Statistica Sinica **14**(2004), 713-729

RESTRICTED LIKELIHOOD RATIO TESTS IN NONPARAMETRIC LONGITUDINAL MODELS

Ciprian M. Crainiceanu and David Ruppert

Johns Hopkins University and Cornell University

Abstract: We assume that repeated measurements are taken on each of several subjects that are randomly sampled from some population. The observations on a particular subject are expressed as the sum of an average curve for the population and a deviation of the subject's curve from the average plus independent errors. Both curves are modeled nonparametrically as splines. We use roughness penalties on the splines, which is equivalent to assuming a linear mixed model. Within this linear mixed model, we consider likelihood ratio tests of several scientifically relevant hypotheses about the two curves, for example, that the subject deviations are all zero or that they are each constant. The large-sample null distributions of the test statistics are shown to be non-standard, but we develop bootstrap techniques that can compute the exact null distributions much more rapidly than a direct application of the bootstrap.

Key words and phrases: Bootstrap, linear mixed models, non-standard asymptotics, penalized splines, subject-specific curves, variance components.

1. Introduction

Brumback and Rice (1998) study an important class of models for longitudinal data. In this paper, we consider a subclass of those models where repeated observations are taken on each of several subjects. Suppose that y_{ij} is the jth observation on the ith subject recorded at time t_{ij}, where $1 \le i \le I$, $1 \le j \le J(i)$, and $n = \sum_{i=1}^{I} J(i)$ is the total number of observations. Consider the nonparametric model

$$y_{ij} = f(t_{ij}) + f_i(t_{ij}) + \epsilon_{ij} \ , \tag{1}$$

where ϵ_{ij} are independent $N(0, \sigma_\epsilon^2)$ errors and both the population curve $f(\cdot)$ and $f_i(\cdot)$, the deviation of the ith subject's curve from the population average, are modeled nonparametrically. Models similar to (1) have been studied by many other authors, e.g., Wang (1998).

A number of simple special cases of (1) are potentially of interest as null hypotheses. For example, we might wish to test that there are no subject effects, that is, that $f_i(t) \equiv 0$. Alternatively, we might wish to test that each $f_i(t)$ is

constant, that is, $f_i(t) \equiv a_{i0}$ for some constants a_{i0} so that the subject curves $f + f_i$ are parallel, or that each f_i is a linear function.

Following Brumback and Rice, we model the population and subject curves as splines, though we do not use smoothing splines as they do. Rather we use penalized splines (P-splines) with a relatively small number of knots. The advantage of P-splines is that the number of parameters can be kept reasonably small, which makes rapid computation feasible, while the accuracy is as good as with smoothing splines (Ruppert (2002)).

Penalized splines can be viewed as BLUPs in linear mixed models (LMM's). The hypotheses that interest us can each be expressed as a constraint that certain variance components are zero. It is natural to test these hypotheses using likelihood ratio tests (LRTs). LRTs in linear mixed models have already been studied by Crainiceanu, Ruppert and Vogelsang (2002), Crainiceanu and Ruppert (2004), Crainiceanu, Ruppert, Claeskens and Wand (2002). However, those papers consider simpler testing situations, for example, testing that a univariate regression is a polynomial versus the alternative that it is a spline. One interesting conclusion of these papers is that the asymptotic null distribution of LRTs in spline mixed models is not a mixture of chi-squared distributions. This conclusion is surprising, since chi-squared mixtures are expected from the classical work of Chernoff (1954) and Self and Liang (1987). As explained in Section 4, the reason that one does not obtain standard large-sample null distributions for spline mixed models is a lack of independence. However, Crainiceanu, Ruppert and Vogelsang (2002) and Crainiceanu and Ruppert (2004) were able to find the large-sample null distribution needed for testing in spline mixed models. This paper extends their work to model (1).

2. Univariate Nonparametric Regression

Consider the regression equation

$$y_i = f(x_i) + \epsilon_i \,,$$

where ϵ_i are i.i.d. $N(0, \sigma_\epsilon^2)$. Let $\kappa_1 < \cdots < \kappa_K$ be fixed knots in the range of x. Let $B_0(x), \ldots, B_{p+K}(x)$ be a basis for the space of pth degree splines with these knots, such that $B_0(x), \ldots, B_p(x)$ span the space of polynomials of degree p — the latter requirement is for convenience when we wish to test the null hypothesis that $f(x)$ is a polynomial. We assume that

$$f(x) = f(x, \boldsymbol{\theta}) = \sum_{j=0}^{p+K} \theta_j B_j(x) = \sum_{j=0}^{p} \beta_j B_j(x) + \sum_{j=p+1}^{p+K} b_j B_j(x),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{b}^T)^T$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ are the coefficients of the polynomial basis functions, and $\boldsymbol{b} = (b_1, \ldots, b_K)^T$ are the coefficients of the other basis

functions. A convenient choice of basis, and the one used in our numerical work, is the set of polynomials plus truncated power functions where

$$f(x,\boldsymbol{\theta}) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} b_k (x-\kappa_k)_+^p \ . \tag{2}$$

The choice of K is discussed in Ruppert (2002) who finds that for P-splines the exact value of K has little effect on the estimator, provided that K is at least a certain minimum value, because the amount of smoothing is determined not by K but rather by the penalty parameter λ that is discussed below. Berry, Carroll and Ruppert (2002) found that P-splines and smoothing splines, which have a knot at each unique value of the covariate, generally give very similar answers, so one can use a large number of knots even if it is not necessary to do so. We consider a number of knots that is large enough (typically 5–20) to ensure the desired flexibility, and κ_k is the sample quantile of x's corresponding to probability $k/(K+1)$. The amount of smoothing depends on the trace of the smoother matrix $\boldsymbol{S}_\lambda$ defined below in (9). This trace is also called the effective degrees of freedom of the fit and denoted by $\mathrm{df}_{\mathrm{fit}}$. For a pth degree spline with K knots, $\mathrm{df}_{\mathrm{fit}}$ increases smoothly from $p+1$ to $K+p+1$ as λ increases from 0 to ∞, so K determines only the maximum value of $\mathrm{df}_{\mathrm{fit}}$; see Ruppert, Wand and Carroll (2003). For example, if $\mathrm{df}_{\mathrm{fit}} = 7.5$ gives a good tradeoff between bias and variance and $p = 2$, then roughly any value of K above 5 would be suitable.

