Journal of Statistical Software

August 2014, Volume 59, Issue 9. *http://www.jstatsoft.org/*

A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest

Ulrich Halekoh
University of Southern Denmark

Søren Højsgaard
Aalborg University

Abstract

When testing for reduction of the mean value structure in linear mixed models, it is common to use an asymptotic χ^2 test. Such tests can, however, be very poor for small and moderate sample sizes. The **pbkrtest** package implements two alternatives to such approximate χ^2 tests: The package implements (1) a Kenward-Roger approximation for performing F tests for reduction of the mean structure and (2) parametric bootstrap methods for achieving the same goal. The implementation is focused on linear mixed models with independent residual errors. In addition to describing the methods and aspects of their implementation, the paper also contains several examples and a comparison of the various methods.

Keywords: adjusted degree of freedom, denominator degree of freedom, F test, linear mixed model, **lme4**, R, parametric bootstrap, Bartlett correction.

1. Introduction

In this paper we address the question of testing for reduction of the systematic components in mixed effects models. Attention is restricted to models which are linear and where all random effects are Gaussian. The focus in this paper is on the implementation of these models in the **lme4** package (Bates, Maechler, Bolker, and Walker 2014a) for R (R Core Team 2014); specifically as implemented in the `lmer()` function. The package **pbkrtest** (Halekoh and Højsgaard 2014) implements the methods described in this paper and the package is available on the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=pbkrtest`.

It is always possible to exploit that the likelihood ratio (LR) test statistic has a limiting

χ^2 distribution as the amount of information in the sample goes to infinity. We shall refer to this test as *the asymptotic* χ^2 *test*. However, the χ^2 approximation can be poor and lead to misleading conclusions for small and moderate sample sizes. For certain types of studies it is possible to base the inference on an F statistic. Such studies generally need to be balanced in some way, for example, the number of observations in each treatment group being the same and so on. These balance requirements can often not be met in practice. Therefore there is a need for tests which, for a large class of linear mixed models, (1) are better than the asymptotic χ^2 test and (2) which are relatively easy to compute in practice.

The paper is structured as follows: Section 2 describes the problem addressed in more detail and sets the notation of the paper. Section 3 illustrates the problems related to tests in mixed models through several examples. In Section 4 we describe the approach taken by Kenward and Roger (1997) to address the inference problem. Section 5 describes an alternative approach based on parametric bootstrap methods. In Section 6 we apply the methods to several data sets. Section 7 contains a discussion and outlines some additional improvements that can be made to the implementation in **pbkrtest**.

Throughout this paper we will use the CRAN version 1.0-5 of the **lme4** package.

2. Preliminaries and notation

In this paper we focus on linear mixed models which, in the formulation of Laird and Ware (1982), are of the form

$$\mathbf{Y}^N = \mathbf{X}^{N\times p}\boldsymbol{\beta}^p + \mathbf{Z}^{N\times u}\mathbf{b}^u + \boldsymbol{\epsilon}^N, \tag{1}$$

where $\mathbf{Y}$ is an N vector of observables. The superscripts in Equation 1 refer to the dimension of the quantities and these superscripts will be omitted whenever possible in the following.

In (1), $\mathbf{X}$ and $\mathbf{Z}$ are design matrices of the fixed and random effect, $\mathbf{b}$ is the random effect vector distributed as $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is the vector of residual errors where $\mathbf{I}$ is the $N \times N$ identity matrix. It is assumed that $\mathbf{b}$ and $\boldsymbol{\epsilon}$ are independent. The covariance matrix of $\mathbf{Y}$ is therefore $\mathbb{V}\mathrm{ar}(\mathbf{Y}) = \boldsymbol{\Sigma}^{N\times N} = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}^\top + \sigma^2\mathbf{I}$. This model is a simplification of the more general model proposed in Laird and Ware (1982), who allow the covariance matrix of $\boldsymbol{\epsilon}$ to be a general positive definite matrix.

We are interested in testing hypotheses about the fixed effects in (1), i.e., testing for the smaller model

$$M_0 : \mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta}_0 + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathcal{C}(\mathbf{X}_0) \subset \mathcal{C}(\mathbf{X})$ with $\mathcal{C}(\mathbf{X})$ denoting the column space of $\mathbf{X}$. Let $d = \dim(\mathcal{C}(\mathbf{X})) - \dim(\mathcal{C}(\mathbf{X}_0))$. Notice that the structural forms of the random components of the two models are identical.

Testing the reduction of $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ to $\mathbb{E}(\mathbf{Y}) = \mathbf{X}_0\boldsymbol{\beta}_0$ can in some cases be made as an F test; one example is given in Section 3. However, in many practical cases, such an exact F test is not available and one often resorts to asymptotic tests. One approach is based on the LR test statistic T which is twice the difference of the maximized log-likelihoods

$$T = 2(\log L - \log L_0). \tag{3}$$

Under the hypothesis, T has an asymptotic χ^2_d distribution (Wilks 1938). The reduction of the large model to the small model can equivalently be expressed by the equation $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$

with a non-singular $d \times p$ restriction matrix $\mathbf{L}$. In Appendix B it is shown how $\mathbf{L}$ can be constructed from $\mathbf{X}$ and $\mathbf{X}_0$.

A test of the more general hypothesis $\mathbf{L}(\boldsymbol{\beta} - \boldsymbol{\beta}_H) = \mathbf{0}$ can be based on the Wald test statistic

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_H)^\top \mathbf{L}^\top (\mathbf{L}\hat{\mathbf{V}}\mathbf{L}^\top)^{-1} \mathbf{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_H), \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is an estimate for $\boldsymbol{\beta}$ and $\hat{\mathbf{V}}$ for the covariance matrix of $\hat{\boldsymbol{\beta}}$. In this paper we focus on the case where $\boldsymbol{\beta}_H = \mathbf{0}$. Under the hypothesis, W also has an asymptotic χ^2_d distribution and the Wald and the LR test are hence asymptotically equivalent.

The approximation of the null distribution of T or W by a χ^2_d distribution can for small samples be quite poor and this can lead to misleading conclusions. An example of this is given in Section 3. Nonetheless, this approximation is often used in practice – mainly because of the lack of attractive alternatives. This paper is aimed at providing some remedies for this.

(a) Kenward and Roger (1997) provide a modification of W given in (4). They also argue that this modified statistic is asymptotically distributed as an $F_{d,m}$ distribution for which they provide a method for estimating the denominator degrees of freedom m. We have implemented their work in the function `KRmodcomp()` for models of the form (1); notice in particular that attention is restricted to models for which the residuals are independent and have constant variance. Throughout this paper we shall refer to Kenward and Roger (1997) as KR.

(b) The second contribution of this paper is to determine either the full null distribution or moments of the null distribution of the LR test statistic (3) by a parametric bootstrap approach (Davison and Hinkley 1997, Chapter 4). This has been implemented in the function `PBmodcomp()`.

3. The degree-of-freedom issue for linear mixed models

In this section we discuss the degree-of-freedom issue on the basis of the `beets` dataset in the **pbkrtest** package. The `beets` data, which to our knowledge have not been published elsewhere, come from a split-plot experiment. Although the classical analysis of split-plot experiments is described in many places in the literature, see e.g., Cochran and Cox (1957, Chapter 7), we treat the topic in some detail in order to put the other parts of the article into context.

3.1. The sugar beets example

The experiment was laid out as follows: The effect of harvesting time and sowing time on (i) yield (in kg) and (ii) sugar percentage of sugar beets is investigated. Five different sowing dates and two different harvesting dates were used and the experiment was laid out in three blocks. The experimental plan is as follows:

```
Experimental plan for sugar beets experiment

Sowing dates:
  1: 4/4, 2: 12/4, 3: 21/4, 4: 29/4, 5: 18/5
```

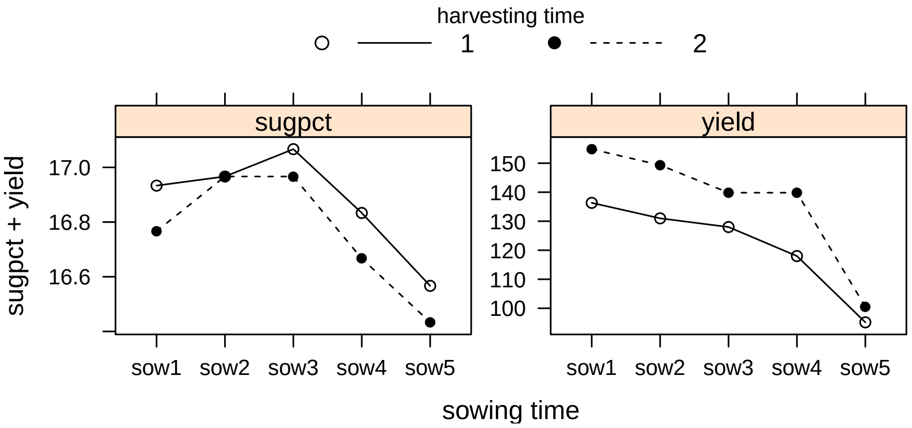

Figure 1: Dependence of sugar percentage and yield [kg] on sowing time and harvesting time.

```
Harvesting dates:
  1: 2/10, 2: 21/10

Plot allocation:
            |  Block 1        |  Block 2        |  Block 3        | Time
            +----------------|----------------|----------------+
Split-plots | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | Harvesting
 1-15       | s3 s4 s5 s2 s1 | s3 s2 s4 s5 s1 | s5 s2 s3 s4 s1 | Sowing
            -----------------|----------------|----------------|
Split-plots | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | Harvesting
16-30       | s2 s1 s5 s4 s3 | s4 s1 s3 s2 s5 | s1 s4 s3 s2 s5 | Sowing
            +----------------|----------------|----------------+
```

Each block is sub-divided into two whole-plots (a term used in the experimental design literature) which are harvested at two different dates. Each whole-plot is further sub-divided into five split-plots and each of the five sowing dates are allocated to one of these split-plots. So, for example, the first split-plot in the upper left corner above has harvest time `h1` (October 2nd) and sowing time `s3` (April 21st). All together there are hence 6 whole-plots and 30 split-plots. The harvesting time is called the whole-plot treatment and the sowing time is called the split-plot treatment. The area of each split-plot was $25m^2$.

In the following i denotes harvesting dates ($i = 1, 2$), j denotes block ($j = 1, 2, 3$) and k denotes sowing dates ($k = 1, \dots, 5$). Let $I = 2$, $J = 3$ and $K = 5$. For simplicity we assume that there is no interaction between sowing and harvesting time (this assumption is supported by Figure 1). A typical model for such an experiment would be

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + U_{ij} + \epsilon_{ijk}, \tag{5}$$

where $U_{ij} \sim N(0, \omega^2)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$. Notice that U_{ij} describes the random variation between whole-plots (within blocks) and the presence of this term implies that measurements on the same split-plot will be positively correlated.

