Lecture Notes in Statistics 192

Edited by P. Bickel, S. Fienberg

David B. Dunson
Editor

Random Effect and Latent Variable Model Selection

Springer

Editor
David B. Dunson
National Institute of Environmental Health Sciences
Research Triangle Park, NC
USA
dunson@stat.duke.edu

ISBN: 978-0-387-76720-8 e-ISBN: 978-0-387-76721-5
DOI: 10.1007/978-0-387-76721-5

Library of Congress Control Number: 2008928920

Cover illustration: Follicles of colloid in thyroid

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface
Random Effect and Latent Variable Model Selection

In recent years, there has been a dramatic increase in the collection of multivariate and correlated data in a wide variety of fields. For example, it is now standard practice to routinely collect many response variables on each individual in a study. The different variables may correspond to repeated measurements over time, to a battery of surrogates for one or more latent traits, or to multiple types of outcomes having an unknown dependence structure. Hierarchical models that incorporate subject-specific parameters are one of the most widely-used tools for analyzing multivariate and correlated data. Such subject-specific parameters are commonly referred to as random effects, latent variables or frailties.

There are two modeling frameworks that have been particularly widely used as hierarchical generalizations of linear regression models. The first is the linear mixed effects model (Laird and Ware , 1982) and the second is the structural equation model (Bollen , 1989). Linear mixed effects (LME) models extend linear regression to incorporate two components, with the first corresponding to fixed effects describing the impact of predictors on the mean and the second to random effects characterizing the impact on the covariance. LMEs have also been increasingly used for function estimation. In implementing LME analyses, model selection problems are unavoidable. For example, there may be interest in comparing models with and without a predictor in the fixed and/or random effects component. In addition, there is typically uncertainty in the subset of predictors to be included in the model, with the number of candidate predictors large in many applications.

To address problems of this type, it is not appropriate to rely on classical methods developed for model selection and inferences in non-hierarchical regression models. For example, the widely used BIC criteria are not valid for random effects models, and likelihood ratio and score tests face difficulties, since the null hypothesis often falls on the boundary of the parameter space. The objective of the first part of this book is to provide an overview of a variety of promising strategies for addressing model selection problems in LMEs and related modeling frameworks.

In the chapter, “Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models,” Ciprian Crainiceanu provides an applications-motivated overview of recent work on likelihood ratio and restricted likelihood ratio tests for testing whether random effects have zero variance. The approaches he describes represent an important advance over the current standard practice in testing for zero variance components in hierarchical models. Such approaches include ignoring the boundary problem and assuming the likelihood ratio test statistic has a chi-square distribution under the null and relying on asymptotic results showing a mixture of chi-squares is more appropriate (Stram and Lee, 1994). Crainiceanu shows that asymptotic approximations may be unreliable in many applications, motivating use of finite sample approaches. He illustrates the ideas through several examples, including applications to nonlinear regression modeling.

Score tests provide a widely-used alternative to likelihood ratio tests, and in the chapter, "Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and Other Related Topics," of this volume Daowen Zhang and Xihong Lin provide an excellent overview of the recent literature on score test-based approaches. In addition, Zhang and Lin consider a broader class of models, which includes GLMMs and generalized additive mixed models (GAMMs). GAMMs provide an extremely rich framework for semiparametric modeling of longitudinal data allowing flexible predictor effects through replacing linear terms in a generalized linear model with unknown non-linear functions, while also including random effects to account for within-subject dependence and heterogeneity.

The first part of the volume is completed with two companion chapters describing Bayesian approaches for variable selection in LMEs and GLMMs. The likelihood ratio and score test methods provide an approach for comparing two nested models with the smaller model having a random effect excluded. However, in many applications one is faced with a set of p candidate predictors, with uncertainty in which subsets should be included in the fixed and random effects components of the model. Clearly, the number of candidate models grows extremely rapidly with p, so that it often becomes impossible to fit each model in the list. One possibility is to use a likelihood ratio test within a stepwise selection procedure. However, the final model selected will depend on the order in which candidate predictors are added or deleted and it is difficult to adjust for uncertainty in subset selection in performing inferences and predictions. In non-hierarchical regression models, Bayesian variable selection implemented with stochastic search algorithms has been very widely used to address this problem. In the chapter, "Bayesian Model Uncertainty in Mixed Effects Models," Satkartar Kinney and I describe an approach for LMEs, while in the chapter, "Bayesian Variable Selection in Generalized Linear Mixed Models," Bo Cai and I describe an alternative for GLMMs.

