Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models
Author(s): Sonja Greven, Ciprian M. Crainiceanu, Helmut Küchenhoff and Annette Peters
Source: *Journal of Computational and Graphical Statistics*, Vol. 17, No. 4 (December 2008), pp. 870-891
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: https://www.jstor.org/stable/25651233
Accessed: 27-08-2018 19:20 UTC

Supplemental materials for this article are available online
www.amstat.org/publications/JCGS

Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models

Sonja GREVEN, Ciprian M. CRAINICEANU,
Helmut KÜCHENHOFF, and Annette PETERS

The goal of our article is to provide a transparent, robust, and computationally feasible statistical platform for restricted likelihood ratio testing (RLRT) for zero variance components in linear mixed models. This problem is nonstandard because under the null hypothesis the parameter is on the boundary of the parameter space. Our proposed approach is different from the asymptotic results of Stram and Lee who assumed that the outcome vector can be partitioned into many independent subvectors. Thus, our methodology applies to a wider class of mixed models, which includes models with a moderate number of clusters or nonparametric smoothing components. We propose two approximations to the finite sample null distribution of the RLRT statistic. Both approximations converge weakly to the asymptotic distribution obtained by Stram and Lee when their assumptions hold. When their assumptions do not hold, we show in extensive simulation studies that both approximations outperform the Stram and Lee approximation and the parametric bootstrap. We also identify and address numerical problems associated with standard mixed model software. Our methods are motivated by and applied to a large longitudinal study on air pollution health effects in a highly susceptible cohort. Relevant software is posted as an online supplement.

Key Words: Parametric Bootstrap; Nonparametric smoothing; Nonregular problem; Penalized splines.

1. INTRODUCTION

Mixed models are a powerful inferential tool with a wide range of applications including longitudinal studies, hierarchical modeling, and smoothing. In a standard linear

Sonja Greven is Postdoctoral Fellow, and Ciprian M. Crainiceanu is Assistant Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205 (E-mail addresses: *sgreven@jhsph.edu* and *ccrainic@jhsph.edu*). Helmut Küchenhoff is Professor, Department of Statistics, Statistical Consulting Unit, Ludwig-Maximilians-Universität München, Akademiestr. 1, 80799 Munich, Germany (E-mail: *Kuechenhoff@stat.uni-muenchen.de*). Annette Peters is Head of Research Units Epidemiology of Air Pollution Health Effects and Epidemiology of Chronic Diseases, Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany (E-mail: *peters@helmholtz-muenchen.de*).

Journal of Computational and Graphical Statistics, Volume 17, Number 4, Pages 870–891
DOI: 10.1198/106186008X386599

mixed model (LMM) (Laird and Ware 1982), variance components control shrinkage of various model components such as random intercepts or slopes towards the population or group mean. In a general design LMM, variance components may also control shrinkage of smooth uni- or multivariate functions towards their prior means. An example of a model incorporating both smooth functions and mixed effects is the one used in the AIRGENE study described in Section 5. In this study, the trajectory of each subject's inflammatory markers is modeled as the sum of an unspecified smooth population function and a subject specific intercept.

Methodological and computational advancements have led to the development of software for mixed model inference such as the `MIXED` procedure in SAS, the `lme` function in R and S+, or the `xtmixed` function in STATA. However, the problem of calculating the RLRT null distributions and p-values for zero variance testing remains an open problem for many, if not most, modern mixed model applications. Our article is designed to improve the current state-of-the-art for zero variance RLRTs in linear mixed models. We also identify numerical problems associated with mixed model software in this context and propose practical solutions.

We focus on the following general form of a mixed model

$$\begin{cases} \boldsymbol{Y} & = & \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{b}_1 + \cdots + \boldsymbol{Z}_S\boldsymbol{b}_S + \boldsymbol{\varepsilon}; \\ \boldsymbol{b}_s & \sim & N(\boldsymbol{0}, \sigma_s^2 \boldsymbol{I}_{K_s}), s = 1, \ldots, S; \\ \boldsymbol{\varepsilon} & \sim & N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n), \end{cases} \tag{1.1}$$

where the random effects $\boldsymbol{b}_s$, $s = 1, \ldots, S$, and the errors $\boldsymbol{\varepsilon}$ are mutually independent, K_s denotes the number of columns in $\boldsymbol{Z}_s$, n is the sample size, and $\boldsymbol{I}_\nu$ denotes the identity matrix with ν columns. This is not the most general form of a linear mixed model as it assumes independence of the random effects, but it is often used in practice and covers many important settings.

We are interested in testing

$$H_{0,s} : \sigma_s^2 = 0 \quad \text{versus} \quad H_{A,s} : \sigma_s^2 > 0, \tag{1.2}$$

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 hypothesis the component $\boldsymbol{Z}_s\boldsymbol{b}_s$ of model (1.1) is zero.

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 likelihood ratio test (LRT) for testing (1.2) 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 hypotheses*. Thus, it could be surprising that in many applications the null distribution of the LRT and RLRT 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. Crainiceanu and Ruppert (2004b) showed that the asymptotic distribution is different from a $0.5\chi^2_0 : 0.5\chi^2_1$ mixture in this case. 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. Furthermore, subvectors may not be identically distributed due to unbalanced designs or missing data. Even though the asymptotic results still hold in this case, the rate of convergence can be seriously affected, which leads to large differences between finite sample and asymptotic distributions. In the case of a LMM with one variance component ($S = 1$) Crainiceanu and Ruppert (2004b) and Crainiceanu, Ruppert, Claeskens, and Wand (2005) have derived the finite sample and asymptotic distribution of the LRT and RLRT showing that, under general conditions, the null distribution for testing $H_{0,S=1}$ is typically different from $0.5\chi^2_0 : 0.5\chi^2_1$.

The methodology developed by Crainiceanu and Ruppert (2004b) could be used to derive the null distribution for the more general case discussed in this article. While the result is theoretically interesting, this distribution is obtained by maximizing a stochastic process over the variance components of model (1.1), which makes the implementation computationally equivalent to the parametric bootstrap. For this reason, Crainiceanu and Ruppert (2004b,a) suggested using the parametric bootstrap in this context. However, in many situations, evaluating the likelihood is computationally expensive and it may not be reasonable to do thousands of simulations. Therefore, we propose two approximations of the finite sample null distribution of the RLRT for testing $H_{0,s}$. The first approximation is practically instantaneous and avoids bootstrap. The second uses a simple parametric approximation that reduces the necessary number of parametric bootstrap samples. In extensive simulation studies we show that both approximations outperform the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation and the parametric bootstrap. Surprisingly, we identified small numerical issues in mixed model software with serious implications for likelihood ratio testing. We propose two simple and effective methods to adjust the calculation of critical values accordingly.

We present the proposed approximations in Section 2 and discuss computational issues and their implications for likelihood ratio testing in linear mixed models in Section 3. In Section 4, we compare the two proposed approximations in extensive simulations to the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation and a parametric bootstrap. Section 5 illustrates the application of our proposed methodology to the AIRGENE study, a large longitudinal study on air pollution health effects in a highly susceptible cohort. Section 6 concludes with a summary of our results and a discussion.

2. TWO APPROXIMATIONS TO THE RLRT NULL DISTRIBUTION

2.1 FAST FINITE SAMPLE APPROXIMATION

Our first approximation of the distribution of the RLRT is inspired by pseudo-likelihood estimation (Gong and Samaniego 1981). Consider the likelihood $L(\boldsymbol{\theta}, \boldsymbol{\phi})$ for independent and identically distributed (iid) random variables $X_1, \ldots, X_n$, where the likelihood depends on the parameters of interest $\boldsymbol{\theta}$ and on nuisance parameters $\boldsymbol{\phi}$. Suppose that $L(.,.)$ is a complicated function of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, but simple as a function of $\boldsymbol{\theta}$ alone when $\boldsymbol{\phi}$ is fixed. In this case, pseudo-likelihood replaces $\boldsymbol{\phi}$ by a consistent estimator $\hat{\boldsymbol{\phi}}$ and maximizes $L^*(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}})$ over $\boldsymbol{\theta}$ to obtain the pseudo maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. The pseudo LRT for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$ is then defined as $\text{LRT}^* = 2 \log L^*(\hat{\boldsymbol{\theta}}) - 2 \log L^*(\boldsymbol{\theta_0})$.

Liang and Self (1996) showed that, under certain regularity assumptions, the asymptotic distribution of LRT* is the same as the distribution of the LRT when $\boldsymbol{\phi}$ is known. For testing one parameter θ, the resulting distribution is χ^2_1 for θ in the interior of the parameter space and $0.5\chi^2_0 : 0.5\chi^2_1$ for θ on the boundary. Although not stated explicitly, the proof is based on Theorem 2.2 in Gong and Samaniego (1981), which assumes a large number of iid observations or subvectors. However, this assumption often fails in modern applications such as nonparametric smoothing or longitudinal studies with imbalanced data or moderate number of clusters.