Powell's (1981) results on the approximation properties of splines help explain why relatively few knots are needed, why their exact locations are not crucial, and why K depends little upon the sample size n. Let $a < b$ be two real numbers and consider spline approximation of the regression function f on $[a,b]$. Powell's Theorem 20.3 gives the error of the best approximation of a $C^l[a,b]$ (l times continuously differentiable) function f by an kth degree spline. For convenience, we restate the theorem.

Theorem 1. (Powell, 1981) *Let $\mathcal{S}$ be the set of pth degree splines on $[a,b]$ with knots $a < \kappa_1 < \cdots < \kappa_K < b$. Define $\kappa_0 = a$ and $\kappa_{k+1} = b$. Let $h = \max\{\kappa_j - \kappa_{j-1} : j = 1,\ldots,K+1\}$. Suppose that f is $C^l[a,b]$. Then*

$$\inf_{s\in\mathcal{S}} \|f-s\|_\infty \le \frac{(k+1)!}{(k+1-j)!}(h/2)^j \|f^{(j)}\|_\infty. \tag{3}$$

for every $j \in \{1,2,\ldots,\min(l,p+1)\}$, where $\|\cdot\|_\infty$ is the L^∞ norm on $[a,b]$.

This bound is independent of the specific knot locations, instead depending only on the maximum distance between any two consecutive knots. For example, if one uses quadratic splines and $f \in C^3[a,b]$, then (3) holds for $j = 3$. If the covariate has a density on $[a,b]$ that is bounded away from zero, then h will be

proportional to $1/K$ for knots at sample quantiles with equal probability spacing. Thus, the bias due to the approximation of f by a spline will be $O(K^{-3})$. If $K \propto n^{1/6}$, then the squared bias due to spline approximation will be $O(1/n)$, the parametric rate for the variance. For most practical purposes, $K \propto n^{1/6}$ is essentially the same as K being independent of n. For example, if for some f using $K = 5$ is sufficient with $n = 100$, we solve for C in $5 = C100^{1/6}$ to find that $C = 2.32$. Then when $n = 30,000$, we might expect that $K = (2.32)(30,000^{1/6}) = 13$ would be sufficient. Since $K = 13$ could also be used when $n = 100$ and even for a smaller sample size, say $n = 50$, we might use $K = 13$ over the entire range $50 \leq n \leq 30,000$.

The criterion to be minimized is a penalized sum of squares

$$\sum_{i=1}^{n} \{y_i - f(x_i, \boldsymbol{\theta})\}^2 + \lambda^{-1} \boldsymbol{\theta}^T \boldsymbol{G} \boldsymbol{\theta} \,, \tag{4}$$

where λ is the smoothing parameter selected by some external criteria, such as ML or REML. $\boldsymbol{G}$ is a positive semi-definite matrix which is determined by the form of the penalty and therefore is *known*, as will soon be illustrated.

Let $f^{(q)}(x, \boldsymbol{\theta})$ be the qst derivative with respect to x. The penalty

$$\lambda^{-1} \int \{f^{(q)}(x, \boldsymbol{\theta})\}^2 dx, \; q \leq p, \tag{5}$$

used for smoothing splines, typically with $q = 2$, can be achieved with $\boldsymbol{G}$ equal to the matrix of sample second moments of the qth derivatives of the spline basis functions; notice that in this case $\boldsymbol{G}$ is known. However, in this paper we focus on matrices $\boldsymbol{G}$ of the form

$$\boldsymbol{G} = \begin{bmatrix} \mathbf{0}_{p+1 \times p+1} & \mathbf{0}_{p+1 \times K} \\ \mathbf{0}_{K \times p+1} & \boldsymbol{\Sigma}^{-1} \end{bmatrix}, \tag{6}$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix and $\mathbf{0}_{m \times l}$ is an $m \times l$ matrix of zeros. When the truncated power basis (2) is used, this choice of the matrix $\boldsymbol{G}$ does not penalize the coefficients of the polynomial basis functions and will be used in the remainder of the paper. With this $\boldsymbol{G}$, (4) can be viewed as minus twice the log-likelihood of a linear mixed model with $\lambda \boldsymbol{\Sigma}^{-1}$ being the covariance matrix of the random effects; see Section 3. Therefore, the reason this covariance matrix is known, at least up to the parameter λ, is that it is part of a Bayesian prior that is equivalent to the roughness penalty having a known form since it is chosen by the user. For example, in the smoothing spline literature, the penalty (5) was apparently chosen because it seemed reasonable and worked well in practice, and then later Wahba (1978) proved that smoothing splines were Bayes estimators for a particular partially improper prior. The prior is improper because the

coefficients of the polynomial basis functions have a prior that is uniform on $\Re^{p+1}$; these coefficients can be interpreted as the fixed effect parameters.

With the truncated power basis, a standard choice is $\mathbf{\Sigma} = \boldsymbol{I}_K$, so that

$$\boldsymbol{G} = \begin{bmatrix} \mathbf{0}_{p+1\times p+1} & \mathbf{0}_{p+1\times K} \\ \mathbf{0}_{K\times p+1} & \boldsymbol{I}_K \end{bmatrix}. \tag{7}$$

This choice of $\mathbf{\Sigma}$ makes the P-spline a close approximation to a smoothing spline, as will now be discussed. A smoothing spline uses the penalty (5), usually with $p = 1$, which causes $f(x, \widehat{\boldsymbol{\theta}})$ to converge to the pth degree polynomial regression fit as $\lambda \to \infty$. The pth degree spline (2) has a pth derivative which is constant between knots and takes a jump of size $p!\, b_k$ at the kth knot. Thus, the $p+1$th derivative of this spline is a generalized function (linear combination of Dirac delta functions) and, with $q = p + 1$, the analogue of (5) is

$$\lambda^{-1}(p!)^2 \sum_{k=1}^{K} (b_k)^2, \tag{8}$$

which, after absorbing $(p!)^2$ into λ^{-1}, is the penalty in (4) if $\boldsymbol{G}$ is given by (7). In particular, $\boldsymbol{G}$ is known. Note that the choice of $\boldsymbol{G}$ depends on the spline basis being used, and the P-spline penalty when $\boldsymbol{G}$ is given by (7) is analogous to a smoothing spline penalty only for the truncated power basis in (2). For another basis, say B-splines, $\boldsymbol{G}$ is obtained from (7) using a known change of basis matrix (Ruppert, Wand and Carroll (2003)), so that $\boldsymbol{G}$ will again be known. Penalty matrix (7) or, equivalently, penalty (8) has the Bayesian interpretation that our prior knowledge about the jumps of $f^{(p)}$ is independent between knots. Since we typically have little prior knowledge of the fine structure of f, this prior seems reasonable. The point is that the form of the penalty is not so much an assumption about f but rather about our prior knowledge of f.