The second part of the book switches gears to focus on structural equation models (SEMs), which have been very widely used in social science applications for assessing relationships among latent variables, such as poverty or violence, that can only be measured indirectly through multiple surrogates. SEMs provide a generalization of factor analysis, which allows for modeling of linear relationships among the latent factors through a linear structural relations (LISREL) model. SEMs are also quite useful outside of traditional application areas for sparse covariance structure modeling of high-dimensional multivariate data. However, one of the main issues in applying SEMs is how to deal with model uncertainty, which commonly arises in deciding on the number of factors to include in each component and the relationships among these factors. In the chapter, “A Unified Approach to Two-Level Structural Equation Models and Linear Mixed Effects Models,” Peter Bentler and Jiajuan Liang provide a bridge between the first and second parts of the volume in linking LMEs and SEMs, while also considering methods for model selection.

In the chapter, “Bayesian Model Comparison of Structural Equation Models,” Sik-Yum Lee and Xin-Yuan Song provide a general Bayesian approach to comparison of SEMs. Typical Bayesian methods for comparing models rely on Bayes factors. However, Bayes factors have proved quite difficult to estimate accurately in SEMs. Lee and Song propose a useful and clever solution to this problem using path sampling. One well-known issue in model selection using Bayes factors is sensitivity to prior selection. This has motivated a rich literature on default priors. In the chapter, “Bayesian Model Selection in Factor Analytic Models” Joyee Ghosh and I build on the approach of Lee and Song, proposing a default prior, and an efficient approach for posterior computation relying on parameter expansion. In addition, an importance sampling algorithm is proposed as an alternative to path sampling.

In summary, this volume provides a practically-motivated overview of a variety of recently proposed approaches for model selection in random effects and latent variable models. The goal is to make these methods more accessible to practitioners, while also stimulating additional research in this important and under-studied area of statistics. There are a number of topics related to model selection in random effects and latent variable models that are in need of new research, with solutions having the potential for substantial applied impact. The first topic is the development of simple methods to calculate model selection criteria, which modify AIC and BIC to incorporate a penalty for model complexity that is appropriate for a hierarchical model. A second topic is the development of efficient methods for simultaneous model search and posterior computation in SEMs. Often, one has a high-dimensional set of SEMs that are plausible a priori and consistent with current scientific or sociologic theories. It is of substantial interest to identify high posterior probability models and to average across models in making predictions. However, typical tricks used in other model classes, such as zeroing out coefficients, do not work in general for SEMs, and efficient alternatives remain to be developed.

References

Bollen, K.A. (1989). *Structural Equation Models with Latent Variables*. New York: Wiley

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974

David B. Dunson

Contents

Part I Random Effects Models

Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models 3
Ciprian M. Crainiceanu

Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and other Related Topics 19
Daowen Zhang and Xihong Lin

Bayesian Model Uncertainty in Mixed Effects Models 37
Satkartar K. Kinney and David B. Dunson

Bayesian Variable Selection in Generalized Linear Mixed Models 63
Bo Cai and David B. Dunson

Part II Factor Analysis and Structural Equations Models

A Unified Approach to Two-Level Structural Equation Models and Linear Mixed Effects Models 95
Peter M. Bentler and Jiajuan Liang

Bayesian Model Comparison of Structural Equation Models 121
Sik-Yum Lee and Xin-Yuan Song

Bayesian Model Selection in Factor Analytic Models 151
Joyee Ghosh and David B. Dunson

Index 165

Part I
Random Effects Models

Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models

Ciprian M. Crainiceanu

Mixed models are a powerful inferential tool with a wide range of applications including longitudinal studies, hierarchical modeling, and smoothing. Mixed models have become the state of the art for statistical information exchange and correlation modeling. Their popularity has been augmented by the availability of dedicated software, e.g., the `MIXED` procedure in SAS, the `lme` function in R and S+, or the `xtmixed` function in STATA.

In this paper, we consider the problem of testing the null hypothesis of a zero variance component in a linear mixed model (LMM). We focus on the likelihood ratio test (LRT) and restricted likelihood ratio test (RLRT) statistics for three reasons. First, (R)LRTs are uniformly most powerful for simple null and alternative hypotheses and have been shown to have good power properties in a variety of theoretical and applied frameworks. Second, given their robust properties, (R)LRTs are the benchmark for statistical testing. Third, (R)LRT can now be used in realistic data sets and applications due to a better understanding of their null distribution and improved computational tools.

The paper is organized as follows. Section 1 describes three applications of testing for a zero variance component. Section 2 contains the model and a description of the testing framework. Section 3 describes standard asymptotic results and provides a short discussion of their applicability. Section 4 presents finite sample and asymptotic results for linear mixed models (LMMs) with one variance component. Section 5 introduces two approximations of the finite sample (R)LRT distribution for testing for zero variance components in LMMs with multiple variance components. Section 6 presents the corresponding testing results for the examples introduced in Sect. 1. Section 7 provides the discussion and practical recommendations.

