J. R. Statist. Soc. B (2004)
66, *Part* 1, *pp.* 165–185

Likelihood ratio tests in linear mixed models with one variance component

Ciprian M. Crainiceanu and David Ruppert

Cornell University, Ithaca, USA

[Received July 2002. Revised May 2003]

Summary. We consider the problem of testing null hypotheses that include restrictions on the variance component in a linear mixed model with one variance component and we derive the finite sample and asymptotic distribution of the likelihood ratio test and the restricted likelihood ratio test. The spectral representations of the likelihood ratio test and the restricted likelihood ratio test statistics are used as the basis of efficient simulation algorithms of their null distributions. The large sample χ^2 mixture approximations using the usual asymptotic theory for a null hypothesis on the boundary of the parameter space have been shown to be poor in simulation studies. Our asymptotic calculations explain these empirical results. The theory of Self and Liang applies only to linear mixed models for which the data vector can be partitioned into a large number of independent and identically distributed subvectors. One-way analysis of variance and penalized splines models illustrate the results.

Keywords: Degrees of freedom; Non-regular problems; Penalized splines

1. Introduction

We consider the problem of testing null hypotheses that include constraints on the variance component in a linear mixed model (LMM) with one variance component. Our main result is the derivation of the finite sample distributions of the likelihood ratio test (LRT) and restricted likelihood ratio test (RLRT). The spectral decomposition of the LRT and RLRT statistics are used to explain the finite sample behaviour that is observed in empirical studies, e.g. Pinheiro and Bates (2000). For the important special case of the null hypothesis that the variance component is 0, we derive the asymptotic distributions of the LRT and RLRT statistics under weak assumptions on the eigenvalues of certain design matrices. We also derive the finite sample and asymptotic probabilities that the estimated variance component is 0 or, equivalently, that the LRT or RLRT is 0.

One possible application is testing for a level (or group) effect in a mixed balanced one-way analysis-of-variance (ANOVA) model. We use this example to compare our results with 'standard' asymptotic results derived for testing null hypotheses on the boundary of the parameter space.

Another application is using LRTs and RLRTs for testing a polynomial fit against a general alternative described by penalized splines (P-splines). As shown in Section 5, this is equivalent to testing whether the variance component of a particular LMM is 0. Our LRT and RLRT approach of using nonparametric regression to test for general departures from a polynomial model is part of a rich literature using kernel estimation (e.g. Azzalini and Bowman (1993)),

Address for correspondence: Ciprian M. Crainiceanu, Department of Statistics, 301 Malott Hall, Cornell University, Ithaca, NY 14853, USA.
E-mail: cmc59@cornell.edu

 1369–7412/04/66165

P-splines (e.g. Hastie and Tibshirani (1990) and Ruppert *et al.* (2003)) or local polynomial regression (e.g. Cleveland and Devlin (1988)).

A third application of our work is testing for a fixed smoothing parameter in a P-spline regression. This is equivalent to testing a fixed number of degrees of freedom of the regression against a general alternative. This application is important because it can be used to derive confidence intervals for a variance component or smoothing parameter.

Our results are different from those derived by Self and Liang (1987), Liang and Self (1996) and Stram and Lee (1994) under the restrictive assumption that the response variable vector can be partitioned into independent and identically distributed (IID) subvectors and the number of independent subvectors tends to ∞. Self and Liang (1987) and Liang and Self (1996) explicitly stated that the data are IID for all values of the parameter (see their introduction). Stram and Lee (1994) assumed that random effects are independent from subject to subject and they implicitly assumed that the number of subjects increases to ∞. Their results would not hold for a fixed number of subjects, even if the number of observations per subject increased to ∞. Feng and McCulloch (1992) showed that for IID data (see their theorems 2.2 and 2.3) the LRT has classical asymptotic properties on an enlarged parameter space. Our results are also different from the results in Andrews (2001), derived for the random-coefficients model.

Consider an LMM with one variance component

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\varepsilon}, \quad E\begin{pmatrix}\mathbf{b}\\ \boldsymbol{\varepsilon}\end{pmatrix} = \begin{pmatrix}\mathbf{0}_K\\ \mathbf{0}_n\end{pmatrix}, \quad \operatorname{cov}\begin{pmatrix}\mathbf{b}\\ \boldsymbol{\varepsilon}\end{pmatrix} = \begin{pmatrix}\sigma_b^2\Sigma & 0\\ 0 & \sigma_\varepsilon^2 I_n\end{pmatrix}, \tag{1}$$

where $\mathbf{0}_K$ is a K-dimensional column of 0s, Σ is a known symmetric positive definite $K \times K$ matrix, $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects parameters, $\mathbf{b}$ is a K-dimensional vector of random effects and $(\mathbf{b}, \boldsymbol{\varepsilon})$ has a normal distribution. Under these conditions it follows that

$$E(\mathbf{Y}) = X\boldsymbol{\beta},$$
$$\operatorname{cov}(\mathbf{Y}) = \sigma_\varepsilon^2 V_\lambda,$$

where $\lambda = \sigma_b^2/\sigma_\varepsilon^2$, $V_\lambda = I_n + \lambda Z\Sigma Z^{\mathrm{T}}$ and n is the length of $\mathbf{Y}$. The parameter λ can be considered a signal-to-noise ratio since σ_b^2 determines the size of the 'signal' given by $E(\mathbf{Y}|\mathbf{b}) = X\boldsymbol{\beta} + Z\mathbf{b}$. Note that $\sigma_b^2 = 0$ if and only if $\lambda = 0$ and the parameter space for λ is $[0, \infty)$. Consider testing the null hypothesis

$$H_0 : \beta_{p+1-q} = \beta_{p+1-q}^0, \ldots, \beta_p = \beta_p^0, \qquad \sigma_b^2 = 0 \text{ (or equivalently } \lambda = 0), \tag{2}$$

against the composite alternative

$$H_{\mathrm{A}} : \beta_{p+1-q} \neq \beta_{p+1-q}^0 \text{ or}, \ldots, \text{ or } \beta_p \neq \beta_p^0, \text{ or } \sigma_b^2 > 0 \text{ (or equivalently } \lambda > 0), \tag{3}$$

for $q > 0$. In Section 5 we show that testing the null hypothesis of a $(p-q)$-degree polynomial against a general alternative modelled by a p-degree spline is a particular case of testing hypothesis (2) *versus* hypothesis (3). If $q = 0$ then we have the important particular case of testing that the variance component is 0:

$$H_0 : \sigma_b^2 = 0 \ (\lambda = 0) \quad \textit{versus} \quad H_{\mathrm{A}} : \sigma_b^2 > 0 \ (\lambda > 0). \tag{4}$$

Testing the null hypothesis (2) against the alternative hypothesis (3) is non-standard because the parameter under the null hypothesis is on the boundary of the parameter space and also because the response variables are not independent under the alternative.

A generalization of expression (4) is

$$H_0 : \lambda = \lambda_0 \quad \textit{versus} \quad H_A : \lambda \in \Lambda \subset [0, \infty), \tag{5}$$

where Λ can be any subset of $[0, \infty)$. We discuss the following cases of Λ: $\{\lambda_1\}$ for fixed $\lambda_1 \neq \lambda_0$, (λ_0, ∞) and $[0, \infty)\backslash\{\lambda_0\}$. In Section 6 we show that testing for a fixed smoothing parameter (or equivalently number of degrees of freedom) of a penalized spline regression is equivalent to testing expression (5). Tests of this type can be inverted to create confidence intervals for λ.

Under the assumption of IID data, Stram and Lee (1994) proved that the LRT for testing the null hypothesis (2) against the alternative hypothesis (3) has a $0.5\chi^2_q + 0.5\chi^2_{q+1}$ asymptotic distribution, where q is the number of fixed effects parameters constrained under H_0. However, the hypothesis of IID data does not hold for most LMMs. Even for the simple one-way balanced ANOVA model the assumption only holds when the number of observations per level is fixed and the number of levels tends to ∞. For the null hypothesis (4), $q = 0$, Crainiceanu, Ruppert and Vogelsang (2003) calculated the finite sample probability mass at 0 of the LRT (and RLRT), proving that, both for simple (one-way ANOVA) and more complex (P-spline regression) models, the $0.5\chi^2_q + 0.5\chi^2_{q+1}$ asymptotic approximations can be very poor.

By finding the null finite sample distributions of the LRT (and RLRT) for testing the hypotheses (2) against hypotheses (4) or (5), we provide an appealing practical testing methodology. The advantage of our work over Stram and Lee's (1994) results is that the hypothesis of IID data is eliminated, which changes the null distribution theory compared with Stram and Lee's. These changes can be severe. In Section 6 we discuss the RLRTs for hypotheses (5). Our results extend previous work by Shephard and Harvey (1990), Shephard (1993) and Kuo (1999) for regression with a stochastic trend. In a related paper Stern and Welsh (2000) provided local asymptotic approximations to construct confidence intervals for the components of the variance that are close to the boundary of the parameter space in LMMs.

Proofs of results in this paper are provided in Crainiceanu and Ruppert (2003).