In our framework, $\boldsymbol{\theta} = (\sigma^2_s, \boldsymbol{\beta}, \boldsymbol{b_s})$ could be viewed as the parameters of interest, and the $\boldsymbol{b_i}, i \neq s$, as nuisance parameters. If the $\boldsymbol{b_i}$'s were known, the outcome vector could be redefined as $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \sum_{i \neq s} \boldsymbol{Z_i b_i}$ and our model could be reduced accordingly. The idea we transfer from pseudo-likelihood estimation is, that under some regularity conditions and with increasing sample size, the prediction of $\sum_{i \neq s} \boldsymbol{Z_i b_i}$ might be good enough to allow the RLRT for testing $\boldsymbol{\theta}$ to be closely approximated by the RLRT when $\sum_{i \neq s} \boldsymbol{Z_i b_i}$ is known. Thus, we propose to approximate the distribution of the RLRT in the original model (1.1) by that of the RLRT in the reduced model

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{X\beta} + \boldsymbol{Z_s b_s} + \boldsymbol{\varepsilon}, \tag{2.1}$$

with the same assumptions for $\boldsymbol{b_s}$ and $\boldsymbol{\varepsilon}$ as in (1.1), and $\widetilde{\boldsymbol{Y}}$ assumed to be known.

As model (2.1) has only one variance component, σ^2_s, the exact null distribution of the RLRT for testing $H_{0,s} : \sigma^2_s = 0$ versus $H_{A,s} : \sigma^2_s > 0$ is (Crainiceanu and Ruppert 2004b)

$$\text{RLRT}_n \overset{d}{=} \sup_{\lambda \geq 0} \left\{ (n-p) \log \left[1 + \frac{N_n(\lambda)}{D_n(\lambda)} \right] - \sum_{l=1}^{K_s} \log(1 + \lambda \mu_{l,n}) \right\}, \tag{2.2}$$

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

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

$w_l, l = 1, \ldots, n-p$, are independent $N(0, 1)$, and $\mu_{l,n}, l = 1, \ldots, K_s$, are the eigenvalues of the $K_s \times K_s$ matrix $\boldsymbol{Z}_s'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Z}_s$.

This distribution can be computed efficiently because it depends only on the eigenvalues $\mu_{l,n}$ of a $K_s \times K_s$ matrix, which need to be computed only once. Simulation from the distribution then effectively reduces to simulation of $(K_s + 1)$ χ^2 variables and a grid search over λ that does not depend on the sample size n and is almost instantaneous (5,000 simulations/second on a PC with 2.66 GHz CPU and 1 GB random access memory).

While we used empirical arguments to motivate this approximation, distribution (2.2) converges weakly to the asymptotic $0.5\chi_0^2 : 0.5\chi_1^2$ distribution obtained by Self and Liang (1987) under the assumption of independence *under null and alternative hypotheses*. However, these assumptions fail in many typical applications when data are not independent or when the number of independent subvectors is small to moderate. For such situations, we show in extensive simulation studies that approximation (2.2) generally outperforms the Self and Liang (1987) approximation. This is good news since calculation speed of critical or p-values using (2.2) is modern-day indistinguishable from using the Self and Liang (1987) approximation.

2.2 Mixture Approximation to the Parametric Bootstrap

In some cases, one might still want to use a simple parametric bootstrap to determine the distribution of the RLRT. However, reliable estimation of $(1 - \alpha)$ quantiles based on parametric bootstrap may be computationally intensive, especially for small α. Consider, for example, the following longitudinal model

$$Y_{ij} = b_i + f(x_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \qquad (2.3)$$
$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), \ \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2),$$

where the b_i's are random subject intercepts and $f(\cdot)$ is an unspecified population mean function. We demonstrate in Section 4.2 how (2.3) can be expressed as a particular mixed model. Then, testing for linearity of $f(\cdot)$ against a nonparametric alternative is equivalent to testing

$$H_0 : \sigma_u^2 = 0 \quad \text{versus} \quad H_A : \sigma_u^2 > 0, \qquad (2.4)$$

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

We performed 10,000 simulations, including computation of both LRT and RLRT, for $I = 6, 10$ subjects and $J = 25, 50, 100$ observations per subject on a server (Intel Xeon 3GHz CPU) and a personal computer (Intel Pentium M 1400 MHz processor). The resulting computation times for R and SAS are shown in Figure 1. For 6 subjects and 50 observations per subject, computation time on the PC was 10.2 hours for R and 1 hour for SAS. Additionally, run time increased steeply with both I and J for R. SAS seems to be faster, but we identified numerical problems that will be discussed in Section 3. Needless to say that in more complex models with larger sample sizes the computational burden is even more serious, in particular when running several tests or performing simulation studies.

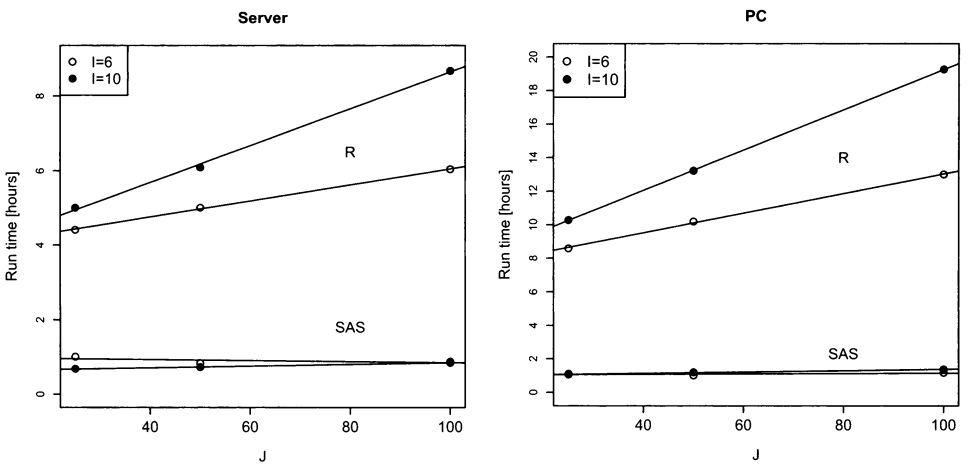

Figure 1. Computation times for 10,000 simulations in R and SAS in the case of testing for linearity of a smooth function in a model including a random intercept. I is the number of subjects and J is the number of observations per subject.

To reduce computation time, we propose to use a parametric approximation to the RLRT distribution. While in the case of iid data the distribution is asymptotically a $0.5\chi^2_0 : 0.5\chi^2_1$ mixture, Crainiceanu and Ruppert (2004b) showed that for correlated responses and finite sample size the distribution can severely deviate from this mixture. We propose to use the following finite sample approximation

$$\text{RLRT} \overset{d}{\approx} aUD, \tag{2.5}$$

where $U \sim \text{Bernoulli}(1-p)$, $D \sim \chi^2_1$, $p = P(U = 0)$ and a are unknown constants, and $\overset{d}{\approx}$ denotes approximate equality in distribution. This approximation has been applied before by Crainiceanu and Ruppert (2004a) in nonlinear regression, but behavior of this approximation in a wide variety of possible settings has not been studied so far.

The class of distributions in (2.5) is flexible and contains as a particular case the iid case asymptotic $0.5\chi^2_0 : 0.5\chi^2_1$ distribution with $a = 1$ and $p = 0.5$, and is just as easy to use. As the point mass at zero, p, and the scaling factor, a, are unknown in all other cases, we propose to estimate them from a parametric bootstrap sample. The idea of the parametric approximation is to use the entire parametric bootstrap sample to fit a flexible two parameter family of distributions, thus reducing the necessary number of simulations required for estimating tail quantiles. We will show in Section 4 that approximation (2.5) generally outperforms the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation. This happens in many applications when the correlation structure imposed by the random effects, $\boldsymbol{b_i}$, cannot be ignored, or when the sample size is small to moderate. Note that both our proposed approximations are asymptotically identical to the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation when the assumptions in Self and Liang (1987) and Stram and Lee (1994) hold, respectively.

