

Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data

BY ZONGHUI HU, NAISYIN WANG AND RAYMOND J. CARROLL

*Department of Statistics, Texas A&M University, College Station, Texas 77843-3143,
U.S.A.*

zhu@stat.tamu.edu nwang@stat.tamu.edu carroll@stat.tamu.edu

SUMMARY

We study the profile-kernel and backfitting methods in partially linear models for clustered/longitudinal data. For independent data, despite the potential root- n inconsistency of the backfitting estimator noted by Rice (1986), the two estimators have the same asymptotic variance matrix, as shown by Opsomer & Ruppert (1999). In this paper, theoretical comparisons of the two estimators for multivariate responses are investigated. We show that, for correlated data, backfitting often produces a larger asymptotic variance than the profile-kernel method; that is, for clustered data, in addition to its bias problem, the backfitting estimator does not have the same asymptotic efficiency as the profile-kernel estimator. Consequently, the common practice of using the backfitting method to compute profile-kernel estimates is no longer advised. We illustrate this in detail by following Zeger & Diggle (1994) and Lin & Carroll (2001) with a working independence covariance structure for nonparametric estimation and a correlated covariance structure for parametric estimation. Numerical performance of the two estimators is investigated through a simulation study. Their application to an ophthalmology dataset is also described.

Some key words: Asymptotic distribution; Bandwidth; Kernel estimation; Local linear estimation; Smoother matrix; Under-smoothing.

1. INTRODUCTION

The partially linear model has been investigated intensively in the literature and various extensions have been proposed; see for example Härdle et al. (2000, § 1.1). There have been two main classes of estimation methods for this model, namely the profile-kernel and backfitting methods. For independent data, Severini & Staniswalis (1994) and Carroll et al. (1997), among others, studied the profile-kernel approach. Buja et al. (1989), Hastie & Tibshirani (1990) and Opsomer & Ruppert (1999) investigated the backfitting approach. For clustered data, Severini & Staniswalis (1994) and Lin & Carroll (2001) extended the profile-kernel method to accommodate multivariate responses, as did N. Wang, R. J. Carroll and X. Lin in an unpublished report, while Zeger & Diggle (1994) studied the backfitting method.

On the theoretical front, the asymptotic properties of profile-kernel estimators were provided by Severini & Staniswalis (1994), Lin & Carroll (2001) and by Wang, Carroll and

Lin in their report. Their results also cover the clustered-data scenario. For independent data the bias problem of backfitting estimation was first noted by Rice (1986); see also Speckman (1988) and Opsomer & Ruppert (1999). Their findings indicate that under-smoothing during nonparametric estimation is required for root- n consistent parametric estimation for the backfitting method. Meanwhile, Opsomer & Ruppert (1999) also showed that the two estimators share the same asymptotic variance matrix.

In contrast to profile-kernel methods, properties of backfitting for clustered data are less well understood. In this paper, we investigate the asymptotic properties of the backfitting method for clustered data. In practice, backfitting is often used as a substitute for profile-kernel estimation, perhaps because of their variance equivalence property in the independent case, as well as its simplicity. However, it is unclear whether or not this equivalence still holds for clustered data. The main purpose of this paper is to investigate this issue.

We will make asymptotic comparisons between profile-kernel and backfitting estimation in two contexts, namely generally under a specific but widely applicable condition on the covariance matrix of the clustered data, and specifically under the scenario considered in Zeger & Diggle (1994) and Lin & Carroll (2001). For the latter, we use a working independence correlation structure for the nonparametric estimation and a moment-based estimated covariance structure in parametric estimation: this estimation scheme is commonly used in practice. We will show that, besides the bias problem, for clustered data, the backfitting estimator tends to have larger variance than the profile-kernel estimator; that is, the asymptotic equivalence in variance no longer holds for the multivariate case.

We discuss the two estimation procedures in § 2 and summarise their asymptotic properties in § 3. We demonstrate our theoretical results with a simulation study in § 4, and an application in ophthalmology is given in § 5. Finally, concluding remarks are given in § 6.

2. ESTIMATION PROCEDURES

The partially linear model is

$$Y_{ij} = X_{ij}^T \beta + \theta(T_{ij}) + \varepsilon_{ij}, \quad (1)$$

where the i th cluster, $i = 1, \dots, n$, has m_i observations, β is a $p \times 1$ vector and $\theta(\cdot)$ is an unknown smooth function. Here, the ε_{ij} are random errors and we assume that the ε_{ij} from different clusters are independent. Without loss of generality, we let $m_i = m$ for all i . As in Lin & Carroll (2001), we assume that $E(Y_{ij}|X_i, T_i) = E(Y_{ij}|X_{ij}, T_{ij})$, where $X_i = (X_{i1}, \dots, X_{im})^T$, $T_i = (T_{i1}, \dots, T_{im})^T$ denote the covariates observed from the i th subject; see also Pepe & Couper (1997). Likewise, we have $E(Y_{ij}|T_i) = E(Y_{ij}|T_{ij})$ and denote it by $m_Y(T_{ij})$; $m_X(T_{ij})$ is defined equivalently.

