

Marginal Longitudinal Nonparametric Regression: Locality and Efficiency of Spline and Kernel Methods

Alan H. Welsh, Xihong Lin and Raymond J. Carroll*

April 30, 2001

Abstract

We consider nonparametric regression in a longitudinal marginal model of GEE-type with a time-varying covariate in the situation where the number of observations per subject is finite, while the number of subjects is large. In such models, the basic shape of the regression function is affected only by the covariate values and not otherwise by the ordering of the observations. Two methods of estimating the nonparametric function can be considered: kernel methods and spline methods. Recently, surprising evidence has emerged that suggests that for kernel methods previously proposed in the literature, it is generally asymptotically preferable to ignore the correlation structure in our marginal model, and instead to assume that the data are independent, i.e., working independence in the GEE jargon. As seen through equivalent kernel results, in univariate independent data problems splines and kernels have similar behavior: smoothing splines are equivalent to kernel regression with a specific higher-order kernel, and hence smoothing splines are local. This equivalence suggests that in our marginal model, working independence might be preferable for spline methods. Our results suggest the opposite: via theoretical and numerical calculations, we provide evidence suggesting that for our marginal model, marginal smoothing and penalized regression splines are not local in their behavior. In contrast to the kernel results, our evidence suggests that when using spline methods, it is worthwhile to account for the correlation structure. In the light of this vastly different behavior, we show that spline methods are more efficient than the previously proposed kernel methods for our marginal model.

KEY WORDS: Clustered data; Generalized estimating equations; Equivalent kernels; Generalized least squares; Kernel regression; Longitudinal models; Nonparametric regression; P-splines; Regression splines; Repeated measures; Smoothing Splines; Weighted least squares.

Short title: Marginal Longitudinal Nonparametric Regression

AUTHOR AFFILIATIONS AND ACKNOWLEDGMENTS

*Alan H. Welsh (alan.welsh@anu.edu.au) is Professor, Faculty of Mathematical Studies, University of Southampton, Southampton, Hants SO17 1BJ, UK. His research at the Australian National University was supported by ARC Large Research Grant A00000506.

Xihong Lin (xlin@sph.umich.edu) is Associate Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA. Her research was supported by a grant from the National Cancer Institute (CA76404).

Raymond J. Carroll (carroll@stat.tamu.edu) is Distinguished Professor, Department of Statistics and Department of Epidemiology and Biostatistics, Texas A&M University, College Station, TX 77843-3143, USA. His research was supported by a grant from the National Cancer Institute (CA57030) and through the Texas A&M Center for Environment and Rural Health by a grant from the National Institute of Environmental Health Sciences (P30-ES09106). Carroll's research took place during a visit to the Centre for Mathematics and its Applications at the Australian National University, with partial support from ARC Large Research Grant A00000506.

We thank the editor, the associate editor and two referees for their helpful comments and suggestions that improved the article.

Marginal Longitudinal Nonparametric Regression: Locality and Efficiency of Spline and Kernel Methods

Abstract

We consider nonparametric regression in a longitudinal marginal model of GEE-type with a time-varying covariate in the situation where the number of observations per subject is finite, while the number of subjects is large. In such models, the basic shape of the regression function is affected only by the covariate values and not otherwise by the ordering of the observations. Two methods of estimating the nonparametric function can be considered: kernel methods and spline methods. Recently, surprising evidence has emerged that suggests that for kernel methods previously proposed in the literature, it is generally asymptotically preferable to ignore the correlation structure in our marginal model, and instead to assume that the data are independent, i.e., working independence in the GEE jargon. As seen through equivalent kernel results, in univariate independent data problems splines and kernels have similar behavior: smoothing splines are equivalent to kernel regression with a specific higher-order kernel, and hence smoothing splines are local. This equivalence suggests that in our marginal model, working independence might be preferable for spline methods. Our results suggest the opposite: via theoretical and numerical calculations, we provide evidence suggesting that for our marginal model, marginal smoothing and penalized regression splines are not local in their behavior. In contrast to the kernel results, our evidence suggests that when using spline methods, it is worthwhile to account for the correlation structure. In the light of this vastly different behavior, we show that spline methods are more efficient than the previously proposed kernel methods for our marginal model.

KEY WORDS: Clustered data; Generalized estimating equations; Equivalent kernels; Generalized least squares; Kernel regression; Longitudinal models; Nonparametric regression; P-splines; Regression splines; Repeated measures; Smoothing Splines; Weighted least squares.

Short title: Marginal Longitudinal Nonparametric Regression

1 INTRODUCTION

Our purpose is to investigate whether smoothing and penalized regression splines have the same type of local behavior that is exhibited by previously proposed kernel methods in marginal models of GEE-type for longitudinal data. We conclude that while spline and previously proposed kernel methods are almost without exception essentially equivalent for nonparametric regression in the usual independent-data case, they have different behavior in the longitudinal context, and this different behavior leads to differences in optimal strategies for estimation and weighting.

In what follows, we consider only the simplest possible marginal model. For $i = 1, \dots, n$ individuals, there are $j = 1, \dots, m$ (with $m \ll n$) observed responses Y_{ij} and predictors X_{ij} . In other words, we consider the case when the number of observations per subject is small while the number of subjects is relatively large, which is common in many longitudinal studies. We assume X_{ij} is a time-varying covariate (i.e., X_{ij} varies within each subject) with marginal continuous density f_j . Let \mathbf{Y}_i be the vector of responses on the i th individual and let \mathbf{X}_i be defined similarly. By a marginal nonparametric regression model, we mean that the mean and covariance of the responses are given by

$$E(Y_{ij}|\mathbf{X}_i) = g(X_{ij}); \tag{1}$$

$$\text{cov}(\mathbf{Y}_i|\mathbf{X}_i) = \Sigma, \tag{2}$$

where $g(\cdot)$ is an unknown function. Note that we here assume $E(Y_{ij}|\mathbf{X}_i) = E(Y_{ij}|X_{ij})$ (Pepe and Couper, 1997). The key feature of the marginal model is that the means in (1) have the same shape and differ only because the covariates differ for the observations within the individual. The model (1)–(2) obviously also applies to more general clustered data and, we believe, is the natural analogue to the usual marginal parametric or GEE-type model, see for example Liang and Zeger (1986) and Diggle, Liang and Zeger (1994).

Nonparametric longitudinal regression in the marginal model using kernel methods has been investigated by a number of authors, see Severini and Staniswalis (1994), Zeger and Diggle (1994), Wu, Chiang and Hoover (1998), and Hoover, et al. (1998), among others. Severini and Staniswalis (1994) estimate the covariance matrix Σ in (2) and use this in their

kernel construction of the nonparametric regression estimate. The other papers effectively ignore the correlation structure entirely and “pretend” that the data are really independent, this being the so-called “working independence” method. Ruckstuhl, Welsh and Carroll (2000) and Lin and Carroll (2000) provided theoretical evidence in support of the working independence method. In fact, they showed that for many situations using the Severini-Staniswalis construction of kernel estimating equations as well as some modifications of it, the working independence method is most efficient in terms of mean squared error. That is, for the kernel methods proposed in the literature, it is generally better to ignore the correlation structure (2) entirely.

In this paper, we use “kernel methods” to describe the previously proposed methods which consist essentially of the method introduced by Severini and Staniswalis (1994) and some natural alternatives. “Kernel methods” should not be read as “all possible kernel-type methods”. Rather than annoying the reader by continually making this distinction, except in the discussion we will adopt the convention that “kernel methods” means “previously proposed kernel methods.”

The purpose of this paper is to investigate whether the kernel result holds for marginal spline methods, especially smoothing splines (see Wahba, 1990; Green and Silverman, 1994 among many others) and penalized regression splines (P-splines, see Eilers and Marx, 1996; Ruppert and Carroll, 2000). One might reasonably suppose that the kernel result will hold for splines because, in the case of independent data, Silverman (1984) showed that smoothing splines are asymptotically equivalent to kernel regression with a particular higher-order kernel. It would not appear to be too great a leap to conclude that if smoothing splines and kernels are equivalent in the usual independence case, then they ought also to be equivalent in marginal longitudinal nonparametric models, and hence with splines it is also better to ignore the correlation structure entirely. For simplicity, we assume in this paper that the true covariance matrix \mathbf{V} is known and compare the finite sample and asymptotic performance of spline and kernel methods when using either working independence or the true covariance.

In this paper, we provide three pieces of evidence to suggest that the kernel results are essentially irrelevant for smoothing and penalized regression splines. First, in Section 2, we construct the equivalent kernels for smoothing splines, penalized regression splines

(P-splines) and kernel methods in the longitudinal marginal model. In Section 3, we use these theoretical results to compute the equivalent kernels numerically in finite sample cases to show splines are non-local. In Section 4, the non-local behavior of smoothing splines exhibited in these finite sample results is shown to persist asymptotically (as $n \rightarrow \infty$ with m fixed). These results indicate that marginal smoothing splines and P-splines do not have the same local behavior as kernel methods, unless the covariance matrix (2) is diagonal. The non-locality of splines depends on the within-cluster correlation of the outcomes Y_{ij} and the within-cluster correlation of the covariates X_{ij} . Stronger correlation among the Y_{ij} and weaker correlation among the X_{ij} make splines more non-local.

Second, in Section 5, we compare the mean squared errors (MSE) of marginal penalized regression, smoothing splines and kernel methods. We show that asymptotically the variances of the smoothing spline and P-spline estimators are minimized when the true covariance is used. This suggests that the mean squared error is minimized when using the true covariance matrix rather than working independence, a result exactly the opposite of what happens for kernel methods. We further show that spline methods give more efficient estimators of the nonparametric function than kernel methods. We then provide support for this finding by comparing the exact MSEs in finite samples.

Finally, in Section 6, we provide simulation results for smoothing splines, P-splines and kernels. The spline work is of independent interest because it includes modified CV and GCV estimates of the smoothing parameter. The results confirm the theory, namely in contrast to kernels, estimating the correlation structure in the data improves the efficiency of the spline estimates. Further, spline methods give more efficient estimators of the nonparametric function than kernel methods. Concluding comments are given in Section 7.

2 Equivalent Kernels: Theory

2.1 Preliminaries

The idea of the *equivalent kernel* of an estimator is to measure the contribution of a single observation located at, say X_{ij} , to the predicted value (based on the estimator) of $g(\cdot)$ at

s . Thus, if we can write the predicted value at s as $\hat{g}(s) = (nm)^{-1} \sum_{i=1}^n \sum_{j=1}^m G_{Nij}(s, \mathbf{X}) Y_{ij}$, where $N = nm$ is the total sample size, then the equivalent kernel is given by $G_{Nij}(s, \mathbf{X})$. Our interest is in exploring the shape of this function for different estimators as we vary s . We classify estimators as local or global according to whether a point at X_{ij} is weighted only when s is near X_{ij} or not. This investigation is carried out in finite samples numerically (Section 3) and in large samples theoretically (Section 4).

