988) studied local constant smoother. We use local linear smoothers with fixed bandwidths in this paper. The same results will apply for any kernel-based method.

We establish the asymptotic normality of $\hat{\beta}_P$ under the following assumptions.

Assumption 2.1

- $\Sigma_{X|T}$ is a positive-definite matrix, $E(\varepsilon|X, T) = 0$, and $E(\varepsilon^2|X, T) < \infty$;
- The bandwidths estimating $m_x(t)$ and $m_y(t)$ are of order $n^{-1/5}$;
- $K(\bullet)$ is a bounded symmetric density function with compact support and satisfies that $\int K(u)du = 1$, $\int K(u)udu = 0$ and $\int u^2 K(u)du = 1$;
- The density function of $T, f(t)$, is bounded away from 0 and has a bounded continuous second derivative;
- $m_y(t)$, $m_x(t)$, and $g(t)$ have bounded and continuous second derivatives.

Theorem 2.1

Under Assumption 2.1, $\sqrt{n}(\hat{\beta}_P - \beta)$ is asymptotically normally distributed with mean zero and covariance matrix $\sum_{X|T}^{-1} E(\varepsilon^2 \tilde{X} \tilde{X}^T) \sum_{X|T}^{-1}$.

The proof is given in the Appendix. The result is the same as that of Speckman (1988). If ε is independent of (X, T) and has variance σ^2 , the covariance matrix is $\sum_{X|T}^{-1} \sigma^2$. Checking the proof of Theorem 2.1, we find that, for any estimators $\hat{m}_x(t)$ and $\hat{m}_y(t)$ that satisfy $\sup_t |\hat{m}_x(t) - m_x(t)| = o_p(n^{-1/4})$ and $\sup_t |\hat{m}_y(t) - m_y(t)| = o_p(n^{-1/4})$, Theorem 2.1 still holds.

Theorem 2.1 indicates that the profile estimator $\hat{\beta}_P$ is asymptotically normal for the usual optimal bandwidth $n^{-1/5}$. This seems to contradict the conclusion made by Opsomer and Ruppert (1999). Before investigating further, we introduce Opsomer and Ruppert's estimators, $\hat{\beta}_{\text{back}}$, of β and $g(t)$ as follows. A detailed discussion of the two estimators are referred to their article. Although the two estimators are gotten from a backfitting algorithm, there are closed forms and calculation is unnecessarily iterative. However, we still call them "backfitting" estimators in what follows.

Denote $\mathbf{Z}_t = \begin{Bmatrix} 1 & (T_1 - t) \\ \vdots & \vdots \\ 1 & (T_n - t) \end{Bmatrix}$, $\mathbf{W}_t = \text{diag}\{K_h(T_1 - t), \dots, K_h(T_n - t)\}$ with $K_h(\cdot) = 1/hK(\cdot/h)$, and $\mathbf{S}_t = \mathbf{e}_1^T (\mathbf{Z}_t^T \mathbf{W}_t \mathbf{Z}_t)^{-1} \mathbf{Z}_t^T \mathbf{W}_t$ with $\mathbf{e}_1 = (1, 0)^T$, and $\mathbf{S} = (\mathbf{S}_{T_1}^T, \dots, \mathbf{S}_{T_n}^T)^T$. The backfitting estimators $\hat{\beta}_{\text{back}}$ and $\hat{g}_{\text{back}}$ are explicitly expressed as follows:

$$\begin{aligned}\widehat{\beta}_{\text{back}} &= \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1} \mathbf{X}^T(\mathbf{I} - \mathbf{S})\mathbf{Y}, \\ \widehat{\mathbf{g}}_{\text{back}} &= \mathbf{S}(\mathbf{Y} - \mathbf{X}^T \widehat{\beta}_{\text{back}}),\end{aligned}$$

where and in the sequel, $\mathbf{I}$ denotes the $n \times n$ identity matrix.

