

Permuting longitudinal data in spite of the dependencies

Sarah Friedrich, Edgar Brunner, Markus Pauly

Angaben zur Veröffentlichung / Publication details:

Friedrich, Sarah, Edgar Brunner, and Markus Pauly. 2017. "Permuting longitudinal data in spite of the dependencies." *Journal of Multivariate Analysis* 153: 255-65. <https://doi.org/10.1016/j.jmva.2016.10.004>.

Permuting longitudinal data in spite of the dependencies

Sarah Friedrich^{a,*}, Edgar Brunner^b, Markus Pauly^a

^a *Ulm University, Institute of Statistics, Germany*

^b *University Medical Center Göttingen, Institute of Medical Statistics, Germany*

1. Motivation and introduction

In many experiments in the life, social or psychological sciences the experimental units (e.g., subjects) are repeatedly observed at different occasions (e.g., at different time points) or under different treatment conditions. This leads to certain dependencies between observations from the same unit and results in a more complicated statistical analysis of such studies. In the context of experimental designs, the repeated measures are considered as levels of the *sub-plot* factor. If several groups are observed, these are considered as levels of the *whole-plot* factor.

Typical questions in repeated measures and profile analysis concern the investigation of a group effect, a non-constant effect of time or different time profiles in the groups; see, e.g., the monographs of Davis [14, Section 4.3] or Johnson and Wichern [25, Section 6.8]. Classical repeated measures models, where hypotheses are tested with Hotelling's T^2 [19] or Wilks's Λ [45], assume normally distributed observation vectors and a common covariance matrix for all groups; see e.g., the monograph of Davis [14]. In medical and biological research, however, the assumptions of equal covariance matrices and multivariate normally distributed outcomes are often not met and a violation of them may inflate the type-I error rates; see the comments in Xu and Cui [46], Suo et al. [40] or Konietschke et al. [28].

Therefore, other procedures have been developed for repeated measures which are based on certain approximation techniques [1,7–10,17,18,21,26,27,30,35,41,44]. However, these papers mainly assume the multivariate normal distribution and only discuss methods for specific models which are also asymptotically only approximations, i.e., they do not even lead to asymptotic exact tests. Another possibility is to apply a specific mixed model in the GEE context, see, e.g., the text books by Verbeke and Molenberghs [42,43]. These methods require that the data stem from a specific exponential family. An

* Corresponding author.

E-mail address: sarah.friedrich@uni-ulm.de (S. Friedrich).

Table 1
Means and empirical standard deviations of oxygen consumption of leukocytes in the presence and absence of inactivated staphylococci.

		O_2 -Consumption [$\mu\ell$]					
		Staphylococci					
		With			Without		
		Time (min)			Time (min)		
		6	12	18	6	12	18
Placebo ($n = 12$)	Mean	1.618	2.434	3.527	1.322	2.430	3.425
	Sd	0.157	0.303	0.285	0.193	0.263	0.339
Verum ($n = 12$)	Mean	1.656	2.799	4.029	1.394	2.57	3.677
	Sd	0.207	0.336	0.256	0.218	0.242	0.340

exception is given by the multivariate Wald-type test statistic (WTS), which is asymptotically exact. However, it is well known that it requires large sample sizes to keep the pre-assigned type-I error level; see, e.g., [6,28,34].

To improve the small-sample behavior of the WTS in a MANOVA setting, Konietzschke et al. [28] proposed different bootstrap techniques. Another possibility would be to apply permutation procedures. It is well known that permutation tests are finitely exact under the assumption of exchangeability; see, e.g., [5,31,36] or [37–39] as well as [2,3,12] for examples. In most of these examples, however, permutation tests are only applied in situations where the null distribution is invariant under the corresponding randomization group.

A modified permutation procedure may also be applied in situations where this invariance does not hold; see, e.g., [11,23,24,33,34]. The main idea in these papers is to apply a studentized test statistic and to use its permutation distribution (based on permuting the pooled sample) for calculating critical values. This leads to particularly good finite-sample properties even in case of general factorial designs with fixed factors [34]. It is the aim of the present paper to extend the concept of permuting all data to the context of longitudinal data in general (not necessarily normal and homoscedastic) split plot designs. Applied to the WTS this generalizes the results of Pauly et al. [34] and leads to astonishingly accurate results despite the dependencies in repeated measurements data.