Define $\mathbf{g}_i = \{g(X_{i1}), \dots, g(X_{im})\}^T$, $\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)^T$ and $\mathbf{g} = (\mathbf{g}_1^T, \dots, \mathbf{g}_n^T)^T$. Then the model in (1)-(2) can be written in matrix notation as $E(\mathbf{Y}) = \mathbf{g}$ and $\text{cov}(\mathbf{Y}) = \mathbf{V}$, where $\mathbf{V} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$. Let \mathbf{W}_{work}^{-1} be the working covariance matrix, e.g., $\mathbf{W}_{indp}^{-1} = \sigma_\epsilon^2 \mathbf{I}_N$ for working independence or $\mathbf{W}^{-1} = \mathbf{V}$ when the covariance structure is modeled. In either case, \mathbf{W}_{work}^{-1} is a block diagonal matrix. For later use, note that with $\mathbf{G}_N(s, \mathbf{X}) = \{G_{N11}(s, \mathbf{X}), \dots, G_{Nnm}(s, \mathbf{X})\}^T$, we can write $\hat{g}(s) = (nm)^{-1} \mathbf{G}_N(s, \mathbf{X})^T \mathbf{Y}$ and its exact mean squared error as

$$\text{MSE}_N(s, \mathbf{W}_{work}) = (nm)^{-2} \mathbf{G}_N(s, \mathbf{X})^T \mathbf{V} \mathbf{G}_N(s, \mathbf{X}) + \{(nm)^{-1} \mathbf{G}_N(s, \mathbf{X})^T \mathbf{g} - g(s)\}^2. \quad (3)$$

2.2 Local Polynomial Estimators

Define the $N \times (1+p)$ matrix \mathbf{X}_s to have entries in the k th column given by $(X_{ij} - s)^{k-1}$, and the $N \times N$ matrix $\mathbf{K} = \text{diag}\{K_h(X_{11} - s), \dots, K_h(X_{nm} - s)\}$, where $K_h(x) = h^{-1}K(x/h)$, $K(\cdot)$ is a kernel function and h is a bandwidth. For parameters $(a, b) = (0, 1)$ (Severini and Staniswalis, 1994; Ruckstuhl, Welsh and Carroll, 2000) and $(a, b) = (0.5, 0.5)$ (Lin and Carroll, 2000), the prediction at s is given by

$$\hat{g}_h(s) = (1, \mathbf{0}_p^T) (\mathbf{X}_s^T \mathbf{K}^a \mathbf{W}_{work} \mathbf{K}^b \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{K}^a \mathbf{W}_{work} \mathbf{K}^b \mathbf{Y}. \quad (4)$$

It follows that the equivalent kernel for Y_{ij} is

$$G_{Nij}(s, \mathbf{X}) = N(1, \mathbf{0}_p^T) (\mathbf{X}_s^T \mathbf{K}^a \mathbf{W}_{work} \mathbf{K}^b \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{K}^a \mathbf{W}_{work} \boldsymbol{\delta}_{ij} \{K_h(X_{ij} - s)\}^b, \quad (5)$$

where $\boldsymbol{\delta}_{ij}$ is the N -vector of zeros with a one in the ij th position.

2.3 P-Splines

Penalized regression splines or P-splines are defined by Eilers and Marx (1996) and Ruppert and Carroll (2000): we use the formulation of the latter authors. Let the order of the P-spline

be p , and consider for $k = 1, \dots, M$ basis functions $B_{kM}(s) = (s - \xi_k)^p I(s \geq \xi_k)$ for fixed knots ξ_k . Denote the block diagonal matrix \mathbf{W}_{work}^{-1} by $\mathbf{W}_{work}^{-1} = \text{diag}\{\boldsymbol{\Sigma}_{work}\}$. Define the $(1 + p + M)$ -vector $\mathbf{q}(s) = \{1, s, \dots, s^p, B_{1M}(s), \dots, B_{MM}(s)\}^T$, the $(1 + p + M) \times (1 + p + M)$ matrix $\mathbf{P} = \text{block diag}(\mathbf{0}_{1+p}, \mathbf{I}_M)$, the $m \times (1 + p + M)$ matrices $\mathbf{Q}_i^T = \{\mathbf{Q}(X_{i1}), \dots, \mathbf{Q}(X_{im})\}$, and the $(1 + p + M) \times (1 + p + M)$ matrix $\mathbf{H}_n = n^{-1} \sum_{i=1}^n \mathbf{Q}_i^T \boldsymbol{\Sigma}_{work}^{-1} \mathbf{Q}_i$. The P-spline minimizes in $\boldsymbol{\Theta}$

$$\sum_{i=1}^n (\mathbf{Y}_i - \mathbf{Q}_i \boldsymbol{\Theta})^T \boldsymbol{\Sigma}_{work}^{-1} (\mathbf{Y}_i - \mathbf{Q}_i \boldsymbol{\Theta}) + n \lambda \boldsymbol{\Theta}^T \mathbf{P} \boldsymbol{\Theta},$$

where λ is a smoothing parameter. The fitted value at s is

$$\hat{g}(s) = \mathbf{q}(s)^T (\mathbf{H}_n + \lambda \mathbf{P})^{-1} n^{-1} \sum_{i=1}^n \mathbf{Q}_i^T \boldsymbol{\Sigma}_{work}^{-1} \mathbf{Y}_i.$$

It follows that the equivalent kernel for Y_{ij} is

$$G_{Nij}(s, \mathbf{X}) = \frac{N}{n} \mathbf{q}(s)^T (\mathbf{H}_n + \lambda \mathbf{P})^{-1} \mathbf{Q}_i^T \boldsymbol{\Sigma}_{work}^j, \quad (6)$$

where $\boldsymbol{\Sigma}_{work}^j$ is the j th column of $\boldsymbol{\Sigma}_{work}^{-1}$.

2.4 Smoothing splines

Let $\mathbf{X}_o = \{X_{(1)}, \dots, X_{(N)}\}^T$ be the ordered values of the X_{ij} . Following Green and Silverman (1994, Chapter 2, Sections 3.5 and 6.3), put $l_i = X_{(i+1)} - X_{(i)}$ for $i = 1, \dots, N - 1$. Define the $N \times (N - 2)$ matrix \mathbf{C} to have entries c_{ij} for $i = 1, \dots, N$ and $j = 2, \dots, N - 1$ given by $c_{j-1,j} = l_{j-1}^{-1}$, $c_{jj} = -l_{j-1}^{-1} - l_j^{-1}$, $c_{j+1,j} = l_j^{-1}$ and $c_{ij} = 0$ if $|i - j| \geq 2$. Also define the $(N - 2) \times (N - 2)$ matrix \mathbf{R} to have entries r_{ij} for $i, j = 2, \dots, N - 1$ given by $r_{ii} = (l_{i-1} + l_i)/3$ for $i = 2, \dots, N - 1$ and $r_{i,i+1} = r_{i+1,i} = l_i/6$ for $i = 2, \dots, N - 2$, and $r_{ij} = 0$ for $|i - j| \geq 2$. Note that the columns of \mathbf{C} and \mathbf{R} are numbered in a non-standard way starting at $j = 2$ such that the top left element of \mathbf{C} and \mathbf{R} are q_{12} and r_{12} . Then put $\boldsymbol{\Psi} = \mathbf{C} \mathbf{R}^{-1} \mathbf{C}^T$.

Define \mathbf{U} to be an incidence matrix relating \mathbf{X} to the ordered values \mathbf{X}_o such that the $\{(i, j), k\}$ th element of \mathbf{U} is 1 if $X_{ij} = X_{(k)}$ and 0 otherwise ($i = 1, \dots, n; j = 1, \dots, m; k = 1, \dots, N$), i.e., $\mathbf{X} = \mathbf{U} \mathbf{X}_o$. Note that $\mathbf{U}^T \mathbf{U} = \mathbf{I}$. Define $\mathbf{g} = g(\mathbf{X}_o)$. The smoothing spline minimizes

$$(\mathbf{Y} - \mathbf{U} \mathbf{g})^T \mathbf{W}_{work} (\mathbf{Y} - \mathbf{U} \mathbf{g}) + \lambda \int g''(t)^2 dt = (\mathbf{Y} - \mathbf{U} \mathbf{g})^T \mathbf{W}_{work} (\mathbf{Y} - \mathbf{U} \mathbf{g}) + \lambda \mathbf{g}^T \boldsymbol{\Psi} \mathbf{g},$$

where \mathbf{W}_{work} was defined in Section 2.3 as $\mathbf{W}_{work}^{-1} = \text{diag}(\boldsymbol{\Sigma}_{work})$. The spline fit can be written as $\hat{\mathbf{g}}_\lambda = (\mathbf{U}^T \mathbf{W}_{work} \mathbf{U} + \lambda \boldsymbol{\Psi})^{-1} \mathbf{U}^T \mathbf{W}_{work} \mathbf{Y}$. The bias and variance at the N knots are

$$\begin{aligned} E(\hat{\mathbf{g}}_\lambda - \mathbf{g}) &= \{(\mathbf{U}^T \mathbf{W}_{work} \mathbf{U} + \lambda \boldsymbol{\Psi})^{-1} \mathbf{U}^T \mathbf{W}_{work} \mathbf{U} - \mathbf{I}_N\} \mathbf{g} \\ \text{cov}(\hat{\mathbf{g}}_\lambda) &= (\mathbf{U}^T \mathbf{W}_{work} \mathbf{U} + \lambda \boldsymbol{\Psi})^{-1} \mathbf{U}^T \mathbf{W}_{work} \mathbf{V} \mathbf{W}_{work} \mathbf{U} (\mathbf{U}^T \mathbf{W}_{work} \mathbf{U} + \lambda \boldsymbol{\Psi})^{-1}. \end{aligned}$$

Denote the second derivatives of g by $\gamma_i = g''(X_{(i)})$. By definition, $\gamma_1 = \gamma_N = 0$ so let $\boldsymbol{\gamma} = (\gamma_2, \dots, \gamma_{N-1})^T$ be the $(N-2)$ -vector of nontrivial second derivatives. We can estimate $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}_\lambda = \mathbf{R}^{-1} \mathbf{C}^T \hat{\mathbf{g}}_\lambda$. Define the $N \times N$ and $(N-2) \times N$ matrices $\mathbf{F} = (\mathbf{U}^T \mathbf{W}_{work} \mathbf{U} + \lambda \mathbf{K})^{-1} \mathbf{U}^T \mathbf{W}_{work} \mathbf{U}$ and $\mathbf{D} = \mathbf{R}^{-1} \mathbf{C}^T \mathbf{F}$, so that $\hat{\mathbf{g}}_\lambda = \mathbf{F} \mathbf{Y}_o$ and $\hat{\boldsymbol{\gamma}}_\lambda = \mathbf{D} \mathbf{Y}_o$, where \mathbf{Y}_o satisfies $\mathbf{Y} = \mathbf{U} \mathbf{Y}_o$, i.e., \mathbf{Y}_o is obtained by ordering \mathbf{Y} according to the order of \mathbf{X}_o . The notation gets complicated so we suppress the dependence of \mathbf{F} and \mathbf{D} on λ . Write $\mathbf{F}^T = (\mathbf{f}_1, \dots, \mathbf{f}_N)$ and $\mathbf{D}^T = (\mathbf{d}_1, \dots, \mathbf{d}_N)$, so $\hat{g}(X_{(i)}) = \mathbf{f}_i^T \mathbf{Y}_o$ etc.