For profile-kernel estimation, for a given β , the estimator of $\theta(T_a)$ is

$$\hat{\theta}(T_a; \beta) = \hat{m}_Y(T_a) - \hat{m}_X(T_a)\beta,$$

where $T_a = (T_1^T, \dots, T_n^T)^T$, and $\hat{m}_Y(T_a)$ and $\hat{m}_X(T_a)$ are nonparametric estimators of $m_Y(T_a)$ and $m_X(T_a)$, respectively. For a function with a scalar argument, such as $\theta(\cdot)$, the notation $\theta(v)$ denotes a vector whose i th element is $\theta(v_i)$. Hereafter, the subscript ‘ a ’ indicates a vector or a matrix being corresponding to all observations.

The parameter β is then estimated by a profile-kernel generalised estimating equation,

$$\sum_{i=1}^n \frac{\partial \{X_i \beta + \hat{\theta}(T_i; \beta)\}^T}{\partial \beta} V_i^{-1}(X_i, T_i) [Y_i - \{X_i \beta + \hat{\theta}(T_i; \beta)\}] = 0,$$

where the V_i 's are the working covariance matrices. The profile-kernel estimators of β and θ are, respectively,

$$\hat{\beta}_P = \left[\sum_{i=1}^n \{X_i - \hat{m}_X(T_i)\}^T V_i^{-1} \{X_i - \hat{m}_X(T_i)\} \right]^{-1} \left[\sum_{i=1}^n \{X_i - \hat{m}_X(T_i)\}^T V_i^{-1} \{Y_i - \hat{m}_Y(T_i)\} \right],$$

$$\hat{\theta}(t) = \hat{m}_Y(t) - \hat{m}_X(t) \hat{\beta}_P.$$

In matrix form, the profile-kernel estimator of β can be written as

$$\hat{\beta}_P = \{X_a^T (I - S_a)^T V_a^{-1} (I - S_a) X_a\}^{-1} X_a^T (I - S_a)^T V_a^{-1} (I - S_a) Y_a, \quad (2)$$

where S_a is a smoother matrix with respect to T_a (Opsomer & Ruppert, 1997) and $V_a = \text{diag}(V_1, \dots, V_n)$ is the block diagonal matrix containing the n working covariance matrices.

For backfitting, at the current value $\hat{\beta}_c$ of β , the updated estimator of θ is

$$\hat{\theta}(T_a; \hat{\beta}_c) = \hat{m}_Y(T_a) - \hat{m}_X(T_a) \hat{\beta}_c,$$

and the updated value of β is obtained by a generalised least squares regression of $Y_i - \hat{\theta}(T_i; \hat{\beta}_c)$ on X_i with the argument β minimising

$$\sum_{i=1}^n \{Y_i - \hat{\theta}(T_i; \hat{\beta}_c) - X_i \beta\}^T V_i^{-1} \{Y_i - \hat{\theta}(T_i; \hat{\beta}_c) - X_i \beta\}.$$

At convergence, the backfitting estimators of β and θ are, respectively,

$$\hat{\beta}_{BF} = \left[\sum_{i=1}^n X_i^T V_i^{-1} \{X_i - \hat{m}_X(T_i)\} \right]^{-1} \left[\sum_{i=1}^n X_i^T V_i^{-1} \{Y_i - \hat{m}_Y(T_i)\} \right],$$

$$\hat{\theta}(t) = \hat{m}_Y(t) - \hat{m}_X(t) \hat{\beta}_{BF}.$$

In matrix form, the backfitting estimator of β is

$$\hat{\beta}_{BF} = \{X_a^T V_a^{-1} (I - S_a) X_a\}^{-1} X_a^T V_a^{-1} (I - S_a) Y_a. \quad (3)$$

For independent data where $V_a = I$ is used, the two estimators for β are

$$\hat{\beta}_P = \{X_a^T (I - S_a)^T (I - S_a) X_a\}^{-1} X_a^T (I - S_a)^T (I - S_a) Y_a,$$

$$\hat{\beta}_{BF} = \{X_a^T (I - S_a) X_a\}^{-1} X_a^T (I - S_a) Y_a. \quad (4)$$

3. ASYMPTOTIC PROPERTIES

Throughout, the number of observations for each subject, m , is regarded as fixed. The usual regularity assumptions on the kernel function are assumed, including that the second moment is assumed to equal 1. We also assume that (Y_i, X_i, T_i) , for $i = 1, \dots, n$, are independent and identically distributed with $f_j(t)$ denoting the marginal density of T_{ij} . Throughout this section, we assume the regularity conditions as in Lin & Carroll (2001) and suppress the index i in the presentation.

The results concerning the comparison of the asymptotic variances of the two estimators can be constructed based on (2) and (3); that is, the results are not restricted to the case of the local linear smoother.

For independent data, as observed in expression (4), the profile-kernel estimator and the backfitting estimator are identical if the smoother matrix S_a is symmetric and idempotent. They are generally different otherwise. However, the two estimators have the same asymptotic variance matrix; see Opsomer & Ruppert (1999). For clustered data, the comparison of the variances of the two estimators can be simplified when V and Σ are functions only of T .