Let $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$, and $\boldsymbol{\mathcal{X}}$ be the design matrix having the ith row $\boldsymbol{\mathcal{X}}_i = \{B_0(x_i), \ldots, B_{p+K}(x_i)\}$. Then, for a fixed λ, the vector of regression coefficients that minimizes (4) is $\hat{\boldsymbol{\theta}}(\lambda) = (\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}} + \lambda^{-1}\boldsymbol{G}^{-1})\boldsymbol{\mathcal{X}}^T\boldsymbol{Y}$, and the estimated smoothed values are given by $\hat{\boldsymbol{Y}}(\lambda) = \boldsymbol{S}_\lambda\boldsymbol{Y}$, where

$$\boldsymbol{S}_\lambda = \boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}} + \lambda^{-1}\boldsymbol{G})^{-1}\boldsymbol{\mathcal{X}}^T \tag{9}$$

is the smoother matrix. Decompose $\boldsymbol{\mathcal{X}}$ as $\boldsymbol{\mathcal{X}} = [\boldsymbol{X}|\boldsymbol{Z}]$ where $\boldsymbol{X}$ is formed with the first $p+1$ columns of $\boldsymbol{\mathcal{X}}$ and corresponds to the polynomial basis functions.

3. Penalized Splines as Linear Mixed Models

Observe that when the penalized spline fitting criterion (4) is divided by σ_ϵ^2, we obtain

$$\sigma_\epsilon^{-2} \left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}\right\|^2 + (\lambda\sigma_\epsilon^2)^{-1}\boldsymbol{b}^T\mathbf{\Sigma}^{-1}\boldsymbol{b}.$$

Define $\sigma_b^2 = \lambda\sigma_\epsilon^2$ and consider the vector $\boldsymbol{\beta}$ as an unknown fixed parameter and the vector $\boldsymbol{b}$ as a random parameter with $E(\boldsymbol{b}) = 0$ and $\mathrm{Cov}(\boldsymbol{b}) = \sigma_b^2\boldsymbol{\Sigma}$. If $(\boldsymbol{b}^T, \boldsymbol{\epsilon}^T)^T$ is a normal random vector and $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent, then one obtains the Linear Mixed Model (LMM) representation (Brumback, Ruppert and Wand (1999)) of the penalized spline model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \;\; \mathrm{Cov}\begin{pmatrix}\boldsymbol{b}\\ \boldsymbol{\epsilon}\end{pmatrix} = \begin{bmatrix}\sigma_b^2\boldsymbol{\Sigma} & 0\\ 0 & \sigma_\epsilon^2\boldsymbol{I}_n\end{bmatrix}. \tag{10}$$

For this model $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{Y}) = \sigma_\epsilon^2\boldsymbol{V}_\lambda$, where $\boldsymbol{V}_\lambda = \boldsymbol{I}_n + \lambda\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^T$ and n is the total number of observations. The fitted spline is the best linear unbiased predictor (BLUP) of $\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{b} = E(\boldsymbol{Y}|\boldsymbol{b})$. In the penalized spline (4), λ is a tuning parameter controlling the amount of smoothing, whereas in the LMM model (10) the ratio of the variance components σ_b^2 and σ_ϵ^2 controls the amount of shrinkage. A standard estimation criterion for model (10) is the Restricted Likelihood

$$\begin{aligned}\mathrm{REL}(\boldsymbol{\beta},\, \sigma_\epsilon^2,\, \lambda) = -\Big[&(n-p-1)\log(\sigma_\epsilon^2) + \log\{\det(\boldsymbol{V}_\lambda)\}\\ &+ \log\left\{\det(\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})\right\} + \frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma_\epsilon^2}\Big].\end{aligned} \tag{11}$$

The joint maximization of this criterion over $(\boldsymbol{\beta},\, \sigma_\epsilon^2,\, \lambda)$ provides the Restricted Maximum Likelihood (REML) estimators. Restricted, or residual, maximum likelihood was introduced in the framework of LMMs by Patterson and Thompson (1971) to take into account the loss in degrees of freedom due to estimation of the $\boldsymbol{\beta}$ parameters and thereby to obtain unbiased variance components estimators. REML consists in maximizing the likelihood function associated with $n - p - 1$ linearly independent error contrasts (Harville (1977)). REML, as a method for choosing the smoothing parameter of a smoothing spline, was studied by Anderssen and Bloomfield (1974), Wecker and Ansley (1983) and Barry (1983); see also Wahba (1990, p.63). A comparison between REML and GCV was made by Wahba (1985) and Kohn, Ansley and Tharm (1991). Wahba (1985) refers to REML as GML (generalized maximum likelihood).

Given the representation (10) of the penalized spline, testing whether the regression function $f(\cdot)$ is a degree p polynomial is equivalent to testing

$$H_0:\; \sigma_b^2 = 0 \quad \text{vs.} \quad H_A:\; \sigma_b^2 > 0\,. \tag{12}$$

Since $\sigma_b^2 = \lambda\sigma_\epsilon^2$ the null is equivalent to $\lambda = 0$ and the alternative is equivalent to $\lambda > 0$. Because the coefficients $\boldsymbol{b}$ have mean zero and covariance matrix $\sigma_b^2\boldsymbol{\Sigma}$, the condition that $\sigma_b^2 = 0$ in H_0 is equivalent to the condition that all truncated

polynomial coefficients b_i are identically zero and the spline is a polynomial. If $\sigma_b^2 > 0$, then any open set of truncated power function coefficients has positive probability so $f(x, \boldsymbol{\theta})$ can be an arbitrary spline with the given knots. As explained in Section 2, for most practical purposes, assuming that f is a spline with the given knots is the same as assuming that f is a smooth function.

