

Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data

BY XIHONG LIN

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
xlin@umich.edu

NAISYIN WANG

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
nwang@stat.tamu.edu

ALAN H. WELSH

*Faculty of Mathematical Studies, University of Southampton, Southampton,
Hampshire SO17 1BJ, U.K.*
a.h.welsh@maths.soton.ac.uk

AND RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
carroll@stat.tamu.edu

SUMMARY

For independent data, it is well known that kernel methods and spline methods are essentially asymptotically equivalent (Silverman, 1984). However, recent work of Welsh et al. (2002) shows that the same is not true for clustered/longitudinal data. Splines and conventional kernels are different in localness and ability to account for the within-cluster correlation. We show that a smoothing spline estimator is asymptotically equivalent to a recently proposed seemingly unrelated kernel estimator of Wang (2003) for any working covariance matrix. We show that both estimators can be obtained iteratively by applying conventional kernel or spline smoothing to pseudo-observations. This result allows us to study the asymptotic properties of the smoothing spline estimator by deriving its asymptotic bias and variance. We show that smoothing splines are consistent for an arbitrary working covariance and have the smallest variance when assuming the true covariance. We further show that both the seemingly unrelated kernel estimator and the smoothing spline estimator are nonlocal unless working independence is assumed but have asymptotically negligible bias. Their finite sample performance is compared through simulations. Our results justify the use of efficient, non-local estimators such as smoothing splines for clustered/longitudinal data.

Some key words: Asymptotic bias and variance; Asymptotic equivalent kernel; Consistency; Kernel regression; Longitudinal data; Non-localness; Nonparametric regression; Smoothing spline regression.

1. INTRODUCTION

Nonparametric regression for clustered/longitudinal data has attracted considerable recent interest. Kernel methods have been considered by Zeger & Diggle (1994), Hoover et al. (1998) and Fan & Zhang (2000), all of whom ignored the within-cluster correlation, and Severini & Staniswalis (1994) and Ruckstuhl et al. (2000), who incorporated the correlation into their kernel estimator. Lin & Carroll (2000) showed that the most efficient conventional local kernel estimator is obtained by ignoring the correlation. Spline methods for clustered/longitudinal data have been investigated by Brumback & Rice (1998), Wang (1998), Zhang et al. (1998), Lin & Zhang (1999) and Verbyla et al. (1999), among others. Most of these authors incorporated the within-cluster correlation into the construction of their spline estimators. An attractive feature of smoothing is that it can be easily obtained by fitting mixed effects models.

For independent data, it is well known (Silverman, 1984) that kernel and spline estimators are asymptotically equivalent and are local in the sense that the estimator at a point gives nonzero weights only to observations whose covariate is in a shrinking neighbourhood of that point. However, the relationship between kernel and spline estimators is not so well understood for clustered/longitudinal data. Welsh et al. (2002) recently found that conventional kernel and spline estimators behave completely differently for clustered/longitudinal data. While the most efficient conventional kernel estimator requires one to ignore the within-cluster correlation completely, the spline estimator with the smallest variance requires one to incorporate the within-cluster correlation. Further, conventional kernel estimators are local asymptotically but spline estimators are not.

These results suggest that Silverman's results about the asymptotic equivalence of spline and kernel estimators do not hold for clustered/longitudinal data. This raises challenging questions. First, what is the relationship between spline and kernel methods for clustered/longitudinal data? Is there a kernel estimator outside the conventional local kernel paradigm that is asymptotically equivalent to a spline estimator? Secondly, it is widely believed in the nonparametric literature that consistency of a nonparametric estimator requires localness. Since spline estimators are not local for clustered data, can they still be consistent? Thirdly, what are the asymptotic properties, such as the asymptotic bias and variance, of a spline estimator for clustered/longitudinal data?

In this paper, we first show in § 3 that, for any working covariance matrix, the spline estimator is asymptotically equivalent to a recently proposed seemingly unrelated kernel estimator (Wang, 2003), which is constructed iteratively in a non-traditional way and is shown to be more efficient for clustered/longitudinal data than the best conventional local kernel estimator. Here the asymptotic equivalence is in a similar sense to that of Silverman (1984), namely that the weights of the smoothing spline estimator asymptotically converge to the weights of the seemingly unrelated kernel estimator using Silverman's kernel function. In § 4, we show that, for any working covariance matrix, both estimators can be obtained iteratively by applying conventional kernel or spline smoothing to pseudo-observations. This result allows us to derive in § 5 the asymptotic bias and variance of a smoothing spline estimator. It is shown that a smoothing spline estimator is consistent for any arbitrary working covariance matrix and has the smallest variance when assuming the true covariance. Section 6 shows that both estimators are non-local and explains how non-local estimators can be consistent. Section 7 contains simulation results illustrating the equivalence we have shown in finite samples, and § 8 contains concluding remarks.

2. THE KERNEL AND SPLINE METHODS FOR NONPARAMETRIC REGRESSION IN CLUSTERED LONGITUDINAL DATA

2.1. *The model*

Consider data from n clusters with m_i observations in the i th cluster ($i = 1, \dots, n$). The j th observation ($j = 1, \dots, m_i$) in the i th cluster consists of an outcome variable Y_{ij} and a single covariate T_{ij} whose values vary within the cluster. For longitudinal data, a cluster refers to a subject and within-cluster observations refer to repeated measures over time, and T_{ij} is a time-varying covariate or time; for familial data, a cluster refers to a family and within-cluster observations refer to different family members and T_{ij} is a member-specific covariate. For simplicity, we assume the number of observations per cluster takes the same value m in all clusters, although the results hold when cluster sizes vary. The outcome Y_{ij} depends on the covariate T_{ij} through

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

where $\theta(t)$ is an unknown smooth function, and the errors $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})^T$ are independent with mean zero and true covariance matrix Σ . We assume in our asymptotic study that the cluster size m is finite but the number of clusters n goes to infinity. For practical applications of model (1), see Zeger & Diggle (1994) and Zhang et al. (1998).

Nonparametric estimation of $\theta(t)$ can proceed using kernel or spline methods. It is of interest to construct nonparametric estimators of $\theta(t)$ that account for the within-cluster correlation. We describe two methods which incorporate the within-cluster correlation, namely the seemingly unrelated kernel method (Wang, 2003) and the smoothing spline method (Wang, 1998; Zhang et al., 1998).

2.2. *The seemingly unrelated kernel estimator*

Define $Y_i = (Y_{i1}, \dots, Y_{im})^T$, $\theta(T_i) = \{\theta(T_{i1}), \dots, \theta(T_{im})\}^T$, $Y = (Y_1, \dots, Y_n)^T$, and T_i , T and $\theta(T)$ similarly. Define $K_h(s) = h^{-1} K(s/h)$, where $K(\cdot)$ is a zero-mean kernel function and h is a bandwidth. For an arbitrary matrix A , denote by a^{jk} the (j, k) th element of A^{-1} . Let V be a working covariance matrix (Liang & Zeger, 1986). Consider a q th-order polynomial kernel estimator. If $\theta(t)$ is estimated at the l th iteration by $\hat{\theta}_K^{(l)}(t)$, one updates $\theta(t)$ at the $(l+1)$ th iteration by $\hat{\theta}_K^{(l+1)}(t) = \hat{\alpha}_0$, where $\hat{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_q)^T$ solves

$$\sum_{i=1}^n \sum_{j=1}^m K_h(T_{ij} - t) B_{ij}(t)^T V^{-1} \{Y_i - \mu_{i(j)}(t)\} = 0, \quad (2)$$

in which $B_{ij}(t)$ is an $m \times (q+1)$ matrix of zeros except that the j th row is

$$\{1, (T_{ij} - t), \dots, (T_{ij} - t)^q\}^T,$$

and

$$\mu_{i(j)}(t) = \left(\hat{\theta}_K^{(l)}(T_{i1}), \dots, \hat{\theta}_K^{(l)}(T_{i,j-1}), \sum_{k=0}^q \alpha_k (T_{ij} - t)^k, \hat{\theta}_K^{(l)}(T_{i,j+1}), \dots, \hat{\theta}_K^{(l)}(T_{im}) \right)^T.$$

The final kernel estimator at convergence is called the seemingly unrelated kernel estimator of $\theta(t)$ and is denoted by $\hat{\theta}_K(t)$.

