BIOMETRICS 56, 31-39
March 2000

Semiparametric Regression for Periodic Longitudinal Hormone Data from Multiple Menstrual Cycles

Daowen Zhang,[1,*] Xihong Lin,[2] and MaryFran Sowers[3]

[1]Department of Statistics, North Carolina State University,
Raleigh, North Carolina 27695-8203, U.S.A.
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.
[3]Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.
\email: dzhang2@unity.ncsu.edu

SUMMARY. We consider semiparametric regression for periodic longitudinal data. Parametric fixed effects are used to model the covariate effects and a periodic nonparametric smooth function is used to model the time effect. The within-subject correlation is modeled using subject-specific random effects and a random stochastic process with a periodic variance function. We use maximum penalized likelihood to estimate the regression coefficients and the periodic nonparametric time function, whose estimator is shown to be a periodic cubic smoothing spline. We use restricted maximum likelihood to simultaneously estimate the smoothing parameter and the variance components. We show that all model parameters can be easily obtained by fitting a linear mixed model. A common problem in the analysis of longitudinal data is to compare the time profiles of two groups, e.g., between treatment and placebo. We develop a scaled chi-squared test for the equality of two nonparametric time functions. The proposed model and the test are illustrated by analyzing hormone data collected during two consecutive menstrual cycles and their performance is evaluated through simulations.

KEY WORDS: Nonparametric regression; Penalized likelihood; Periodic smoothing spline; Restricted maximum likelihood; Test for equality of functions.

1. Introduction

In many longitudinal studies, it is of interest to model the time and covariate effects on an outcome variable. A common difficulty is that the time course is often too complicated to model parametrically. Examples include studies on growth (Donnelly, Laird, and Ware, 1995), HIV research on CD4 counts (Zeger and Diggle, 1994), and hormone research (Zhang et al., 1998). Several authors have hence considered semiparametric and nonparametric regression using kernel smoothing (Altman, 1991; Zeger and Diggle, 1994) and spline smoothing (Rice and Silverman, 1991; Zhang et al., 1998). Specifically, Zeger and Diggle (1994) and Zhang et al. (1998) proposed a semiparametric stochastic mixed model, which assumes parametric covariate effects and a nonparametric time effect and accounts for the within-subject correlation using random effects and a stochastic process. Unlike Zeger and Diggle (1994), a key feature of the Zhang et al. (1998) approach is that inference for all model components can easily proceed in a unified linear mixed model framework.

In some situations, especially in hormone research, nonparametric modeling of the time profile is further complicated by the fact that the data are collected periodically over time for each subject and both the mean and the variance of the outcome variable demonstrate periodic patterns. Another common problem in longitudinal data analysis is to compare the time profiles of two groups. For example, Sowers et al. (1998) conducted a longitudinal study to examine the effect of bone mineral density on reproductive hormone progesterone profiles. Urine samples were collected from 34 women with normal bone mineral density (controls) and 31 women with low bone mineral density (cases) on alternative days in two consecutive menstrual cycles. The investigators were interested in estimating the time courses of progesterone for both controls and cases and testing whether the time courses are the same in the two groups. The covariates of interest included age and body mass index (BMI).

It is meaningful biologically that the progesterone level changes periodically from one menstrual cycle to another. This view is supported by examining the raw data for both controls and cases (Figure 1a and 1b). Another feature of the data is that the variance of the progesterone level also changes periodically over time during the two consecutive menstrual cycles (Figure 2a and 2b). It is hence necessary to take into account these features when analyzing the progesterone data.

For independent data, a periodic cubic smoothing spline has often been used for estimating a periodic function nonparametrically (Wahba, 1980; Eubank, 1988; Wang and Brown, 1996). The smoothing paraṃeter is often estimated by minimizing integrated mean squared error or by cross-validation. Several authors also considered testing the equality

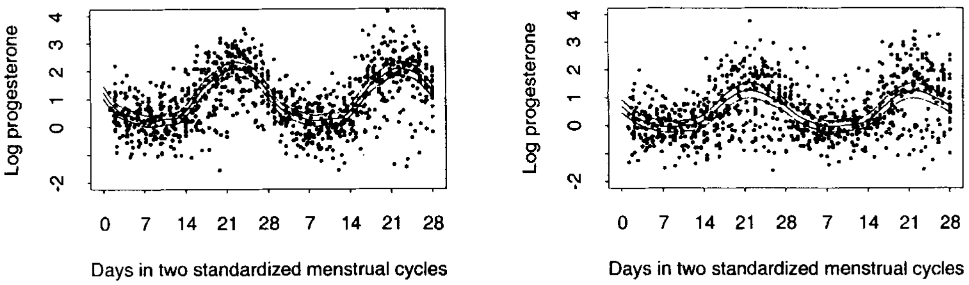

Figure 1. Log progesterone levels plotted against the number of days in two standardized menstrual cycles, with estimated population mean curve $\hat{f}(t)$ and 95% pointwise frequentist and Bayesian confidence intervals (CIs). —— $\hat{f}(t)$; - - - - frequentist CI; – – – Bayesian CI. The two plots are for (a) controls and (b) cases.

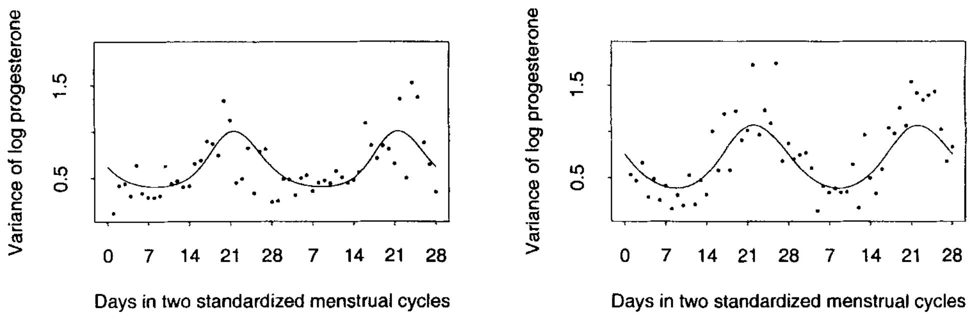

Figure 2. Sample variances of the log progesterone values calculated by grouping the data into 56 1-day intervals. The solid line is the estimated variance function curve obtained from fitting model (8). The two plots are for (a) controls and (b) case.

of two nonparametric functions for independent data. Härdle and Marron (1990) developed a test for comparing two functions that are the same up to a parametric transformation. Hall and Hart (1990) constructed a test statistic using the differences of the outcome variables of two groups by assuming they have the same design points and then applying a bootstrap test. Kulasekera (1995) used quasi-residuals to construct test statistics. Young and Bowman (1995) considered testing equality and parallelism of response curves from several groups.

