Likelihood ratio tests in linear mixed models with one variance component

March 31, 2003

Ciprian M. Crainiceanu

Cornell University, Ithaca, USA

and David Ruppert

Cornell University, Ithaca, USA

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. We derive the finite sample and asymptotic distribution of the likelihood ratio test (LRT) and the restricted likelihood ratio test (RLRT). The spectral representations of the LRT and RLRT statistics are used as the basis of an efficient simulation algorithm of these null distributions. The large sample chi-square mixture approximations using the usual asymptotic theory for a null hypothesis on the boundary of the parameter space (e.g., Self and Liang 1987, 1995), has been shown to be poor in simulation studies. Our asymptotic calculations explain these empirical results. The theorems of Self and Liang are perfectly correct, but we show that their assumptions do not, in general, hold for linear mixed models. Moreover, the hope that these assumptions could be weakened so that the Self and Liang theory would apply to linear mixed models proved false, since these asymptotic distributions do not hold for linear mixed models. One-way ANOVA and penalized splines models illustrate the results.

Keywords: Null hypothesis on boundary, Penalized splines, Degrees of freedom

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

1 Introduction

We consider the problem of testing null hypotheses that include constraints on the variance component in a linear mixed model (LMMs) 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 finite sample behavior observed in empirical studies, e.g., Pinheiro and Bates (2000). For the important special case of the null hypothesis that the variance component is zero, we also derive the asymptotic distributions of (R)LRT 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 zero or, equivalently, that the (R)LRT is zero.

One possible application is testing for a level (or group) effect in a mixed balanced one-way 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. Given its simplicity, the one-way ANOVA model will be used as a basis of comparison with more complex models.

Another application is using (R)LRTs for testing a polynomial fit versus a general alternative described by penalized splines (P-spline). As shown in section 5, this is equivalent to testing whether the variance component of a particular LMM is zero. Our (R)LRT 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), penalized splines (e.g. Hastie and Tibshirani; Ruppert, Wand and Carroll, 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 the ones derived by Self and Liang (1987, 1995) and Stram and Lee (1994) under the restrictive assumption that the response variable vector can be partitioned into i.i.d. subvectors and the number of independent subvectors tends to infinity — we will call this 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 subjects increases to infinity. Their results would not hold for a fixed number of subjects, even if the number of observations per subject increased to infinity. Feng and McCulloch (1992) show that for i.i.d. data (see their Theorems 2.2 and 2.3) the likelihood ratio test 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 a LMM with one variance component

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\epsilon}, \;\; \mathrm{E}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_K \\ \mathbf{0}_n \end{bmatrix}, \;\; \mathrm{Cov}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 \mathit{\Sigma} & \mathit{0} \\ \mathit{0} & \sigma_\epsilon^2 I_n \end{bmatrix}, \tag{1}$$

where $\mathit{0}_K$ is a K dimensional column of zeros, $\mathit{\Sigma}$ is a known symmetric positive definite $K \times K$ matrix, $\boldsymbol{\beta}$ is a p-dimensional vector of fixed-effects parameters, $\boldsymbol{b}$ is a K dimensional vector of random effects, and $(\boldsymbol{b}, \boldsymbol{\epsilon})$ is normal distributed. Under these conditions it follows that

$$E[\boldsymbol{Y}] = X\boldsymbol{\beta} \text{ and } \mathrm{Cov}(\boldsymbol{Y}) = \sigma_\epsilon^2 V_\lambda,$$

where $\lambda = \sigma_b^2/\sigma_\epsilon^2$, $V_\lambda = I_n + \lambda Z \mathit{\Sigma} Z^T$, and n is the length of $\boldsymbol{Y}$. The parameter λ can be considered a signal-to-noise ratio since σ_b^2 determines the size of the "signal" given by $E[\boldsymbol{Y}|\boldsymbol{b}] = X\boldsymbol{\beta} + Z\boldsymbol{b}$. Note that $\sigma_b^2 = 0$ if and only if $\lambda = 0$ and the parameter space for λ is $[0, \infty)$. Consider testing for the null hypothesis

$$\mathrm{H}_0 : \; \beta_{p+1-q} = \beta_{p+1-q}^0, \; \ldots, \; \beta_p = \beta_p^0, \; \sigma_b^2 = 0 \; (\text{equivalently, } \lambda = 0) \; , \tag{2}$$

versus the composite alternative

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

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

$$\mathrm{H}_0 : \; \sigma_b^2 = 0 \; (\lambda = 0) \; , \;\; \text{vs.} \;\; \mathrm{H}_\mathrm{A} : \; \sigma_b^2 > 0 \; (\lambda > 0) \; . \tag{4}$$

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

A generalization of (4) is

$$\mathrm{H}_0 : \; \lambda = \lambda_0 \; , \;\; \text{vs.} \;\; \mathrm{H}_\mathrm{A} : \; \lambda \in \Lambda \subset [0, \infty) \; , \tag{5}$$

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 $\{\lambda \neq \lambda_0 : \lambda \in [0, \infty)\}$. In section 6 show that testing for a fixed smoothing parameter (or equivalently number of degrees of freedom) of a penalized spline regression is equivalent to testing (5). Tests of this type can be inverted to create confidence intervals for λ.

Under the i.i.d. assumption, Stram and Lee (1994) proved that the LRT for testing the null (2) against the alternative (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 i.i.d. hypothesis does not hold for most LMMs. Even for 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 infinity. For the null hypothesis (4), $q = 0$, Crainiceanu, Ruppert and Vogelsang (2002) calculated the finite sample probability mass at zero of 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 LRT (and RLRT) for testing the hypotheses (2) against (4) or (5), we provide an appealing practical testing methodology. The advantage of our work over the Stram and Lee results (1994) is that the i.i.d. hypothesis is eliminated, which changes the null distribution theory compared to 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) provide 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.

2 Finite sample distribution of LRT and RLRT

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

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

and the LRT is defined as

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

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

$$\widehat{\boldsymbol{\beta}}(\lambda) = \left(X^T V_\lambda^{-1} X\right)^{-1} X^T V_\lambda^{-1}\boldsymbol{Y}\ ,\ \ \widehat{\sigma}^2_\epsilon(\lambda) = \frac{\{\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}(\lambda)\}^T V_\lambda^{-1}\{\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}(\lambda)\}}{n}\ .$$

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

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

where $P_\lambda = I_n - X(X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1}$. 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 $S_1 = I_n - X_1(X_1^T X_1)^{-1} X_1^T$ the LRT statistic is

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

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

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

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

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

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

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

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

The distribution described in (7) depends only on the eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$ of two $K \times K$ matrices. Once these have been calculated, simulations from this distribution can be done extremely 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.