3. IMPLEMENTATION AND COMPUTATIONAL ISSUES IN MIXED MODEL SOFTWARE

In this section, we focus on the aUD approximation of the RLRT null distribution given in (2.5). One obvious approach to estimate p and a could be to use maximum likelihood (ML). If the simulated RLRT values $t_i, i = 1, \ldots, N$, are an iid sample from the aUD distribution in (2.5), the corresponding likelihood is

$$L(a, p) = p^{N_0}(1-p)^{N-N_0} \prod_{i:t_i>0} \frac{1}{a} \frac{1}{\sqrt{2\pi}} \left(\frac{t_i}{a}\right)^{-\frac{1}{2}} e^{-\frac{t_i}{2a}},$$

where $N_0 = |\{i : t_i = 0\}|$. Thus, the ML estimates for p and a are $\hat{p}_{\mathrm{ML}} = N_0/N$ and $\hat{a}_{\mathrm{ML}} = \bar{t}/(1 - \hat{p}_{\mathrm{ML}})$, with $\bar{t}$ the mean of $t_i, i = 1, \ldots, N$.

While this is the obvious theoretical solution to the estimation problem, as it provides asymptotically unbiased and efficient estimators for p and a, numerical problems with standard mixed model software render the implementation of these estimators difficult. Because $\hat{p}_{\mathrm{ML}}$ uses the proportion of RLRT values that are exactly zero, numerical imprecisions can lead to serious under- or over-estimation of p and a.

For example, when testing for a zero variance component in a linear mixed model with two variance components, standard statistical software (the `MIXED` procedure in SAS and the `lme` function in R) typically estimates a probability mass at zero between 0 and 0.05. In contrast, the asymptotic distribution for the iid case has 0.5 probability mass at zero, and we show in our simulations in Section 4 that, often, it is even higher. There are two important reasons for this surprising numerical result.

First, in many simulations the likelihood of the alternative model is estimated to be slightly smaller than that of the null model. This leads to a large proportion of negative values for the RLRT and could be corrected by setting all these values to zero. This behavior is especially pronounced in R `lme`, where the (restricted) log-likelihood is maximized with respect to the scaled logarithms of the variances (see Pinheiro and Bates 2000, sect. 2.2) and maxima at zero cannot be found.

Second, in many other simulations the value of the RLRT statistic is estimated to be positive, but very small. Potential explanations for these results are numerical imprecisions. The log-likelihood values calculated for the two models under null and alternative are the result of floating point computations. Additionally, these values are the result of numerical optimization, so that the values calculated are even more approximate than with direct floating point computation. Different starting estimates or different convergence criteria in the optimization can lead to subtly different values. This problem is especially serious in SAS. When testing for linearity in model (2.3), decreasing the default criterion in SAS (corresponding to 10^{-8}, SAS Institute Inc. 2004) results in a pronounced increase of the proportion of values smaller or equal to zero, for example from 0.33 to 0.56 for $I = 10$, $J = 100$ and a convergence criterion value of 10^{-16} (see Table 1). However, this improved precision comes at the expense of an increase of the nonconvergence proportion from 0% to as much as 99%, with corresponding serious increase in simulation time. This

Table 1. ML estimates of p and a for testing for linearity in (2.3) using RLRT, based on 1,000 simulations in SAS with different values for the convergence criterion c and the maximum number of iterations m. Negative values were first set to zero.

		Default					
		$c = 10^{-8}$, m=50		$c = 10^{-12}$, m=100		$c = 10^{-16}$, m=150	
I	J	$\hat{p}_{\mathrm{ML}}$	$\hat{a}_{\mathrm{ML}}$	$\hat{p}_{\mathrm{ML}}$	$\hat{a}_{\mathrm{ML}}$	$\hat{p}_{\mathrm{ML}}$	$\hat{a}_{\mathrm{ML}}$
6	25	0.28	0.37	0.41	0.50	0.53	0.53
	100	0.68	0.78	0.23	0.37	0.56	0.45
10	25	0.28	0.39	0.42	0.48	0.56	0.49
	100	0.33	0.45	0.48	0.59	0.56	0.54

behavior raises a qualitatively different problem because the user cannot distinguish between true very small, but positive, values and exact zeros that are estimated to be small and positive.

Thus, calculating the ML estimators of p and a is practically impossible with current software. Instead, we propose two robust alternatives for estimation of p and a. Application of either method begins by setting negative values to zero.

3.1 METHOD OF MOMENTS

The first alternative approach uses the method of moments (MoM). The first and second moments of the approximate null distribution of the RLRT statistic, T, described in (2.5) are:

$$\mathrm{E}(T) = (1-p)a, \quad \mathrm{E}(T^2) = 3(1-p)a^2.$$

Thus, the MoM estimators of p and a are $\hat{p}_{\mathrm{MoM}} = 1 - 3\bar{t}^2/\overline{t^2}$ and $\hat{a}_{\mathrm{MoM}} = \bar{t}/(1-\hat{p}_{\mathrm{MoM}})$, where $\overline{t^2} = \sum_{i=1}^{n} t_i^2/n$. This estimator circumvents the problems related to probability mass at zero estimation by using sample moments that are relatively robust to small variations of values around zero.

3.2 QUANTILE REGRESSION

The second alternative approach uses quantile regression (QR) of observed on theoretical quantiles. Let

$$T_r = aF_{\chi_1^2}^{-1}\left\{I(r > p)\frac{r-p}{1-p}\right\} \tag{3.1}$$

be the theoretical r quantile of the aUD approximation in (2.5), where $F_{\chi_1^2}$ is the cumulative distribution function of the χ_1^2 distribution, and $I(\cdot)$ is the indicator function. To estimate p, we substitute the MoM estimator $\bar{t}/(1-p)$ for a into Equation (3.1) and min-

Table 2. p- and a-values for testing (2.4) in (2.3) using RLRT, estimated by ML, MoM, and QR using 10,000 simulations in SAS and R with default settings.

		SAS						R					
		ML		MoM		QR		ML		MoM		QR	
I	J	$\hat{p}$	$\hat{a}$	$\hat{p}$	$\hat{a}$	$\hat{p}$	$\hat{a}$	$\hat{p}$	$\hat{a}$	$\hat{p}$	$\hat{a}$	$\hat{p}$	$\hat{a}$
6	25	0.29	0.41	0.66	0.86	0.66	0.86	0.68	0.76	0.72	0.85	0.72	0.86
	100	0.65	0.84	0.67	0.91	0.67	0.91	0.68	0.86	0.71	0.94	0.71	0.95
10	25	0.28	0.41	0.69	0.94	0.69	0.95	0.64	0.81	0.67	0.90	0.68	0.90
	100	0.33	0.43	0.68	0.92	0.68	0.93	0.68	0.76	0.71	0.83	0.71	0.83

imize the least squares function

$$\min_{p\in[0,1)} \sum_{j=1}^{N} \left[t_{(j)} - \frac{\bar{t}}{1-p} F_{\chi_1^2}^{-1} \left\{ I\left(\frac{j-0.5}{N} > p \right) \frac{\frac{j-0.5}{N} - p}{1-p} \right\} \right]^2, \tag{3.2}$$

where $t_{(j)}$ indicates the jth ordered observation. This is a nonlinear optimization problem in p that can be solved using standard software (such as the function `optim` in R) or straightforward grid minimization. We then use the resulting quantile regression estimate, $\hat{p}_{\mathrm{QR}}$, to calculate $\hat{a}_{\mathrm{QR}} = \bar{t}/(1-\hat{p}_{\mathrm{QR}})$.

Table 2 compares all three estimation methods for p and a for testing (2.4) in (2.3) using RLRT, based on the default settings for SAS and R, respectively. Negative values are first set to zero. Results from the MoM and QR methods are similar, while ML seriously underestimates p and a in SAS, and slightly in R. Note that while differences in $\hat{p}$ and $\hat{a}$ between R and SAS become much smaller using MoM or QR rather than ML, some discrepancies remain, with $\hat{p}$ being typically higher in R. Thus, simulated RLRT distributions might depend on the particular software platform used.

4. SIMULATIONS

To address testing for zero variance components for a range of models that reasonably represent problems and situations encountered in practice, our simulation studies are structured as follows. The first two models are LMMs with one variance component: one-way analysis of variance (ANOVA) with a covariate, and univariate smoothing. The next four models are LMMs with two variance components: a model with a random intercept and a random slope, one with a random intercept and a smooth function, one with two smooth functions, and a model with a random intercept and a bivariate smooth function.

