Statistical Papers (2024) 65:335–356
https://doi.org/10.1007/s00362-023-01396-2

REGULAR ARTICLE

A Monte Carlo permutation procedure for testing variance components using robust estimation methods

Yahia S. El-Horbaty[1] · **Eman M. Hanafy**[1]

Received: 7 June 2022 / Revised: 8 January 2023 / Accepted: 23 January 2023 /
Published online: 3 February 2023

Abstract
Permutation methods offer an acceptable and convenient tool for inferring zero variance components in linear mixed models using the likelihood ratio test. However, when data exhibit heavy-tailed distribution, heavy-skewed distribution or outliers, maximum likelihood estimation may not be the best choice in constructing useful test statistics. In this article, we propose the use of robust rank-based estimation as an alternative. The finite sample distribution of our test statistic is well approximated using suitable permutations of the cluster indices that are exchangeable when the null hypothesis is true. Empirical results, comparing the new test to existing tests, indicate that all tests maintain acceptable Type I error rates when data exhibit heavy-tailed or heavy-skewed distributions. However, only our new test remains robust against the presence of outlier in the response space. Besides, it is only the latter case where other tests could show a competing power to our test. Otherwise, the new test is superior with an outstanding power under the remaining settings.

Keywords Exchangeability · Robustness · Rank-based estimation · Permutation test · Outliers

1 Introduction

Statistical inference using linear mixed-effects (LME) models is usually encountered in many applications where the structure of the data exhibits a clustering nature (Fitzmaurice et al. 2007), accounts for blocking factors (Kloke et al. 2009), or is delivered from a two-stage sampling design (Pfeffermann 2013). Inferring the need for random effects or equivalently testing the nullity of variance components is an essential task in LME models. Assuming the familiar chi-square distribution of the likelihood ratio

✉ Yahia S. El-Horbaty
yahia_mohamed@commerce.helwan.edu.eg

[1] Helwan University Faculty of Commerce, Helwan, Egypt

test (LRT) statistic is usually criticized because the null value of the variance components lies on the boundary of the parameter space (Self and Liang 1987). The limiting distribution of the LRT statistic is derived in Self and Liang (1987) as a mixture of chi-square distributed random variables for models involving one variance component. For models containing multiple random effects, Stram and Lee (1994) concluded that the asymptotic distribution of the LRT statistic can be affected by the correlation between the random effects. Investigation of this variance boundary problem is considered in various studies involving Shapiro (1985, 1988) and Stoel et al. (2006). Using numerical simulations, Fitzmaurice et al. (2007) suggested that even with large number of clusters, the mixture chi-square distribution is a poor approximation. Crainiceanu and Ruppert (2004) used a simulation-based algorithm to generate the finite sample distribution using an eigen-decomposition of the LRT statistic. Recent tests for zero variance components tend to approximate the finite sample distribution of the LRT statistic using permutations methods (Arboretti et al. 2015). See for example Fitzmaurice et al. (2007) and Lee and Braun (2012). Permutation methods have been used also in the tests proposed in Drikvandi et al. (2013) and Du and Wang (2020).

In practice, the presence of outliers, heavy-tailed distributions, or heavy-skewed distributions is evident in various applications exhibiting hierarchical data structures. In such cases, the superiority of likelihood-based estimation is questionable and hence the use of the LRT. On another hand, to our knowledge, neither robust variance components test procedures nor a relevant empirical assessment of the robustness of the LRT has yet been considered in the literature under such distributional violations. Robust rank-based estimation of LME models offers an attractive alternative to maximum likelihood estimation (Hettmansperger and McKean 2010; Liu and McKean 2015). Original developments for regression models with identically and independently distributed (iid) errors were considered in Jureckova (1971) and Jaeckel (1972). Kloke et al. (2009) developed the theory for obtaining robust joint-rank (JR) estimators of the unknown parameters under LME models with one variance component. The development therein provides protection against outlying responses, heavy-tailed symmetric distributions, and heavy-skewed distributions of the error components. Of note, robust rank-based estimation has not been used in constructing test statistics for testing zero variance components. Bridging this gap provides a reasonable alternative to the LRT when it does not offer the best choice.

The objective of this article is to introduce a robust test that also does not suffer from the variance boundary problem. To achieve this task, we use the robust rank-based estimation method under LME models (Hettmansperger and McKean 2010; Kloke et al. 2009). The task is fulfilled by introducing a test statistic with a well-approximated finite sample distribution, i.e. controllable Type-I error rate, using a permutation method. In other words, we propose a permutation test where calculation of the test statistic is based on the robust rank-based parameter estimation theory. We shall base the calculation of our test statistic on the estimators of the fixed effects and the variance components as prescribed in Kloke et al. (2009). Under the null hypothesis of zero variance components, the cluster indices are simply random labels. Thus, any permutation of those indices is just equally likely, ensuring their exchangeability (Fitzmaurice et al. 2007). As such permutation of the indices is nothing but permuting the pairs (y, x) that include the response and the associated set of explanatory variables,

then it is also a permutation of the residual errors that are iid when the null hypothesis holds. Hence, the necessary condition of the exchangeability of the residual errors is satisfied. An approximate finite sample distribution of the proposed test statistic is then obtained using the permutation distribution (Pesarin and Salmaso 2010) generated in conjunction with the robust estimation of the parameters of the LME model.

We shed light on situations where rank-based estimation is more efficient (produces smaller standard errors) than maximum likelihood estimation (Kloke et al. 2009; McKean and Kloke 2014; McKean and Hettmansperger 2016). In such situations, our empirical results show that robust rank-based estimation empowers the use of permutation tests for testing zero variance components. We emphasize that our development applies under LME models involving a single variance component. We rely on simulation experiments via which we highlight the superiority of the proposed test under all chosen schemes for comparisons. Simulation schemes are chosen such that we violate many of the standard assumptions upon which maximum likelihood estimation is known to lose efficiency. Using the proper score function for calculating the robust rank-based estimates, the proposed permutation test can be as doubly powerful (or even more) as the remaining tests.

The rest of this paper is organized as follows. Section 2 introduces the LME model. The proposed test statistic is considered in Sect. 3. In Sect. 4, the results of the simulation study are presented and a summary of the performance of the proposed test is provided. An application to a real dataset is given in Sect. 5. Conclusions of this study are summarized in Sect. 6.

2 Linear mixed-effects model

Consider a data set of m clusters, with n_k observations in the kth cluster, $k = 1, \ldots, m$. Let $\boldsymbol{Y}_k$ and $\boldsymbol{X}_k$, denote, respectively, the $n_k \times 1$ vector of responses and the $n_k \times p$ design matrix. Let b_k denotes the kth random cluster effect, and $\boldsymbol{\epsilon}_k$ the $n_k \times 1$ vector of errors. The model for $\boldsymbol{Y}_k$ is

$$\boldsymbol{Y}_k = \boldsymbol{X}_k\boldsymbol{\beta} + b_k 1_{n_k} + \boldsymbol{\epsilon}_k, \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients that usually contains an intercept term. Alternatively, the model can be written in a compact form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_m')'$, $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \ldots, \boldsymbol{\epsilon}_m')'$, $\boldsymbol{b} = (b_1, \ldots, b_m)'$, and $\boldsymbol{Z} = diag(1_1, \ldots, 1_m)$ such that 1_k denotes an $n_k \times 1$ vector of ones. Further, denote by $N = \sum_{k=1}^{m} n_k$ the total sample size and let $E(\boldsymbol{\epsilon}) = 0$, $var(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 \boldsymbol{I}$, $E(\boldsymbol{b}) = 0$, $var(\boldsymbol{b}) = \sigma_b^2 \boldsymbol{I}$, and $cov(\boldsymbol{\epsilon}, \boldsymbol{b}) = 0$. Independence is assumed among the random effects in $\boldsymbol{b}$, among the residual errors in $\boldsymbol{\epsilon}$, and between $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$.

