Probability that the MLE of a Variance Component is Zero With Applications to Likelihood Ratio Tests

Ciprian M. Crainiceanu* David Ruppert† Timothy J. Vogelsang‡

April 2, 2002

Abstract

We compute the finite sample probability that the maximum likelihood or residual maximum likelihood estimate of the variance component is zero in a linear mixed model with one variance component. The calculations are expedited by simple matrix diagonalization techniques. Application of these calculations to likelihood ratio tests (LRT) and residual likelihood ratio tests (RLRT) for a zero variance of the random effects yields the finite sample probability that the log of the likelihood ratio, or residual likelihood ratio, is zero. The asymptotic behavior of the probability mass at zero of log of the likelihood ratio (log-LR) and the log of the residual likelihood ratio (log-RLR) statistics is derived for two models, one-way ANOVA and penalized splines. The large sample chi-square mixture approximation to the distribution of the log-likelihood ratio, using the usual asymptotic theory for when a parameter is on the boundary, has been shown to be poor in simulations studies. Our calculations explain these empirical results.

Short title: Probability a variance component's MLE is zero

Keywords: Effects of dependence, Linear mixed models, Likelihood ratio tests, Penalized splines, Testing polynomial regression.

*Department of Statistical Science, Cornell University, Malott Hall, NY 14853, USA. E-mail: cmc59@cornell.edu

†School of Operational Research and Industrial Engineering, Cornell University, Rhodes Hall, NY 14853, USA. E-mail: ruppert@orie.cornell.edu

‡Departments of Economics and Statistics, Cornell University, Uris Hall, NY 14853-7601, USA. E-mail:tjv2@cornell.edu

1 INTRODUCTION

Linear mixed models (LMMs) are now widely used for, inter alia, modeling longitudinal data, and MLE and REML estimation of variance components is becoming commonplace. It is, therefore, natural to consider likelihood ratio tests (LRTs) to test the null hypothesis that a variance component is zero. In this paper we show that asymptotic theory developed for independent data cannot be applied blindly to tests in LMMs. For example, we show that applying asymptotics to even the simple case of balanced one-way ANOVA is subtle. If the number of observations per level is fixed and the number of levels tends to infinity, then i.i.d. theory applies. But, if the number of levels is fixed and the number of observations per level tends to infinity, then we get different asymptotic behavior of the LRT. The disagreement between i.i.d. theory and the actual asymptotic distribution is even worse for spline models.

This work was motivated by our research in testing parametric regression models versus nonparametric alternatives. It is becoming more widely appreciated that penalized splines and other penalized likelihood models can be viewed as LMMs and the fitted curves as a BLUPs (e.g., Brumback, Ruppert, and Wand, 1999). In this framework the smoothing parameter is a ratio of variance components. The latter can be estimated by the MLE or REML estimate. REML is often called generalized maximum likelihood (GML) in the smoothing spline literature. Within the random effects framework, it is natural to consider likelihood ratio tests and residual likelihood ratio tests (RLRTs) for testing null hypotheses such as no covariate effect or the regression function being a polynomial. These null hypotheses are equivalent to the hypothesis of a variance component being zero. LRTs for null variance components are non-standard for two reasons. First, the null value of the parameter is on the boundary of the parameter space. Second, the data are dependent, at least under the alternative hypothesis. These problems led us to investigate further the asymptotic distribution of LRTs that a variance component is zero.

Our major result is that present asymptotic theory is inadequate because it cannot take into account dependencies in the data. We conclude that the RLRT is much preferred to the LRT and that simulation is, at present, the best method for establishing critical values of the RLRT.

Our work is applicable to many LMMs, not to just penalized likelihood models. The major restriction on our results is that we only consider models with a single variance component, though we intend to look at a wider class of LMMs in the future.

Consider a LMM with one variance component

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}, \quad \mathrm{E}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0}_K \\ \boldsymbol{0}_n \end{bmatrix}, \quad \mathrm{Cov}\begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 \boldsymbol{I_K} & 0 \\ 0 & \sigma_\epsilon^2 \boldsymbol{I_n} \end{bmatrix}, \tag{1}$$

where $\boldsymbol{0}_K$ is a K dimensional column of zeros, $\boldsymbol{I_K}$ is a K dimensional identity matrix, $\boldsymbol{\beta}$ is a $p+1$ dimensional vector of parameters corresponding to fixed effects, $\boldsymbol{b}$ is a K dimensional vector of exchangeable random effects, and $(\boldsymbol{b}, \boldsymbol{\epsilon})$ is a normal distributed random vector. Under these conditions it follows that

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

where $\lambda = \sigma_b^2/\sigma_\epsilon^2$ is the ratio between the variance of random effects $\boldsymbol{b}$ and the variance of the error variables $\boldsymbol{\epsilon}$, $\boldsymbol{V}_\lambda = I_n + \lambda \boldsymbol{Z}\boldsymbol{Z}^T$, and n is the size of the vector $\boldsymbol{Y}$ of the response variable. Note that $\sigma_b^2 = 0$ if and only if $\lambda = 0$ and the parameter space for λ is $[0, \infty)$.

The LMM described by equation (1) contains standard regression fixed effects $\boldsymbol{X\beta}$ specifying the conditional response mean and random effects $\boldsymbol{Zb}$ that account for correlation. Testing whether this extra structure is necessary is equivalent to the hypotheses

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

If one uses the LRT, then what seems a typical testing problem proves to be non-standard for several reasons. First, under the null hypothesis, the parameter is on the boundary of the parameter space. Using non-standard asymptotic theory developed by Self and Liang (1987) for independent data, one may be tempted to conclude that the finite samples distribution of the log likelihood ratio (log-LR) and of the residual likelihood ratio (log-RLR) could be approximated by a $0.5\chi_0^2 + 0.5\chi_1^2$ mixture. Here χ_k^2 is the chi-square distribution with k degrees of freedom and χ_0^2 means point probability mass at 0. However, a second problem is lack of independence, at least under the alternative. Because the response variable $\boldsymbol{Y}$ in model (1) is not a vector of independent random variables, the Self and Liang theory does not apply. Stram and Lee (1994) showed that the Self and Liang result can still be applied to testing for the zero variance of random effects in Linear Mixed Models in which the response variable $\boldsymbol{Y}$ can be partitioned into independent vectors and the number of independent subvectors tends to infinity.