C.M. Crainiceanu
Department of Biostatistics, Johns Hopkins University
ccrainic@jhsph.edu

D. B. Dunson (ed.) *Random Effect and Latent Variable Model Selection*,
DOI: 10.1007/978-0-387-76721-5, © Springer Science+Business Media, LLC 2008

1 Examples

The three examples in this section illustrate the wide variety of applications of testing for zero variance components in LMMs. This list is far from being exhaustive but provides a foretaste of what is possible and needed in this framework.

1.1 Loa loa *Prevalence in West Africa*

Figure 1 displays village locations from one of the several parasitological survey location in West Africa. In all these villages parasitological sampling was conducted to assess the prevalence of Loaisis. Here we provide a short summary, but a complete description of the problem can be found in Crainiceanu et al. (2007). Loaisis, or eyeworm, is an endemic disease of the wet tropics, caused by *Loa loa*, a filarial parasite which is transmitted to humans by the bite of an infected *Chrysops* fly. In Fig. 1 the empirical prevalence rates at location $\boldsymbol{x}$, $\widehat{p}(\boldsymbol{x})$, are indicated as dots coded according to their size: small $\widehat{p}(\boldsymbol{x}) < 0.18$, medium $0.18 \leq \widehat{p}(\boldsymbol{x}) < 0.20$, large $0.20 \leq \widehat{p}(\boldsymbol{x}) < 0.25$, and very large $\widehat{p}(\boldsymbol{x}) > 0.30$.

A complete bivariate binomial analysis of this data set can be found in Crainiceanu et al. (2007). Here, we consider the following simpler univariate model for the logit prevalence at the spatial location $\boldsymbol{x}$

$$\text{logit}\{\widehat{p}(\boldsymbol{x})\} = \alpha_0 + \alpha_1 g(\boldsymbol{x}) + \alpha_2 s(\boldsymbol{x}) + \alpha_3 e(\boldsymbol{x}) + \alpha_4 \{e(\boldsymbol{x}) - 800\}_+ + S(\boldsymbol{x}) + \epsilon(\boldsymbol{x}), \quad (1)$$

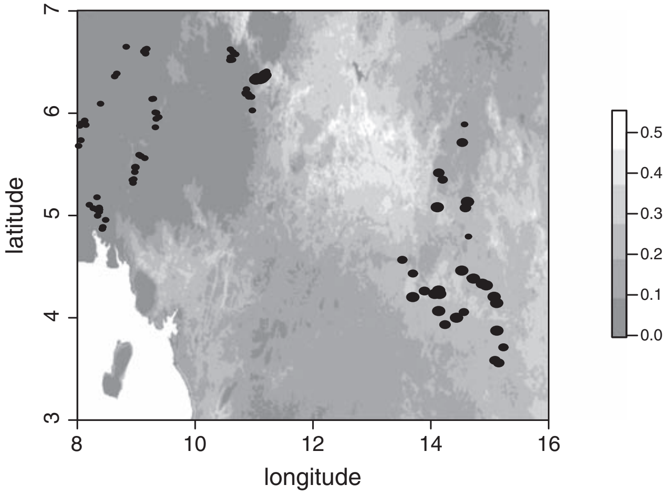

Fig. 1 Village sampling locations in one subregion from West Africa. The empirical prevalence rates are indicated as *dots* coded according to their size: small $\widehat{p}(\boldsymbol{x}) < 0.18$, medium $0.18 \leq \widehat{p}(\boldsymbol{x}) < 0.20$, large $.20 \leq \widehat{p}(\boldsymbol{x}) < 0.25$, very large $\widehat{p}(\boldsymbol{x}) > 0.30$. The estimated mean prevalence based on model (1) is grey-scale coded according to the legend

where $g(\boldsymbol{x})$ is an annual average measure of greenness, $s(\boldsymbol{x})$ the standard deviation of greenness, $e(\boldsymbol{x})$ the elevation in meters, $S(\boldsymbol{x})$ a spatial component, and $\epsilon(\boldsymbol{x}) \sim \text{Normal}(0, \sigma_\epsilon^2)$ are the independent errors. Here a_+ is equal to a if $a > 0$ and 0 otherwise, so that $\{e(\boldsymbol{x}) - 800\}_+$ represents the elevation at location $\boldsymbol{x}$ truncated below 800 m. If the spatial component $S(\boldsymbol{x})$ is modeled as a low rank penalized thin plate spline then

$$\begin{cases} S(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{x})\boldsymbol{b}, \\ \boldsymbol{b} \quad\;\; \sim \text{Normal}(0, \sigma_b^2 I_K), \end{cases} \tag{2}$$

where $Z(\boldsymbol{x})$ is the low rank specific design vector (for details see (Ruppert et al., 2003; Kammann and Wand, 2003)), b the thin plate spline coefficients describing the spatial features of $S(\boldsymbol{x})$, σ_b^2 the smoothing parameter controlling the amount of smoothing, and I_K is the identity matrix where K is the number of spatial knots.