4.1 ONE VARIANCE COMPONENT

Case (RI): Consider the balanced one-way ANOVA model with an additional covariate

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, I;\ j = 1, \ldots, J, \tag{4.1}$$

where $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$ independently of the $b_{0i} \overset{\text{iid}}{\sim} N(0, \sigma_{b_0}^2)$. Testing for equality of individual intercepts $\beta_0 + b_{0i}$ is equivalent to testing

$$H_0 : \sigma_{b_0}^2 = 0 \quad \text{versus} \quad H_A : \sigma_{b_0}^2 > 0. \tag{4.2}$$

Case (SMO): The univariate smoothing model is given by

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.3}$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$, and $f(\cdot)$ is an unspecified function of x. Suppose that we are interested in testing whether $f(\cdot)$ is a linear function $f(x) = \beta_0 + \beta_1 x$. To allow sufficient flexibility for the alternative, we use low-rank thin plate splines, radial smoothers, which extend easily to the multivariate case (French, Kammann, and Wand 2001; Ruppert, Wand, and Carroll 2003). Specifically, we represent (4.3) as the mixed model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z_1 u_1} + \boldsymbol{\varepsilon}, \tag{4.4}$$

where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I_n})$ independently of $\boldsymbol{u_1} \sim N(\boldsymbol{0}, \sigma_{u_1}^2 \boldsymbol{I_K})$, with

$$\boldsymbol{X} = [1\ x_i]_{1 \le i \le n}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1), \quad \boldsymbol{Z_1} = \boldsymbol{Z_K \Omega_K^{-1/2}}, \quad \boldsymbol{Z_K} = [|x_i - \kappa_k|^3]_{\substack{1 \le i \le n \\ 1 \le k \le K}}$$
$$\text{and} \quad \boldsymbol{\Omega_K} = [|\kappa_k - \kappa_{k'}|^3]_{\substack{1 \le k \le K \\ 1 \le k' \le K}},$$

where the $\kappa_k, k = 1, \ldots, K$, are at least five knots placed at every 4th or 5th unique x_i, up to a maximum of 35 knots (Ruppert 2002; Ngo and Wand 2004). In this representation, the shrinkage parameter $\sigma_{u_1}^2$ controls deviations from linearity. Testing for linearity is then equivalent to testing

$$H_0 : \sigma_{u_1}^2 = 0 \quad \text{versus} \quad H_A : \sigma_{u_1}^2 > 0. \tag{4.5}$$

4.2 TWO VARIANCE COMPONENTS

Case (RS|RI): If individual slopes are allowed to differ in the balanced ANOVA case (4.1), the model expands to

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + b_{1i} x_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, I;\ j = 1, \ldots, J, \tag{4.6}$$

where $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$, $b_{0i} \overset{\text{iid}}{\sim} N(0, \sigma_{b_0}^2)$, and $b_{1i} \overset{\text{iid}}{\sim} N(0, \sigma_{b_1}^2)$ are mutually independent. Testing for equality of individual slopes $\beta_1 + b_{1i}$ is equivalent to testing

$$H_0 : \sigma_{b_1}^2 = 0 \quad \text{versus} \quad H_A : \sigma_{b_1}^2 > 0. \tag{4.7}$$

Note that within the framework of model (1.1), we assume independence of the random intercept b_{0i} and slope b_{1i}. In practice, this assumption may not be appropriate, especially when the x vector is not centered.

Cases (SMO|RI) and (RI|SMO): The next model incorporates a random intercept and a univariate smooth function

$$Y_{ij} = b_{0i} + f(x_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, I;\ j = 1, \ldots, J, \tag{4.8}$$

with ε_{ij} and b_{0i} as before. Following (4.4) we represent model (4.8) as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z_0 b_0} + \boldsymbol{Z_1 u_1} + \boldsymbol{\varepsilon}, \tag{4.9}$$

where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I_n})$, $\boldsymbol{b_0} \sim N(\boldsymbol{0}, \sigma_{b_0}^2 \boldsymbol{I_I})$, and $\boldsymbol{u_1} \sim N(\boldsymbol{0}, \sigma_{u_1}^2 \boldsymbol{I_K})$ are mutually independent, $\boldsymbol{X}$ and $\boldsymbol{Z_1}$ are set up as in (4.4), and where $\boldsymbol{Z_0} = [\delta_{ii'}]_{1 \le i \le I, 1 \le j \le J}$ is the design matrix for the random intercept, with the Kronecker delta $\delta_{ii'} = 1$ for $i = i'$ and $= 0$ for $i \neq i'$. Testing for equality of individual intercepts (Case $RI|SMO$) and testing for linearity of $f(\cdot)$ (Case $SMO|RI$) then reduce to testing (4.2) and (4.5), respectively.

Case (SMO|SMO): Consider a model with two smooth univariate functions

$$Y_i = f_1(x_{i1}) + f_2(x_{i2}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.10}$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$, and $f_1(\cdot)$ and $f_2(\cdot)$ are two unspecified functions of x_1 and x_2, respectively. Following (4.4) we represent model (4.10) as

$$\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{Z_1 u_1} + \boldsymbol{Z_2 u_2} + \boldsymbol{\varepsilon}, \tag{4.11}$$

where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I_n})$, $\boldsymbol{u_1} \sim N(\boldsymbol{0}, \sigma_{u_1}^2 \boldsymbol{I_{K_1}})$ and $\boldsymbol{u_2} \sim N(\boldsymbol{0}, \sigma_{u_2}^2 \boldsymbol{I_{K_2}})$, $\boldsymbol{X} = [1 x_{i1} x_{i2}]_{1 \le i \le n}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$, and $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are both set up analogously to (4.4). Testing for linearity of $f_1(\cdot)$ then is equivalent to testing (4.5).

Cases (BIV|RI) and (RI|BIV): In the situation of (4.8), let the smooth function $f(\cdot)$ depend on two variables

$$Y_{ij} = b_{0i} + f(x_{ij1}, x_{ij2}) + \varepsilon_{ij}, \quad i = 1, \ldots, I;\ j = 1, \ldots, J. \tag{4.12}$$

Using the thin plate spline modeling approach, bivariate smoothing is a straight forward extension from (4.4) (see, e.g., Nychka 2000), and (4.12) can be represented as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z_0 b_0} + \boldsymbol{Z_3 u_3} + \boldsymbol{\varepsilon}, \tag{4.13}$$

where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I_n})$, $\boldsymbol{b_0} \sim N(\boldsymbol{0}, \sigma_{b_0}^2 \boldsymbol{I_I})$ and $\boldsymbol{u_3} \sim N(\boldsymbol{0}, \sigma_{u_3}^2 \boldsymbol{I_K})$, where $\boldsymbol{X} = [1\ x_{ij1}\ x_{ij2}]_{1 \le i \le I, 1 \le j \le J}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$, $\boldsymbol{Z_0}$ as in (4.9), and

$$\boldsymbol{Z_3} = [C(\boldsymbol{x_{ij}} - \boldsymbol{\kappa_k})]_{\substack{1 \le i \le I, 1 \le j \le J \\ 1 \le k \le K}} [C(\boldsymbol{\kappa_k} - \boldsymbol{\kappa_{k'}})]^{-1/2}_{\substack{1 \le k \le K \\ 1 \le k' \le K}} \tag{4.14}$$
$$\text{with } C(\boldsymbol{r}) = \| \boldsymbol{r} \|^2 \log(\| \boldsymbol{r} \|),$$

$\boldsymbol{x_{ij}} = (x_{ij1}, x_{ij2})'$, and where the $\boldsymbol{\kappa_k}$ are K knots positioned across the range of $\boldsymbol{x}$ using a space-filling algorithm (Nychka and Saltzman 1998). Testing for equality of individual intercepts (Case $RI|BIV$) then reduces to testing (4.2), and testing for additivity and linearity $f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (Case $BIV|RI$) is equivalent to testing

$$H_0 : \sigma_{u_3}^2 = 0 \quad \text{versus} \quad H_A : \sigma_{u_3}^2 > 0. \tag{4.15}$$

Table 3. Parameter settings for the RLRT null distribution for all simulations. 0 indicates a zero under the null hypothesis; no entry indicates nonpresence in the full model. *This value was also set to 0, 0.1, 10 and 100. ♯This distribution was also set to χ^2_1.