If Assumption 2.1 holds, except replacement of (b) by $h \approx n^s$ for $-1 < s < -1/4$, β_{back} is still asymptotically normal with the same limit distribution as that of β_P . The proof and discussion are given in Opsomer and Ruppert (1999). This statement indicates that to be root- n consistent, undersmoothing is unavoidable. We analyze this necessity and the unrestriction for β_P as follows.

When we calculate the bias of $\hat{\beta}_{\text{back}}$, the essential term is $\mathbf{X}^T(\mathbf{I} - \mathbf{S})\mathbf{g}/n$. Each element of this vector is the same order as $h^2 + o_P(h^2)$. However, the bias of $\hat{\beta}_P$ is determined by $\mathbf{X}^T(\mathbf{I} - \mathbf{S}_1)(\mathbf{I} - \mathbf{S}_2)\mathbf{g}/n$, where $\mathbf{S}_1$ and $\mathbf{S}_2$ are two projection smoother matrices like $\mathbf{S}$ in $\hat{\beta}_{\text{back}}$. Note that each element of the $d \times n$ matrix $\mathbf{X}^T(\mathbf{I} - \mathbf{S}_1) - \{\mathbf{X} - E(\mathbf{X}|\mathbf{T})\}^T$ is of order $h^2 + o_P(h^2)$, and each element of the n -vector $(\mathbf{I} - \mathbf{S}_2)\mathbf{g}$ is of order $h^2 + o_P(h^2)$. A direct derivation yields that each element of the d -vector $\mathbf{X}^T(\mathbf{I} - \mathbf{S}_1)(\mathbf{I} - \mathbf{S}_2)\mathbf{g}/n$ is of order $h^4 + o_P(h^4)$. The biases of $\hat{\beta}_{\text{back}}$ and $\hat{\beta}_P$ are therefore of orders h^2 and h^4 , respectively. The usual “optimal” bandwidth $n^{-1/5}$ therefore results that $\hat{\beta}_{\text{back}}$ is not consistent, but $\hat{\beta}_P$ is. That is why we need to undersmooth for $\hat{\beta}_{\text{back}}$, but undersmoothing is unnecessary for $\hat{\beta}_P$.

The statements on bandwidth selection indicate that after obtaining the estimators $\hat{\beta}_P$ with optimal bandwidths and the estimator $\hat{\beta}_{\text{back}}$ with undersmoothing, if we directly estimate $g(t)$ by $\hat{g}_P(t) = \hat{m}_y(t) - \{\hat{m}_x(t)\}^T \hat{\beta}_P$ and $\hat{g}_{\text{back}}(t) = \mathbf{S}_t(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{back}})$ with the corresponding bandwidths, the mean squared errors of two nonparametric estimators are of the same order $h^4 + O((nh)^{-1})$. The squared bias and variance of $\hat{g}_P(t)$ are of the same order as $O(n^{-4/5})$, whereas the variance of $\hat{g}_{\text{back}}(t)$ dominates the squared bias of $\hat{g}_{\text{back}}(t)$ and is of higher order than $O(n^{-4/5})$ because the corresponding bandwidth is n^s for $s < -1/4$. What we suggest here is that regardless whether $\hat{\beta}_P$ or $\hat{\beta}_{\text{back}}$ is used, we estimate $g(t)$ by regressing $y - \mathbf{X}^T \hat{\beta}_P$ (or $\hat{\beta}_{\text{back}}$) on T with optimal bandwidths. This leads to an appropriate estimator of $g(t)$.

When implementing backfitting estimator $\hat{\beta}_{\text{back}}$ with local linear regression, we adopt the EBBS methods proposed by Ruppert (1997). Its implementation in partially linear models has been suggested by Opsomer and Ruppert (1999) and can be described as follows.