Algorithm

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 $S_K = \sum_{s=1}^{K} w_s^2$

Step 3. Independently of step 1, simulate $X_{n,K,p} = \sum_{s=K+1}^{n-p} w_s^2$ with χ_{n-p-K}^2 distribution

Step 4. Independently of steps 1 and 2, simulate $X_q = \sum_{s=1}^{q} u_s^2$ with χ_q^2 distribution

Step 5. For every grid point λ_i compute

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

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

Step 7. Compute

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

Step 8. Repeat **Steps 2-7**

An important feature of the algorithm is that its speed depends on the number of random effects K but not on the number of observations n. As an example, for $K = 20$ knots we obtained $5,000$ simulations/second (2.66GHz CPU, 1Mb RAM) using efficient matrix manipulations in `Matlab`. The algorithm remains feasible as long as we can obtain the eigenvalues of $K \times K$ matrices and simulation time remains in the order of seconds for K as large as 500.

Direct bootstrap simulation of the null finite sample distribution increases with the number of observations n and with the number of knots K. Some examples of simulation times for 5,000 bootstrap simulations from the null distribution with an optimized Monte Carlo program are: 1.5 minutes for $K = 20$ and $n = 500$, 10 minutes for $K = 100$ and $n = 1,000$, and 60 minutes for $K = 300$ and $n = 2000$. When $K > 400$, computation times for direct bootstrap simulation become prohibitive. A similar algorithm can be given for RLRT_n 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 $\widehat{\lambda}_{\mathrm{ML}}$. Because for the null distributions of interest $\widehat{\lambda}_{\mathrm{ML}}$ is zero, or very close to zero, the grid needs to be very fine in a neighborhood of zero 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 ($\lambda \geq 0$) we also simulated the exact null distribution. This algorithm was slower than the algorithm 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 zero of the (R)LRT statistic for testing that the variance component is zero with no constraints on the fixed effects ($q = 0$, $\sigma_b^2 = 0$). The first order condition for $f_n(\lambda)$ to have a local maximum at $\lambda = 0$ is

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

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

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

This is the exact probability of a local maximum at zero and provides an excellent approximations for the probability of a global maximum at $\lambda = 0$ (Crainiceanu, Ruppert and Vogelsang, 2002). This probability can be easily obtained by simulation.

Because the finite sample distribution of the (R)LRT 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" chi-square 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 behavior of the LRT_n distribution depends essentially on the asymptotic behavior 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 $\boldsymbol{\beta}$ parameters and thereby to obtain unbiased variance components estimators. REML consists of maximizing the likelihood function associated with $n-p$ linearly independent contrasts. It makes no difference which $n-p$ contrasts are used because the likelihood function for any such set differs by no more than an additive constant (Harville, 1977). The restricted profile likelihood log-likelihood function is (Harville, 1977)

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

The RLRT_n is defined like the LRT_n 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 if the fixed effects are the same under null and alternative. Therefore 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 described in (4)

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

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

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

where $w_1, \ldots, w_K$ are independent $N(0,1)$ random variables. The notations here are 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 to 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 hypothesis (2) and (4) respectively. Because the finite sample results in section 2 depend essentially on the eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$ one may expect to have a relationship between the asymptotic distributions of test statistics and the asymptotic behavior of these eigenvalues. The following theorem provides the formal description of this relationship.

Theorem 2 *Assume that hypothesis H_0 in (2) is true. Suppose that there exists an $\alpha \geq 0$ so that for every s the K eigenvalues $\mu_{s,n}$ and $\xi_{s,n}$ of matrices $\Sigma^{1/2}Z^TP_0Z\Sigma^{1/2}$ and $\Sigma^{1/2}Z^TZ\Sigma^{1/2}$ respectively satisfy $\lim_{n\to\infty}\mu_{s,n} = \mu_s$, $\lim_{n\to\infty}\xi_{s,n} = \xi_s$, where not all μ_s are zero. Then*

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

where the two terms in the asymptotic distribution are independent, w_s are i.i.d. 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 and is presented in the Appendix. The first part of the asymptotic distribution is a χ_q^2 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 i.i.d. data suggests the following asymptotic result

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

where U is a $N(0,1)$ random variable independent of all w_s, $U_+^2 = U^2 I(U > 0)$ 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 $\sup_{d \geq 0} \mathrm{LRT}_\infty$ 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 (2002) show that $\sup_{d \geq 0} \mathrm{LRT}_\infty$ is essentially the Dirac measure at zero 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 behavior, but i.i.d. asymptotic theory gives no indication of the lack of power. Crainiceanu, Ruppert and Vogelsang (2002) show that the RLRT_n does not have this problem of a near degenerate asymptotic null distribution and suggest using the RLRT to increase power.

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

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

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 zero. Crainiceanu, Ruppert and Vogelsang (2002) 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 zero. The latter is well approximated by the null probability of having a local maximum at $\lambda = 0$. The first order conditions for local maximum at $\lambda = 0$ are

$$\left. \frac{\partial}{\partial d} \mathrm{LRT}_\infty(d) \right|_{d=0} \leq 0 \quad \text{or} \quad \left. \frac{\partial}{\partial d} \mathrm{RLRT}_\infty(d) \right|_{d=0} \leq 0 \,.$$

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

$$\mathrm{pr}\left(\sum_{s=1}^{K} \mu_s w_s^2 \leq \sum_{s=1}^{K} \xi_s \right) \quad \text{or} \quad \mathrm{pr}\left(\sum_{s=1}^{K} \mu_s w_s^2 \leq \sum_{s=1}^{K} \mu_s \right) \,,$$

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

While the asymptotic distributions are not needed to approximate the finite sample distributions, two important conclusions follow from the result of this section. The first is that the usual boundary asymptotic theory for i.i.d. 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 i.i.d. data and our results. We also compare finite sample and asymptotic distributions of (R)LRT statistics.