The objective of this article is to test whether the random effects are needed in model (1). Thus, the hypothesis of interest can be formulated as

$$H_0 : \sigma_b^2 = 0 \text{ versus } H_1 : \sigma_b^2 > 0. \tag{2}$$

Let l_{H_0} and l_{H_1} denote, respectively, the log-likelihood functions maximised over H_0 and H_1. The LRT statistic is given by

$$LRT = -2\left[l_{H_0} - l_{H_1}\right] \tag{3}$$

Crainiceanu and Ruppert (2004) proposed a finite sample distribution of the LRT in (3) under null hypotheses and provided an algorithm for simulating that distribution. Fitzmaurice et al. (2007) proposed a permutation test for variance components using (3), which provides a one-sided p-value and has the correct empirical size regardless of the number of clusters or the cluster size. The latter test randomly permutes the cluster indices, holding the number of observations within each cluster as structured in the original dataset. The authors showed, using simulation studies, that this permutation test controls the Type-I error rate when the null hypothesis holds. We shall follow the same permutation method given therein.

As we focus on situations where the common assumptions underlying maximum likelihood estimation are severely violated, one immediately thinks of robust estimation methods. We mainly consider robust rank-based estimation. The statistical theory for rank-based estimation under (1) is developed in Kloke et al. (2009). We provide a brief overview of this method. The subsequent steps to generate the finite sample distribution of the proposed test statistic are given in Sect. 3.

For notational convenience, let η denote the intercept term to be excluded from $\boldsymbol{\beta}$ and rewrite model (1), following the notations in Kloke et al. (2009), such that

$$\boldsymbol{Y}_k = \eta 1_{n_k} + \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{e}_k \tag{4}$$

where

$$\boldsymbol{e}_k = b_k 1_{n_k} + \boldsymbol{\epsilon}_k. \tag{5}$$

Combining (4) and (5) for all clusters, then

$$\boldsymbol{Y} = \eta 1_N + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{6}$$

where $\boldsymbol{e} = (\boldsymbol{e}_1', \ldots, \boldsymbol{e}_m')'$. The following assumptions are needed. The random vectors in $\boldsymbol{e}$ are independent and the univariate marginal distribution of $\boldsymbol{e}_k$ is continuous and is the same for all k. Let $F_{\boldsymbol{e}}(.)$ and $f_{\boldsymbol{e}}(.)$ denote, respectively, this common distribution function and density function about $\boldsymbol{e}_k$. Further, assume that $f_{\boldsymbol{e}}(.)$ is absolutely continuous and that the usual regularity (likelihood) conditions hold. Assume further that Huber's condition holds for the design matrix $\boldsymbol{X}$ [i.e. the leverage values get uniformly small as N goes large (Kloke et al. 2009)]. Under a LME modelling framework, the ordinary rank-based estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\varphi} = Argmin\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\varphi}, \tag{7}$$

where $\|\boldsymbol{v}\|_{\varphi} = \sum_{t=1}^{N}\{a[R(v_t)]v_t\}$ for $\boldsymbol{v} \in \mathbb{R}^N$, $R(v_t)$ denotes the rank of v_t among $v_1, \ldots, v_N$ and the scores $a[.]$ are generated as $a[t] = \varphi[t/(N+1)]$ for $\varphi(u)$ a

nondecreasing bounded square-integrable function defined on the interval (0,1) such that $\sum_t a[t] = 0$, $\int_0^1 \varphi(u)du = 0$ and $\int_0^1 \varphi^2(u)du = 1$. The estimator in (7) satisfies the solution to $\boldsymbol{S}_X(\boldsymbol{\beta}) = 0$ where

$$\boldsymbol{S}_X(\boldsymbol{\beta}) = \boldsymbol{X}' a[R(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})]. \tag{8}$$

The estimator of the intercept term η, denoted by $\widehat{\eta}$, is given by the median over the residuals where

$$\widehat{\eta} = median_{kj}\left\{y_{kj} - \boldsymbol{x}'_{kj}\widehat{\boldsymbol{\beta}}_{\varphi}\right\}. \tag{9}$$

Consequently, the residuals are defined as

$$\widehat{\boldsymbol{e}}_{JR} = \boldsymbol{Y} - \left(1_N\widehat{\eta} + \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\varphi}\right). \tag{10}$$

The estimate of σ_b^2 using these residuals can be calculated as follows. Rewrite model (4) in element-wise form as

$$y_{kj} - \left(\eta + \boldsymbol{x}'_{kj}\boldsymbol{\beta}\right) = b_k + \epsilon_{kj} \tag{11}$$

for $j = 1, \ldots, n_k$. Since the residuals $\widehat{e}_{kj}$ in (10) provide estimates of the left side in (11), a predictor of b_k for a given cluster, say k, is the median over the n_k residuals in that cluster. That is, $\widehat{b}_k = median_{1\le j\le n_k}\{\widehat{e}_{kj}\}$. The robust estimator of σ_b^2 is given by

$$\widehat{\sigma}_b^2 = \left(1.483 median_{1\le k\le m}|\widehat{b}_k - median_{1\le r\le m}\{\widehat{b}_r\}|\right)^2$$

The last formula for $\widehat{\sigma}_b^2$ denotes the squared scaled median absolute deviations of $\widehat{b}_k$'s from their overall median. See Kloke et al. (2009) and Liu and McKean (2015) for thorough details and references on the derivation of $\widehat{\sigma}_b^2$ and the rationale behind it.

3 New test based on robust estimation

Permutation tests (Pesarin and Salmaso 2010, 2012; Hahn and Salmaso 2017) are non-parametric computationally intensive tests. In regression contexts, permutation tests possess the nominal size (Schmoyer 1994) when the sample data are correctly permuted such that the null distribution of the test statistic is approximated by repeatedly computing its values using each permuted sample. Specifically, those tests assume the exchangeability of the values being permuted (Basso et al. 2009) where exchangeability is less stringent than being iid.

We propose a robust permutation test for (2), utilizing the fact that permutation tests are distribution free. To investigate the robustness, we consider the error components in (1) to follow a symmetric distribution with heavy tails, a heavy skewed distribution, or to contain outliers. To fulfill this proposal, we replace the unknown variance component

σ_b^2 by its robust rank-based estimator $\widehat{\sigma}_b^2$ as described in Sect. 2, which can be calculated from the available data $(\boldsymbol{Y}, \boldsymbol{X})$. Letting $\boldsymbol{Z} = diag(1_{n_1}, \ldots, 1_{n_m})$, the proposed test statistic is given by

$$T_{JR} = trace(\widehat{\sigma}_b^2 \boldsymbol{Z}\boldsymbol{Z}') \tag{12}$$

where the test offers the calculation of a one-sided p-value in a way that yields the correct Type-I error rate under the null hypothesis. As the expression in (12) will be applied to random intercept models, T_{JR} is simply proportional to $\widehat{\sigma}_b^2$ since $T_{JR} = \widehat{\sigma}_b^2 \sum_{k=1}^{m} n_k$.