Wang (2003) gave the asymptotic bias and variance of $\hat{\theta}_K(t)$ as $n \rightarrow \infty$, $h \rightarrow 0$ and $nh \rightarrow \infty$ for the linear kernel estimator, corresponding to $q = 1$, as

$$E\{\hat{\theta}_K(t)\} - \theta(t) = h^2 \phi b_K(t) + o(h^2), \quad (3)$$

$$\text{var}\{\hat{\theta}_K(t)\} = \frac{\gamma}{nh} \frac{\tau(t)}{\eta^2(t)} + o\{(nh)^{-1}\}, \quad (4)$$

where $\phi = \int s^2 K(s) ds$, $\gamma = \int K^2(s) ds$, $\eta(t) = \sum_{j=1}^m v^{ij} f_j(t)$, $\tau(t) = \sum_{j=1}^m c_{jj} f_j(t)$ with c_{jj} being the (j, j) th element of $C = V^{-1} \Sigma V^{-1}$, $f_j(t)$ is the marginal density of T_j , and $b_K(t)$ solves

$$\sum_{j=1}^m \sum_{k=1}^m v^{jk} E\{b_K(T_k) | T_j = t\} f_j(t) = \frac{1}{2} \left\{ \sum_{j=1}^m v^{jj} f_j(t) \right\} \theta^{(2)}(t).$$

Equivalently, $b_K(t)$ satisfies the following Fredholm integral equation of the second kind:

$$b_K(t) + \int b_K(u) \zeta(u, t) du = \frac{1}{2} \theta^{(2)}(t), \quad (5)$$

where $\zeta(u, t) = \sum_{j=1}^m \sum_{k \neq j} v^{jk} f_{jk}(u, t) / \{\sum_{j=1}^m v^{jj} f_j(t)\}$ and $f_{jk}(u, t)$ is the joint density of (T_j, T_k) ; see equation (17) of Lin & Carroll (2001). Wang (2003) showed that $\hat{\theta}_K(t)$ with the smallest variance for a fixed bandwidth is obtained by accounting for the within-cluster correlation with $V = \Sigma$, and that it is more efficient than the best local kernel estimator of Lin & Carroll (2000).

2.3. The generalised least squares smoothing spline estimator

An alternative method for estimating $\theta(t)$ nonparametrically is to use smoothing splines. If we assume a working covariance matrix V , the p th-order smoothing spline estimator minimises

$$\frac{1}{n} \sum_{i=1}^n \{Y_i - \theta(T_i)\}^T V^{-1} \{Y_i - \theta(T_i)\} + \lambda \int \{\theta^{(p)}(t)\}^2 dt,$$

where λ is a smoothing parameter controlling the trade-off between the goodness of fit and the smoothness of the curve. The resulting p th-order smoothing spline estimator of $\theta(t)$ is

$$\hat{\theta}_S(T) = (\tilde{V}^{-1} + n\lambda\Psi)^{-1} \tilde{V}^{-1} Y, \quad (6)$$

where Ψ is the smoothing matrix (Green & Silverman, 1994, p. 13) and $\tilde{V} = \text{diag}(V, \dots, V)$. We call $\hat{\theta}_S(t)$ the generalised least squares smoothing spline estimator. Welsh et al. (2002) showed $\text{var}\{\hat{\theta}_S(t)\}$ is minimised for a fixed λ by accounting for the within-cluster correlation with $V = \Sigma$.

The generalised least squares spline smoother has attracted considerable recent attention because of its close connection with mixed effects models (Wang, 1998; Zhang et al., 1998). In particular, it has an attractive feature that it can be easily fitted using mixed effect models if we write $\hat{\theta}_S(T)$ as a linear combination of fixed effects and random effects and calculate it using the best linear unbiased predictors from mixed models. However, its theoretical properties are not known, and we investigate them in this paper.

3. ASYMPTOTIC EQUIVALENCE OF THE SPLINE AND THE SEEMINGLY UNRELATED KERNEL ESTIMATORS IN THE SENSE OF SILVERMAN

We show in this section that the spline estimator $\hat{\theta}_S(t)$ and the seemingly unrelated kernel estimator $\hat{\theta}_K(t)$ are asymptotically equivalent in the sense of Silverman (1984), i.e. the weight functions used to calculate them are asymptotically equivalent when Silverman's (1984) kernel function is used for $\hat{\theta}_K(t)$. It is difficult to relate $\hat{\theta}_S(t)$ and $\hat{\theta}_K(t)$ by comparing (6) and (2) directly. Our strategy is to obtain in Proposition 1 a closed-form expression

for $\hat{\theta}_K(t)$. We then use this expression to relate $\hat{\theta}_S(t)$ and $\hat{\theta}_K(t)$. For simplicity, we state in Proposition 1 the results for average kernels, $q = 0$. Results for a general q th-order polynomial kernel are similar and are briefly stated after Proposition 1; the proposition is proved in the Appendix.

PROPOSITION 1. (i) *For any working covariance matrix V , the seemingly unrelated average kernel estimator at convergence $\hat{\theta}_K(t)$ is given by*

$$\hat{\theta}_K(t) = K_{wh}^T(t) \{I + (\tilde{V}^{-1} - \tilde{V}^d)K_w\}^{-1} \tilde{V}^{-1} Y, \quad (7)$$

where $K_{wh}(t) = \{\sum_{i=1}^n \sum_{j=1}^m K_h(T_{ij} - t)v^{ij}\}^{-1} \{K_h(T_{11} - t), \dots, K_h(T_{nm} - t)\}^T$ is an $nm \times 1$ vector, $K_w = \{K_{wh}(T_{11}), \dots, K_{wh}(T_{nm})\}^T$ is an $nm \times nm$ matrix, $\tilde{V}^d = \text{diag}(V^d, \dots, V^d)$ and $V^d = \text{diag}(V^{-1})$.

(ii) *Let $\hat{\theta}_K(T)$ be an $nm \times 1$ vector containing the evaluations of $\hat{\theta}_K(t)$ at the vector of all the design points T , that is $\hat{\theta}_K(T) = \{\hat{\theta}_K(T_{11}), \dots, \hat{\theta}_K(T_{nm})\}^T$. Then*

$$\hat{\theta}_K(T) = \{I + K_w(\tilde{V}^{-1} - \tilde{V}^d)\}^{-1} K_w \tilde{V}^{-1} Y. \quad (8)$$

For the q th-order polynomial seemingly unrelated kernel estimator, one modifies $K_{wh}(t)$ in (7) to $K_{wh}(t)^T = \delta_1^T \{\tilde{T}(t)^T K_{ah}(t) \tilde{V}^d \tilde{T}(t)\}^{-1} \tilde{T}(t)^T K_{ah}(t)$, where $\delta_1 = (1, 0, \dots, 0)^T$, $\tilde{T}(t)$ is an $nm \times (q+1)$ matrix with $\{(n-1)i+j\}$ th row given by $\{1, (T_{ij}-t), \dots, (T_{ij}-t)^q\}$, and $K_{ah}(t) = \text{diag}\{K_h(T_{11}-t), \dots, K_h(T_{nm}-t)\}$.

Now consider the smoothing spline estimator $\hat{\theta}_S(T)$. For any working covariance matrix V , simple calculations show that we can rewrite the spline estimator $\hat{\theta}_S(T)$ in (6) as

$$\hat{\theta}_S(T) = \{I + (\tilde{V}^d + n\lambda\Psi)^{-1}(\tilde{V}^{-1} - \tilde{V}^d)\}^{-1}(\tilde{V}^d + n\lambda\Psi)^{-1} \tilde{V}^{-1} Y. \quad (9)$$

A comparison of (8) and (9) suggests that, to show that the weight function of $\hat{\theta}_S(t)$ is asymptotically equivalent to that of $\hat{\theta}_K(t)$, we need to show that $K_w = (\tilde{V}^d + n\lambda\Psi)^{-1}$ asymptotically. Since \tilde{V}^d is a diagonal matrix, we just need to show that, under working independence, a weighted smoothing spline estimator is asymptotically equivalent to a conventional weighted kernel estimator in the sense of Silverman (1984). This is done in Proposition 2. Its proof is in the Appendix.

PROPOSITION 2. *For any working covariance matrix V , denote the smoothing spline estimator by $\hat{\theta}_S(t) = n^{-1} \sum_{i=1}^n \sum_{j=1}^m W_{S,ij}(t, T) Y_{ij}$ and the seemingly unrelated kernel estimator by $\hat{\theta}_K(t) = n^{-1} \sum_{i=1}^n \sum_{j=1}^m W_{K,ij}(t, T) Y_{ij}$. Then $\hat{\theta}_S(t)$ with smoothing parameter λ is asymptotically equivalent to $\hat{\theta}_K(t)$ with the effective bandwidth $h(t) = \{\lambda / \sum_{j=1}^m v^{jj} f_j(t)\}^{1/(2p)}$ and the kernel function $K(t)$ solving the differential equation (Silverman, 1984)*

$$(-1)^p K^{(2p)}(t) + K(t) = \Delta(t), \quad (10)$$

where $\Delta(t)$ is the Dirac delta function, in the sense that their weight functions $W_S(t, T)$ and $W_K(t, T)$ are asymptotically equivalent.

Remark 1. The asymptotic equivalence here is in the sense of Silverman (1984), and the kernel function $K(t)$ in (10) is identical to that given by Silverman (1984). For clustered data, a smoothing spline is thus effectively a seemingly unrelated kernel estimator with a varying bandwidth that depends not only on the marginal densities of the T_j but also on the working covariance matrix V .