2. Finite sample distribution of likelihood ratio test and restricted likelihood ratio test

For simplicity, we first focus on testing the null hypothesis (2) by using the LRT. Similar reasoning holds for the RLRT. Twice the log-likelihood function for model (1) is

$$2\log\{L(\boldsymbol{\beta}, \lambda, \sigma^2_\varepsilon)\} = -\log(\sigma^2_\varepsilon) - \log|V_\lambda| - \frac{(\mathbf{Y} - X\boldsymbol{\beta})^{\mathrm{T}} V_\lambda^{-1} (\mathbf{Y} - X\boldsymbol{\beta})}{\sigma^2_\varepsilon} \tag{6}$$

and the LRT is defined as

$$\mathrm{LRT}_n = 2\sup_{H_A}\{L(\boldsymbol{\beta}, \lambda, \sigma^2_\varepsilon)\} - 2\sup_{H_0}\{L(\boldsymbol{\beta}, \lambda, \sigma^2_\varepsilon)\}.$$

Under the alternative hypothesis, by fixing λ and solving the first-order maximum conditions for $\boldsymbol{\beta}$ and σ^2_ε, we obtain the profile likelihood estimates

$$\hat{\boldsymbol{\beta}}(\lambda) = (X^{\mathrm{T}} V_\lambda^{-1} X)^{-1} X^{\mathrm{T}} V_\lambda^{-1} \mathbf{Y},$$

$$\hat{\sigma}^2_\varepsilon(\lambda) = \frac{\{\mathbf{Y} - X\,\hat{\boldsymbol{\beta}}(\lambda)\}^{\mathrm{T}} V_\lambda^{-1} \{\mathbf{Y} - X\,\hat{\boldsymbol{\beta}}(\lambda)\}}{n}.$$

Plugging these expressions into equation (6) we obtain (up to a constant that does not depend on the parameters) the profile log-likelihood function

$$L^{K,n}(\lambda) = -\log|V_\lambda| - n\log(\mathbf{Y}^{\mathrm{T}} P_\lambda^{\mathrm{T}} V_\lambda^{-1} P_\lambda \mathbf{Y}),$$

where

$$P_\lambda = I_n - X(X^{\mathrm{T}} V_\lambda^{-1} X)^{-1} X^{\mathrm{T}} V_\lambda^{-1}.$$

Under the null hypothesis the model becomes a standard linear regression. If X_1 is the matrix formed with the first $p-q$ columns of X and

$$S_1 = I_n - X_1(X_1^{\mathrm{T}} X_1)^{-1} X_1^{\mathrm{T}}$$

the LRT statistic is

$$\mathrm{LRT}_n = \sup_{\lambda \geqslant 0}\{n\log(\mathbf{Y}^{\mathrm{T}} S_1 \mathbf{Y}) - n\log(\mathbf{Y}^{\mathrm{T}} P_\lambda^{\mathrm{T}} V_\lambda^{-1} P_\lambda \mathbf{Y}) - \log|V_\lambda|\}.$$

The following theorem gives the spectral decomposition of the LRT_n statistic for testing the null hypothesis (2) against the alternative hypothesis (3). Define

$$f_n(\lambda) = n\log\left\{1 + \frac{N_n(\lambda)}{D_n(\lambda)}\right\} - \sum_{s=1}^{K} \log(1+\lambda\xi_{s,n}),$$

where $N_n(\lambda)$ and $D_n(\lambda)$ are defined in theorem 1.

Theorem 1. If $\mu_{s,n}$ and $\xi_{s,n}$ are the K eigenvalues of the $K \times K$ matrices $\Sigma^{1/2} Z^{\mathrm{T}} P_0 Z \Sigma^{1/2}$ and $\Sigma^{1/2} Z^{\mathrm{T}} Z \Sigma^{1/2}$ respectively, where $P_0 = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, then

$$\mathrm{LRT}_n \stackrel{\mathcal{D}}{=} n\left(1 + \sum_{s=1}^{q} u_s^2 \Big/ \sum_{s=1}^{n-p} w_s^2\right) + \sup_{\lambda \geqslant 0}\{f_n(\lambda)\}, \tag{7}$$

where u_s for $s = 1, \ldots, K$ and w_s for $s = 1, \ldots, n-p$ are independent $N(0,1)$, the notation $=^{\mathcal{D}}$ denotes equality in distribution and

$$N_n(\lambda) = \sum_{s=1}^{K} \frac{\lambda\mu_{s,n}}{1+\lambda\mu_{s,n}} w_s^2,$$

$$D_n(\lambda) = \sum_{s=1}^{K} \frac{w_s^2}{1+\lambda\mu_{s,n}} + \sum_{s=K+1}^{n-p} w_s^2.$$

The distribution described in equation (7) depends only on the eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$ of two $K \times K$ matrices. Once they have been calculated, simulations from this distribution can be done rapidly, much faster than direct bootstrapping which would not take advantage of the spectral decomposition (7). The following algorithm provides a simple way to simulate the null finite sample distribution of LRT_n.

Step 1: define a grid $0 = \lambda_1 < \lambda_2 < \ldots < \lambda_m$ of possible values for λ.
Step 2: simulate K independent χ_1^2 random variables $w_1^2, \ldots, w_K^2$. Set $S_K = \Sigma_{s=1}^{K} w_s^2$.
Step 3: independently of step 1, simulate $X_{n,K,p} = \Sigma_{s=K+1}^{n-p} w_s^2$ with a χ^2_{n-p-K}-distribution.
Step 4: independently of steps 1 and 2, simulate $X_q = \Sigma_{s=1}^{q} u_s^2$ with a χ_q^2-distribution.
Step 5: for every grid point λ_i compute

$$N_n(\lambda_i) = \sum_{s=1}^{K} \frac{\lambda_i \mu_{s,n}}{1 + \lambda_i \mu_{s,n}} w_s^2,$$

$$D_n(\lambda_i) = \sum_{s=1}^{K} \frac{w_s^2}{1 + \lambda_i \mu_{s,n}} + X_{n,K,p}.$$

Step 6: determine $\lambda_{\max}$ which maximizes $f_n(\lambda_i)$ over $\lambda_1, \ldots, \lambda_m$.
Step 7: compute

$$\mathrm{LRT}_n = f_n(\lambda_{\max}) + n \log\left(1 + \frac{X_q}{S_K + X_{n,K,p}}\right).$$

Step 8: repeat steps 2–7 until the desired number of simulations is achieved.

An important feature of the algorithm is that the eigenvalues need to be computed only once before simulation begins and its speed depends on the number of random effects K but not on the number of observations n. As an example, for $K = 20$ we obtained 5000 simulations per second (with a 2.66 GHz central processor unit and 1 Mbyte random access memory) using efficient matrix manipulations in MATLAB (Hanselman and Littlefield, 2000). The algorithm remains feasible as long as we can obtain the eigenvalues of $K \times K$ matrices and the simulation time remains of the order of seconds for K as large as 500. A similar algorithm can be given for RLRT_n by using the spectral representation (9) below.

Using a grid of values for λ as in step 1 provides a discrete and bounded approximation for the true distribution of $\hat{\lambda}_{\mathrm{ML}}$. Because for the null distributions of interest $\hat{\lambda}_{\mathrm{ML}} = 0$ or is very close to 0, the grid needs to be very fine in a neighbourhood of 0 but can be rougher for larger values. For $\lambda_i > 0$ we used 200 grid points equally spaced on the natural log-scale between $[-12, 12]$. Using a maximization algorithm with linear constraints ($\lambda \geqslant 0$) we also simulated the exact null distribution. This algorithm was slower than the algorithm that is proposed in this paper but the results were practically indistinguishable, indicating that our choice of grid provides an accurate approximation.

The spectral representations (7) and (9) below can be used to compute the probability mass at 0 of the LRT and RLRT statistic for testing that the variance component is 0 with no constraints on the fixed effects ($q = 0$ and $\sigma_b^2 = 0$). The first-order condition for $f_n(\lambda)$ to have a local maximum at $\lambda = 0$ is

$$\left.\frac{\partial f_n(\lambda)}{\partial \lambda}\right|_{\lambda=0} \leqslant 0.$$

It follows that the probability of having a local maximum of the profile likelihood at $\lambda = 0$ is

$$\mathrm{pr}\left(\frac{\sum_{s=1}^{K} \mu_{s,n} w_s^2}{\sum_{s=1}^{n-p} w_s^2} \leqslant \frac{1}{n} \sum_{s=1}^{K} \xi_{s,n}\right). \tag{8}$$

This is the exact probability of a local maximum at 0 and provides an excellent approximation for the probability of a global maximum at $\lambda = 0$ (Crainiceanu, Ruppert and Vogelsang, 2003). This probability can be easily obtained by simulation.

Because the finite sample distribution of the LRT and RLRT statistics can be simulated so easily using theorem 1, there is no practical need for asymptotic results. However, since practitioners will be tempted to use 'standard' χ^2 mixture asymptotic approximations, it is important to study the accuracy of these approximations. This is done in this paper, and the

accuracy is found often to be poor. From theorem 1 it follows that the finite sample distribution of LRT_n depends on the eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$. In Section 3 we show that the asymptotic behaviour of the LRT_n distribution depends essentially on the asymptotic behaviour of these eigenvalues. In Section 7 we discuss the relationship between the types of spectra and distribution theory.