Case	β_0	β_1	β_2	σ^2_ε	x	(x_1, x_2)	$\sigma^2_{b_0}$	$\sigma^2_{b_1}$	$\sigma^2_{u_1}$	$\sigma^2_{u_2}$	$\sigma^2_{u_3}$
(RI)	1	−1		1	$N(0,1)$		0				
(SMO)	1	−1		1	$N(0,1)$				0		
(RS|RI)	1	−1		1	$N(0,1)$		1*	0			
(RI|SMO)	1	−1		1	$N(0,1)$		0		1*		
(SMO|RI)	1	−1		1	$N(0,1)^\sharp$		1*		0		
(SMO|SMO)	1	−1	−1	1		$N(\mathbf{0}, \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right))$			0	1*	
						$\rho = 0, 0.3, 0.6, 0.9$					
(BIV|RI)	1	−1	−1	1		$N(\mathbf{0}, \boldsymbol{I_2})$	1				0
(RI|BIV)	1	−1	−1	1		$N(\mathbf{0}, \boldsymbol{I_2})$	0				1

4.3 SIMULATION SET-UP AND RESULTS

For each of the eight cases introduced in Sections 4.1 and 4.2, we simulated 10,000 samples from the null distribution of the RLRT (henceforth called "simulated samples"). The parameters used are provided in Table 3. A zero variance component indicates that in the null model used for simulations this variance component is equal to zero, but the alternative model contains it. In contrast, an empty entry means that this particular variance component is not part of the full model. For setting $(SMO|RI)$, for example, testing for linearity in a model with a random intercept, the RLRT distribution was simulated from the null model

$$Y_{ij} = 1 + b_{0i} - x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0,1),\ b_{0i} \overset{\text{iid}}{\sim} N(0,1),$$

with the x_{ij} being a random sample from a $N(0,1)$ distribution, and the models fitted under null and alternative hypotheses were

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon),\ b_{0i} \overset{\text{iid}}{\sim} N(0, \sigma^2_{b_0}),$$

and

$$Y_{ij} = b_{0i} + f(x_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon),\ b_{0i} \overset{\text{iid}}{\sim} N(0, \sigma^2_{b_0}),$$

respectively. The sample size used was $n = I \times J$ with $I = 6, 10$ subjects and $J = 5, 25, 50, 100$ observations per subject for all simulations.

We compared our two approximation methods introduced in Section 2 with the parametric bootstrap and the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation of Stram and Lee (1994). Our first method was the fast finite sample approximation. We simulated 100,000 samples (henceforth called "CR samples") from (2.2) using efficient MATLAB code (Crainiceanu and Ruppert 2004b, *www.biostat.jhsph.edu/~ccrainic*). The second method used the aUD approximation described in (2.5), based on $N = 200$, 500, 1000, and 2000 out of the 10,000 simulated samples. For the aUD approximation, we report only results based on MoM

estimators of p and a because the QR method provides almost identical results. We considered smaller simulated subsample sizes, N, to investigate whether the aUD approximation provides accurate approximations of the parametric bootstrap distribution based on 10,000 samples. This is relevant in many applications where evaluating the RLRT is computationally expensive.

To determine Type I error rates for each method, we calculated critical values as empirical quantiles from the 100,000 CR samples for the fast finite sample approximation, and from $N = 500$ simulated samples for the parametric bootstrap. For the aUD and $0.5\chi^2_0 : 0.5\chi^2_1$ approximations, critical values were calculated as theoretical quantiles of the respective mixture distribution, using the estimated p and a values from $N = 500$ simulated samples in the aUD case. Empirical Type I error rates were then calculated as the proportion of rejections from the remaining 9,500 simulated samples for all methods. The $N = 500$ samples used for the aUD method and the parametric bootstrap were randomly drawn out of all 10,000 simulated samples 100 times to evaluate estimation variability. Note that this approach overestimates the performance of the bootstrap-based procedures, as it uses the true parameters rather than estimated parameters in the simulations; see Scheipl et al. (2008) for a discussion. However, it preserves the relationship between the performances of the parametric bootstrap and the aUD approximation, which is of interest to us

Because results were remarkably similar across simulation frameworks, we provide results from our simulations in SAS for two cases. Additional extensive simulation results and user friendly software are available in an online supplement. The fast finite sample approximation was also implemented by Fabian Scheipl in the R package `RLRsim` available from *www.r-project.org*. Figure 2(a) displays the results for case (RI), testing a random intercept. Figure 2(b) shows results for case $(SMO|RI)$, testing a smooth function for linearity (with random intercept). Estimated Type I error rates for the aUD approximation and the parametric bootstrap are shown as boxplots of values for the 100 draws, with a dotted line at the nominal Type I error rate of 0.05. A solid line indicates the estimated Type I error rate using the fast finite sample approximation, and a dashed line the one for the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation. In the (RI) case the fast finite sample approximation is, in fact, the exact distribution. Thus, the estimated Type I error rates based on 9,500 simulations from the null distribution are very near the nominal level (Figure 2(a)). For the smooth $(SMO|RI)$ case, the fast finite sample approximation gives equally good results (Figure 2(b)).

An important conclusion is that the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation is very conservative in all cases, as expected from Crainiceanu and Ruppert (2004b). Thus, it provides tests that are undersized and less powerful than the approximations proposed in Section 2. For case (RI) of testing for the zero variance of random intercepts, the iid assumption holds and the Stram and Lee (1994) result applies asymptotically. The approximation is less conservative and problems are mitigated as the number of clusters increases.

In contrast, the RLRT_{CR} approximation of the null distribution provides good results regardless of the sample size and correlation structure. Only in one extreme simulation case is a moderate sample size necessary to achieve a good approximation. This happens

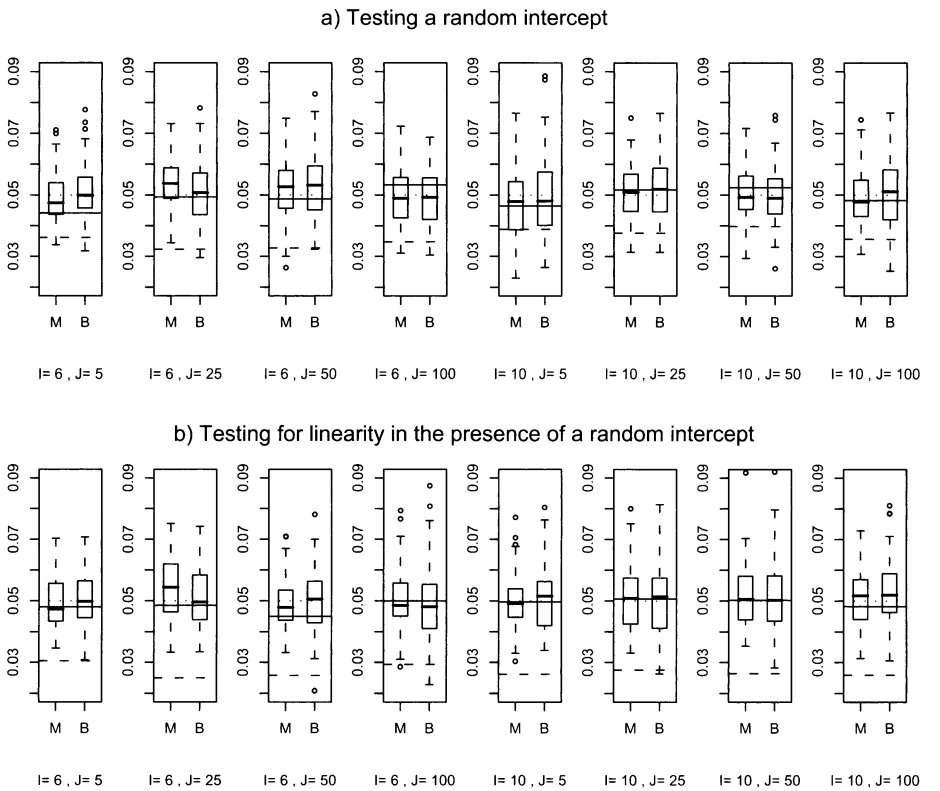

Figure 2. Type I error rates for testing a random intercept (case RI) and testing for linearity in the presence of a random intercept (case $SMO|RI$), estimated from 9,500 simulated samples for a nominal level of α=0.05 (indicated by a dotted line). Estimated type I error rates for the aUD approximation (MoM estimator, $N = 500$) and the parametric bootstrap ($N = 500$) are shown as boxplots of values for the 100 draws, and denoted by M and B, respectively. Dashed and solid lines indicate estimated type I error rates for the the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation and the fast finite sample approximation (100, 000 CR samples), respectively. I is the number of subjects and J is the number of observations per subject.

in the $(SMO|SMO)$ case, where one of two smooth functions is tested for linearity. The extreme case is obtained by increasing the correlation, ρ, between the arguments x_1 and x_2 to $\rho = 0.9$. Even in this case the empirical rejection probability at level $\alpha = 0.05$ for $\sigma^2_{u_2} = 1$ is estimated to be 0.072 and 0.056 for $n = 50$ and 150, respectively. This is probably due to the fact that it is relatively difficult to estimate the functions under the alternative with very highly correlated arguments and small sample sizes. However, the necessary sample sizes are reasonably small. The aUD approximation continues to perform well even in this very special case and could be used as an alternative if such a case were of interest.