In this paper, we propose a periodic semiparametric stochastic mixed model for periodic longitudinal data, such as the progesterone data. We use parametric functions to model the covariate effects and a periodic smooth nonparametric function to model the underlying complex periodic time course. The within-subject covariance is modeled using a random intercept and a stochastic process with periodic variance function. We use maximum penalized likelihood to estimate the regression coefficients and the periodic nonparametric function. The penalty is chosen in such a way that the resulting estimator of the nonparametric function is a periodic cubic smoothing spline. This formulation enables us to adapt the estimation procedure of Zhang et al. (1998) by casting our periodic nonparametric regression problem in a modified linear mixed model framework. Specifically, we write a periodic cubic smoothing spline estimator as a linear combination of a fixed effect and random effects and treat the inverse of the smoothing parameter as an extra variance component. The linear mixed model formulation of a smoothing spline was also used by Brumback and Rice (1998), Verbyla (1995), Verbyla et al. (1998), and Wang (1998a,b). The second objective of this paper is to propose a scaled chi-squared test for testing the equality of two nonparametric time functions. We illustrate the proposed model and the test by analyzing the progesterone data and evaluating their performance through simulations.

2. The Statistical Model

We present in this section a periodic semiparametric stochastic mixed model for periodic longitudinal data. Let the data consist of m subjects with the ith subject having n_i observations over time. Suppose Y_{ij} $(i = 1, \ldots, m,\ j = 1, \ldots, n_i)$ is the response for the ith subject at time point t_{ij} and satisfies

$$Y_{ij} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f(t_{ij}) + b_i + U_i(t_{ij}) + \epsilon_{ij}, \qquad (1)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients associated with the subject-level covariate vector $\mathbf{x}_i$; $f(t)$ is a twice-differentiable periodic function with the period length equal to T; the $b_i \sim \mathrm{N}(0, \phi)$ are independent subject-specific random intercepts; the $U_i(t)$ are independent and normally distributed mean-zero stochastic processes with periodic variance function $\xi(t)$ and correlation function $\mathrm{corr}(U_i(t), U_i(s)) = \eta(\rho; t, s)$, where $0 \leq \rho \leq 1$ is a correlation parameter; and the $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ are independent measurement errors. We further assume that $b_i, U_i(t)$ and ϵ_{ij} are mutually independent. For the sake of identifiability, we assume $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ does not contain an intercept.

Compared to the semiparametric stochastic mixed model of Zeger and Diggle (1994) and Zhang et al. (1998), a key feature of model (1) is that both the nonparametric function $f(t)$ and the variance function of $U_i(t)$ are constrained to be periodic functions.

3. Estimation Procedure

3.1 *Some Notation*

Without loss of generality, we assume $t_{ij} \geq 0$ and $\min\{t_{ij}\} = 0$. Due to the periodicity of $f(t)$, we only need to estimate $f(t)$ for $t \in [0, T)$. Let $\mathbf{t}^0 = (t_1^0, \ldots, t_r^0)^{\mathrm{T}}$ $(t_j^0 \in [0, T))$ be a vector of r ordered, distinct values of $t_{ij}^0 = \mathrm{mod}(t_{ij}, T)$ for $i = 1, \ldots, m, j = 1, \ldots, n_i$, and let $\mathbf{N}_i$ be an $n_i \times r$ incidence matrix for the ith subject connecting $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})^{\mathrm{T}}$ and $\mathbf{t}^0$ such that the (j, l)th element of $\mathbf{N}_i$ is one if $t_{ij}^0 = t_l^0$ and zero otherwise $(j = 1, \ldots, n_i, l = 1, \ldots, r)$. Denote $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}$, $\mathbf{Y} = (\mathbf{Y}_1^{\mathrm{T}}, \ldots, \mathbf{Y}_m^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{X}, \mathbf{N}, \mathbf{Z}, \mathbf{U}, \boldsymbol{\epsilon}$ similarly. We can write model (1) as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{f} + \mathbf{Z}\mathbf{b} + \mathbf{U} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{f} = (f(t_1^0), \ldots, f(t_r^0))^{\mathrm{T}}$; $\mathbf{b} = (b_1, \ldots, b_m)^{\mathrm{T}}$ is distributed as $\mathrm{N}(0, D(\phi))$, with $D(\phi) = \mathrm{diag}(\phi, \ldots, \phi)$; $\mathbf{U} = (\mathbf{U}_1^{\mathrm{T}}, \ldots, \mathbf{U}_m^{\mathrm{T}})^{\mathrm{T}}$ is distributed as $\mathrm{N}(0, \boldsymbol{\Gamma}(\boldsymbol{\xi}, \rho))$, with $\boldsymbol{\Gamma}(\boldsymbol{\xi}, \rho) = \mathrm{diag}(\boldsymbol{\Gamma}_1(\mathbf{t}_1, \mathbf{t}_1), \ldots, \boldsymbol{\Gamma}_m(\mathbf{t}_m, \mathbf{t}_m))$ and the (j, j')th element $(j, j' = 1, \ldots, n_i)$ of $\boldsymbol{\Gamma}_i(\mathbf{t}_i, \mathbf{t}_i)$ being $(\xi(t_{ij})\xi(t_{ij'}))^{1/2}\eta(\rho; t_{ij}, t_{ij'})$; and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_m^{\mathrm{T}})^{\mathrm{T}}$ is distributed as $\mathrm{N}(0, \sigma^2\mathbf{I})$, with $\mathbf{I}$ being an identity matrix of dimension $n = \Sigma_{i=1}^n n_i$.