Residual, or restricted, maximum likelihood (REML) was introduced by Patterson and Thompson (1971) to take into account the loss in degrees of freedom due to estimation of β-parameters. REML consists of maximizing the likelihood function that is associated with $n - p$ linearly independent contrasts, and the log-likelihood function for any such set of contrasts differs by no more than an additive constant (Harville, 1977). Twice the restricted profile log-likelihood function is (Harville, 1977)

$$2\,l^{K,n}(\lambda) = -\log|V_\lambda| - \log|X^{\mathrm{T}}V_\lambda^{-1}X| - (n-p)\log(\mathbf{Y}^{\mathrm{T}}P_\lambda^{\mathrm{T}}V_\lambda^{-1}P_\lambda\mathbf{Y}).$$

The RLRT_n is defined like the LRT_n by using the restricted likelihood instead of the likelihood function. Because the RLRT_n uses the likelihood of residuals after fitting the fixed effects, the RLRT is appropriate for testing only when the fixed effects are the same under the null and alternative hypotheses. Therefore the RLRT_n will be used only when the number of fixed effects constrained under H_0 is $q = 0$ and we test for $\sigma_b^2 = 0$ only. Then, under the null hypothesis described in equation (4),

$$\mathrm{RLRT}_n \stackrel{\mathcal{D}}{=} \sup_{\lambda \geqslant 0}\left[(n-p)\log\left\{1 + \frac{N_n(\lambda)}{D_n(\lambda)}\right\} - \sum_{s=1}^{K}\log(1+\lambda\mu_{s,n})\right], \tag{9}$$

and the probability of having a local maximum at $\lambda = 0$ is

$$\mathrm{pr}\left(\frac{\sum_{s=1}^{K}\mu_{s,n}w_s^2}{\sum_{s=1}^{n-p}w_s^2} \leqslant \frac{1}{n-p}\sum_{s=1}^{K}\mu_{s,n}\right),$$

where $w_1, \ldots, w_K$ are independent $N(0,1)$ random variables. The notation here is the same as in theorem 1. An efficient simulation algorithm for the finite sample distribution of RLRT_n can be easily obtained by direct analogy with the LRT_n case.

3. Asymptotic results

This section presents the asymptotic distributions of the LRT_n and RLRT_n statistics for testing the null hypotheses (2) and (4) respectively. Because the finite sample results in Section 2 depend essentially on the eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$ we may expect to have a relationship between the asymptotic distributions of test statistics and the asymptotic behaviour of these eigenvalues. The following theorem provides the formal description of this relationship.

Theorem 2. Assume that hypothesis H_0 in expression (2) is true. Suppose that there is an $\alpha \geqslant 0$ so that for every s the K eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$ of matrices $\Sigma^{1/2}Z^{\mathrm{T}}P_0Z\Sigma^{1/2}$ and $\Sigma^{1/2}Z^{\mathrm{T}}Z\Sigma^{1/2}$ respectively satisfy $\lim_{n\to\infty}(n^{-\alpha}\mu_{s,n}) = \mu_s$ and $\lim_{n\to\infty}(n^{-\alpha}\xi_{s,n}) = \xi_s$, where not all ξ_s are 0. Then

$$\mathrm{LRT}_n \Rightarrow \sum_{s=1}^{q}u_s^2 + \sup_{d\geqslant 0}\{\mathrm{LRT}_\infty(d)\},$$

where the two terms in the asymptotic distribution are independent, u_s, w_s are IID $N(0,1)$ and

$$\mathrm{LRT}_\infty(d) = \sum_{s=1}^{K} \frac{d\mu_s}{1+d\mu_s} w_s^2 - \sum_{s=1}^{K} \log(1+d\xi_s).$$

Here '$\Rightarrow$' denotes weak convergence. The proof of this result is based on the weak convergence of the profile LRT in the space $\mathcal{C}[0,\infty)$ of continuous functions on $[0,\infty)$ and a continuous mapping theorem. The first part of the asymptotic distribution is a χ^2_q-distribution corresponding to testing for q fixed effects and the second part corresponds to testing for $\sigma_b^2 = 0$.

Asymptotic theory for a null hypothesis on the boundary of the parameter space developed for IID data suggests the asymptotic result

$$\mathrm{LRT}_n \Rightarrow \sum_{s=1}^{q} u_s^2 + U_+^2,$$

where U is an $N(0,1)$ random variable independent of all u_s and $U_+^2 = U^2\, I(U>0)$ with $I(\cdot)$ the indicator function. The distribution of U_+^2 is a $0.5\chi_0^2{:}0.5\chi_1^2$ mixture between a χ_0^2- (Dirac distribution at 0) and a χ_1^2-distribution. In contrast, the distribution of $\sup_{d\geqslant 0}\{\mathrm{LRT}_\infty(d)\}$ depends essentially on the asymptotic eigenvalues μ_s and ξ_s and is generally different from the $0.5\chi_0^2{:}0.5\chi_1^2$ mixture, and differences can be severe. In one example, Crainiceanu, Ruppert and Vogelsang (2003) showed that $\sup_{d\geqslant 0}\{\mathrm{LRT}_\infty(d)\}$ is essentially the Dirac measure at 0 when testing for a linear model against a general alternative modelled by a penalized spline. Obviously, the LRT will have little power because of this behaviour, but asymptotic theory for IID data gives no indication of the lack of power. Crainiceanu, Ruppert and Vogelsang (2003) showed that the RLRT_n does not have this problem of a near degenerate asymptotic null distribution and suggested using the RLRT to increase the power.

Under the same assumptions as in theorem 2, if the null hypothesis H_0 in equation (4) is true then

$$\mathrm{RLRT}_n \Rightarrow \sup_{d\geqslant 0}\{\mathrm{RLRT}_\infty(d)\},$$

where

$$\mathrm{RLRT}_\infty(d) = \sum_{s=1}^{K} \frac{d\mu_s}{1+d\mu_s} w_s^2 - \sum_{s=1}^{K} \log(1+d\mu_s).$$

If $q = 0$, then the LRT and RLRT statistics have probability mass at 0. Crainiceanu, Ruppert and Vogelsang (2003) showed that this mass can be very large for LRT_∞ and RLRT_∞ and is equal to the null probability that $\mathrm{LRT}_\infty(\cdot)$ and $\mathrm{RLRT}_\infty(\cdot)$ have a global maximum at 0. The latter is well approximated by the null probability of having a local maximum at $\lambda = 0$. The first-order conditions for a local maximum at $\lambda = 0$ are

$$\frac{\partial}{\partial d}\mathrm{LRT}_\infty(d)\big|_{d=0} \leqslant 0$$

or

$$\frac{\partial}{\partial d}\mathrm{RLRT}_\infty(d)\big|_{d=0} \leqslant 0.$$

Therefore, the probability of having a local maximum at $d = 0$ for $\mathrm{LRT}_\infty(\cdot)$ or $\mathrm{RLRT}_\infty(\cdot)$ is

$$\mathrm{pr}\left(\sum_{s=1}^{K} \mu_s w_s^2 \leqslant \sum_{s=1}^{K} \xi_s\right)$$

or

$$\mathrm{pr}\left(\sum_{s=1}^{K}\mu_s w_s^2 \leqslant \sum_{s=1}^{K}\mu_s\right),$$

where w_s are IID $N(0,1)$ random variables.

Although the asymptotic distributions are not needed to approximate the finite sample distributions, two conclusions follow from the results of this section. The first is that the usual boundary asymptotic theory for IID data does not apply to testing a hypothesis about the variance component. The second is that the asymptotic distribution depends on the model through the eigenvalues μ_s and ξ_s.

In Sections 4 and 5 we investigate further the differences between standard boundary asymptotic theory for IID data and our results. We also compare finite sample and asymptotic distributions of LRT and RLRT statistics.

4. Balanced one-way analysis of variance

Consider the balanced one-way ANOVA model with K levels and J observations per level

$$Y_{kj} = \mu + b_k + \varepsilon_{kj}, \qquad k = 1,\ldots,K \text{ and } j = 1,\ldots,J,$$

where ε_{kj} are IID $N(0,\sigma_\varepsilon^2)$, b_k are IID random effects distributed $N(0,\sigma_b^2)$ independent of the ε_{kj} and μ is a fixed unknown intercept. Define $\lambda = \sigma_b^2/\sigma_\varepsilon^2$. The matrix X for fixed effects is simply a $JK \times 1$ column of 1s, the matrix Z is a $JK \times K$ matrix with every column containing only 0s with the exception of a J-dimensional vector of 1s corresponding to the level parameter and the matrix Σ is the identity matrix I_K. The size of the response vector $\mathbf{Y}$ is $n = JK$.

Consider the test for $\sigma_b^2 = 0$. To find the finite sample distributions of the LRT_n or RLRT_n we need to determine the eigenvalues of $Z^{\mathrm{T}}Z$ and $Z^{\mathrm{T}}P_0Z$. In this simple model we can find them explicitly. All K eigenvalues of the matrix $Z^{\mathrm{T}}Z$ are equal and $\xi_{s,n} = J$. Also, one eigenvalue of the matrix $Z^{\mathrm{T}}P_0Z$ is equal to 0 and the remaining $K-1$ eigenvalues are equal and $\mu_{s,n} = J$. Using theorem 1 it follows that

$$\mathrm{LRT}_n \overset{\mathcal{D}}{=} n\log(X_{K-1} + X_{n-K}) - \inf_{d\geqslant 0}\left\{n\log\left(\frac{X_{K-1}}{1+d} + X_{n-K}\right) + K\log(1+d)\right\},$$

where X_{K-1} and X_{n-K} are independent random variables with distributions χ^2_{K-1} and χ^2_{n-K} respectively. This distribution can be obtained explicitly or simulated by using a simpler version of the algorithm in Section 2. The finite sample probability mass at 0 given by equation (8) is

$$\mathrm{pr}\{F_{K-1,n-K} \leqslant K/(K-1)\},$$

where $F_{K-1,n-K}$ has an F-distribution with $(K-1, n-K)$ degrees of freedom. The probability mass at 0 for this special case is similar to that obtained by Searle *et al.* (1992), page 137, for the ANOVA estimator of σ_b^2.