The smoothing spline fit at an arbitrary point s is given by

$$\hat{g}(s) = \begin{cases} \left\{ \mathbf{f}_1 - (X_{(1)} - s) \left(\frac{\mathbf{f}_2 - \mathbf{f}_1}{l_1} - \frac{1}{6} l_1 \mathbf{d}_2 \right) \right\}^T \mathbf{Y}_o & \text{if } s \leq X_{(1)} \\ \left\{ \frac{(s - X_{(j)}) \mathbf{f}_{j+1} + (X_{(j+1)} - s) \mathbf{f}_j}{l_j} - \frac{1}{6} (s - X_{(j)}) (X_{(j+1)} - s) \right. \\ \quad \times \left[\left(1 + \frac{s - X_{(j)}}{l_j} \right) \mathbf{d}_{j+1} + \left(1 + \frac{X_{(j+1)} - s}{l_j} \right) \mathbf{d}_j \right] \left. \right\}^T \mathbf{Y}_o & \text{if } X_{(j)} \leq s < X_{(j+1)} \\ \left\{ \mathbf{f}_N + (s - X_{(N)}) \left(\frac{\mathbf{f}_N - \mathbf{f}_{N-1}}{l_{N-1}} + \frac{1}{6} l_{N-1} \mathbf{d}_{N-1} \right) \right\}^T \mathbf{Y}_o & \text{if } X_{(N)} < s. \end{cases}$$

It follows that the equivalent kernel is, apart from a factor N ,

$$G_{Ni}(s, \mathbf{X}) = \begin{cases} f_{1i} - (X_{(1)} - s) \left(\frac{f_{2i} - f_{1i}}{l_1} - \frac{1}{6} l_1 d_{2i} \right) & \text{if } s \leq X_{(1)} \\ \frac{(s - X_{(j)}) f_{j+1,i} + (X_{(j+1)} - s) f_{ji}}{l_j} - \frac{1}{6} (s - X_{(j)}) (X_{(j+1)} - s) \\ \quad \times \left[\left(1 + \frac{s - X_{(j)}}{l_j} \right) d_{j+1,i} + \left(1 + \frac{X_{(j+1)} - s}{l_j} \right) d_{ji} \right] & \text{if } X_{(j)} \leq s < X_{(j+1)} \\ f_{Ni} + (s - X_{(N)}) \left(\frac{f_{Ni} - f_{N-1,i}}{l_{N-1}} + \frac{1}{6} l_{N-1} d_{N-1,i} \right) & \text{if } X_{(N)} < s. \end{cases} \quad (7)$$

3 Equivalent Kernels: Numerical Results

To examine the locality of splines and kernels, we first study numerically their equivalent kernels obtained in Section 2. We generated the predictors X_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, m$ as independent uniform random variables on the interval $[-2, 2]$ and computed the equivalent kernels at $X_{ij} = 0.25$. The common variance of the Y_{ij} was $\sigma_\epsilon^2 = 4$. We considered three correlation structures: (1) autoregressive with correlation ρ ; (2) exchangeable with common correlation ρ ; (3) an unstructured and nearly singular correlation matrix. We used a fixed value of the smoothing parameter λ , although sensitivity analyses not reported here show that the choice of λ does not change the qualitative conclusions. In the following figures, we used $\lambda = 0.01$.

Figure 1 displays the equivalent kernels (5) for the kernel method in the exchangeable case with $\rho = 0.0, 0.4, 0.8$, $n = 35$, $m = 3$, $h = 0.25$, $a = b = 0.5$ and the Gaussian kernel. We assumed the true exchangeable correlation in our kernel estimation. Clearly, the kernel is local even when the correlation structure is taken into account.

In Figures 2 and 3, we display the equivalent kernels (7) for smoothing splines in the exchangeable case with $\rho = 0.0, 0.4, 0.8$, $m = 3$ and $n = 35$ or $n = 50$ respectively. We see from these figures that under independence $\rho = 0.0$, the equivalent kernel is truly local, giving essentially no influence to X when predicting outside the range $[-1, 1]$. However, by the time $\rho = 0.8$, there is considerable weight given to X when predicting in the more distant range $[-2, -1]$. This figure shows clearly that as the correlation increases, the smoothing spline becomes increasingly non-local.

A somewhat more striking case of smoothing splines is given in Figure 4. Here $n = 35$, $m = 3$, and we contrasted the independence case, autocorrelation with $\rho = 0.8$, and an unstructured and nearly singular correlation matrix which has correlation between measurements 1 and 2 and between measurements 2 and 3 equal to 0.80, and correlation between measurements 1 and 3 equal to 0.3. The non-local behavior of smoothing splines is clear even for the autocorrelation case, and remarkably large for the unstructured case. While not displayed here, the equivalent kernel (6) for the P-spline of order $p = 2$ with 35 equally spaced fixed knots is a quadratic function which is non-local.

4 Equivalent Kernels: Asymptotic Results

To gain insight into the finite sample numerical results in Section 3, we provide in this section theoretical justification for the different local behavior of smoothing splines and kernels observed in Section 3. Without loss of generality, we assume Σ is a correlation matrix when stating the results.

We first show that the kernel estimator $\hat{g}_h(s)$ in equation (4) is local asymptotically. The kernel estimator is clearly local under working independence. Hence suppose that the true covariance Σ is used in $\hat{g}_h(s)$ and let $a = b = 1/2$. Denote by \mathbf{E} the $(p+1) \times (p+1)$ matrix with (j, k) th element $\int z^{j+k-2} K(z) dz$. Some calculations using Theorem 4 of Lin and Carroll (2000) show that the kernel estimator $\hat{g}_h(s)$ in (4) is asymptotically equivalent to

$$\hat{g}_h(s) = n^{-1} \sum_{i=1}^n \sum_{j=1}^m \frac{\left[\sum_{k=1}^{p+1} e^{1k} \{(X_{ij} - s)/h\}^{k-1} \right] K_h(X_{ij} - s) \sigma^{jj}}{\sum_{k=1}^m \sigma^{kk} f_k(s)} Y_{ij},$$

where e^{1k} is the $(1, k)$ th element of \mathbf{E}^{-1} , σ^{jj} is the j th diagonal element of Σ^{-1} and $f_k(s)$ is the density of X_{ik} evaluated at s . It follows that the weight function $G_{N,ij}(s, \mathbf{X})$ in (5) is asymptotically equivalent to

$$G_{N,ij}(s, \mathbf{X}) \approx n^{-1} \sum_{i=1}^n \sum_{j=1}^m \frac{\left[\sum_{k=1}^{p+1} e^{1k} \{(X_{ij} - s)/h\}^{k-1} \right] K_h(X_{ij} - s) \sigma^{jj}}{\sum_{k=1}^m \sigma^{kk} f_k(s)}.$$

For average and linear kernels ($p = 0, 1$), the asymptotic weight function of $G_{N,ij}(s, \mathbf{X})$ can be simplified as

$$G_{N,ij}(s, \mathbf{X}) \approx n^{-1} \sum_{i=1}^n \sum_{j=1}^m \frac{K_h(X_{ij} - s) \sigma^{jj}}{\sum_{k=1}^m \sigma^{kk} f_k(s)}.$$

It follows that the kernel estimator $\hat{g}_h(s)$ is local asymptotically. The results of Ruckstuhl, Welsh and Carroll (2000) and Lin and Carroll (2000) further show that assuming working independence gives the most efficient kernel estimator $\hat{g}_h(s)$.

For splines, it is unfortunately much more difficult to obtain the asymptotic form of the equivalent kernel (7). However, we can still obtain valuable insights from asymptotic results. Suppose first that we use working independence. In this case, to prove that smoothing splines are local asymptotically, we only need to show that a spline is asymptotically equivalent to a kernel estimator. The proof is given in Appendix A and is similar to Silverman (1984).

Next suppose that the true covariance Σ is used in spline estimation. We study the asymptotic locality of smoothing splines under the same asymptotic framework as Silverman (1984), i.e., $\lambda \rightarrow 0$ and $\lambda^{-1/4}a(n) \rightarrow 0$ for some $a(\cdot)$ as $n \rightarrow \infty$. Without obtaining an explicit expression for the asymptotic form of the equivalent kernel (7), we show in Appendix B that

(1) The smoothing spline estimator $\hat{g}(s)$ is generally not local asymptotically in the sense that for any given bandwidth h and $|X_{ij} - s| > h$, there exists a constant c such that $\lim_{n \rightarrow \infty} \Pr\{|G_{Nij}(s, \mathbf{X})| > c\} > 0$.

(2) As the within-cluster correlation of the Y_{ij} increases, $\hat{g}(s)$ becomes more non-local in the sense that the value of the constant c becomes larger. The worst case is when Σ is nearly singular as c can be arbitrarily large.

(3) As the within cluster correlation of the X_{ij} decreases, $\hat{g}(s)$ becomes more non-local in the sense that asymptotically the probability $\Pr\{|G_{Nij}(s, \mathbf{X})| > c\}$ (given the constant c) becomes larger. The worst case is when the X_{ij} are independent.

5 MSE Comparisons

As shown in Sections 2-4, splines and kernels have different local behavior. This is likely to result in differences in the asymptotic efficiency of these methods. We therefore investigate in this section the mean squared error properties of these methods. Specifically, we show in Section 5.1 that asymptotically the variances of smoothing splines and P-splines are minimized when the working covariance is equal to the true covariance. This suggests that accounting for correlation gives the most efficient spline estimators asymptotically, exactly the opposite of the kernel methods we are studying. We then conduct in Sections 5.2 and 5.3 numerical MSE comparisons using the exact finite sample MSE formulas given in Section 2.

5.1 Optimal Kernel, P-Spline and Smoothing Spline Estimators

We first consider the optimal kernel estimator in the form of equation (4). Lin and Carroll (2000) showed that the asymptotic bias does not depend on the working covariance matrix and that the asymptotic variance is minimized under working independence. That is, the

asymptotically efficient kernel estimator is obtained by ignoring the within-cluster correlation and assuming working independence.

We next study the asymptotically optimal choice of the working covariance matrix for P-spline and smoothing spline estimators. Let $\mathbf{W}_{work}^{-1} = \text{diag}(\boldsymbol{\Sigma}_{work})$ be any working covariance matrix. Let $\hat{\boldsymbol{\alpha}}_\lambda = \widehat{\boldsymbol{\Theta}}$ for the P-spline, with $\widehat{\boldsymbol{\Theta}}$ defined in Section 2.3 and $\hat{\boldsymbol{\alpha}}_\lambda = \hat{\mathbf{g}}_\lambda$ for the smoothing spline, with $\hat{\mathbf{g}}_\lambda$ defined in Section 2.4. Using the results in Sections 2.3 and 2.4, both estimators take the same form

$$\hat{\boldsymbol{\alpha}}_\lambda = (\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z} + \lambda \mathbf{S})^{-1} \mathbf{Z}^T \mathbf{W}_{work} \mathbf{Y},$$

and their covariances also take the same form

$$\text{cov}(\hat{\boldsymbol{\alpha}}_\lambda) = (\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z} + \lambda \mathbf{S})^{-1} \mathbf{Z}^T \mathbf{W}_{work} \mathbf{V} \mathbf{W}_{work} \mathbf{Z} (\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z} + \lambda \mathbf{S})^{-1}, \quad (8)$$

where \mathbf{Z} equals $\mathbf{Q} = (\mathbf{Q}_1^T, \dots, \mathbf{Q}_n^T)^T$ for P-splines and \mathbf{U} for smoothing splines, and the penalty matrix \mathbf{S} equals \mathbf{P} for P-splines and $\boldsymbol{\Psi}$ for smoothing splines.