PROPOSITION 1. *Under the assumption that both the working covariance matrix V and the true covariance matrix Σ depend only on T , the asymptotic variance matrix of the backfitting estimator is at least as large as that of profile-kernel estimator; that is, $V_{\text{BF}} - V_{\text{P}}$ is positive semidefinite.*

A sketch proof is given in the Appendix. Proposition 1 shows that, for clustered data, the two estimators may not share the same asymptotic variance matrix, in contrast to the independent case. This result is completely general and does not require a specific choice of working covariance matrix beyond that it does not depend on X . The result also applies to general nonparametric smoothers.

To appreciate better the differences between the two estimators, we now concentrate on the following commonly-used estimation scheme. For nonparametric estimation, we assume a working independence correlation matrix, and, for parametric estimation, we use a working covariance matrix V_i estimated by data. Wang & Wang (2001) and Lin & Carroll (2001) discuss the advantage of variance reduction in using the correlation for parametric estimation versus ignoring the correlation.

The following proposition concerning the profile-kernel method is given in Lin & Carroll (2001). We quote it here to ease comparison with properties of the backfitting method given in Proposition 3. In Propositions 2 and 3, the results are based on using a local linear smoother with working independence in nonparametric estimation. This estimation scheme is also taken for the numerical studies in the following sections.

PROPOSITION 2 (Lin & Carroll, 2001). *Suppose that $h \propto n^{-\alpha}$, $\frac{1}{5} \leq \alpha \leq \frac{1}{3}$ and $n \rightarrow \infty$ and define*

$$\tilde{X} = X = \lim_{n \rightarrow \infty} \partial \hat{\theta}(T; \beta) / \partial \beta.$$

Then $\hat{\beta}_{\text{P}}$ converges in distribution: $\sqrt{n}\{\hat{\beta}_{\text{P}} - \beta + h^2 b_{\text{P}}(\beta, \theta)/2\} \rightarrow N(0, V_{\text{P}})$, where

$$b_{\text{P}}(\beta, \theta) = E(\tilde{X}^T V^{-1} \tilde{X})^{-1} E\{\tilde{X}^T V^{-1} \theta^{(2)}(T)\},$$

$$V_{\text{P}} = E(\tilde{X}^T V^{-1} \tilde{X})^{-1} E\{(Z_1 - Z_2)^T \Sigma (Z_1 - Z_2)\} E(\tilde{X}^T V^{-1} \tilde{X})^{-1}.$$

Here $\tilde{X} = \{X - m_X(T)\}$, $\Sigma = \text{var}(Y|X, T)$, $Z_1 = V^{-1} \tilde{X}$, $Z_2 = (Z_2^1, \dots, Z_2^m)^T$, with

$$Z_2^j = \left\{ \sum_{k=1}^m \sum_{l=1}^m E(\tilde{X}^k V^{kl} | T^l = T^j) \right\} f_j(T^j) / \sum_{l=1}^m f_l(T^j),$$

and V^{kl} denotes the (k, l) entry of V^{-1} .

PROPOSITION 3. Under the same conditions as those of Proposition 2, the backfitting estimator $\hat{\beta}_{\text{BF}}$ converges in distribution: $\sqrt{n}\{\hat{\beta}_{\text{BF}} - \beta + h^2 b_{\text{BF}}(\beta, \theta)/2\} \rightarrow N(0, V_{\text{BF}})$, where

$$b_{\text{BF}}(\beta, \theta) = E(\tilde{X}^T V^{-1} \tilde{X})^{-1} E\{X^T V^{-1} \theta^{(2)}(T)\},$$

$$V_{\text{BF}} = E(\tilde{X}^T V^{-1} \tilde{X})^{-1} E\{(Z_1^* - Z_2^*)^T \Sigma (Z_1^* - Z_2^*)\} E(\tilde{X}^T V^{-1} \tilde{X})^{-1},$$

and $Z_1^* = V^{-1} X$, $Z_2^* = (Z_2^{*1}, \dots, Z_2^{*m})^T$, with

$$Z_2^{*j} = \left\{ \sum_{k=1}^m \sum_{l=1}^m E(X^k V^{kl} | T^l = T^j) \right\} f_j(T^j) / \sum_{l=1}^m f_l(T^j).$$

A sketch proof of Proposition 3 is provided in the Appendix.

For clustered data under the estimation scheme considered, the profile-kernel estimator is in general root- n inconsistent. An exception occurs when working independence is assumed throughout (Lin & Carroll, 2001).

COROLLARY 1. Under the assumption that the working covariance matrix V depends only on T , when h is of regular order $n^{-1/5}$, the profile-kernel estimator is root- n consistent, while the backfitting estimator is root- n inconsistent; under the assumed conditions, $E\{X^T V^{-1} \theta^{(2)}(T)\}$ in b_{BF} remains nonzero.

Corollary 1 is a direct consequence of (A3) with straightforward conditional expectation calculations.