3.2 *The Penalized Likelihood*

Denote $\mathbf{V} = \mathrm{cov}(\mathbf{Y})$. For given variance components $\boldsymbol{\theta} = (\phi, \boldsymbol{\xi}^{\mathrm{T}}, \rho, \sigma^2)^{\mathrm{T}}$, the log-likelihood function of $(\boldsymbol{\beta}, \mathbf{f})$ is, apart from a constant,

$$\ell(\boldsymbol{\beta}, \mathbf{f}; \mathbf{Y}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f}).$$

Since $f(t)$ is a periodic nonparametric function, we consider the penalized log likelihood

$$\ell(\boldsymbol{\beta}, \mathbf{f}; \mathbf{Y}) - \frac{\lambda}{2}\int_0^T [f''(t)]^2 dt = \ell(\boldsymbol{\beta}, \mathbf{f}; \mathbf{Y}) - \frac{\lambda}{2}\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f}, \tag{3}$$

where $\lambda > 0$ is a smoothing parameter controlling the balance between goodness-of-fit and smoothness and $f(t)$ is a twice-differentiable function subject to the periodic constraints $f(0) = f(T), f'(0) = f'(T)$, and $f''(0) = f''(T)$. The matrix $\mathbf{K}$ is a nonnegative definite periodic smoothing matrix defined in Appendix A. Note that matrix $\mathbf{K}$ defined herein differs from the conventional smoothing matrix given in Green and Silverman (1994, p. 12) in that it is constructed using the periodic constraints of $f(t)$ and the modified knot vector $\mathbf{t}^0$ and has only one 0 eigenvalue. The proof of the equality in (3) is given in Appendix A. The maximizer $(\hat{\boldsymbol{\beta}}, \hat{f}(t))$ of equation (3) is defined as the maximum penalized likelihood estimator (MPLE) and can be easily shown to be a periodic cubic smoothing spline (Appendix A).

Equation (3) has exactly the same form as equation (8) of Zhang et al. (1998) except that $\mathbf{K}$ now is a periodic smoothing matrix. Therefore, we can adapt their approach for inference on all model components in periodic model (1) within a unified framework by representing model (2) as a linear mixed model.

3.3 *Estimation of Model Components Using a Linear Mixed Model Representation*

Following Zhang et al. (1998), we show in this section how to make inference within a linear mixed model framework on all model components of model (1), including the mean parameters $\boldsymbol{\beta}$ and $\mathbf{f}$, the random intercept b_i and the stochastic process $U_i(t)$, and the smoothing parameter λ and the variance components $\boldsymbol{\theta}$.

From Appendix A, $\mathbf{K}$ has rank $r - 1$ and satisfies $\mathbf{K}\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ is an $r \times 1$ vector of ones. Similar to Green (1987) and Zhang et al. (1998), it can be shown that there exists an $r \times (r - 1)$ full rank matrix $\mathbf{B}$ such that $\mathbf{f} = \mathbf{1}\delta + \mathbf{B}\mathbf{a}$ and $\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f} = \mathbf{a}^{\mathrm{T}}\mathbf{a}$ for a scalar δ and a vector $\mathbf{a}$ of dimension $(r-1)$. Using the equality $\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f} = \mathbf{a}^{\mathrm{T}}\mathbf{a}$, the penalized log likelihood (3) becomes

$$\begin{aligned} &-\frac{1}{2}\log|\mathbf{V}| \\ &\quad -\frac{1}{2}(\mathbf{Y} - \mathbf{1}\delta - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{B}\mathbf{a})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{1}\delta - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{B}\mathbf{a}) \\ &\quad -\frac{1}{2\tau}\mathbf{a}^{\mathrm{T}}\mathbf{a}, \end{aligned}$$

where $\tau = 1/\lambda$. It follows that the periodic semiparametric mixed model (2) can be written as a modified linear mixed model,

$$\mathbf{Y} = \mathbf{1}\delta + \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{B}\mathbf{a} + \mathbf{Z}\mathbf{b} + \mathbf{U} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones, $\boldsymbol{\beta}_* = (\delta, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ are regression coefficients, $\mathbf{b}_* = (\mathbf{a}^{\mathrm{T}}, \mathbf{b}^{\mathrm{T}}, \mathbf{U}^{\mathrm{T}})^{\mathrm{T}}$ are mutually independent random effects with $\mathbf{a}$ distributed as $\mathrm{N}(0, \tau\mathbf{I})$ and $(\mathbf{b}, \mathbf{U})$ having the same distributions as those given in Section 3.1.

Using the results of Zhang et al. (1998), one can easily show that the MPLEs $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}})$ correspond to the linear combinations of the best linear unbiased predictors (BLUPs) $\hat{\boldsymbol{\beta}}_*$ and $\hat{\mathbf{a}}$ under the linear mixed model (4) with $\hat{\mathbf{f}} = \mathbf{1}\hat{\delta} + \mathbf{B}\hat{\mathbf{a}}$. The estimators of the random effect b_i and the stochastic process $\mathbf{U}_i(t)$ can also be obtained as BLUPs under the linear mixed model (4). The bias expressions of these MPLEs have the same form as given in Zhang et al. (1998). (For more details, see Zhang et al. [1998].)

The standard errors of $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}})$ and $(\hat{b}_i, \hat{\mathbf{U}}_i(t))$ can be calculated using either the frequentist or the Bayesian approach, where the frequentist standard errors are calculated by assuming the true $f(t)$ is a fixed smooth function, while the Bayesian standard errors are calculated by assuming a flat prior for $\boldsymbol{\beta}$ and a periodic cubic smoothing spline prior for $\mathbf{f}$, whose log density is $-\lambda\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f}/2$, or, equivalently, a periodic improper integrated Wiener prior for $\mathbf{f}$ (Wahba, 1980). We evaluate through simulations the performance of both standard errors in Section 6.1.