4 Balanced one-way ANOVA

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

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

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

Consider the test for $\sigma_b^2 = 0$. To find the finite sample distributions of (R)LRT$_n$ one needs to determine the eigenvalues of $Z^T Z$ and $Z^T P_0 Z$. In this simple model we can actually find them explicitly. All K eigenvalues of the matrix $Z^T Z$ are equal to $\xi_{s,n} = J$. Also, one eigenvalue of the matrix $Z^T P_0 Z$ is equal to zero and the remaining $K - 1$ eigenvalues are equal to $\mu_{s,n} = J$. Using Theorem 1 it follows that

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

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 using a simpler version of the algorithm in section 2. In particular, the finite sample probability mass at zero given by equation (8) is

$$\text{pr}\left\{ F_{K-1,n-1} \leq \frac{K(n-1)}{(K-1)n} \right\}\ ,$$

where $F_{K-1,n-1}$ has an F distribution with $(K-1, n-1)$ degrees of freedom. The probability mass at zero for this special case is similar to that obtained by Searle, Casella and McCulloch (1992, p. 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

$$n^{-1}\xi_{s,n} \to \frac{1}{K},\ s = 1, \ldots, K,\ \ n^{-1}\mu_{s,n} \to \frac{1}{K},\ s = 1, \ldots, K-1,\ \ n^{-1}\mu_{s,K} \to 0\ .$$

(In each case, "→" is, in fact, equality.) Using these expressions of μ_s and ξ_s in Theorem 2 the following result holds

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

where X_{K-1} denotes a random variable with a χ^2_{K-1} distribution, and I is the indicator function. Note that this asymptotic distribution has mass at zero equal to $P\left(X_{K-1} < K\right)$. The usual nonstandard asymptotic results requires K to increase to infinity, with J either fixed or also increasing to infinity, so that $P\left(X_{K-1} < K\right) \to 1/2$.

The balanced one-way ANOVA model is one of the few possible cases when the observed response vector $\boldsymbol{Y}$ can be partitioned into i.i.d. sub-vectors 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 ∞.

Figure 1 displays the 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). Note that the finite sample distributions converge quickly to the K-fixed asymptotic distribution described in (10) and away from the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution. The $0.5\chi^2_0 : 0.5\chi^2_1$ is a conservative approximation of 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 zero and to differences in the non-zero parts of the distributions. For $K = 5$, 10, 20, and 100, the probability mass at zero of the LRT statistic is 0.71, 0.65, 0.61, and 0.55, respectively, showing that unless the number of levels K is very large, the 0.5 asymptotic value is inaccurate. Figure 2 shows the QQ-plots of the χ^2_1 distribution versus the K-fixed asymptotic distribution of LRT_n conditional on $\mathrm{LRT}_n > 0$ for $K = 3$, 5, and 20. The quantiles were obtained using 1 million simulations from these conditional distributions. When K goes to infinity, 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 is 4.99. Because the curves in Figure 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

$$\text{RLRT}_n \Rightarrow \left\{ X_{K-1} - (K-1) - (K-1)\log\left(\frac{X_{K-1}}{K-1}\right) \right\} I\left\{X_{K-1} > K-1\right\} \ , \qquad (11)$$

and the asymptotic probability mass at zero 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 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 LRT and RLRT statistics in several important cases. While we focus on penalized splines, results are very general and can be used for any type of basis function (truncated polynomials, B-splines, trigonometric polynomials) and for any type of quadratic penalty.

5.1 P-Splines regression and linear mixed models

Consider the following regression equation

$$y_i = m\left(x_i\right) + \epsilon_i,$$

where ϵ_i are i.i.d. $N\left(0, \sigma^2_\epsilon\right)$ and $m(\cdot)$ is the unknown mean function. Suppose that we are interested in testing if $m(\cdot)$ is a $p-q$ degree polynomial:

$$H_0 : \ m\left(x\right) = \beta_0 + \beta_1 x + \ldots + \beta_{p+1-q} x^{p-q}. \qquad (12)$$

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

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

where $\boldsymbol{\theta} = \left(\beta_0, \ldots, \beta_p, b_1, \ldots, b_K\right)^T$ is the vector of regression coefficients, $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_p\right)^T$ is the vector of polynomial parameters, $\boldsymbol{b} = \left(b_1, \ldots, b_K\right)^T$ 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 large enough (e.g., 20) to ensure the desired flexibility. The knots can be taken to be equally-spaced quantiles of the x's, that is, the $1/(K+1)$,..., $K/(K+1)$ sample quantiles. To avoid overfitting, the criterion to be minimized is a penalized sum of squares

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

where $\lambda \geq 0$ is the smoothing parameter and L is a positive semi-definite matrix. The fitted function is called a penalized spline, or simply P-spline. A common choice of L is

$$L = \begin{bmatrix} 0 & 0 \\ 0 & \varSigma^{-1} \end{bmatrix},$$

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

$$\frac{1}{\sigma_\epsilon^2} \|\boldsymbol{Y} - X\boldsymbol{\beta} - Z\boldsymbol{b}\|^2 + \frac{1}{\lambda\sigma_\epsilon^2} \boldsymbol{b}^T \varSigma^{-1} \boldsymbol{b}\,. \tag{14}$$

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

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

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

We will consider the case $\varSigma = I_K$, penalizing the sums of squares of the jumps at the knots of the pth derivative of the fitted curve. While theoretical results in section 3 hold for a general known symmetric positive definite penalty matrix $\varSigma$, the choice $\varSigma = I_K$ will be used in the remainder of the paper 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

$$\mathrm{H}_0 : \beta_{p+1-q} = \beta_{p+1-q}^0,\ \ldots,\ \beta_p = \beta_p^0,\ \sigma_b^2 = 0\ (\lambda = 0)$$

versus the alternative

$$\mathrm{H_A}:\ \beta_{p+1-q} \neq \beta^0_{p+1-q} \text{ or, } \dots, \text{ or } \beta_p \neq \beta^0_p, \text{ or } \sigma^2_b > 0\ (\lambda > 0).$$

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

There is one very important point that we wish to emphasize. The assumption that the b_i's are i.i.d. normal random variables is, of course, a convenient fiction that converts nonparametric regression into a LMM. There are legitimate concerns about making inferences based on this assumption. However, if the null hypothesis (12) that m is a $p-q$ degree polynomial is true, then the b_i are all zero and therefore they are, in fact, i.i.d. $N(0,\sigma^2_b)$ with $\sigma^2_b = 0$! 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 (12) against a general alternative. Moreover, by using simulated quantiles from the finite-sample null distribution of the (R)LRT, 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 will model the alternative as a piecewise constant spline with K knots and test

$$H_0:\ m(x) = \beta_0 \ \text{ vs. }\ H_A:\ m(x) = \beta_0 + \sum_{k=1}^{K} b_k I\left\{x > \kappa_k\right\}\ .$$