As shown in Proposition 1, the results concerning asymptotic variance matrices of the two estimators apply to general nonparametric smoothers. For independent data, Opsomer & Ruppert (1999) point out that the two estimators have the same asymptotic variance matrix. This is also an easy consequence of Propositions 2 and 3. To see this, note that, for independent data, both Σ and V equal $\sigma^2 I$. In this case, $Z_1 = \sigma^{-2} \tilde{X}$, $Z_2 = \sigma^{-2} E(\tilde{X}|T) = 0$, $Z_1^* = \sigma^{-2} X$ and $Z_2^* = \sigma^{-2} E(X|T)$. Consequently, $Z_1 - Z_2 = Z_1^* - Z_2^* = \sigma^{-2} \tilde{X}$ and the asymptotic variance matrices of the two estimators are $V_p = V_{\text{BF}} = \sigma^2 [E\{\text{cov}(X|T)\}]^{-1}$.

For clustered data, the results in the Appendix indicate that the two asymptotic variance matrices will be the same if and only if

$$E\{m_X(T_a)^T V_a^{-1} (I - S_a) \Sigma_a (I - S_a)^T V_a^{-1} m_X(T_a)\}$$

is zero; that is, a specific structure is required of the smoother matrix. In Lemma 1 of Wang, Carroll and Lin's report, it is shown that the nonparametric smoother of Wang (2003) possesses such a property. The above propositions and Corollary 1 clearly indicate that, under the currently most commonly used estimation scheme, backfitting in general has a larger asymptotic variance than the profile-kernel estimator and is often more biased.

4. SIMULATION STUDY

We conducted a simulation study to evaluate the finite-sample performance of the profile-kernel method versus the backfitting method, again in the specific context that the nonparametric estimation uses working independence. Of course, from our results, we expect the profile-kernel method to have smaller variance in general, not just for this particular choice of smoother.

For the case of clustered data, we generated 500 datasets, each comprising $n = 100$ subjects with $m = 5$ observations per subject. The covariate vectors (T_{ij}, X_{ij}) , for $j = 1, \dots, m$, were independently generated from the bivariate normal distribution with mean 0,

variance 1 and correlation coefficient $0.75^{\frac{1}{2}}$. The Y_{ij} were generated from the partially linear model (1), where $\theta(t) = \sin(2t)$ and $\beta = 1$, with normally distributed error with variance 1 and exchangeable correlation 0.4. For nonparametric estimation, we used local linear kernel estimation with the bandwidth choices 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6, and we assumed working independence. For parametric estimation, the working covariance V_i was set to be the true within-subject covariance of Y_i .

Table 1(a) reports the empirical biases and standard deviations of the estimated β from the profile-kernel and backfitting methods. It shows that the bias of the profile-kernel estimator is negligible over the range of bandwidths, but the bias of the backfitting estimator increases sharply as the bandwidth gets larger. This observation implies that backfitting estimator is more sensitive to bandwidth selection, as suggested by our theory. Table 1(a) also shows that the backfitting estimator has larger empirical standard deviations, about twice the size of the profile-kernel standard deviations. This observation agrees with our general theoretical result in Proposition 1.

Table 1. *Simulation results for 500 sets of clustered data and 500 sets of independent data*

(a) *Clustered data*

Estimator		Bandwidth					
		$h = 0.1$	$h = 0.2$	$h = 0.3$	$h = 0.4$	$h = 0.5$	$h = 0.6$
$\hat{\beta}_P$	Bias	-0.0014	-0.0022	-0.0016	-0.0015	-0.0019	-0.0022
	SD	0.1608	0.1563	0.1539	0.1534	0.1533	0.1530
$\hat{\beta}_{BF}$	Bias	0.0147	0.0641	0.1412	0.2473	0.3801	0.5385
	SD	0.3801	0.3730	0.3671	0.3610	0.3609	0.3625

(b) *Independent data*

Estimator		Bandwidth					
		$h = 0.1$	$h = 0.2$	$h = 0.3$	$h = 0.4$	$h = 0.5$	$h = 0.6$
$\hat{\beta}_P$	Bias	0.0015	0.0078	0.0090	0.0091	0.0096	0.0100
	SD	0.2444	0.2351	0.2328	0.2320	0.2320	0.2316
$\hat{\beta}_{BF}$	Bias	0.0160	0.0802	0.1695	0.2783	0.4153	0.5802
	SD	0.2710	0.2555	0.2548	0.2603	0.2615	0.2663

$\hat{\beta}_P$, profile-kernel estimator; $\hat{\beta}_{BF}$, backfitting estimator.
SD, standard deviation.

As a contrast, a numerical study was also carried out on independent data, where 500 datasets were generated, each comprising 300 subjects. Variables (T_i, X_i) and Y_i were generated in the same way as in the clustered-data case, except that the responses Y_i are independent of each other. The empirical biases and standard deviations from the two methods are reported in Table 1(b).

Table 1(b) shows a similar pattern in bias to that for clustered data, but the backfitting estimator has very similar standard deviations to those of the profile-kernel estimator. This indicates that the two estimators are nearly equally efficient for independent data, which is consistent with the traditional finding.