3.2. The asymptotic χ^2 test

We can fit the models and test for no effect of sowing and harvesting time using the `lmer()` function from **lme4** (Bates *et al.* 2014a).

```
R> library("lme4")
R> data("beets", package = "pbkrtest")
R> sug <- lmer(sugpct ~ block + sow + harvest + (1 | block:harvest),
+    data = beets, REML = FALSE)
R> sug_no.harv <- update(sug, . ~ . - harvest)
R> sug_no.sow <- update(sug, . ~ . - sow)
```

We then proceed by testing for no effect of sowing and of harvesting time:

```
R> anova(sug, sug_no.sow)

Data: beets
Models:
sug_no.sow: sugpct ~ block + harvest + (1 | block:harvest)
sug: sugpct ~ block + sow + harvest + (1 | block:harvest)
           Df     AIC     BIC logLik deviance  Chisq Chi Df
sug_no.sow  6  -2.795   5.612  7.398  -14.795
sug        10 -79.998 -65.986 49.999  -99.998 85.203      4
           Pr(>Chisq)
sug_no.sow
sug         < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> anova(sug, sug_no.harv)

Data: beets
Models:
sug_no.harv: sugpct ~ block + sow + (1 | block:harvest)
sug: sugpct ~ block + sow + harvest + (1 | block:harvest)
            Df     AIC     BIC logLik deviance  Chisq Chi Df
sug_no.harv  9 -69.084 -56.473 43.542  -87.084
sug         10 -79.998 -65.986 49.999  -99.998 12.914      1
            Pr(>Chisq)
sug_no.harv
sug          0.0003261 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These tests are based on the limiting χ^2 distribution of the LR test statistic and suggest a highly significant effect of both sowing and harvesting time. Notice that above we have fitted the models with `REML = FALSE`, i.e., by maximum likelihood rather than by restricted maximum likelihood. We must do so for the following asymptotic χ^2 tests to make sense. However the test for no effect of harvesting time is misleading because the hierarchical structure of the data has not been appropriately accounted for. We shall discuss this important issue in detail below.

3.3. The exact F test

Consider a comparison of two sowing dates and of two harvesting dates. From (5) we get:

$$\begin{aligned} y_{ij1} - y_{ij2} &= \delta_1 - \delta_2 + \epsilon_{ij1} - \epsilon_{ij2} \sim N(\delta_1 - \delta_2, 2\sigma^2) && (6) \\ y_{1jk} - y_{2jk} &= \alpha_1 - \alpha_2 + U_{1j} - U_{2j} + \epsilon_{1jk} - \epsilon_{2jk} \sim N(\alpha_1 - \alpha_2, 2\omega^2 + 2\sigma^2). && (7) \end{aligned}$$

For the sowing dates the whole plot variation cancels out whereas the whole-plot variation prevails for the harvesting dates. This means that the effect of whole-plot treatments are determined with smaller precision than the effect of split-plot treatments. In some applications (for example if whole-plots are animals and split plots correspond to an application of a treatment at different time points) it is often the case that ω^2 is considerably larger than σ^2. Estimated contrasts for sowing dates and harvesting dates hence become

$$\begin{aligned} \frac{1}{IJ}\sum_{ij}(y_{ij1} - y_{ij2}) &\sim N(\delta_1 - \delta_2, \frac{2}{J}\{\sigma^2/I\}) && (8) \\ \frac{1}{JK}\sum_{jk}(y_{1jk} - y_{2jk}) &\sim N(\alpha_1 - \alpha_2, \frac{2}{J}\{\omega^2 + \sigma^2/K\}). && (9) \end{aligned}$$

Test for no effect of harvesting time

Next we consider test statistics. We shall use the notation $y_{i++} = \sum_{jk} y_{ijk}$ and $\bar{y}_{i++} = y_{i++}/(JK)$ etc. Also we let $\tilde{\sigma}^2 = \omega^2 + \sigma^2/K$. The test for no effect of harvesting time is based on the marginal model obtained after averaging over the sowing dates, i.e.,

$$\bar{y}_{ij+} = \mu + \alpha_i + \beta_j + \bar{\delta}_+ + \bar{U}_{ij} + \bar{\epsilon}_{ij+} \sim N(\mu + \alpha_i + \beta_j + \bar{\delta}_+, \tilde{\sigma}^2). \quad (10)$$

Observe that $\bar{y}_{ij+}$ in (10) has the structure of a model for a balanced two-way layout (with factors harvesting times and block) without replicates.

Let $SS_I = \sum_{ijk}(\bar{y}_{i++} - \bar{y}_{+++})^2$ be the sums of squares associated with harvesting time. Direct calculation shows that $\mathbb{E}(SS_I) = Q_I + (I-1)K\tilde{\sigma}^2$ where $Q_I = JK\sum_i(\alpha_i - \bar{\alpha}_+)^2$. The corresponding mean squares $MS_I = SS_I/(I-1)$ then has expectation $\mathbb{E}(MS_I) = Q_I/(I-1) + K\tilde{\sigma}^2$. Since $Q_I \geq 0$ and $Q_I = 0$ iff all α_i are identical, MS_I can be used for constructing a test for no effect of harvesting time.

The relevant error sum of squares becomes the residual sum of squares in the marginal model (10), i.e., $SS_{I+J} = \sum_{ijk}(\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2$. Direct calculation shows that $\mathbb{E}(SS_{I+J}) = (I-1)(J-1)K\tilde{\sigma}^2$. Define the mean squares as $MS_{I+J} = SS_{I+J}/[(I-1)(J-1)]$. Then $\mathbb{E}(MS_{I+J}) = K\tilde{\sigma}^2$. From this we obtain the F statistic for testing for no effect of harvesting time:

$$F = \frac{MS_I}{MS_{I+J}} \sim F_{(I-1),(I-1)(J-1)} \text{ under the hypothesis.} \quad (11)$$

Test for no effect of sowing time

Let $SS_K = \sum_{ijk}(y_{++k} - \bar{y}_{+++})^2 = \sum_{ijk}\{(\delta_k - \bar{\delta}_+) + (\bar{\epsilon}_{++k} - \bar{\epsilon}_{+++})\}^2$ be the sum of squares associated with sowing time and let $MS_K = SS_K/(K-1)$ be the corresponding mean squares. Defining $Q_K = IJ\sum_k(\delta_k - \bar{\delta}_+)^2$, a direct calculation shows that $\mathbb{E}(MS_K) = Q_K/(K-1) + \sigma^2$. The corresponding error term becomes $SS_\epsilon = \sum_{ijk}(y_{ijk} - y_{ij+} - y_{++k} + y_{+++})^2$ which is the residual sum of squares for a linear normal model with an effect of sowing time plus an interaction between harvesting time and block. Define the mean squares as $MS_\epsilon = SS_\epsilon/(IJ-1)(K-1)$ and direct calculation shows that $\mathbb{E}(MS_\epsilon) = \sigma^2$. From this we obtain the F statistic for testing of no effect of sowing times as

$$F = \frac{MS_K}{MS_\epsilon} \sim F_{(K-1),(IJ-1)(K-1)} \text{ under the hypothesis.} \tag{12}$$

Making the relevant F tests with `aov()`

The `aov()` function makes the tests in (11) and (12) as follows:

```
R> beets$bh <- with(beets, interaction(block, harvest))
R> summary(aov(sugpct ~ block + sow + harvest + Error(bh), data = beets))
```

```
Error: bh
          Df  Sum Sq Mean Sq F value Pr(>F)
block      2 0.03267 0.01633   2.579 0.2794
harvest    1 0.09633 0.09633  15.211 0.0599 .
Residuals  2 0.01267 0.00633
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
          Df Sum Sq Mean Sq F value   Pr(>F)
sow        4   1.01  0.2525     101 5.74e-13 ***
Residuals 20   0.05  0.0025
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence we get $\hat{\tilde{\sigma}}^2 = 0.00633$ and $\hat{\sigma}^2 = 0.0025$. From $\tilde{\sigma}^2 = \omega^2 + \sigma^2/K$ we obtain the estimate of ω^2 as $\hat{\omega}^2 = \hat{\tilde{\sigma}}^2 - \hat{\sigma}^2/K = 0.00633 - 0.0025/5 = 0.00583$.

Hence, when the hierarchical structure of the experiment has been accounted for, the effect of harvesting time is not significant at the 5% level.

3.4. The Mississippi influents example

The `Mississippi` dataset in the **SASmixed** package (Bates 2011) contains the nitrogen concentration (in PPM) from several sites at six randomly selected influents of the Mississippi river.

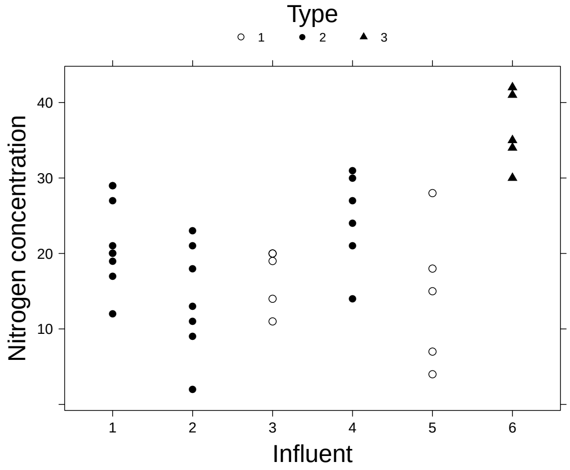

Figure 2: Nitrogen concentration in PPM at six different influents of the Mississippi differentiated for three types of watershed.

```
R> data("Mississippi", package = "SASmixed")
R> Mississippi$influent <- factor(Mississippi$influent)
R> Mississippi$Type <- factor(Mississippi$Type)
R> head(Mississippi)
```

```
  influent  y Type
1        1 21    2
2        1 27    2
3        1 29    2
4        1 17    2
5        1 19    2
6        1 12    2
```

The influents were characterized according to watersheds as follows. Type = 1: No farmland in watershed (influents no. 3 and 5); Type = 2: Less than 50% farmland in watershed (influents no. 1, 2 and 4); Type = 3: More than 50% farmland in watershed (influent no. 6). Measurements from the same influent are expected to be similar and there is no particular interest in the individual influents. It is more interesting to investigate the effect of the watershed type on the nitrogen concentration.