As shown in section 2, the finite sample distribution of the $(\mathrm{R})\mathrm{LRT}_n$ statistics depend on the eigenvalues of the $K \times K$ matrices $Z^T P_0 Z$ and $Z^T Z$. Penalized splines use a moderately large number of knots K, with $K \leq 20$ in most applications and $K = 100$ being a rather extreme choice. With K in this range, these eigenvalues can be computed numerically using an efficient matrix diagonalization algorithm, such as the one implemented in `Matlab` (function `eig`) and need to be computed only once before the simulation algorithm described in section 2. We recommend using 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, to show the inaccuracy of "standard" asymptotics assuming i.i.d. data. If we want the asymptotic distribution to approximate accurately the finite-sample distribution, then K should be keep fixed at its actual in a given application. Therefore, we are interested in asymptotic results when the number of observations n tends to infinity and the number of knots K is kept constant. For this case, the asymptotic behavior of matrices Z^TZ and Z^TP_0Z was studied by Crainiceanu, Ruppert and Vogelsang (2002). If we denote by $n_s(n)$ the number of x's larger than the s-th knot and assume that $n_s(n)/n \to p_s$ as $n \to \infty$ one can show that

$$\lim_{n\to\infty} \frac{Z^TZ}{n} = R \text{ and } \lim_{n\to\infty} \frac{Z^TP_0Z}{n} = M ,$$

where the (i,j)-th entries for matrices R and M are $r_{ij} = p_{\max(i,j)}$ and $m_{ij} = p_{\max(i,j)} - p_ip_j$ 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

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

Substituting these values into $\text{LRT}_\infty(d)$ 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 zero is 0.66 for RLRT and 0.95 for LRT.

Figure 3 shows the QQ-plots for comparing quantiles of the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution (horizontal axis) with quantiles of the RLRT_n null distributions when testing for a constant mean versus a general alternative modeled by a piecewise constant spline with K-knots. We used the case of equally spaced x's in [0,1] with the k-th knot being the empirical quantiles of the x's corresponding to probability $k/(K+1)$. We consider $K = 10$ and $K = 20$. The solid line is the 45 degree line corresponding to the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution, the dotted lines correspond to finite sample distributions for $n = 50$ and $n = 100$ respectively, and the dashed line corresponds to the asymptotic distribution. One million simulations (taking approximately 3.5 minutes) 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.

The $0.5\chi^2_0 + 0.5\chi^2_1$ distribution represents a very conservative approximation of the finite sample and asymptotic distributions. Therefore, using the $0.5 : 0.5$ mixture can result in severe losses in power. Using an analogy with the ANOVA case, one 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 the case here, since the probability mass at zero remains practically unchanged, 0.65, for K between 2 and 100.

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

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

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

As in the previous case, it is easy to obtain the finite sample distribution of the (R)LRT$_n$ test by diagonalizing the corresponding matrices $Z^T P_0 Z$ and $Z^T Z$ and using the simulation algorithm described in section 2. To obtain the asymptotic distribution note that

$$\lim_{n\to\infty} \frac{Z^T Z}{n} = U \ \text{ and } \ \lim_{n\to\infty} \frac{Z^T P_0 Z}{n} = V,$$

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

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

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 zero of these distributions when K is 20 is 0.68 for RLRT and > 0.99 for LRT. Hence, the asymptotic distribution of the LRT is practically a point mass at zero, making the LRT impractical for this case. This is due to the downward bias of ML variance estimation.

Simulations were used to obtain the finite sample and asymptotic distributions of RLRT$_n$. Similar patterns to the ones presented in Figure 3 for testing for a constant mean were obtained for testing for a linear mean.

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

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

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

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

Theorem 1 provides the finite sample distribution of the LRT_n statistic. Using theorem 2, the asymptotic distribution for LRT is $\sup_{d\geq 0}\mathrm{LRT}_\infty(d) + Z^2$, where $\sup_{d\geq 0}\mathrm{LRT}_\infty(d)$ is the asymptotic distribution for LRT for testing a linear polynomial versus a piecewise linear spline, and Z is a standard normal random variable. Because we already proved that $\sup_{d\geq 0}\mathrm{LRT}_\infty(d)$ is practically the point mass at zero, we conclude that the asymptotic distribution for testing a constant mean versus a piecewise linear spline (two constrained parameters under the null) is practically a χ^2_1 while the "standard" i.i.d. asymptotic distribution is $.5\chi^2_1 + .5\chi^2_2$! For RLRT the finite sample and asymptotic distributions are different from a χ^2_1 because $\sup_{d\geq 0}\mathrm{RLRT}_\infty(d)$ for testing a linear polynomial versus a piecewise linear spline does not have the entire mass at zero. Nonetheless, the "standard" $.5\chi^2_1 + .5\chi^2_2$ approximation is not very accurate for the RLRT.

5.3 Upper Cape Cod birth weight data

Figure 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. Birthweight 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 commissioned in the late 1990s by the Department of Public Health, Commonwealth of Massachusetts (Ruppert, Wand and Caroll, 2003). Figure 4 also shows the no-effect and the ML linear penalized splines fits ($K = 20$). The ML estimator is very close to the overall mean (the smoothing parameter was estimated to be zero).

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

$$\begin{array}{l} H_0 : \; E\,[\text{birth weight}|\text{maternal.age}] = \text{constant} \\ H_A : \; E\,[\text{birth weight}|\text{maternal.age}] = f(\text{maternal.age})\;. \end{array}$$

f can be modeled either as a piecewise constant spline

$$f_c(x) = \beta_0 + \sum_{k=1}^{K} b_k I_{x>\kappa_k}\;,\quad b_k \text{ i.i.d } N(0,\sigma_b^2)\;,$$

or as a linear spline

$$f_l(x) = \beta_0 + \beta_1 + \sum_{k=1}^{K} b_k (x-\kappa_K)_+\;,\quad b_k \text{ i.i.d } N(0,\sigma_b^2)\;.$$

In the first case the hypotheses become

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

and we can use RLRT because the fixed effects are the same under the null and alternative.