Remark 2. The kernel function $K(t)$ in (10) has Fourier transform $(1 + t^{2p})^{-1}$ (Silverman, 1984) and satisfies $\int K(t)dt = 1$, $\int t^q K(t)dt = 0$ ($0 < q < 2p$) and $\int t^{2p} K(t)dt = (-1)^{p-1}$; that

is, $K(t)$ is a $2p$ th-order kernel. For a linear spline, $p = 1$ and $K(t)$ is the Laplace density $\frac{1}{2} \exp(-|t|)$ and is the traditional second-order kernel. For a cubic spline, $p = 2$ and $K(t)$ is a fourth-order kernel,

$$K(t) = \frac{1}{2} \exp(-|t|/\sqrt{2}) \sin(|t|\sqrt{2} + \pi/4). \tag{11}$$

4. UNDERSTANDING THE RELATIONSHIP BETWEEN THE SPLINE AND SEEMINGLY UNRELATED KERNEL ESTIMATORS

The results in § 3 do not give us a good insight into why the asymptotic equivalence of the two estimators holds. We show in this section that both estimators can be obtained in the same iterative fashion using pseudo-observations. They differ only in that standard weighted kernel smoothing is used at each iteration for the seemingly unrelated kernel estimator, while standard weighted spline smoothing is used at each iteration for the smoothing spline estimator.

We start with the seemingly unrelated kernel estimator $\hat{\theta}_K(t)$. For simplicity, consider the average kernel, $p = 0$. From equation (2), for any working covariance matrix V , simple calculations show that the seemingly unrelated kernel estimator at the $(l + 1)$ th iteration can be rewritten as

$$\hat{\theta}_K^{(l+1)}(t) = \frac{n^{-1} \sum_{i=1}^n \sum_{j=1}^m K_h(T_{ij} - t) v^{jj} Y_{ij}^{(l+1)}}{n^{-1} \sum_{i=1}^n \sum_{j=1}^m v^{jj} K_h(T_{ij} - t)}, \tag{12}$$

where the pseudo-observation at the $(l + 1)$ th iteration is defined by

$$Y_{ij}^{(l+1)} = Y_{ij} + (v^{jj})^{-1} \sum_{k \neq j} v^{jk} \{Y_{ik} - \hat{\theta}_K^{(l)}(T_{ik})\}. \tag{13}$$

Equation (12) shows that at each iteration conventional weighted Nadaraya–Watson kernel smoothing is applied to the pseudo-observations $Y_{ij}^{(l+1)}$ with the weights $\{v^{jj}\}$.

We now show that $\hat{\theta}_S(t)$, given in (6), can also be obtained iteratively by applying conventional weighted spline smoothing to similar pseudo-observations $Y_{ij}^{(l+1)}$ with the weights $\{v^{jj}\}$ at each iteration. Consider calculating a smoothing spline estimator $\hat{\theta}_S^*(t)$ in the following fashion. Set the initial estimator $\hat{\theta}_S^{(0)}(t)$ of $\theta(t)$ to be the standard p th-order smoothing spline estimator obtained by assuming independence among all observations. If $\hat{\theta}_S^{(l)}(t)$ is the smoothing spline estimator at the l th iteration, one updates $\theta(t)$ at the $(l + 1)$ th iteration by minimising

$$\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m v^{jj} \{Y_{ij}^{(l+1)} - \theta(T_{ij})\}^2 + \lambda \int \{\theta^{(p)}(t)\}^2 dt$$

with respect to $\theta(\cdot)$, where the $Y_{ij}^{(l+1)}$ are the pseudo-observations defined in (13) except that $\hat{\theta}_K^{(l)}(t)$ is replaced by $\hat{\theta}_S^{(l)}(t)$. This gives the conventional weighted p th-order smoothing spline estimator

$$\hat{\theta}_S^{(l+1)}(T) = (\tilde{V}^d + n\lambda\Psi)^{-1} \tilde{V}^d Y^{(l+1)}, \tag{14}$$

where $Y^{(l+1)} = \{Y_1^{(l+1)\top}, \dots, Y_n^{(l+1)\top}\}^\top$ with $Y_i^{(l+1)} = \{Y_{i1}^{(l+1)}, \dots, Y_{im}^{(l+1)}\}^\top$.

This means that, at the $(l + 1)$ th iteration, we update $\theta(t)$ by applying conventional weighted p th-order spline smoothing to the pseudo-observations $Y_{ij}^{(l+1)}$ with the weights v^{jj} . Following Silverman (1984) and Nychka (1995), one can easily show further that, for any t ,

the estimator at the $(l + 1)$ th iteration can be written asymptotically as

$$\hat{\theta}_S^{(l+1)}(t) = \left(n \sum_{j=1}^m v^{jj} \right)^{-1} \sum_{i=1}^n \sum_{j=1}^m G_{\lambda^*}(t, T_{ij}) v^{ij} Y_{ij}^{(l+1)} + o_p(1), \quad (15)$$

where $G_{\lambda^*}(t, s)$ is the Green's function defined in (A10) with $\lambda^* = \lambda / (\sum_{j=1}^m v^{jj})$ and $f(t) = \{ \sum_{j=1}^m v^{jj} f_j(t) \} / (\sum_{j=1}^m v^{jj})$. Denote by $\hat{\theta}_S^*(t)$ the estimator $\hat{\theta}_S^{(l+1)}(t)$ at convergence. The relationship between $\hat{\theta}_S^*(t)$ and $\hat{\theta}_S(t)$ is given in Proposition 3.

PROPOSITION 3. *Denote by $\hat{\theta}_S^*(t)$ the iterative weighted smoothing spline estimator (15) at convergence. Then $\hat{\theta}_S^*(t)$ has a closed-form expression and equals the generalised least squares smoothing spline estimator $\hat{\theta}_S(t)$ in (6).*

The proof is given in the Appendix. The relationship between $\hat{\theta}_K(t)$ and $\hat{\theta}_S(t)$ is now transparent, and we can see why they are asymptotically equivalent. Both estimators can be obtained iteratively. At each iteration, the seemingly unrelated kernel estimator applies standard weighted kernel smoothing to the pseudo-observations $Y_{ij}^{(l+1)}$, while the generalised least squares smoothing spline estimator applies standard weighted spline smoothing to the pseudo-observations $Y_{ij}^{(l+1)}$. Since standard weighted spline and kernel smoothing are asymptotically equivalent (Silverman, 1984) at each iteration, they should be asymptotically equivalent at convergence.

5. THE ASYMPTOTIC BIAS AND VARIANCE OF THE GENERALISED LEAST SQUARES SMOOTHING SPLINE ESTIMATOR

Analogously to the independent data case pointed out by Nychka (1995), the results in § 3 are not strong enough to establish the asymptotic bias and variance of the generalised least squares smoothing spline estimator $\hat{\theta}_S(t)$. Such calculations are of substantial interest, since they provide the asymptotic properties of $\hat{\theta}_S(t)$ and allow us to investigate whether or not $\hat{\theta}_S(t)$ is consistent.

It is difficult to study the asymptotic bias and variance of $\hat{\theta}_S(t)$ using its closed-form expression in equation (6). Using the fact in Proposition 3 that $\hat{\theta}_S(t)$ can be obtained iteratively by standard weighted spline smoothing at each iteration, we can calculate the asymptotic bias and variance of $\hat{\theta}_S(t)$ in a much easier way by iteratively applying the asymptotic bias and variance results of the smoothing spline estimator for independent data (Nychka, 1995) at each iteration. These results are provided in Proposition 4 and its proof is given in the Appendix.

PROPOSITION 4. *Denote by $\hat{\theta}_S(t)$ the p th-order generalised least squares smoothing spline estimator given in (6) using any given working covariance matrix V . Assume that the marginal densities $f_j(t)$ of the T_{ij} have uniformly continuous derivatives. We make the same asymptotic assumptions about n and λ as in Nychka (1995), and we assume that Nychka's bias and variance results for the p th-order spline in his equations (1·9) and (1·10) hold for independent data and that the properties of the Green's function in (A11) and (A14) hold.*

(i) *The asymptotic bias of $\hat{\theta}_S(t)$ is*

$$E\{\hat{\theta}_S(t)\} - \theta(t) = (-1)^{p-1} \frac{\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} b_S(t) + o(\lambda),$$

where $b_S(t)$ satisfies

$$\sum_{j=1}^m \sum_{k=1}^m v^{jk} E\{b_S(T_k) | T_j = t\} f_j(t) = \frac{1}{a_p} \left\{ \sum_{j=1}^m v^{jj} f_j(t) \right\} \theta^{(2p)}(t),$$

where a_p is a constant. Equivalently, $b_S(t)$ solves the Fredholm integral equation of the second kind, with the right-hand side of (5) replaced by $\theta^{(2p)}(t)/a_p$.

(ii) The asymptotic variance of $\hat{\theta}_S(t)$ is

$$\text{var} \{ \hat{\theta}_S(t) \} = \frac{\gamma_p}{n} \left\{ \frac{\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} \right\}^{-1/(2p)} \frac{\tau(t)}{\eta^2(t)} + o_p \{ (n\lambda^{1/(2p)})^{-1} \},$$

where γ_p is a constant and $\tau(t)$ and $\eta(t)$ are defined below equation (4). We believe that $\gamma_p = \int K^2(t) dt$, where $K(t)$ satisfies (10).