The testing problem is non-standard because the parameter vector is on the boundary of the parameter space ($\sigma_b^2 = 0$) under the null; however, likelihood ratio testing with the null hypothesis on the boundary has been investigated by Chernoff (1954), Self and Liang (1985) and others. A more serious difficulty, which has not been investigated until recently, is that for tests that variance components are zero in a LMM, observations $\boldsymbol{Y}$ are not independent, at least not under the alternative and often not under the null. Nonstandard asymptotic theory developed by Self and Liang require independence both under the null and alternative hypotheses. Crainiceanu and Ruppert (2004) find that the asymptotic *null* distribution given by the Self and Liang theory need not hold *even when the data are independent under the null hypothesis* if they are not independent under the alternative. The likelihood ratio statistic depends upon the alternative, so its null distribution also depends on the alternative.

The same type of equivalence with standard mixed models can be obtained more generally for penalized likelihood models. Natural extensions to semiparametric models are discussed in detail in Ruppert, Wand and Carroll (2003). In Section 5 we discuss the extension to nonparametric longitudinal models.

4. RLRT Tests for Polynomial Regression

Define the (log) Restricted Likelihood Ratio Test (RLRT) statistic as

$$\text{RLRT} = \sup_{H_A} \text{REL}(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) - \sup_{H_0} \text{REL}(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) \ ,$$

where H_0 and H_A are given by (12). Because REML uses the likelihood of residuals after fitting the fixed effects, it is appropriate for testing only if the fixed effects are the same under the null and the alternative hypotheses.

Computing RLRT is very simple. Indeed, under H_0 one need only compute the REML for a polynomial regression. Under H_A one need only compute the REML for a LMM with one variance component. Available software, such as S-PLUS (lme function) or SAS (MIXED procedure), provide excellent tools for this type of calculations.

Finding the finite sample or asymptotic null distribution of the RLRT is a more challenging problem. Self and Liang (1987, 1995), Stram and Lee (1994) and Andrews (2001) assume that the data can be represented as an i.i.d. sequence, but this assumption does not hold in general for LMM's, at least not under

the alternative hypothesis. Crainiceanu, Ruppert and Vogelsang (2002) show that the asymptotic null distributions for LRT's and RLRT's in penalized spline LMM's are different from the chi-squared mixture limits derived by Self and Liang (1987, 1995), Stram and Lee (1994) and Andrews (2001) for certain other LMM's. The reason one does not in general get chi-squared mixtures as the asympotic null distribution is the violation of the i.i.d. assumption. Self and Liang (1987, 1995) explicitly state that the data are i.i.d. for all values of the parameter (see their introduction). Stram and Lee (1994) assume that random effects are independent from subject to subject and they implicitly assume that the number of subject increases to infinity. Their results would not hold for a fixed number of subjects, even if the number of observations per subject increases to infinity. Andrews's (2001) results for the random coefficients model are derived under the independence of data assumption.

The null finite sample and asymptotic distribution for RLRT are derived by Crainiceanu and Ruppert (2004) for the case of a LMM with one variance component, and by Crainiceanu, Ruppert, Claeskens and Wand (2002) for testing polynomial regression against a general alternative modeled by a penalized spline.

Score tests have been proposed for testing variance components in linear mixed models and it has been assumed that the asymptotic theory of score tests for independent data is applicable, e.g., by Verbeke and Molenberghs (2003). However, there seems to be have been no investigation of when this theory holds for mixed models, and we are reluctant to use score tests for P-spline models until the relevant asymptotic theory or, better, exact distributions have been developed.

5. Nonparametric Models for Longitudinal Data

Consider (1) and assume that both the population curve $f(\cdot)$ and the ith subject effect (deviation from the population curve) $f_i(\cdot)$ are modeled nonparametrically as degree p splines

$$f(t) = \sum_{k=0}^{p} \beta_k t^k + \sum_{k=1}^{K_1} b_k (t - \kappa_{1,k})_+^p \,, \quad f_i(t) = \sum_{k=0}^{p} a_{ik} t^k + \sum_{k=1}^{K_2} u_{ik} (t - \kappa_{2,k})_+^p \,.$$

Here, for concreteness, we are using the truncated power basis, but other bases could be used. The knots $\kappa_{1,1}, \ldots, \kappa_{1,K_1}$ are for the population curve, and the knots $\kappa_{2,1} \ldots, \kappa_{2,K_2}$ are for the subject curves. The numbers of knots K_1 and K_2 are fixed, but large enough to ensure the desired flexibility. In some applications one would use $K_1 = K_2$, but often using $K_2 < K_1$ would be sensible since the individual subject curves are estimated with less data than the population curve. To simplify the notation, we take the same spline degree, number of knots,

and knots for $f(\cdot)$ and $f_i(\cdot)$. To avoid overfitting, penalties are imposed on the truncated spline basis coefficients.

For the population curve, $\beta_0, \ldots, \beta_p$ will be fixed effects and $b_1, \ldots, b_K$ will be independent random coefficients distributed as $N(0, \sigma_1^2)$. In LMM's, parameters are usually treated as random effects because they are subject-specific and the subjects have been sampled randomly. Here, $b_1, \ldots, b_K$ are treated as random effects for an entirely different reason. Modeling them as random specifies a Bayesian prior and allows for shrinkage that assures smoothness of the fit. For the subject curves, the vectors of polynomial coefficients $(a_{i0}, \ldots, a_{ip})$ are assumed independent across subjects with a $N(\mathbf{0}, \boldsymbol{C}_a)$ distribution, where $\boldsymbol{C}_a$ is an unknown covariance matrix. This is a typical random effects assumption since the subjects are sampled randomly. However, some other authors, e.g., Brumback and Rice (1998), treat these effects as fixed, presumably because they do not view the subjects as forming a *random* sample. When the polynomial coefficients a_{ik} are modeled as random, one issue is how to model the within-subject correlations between the a_{ij}. To reduce the number of unknown parameters, it would convenient to make the working assumption that there is no within-subject correlation between these coefficients. However, this assumption is unlikely to be true and even if true in one parameterization, it would not be true in others, e.g., it would not hold if t were uncentered even if it did with t mean-centered. In addition, u_{ik} will be treated as independent random coefficients distributed $N(0, \sigma_{p+3}^2)$. The nonparametric model (1) can be rewritten as a LMM with $p+3$ variance components