We originally conjectured that these results would still hold, that is, that their assumptions could be weakened to allow dependence. However, in empirical studies this $0.5\chi_0^2 + 0.5\chi_1^2$ mixture approximation has been shown to be conservative though *sometimes* it can be rather accurate. Indeed, in a simulation study for a related model, Pinheiro and Bates (2000) found that a $0.5\chi_0^2 + 0.5\chi_1^2$ mixture distribution approximates well the finite sample distribution of log-RLR. They also found that a $0.65\chi_0^2 + 0.35\chi_1^2$ mixture approximates better the finite sample distribution of log-LR. Although the Pinheiro and Bates approximations work well in the situations they simulated, these approximation do not work in all settings. By computing the finite sample probability at zero of the log-LR and log-RLR statistics, we show that both these approximations can be very poor, even for some very simple LMMs.

A case where it has been shown that the asymptotic mixture probabilities differ from $0.5\chi_0^2 + 0.5\chi_1^2$ is regression with a stochastic trend analyzed by Shephard and Harvey (1990) and Shephard (1993). They consider the particular case of model (1) where the random effects $\boldsymbol{b}$ are modeled as random walk. They show that the asymptotic mass at zero can be as large as 0.96 for log-LR and 0.65 for log-RLR. This shows that the 0.65 : 0.35 mixture for log-LR and 0.5 : 0.5 mixture for the log-RLR cannot be generalized to all LMMs.

Because it is not clear when different approximations can be used, we compute the finite samples probability at zero for the log-LR and log-LRT as a function of $\lambda = \sigma_b^2/\sigma_\epsilon^2$. This probability is equal to the probability of estimating the random effects variance to be zero ($\sigma_b^2 = 0$). In particular, the finite samples probability mass at zero of the log-LR and log-RLR under the null hypothesis is obtained when $\lambda = 0$. The computation procedure is very fast because it can be reduced to simple matrix diagonalization techniques. We derive the asymptotic behavior of the null probability at zero of log-LR and log-RLR for two LMMs, balanced one-way ANOVA and penalized splines. It is shown that this asymptotic probabilities are not 1/2.

2 MAXIMUM LIKELIHOOD APPROACH

We consider maximum likelihood estimation (MLE) for model (1). Minus twice the log-likelihood of $\boldsymbol{Y}$ given the parameters $\boldsymbol{\beta}$, σ_ϵ^2, and λ is, up to a constant that does not depend on the parameters,

$$f(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) = n \log \sigma_\epsilon^2 + \log |\boldsymbol{V}_{\boldsymbol{\lambda}}| + \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}_\lambda^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{\sigma_\epsilon^2} \tag{3}$$

A local minimum of the function $f(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda)$ can occur on the boundary of the parameter space where $\lambda = 0$ if and only if there exists $\boldsymbol{\beta}$ and σ_ϵ^2 so that the following system is satisfied at $\lambda = 0$

$$\left\{\begin{array}{lcl} \frac{\partial f}{\partial \beta}(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) & = & 0 \\ \frac{\partial f}{\partial \sigma_\epsilon^2}(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) & = & 0 \\ \frac{\partial f}{\partial \lambda}(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) & \geq & 0 \end{array}\right. \tag{4}$$

where the last derivative is from the right. Solving the first two equations in (4) is standard and gives immediately

$$\widehat{\boldsymbol{\beta}}(\lambda) = \left(\boldsymbol{X}^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{Y}\,, \tag{5}$$

and

$$\widehat{\sigma}_\epsilon^2(\lambda) = \frac{\{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}^T \boldsymbol{V}_\lambda^{-1} \{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}}{n}\,. \tag{6}$$

However, the last relationship in (4) is not standard and will be treated separately. In the right hand side of equation (3) only the second and the third term depend on λ. To determine the derivatives of these terms with respect to λ we first denote by d_i the i-th eigenvalue of the symmetric semi-positive definite matrix $\boldsymbol{Z}\boldsymbol{Z}^T$, and let $\boldsymbol{D}$ be the diagonal matrix having d_i as the i-th diagonal entry. Let $\boldsymbol{U}$ be the orthonormal matrix of eigenvectors of $\boldsymbol{Z}\boldsymbol{Z}^T$, so that $\boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$. It follows that $\boldsymbol{V}_{\boldsymbol{\lambda}} = \boldsymbol{U}(\boldsymbol{I}_n + \lambda \boldsymbol{D})\boldsymbol{U}^T$, that $|\boldsymbol{V}_\lambda| = \prod_{i=1}^n (1 + \lambda d_i)$, and that

$$\frac{\partial}{\partial \lambda} \log |\boldsymbol{V}_\lambda| = \sum_{i=1}^n \frac{d_i}{1 + \lambda d_i}\,. \tag{7}$$

Also, the following equality holds (Harville, 1977)

$$\frac{\partial}{\partial \lambda}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}_\lambda^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = -(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{V}_\lambda^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\,. \tag{8}$$

The conditions in equation (4) are equivalent to the set of equations (5), (6), and

$$-n \left[\frac{\{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{V}_\lambda^{-1} \{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}}{(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda))^T \boldsymbol{V}_\lambda^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda))}\right] + \sum_{i=1}^n \frac{d_i}{1 + \lambda d_i} \geq 0\,. \tag{9}$$

If $\boldsymbol{P}(\lambda) = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{V}_\lambda^{-1}$, then equation (9) can be rewritten as

$$-n \left\{\frac{\boldsymbol{Y}^T \boldsymbol{P}^T(\lambda) \boldsymbol{V}_\lambda^{-1} \boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{P}(\lambda) \boldsymbol{Y}}{\boldsymbol{Y}^T \boldsymbol{P}^T(\lambda) \boldsymbol{V}_\lambda^{-1} \boldsymbol{P}(\lambda) \boldsymbol{Y}}\right\} + \sum_{i=1}^n \frac{d_i}{1 + \lambda d_i} \geq 0\,. \tag{10}$$