(iii) With the effective bandwidth held fixed as $h(t) = \{\lambda / \sum_{j=1}^m \sigma^{jj} f_j(t)\}^{1/(2p)}$, the generalised least squares smoothing spline estimator with the smallest variance is obtained by assuming that the working covariance matrix V equals the true covariance Σ , that is $V = \Sigma$. Its variance is

$$\text{var}_{\min} \{ \hat{\theta}_S(t) \} = \left\{ \frac{\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} \right\}^{-1/(2p)} \frac{1}{\sum_{j=1}^m \sigma^{jj} f_j(t)} + o_p \{ (n\lambda^{1/(2p)})^{-1} \}.$$

(iv) The generalised least squares smoothing spline estimator has the asymptotic expansion

$$\begin{aligned} \hat{\theta}_S(t) - \theta(t) &= \frac{1}{n \sum_{j=1}^m v^{jj}} \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m G_{\lambda^*}(t, T_{ij}) v^{jk} \{ Y_{ik} - \theta(T_{ik}) \} + (-1)^{p-1} h^{2p}(t) b_S(t) \\ &\quad + o_p [\{ nh(t) \}^{-\frac{1}{2}} + h^{2p}(t)], \end{aligned} \quad (16)$$

where $G_{\lambda^*}(t, s)$ is the Green's function defined in (A10) and $h(t) = [\lambda / \{\sum_{j=1}^m v^{jj} f_j(t)\}]^{1/(2p)}$ denotes the effective bandwidth; see Proposition 2.

Remark 3. The bias results in part (i) of Proposition 4 show that, for any working covariance matrix V , the generalised least squares smoothing spline estimator is consistent.

Remark 4. The expressions for the asymptotic bias and variance of the generalised least squares smoothing spline estimator closely resemble the forms of the asymptotic bias and variance of the seemingly unrelated kernel estimator given in (3) and (4). With $h(t) = [\lambda / \{\sum_{j=1}^m v^{jj} f_j(t)\}]^{1/(2p)}$, the asymptotic bias and variance of $\hat{\theta}_S(t)$ are

$$\begin{aligned} E\{ \hat{\theta}_S(t) \} - \theta(t) &= (-1)^{p-1} h^{2p}(t) b_S(t) + o_p \{ h^{2p}(t) \}, \\ \text{var} \{ \hat{\theta}_S(t) \} - \theta(t) &= \{ nh(t) \}^{-1} \{ \gamma_p \tau(t) / \eta^2(t) \} + o_p [\{ nh(t) \}^{-1}]. \end{aligned}$$

This clearly shows that the p th-order generalised least squares smoothing spline estimator behaves asymptotically similarly to the $q = (2p - 1)$ th-order polynomial seemingly unrelated kernel estimator with a $(2p)$ th-order kernel function defined in (10). Note that Wang (2003) focused on the second-order seemingly unrelated kernels. Extensions of her results to higher-order kernels are straightforward along the lines of Wand & Jones (1995, § 5.4).

Remark 5. The asymptotic bias of the linear generalised least squares smoothing spline estimator ($p = 1$) is $h^2(t) b_S(t)$, where $b_S(t)$ solves (5) with the right-hand side of the equation equal to $\theta^{(2)}(t)/a_1$, where a_1 is some constant. Hence it corresponds to the second-order seemingly unrelated kernel. The asymptotic bias of the cubic generalised least squares

smoothing spline is $-h^4(t)b_S(t)$, where $b_S(t)$ solves (5) with the right-hand side of the equation equal to $\theta^{(4)}(t)/a_2$, where a_2 is a constant. Hence it corresponds to a higher, fourth-order seemingly unrelated kernel.

Remark 6. When $m = 1$, we have cross-sectional independent data following $Y_i = \theta(T_i) + \varepsilon_i$, where the ε_i are independent and identically distributed as $N(0, \sigma^2)$. The results in Proposition 4 reduce to those given in Nychka (1995), namely

$$E\{\hat{\theta}_S(t)\} - \theta(t) = (-1)^{p-1} h^{2p}(t) \theta^{(2p)}(t) / a_p + o_p\{h^{2p}(t)\}, \quad (17)$$

$$\text{var}\{\hat{\theta}_S(t)\} \asymp \{nh(t)\}^{-1} \{\gamma_p \sigma^2 / f(t)\} + o_p[\{nh(t)\}^{-1}], \quad (18)$$

where $h(t) = \{\lambda \sigma^2 / f(t)\}^{1/(2p)}$ and $f(t)$ is the density of T_i .

Remark 7. For clustered/longitudinal data, the working independence smoothing spline estimator is calculated by treating the data as being independent with error variances σ_{jj} . Its asymptotic bias and variance takes the same form as (17) and (18) except that $\sigma^2 / f(t)$ is replaced by $\{\sum_{j=1}^m \sigma_{jj}^{-1} f_j(t)\}^{-1}$ in (18) and $h(t)$ replaced by $h(t) = [\lambda / \{\sum_{j=1}^m \sigma_{jj}^{-1} f_j(t)\}]^{1/(2p)}$.

6. NON-LOCALNESS AND CONSISTENCY OF GENERALISED LEAST SQUARES SPLINES AND SEEMINGLY UNRELATED KERNELS

6.1. Observation-level non-localness

For independent data, both kernel and spline estimators are local in the sense that, for any t , only observations whose covariate values are in the neighbourhood of t are used asymptotically to estimate $\theta(t)$. For clustered/longitudinal data, the same is true for working independence kernel and spline estimators (Welsh et al., 2002). If m is finite, this means that only observations from different clusters contribute to estimation of $\theta(t)$ at any t . Such localness is widely presumed to be necessary to ensure consistency of a nonparametric estimator. However, Welsh et al. (2002) observed that the generalised least squares spline is not local. This non-localness of the spline is supported by the asymptotic spline expansion in (16). A similar asymptotic expansion holds for the seemingly unrelated kernel estimator if we replace $G_{\lambda*}(t, T_{ij})$ in (16) by $K_h(T_{ij} - t)$ (Wang, 2003). It follows that the seemingly unrelated kernel estimator is not local either.

The non-localness can also be demonstrated for seemingly unrelated kernel estimators via (12) and for the generalised least squares spline estimators via (15). Here we see clearly that both methods are local in the pseudo-observations (13): the seemingly unrelated kernel pseudo-observation at convergence is

$$Y_{ij}^* = Y_{ij} + (v^{jj})^{-1} \sum_{j=1}^m \sum_{k \neq j} v^{jk} \{Y_{ik} - \hat{\theta}_K(T_{ik})\},$$

and a similar form holds for the generalised least squares spline. Clearly Y_{ij}^* depends on all the responses in the cluster, not just Y_{ij} . This discussion suggests that, if any observation in a cluster has a T_{ij} near the value at which the function is to be fitted, then all the observations in the cluster contribute to the fit; that is both methods are not local at the observation-level.

To illustrate this non-localness, we provide a numerical example. For $n = 50$ and $m = 3$, we generated the covariate T_{ij} from the $\text{Un}(-2, 2)$ distribution and assumed that Y_{ij} followed model (1) with $\theta(t) = \sin(2t)$ and an exchangeable covariance matrix $\Sigma = 4(0.2I + 0.8J)$, where J is an $m \times m$ matrix of ones; i.e. the pairwise within-cluster correlation is 0.8. Both

$\hat{\theta}_K(t)$ and $\hat{\theta}_S(t)$ can be written in the form $\hat{\theta}(t) = \sum_{i=1}^n \sum_{j=1}^m W_{ij}(t, T) Y_{ij}$, where the expression for the weights $W_{ij}(t, T)$ is given by equation (7) for estimator $\hat{\theta}_K(t)$ and by equation (7) of Welsh et al. (2002) for $\hat{\theta}_S(t)$. Using these weight functions, we investigate how the observation at a fixed point t is weighted when we estimate $\theta(s)$ for a series of values of s .

Figure 1(a) plots the weights for $\hat{\theta}_K(t)$ at $t = 0.25$ with the kernel function (11) and bandwidth $h = 0.4$ for both working independence, which is identical to the classical kernel estimator, and assuming the true covariance matrix. A similar plot for the cubic smoothing spline estimator $\hat{\theta}_S(t)$ with $\lambda = 0.1$ is given in Fig. 1(b). One can easily see that the working independence kernel and spline estimates are local, while the seemingly unrelated kernel estimate and the generalised least squares cubic smoothing spline estimate assuming the true covariance matrix are not local. Furthermore, since the kernel function (11) is the equivalent kernel of the cubic smoothing spline estimator, the shapes of the two weight curves are nearly identical. Figure 1 hence also supports the theoretical results in §§ 3 and 4 and provides numerical evidence that the asymptotic equivalence of $\hat{\theta}_K(t)$ and $\hat{\theta}_S(t)$ carries over quite well to finite samples.