Following Zhang et al. (1998), we simultaneously estimate the smoothing parameter and variance components using restricted maximum likelihood (REML) (Harville, 1977) under the mixed model representation (4) by treating both $\tau = 1/\lambda$ and $\boldsymbol{\theta}$ as variance components. (For more details and justification of REML estimation, see Zhang et al. [1998].) We evaluate through simulations the performance of the proposed estimation procedure in Section 6.1.

4. The Global Test for Equality of Two Nonparametric Functions

A common problem in the analysis of longitudinal data is to compare the covariate-adjusted nonparametric time courses of two groups, as illustrated by the progesterone data introduced in Section 1. A simple method is to construct pointwise confidence intervals of the differences of the estimated nonparametric time functions of the two groups. However, it is often of interest to develop a global test for the equality of the two time functions. We hence consider in this section such a global test.

4.1 *The Global Test Statistic*

Suppose the two comparison groups consist of m_1 and m_2 subjects, respectively, and the outcome variable Y for group k ($k = 1, 2$) satisfies the following model:

$$Y_{kij} = \mathbf{x}_{ki}^{\mathrm{T}}\beta + f_k(t_{kij}) + b_{ki} + U_{ki}(t_{kij}) + \epsilon_{kij}, \qquad (5)$$

where all model components have specifications similar to those given in model (1). A natural global measure of the difference between $f_1(t)$ and $f_2(t)$ is

$$\Delta(f_1(\cdot), f_2(\cdot)) = \int_0^T [f_1(t) - f_2(t)]^2\, dt.$$

To test the null hypothesis H_0: $f_1(t) = f_2(t)$, we hence construct the test statistic

$$S\left(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}\right) = \int_0^T \left[\hat{f}_1(t) - \hat{f}_2(t)\right]^2 dt, \qquad (6)$$

where $\hat{f}_k(t)$ ($k = 1, 2$) is the MPLE of $f_k(t)$ obtained by separately fitting model (5) to the kth group data $\mathbf{Y}_{(k)}$. A large value of $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ would provide evidence against the null hypothesis H_0. However, the exact null distribution of S is difficult to evaluate. A simple approximation is proposed in the next section.

Note that, unlike the test statistic considered by Hall and Hart (1990), the proposed test statistic $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ does not require the two comparison groups to have the same design points. It can also be extended to $g > 2$ groups by using the sum of the test statistics $S(\mathbf{Y}_{(k)}, \mathbf{Y}_{(k+1)})$ for $k = 1, \ldots, g-1$. Another feature is that this test statistic and the test procedures proposed in the ensuing section can easily be extended to test the equivalence of two nonparametric functions in any subinterval of $[0, T]$ of interest.

4.2 *The Scaled Chi-Squared Test*

Denote by λ_k the smoothing parameter and by $\boldsymbol{\theta}_k$ the variance components under model (5) for group $k = 1, 2$. Using the results in Section 3.2 and Appendix A, it can be easily shown that, for given λ_k and $\boldsymbol{\theta}_k$, there exists a vector function $\mathbf{c}_k(t)$ such that the MPLE $\hat{f}_k(t)$ can be written as $\hat{f}_k(t) = \mathbf{c}_k^{\mathrm{T}}(t)\mathbf{Y}_{(k)}$. Let $\mathbf{c}(t) = [\mathbf{c}_1(t)^{\mathrm{T}}, -\mathbf{c}_2(t)^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{Y}_{(0)} = [\mathbf{Y}_{(1)}^{\mathrm{T}}, \mathbf{Y}_{(2)}^{\mathrm{T}}]^{\mathrm{T}}$. It follows that the test statistic $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ can be written as a quadratic function of the data $\mathbf{Y}_{(0)}$,

$$S\left(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}\right) = \int_0^T \mathbf{Y}_{(0)}^{\mathrm{T}}\mathbf{c}(t)\mathbf{c}(t)^{\mathrm{T}}\mathbf{Y}_{(0)}dt = \mathbf{Y}_{(0)}^{\mathrm{T}}\mathbf{C}\mathbf{Y}_{(0)}, \qquad (7)$$

where $\mathbf{C} = \int_0^T \mathbf{c}(t)\mathbf{c}(t)^{\mathrm{T}}dt$ and the integration is evaluated for each element of $\mathbf{C}$.

Equation (7) suggests that $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ follows a mixture of chi-squared distributions. We hence propose to use Satterthwaite's (1946) method to approximate the distribution of $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ under H_0: $f_1(t) = f_2(t)$ by a scaled chi-squared distribution $\kappa\chi_\nu^2$. Denote by $\mathbf{E}_0$ and $\mathbf{V}_0$ the mean and covariance of $\mathbf{Y}_{(0)}$ under H_0. Then the mean e and variance ψ of the test statistic $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ under H_0 can be calculated as

$$e = \mathbf{E}_0^{\mathrm{T}}\mathbf{C}\mathbf{E}_0 + \mathrm{tr}(\mathbf{C}\mathbf{V}_0),$$
$$\psi = 2\mathrm{tr}(\mathbf{C}\mathbf{V}_0)^2 + 4\mathbf{E}_0^{\mathrm{T}}\mathbf{C}\mathbf{V}_0\mathbf{C}\mathbf{E}_0.$$

In practice, e and ψ are evaluated at the MPLEs of $\boldsymbol{\beta}_k$ and f_k and the REML estimates of λ_k and $\boldsymbol{\theta}_k$ under H_0. Since $\mathbf{C}\mathbf{E}_0$ is negligible under H_0, e and ψ can often be approximated by $e \approx \mathrm{tr}(\mathbf{C}\mathbf{V}_0)$ and $\psi \approx 2\mathrm{tr}(\mathbf{C}\mathbf{V}_0)^2$. Equating e and ψ to the mean and the variance of $\kappa\chi_\nu^2$ gives $\kappa = \psi/(2e)$ and $\nu = 2e^2/\psi$. Denote $\chi_{\mathrm{obs}}^2 = S_{\mathrm{obs}}/\kappa$. The p-value of the test statistic $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ can be approximated by p-value $=$ $\mathrm{Prob}[\chi_\nu^2 > \chi_{\mathrm{obs}}^2]$.