To find the probability of estimating σ_b^2 to be 0 we compute the probability

$$P\left(\frac{\boldsymbol{Y}^T\boldsymbol{P}^T(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Y}}{\boldsymbol{Y}^T\boldsymbol{P}(0)\boldsymbol{Y}} \leq \frac{1}{n}\sum_{i=1}^{n} d_i\right), \tag{11}$$

where we used the idempotency of $\boldsymbol{P}(0)$ and $\boldsymbol{V}_0 = \boldsymbol{I}_n$. For a fixed λ, the random variable $\boldsymbol{P}(0)\boldsymbol{Y}$ is normally distributed with mean zero and covariance matrix $\sigma_\epsilon^2\boldsymbol{P}(0)\boldsymbol{V}_\lambda\boldsymbol{P}^T(0)$. Also observe that $\sum_{i=1}^n d_i = \text{tr}(\boldsymbol{Z}\boldsymbol{Z}^T)$ is equal to $\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z})$, where "tr" denotes the trace of the matrix. We use $\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z})$ for computations because the dimension of $\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z})$ is much smaller than that of $\text{tr}(\boldsymbol{Z}\boldsymbol{Z}^T)$. If we denote by $\boldsymbol{u}$ an $N_n(0, \boldsymbol{I}_n)$ random vector, then the probability in (11) equals

$$P\left(\frac{\boldsymbol{u}^T\boldsymbol{V}_\lambda^{1/2}\boldsymbol{P}^T(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{V}_\lambda^{1/2}\boldsymbol{u}}{\boldsymbol{u}^T\boldsymbol{V}_\lambda^{1/2}\boldsymbol{P}(0)\boldsymbol{V}_\lambda^{1/2}\boldsymbol{u}} \leq \frac{1}{n}\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z})\right), \tag{12}$$

under the assumption that $\sigma_b^2 = \lambda\sigma_\epsilon^2$. Denoting by $\boldsymbol{A} = \boldsymbol{V}_\lambda^{1/2}\boldsymbol{P}^T(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{V}_\lambda^{1/2}$, by $\boldsymbol{B} = \boldsymbol{V}_\lambda^{1/2}\boldsymbol{P}(0)\boldsymbol{V}_\lambda^{1/2}$, by $\bar{d} = \frac{1}{n}\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z})$, and taking into account that $\boldsymbol{B}$ is semi-positive definite, the probability in (12) can be rewritten as

$$P\left(\boldsymbol{u}^T(\boldsymbol{A} - \bar{d}\boldsymbol{B})\boldsymbol{u} \leq 0\right) \tag{13}$$

Denoting by ψ_i the i-th eigenvalue of the matrix $\boldsymbol{A} - \bar{d}\boldsymbol{B}$, the probability in (13) is

$$P\left(\sum_{i=1}^{n}\psi_i v_i^2 \leq 0\right), \tag{14}$$

where v_i are i.i.d. N(0,1) random variables. The last quantity can be calculated using exact algorithms, such as Davies (1980) or Farebrother (1990), or estimated by simulation.

In the important particular case when $\lambda = 0$ we obtain $\boldsymbol{A} = \boldsymbol{P}^T(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)$ and $B = \boldsymbol{P}(0)$, where $\boldsymbol{P}(0) = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Observe that if we require $\boldsymbol{X}^T\boldsymbol{Z} = 0$ then $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^T$.

3 RESTRICTED MAXIMUM LIKELIHOOD APPROACH

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 $n - p - 1$ linearly independent error contrasts. It makes no difference which $n-p-1$ contrasts are used because the likelihood function for any such set differs by no more than an additive constant (Harville, 1977). In particular, if $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is the linear model, then one is interested in the likelihood of $\boldsymbol{AY}$, where $\boldsymbol{A}$ is an $(n-p-1) \times n$ matrix whose rows are any $n-p-1$ linearly independent rows of the matrix $\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. For a LMM described in equation (1), minus twice the log-likelihood of $\boldsymbol{AY}$ was derived by Harville (1974) and is, up to a constant independent of the parameters, the log residual likelihood (log-REL)

$$g(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) = f(\boldsymbol{\beta}, \sigma_\epsilon^2, \lambda) - (p+1)\log(\sigma_\epsilon^2) + \log(|\boldsymbol{X}^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{X}|) \, . \tag{15}$$

As with the MLE we are interested in the probability that the function g has a local minimum corresponding to $\lambda = 0$. This is equivalent to the three conditions in (4) holding for g instead of f. Under these conditions, $\widehat{\boldsymbol{\beta}}(\lambda)$ is still given by equation (5) but

$$\widehat{\sigma}_\epsilon^2(\lambda) = \frac{\{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}^T \boldsymbol{V}_\lambda^{-1} \{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\}}{n-p-1} \, . \tag{16}$$

Note that in the right hand side of the equation (15) only the first and the third term depend on λ, that the derivative of f with respect to λ was already computed in section 2, and that (see Appendix A1)

$$\left. \frac{\partial}{\partial \lambda} \log(|\boldsymbol{V}_\lambda|) + \frac{\partial}{\partial \lambda} \log(|\boldsymbol{X}^T \boldsymbol{V}_\lambda^{-1} \boldsymbol{X}|) \right|_{\lambda=0} = \operatorname{tr}\left(\boldsymbol{Z}^T \boldsymbol{P}(0) \boldsymbol{Z}\right) \, , \tag{17}$$

which depends only on the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$. Using a calculation similar that in Section 2 for maximum likelihood, the probability that $\lambda = 0$ is a local minimum for log-REL is

$$P\left(\frac{\boldsymbol{u}^T \boldsymbol{V}_\lambda^{1/2} \boldsymbol{P}^T(0) \boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{P}(0) \boldsymbol{V}_\lambda^{1/2} \boldsymbol{u}}{\boldsymbol{u}^T \boldsymbol{V}_\lambda^{1/2} \boldsymbol{P}(0) \boldsymbol{V}_\lambda^{1/2} \boldsymbol{u}} \leq \frac{1}{n-p-1} \operatorname{tr}\left(\boldsymbol{Z}^T \boldsymbol{P}(0) \boldsymbol{Z}\right) \right) \, , \tag{18}$$

where $\boldsymbol{u}$ is $N(0, \boldsymbol{I}_n)$. When the true value is $\lambda = 0$ the probability in equation (18) becomes

$$P\left(\frac{\boldsymbol{u}^T \boldsymbol{P}(0) \boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{P}(0) \boldsymbol{u}}{\boldsymbol{u}^T \boldsymbol{P}(0) \boldsymbol{u}} \leq \frac{1}{n-p-1} \operatorname{tr}\left(\boldsymbol{Z}^T \boldsymbol{P}(0) \boldsymbol{Z}\right) \right) \, . \tag{19}$$

Probabilities described by equations (18) and (19) can be computed as for maximum likelihood. By directly comparing equations (12) and (18) we observe that the probability of finding a local minimum at $\lambda = 0$ is smaller when maximum likelihood is used if and only if

$$\frac{\operatorname{tr}(\boldsymbol{Z}^T \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Z})}{\operatorname{tr}(\boldsymbol{Z}^T \boldsymbol{Z})} \leq \frac{p+1}{n} \, . \tag{20}$$

In the particular case when $\boldsymbol{Z}^T\boldsymbol{X} = 0$ equation (20) always holds true because the left hand side of the equation is zero. Consider another particular case, when $\boldsymbol{X}$ is an $n \times 1$ vector of 1s, corresponding to having only an intercept in the fixed effects part of model (1). In this case $p = 0$ and if z_{ik} is the (i,k)-th entry of matrix $\boldsymbol{Z}$ then the left hand side of equation (20) is equal to

$$\frac{\sum_{k=1}^K \left(\sum_{i=1}^n z_{ik}\right)^2}{n\sum_{k=1}^K \sum_{i=1}^n z_{ik}^2},$$

which is greater than $1/n$ if all entries of matrix $\boldsymbol{Z}$ are positive. The inequality is strict if at least one column of $\boldsymbol{Z}$ there are two non-zero entries. This shows that, in this case, the probability of having a local minimum at $\lambda = 0$ is greater for maximum likelihood than for REML.