Give a set of bandwidth candidates $H = \{h_1, \dots, h_M\}$ and a d -vector $\mathbf{c}$. Approximate the bias of $\mathbf{c}^T \hat{\beta}$ by a polynomial of h , and regress these M biases at H on $h_1, \dots, h_M$ to obtain the bias of $\mathbf{c}^T \hat{\beta}$, say $b(h)$. The final bandwidth is the solution of

$$\min_{h \in \{h_{k_1}, \dots, h_{M-k_2}\}} \{b^2(h) + \mathbf{c}^T \widehat{\sigma}^2(h) \sum (h) \mathbf{c}\},$$

where $\widehat{\sigma}^2(h) = n^{-1} \sum_{i=1}^n \{Y_i - X_i^T \widehat{\beta}_{\text{back}}(h) - \widehat{g}_{\text{back}}(T_i, h)\}^2$ and $\Sigma(h) = \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1} \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})\mathbf{X}\} \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1}$; k_1 and k_2 are two tuning parameters. More details refer to Opsomer and Ruppert (1999).

3 The Penalized Spline

The penalized spline method for nonparametric regression was developed by Eilers and Marx (1996) and Ruppert and Carroll (2000). Brumback, Ruppert and Wand (1999) connected this idea with LME framework. See also Coull et al. (2001), Kammann and Wand (2003), and Wand (2003). In this section, we employ the penalized spline method for partially linear models.

We approximate $g(t)$ by $g(t, \gamma) = \gamma_0 + \gamma_1 t + \cdots + \gamma_p t^p + \sum_{k=1}^K b_k(t - \zeta_k)_+^p$, where $p \geq 1$ is an integer and $\zeta_1 < \cdots < \zeta_K$ are fixed knots, $a_+ = \max(a, 0)$. Denote $\gamma = (\gamma_0, \dots, \gamma_p)^T$. Consider $Y = X^T \beta + g(T, \gamma) + \varepsilon$. The penalized spline estimator $(\hat{\beta}_{ps}^T, \hat{\gamma}_{ps}^T)^T$ of $(\beta^T, \gamma^T)^T$ is defined as the minimizer of

$$\sum_{i=1}^n \{Y_i - X_i^T \beta - g(T_i, \gamma)\}^2 + \alpha \sum_{k=1}^K b_k^2, \quad (3)$$

where α is a smoothing parameter.

As shown by Brumback et al. (1999), the estimator $(\hat{\beta}_{ps}^T, \hat{\gamma}_{ps}^T)^T$ based on equation (3) is equivalent to the estimator of β based on an LME model,

$$\text{where } \Lambda = \begin{pmatrix} x_{11} & \cdots & x_{1d} & 1 & T_1 & \cdots & T_1^p \\ x_{21} & \cdots & x_{2d} & 1 & T_2 & \cdots & T_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nd} & 1 & T_n & \cdots & T_n^p \end{pmatrix}, \quad Z = \begin{pmatrix} (T_1 - \zeta_1)_+^p & \cdots & (T_1 - \zeta_K)_+^p \\ (T_2 - \zeta_1)_+^p & \cdots & (T_2 - \zeta_K)_+^p \\ \vdots & \ddots & \vdots \\ (T_n - \zeta_1)_+^p & \cdots & (T_n - \zeta_K)_+^p \end{pmatrix},$$

where $b = (b_1, \dots, b_K)^T \sim (0, \sigma_b^2)$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T \sim (0, \sigma_\varepsilon^2)$, and $\alpha = \sigma_\varepsilon^2 / \sigma_b^2$. This fact implies that the penalized spline smoother in the framework of equation (3) is equivalent to a standard LME model. The solution can be obtained through the use of an LME macro available in S-PLUS software.

Assume that Λ is of full rank and we denote it by $(\mathbf{X}, \Lambda_2)$, a direct calculation yields that

$$\hat{\beta}_{ps} = [\mathbf{X}^T \mathbf{V}^{-1/2} \{\mathbf{I} - \mathbf{V}^{-1/2} \Lambda_2 (\Lambda_2^T \mathbf{V}^{-1} \Lambda_2)^{-1} \Lambda_2^T \mathbf{V}^{-1/2}\} \mathbf{V}^{-1/2} \mathbf{X}]^{-1} \cdot [\mathbf{X}^T \mathbf{V}^{-1/2} \{\mathbf{I} - \mathbf{V}^{-1/2} \Lambda_2 (\Lambda_2^T \mathbf{V}^{-1} \Lambda_2)^{-1} \Lambda_2^T \mathbf{V}^{-1/2}\} \mathbf{V}^{-1/2} \mathbf{Y}],$$

where $\mathbf{V} = \mathbf{Z} \mathbf{Z}^T + \alpha \mathbf{I}$.