A typical model for such data would be

$$y_i = \alpha_{\mathit{Type}(i)} + U_{\mathit{influent}(i)} + \epsilon_i,$$

where $U_l \sim N(0, \omega^2)$ and $\epsilon_i \sim N(0, \sigma^2)$. The χ^2 test suggests that the effect of `Type` is highly significant:

```
R> miss1 <- lmer(y ~ Type + (1 | influent), data = Mississippi, REML = FALSE)
R> miss0 <- update(miss1, . ~ . - Type)
R> anova(miss1, miss0)
```

```
Data: Mississippi
Models:
miss0: y ~ (1 | influent)
miss1: y ~ Type + (1 | influent)
      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
miss0  3 262.56 267.39 -128.28   256.56
miss1  5 256.57 264.63 -123.29   246.57 9.9834      2   0.006794 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Trusting large sample asymptotic results is questionable. If the data had been balanced such that there were the same number of influents for each watershed type and the same number of recordings for each influent, then we could have made a proper F test along the lines of Section 3.1.

An alternative is to analyze the means for each influent and this yields a much less clear indication of an effect of watershed type. To calculate the means we employ the **doBy** package (Højsgaard and Halekoh 2013)

```
R> library("doBy")
R> Miss.mean <- summaryBy(y ~ influent + Type, data = Mississippi,
+    FUN = mean)
R> detach("package:doBy")
R> miss1_lm <- lm(y.mean ~ Type, data = Miss.mean)
R> anova(miss1_lm)
```

```
Analysis of Variance Table

Response: y.mean
          Df  Sum Sq Mean Sq F value  Pr(>F)
Type       2 298.276 149.138  7.0702 0.07322 .
Residuals  3  63.282  21.094
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Approximate F statistic and the KR approximation

In this section we describe first the KR approach of testing the hypothesis $\mathbf{L}(\boldsymbol{\beta}-\boldsymbol{\beta}_H)=\mathbf{0}$ for a more general model than (1). We describe then the class of linear mixed models fitted with `lmer()` for which function `KRmodcomp()` of the package **pbkrtest** provides the KR approach.

4.1. A multivariate normal model

KR consider for $\mathbf{Y}$ the multivariate normal model

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

The covariance-matrix $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\gamma})$ is assumed to be a function of M parameters collected in the vector $\boldsymbol{\gamma}$. We denote the REML estimates of these parameters with $\hat{\boldsymbol{\gamma}}$. The unbiased REML estimate of $\boldsymbol{\beta}$ is then, see Kackar and Harville (1984),

$$\hat{\boldsymbol{\beta}} = \mathbf{\Phi}(\hat{\boldsymbol{\gamma}})\mathbf{X}^\top \mathbf{\Sigma}(\hat{\boldsymbol{\gamma}})^{-1}\mathbf{Y} \text{ with } \mathbf{\Phi}(\hat{\boldsymbol{\gamma}}) = \left(\mathbf{X}^\top \mathbf{\Sigma}(\hat{\boldsymbol{\gamma}})^{-1}\mathbf{X}\right)^{-1}, \tag{13}$$

where $\mathbf{\Phi}$ is the covariance matrix of the asymptotic distribution of $\boldsymbol{\beta}$ and $\mathbf{\Phi}(\hat{\boldsymbol{\gamma}})$ is a consistent estimate of $\mathbf{\Phi}$.

A scaled Wald-type statistics for testing the hypothesis $\mathbf{L}(\boldsymbol{\beta} - \boldsymbol{\beta}_H) = \mathbf{0}$ is

$$F = \frac{1}{d}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_H)^\top \mathbf{L}^\top (\mathbf{L}\hat{\mathbf{V}}\mathbf{L}^\top)^{-1}\mathbf{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_H), \tag{14}$$

where $\hat{\mathbf{V}}$ is some positive definite symmetric matrix. The usual Wald test statistic uses $\hat{\mathbf{V}} = \mathbf{\Phi}(\hat{\boldsymbol{\gamma}})$. In this case F has asymptotically a $\frac{1}{d}\chi^2_d$ distribution (which can be thought of as the limiting distribution of an $F_{d,m}$ distribution when $m \to \infty$.) For some models, F has an exact F distribution under the hypothesis. One example of this is a balanced one-way analysis of variance.

4.2. The approach of Kenward and Roger

KR modify the statistic F in (14) to improve the small sample properties by approximating the distribution of F by an $F_{d,m}$ distribution, and they also provide a method for calculating the denominator degrees of freedom m. The fundamental idea is to calculate the approximate mean and variance of their statistic and then match moments with an F distribution to obtain the denominator degrees of freedom. KR left out some detail in the derivation of their method. Alnosaier (2007) provides more details, weakens some of the assumptions for the approach, and extends the list of models for which it is known that the approach yields exact F tests.

KR take two steps to improve the small sample distributional properties of F. Firstly, Kackar and Harville (1984) showed that the covariance matrix of $\hat{\boldsymbol{\beta}}$ can be written as the sum $\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}}) = \mathbf{\Phi} + \mathbf{\Lambda}$ where $\mathbf{\Lambda}$ expresses the bias by which the asymptotic covariance matrix $\mathbf{\Phi}$ underestimates $\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}})$. KR combine a Taylor approximation to $\mathbf{\Lambda}$ with a bias-corrected modification of $\mathbf{\Phi}(\hat{\boldsymbol{\gamma}})$ using second order Taylor expansion to derive a new estimate $\mathbf{\Phi}_A(\hat{\boldsymbol{\gamma}})$. In the statistic F in (14), KR replace the matrix $\hat{\mathbf{V}}$ with $\hat{\mathbf{V}} = \mathbf{\Phi}_A(\hat{\boldsymbol{\gamma}})$. Secondly, KR derive a scaling factor λ (such that the statistic they consider is λF) and a denominator degree of freedom m by matching approximations of the expectation and variance of λF with the moments of an $F_{d,m}$ distribution. In more detail, KR derive an approximation for the expectation $E^\star$ and variance $V^\star$ of F based on a first order Taylor expansion. Then they solve the system of equations

$$\mathbb{E}(F) \approx \lambda E^\star = \mathbb{E}(F_{d,m}) = \frac{m}{m-2}, \tag{15}$$

$$\mathbb{V}\mathrm{ar}(F) \approx \lambda^2 V^\star = \mathbb{V}\mathrm{ar}(F_{d,m}) = \frac{2m^2(d+m-2)}{d(m-2)^2(m-4)} = \{\mathbb{E}(F_{d,m})\}^2 \frac{2(d+m-2)}{d(m-4)}, \tag{16}$$

where $\mathbb{E}(F_{d,m})$ and $\mathbb{V}\mathrm{ar}(F_{d,m})$ denote expectation and variance of an $F_{d,m}$ distributed random variable. The $E^\star$ and $V^\star$ are slightly modified without changing the order of approximation such that for the balanced one-way ANOVA model and the Hoteling's T^2 model the exact

F tests are reproduced (Alnosaier 2007, Chapters 4.1 and 4.2). We shall refer to these two steps as *the Kenward-Roger approximation* (or KR approximation in short). Details of the computations are provided in Appendix A.1. In particular, the solution to the equations above is given in (27). Recall that the mean of an $F_{d,m}$ distribution exists provided that $m > 2$ and the variance exists provided that $m > 4$. The moment matching method does however not prevent estimates of m that are less than or equal to 2. KR did not address this problem and neither did we in our implementation.

4.3. Models for which `KRmodcomp()` provides tests

The `KRmodcomp()` function of the **pbkrtest** package provides the KR approximation for linear mixed models of the form (1) where $\mathbf{\Sigma}$ is a sum of known matrices

$$\mathbf{\Sigma} = \sum_r \gamma_r \mathbf{G}_r + \sigma^2 \mathbf{I}. \tag{17}$$

The matrices $\mathbf{G}_r$ are usually very sparse matrices. Variance component models and random coefficient models are models which have this simplified covariance structure. For details we refer to Appendix A.1.

5. Parametric bootstrap

An alternative approach is based on parametric bootstrap, and this is also implemented in **pbkrtest**. The setting is the LR test statistic T for which we have an observed value t_{obs}. The question is in which reference distribution t_{obs} should be evaluated; i.e., what is the null distribution of T. Instead of relying on the approximation of the null distribution by a χ^2_d distribution one can use parametric bootstrap:

First, create B (e.g., $B = 1000$) bootstrap samples $y^1, \dots, y^B$ by simulating from $\hat{f}_0(y)$ (where $\hat{f}_0$ denotes the fitted distribution under the hypothesis). Next, calculate the corresponding values $T^* = \{t^1, \dots, t^B\}$ of the LR test statistic. For what follows, let E^*_T and V^*_T denote sample mean and sample variance of T^*. These simulated values can then be regarded as samples from the null distribution and these values can be used in different ways which are implemented in the `PBmodcomp()` function. The labels below refer to the output from `PBmodcomp()`, see Section 6:

`PBtest:` Direct calculation of tail probabilities: The values T^* provide an empirical null distribution in which t_{obs} can be evaluated. Let $I(x)$ be an indicator function which is 1 if x is true and 0 otherwise. Following Davison and Hinkley (1997, Chapter 4), the p value then becomes

$$p = \frac{n_{extreme} + 1}{B + 1}, \quad \text{where} \quad n_{extreme} = \sum_{k=1}^{B} I(t^k \geq t_{obs}). \tag{18}$$

`Gamma:` Approximate the null distribution by a gamma distribution with mean E^*_T and variance V^*_T.

`Bartlett:` Improve the LR test statistic by a Bartlett type correction: The LR test statistic T can be scaled to better match the χ^2_d distribution as $T_B = Td/E^*_T$.

`F`: Approximate the null distribution of T/d by an $F_{d,m}$ distribution with mean E^*/d. This yields a single equation for deriving m, namely $m = 2E^*_T/(E^*_T - d)$.

Notice that the parametric bootstrap approach is not restricted to linear mixed models of the type discussed in this paper. In fact, **pbkrtest** implements parametric bootstrap also for generalized linear and for generalized linear mixed models.

We shall make the following remarks about the quantities mentioned in the listing above (in Section 6 we also provide graphical illustrations of these approaches):

1. Regarding `PBtest` recall that the definition of a p value for a composite hypothesis is (see e.g., Casella and Berger 2002, p. 397)

   $$p = \sup_{\boldsymbol{\theta}} \mathsf{P}_{\boldsymbol{\theta}}(T > t_{obs}),$$

   where the supremum is taken over all possible values $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \boldsymbol{\gamma})$ under the hypothesis. When this supremum can not be evaluated in practice, it is often exploited that for large samples $\mathsf{P}_{\boldsymbol{\theta}}$ is approximately the distribution function for a χ^2_d distribution which is independent of $\boldsymbol{\theta}$. Implicit in (18) is therefore a definition of a bootstrapped p value to be $p = \mathsf{P}_{\hat{\boldsymbol{\theta}}}(T > t_{obs})$ and then (18) is used for the calculation. Determining the tail of a distribution as in (18) by sampling requires a large number of samples B (but how large B must be depends in practice on the size of t_{obs}).

2. The quantities `Gamma`, `Bartlett` and `F` are based on assuming a parametric form of the null distribution such that the null distribution can be determined from at most the first two sample moments of T^*. It requires in general fewer samples to obtain credible estimates for these moments than for obtaining the tail probabilities in (18). We have no compelling mathematical argument why T^* should be well approximated by a gamma distribution, but since a χ^2_d distribution is also a gamma distribution, it is appealing to approximate T^* by a gamma distribution where we match the first two moments. In practice this means that we obtain a distribution which can have a heavier tail than the χ^2_d distribution. The idea behind adjusting the LR test statistic by a Bartlett type correction as in $T_B = \frac{T}{E^*_T/d}$ is to obtain a a statistic whose distribution becomes closer to a χ^2_d distribution (cfr. Cox 2006, p. 130). See also e.g., Jensen (1993) for a more comprehensive treatment of Bartlett corrections. Approximating the distribution of T/d by an $F_{d,m}$ distribution can be motivated as follows: Under the hypothesis, T is in the limit χ^2_d distributed so T/d has in the limit a χ^2_d/d distribution with expectation 1 and variance $2/d$. This is, loosely speaking, the same as an $F_{d,m}$ distribution with an infinite number of denominator degrees of freedom m. By estimating m as $m = 2E^*_T/(E^*_T - d)$ we obtain the increased flexibility of an F distribution with a larger variance than $2/d$, i.e., a distribution with a heavier tail than that of a χ^2_d/d distribution.