There is a regularity assumption for the RLRT_{CR} approximation worth noting. Self and Liang (1987), in their case 8 for iid data, show that the geometry of the problem changes and the asymptotic distribution of the LRT can be different when a nuisance parameter is on the boundary of the parameter space. This occurs when the nuisance parameter is not independent of the tested parameter, in the sense that the corresponding Fisher information

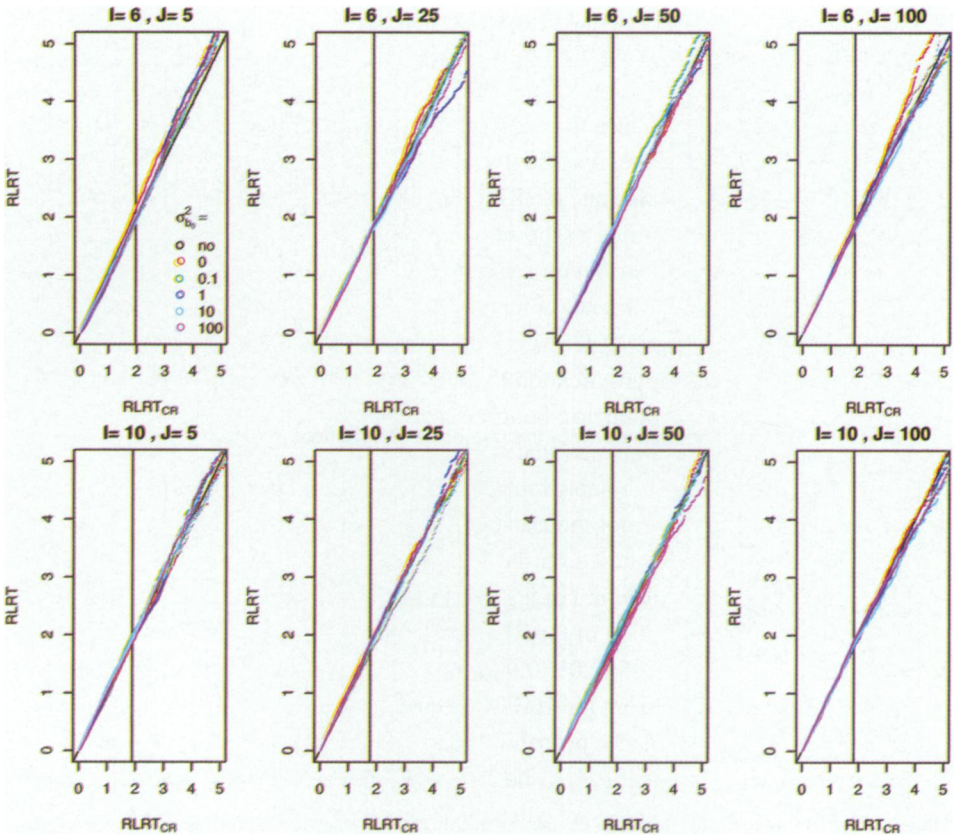

Figure 3. q-q-Plots of simulated RLRT samples against the fast finite sample approximation RLRT_{CR} (10,000 samples each). The RLRT distribution is plotted for case $(SMO|RI)$, testing for linearity in the presence of a random intercept. Different values of the second variance component $\sigma^2_{b_0}$, which controls deviations of individual from overall mean, are indicated by different colors (red=0, green=0.1, blue=1, cyan=10, magenta=100); the case (SMO) where no random intercept is part of the full model is denoted by "no" and shown in black. A vertical line gives the 95th percentile of the RLRT_{CR} distribution and a diagonal line the bisecting line. I is the number of subjects and J is the number of observations per subject.

is not diagonal. Thus, when using the asymptotic $0.5\chi^2_0 : 0.5\chi^2_1$ mixture for iid data, a necessary regularity assumption is that all nuisance parameters are in the interior of the parameter space, or are independent of the tested parameter. We observe similar results for the non-iid case. When in the $(SMO|SMO)$ case $f_2(\cdot)$ is also linear, with the nuisance parameter $\sigma^2_{u_2}$ on the boundary, the RLRT_{CR} approximation works well for $\rho = 0.0$ and 0.3, but becomes conservative for higher correlations. Indeed, for $\rho = 0.6$ empirical rejection probabilities range from 0.039 to 0.054 and for $\rho = 0.9$ they range from 0.012 to 0.031. Not surprisingly, the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation is even more conservative in this case and performs worse than the RLRT_{CR} approximation. The aUD approximation continues to produce reliable and stable results even in this rather exotic case.

We also investigated the agreement between the entire simulated distribution and the fast finite sample approximation RLRT_{CR}. Figure 3 displays corresponding q-q-plots for testing the linearity of a smooth function against a general alternative in a model incorporat-

ing a smooth population function and subject specific random intercepts (case $SMO|RI$). We varied the variance controlling the shrinkage of the random subject specific intercepts, $\sigma^2_{b_0} = 0, 0.1, 1, 10, 100$, and included the case when $\sigma^2_{b_0}$ was not part of the model (indicated by "no"). This corresponds to case (SMO), testing for linearity of a smooth function against a general alternative. A close inspection of Figure 3 reveals that there is good agreement between the simulated and the RLRT_{CR} distributions and that the simulated distribution is insensitive to the value of the second variance component, $\sigma^2_{b_0}$. Moreover, the variability of the simulated distribution relative to RLRT_{CR} is comparable to that exhibited in the corresponding one variance component case without a random intercept. This is reassuring, because in this case RLRT_{CR} is the exact distribution.

We proposed the aUD approximation as a practical and accurate alternative to simulating tens of thousands of parametric bootstrap samples from the null distribution of the RLRT. Our implicit anticipation was that the aUD approximation provides an accurate approximation to the parametric bootstrap distribution even when the a and p parameters are estimated using hundreds, not thousands, of bootstrap samples. To compare the precision of p-value calculations based on the various methods discussed in this article, we consider the following comparison. For a given testing case, we use the 10,000 simulated samples from the null distribution of the RLRT to estimate the empirical quantile, $q_{E,1-\alpha}$, corresponding to $\alpha = 0.1, 0.05, 0.01, 0.005$, and 0.001. We then subsample $N = 200$, 500, 1000, 2000 simulations from the 10,000 original simulations 100 times and use each method to evaluate the exceedance probabilities of $q_{E,1-\alpha}$. We denote by $p^l_{\alpha,M,N}$ the estimated exceedance probability of $q_{E,1-\alpha}$ based on method M and the lth sample of size N, $l = 1, \ldots, 100$. For each sample l we calculate $d_l(\alpha, M, N) = \text{logit}(p^l_{\alpha,M,N}) - \text{logit}(\alpha)$, which is a measure of discrepancy between the observed and expected proportion of rejections.

Figure 4 displays the boxplots of discrepancy for the parametric bootstrap and the aUD approximation based on the MoM estimator. The sample size for each method is varied from 200 to 2000 (left to right). Plots correspond to testing for a linear trend against a general alternative in the presence of subject specific intercepts in the model (case $SMO|RI$). The discrepancy of the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation is indicated by a dashed line, while the discrepancy of the RLRT_{CR} approximation is indicated by a solid line. The two left plots correspond to $I = 6$, $J = 50$ and the two right plots correspond to $I = 10$, $J = 25$, where I is the number of subjects and J is the number of observations per subject. As shown in Figure 2, the RLRT_{CR} approximation outperforms the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation. Indeed, additional simulations indicate that for the RLRT_{CR}, most of the variation around zero is due to the finite sample size of the parametric bootstrap sample that is used to estimate the discrepancy. In addition, Figure 4 shows that for $\alpha = 0.05$ the parametric bootstrap (B) would require roughly 500 simulations to achieve the same level of accuracy as the one of the aUD approximation (M) based on 200 simulations. For $\alpha = 0.001$ using the parametric bootstrap often results in discrepancy measures of minus infinity and more than 2,000 simulations would be needed to achieve a level of accuracy similar to the aUD approximation based on 200 simulations.