To study the property of the approximation, one can easily see that ignoring $\mathbf{C}\mathbf{E}_0$ is equivalent to assuming that the biases in $\hat{f}_1(t)$ and $\hat{f}_2(t)$ cancel under H_0. For independent data with a single nonparametric function, Young and Bowman (1995) proposed a test statistic similar to $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$, where they estimated the nonparametric functions using the kernel method. They showed that the biases in the estimated nonparametric functions canceled asymptotically under the null hypothesis when the Gasser–Müller (1979) kernel estimator was used.

For longitudinal data, when the two groups have the same values of $(t_{ij}, \mathbf{x}_i)$ and $(\boldsymbol{\theta}, \lambda)$, the bias results in Zhang et al. (1998) show that the biases in the smoothing spline estimates $\hat{f}_1(t)$ and $\hat{f}_2(t)$ cancel under H_0. We hence have $\mathbf{C}\mathbf{E}_0 = 0$ and can ignore $\mathbf{C}\mathbf{E}_0$ in calculating e and ϕ. When the t_{ij}'s differ between the two groups, using the equivalent kernel results of Silverman (1984) and the results of Young and Bowman (1995), we expect the biases in the smoothing spline estimates $\hat{f}_1(t)$ and $\hat{f}_2(t)$ could also cancel asymptotically. In more general situations where the two groups have different values of $(t_{ij}, \mathbf{x}_i)$ and $(\boldsymbol{\theta}, \lambda)$, we expect that the biases in $\hat{f}_1(t)$ and $\hat{f}_2(t)$ could partially cancel out under H_0 and $\mathbf{C}\mathbf{E}_0$ is negligible. Alternatively, bias correction techniques such as undersmoothing could be used to reduce the biases in $\hat{f}_1(t)$ and $\hat{f}_2(t)$ when calculating $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$.

One can use higher moments in matching to improve the approximation of the distribution of the test statistic $S(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)})$ (Young and Bowman, 1995). We illustrate this scaled chi-squared test by application to the progesterone data in Section 5 and evaluate its performance through simulations in Section 6.2.

5. Application to the Progesterone Data

We applied the proposed periodic semiparametric stochastic mixed model in analyzing the progesterone data introduced in Section 1. Progesterone is a reproductive hormone responsible for normal fertility and menstrual cycling. The study sample consisted of 65 premenopausal women aged 29–46 years, 34 with normal bone mass density (controls) and 31 with low bone mass density (cases). Urine samples were collected and analyzed on alternative days during two consecutive menstrual cycles. The objectives of the study were to examine the time courses of the progesterone levels in a menstrual cycle for controls and cases and to compare the progesterone time courses between these two groups while accounting for the possible effects of age and body mass index (BMI). (See Sowers et al. [1998] for more details.)

Controls had 23–42 observations over time, with an average of about 28 observations. Their menstrual cycle lengths ranged from 23 to 56 days, with an average of 28.5 days. Cases had 21–42 observations over time, with an average of 28.5 observations. Their menstrual cycle lengths ranged from 21 to 56 days, with an average of 28.8 days. To overcome the problem of unequal cycle lengths among the study participants, each woman's menstrual length was standardized uniformly to a reference 28-day menstrual cycle (Sowers et al., 1998; Zhang et al., 1998). A log transformation was applied to the progesterone level to make the normality assumption more plausible.

Figure 1a and 1b displays the log-transformed progesterone values in the two consecutive menstrual cycles for controls and cases, respectively. Figure 2a and 2b shows their empirical sample variances calculated by grouping the data into 56 1-day intervals. Note that, since urine samples were analyzed on alternative days, the observations within every 1-day interval came from different subjects and were hence independent. These figures suggest that both the mean and the variance of the progesterone level vary over time periodically from one menstrual cycle to another.

Denote by Y_{ij} the jth log-transformed progesterone value measured on standardized day t_{ij} since the menstruation of the first cycle for the ith control (or case) and by AGE_i and BMI_i her age and body mass index, respectively. We fit the following periodic semiparametric stochastic mixed model for controls and cases separately:

$$Y_{ij} = \beta_1 \text{AGE}_i + \beta_2 \text{BMI}_i + f(t_{ij}) + b_i + U_i(t_{ij}) + \epsilon_{ij}, \quad (8)$$

where $f(t)$ is a periodic function with the period length T equal to 28 days, the b_i are independent random intercepts following $\text{N}(0, \phi)$, the $U_i(t)$ are random mean-zero nonhomogeneous Ornstein–Uhlenbeck (NOU) processes modeling serial correlation, and the ϵ_{ij} are independent measurement errors following $\text{N}(0, \sigma^2)$. The NOU process $U_i(t)$ has an exponentially decaying correlation coefficient $\text{corr}(U_i(t), U_i(s)) = \rho^{|t-s|}$. To allow the variance of Y_{ij} to vary periodically over time, we assumed a periodic variance function of $U_i(t)$ as $\text{var}(U_i(t)) = \exp(\xi(t))$, where $\xi(t)$ is a periodic cubic spline with some fixed knots in $[0, T]$. (See Appendix B for the functional form of a periodic cubic spline.) After some exploration of $\xi(t)$, we found that a periodic cubic spline with two equally spaced interior knots fit the empirical variances very well for both controls and cases. For the sake of computational stability and the ease of interpretation, time since first menstruation t_{ij} was divided by 10 and age and BMI were centered at the medians 37 years and 25 kg/m^2 and were divided by 100. Therefore, $f(t)$ represents the progesterone profile for 37-year-old women with BMI = 25 kg/m^2 for controls (or cases).

Figure 1a and 1b superimposes the MPLEs of $f(t)$ and their pointwise 95% frequentist and Bayesian confidence intervals for controls and cases, respectively. A common feature of controls and cases is that the progesterone levels remain relatively low and stable in the first half of a menstrual cycle and increase markedly after ovulation. They reach a peak around the 23rd reference day and then decrease. A comparison of Figure 1a and 1b shows that controls have a much higher peak value than cases. This suggests that bone mineral density might affect the progesterone profile.