In the case of low rank smoothers the set of K knots for the covariates have to be chosen. One possibility is to use equally spaced knots. Another possibility is to select the knots and subknots using the *space filling* design (Nychka and Saltzman, 1998), which is based on the maximal separation principle. This avoids wasting knots and is likely to lead to better approximations in sparse regions of the data. The `cover.design()` function from the R package `Fields` (Fields Development Team, 2006) provides software for space filling knot selection.

If the smoothing parameter is estimated by restricted maximum likelihood (REML), then the model described in (1) and (2) is equivalent to a particular LMM with one variance component. Figure 1 displays the estimated mean prevalence at all locations in the map coded according to the legend. In this context testing whether the nonlinear spatial component of $S(\boldsymbol{x})$ is necessary to explain the residual variability after fitting the scientifically available covariates is equivalent to testing

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

From a scientific perspective testing H_0 is equivalent to testing whether simpler models including only covariates could capture the complex stochastic nature of the spatial data and have good predictive power.

1.2 Onion Density in Australia

Figure 2 contains data on yields (grams/plant) of white Spanish onions in two locations: Purnong Landing and Virginia, South Australia (Ratkowsky, 1983). The horizontal axis corresponds to areal density of plants (plants/m^2). Detailed analyses of these data are given by Ruppert et al. (2003) and Crainiceanu (2003). Denote by (y_i, x_i, s_i) the yield, density of plants and location for the ith observation. Here, $s_i = 1$ corresponds to Purnong Landing and $s_i = 0$ corresponds to Virginia. The solid lines in Fig. 2 correspond to fitting the linear additive model

$$\log(y_i) = \beta_0 + \beta_1 s_i + \beta_2 d_i + \epsilon_i \,. \tag{3}$$

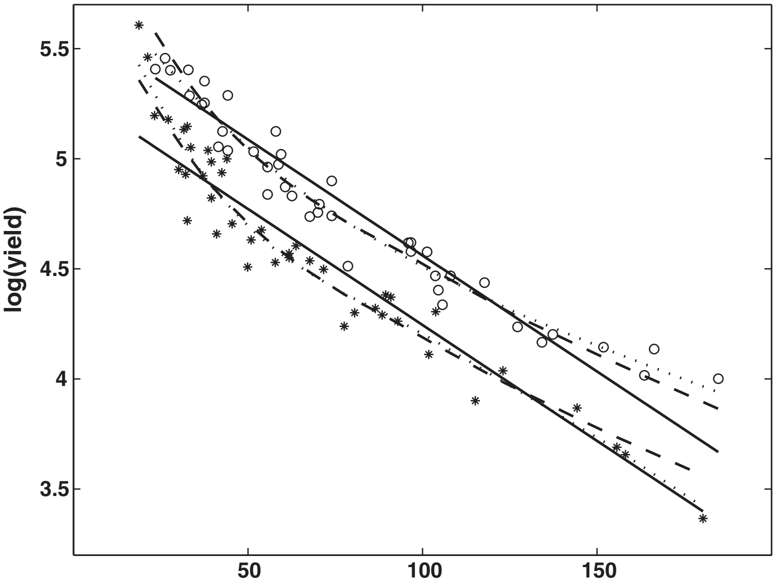

Fig. 2 Log yield for the onion data plotted against density (*circle* Purnong Landing; *asterisk* Virginia), straight line fit (*solid line*), binary offset model using a penalized linear spline fit with $K = 15$ knots and REML estimation of smoothing parameter (*dashed line*), discrete by continuous interaction model (*dotted line*)

The dashed lines represent the mean fit using a *semiparametric binary offset model* (Ruppert et al., 2003)

$$\log(y_i) = \beta_1 s_i + f(d_i) + \epsilon_i, \tag{4}$$

which contains a parametric component, $\beta_1 s_i$, and a nonparametric component, $f(d_i)$. The binary variable s vertically offsets the relationship between $E[\log(y)]$ and density according to location. By specifying a linear penalized spline model for $f(d_i)$ the model becomes

$$\log(\mathrm{y}_i) = \beta_0 + \beta_1 s_i + \beta_2 d_i + \sum_{k=1}^{K} b_k (d_i - \kappa_k)_+ + \epsilon_i,$$

where b_k are i.i.d. $N(0, \sigma_b^2)$ and ϵ_i are i.i.d. $N(0, \sigma_\epsilon^2)$. Following Ruppert et al. (2003), we use $K = 15$ knots chosen at the sample quantiles of density corresponding to frequencies $1/(K+1), \ldots, K/(K+1)$.