4 LIKELIHOOD RATIO TESTING

The probability that the Ml or REML estimator is on the boundary is needed to specify the distribution of the log-LR or log-RML test statistics.

Note that the probability mass at zero for log-LR and log-RLR equals the probability that the functions f and g described in equations (3) and (15), respectively, have a global minimum at $\lambda = 0$. For a given sample size we compute the exact probability of having a local minimum of the log-LR (or log-RLR) at $\lambda = 0$. This probability is an upper bound for the probability of having a global minimum at zero but, as we will show, it provides an excellent approximation. For two models we show that the asymptotic probability mass at zero for log-LR and log-RLR is not 0.5. Also, in finite samples the 0.5 : 0.5 approximation is very poor. We also show that the approximation proposed by Pinheiro and Bates (2000) cannot be applied to all LMMs.

A different problem is the use of the χ_1^2 distribution to approximate the non-zero part of the distribution. There is evidence that even this approximation does not always hold, but we will discuss this in another paper.

In the following we apply our results to LRTs in two rather different LMMs, balanced one-way ANOVA and penalized splines.

4.1 One-way ANOVA

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

$$Y_{ij} = \mu + b_i + \epsilon_{ij}\ ,\ i = 1, \ldots, K \ \text{ and } \ j = 1, \ldots, n. \tag{21}$$

where ϵ_{ij} are i.i.d. random variables $N(0, \sigma_\epsilon^2)$, b_i are i.i.d. random effects distributed $N(0, \sigma_b^2)$ independent of ϵ_{ij}, μ is a fixed unknown intercept, and denote by $\lambda = \sigma_b^2/\sigma_\epsilon^2$. The matrix $\boldsymbol{X}$ for fixed effects is simply an $nK \times 1$ column of ones and the matrix $\boldsymbol{Z}$ is an $nK \times K$ matrix with every column containing only zeros with the exception of an n-dimensional vector of 1's corresponding to the level parameter. It is easy to prove that $\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z}) = nK$ and $\text{tr}(\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Z}) = nK - n$.

If the true λ is 0, then the probability that the maximum likelihood criterion in (3) has a local minimum at $\lambda = 0$ is given by equation (14) where ψ_i are the eigenvalues of the matrix $\boldsymbol{A} - (nK)^{-1}\text{tr}(\boldsymbol{Z}^T\boldsymbol{Z})\boldsymbol{B} = \boldsymbol{A} - \boldsymbol{B}$. Here $\boldsymbol{A} = \boldsymbol{P}(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)$ and $\boldsymbol{B} = \boldsymbol{P}(0)$. In Appendix A2 it is shown that the first $K-1$ eigenvalues of this matrix are equal to $n-1$, the following $nK - K$ are equal to -1, and the last eigenvalue is 0. Therefore the probability of having a local minimum at $\lambda = 0$ for maximum likelihood is

$$p_{\text{ML}}(n, K) = P\left(\frac{1}{K}\sum_{i=1}^{K-1} v_i^2 < \frac{1}{nK-K}\sum_{i=K}^{nK-1} v_i^2\right), \tag{22}$$

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

For residual maximum likelihood ψ_i are the eigenvalues of the matrix $\boldsymbol{A} - \bar{d}\boldsymbol{B}$, where $\bar{d} = \frac{nK-n}{nK-1}$. The first $K-1$ eigenvalues of this matrix are equal to $n - \bar{d}$, the following $nK - K$ are equal to $-\bar{d}$, and the last one is equal to 0. Therefore the probability of having a local minimum at $\lambda = 0$ for residual maximum likelihood is

$$p_{\text{REML}}(n, K) = P\left(\frac{n-\bar{d}}{nK-K}\sum_{i=1}^{K-1} v_i^2 < \frac{\bar{d}}{nK-K}\sum_{i=K}^{nK-1} v_i^2\right). \tag{23}$$

The probabilities in equations (22) and (23) are easy to compute. For now we only study their asymptotic behavior when the number of observations per level n tends to ∞ and the number of levels K is fixed. Using the Law of Large Numbers it follows that for maximum likelihood

$$p_{\text{ML}}(K) = \lim_{n\to\infty} p_{\text{ML}}(n, K) = P\left(\sum_{i=1}^{K-1} v_i^2 < K\right), \tag{24}$$

which is the c.d.f. of a χ^2_{K-1} distribution evaluated at K. For the residual maximum likelihood we get

$$p_{\text{REML}}(K) = \lim_{n\to\infty} p_{\text{REML}}(n, K) = P\left(\sum_{i=1}^{K-1} v_i^2 < K-1\right), \tag{25}$$

which is the c.d.f. of a χ^2_{K-1} distribution evaluated at $K-1$.

Figure 1 shows the two probabilities (24) and (25) versus K. They represent the asymptotic probabilities when K fixed and n tends to ∞. By the Central Limit Theorem, when $K \to \infty$, both $p_{\mathrm{ML}}(K)$ and $p_{\mathrm{REML}}(K)$ converge to 1/2. However, in many applications, $K \leq 20$, and $p_{\mathrm{ML}}(5) = 0.7127$, $p_{\mathrm{ML}}(10) = 0.6495$, and $p_{\mathrm{ML}}(20) = 0.6054$. Therefore, the conjectured asymptotic value of 1/2 is incorrect when the number of levels K is fixed. Similar results can easily be derived for REML. The finite samples probability mass at zero of the random effects variance estimator using maximum likelihood was also reported by Yu et al. (1994).

It is interesting that the probability $\widehat{\sigma}_b^2$ is 0 for the balanced one-way ANOVA model converges to 0.5 as prescribed by Self and Liang (1987) for independent data. This can easily be explained because the response variable $\boldsymbol{Y}$ can be partitioned in n-dimensional independent blocks corresponding to each level. However, this is not likely of practical importance since, in general, the number of levels is fixed and small and then the 0.5 approximation is poor. Moreover, the one-way ANOVA model is probably the only non-trivial example of a LMM where we can do this partition into independent blocks!