The methodology derived in the present paper is motivated by the following data example on the O_2 consumption of leukocytes. To examine the breathability of leukocytes, an experiment with 44 HSD-rats was conducted. A group of 22 rats was treated with a placebo, while the other 22 rats were treated with a substance supposed to enhance the humoral immunity. 18 h prior to the opening of the abdominal cavity, all animals received 2.4 g sodium-caseinate for the production of a peritoneal exudate rich on leukocytes. In order to obtain a sufficient amount of material the peritoneal liquid of 3–4 animals was mixed and the leukocytes therein were rehashed in an experimental batch. One half of the experimental batch was mixed with inactivated staphylococci in a ratio of 100:1, the other half remained untreated and served as a control. Then, the oxygen consumption of the leukocytes was measured with a polarographic electrode after 6, 12 and 18 min, respectively. For each group separately, 12 experimental batches were carried out. Some descriptive statistics of the experimental batches in both treatment groups are listed in Table 1.

Questions of interest in this example concern the effect of the whole-plot factor ‘treatment’, the effect of the sub-plot factors ‘staphylococci’ and ‘time’ as well as interactions between these effects. We note that the empirical 6×6 covariance matrices of the two groups appear to be quite different (see the supplement (see Appendix A) for details). This also motivates the inclusion of unequal covariance matrices in our model. For such experimental designs, procedures are derived in this paper that lead to good small-sample control of the type-I error while being asymptotically exact.

The paper is organized as follows. The underlying statistical model is described in Section 2, where we also introduce the Wald-type (WTS) as well as the ANOVA-type statistic (ATS) and state their asymptotic behavior. In Section 3, we describe the novel permutation procedure used to improve the small sample behavior of the WTS. Afterwards, we present the results of extensive simulation studies in Section 4, analyzing the behavior of the permuted test statistic in different simulation designs with certain competitors. Additional simulation results have also been run for several other resampling schemes. They did not show a better performance than the permutation procedure and are only reported in the supplementary material, where also various power simulations can be found. The motivating data example is analyzed in detail in Section 5. The paper closes with a brief discussion of our results in Section 6. All proofs are given in the supplementary material (see Appendix A).

2. Statistical model, hypotheses and statistics

2.1. Statistical model and hypotheses

To establish the general model, let

$$\mathbf{Y}_{ik} = (Y_{ik1}, \dots, Y_{ikt_i})^\top, \quad i = 1, \dots, a; \quad k = 1, \dots, n_i \quad (2.1)$$

denote independent random vectors with distribution F_i and expectation $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{it_i})^\top = E(\mathbf{Y}_{i1})$ in treatment group i . The underlying dependency structure is regulated by pairwise correlations. In particular, we do not assume any special

structure of the group-specific covariance matrix $\mathbf{V}_i = \text{cov}(\mathbf{Y}_{i1}) > \mathbf{0}$ which may even differ between groups $i \in \{1, \dots, a\}$. Note that we also allow the number of time points t_i to differ between groups. The most common case where $t_i = t$ for all $i \in \{1, \dots, a\}$ is thus a special case of model (2.1). Here the time points $t_i \in \mathbb{N}$ are fixed. For convenience, we collect the observation vectors \mathbf{Y}_{ik} in

$$\mathbf{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_a^\top)^\top, \quad \mathbf{Y}_i = (\mathbf{Y}_{i1}^\top, \dots, \mathbf{Y}_{in_i}^\top)^\top. \quad (2.2)$$

In this set-up, hypotheses are formulated as $\mathcal{H}_0^\mu : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_a^\top)^\top$ denotes the vector of all expectations $\mu_{is} = E(Y_{is})$, $i \in \{1, \dots, a\}$, $s \in \{1, \dots, t_i\}$ and \mathbf{H} is a suitable contrast matrix, i.e., its rows sum up to zero. Examples of \mathbf{H} are presented in Section 4.

Throughout the paper, we will use the following notation. We denote by \mathbf{I}_t the t -dimensional unit matrix and by \mathbf{J}_t the $t \times t$ matrix of 1's, i.e., $\mathbf{J}_t = \mathbf{1}_t \mathbf{1}_t^\top$, where $\mathbf{1}_t = (1, \dots, 1)^\top$ is the t -dimensional column vector of 1's. Furthermore, let $\mathbf{P}_t = \mathbf{I}_t - 1/t \cdot \mathbf{J}_t$ denote the t -dimensional centering matrix. By \oplus and \otimes we denote the direct sum and the Kronecker product, respectively.