To obtain the asymptotic distribution when $J \to \infty$ and K is constant note that if $\alpha = 1$ then

$$\begin{aligned} n^{-1}\xi_{s,n} &\to 1/K, \qquad s = 1,\ldots,K,\\ n^{-1}\mu_{s,n} &\to 1/K, \qquad s = 1,\ldots,K-1,\\ n^{-1}\mu_{s,K} &\to 0. \end{aligned}$$

(In each case, '$\to$' is, in fact, equality.) Using these expressions for μ_s and ξ_s in theorem 2 the following result holds:

$$\mathrm{LRT}_n \Rightarrow \{X_{K-1} - K - K\log(X_{K-1}/K)\}I(X_{K-1} > K), \tag{10}$$

where X_{K-1} denotes a random variable with a χ^2_{K-1}-distribution. This asymptotic distribution has mass at 0 equal to $\mathrm{pr}(X_{K-1} < K)$. The usual non-standard asymptotic results require K to increase to ∞, with J either fixed or also increasing to ∞, so that $\mathrm{pr}(X_{K-1} < K) \to 0.5$.

The balanced one-way ANOVA model is one of the few possible cases when the observed response vector $\mathbf{Y}$ can be partitioned into a large number of IID subvectors corresponding to each level. Moreover, both the finite sample and the asymptotic distributions can be obtained explicitly. Therefore, it may be interesting to compare these distributions with the $0.5\chi^2_0{:}0.5\chi^2_1$ mixture, which is the asymptotic distribution when $K \to \infty$, with J either fixed or also increasing to ∞.

Fig. 1 displays QQ-plots of the $0.5\chi^2_0{:}0.5\chi^2_1$ mixture *versus* the asymptotic distribution of the LRT_n for a fixed number of levels, $K = 5$, and two finite sample distributions corresponding to $n = 50$ ($J = 10$ observations per level) and $n = 100$ ($J = 20$ observations per level). The finite sample distributions converge quickly to the K-fixed asymptotic distribution that is described in expression (10) and away from the $0.5\chi^2_0{:}0.5\chi^2_1$ distribution. The $0.5\chi^2_0{:}0.5\chi^2_1$ mixture is a conservative approximation to this asymptotic distribution. For example, the quantile corresponding to probability 0.99 is 5.41 for the $0.5\chi^2_0{:}0.5\chi^2_1$ distribution and 3.48 for the asymptotic distribution.

The difference between the $0.5\chi^2_0{:}0.5\chi^2_1$ asymptotic distribution and either the K-fixed asymptotic distribution or the finite sample distribution is due both to different probability masses at 0 and to differences in the non-zero parts of the distributions. For $K = 5, 10, 20, 100$, the probability mass at 0 of the LRT statistic is 0.71, 0.65, 0.61, 0.55 respectively, showing that, unless the number of levels K is very large, the 0.5 asymptotic value is inaccurate.

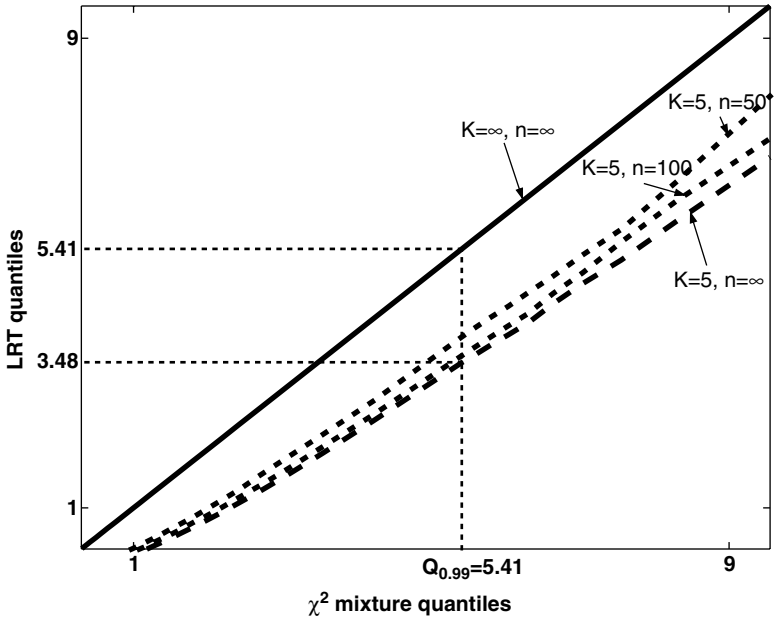

Fig. 1. *QQ* -plots for comparing quantiles of the $0.5\chi^2_0{:}0.5\chi^2_1$ mixture distribution (horizontal axis) with quantiles of the LRT_n null distributions when testing for level or subject effect in a balanced one-way ANOVA with $K = 5$ subjects: ———, 45° line corresponding to the $0.5\chi^2_0{:}0.5\chi^2_1$ distribution; -------, finite sample distributions for $n = 50$ and $n = 100$; – – –, asymptotic distribution

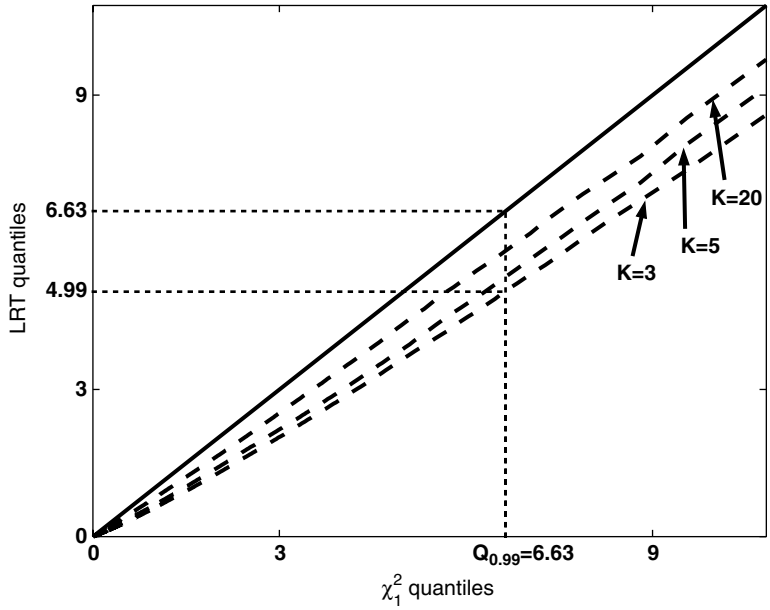

Fig. 2. *QQ* -plots for comparing the χ^2_1-distribution (horizontal axis) with the asymptotic distributions of LRT_n conditional on $\text{LRT}_n > 0$ (equation (10)) for a balanced one-way ANOVA model: ———, χ^2_1-distribution; - - - - - - -, three levels in the ANOVA model, $K = 3, 5, 20$

Fig. 2 shows QQ-plots of the χ^2_1-distribution *versus* the K-fixed asymptotic distribution of LRT_n conditional on $\text{LRT}_n > 0$ for $K = 3, 5, 20$. The quantiles were obtained by using 1 million simulations from these conditional distributions. When $K \to \infty$, the conditional distributions converge to the χ^2_1-distribution. However, for moderately large K, χ^2_1 is conservative relative to these distributions. For example, for $K = 3$ the 0.99-quantile for the χ^2_1-distribution is 6.63 and for the LRT distribution it is 4.99. Because the curves in Fig. 2 are nearly straight lines, the asymptotic distribution LRT_n can be closely approximated by a scalar multiple of a χ^2_1 random variable, with the scalar depending on K and increasing to 1 as $K \to \infty$.

Similar results, but with less severe differences, can be obtained for the RLRT_n statistic. For example, the asymptotic distribution is

$$\text{RLRT}_n \Rightarrow [X_{K-1} - (K-1) - (K-1)\log\{X_{K-1}/(K-1)\}]\, I(X_{K-1} > K-1), \tag{11}$$

and the asymptotic probability mass at 0 is $\text{pr}(X_{K-1} < K-1)$.

5. Nonparametric testing for polynomial regression against a general alternative

In this section we show that nonparametric regression by using penalized splines is equivalent to a particular LMM and that testing for a polynomial regression *versus* a general alternative can be viewed as testing for a zero-variance component in this LMM. We then compute the finite sample and asymptotic distribution of LRT and RLRT statistics in several important cases. Although we focus on penalized splines, the results are very general and can be used for any type of basis function (truncated polynomials, B-splines or trigonometric polynomials) and for any type of quadratic penalty.