The difficulty of obtaining explicit expressions for the asymptotic form of the equivalent kernel alluded to in Section 4, make it difficult to study the asymptotic bias of spline estimators assuming a general working covariance matrix. We study in Appendix A the asymptotic bias and variance of spline estimators assuming working independence. This result is a refinement of our findings in Section 4 of the asymptotic equivalence of working independence spline and kernel estimators in the spirit of Silverman (1984). Specifically, we follow the proof of Nychka (1995) to show that the asymptotic bias and variance of a spline estimator under working independence are the same as those of a kernel estimator. While it is quite plausible that the asymptotic bias of splines does not depend on the working covariance matrix, we cannot provide a formal proof. However, our working independence spline and kernel results provide some evidence that it is probably reasonable to ignore the bias when comparing MSEs.

We can show in Appendix C that when the smoothing parameter λ converges to 0 as a function of n at some specific rate, the covariance matrix $\text{cov}(\hat{\boldsymbol{\alpha}}_\lambda)$ is minimized when the working covariance matrix \mathbf{W}_{work}^{-1} is equal to the true covariance \mathbf{V} , and the minimized $\text{cov}(\hat{\boldsymbol{\alpha}}_\lambda)$ is

$$\text{cov}(\hat{\boldsymbol{\alpha}}_\lambda) = (\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} + \lambda \mathbf{S})^{-1} \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} (\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} + \lambda \mathbf{S})^{-1}.$$

This result suggests that in contrast to kernels, the asymptotically efficient smoothing spline and P-spline estimators are obtained by accounting for the within-cluster correlation using the true covariance. This result can be partly explained by the asymptotic non-local behavior of splines shown in Sections 2-4.

We finally compare spline estimators and kernel estimators. Modulo the asymptotic bias, for longitudinal data, the most efficient spline estimator is obtained by using the true covariance and the most efficient kernel estimator is obtained by using working independence. Since under working independence a spline is asymptotically equivalent to a kernel estimator with a specific higher-order kernel (see Section 4 and above) and choices of different kernel functions often have small effects (Eubank, 1999, p177), the results in this section suggest that for the longitudinal marginal model, spline methods may be preferable to the kernel methods studied in this paper.

A technically fully rigorous theoretical comparison of the MSEs between splines and kernels seems difficult, e.g., in view of the difficulty in the bias calculations of splines in general cases. As a remedy for lacking a fully rigorous theoretical investigation on MSEs, we compare numerically the exact finite sample MSEs of splines and kernels using the analytic MSE expression (3) in the next two sections and simulated MSEs in Section 6. These empirical results incorporate the biases when comparing the MSEs.

5.2 MSE Comparisons of P–Splines and Kernels: Numerical Results

In this section, we compare numerically the exact finite sample mean squared errors (3) of P–splines and kernel methods. From the results of Sections 2-4, we expect the different local behavior of P–splines and kernels to lead to different behavior in the function estimates. In particular, we expect that the kernel result which says that using working independence is preferable to estimating the covariance structure will not hold for P–splines. Furthermore, based on the results in Section 5.1, we expect the P–spline estimates assuming the true covariance to be more efficient than those assuming working independence. The numerical results of this section show that at least in the cases we consider, this logic is correct.

We again consider the situation where the predictors X_{ij} for $i = 1, \dots, n = 50$ and $j = 1, \dots, m = 3$ are independent uniform random variables on the interval $[-2, 2]$. The theory in Section 4 suggests that the independent X case is the case in which splines and kernels differ most in their local behavior. We considered the function $g(x) = \sin(2x)$, which is not well-modeled by a simple polynomial, and the three correlation structures: exchangeable with common correlation ρ , autoregressive with correlation ρ , and unstructured with correlation between measurements 1 and 2 and between measurements 2 and 3 equal to 0.80, and correlation between units 1 and 3 equal to ρ .

Using equation (3) in Section 2.3, we computed the mean squared error of the predictions of the P-spline of order $p = 2$ with 35 equally spaced knots for a fixed smoothing parameter λ . Using the notation in Section 2.3, the first term in the mean squared error formula (3) can be written as $\mathbf{q}^T(s) (\mathbf{H}_n + \lambda \mathbf{P})^{-1} n^{-1} \mathbf{G}_n (\mathbf{H}_n + \lambda \mathbf{P})^{-1}$, where $\mathbf{G}_n = n^{-1} \sum_{i=1}^n \mathbf{Q}_i^T \boldsymbol{\Sigma}_{work}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_{work}^{-1} \mathbf{Q}_i$. We replaced \mathbf{H}_n and \mathbf{G}_n by their asymptotic limits, then averaged these mean squared errors over the interval $[-1.8, 1.8]$, and selected λ to minimize the mean squared errors.

The results are given in Table 1, where we list the efficiency of using working independence relative to the true covariance based on our calculations for P-splines. We also list the efficiency for a purely parametric model with no smoothing, i.e., $\lambda = 0$, and the efficiency for kernels as described above. The results are striking and consistent with the theory. The kernel method of course always favors working independence, while the P-spline results suggest that using the true covariance matrix is more efficient, and sometimes much more efficient, a finding that is illustrated in the simulations of Section 6.

5.3 MSE Comparison of Smoothing Splines: Numerical Results

In this section, we compare numerically the exact finite sample MSE efficiencies of smoothing splines using the true covariance relative to using working independence. Specifically, we computed for various sample sizes the exact mean squared errors of the smoothing spline estimates using the bias and the variance formulas in Section 2.4 for two functions: $g(x) = \exp(x)$ and $g(x) = \sin(2x)$, for different values of the correlation parameter ρ under the exchangeable

and autoregressive correlation structures, and a fixed smoothing parameter. We generated 1000 sets of X 's as independent random variables on the interval $[-2,2]$, calculated the exact MSEs for each set of X 's, and then averaged these MSEs. The results are displayed in Table 2. Just as for P-splines, for smoothing splines, it is more efficient to estimate the covariance matrix than to use working independence.

6 Simulations

In this section, we present the results of simulations for P-splines, smoothing splines and kernels. We first investigate the efficiency of nonparametric estimates using the true covariance relative to working independence for each of the three methods, and then investigate the efficiency of P-splines and smoothing splines relative to kernels. The P-spline work includes estimation of the smoothing parameter λ via CV and GCV arguments. The purpose of this section is to show that the theoretical work presented in Sections 3–5 is indeed a guide to what will happen with actual data.

The situation we consider is that the predictors X_{ij} for $i = 1, \dots, n = 50, 100$ and $j = 1, \dots, m = 3$ are independent uniform random variables on the interval $[-2, 2]$. The common variance of the Y_{ij} was $\sigma_\epsilon^2 = 1$. We considered three correlation structures: exchangeable with common correlation $\rho = 0.6$, autoregressive with correlation $\rho = 0.6$, and unstructured where the correlation between measurements 1 and 2 and between measurements 2 and 3 is 0.80, and the correlation between units 1 and 3 is $\rho = 0.5$.

Let $z = (x + 2)/4$. The functions chosen were model = 1 if $g(x) = \sin(x)$; model = 2 if $g(x) = \exp(x)$; model = 3 if $g(x) = \sin(2x)$; model = 4 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1 + 2^{-3/5})/(z + 2^{-3/5})\}$; model = 5 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1 + 2^{-7/5})/(z + 2^{-7/5})\}$; and model = 6 if $g(z) = \sin(8z - 4) + 2\exp\{-256(z - .5)^2\}$. Models 1–2 are almost perfectly fit by a quadratic polynomial, while models 3–6 are poorly fit by a quadratic polynomial but are fit well by splines and kernels.

We used the Epanechnikov kernel function when calculating the kernel estimates. This kernel function minimizes the MSE and seems to be a best choice to dampen the effects of the choices of kernel functions when comparing with the spline estimates. Additional

simulations were run using different kernel functions, e.g., a uniform kernel, and the results were similar although the kernel estimates performed slightly worse in MSEs than those using the Epanechnikov kernel function. The P-spline specifics were the same as in Section 5, except that the knots were chosen to be the sample quantiles of the observed X 's. We assumed in our simulations that the true covariance was known. For each configuration, we generated 200 simulated data sets. For each simulated data set, we estimated $g(\cdot)$ using a P-spline, a smoothing spline and a kernel estimator assuming the true covariance or working independence. We estimated the smoothing parameter λ for the P-spline and the smoothing spline and the bandwidth parameter h for the kernel by minimizing the MSEs $(\hat{\mathbf{g}} - \mathbf{g})^T(\hat{\mathbf{g}} - \mathbf{g})$ on a grid of values for λ and h .

Table 3 compares the mean squared error efficiency of using the true covariance as opposed to working independence for each of the three methods: kernel, P-splines, smoothing splines. For both P-splines and smoothing splines, the efficiency gain of using the true covariance is substantial in all cases, especially in the nearly singular covariance case, where the efficiency gain doubles. However, for kernels, there is little efficiency gain from using the true covariance relative to working independence. These results support the theory: in contrast to kernels, it is advantageous to account for the within-cluster correlation for P-splines and smoothing splines.

Table 4 compares the mean squared error efficiency of P-splines and smoothing splines with kernels when using the true covariance and working independence. When the true covariance is used, both P-splines and smoothing splines are more efficient than kernels, especially in the nearly singular covariance case, where the efficiency gain doubles. When working independence is assumed, the performance of splines and kernels are about the same. These results suggest that the spline methods are more efficient than kernels in longitudinal data when we account for the within-cluster correlation.

Since our purpose was to compare the performance of splines and kernels, for simplicity and for the sake of consistency when comparing the three methods, we assumed the covariance matrix was known and chose the smoothing parameter and the bandwidth parameter using the same criteria by minimizing the MSEs. We also conducted simulations using P-splines to show that the results persist when we estimated the covariance matrix Σ and

estimated the smoothing parameter using modified GCV.

Specifically, for P-splines, we estimated Σ by the following simple device: (a) run an ordinary unweighted least squares regression of the Y_{ij} on the $\mathbf{q}(X_{ij})$ to form an unsmoothed estimate of Θ ; (b) form the residuals r_{ij} ; (c) arrange them into the vector \mathbf{r}_i ; and (d) compute the covariance matrix of the residual vectors. Ordinary least squares can be replaced by a standard P-spline with a small degree of smoothing, although in our examples this was not necessary. If $\hat{\Sigma}$ is the estimate so formed, the independence covariance matrix is simply $\hat{\Sigma}_{indp} = \text{diag}(\hat{\Sigma})$.