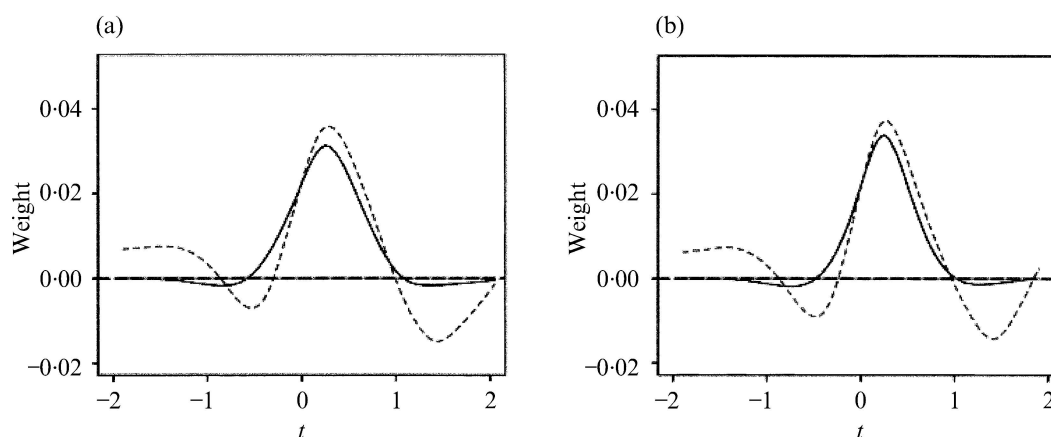


Fig. 1. Observation-level non-localness. The weights for the iterative kernel estimate and the smoothing spline estimate when $n = 50$ and $m = 3$ in the exchangeable case with $\rho = 0.8$. The solid line assumes working independence and dotted line assumes the true covariance structure: (a) the seemingly unrelated kernel estimate, (b) the smoothing spline estimate.

6.2. Cluster-level localness and consistency

The discussion in § 6.1 together with the earlier results in §§ 2.1 and 4 indicate that an efficient kernel/spline estimator that effectively accounts for the within-cluster correlation requires using all observations within the same cluster and hence has to be non-local at the observation level. This raises an intriguing question of how a non-local estimator can be consistent: how can observations whose covariate values are outside the neighbourhood of t still contribute to estimation of $\theta(t)$ without introducing asymptotic bias? This is often viewed as impossible for independent data.

We investigate this issue by examining the pseudo-observations used at each iteration of the estimators $\hat{\theta}_K(t)$ and $\hat{\theta}_S(t)$. From equations (12) and (15), conventional weighted kernels and splines are applied to the pseudo-observations $Y_{ij}^{(l+1)}$ in (13) at each iteration with $\hat{\theta}^{(l)}(t)$ estimated using the kernel method and the spline method at the previous iteration respectively. Hence both methods are local in the pseudo-observations. Each

pseudo-observation $Y_{ij}^{(l+1)}$ is a weighted average of Y_{ij} and the residuals of the other observations within the same cluster $\{Y_{ik} - \hat{\theta}^{(l)}(T_{ik})\}$ calculated by centering Y_{ik} around their means estimated consistently using either kernel or spline methods at the previous iteration $\hat{\theta}^{(l)}(T_{ik})$. This means that, although both methods are non-local at the observation level, they are local at the cluster level. Furthermore, although $Y_{ij}^{(l+1)}$ uses all observations within the same cluster, $Y_{ij}^{(l+1)}$ is asymptotically unbiased for $\theta(T_{ij})$, that is $E\{Y_{ij}^{(l+1)}\} = \theta(T_{ij})$ asymptotically. Hence by re-centring at each iteration, observations within the same cluster whose covariate values are not in the neighbourhood of t are used to improve the efficiency of estimation of $\theta(t)$ without introducing bias. When the iterative procedures are used, the re-centring is done by updating $\hat{\theta}(T_{ij})$ at each iteration. When their closed-form expressions in (7) and (6) are used, the re-centering is done implicitly for all design points simultaneously.

7. SIMULATION STUDY

In this section, we present simulation results which compare the finite sample efficiency of the kernel and spline estimators $\hat{\theta}_K(t)$ and $\hat{\theta}_S(t)$. We focus on using the true covariance in estimation. Under working independence, they reduce to conventional kernel and spline estimators respectively, and their final sample performance has already been compared in Welsh et al. (2002). We consider the seemingly unrelated linear kernel estimator using the Epanechnikov kernel, and the cubic smoothing spline estimator.

We assumed in our simulation that the number of clusters is $n = 50$ or $n = 100$, and the cluster size is $m = 3$. The covariate T_{ij} was generated independently from the $\text{Un}[-2, 2]$ distribution. The outcome Y_{ij} followed model (1) with $\theta(t)$ specified by each of the following four functions with different curvatures:

in Model 1, $\theta(t) = \sin(2t)$;

in Model 2, $\theta(t) = \{z(1-z)\}^{\frac{1}{2}} \sin\{2\pi(1+2^{-3/5})/(z+2^{-3/5})\}$;

in Model 3, $\theta(t) = \{z(1-z)\}^{\frac{1}{2}} \sin\{2\pi(1+2^{-7/5})/(z+2^{-7/5})\}$;

in Model 4, $\theta(t) = \sin(8z-4) + 2 \exp\{-256(z-0.5)^2\}$, where $z = (t+2)/4$.

We assumed that the marginal variances of the Y_{ij} were 1, and considered three true covariance structures, namely exchangeable with common correlation $\rho = 0.6$, autoregressive with correlation $\rho = 0.6$ and unstructured with $\rho_{12} = \rho_{23} = 0.8$ and $\rho_{13} = 0.5$.

For each configuration, we generated 200 simulated datasets and estimated $\theta(t)$ using the seemingly unrelated linear kernel estimator and the generalised least squares cubic smoothing spline estimator, where the true covariance structure was assumed. For simplicity and for the sake of consistency when comparing different methods, we estimated the bandwidth parameter and the smoothing parameter for each simulated dataset as follows. We selected 35 equally-spaced grid points in $(-2, 2)$. For each simulated dataset, at each grid point t_k ($k = 1, \dots, 35$), we estimated the bandwidth parameter h for $\hat{\theta}_K(t_k)$ and the smoothing parameter λ for $\hat{\theta}_S(t_k)$ by minimising the mean squared errors of the estimators $\sum_{i=1}^n \sum_{j=1}^m \{\hat{\theta}(t_k) - \theta(t_k)\}^2$ by using our knowledge of the actual function.

Table 1 compares the average mean squared error efficiencies of two estimators over the grid points assuming the true covariance matrix. It shows that the relative efficiencies of generalised least squares spline estimators to seemingly unrelated kernel estimators are close to one, indicating their similar behaviour in finite samples, consistent with the asymptotic equivalence of the two methods.

Table 1. Simulated relative mean squared error (MSE) efficiencies comparing smoothing splines with seemingly unrelated (SUR) kernels assuming the true covariance structure

Model	Corr	MSE efficiency of splines relative to SUR kernels		Model	Corr	MSE efficiency of splines relative to SUR kernels	
		$n = 50$	$n = 100$			$n = 50$	$n = 100$
1	1	1.15	1.22	3	1	0.98	1.00
1	2	1.16	1.21	3	2	0.98	1.00
1	3	0.99	1.07	3	3	0.92	0.98
2	1	0.99	1.03	4	1	1.09	1.07
2	2	0.98	1.01	4	2	1.08	1.06
2	3	0.94	0.99	4	3	1.25	1.13

The four models correspond to:

Model 1, $\theta(t) = \sin(2t)$;

Model 2, $\theta(t) = \{z(1-z)\}^{\frac{1}{2}} \sin \{2\pi(1+2^{-3/5})/(z+2^{-3/5})\}$;

Model 3, $\theta(t) = \{z(1-z)\}^{\frac{1}{2}} \sin \{2\pi(1+2^{-7/5})/(z+2^{-7/5})\}$;

Model 4, $\theta(t) = \sin(8z-4) + 2 \exp\{-256(z-0.5)^2\}$.

The three correlation structures correspond to: Corr = 1, the autoregressive case with $\rho = 0.6$; Corr = 2, the exchangeable case with $\rho = 0.6$; Corr = 3, the unstructured case with $\rho_{12} = \rho_{23} = 0.8$ and $\rho_{13} = 0.5$.

8. DISCUSSION

We have assumed in this paper that the number of observations per cluster is finite when the number of clusters goes to infinity. This assumption often holds for most common clustered/longitudinal studies. The case where both the number of observations and the number of clusters go to infinity is also of interest. For example, electroencephalography recording is done every 30 seconds overnight for patients in neurological research to study brain activities (Malow et al., 1996). Examination of the proof of Proposition 1 in the Appendix shows that the asymptotic equivalence of smoothing splines and seemingly unrelated kernels in the sense of Silverman (1984) still holds even when both the number of observations and the number of clusters go to infinity. However, the asymptotic properties such as biases and variances of seemingly unrelated kernel estimators and smoothing splines in this scenario are not well understood and are currently under investigation.

ACKNOWLEDGEMENT

This work was supported in part by grants from the National Cancer Institute and the National Institute of Environmental Health Sciences.