We now focus on computing exact probabilities of having a local or global minimum at $\lambda = 0$ for different values λ. We set the number of levels $K = 5$ and vary the number of observations at each level. Results are reported in Table 1 for the log-LR and in Table 2 for the log-RLR. To obtain the probability of having a global minimum at zero we used 10,000 simulations from the model described in equation (21) for different values of λ and calculated the frequency of estimating zero-variance. These frequencies are reported between brackets in Tables 1 and 2.

As described in section (2) we need to compute probabilities given in equation (14), where v_i are i.i.d. N(0,1). Once the eigenvalues ψ_i are computed, we can either directly compute these probabilities, using an algorithm such as Farebrother (1990), or using simulations. We chose to simulate one million realizations of the the random variable $\sum_{i=1}^{n} \psi_i v_i^2$ because it is simple, very rapid, and accurate.

There is strong agreement between the exact probability of having a local minimum at $\lambda = 0$ and the estimated probability of having a global minimum at $\lambda = 0$, both for the log-LR and log-RLR. The columns corresponding to $\lambda = 0$ in Tables 1 and 2 show that, in this case, the asymptotic approximation of these probabilities by 0.5 is poor for both log-LR and log-RLR. The Pinheiro and Bates 0.65 : 0.35 approximation for log-LR is better but still off. Increasing n with K fixed does not solve the problem because the probability of having a local minimum at $\lambda = 0$ converges very quickly to the asymptotic value, e.g., for $K = 5$, $p_{\text{ML}}(5) = 0.7127$ or $p_{\text{REML}}(5) = 0.5940$.

Table 1: Probability of having a local (global) minimum at $\lambda = 0$ for the log-LR. The number of levels is $K = 5$.

n	$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$
5	0.6771 (0.6732)	0.6548 (0.6512)	0.4802 (0.4733)	0.0697 (0.0669)
10	0.6967 (0.6972)	0.6480 (0.6456)	0.3525 (0.3562)	0.0228 (0.0211)
20	0.7048 (0.7008)	0.6091 (0.6004)	0.2032 (0.2027)	0.0066 (0.0066)
40	0.7086 (0.7106)	0.5307 (0.5348)	0.0910 (0.0920)	0.0019 (0.0011)

Table 2: Probability of having a local and global minimum at $\lambda = 0$ for log-RLR. $K = 5$.

n	$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$
5	0.5707 (0.5727)	0.5447 (0.5440)	0.3777 (0.3780)	0.0471 (0.0467)
10	0.5822 (0.5822)	0.5331 (0.5282)	0.2639 (0.2629)	0.0153 (0.0154)
20	0.5885 (0.5880)	0.4932 (0.4945)	0.1452 (0.1471)	0.0044 (0.0036)
40	0.5914 (0.5920)	0.4167 (0.4198)	0.0619 (0.0621)	0.0012 (0.0012)

Notes: The finite sample probability of having a global maximum (probability mass at zero of log-LR and log-RLR respectively) is reported within brackets. It represents the frequency of estimating $\lambda = 0$ for different true values λ in 10,000 simulations from the balanced ANOVA model described in (21) for $K = 5$ levels and different number of observations n per level.

4.2 TESTING POLYNOMIAL REGRESSION VERSUS A NONPARAMETRIC ALTERNATIVE

In this section we show that nonparametric regression using P-splines is equivalent to a particular LMM. We then prove that testing for a polynomial regression versus a general alternative can be viewed as testing for a zero variance component in this LMM. Finally, we compute the finite sample and asymptotic probability at zero of the log-LR and log-RLR statistics and compare them with the two approximations discussed in Section 1.

4.2.1 P-SPLINES REGRESSION AND LINEAR MIXED MODELS

Consider the following regression equation

$$y_i = m\left(x_i\right) + \epsilon_i \;, \tag{26}$$

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

$$H_0:\; m\left(x\right) = \beta_0 + \beta_1 x + ... + \beta_p x^p \;.$$

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

$$H_A:\; m(x) = m\left(x, \boldsymbol{\Theta}\right) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} b_k \left(x - \kappa_k\right)_+^p \;, \tag{27}$$

where $\boldsymbol{\Theta} = \left(\beta_0, ..., \beta_p, b_1, ..., b_K\right)^T$ is the vector of regression coefficients, $\boldsymbol{\beta} = \left(\beta_0, ..., \beta_p\right)^T$ is the vector of polynomial parameters, $\boldsymbol{b} = \left(b_1, ..., b_K\right)^T$ is the vector of spline coefficients, and $\kappa_1 < \kappa_2 < ... < \kappa_K$ 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 are taken to be equally spaced quantiles on the range of the x's. To avoid overfitting, the criterion to be minimized is a penalized sum of squares

$$\sum_{i=1}^{n} \left\{y_i - m\left(x_i; \boldsymbol{\Theta}\right)\right\}^2 + \lambda \boldsymbol{\Theta}^T \boldsymbol{W} \boldsymbol{\Theta} \;, \tag{28}$$

where $\lambda \geq 0$ is the smoothing parameter and $\boldsymbol{W}$ is a positive semi-definite matrix.

Denote $\boldsymbol{Y} = (y_1, y_2, ..., y_n)^T$, $\boldsymbol{X}$ the matrix having the i-th row $\boldsymbol{X}_i = \left(1, x_i, \ldots, x_i^p\right)$, $\boldsymbol{Z}$ the matrix having the i-th row $\boldsymbol{Z}_i = \left((x_i - \kappa_1)_+^p, (x_i - \kappa_2)_+^p, \ldots, (x_i - \kappa_K)_+^p\right)$, and $\boldsymbol{\mathcal{X}} = [\boldsymbol{X}|\boldsymbol{Z}]$.

For the remainder of this paper $\boldsymbol{W}$ will denote a diagonal matrix with the first $p+1$ diagonal elements 0 and the remaining K diagonal elements 1. This choice of $\boldsymbol{W}$ penalizes the sums of squares of the jumps at the knots of the pth derivative of the fitted curve so that this curve is shrunk towards a pth degree polynomial fit. If criterion (28) is divided by σ_ϵ^2 one obtains

$$\frac{1}{\sigma_\epsilon^2}\left\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b}\right\|^2+\frac{\lambda}{\sigma_\epsilon^2}\boldsymbol{b}^T\boldsymbol{b}\,. \tag{29}$$

Consider the vector $\boldsymbol{\beta}$ as fixed effects and the vector $\boldsymbol{b}$ as a set of random effects with $E[\boldsymbol{b}]=0$ and $\text{Cov}(\boldsymbol{b})=\sigma_b^2\boldsymbol{I}_K$. Then (29) is, up to a constant, minus twice the log of the joint density of $(\boldsymbol{Y},\boldsymbol{\beta})$.