Table 1 presents the estimates of the regression coefficients, the variance components, and the smoothing parameter for controls and cases separately. Note that the frequentist and Bayesian standard errors of the estimates of the regression coefficients are almost identical, which is consistent with the results of Zhang et al. (1998). The estimates of the coefficients ξ_0, ξ_1, and ξ_2 of the log-variance function $\xi(t)$ (see Appendix B for their definition) indicate that the variance of the progesterone level varies strongly over time. The estimated variance curves are superimposed in Figure 2a and 2b and agree very well with the empirical variances for both groups. These results suggest that controls and cases have variance functions of similar pattern. However, controls seem to have higher between-subjects variability and lower within-subject variability than cases.

Figure 3 shows the difference of the MPLEs $\hat{f}(t)$'s between controls and cases and its 95% pointwise frequentist and Bayesian confidence intervals. The two progesterone profiles seem not statistically significantly different before ovulation. However, controls have significantly higher progesterone levels after ovulation. This again indicates the effect of bone mineral density on the progesterone profile.

We used the scaled chi-squared test proposed in Section 4 to test whether or not controls and cases have the same progesterone profile globally. The observed scaled chi-squared test statistic was 18.8 with 1.69 d.f., which strongly suggests that the controls and cases have significantly different overall progesterone profiles (p-value = 0.000).

6. Simulation Studies

6.1 *Evaluation of the Estimation Procedure*

We conducted a simulation study to evaluate the performance of the MPLEs of the regression coefficients and the nonparametric function and the REML estimates of the smoothing parameter and the variance components in periodic semiparametric stochastic mixed models. The design of the simulation study was identical to that of the original progesterone data for controls. We generated data in two consecutive cycles from model (8), where the true model parameters and $f(t)$ were set equal to the estimates obtained from the analysis of the observed control data given in Table 1 and Figure 1a. Five hundred simulation data sets were generated.

Table 2 gives the relative biases, empirical standard errors, and model-based frequentist and Bayesian standard errors of the parameter estimates. All estimates have minimal biases. The model-based standard errors of the parameter estimates agree very well with the empirical standard errors. The frequentist and Bayesian standard errors of the regression coefficient estimates are almost identical. Figure 4 shows that

Table 1
Estimates of the regression coefficients, variance components, and smoothing parameter for controls and cases in the progesterone data

	Controls			Cases		
Parameter	Estimate	Bayesian SE	Frequentist SE	Estimate	Bayesian SE	Frequentist SE
β_1	1.0428	1.9007	1.9006	1.2381	1.9495	1.9494
β_2	−2.2412	2.3512	2.3512	−1.6870	2.3650	2.3649
ϕ	0.2563	0.0708		0.0914	0.0563	
ρ	0.0706	0.0334		0.3571	0.0732	
ξ_0	−1.3204	0.1615		−0.5333	0.1601	
ξ_1	−4.7935	0.7282		−2.7229	0.4309	
ξ_2	0.5945	0.5942		1.8300	0.3935	
σ^2	0.0985	0.0161		0.0783	0.0127	
τ	9.3669	4.9036		3.6821	2.2530	

the bias in the estimated nonparametric function $f(t)$ is minimal. Figure 5 compares the pointwise model-based frequentist and Bayesian standard errors with the empirical standard errors. They both agree quite well with the empirical standard errors. Figure 6 compares the estimated pointwise coverage probabilities of the frequentist and Bayesian confidence intervals of $f(t)$. The nominal coverage probability is 95%. Their overall performance is similar and the averages of the estimated frequentist and Bayesian coverage probabilities over time are 94.2 and 94.9%, respectively. These results are consistent with those given in Zhang et al. (1998).

6.2 *Evaluation of the Scaled Chi-Squared Test*

We conducted a separate simulation study to evaluate the performance of the scaled chi-squared test proposed in Section 4.2 for testing the equality of two (periodic) nonparametric functions. The simulation design was the same as that of the progesterone data except that only the first menstrual cycle design was used for the sake of computational simplicity. The data Y were generated separately for controls and cases using model (8), where the true $f(t)$ was set to be $f_{dk}(t) = (d/4)\hat{f}_k(t) + [1 - (d/4)]\hat{f}(t)$, where $d = 0, 1, 2, 3, 4$ and $k = 1$ for controls and $k = 2$ for cases. Here $\hat{f}_1(t)$ and $\hat{f}_2(t)$ are estimated $f(t)$ for controls and cases, respectively, from the real data and $\hat{f}(t)$ is the estimated common $f(t)$ by pooling the two group data assuming $f_1(t) = f_2(t)$. The other true model parameters were set to be their estimates given in Table 1.

The size of the scaled chi-squared test was studied using the data generated with $d = 0$ and using 1000 simulated data sets. The power of the test was studied by setting $d = 1, 2, 3, 4$ and using 500 simulated data sets. Table 3 reports the empirical size and power. The observed size of the test is very close to the nominal value. As the difference between the two nonparametric functions becomes larger, the power of the test increases. These results show that the scaled chi-squared test performs very well in terms of both size and power.

7. Discussion

In this paper, we proposed a semiparametric stochastic mixed model for periodic longitudinal data. Maximum penalized

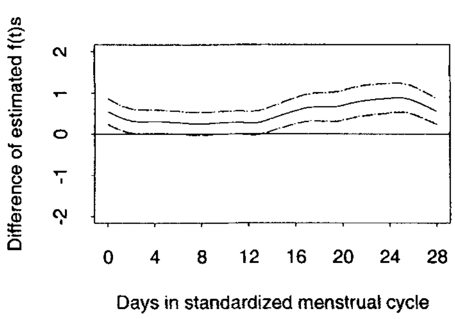

Figure 3. The difference between two estimated population mean curves $\hat{f}(t)$'s for controls and cases of the progesterone study and its 95% pointwise frequentist and Bayesian confidence intervals (CIs). —— Difference of $\hat{f}(t)$'s; - - - - frequentist CI; — — — Bayesian CI.