In the second case we use LRT to test the hypotheses

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

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 i.i.d. data for $K = 10$ and $K = 20$ knots. Knots are equally spaced quantiles on the space of observed maternal age, and $100,000$ simulation of the null distribution have been used for each test statistic. Results show that neither RLRT_n nor LRT_n can reject the null 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 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 typical critical values such as .05.

Table 1: Testing for no-effect of maternal age on child birth weight

	$K = 10$			$K = 20$		
	value	p-value	$\approx$ p-value	value	p-value	$\approx$ 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

Note: "p-value" is the exact p-value, "$\approx$ p-value" is the approximate p-value under the i.i.d. assumption, and "value" is the value of the test statistic.

6 Testing for a fixed signal-to-noise ratio

We now focus on the signal-to-noise ratio $\lambda = \sigma_b^2/\sigma_\epsilon^2$ in a LMM with one variance component. We are interested in testing hypothesis described in equation (5) and obtaining confidence intervals by inverting the (R)LRT statistic. The following theorem provides the spectral decomposition of the RLRT_n statistic for testing hypotheses (5) with $\Lambda = [0, \infty) - \{\lambda_0\}$.

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

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

where

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

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

The finite sample distributions of RLRT_n and $\widehat{\lambda}_{\mathrm{REML}}$ depend essentially on λ_0, $\mu_{s,n}$ and Λ. This shows that their distributions are invariant only to reparameterizations that leave $Z^T P_0 Z$ invariant. For $\lambda_0 = 0$ and $\Lambda = (0, \infty)$ we obtain the result in (9). An algorithm similar to the one described in section 2 can be developed to simulate the finite sample distribution of RLRT_n and REML estimator $\widehat{\lambda}_{\mathrm{REML}}$. Because the distribution of $\widehat{\lambda}_{\mathrm{REML}}$ is concentrated around the true value λ_0 the grid used for λ in $\{\lambda_0\} \cup \Lambda$ has to be finer around λ_0 but can be coarser farther away. Needless to say that the finite sample distributions are not χ_1^2. Crainiceanu, Ruppert and Vogelsang (2002), using the first order conditions for a maximum at $\lambda = 0$, compute the finite sample probability mass at zero of the RLRT_n for $\lambda_0 \in [0, \infty)$. They show that this probability is not zero 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

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

where p-value(λ_0, Λ) denotes the p-value for the RLRT_n statistic for testing the null $\lambda = \lambda_0$ versus the alternative $\lambda \in [0, \infty) - \{\lambda_0\}$. Because p-value$(\lambda_0, \Lambda)$ 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 using a direct bootstrap instead of taking advantage of the spectral decomposition (16).

In the case of balanced one-way ANOVA model this procedure provides an alternative α-level confidence interval to the one based on F-statistic obtained by Searle, Casella and McCulloch (1992). In the case of penalized 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 penalized splines regression as described by Cantoni and Hastie (2002). In the framework described in section 5.1 they were interested in testing

$$\mathrm{H}_0 : \lambda = \lambda_0 \, , \;\; \text{vs.} \;\; \mathrm{H}_\mathrm{A} : \lambda = \lambda_1 \, ,$$

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

$$H_0 : \lambda = 0 \quad \text{vs.} \quad H_A : \lambda > 0 \, ,$$

by using the $\widehat{\lambda}_{\mathrm{REML}}$ estimator instead of λ_1, ignoring the estimation variability in $\widehat{\lambda}_{\mathrm{REML}}$ and using the finite sample distribution of $\mathrm{R}(0, \widehat{\lambda}_{\mathrm{REML}})$ as if $\widehat{\lambda}_{\mathrm{REML}}$ 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 (2002) show that the null probability mass at zero of $\mathrm{R}(0, \widehat{\lambda}_{\mathrm{REML}})$ is equal to the probability mass at zero of RLRT_n and is generally very large (> 0.5). In contrast, the distribution of $\mathrm{R}(0, \lambda_1)$ has no mass at zero for any $\lambda_1 > 0$.

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 the one described in section 2.

Properties of the REML estimator of the smoothing parameter can be obtained as a byproduct of the spectral decomposition in equation (16). Indeed, denote by $f_n(\lambda, \lambda_0)$ the quantity to be maximized in 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 or equal to

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

By directly calculating this derivative we obtain that this probability is equal to

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

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

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

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

$$\text{pr}\left(\widehat{\lambda}^1_{REML} < \lambda_0\right) \geq \text{pr}\left\{\sum_{s=1}^{K} c_{s,n}(\lambda_0) w_s^2 \leq \frac{\sum_{s=1}^{n-p} w_s^2}{n-p}\right\} .$$

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

$$\text{pr}\left(\widehat{\lambda}^1_{REML} < \lambda_0\right) \geq 0.5 ,$$

for every λ_0. This is an important property of the REML estimator which tends to oversmooth the data. Corresponding asymptotic results were obtained for smoothing splines under additional assumptions by Kauerman (2002).

7 Relationship between distribution theory and eigenvalues

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 $Z^T P_0 Z$ and $Z^T Z$ 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, that is 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 modeled 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 x's. The first case considers x's equally spaced in $[0, 1]$ and the other three cases correspond to x's simulated from the BETA$(1, 1)$, BETA$(20, 1)$ and BETA$(1, 20)$ distributions respectively. We also considered two cases for the number of knots, $K = 20$ which is often used in penalized splines framework and $K = 100$ which corresponds to smoothing splines (one knot at each observation).

Figure 8 displays all the standardized eigenvalues $\mu_{s,n}$ of $Z^T P_0 Z$ 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 zero. Figure 8 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 the case when the usual asymptotic theory holds if the number of random effects (or levels) goes to infinity.

While for the ANOVA model the standardized eigenvalues are constant, for the penalized spline model they decrease rapidly to zero, 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 i.i.d. data. Another important feature is that the standardized eigenvalues are practically the same for $K = 20$ and $K = 100$ in all cases considered, indicating that the finite-sample distributions of the RLRT_n in the two models are indistinguishable. The same type of result was found in many other cases that we do not report here.

Liu and Wang, 2002 compare the power properties of several tests for polynomial regression versus a general alternative modelled by smoothing splines including RLRT_n (GML in their paper). They acknowledge that the null distribution is difficult to derive and use 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 zero and these eigenvalues are enough to simulate the finite sample distribution in the case of smoothing splines. Penalized splines with a reasonably large number of knots (say $K = 20$) avoids this problem because they only require the diagonalization of small dimension matrices.