Construction of the permutation distribution of T_{JR} is needed to calculate the p-value. To do so, The marginal errors in (6) are permuted where, under the null hypothesis, the errors $\boldsymbol{e}$ are iid with zero mean and variance equal to σ_ϵ^2 and thus they are exchangeable. Note that the subtraction of the fixed effects term in (6) from $\boldsymbol{Y}$ resolves the problems of requiring the continuous covariates to be identical among the clusters and the necessity of having equal number of observations per cluster. Hence, the errors can be permuted within and between clusters. Since η and $\boldsymbol{\beta}$ need to be replaced by their estimates in practice, the estimated errors are calculated from the alternative model. It is shown by Schmoyer (1994) that, under the null hypothesis, the residuals are also asymptotically exchangeable both within and among clusters. Since $\widehat{\sigma}_b^2$ is a function of the residuals $\widehat{e}_{kj}$, as shown below (11), a straightforward permutation distribution for T_{JR} can be generated.

Since the number of permutations grows with $N = \sum_{k=1}^{m} n_k$, we use a general algorithm for obtaining a Monte Carlo estimate of the permutation p-value as follows:

(i) Under $H_0 : \sigma_b^2 = 0$, calculate T_{JR} from the original sample.
(ii) Randomly permute the cluster indices over all clusters, holding fixed the cluster sizes as n_k in the new permuted sample. Then, recalculate the test statistic, say $T_{JR}^{(r)}$ where the superscript r denotes that the rth permutation sample has been constructed.
(iii) Repeat the process a large number of times, say $\widetilde{R}$ times, producing $\widetilde{R}$ test statistics $T_{JR}^{(r)}, r = 1, \ldots, \widetilde{R}$.
(iv) The one-sided p-value, according to steps (i)–(iii), is calculated as the proportion of permutation samples (out of $\widetilde{R}$) such that $T_{JR}^{(r)}$ exceeds the original sample value of the test statistic.

In implementing of the Monte Carlo algorithm, the pooled set of pairs $\left\{\left(y_{kj}, \boldsymbol{x}_{kj}\right); k = 1, \ldots, m; j = 1, \ldots, n_k\right\}$ are exchangeable when the null hypothesis in (2) is true. The set of all residuals $\left\{\widehat{e}_{kj}; k = 1, \ldots, m; j = 1, \ldots, n_k\right\}$ are also exchangeable under the null hypothesis because both $\widehat{\eta}$ and $\widehat{\boldsymbol{\beta}}_\varphi$ are permutation invariant. Indeed, this invariance applies under any suitable regression estimation method when $\sigma_b^2 = 0$. When the distribution of the error components in the right-hand side of (5) is contaminated, our proposed test is thus based on the invariant values of $\widehat{\eta}$ and $\widehat{\boldsymbol{\beta}}_\varphi$ using robust rank-based estimation of $\widehat{\sigma}_b^2$. The generated permutation distribution is valid regardless of (i) the distributional assumptions that are made about the error components in model (1) except for the first two moments, (ii) the estimation

method that can be used to fit the model provided that the estimator is invariant to data permutations when the null hypothesis is true, and (iii) the cluster size, n_k, which may change from one cluster to another in unbalanced data. Beside T_{JR}, the above algorithm also applies to obtain the sampling distribution of $\widehat{\sigma}_b^2 = \left(\sum_{k=1}^{m} n_k\right)^{-1} T_{JR}$.

4 Simulation study

Simulation experiments are conducted to investigate the performance of the proposed test (T_{JR}-test hereafter). The empirical size and power are evaluated and compared to the permutation LRT (pLRT) (Fitzmaurice et al. 2007), the LRT and the restricted LRT (RLRT) (Crainiceanu and Ruppert 2004). The simulation setup covers various schemes such that focus is on the violations of the standard distributional assumptions about the error terms that are known to reduce the efficiency of the maximum likelihood estimators.

4.1 Simulation setup

Let the model for the response variable y_{kj} given the random effect b_k be given by

$$y_{kj} = \eta + b_k + \epsilon_{kj} \quad j = 1, \ldots, n_k, \ k = 1, \ldots, m \tag{13}$$

where we choose $m = 30, 40$ clusters, $n_k = 3, 10$ observations within a cluster and $\eta = 2$. Assume that the intra-cluster correlation (ICC) takes on the values 0.10, 0.20, and 0.30 where ICC $= \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2)$. For every test under consideration, the value of ICC $= 0$ is used to examine the empirical size (Type-I error) while the empirical power (ICC > 0). Both size and power are explored under the violation schemes given next. Assume that $b_k \sim N(0, \sigma_b^2)$ and that the residual error term ϵ_{kj} follows a symmetric contaminated normal distribution, a skewed contaminated normal distribution, a normal distribution while allowing for the presence of outliers, and a skewed distribution. The detailed setup under each scheme, involving the value of σ_ϵ^2, is given below.

4.1.1 Symmetric contaminated normal distribution

A symmetric contaminated normal distribution is a mixture of two normal distributions with mixing probabilities $(1 - \delta)$ and δ where $0 < \delta < 1$. For any random variable, say ϵ, that follows a normal distribution with density function $g(\epsilon; \mu, \sigma_\epsilon)$ where μ and σ denote, respectively, the mean and the standard deviation of the distribution, the contaminated normal density can be expressed as $f^*(\epsilon) = (1 - \delta)g(\epsilon; \mu, \sigma_\epsilon) + \delta g(\epsilon; \mu, \lambda\sigma_\epsilon)$ where $\lambda > 1$ is a parameter that determines the standard deviation of the wider component. In the simulations, we apply the definition of $f^*(.)$ to the residual errors ϵ_{kj} in (13). We consider $\delta = 20\%$ as a commonly used level of contamination in the distribution of ϵ_{kj} (Kloke et al. 2009), $\lambda = 5$, $\mu = 0$ and $\sigma_\epsilon^2 = 1$. Table 1 summarizes the simulation results of this scheme.

Table 1 Empirical rejection rates of tests when the residual errors are generated from symmetric contaminated normal distribution

m	n_k	ICC	Level of contamination (20%)			
			$T_{JR}(\%)$	pLRT (%)	LRT (%)	RLRT (%)
30	3	0.00	04.60	07.00	04.40	04.60
		0.1 0	07.20	04.00	05.20	05.60
		0.2 0	11.60	04.00	07.40	07.00
		0.3 0	14.80	05.00	09.80	09.60
	10	0.00	05.40	03.00	06.20	05.60
		0.1 0	14.60	04.00	08.80	13.40
		0.2 0	32.40	06.00	20.40	22.00
		0.3 0	52.80	08.00	36.00	36.80
40	3	0.00	05.80	03.50	06.20	06.00
		0.1 0	08.00	05.50	06.00	05.80
		0.2 0	10.00	05.50	07.60	07.40
		0.3 0	13.60	07.00	10.20	11.40
	10	0.00	05.60	05.50	04.60	03.00
		0.1 0	18.40	07.00	13.20	10.40
		0.2 0	36.20	07.50	25.20	20.80
		0.3 0	58.80	13.50	43.60	37.60

4.1.2 Skewed contaminated normal distribution

Here, we investigate the performance of the tests when ϵ_{kj} are generated from a skewed normal distribution which can be defined as

$$f(\epsilon) = 2\phi(\epsilon)\mathbf{\Phi}(\mathrm{s}\epsilon), \tag{14}$$

where $\phi(\epsilon)$ and $\mathbf{\Phi}(\mathrm{s}\epsilon)$ denote the standard normal density function and its distribution function that are defined at point $\mathrm{s}\epsilon$ respectively (Azzalini and Valle 1996). The component s represents the *shape/skewness parameter* because it regulates the shape of the density function. In the empirical study, ϵ_{kj} are generated from a skewed normal distribution that is contaminated, as defined in Sect. 4.1.1, with level of contamination being equal to $\delta = 20\%$, where $\lambda = 5$, $\mu = 0$, $\sigma_\epsilon^2 = 1$ and skewness parameter equal to 10 (McKean and Kloke 2014). The simulation results of this scheme are given in Table 2.