3. A general problem with the parametric bootstrap approach is that it is computationally intensive. However the **pbkrtest** package allows for the samples to be drawn in parallel by utilizing several processors on the computer.

4. The parametric bootstrap approach may be modified into a sequential scheme as follows: Instead of fixing the number of parametric bootstrap samples B in advance, on may

draw samples until h (e.g., $h = 20$) values of the test statistic which are more extreme than the observed test statistic have been obtained. If this takes B' samples then the p value to report is $(h+1)/(B'+1)$. If there is little evidence against the hypothesis then only a small number B' of simulations would be needed. This idea is the parametric bootstrap version of the approach of Besag and Clifford (1991) for calculating sequential Monte Carlo p values. This idea is illustrated below.

6. Applications of the methods

This section contains applications of the methods described in Sections 4 and 5 to the examples in Section 3. This section also contains additional examples. In connection with parametric bootstrap, **pbkrtest** allows for samples to be drawn in parallel by utilizing several processors on the computer via the facilities provided in the **parallel** package. To do so we create clusters:

```
R> nc <- detectCores()
R> clus <- makeCluster(rep("localhost", nc))
```

6.1. The sugar beets example

For the sugar beets example of Section 3.1, the KR approximation provides the following results.

Harvesting time

The test for harvesting time yields

```
R> (sug.kr.h <- KRmodcomp(sug, sug_no.harv))

F-test with Kenward-Roger approximation; computing time: 0.16 sec.
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
       stat   ndf   ddf F.scaling p.value
Ftest 15.21  1.00  2.00         1  0.0599 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> (sug.pb.h <- PBmodcomp(sug, sug_no.harv, cl = clus))

Parametric bootstrap test; time: 66.09 sec; samples: 1000 extremes: 37;
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
         stat df   p.value
LRT    12.914  1 0.0003261 ***
PBtest 12.914    0.0379620 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sowing time

The test for sowing time yields

```
R> (sug.kr.s <- KRmodcomp(sug, sug_no.sow))

F-test with Kenward-Roger approximation; computing time: 0.17 sec.
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + harvest + (1 | block:harvest)
      stat ndf ddf F.scaling   p.value    
Ftest  101   4  20         1 5.741e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> (sug.pb.s <- PBmodcomp(sug, sug_no.sow, cl = clus))

Parametric bootstrap test; time: 46.34 sec; samples: 1000 extremes: 0;
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + harvest + (1 | block:harvest)
         stat df   p.value    
LRT    85.203  4 < 2.2e-16 ***
PBtest 85.203     0.000999 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First, it is noted that the p values reported from both `KRmodcomp()` and `PBmodcomp()` generally are (1) within the same order of magnitude and (2) close to the results of the exact F test of Section 3.1. Hence the results would all suggest the same qualitative conclusion, namely that there is little (if any) evidence for an effect of harvesting time and strong evidence for an effect of sowing time. Secondly, it is noticed that `KRmodcomp()` is much faster than `PBmodcomp()` in these examples. However the difference in computing time is much smaller for other types of models/datasets; for example for certain random regression models (not reported in this paper).

6.2. Warnings from the optimizers

The built-in optimizers for `lmer()` are the `"bobyqa"` method for bound constrained optimization without derivatives from the **minqa** package, (Bates, Mullen, Nash, and Varadhan 2014b), see also Powell (2009) and Nelder-Mead optimization as implemented in the `Nelder_Mead()` function in the **lme4** package.

The default optimization method in `lmer()` is the `"bobyqa"` method. When using this method in connection with parametric bootstrap, as for example in

```
R> PBmodcomp(sug, sug_no.sow)
```

on may encounter warnings of the following form:

```
Warning in optwrap(object@optinfo$optimizer, ff, x0, lower = lower,
  control = control$optCtrl,  :
  convergence code 3 from bobyqa: bobyqa -- a trust region step failed
  to reduce q
```

For the specific case above, this happens in a small number of simulations, say in up to 5% of the cases. One may alternatively use Nelder-Mead optimization as:

```
R> sugNM <- lmer(sugpct ~ block + sow + harvest + (1 | block:harvest),
+    data = beets, REML = FALSE,
+    control = lmerControl(optimizer = "Nelder_Mead"))
R> sugNM_no.sow <- update(sugNM, . ~ . - sow)
R> PBmodcomp(sugNM, sugNM_no.sow)
```

Nelder-Mead optimization, on the other hand, can result in the following warning (which for the specific case above happens very rarely, say in 5 out of 1000 simulations):

```
Warning message:
In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
  Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

In either case, the warnings indicate convergence problems and in practical use of `PBmodcomp()` one must check that these do not happen in too many simulations.

6.3. How good are the parametric reference distributions?

In Section 5 the idea of approximating the bootstrap distribution by an F distribution, a gamma distribution and a scaled χ^2 distribution (Bartlett correction) was introduced. In this section we illustrate how well these approximations work.

The results of these approximations are obtained using `summary()`:

```
R> summary(sug.pb.h)


Parametric bootstrap test; time: 66.09 sec; samples: 1000 extremes: 37;
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
            stat      df    ddf   p.value
PBtest   12.9142               0.0379620 *
Gamma    12.9142               0.0321151 *
Bartlett  4.1764  1.0000       0.0409900 *
F        12.9142  1.0000 2.9559 0.0378266 *
LRT      12.9142  1.0000       0.0003261 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> summary(sug.pb.s)
```

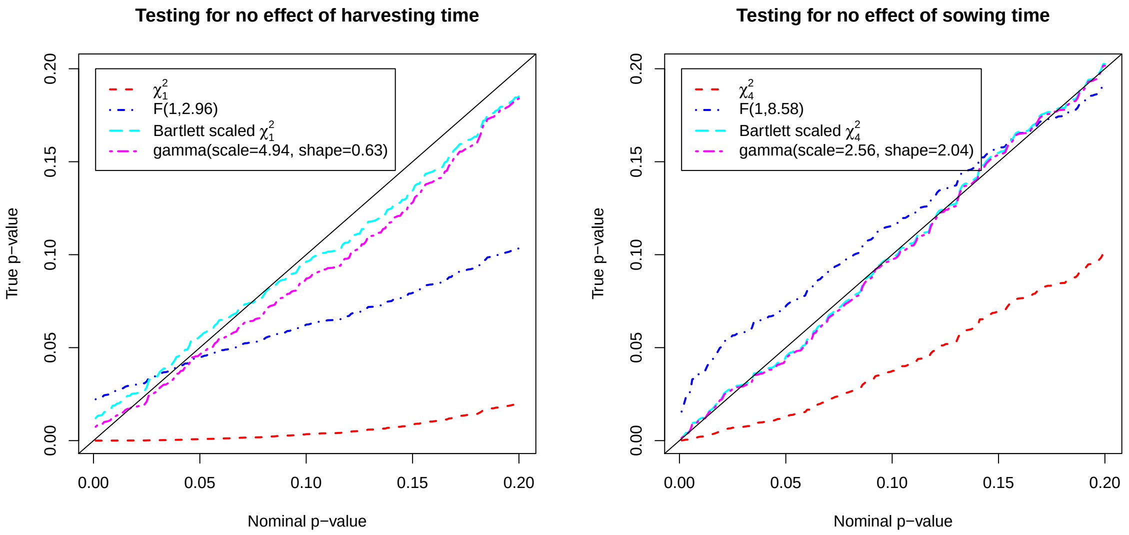

Figure 3: Comparisons of the p values of the bootstrapped null distribution with the p values of the approximating parametric distributions. Left: Testing for no effect of harvesting time. Right: Testing for no effect of sowing time.

```
Parametric bootstrap test; time: 46.34 sec; samples: 1000 extremes: 0;
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + harvest + (1 | block:harvest)
           stat     df    ddf   p.value
PBtest   85.203                0.0009990 ***
Gamma    85.203                1.393e-13 ***
Bartlett 65.343  4.000         2.179e-13 ***
F        21.301  4.000 8.5803 0.0001714 ***
LRT      85.203  4.000         < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For a range of p values from 0.0001 to 0.200, i.e., those p values which are typically of practical relevance, we have calculated the quantiles q in the bootstrap reference distribution. We have then calculated the tail probabilities corresponding to these quantiles in the approximating gamma distribution, F distribution, the Bartlett scaled distribution, and the asymptotic χ^2 distribution of the LR statistic. The results are shown in Figure 3. It is clear form these plots that the Bartlett scaled distribution, the gamma distribution and (to a lesser extent) the F distribution approximates the bootstrap distribution quite well whereas the χ^2 distributions approximate the reference distribution poorly.

6.4. A sequential version of parametric bootstrap

As mentioned above, one may create a sequential version of the parametric bootstrap scheme as follows (where the aim is to speed up computations): Instead of fixing the number of samples B in advance, on may draw samples until h (e.g., $h = 20$) values of the test statistic which are more extreme than the observed test statistic have been obtained. If this takes B'

samples then the p value to report is $(h+1)/(B'+1)$. If there is little evidence against the hypothesis then only a small number B' of simulations would be needed. This functionality is not implemented in **pbkrtest** at the time of writing, but it is straight forward to create a minimal implementation:

```
R> seqPBmodcomp <- function(largeModel, smallModel, h = 20, nsim = 1000) {
+    t.start <- proc.time()
+    chunk.size <- 50
+    nchunk <- nsim %/% chunk.size
+    LRTstat <- getLRT(largeModel, smallModel)
+    ref <- NULL
+    for (ii in 1:nchunk) {
+      ref <- c(ref, PBrefdist(largeModel, smallModel, nsim = chunk.size))
+      n.extreme <- sum(ref > LRTstat["tobs"])
+      if (n.extreme >= h)
+        break
+    }
+    ans <- PBmodcomp(largeModel, smallModel, ref = ref)
+    ans$ctime <- (proc.time() - t.start)[3]
+    ans
+  }
R> seqPBmodcomp(sug, sug_no.harv, h = 10)

Parametric bootstrap test; time: 20.34 sec; samples: 300 extremes: 10;
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
         stat df   p.value
LRT    12.914  1 0.0003261 ***
PBtest 12.914    0.0365449 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running `PBmodcomp()` without parallel computing takes about a minute so the saving in computing time is significant.

6.5. The Mississippi influents example

For the `Mississippi` data of Section 3.4 our methods provide the following results:

```
R> KRmodcomp(miss1, miss0)

F-test with Kenward-Roger approximation; computing time: 0.18 sec.
large : y ~ Type + (1 | influent)
small : y ~ (1 | influent)
        stat    ndf    ddf F.scaling p.value
Ftest 6.3690 2.0000 3.3195   0.99967 0.07307 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R> summary(PBmodcomp(miss1, miss0, cl = clus))

Parametric bootstrap test; time: 64.72 sec; samples: 1000 extremes: 66;
large : y ~ Type + (1 | influent)
small : y ~ (1 | influent)
           stat     df    ddf  p.value
PBtest   9.9834               0.066933 .
Gamma    9.9834               0.056237 .
Bartlett 5.4047 2.0000        0.067046 .
F        4.9917 2.0000 4.3608 0.074557 .
LRT      9.9834 2.0000        0.006794 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence we obtain p values which are in the order of 10 times the p value provided by the χ^2 approximation. The p values we obtain are in good accordance with the p value obtained when analyzing the means as done in Section 3.4.