Testing model (3) corresponding to the solid line fits in Fig. 2 versus model (4) corresponding to dashed lines in Fig. 2 corresponds to testing $H_0 : \sigma_b^2 = 0$ vs. $H_A : \sigma_b^2 > 0$. For these data and hypothesis testing framework, Crainiceanu (2003) calculated RLRT $= 35.93$ with a corresponding p-value < 0.001. The calculation of the p-value was based on the exact distribution of the RLRT as obtained by Crainiceanu and Ruppert (2004b). This result is not surprising, given the large discrepancies between the two model fits in Fig. 2. In fact, results would not change even if one used the more conservative (but incorrect in this case) $0.5\chi^2_0 : 0.5\chi^2_1$ approximation to the null RLRT distribution (Self and Liang, 1987).

It is natural, however, to ask whether the binary offset model accurately represents the data. To address this question we nest model (4) into the following discrete by continuous interaction model

$$E\{\log(y_i)\} = \begin{cases} f_{\text{PL}}(d_i) & \text{if } s_i = 1; \\ f_{\text{VA}}(d_i) & \text{if } s_i = 0, \end{cases}$$

where the subscripts PL and VA denote the Purnong Landing and Virginia locations, respectively. The basic idea is to model the mean response at one of the locations, say Purnong Landing, as a nonparametric spline and the deviations from this function corresponding to the other location, say Virginia, as another nonparametric spline. The discrete by continuous interaction model is

$$\log(y_i) = \beta_0 + \beta_1 d_i + \sum_{k=1}^{K} b_k (d_i - \kappa_k)_+ + \{\gamma_0 + \gamma_1 d_i + \sum_{k=1}^{K} \upsilon_k (d_i - \kappa_k)_+\} I(i \in \text{PL}) + \epsilon_i \tag{5}$$

for Virginia ($s = 0$), where β_0, β_1, γ_0, and γ_1 are fixed unknown parameters, b_k are i.i.d. $N(0, \sigma^2_b)$, υ_k are i.i.d. $N(0, \sigma^2_\upsilon)$, and $I(i \in \text{PL})$ is 1 if the observation i is from Purnong Landing and 0 otherwise. The model (5) is an LMM with two random effects variance components, σ^2_b and σ^2_υ, and the fit to the data is depicted by the two dotted curves in Fig. 2. Testing for linear versus nonlinear deviations from the smooth regression function corresponding to the Purnong Landing location reduces in this model to testing

$$H_0 : \sigma^2_\upsilon = 0 \quad \text{vs.} \quad \sigma^2_\upsilon > 0,$$

which is equivalent to testing for a zero variance component in an LMM with two variance components. After discussing the state of the art in statistical testing in this framework we will revisit this example in Sect. 6.

1.3 Coronary Sinus Potassium

We consider the coronary sinus potassium concentration data measured on 36 dogs published by Grizzle and Allan (1969) and Wang (1998). The measurements on each dog were taken every 2 min from 1 to 13 min (seven observations per dog). The 36 dogs come from four treatment groups. Figure 3 displays the data for the nine dogs in the first treatment group (dotted lines).

If y_{ij} denotes the jth concentration for the ith dog at time $t_{ij} = 1 + 2j$ then a reasonable LMM model for the first treatment group is

$$y_{ij} = \beta_0 + u_i + \beta_1 t_{ij} + \beta_2 t^2_{ij} + \epsilon_{ij}, \tag{6}$$

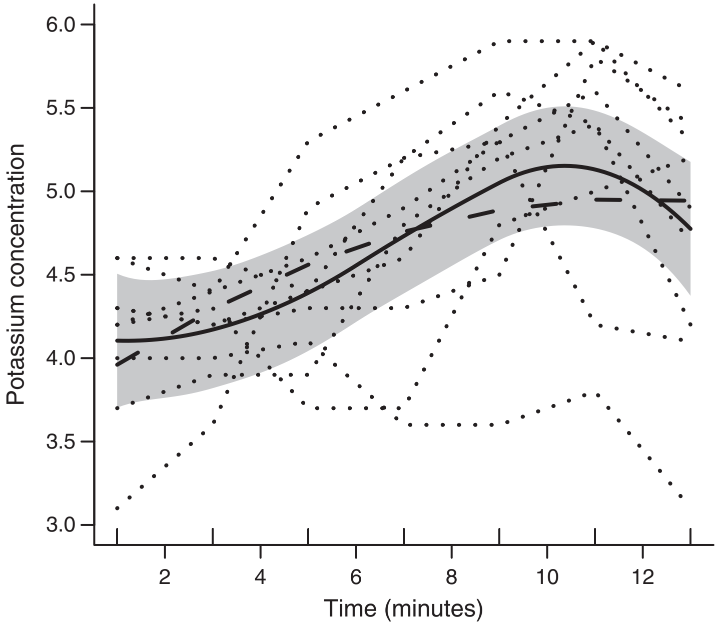

Fig. 3 Sinus potassium concentration for nine dogs in the first treatment group (*dotted lines*)