APPENDIX

Proofs

Proof of Proposition 1. Consider the average kernel estimator ($q = 0$). Denote by $\hat{\theta}_K^{(l)}(T) = \{\hat{\theta}_K^{(l)}(T_{11}), \dots, \hat{\theta}_K^{(l)}(T_{mm})\}^T$ the seemingly unrelated kernel estimator at the l th iteration. From equation (2), some calculations lead to

$$\hat{\theta}_K^{(l+1)}(t) = (1^T K_{ah} \tilde{V}^d 1)^{-1} [1^T K_{ah} \tilde{V}^d \hat{\theta}_K^{(l)}(T) + 1^T K_{ah} \tilde{V}^{-1} \{Y - \hat{\theta}_K^{(l)}(T)\}], \quad (\text{A1})$$

where $K_{dh} = \text{diag} \{K_h(T_{11} - t), \dots, K_h(T_{nm} - t)\}$. Denote by

$$K_{wh} = \left\{ \sum_{i=1}^n \sum_{j=1}^m K_h(T_{11} - t) v^{ij} \right\}^{-1} \{K_h(T_{11} - t), \dots, K_h(T_{nm} - t)\}^T$$

the vector of standardised kernel weights at t , and write $\hat{\theta}_K^{(l)}(T) = K^{(l)} Y$. We need to find the relationship between $K^{(l+1)}$ and $K^{(l)}$. From (A1), simple calculations give

$$\hat{\theta}_K^{(l+1)}(t) = K_{wh}(t)^T \{\tilde{V}^d K^{(l)} + \tilde{V}^{-1}(I - K^{(l)})\} Y, \quad (\text{A2})$$

$$\hat{\theta}_K^{(l+1)}(T) = K_w \{\tilde{V}^d K^{(l)} + \tilde{V}^{-1}(I - K^{(l)})\} Y, \quad (\text{A3})$$

and $K^{(l+1)} = K_w \{\tilde{V}^d K^{(l)} + \tilde{V}^{-1}(I - K^{(l)})\}$, where $K_w = \{K_{wh}(T_{11}), \dots, K_{wh}(T_{nm})\}^T$. At convergence, $K^{(l+1)} = K^{(l)} = K_*$. Hence K_* solves $K_* = K_w \{\tilde{V}^d K_* + \tilde{V}^{-1}(I - K_*)\}$; that is,

$$K_* = \{I + K_w(\tilde{V}^{-1} - \tilde{V}^d)\}^{-1} K_w \tilde{V}^{-1}. \quad (\text{A4})$$

If we substitute (A4) into (A2) and (A3), some simple algebra gives (7) and (8).

Proof of Proposition 2. If we use the results in the paragraph before Proposition 2, it suffices to show that $K_w = (\tilde{V}^d + n\lambda\Psi)^{-1}$ asymptotically. Since \tilde{V}^d is a diagonal matrix, we simply need to show that, under working independence, the weighted spline estimator is asymptotically equivalent to the conventional kernel estimator. This can be shown using the results in § 6 of Silverman (1984).

To be specific, for any weights w_{ij} , the weighted kernel estimator under working independence is

$$\hat{\theta}_{\text{WK}}(t) = \{C_w(t)\}^{-1} \sum_{i=1}^n \sum_{j=1}^m w_{ij} K_h(T_{ij} - t) Y_{ij},$$

where $C_w(t) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} K_h(T_{ij} - t)$ and the subscript WK denotes the weighted kernel under working independence. Write $C_w = \text{diag} \{C_w(T_{11}), \dots, C_w(T_{nm})\}$. Then $\hat{\theta}_{\text{WK}}(t)$ evaluated at the vector of all design points T is

$$\hat{\theta}_{\text{WK}}(T) = \{\hat{\theta}_{\text{WK}}(T_{11}), \dots, \hat{\theta}_{\text{WK}}(T_{nm})\}^T = K_w W Y, \quad (\text{A5})$$

where $W = \text{diag} \{w_{11}, \dots, w_{nm}\}$, $K_w = C_w^{-1} K_h$ and K_h is a $nm \times nm$ matrix with elements

$$K_h(T_{ij} - T_{i'j'}) \quad (i, i' = 1, \dots, n, j, j' = 1, \dots, m).$$

Denote by $\hat{\theta}_{\text{WS}}(T)$ the weighted smoothing spline estimator under working independence evaluated at all the design points T . We have

$$\hat{\theta}_{\text{WS}}(T) = (W + n\lambda\Psi)^{-1} W Y. \quad (\text{A6})$$

The results in § 6 of Silverman (1984) show that the weights of $\hat{\theta}_{\text{WK}}(T)$ and those of $\hat{\theta}_{\text{WS}}(T)$ are asymptotically equivalent. Now let $W = \tilde{V}^d$. A comparison of (A5) and (A6) gives $K_w = (\tilde{V}^d + n\lambda\Psi)^{-1}$ asymptotically with the kernel function defined in (10).

We now study how the bandwidth h is related to the smoothing parameter λ . Following Silverman (1984), we first standardise the weights as $w_{ij} = v^{ij}/n \sum_{j=1}^m v^{jj}$ such that $\sum_{i=1}^n \sum_{j=1}^m w_{ij} = 1$, and calculate the weighted cumulative distribution function and its limit as

$$F_n^w(t) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} I(T_{ij} \leq t) \rightarrow F(t) = \sum_{j=1}^m v^{jj} F_j(t) / \sum_{j=1}^m v^{jj}. \quad (\text{A7})$$

Using the results in Theorem A of Silverman (1984) and replacing λ and $f(t)$ by $\lambda/\sum_{j=1}^m v^{jj}$ and $\sum_{j=1}^m v^{jj} f_j(t)/\sum_{j=1}^m v^{jj}$, respectively, we find that the bandwidth $h(t)$ is related to λ by

$$h(t) = \left(\frac{\lambda}{\sum_{j=1}^m v^{jj}} \right)^{1/(2p)} \left\{ \frac{\sum_{j=1}^m v^{jj} f_j(t)}{\sum_{j=1}^m v^{jj}} \right\}^{-1/(2p)} = \left\{ \frac{\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} \right\}^{1/(2p)}.$$

Proof of Proposition 3. The pseudo-observations at the $(l+1)$ th iteration $Y_{ij}^{(l+1)}$ in (13) can be written in a vector form as

$$Y^{(l+1)} = \{\tilde{V}^d\}^{-1} [\tilde{V}^d \hat{\theta}_S^{(l)}(T) + \tilde{V}^{-1} \{Y - \hat{\theta}_S^{(l)}(T)\}]. \quad (\text{A8})$$

If we plug (A8) into (14), some calculations give

$$\hat{\theta}_S^{(l+1)}(T) = (\tilde{V}^d + n\lambda\Psi)^{-1} \{(\tilde{V}^d - \tilde{V}^{-1})\hat{\theta}_S^{(l)}(T) + \tilde{V}^{-1}Y\}. \quad (\text{A9})$$

Write $\hat{\theta}_S^{(l)}(T) = S^{(l)}Y$. From (A9), we have $S^{(l+1)} = (\tilde{V}^d + n\lambda\Psi)^{-1} \{(\tilde{V}^d - \tilde{V}^{-1})S^{(l)} + \tilde{V}^{-1}\}$. At convergence $S^{(l+1)} = S^{(l)} = S_*$. Hence S_* solves $S_* = (\tilde{V}^d + n\lambda\Psi)^{-1} \{(\tilde{V}^d - \tilde{V}^{-1})S_* + \tilde{V}^{-1}\}$. Simple calculations give $S_* = (\tilde{V}^{-1} + n\lambda\Psi)^{-1} \tilde{V}^{-1}$.

Proof of Proposition 4. We assume that $n \rightarrow \infty$ and $\lambda \rightarrow 0$, and λ and t are chosen similarly to Nychka (1995). We first state a few facts about the Green's function, which is associated with the solution of the differential equation of $\alpha(t)$,

$$\lambda^*(-1)^p \frac{d^{2p}\alpha(t)}{dt^{2p}} + f(t)\alpha(t) = f(t)g(t), \quad (\text{A10})$$

for any functions $f(t)$ and $g(t)$ and any arbitrary positive constant λ^* . The solution of (A10) is $\alpha(t) = \int G_{\lambda^*}(t, \tau)f(\tau)g(\tau)d\tau$, where $G_{\lambda^*}(t, \tau)$ is called the Green's function and satisfies

$$\int G_{\lambda^*}(t, \tau)g(\tau)f(\tau)d\tau - g(t) = \frac{(-1)^{p-1}\lambda^*}{a_p f(t)} g^{(2p)}(t) + o\{\lambda^*/f(t)\}, \quad (\text{A11})$$

$$\int G_{\lambda^*}^2(t, \tau)f(t)d\tau = \frac{\gamma_p}{f(t)} \left\{ \frac{\lambda^*}{f(t)} \right\}^{-1/(2p)} + o[\{\lambda^*/f(t)\}^{-1/(2p)}], \quad (\text{A12})$$

$$\iint G_{\lambda^*}(t, \tau_1)G_{\lambda^*}(t, \tau_2)r(\tau_1, \tau_2)d\tau_1d\tau_2 = \frac{r(t, t)}{f^2(t)} + o(1), \quad (\text{A13})$$

$$\int d(\tau)G_{\lambda^*}^2(t, \tau)f(\tau)d\tau = d(t) \frac{\gamma_p}{f(t)} \left\{ \frac{\lambda^*}{f(t)} \right\}^{-1/(2p)} + o[\{\lambda^*/f(t)\}^{-1/(2p)}], \quad (\text{A14})$$

where a_p and γ_p are certain constants and $d(\tau)$ and $r(\tau_1, \tau_2)$ are arbitrary functions.