Table 2 Empirical rejection rates of tests when the residual errors are generated from skewed contaminated distribution

m	n_k	ICC	Level of contamination (20%)			
			T_{JR}(%)	pLRT (%)	LRT (%)	RLRT (%)
30	3	0.00	04.20	05.00	07.80	08.00
		0.10	13.60	08.50	06.60	05.60
		0.20	19.60	10.50	09.80	07.20
		0.30	27.60	14.50	15.40	11.80
	10	0.00	05.00	06.50	05.40	04.20
		0.10	37.00	16.00	14.20	13.80
		0.20	66.20	29.00	30.20	33.40
		0.30	80.20	53.00	51.60	57.00
40	3	0.00	04.70	05.00	04.60	04.60
		0.10	13.80	07.00	07.80	08.20
		0.20	25.00	08.50	11.00	11.00
		0.30	32.40	12.00	16.20	16.20
	10	0.00	05.80	06.50	03.20	05.00
		0.10	38.80	13.50	16.20	14.00
		0.20	73.80	30.50	35.20	33.80
		0.30	87.60	64.00	65.20	62.00

4.1.3 Outliers

Assuming that $\epsilon_{kj} \sim N(0, \sigma_\epsilon^2)$ where $\sigma_\epsilon^2 = 0.5$, under this scheme we replace 5% of the residual errors by residual errors drawn from $N(5, 15^2)$. We adopt this replacement for ϵ_{kj} while maintaining $b_k \sim N(0, \sigma_b^2)$. Maximum likelihood estimation is known to produce inefficient estimates under the presence of outliers of this form. Table 3 emphasizes the consequences of this fact by displaying the empirical Type-I error rates that are achieved by each of the competing tests. The corresponding empirical power results are also reported.

4.1.4 Skewed distribution

We also investigate the performance of the competing tests when ϵ_{kj} are generated from heavily skewed distributions such as the Cauchy distribution with location parameter zero and scale parameter 0.5 [i.e. C(0, 0.5)], the chi-square distribution with 1 degree of freedom and the log-normal distribution with parameters ($\mu = 0$, $\sigma = 1$). The results of Cauchy distribution are provided in Table 4 while those for the chi-square and log-normal distributions are provided in Table 5.

Table 3 Empirical rejection rates of tests when data involved outliers

m	n_k	ICC	Outliers scheme			
			$T_{JR}(\%)$	pLRT (%)	LRT (%)	RLRT (%)
30	3	0.00	06.40	01.60	00.60	00.40
		0.1 0	13.00	04.00	02.00	01.40
		0.2 0	15.30	11.50	07.20	06.80
		0.3 0	28.30	21.50	15.40	15.00
	10	0.00	05.30	00.30	00.40	00.20
		0.1 0	38.00	13.50	16.20	14.20
		0.2 0	68.70	50.50	58.80	57.40
		0.3 0	84.70	78.50	87.80	83.80
40	3	0.00	06.30	00.60	00.20	00.40
		0.1 0	12.00	08.50	04.60	03.80
		0.2 0	20.70	19.50	13.00	13.40
		0.3 0	31.70	31.00	32.60	31.20
	10	0.00	06.60	01.00	00.40	00.20
		0.1 0	47.30	01.50	26.20	20.40
		0.2 0	78.30	28.00	74.20	69.20
		0.3 0	91.30	70.50	93.80	94.60

4.2 Simulation results

Though not restricted to, the simulation outcomes obtained for the proposed test are based on defining $\varphi(u) = \sqrt{12}[u - (1/2)]$ where $\varphi(u)$ is mentioned below (7) which denotes the Wilcoxon score function (Hettmansperger and McKean 2010; Kloke et al. 2009). Applying JR estimation, presented in Sect. 2, to calculate $\widehat{\sigma}_b^2$ under the working model (13) is essential for computing T_{JR} as given in (12). Note that the vector of residuals $\widehat{\boldsymbol{e}}_{JR}$ is calculated under the working model as $\widehat{\boldsymbol{e}}_{JR} = \boldsymbol{Y} - 1_N\widehat{\eta}$, where $\widehat{\eta} = median_{kj}\{y_{kj}\}$. For the remaining tests, we use maximum likelihood estimation as recommended in their corresponding references. To evaluate the size or the power of each test, we generate 10,000 original samples. Besides, 10,000 permutation samples per each original sample are generated to test the null hypothesis and obtain the p-values using the T_{JR}-test and the pLRT. The empirical size is calculated as the proportion of times in which a given p-value is less than or equal the nominal level $\alpha = 5\%$.

Under the first contamination scheme, Table 1 summarizes the empirical sizes (ICC $= 0$) of the proposed T_{JR}-test, which are close to the nominal level $\alpha = 5\%$. The LRT is the next closest test to the nominal level followed by RLRT. The empirical power (ICC > 0) of the T_{JR}-test exceeds the power of the remaining tests where the poorest performance is provided by pLRT. We can see that when $m = 30, 40$ and $n_k = 3$, the power (as the ICC departs from zero) of the T_{JR}-test increases, though not with high

Table 4 Empirical rejection rates of tests when the residual errors are generated from Cauchy distribution

m	n_k	ICC	C(0, 0.5)			
			T_{JR}(%)	pLRT (%)	LRT (%)	RLRT (%)
30	3	0.00	06.50	10.00	02.00	02.50
		0.1 0	08.50	07.50	01.50	02.00
		0.2 0	14.50	08.50	01.50	02.00
		0.3 0	20.00	11.50	01.50	02.00
	10	0.00	05.40	07.00	03.00	01.50
		0.1 0	42.00	08.00	02.50	01.50
		0.2 0	62.00	10.00	04.00	02.00
		0.3 0	79.00	16.50	05.50	03.50
40	3	0.00	05.50	09.50	02.50	01.50
		0.1 0	11.00	10.50	02.00	01.50
		0.2 0	14.00	11.00	02.60	02.00
		0.3 0	18.50	13.50	02.80	02.90
	10	0.00	05.00	06.50	03.00	00.50
		0.1 0	48.00	07.50	01.50	02.00
		0.2 0	75.50	08.50	02.50	03.00
		0.3 0	90.00	09.50	04.50	03.90

jumps, at faster rate compared to the remaining three tests. However, as the cluster size increases ($n_k = 10$), both the rate of increase in the power of the T_{JR}-test and the gap from the other tests increase, confirming the superiority of the proposed test. It is obvious that the increase in the cluster size is the factor that most discriminates the performance of the competing tests where the best performance is always dedicated to the proposed T_{JR}-test.

Table 2 presents the results under the second scheme in where the residual errors have a skewed contaminated normal distribution. The size of each of the four competing tests remains not too distant from the nominal level. The T_{JR}-test, in particular, preserves an acceptable performance along with the chosen cluster sizes and number of clusters. The power of the T_{JR}-test remains the highest in all experiments. We also note that the power performance of the other three tests remains very close to each other as the value of the ICC increases. Unlike the comparisons made under the first scheme, the pLRT here possesses a competitive power to the LRT and the RLRT. Maintaining all other factors fixed at their level under this scheme, we note that the imposed skewness on the distribution of the residual error widens the gap between the T_{JR}-test and the remaining tests if compared to the situation when residual errors follow a symmetric contaminated distribution (i.e. Table 1). This considerable discrimination holds for every power comparison (i.e. for every ICC > 0).