An estimator of $\boldsymbol{\mu}$ is given by $\bar{\mathbf{Y}}_\bullet = (\bar{\mathbf{Y}}_{1\bullet}^\top, \dots, \bar{\mathbf{Y}}_{a\bullet}^\top)^\top$, where, for each $i \in \{1, \dots, a\}$ and $s \in \{1, \dots, t_i\}$,

$$\bar{\mathbf{Y}}_{i\bullet} = (Y_{i,1}, \dots, Y_{i,t_i})^\top, \quad \bar{Y}_{i,s} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{iks},$$

and the covariance matrix \mathbf{V}_i in treatment group i is estimated by the sample covariance matrix

$$\hat{\mathbf{V}}_i = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (\mathbf{Y}_{ik} - \bar{\mathbf{Y}}_{i\bullet})(\mathbf{Y}_{ik} - \bar{\mathbf{Y}}_{i\bullet})^\top.$$

Let $N = n_1 + \dots + n_a$ denote the total number of subjects in the trial, $T = t_1 + \dots + t_a$ the total number of time points and $\tilde{N} = n_1 t_1 + \dots + n_a t_a$ the total number of observations. Then the asymptotic results are derived under the following two assumptions:

- (1) $n_i/N \rightarrow \kappa_i \in (0, 1)$ as $\min(n_1, \dots, n_a) \rightarrow \infty$,
- (2) $\sup_i E(\|\mathbf{Y}_{i1}\|^4) < \infty$.

2.2. Statistics and asymptotics

We consider two commonly used test statistics for repeated measures and multivariate data. First, the so-called ANOVA-type statistic (ATS), introduced in [6], is given as:

$$\tilde{Q}_N = N \bar{\mathbf{Y}}_\bullet^\top \mathbf{H}^\top (\mathbf{H} \mathbf{H}^\top)^- \mathbf{H} \bar{\mathbf{Y}}_\bullet = N \bar{\mathbf{Y}}_\bullet^\top \mathbf{T} \bar{\mathbf{Y}}_\bullet, \quad (2.3)$$

where $(\cdot)^-$ denotes some generalized inverse. Note that the test statistic does not depend on the special choice of the generalized inverse. Its asymptotic distribution is established in the next theorem.

Theorem 1. *Under the null hypothesis $\mathcal{H}_0^\mu : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, the ATS in (2.3) has, asymptotically, the same distribution as the random variable*

$$X = \sum_{i=1}^a \sum_{s=1}^{t_i} \lambda_{is} X_{is},$$

where $X_{is} \stackrel{i.i.d.}{\sim} \chi_1^2$ and the weights λ_{is} are the eigenvalues of $\mathbf{T}\boldsymbol{\Sigma}$ for $\boldsymbol{\Sigma} = \bigoplus_{i=1}^a \kappa_i^{-1} \mathbf{V}_i$. Moreover, for local alternatives $\mathbf{T}\boldsymbol{\mu} = 1/\sqrt{N} \cdot \mathbf{T}\mathbf{v}$, $\mathbf{v} \in \mathbb{R}^T$, the ATS has, asymptotically, the same distribution as $\mathbf{Z}^\top \mathbf{T} \mathbf{Z}$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{v}, \boldsymbol{\Sigma})$. If additionally $\boldsymbol{\Sigma} > \mathbf{0}$, the ATS has the same distribution as a weighted sum of $\chi_1^2(\delta)$ distributed random variables, where the weights are again the eigenvalues λ_{is} and $\delta = \mathbf{v}^\top \boldsymbol{\Sigma}^{-1} \mathbf{v}$.