When working independence was assumed, we used the standard GCV (Green and Silverman, 1994) to estimate the smoothing parameter. In the presence of within-cluster correlation, standard GCV could incur a bias, which can be overcome by deleting one subject at a time and then using CV. Some calculations show that CV can be written as

$$\text{CV}(\lambda) = n^{-1} \sum_{i=1}^n (\mathbf{Y}_i - \hat{\mathbf{g}}_i)^T [(\mathbf{I} - \mathbf{A}_{ii})^{-1}]^T (\mathbf{I} - \mathbf{A}_{ii})^{-1} (\mathbf{Y}_i - \hat{\mathbf{g}}_i),$$

where writing $\hat{\mathbf{g}} = \mathbf{A}\mathbf{Y}$ and partitioning \mathbf{A} into n^2 block matrices of size $m \times m$, \mathbf{A}_{ii} is the i th diagonal block matrix corresponding to \mathbf{Y}_i .

When the true covariance matrix was used, i.e., $\Sigma_{\text{work}} = \Sigma$, a modified GCV was used. Define $\mathbf{Y}_i^* = \Sigma^{-1/2}\mathbf{Y}_i$ and $\mathbf{Q}_i^* = \Sigma^{-1/2}\mathbf{Q}_i$. Since we are pretending that $E(\mathbf{Y}_i|\mathbf{Q}_i) = \mathbf{Q}_i\Theta$, we are similarly pretending that $E(\mathbf{Y}_i^*|\mathbf{Q}_i^*) = \mathbf{Q}_i^*\Theta$. Further, the covariance matrix of \mathbf{Y}_i^* is the identity matrix, i.e., we have converted the problem to one of independence, with $N = nm$ observations. It is easily seen from standard calculations that the GCV estimate of λ minimizes

$$\text{GCV}(\lambda) = \frac{\sum_{i=1}^n \sum_{j=1}^m \{Y_{ij}^* - \mathbf{Q}_{ij}^* \widehat{\Theta}(\lambda)\}^2}{\{1 - (nm)^{-1}T_n(\lambda)\}^2} = \frac{\sum_{i=1}^n \{\mathbf{Y}_i - \mathbf{Q}_i \widehat{\Theta}(\lambda)\}^T \Sigma^{-1} \{\mathbf{Y}_i - \mathbf{Q}_i \widehat{\Theta}(\lambda)\}}{\{1 - (nm)^{-1}T_n(\lambda)\}^2}, \quad (9)$$

where $\mathbf{D}(\lambda) = \sum_{i=1}^n \mathbf{Q}_i^{*T} \mathbf{Q}_i^* = \sum_{i=1}^n \mathbf{Q}_i^T \Sigma^{-1} \mathbf{Q}_i$, $\mathbf{B}(\lambda) = (\mathbf{D}(\lambda) + n\lambda\mathbf{P})^{-1}$, $T_n(\lambda) = \text{trace}\{\mathbf{B}(\lambda)\mathbf{D}(\lambda)\}$ and $\widehat{\Theta}(\lambda) = \mathbf{B}(\lambda) \sum_{i=1}^n \mathbf{Q}_i^{*T} \mathbf{Y}_i^* = \mathbf{B}(\lambda) \sum_{i=1}^n \mathbf{Q}_i^T \Sigma^{-1} \mathbf{Y}_i$. It can be shown that this modified GCV approximates the weighted MSE. We estimated the smoothing parameter λ by minimizing GCV on a grid of values of λ . The results are displayed in Table 5, where we see that estimating the covariance matrix leads to a more efficient estimate of

the function $g(x)$ than using working independence. These results are consistent with those in Table 3 when Σ was assumed known and λ was estimated by minimizing the MSE.

7 Discussion

We have discussed a longitudinal marginal nonparametric regression problem, and in particular we have contrasted the strategies of modeling the covariance structure and using working independence (i.e., ignoring the correlation structure entirely) when using splines and kernel methods. Once again, we emphasize that our results pertain to the kernel methods previously proposed in the literature, and not to the totality of all possible kernel methods that could be proposed. In this context, it is well-known that parametric model fitting using the covariance structure leads to more efficient model fitting than using working independence. In the nonparametric problem, kernel methods have the unusual property that using working independence usually leads to a more efficient estimate.

Our purpose in this paper was to try to understand whether the kernel results are relevant to smoothing and regression P-spline methods. This is not an unusual question, because in the independent data case smoothing splines are asymptotically equivalent to a particular form of kernel regression.

When X_{ij} is a subject-level covariate, i.e., $X_{ij} = X_i$, simple calculations show that, when calculating the kernel and spline estimators of $g(\cdot)$, we can create a univariate response for the i th subject by pooling the responses Y_{ij} as $\mathbf{1}^T \Sigma^{-1} \mathbf{Y}_i$. Splines are then local and asymptotically equivalent to kernel regression with the same higher-order kernel defined by Silverman (1984), and the efficient spline and kernel estimators both require the within-cluster correlation to be taken into account.

When the covariate X is a time-varying covariate, our results suggest that the kernel results are *irrelevant* for smoothing and regression P-splines. Asymptotic theoretical and exact calculations (Section 5) and simulations (Section 6) suggest that splines operate more like parametric estimators, and using working independence leads to less efficient model fitting. This difference in behavior may be partially explained by the results on equivalent kernels given in Sections 2–4, where we show that with dependent data, splines are less local

than kernel methods.

Given the surprising nature of our results, it is perhaps useful to emphasize what they do not say. First, the results apply to our model (1)–(2) and not necessarily to other models for longitudinal data. Our model has been chosen for simplicity of exploration and exposition to represent a situation which is common in many longitudinal studies. In particular, we would like to emphasize that our results apply to the case when the number of observations per subject m is finite while the number of subjects n is relatively large. Different models or even the same model with different asymptotics (e.g. with both m and $n \rightarrow \infty$ or $\Sigma \rightarrow \sigma^2 \mathbf{I}$) which produce different results do not of course contradict our results. Fan and Zhang (2000) considered the case when both m and $n \rightarrow \infty$. We do not expect our results to apply to this case, since the early results in nonparametric regression in time series suggest that one needs to account for correlation in kernel regression (Hart, 1991). Second, in all our references to kernel estimators, we mean previously proposed kernel estimators. This does not mean that no “kernel estimator” can improve on the standard kernel estimators: Ruckstuhl, Welsh and Carroll (2000) constructed a two-step kernel estimator which has a smaller asymptotic variance than standard kernel estimators when the dependence structure is taken into account. Finally, we have treated the question of whether or not it is worth taking the trouble to model the correlation structure in (1)–(2) when using P-splines or smoothing splines. We have not discussed the difficulties of identifying and estimating the correlation structure. Moreover, we have not explored the consequences of misspecification of the covariance except that it is obvious that this will increase the mean squared error of the estimator.

Nonetheless, one major consequence of our work is the suggestion that, for time-varying covariates, spline methods which account for the correlation in the marginal model (1)–(2) for longitudinal data will yield more efficient function estimates than previously proposed kernel methods, the known forms of which are optimized at working independence. This is a powerful argument for the use of splines in this context, and an argument against the use of such kernel methods.

REFERENCES

- Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford: Oxford University Press.
- Eilers, P. H. C. and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties (with discussion)," *Statistical Science*, 11, 89–121.
- Fan, J. and Zhang, J.T. (2000), "Two-step Estimation of Functional Linear Models With Applications to Longitudinal Data," *Journal of the Royal Statistical Society*, Ser. B, 62, 303-322.
- Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*, Marcel Dekker, New York.
- Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.
- Hart, J. D. (1991), "Kernel Regression Estimation with Time Series Errors," *Journal of the Royal Statistical Society, Series B*, 53, 173-187.
- Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, Y. (1998), "Nonparametric Smoothing Estimates of Time-varying Coefficient Models with Longitudinal Data," *Biometrika*, 85, 809–822.
- Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis using Generalized Linear Models," *Biometrika*, 73, 13-22.
- Lin, X. and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error," *Journal of the American Statistical Association*, 95, 520-534.
- Nychka, D. (1995), "Splines as Local Smoothers," *Annals of Statistics*, 22, 1175-1197.
- Pepe, M. S. and Couper, D. (1997), "Modeling Partly Conditional Means with Longitudinal Data", *Journal of the American Statistical Association*, 92, 991- 998.
- Ruckstuhl, A., Welsh, A. H. and Carroll, R. J. (2000), "Nonparametric Function Estimation of the Relationship Between Two Repeatedly Measured Variables," *Statistica Sinica*, 10, 51–71.
- Ruppert, D. and Carroll, R. J. (2000), "Spatially-adaptive Penalties for Spline Smoothing," *Australian and New Zealand Journal of Statistics*, 2, 205–224.
- Severini, T. A. and Staniswalis, J. G. (1994), "Quasilikelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.
- Silverman, B. (1984). "Spline Smoothing: The Equivalent Variable Kernel Method," *Annals of Statistics*, 12, 898-916.
- Wahba, G. (1990), *Spline Models for Observational Data*, SIAM Press, Philadelphia.
- Wu, C. O., Chiang, C. T. and Hoover, D. R. (1998), "Asymptotic Confidence Regions for Kernel Smoothing of a Varying Coefficient Model with Longitudinal Data," *Journal of the American Statistical Association*, 93, 1388–1402.
- Zeger, S. L. and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689-699.

Appendix A Equivalence of Spline and Kernel Estimators Under Working Independence

We prove in this section that, under working independence, spline and kernel estimators are asymptotically equivalent. We assume the same setup as Silverman (1984) except that we assume that the \mathbf{X}_i rather than the X_{ij} are independent and identically distributed. Note that the X_{ij} and Y_{ij} may be correlated within each subject.

Under working independence, the smoothing spline estimator $\hat{g}(\cdot)$ which minimizes

$$n^{-1} \sum_{i=1}^n \sum_{j=1}^m \{Y_{ij} - g(X_{ij})\}^2 + \lambda \int \{g''(x)\}^2 dx. \quad (\text{A.1})$$

can be written at any s as

$$\hat{g}(s) = \sum_{i=1}^n \sum_{j=1}^m G(s, X_{ij}) Y_{ij},$$

where $G(s, X_i)$ is a weight function depending on the design points (X_{11}, \dots, X_{nm}) and the smoothing parameter λ . Following Silverman (1984), define a functional $A_x(g)$ by

$$A_x(g) = \frac{1}{2} \int g^2(u) dF_N(u) - g(x) + \frac{1}{2} \lambda \int \{g''(u)\}^2 du, \quad (\text{A.2})$$

where $F_N(\cdot)$ is the empirical distribution function of (X_{11}, \dots, X_{nm}) , i.e.,

$$F_N(x) = N^{-1} \sum_{i=1}^n \sum_{j=1}^m I(X_{ij} \leq x),$$

where $I(\cdot)$ is the indicator function. Let $g_x(s)$ be the minimizer of (A.2). Silverman's (1984, p901) argument that $G(s, x) = g_x(s)$ does not require the Y_{ij} to be independent and identically distributed so also applies in our context. Rewriting $F_N(x)$ as

$$F_N(x) = m^{-1} \{F_{n1}(x) + \dots + F_{nm}(x)\},$$

where $F_{nj}(x) = n^{-1} \sum_{i=1}^n I(X_{ij} \leq x)$ ($j = 1, \dots, m$) is the empirical distribution function of X_{ij} ($i = 1, \dots, n$), we see that $F_{nj}(x)$ converges to $F_j(x)$ uniformly in x . Since m is finite, it follows that $F_N(x)$ converges to $F(x) = m^{-1} \sum_{j=1}^m F_j(x)$ uniformly in x . Hence Silverman's condition (2.5) is satisfied and we can complete the proof as in Silverman (1984).