8 Discussion

We derived the finite sample and asymptotic distribution of the (restricted) likelihood ratio tests for null hypotheses that include constraints on variance components in LMM with one variance component. The distributions depend essentially on the eigenvalues of some design matrices. Once they are 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 effect in a balanced one-way ANOVA, testing for polynomial regression versus a general alternative modelled by penalized 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 infinity but provides conservative approximations of the finite sample distributions for a fixed number of subjects. In the case of testing for a polynomial regression the asymptotic theory for i.i.d. data does not hold even if the number of knots used to fit the penalized spline under the alternative increases to infinity. Using the same idea our results can be used to testing in other penalized likelihood models.

While our results provide solutions to the problems considered, we only consider the case of LMMs with one random effects variance component. Crainiceanu, Ruppert, Claeskens and Wand (2002) 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.

Acknowledgements

We wish to 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(3), 683–734.

Azzalini, A. and Bowman, A.W. (1993). On the use of nonparametric regression for checking linear relationships, *Journal of the Royal Statistical Society 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. *Journal of the American Statistical Association*, 94, 794–797.

Cantoni, E., and Hastie, T.J. (2000). 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, *Journal of the American Statistical Association*, 83, 597–610.

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. *submitted to Journal of the American Statistical Association*, available at www.orie.cornell.edu/ ~davidr/papers.

Crainiceanu, C. M., and Ruppert, D., and Claeskens, G., Wand, M.P., (2002). Likelihood Ratio Tests of Polynomial Regression Against a General Nonparametric Alternative. *submitted*, available at www.orie.cornell.edu/~davidr/papers.

Gray, R. J. (1994) Spline-Based Tests in Survival Analysis. *Biometrics*, 50, 640–652.

Feng, Z., 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, *Statistics & Probability Letters*, 13, 325–332.

Harville, D.A., (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.

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

Kauerman, G., A note on bandwidth selection for penalised spline smoothing, Manuscript.

Kuo, B.S., (1999) Asymptotics of ML estimator for regression models with a stochastic trend component. *Econometric Theory*, 15, 24-49.

Liu, A., Wang, Y., (2002). Hypothesis testing in smoothing spline Models, *Manuscript.*

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. *Journal of Computational and Graphical Statistics*, 11, 735–757.

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

Searle, S.R., Casella, G., McCulloch, C.E., (1992). *Variance components.* Wiley, New York.

Self, S.G., and Liang, K.Y., (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association*, 82, 605-610.

Self, S.G., and Liang, K.Y., (1995) On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic. *Journal of Royal Statistical Society B*, 58, 785-796.

Shephard, N.G., Harvey, A.C., (1990) On the probability of estimating a deterministic component in the local level model. *Journal of Time Series Analysis*, 4, 339-347.

Shephard, N.G., (1993) Maximum Likelihood Estimation of Regression Models with Stochastic Trend Components. *Journal of the American Statistical Association*, 88, 590-595.

Stern, S.E., Welsh, A.H., Likelihood inference for small variance components. *The Canadian Journal of Statistics* 28(3), 517–532.

Stram, D.O., Lee, J.W., (1994) Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, 50, 1171-1177.

Figure 1: QQ-plots for comparing quantiles of the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution (horizontal axis) with quantiles of the LRT$_n$ null distributions when testing for level or subject effect ($K = 5$ subjects). The solid line is the 45 degree line corresponding to the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution, the dotted lines correspond to finite sample distributions for $n = 50$ and $n = 100$ respectively, and the dashed line corresponds to the asymptotic distribution.

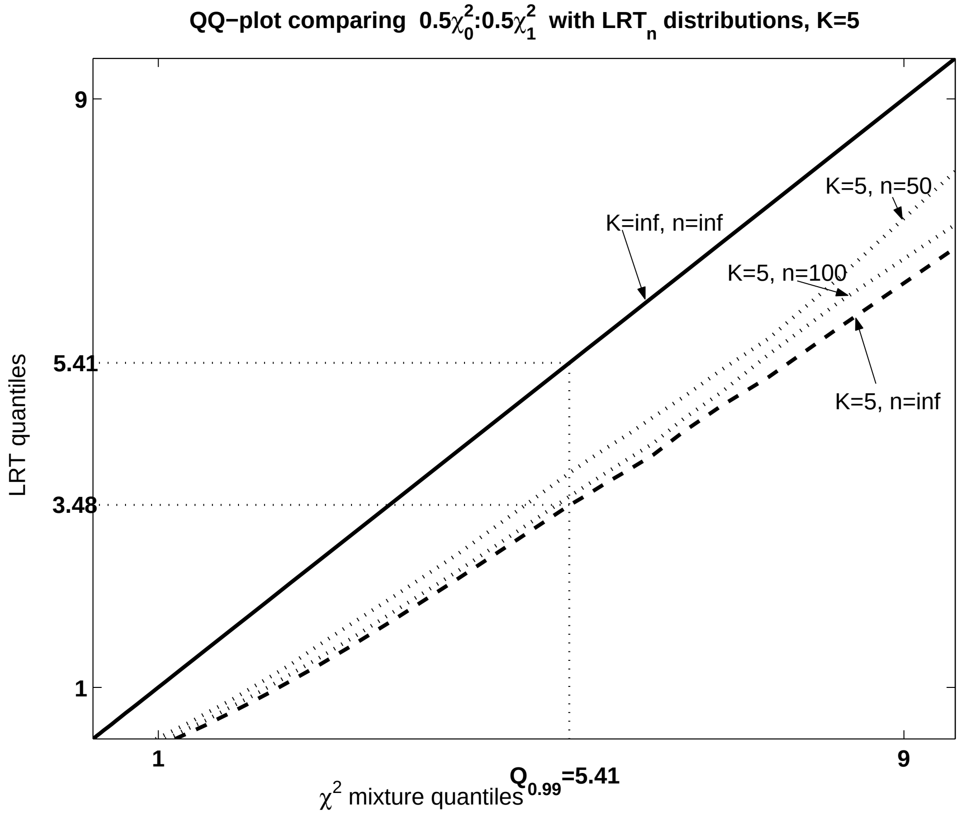


Figure 2: QQ-plots for comparing the χ_1^2 (horizontal axis) distribution with the asymptotic distributions of LRT_n conditional on $\mathrm{LRT}_n > 0$ {equation (10)} for balanced one-way ANOVA model. The solid line represents the χ_1^2 distribution. The dashed curves correspond to three number of levels in the ANOVA model, $K = 3, 5, 20$.