It follows (Brumback et al., 1999) that the regression spline model is also a LMM. Denote by $\lambda=\sigma_b^2/\sigma_\epsilon^2$ and by $\boldsymbol{V_\lambda}=\boldsymbol{I_n}+\lambda\boldsymbol{Z}\boldsymbol{Z}^T$ where n is the total number of observations. It is easy to see that $E[\boldsymbol{Y}]=\boldsymbol{X}\boldsymbol{\beta}$ and $\text{Cov}(\boldsymbol{Y})=\sigma_\epsilon^2\boldsymbol{V_\lambda}$. Minus twice the log-likelihood and residual log-likelihood of the model are described by equations (3) and (15) respectively.

4.2.2 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 the following hypotheses

$$H_0:\lambda=0\ \left(\sigma_b^2=0\right)\quad\text{vs.}\quad H_A:\lambda>0\ \left(\sigma_b^2>0\right)\,.$$

Because the b_i have mean zero, $\sigma_b^2=0$ 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.

Denote by $\text{ML}(p,0)$ the minimum of the function $f(\boldsymbol{\beta},\sigma_\epsilon^2,\lambda=0)$ in equation (3) and by $\text{ML}(p,\widehat{\lambda})$ the minimum of the same function with respect to all parameters. Then the log likelihood ratio is

$$\log\text{LR}=\text{ML}(p,\widehat{\lambda})-\text{ML}(p,0)\,. \tag{30}$$

Denote by $\text{REML}(p,0)$ the minimum of the function $g(\boldsymbol{\beta},\sigma_\epsilon^2,\lambda=0)$ described in equation (15) and by $\text{REML}(p,\widehat{\lambda})$ the minimum of the same function with respect to all parameters. Then the residual log likelihood ratio for testing H_0 against H_A is

$$\log\text{RLR}=\text{REML}(p,\widehat{\lambda})-\text{REML}(p,0)\,. \tag{31}$$

Note that the probability at zero of the log-LR (or log-RLR) test statistics is equal to the probability that the function f (or g) has at global minimum for $\lambda=0$. The probability of having a local minimum at $\lambda = 0$ can be calculated using either equation (12) for log-LR or equation (18) for log-RLR. This probability an upper bound for the probability of having global minima at zero. In the following we compute exact probabilities of having a local minimum and compare them with the frequency of having a global minimum at zero in simulations.

4.2.3 PROBABILITY OF ZERO FOR log-LR AND log-RLR

In this section we compute the probability that the log-LR and log-RLR is 0. We investigate the case when $p = 0$ corresponding to a constant mean and $p = 1$ corresponding to a linear polynomial under the null. When $p = 0$ the alternative is a piecewise constant spline and when $p = 1$ the alternative is a linear spline. We analyze the case when the x's are equally spaced on [0,1], but the same procedure can be applied to the more general case.

The probability of having a local minimum at zero for log-LR or log-RLR can be computed as in Section 4.1 using different design matrices and the results are reported in Tables 3 and 4. We also report between brackets the estimated probabilities of having a global maximum at zero. As for one-way ANOVA, we used 10,000 simulations from the model (1) for fixed λ. Again there is close agreement between these probabilities, especially when λ is close to 0.

An important observation is that the log-LR has almost all its mass at zero, that is $\simeq 0.925$, for $p = 0$, and 0.995, for $p = 1$. This makes the construction of a LRT very difficult, if not impossible, especially when we test for the linearity against a general alternative. Also the very high probability of estimating a zero smoothing parameter λ even when the true parameter is large, e.g. $\lambda = 1$ and $p = 1$, suggests that the power of this test can be poor. It is clear that the asymptotic approximation of the finite sample probability mass at zero of the log-LR by 0.5 is very poor in this case.

The log-RLR has less mass at zero $\simeq 0.66$, for $p = 0$, and $\simeq 0.67$, for $p = 1$, thus allowing the construction of tests. Also, the probability of estimating zero smoothing parameter when the true parameter λ is positive is much smaller (note the different scales in Tables 3 and 4) indicating that the RLRT is probably more powerful than the LRT. Also, the 0.5 asymptotic approximation found by Pinheiro and Bates to work well in a different example is very poor in this situation.

When testing for a constant mean versus a general alternative modeled by a piecewise constant spline, the asymptotic probability of having a local minimum at $\lambda = 0$ for log-LR and log-RLR is derived in Appendix A3. Figure 2 shows these probabilities. It is interesting that over a wide range of number of knots the probabilities are practically constant, ≈ 0.95 for ML and ≈ 0.65 for REML. Approximating the probability of estimating λ as 0 by 1/2 is very inaccurate, since 1/2 is not even correct asymptotically.

When the number of knots of the spline under the alternative is $K = 20$ and the true $\lambda = 0$ the asymptotic probabilities of having a local minimum at $\lambda = 0$ are $p_{\mathrm{ML}}(20) = 0.9545$ and $p_{\mathrm{REML}}(20) = 0.6567$. This proves that the asymptotic probability mass at zero for log-LR and RLRT for a fixed number of knots is not 0.5 and differences are very large. Comparing these results with results from the first columns of Tables 3 and 4 it follows that the finite samples probability of having a local minimum at $\lambda = 0$ is very well approximated by the correct asymptotic probability (not 1/2) even for a small number of observations.

Table 3: Probability of having a local and global minimum at $\lambda = 0$ for log-LR

	$p = 0$			$p = 1$		
n	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 0$	$\lambda = 1$	$\lambda = 10$
50	0.9532 (0.9139)	0.4298 (0.3137)	0.1145 (0.0324)	> 0.9999 (0.9939)	> 0.9999 (0.8919)	0.9941 (0.4217)
100	0.9536 (0.9233)	0.2709 (0.1599)	0.0365 (0.002)	> 0.9999 (0.9941)	0.9990 (0.7709)	0.7736 (0.2640)
200	0.9536 (0.9244)	0.1455 (0.0649)	0.0079 (0)	> 0.9999 (0.9947)	0.9721 (0.6161)	0.5448 (0.1393)
400	0.9543 (0.9250)	0.0638 (0.0155)	0.0011 (0)	> 0.9999 (0.9950)	0.8741 (0.4516)	0.3805 (0.0647)

5 DISCUSSION

We have found that the probability that the MLE or REML estimate of a variance component is zero is not well approximated by the theory of Self and Liang because their assumptions do not apply to most LMMs. They assume independence both under the null and under the alternative. For LMMs, the data are dependent, at least under the alternative.