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{b} + \sum_{k=0}^{p} \boldsymbol{Z}_{k+2}\boldsymbol{a}_k + \boldsymbol{Z}_{p+3}\boldsymbol{u} + \boldsymbol{\epsilon}\,, \tag{13}$$

where $\boldsymbol{b} \sim N(\mathbf{0}, \sigma_1^2 \boldsymbol{I}_K)$, $\boldsymbol{a}_k = (\alpha_{1k}, \ldots, \alpha_{Ik})^T \sim N(\mathbf{0}, \sigma_{k+2}^2 \boldsymbol{I}_I)$ for $k = 0, \ldots, p$, $\boldsymbol{u} = (u_{11}, \ldots, u_{IK})^T \sim N(\mathbf{0}, \sigma_{p+3}^2 \boldsymbol{I}_{IK})$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_n)$. Here $\boldsymbol{I}_s$ denotes the $s \times s$ identity matrix and $\mathbf{0}$ is a column vector of zeros matching the size for each normal vector. For details on the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}_k$, $1 \le k \le p+3$, see the Appendix A1.

Using the equivalence between the Linear Mixed Model (13) and the nonparametric model (1), the parameters in the latter can be estimated using available software, such as S-PLUS (lme function) or SAS (MIXED procedure). The degree of the penalized spline is generally small, typically 1, 2, or 3 (Ruppert, Wand and Carroll, 2003). We take $p = 1$ for simplicity. In this case (a_{i0}, a_{i1}) has a normal distribution with mean zero and covariance matrix

$$\boldsymbol{C}_a = \begin{pmatrix} \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix},$$

where σ_2^2 and σ_3^2 are the random intercept and slope variances respectively, and ρ is the within subject correlation parameter.

Testing for simplifying assumptions, such as linear or constant individual deviations from the population curve against a general alternative are tests for zero variance of random effects. For example, consider testing $H_0: \sigma_4^2 = 0$ vs. $H_A: \sigma_4^2 > 0$. Because the u_{ik} are distributed $N(0, \sigma_4^2)$, under the null hypothesis all $u_{ik} = 0$ and the model becomes $y_{ij} = f(t_{ij}) + \alpha_{i0} + \alpha_{i1} t_{ij} + \epsilon_{ij}$, which assumes random linear individual deviations from the population curve $f(\cdot)$. Testing for random constant deviations can be achieved by testing $H_0: \sigma_3^2 = 0,\ \sigma_4^2 = 0$ vs. $H_A: \sigma_3^2 > 0$ or $\sigma_4^2 > 0$. In this case, the null hypothesis corresponds to the model $y_{ij} = f(t_{ij}) + \alpha_{i0} + \epsilon_{ij}$. Hence, testing for individual polynomial deviations from the population mean is equivalent to testing for zero variance of random effects in a particular LMM.

6. Restricted Likelihood Ratio Tests for Zero Variance in Linear Mixed Models

Consider a LMM with S variance components

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{s=1}^{S} \boldsymbol{Z}_s \boldsymbol{u}_s + \boldsymbol{\epsilon}\ , \tag{14}$$

where $\boldsymbol{\beta}$ captures p fixed effects, $\boldsymbol{u}_s$ are random effects, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}_n)$. We assume that $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_S)$ and $\boldsymbol{\epsilon}$ are independent, that the $\boldsymbol{u}_s$ have mean zero and covariance matrix $\sigma_s^2 \boldsymbol{C}_s$. Suppose that for some $S_0 \in \{0, 1, \ldots, S-1\}$ we want to test

$$H_0:\ \sigma_{S_0+1}^2 = 0, \ldots, \sigma_S^2 = 0 \quad \text{vs.} \quad H_0:\ \sigma_{S_0+1}^2 > 0,\ \text{or} \ldots, \text{or}\ \sigma_S^2 > 0\ .$$

Let $\boldsymbol{Z} = [\boldsymbol{Z}_1 | \ldots, | \boldsymbol{Z}_S]$, $\boldsymbol{u} = (\boldsymbol{u}_1^T, \ldots, \ldots, \boldsymbol{u}_S^T)^T$, $\lambda_s = \sigma_s^2 / \sigma_\epsilon^2$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S,$ $\boldsymbol{\rho}^T)^T$, where $\boldsymbol{\rho}$ is the vector containing all correlation parameters. Then (14) can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}\ ,$$

with $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{Cov}\,(\boldsymbol{Y}) = \sigma_\epsilon^2 \boldsymbol{V}_{\boldsymbol{\lambda}}$, where $\boldsymbol{V}_{\boldsymbol{\lambda}} = \boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{\lambda}}\boldsymbol{Z}^T$. Let $M = \sum_{s=1}^{S} M_s$ be the size of the matrix $\boldsymbol{D}_{\boldsymbol{\lambda}}$. For estimation we use (11) with the difference that the smoothing parameter λ is replaced by the vector $\boldsymbol{\lambda}$.

The Restricted Likelihood Ratio statistic for testing H_0 versus H_A is

$$\mathrm{RLRT} = \sup_{H_A} \mathrm{REL}(\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\lambda}) - \sup_{H_0} \mathrm{REL}(\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\lambda})\ .$$

Computing RLRT is easy using available software but determining the null finite sample or asymptotic distribution is not. Standard asymptotic theory does not

apply in this case because the response variable vector cannot be partitioned into i.i.d. components.

Good starting points have proved to be important in our simulations. We constructed a Latin hypercube by taking a matrix with columns corresponding to grid points for each parameter, and then randomly permuted each column. Thus, each row of the matrix provides one point in the parameter space, and we chose the point that minimizes the objective function over these points as the initial point in the nonlinear minimization algorithm. We used equally spaced points on the log scale for the λ_i parameters and equally spaced points for the correlation parameters.

For the case of LMM's with one variance component, Crainiceanu and Ruppert (2004) give the finite sample and asymptotic distributions of RLRT. For LMM with more than one variance component, Crainiceanu, Ruppert, Claeskens and Wand (2002) give the spectral decomposition of RLRT and fast simulation algorithms in some particular cases. Their results suggest that, in the general case, a good strategy (which we use here) is to use parametric bootstrap.