Table 2
Relative biases and standard error estimates from simulation study of 500 replications based on the model for controls

Model parameter	Relative bias	Empirical SE	Model-based Bayesian SE	Model-based frequentist SE
β_1	−0.0841	1.8968	1.8610	1.8610
β_2	0.0461	2.3386	2.3021	2.3021
ϕ	−0.0266	0.0692	0.0689	
ρ	0.0651	0.0352	0.0337	
ξ_0	0.0112	0.1770	0.1640	
ξ_1	0.0111	0.7551	0.7366	
ξ_2	0.0070	0.6262	0.5997	
σ^2	−0.0101	0.0168	0.0163	

Table 3
Empirical size and power of the scaled chi-squared test based on 500 replications

Size[a]	Power			
$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d = 4$
0.052	0.194	0.442	0.770	0.964

[a] Nominal size is 0.05 and 1000 replications were used.

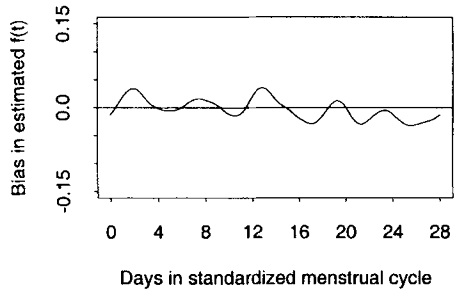

Figure 4. Empirical biases in the estimated nonparametric function $\hat{f}(t)$ based on 500 replications using the design of controls.

likelihood (MPL) is used to estimate the regression coefficients and the nonparametric function. Restricted maximum likelihood (REML) is used to estimate the smoothing parameter and the variance components simultaneously. A key feature of this approach is that all model components can be easily estimated by fitting a modified linear mixed model. Our simulation study results show that MPL and REML estimation performs well.

We applied the periodic semiparametric mixed model to the analysis of the progesterone data. The choice of this periodic model is natural, given the scientific knowledge of how hormone level changes in each cycle and the features of the data. Compared to the results obtained by using a nonperiodic model in Zhang et al. (1998), the values of the estimated periodic nonparametric function have smaller biases and variances at $t = 0$ and $t = T(28)$, the two boundary points of the nonparametric function in Zhang et al. (1998). This is because these two points are no longer boundary points of the periodic nonparametric function in our periodic model. Therefore, better inference can be made on the periodic nonparametric function in the neighborhood of these two points.

We proposed in this paper a scaled chi-squared test for the equality of two nonparametric functions. Our simulation study results show that this simple scaled chi-squared test performs well. We studied in special cases the impact of the biases in the estimated nonparametric functions on the test statistic. However, its asymptotic property under the general semiparametric model (1) still requires further research.

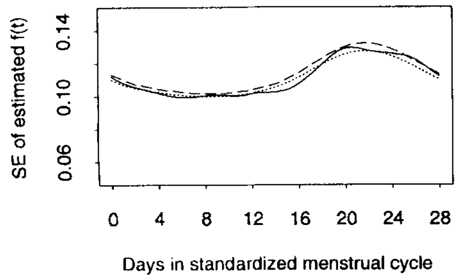

Figure 5. Empirical, frequentist, and Bayesian pointwise standard errors of the estimated nonparametric time functions $\hat{f}(t)$ based on 500 replications. —— empirical SE; - - - - frequentist SE; – – – Bayesian SE.

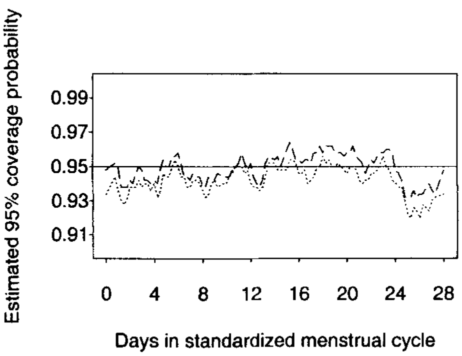

Figure 6. Estimated pointwise frequentist and Bayesian 95% coverage probabilities of the values of the true fixed function $f(t)$ based on 500 replications. —— Nominal level (95%); - - - - frequentist; – – – Bayesian.

We estimated in this paper the nonparametric time function using a periodic smoothing spline. Alternatively, one can use the kernel method to estimate $f(t)$. However, if kernel estimation is used, one cannot cast the semiparametric model (1) within a unified linear mixed model framework. The subsequent inference for the other model parameters, especially the variance components, might be difficult.

Acknowledgements

This work was supported in part by National Cancer Institute grant R29 CA76404 and National Institute of Health grant R01 NIAMS 40888. The authors are grateful to the editor, the associate editor, and two anonymous referees for comments that helped greatly improve the paper.

Résumé

On considère la régression semi-paramétrique pour des données longitudinales périodiques. On utilise un modèle paramétrique à effets fixes pour décrire les effets des covariables et une fonction périodique non-paramétrique lissée pour décrire l'effet du temps. La corrélation intra-sujet est décrite par un modèle aléatoire spécifique au sujet combiné à un processus stochastique avec variance périodique. L'estimation simultanée du paramètre de lissage et des composantes de la variance est réalisée par maximum de vraisemblance restreint. Nous montrons que tous les paramètres du modèles sont aisément obtenus avec l'ajustement d'un modèle linéaire mixte. Un problème courant dans l'analyse de données longitudinales est la comparaison inter-groupe de profils, par exemple entre un groupe placebo et un groupe traitement. On développe un test de forme Khi Deux pour la comparaison de deux fonctions du temps non-paramétriques. Le modèle proposé et le test sont illustrés en analysant des données de dosage d'hormone réalisé au cours de deux cycles menstruels consécutifs, leur performance est évaluée à l'aide de simulations.

References

Altman, N. S. (1991). Kernel smoothing of data with correlated error. *Journal of the American Statistical Association* **85**, 749–759.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association* **93**, 961–994.

Donnelly, C. A., Laird, N. M., and Ware, J. H. (1995). Prediction and creation of smooth curves for temporally correlated longitudinal data. *Journal of the American Statistical Association* **90**, 984–989.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Decker.

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions. In *Lectures Notes in Mathematics,* Volume 757, T. Gasser and M. Rosenblastt (eds), 23–68. New York: Springer-Verlag.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55,** 245–260.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman and Hall.

Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* **85**, 1039–1049.

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *Annals of Statistics* **18,** 63–89.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72,** 320–340.