where $u_i \sim N(0, \sigma_u^2)$ are independent dog specific intercepts and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ are independent errors. Figure 3 displays the fit of model (6) as a dashed line. It is natural to ask the question whether model (6) is enough to capture the complexity of the population mean function. One way to answer this question is by embedding model (6) into the following more general model

$$y_{ij} = \beta_0 + u_i + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \sum_{i=1}^{K} b_k (t_{ij} - \kappa_k)_+^2 + \epsilon_{ij}, \tag{7}$$

where $b_k \sim N(0, \sigma_b^2)$ are independent truncated spline coefficients, K the number of knots and $\kappa_k, k = 1, \ldots, K$ are the knots. All the other assumptions are the same as in model (6). Note that model (6) is an LMM with two variance components: one, σ_u^2, controlling the shrinkage of random intercepts towards their mean and the other one, σ_b^2, controlling the shrinkage of the population function towards a quadratic polynomial. Figure 3 displays the fit of this model as a solid line together with 95% pointwise confidence intervals (shaded area).

Testing the null hypothesis described by model (6) versus the alternative described by model (7) is equivalent to testing for

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

Similarly, testing for dog response homogeneity is equivalent to testing

$$H_0:\ \sigma_u^2 = 0 \quad \text{vs.} \quad \sigma_u^2 > 0.$$

Both frameworks correspond to testing for a zero variance component in an LMM with two variance components.

As the last point for this example, note that a naive way to test for $H_0 : \sigma_b^2 = 0$ is to check whether the null fit is contained in the shaded area. This may seem like a good idea, but leads to incorrect inferences. Indeed, all the confidence intervals for the mean function based on model (7) contain the fit based on model (6). However, as we show in Sect. 6, the RLRT indicates strong evidence against the null hypothesis of a quadratic population curve.

2 Model and Testing Framework

All examples in Sect. 1, and many others, involve testing for a zero variance component as the methodological answer to important scientific questions. To formalize the framework, let us assume that the outcome vector, $\boldsymbol{Y}$, is modeled as an LMM

$$\begin{cases} \boldsymbol{Y} &= \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{b}_1 + \cdots + \boldsymbol{Z}_S\boldsymbol{b}_S + \boldsymbol{\epsilon}, \\ \boldsymbol{b}_s &\sim N(\boldsymbol{0}, \sigma_s^2\boldsymbol{I}_{K_s}), \quad s = 1, \ldots, S, \\ \boldsymbol{\epsilon} &\sim N(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{I}_n). \end{cases} \tag{8}$$

Here the random effects $\boldsymbol{b}_s$, $s = 1, \ldots, S$, and the error vector $\boldsymbol{\epsilon}$ are mutually independent, K_s denotes the number of columns in $\boldsymbol{Z}_s$, n the sample size, and $\boldsymbol{I}_\nu$ denotes the identity matrix with ν columns. This is not the most general form of an LMM, but it is often used in practice and keeps the presentation simple.

We are interested in testing

$$H_{0,s} : \sigma_s^2 = 0 \quad \text{vs.} \quad H_{A,s} : \sigma_s^2 > 0, \tag{9}$$

where the hypotheses are indexed by $s = 1, \ldots, S$ to emphasize that these are distinct and not joint hypotheses for all variance components. Note that because $\boldsymbol{b}_s \sim N(\boldsymbol{0}, \sigma_s^2\boldsymbol{I}_{K_s})$, the null hypothesis is equivalent to $\boldsymbol{b}_s = \boldsymbol{0}$, indicating that under the null the component $\boldsymbol{Z}_s\boldsymbol{b}_s$ of model (8) is zero.

Denote by $\boldsymbol{\theta}_{-s}$ all the parameters in model (8) with the exception of σ_s^2. The RLRT for testing $H_{0,s}$ is then defined as

$$\text{RLRT} = 2\text{sup}_{\theta_{-s},\sigma_s^2}\{\log L(\theta_{-s}, \sigma_s^2)\} - 2\text{sup}_{\theta_{-s}}\{\log L(\theta_{-s}, 0)\},$$

where $L(\theta_{-s}, \sigma_s^2)$ is the restricted likelihood function for model (8). A similar definition holds for LRT using the likelihood instead of the restricted likelihood function.