The penalty parameter α and the number of knots K must be selected in implementing the penalized spline. For these two parameters, α plays a more important role. However, the formulation of a mixed-effects model automatically derives an estimated value of α . We only need to specify the number of knots K . On the other hand, the estimator $\hat{\beta}_{ps}$ is not so sensitive to K as to α . Our computation experience indicates that $\max(10, n=4)$ is a good choice of the value of K and that the results are insensitive to different values of K . With this value of K , the

knots are at equally spaced sample quantiles of $\{T_i\}$. See Ruppert (2002) for a detailed discussion of knot selection.

4 Simulation Experiment

To evaluate the three estimators $\hat{\beta}_{\text{back}}$, $\hat{\beta}_{\text{P}}$, and $\hat{\beta}_{\text{PS}}$, we conduct an intensive simulation experiment to explore their numerical performance. We generate $n = 300$ data from model (1), and assume that $\beta = (1.3, 0.45)^T$. Let $T \sim \text{Uniform}(0, 1)$, and $g(t) = 4.26\{\exp(-3.25t) - 4 \exp(-6.5t) + 3 \exp(-9.75t)\}$.

We consider three cases to see the numerical performance of three methods.

Case 1

$X \sim \text{Normal}(\mathbf{0}, \text{diag}(0.3, 0.4))$ and $\varepsilon \sim \text{Normal}(0, 1)$. X and T are independent.

Case 2

$X \sim \text{Normal}(\mathbf{0}, \text{diag}(0.3, 0.4))$ and $\varepsilon \sim \text{Normal}(0, |t|)$. X and T are independent.

We use this case to show the effect of heteroscedasticity on the three estimators.

Case 3

$x_{ij} = 0.4T_i + 0.6U_i$ for $j = 1, 2$, $U \sim \text{Uniform}(0, 1)$, and $\varepsilon \sim \text{Normal}(0, 1)$. We use this case to explore the effect of correlation between X and T on the three estimators.

At each configuration, 500 independent sets of data are generated. In our simulation and consideration of a real data example later, we use the Epanechnikov kernel function $k(u) = 15/16(1 - u^2)^2 I_{(|u| \leq 1)}$. We obtain the estimates of the parametric and nonparametric components for each of three estimation methods. When implementing $\hat{\beta}_{\text{P}}$, we estimate $m_x(t)$ and $m_y(t)$ by using the usual local linear regression and select bandwidths by using the plug-in method (Ruppert et al. 1995). When implementing $\hat{\beta}_{\text{back}}$, we select bandwidths by using the EBBS method. The estimates of β are summarized in Table 1. We also present the box plots of the parametric estimates for $N = 500$ in case 1 in Figure 1. Outcomes from case 2 is similar and the boxplot is not reported. The estimates $\hat{\beta}_{\text{back}}$ from case 3 are more stable than those from cases 1 and 2 and the boxplot is omitted. These results indicate that the estimated values are highly exact. The variance of $\hat{\beta}_{\text{P}}$ is slightly larger than those of $\hat{\beta}_{\text{back}}$ and $\hat{\beta}_{\text{PS}}$. When the error is not homogeneous (error distribution $\text{Normal}(0, |t|)$), the estimators given in Section 2 are theoretically not efficient. However, the simulation experiment results are still satisfactory.

In most cases, we are more interested in the parameter β and take the $g(t)$ as a nuisance parameter. However, the estimates of the nonparametric function $g(t)$ sometimes provide useful information in practice. A question therefore arises how about the estimates of $g(\cdot)$ obtained according to our suggestion. Figure 2 shows us the estimates of $g(t)$ after selecting appropriately smoothing parameters for estimating β in case 1. The solid, dotted, dashed, and long-dashed lines indicate the true curve and the estimates based on the backfitting, profile, and penalized spline techniques. The estimated curves are indistinguishable from the true curve. Outcomes from cases 2 and 3 are similar and are not reported. When we estimate $g(t)$ by $\hat{g}_{\text{back}}(t)$ or $\hat{g}_{\text{P}}$ directly, their biases are larger than those obtained in the present context, especially for the backfitting method (not shown here).