Our proof starts with deriving under model (1) an asymptotic expansion of $\hat{\theta}_{\text{SI}}(t)$, the smoothing spline estimator assuming working independence with $V = \text{diag}\{v^{jj}\}$. We can show that the results in § 6 of Silverman (1984) still hold for $\hat{\theta}_{\text{SI}}(t)$ and asymptotically, at any given t ,

$$\hat{\theta}_{\text{SI}}(t) = \frac{1}{\sum_{j=1}^m v^{jj}} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m G_{\lambda^*}(t, T_{ij})v^{jj}Y_{ij} \right\} + o_p(1),$$

where $G_{\lambda^*}(t, s)$ is the Green's function associated with (A10) with $\lambda^* = \lambda/(\sum_{j=1}^m v^{jj})$ and $f(t) = \sum_{j=1}^m v^{jj}f_j(t)/\sum_{j=1}^m v^{jj}$.

Using (A11)–(A14) and the definitions of $f(t)$ and λ^* , we can show that

$$E\{\hat{\theta}_{\text{SI}}(t)\} - \theta(t) = \frac{(-1)^{p-1}\lambda}{\sum_{j=1}^m v^{jj}f_j(t)} \frac{\theta^{(2p)}(t)}{a_p} + o_p\{h^{2p}(t)\}, \quad (\text{A15})$$

$$\text{var}\{\hat{\theta}_{\text{SI}}(t)\} = \frac{\sum_{j=1}^m (v^{jj})^2 \sigma_{jj} f_j(t)}{\{\sum_{j=1}^m v^{jj} f_j(t)\}^2} \frac{\gamma_p}{n} \left\{ \frac{\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} \right\}^{-1/(2p)} + o_p[\{nh(t)\}^{-1}], \quad (\text{A16})$$

where $h(t) = [\lambda/\sum_{j=1}^m v^{jj} f_j(t)]^{1/(2p)}$. Furthermore, $\hat{\theta}_{\text{SI}}(t)$ has the expansion

$$\hat{\theta}_{\text{SI}}(t) - \theta(t) = \frac{1}{n \sum_{j=1}^m v^{jj}} \sum_{i=1}^n \sum_{j=1}^m v^{jj} G_{\lambda^*}(t, T_{ij}) \{Y_{ij} - \theta(T_{ij})\} + a(t) + o_p[\{nh(t)\}^{-\frac{1}{2}} + h^{2p}(t)], \quad (\text{A17})$$

where

$$a(t) = \frac{(-1)^{p-1}\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} \frac{\theta^{(2p)}(t)}{a_p}. \quad (\text{A18})$$

We now derive the asymptotic bias and variance of the generalised least squares smoothing spline estimator $\hat{\theta}_S(t)$. From Proposition 3, $\hat{\theta}_S(t)$ can be obtained iteratively by applying standard weighted spline smoothing to the pseudo-observations. We hence proceed by deriving the asymptotic properties of $\hat{\theta}_S^{(l)}(t)$ at each iteration and then those at convergence. This strategy allows us to apply properties of the working independence smoothing spline estimator at each iteration.

We first study the asymptotic properties of the one-step estimator $\hat{\theta}_S^{(1)}(t)$ by setting the initial estimator $\hat{\theta}_S^{(0)}(t)$ as the working independence estimator. Using (15) and setting $l=1$, we have $\hat{\theta}_S^{(1)}(t) = D_{1n} + D_{2n} + D_{3n} + o_p(1)$, where

$$D_{1n} = \frac{1}{n \sum_{j=1}^m v^{jj}} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m G_{\lambda_*}(t, T_{ij}) v^{ik} \{Y_{ik} - \theta(T_{ik})\}, \quad (\text{A19})$$

$$D_{2n} = \frac{1}{n \sum_{j=1}^m v^{jj}} \sum_{i=1}^n \sum_{j=1}^m G_{\lambda_*}(t, T_{ij}) v^{jj} \theta(T_{ij}), \quad (\text{A20})$$

$$D_{3n} = -\frac{1}{n \sum_{j=1}^m v^{jj}} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j} v^{jk} G_{\lambda_*}(t, T_{ik}) \{\hat{\theta}_S^{(0)}(T_{ik}) - \theta(T_{ik})\}. \quad (\text{A21})$$

First examine D_{3n} . Using the asymptotic expansion $\hat{\theta}_S^{(0)}(t)$ in (A17), we can write D_{3n} as $D_{3n} = D_{3n,1} + D_{3n,2}$, where

$$\begin{aligned} D_{3n,1} &= -\frac{1}{\sum_{j=1}^m v^{jj}} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j} v^{jk} G_{\lambda_*}(t, T_{ij}) \left[\frac{1}{\sum_{j=1}^m v^{jj}} \frac{1}{n} \sum_{r=1}^n \sum_{s=1}^m G_{\lambda_*}(T_{ik}, T_{rs}) v^{ss} \{Y_{rs} - \theta(T_{rs})\} \right] \\ &\quad + o_p[\{nh(t)\}^{-\frac{1}{2}}] \\ &= -\frac{1}{(\sum_{j=1}^m v^{jj})^2} \frac{1}{n} \sum_{r=1}^n \sum_{s=1}^m v^{ss} \{Y_{rs} - \theta(T_{rs})\} \sum_{j=1}^m \sum_{k \neq j} v^{jk} \iint G_{\lambda_*}(t, \tau) G_{\lambda_*}(T_{rs}, s) f_{jk}(\tau, s) d\tau ds \\ &\quad + o_p[\{nh(t)\}^{-\frac{1}{2}}], \end{aligned} \quad (\text{A22})$$

$$\begin{aligned} D_{3n,2} &= -\frac{1}{\sum_{j=1}^m v^{jj}} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j} v^{jk} G_{\lambda_*}(t, T_{ij}) b_0(T_{ik}) + o_p\{h^{2p}(t)\} \\ &= -\sum_{j=1}^m \sum_{k \neq j} v^{jk} \int G_{\lambda_*}(t, \tau) \left[E\{b_0(T_k) | T_j = \tau\} \frac{f_j(\tau)}{\sum_{j=1}^m v^{jj} f_j(\tau)} \right] f(\tau) d\tau + o_p\{h^{2p}(t)\} \\ &= -\sum_{j=1}^m \sum_{k \neq j} v^{jk} E\{b_0(T_k) | T_j = t\} f_j(t) \Big/ \left\{ \sum_{j=1}^m v^{jj} f_j(t) \right\} + o_p\{h^{2p}(t)\}, \end{aligned} \quad (\text{A23})$$

in which $b_0(\cdot) = a(\cdot)$, and we have used (A11) by setting $g(\tau) = E\{a(T_k) | T_j = \tau\} f_j(\tau) / f(\tau)$ in deriving the last equation. From (A13), $D_{3n,1} = o_p(1)$ and is negligible. Hence D_{3n} contributes to the leading asymptotic bias of $\hat{\theta}_S^{(1)}(t)$, but not to the leading asymptotic variance of $\hat{\theta}_S^{(1)}(t)$.

The term D_{1n} has mean 0 and does not contribute to the bias of $\hat{\theta}_S^{(1)}(t)$, while D_{2n} only contributes to the bias of $\hat{\theta}_S^{(1)}(t)$. Using (A15) with $a(t)$ defined in (A18), we have $D_{2n} - \theta(t) = a(t) + o_p\{h^{2p}(t)\}$. The asymptotic bias of $\hat{\theta}_S^{(1)}(t)$ hence is $E\{\hat{\theta}_S^{(1)}(t)\} - \theta(t) = b_1(t) + o_p\{h^{2p}(t)\}$, where

$$b_1(t) = a(t) - \left[\sum_{j=1}^m \sum_{k \neq j} v^{jk} E\{b_0(T_k) | T_j = t\} f_j(t) \Big/ \left\{ \sum_{j=1}^m v^{jj} f_j(t) \right\} \right]. \quad (\text{A24})$$