3 Standard Asymptotic Results for LMMs

Testing for zero variance components is not new in mixed models. Using theory originally developed by Chernoff (1954), Moran (1971), and Self and Liang (1987), Stram and Lee (1994) proved that the LRT for testing (9) has an asymptotic $0.5\chi_0^2 : 0.5\chi_1^2$ mixture distribution under the null hypothesis $H_{0,s}$ if data are independent and identically distributed *both under the null and alternative hypothesis*. For more details on standard asymptotic results, see the chapter by Zhang and Lin (2007) in this book. Thus, it could be surprising that in many applications the null distribution of the LRT using simulations is far from being a $0.5\chi_0^2 : 0.5\chi_1^2$ mixture.

There are several reasons for these inconsistencies. First, the Laird and Ware (1982) model used by Stram and Lee (1994) allows the partition of the outcome vector $\boldsymbol{Y}$ into independent subvectors. This could be revealed by close inspection of this model, which is typically described in terms of the subject-level vector $\boldsymbol{Y}_i$ and not in terms of the data vector $\boldsymbol{Y}$. The independence assumption is violated, for example, when representing nonparametric smoothing as a particular LMM. Second, even when the outcome vector can be partitioned into independent subvectors, the number of subvectors may not be sufficient to ensure an accurate asymptotic approximation. Third, subvectors may not be identically distributed due to unbalanced designs or missing data. In the case of an LMM with one variance component ($S = 1$) Crainiceanu and Ruppert (2004b) and Crainiceanu et al. (2005) have derived the finite sample and asymptotic distribution of the LRTs showing that, under general conditions, the null distribution for testing $H_{0,s}$ is typically different from $0.5\chi_0^2 : 0.5\chi_1^2$. In the following section, we provide a summary of these results and discuss the implications for applied statistical inference.

4 Finite Sample and Asymptotic Results for General Design LMMs with One Variance Component

Consider the particular case of model (8) with Gaussian outcome vector and one variance component

$$\begin{cases} \boldsymbol{Y} &= \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{b}_1 + \boldsymbol{\varepsilon}, \\ \boldsymbol{b}_1 &\sim N(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I}_{K_1}), \\ \boldsymbol{\varepsilon} &\sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n), \end{cases} \tag{10}$$

where $\boldsymbol{b}_1$ and $\boldsymbol{\varepsilon}$ as mutually independent.

As model (10) has only one variance component, σ_1^2, the exact null distribution of the RLRT for testing $H_{0,1} : \sigma_1^2 = 0$ versus $H_{A,1} : \sigma_1^2 > 0$ is Crainiceanu and Ruppert (2004b)

$$\mathrm{RLRT}_n \overset{d}{=} \sup_{\lambda \geq 0} \left\{ (n-p)\log\left[1 + \frac{N_n(\lambda)}{D_n(\lambda)}\right] - \sum_{l=1}^{K_1} \log(1 + \lambda\mu_{l,n}) \right\}, \tag{11}$$

where "$\overset{d}{=}$" denotes equality in distribution, p is the number of columns in X,

$$N_n(\lambda) = \sum_{l=1}^{K_1} \frac{\lambda \mu_{l,n}}{1 + \lambda \mu_{l,n}} w_l^2, \quad D_n(\lambda) = \sum_{l=1}^{K_1} \frac{w_l^2}{1 + \lambda \mu_{l,n}} + \sum_{l=K_1+1}^{n-p} w_l^2,$$

$w_l, l = 1, \ldots, n - p$, are independent $N(0, 1)$, and $\mu_{l,n}, l = 1, \ldots, K_1$, are the eigenvalues of the $K_1 \times K_1$ matrix $\boldsymbol{Z_1}'(\boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Z_1}$. The asymptotic distribution of the LRT was also derived by Crainiceanu and Ruppert (2004b) and depends essentially on the asymptotic geometry of the eigenvalues $\mu_{l,n}$. This distribution may or may not be equal to the $0.5\chi_0^2 : 0.5\chi_1^2$ mixture, depending on the asymptotic behavior of these eigenvalues. A similar result for LRT can be found in Crainiceanu and Ruppert (2004b).