There are a number of open problems that we intend to investigate. In particular, these results should be extended to models with more than one variance component and to null hypotheses that involve fixed effects as well variance components. For example, to test for no effect for one

Table 4: Probability of having a local and global minimum at $\lambda = 0$ for RLRT

	$p = 0$			$p = 1$		
n	$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$
50	0.6524 (0.6367)	0.5066 (0.4626)	0.1868 (0.1484)	0.6717 (0.6647)	0.6685 (0.6557)	0.6358 (0.6130)
100	0.6537 (0.6470)	0.3755 (0.3696)	0.0738 (0.0651)	0.6735 (0.6664)	0.6633 (0.6590)	0.5888 (0.5813)
200	0.6559 (0.6514)	0.2665 (0.2591)	0.0285 (0.0222)	0.6746 (0.6688)	0.6551 (0.6431)	0.5276 (0.5203)
400	0.6559 (0.6522)	0.1652 (0.1593)	0.0083 (0.0046)	0.6746 (0.6722)	0.6373 (0.6316)	0.4472 (0.4343)

Notes: The finite sample probability of having a global maximum (probability mass at zero of log-LR and RLRT respectively) is reported within brackets. It represents the frequency of estimating $\lambda = 0$ for different true values λ in 10,000 simulations from the model (26) where the mean regression function is modeled as a spline function with a fixed number of knots $K = 20$ as described by equation 27.

covariate in an additive model, we could model each covariate as a linear spline with its own variance component. The null hypothesis would be that the linear fixed effect and the variance component for the covariate of interest are both zero.

There are at least two practical implications of our work. First, the RLRT is preferred to the LRT because log-LR has so much probability mass at zero. Second, the chi-square mixture approximations that Self and Liang derived under independence should, in general, not be used for the distribution of log-RLR. At present, we recommend that critical values for RLRTs be calculated by simulation. Monte Carlo is, of course, somewhat time consuming and we are working on alternatives to simulation.

REFERENCES

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.

Davies, R.B., (1980). "The distribution of a linear combination of χ^2 random variables," *Applied Statistics*, 29, 323-333.

Farebrother, R.W., (1990). "The distribution of a quadratic form in normal variables," *Applied*

Statistics, 39, 294-309.

Harville, D.A., (1974). "Bayesian inference for variance components using only error contrasts," *Biometrika*, 61, 383-385.

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

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

Hoffman, K., Kunze, R., (1971). *Linear Algebra,* Prentice-Hall, Englewood Cliffs, New Jersey.

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

Pinheiro, J.C., Bates, D.M., (2000). *Mixed-Effects Models in S and S-Plus,* Springer-Verlag, New York.

Ruppert, D., 2002. "Selecting the number of knots for penalized splines," *Journal of Computational and Graphical Statistics*, to appear.

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.

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

Yu, H., Searle, S.R., and McCulloch, (1994). "Properties of maximum likelihood estimators of variance components in the one-way classification model, balanced data," *C.E. Communications in Statistics - Simulation and Computation*, 23, 897-914.

APPENDIX A1

Taking logarithms in the following identity (Harville, 1977)

$$|\boldsymbol{V}_\lambda|\,|\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}| = |\boldsymbol{X}^T\boldsymbol{X}|\,|\boldsymbol{I}_K + \lambda\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Z}|,$$

denoting by ω_i the i-th eigenvalue of the matrix $\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Z}$ we get, as in the proof for equation (7),

$$\frac{\partial}{\partial\lambda}\log|\boldsymbol{V}_\lambda| + \frac{\partial}{\partial\lambda}\log(|\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}|) = \sum_{i=1}^{n}\frac{\omega_i}{1+\lambda\omega_i}\,.$$

When we plug-in $\lambda = 0$ the right hand side of the equation becomes $\sum_{i=1}^n \omega_i$ which is equal to the trace of matrix $\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Z}$.

APPENDIX A2

If $\boldsymbol{Z}$ is the matrix corresponding to random effects in model described in equation (21) then $\boldsymbol{Z}\boldsymbol{Z}^T$ is an $nK \times nK$ having K $n \times n$ dimensional matrices of ones on the first diagonal, all other entries being zero. For any constant $\bar{d}$ one needs to compute the eigenvalues of the matrix $\boldsymbol{P}(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0) - \bar{d}\boldsymbol{P}(0)$ which are identical to the eigenvalues of $(\boldsymbol{Z}\boldsymbol{Z}^T - \bar{d}\boldsymbol{I}_{nK})\boldsymbol{P}(0)$, because $\boldsymbol{P}(0)$ is idempotent.

It is easy to see that $\boldsymbol{P}(0)$ has $nK - 1$ eigenvalues equal to one and 1 eigenvalue equal to zero. Also $\boldsymbol{Z}\boldsymbol{Z}^T - \bar{d}\boldsymbol{I}_{nK}$ has K eigenvalues equal to $n - \bar{d}$ and $nK - K$ eigenvalues equal to $-\bar{d}$. Because $\boldsymbol{P}(0)$ and $\boldsymbol{Z}\boldsymbol{Z}^T - \bar{d}\boldsymbol{I}_{nK}$ commute they are simultaneously diagonalizable and the eigenvalues of their product are the product of their eigenvalues (Hoffman and Kunze, 1971). But the null space of $\boldsymbol{P}(0) - 0\boldsymbol{I}_{nK}$ is one dimensional and an $nK \times 1$ dimensional vector v of ones is a basis in this space. Moreover, the vector v is an eigenvector of the matrix $\boldsymbol{Z}\boldsymbol{Z}^T - \bar{d}\boldsymbol{I}_{nK}$ corresponding to the eigenvalue $n - \bar{d}$. It follows that the the matrix of interest has $K - 1$ eigenvalues equal to $n - \bar{d}$, one eigenvalue equal to zero, and $nK - K$ eigenvalues equal to $-\bar{d}$.