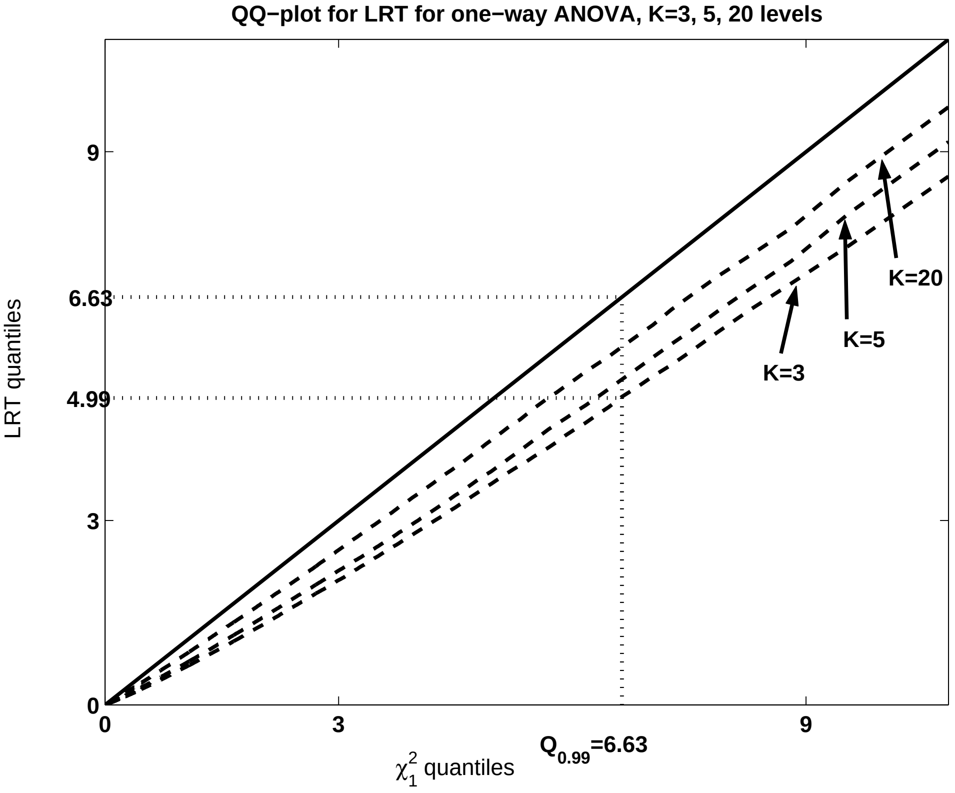
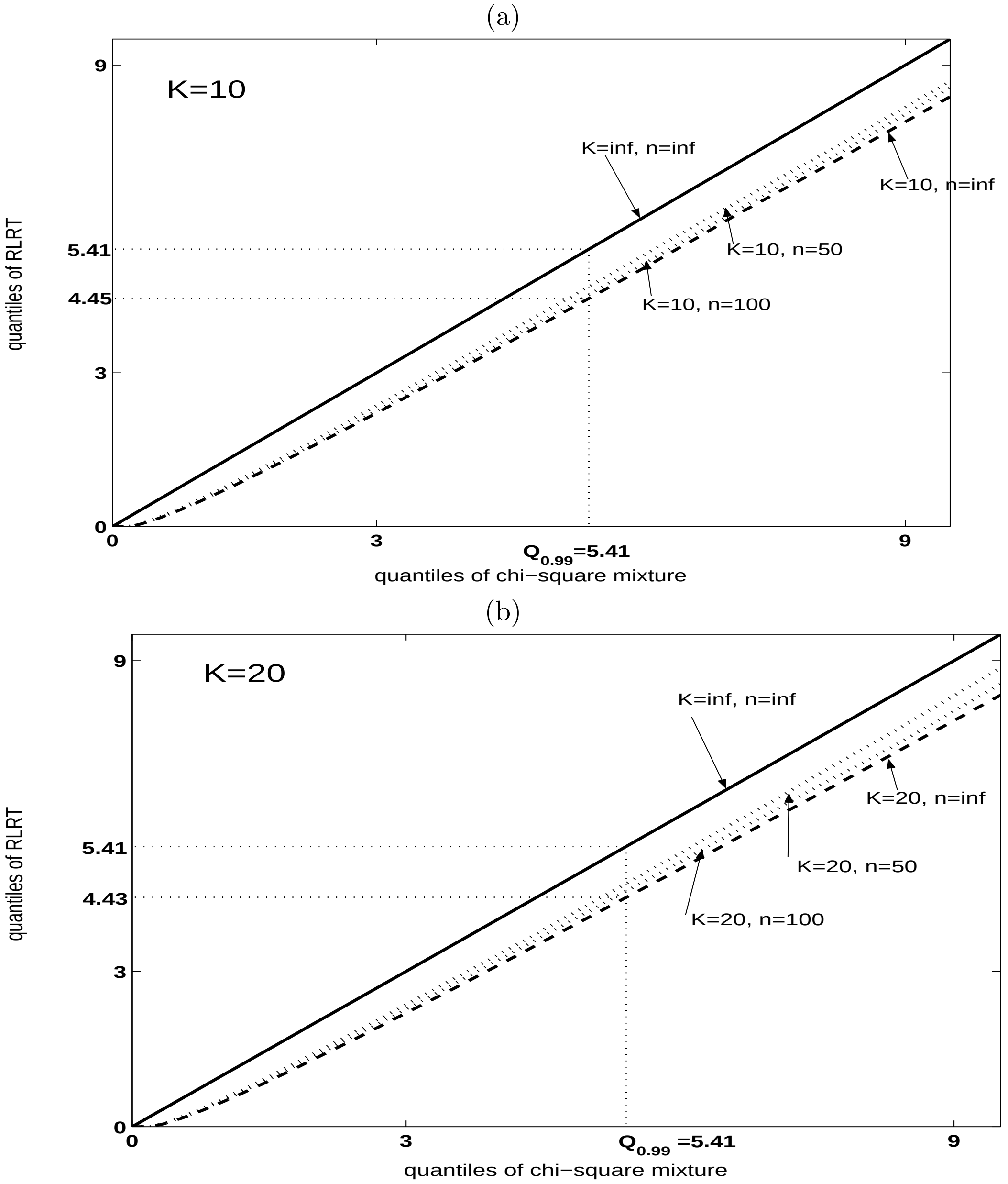


Figure 3: QQ-plots for comparing quantiles of the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution (horizontal axis) with quantiles of the RLRT_n null distributions when testing for a constant mean versus a general alternative modelled by a piecewise constant spline with K-knots. Equally spaced x's in [0,1], the knots are equally spaced quantiles of the x's. (a)- $K = 10$, (b)- $K = 20$. The solid line is the 45 degree line corresponding to the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution, the dotted lines correspond to finite sample distributions for $n = 50$ and $n = 100$ respectively, and the dashed line corresponds to the asymptotic distribution.