Since the λ_{is} are unknown, the result cannot be applied directly. Nevertheless, Brunner [6] proposed to approximate the distribution of X by the distribution of a scaled χ^2 -distribution, i.e., by $g\tilde{X}_\nu$, where $\tilde{X}_\nu \sim \chi_\nu^2$. The constants g and ν are estimated from the data such that the first two moments of X and $g\tilde{X}_\nu$ coincide; see [4]. This leads to approximating the statistic

$$F_N = \frac{N}{\text{tr}(\mathbf{T}\hat{\boldsymbol{\Sigma}})} \bar{\mathbf{Y}}_\bullet^\top \mathbf{T} \bar{\mathbf{Y}}_\bullet. \quad (2.4)$$

by an $\mathcal{F}(\hat{\nu}, \infty)$ -distribution with estimated degree of freedom $\hat{\nu} = \text{tr}^2(\mathbf{T}\hat{\boldsymbol{\Sigma}})/\text{tr}(\mathbf{T}\hat{\boldsymbol{\Sigma}})^2$, where $\hat{\boldsymbol{\Sigma}} = N \bigoplus_{i=1}^a 1/n_i \hat{\mathbf{V}}_i$. The corresponding ATS test $\varphi_{\text{ATS}} = \mathbf{1}\{\tilde{Q}_N > \mathcal{F}_\alpha(\hat{\nu}, \infty)\}$, where $\mathcal{F}_\alpha(\hat{\nu}, \infty)$ denotes the $(1 - \alpha)$ -quantile of the $\mathcal{F}(\hat{\nu}, \infty)$ -distribution, leads to consistent test decisions for fixed alternatives. However, it is in general no asymptotic level α test

Table 2

Simulated type-I error rates (10 000 simulations) in a repeated measures design with $n = 10, 20, 50, 100$ individuals and $t = 4, 8$ repeated measures. The ATS is compared to the upper 5% quantile of the $\mathcal{F}(\hat{\nu}, \infty)$ -distribution, the WTS to the upper 5% quantile of the χ_{t-1}^2 -distribution.

n	Type-I error rates ($\alpha = 0.05$)			
	ATS: F -quantile		WTS: χ^2 -quantile	
	t = 4	t = 8	t = 4	t = 8
10	0.025	0.012	0.223	0.776
20	0.026	0.014	0.126	0.388
50	0.030	0.021	0.081	0.166
100	0.035	0.025	0.067	0.111

under the null hypothesis, which is a severe drawback of this procedure. Thus, we discuss a second statistic, the so-called Wald-type statistic (WTS) given as

$$Q_N = N\bar{\mathbf{Y}}_{\bullet}^{\top} \mathbf{H}^{\top} (\mathbf{H}\widehat{\Sigma}\mathbf{H}^{\top})^+ \mathbf{H}\bar{\mathbf{Y}}_{\bullet}. \quad (2.5)$$

Here $(\mathbf{H}\widehat{\Sigma}\mathbf{H}^{\top})^+$ denotes the Moore–Penrose inverse of $(\mathbf{H}\widehat{\Sigma}\mathbf{H}^{\top})$. In order to test the general linear hypotheses $\mathcal{H}_0^{\mu} : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$ critical values are taken from the asymptotic distribution of Q_N under the null hypothesis stated below.

Theorem 2. *Under the null hypothesis $\mathcal{H}_0^{\mu} : \mathbf{H}\boldsymbol{\mu} = \mathbf{0}$, the WTS in (2.5) has, asymptotically, a central χ_f^2 -distribution with $f = \text{rank}(\mathbf{H})$. The corresponding test is given by $\varphi_{WTS} = \mathbf{1}\{Q_N > \chi_{f,1-\alpha}^2\}$, where $\chi_{f,1-\alpha}^2$ denotes the $(1 - \alpha)$ -quantile of the χ_f^2 distribution. This test is an asymptotic level α test and is consistent for general fixed alternatives $\mathbf{H}\boldsymbol{\mu} \neq \mathbf{0}$. Moreover, for local alternatives $\mathbf{H}\boldsymbol{\mu} = 1/\sqrt{N}\mathbf{v}$, $\mathbf{v} \in \mathbb{R}^T$, Q_N has asymptotically a non-central $\chi_f^2(\tilde{\delta})$ distribution where $\tilde{\delta} = (\mathbf{H}\mathbf{v})^{\top} (\mathbf{H}\Sigma\mathbf{H}^{\top})^+ \mathbf{H}\mathbf{v}$. This implies that $E_{\mathcal{H}_1}(\varphi_{WTS}) \rightarrow \Pr(Z > \chi_{f,1-\alpha}^2)$ with $Z \sim \chi_f^2(\tilde{\delta})$.*

Although φ_{WTS} possesses these nice asymptotic properties, it is well-known that very large sample sizes n_i are necessary to maintain the pre-assigned level α using quantiles of the limiting χ^2 -distribution; see [6,28,34] as well as Table 2. This leads to a limited applicability of the WTS in practice.