We next show that the asymptotic bias and variance of a spline estimator under working independence take the same forms as a kernel estimator. Our proof is similar to Nyshka

(1995). The first key requirement in Nyshka’s proof is the equality $G(s, x) = g_x(s)$. Our above argument shows that this equality holds for a spline estimator under working independence. Hence the Cox series representation of the spline weight function in equation (3.2) of Nyshka (1995) still holds. It follows that the spline weight function can be approximated by the same Green’s function, which plays a fundamental role in deriving the asymptotic bias and variance. The second key assumption in Nyshka (1995) is that $F_N(x)$ converges to $F(x)$ uniformly in x , which was proved in the previous paragraph. The rest proof is identical to Nyshka (1995). Assuming the Y_{ij} have the same marginal variance σ^2 , the bias and variance of the working independence spline estimator take the same form as those given in Theorem 2.2 of Nyshka with $f(t) = \partial F(t)/\partial t$ and n replaced by nm .

Appendix B Asymptotic Locality of Smoothing Splines

We first review the results for the independent data case. The smoothing spline estimator $\hat{\mathbf{g}}_\lambda$ can be written as $\hat{\mathbf{g}}_\lambda = \mathbf{G}\mathbf{Y}$, where $\mathbf{G} = (\mathbf{I} + \lambda\mathbf{\Psi})^{-1}$. Silverman (1984) showed that $\hat{g}(s)$ is asymptotically equivalent to a kernel estimator,

$$\hat{g}(s) = \sum_{i=1}^N \frac{K_h(X_i - s)}{f(s)} Y_i,$$

where the kernel function $K_h(\cdot)$ is defined in Silverman (1984). This implies that for large n and small λ , for any given small $\epsilon > 0$, there exists a constant M_1 such that the elements G_{jk} of the hat matrix \mathbf{G} satisfy

$$|G_{jk}| < \epsilon \text{ for } |j - k| > M_1. \tag{A.3}$$

In other words, the hat matrix \mathbf{G} is asymptotically “equivalent” to a banded matrix with bandwidth M_1 . Further, using the definition of $\mathbf{\Psi}$ given in Section 2.5, some calculations show that \mathbf{R}^{-1} defined in Section 2.5 is asymptotically “equivalent” to (in the sense of (A.3)) a banded matrix with some bandwidth M_2 . Since the matrix \mathbf{C} is banded, it follows that $\mathbf{\Psi}$ is asymptotically “equivalent” to a banded matrix with bandwidth M_2 .

Now we study the hat matrix in the longitudinal data case for large n and small λ . Suppose that $(\mathbf{X}_i, \mathbf{Y}_i)$ are defined as in Section 2.1, i.e., the data are entered by subjects

so that $\mathbf{V} = \text{diag}(\boldsymbol{\Sigma})$ is a block diagonal matrix. For simplicity, assume $\boldsymbol{\Sigma}$ is a correlation matrix and there are no replicates of the X_{ij} . Suppose that the true correlation matrix $\boldsymbol{\Sigma} \neq \mathbf{I}$ is used in calculating the smoothing spline estimator $\hat{\mathbf{g}}_\lambda$.

Let \mathbf{U} be the incidence matrix defined in Section 2.4. Recall that \mathbf{U} permutes the rows of an identity matrix and the permutation distribution depends on the distribution of the X_{ij} , and \mathbf{U} satisfies $\mathbf{U}^T \mathbf{U} = \mathbf{I}$. Using the results in Section 2.4, we have $\hat{\mathbf{g}}_\lambda = \mathbf{G} \mathbf{Y}_o$, where the hat matrix \mathbf{G} is

$$\mathbf{G} = (\mathbf{U}^T \mathbf{V}^{-1} \mathbf{U} + \lambda \boldsymbol{\Psi})^{-1} \mathbf{U}^T \mathbf{V}^{-1} \mathbf{U} = \mathbf{I} - (\mathbf{U}^T \mathbf{V}^{-1} \mathbf{U} + \lambda \boldsymbol{\Psi})^{-1} \lambda \boldsymbol{\Psi}. \quad (\text{A.4})$$

To show the smoothing spline estimator $\hat{g}(s)$ is not local at any given s asymptotically, it is sufficient to show that $\hat{\mathbf{g}}_\lambda$ is not local asymptotically. This implies that we only need to show that the hat matrix \mathbf{G} in (A.4) is not “equivalent” to a banded matrix asymptotically. Let $\tilde{\mathbf{V}} = \mathbf{U}^T \mathbf{V}^{-1} \mathbf{U}$. Then $\mathbf{G} = \mathbf{I} - (\tilde{\mathbf{V}} + \lambda \boldsymbol{\Psi})^{-1} \lambda \boldsymbol{\Psi}$. Since $\boldsymbol{\Psi}$ is “equivalent” to a banded matrix with bandwidth M_2 asymptotically, we only need to show $\tilde{\mathbf{V}}$ is asymptotically not “equivalent” to a banded matrix and its structure can be arbitrary, i.e., for any given fixed bandwidth M_3 , there exists a constant T such that asymptotically

$$\Pr(|\tilde{v}_{jk}| > T : |j - k| > M_3) > 0,$$

where \tilde{v}_{jk} is the (j, k) th element of $\tilde{\mathbf{V}}$. In other words, the elements of $\tilde{\mathbf{V}}$ outside any fixed bandwidth M_3 have positive probability of being bounded away from 0.

To proceed, first note that $\mathbf{V}^{-1} = \text{diag}(\boldsymbol{\Sigma}^{-1})$ is a block diagonal matrix with block size m , and the matrix $\tilde{\mathbf{V}}$ permutes the rows and the columns of \mathbf{V}^{-1} according to the distribution of the \mathbf{X}_{ij} . Hence $\tilde{\mathbf{V}}$ is generally not a block diagonal matrix. For example, in the case that the X_{ij} are independent and identically distributed within each subject, one would permute the rows and the columns of \mathbf{V}^{-1} with equal probabilities. The non-zero off-diagonal elements of $\tilde{\mathbf{V}}$, i.e., the elements $\sigma^{jk} (j \neq k)$ of $\boldsymbol{\Sigma}^{-1}$, can hence be permuted to any place of $\tilde{\mathbf{V}}$ with equal probabilities and the structure of $\tilde{\mathbf{V}}$ can be arbitrary. In other words, for any given bandwidth $M_3 > \min(M_2, m)$, asymptotically,

$$\Pr(|\tilde{v}_{jk}| > T : |j - k| > M_3) > 0,$$

where $T = \min_{j \neq k} \{|\sigma^{jk}| : |\sigma^{jk}| > 0\}$. This suggests that one would expect a large number of values of $\widetilde{\mathbf{V}}$ bounded away from 0 (i.e., $> T$) in arbitrary locations outside the bandwidth M_3 . It follows that the hat matrix \mathbf{G} is not “equivalent” to a banded matrix asymptotically. Since there is a one-to-one-correspondence between $\widetilde{\mathbf{V}}$ and \mathbf{G} , there exists a constant c such that asymptotically, the elements of \mathbf{G} satisfy

$$\Pr(|G_{jk}| > c : |j - k| > M_3) > 0.$$

Hence the smoothing spline estimator $\widehat{\mathbf{g}}_\lambda$ is not local asymptotically. As shown below, the non-locality of $\widehat{\mathbf{g}}_\lambda$ increases with the within-subject correlation among the Y_{ij} and decreases with the within-subject correlation among the X_{ij} .

Since the value of T increases with the within-subject correlation among the Y_{ij} , i.e.,

$$\text{corr}(Y_{ij}, Y_{ij'}) \uparrow \Rightarrow T \uparrow,$$

the non-zero elements of $|\tilde{v}_{jk}|$ ($|j - k| > M_3$) become larger as the correlation among the Y_{ij} increases, i.e.,

$$\text{corr}(Y_{ij}, Y_{ij'}) \uparrow \Rightarrow \{|\tilde{v}_{jk}| \uparrow : |\tilde{v}_{jk}| > 0, |j - k| > M_3\}.$$

Hence more weight is given to observations outside the bandwidth M_3 when calculating each component of $\widehat{\mathbf{g}}_\lambda$, i.e., the smoothing spline estimator $\widehat{\mathbf{g}}_\lambda$ is more non-local. The worst case is when Σ is nearly singular, because elements of Σ^{-1} can be arbitrarily large and hence the nonzero elements of \tilde{v}_{jk} ($|j - k| > M_3$) outside the bandwidth can be arbitrarily large.

The locality of $\widehat{\mathbf{g}}_\lambda$ is also affected by the within-cluster correlation among the X_{ij} . This is because one can easily show that

$$\text{corr}(X_{ij}, X_{ij'}) \downarrow \Rightarrow \Pr(|\tilde{v}_{jk}| > T : |j - k| > M_3) \uparrow.$$

The worst case is when the X_{ij} are independent. The core intuition is that when the X are independent, ordering brings about the permutation of X 's from different subjects and so breaks the block diagonal structure, but when they are dependent, with high probability, ordering involves only permutation within subjects which preserves the block diagonal structure.

Appendix C Asymptotically Optimal Smoothing Spline and P-spline Estimators

Let $\hat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\Theta}}$ for the P-spline with $\widehat{\boldsymbol{\Theta}}$ defined in Section 2.3 and $\hat{\boldsymbol{\alpha}} = \hat{\mathbf{g}}_\lambda$ for the smoothing spline with $\hat{\mathbf{g}}_\lambda$ defined in Section 2.4. Denote by $\hat{\boldsymbol{\alpha}}_W$ and $\hat{\boldsymbol{\alpha}}_T$ the estimators $\hat{\boldsymbol{\alpha}}$ when the working covariance \mathbf{W}_{work}^{-1} and the true covariance \mathbf{V} are assumed respectively. We need to show that asymptotically $\text{cov}(\hat{\boldsymbol{\alpha}}_W) - \text{cov}(\hat{\boldsymbol{\alpha}}_T) > 0$ (i.e., positive definite) for any \mathbf{W}_{work} . Equivalently, we need to show $\text{cov}^{-1}(\hat{\boldsymbol{\alpha}}_T) - \text{cov}^{-1}(\hat{\boldsymbol{\alpha}}_W) > 0$. Using equation (8), we have

$$\begin{aligned} \mathbf{D}(\lambda) &= \text{cov}^{-1}(\hat{\boldsymbol{\alpha}}_T) - \text{cov}^{-1}(\hat{\boldsymbol{\alpha}}_W) \\ &= (\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} + \lambda \mathbf{S})(\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z})^{-1}(\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} + \lambda \mathbf{S}) \\ &\quad - (\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z} + \lambda \mathbf{S})(\mathbf{Z}^T \mathbf{W}_{work} \mathbf{V} \mathbf{W}_{work} \mathbf{Z})^{-1}(\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z} + \lambda \mathbf{S}) \\ &= \mathbf{B}_1 + \lambda \mathbf{B}_2 + \lambda^2 \mathbf{B}_3, \end{aligned}$$

where

$$\begin{aligned} \mathbf{B}_1 &= \mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} - (\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z})(\mathbf{Z}^T \mathbf{W}_{work} \mathbf{V} \mathbf{W}_{work} \mathbf{Z})^{-1}(\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z}) \\ \mathbf{B}_2 &= 2\mathbf{S} - \mathbf{S}(\mathbf{Z}^T \mathbf{W}_{work} \mathbf{V} \mathbf{W}_{work} \mathbf{Z})^{-1}(\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z}) \\ &\quad - (\mathbf{Z}^T \mathbf{W}_{work} \mathbf{Z})(\mathbf{Z}^T \mathbf{W}_{work} \mathbf{V} \mathbf{W}_{work} \mathbf{Z})^{-1} \mathbf{S} \\ \mathbf{B}_3 &= \mathbf{S} \left\{ (\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z})^{-1} - (\mathbf{Z}^T \mathbf{W}_{work} \mathbf{V} \mathbf{W}_{work} \mathbf{Z})^{-1} \right\} \mathbf{S}. \end{aligned}$$

One can easily show that $\mathbf{B}_1 > 0$ (positive definite), since the second term of \mathbf{B}_1 is a standard sandwich covariance matrix. The matrices \mathbf{B}_2 and \mathbf{B}_3 are symmetric but might not be positive definite.