Another observation from Table 1 is that, since the working covariance matrix V_i used in the clustered-data simulation does not depend on X , the profile-kernel estimator is

actually root- n consistent. This is the situation in Corollary 1. Thus, it is natural that we observe negligible bias from profile-kernel estimation in both the clustered and independent cases.

5. AN APPLICATION IN OPHTHALMOLOGY

In this section we analyse data from a prospective ophthalmology study on the use of intraocular gas in retinal repair surgeries (Meyers et al., 1992; Song & Tan, 2000). Three different volumes of gas were injected into the eye before surgery in a total of 31 patients. The patients were then followed up 3 to 8 times over a 60-day period, and the volume of the gas left in the eye at the follow-up times were recorded as a percentage of the initial gas level in that eye. The issue was to estimate the kinetics of the disappearance of the gas with respect to time. We let the response variable be the arcsin square root transformed percentage of gas left in the eye. The covariates are the initial level of gas concentration in the eye, denoted by X , and the follow-up observation time T , in the unit of days. We then assume that the transformed response follows the partially linear model (1).

Since there seems to exist a positive correlation among responses from the same patient, we need to incorporate a correlation structure into the estimation scheme. From the analysis of the residuals from the initial estimate assuming working independence (Diggle et al., 2002, Ch. 3), we found that the compound symmetry covariance matrix fitted the data reasonably well. The estimated correlation is $\rho = 0.5442$, and the estimated variance is $\sigma^2 = 0.0678$.

The bandwidth h was chosen by ‘leaving one subject out’ crossvalidation (Rice & Silverman, 1991; Härdle et al., 2000, § 2.1.3) using the profile-kernel method. The exact procedure and a short justification of the use of this bandwidth selection method are given in the Appendix. We found that estimates with bandwidth ranging from 6 to 7 performed best and that differences among them were negligible. To ensure that the conclusion was not bandwidth dependent, we carried out the estimation for the bandwidth choices 6, 6.5, 7 and 8. We then applied the profile-kernel and the backfitting estimation methods as described in § 2 to these data, where the estimated compound symmetry working covariance matrix was assumed in the parametric estimation and the local linear smoother was used for nonparametric estimation. The results are given in Table 2.

Based on the results, we see that the percentage of gas volume left in the eye depends positively on the original gas concentration in the eye. The positive estimated values of β indicate that the percentage of gas volume left in the eye is high when the original level

Table 2: *Ophthalmology example. Estimates and standard errors of the parametric coefficient using profile-kernel and backfitting methods*

Estimator		Bandwidth			
		$h = 6.0$	$h = 6.5$	$h = 7.0$	$h = 8.0$
$\hat{\beta}_P$	Estimate	0.1037	0.1024	0.1014	0.1041
	SE	0.0080	0.0072	0.0070	0.0063
$\hat{\beta}_{BF}$	Estimate	0.0898	0.0890	0.0884	0.0879
	SE	0.0118	0.0119	0.0117	0.0151

$\hat{\beta}_P$, profile-kernel estimator; $\hat{\beta}_{BF}$, backfitting estimator.
SE, standard error.

is high. This result is consistent with the findings of Song & Tan (2000). Moreover, both profile-kernel and backfitting estimation show a significant effect from the original gas concentration for all bandwidths considered. Regarding this aspect, the semiparametric model and estimation scheme considered here is different from the result in Song & Tan (2000), where a more complex model involving the same response and covariates suggests that the effect from the original gas concentration is insignificant. Our graphical diagnosis indicates that modelling the transformed responses with a semiparametric partially linear model provides sufficient flexibility to model the data reasonably well. The assumption violation observed in the parametric model considered in Song & Tan, which motivated their model, no longer exists.

The time profile of the percentage of gas left in the eye is reflected by $\theta(t)$ in the semiparametric model, and we plot the estimated curve of $\theta(t)$ based on bandwidth $h = 7$ in Fig. 1. The plots from profile-kernel and backfitting estimation are almost identical and indicate the same decreasing trend.

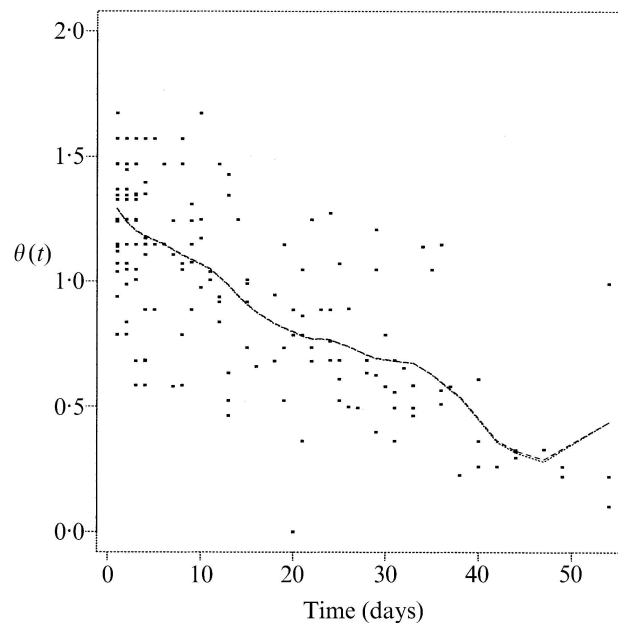


Fig. 1: Ophthalmology example. Fitted curve for $\theta(t)$ by profile-kernel and backfitting methods, shown by dotted and dashed lines respectively, at bandwidth $h = 7$.