5.1. P-splines regression and linear mixed models

Consider the regression equation

$$y_i = m(x_i) + \varepsilon_i,$$

where the ε_i are IID $N(0, \sigma_\varepsilon^2)$ and $m(\cdot)$ is the unknown mean function. Suppose that we are interested in testing whether $m(\cdot)$ is a $(p-q)$-degree polynomial:

$$H_0 : \ m(x) = \beta_0 + \beta_1 x + \ldots + \beta_{p-q} x^{p-q}. \tag{12}$$

To define an alternative that is sufficiently flexible to describe a large class of functions, we consider the class of splines

$$H_A : m(x) = m(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} b_k (x - \kappa_k)_+^p,$$

where $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_p, b_1, \ldots, b_K)^{\mathrm{T}}$ is the vector of regression coefficients, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of polynomial parameters, $\mathbf{b} = (b_1, \ldots, b_K)^{\mathrm{T}}$ is the vector of spline coefficients and $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ are fixed knots. Following Gray (1994) and Ruppert (2002), we consider a number of knots that is sufficiently large (e.g. 20) to ensure the desired flexibility. The knots can be taken to be equally spaced quantiles of the xs, i.e. the $1/(K+1), \ldots, K/(K+1)$ sample quantiles. To avoid overfitting, the criterion to be minimized is a penalized sum of squares:

$$\sum_{i=1}^{n} \{y_i - m(x_i; \boldsymbol{\theta})\}^2 + \frac{1}{\lambda} \boldsymbol{\theta}^{\mathrm{T}} L \boldsymbol{\theta}, \tag{13}$$

where $\lambda \geqslant 0$ is the smoothing parameter and L is a positive semidefinite matrix. The fitted function is called a P-spline. A common choice of L is

$$L = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma^{-1} \end{pmatrix},$$

where Σ is a known $K \times K$ positive definite matrix, with $\Sigma = I_K$ being a standard choice. Define $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, let X be the matrix having the ith row $\mathbf{X}_i = (1, x_i, \ldots, x_i^p)$, let Z be the matrix having the ith row $\mathbf{Z}_i = \{(x_i - \kappa_1)_+^p, \ldots, (x_i - \kappa_K)_+^p\}$ and define $\mathcal{X} = [X|Z]$. If criterion (13) is divided by σ_ε^2 we obtain

$$\frac{1}{\sigma_\varepsilon^2} \|\mathbf{Y} - X\boldsymbol{\beta} - Z\mathbf{b}\|^2 + \frac{1}{\lambda \sigma_\varepsilon^2} \mathbf{b}^{\mathrm{T}} \Sigma^{-1} \mathbf{b}. \tag{14}$$

Minimizing this expression shrinks the curve towards a pth degree polynomial fit. Define $\sigma_b^2 = \lambda \sigma_\varepsilon^2$, consider the vector $\boldsymbol{\beta}$ as unknown fixed parameters and consider the vector $\mathbf{b}$ as a set of random parameters with $E(b) = 0$ and $\mathrm{cov}(b) = \sigma_b^2 \Sigma$. If $(b^{\mathrm{T}}, \varepsilon^{\mathrm{T}})^{\mathrm{T}}$ is a normal random vector and if b and ε are independent, then we obtain an equivalent model representation of the P-spline in the form of an LMM (Brumback *et al.*, 1999). Specifically, the P-spline is equal to the best linear unbiased predictor (BLUP) of $X\boldsymbol{\beta} + Z\mathbf{b}$ in the LMM

$$\begin{aligned} Y &= X\beta + Zb + \varepsilon, \\ \mathrm{cov}\begin{pmatrix} b \\ \varepsilon \end{pmatrix} &= \begin{pmatrix} \sigma_b^2 \Sigma & 0 \\ 0 & \sigma_\varepsilon^2 I_n \end{pmatrix}. \end{aligned} \tag{15}$$

For this model $E(Y) = X\beta$ and $\mathrm{cov}(Y) = \sigma_\varepsilon^2 V_\lambda$, where $V_\lambda = I_n + \lambda Z \Sigma Z^{\mathrm{T}}$ and n is the total number of observations.

We shall consider the case $\Sigma = I_K$, penalizing the sums of squares of the jumps at the knots of the pth derivative of the fitted curve. Although the results hold for a general known symmetric positive definite penalty matrix Σ, the choice $\Sigma = I_K$ will be used to illustrate the results.

5.2. Tests for polynomial regression

We have transformed our problem of testing for a polynomial fit against a general alternative described by a P-spline to testing the null hypothesis

$$H_0 : \beta_{p+1-q} = 0, \ldots, \beta_p = 0, \quad \sigma_b^2 = 0 \ (\lambda = 0)$$

against the alternative

$$H_{\mathrm{A}} : \beta_{p+1-q} \neq 0 \text{ or }, \ldots, \text{ or } \beta_p \neq 0, \text{ or } \sigma_b^2 > 0 \ (\lambda > 0).$$

This is a particular case of the null and alternative hypotheses that are described in equations (2) and (3). Because the b_i have mean 0, $\sigma_b^2 = 0$ in H_0 is equivalent to the condition that all coefficients b_i of the truncated power functions are identically 0. These coefficients account for departures from a polynomial.

There is one very important point that we wish to emphasize. The assumption that the b_is are IID normal random variables is, of course, a convenient fiction that converts nonparametric regression into an LMM. There are legitimate concerns about making inferences based on this assumption. However, if the null hypothesis (12) that $m(\cdot)$ is a $(p-q)$-degree polynomial is true, then the b_i are all 0 and therefore they are, in fact, IID $N(0, \sigma_b^2)$ with $\sigma_b^2 = 0$. In other words, the LMM holds exactly under the null hypothesis (12). Therefore, a test that is exact within the LMM framework is, in fact, exact more generally for testing hypothesis (12) against a general alternative. Moreover, by using simulated quantiles from the finite sample null distribution of the LRT or RLRT, we do obtain exact tests.

As a first example, let us consider the problem of testing for a constant mean regression *versus* a general alternative. We shall model the alternative as a piecewise constant spline with K knots and test

$$H_0 : \ m(x) = \beta_0 \quad \textit{versus} \quad H_A : \ m(x) = \beta_0 + \sum_{k=1}^{K} b_k I(x > \kappa_k).$$

As shown in Section 2, the finite sample distribution of the LRT_n and RLRT_n statistics depends on the eigenvalues of the $K \times K$ matrices $Z^{\mathrm{T}} P_0 Z$ and $Z^{\mathrm{T}} Z$. P-splines use a moderately large number of knots K, with $K \leqslant 20$ in most applications and $K = 100$ being a rather extreme choice. With K in this range, these eigenvalues can be computed numerically by using an efficient matrix diagonalization algorithm, such as that implemented in MATLAB (function `eig`). We recommend the finite sample distribution because it is exact, easy to simulate and does not require additional assumptions.

Although finite sample distributions are recommended for practical testing problems, asymptotics are of interest, if for no other reason than to show the inaccuracy of 'standard' asymptotics assuming IID data. If we want the asymptotic distribution to approximate the finite sample distribution accurately, then K should be kept fixed at its actual value in a given application. Therefore, we are interested in asymptotic results when the number of observations n tends to ∞ and the number of knots K is kept constant. For this case, the asymptotic behaviour of matrices $Z^{\mathrm{T}} Z$ and $Z^{\mathrm{T}} P_0 Z$ was studied by Crainiceanu, Ruppert and Vogelsang (2003). If we denote by $n_s(n)$ the number of xs that are larger than the sth knot and assume that $n_s(n)/n \to p_s$ as $n \to \infty$ we can show that

$$\lim_{n\to\infty}(Z^{\mathrm{T}}Z/n) = R,$$
$$\lim_{n\to\infty}(Z^{\mathrm{T}}P_0Z/n) = M,$$

where the (i, j)th entries for matrices R and M are $r_{ij} = p_{\max(i,j)}$ and $m_{ij} = p_{\max(i,j)} - p_i p_j$ respectively. Denoting by $\xi_1, \ldots, \xi_K$ the eigenvalues of R and $\mu_1, \ldots, \mu_K$ the eigenvalues of M, it follows that

$$\lim_{n\to\infty}(n^{-1}\xi_{s,n}) = \xi_s,$$
$$\lim_{n\to\infty}(n^{-1}\mu_{s,n}) = \mu_s.$$

Substituting these values into $\mathrm{LRT}_\infty(d)$ of theorem 2, the asymptotic distributions of interest can easily be simulated. In particular, when the x-values are equally spaced and $K = 20$ equally spaced knots are used, the asymptotic probability mass at 0 is 0.65 for RLRT and 0.95 for LRT.

Fig. 3 shows QQ-plots for comparing quantiles of the $0.5\chi^2_0{:}0.5\chi^2_1$ mixture distribution (horizontal axis) with quantiles of the RLRT_n null distributions when testing for a constant mean *versus* a general alternative modelled by a piecewise constant spline with K-knots. We used the case of equally spaced xs in $[0, 1]$ with the kth knot being the empirical quantiles of the xs corresponding to probability $k/(K + 1)$. We consider $K = 20$, but similar results were obtained for other values of K. The full line is the 45° line corresponding to the $0.5\chi^2_0{:}0.5\chi^2_1$ mixture distribution, the dotted lines correspond to finite sample distributions for $n = 50$ and $n = 100$ and the broken line corresponds to the asymptotic distribution. 1 million simulations (taking approximately 3.5 min) were used for each distribution. These many simulations are generally not necessary, but we wanted to emphasize again the speed of the simulation algorithm and to ensure that extreme quantiles were estimated accurately.

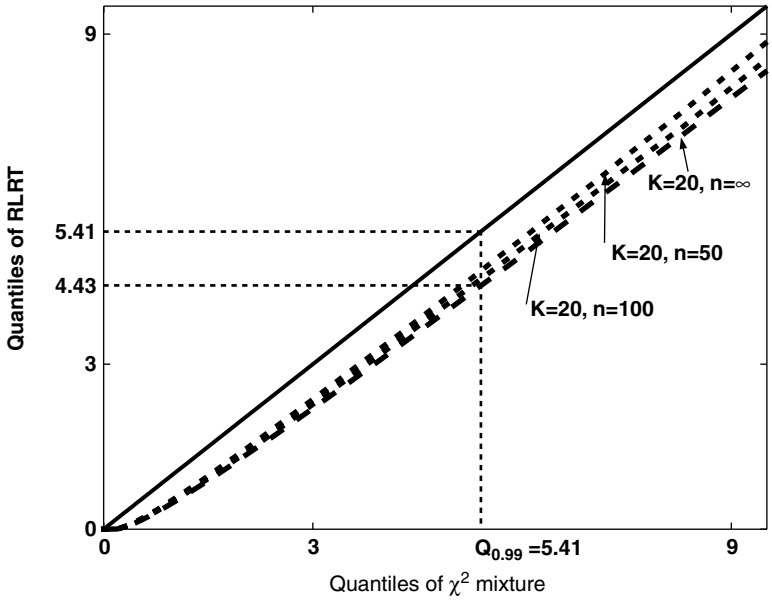

Fig. 3. *QQ* -plots for comparing quantiles of the $0.5\chi^2_0{:}0.5\chi^2_1$ mixture distribution (horizontal axis) with quantiles of the RLRT_n null distributions when testing for a constant mean *versus* a general alternative modelled by a piecewise constant spline with K knots (equally spaced xs in [0,1]; the knots are equally spaced quantiles of the xs; $K = 20$): ———, 45° line corresponding to the $0.5\chi^2_0{:}0.5\chi^2_1$ distribution: -------, finite sample distributions for $n = 50$ and $n = 100$; – – –, asymptotic distribution

The $0.5\chi^2_0 + 0.5\chi^2_1$ mixture distribution represents a conservative approximation to the finite sample and asymptotic distributions and using the 0.5:0.5 mixture can result in severe losses in power. Using an analogy with the ANOVA case, we may hope that if we increased the number of knots then the asymptotic distribution would tend to the 0.5:0.5 mixture. However, this is not so here, since the probability mass at 0 remains practically unchanged, 0.65, for K between 2 and 100.