5 Real Data Example

In this section we analyze a real data example by using the three estimation methods. The real data are from a study of the relation between the log-earnings of an individual and personal

characteristics (sex, marital status) and measures of a person's human capital, such as schooling and labor market experience. See Rendtel and Schwarze (1995) for a detailed description of the data. Experience suggests that there is a non-linear relation between log-earnings and labor market experience, which therefore plays the role of the variable T in (1). The wage curve is obtained by including the local unemployment rate as an additional regressor, that may have a nonlinear influence. Rendtel and Schwarze (1995), for instance, estimate $g(\bullet)$ as a function of the local unemployment rate by using smoothing-splines. We study the relation by using the model

$$\ln(Y) = X\beta + g(T) + \varepsilon,$$

where X is a dummy variable indicating the level of secondary schooling a person has completed and T is a measure of labor market experience defined as the number of years spent in the labor market and approximated by subtracting (years of schooling + 6) from a person's age. The estimate of β can be interpreted as the rate of return from obtaining the respective level of secondary schooling. Human capital theory suggests that $g(T)$ should take concave form: rapid human capital accumulation in the early stage of one's labor market career is associated with rising earnings that peak somewhere during mid-life and flatten thereafter as hours worked and the incentive to invest in human capital decrease.

The estimates of β based on the backfitting, profile-kernel, and penalized spline methods are $\hat{\beta}_{\text{back}} = 0.0942$ (s.e. 0.0085), $\hat{\beta}_P = 0.0914$ (s.e. 0.0089), and $\hat{\beta}_{\text{PS}} = 0.0912$ (s.e. 0.0083). Standard errors are obtained by 500 bootstrap resample. The estimates of $g(T)$ are reported in Figure 3. The results from three methods are similar. We also fit a linear regression model to the data and obtain a mean function $\ln(y) = 0.594 + 0.0964x + 0.0118t$. The estimated coefficient of x is close to those obtained from partially linear modeling. But this linear regression does not reflect the changing trend of the log-earnings over the labor market experience. To investigate the effects of different values of K on the $\hat{\beta}_{\text{PS}}$, we also run this data set for $K = 2, 5, 8$, and 15 , and obtain $\hat{\beta}_{\text{PS}; K=2} = 0.09108$ (s.e. 0.00831); $\hat{\beta}_{\text{PS}; K=5;8;15} = 0.09121$ (s.e. 0.00832). The corresponding estimates of $g(T)$ are described in Figure 4. We therefore conclude that the penalized spline estimates are very robust to differing numbers of knots.

6 Conclusions and Discussion

In this paper we reviewed the profile-kernel based and backfitting methods and developed a penalized spline method for partially linear models. We analyzed the essential reasons that why the backfitting estimation needs undersmoothing, whereas the profile-kernel based method avoids this restriction, in order to have asymptotically normal estimators. We also suggested a strategy for effectively estimating the nonparametric component. We numerically compared the estimators obtained with backfitting, profile-kernel based, and penalized spline methods.

The plug-in and EBBS methods were used to select bandwidths in the different situations. The main advantages of the penalized spline are that selection of smoothing parameters can be avoided, computation is fast, its numerical performance is comparable to that of other methods. In conclusion, we recommend the penalized spline method because of its robustness in implementation, its computational speediness, and its comparability to other methods. We believe that the penalized spline is methodologically and practically valuable.