Finally, we note that in Table 2, for all bandwidths, the backfitting estimator had larger estimated standard error than the profile-kernel estimator. This observation agrees with the asymptotic properties and the simulation results in §§ 3 and 4. It also suggests that for multivariate data one should no longer use backfitting as a substitute for the profile-kernel method.

6. CONCLUDING REMARKS

The associate editor asked us to comment on the use of kernel methods versus penalised spline approaches as a general statistical methodology, and in particular the implementation of penalised splines via variance-component-model representations. We

will let others comment on the somewhat controversial nature of penalised low-order basis splines versus smoothing splines, knot selection methods without penalties and estimation of smoothing parameters, the spline literature being in no agreement on these points.

The advantages and disadvantages of kernel methods and penalised splines using variance-component-model representations are fairly well known. As made clear by Ruppert et al. (2003, Ch. 1–2), penalised splines have the advantage that they are easily adopted into a wide variety of likelihood-type problems, by incorporating the penalties via a variance-components representation.

However, a variance-component-model representation of penalised splines may not always make sense, as for example in the marginal generalised partially linear model in Lin & Carroll (2001) when the responses were non-Gaussian. There is no likelihood function for such problems in general, so that the penalised-spline method would have to abandon the variance-components representation in favour of ad hoc approaches or alternatives which are known to have nontrivial computation and marginalisation problems.

While variance-component-model representation of penalised splines can have certain advantages over kernels in terms of ease of method development, the opposite is true in terms of theoretical development. It is generally easy to analyse kernel methods, to develop appropriate bandwidths and to estimate these bandwidths in such a way that theoretical properties are ensured. In our Propositions 2 and 3, for example, we see that a standard bandwidth of order $n^{-1/5}$ will not result in \sqrt{n} -convergence rates for estimated β in general, while one of order $n^{-1/3}$ will do so. In contrast, the variance-component-model representation of penalised splines results in an estimated smoothing parameter, but it is generally unknown whether or not that smoothing parameter is estimated at rates that ensure asymptotic properties, especially for example for low-order basis representations where the number of knots is allowed to grow with the sample size.

Other examples of this difference in ease of theoretical development are available, such as in partially linear single-index models; Carroll et al. (1997) develop a semiparametric efficient kernel method for estimating the parameters in the model. We conjecture that the method of penalised low-order basis splines of Yu & Ruppert (2002) is also semi-parametric efficient if the number of knots grows at an appropriate rate and if the smoothing parameter is appropriately selected, but deriving these two items in generality may well prove to be extremely challenging.

ACKNOWLEDGEMENT

This work was supported by grants to N. Wang and R. J. Carroll from the National Cancer Institute and by the Texas A&M Center for Environmental and Rural Health through the National Institute for Environmental Health Sciences. It was also supported by the Texas A&M Foundation LINK program and the Texas Advanced Research Program.

APPENDIX

Proofs

In the following proofs, recall that T_a , X_a and Y_a denote the observations over all the clusters; that is $T_a = (T_1^T, \dots, T_n^T)^T$, and similarly for X_a and Y_a . Also, V_a and Σ_a stand for the $nm \times nm$ assumed and true covariance matrices for all data, respectively.

Sketch proof of Proposition 1. For clustered data, the asymptotic variance V_{BF} has its central component generated from $n^{-1}X_a^T V_a^{-1}(I - S_a)\varepsilon_a$, as we will show in (A5). Similarly, the central component in the asymptotic variance V_p is from $n^{-1}X_a^T(I - S_a)^T V_a^{-1}(I - S_a)\varepsilon_a$, which is $n^{-1}\tilde{X}_a^T V_a^{-1}(I - S_a)\varepsilon_a$ asymptotically. To compare V_p and V_{BF} , it is thus sufficient to compare the variances of the two central terms.