7. RLRT Simulation

The basic idea is to estimate the model under the null hypothesis and then use the parametric bootstrap to obtain the finite sample distribution of RLRT. This may be computationally intensive, but the algorithm can be made more efficient using basic matrix computation techniques.

For example, at each step, simulate $\boldsymbol{Y}$ under H_0 and maximize $\mathrm{REL}(\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\lambda})$ under H_0 and H_A. Maximizing under H_A with respect to $(\boldsymbol{\beta}, \sigma_\epsilon^2)$, one obtains the profile restricted likelihood $\mathrm{REL}(\boldsymbol{\lambda})$ for any fixed $\boldsymbol{\lambda}$. Using results from Harville (1977) we obtain that, up to a constant that does not depend on the parameters, the profile restricted likelihood is

$$\begin{aligned}\mathrm{REL}(\boldsymbol{\lambda}) = (n-p-1)\log\left\{a_{\boldsymbol{Y}} - \boldsymbol{b_Y}^T \boldsymbol{D_\lambda}\left(\boldsymbol{I}_M + \boldsymbol{C}\boldsymbol{D_\lambda}\right)^{-1}\boldsymbol{b_Y}\right\} \\ + \log\left\{\det(\boldsymbol{I}_M + \boldsymbol{C}\boldsymbol{D_\lambda})\right\}\ , \end{aligned} \tag{15}$$

where $a_{\boldsymbol{Y}} = \boldsymbol{Y}^T\boldsymbol{P_0}\boldsymbol{Y}$, $\boldsymbol{b_Y} = \boldsymbol{Z}^T\boldsymbol{P_0}\boldsymbol{Y}$, $\boldsymbol{C} = \boldsymbol{Z}^T\boldsymbol{P_0}\boldsymbol{Z}$ and $\boldsymbol{P_0} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. For details on derivation of (15) see Appendix A2. To obtain the maximum restricted likelihood under the alternative, one needs to maximize $\mathrm{REL}(\boldsymbol{\lambda})$ subject to $\boldsymbol{\lambda} \geq \boldsymbol{0}$. Under the null hypothesis the first S_0 entries of $\boldsymbol{\lambda}$ are restricted to be non-negative and the rest are set equal to zero.

The function $\mathrm{REL}(\boldsymbol{\lambda})$, its Jacobian and Hessian matrices will depend on $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ only through $a_{\boldsymbol{Y}}$, $\boldsymbol{b_Y}$ and $\boldsymbol{C}$. Because it does not depend on $\boldsymbol{\lambda}$ or the simulated vector $\boldsymbol{Y}$, the $M \times M$ matrix $\boldsymbol{C}$ needs to be computed only once, before simulation begins. Moreover, because the scalar $a_{\boldsymbol{Y}}$ and the $M \times 1$ vector $\boldsymbol{b_Y}$ do not depend on $\boldsymbol{\lambda}$, they need to be computed only once per simulation.

A second advantage of (15) is that the matrices $\boldsymbol{D}_{\boldsymbol{\lambda}}$ and $\boldsymbol{I}_M + \boldsymbol{C}\boldsymbol{D}_{\boldsymbol{\lambda}}$ are $M \times M$ matrices and not $n \times n$. In general, and for penalized splines in particular, the number of random effects M is much smaller than the number of observations n. In our experience, the implementation of the bootstrap depends on the efficient computation of $(\boldsymbol{I}_M + \boldsymbol{C}\boldsymbol{D}_{\boldsymbol{\lambda}})^{-1}$ and its first and second order derivatives. This can be done relatively easily when M is small to moderate in size, or the eigenvalues of $\boldsymbol{I}_M + \boldsymbol{C}\boldsymbol{D}_{\boldsymbol{\lambda}}$ are explicit functions of the eigenvalues of $\boldsymbol{C}$. For example, in the case of one variance component ($S = 1$) the eigenvalues of $\boldsymbol{I}_M + \boldsymbol{C}\boldsymbol{D}_{\boldsymbol{\lambda}}$ are $\xi_s(\lambda) = 1 + \lambda\mu_s$, where μ_s are the eigenvalues of $\boldsymbol{C}$. For a complete analysis of this case see Crainiceanu and Ruppert (2004) and Crainiceanu, Ruppert and Vogelsang (2002).

8. Example

To illustrate our methodology we consider a data set from Grizzle and Allen (1969) and Wang (1998). Data are coronary sinus potassium concentrations measured on 36 dogs. The measurements on each dog were taken every two minutes, seven observations for each dog.

Wang (1998) presents four smoothing spline analysis of variance models similar to our model (1). He shows how to use nonparametric mixed effects models for estimating the treatment effects and population mean concentration. This paper is concerned with testing simplifying assumptions such as random constant or linear individual deviations from the population mean.

The 36 dogs come from 4 treatments. To illustrate our methodology we ignore the treatments effects, these effects being subsumed within the subject (dog) effects.

We consider (1) where both the individual and the population response curves are modeled as piecewise linear splines with $K = 3$ knots. The knots are at 3, 7, 9, and the penalty is the sum of squares of the spline coefficients. As shown in Section 5, this model is equivalent to a Linear Mixed Model with

$$\boldsymbol{D}_{\boldsymbol{\lambda}} = \begin{pmatrix} \lambda_1 \boldsymbol{I}_K & \boldsymbol{0}_{K\times I} & \boldsymbol{0}_{K\times I} & \boldsymbol{0}_{K\times KI} \\ \boldsymbol{0}_{I\times K} & \lambda_2 \boldsymbol{I}_I & \rho\sqrt{\lambda_2\lambda_3}\boldsymbol{I}_I & \boldsymbol{0}_{I\times KI} \\ \boldsymbol{0}_{I\times K} & \rho\sqrt{\lambda_2\lambda_3}\boldsymbol{I}_I & \lambda_3 \boldsymbol{I}_I & \boldsymbol{0}_{I\times KI} \\ \boldsymbol{0}_{KI\times K} & \boldsymbol{0}_{KI\times I} & \boldsymbol{0}_{KI\times I} & \lambda_4 \boldsymbol{I}_{KI} \end{pmatrix}.$$

According to the parameters set to zero, we define 5 nested models corresponding to increasing degrees of complexity. Model 1 is obtained for $\sigma_1^2 = 0, \ldots, \sigma_4^2 =$

$0, \rho = 0$, and corresponds to a linear population curve with no subject effects. Model 2 is obtained for $\sigma_2^2 = 0, \sigma_3^2 = 0, \sigma_4^2 = 0, \rho = 0$, and corresponds to a nonparametric population curve with no subject effects. Model 3 is obtained for $\sigma_3^2 = 0, \sigma_4^2 = 0, \rho = 0$, and corresponds to a nonparametric population curve and subject random intercepts. Model 4 is obtained for $\sigma_4^2 = 0$, and corresponds to a nonparametric population curve with random subject linear deviations. Model 5 is the full model where none of the parameters is set to zero and corresponds to a nonparametric population curve with nonparametric subject deviations. Note that $\sigma_i^2 = 0$ is equivalent to $\lambda_i = 0$.