References

- Brumback BA, Ruppert D, Wand M. Comment on Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-based Prior by Shively, Kohn, and Wood. *Journal of the American Statistical Association* 1999;94:794–797.
- Carroll RJ, Fan JQ, Gijbels I, Wand MP. Generalized Partially Linear Single-index Models. *Journal of the American Statistical Association* 1997;92:477–489.
- Coull BA, Schwartz J, Wand MP. Respiratory Health and Air Pollution: Additive Mixed Model Analyses. *Biometrics* 2001;2:337–349.
- Cuzick J. Semiparametric Additive Regression. *Journal of the Royal Statistical Society, Series B* 1992;54:831–843.
- Engle RF, Granger CWJ, Rice J, Weiss A. Semiparametric Estimates of the Relation between Weather and Electricity Sales. *Journal of the American Statistical Association* 1986;81:310–320.
- Eilers PHC, Marx BD. Flexible Smoothing with B-splines and Penalties (with comments). *Statistical Science* 1996;11:89–121.
- Gao JT, Anh V. Semiparametric regression under long-range dependent errors. *Journal of Statistical Planning and Inference* 1999;80:37–57.
- Green P, Jennison C, Seheult A. Analysis of Field Experiments by Least-squares Smoothing. *Journal of the Royal Statistical Society, Series B* 1985;47:299–315.
- Hamilton SA, Truong YK. Local Linear Estimation in Partly Linear Models. *Journal of Multivariate Analysis* 1997;60:1–19.
- Härdle, W.; Liang, H.; Gao, JT. Partially Linear Models. Heidelberg: Springer Physica-Verlag; 2000.
- Hastie, TJ.; Tibshirani, RJ. Generalized Additive Models. Vol. 43. London: Chapman and Hall; 1990.
- Heckman NE. Spline Smoothing in Partly Linear Models. *Journal of the Royal Statistical Society, Series B* 1986;48:244–248.
- Li Q. Efficient Estimation of Additive Partially Linear Models. *International Economic Review* 2000;41:1073–1092.
- Lin XH, Carroll RJ. Semiparametric Regression for Clustered Data. *Biometrika* 2001;88:1179–1185.
- Linton OB, Nielsen JP. A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration. *Biometrika* 1995;82:93–101.
- Kammann EE, Wand MP. Geoadditive Models. *Journal of the Royal Statistical Society, Series C* 2003;52:1–18.
- Newey WK. The Asymptotic Variance of Semiparametric Estimators. *Econometrica* 1994;62:1349–1382.
- Opsomer JD, Ruppert D. A Root- n Consistent Backfitting Estimator for Semiparametric Additive Modelling. *Journal of Computational and Graphical Statistics* 1999;8:715–732.
- Prada-Sánchez JM, Febrero-Bande M, Cotos-Yáñez T, González-Manteiga W, Bermúdez-Cela JL, Lucas-Dominguez T. Prediction of SO₂ Pollution Incidents near a Power Station Using Partially Linear Models and an Historical Matrix of Predictor-response Vectors. *Environmetrics* 2000;11:209–225.
- Rendtel, U.; Schwarze, J. Discussion Paper 118. German Institute for Economic Research (DIW); Berlin: 1995. Zum Zusammenhang zwischen Lohnhöhe und Arbeitslosigkeit: Neue Befunde auf Basis Semiparametrischer Schätzungen und eines Verallgemeinerten Varianz-Komponenten Modells.
- Rice JA. Convergence Rates for Partially Splined Models. *Statistics and Probability Letters* 1986;4:203–208.
- Robinson PM. Root- n -consistent Semiparametric Regression. *Econometrica* 1988;56:931–954.
- Ruppert D. Empirical-bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation. *Journal of the American Statistical Association* 1997;92:1049–1062.
- Ruppert D. Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics* 2002;11:735–757.
- Ruppert D, Carroll R. Spatially-adaptive Penalties for Spline Fitting. *Australian and New Zealand Journal of Statistics* 2000;42:205–223.