As mentioned in Sect. 4.1.3, the third scheme in our simulation experiments is concerned with the presence of outliers in the y-space and its implications on the

Table 5 Empirical rejection rates of tests when the residual errors are generated from chi-square and log-normal distributions

m	n_k	ICC	$(\chi^2_{(1)})$				Log-normal (0, 1)			
			$T_{JR}(\%)$	pLRT (%)	LRT (%)	RLRT (%)	$T_{JR}(\%)$	pLRT (%)	LRT (%)	RLRT (%)
30	3	0.00	04.50	05.50	05.00	06.50	04.50	04.50	07.00	04.50
		0.1 0	16.50	08.00	08.50	05.00	16.00	06.50	07.00	04.70
		0.2 0	25.00	14.00	13.50	10.00	22.00	10.00	09.50	05.00
		0.3 0	34.00	22.50	19.00	23.50	25.00	14.50	14.50	08.00
	10	0.00	04.50	005.50	03.50	07.00	06.50	1.50	01.00	02.50
		0.1 0	45.50	27.50	26.00	29.00	27.50	14.00	09.00	07.50
		0.2 0	76.00	60.00	56.50	60.50	56.50	30.00	22.00	21.50
		0.3 0	90.50	81.50	87.50	86.00	77.00	48.50	46.50	40.50
40	3	0.00	05.50	04.50	06.00	05.00	06.00	05.50	06.00	02.50
		0.1 0	21.00	09.00	08.50	12.00	12.00	7.50	05.50	04.00
		0.2 0	25.50	13.00	14.50	20.00	19.50	13.00	10.50	08.50
		0.3 0	42.00	26.00	30.00	32.50	30.00	19.00	15.50	15.50
	10	0.00	05.90	05.00	06.50	07.00	06.00	04.50	03.00	02.50
		0.1 0	59.00	32.50	30.00	29.50	33.50	13.50	16.00	08.00
		0.2 0	85.00	64.50	63.00	67.50	64.00	38.00	30.50	16.50
		0.3 0	95.00	93.00	88.50	91.00	84.00	60.00	63.00	43.50

performance of the competing tests. Table 3 provides the empirical sizes and powers of the four tests. We observe that the presence of outliers has a dramatic effect on Type-I error rates produced by the pLRT, LRT and RLRT (i.e. when ICC = 0). Obviously, the T_{JR}-test is the only robust test with reasonable rates that are close to the nominal level of 5%. The empirical sizes of the remaining three tests are far distant from this nominal level, indicating how poor and unreliable might the performance of these tests be when outliers are suspected in the available data.

Although the three tests (pLRT, LRT, and RLRT) do not possess correct error rates under null hypothesis when outliers are present, results on their rejection rates are reported when the alternative hypothesis in (2) holds. It is obvious that as any of the three factors (i.e. ICC level, the cluster size, and the number of clusters) increases, the corresponding rejection rates increase. Noticeably, when ICC = 0.30, the proportion of rejecting the nullity of the variance component using the LRT and the RLRT is either close to the power of the T_{JR}-test or even higher. Nevertheless, we recommend the use of the T_{JR}-test due to its robust performance in the presence of outliers.

The results of the fourth scheme are provided in Tables 4 and 5. Assuming the Cauchy distribution for ϵ_{kj}, we conclude from Table 4 that the T_{JR}-test proceeds to control Type-I error rates when ICC = 0. As in the previous scheme, the other three tests do not guarantee an acceptable rejection rates under the null hypothesis. The

T_{JR}-test proceeds to outperform the remaining tests in terms of its power under the alternative hypothesis. Indeed, the remaining tests fail to reject the null hypothesis due to the poor estimates produced using the maximum likelihood method under this scheme.

Further investigation under the fourth scheme is provided where ϵ_{kj} are generated from two heavily skewed distributions, namely the $\chi^2_{(1)}$ and $lognormal(0,1)$ distributions. In Table 5, the empirical sizes of the three competing tests remain unstable but generally improve over their corresponding performance in Table 4. Noticeably, their power improves as we depart from the null hypothesis. The proposed T_{JR}-test remains the champion in terms of power comparisons, as is the case in all previous settings.

To sum up, the simulation experiments that are conducted in this section show a strong evidence that favors the use of the proposed T_{JR}-test, based on size-power comparisons, to the other three tests. Our proposal remains robust when the other tests fail to do so, preserving a considerable power increase in all the schemes under consideration as we depart form the nullity of the ICC.

5 Rat pup data

In this section, the rat pup dataset (Pinheiro and Bates 2006) is used. The study considers the experimental compound effects on the birth weights of 322 pups for 30 mother rats. The data consists of 27 litters, which were randomly assigned to a specific level of treatment (high, low, control), and 322 rat pups were nested within these litters. The study had an unbalanced design such that the number of pups per litter is not the same. The smallest litter had a size of 2 pups while the largest litter had a size of 18 pups. In addition, the number of litters per treatment is not the same (i.e. 10 litters were assigned to the control treatment, 7 to the high dose treatment and 10 litters were assigned to the low dose treatment).

A summary of the weights-by-treatment and sex is provided in Table 6 and Fig. 1. We note that the experimental treatments (high and low) appear to have a negative effect on mean birth weight. The averages (also the medians) of the birth weights for the pups born in litters that received high and low treatments are lower than the those of the birth weights for rats born in litters that received the control dose. Besides, the

Table 6 Summary statistics for rat pup birth weights by treatment and sex

Treatment	Sex	Number of observation	Mean	Standard deviation	Minimum	Maximum
Control	Female	54	6.12	0.69	3.68	7.57
	Male	77	6.47	0.75	4.57	8.33
High	Female	32	5.85	0.60	4.48	7.68
	Male	33	5.92	0.69	5.01	7.70
Low	Female	65	5.84	0.45	4.75	7.73
	Male	61	6.03	0.38	5.25	7.13

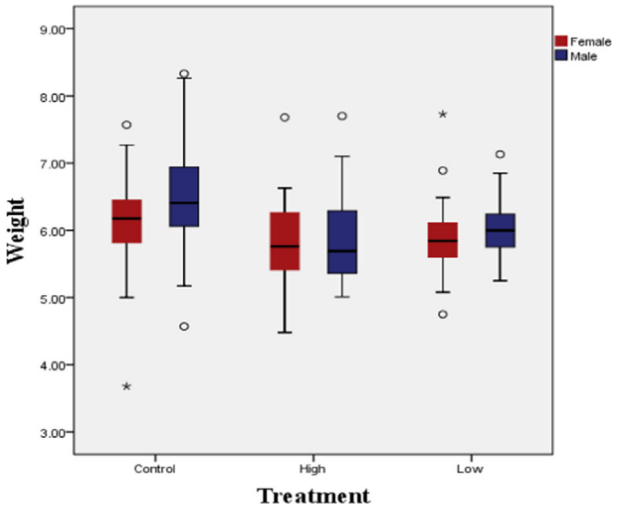

Fig. 1 Box plots for rat pup birth weights by treatment and sex

sample means of birth weights of male pups are higher than those of females within all levels of treatment.

Figure 2 describes the litter effect on the rat pup birth weights using 27 box plots such that, from left to right, the first 10 belong to control level followed by 7 box plots that belong to a high level and the last 10 belong to the low level of treatment. It is obvious that the means/medians of the 27 box plots are not same where the largest means/medians appear in litters 8, 17 and 27 and the smallest means/medians are in litters 1, 11, 12 and 18. Potential outliers are also recognized in both Figs. 1 and 2 since some pups appear to have either lower or higher weights than the other pups that belong to the same group (treatment/litter).