We now show that, under the condition that V_a and Σ_a depend only on T_a ,

$$\text{cov}\{X_a^T V_a^{-1}(I - S_a)\varepsilon_a\} \geq \text{cov}\{\tilde{X}_a^T V_a^{-1}(I - S_a)\varepsilon_a\}.$$

For the backfitting estimator,

$$\begin{aligned} \text{cov}\{X_a^T V_a^{-1}(I - S_a)\varepsilon\} &= E\{X_a^T V_a^{-1}(I - S_a)\Sigma_a(I - S_a)^T V_a^{-1}X_a\} \\ &= E\{m_X^T(T_a)V_a^{-1}(I - S_a)\Sigma_a(I - S_a)^T V_a^{-1}m_X(T_a)\} \\ &\quad + E[\text{tr}\{V_a^{-1}(I - S_a)\Sigma_a(I - S_a)^T V_a^{-1}\text{cov}(X_a|T_a)\}]. \end{aligned} \quad (\text{A1})$$

In this expression, $m_X(T_a)$ is generally nonzero and the first term is positive semidefinite because $V_a^{-1}(I - S_a)\Sigma_a(I - S_a)^T V_a^{-1}$ is positive semidefinite. Also,

$$\begin{aligned} \text{cov}\{\tilde{X}_a^T V_a^{-1}(I - S_a)\varepsilon_a\} &= E\{\tilde{X}_a^T V_a^{-1}(I - S_a)\Sigma_a(I - S_a)^T V_a^{-1}\tilde{X}_a\} \\ &= E\{E(\tilde{X}_a|T_a)^T V_a^{-1}(I - S_a)\Sigma_a(I - S_a)^T V_a^{-1}E(\tilde{X}_a|T_a)\} \\ &\quad + E[\text{tr}\{V_a^{-1}(I - S_a)\Sigma_a(I - S_a)^T V_a^{-1}\text{cov}(\tilde{X}_a|T_a)\}]. \end{aligned} \quad (\text{A2})$$

Note that

$$E(\tilde{X}_i|T_i) = E\{X_i - m_X(T_i)|T_i\} = 0, \quad \text{cov}(\tilde{X}_i|T_i) = \text{cov}\{X_i - m_X(T_i)|T_i\} = \text{cov}(X_i|T_i). \quad (\text{A3})$$

Therefore, the first term in (A2) is 0, and the second terms in (A1) and (A2) are identical. It follows that $\text{cov}\{X_a^T V_a^{-1}(I - S_a)\varepsilon_a\} \geq \text{cov}\{\tilde{X}_a^T V_a^{-1}(I - S_a)\varepsilon_a\}$, and consequently $V_{\text{BF}} \geq V_p$. \square

Sketch proof of Proposition 3. For the backfitting estimator, based on expression (3),

$$\hat{\beta}_{\text{BF}} - \beta = \{n^{-1}X_a^T V_a^{-1}(I - S_a)X_a\}^{-1}[n^{-1}X_a^T V_a^{-1}(I - S_a)\{\theta(T_a) + \varepsilon_a\}]. \quad (\text{A4})$$

In the first term of (A4),

$$\frac{1}{n}X_a^T V_a^{-1}(I - S_a)X_a \rightarrow E[X_i^T V_i^{-1}\{X_i - m_X(T_i)\}]$$

with probability 1, where

$$E[X_i^T V_i^{-1}\{X_i - m_X(T_i)\}] = E[\{X_i - m_X(T_i)\}^T V_i^{-1}\{X_i - m_X(T_i)\}] = E(\tilde{X}_i^T V_i^{-1}\tilde{X}_i).$$

In the second term of (A4),

$$\frac{1}{n}X_a^T V_a^{-1}(I - S_a)\{\theta(T_a) + \varepsilon_a\} = \frac{1}{n}X_a^T V_a^{-1}(I - S_a)\theta(T_a) + \frac{1}{n}X_a^T V_a^{-1}(I - S_a)\varepsilon_a, \quad (\text{A5})$$

where the first term determines the bias of the backfitting estimator in Proposition 3:

$$\frac{1}{n}X_a^T V_a^{-1}(I - S_a)\theta(T_a) = -\frac{h^2}{2}E\{X_i^T V_i^{-1}\theta^{(2)}(T_i)\} + o_p(h^2);$$

see Opsomer & Ruppert (1997). The second term in (A5) determines the ‘centred’ asymptotic distribution of the backfitting estimator and can be written as

$$\frac{1}{n}\sum_{i=1}^n X_i^T V_i^{-1}\varepsilon_i - \frac{1}{n}\sum_{i=1}^n X_i^T V_i^{-1}\{\hat{m}_\varepsilon(T_i) - m_\varepsilon(T_i)\},$$

where $\hat{m}_\varepsilon(t)$ is the nonparametric smooth of ε at t and $m_\varepsilon(t)$ is its expectation.

Recalling that $K_h(s) = h^{-1}K(s/h)$, where K is a kernel function in nonparametric estimation, we have

$$\hat{m}_\epsilon(t; \beta) - m_\epsilon(t) = w_2^{-1}(t) \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m K_h(T_{ij} - t) \epsilon_{ij} + o_p(n^{-1/2}),$$

where $w_2(t) = \sum_{i=1}^m f_i(t)$. Proposition 3 follows by substituting this expression back into (A4) and carrying out the expectation calculation. \square