We also investigated the null distribution of the LRT and compared it with that of the

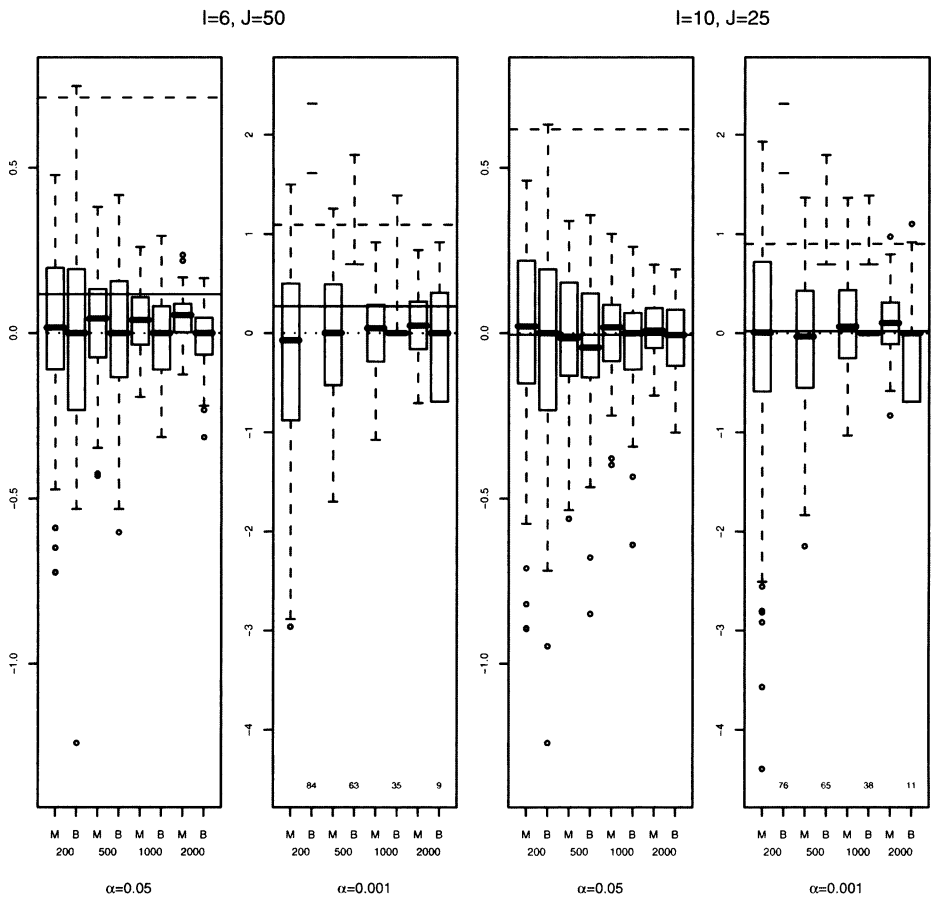

Figure 4. Boxplots of discrepancy measure values for 100 draws for the parametric bootstrap (B) and the aUD approximation based on the MoM estimator (M). The sample size N for each method is varied from 200 to 2000 (left to right). Plots correspond to testing for a linear trend against a general alternative in a model including subject specific intercepts (case $SMO|RI$). The discrepancy of the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation is indicated by a dashed line, while the discrepancy of the RLRT_{CR} approximation (100,000 samples) is indicated by a solid line. The two left plots correspond to $I = 6$, $J = 50$ and the two right plots correspond to $I = 10$, $J = 25$, where I is the number of subjects and J is the number of observations per subject. Numbers on the bottom indicate the number of minus infinities.

RLRT. The probability mass at zero for the null distribution of the LRT is typically above 0.95 when testing for linearity of a univariate, or additivity and linearity of a bivariate smooth. When testing for a random intercept or slope, the estimated mass at zero $\hat{p}_{\text{MoM}}$ for the LRT varies between 0.65 and 0.87 across different cases. The large probability mass at zero is probably due to the property of ML estimation to severely underestimate variance components, which would also put the applicability of the fast finite sample approximation into question. While the power properties were not investigated in this article, Crainiceanu et al. (2005) have reported severe loss of power for the LRT. For these reasons, we recommend using RLRT.

Scheipl et al. (2008) also compare the RLRT_{CR} approximation to several F-type tests for zero variance components or for polynomial regression. They find the RLRT_{CR} to be

comparable to the best bootstrap-based competitors with regard to power and adherence to the nominal level, while reducing computation time from hours to seconds.

5. THE AIRGENE STUDY ON AIR POLLUTION AND INFLAMMATION

The AIRGENE study was conducted in six European cities: Helsinki (Finland), Stockholm (Sweden), Augsburg (Germany), Barcelona (Spain), Rome (Italy), and Athens (Greece) between May 2003 and July 2004. While several previous epidemiological studies have demonstrated effects of ambient air pollution on morbidity and mortality due to cardio-pulmonary diseases (Dominici et al. 2006; Pope et al. 2002), this study investigates possible causal pathways for the observed effects. One of the aims of the AIRGENE study is to assess the association between inflammatory responses and ambient air pollution concentrations in myocardial infarction survivors. Three inflammatory blood markers (CRP, Fibrinogen and IL-6) were measured every month up to eight times in 1,003 patients. Air pollution and weather data was collected concurrently in each city. Patients were genotyped and additional information was collected at baseline. A full description of the study design can be found in Peters et al. (2007). A restricted anonymized data sample is available on request through the principal investigator of the AIRGENE study, Annette Peters.

Analyses had to account for longitudinal data structure and potential nonlinearity of weather and trend variables, with smooth effects estimated in the mixed model framework. As the shape of the air pollution dose–response functions has important policy implications, one aim of the study was to investigate the functional form of the air pollution effects on the inflammatory blood markers. To illustrate our approach, we focus on the association between PNC levels 12–17 hours before blood withdrawal, ultrafine particles with diameter less than 0.1 μm, and IL-6 in Athens. A total of 440 valid blood samples and PNC exposures were available for 108 patients, with an average of 4.1 observations per patient. Figure 5 displays the longitudinal log(IL-6) measurements for five patients, illustrating the large within- and between-patient variability. IL-6 needed to be log-transformed to fulfill the model assumption of residual normality.

The model used for estimation of the PNC-IL-6 dose–response function is

$$\log(\mathrm{IL6}_{ij}) = u_i + f(\mathrm{PNC}_{ij}) + \sum_{l=2}^{L} \beta_l x_{ijl} + \varepsilon_{ij}, \tag{5.1}$$

where $\mathrm{IL6}_{ij}$ is the jth IL-6 value of the ith patient, $u_i \overset{\text{iid}}{\sim} N(0, \sigma^2_{S=2})$ is a random patient intercept and PNC indicates the 12–17-hour-average PNC exposure before blood withdrawal. $f(\cdot)$ is a smooth, unspecified function estimated using penalized cubic B-Splines based on 20 equidistant knots in the range of PNC, 4, 625 to 83, 647 n/cm^3, and penalizing deviations from linearity (Greven, Küchenhoff, and Peters 2006). Other covariates that enter the model linearly or categorically are patient's body mass index, if the MI was a reinfarction, diagnosis of arrhythmia or congestive heart failure, long-term time trend, average temperature during the 48 hours prior to blood withdrawal (quadratic) and relative

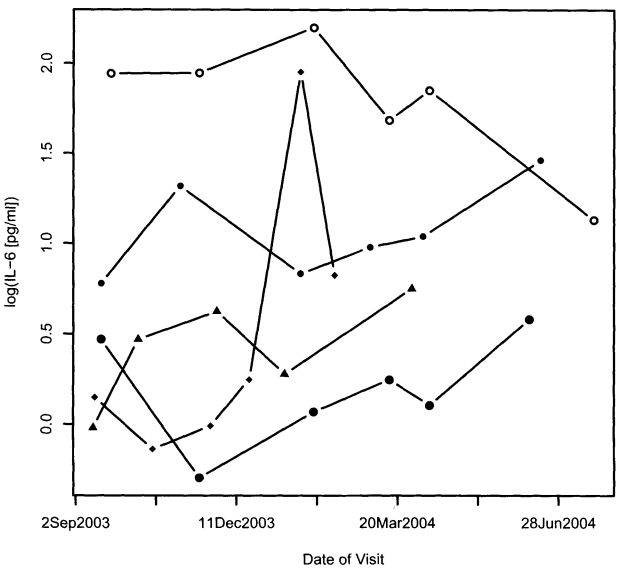

Figure 5. log(IL-6) values over time of five example patients in Athens.

humidity on day of blood withdrawal. We assume that $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$ due to the short half-life of IL-6, which seemed reasonable after standard residual analysis.

Figure 6 shows the estimated smooth PNC effect on log(IL-6) in Athens with approximate pointwise 95% confidence bands. An important scientific question is whether the dose–response function is linear. This is equivalent to testing (1.2) in (5.1), where $\sigma_{s=1}^2$ is a variance component controlling the smoothness of $f(\cdot)$. Note that the iid assumption is violated in this model due to unbalanced data and nonparametric smoothing, and that the model includes at least two variance components, one for the random intercept and one for $f(\cdot)$.

The restricted likelihood ratio test statistic for testing linearity of $f(\cdot)$ against a general alternative modeled by penalized splines takes the value RLRT $= 3.5$. The test results using all four approximations to the RLRT null distribution discussed in Sections 2 to 4 are given in Table 4. Even for this relatively small dataset, the fast finite sample approximation reduces computation time by more than three orders of magnitude, with 100,000 samples easily obtainable in half a minute, while the result is close to the parametric bootstrap. The aUD approximation gives similar results for as little as 1,000 parametric bootstrap samples, which could reduce the computation time by as much as an order of magnitude. The $0.5\chi_0^2 : 0.5\chi_1^2$ approximation is conservative. These results indicate that the shape of the function $f(\cdot)$ in Figure 6 is statistically different from a linear trend. Moreover, this dose-response function shows no effect up to about 40,000 n/cm^3, and is roughly linear for higher PNC values.