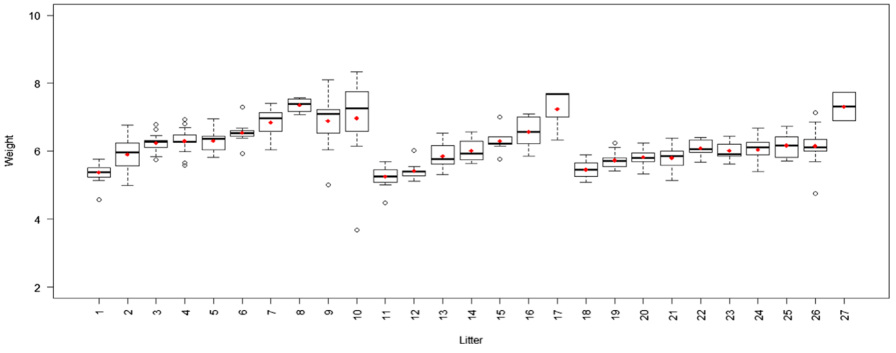

Fig. 2 Box plots for rat pup birth weights by litter

5.1 LME model for the rat pup data

Figure 2 indicates a potential varying litter effect on the distribution of the values of the rat pup birth weights in each litter. Considering this effect to be random, the individual birth weight observation ($WEIGHT_{kj}$) of the jth rat pup within the kth litter can be modeled using the following two-level random intercept regression model:

$$\begin{aligned} WEIGHT_{kj} =& \beta_0 + \beta_1 TREAT1_k + \beta_2 TREAT2_k + \beta_3 SEX_{kj} + \beta_4 LITSIZE_k \\ &+ \beta_5 TREAT1_k SEX_{kj} + \beta_6 TREAT2_k SEX_{kj} + b_k + \epsilon_{kj} \\ & j = 1, \ldots, n_k, k = 1, \ldots, 27 \end{aligned} \tag{15}$$

where n_k refers to the litter size that ranges between 2 and 18 pups per litter, $WEIGHT_{kj}$ is the response variable, $TREAT1_k$ and $TREAT2_k$ denote respectively level-2 indicator variables for receiving the high and low levels of treatment, SEX_{kj} is a level-1 indicator variable for female rat pup and, $LITSIZE_k$ refers to the size of litter k, where $k = 1, \ldots, 27$. The random litter effect, b_k, is assumed to have normal distribution with mean zero and constant variance σ^2_{litter} and the residual error term, ϵ_{kj}, is also assumed to have a normal distribution with mean zero and constant variance $\sigma^2_{residuals}$ (Pinheiro and Bates 2006).

5.2 Parameter estimation

Former analyses of this dataset focused on using the restricted maximum likelihood (REML) estimation method to infer about the effect of the different treatment levels on the birth weight (Pinheiro and Bates 2006). REML estimation also represents the basic method on which the competing tests were based, and is preferred to maximum likelihood estimation as it takes into account the loss in degrees of freedom due to estimation of fixed effect parameters (Patterson and Thompson 1971). Nevertheless, REML estimation does not figure out the potential effect of outliers and other violations of the distributional assumptions on the efficiency of the estimates and the consequent inference under the LME framework. In the remainder of this section, we highlight the gains from using the robust rank-based estimation method in terms of estimating both the fixed effects and the variance components with higher efficiency when compared to likelihood-based estimates.

The results of fitting model (15) are reported in Table 7 using the REML method versus the robust non-parametric JR method. The main effects (high vs. control) and (low vs. control) have a significant negative magnitude, indicating a negative effect on the birth weights of rat pups. The litter size is also found to have a significant negative effect on the birth weights of rat pup. The study shows a strong tendency for birth weights to decrease as a function of litter size in all litters.

Estimates of the variance components are also given in Table 7. We note that the JR estimate of σ^2_{litter} has smaller standard error compared to the corresponding REML estimator. The same conclusion holds for the estimated value of $\sigma^2_{residuals}$. Next, we examine the effect of the outliers and the distributional assumptions on each estimation method.

Table 7 REML and JR estimates and standard errors of effects for rat pup data

Estimation method	REML			JR		
	Estimate	SE	p value	Estimate	SE	p value
Fixed effects						
β_0(intercept)	8.32	0.27	0.00	8.39	0.07	0.00
β_1(high vs. control)	− 0.91	0.19	0.00	− 1.00	0.06	0.00
β_2(low vs. control)	− 0.47	0.16	0.01	− 0.46	0.04	0.00
β_3(female vs. male)	− 0.41	0.07	0.00	− 0.23	0.06	0.00
β_4(litter size)	− 0.13	0.02	0.00	− 0.14	0.01	0.00
β_5(high × female)	0.11	0.13	0.42	0.05	0.10	0.59
β_6(low × female)	0.08	0.11	0.43	− 0.03	0.09	0.75
Random effects						
σ^2_{litter}	0.0965	0.3107		0.0018	0.0431	
$\sigma^2_{residuals}$	0.1635	0.4043		0.0863	0.2938	

5.3 Robustness of estimation methods

Here, we explore whether two features might have led to the superiority of the JR estimators in Table 7 over the REML estimators. First, we test the assumption of normality of data using Shapiro–Wilk test. Based on the original data, the Shapiro–Wilk test produces a test statistic of 0.8448 with $p\text{-}value < 0.001$, which reveals a violation of the normality assumption. This result asserts the tendency of the JR method to outperform the REML method as concluded from Table 7 where the considerable departure from the normality assumption can be one of the reasons that favors the use of the JR fit.

The second feature of concern is the presence of potential outliers in the rat pup data as concluded from Fig. 2. In exploring the second feature, we follow the procedures in Kloke et al. (2009) to study the effect of changing the magnitude of the suspicious outliers on the efficiency of the REML and JR fits. The results are provided in Table 8. Moreover, we study the effect of removing these potential outliers, hence reducing the total sample size, by refitting the model to the reduced dataset. The corresponding results are provided in Table 10.

In order to assess the effect of the presence of the potential outliers in the rat pup data, we change their magnitudes in two dimensions as follow. For pups with weights larger than the majority of the other pups in the same litter, their magnitudes have been doubled. For those with weights less than the majority of the other pups in the same litter, their values have been divided by 2. From the results in Table 8, we note that according to each estimation method, the significance/insignificance status of fixed effects estimates remained unchanged. However, for the variances components, the REML standard errors became less efficient than their corresponding values using the

Table 8 REML and JR estimates and standard errors of effects for the changed rat pup data

Estimation method	REML			JR		
	Estimate	SE	p value	Estimate	SE	p value
Fixed effects						
β_0(intercept)	13.25	0.77	0.00	10.57	0.08	0.00
β_1(high vs. control)	− 2.37	0.53	0.00	− 1.39	0.06	0.00
β_2(low vs. control)	− 1.28	0.43	0.00	− 0.57	0.05	0.00
β_3(female vs. male)	− 1.06	0.26	0.00	− 0.21	0.06	0.00
β_4(litter size)	− 0.42	0.05	0.00	− 0.28	0.01	0.00
β_5(high × female)	0.50	0.47	0.29	0.17	0.10	0.10
β_6(low × female)	0.71	0.37	0.54	− 0.05	0.09	0.57
Random effects						
σ^2_{litter}	0.5906	0.7685		0.0035	0.0594	
$\sigma^2_{residuals}$	2.0695	1.4386		0.0879	0.2965	

Table 9 Summary of variance components parameter estimates

Method	Original data				Changed data			
	σ^2_{litter}	$\sigma^2_{residuals}$	σ^2	ICC	σ^2_{litter}	$\sigma^2_{residuals}$	σ^2	ICC
REML	0.0965	0.16349	0.2599	0.371	0.59059	2.0695	2.6601	0.222
JR	0.0018	0.0863	0.0881	0.020	0.0035	0.0879	0.0914	0.038

original data. The JR standard errors remain approximately unchanged, confirming that their robustness to the presence of the outliers.