6.6. Random coefficient regression – A simulation study

KR perform a small simulation study on a simple random coefficient regression model. We made a simulation using the same model set-up and use it to compare the results between the different tests we provide and to the KR approach as implemented by the `MIXED` procedure of the SAS software system (SAS Institute Inc. 2013).

Kenward and Roger (1997, Table 4) consider the following random coefficient model

$$y_{jt_j} = \beta_0 + \beta_1 \cdot t_j + A_j + B_j \cdot t_j + \epsilon_{jt_j}$$

with

$$\mathbb{C}\mathrm{ov}(A_j, B_j) = \begin{bmatrix} 0.250 & -0.133 \\ -0.133 & 0.250 \end{bmatrix} \qquad \text{and} \qquad \mathbb{V}\mathrm{ar}(\epsilon_{jt}) = 0.25. \tag{19}$$

There are $j = 1, \ldots, 24$ observed subjects divided into three groups of eight subjects. For each group observations are made at the non overlapping times $t = 0, 1, 2; t = 3, 4, 5$ and $t = 6, 7, 8$. The data for the simulation were generated under the assumption that $\beta_0 = \beta_1 = 0$, (A_j, B_j) and ϵ_{jt} are normally distributed with zero expectation, (A_j, B_j) are independent from ϵ_{tj} and observations from different subjects are independent.

The full model and the reduced models are fitted by:

```
R> Mod <- lmer(y ~ 1 + t + (1 + t | subject))
R> Mod_no.int <- lmer(y ~ 0 + t + (1 + t | subject))
R> Mod_no.slope <- lmer(y ~ 1 + (1 + t | subject))
```

The results are shown in Table 1. The LR test gives for both parameters and for all significance levels (the α's) anti-conservative p values as expected. For example, consider testing $\beta_0 = 0$ on the 5% significance level. The LR test rejects the hypothesis in 6.7% of the simulations, i.e., the tests overstate the evidence against the hypothesis that $\beta_0 = 0$. For all other approaches,

Parm	$\alpha \times 100$	LR	KR (R)	KR (SAS)	`PBtest`	`Bartlett`	`Gamma`	`F`
β_0	1	1.7	0.7	1.4	0.9	1.0	1.2	0.7
β_1	1	1.4	1.0	1.0	0.8	0.9	1.0	0.7
β_0	5	6.7	4.4	5.2	5.2	5.2	5.6	5.1
β_1	5	6.1	5.1	5.1	4.9	4.9	5.1	4.7
β_0	10	12.7	9.2	10.0	10.3	10.4	10.8	11.1
β_1	10	11.5	10.1	10.0	9.8	9.8	10.0	10.0

Table 1: Observed test sizes (×100) for three test levels $\alpha = 0.01, 0.05, 0.1$ for $H_0 : \beta_k = 0$ from the random coefficient model. The results are based on 20000 simulations, for the bootstrapped p values 500 subsamples were taken. KR (R) and KR (SAS) are the KR approximations as implemented in `KRmodcomp()` and in SAS; the other results refer to the null distribution of the LR test statistic, either the χ^2 approximation (LR) or bootstrapped values. `PBtest` relates to the raw parametric bootstrap p value. The other p values are based on approximations to the bootstrap distribution either via a Bartlett correction, a gamma or an F distribution.

the observed test-levels are closer to the nominal levels than for the LR test and in most cases the p values are anti-conservative.

The KR approach from our implementation yields slightly conservative results for the tests on the intercept parameter β_0 and the tests in column `F` yield conservative results for the lowest nominal level. The difference of the results of the KR approach between our implementation and that of SAS may lie in the different treatment of cases where the covariance matrix $\mathbf{\Gamma}$ is singular.

6.7. Testing a hypothesis $\mathbf{L}\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\beta}_H$

We present now an example on testing a hypothesis $\mathbf{L}(\boldsymbol{\beta} - \boldsymbol{\beta}_H) = \mathbf{0}$ with `KRmodcomp()` via the specification of a matrix $\mathbf{L}$. We illustrate this with the sugar beets data. Assume that one wants to test the hypothesis that the difference between the second and first sowing date and between the third and second sowing date are equal to 0.1. The parameter vector $\boldsymbol{\beta}$ from the model fit `sug` in Section 3.2 has the entries

```
R> names(fixef(sug))

[1] "(Intercept)"  "blockblock2"  "blockblock3"  "sowsow2"
[5] "sowsow3"      "sowsow4"      "sowsow5"      "harvestharv2"
```

The restriction matrix $\mathbf{L}$ is then given as

```
R> L <- matrix(0, nrow = 2, ncol = 8)
R> colnames(L) <- names(fixef(sug))
R> L[1, "sowsow2"] <- 1
R> L[2, c("sowsow2", "sowsow3")] <- c(-1, 1)
R> t(L)

             [,1] [,2]
(Intercept)     0    0
```

```
blockblock2     0    0
blockblock3     0    0
sowsow2         1   -1
sowsow3         0    1
sowsow4         0    0
sowsow5         0    0
harvestharv2    0    0
```

With $\mathbf{c} = (0.1, 0.1)^\top$, a vector $\boldsymbol{\beta}_H$ to be used in $\mathbf{L}(\boldsymbol{\beta} - \boldsymbol{\beta}_H) = \mathbf{0}$ is $\boldsymbol{\beta}_H = \mathbf{L}^- \mathbf{c}$ where $\mathbf{L}^-$ is a generalized inverse of $\mathbf{L}$. A generalized inverse can be obtained with the function `ginv()` of the package **MASS** (Venables and Ripley 2002).

```
R> library("MASS")
R> beta_H <- ginv(L) %*% c(0.1, 0.1)
R> t(beta_H)

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    0    0    0  0.1  0.2    0    0    0
```

The hypothesis is then tested with

```
R> KRmodcomp(sug, L, betaH = beta_H)

F-test with Kenward-Roger approximation; computing time: 0.14 sec.
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : L beta = L betaH
L=
2 x 8 sparse Matrix of class "dgCMatrix"

[1,] . . .  1 . . . .
[2,] . . . -1 1 . . .
betaH=
     [,1]
[1,]  0.0
[2,]  0.0
[3,]  0.0
[4,]  0.1
[5,]  0.2
[6,]  0.0
[7,]  0.0
[8,]  0.0
         stat     ndf     ddf F.scaling p.value
Ftest  1.5556  2.0000 20.0000         1  0.2356
```

If the restriction matrix is not of full row rank it will be replaced by a matrix of full row rank using Gram-Schmidt orthogonalization as in the following example on the difference between the third and second sowing date.

```
R> L[1, c("sowsow2", "sowsow3")] <- c(1, -1)
R> L[2, c("sowsow2", "sowsow3")] <- c(-1, 1)
R> KRmodcomp(sug, L)

F-test with Kenward-Roger approximation; computing time: 0.14 sec.
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : L beta = L betaH
L=
1 x 8 sparse Matrix of class "dgCMatrix"

[1,] . . . -0.7071068 0.7071068 . . .
betaH=
[1] 0
      stat ndf ddf F.scaling p.value
Ftest    3   1  20         1 0.09866 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.8. Constructing tests manually

It is illustrative to see how to construct such tests manually using tools provided in **pbkrtest**. We construct the t test statistic for testing $\mathbf{L}\boldsymbol{\beta} = 0$ for a specific choice of $\mathbf{L}$. Consider an unbalanced version of the `beets` data:

```
R> beetsUB <- subset(beets, !(harvest == "harv1" & block == "block1" &
+    (sow %in% c("sow1", "sow2"))))
R> ftable(xtabs(~ harvest + block + sow, data = beetsUB))

               sow sow1 sow2 sow3 sow4 sow5
harvest block
harv1   block1        0    0    1    1    1
        block2        1    1    1    1    1
        block3        1    1    1    1    1
harv2   block1        1    1    1    1    1
        block2        1    1    1    1    1
        block3        1    1    1    1    1
```

A comparison of sowing times `sow2` and `sow3` can under an additive model be made with the following choice of $\mathbf{L}$:

```
R> sugUB <- lmer(sugpct ~ block + sow + harvest + (1 | block:harvest),
+    data = beetsUB)
R> L <- c(0, 0, 0, 1, -1, 0, 0, 0)
```

pbkrtest provides an adjusted variance-covariance matrix for the regression parameters and the estimated degrees of freedom with:

```
R> Va <- vcovAdj(sugUB)
R> ddf <- get_ddf_Lb(sugUB, L)
R> ddf

[1] 17.77283
```

From this we can construct the usual t statistic and corresponding p value:

```
R> b.hat <- fixef(sugUB)
R> Lb.hat <- sum(L * b.hat)
R> Va.Lb.hat <- t(L) %*% Va %*% L
R> t.stat <- as.numeric(Lb.hat / sqrt(Va.Lb.hat))
R> t.stat

[1] -1.191398

R> p.value  <- 2 * pt(abs(t.stat), df = ddf, lower.tail = FALSE)
R> p.value

[1] 0.2491653
```

Notice that the same result (in terms of p value) is obtained with the following (where the F statistic is the squared t statistic from above):

```
R> KRmodcomp(sugUB, matrix(L, nrow = 1))
```

6.9. Computing least-squares means with adjusted degrees of freedom

The **doBy** package (Højsgaard and Halekoh 2013) provides methods for computing least-squares means (sometimes also called LS means) using adjusted degrees of freedom. Using the model defined in Section 6.8, the least-squares means for `harvest` is obtained with

```
R> library("doBy")
R> LSmeans(sugUB, effect = "harvest")

  estimate         se       df   t.stat      p.value      lwr
1 16.87591 0.02228984 2.090753 757.1123 1.015562e-06 16.78389
2 16.76000 0.02127864 1.841845 787.6443 4.191143e-06 16.66048
       upr harvest
1 16.96794   harv1
2 16.85952   harv2
```

If one wants to test all contrasts with the control, one can use the `lsmeans()` function of the **lsmeans** package (Lenth 2013). The following code shows how to compute these comparisons to the control `sow = 1`:

```
R> library("lsmeans")
R> lsmeans(sugUB , spec = trt.vs.ctrl1 ~ sow)[[2]]

              estimate         SE       df   t.ratio p.value
sow2 - sow1  0.1400000 0.03022757 18.00368   4.63153 0.00083
sow3 - sow1  0.1751075 0.02946746 17.77283   5.94240 0.00005
sow4 - sow1 -0.0915592 0.02946746 17.77283  -3.10713 0.02439
sow5 - sow1 -0.3415592 0.02946746 17.77283 -11.59106 0.00000
    p values are adjusted using the sidak method for 4 tests
```

7. Discussion

In this paper we have presented our implementation of a KR approximation for tests in linear mixed models. In the implementation, there are several matrices of the order $N \times N$ where N is the number of observations. We have exploited that several of the matrices involved in the computations in many cases will be sparse via the facilities in the **Matrix** package (Bates and Maechler 2014). Nonetheless, the current implementation of the KR approximation does not always scale to large datasets. As an example, consider a repeated measurement problem in which repeated measurements are made on a collection of subjects. If there are many subjects and the time series for each subject is short then there is a sparseness to be exploited. On the other hand, if there are a few long time series then the matrices involved will have a non-negligible number of non-zero elements. One approach to speed up the computations is to compute the average of the observed and expected information matrices rather than the expected information matrix. This can lead to substantial improvements in computing time because some of the computationally most intractable terms vanish in the average information. See Gilmour, Thompson, and Cullis (1995) and Jensen, Mantysaari, Madsen, and Thompson (1996) for details. This may become available in later versions of **pbkrtest**. A very specific issue which we have no clear answer to is how the KR approximation should be modified in case of a singular estimate of the covariance matrix.

Contrary to the KR approximation, the parametric bootstrap approach has the advantage that it is easy to implement; all that is required is a way of sampling data from the fitted model under the hypothesis. Furthermore, parametric bootstrap is straightforward to implement for many types of models. Parametric bootstrap is already implemented for generalized linear models and for generalized linear mixed models in **pbkrtest**.

A problem with the parametric bootstrap approaches is the randomness of the results; repeated applications to the same dataset do not give entirely identical results. Moreover, calculating the reference distribution by sampling is computationally demanding. However, **pbkrtest** implements the possibility of parallel computing of the reference distribution using multiple processors via the **parallel** package. There are various possibilities for speeding up the parametric bootstrap computations: (1) Instead of fixing the number of parametric bootstrap samples B in advance, one may adopt a sequential scheme in which sampling continues until a pre-specified number of extreme samples have been obtained. This idea, which is closely related to the approach of Besag and Clifford (1991) for calculating sequential Monte Carlo p values, has been illustrated in the paper. (2) The Bartlett type correction we implemented is such a possibility because the correction depends only on the mean of the simulated null distribution and the gamma approximation depends only on the mean and variance of the simulated null distribution. Estimating these two moments will in general require fewer simulations than estimating the tail of the null distribution. Hence, if one chooses to focus on these two distributions then one *may* get credible results with fewer samples. (3) It may also be possible to devise a sequential sampling scheme such that sampling stops when the estimates of the first or the first two moments have stabilized.

In the beginning of the paper it is stated that we consider models which are nested with respect to the structure of the mean values but where the random effects are the same. For the KR approach, this is formally not a requirement, because it is only the random structure of the large model that matters. The small model is only used for the construction of the restriction matrix. In contrast, for the parametric bootstrap approach, it is only the structure of the random effect of the smaller model that matters.

An important final comment is that we do not in any way claim to have an omnibus panacea solution to a difficult problem. Instead we have provided two practically applicable alternatives to relying on large sample asymptotics when testing for the reduction of the mean value in general linear mixed models.

Acknowledgments

This work was supported by the Danish National Advanced Technology Foundation through the ILSORM project. The authors would like to thank Steffen Lauritzen (Oxford University), Jens Ledet Jensen (Aarhus University) and Poul Svante Eriksen (Aalborg University) and the referees for very constructive comments.

References

Alnosaier WS (2007). *Kenward-Roger Approximate F Test for Fixed Effects in Mixed Linear Models*. Ph.D. thesis, Oregon State University. URL `http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/5262/mydissertation.pdf?sequence=1`.

Bates D (2011). ***SASmixed**: Data Sets from 'SAS System for Mixed Models'*. R package version 1.0-4, URL `http://CRAN.R-project.org/package=SASmixed`.

Bates D (2013). "Linear Mixed Model Implementation in **lme4**." *Vignette "Implementation – Implementation Details" of R package **lme4** version 0.999999-2*, University of Wisconsin-Madison.

Bates D, Maechler M (2014). ***Matrix**: Sparse and Dense Matrix Classes and Methods*. R package version 1.1-4, URL `http://CRAN.R-project.org/package=Matrix`.

Bates D, Maechler M, Bolker B, Walker S (2014a). ***lme4**: Linear Mixed-Effects Models Using **Eigen** and S4*. R package version 1.1-7, URL `http://CRAN.R-project.org/package=lme4`.

Bates D, Mullen KM, Nash JC, Varadhan R (2014b). ***minqa**: Derivative-free optimization algorithms by quadratic approximation*. R package version 1.2.3, URL `http://CRAN.R-project.org/package=minqa`.

Besag J, Clifford P (1991). "Sequential Monte Carlo p-Values." *Biometrika*, **78**(2), 301–304.

Casella G, Berger RL (2002). *Statistical Inference*. 2nd edition. Duxbury.

Cochran WG, Cox GM (1957). *Experimental Design*. 2nd edition. Chapman and Hall.

Cox DR (2006). *Principles of Statistical Inference*. Cambridge.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Gilmour AR, Thompson R, Cullis BR (1995). "Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models." *Biometrics*, **51**(4), 1440–1450.

Halekoh U, Højsgaard S (2014). ***pbkrtest**: Parametric Bootstrap and Kenward Roger Based Methods for Mixed Model Comparison*. R package version 0.4-0, URL http://CRAN.R-project.org/package=pbkrtest.

Harville DA (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag.

Højsgaard S, Halekoh U (2013). ***doBy**: Groupwise Summary Statistics, General Linear Contrasts, Population Means (Least-Squares-Means), and other Utilities*. R package version 4.5-10, URL http://CRAN.R-project.org/package=doBy.

Jensen J, Mantysaari EA, Madsen P, Thompson R (1996). "Residual Maximum Likelihood Estimation of (Co)Variance Components in Multivariate Mixed Linear Models using Average Information." *Journal of the Indian Society of Agricultural Statistics*, **49**, 215–236.

Jensen JL (1993). "A Historical Sketch and Some New Results on the Improved Log Likelihood Ratio Statistic." *Scandinavian Journal of Statistics*, **20**(1), 1–15.

Kackar RN, Harville DA (1984). "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models." *Journal of the American Statistical Association*, **79**(388), 853–862.

Kenward MG, Roger JH (1997). "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics*, **53**(3), 983–997.

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

Lenth RV (2013). ***lsmeans**: Least-Squares Means*. R package version 2.11, URL http://CRAN.R-project.org/package=lsmeans.

Powell MJD (2009). "The BOBYQA algorithm for bound constrained optimization without derivatives." *Technical Report Report No. DAMTP 2009/NA06*, Centre for Mathematical Sciences, University of Cambridge, UK.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

SAS Institute Inc (2013). *SAS/STAT Software, Version 9.4*. Cary, NC. URL `http://www.sas.com/`.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York.

Wilks SS (1938). "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics*, **9**(1), 60–62.

A. Technical details for the KR approximation

A.1. Computations related to the KR approximation

In this appendix more details of the implementation of the approach of KR in `KRmodcomp()` are given. First we describe the structure of the design matrix of the random effects $\mathbf{Z}$ and the related structure of the covariance matrix $\mathbf{\Gamma}$. Secondly, the sequence of computations with the matrices available from a fitted model object from `lmer()` and the derived matrices are given.

Structure for $\mathbf{Z}$ *and* $\mathbf{\Gamma}$