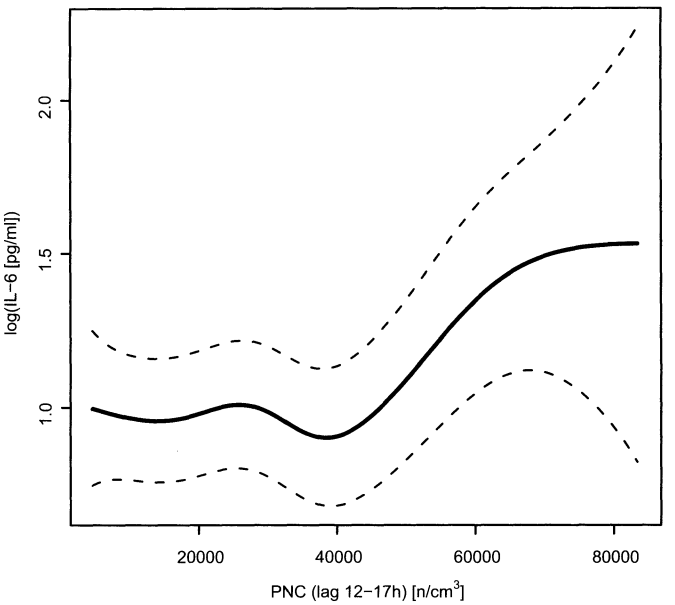

Figure 6. Estimated smooth PNC-log(IL-6) dose-response function in Athens with approximate pointwise 95% confidence intervals.

6. DISCUSSION

This article discusses restricted likelihood ratio testing (RLRT) for zero variance components in linear mixed models. Possible applications include, but are not limited to, testing for zero random intercepts or slopes and testing for linearity of a smooth function against a general alternative.

For models with one variance component, we recommend the exact finite sample null distribution of the RLRT statistic derived by Crainiceanu and Ruppert (2004b), which can be simulated efficiently. For models with more than one variance component, we present two approximations of the finite sample null distribution. The first approximation uses an

Table 4. Results for testing the PNC-log(IL-6) dose–response function in Athens for linearity using restricted likelihood ratio testing. Computation time was determined on a standard PC using MATLAB and R for the fast finite sample approximation, and SAS for the parametric bootstrap and aUD approximation.

Approximation	Samples	Computation time	p-value
Fast finite sample (Matlab)	100,000	33sec	0.014
Fast finite sample (R)	100,000	88sec	0.016
aUD	1,000	19min	0.017
aUD	10,000	3.4h	0.019
$0.5\chi^2_0 : 0.5\chi^2_1$	–	–	0.031
Bootstrap	10,000	3.4h	0.017

idea similar to pseudo-likelihood to obtain a distribution that can be simulated efficiently and avoids the parametric bootstrap. The second uses a simple parametric approximation to reduce the necessary number of bootstrap samples. Both approximations converge weakly to the $0.5\chi^2_0 : 0.5\chi^2_1$ distribution when the assumptions in Stram and Lee (1994) apply. In many typical applications these assumptions do not hold and the proposed approximations markedly outperform the $0.5\chi^2_0 : 0.5\chi^2_1$ approximation.

We report numerical imprecisions in current mixed model software that have serious implications for likelihood ratio testing. We propose two parametric approximation estimation methods (MoM and QR) designed to avoid these imprecisions. As the results of the two methods are similar, we suggest using the MoM because it is simpler.

Our article provides practical solutions to the variance testing problem for the general mixed effects model (1.1). This methodology applies to many data structures and scientific problems. However, the general area of testing for zero variance components remains a rich and challenging area of research. For example, we did not address the problem of testing for a random slope in a model with correlated random intercepts and slopes. This problem raises additional challenges because a nuisance parameter, the correlation, disappears under the null.

Relevant software is posted as an online supplement to this article. The R package `RLRsim` developed by Fabian Scheipl and available from CRAN provides the R interface for our first proposed approximation.

ACKNOWLEDGMENTS

We thank two reviewers and an associate editor for their careful reading of the original manuscript and for their comments, which helped to improve the article. This work was conducted while the first author was visiting Johns Hopkins University on a fellowship from the German Academic Exchange Service (DAAD). Ciprian Crainiceanu's work was supported by NIH grant AG025553-02 on the Effects of Aging on Sleep Architecture. We also thank Yu-Jen Cheng and Fabian Scheipl for discussions related to the topic of this article. The AIRGENE study was funded as part of the European Union's 5th Framework Programme, key action number 4: " Environment and Health," contract number QLRT-2002-02236. We would like to thank all members of the AIRGENE study group.

[Received March 2007. Revised January 2008.]

REFERENCES

Chernoff, H. (1954), "On the Distribution of the Likelihood Ratio," *Annals of Mathematical Statistics*, 25, 573–578.

Crainiceanu, C., and Ruppert, D. (2004a), "Likelihood Ratio Tests for Goodness-of-Fit of a Nonlinear Regression Model," *Journal of Multivariate Analysis*, 91, 35–52.

——— (2004b), "Likelihood Ratio Tests in Linear Mixed Models with One Variance Component," *Journal of the Royal Statistical Society*, Series B, 66, 165–185.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. (2005), "Exact Likelihood Ratio Tests for Penalised Splines," *Biometrika*, 92, 91–103.

Dominici, F., Peng, R., Bell, M., Pham, L., McDermott, A., Zeger, S., and Samet, J. (2006), "Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases," *Journal of the American Medical Association*, 295, 1127–1134.

French, J. L., Kammann, E. E., and Wand, M. P. (2001), Comment on "Semiparametric Nonlinear Mixed-Effects Models and Their Applications," *Journal of the American Statistical Association*, 96, 1285–1288.

Gong, G., and Samaniego, F. J. (1981), "Pseudo Maximum Likelihood Estimation: Theory and Applications," *The Annals of Statistics*, 9, 861–869.

Greven, S., Küchenhoff, H., and Peters, A. (2006), "Additive Mixed Models with P-Splines," in *Proceedings of the 21st International Workshop on Statistical Modelling*, eds. J. Hinde, J. Einbeck, and J. Newell, Statistical Modelling Society, pp. 201–207.

Laird, N., and Ware, J. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38(4), 963–974.

Liang, K.-Y., and Self, S. G. (1996), "On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic," *Journal of the Royal Statistical Society*, Series B, 58, 785–796.

Moran, P. (1971), "Maximum Likelihood Estimators in Non-Standard-Conditions," *Proceedings of the Cambridge Philosophical Society - Mathematical and Physical Sciences*, 70, 441–450.

Ngo, L., and Wand, M. (2004), "Smoothing With Mixed Model Software," *Journal of Statistical Software*, 9, 1–54.

Nychka, D. (2000), "Spatial Process Estimates as Smoothers," in *Smoothing and Regression*, ed. M. Schimek, New York: Springer-Verlag.

Nychka, D., and Saltzman, N. (1998), "Design of Air Quality Monitoring Networks," in *Case Studies in Environmental Statistics*, eds. D. Nychka, L. Cox, and W. Piegorsch, New York: Springer-Verlag.

Peters, A., Schneider, A., Greven, S., Bellander, T., Forastiere, F., Ibald-Mulli, A., Illig, T., Jacquemin, B., Katsouyanni, K., Koenig, W., Lanki, T., Pekkanen, J., Pershangen, G., Piccioto, S., Rückerl, R., Schaffrath Rosario, A., Stefanadis, C., and Sunyer, J. (2007), "Air Pollution and Inflammatory Response in Myocardial Infarction Survivors: Gene-Environment-Interactions in a High-Risk Group," *Inhalation Toxicology*, 19, 161–175.

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

Pope, C., Burnett, R., Thun, M., Calle, E., Krewski, D., Ito, K., and Thurston, G. (2002), "Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution," *Journal of the American Medical Association*, 287, 1132–1141.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.

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

SAS Institute Inc. (2004), *SAS/STAT©9.1 User's Guide.* Cary, NC: SAS Institute, Inc.

Scheipl, F., Greven, S., and Küchenhoff H. (2008), "Size and Power of Tests for a Zero Random Effect Variance or Polynomial Regression in Additive and Linear Mixed Models," *Computational Statistics & Data Analysis*, 52, 3283–3299.

Self, S., and Liang, K.-Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.

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