The leading variance term of $\hat{\theta}_S^{(1)}(t)$ is determined by D_{1n} . Let $G_{\lambda_*}(t, T_i) = \{G_{\lambda_*}(t, T_{i1}), \dots, G_{\lambda_*}(t, T_{im})\}^T$ and rewrite $D_{1n} = (n \sum_{j=1}^m v^{jj})^{-1} \sum_{i=1}^n G_{\lambda_*}(t, T_i)^T V^{-1} \{Y_i - \theta(T_i)\}$. We have

$$\begin{aligned} \text{var} \{\hat{\theta}_S^{(1)}(t)\} &= \frac{1}{(n \sum_{j=1}^m v^{jj})^2} \sum_{i=1}^n G_{\lambda_*}(t, T_i)^T V^{-1} \Sigma V^{-1} G_{\lambda_*}(t, T_i) + o_p[\{nh(t)\}^{-1}] \\ &= F_{1n} + F_{2n} + o_p[\{nh(t)\}^{-1}], \end{aligned}$$

where, denoting by c_{jk} the (j, k) th element of $C = V^{-1}\Sigma V^{-1}$ and using (A13)–(A14), we have

$$\begin{aligned}
F_{1n} &= \frac{1}{(n \sum_{j=1}^m v^{jj})^2} \sum_{i=1}^n \sum_{j=1}^m G_{\lambda*}^2(t, T_{ij}) c_{jj} \\
&= \frac{1}{n \sum_{j=1}^m v^{jj}} \int \left\{ \frac{\sum_{j=1}^m c_{jj} f_j(\tau)}{\sum_{j=1}^m v^{jj} f_j(\tau)} \right\} G_{\lambda*}^2(t, \tau) f(\tau) d\tau + o_p[\{nh(t)\}^{-1}] \\
&= \frac{\sum_{j=1}^m c_{jj} f_j(t)}{\{\sum_{j=1}^m v^{jj} f_j(t)\}^2} \frac{\gamma_p}{n} \left\{ \frac{\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} \right\}^{-1/(2p)} + o_p[\{nh(t)\}^{-1}], \\
F_{2n} &= \frac{1}{(n \sum_{j=1}^m v^{jj})^2} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j} G_{\lambda*}(t, T_{ij}) G_{\lambda*}(t, T_{ik}) c_{jk} \\
&= \frac{1}{n \{\sum_{j=1}^m v^{jj}\}^2} \sum_{j=1}^m \sum_{k \neq j} c_{jk} \iint G_{\lambda*}(t, \tau) G_{\lambda*}(t, s) f_{jk}(\tau, s) d\tau ds + o_p(n^{-1}) \\
&= \frac{1}{n \{f(t) \sum_{j=1}^m v^{jj}\}^2} \sum_{j=1}^m \sum_{k \neq j} c_{jk} f_{jk}(t, t) + o_p(n^{-1}) = O_p(n^{-1}).
\end{aligned}$$

It follows that

$$\text{var} \{\hat{\theta}_S^{(1)}(t)\} = \frac{\sum_{j=1}^m c_{jj} f_j(t)}{\{\sum_{j=1}^m v^{jj} f_j(t)\}^2} \frac{\gamma_p}{n} \left\{ \frac{\lambda}{\sum_{j=1}^m v^{jj} f_j(t)} \right\}^{-1/(2p)} + o_p[\{nh(t)\}^{-1}]. \quad (\text{A25})$$

The one-step estimator $\hat{\theta}_S^{(1)}(t)$ has the asymptotic expansion

$$\hat{\theta}_S^{(1)}(t) - \theta(t) = \frac{1}{n \sum_{j=1}^m v^{jj}} \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m G_{\lambda*}(t, T_{ij}) v^{jk} \{Y_{ik} - \theta(T_{ik})\} + b_1(t) + o_p[\{nh(t)\}^{-\frac{1}{2}} + h^{2p}(t)], \quad (\text{A26})$$

where the bias term $b_1(t)$ is defined in (A24).

Now consider the estimator of $\theta(t)$ at the second iteration $\hat{\theta}_S^{(2)}(t)$. Using (15), we have $\hat{\theta}_S^{(2)}(t) = D_{1n} + D_{2n} + D_{3n} + o_p(1)$, where D_{1n} , D_{2n} and D_{3n} are as given in (A19)–(A21) except that $\hat{\theta}_S^{(0)}(T_{ik})$ is replaced by $\hat{\theta}_S^{(1)}(T_{ik})$ in (A21). Using (A26), we write $D_{3n} = D_{3n,1} + D_{3n,2}$, where $D_{3n,2}$ is the same as (A23) except that $b_0(T_{ik})$ is replaced by $b_1(T_{ik})$, and $D_{3n,1}$ is slightly different from (A22) and is

$$\begin{aligned}
D_{3n,1} &= -\frac{1}{\sum_{j=1}^m v^{jj}} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \sum_{k \neq j} v^{jk} G_{\lambda*}(t, T_{ij}) \left[\frac{1}{\sum_{j=1}^m v^{jj}} \frac{1}{n} \sum_{r=1}^n \sum_{s=1}^m \sum_{\ell=1}^m G_{\lambda*}(T_{ik}, T_{rs}) v^{s\ell} \{Y_{r\ell} - \theta(T_{r\ell})\} \right] \\
&\quad + o_p[\{nh(t)\}^{-\frac{1}{2}}].
\end{aligned}$$

Similar calculations to those in (A22)–(A23) show that $D_{3n,1} = o_p(1)$ and $D_{3n,2}$ is the same as (A23) except that $b_0(\cdot)$ is replaced by $b_1(\cdot)$. Hence the asymptotic variance of $\hat{\theta}_S^{(2)}(t)$ is exactly the same as (A25) and the asymptotic bias is given by $\hat{\theta}_S^{(2)}(t) = b_2(t) + o_p\{h^{2p}(t)\}$, where $b_2(t)$ is the same as the right-hand side of (A24) except that $b_0(\cdot)$ is replaced by $b_1(\cdot)$. Hence the second iteration does not change the variance but only makes a refinement to the bias. The asymptotic expansion of $\hat{\theta}_S^{(2)}(t)$ is the same as (A26) except that $b_1(t)$ is replaced by $b_2(t)$.

Using induction, one can easily see that the expansion of $\hat{\theta}_S^{(l+1)}(t)$ at the $(l+1)$ th iteration ($l > 2$) is exactly the same as for $\hat{\theta}_S^{(2)}(t)$ except that $b_2(t)$ is replaced by $b_{l+1}(t)$, which satisfies

$$b_{l+1}(t) = a(t) - \left[\sum_{j=1}^m \sum_{k \neq j} v^{jk} E\{b_l(T_k) | T_j = t\} f_j(t) \right] / \left\{ \sum_{j=1}^m v^{jj} f_j(t) \right\}.$$

Its asymptotic variance takes the same form as that in (4). In other words, further iterations only make refinements to the bias but not the variance.

At convergence, $\hat{\theta}_S^{(l+1)}(t)$ converges to $\hat{\theta}_S(t)$, from Proposition 3, and $b_{l+1}(t)$ to $b_S(t)$, which satisfies

$$b_S(t) = a(t) - \left[\sum_{j=1}^m \sum_{k \neq j} v^{jk} E\{b_S(T_k) | T_j = t\} f_j(t) \right] / \left\{ \sum_{j=1}^m v^{jj} f_j(t) \right\},$$

and is the asymptotic bias of $\hat{\theta}_S(t)$. Part (i) of Proposition 4 follows immediately. The asymptotic variance of $\hat{\theta}_S(t)$ is the same as the right-hand side of equation (A25). This gives part (ii) of Proposition 4. A direct application of the Cauchy–Schwartz inequality gives part (iii). One can easily see that the asymptotic expansion of $\hat{\theta}_S(t)$ given in part (iv) holds. It should be noted that the asymptotic expansion, bias and variance do not depend on choice of the initial consistent estimator $\hat{\theta}_S^{(0)}(t)$.

REFERENCES

- BRUMBACK, B. & RICE, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with Discussion). *J. Am. Statist. Assoc.* **93**, 961–1006.
- GREEN, P. J. & SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalised Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.
- FAN, J. & ZHANG, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Statist. Soc. B* **62**, 303–22.
- HOOVER, D. R., RICE, J. A., WU, C. O. & YANG, Y. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–22.
- LIANG, K. Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13–22.
- LIN, X. & CARROLL, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Assoc.* **95**, 520–34.
- LIN, X. & CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalised estimating equations. *J. Am. Statist. Assoc.* **96**, 1045–56.
- LIN, X. & ZHANG, D. (1999). Inference in generalised additive mixed models using smoothing splines. *J. R. Statist. Soc. B* **61**, 381–400.
- MALOW, B. A., KUSHWAHA, R., LIN, X., MORTON, K. L. & ALDRICH, M. S. (1997). Relationship of interictal epileptiform discharges to sleep depth in temporal lobe epilepsy. *EEG Clin. Neurophysiol.* **102**, 20–6.
- NYCHKA, D. (1995). Splines as local smoothers. *Ann. Statist.* **23**, 1175–97.
- RUCKSTUHL, A., WELSH, A. H. & CARROLL, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statist. Sinica* **10**, 51–71.
- SEVERINI, T. A. & STANISWALIS, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Assoc.* **89**, 501–11.
- SILVERMAN, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.
- VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G. & WELHAM, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with Discussion). *Appl. Statist.* **48**, 269–311.
- WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.
- WANG, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90**, 43–52.
- WANG, Y. (1998). Mixed effects smoothing spline analysis of variance. *J. R. Statist. Soc. B* **60**, 159–74.
- WELSH, A. H., LIN, X. & CARROLL, R. J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *J. Am. Statist. Assoc.* **97**, 482–93.
- ZEGER, S. L. & DIGGLE, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.
- ZHANG, D., LIN, X., RAZ, J. & SOWERS, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *J. Am. Statist. Assoc.* **93**, 710–9.

[Received December 2002. Revised July 2003]