Table 1 reports the values of the Restricted Likelihood Ratio statistics corresponding to adding successively one variance component:

$$\text{RLRT} = \sup_{M_{i+1}} \text{REL}(\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\lambda}) - \sup_{M_i} \text{REL}(\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\lambda}) \, .$$

Here σ_1^2 is the variance of random coefficients for the population spline coefficients, σ_2^2 and σ_3^2 are the variances of individual random intercepts and slopes, and σ_4^2 is the variance random coefficients for the individual spline coefficients. The p-values were obtained by bootstrapping the null distribution of the RLRT. We used 2,000 simulations for each distribution. The p-value is 0.05 for testing M_1 versus M_2 (RLRT = 1.72). In all other cases, the p-value is much smaller.

Table 1. Longitudinal data nonparametric analysis of variance using restricted likelihood.

Model	Hypothesis	RLRT(M_i/M_{i+1})	p-value* (M_i/M_5)
M_1	$\sigma_1^2 = 0, \ldots, \sigma_4^2 = 0, \rho = 0$	1.72	0
M_2	$\sigma_2^2 = 0, \ldots, \sigma_4^2 = 0, \rho = 0$	182.71	0
M_3	$\sigma_3^2 = 0, \sigma_4^2 = 0, \rho = 0$	48.52	0
M_4	$\sigma_4^2 = 0$	12.4	0
M_5	no restrictions	–	–

The RLRT values for testing M_i versus M_5 can be obtained by simply adding the values of RLRT between model M_i and M_5. Table 1 also shows the p-values for testing all models M_i versus the full model M_5. Simulation of the null distribution of RLRT statistic in this case is computationally intensive due to the large number of random effects under the full model ($M = 183$), which is

also the dimension of the matrix $\boldsymbol{I}_M + \boldsymbol{C}\boldsymbol{D}_{\boldsymbol{\lambda}}$ discussed in Section 7. Moreover, the complexity of the simulation algorithm increases linearly with the number of subjects monitored because five additional random effects (α_{i0}, α_{i1}, u_{ik}, $1 \le k \le 3$) are needed to model each new subject.

As we have discussed, the null distribution theory is complex because the parameter is on the boundary under the null, and the data cannot be partitioned into i.i.d. subvectors. If the data could be so partitioned, then the asymptotic null distribution for testing that a single variance component is zero would be a $0.5\chi_0^2 + 0.5\chi_1^2$ mixture. To illustrate deviations from this distribution, Table 2 presents the probability mass at zero (p_0) and representative quantiles of the RLRT null distributions for testing M_1 versus M_2 (no subject effects and testing a linear versus a nonparametric population curve) and M_4 versus M_5 (linear versus nonparametric subject curve). In both situations we are testing whether one variance component is zero, so we compare these distributions with the $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution obtained for independent data. Such an approximation cannot be justified by asymptotic theory because of a lack of independence, but the approximation may be satisfactory, e.g., for testing M_4/M_5 but not for testing M_1/M_2.

Table 2. Longitudinal data nonparametric analysis of variance using restricted likelihood. Here p_0 is the null probability the log-likelihood ratio statistic is 0 and q_α is the null αth quantile of the log-likelihood ratio statistic. M_i/M_j uses the exact null distribution for testing M_i versus M_j and $0.5\chi_0^2 + 0.5\chi_1^2$ is an approximation using a chi-squared mixture.

Distribution	p_0	$q_{0.70}$	$q_{0.80}$	$q_{0.85}$	$q_{0.90}$	$q_{0.95}$
M_1/M_2	0.68	0.00	0.19	0.40	0.84	1.70
M_4/M_5	0.52	0.22	0.59	0.86	1.34	2.22
$0.5\chi_0^2 + 0.5\chi_1^2$	0.50	0.28	0.71	1.07	1.64	2.71

The $0.5\chi_0^2 + 0.5\chi_1^2$ distribution is a conservative approximation for the finite sample distribution of RLRT for both cases. This is not surprising given the results in Crainiceanu and Ruppert (2004) and Crainiceanu, Ruppert and Vogelsang (2002). However, $0.5\chi_0^2 + 0.5\chi_1^2$ is a better approximation of the null distribution when testing M_4 versus M_5. This is somewhat surprising because the asymptotic theory developed by Stram and Lee (1994) and Self and Liang (1987, 1995) does not directly apply to this case, given the dependence introduced by the population curve model.

There are two possible reasons why $0.5\chi^2_0 + 0.5\chi^2_1$ is a better approximation when testing M_4 versus M_5. The first could be that because, conditional on the population curve, the subjects are independent, a result similar to the one derived by Self and Liang (1987) for independent observations still holds for conditional independence. The second reason could be that, under M_5, the nonparametric population curve $f(\cdot)$ is close to being statistically indistinguishable from its linear component. This is suggested by the small value of RLRT = 1.72 (p-value=0.05) when testing M_1 versus M_2 and by the small value of $\lambda_1 = 0.04$ of the estimated smoothing parameter under M_5. If $f(\cdot)$ were replaced by a linear deterministic function, then the data $\boldsymbol{Y}$ would be partitioned into i.i.d. subvectors corresponding to each subject and standard asymptotic distribution would apply. These are interesting research problems that we intend to address in the future.