We do not show QQ-plots for LRT_n because of the large probability mass at 0 (approximately 0.95) of the finite sample and asymptotic distributions, which makes the construction of a test impractical.

We now consider the problem of testing for a linear polynomial *versus* a general alternative modelled by a piecewise linear spline with K knots. Suppose that $x_i = i/(n+1)$, $i = 1, \ldots, n$, are equally spaced points in $[0, 1]$ and $\kappa_k = k/(K+1)$, $k = 1, \ldots, K$, are fixed equally spaced knots. We want to test

$$H_0 : m(x) = \beta_0 + \beta_1(x - \bar{x}) \quad \textit{versus} \quad H_A : m(x) = \beta_0 + \beta_1(x - \bar{x}) + \sum_{k=1}^{K} b_k(x - \kappa_k)_+.$$

As in the previous case, it is easy to obtain the finite sample distribution of the LRT_n and RLRT_n test by diagonalizing the corresponding matrices $Z^{\mathrm{T}} P_0 Z$ and $Z^{\mathrm{T}} Z$ and using the simulation algorithm described in Section 2. To obtain the asymptotic distribution note that

$$\lim_{n\to\infty} (Z^{\mathrm{T}} Z/n) = U,$$
$$\lim_{n\to\infty} (Z^{\mathrm{T}} P_0 Z/n) = V,$$

where both limit matrices are symmetric and for $k \geqslant l$ the (l, k)th entry of U is

$$u_{lk} = \tfrac{1}{3}(1 - \kappa_k^3) - \tfrac{1}{2}(\kappa_l + \kappa_k)(1 - \kappa_k^2) + \kappa_l \kappa_k (1 - \kappa_k),$$

and the (i, j)th entry of V is

$$v_{lk} = u_{lk} - \frac{1}{12}(1 - \kappa_l)^2 (1 - \kappa_k)^2 \{3 + (2\kappa_l + 1)(2\kappa_k + 1)\}.$$

Computing the eigenvalues of U and V is all that is needed to simulate the asymptotic distributions of RLRT and LRT. The mass at 0 of these distributions when K is 20 is 0.68 for RLRT and greater than 0.99 for LRT. Hence, the asymptotic distribution of the LRT is practically a point mass at 0, making the LRT impractical for this case. This is due to the downward bias of maximum likelihood variance estimation.

Simulations were used to obtain the finite sample and asymptotic distributions of RLRT_n. Similar patterns to those that are presented in Fig. 3 for testing for a constant mean were obtained for testing for a linear mean.

Under the same assumptions on the xs and knots, consider testing for a constant mean *versus* a general alternative modelled by a piecewise linear spline

$$H_0 : m(x) = \beta_0 \quad \textit{versus} \quad H_A : m(x) = \beta_0 + \beta_1(x - \bar{x}) + \sum_{k=1}^{K} b_k(x - \kappa_k)_+.$$

As in the previous cases, this can be reduced to testing

$$H_0 : \beta_1 = 0, \sigma_b^2 = 0 \ (\lambda = 0) \quad \textit{versus} \quad H_A : \beta_1 \neq 0, \text{ or } \sigma_b^2 > 0 \ (\lambda > 0).$$

Theorem 1 provides the finite sample distribution of the LRT_n statistic. Using theorem 2, the asymptotic distribution for LRT is $\sup_{d\geqslant 0}\{\mathrm{LRT}_\infty(d)\} + Z^2$, where $\sup_{d\geqslant 0}\{\mathrm{LRT}_\infty(d)\}$ is the asymptotic distribution for LRT for testing a linear polynomial against a piecewise linear spline and Z is a standard normal random variable. Because we have already proved that $\sup_{d\geqslant 0}\{\mathrm{LRT}_\infty(d)\}$ is practically the point mass at 0, we conclude that the asymptotic distribution for testing a constant mean against a piecewise linear spline (two constrained parameters under the null hypothesis) is practically a χ^2_1-distribution whereas the 'standard' IID asymptotic distribution is $0.5\chi^2_1 + 0.5\chi^2_2$. For the RLRT the finite sample and asymptotic distributions are different from a χ^2_1-distribution because $\sup_{d\geqslant 0}\{\mathrm{RLRT}_\infty(d)\}$ for testing a linear polynomial against a piecewise linear spline does not have the entire mass at 0. None-the-less, the standard $0.5\chi^2_1 + 0.5\chi^2_2$ approximation is not accurate for the RLRT.

5.3. Upper Cape Cod birth weight data

Fig. 4 shows the child birth weight for all 1630 births in 1990 across five towns in the Upper Cape Cod region of Massachusetts, USA: Barnstable, Bourne, Falmouth, Mashpee and Sandwich. Birth weight is sensitive to recent exposures, thus facilitating the determination of exposures of biological importance for human health. The predictor variable is maternal age. These data were obtained as part of a study into geographical variation of health outcomes that was commissioned in the late 1990s by the Department of Public Health, Commonwealth of Massachusetts (Ruppert *et al.*, 2003). Fig. 4 also shows the no-effect and the maximum likelihood linear P-splines fits ($K = 20$). The maximum likelihood estimator is close to the overall mean (the smoothing parameter was estimated to be 0).

We would like to test whether the maternal age has no effect on the child birth weight. In this case the hypotheses are

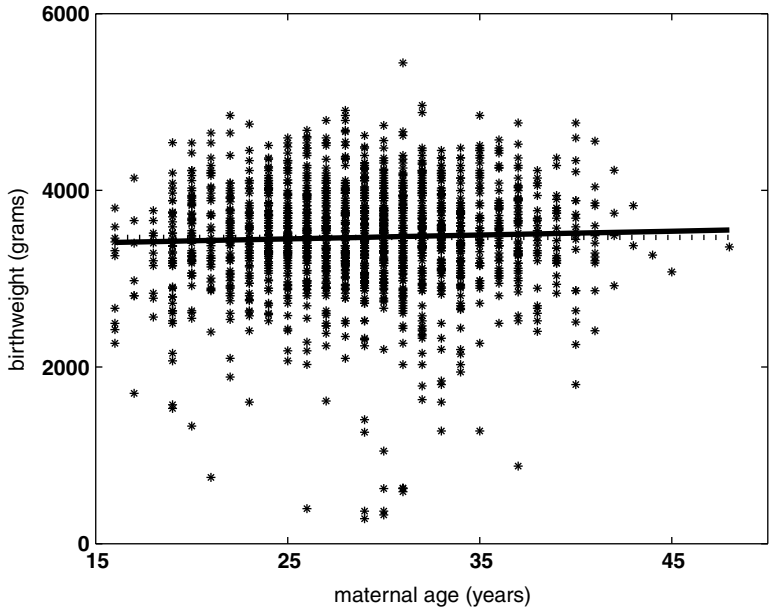

Fig. 4. Child birth weight *versus* maternal age, linear penalized splines estimators with $K = 20$ knots of the mean function by using maximum likelihood: $*$, data; ———, maximum likelihood fit; · · · · · · · , constant

Table 1. Testing for no effect of maternal age on child birth weight†

Test	*Results for the following values of K:*					
	K = 10			*K = 20*		
	Value	*p-value*	*Approximate p-value*	*Value*	*p-value*	*Approximate p-value*
RLRT	0	0.35	0.50	0.04	0.29	0.49
LRT	2.46	0.12	0.20	2.43	0.11	0.21

†Value is the value of the test statistic, p-value is the exact p-value and approximate p-value is the approximate p-value under the assumption of IID data.

$$H_0 : E(\text{birth weight} \,|\, \text{maternal age}) = \text{constant},$$
$$H_A : E(\text{birth weight} \,|\, \text{maternal age}) = f(\text{maternal age}).$$

f can be modelled either as a piecewise constant spline

$$f_c(x) = \beta_0 + \sum_{k=1}^{K} b_k \, I(x > \kappa_k), \qquad b_k \text{ IID } N(0, \sigma_b^2),$$

or as a linear spline

$$f_1(x) = \beta_0 + \beta_1 + \sum_{k=1}^{K} b_k (x - \kappa_K)_+, \qquad b_k \text{ IID } N(0, \sigma_b^2).$$

In the first case the hypotheses become

$$H_0 : \sigma_b^2 = 0 \quad \textit{versus} \quad H_A : \sigma_b^2 > 0,$$

and we can use the RLRT because the fixed effects are the same under the null and alternative hypotheses. In the second case we use the LRT to test the hypotheses

$$H_0 : \beta_1 = 0, \ \sigma_b^2 = 0 \quad \textit{versus} \quad H_A : \beta_1 \neq 0 \text{ or } \sigma_b^2 > 0.$$

Table 1 shows the values of all these statistics, the corresponding finite sample p-value and the approximate p-values using standard boundary asymptotics for IID data for $K = 10$ and $K = 20$ knots. The knots are equally spaced quantiles on the space of observed maternal age, and 100 000 simulations from the null distribution have been used for each test statistic. The results show that neither RLRT_n nor LRT_n can reject the null hypothesis at the level $\alpha = 0.10$. The approximate p-values are much larger than the true p-values, which is in accordance with our theoretical results. For example, for $K = 20$ when the LRT is used the exact p-value is 0.11 and the approximate p-value is 0.21. The severe inaccuracy of the approximate p-value could be a serious problem in examples where the exact p-value is closer to the typical critical values such as 0.05.