- Ruppert D, Sheather SJ, Wand MP. An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association* 1995;90:1257–1270.
- Schmalensee R, Stoker TM. Household Gasoline Demand in the United States. *Econometrica* 1999;67:645–662.
- Severini TA, Staniswalis JG. Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association* 1994;89:501–511.
- Speckman P. Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society, Series B* 1988;50:413–436.
- Stone CJ. Additive Regression and Other Nonparametric Models. *The Annals of Statistics* 1985;13:689–705.
- Wand MP. Smoothing and Mixed Models. *Computational Statistics* 2003;18:223–249.
- Zeger SL, Diggle PJ. Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters. *Biometrics* 1994;50:689–699. [PubMed: 7981395]

Appendix: Proof of Theorems 2.1

We finish the proof of the theorem by using Newey's (1994) results, although a direct proof is also available. Denote $\|\bullet\|$ a norm for a function, such as Sobolev norm, that is a supremum norm for a function and its derivatives. It follows from Assumption 2.1 that $\|\hat{m}_x(\bullet) - m_x(\bullet)\| = o_p(n^{-1/4})$ and $\hat{m}_y(\bullet) - m_y(\bullet) = o_p(n^{-1/4})$. This is assumption 5.1(ii) of Newey (1994). Let

$$D(m_x^* - m_x, m_y^* - m_y, \beta, Y, X, T) = \frac{\partial \Psi}{\partial m_x}(m_x^* - m_x) + \frac{\partial \Psi}{\partial m_y}(m_y^* - m_y).$$

where $\frac{\partial \Psi}{\partial m_x}$ and $\frac{\partial \Psi}{\partial m_y}$ are the Frechet derivatives. A direct calculation derives that $E\left(\frac{\partial \Psi}{\partial m_x}\right) = 0$ and $E\left(\frac{\partial \Psi}{\partial m_y}\right) = 0$. Furthermore,

$$\begin{aligned} & \left\| \Psi(m_x^*, m_y^*, \beta, Y, X, T) - \Psi(m_x, m_y, \beta, Y, X, T) - D(m_x^* - m_x, m_y^* - m_y, \beta, Y, X, T) \right\| \\ &= O_p\left(\left\|m_x^* - m_x\right\|^2 + \left\|m_y^* - m_y\right\|^2\right). \end{aligned} \quad (\text{A.1})$$

Equation (A.1) is Newey's assumption 5.1(i). His assumption 5.2 holds by the expression of $D(\bullet, \beta, Y, X, T)$. In addition, it follows from the above statements that for any (m_x^*, m_y^*) ,

$$E\left\{D(m_x^* - m_x, m_y^* - m_y, \beta, Y, X, T)\right\} = 0,$$

thus verifying Newey's assumption 5.3: his $\alpha(T) = 0$, according to his discussion just above his Lemma 5.1. By that lemma, it follows that $\hat{\beta}_P$ has the same distribution as the solution to the equation

$$0 = \sum_{i=1}^n \Psi(m_x, m_y, \beta, Y_i, X_i, T_i). \quad (\text{A.2})$$

It is easy to show that the solution of (A.2) has the same limit distribution as described in the statement of Theorem 2.1. This completes the proof.

Figure 1.

Box plots of the estimates ($N=500$ simulation runs) for backfitting (back), profile, and penalized spline (PS) methods for homogeneous errors. Upper panel: the first element of β ; lower panel: the second element of β . The long horizontal lines are the true values.

Figure 2. Estimated curves of the nonparametric components for simulated data. The true curve is shown as an unbroken line and the curves estimated with backfitting, profile, and penalized spline methods are shown by dotted, dashed, long-dashed lines, respectively.

Figure 3.
Relation between log-earnings and labor-market experience.

Figure 4. Penalized spline estimates of the nonparametric component from real data with different numbers of knots.

Table 1

Results of the simulation study. “estimates” is the simulation mean, “s.e.” is the simulation standard error.

Case		β_{back}	β_P	β_{PS}	
1	estimates	1.299	0.452	1.297	0.451
	s.e.	0.0189	0.0142	0.0304	0.0167
2	estimates	1.299	0.450	1.296	0.451
	s.e.	0.0259	0.0201	0.0333	0.0296
3	estimates	1.296	0.448	1.32	0.459
	s.e.	0.1085	0.0849	0.101	0.082