(a)

(b)

Figure 4: Child birth weight versus maternal age, linear penalized splines estimators with $K = 20$ knots of the mean function using ML.

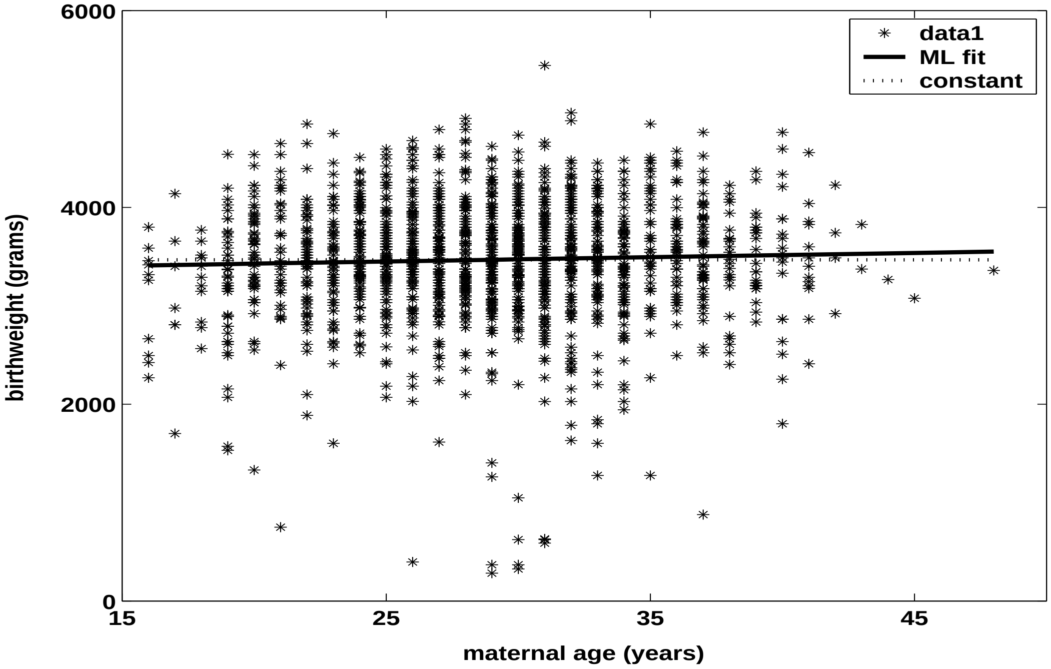
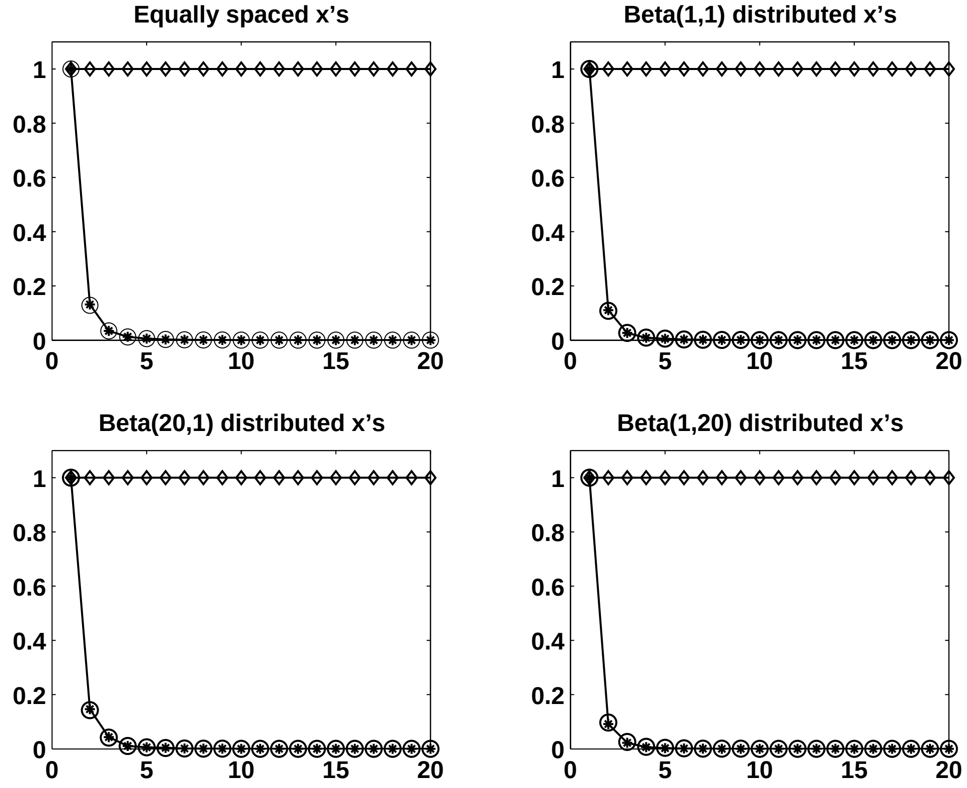


Figure 5: Standardized eigenvalues of the matrix $Z^T P_0 Z$ when testing for a linear polynomial versus a penalized spline with $n = 100$ observations and K knots. "*"- $K = 20$, "o"- $K = 100$. Four cases are considered for x's: equally spaced, BETA$(1, 1)$, BETA$(20, 1)$ and BETA$(1, 20)$. We also present the standardized eigenvalues for a one way ANOVA model ("$\diamond$").


Equally spaced x's
Beta(1,1) distributed x's
Beta(20,1) distributed x's
Beta(1,20) distributed x's