6. Testing for a fixed signal-to-noise ratio

We now focus on the signal-to-noise ratio $\lambda = \sigma_b^2/\sigma_\varepsilon^2$ in an LMM with one variance component. We are interested in testing the hypotheses described in equation (5) and obtaining confidence intervals by inverting the LRT or RLRT statistic. The following theorem provides the spectral decomposition of the RLRT_n statistic for testing hypotheses (5).

Theorem 3. If $\mu_{s,n}$ are the K eigenvalues of the $K \times K$ matrix $\Sigma^{1/2} Z^{\mathrm{T}} P_0 Z \Sigma^{1/2}$ then

$$\mathrm{RLRT}_n \overset{\mathcal{D}}{=} \sup_{\lambda \in \Lambda} \left[(n-p) \log \left\{ 1 + \frac{N_n(\lambda, \lambda_0)}{D_n(\lambda, \lambda_0)} \right\} - \sum_{s=1}^{K} \log \left(\frac{1 + \lambda \mu_{s,n}}{1 + \lambda_0 \mu_{s,n}} \right) \right], \tag{16}$$

where

$$N_n(\lambda, \lambda_0) = \sum_{s=1}^{K} \frac{(\lambda - \lambda_0)\mu_{s,n}}{1 + \lambda \mu_{s,n}} w_s^2,$$

$$D_n(\lambda, \lambda_0) = \sum_{s=1}^{K} \frac{1 + \lambda_0 \mu_{s,n}}{1 + \lambda \mu_{s,n}} w_s^2 + \sum_{s=K+1}^{n-p} w_s^2,$$

and w_s, for $s = 1, \ldots, n-p$, are independent $N(0,1)$.

The finite sample distributions of RLRT_n and $\hat{\lambda}_{\mathrm{REML}}$ depend essentially on λ_0, $\mu_{s,n}$ and Λ. This shows that their distributions are invariant only to reparameterizations that leave $Z^{\mathrm{T}} P_0 Z$ invariant. For $\lambda_0 = 0$ and $\Lambda = (0, \infty)$ we obtain result (9). An algorithm similar to that described in Section 2 can be developed to simulate the finite sample distribution of RLRT_n and REML estimator $\hat{\lambda}_{\mathrm{REML}}$. Because the distribution of $\hat{\lambda}_{\mathrm{REML}}$ is concentrated around the true value λ_0 the grid that is used for λ in $\{\lambda_0\} \cup \Lambda$ must be finer around λ_0 but can be coarser further away. Needless to say the finite sample distributions are not χ_1^2 and, in fact, have positive probability mass at 0. Crainiceanu, Ruppert and Vogelsang (2003), using the first-order conditions for a maximum at $\lambda = 0$, computed the finite sample probability mass at 0 of the RLRT_n for $\lambda_0 \in [0, \infty)$. They showed that this probability is not 0 even when λ_0 is relatively large.

Theorem 3 allows the construction of confidence intervals by inverting the RLRT. Indeed, if $\Lambda = [0, \infty) - \{\lambda_0\}$ we define the α-level restricted likelihood confidence interval for λ

$$\mathrm{CI}_\lambda = \{\lambda_0 | p\text{-value}(\lambda_0, \Lambda) \geqslant \alpha\}, \tag{17}$$

where p-value(λ_0, Λ) denotes the p-value for the RLRT_n statistic for testing the null hypothesis $\lambda = \lambda_0$ against the alternative $\lambda \in [0, \infty) - \{\lambda_0\}$. Because p-value(λ_0, Λ) for any given λ_0 can be obtained in seconds, CI_λ can be obtained by simply computing p-value(λ_0, Λ) on a relatively fine grid. This procedure would be very computationally intensive by using a direct bootstrap instead of taking advantage of the spectral decomposition (16).

In the balanced one-way ANOVA model this procedure provides an alternative α-level confidence interval to that based on the F-statistic that was obtained by Searle *et al.* (1992). In P-spline regression C_λ is an α-level confidence interval for the smoothing parameter λ.

Theorem 3 provides a natural generalization for testing for a fixed degrees of freedom in a P-spline regression as described by Cantoni and Hastie (2002). In the framework that is described in Section 5.1 they were interested in testing

$$H_0 : \lambda = \lambda_0 \quad \textit{versus} \quad H_{\mathrm{A}} : \lambda = \lambda_1 > \lambda_0,$$

which is a particular case of testing described in expression (5) for $\Lambda = \{\lambda_1\}$. Cantoni and Hastie (2002) proposed a test (called R in their paper) that is derived from the LRT in the LMM representation of a smoothing spline. They derived its finite sample distribution under the additional assumption that $X^{\mathrm{T}} Z = 0$. We denote this test by $R(\lambda_0, \lambda_1)$ to indicate the null and the alternative hypothesis. They conjectured that this statistic can be extended to test

$$H_0 : \lambda = \lambda_0 \quad \textit{versus} \quad H_A : \lambda > \lambda_0,$$

by using the $\hat{\lambda}_{\mathrm{REML}}$ estimator instead of λ_1, ignoring the estimation variability in $\hat{\lambda}_{\mathrm{REML}}$ and using the finite sample distribution of $R(\lambda_0, \hat{\lambda}_{\mathrm{REML}})$ as if $\hat{\lambda}_{\mathrm{REML}}$ were fixed. However, replacing a fixed λ_1 by an estimator has severe effects on the finite sample distribution of the test statistic. Crainiceanu, Ruppert, Claeskens and Wand (2003) showed that the null probability mass at λ_0 of $R(\lambda_0, \hat{\lambda}_{\mathrm{REML}})$ is generally very large (greater than 0.5). In contrast, the distribution of $R(\lambda_0, \lambda_1)$ has no mass at 0 for any $\lambda_1 > \lambda_0$.

The general version of this problem is solved by simply taking $\Lambda = [0, \infty) - \{\lambda_0\}$, or $\Lambda = (\lambda_0, \infty)$, in theorem 3 and using fast simulation algorithms similar to that described in Section 2.

Properties of the REML estimator of the smoothing parameter can be obtained as a by-product of the spectral decomposition in equation (16). Indeed, denote by $f_n(\lambda, \lambda_0)$ the quantity to be maximized on the right-hand side of equation (16). It is clear that the probability of having a local maximum of $f_n(\lambda, \lambda_0)$ in $[0, \lambda_0)$ is greater than or equal to

$$\mathrm{pr}\left\{ \frac{\partial}{\partial \lambda} f_n(\lambda, \lambda_0) \Big|_{\lambda=\lambda_0} \leqslant 0 \right\}.$$

Calculating this derivative we obtain that this probability is equal to

$$\mathrm{pr}\left\{ \sum_{s=1}^{K} c_{s,n}(\lambda_0) w_s^2 \leqslant \sum_{s=1}^{n-p} w_s^2/(n-p) \right\}, \tag{18}$$

where $\Sigma_{s=1}^{K} c_{s,n}(\lambda_0) = 1$ and

$$c_{s,n}(\lambda_0) = \frac{\mu_{s,n}}{1 + \lambda_0 \mu_{s,n}} \Big/ \sum_{s=1}^{K} \frac{\mu_{s,n}}{1 + \lambda_0 \mu_{s,n}}.$$

Therefore if we define $\hat{\lambda}^1_{\mathrm{REML}}$ to be the first maximum of $f_n(\lambda, \lambda_0)$ we obtain that under the null hypothesis that $\lambda = \lambda_0$

$$\mathrm{pr}(\hat{\lambda}^1_{\mathrm{REML}} < \lambda_0) \geqslant \mathrm{pr}\left\{ \sum_{s=1}^{K} c_{s,n}(\lambda_0) w_s^2 \leqslant \sum_{s=1}^{n-p} w_s^2/(n-p) \right\}.$$

The probability described in equation (18) can be easily obtained by using simulations. In standard scenarios this probability is greater than 0.5 (for example for $\lambda_0 = 0$ the probability is approximately 0.65). In these scenarios we obtain in finite samples

$$\mathrm{pr}(\hat{\lambda}^1_{\mathrm{REML}} < \lambda_0) \geqslant 0.5,$$

for every λ_0. This shows that the REML estimator tends to oversmooth the data. Corresponding asymptotic results were obtained for smoothing splines under additional assumptions by Kauermann (2003).

7. Distribution theory and geometry of spectra

Equations (7) and (9) provide the finite sample distributions of the LRT_n and RLRT_n statistics in terms of the eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$ of the matrices $Z^{\mathrm{T}} P_0 Z$ and $Z^{\mathrm{T}} Z$ respectively. Therefore investigating the spectra of these matrices could provide more insight into the distribution theory and the differences from standard asymptotics.