Leave one subject out crossvalidation. Let $\hat{\beta}_{P[i]}$ and $\hat{\theta}_{h[i]}(t) = \hat{\theta}_{h[i]}(t, \hat{\beta}_{P[i]})$ be the profile-kernel estimators of β and $\theta(T)$ without observations from subject i . We let $\text{cv}(h)$ be $n^{-1} \sum_i \{Y_i - X_i \hat{\beta}_{P[i]} - \hat{\theta}_{h[i]}(T_i)\}^{\otimes 2}$, where $v^{\otimes 2} = v^T v$, and consider the following decomposition:

$$\begin{aligned} \text{cv}(h) = n^{-1} & \left(\sum_{i=1}^n \epsilon_i^{\otimes 2} + \sum_{i=1}^n \{X_i(\hat{\beta}_{P[i]} - \beta)\}^{\otimes 2} + \sum_{i=1}^n \{\hat{\theta}_{h[i]}(T_i) - \theta(T_i)\}^{\otimes 2} \right. \\ & - \sum_{i=1}^n [\epsilon_i^T \{X_i(\hat{\beta}_{P[i]} - \beta) + \hat{\theta}_{h[i]}(T_i) - \theta(T_i)\} + \{X_i(\hat{\beta}_{P[i]} - \beta) + \hat{\theta}_{h[i]}(T_i) - \theta(T_i)\}^T \epsilon_i] \\ & \left. + \sum_{i=1}^n [\{X_i(\hat{\beta}_{P[i]} - \beta)\}^T \{\hat{\theta}_{h[i]}(T_i) - \theta(T_i)\} + \{\hat{\theta}_{h[i]}(T_i) - \theta(T_i)\}^T \{X_i(\hat{\beta}_{P[i]} - \beta)\}] \right). \quad (\text{A6}) \end{aligned}$$

We select the bandwidth to be h^* which minimises $\text{cv}(h)$ in an interval $[b_1 n^{-1/5}, b_2 n^{-1/5}]$, where $0 < b_1 < b_2 < \infty$. The first term in the right-hand side of (A6) does not depend on h , while, under the conditions of Proposition 1 and for $h = O_p(n^{-1/5})$, the second term is negligible when compared to the third term. Direct derivations also show that, for $h = O_p(n^{-1/5})$, all other terms in (A6) converge to 0 faster than the third term; that is, the bandwidth selection criterion that minimises $\text{cv}(h)$ is asymptotically equivalent to the criterion that minimises

$$n^{-1} \sum_{i=1}^n \{\hat{\theta}_{h[i]}(T_i) - \theta(T_i)\}^{\otimes 2} = n^{-1} \sum_{i=1}^n \sum_{j=1}^m \{\hat{\theta}_{h[i]}(T_{ij}) - \theta(T_{ij})\}^2.$$

The asymptotic bias and variance structures in Lin & Carroll (2001) and Wang (2003) can be used to show that the selected optimal h is of order $n^{-1/5}$, as in the independent case.

REFERENCES

BUJA, A., HASTIE, T. J. & TIBSHIRANI, R. J. (1989). Linear smoothers and additive models (with Discussion). *Ann. Statist.* **17**, 453–555.

CARROLL, R. J., FAN, J., GJEBELS, I. & WAND, M. P. (1997). Generalised linear single-index models. *J. Am. Statist. Assoc.* **92**, 477–89.

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. & ZEGER, S. L. (2002). *The Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press.

HÄRDLE, W., LIANG, H. & GAO, J. (2000). *Partially Linear Models*. Heidelberg: Physica-Verlag.

HASTIE, T. & TIBSHIRANI, R. J. (1990). *Generalised Additive Models*. London: Chapman and Hall.

LIN, X. & CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalised estimating equations. *J. Am. Statist. Assoc.* **96**, 1045–56.

MEYERS, S. M., AMBLER, J. S., TAN, M., WERNER, J. C. & HUANG, S. S. (1992). Variation of perfluoropropane disappearance after vitrectomy. *Retina* **12**, 359–63.

OPSOMER, J. D. & RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25**, 186–211.

OPSOMER, J. D. & RUPPERT, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comp. Graph. Statist.* **8**, 715–32.

PEPE, M. S. & COUPER, D. (1997). Modeling partly conditional means with longitudinal data. *J. Am. Statist. Assoc.* **92**, 991–8.

RICE, J. A. (1986). Convergence rates for partially splined models. *Statist. Prob. Lett.* **4**, 204–8.

RICE, J. A. & SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc. B* **53**, 233–43.

- RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.
- SEVERINI, T. A. & STANISWALIS, J. G. (1994). Quasi-likelihood estimation in semi-parametric models. *J. Am. Statist. Assoc.* **89**, 501–12.
- SONG, X.-K. & TAN, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics* **56**, 496–502.
- SPECKMAN, P. E. (1988). Regression analysis for partially linear models. *J. R. Statist. Soc. B* **50**, 413–36.
- WANG, J. L. & WANG, W. (2001). Comment on ‘Semiparametric and nonparametric regression analysis of longitudinal data’. *J. Am. Statist. Assoc.* **96**, 119–23.
- WANG, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90**, 43–52.
- YU, Y. & RUPPERT, D. (2002). Penalised spline estimation for partially linear single index models. *J. Am. Statist. Assoc.* **97**, 1042–54.
- ZEGER, S. L. & DIGGLE, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.

[Received January 2003. Revised August 2003]