Denote by ω_j the smallest eigenvalue of \mathbf{B}_j ($j = 1, 2, 3$). Then $\omega_1 > 0$, but ω_2 and ω_3 can be negative. For arbitrary \mathbf{x} satisfying $\mathbf{x}^T \mathbf{x} = 1$,

$$\mathbf{x}^T \mathbf{D}(\lambda) \mathbf{x} = \mathbf{x}^T (\mathbf{B}_1 + \lambda \mathbf{B}_2 + \lambda^2 \mathbf{B}_3) \mathbf{x} \geq \omega_1 + \lambda \omega_2 + \lambda^2 \omega_3.$$

One can easily show that if

$$\lambda < (\sqrt{\omega_2^2 + 4\omega_1|\omega_3|} - |\omega_2|)/(2|\omega_3|), \tag{A.5}$$

then $\omega_1 + \lambda\omega_2 + \lambda^2\omega_3 > 0$ and hence $\mathbf{x}^T \mathbf{D}(\lambda) \mathbf{x} > 0$ for any \mathbf{x} , i.e, $\mathbf{D}(\lambda) > 0$ (positive definite).

Note that the condition for λ in (A.5) assumes that λ is a function of n when it converges to 0 asymptotically, and poses some specific rate of convergence of λ as a function of n . It is of substantial interest to study in future research whether the MSE based smoothing parameter selection criteria for λ such as CV and GCV satisfy this condition.

ρ	Spline Working Efficiency	Parametric Working Efficiency	Kernel Working Efficiency	$\det(\Sigma)$
Exchangeable Case				
0.00	1.00	1.00	1.00	1.00
0.10	0.99	0.98	1.00	0.97
0.20	0.96	0.94	1.00	0.90
0.30	0.91	0.87	1.00	0.78
0.40	0.86	0.79	1.00	0.65
0.50	0.80	0.70	1.00	0.50
0.60	0.73	0.60	1.00	0.35
0.70	0.66	0.49	1.00	0.22
0.80	0.59	0.37	1.00	0.10
0.90	0.52	0.25	1.00	0.03
Autoregressive Case				
0.00	1.00	1.00	1.00	1.00
0.10	0.99	0.99	1.00	0.98
0.20	0.96	0.95	1.00	0.92
0.30	0.92	0.89	1.00	0.83
0.40	0.87	0.81	1.01	0.71
0.50	0.80	0.72	1.01	0.56
0.60	0.73	0.61	1.02	0.41
0.70	0.66	0.50	1.04	0.26
0.80	0.59	0.37	1.05	0.13
0.90	0.51	0.25	1.07	0.04
Unstructured Case				
0.80	0.59	0.37	1.00	0.10
0.75	0.60	0.39	1.01	0.12
0.70	0.60	0.39	1.03	0.13
0.65	0.59	0.38	1.05	0.13
0.60	0.57	0.36	1.06	0.13
0.55	0.54	0.33	1.09	0.12
0.50	0.51	0.30	1.11	0.11
0.45	0.46	0.26	1.14	0.09
0.40	0.39	0.21	1.15	0.07
0.35	0.29	0.14	1.16	0.05
0.30	0.12	0.05	1.18	0.01

Table 1: Exact MSE efficiencies of P-spline and kernel estimators. In the unstructured case, the correlation between measurements 1 and 2 and between measurements 2 and 3 is 0.80, and the correlation between units 1 and 3 is ρ . The “Spline Working Efficiency” is the efficiency of the working method using the penalized regression P-spline. The “Parametric Working Efficiency” is the efficiency of the working method using the penalized regression P-spline but without any penalty ($\lambda = 0$). The “Kernel Working Efficiency” is the MSE efficiency of the working method from kernel regression theory.

n	m	Function	Efficiency for Exchangeable	Efficiency for Autoregressive
20	2	1	1.94	1.94
		2	1.94	1.94
15	3	1	1.93	1.93
		2	1.93	1.93
3	15	1	1.23	1.59
		2	1.23	1.59
35	3	1	2.11	2.12
		2	2.11	2.12
7	15	1	1.36	1.76
		2	1.36	1.76
50	3	1	2.18	2.19
		2	2.18	2.19

Table 2: Average exact MSE efficiency of smoothing splines using the true covariance compared to using working independence. We used 1000 simulated sets of covariates and a fixed smoothing parameter $\lambda = 0.01$. Here function = 1 if $g(x) = \exp(x)$ and function = 2 if $g(x) = \sin(2x)$. Also, $\sigma_\epsilon^2 = 4$ and $\rho = 0.8$. throughout.

Model	Corr	Efficiency of True vs Independence (n=50)			Efficiency of True vs Independence (n=100)		
		Kernel	P-spline	S-spline	Kernel	P-spline	S-spline
1	1	1.12	1.35	1.28	1.08	1.29	1.27
	2	1.13	1.35	1.28	1.11	1.29	1.28
	3	1.08	1.84	1.71	1.11	1.73	1.72
2	1	1.07	1.35	1.32	1.03	1.26	1.25
	2	1.10	1.35	1.32	1.05	1.27	1.27
	3	1.04	1.84	1.85	0.95	1.71	1.75
3	1	1.03	1.36	1.37	1.00	1.33	1.32
	2	1.08	1.36	1.37	1.04	1.34	1.32
	3	0.96	2.03	1.97	0.91	1.96	1.90
4	1	1.07	1.33	1.31	1.02	1.33	1.30
	2	1.12	1.34	1.32	1.06	1.35	1.31
	3	1.04	1.97	1.85	0.95	1.98	1.90
5	1	1.05	1.31	1.28	1.00	1.31	1.28
	2	1.08	1.31	1.28	1.03	1.34	1.31
	3	1.03	2.02	1.92	0.95	2.07	1.97
6	1	1.04	1.44	1.40	1.01	1.42	1.38
	2	1.07	1.45	1.41	1.03	1.45	1.41
	3	0.98	2.39	2.25	0.93	2.35	2.27

Table 3: Simulation results of the MSE efficiency when using the true covariance as opposed to using working independence for each of the three methods: P-spline, Smoothing spline (S-spline), and kernel. The marginal variance = 1. Two values of the sample size n with three observations per cluster were used. The P-splines used 35 knots. Here Corr = 1 in the autoregressive case with $\rho = 0.6$; Corr = 2 in the exchangeable case with $\rho = 0.6$; Corr = 3 in the unstructured case with $\rho_{12} = \rho_{23} = 0.8$ and $\rho_{13} = 0.5$. Model = 1 if $g(x) = \sin(x)$; model = 2 if $g(x) = \exp(x)$; model = 3 if $g(x) = \sin(2x)$; model = 4 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-3/5})/(z+2^{-3/5})\}$; model = 5 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-7/5})/(z+2^{-7/5})\}$; and model = 6 if $g(z) = \sin(8z-4) + 2\exp\{-256(z-.5)^2\}$, where $z = (x+2)/4$.

		Efficiency Compared to Kernel ($n = 50$)				Efficiency Compared to Kernel ($n = 100$)			
		True Cov		Independence		True Cov		Independence	
Model	Corr	P-spline	S-spline	P-spline	S-spline	P-spline	S-spline	P-spline	S-spline
1	1	1.18	1.12	0.98	0.99	1.35	1.24	1.14	1.06
	2	1.17	1.12	0.98	0.99	1.31	1.21	1.13	1.05
	3	1.66	1.57	0.97	0.99	1.76	1.64	1.13	1.05
2	1	1.63	1.39	1.29	1.13	1.66	1.40	1.35	1.15
	2	1.57	1.35	1.28	1.12	1.62	1.39	1.34	1.15
	3	2.26	2.00	1.27	1.12	2.34	2.07	1.31	1.13
3	1	1.51	1.60	1.14	1.20	1.59	1.67	1.20	1.26
	2	1.45	1.54	1.15	1.21	1.54	1.60	1.19	1.25
	3	2.40	2.46	1.14	1.19	2.57	2.60	1.19	1.24
4	1	1.21	1.26	0.98	1.04	1.31	1.31	1.00	1.03
	2	1.18	1.23	0.98	1.04	1.28	1.28	1.00	1.03
	3	1.84	1.84	0.97	1.04	2.07	2.03	1.00	1.02
5	1	1.17	1.26	0.94	1.04	1.25	1.31	0.96	1.02
	2	1.15	1.23	0.94	1.04	1.25	1.30	0.96	1.02
	3	1.84	1.93	0.94	1.04	2.10	2.12	0.96	1.02
6	1	1.39	1.43	1.01	1.06	1.41	1.43	1.00	1.04
	2	1.38	1.40	1.02	1.06	1.41	1.43	1.00	1.04
	3	2.46	2.43	1.02	1.06	2.54	2.54	1.01	1.04

Table 4: Simulation results comparing the MSE efficiency of P-splines and smoothing splines to kernels using the true covariance and working independence. The marginal variance = 1. Two values of the sample size n with three observations per cluster were used. The P-splines used 35 knots. Here Corr = 1 in the autoregressive case with $\rho = 0.6$; Corr = 2 in the exchangeable case with $\rho = 0.6$; Corr = 3 in the unstructured case with $\rho_{12} = \rho_{23} = 0.8$ and $\rho_{13} = 0.5$. model = 1 if $g(x) = \sin(x)$; model = 2 if $g(x) = \exp(x)$; model = 3 if $g(x) = \sin(2x)$; model = 4 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-3/5})/(z+2^{-3/5})\}$; model = 5 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-7/5})/(z+2^{-7/5})\}$; and model = 6 if $g(z) = \sin(8z-4) + 2\exp\{-256(z-.5)^2\}$, where $z = (x+2)/4$.