The description of the structure of $\mathbf{Z}$ and $\mathbf{\Gamma}$ draws on the description given in a vignette of the **lme4** package for versions prior to 1.0 (Bates 2013). Structural changes in these matrices for later versions have been accounted for in the following.

For a linear mixed model fitted with `lmer()` it is assumed that we have $i = 1, \dots, f$ grouping factors denoted by $\mathbf{f}_i$. It is allowed that $\mathbf{f}_i = \mathbf{f}_{i'}$ for $i \neq i'$. The ith grouping factor $\mathbf{f}_i$ has g_i levels and there are q_i random effects for each level. The random effects for group level j are collected in the vector $\mathbf{b}_{ij} = (b_{ij1}, \dots, b_{ijq_i})^\top$ and the random effects of $\mathbf{f}_i$ are $\mathbf{b}_i^\top = (\mathbf{b}_{ij}^\top)$.

It is assumed that the random effects from different grouping factors and from different levels of a grouping factor are independent, i.e.,

$$\mathbb{C}\text{ov}(\mathbf{b}_i, \mathbf{b}_{i'}) = 0 \text{ for } i \neq i' \qquad \text{and} \qquad \mathbb{C}\text{ov}(\mathbf{b}_{ij}, \mathbf{b}_{ij'}) = 0 \text{ for } j \neq j'.$$

The covariance matrix of the random effects for grouping level j of factor $\mathbf{f}_i$ is independent of the grouping level and is denoted by

$$\mathbb{V}\text{ar}(\mathbf{b}_{ij}) = \mathbf{\Gamma}_i^{q_i \times q_i} = (\gamma_{i;rr'}).$$

We assume that all of the elements of $\mathbf{\Gamma}_i$ are parameters that vary freely except that $\mathbf{\Gamma}_i$ must be positive definite. Hence $\mathbb{V}\text{ar}(\mathbf{b}_i) = \mathbf{I}^{g_i \times g_i} \otimes \mathbf{\Gamma}_i$ where $\otimes$ denotes the Kronecker product and $\mathbf{I}^{g_i \times g_i}$ the identity matrix of dimension g_i.

For the sugar beets example there is one factor, the interaction $U_{i'j'}$ between block and harvest. In the present notation $f = 1, g_1 = 6, q_1 = 1$, $\mathbf{b}_1 = (b_{1,1}, \dots, b_{1,6})^\top$ and $\mathbf{Z}_1 = \mathbf{I}^6 \otimes \mathbf{1}^5$ where $\mathbf{1}^5$ is a vector of ones.

For the random coefficient model of the simulation example there is one grouping factor, subject, with 24 levels, hence $f = 1, g_1 = 24$ and $q_1 = 2$ random effects (A_j, B_j) for subject j such that $\mathbf{b}_1 = (A_1, B_1, \dots, A_{24}, B_{24})$,

$$\mathbf{Z}_1^{72 \times 48} = \begin{bmatrix} 1 & 0 & & & \\ 1 & 1 & & & \\ 1 & 2 & & & \\ & & \cdots & & \\ & & & 1 & 6 \\ & & & 1 & 7 \\ & & & 1 & 8 \end{bmatrix} \tag{20}$$

and $\mathbf{\Gamma}_1$ is the matrix in (19). If in the simulation example the two random effects A_j and B_j were assumed to be uncorrelated the model would be specified with `lmer()` as `lmer(y ~ A + t + (1 | subject) + (0 + t | subject))`. Now there are two grouping factors, both are equal to subject, hence $f = 2, g_1 = g_2 = 24, q_1 = q_2 = 1$, $\mathbf{b}_1 = (A_1, \dots, A_{24})^\top$, $\mathbf{b}_2 = (B_1, \dots, B_{24})^\top$ and

$$\mathbf{Z}_1^{72\times 24} = \begin{bmatrix} 1 & & \\ 1 & & \\ 1 & & \\ & \dots & \\ & & 1 \\ & & 1 \\ & & 1 \end{bmatrix}, \qquad \mathbf{Z}_2^{72\times 24} = \begin{bmatrix} 0 & & \\ 1 & & \\ 2 & & \\ & \dots & \\ & & 6 \\ & & 7 \\ & & 8 \end{bmatrix}. \tag{21}$$

Let $\boldsymbol{\gamma}_i = (\gamma_{i;11}, \gamma_{i;2,1}, \dots, \gamma_{i;q_i1}, \gamma_{i;22}, \dots, \gamma_{i;q_iq_i})^\top$ denote the $s_i = q_i(q_i+1)/2$ vector of the elements of the lower triangular $\mathbf{\Gamma}_i$. For the kth element $\gamma_{i;k}$ of $\boldsymbol{\gamma}_i$ it holds that $\gamma_{i;k} = \gamma_{i;rr'}$ where $k = (r-1)(q_i - r/2) + r'$. Then we may write

$$\mathbf{\Gamma}_i = \sum_{k=1}^{s_i} \gamma_{i;k}\mathbf{E}_{i;k}.$$

The $\mathbf{E}_{i;k}$ are the $q_i \times q_i$ symmetric incidence matrices with ones at the position (r, r') and (r', r). Now,

$$\begin{aligned} \mathbb{V}\mathrm{ar}(\mathbf{Z}_i\mathbf{b}_i) &= \mathbf{Z}_i \, \mathbb{V}\mathrm{ar}(\mathbf{b}_i)\mathbf{Z}_i^\top = \mathbf{Z}_i(\mathbf{I}^{g_i\times g_i} \otimes \mathbf{\Gamma}_i)\mathbf{Z}_i^\top \\ &= \mathbf{Z}_i(\mathbf{I}^{g_i\times g_i} \otimes \sum_{k=1}^{s_i} \gamma_{i;k}\mathbf{E}_{i;k})\mathbf{Z}_i^\top = \sum_{k=1}^{s_i} \gamma_{i;k}\mathbf{Z}_i(\mathbf{I}^{g_i\times g_i} \otimes \mathbf{E}_{i;k})\mathbf{Z}_i^\top. \end{aligned}$$

With $\mathbf{D}_i = \sum_{k=1}^{s_i} \gamma_{i:k}\mathbf{Z}_i(\mathbf{I}^{g_i\times g_i} \otimes \mathbf{E}_{i;k})\mathbf{Z}_i^\top$ the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{Y}$ is

$$\mathbf{\Sigma} = \sum_{i=1}^{f} \mathbb{V}\mathrm{ar}(\mathbf{Z}_i\mathbf{b}_i) + \mathbb{V}\mathrm{ar}\,(\boldsymbol{\epsilon}) = \sum_{i=1}^{f} \mathbf{D}_i + \sigma^2\mathbf{I}^{N\times N}, \tag{22}$$

where f is the number of grouping factors. Let $\boldsymbol{\gamma}$ denote the vector of length M made by concatenation of the vectors $\boldsymbol{\gamma}_i$ and, as the last element, the σ^2. Let $\mathbf{G}_r = \mathbf{Z}_i(\mathbf{I}^{g_i\times g_i} \otimes \mathbf{E}_{i;r})\mathbf{Z}_i^\top$ where r refers to the rth element in $\boldsymbol{\gamma}$ and i is the group factor $\mathbf{f}_i$ related to the covariance parameter γ_r. Note that $\mathbf{G}_M = \mathbf{I}^{N\times N}$.

Then $\mathbf{\Sigma}$ can be written as a linear combination of known matrices

$$\mathbf{\Sigma} = \sum_{r=1}^{M} \gamma_r\mathbf{G}_r. \tag{23}$$

For the sugar beets example $\mathbf{G}_1 = \mathbf{I}^{6\times 6} \otimes \mathbf{J}^{5\times 5}$ where $\mathbf{J} = \mathbf{1}^5\mathbf{1}^{5^\top}$. For the simulation example $\mathbf{G}_1 = \mathbf{I}^{24\times 24} \otimes \mathbf{J}^3$ is related to $\gamma_1 = 0.25$ and $\mathbf{G}_2$ is related to the covariance $\gamma_2 = -0.133$, with $\mathbf{G}_2 = \mathrm{diag}(\mathbf{I}^{3\times 3} \otimes \mathbf{A}, \mathbf{I}^{3\times 3} \otimes \mathbf{B}, \mathbf{I}^{3\times 3} \otimes \mathbf{C})$ and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 6 & 7 & 8 \\ 7 & 8 & 9 \\ 8 & 9 & 10 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 12 & 13 & 14 \\ 13 & 14 & 15 \\ 14 & 15 & 16 \end{bmatrix}. \tag{24}$$

The representation (23) has two simplifying consequences. Firstly, the derivative of $\mathbf{\Sigma}$ with respect to $\boldsymbol{\gamma}$ is (see e.g., Harville 1997, Equation 8.15)

$$\frac{\partial \mathbf{\Sigma}^{-1}}{\partial \gamma_r} = -\mathbf{\Sigma}^{-1}\mathbf{G}_r\mathbf{\Sigma}^{-1}.$$

Secondly, the estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$ can be expressed without using higher derivatives of $\mathbf{\Sigma}^{-1}$ (cf. Kenward and Roger 1997, Equation 5).

Implementation of the KR approach in the `KRmodcomp()` *function*

The following estimates are directly provided by `lmer()`: (1) the parameter estimate $\hat{\boldsymbol{\beta}}$, (2) the vector $\hat{\boldsymbol{\gamma}}$ of the REML estimated covariance parameters and (3) the estimate $\mathbf{\Phi}(\hat{\boldsymbol{\gamma}})$ of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

The estimate of the covariance matrix for $\hat{\boldsymbol{\gamma}}$

$$\mathbb{C}\text{ov}(\hat{\boldsymbol{\gamma}}) = \mathbf{W}^{M\times M}$$

is not directly available from `lmer()`, but is estimated in (26) from the inverse information matrix, (cf. also Kenward and Roger 1997, Equations 4 and 5).

The implementation of the KR approximation in **pbkrtest** is based on the following quantities.

1. For each covariance parameter γ_r in $\boldsymbol{\gamma}$ we use

$$\mathbf{G}_r^{N\times N} = \mathbf{Z}_i(\mathbf{I}^{g_i\times g_i}\otimes \mathbf{E}_r)\mathbf{Z}_i^\top, \tag{25}$$

   where i refers to the group for the covariance parameter γ_r.

2. Then the estimated covariance matrix for $\mathbf{Y}$ becomes $\hat{\mathbf{\Sigma}} = \sum_r^M \hat{\gamma}_r\mathbf{G}_r$.

3. For the computations to follow, we define the following auxiliary matrices:

   - $\mathbf{T}^{N\times p} = \mathbf{\Sigma}^{-1}\mathbf{X}$
   - $\mathbf{H}_r^{N\times N} = \mathbf{G}_r\mathbf{\Sigma}^{-1}$, $r = 1,\dots,M$
   - $\mathbf{O}_r^{N\times p} = \mathbf{G}_r\mathbf{\Sigma}^{-1}\mathbf{X} = \mathbf{H}_r\mathbf{X}$, $r = 1,\dots,M$
   - $\mathbf{\Omega}_r^{N\times N} = \frac{\partial\mathbf{\Sigma}^{-1}}{\partial\gamma_r} = -\mathbf{\Sigma}^{-1}\mathbf{G}_r\mathbf{\Sigma}^{-1}$, $r = 1,\dots,M$. Notice that $\mathbf{\Omega}_r$ is not used in any computation in the implementation in **pbkrtest** but $\mathbf{\Omega}_r$ appears in the derivations below.

4. For each covariance parameter γ_r let

$$\mathbf{P}_r^{p\times p} = \mathbf{X}^\top\mathbf{\Omega}_r\mathbf{X} = -\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{G}_r\mathbf{\Sigma}^{-1}\mathbf{X} = -\mathbf{T}^\top\mathbf{G}_r\mathbf{T} = -\mathbf{T}^\top\mathbf{O}_r.$$

5. For each pair (γ_r,γ_s) of covariance parameters let

$$\begin{aligned}\mathbf{Q}_{rs}^{p\times p} &= \mathbf{X}^\top\mathbf{\Omega}_r\hat{\mathbf{\Sigma}}\mathbf{\Omega}_s\mathbf{X} = \mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{G}_r\mathbf{\Sigma}^{-1}\mathbf{G}_s\mathbf{\Sigma}^{-1}\mathbf{X}\\ &= \mathbf{T}^\top\mathbf{G}_r\mathbf{\Sigma}^{-1}\mathbf{G}_s\mathbf{T} = \mathbf{O}_r^\top\mathbf{\Sigma}^{-1}\mathbf{O}_s.\end{aligned}$$

   Notice that $\mathbf{Q}_{rs}$ is generally not symmetric but $\mathbf{Q}_{rs} = \mathbf{Q}_{sr}^\top$ and hence $\mathbf{Q}_{rs} + \mathbf{Q}_{sr}$ is symmetric. This symmetry property is exploited below. Moreover, $\text{tr}(\mathbf{Q}_{rs}) = \text{tr}(\mathbf{Q}_{sr})$.