There are several reasons for preferring the distribution in (11) over the $0.5\chi_0^2 : 0.5\chi_1^2$ of Stram and Lee (1994). First, this is the finite sample distribution of the RLRT. Second, the $0.5\chi_0^2 : 0.5\chi_1^2$ asymptotic distribution can be inaccurate when the number of independent sub-vectors of $\boldsymbol{Y}$ is small to moderate or when designs are unbalanced. Typically, the $0.5\chi_0^2 : 0.5\chi_1^2$ provides a conservative approximation of the finite sample distribution with considerable associated losses in power. Third, calculating the distribution in (11) is very fast. Indeed, the distribution in (11) depends only on the eigenvalues $\mu_{l,n}$ of a $K_1 \times K_1$ matrix, which need to be computed only once. Simulation effectively reduces to simulation of $(K_1 + 1)$ χ^2 variables and a grid search over λ. This simulation does not depend on the sample size, n, and is fast (5,000 simulations per second with a 2.66 GHz CPU and 1 Mbyte random access memory). Fourth, when assumptions in Stram and Lee (1994) hold the distribution in (11) converges weakly to the asymptotic $0.5\chi_0^2 : 0.5\chi_1^2$.

5 Linear Mixed Models with Multiple Variance Components

The results in Crainiceanu and Ruppert (2004b) have solved the problem for mixed models with Gaussian outcomes and one variance component. However, in many practical applications there are multiple variance components controlling shrinkage. Two such examples are the onion density and the coronary sinus potassium models in Sects. 1.2 and 1.3, respectively.

The methodology developed by Crainiceanu and Ruppert (2004b) could be used to derive the null distribution for the more general case discussed in this paper. While the result is theoretically interesting, this distribution is obtained by maximizing a stochastic process over the variance components of model (8), which makes the implementation computationally equivalent to the parametric bootstrap. For this reason, Crainiceanu (2003) and Crainiceanu and Ruppert (2004a) suggest using the parametric bootstrap in this context. One could debate the elegance of this approach, but the parametric bootstrap is a practical and robust alternative to the $0.5\chi_0^2 : 0.5\chi_1^2$ approximation.

One problem with the parametric bootstrap is that, in many applications, evaluating the likelihood is computationally expensive and it may not be reasonable to perform thousands of simulations. To illustrate this problem, consider the following simple longitudinal model:

$$Y_{ij} = u_i + f(x_{ij}) + \epsilon_{ij}, \tag{12}$$

where $u_i \sim N(0, \sigma_u^2)$ are random independent subject specific intercepts, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ are independent errors, $i = 1, \ldots, I$, $j = 1, \ldots, J$, I is the number of subjects and J is the number of observations per subject. Here $f(.)$ is an unspecified population mean function. If the function $f(.)$ is modeled as a linear penalized spline, then testing for linearity of $f(.)$ against a nonparametric alternative is equivalent to testing

$$H_0 : \sigma_b^2 = 0 \quad \text{vs.} \quad H_A : \sigma_b^2 > 0, \tag{13}$$

where σ_b^2 is a variance component controlling the degree of smoothness of $f(.)$.

Computation times both for LRT and RLRT were very long even for small sample sizes. For example, for six subjects and 50 observations per subject, computation time for 10,000 simulations was 4.5 h for R and 1 h for SAS on a server (Intel Xeon 3 GHz CPU). Additionally, run time increased steeply with both I and J for R. For R significant reduction of computation times could be achieved by interfacing it with C or FORTRAN. SAS is faster with its default convergence criterion, but we found numerical imprecisions, especially when estimating the probability mass at zero. These problems were mitigated when the convergence criterion was more stringent, but was accompanied by an increasing proportion of unsuccessful model fits. For more details see the extensive simulation study in Greven et al. (2008). Needless to say that in more complex models with larger sample sizes the computational burden is even more serious, especially when running several tests or performing simulation studies.

Therefore, for many applications there is a need for fast and accurate approximations of the null finite sample distribution of the RLRT for testing $H_{0,s}$. We describe two such approximations. The first approximation was introduced by Greven et al. (2008), is practically instantaneous, and avoids bootstrap. The second approximation was introduced by Crainiceanu (2003) and Crainiceanu and Ruppert (2004a) and uses a simple parametric approximation that reduces the necessary number of bootstrap samples. In extensive simulation studies, Greven et al. (2008) show that both methods outperform the $0.5\chi_0^2 : 0.5\chi_1^2$ approximation and the parametric bootstrap. The approximation used by standard software is the $0.5\chi_0^2 : 0.5\chi_1^2$ approximation. The necessary regularity conditions for this approximation to be asymptotically valid are independence under null and alternative hypothesis, large number of subvectors, and balanced designs. When these conditions are met both approximated distributions discussed in the following converge weakly to $0.5\chi_0^2 : 0.5\chi_1^2$ distribution. However, when conditions are not met, both approximate distributions agree with each other, are different from the $0.5\chi_0^2 : 0.5\chi_1^2$ distribution, and better fit the finite sample distribution of the RLRT.