Appendix A1

We describe the design matrices from the Linear Mixed Model in (13). For concreteness, we use the truncated power basis, but other bases could be used instead. The $n \times (p+1)$ matrix $\boldsymbol{X}$ has the (i,j)th row (the row corresponding to the jth observation on the ith subject) $\boldsymbol{X}_{ij} = (1, t_{ij}, \ldots, t_{ij}^p)$. The $n \times K$ matrix $\boldsymbol{Z}_1$ has the (i,j)th row $\boldsymbol{Z}^1_{ij} = [(t_{ij} - \kappa_1)^p_+, \ldots, (t_{ij} - \kappa_K)^p_+]$. For $2 \le k \le p+3$ the $n \times I$ matrix $\boldsymbol{Z}_k$ has the form

$$\boldsymbol{Z}_k = \begin{bmatrix} \boldsymbol{Z}_{1k} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_{2\,k} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{Z}_{Ik} \end{bmatrix},$$

where for $2 \le k \le p+2$, $\boldsymbol{Z}_{i\,k}$ is a $J(i) \times 1$ column vector $\boldsymbol{Z}_{i\,k} = (t_{i1}^{k-2}, \ldots, t_{iJ(i)}^{k-2})^T$ and

$$\boldsymbol{Z}_{i\,p+3} = \begin{pmatrix} (t_{i1} - \kappa_1)^p_+ & \ldots & (t_{i1} - \kappa_K)^p_+ \\ \vdots & \ddots & \vdots \\ (t_{i\,J(i)} - \kappa_1)^p_+ & \ldots & (t_{i\,J(i)} - \kappa_K)^p_+ \end{pmatrix}.$$

Appendix A2

Let $\boldsymbol{P}_{\boldsymbol{\lambda}} = \boldsymbol{V}_{\boldsymbol{\lambda}}{}^{-1} - \boldsymbol{V}_{\boldsymbol{\lambda}}{}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_{\boldsymbol{\lambda}}{}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}_{\boldsymbol{\lambda}}{}^{-1}$ and note that the profile restricted likelihood is, up to a constant that does not depend on the parameters,

$$\mathrm{REL}(\boldsymbol{\lambda}) = -\{(n-p-1)\log(\boldsymbol{Y}^T\boldsymbol{P}_{\boldsymbol{\lambda}}\boldsymbol{Y}) + \log(\det \boldsymbol{V}_{\boldsymbol{\lambda}}) + \log(\det \boldsymbol{X}^T\boldsymbol{V}_{\boldsymbol{\lambda}}{}^{-1}\boldsymbol{X})\}\,.$$

The result in (15) can be obtained by using the following results from Harville (1977): $\boldsymbol{P}_{\boldsymbol{\lambda}} = \boldsymbol{P}_{\boldsymbol{0}} - \boldsymbol{P}_{\boldsymbol{0}}\boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{\lambda}}(\boldsymbol{I}_M + \boldsymbol{Z}^T\boldsymbol{P}_{\boldsymbol{0}}\boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{\lambda}})^{-1}\boldsymbol{Z}^T\boldsymbol{P}_{\boldsymbol{0}}$ and $\det(\boldsymbol{V}_{\boldsymbol{\lambda}})\det(\boldsymbol{X}^T\boldsymbol{V}_{\boldsymbol{\lambda}}{}^{-1}\boldsymbol{X})\det(\boldsymbol{X}^T\boldsymbol{X})\det(\boldsymbol{I}_M + \boldsymbol{Z}^T\boldsymbol{P}_{\boldsymbol{0}}\boldsymbol{Z}\boldsymbol{D}_{\boldsymbol{\lambda}})$.

References

Aerts, M., Claeskens, G. and Wand, M. P. (2002). Some theory for penalized spline additive models. *J. Statist. Plann. Inference* **103**, 455-470.

Anderssen, R. and Bloomfield, P. (1974). A time series approach to numerical differentiation. *Technometrics* **16**, 69-75.

Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica* **69**, 683-734.

Barry, D. (1983). Nonparametric Bayesian regression, Ph.D. thesis, Yale University, New Haven, Connecticut.

Berry, S. A., Carroll, R. J. and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.* **97**, 160-169.

Brumback, B. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.* **93**, 961-994.

Brumback, B., Ruppert, D. and Wand. M. P. (1999). Comment on "Variable selection and function estimation in additive nonparametric regression using data-based prior" by Shively, Kohn, and Wood. *J. Amer. Statist. Assoc.* **94**, 794-797.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25**, 573-578.

Crainiceanu, C. M. and Ruppert, D. (2004). Asymptotic distribution of likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Statist. Soc. Ser. B* **94**, 165-185.

Crainiceanu, C. M., Ruppert, D., Claeskens, G. and Wand, M. P. (2002). Likelihood ratio tests of polynomial regression against a general nonparametric alternative. Available at www.orie.cornell.edu/∼davidr/papers. *Biometrika*. To appear.

Crainiceanu, C. M., Ruppert, D. and Vogelsang, T. J. (2002). Probability that the MLE of a variance component is zero with applications to likelihood ratio tests. Available at www.orie.cornell.edu/∼davidr/papers.

Grizzle, J. E. and Allan, D. M. (1969). Analysis of dose and dose response curves. *Biometrics* **25**, 357-381.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72**, 320-338.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Kohn, R., Ansley, C. F. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Amer. Statist. Assoc.* **86**, 1042-1050.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Comput. Statist. Data Anal.* **28**, 193-209.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Powell, M. J. D. (1981). *Approximation Theory and Methods*. Cambridge University Press, Cambridge, UK.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11**, 735-757.

Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Austral. and N.Z. J. Statist.* **42**, 205-223.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge, UK.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Amer. Statist. Assoc.* **82**, 605-610.

Self, S. G. and Liang, K. Y. (1995). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J. Roy. Statist. Soc. Ser. B* **58**, 785-796.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171-1177.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254-262.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40**, 364-372.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378-1402.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *J. Roy. Statist. Soc. Ser. B* **60**, 159-174.

Wecker, W. and Ansley, C. (1983). The signal extraction approach to non-linear regression and spline smoothing. *J. Amer. Statist. Assoc.* **78**, 81-89.

Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. E3037 Baltimore, MD 21205 U.S.A.

E-mail: ccrainic@jhsph.edu

School of Operational Research and Industrial Engineering, Cornell University, Rhodes Hall, NY 14853, U.S.A.

E-mail: dr24@cornell.edu

(Received January 2003; accepted March 2004)