For simplicity of presentation we focus on RLRT_n whose distribution depends only on $\mu_{s,n}$. We can arrange the eigenvalues $\mu_{s,n}$ in decreasing order because the distribution of RLRT_n is invariant to permutations of $\mu_{s,n}$. This distribution is also invariant to rescaling the eigenvalues, i.e. the distribution remains unchanged if we replace all $\mu_{s,n}$ by $\mu_{s,n}/c$, where c is a fixed

constant. By choosing $c = \max_s(\mu_{s,n})$ we standardize eigenvalues such that $\mu_{1,n} = 1$ and compare eigenvalues across models without changing the null finite sample distributions.

Consider the case of testing for a linear regression *versus* a general alternative modelled by a linear spline with K knots as described in Section 5.2. We consider $n = 100$ observations and four cases for the distribution of xs. The first case considers xs equally spaced in $[0, 1]$ and the other three cases correspond to xs simulated from the beta$(1, 1)$, beta$(20, 1)$ and beta$(1, 20)$ distributions. We also considered two cases for the number of knots $K = 20$, which is often used in P-spline frameworks, and $K = 100$, which corresponds to smoothing splines (one knot at each observation).

Fig. 5 displays all the standardized eigenvalues $\mu_{s,n}$ of $Z^{\mathrm{T}} P_0 Z$ for the case $K = 20$ (described by the asterisk) and only the first 20 eigenvalues for the case $K = 100$ (described by circles), with the remaining eigenvalues being practically 0. Fig. 5 also displays the standardized eigenvalues corresponding to a balanced one-way ANOVA model. As we showed in Section 4, in one-way ANOVA all eigenvalues are equal, and this is so when the usual asymptotic theory holds if the number of random effects (or levels) goes to ∞.

Whereas for the ANOVA model the standardized eigenvalues are constant, for the P-spline model they decrease rapidly to 0, showing that in these cases the distributions depend practically only on the first 10 eigenvalues. It is the skewness of the eigenvalues $\mu_{s,n}$ that determines the differences from the usual asymptotic distribution for IID data. Another important feature is that the standardized eigenvalues are practically the same for $K = 20$ and $K = 100$ in all the cases considered, indicating that the finite sample distributions of the RLRT_n in the two models

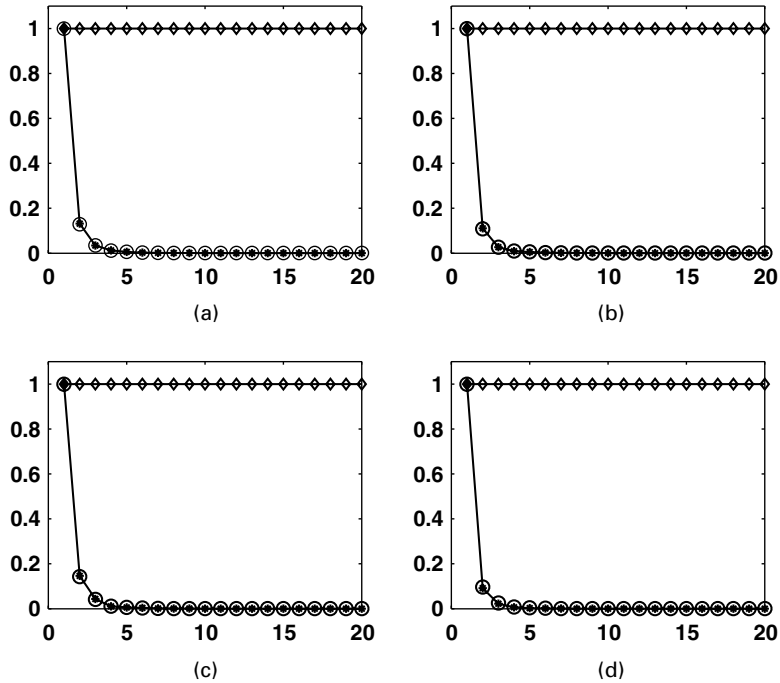

Fig. 5. Standardized eigenvalues of the matrix $Z^{\mathrm{T}} P_0 Z$ when testing for a linear polynomial *versus* a penalized spline with $n = 100$ observations and K knots ($*$, $K = 20$; $\circ$, $K = 100$; $\diamond$, standardized eigenvalues for a one-way ANOVA model): four cases are considered for xs—(a) equally spaced, (b) beta(1,1), (c) beta(20,1) and (d) beta(1,20)

are indistinguishable. The same type of result was found in many other cases that we do not report here.

Liu and Wang (2002) compared the power properties of several tests for polynomial regression *versus* a general alternative modelled by smoothing splines including RLRT_n (GML in Liu and Wang (2002)). They acknowledged that the null distribution is difficult to derive and used direct Monte Carlo simulation to derive it. Our result (9) provides the finite sample distribution of RLRT_n for any number of knots K, including the case $K = n$ of smoothing splines. It is true that if n is very large it may be difficult to diagonalize $n \times n$ matrices. However, in general, only several eigenvalues are essentially different from 0 and these eigenvalues are enough to simulate the finite sample distribution in the case of smoothing splines. P-splines with a reasonably large number of knots (say $K = 20$) avoid this problem because they only require the diagonalization of small dimension matrices.

8. Discussion

We derived the finite sample and asymptotic distribution of the LRTs and RLRTs for null hypotheses that include constraints on variance components in LMMs with one variance component. The distributions depend essentially on the eigenvalues of some design matrices. Once they have been computed explicitly or numerically, an efficient simulation algorithm can be used to derive the distributions of interest.

Three applications were considered: testing for level or subject effects in a balanced one-way ANOVA, testing for polynomial regression *versus* a general alternative modelled by P-splines and testing for a fixed number of degrees of freedom *versus* the alternative. In the ANOVA case the usual asymptotic theory for a parameter on the boundary holds if the number of subjects goes to ∞ but provides conservative approximations to the finite sample distributions for a fixed number of subjects. In the case of testing for a polynomial regression the asymptotic theory for IID data does not hold even if the number of knots that are used to fit the P-spline under the alternative hypothesis increases to ∞. Using the same idea our results can be used for testing in other penalized likelihood models that are equivalent to LMMs.

Although our results provide solutions to the problems that we considered, we only considered the case of LMMs with one random-effects variance component. Crainiceanu, Ruppert, Claeskens and Wand (2003) provide the spectral decomposition of the RLRT distribution for more than one variance component. They also discuss cases when this decomposition can be used efficiently for simulation of the null distribution.

Acknowledgement

We thank two reviewers for their careful reading of the original manuscript and for their comments that significantly improved the paper.

References

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

Azzalini, A. and Bowman, A. (1993) On the use of nonparametric regression for checking linear relationships. *J. R. Statist. Soc.* B, **55**, 549–557.

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. Am. Statist. Ass.*, **94**, 794–797.

Cantoni, E. and Hastie, T. J. (2002) Degrees of freedom tests for smoothing splines. *Biometrika*, **89**, 251–263.

Cleveland, W. S. and Devlin, S. J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Ass.*, **83**, 597–610.
Crainiceanu, C. M. and Ruppert, D. (2003) Proofs of theorems for the paper "Likelihood ratio tests in linear mixed models with one variance component". *Technical Report TR1389*. Department of Statistics, Cornell University, Ithaca. (Available from `http://www.orie.cornell.edu/orie_template.php3?select=TechReps`.)
Crainiceanu, C. M., Ruppert, D., Claeskens, G. and Wand, M. P. (2003) Exact likelihood ratio tests for penalized splines. To be published. (Available from `www.orie.cornell.edu/~davidr/papers`.)
Crainiceanu, C. M., Ruppert, D. and Vogelsang, T. J. (2003) Some properties of likelihood ratio tests in linear mixed models. To be published. (Available from `www.orie.cornell.edu/~davidr/papers`.)
Feng, Z. and McCulloch, C. E. (1992) Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the parameter is on the boundary of the parameter space. *Statist. Probab. Lett.*, **13**, 325–332.
Gray, R. J. (1994) Spline-based tests in survival analysis. *Biometrics*, **50**, 640–652.
Hanselman, D. and Littlefield, B. (2000) *Mastering MATLAB 6*. Englewood Cliffs: Prentice Hall.
Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Statist. Ass.*, **72**, 320–338.
Hastie, T. J. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Kauermann, G. (2003) A note on bandwidth selection for penalised spline smoothing.
Kuo, B. S. (1999) Asymptotics of ML estimator for regression models with a stochastic trend component. *Econometr. Theory*, **15**, 24–49.
Liang, K.-Y. and Self, S. G. (1996) On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J. R. Statist. Soc.* B, **58**, 785–796.
Liu, A. and Wang, Y. (2002) Hypothesis testing in smoothing spline models. (Available from `http://www.pstat.ucsb.edu/faculty/yuedong/papers/tests.pdf`.)
Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.
Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS*. New York: Springer.
Ruppert, D. (2002) Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.*, **11**, 735–757.
Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.
Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components*. New York: Wiley.
Self, S. G. and Liang, K. Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Am. Statist. Ass.*, **82**, 605–610.
Shephard, N. G. (1993) Maximum likelihood estimation of regression models with stochastic trend components. *J. Am. Statist. Ass.*, **88**, 590–595.
Shephard, N. G. and Harvey, A. C. (1990) On the probability of estimating a deterministic component in the local level model. *J. Time Ser. Anal.*, **4**, 339–347.
Stern, S. E. and Welsh, A. H. (2000) Likelihood inference for small variance components. *Can. J. Statist.*, **28**, 517–532.
Stram, D. O. and Lee, J. W. (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.