6. For each pair (γ_r, γ_s) of covariance parameters let

$$\begin{aligned} K_{rs} &= \operatorname{tr}(\mathbf{\Omega}_r \mathbf{\Sigma} \mathbf{\Omega}_s \mathbf{\Sigma}) \\ &= \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{G}_r\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{G}_s\mathbf{\Sigma}^{-1}\mathbf{\Sigma}) = \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{G}_r\mathbf{\Sigma}^{-1}\mathbf{G}_s). \end{aligned}$$

7. Twice the expected information matrix for $\hat{\boldsymbol{\gamma}}$ then becomes:

$$2 \cdot \{\mathbf{I}_E\}_{rs} = K_{rs} - 2 \cdot \operatorname{tr}(\mathbf{\Phi}\mathbf{Q}_{rs}) + \operatorname{tr}(\mathbf{\Phi}\mathbf{P}_r\mathbf{\Phi}\mathbf{P}_s).$$

   Notice that $\operatorname{tr}(\mathbf{\Phi}\mathbf{Q}_{rs}) = \operatorname{tr}(\mathbf{\Phi}\mathbf{Q}_{sr})$ and $\operatorname{tr}(\mathbf{\Phi}\mathbf{P}_r\mathbf{\Phi}\mathbf{P}_s) = \operatorname{tr}(\mathbf{\Phi}\mathbf{P}_s\mathbf{\Phi}\mathbf{P}_r)$.

8. The asymptotic covariance matrix of the random effects parameters becomes

$$\mathbb{C}\mathrm{ov}(\hat{\gamma}) = \mathbf{W}^{M\times M} = 2 \cdot \mathbf{I}_E^{-1}. \tag{26}$$

9. Define

$$\begin{aligned} \mathbf{U}^{p\times p} &= \sum_{r=1}^{M}\sum_{s=1}^{M} W_{rs}(\mathbf{Q}_{rs} - \mathbf{P}_r\mathbf{\Phi}\mathbf{P}_s) \\ &= \sum_{1\le r<s\le M} W_{rs}(\mathbf{Q}_{rs} + \mathbf{Q}_{rs}^\top - \mathbf{P}_r\mathbf{\Phi}\mathbf{P}_s - \mathbf{P}_s\mathbf{\Phi}\mathbf{P}_r) + \sum_{r=1}^{M} W_{rr}(\mathbf{Q}_{rr} - \mathbf{P}_r\mathbf{\Phi}\mathbf{P}_r). \end{aligned}$$

   Notice that the last equation holds because of $\mathbf{Q}_{sr} = \mathbf{Q}_{rs}^\top$.
   Letting $\tilde{\mathbf{U}} = \sum_{1\le r<s\le M} W_{rs}(\mathbf{Q}_{rs} - \mathbf{P}_r\mathbf{\Phi}\mathbf{P}_s)$, one can write alternatively

$$\mathbf{U} = \tilde{\mathbf{U}} + \tilde{\mathbf{U}}^\top + \sum_{r=1}^{M} W_{rr}(\mathbf{Q}_{rr} - \mathbf{P}_r\mathbf{\Phi}\mathbf{P}_r).$$

10. The adjusted estimate of $\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\beta}})$ is then

$$\hat{\mathbf{\Phi}}_A = \mathbf{\Phi}(\hat{\gamma}) + 2 \cdot \hat{\mathbf{\Lambda}}, \quad \text{where } \hat{\mathbf{\Lambda}}^{p\times p} = \hat{\mathbf{\Phi}}\mathbf{U}\hat{\mathbf{\Phi}},$$

   and the adjusted test statistic is (where d is the rank of $\mathbf{L}$)

$$F = \frac{1}{d}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_H)^\top \mathbf{L}^\top (\mathbf{L}\hat{\mathbf{\Phi}}_A\mathbf{L}^\top)^{-1}\mathbf{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_H).$$

11. KR derive a scaling factor λ for the F statistic given above (such that the statistic they finally propose is λF) and a denominator degrees of freedom m by matching approximate first and second moments of the λF statistic with the moments of an $F_{d,m}$ distribution. In this connection KR use the following quantities:

   (a) $\mathbf{\Theta} = \mathbf{L}^\top(\mathbf{L}\mathbf{\Phi}\mathbf{L}^\top)^{-1}\mathbf{L}$

   (b) $A_1 = \sum_{r=1}^{M}\sum_{s=1}^{M} W_{rs} \operatorname{tr}(\mathbf{\Theta}\mathbf{\Phi}\mathbf{P}_i\mathbf{\Phi}) \operatorname{tr}(\mathbf{\Theta}\mathbf{\Phi}\mathbf{P}_j\mathbf{\Phi})$ (where W_{rs} are the elements of the covariance matrix $\mathbf{W}$ from Equation 26).

(c) With $\circ$ denoting the Hadamard product,

$$\begin{aligned} A_2 &= \sum_s^M \sum_s^M W_{rs} \operatorname{tr}(\mathbf{\Theta\Phi P}_i \mathbf{\Phi\Theta\Phi P}_j \mathbf{\Phi}) \\ &= \sum_r^M \sum_s^M W_{rs} \mathbf{1}^\top \Big[(\mathbf{\Phi\Theta\Phi P}_i) \circ (\mathbf{\Phi\Theta\Phi P}_j) \Big] \mathbf{1}. \end{aligned}$$

(d) $B = \frac{1}{2d}(A_1 + 6A_2)$

(e) $E^* = 1/(1 - \frac{A_2}{d})$

(f) $V^\star = \frac{2}{d}\left(\frac{1+c_1B}{(1-c_2B)^2(1-c_3B)}\right)$. The c_is are simple functions of A_1, A_2 and d.

(g) $\rho = V^*/(2[E^*]^2)$

$E^\star$ and $V^\star$ are approximate expectation and variance of F based on the first order Taylor expansion of F.

12. Then KR end up with the following values for m and λ:

$$m = 4 + \frac{d+2}{d\rho - 1} \qquad \text{and} \qquad \lambda = \frac{m}{E^*(m-2)}. \tag{27}$$

A.2. Some numerical issues

In the computation of ρ we encountered numerical problems in the calculation of ρ for some models where the division of two numbers both equal to zero are encountered. One can write ρ as

$$\rho = \frac{1}{2}\left(\frac{D}{V_1}\right)^2 \cdot \frac{V_0}{V_2},$$

where $V_0 = 1 + c_1B$, $V_1 = 1 - c_2B$, $V_2 = 1 - c_3B$ and $D = 1 - A_2/d$. V_1 and D can become simultaneously very small yielding an unreliable ratio D/V_1. We resolve this problem by setting the ratio to 1 if $\max(|D|, |V_1|) < 10^{-11}$.

For example, for a simple block design,

$$Y_{bt} = \mu + \alpha_t + \epsilon_b + \epsilon_{bt}, \quad b = 1, \ldots, n_b, \quad t = 1, \ldots, n_t, \tag{28}$$

one has for $n_t = 2$ and $n_b = 3$ or for $n_t = 3$ and $n_b = 2$ an exact F test with $m = 2$ denominator degrees of freedom. We have for a specific application of this model found that D and V_1 were very close to zero. If we define the ratio to be 1 in this case then we end up with the correct answer, i.e., with $m = 2$. For the same design but for $n_t = 2$ and $n_b = 5$ or $n_t = 3$ and $n_b = 3$ we have for a specific application found that $V_2 = 0$ which leads to $\rho = \infty$. This caused no problem since the correct $m = 4$ degrees of freedom are obtained from Equation 27.

B. From model matrix to restriction matrix and vice versa

Referring to (1) and (2), let $\mathbf{X}$ and $\mathbf{X}_0$ be model matrices with full column rank and dimensions $N \times p$ and $N \times p_0$. We wish to construct a restriction matrix $\mathbf{L}$ such that $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \wedge \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ is equivalent to $\mathbb{E}(\mathbf{Y}) = \mathbf{X}_0\boldsymbol{\beta}_0$.

Let $\mathbf{P}$ and $\mathbf{P}_0$ denote the orthogonal projection matrices onto $\mathcal{C}(\mathbf{X})$ and $\mathcal{C}(\mathbf{X}_0)$. One choice of $\mathbf{L}$ is $\mathbf{L} = (\mathbf{I} - \mathbf{P}_0)\mathbf{P}\mathbf{X} = (\mathbf{P} - \mathbf{P}_0)\mathbf{X}$, where $(\mathbf{P} - \mathbf{P}_0)$ is the orthogonal projection onto the orthogonal complement of $\mathcal{C}(\mathbf{X}_0)$ in $\mathcal{C}(\mathbf{X})$.

Any element of $\mathcal{C}(\mathbf{X})$ can be written as $\mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$ such that

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = (\mathbf{I} - \mathbf{P}_0)\mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}_0\mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\beta} + \mathbf{P}_0\mathbf{X}\boldsymbol{\beta}. \tag{29}$$

If $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ then $\mathbf{X}\boldsymbol{\beta} = \mathbf{P}_0\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X}_0)$. On the other hand, if $\mathbb{E}(\mathbf{Y}) = \mathbf{X}_0\boldsymbol{\beta}_0$, then there exists a $\boldsymbol{\beta}$ such that $\mathbf{X}_0\boldsymbol{\beta}_0 = \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_0\mathbf{X}\boldsymbol{\beta}$ and hence from (29) $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Now $\mathbf{L}$ is an $N \times p$ matrix but it only contains $d = p - p_0$ linearly independent rows, and we only need to extract these to obtain a valid restriction matrix.

The computations in **pbkrtest** are done via the QR decomposition of the augmented matrix $\mathbf{D} = [\mathbf{X}_0 : \mathbf{X}]$, i.e. $\mathbf{D} = \mathbf{Q}\mathbf{R}$. The matrix $\mathbf{Q}_0$ of the first p_0 columns of $\mathbf{Q}$ has $\mathcal{C}(\mathbf{Q}_0) = \mathcal{C}(\mathbf{X}_0)$. The matrix $\mathbf{Q}_1$ of the following $p - p_0$ columns of $\mathbf{Q}$ is a basis for the orthogonal complement of $\mathcal{C}(\mathbf{X}_0)$ in $\mathcal{C}(\mathbf{X})$. Hence $\mathbf{Q}_1\mathbf{Q}_1^\top$ is the orthogonal projection onto this complement and therefore $\mathbf{L} = \mathbf{Q}_1\mathbf{Q}_1^\top\mathbf{X}$. Since $\mathbf{Q}_1\mathbf{Q}_1^\top$ and $\mathbf{Q}_1^\top$ have the same nullspace, a $(p - p_0) \times p$ restriction matrix $\mathbf{L}$ is obtained as $\mathbf{L} = \mathbf{Q}_1^\top\mathbf{X}$.

Next consider the opposite situation: Given $\mathbf{X}$ and $\mathbf{L}$ we want to derive $\mathbf{X}_0$. Let $\mathbf{W}$ denote a $p \times p_0$ matrix such that $\mathcal{C}(\mathbf{W})$ is equal to the nullspace of $\mathbf{L}$. In **pbkrtest**, $\mathbf{W}$ is found from a QR decomposition of $\mathbf{L}^\top$. Hence for $\mathbf{m} = \mathbf{X}\boldsymbol{\beta} \wedge \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ we have $\boldsymbol{\beta} = \mathbf{W}\mathbf{z}$, say, and hence $\mathbf{m} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{W}\mathbf{z}$ so that $\mathbf{X}_0$ can be taken as $\mathbf{X}_0 = \mathbf{X}\mathbf{W}$.

Affiliation:

Ulrich Halekoh
Department of Epidemiology, Biostatistics and Biodemography
University of Southern Denmark
J. B. Winsløws Vej, 5000 Odense C, Denmark
E-mail: uhalekoh@health.sdu.dk

Søren Højsgaard
Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7G, 9220 Aalborg Ø, Denmark
E-mail: sorenh@math.aau.dk
URL: http://people.math.aau.dk/~sorenh/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 59, Issue 9 *Submitted:* 2012-01-11
August 2014 *Accepted:* 2013-12-01