APPENDIX A3

To compute the asymptotic probability of having a local minimum at $\lambda = 0$ one needs to determine the asymptotic behavior the eigenvalues of the matrix $\boldsymbol{M} = \boldsymbol{P}(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0) - \bar{d}\boldsymbol{P}(0)$. $\boldsymbol{P}(0)$ is the projection matrix on the space $\mathrm{col}(\boldsymbol{X})^\perp$ which is the orthogonal space on the space $\mathrm{col}(\boldsymbol{X})$ spanned by the columns of the matrix $\boldsymbol{X}$ corresponding to fixed effects. Clearly

$$\boldsymbol{P}(0)u = 0 \ \text{ if } \ u \in \mathrm{col}(\boldsymbol{X}) \quad \text{and} \quad \boldsymbol{P}(0)u = u \ \text{ if } \ u \in \mathrm{col}(\boldsymbol{X})^\perp\,.$$

Assuming that the columns of matrix $\boldsymbol{X}$ are linearly independent, it follows that 0 is an eigenvalue of the matrix $\boldsymbol{M}$ of algebraic multiplicity $p+1$ corresponding to a set of $p+1$ orthonormal eigenvectors in the space $\text{col}(\boldsymbol{X})$. Suppose now that $u \in \text{col}(\boldsymbol{X})^{\perp}$ is an eigenvector corresponding to a eigenvalue $\nu \neq 0$. It follows that

$$\boldsymbol{P}(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)u = (\nu + \bar{d})u \ ,$$

which shows that u is an eigenvector of $\boldsymbol{P}(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)$ corresponding to the eigenvalue $\nu + \bar{d}$. But the rank of the matrix $\boldsymbol{P}(0)\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}(0)$ is K meaning that exactly K eigenvalues are different from zero and $n-K$ are zero. Denoting by $\mu_1(n), \dots, \mu_K(n)$ the nonzero eigenvalues it follows that u is the eigenvector corresponding to one of the following eigenvalues $\mu_1(n) - \bar{d}, \dots, \mu_K(n) - \bar{d}$ or $-\bar{d}$. If $\mu_1, \dots, \mu_K$ are distinct it can be proved that the algebraic multiplicity of $-\bar{d}$ is $n - p - K - 1$. The advantage of this decomposition is that $\mu_1(n), \dots, \mu_K(n)$ are also the eigenvalues of the $K \times K$ dimensional matrix $\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Z}$. The asymptotic behavior of this matrix determines the asymptotic behavior of its eigenvalues.

When we test for a constant mean against a general alternative modeled by a piecewise constant spline with K knots, the matrix $\boldsymbol{M}$ has 0 as eigenvalue with multiplicity 1, $-\bar{d}$ with multiplicity $n - K - 1$, and $\mu_1(n), \dots, \mu_K(n)$ eigenvalues with multiplicity 1 of the matrix $\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Z}$. If we denote by $n_k(n)$ the number of points bigger than the k-th knot and assume that

$$\lim_{n\to\infty} \frac{n_k(n)}{n} = p_k$$

one can show that

$$\lim_{n\to\infty} \frac{\boldsymbol{Z}^T\boldsymbol{P}(0)\boldsymbol{Z}}{n} = \boldsymbol{M}(0) \ ,$$

where the (i, j)-th entry of the matrix $\boldsymbol{M}(0)$ is

$$m_{ij} = p_{i \vee j} - p_i p_j \ .$$

Denoting by $\mu_1(0), \dots, \mu_K(0)$ the eigenvalues of $\boldsymbol{M}(0)$ it follows that

$$\lim_{n\to\infty} \frac{\mu_k(n)}{n} = \mu_k(0) \ .$$

The asymptotic probability of having a local minimum at 0 is given by

$$p(n, K) = P\left(\sum_{k=1}^{K} \frac{\mu_k(n) - \bar{d}}{n} v_k^2 \leq \frac{\bar{d}}{n} \sum_{k=K+1}^{n} v_k^2 \right) \ ,$$

where v_k are i.i.d $N(0,1)$ random variables. Assuming that $\lim_{n\to\infty} \bar{d}/n = 0$ we get that

$$\lim_{n\to\infty} p(n,K) = P\left(\sum_{k=1}^{K} \mu_k(0) v_k^2 \leq \bar{d}\right) .$$

Figure 1: one-way ANOVA model: asymptotic probability of having a local minimum at $\lambda = 0$ when the true value of λ is $\lambda = 0$. The number of levels is constant and the number of observations per level goes to infinity.

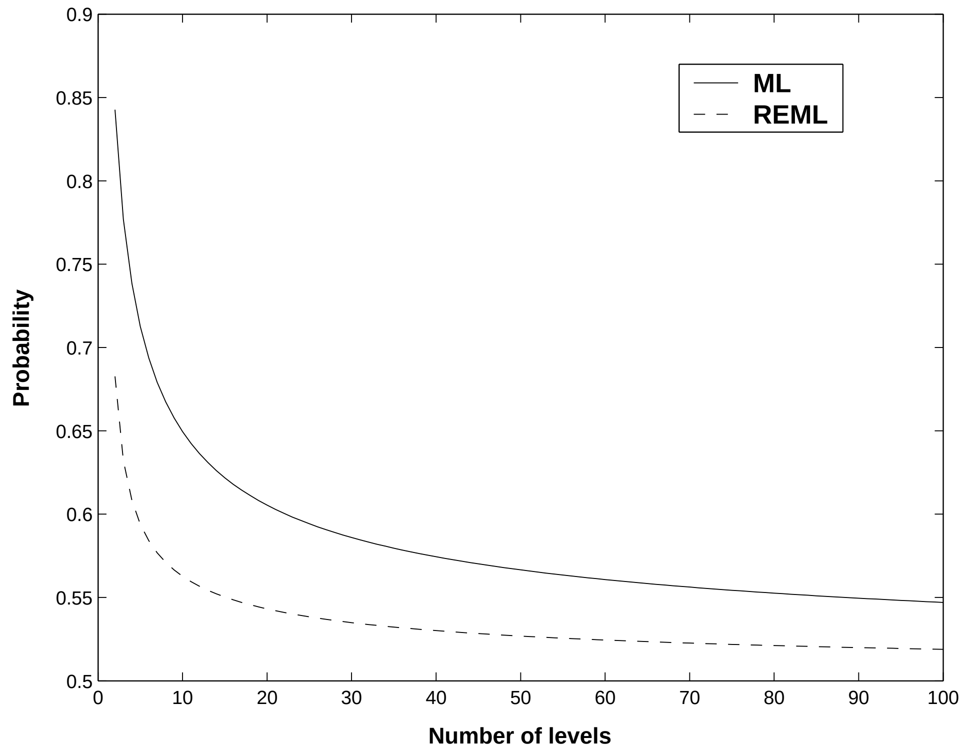

Figure 2: Non-parametric testing for constant mean vs. a general alternative modeled by a piecewise constant spline: asymptotic probability of having a local minimum at $\lambda = 0$ when the true value of λ is $\lambda = 0$. The number of knots is constant and the number of observations goes to infinity.