Table 9 provides a summary of the estimates of variance components and interclass correlation coefficients under REML and JR estimation methods for the original and changed rat pup datasets. The results show that, the JR variance components estimates under the changed data are $\widehat{\sigma}^2_{litter} = 0.0035$, $\widehat{\sigma}^2_{residuals} = 0.0879$ and the estimate of the total model variance is $\widehat{\sigma}^2 = 0.0914$, where $\widehat{\sigma}^2 = \widehat{\sigma}^2_{litter} + \widehat{\sigma}^2_{residuals}$, and ICC = 0.038. These are essentially unchanged compared to their corresponding values in the original data and remain smaller than their corresponding results produced by REML estimation.

Model (15) has been refitted using the REML and JR methods to the reduced data, i.e. after removing the potential outliers from the original data. From Table 10, we conclude that the JR results remain better (in terms of the standard errors of the variance components) than their corresponding REML results. The conclusions made about the estimated fixed effects using both estimation methods do not change.

To sum up, it seems that the violation of the normality assumption was the main cause to advocate the use of the JR method in obtaining the results of the original data (Table 7) rather than the presence of potential outliers (Fig. 2). This conclusion

Table 10 REML and JR estimates and standard errors of effects after removing the potential outliers from the original rat pup data

Estimation method	REML			JR		
	Estimate	SE	p value	Estimate	SE	p value
Fixed effects						
β_0(intercept)	7.86	0.27	0.00	9.83	0.08	0.00
β_1(high vs. control)	− 0.80	0.18	0.00	− 1.17	0.06	0.00
β_2(low vs. control)	− 0.38	0.15	0.02	− 0.52	0.04	0.00
β_3(female vs. male)	− 0.28	0.06	0.00	− 0.23	0.06	0.00
β_4(litter size)	− 0.10	0.02	0.00	− 0.23	0.01	0.00
β_5(high × female)	0.01	0.11	0.90	0.23	0.10	0.22
β_6(low × female)	− 0.06	0.09	0.53	− 0.05	0.08	0.54
Random effects						
σ^2_{litter}	0.0899	0.2998		0.0033	0.0572	
$\sigma^2_{residuals}$	0.1077	0.3282		0.0879	0.2965	

has been enhanced by investigating the original data after changing the magnitude of these outliers (Table 8) and after their exclusion (Table 10).

5.4 Testing litter effect

Testing the need of random effect is conducted to decide whether the random effects that are associated with the intercepts for each litter can be omitted from model (15). Based on the original rat pup dataset, the proposed T_{JR}-test is calculated with 5000 permutation samples. The test produces a test statistic of 0.5796 with a $p\text{-}value = 0.001$. The competing tests are also conducted such that the test statistics pLRT, LRT, and RLRT are 84.213 ($p\text{-}value = 0.001$), 89.406 ($p\text{-}value = 0.0001$) and 84.461 ($p\text{-}value = 0.002$), respectively. Thus, we reject the null hypothesis at the 5% nominal level which allows the random effect b_k ($k = 1, \ldots, 27$) interpretation. This recommends retaining the random litter effects in this model. It should be emphasized that the role of the test is to decide about the need for the variance components in any further inferential procedures about the fixed effects under the potential presence of outliers or the absence of the normality. Retaining the variance components also validates the recommendation of using the JR estimation method. For further inferential procedures about the fixed effects under this method, the reader is then referred to Kloke et al. (2009).

6 Conclusion

In this article, our proposed variance components test is provided via a novel combination of tools that can play an important role in preserving a correct size meanwhile

producing a competitive power using a permutation test. The exchangeability of the cluster indices, hence of the estimated residuals, along with the robustness of the estimation of both fixed effects and variance of the random effects are jointly utilized. This combination seems to be overlooked or not recognized in the literature. Our test statistic seems to be a natural choice for evaluating the nullity of the variance components in the LME model using a permutation-based test. The robust estimation theory for obtaining the test statistic is readily available when the model involves a single variance component. Particularly, the robustness of the underlying parameter estimation method controls the size of the proposed test to remain at an acceptable level compared to the poor size (invalidity) of the competing tests under the presence of outliers. Aside from outliers, the power of the proposed T_{JR}-test always exceeds its competitors under the remaining simulation schemes.

Needless to say, the proposed test remains limited to LME models involving one random effect per cluster. The lack of robust rank-based estimation theory under general linear mixed models with complex/unknown covariance structures restricts our proposal from potential extensions to test multiple variance components. This includes the challenging problem of testing a subset of them. It shall be a demanding point for future research. Extensions should at least cover the cases where the present subset of random effects under the null hypothesis possess the nonstandard properties considered in our simulation schemes.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Appendix

Validity of the permutation test

The test statistic T_{JR} is computed based on $\widehat{\sigma}_b^2$ which is a function of the estimated residuals $\widehat{e}_{kj}$. Under the null hypothesis of zero variance components, the order by which the cluster indices are arranged in the sampled dataset is just one possible arrangement (permutation) $\pi \in \Sigma$ where Σ denotes the set of all $N!$ permutations of cluster indices. Denote by T_{JR}^{π} the value of T_{JR} under permutation π. For each permutation π, the parameters η and $\boldsymbol{\beta}$ are estimated and can be represented by $\widehat{\eta}^{\pi}$ and $\widehat{\boldsymbol{\beta}}_{\varphi}^{\pi}$. Interestingly, under the null hypothesis we have

$$\widehat{\boldsymbol{\beta}}_{\varphi}^{\pi} = \widehat{\boldsymbol{\beta}}_{\varphi} \tag{A1}$$

and

$$\widehat{\eta}^{\pi} = \widehat{\eta} \tag{A2}$$

for all $\pi \in \Sigma$ where $\widehat{\boldsymbol{\beta}}_{\varphi}$ and $\widehat{\eta}$ are given in (7) and (9).

The exchangeability requirement for running a permutation test based on T_{JR} can be proved based on the assumption of independence (hence the exchangeability) of the errors

$$e_{kj} = y_{kj} - \left(\eta + \boldsymbol{x}'_{kj}\boldsymbol{\beta}\right)$$

under the null hypothesis where (11) reduces to

$$y_{kj} - \left(\eta + \boldsymbol{x}'_{kj}\boldsymbol{\beta}\right) = \epsilon_{kj}.$$

Unfortunately, the errors e_{kj} are unobservable random variables. However, the estimates $\widehat{e}_{kj}$ can replace the corresponding errors in approximating the permutation distribution of T_{JR} if those estimates are exchangeable too. Observing that

$$\begin{aligned} \hat{e}^{\pi}_{kj} &= y^{\pi}_{kj} - \left(\hat{\eta}^{\pi} + \boldsymbol{x}'_{kj}\hat{\boldsymbol{\beta}}^{\pi}_{\varphi}\right) \\ &= y^{\pi}_{kj} - \left(\hat{\eta} + \boldsymbol{x}'_{kj}\hat{\boldsymbol{\beta}}_{\varphi}\right). \end{aligned} \tag{A3}$$

Then, the possible permutations of $\widehat{e}^{\pi}_{kj}$ are equiprobable because the cluster indices (i.e. over k and j) are equiprobable when the null hypothesis is true. Hence, it suffices to prove the exchangeability of $\widehat{e}_{kj}$ to validate the approximation of the permutation test using $\widehat{e}^{\pi}_{kj}$ for all $\pi \in \Sigma$. That is, their joint distribution is the same irrespective of their existing order.