Model	Corr	n = 50, Efficiency	n = 100, Efficiency	P-Spline no Bias Asymptotics
1	1	1.31	1.39	
	2	1.33	1.42	
	3	2.00	2.00	
2	1	1.36	1.38	
	2	1.38	1.41	
	3	2.02	1.96	
3	1	1.41	1.36	1.37
	2	1.41	1.42	1.37
	3	2.01	2.08	1.96
4	1	1.36	1.42	
	2	1.36	1.47	
	3	1.93	2.16	
5	1	1.29	1.37	
	2	1.34	1.40	
	3	1.82	2.09	
6	1	1.35	1.41	
	2	1.42	1.48	
	3	2.01	2.25	

Table 5: Simulation results of the MSE efficiency of P-splines when using an unstructured estimate of the correlation matrix as opposed to using working independence. The marginal variance = 1. Two values of the sample size n with three observations per cluster were used. The P-splines used 35 knots. Here Corr = 1 in the autoregressive case with $\rho = 0.6$; Corr = 2 in the exchangeable case with $\rho = 0.6$; Corr = 3 in the unstructured case with $\rho_{12} = \rho_{23} = 0.8$ and $\rho_{13} = 0.5$. The “P-Spline No Bias Asymptotics” is the efficiency when the correlation matrix is known and the estimand is a P-spline so is estimated without bias. model = 1 if $g(x) = \sin(x)$; model = 2 if $g(x) = \exp(x)$; model = 3 if $g(x) = \sin(2x)$; model = 4 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-3/5})/(z+2^{-3/5})\}$; model = 5 if $g(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-7/5})/(z+2^{-7/5})\}$; and model = 6 if $g(z) = \sin(8z-4) + 2\exp\{-256(z-.5)^2\}$, where $z = (x+2)/4$.

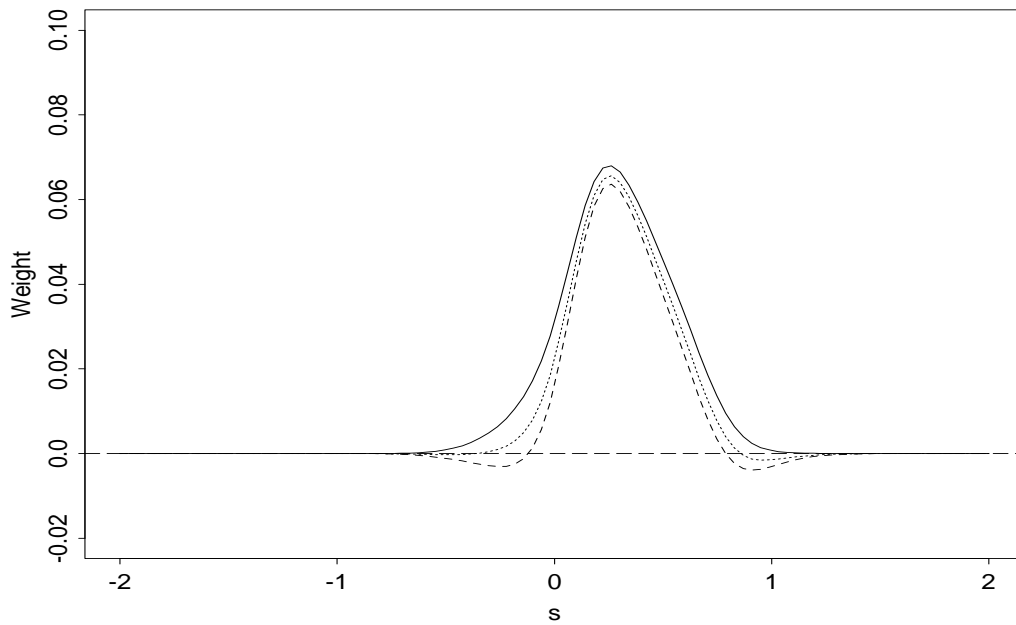


Figure 1: The equivalent kernel for the kernel method with $a = b = 0.5$ and $h = 0.25$, $n = 35$, $m = 3$, in the exchangeable case with $\rho = 0.0$ (solid line); $\rho = 0.4$ (dotted line) and $\rho = 0.8$ (dashed line).

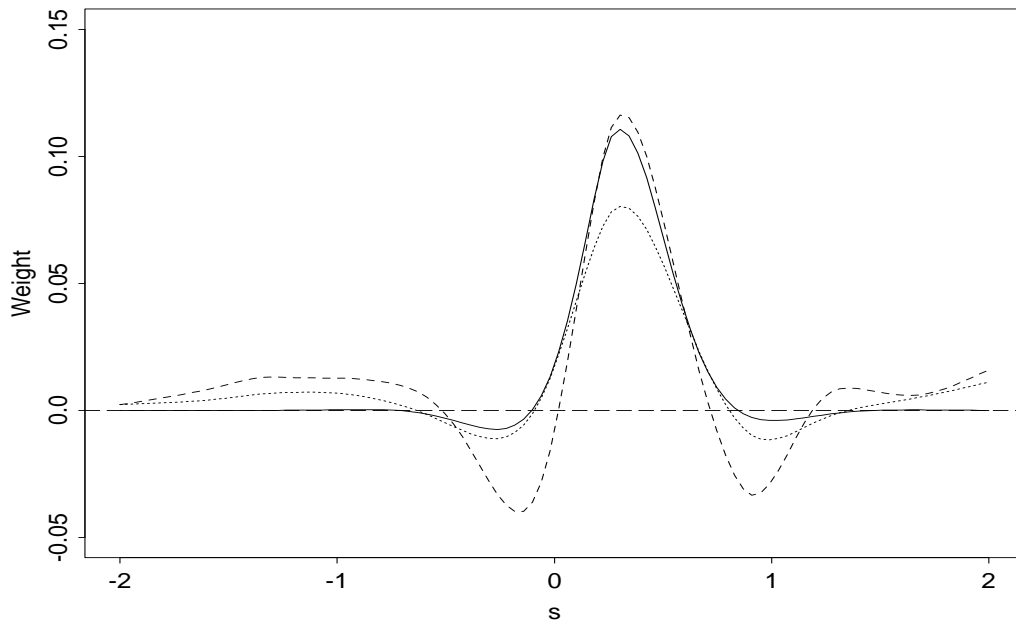


Figure 2: The equivalent kernel for the smoothing spline when $n = 35$, $m = 3$, in the exchangeable case with $\rho = 0.0$ (solid line); $\rho = 0.4$ (dotted line) and $\rho = 0.8$ (dashed line).

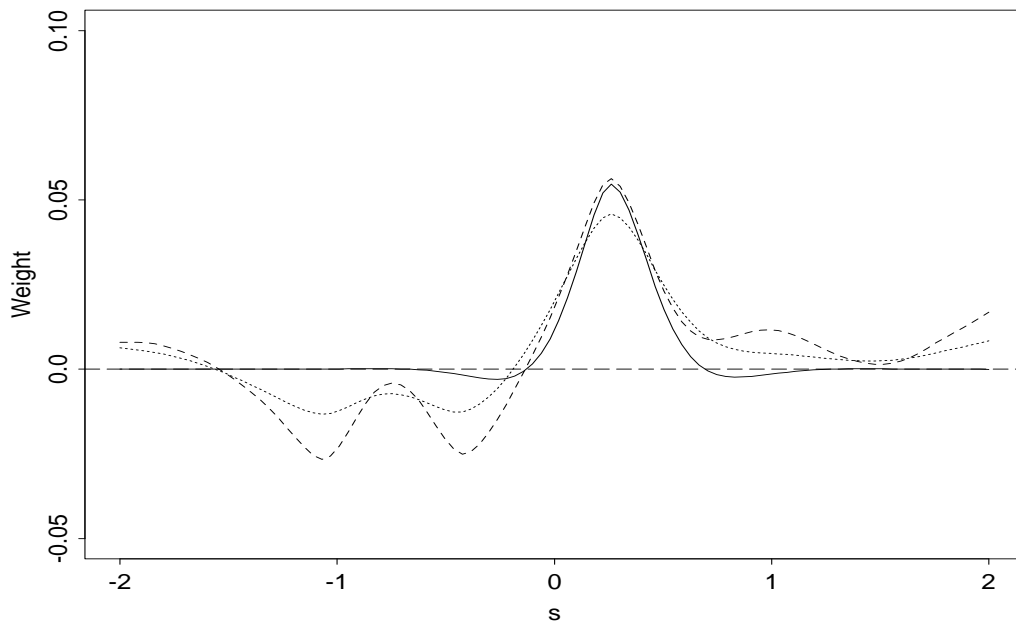


Figure 3: The equivalent kernel for the smoothing spline when $n = 50$, $m = 3$, in the exchangeable case with $\rho = 0.0$ (solid line); $\rho = 0.4$ (dotted line) and $\rho = 0.8$ (dashed line).

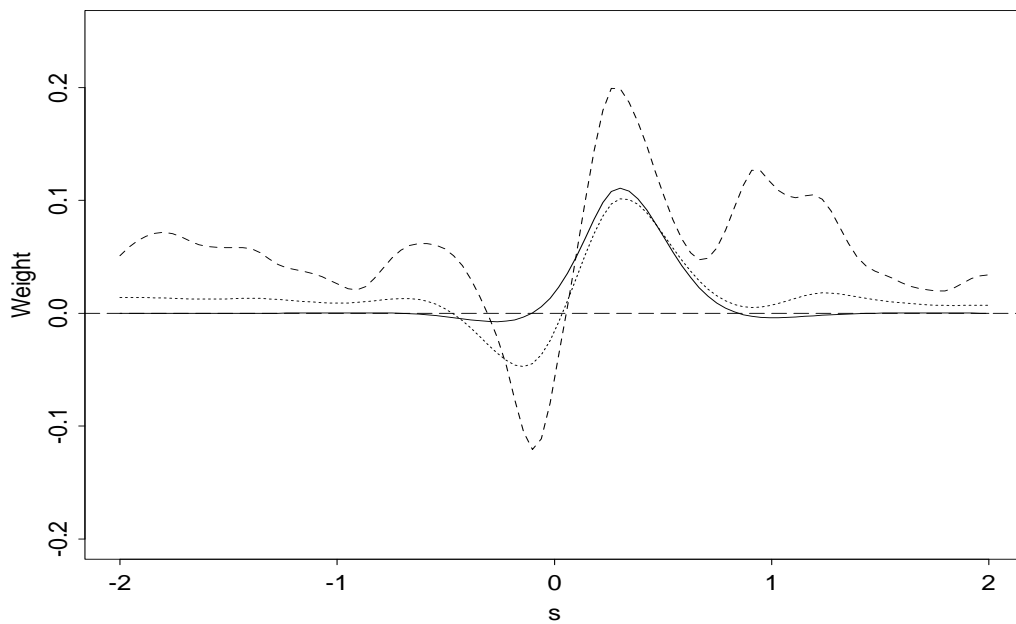


Figure 4: The equivalent kernel for the smoothing spline when $n = 35$, $m = 3$. The solid line is working independence, the dotted autocorrelation $\rho = 0.8$, and the dashed line is an unstructured and nearly singular correlation matrix as described in the text.