Kulasekera, K. B. (1985). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association* **90,** 1085–1093.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* **53,** 233–243.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrika* **2,** 110–114.

Silverman, B. (1984). Spline smoothing: The equivalent variable kernel method. *Annals of Statistics* **12,** 898–916.

Sowers, M. F., Crutchfield, M., Shapiro, B., Zhang, B., Pietra, M. L., Randolph, J. F., and Schork, M. A. (1998). Urinary ovarian and gonadotrophin hormone levels in premenopausal women with low bone mass. *Journal of Bone and Mineral Research* **13,** 1191–1202.

Verbyla, A. P. (1995). *A mixed model formulation of smoothing splines and testing linearity in generalized linear models.* Research Report 95/5, Department of Statistics, University of Adelaide, Adelaide.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1998). Analysis of designed experiments and longitudinal data using smoothing splines. *Applied Statistics* **47,** in press.

Wahba, G. (1980). Automatic smoothing of the log periodogram. *Journal of the American Statistical Association* **75,** 122–132.

Wang, Y. (1998a). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B* **60,** 159–174.

Wang, Y. (1998b). Smoothing spline models with correlated errors. *Journal of the American Statistical Association* **93,** 341–348.

Wang, Y. and Brown, M. B. (1996). A flexible model for human circadian rhythms. *Biometrics* **52,** 588–596.

Young, S. G. and Bowman, A. W. (1995). Non-parametric analysis of covariance. *Biometrics* **51,** 920–931.

Zeger, S. L. and Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50,** 689–699.

Zhang, D., Lin, X., Raz, J., and Sowers, M. F. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93,** 710–719.

Received February 1998. *Revised April* 1999.
Accepted June 1999.

Appendix A

Proof of $\int_0^T [f''(t)]^2 dt = \mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f}$ *for a Periodic Cubic Smoothing Spline* $f(t)$

Let $t^0_{r+1} = T$, $f_l = f(t^0_l), \gamma_l = f''(t^0_l)$ for $l = 1, \ldots, r+1$, and $h_l = t^0_{l+1} - t^0_l$ for $l = 1, \ldots, r$. Following Green and Silverman (1994), we have

$$\begin{aligned}\int_0^T [f''(t)]^2 dt &= f'(t)f''(t)|_0^T - \int_0^T f'''(t)f'(t)dt \\ &= -\sum_{l=1}^{r} \int_{t_l}^{t_{l+1}} f'''(t)f'(t)dt \\ &= -\sum_{l=1}^{r} \frac{\gamma_{l+1} - \gamma_l}{h_l}(f_{l+1} - f_l).\end{aligned}$$

The periodicity of $f(t)$ implies $f_1 = f_{r+1}$ and $\gamma_1 = \gamma_{r+1}$. It follows that $\int_0^T [f''(t)]^2 dt = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Q}\mathbf{f}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)^{\mathrm{T}}$, $\mathbf{f} = (f_1, \ldots, f_r)^{\mathrm{T}}$, and $\mathbf{Q} = \{q_{ij}\}$ is an $r \times r$ symmetric matrix of rank $(r-1)$ whose nonzero elements q_{ij} are

$$\begin{aligned} q_{11} &= -h_1^{-1} - h_r^{-1}, \\ q_{1r} &= q_{r1} = h_r^{-1}, \\ q_{ll} &= -h_{l-1}^{-1} - h_l^{-1}, \\ q_{l-1,l} &= q_{l,l-1} = h_{l-1}^{-1} \qquad (l = 2, \ldots, r). \end{aligned}$$

Using the results of Green and Silverman (1994, p. 24), we can show that $\boldsymbol{\gamma}$ and $\mathbf{f}$ are related by $\mathbf{R}\boldsymbol{\gamma} = \mathbf{Q}\mathbf{f}$, where $\mathbf{R} = \{r_{ij}\}$ is an $r \times r$ positive definite matrix whose nonzero elements r_{ij} are

$$\begin{aligned} r_{11} &= \frac{1}{3}(h_1 + h_r), \\ r_{1r} &= r_{r1} = \frac{1}{6}h_r, \\ r_{ll} &= \frac{1}{3}(h_{l-1} + h_l), \\ r_{l-1,l} &= r_{l,l-1} = \frac{1}{6}h_{l-1} \qquad (l = 2, \ldots, r). \end{aligned}$$

We hence have $\int_0^T [f''(t)]^2 dt = \mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f}$, with $\mathbf{K} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}$. Note the rank of the periodic smoothing matrix $\mathbf{K}$ is $r-1$ instead of $r-2$, which is the rank of conventional smoothing matrix $\mathbf{K}$ defined by Green and Silverman (1994).

Appendix B

Functional Form of a Periodic Cubic Spline with Fixed Interior Knots

We here derive the functional form of a periodic cubic spline with fixed interior knots that is used in Section 5. Suppose $0 < t_1 < \ldots < t_k < T$ are k interior knots in $[0,T]$. Then a periodic cubic spline $\xi(t)$ with these knots is

$$\xi(t) = \xi_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \sum_{j=1}^{k} \xi_j (t-t_j)_+^3,$$

where a_+ denotes the positive part of a and $\xi(t)$ needs to satisfy the periodic constraints $\xi(0) = \xi(T), \xi'(0) = \xi'(T)$, and $\xi''(0) = \xi''(T)$. Some algebra shows that the resulting periodic cubic spline $\xi(t)$ for $t \in [0,T]$ takes the form

$$\xi(t) = \xi_0 + \sum_{j=1}^{k} \xi_j s_j(t),$$

where $s_j(t) = a_j t + b_j t^2 + c_j t^3 + (t-t_j)_+^3$, whose coefficients a_j, b_j, and c_j are

$$\begin{aligned} a_j &= -\frac{T(T-t_j)}{2} + \frac{3(T-t_j)^2}{2} - \frac{(T-t_j)^3}{T}, \\ b_j &= \frac{3(T-t_j)}{2} - \frac{3(T-t_j)^2}{2T}, \\ c_j &= \frac{T-t_j}{T}. \end{aligned}$$

A periodic cubic spline is obtained by replicating $\xi(t)$ $(t \in [0,T])$ periodically outside $[0,T]$.