For fixed j and $k = 1, \ldots, m$, the joint distribution of $\widehat{e}^{\pi}_{kj}$ over all clusters is given by

$$\begin{aligned} f\left(\hat{e}^{\pi}_{1j}, \ldots, \hat{e}^{\pi}_{mj}\right) &= \iint f\left(\hat{e}^{\pi}_{1j}, \ldots, \hat{e}^{\pi}_{mj} | \hat{\eta}^{\pi}, \hat{\boldsymbol{\beta}}^{\pi}_{\varphi}\right) dF\left(\hat{\eta}^{\pi}, \hat{\boldsymbol{\beta}}^{\pi}_{\varphi}\right) \\ &= \iint f\left(\hat{e}^{\pi}_{1j}, \ldots, \hat{e}^{\pi}_{mj} | \hat{\eta}^{\pi}, \hat{\boldsymbol{\beta}}^{\pi}_{\varphi}\right) dF\left(\hat{\eta}^{\pi} | \hat{\boldsymbol{\beta}}^{\pi}_{\varphi}\right) dF\left(\hat{\boldsymbol{\beta}}^{\pi}_{\varphi}\right) \\ &= \iint f\left(\hat{e}^{\pi}_{1j}, \ldots, \hat{e}^{\pi}_{mj} | \hat{\eta}, \hat{\boldsymbol{\beta}}_{\varphi}\right) dF\left(\hat{\eta} | \hat{\boldsymbol{\beta}}_{\varphi}\right) dF\left(\hat{\boldsymbol{\beta}}_{\varphi}\right) \\ &= \iint \prod_{k=1}^{m} f\left(\hat{e}^{\pi}_{kj} | \hat{\eta}, \hat{\boldsymbol{\beta}}_{\varphi}\right) dF\left(\hat{\eta} | \hat{\boldsymbol{\beta}}_{\varphi}\right) dF\left(\hat{\boldsymbol{\beta}}_{\varphi}\right), \end{aligned} \tag{A4}$$

where under the null hypothesis, the last equation (A4) indicates that the estimated residuals $\widehat{e}^{\pi}_{kj}$ are independent and identically distributed given $\widehat{\eta}$ and $\widehat{\boldsymbol{\beta}}_{\varphi}$. Let $\pi^* \in \Sigma$ be another permutation. Then,

$$f\left(\widehat{e}^{\pi}_{1j}, \ldots, \widehat{e}^{\pi}_{mj}\right) = \iint \prod_{k=1}^{m} f\left(\widehat{e}^{\pi}_{kj} | \widehat{\eta}^{\pi^*}, \widehat{\boldsymbol{\beta}}^{\pi^*}_{\varphi}\right) dF\left(\widehat{\eta}^{\pi^*}, \widehat{\boldsymbol{\beta}}^{\pi^*}_{\varphi}\right) dF\left(\widehat{\boldsymbol{\beta}}^{\pi^*}_{\varphi}\right)$$

and due to the conditional independence, then

$$= \iint \prod_{k=1}^{m} f\left(\widehat{e}_{kj}^{\pi^*} | \widehat{\eta}^{\pi^*}, \widehat{\boldsymbol{\beta}}_{\varphi}^{\pi^*}\right) dF\left(\widehat{\eta}^{\pi^*}, \widehat{\boldsymbol{\beta}}_{\varphi}^{\pi^*}\right) dF\left(\widehat{\boldsymbol{\beta}}_{\varphi}^{\pi^*}\right) \tag{A5}$$

where the last equation (A5) implies $f\left(\widehat{e}_{1j}^{\pi}, \ldots, \widehat{e}_{mj}^{\pi}\right) = f\left(\widehat{e}_{1j}^{\pi^*}, \ldots, \widehat{e}_{mj}^{\pi^*}\right)$ for any nonidentical permutations $\pi \neq \pi^*$.

References

Arboretti R, Corain L, Salmaso L, Melas VB, Pepelyshev A, Shpilev P (2015) On the optimal choice of the number of empirical Fourier coefficients for comparison of regression curves. Stat Pap 56(4):981–997

Azzalini A, Valle AD (1996) The multivariate skew-normal distribution. Biometrika 83(4):715–726

Basso D, Pesarin F, Salmaso L, Solari A (2009) Permutation tests for stochastic ordering and ANOVA: theory and applications in R. Springer, New York

Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. J R Stat Soc Ser B 66:165–185

Drikvandi R, Verbeke G, Khodadadi A et al (2013) Testing multiple variance components in linear mixed-effects models. Biostatistics 14:144–159

Du H, Wang L (2020) Testing variance components in linear mixed modeling using permutation. Multivar Behav Res 55(1):120–136

Fitzmaurice GM, Lipsitz SR, Ibrahim JG (2007) A note on permutation tests for variance components in multilevel generalized linear mixed models. Biometrics 63:942–946

Hahn S, Salmaso L (2017) A comparison of different synchronized permutation approaches to testing effects in two-level two-factor unbalanced ANOVA designs. Stat Pap 58(1):123–146

Hettmansperger TP, McKean JW (2010) Robust nonparametric statistical methods. CRC Press, Boca Raton

Jaeckel LA (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. Ann Math Stat 43:1449–1458

Jureckova J (1971) Nonparametric estimate of regression coefficients. Ann Math Stat 42:1328–1338

Kloke JD, McKean JW, Rashid MM (2009) Rank-based estimation and associated inferences for linear models with cluster correlated errors. J Am Stat Assoc 104:384–390

Lee OE, Braun TM (2012) Permutation tests for random effects in linear mixed models. Biometrics 68:486–493

Liu R, McKean JW (2015) Robust rank-based and nonparametric methods. Springer, Cham

McKean JW, Hettmansperger TP (2016) Rank-based analysis of linear models and beyond: a review. Robust Rank-Based and Nonparametric Methods: Michigan, USA, April 2015: Selected, Revised, and Extended Contributions, pp 1–24

McKean JW, Kloke JD (2014) Efficient and adaptive rank-based fits for linear models with skew-normal errors. J Stat Distrib Appl 1:1–18

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58(3):545–554

Pesarin F, Salmaso L (2010) Permutation tests for complex data: theory, applications and software. Wiley, Hoboken

Pesarin F, Salmaso L (2012) A review and some new results on permutation testing for multivariate problems. Stat Comput 22(2):639–646

Pfeffermann D (2013) New important developments in small area estimation. Stat Sci 28:40–68

Pinheiro J, Bates D (2006) Mixed-effects models in S and S-PLUS. Springer Science & Business Media, New York

Schmoyer RL (1994) Permutation tests for correlation in regression errors. J Am Stat Assoc 89:1507–1516

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610

Shapiro A (1985) Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. Biometrika 72(1):133–144

Shapiro A (1988) Towards a unified theory of inequality constrained testing in multivariate analysis. Int Stat Rev 56(1):49–62
Stoel RD, Garre FG, Dolan C et al (2006) On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. Psychol Methods 11(4):439–455
Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171–1177