To accept the need for a novel procedure, we investigate the accuracy of the two test statistics in a one-sample repeated measures design with n subjects and t repeated measures Y_{ks} . The null hypothesis $\mathcal{H}_0^{\mu} : \{\mu_1 = \dots = \mu_t\} = \{\mathbf{P}_t\boldsymbol{\mu} = \mathbf{0}\}$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_t)^{\top}$ is considered and the components of \mathbf{Y}_k are selected as standardized log-normally distributed random variables, i.e.,

$$Y_{ks} = \frac{ks - E(\epsilon_{ks})}{\sqrt{\text{var}(\epsilon_{ks})}}$$

for i.i.d. log-normally distributed ϵ_{ks} for all $k \in \{1, \dots, n\}$ and $s \in \{1, \dots, t\}$. The results are displayed in Table 2, where the simulated type-I error rates of the WTS and ATS are given. It is readily seen that the test based on the WTS considerably exceeds the nominal level of 5%, while the ATS leads to rather conservative decisions.

Thus, to enhance the small-sample properties of the above tests we have compared different resampling approaches in an extensive simulation study, presented in Section 9 of the supplementary material [15]. The resampling approaches considered there are a nonparametric and a parametric bootstrap approach (described in detail in the supplementary material) as well as a permutation procedure. Surprisingly, the best procedure in terms of type-I error control turned out to be a permutation technique that randomly permutes the pooled univariate observations without taking into account the existing dependencies for calculating critical values. Motivation for this seemingly counter-intuitive method stems from [29], where a similar approach has been applied in the paired two-sample case. Moreover, the current procedure generalizes the permutation test on independent observations by Pauly et al. [34] and implemented in the R package GFD [16] to the case of repeated measures and multivariate data. The details are explained in the next section.

3. The permutation procedure

Let $\mathbf{Y}^{\pi} = \pi(Y_{111}, \dots, Y_{anata})^{\top} = (Y_{111}^{\pi}, \dots, Y_{anata}^{\pi})^{\top}$ denote a fixed but arbitrary permutation of all \tilde{N} elements of \mathbf{Y} in (2.2), i.e., $\pi \in \mathcal{S}_{\tilde{N}}$. In this notation, Y_{iks}^{π} denotes the (i, k, s) -component of the permuted vector \mathbf{Y} . Furthermore, let $\bar{\mathbf{Y}}_{\bullet}^{\pi}$ denote the vector of the means under this permutation and $\widehat{\Sigma}^{\pi} = \bigoplus_{i=1}^a N/n_i \widehat{\mathbf{V}}_i^{\pi}$ the empirical covariance matrix of the permuted observations.

It is obvious, that \mathbf{Y} and \mathbf{Y}^{π} only have the same distribution whenever the components of \mathbf{Y} are exchangeable. However, this is not the case in general two- and higher way layouts, even in the case of independent observations; see, e.g., [20].

Following the approach of [11,22,23,32–34] in the case of independent observations, the idea is to studentize the statistic $\sqrt{N}\bar{\mathbf{Y}}^\pi$ and consider its projection into the hypothesis space, resulting in the WTS of the permuted observations, namely

$$Q_N^\pi = N(\bar{\mathbf{Y}}^\pi)^\top \mathbf{H}^\top (\mathbf{H}\hat{\Sigma}^\pi \mathbf{H}^\top)^{-1} \mathbf{H}\bar{\mathbf{Y}}^\pi. \quad (3.1)$$

In the sequel we will denote Q_N^π as the WTPS. Note that the question of how to permute is more involved here than in the case of independent univariate observations. A heuristic reason why the above approach might work is as follows: Unconditionally, all permuted components possess the same mean. Thus, when multiplied by a contrast matrix the permuted means vector always mimics the null situation, i.e., $\mathbf{H}\mathbf{E}(\bar{\mathbf{Y}}^\pi) = \mathbf{0}$ always holds. In particular, it can be shown that the conditional distribution of the WTPS Q_N^π in (3.1) always approximates the null distribution of Q_N in (2.5) in the general repeated measures design under study; thus leading to an asymptotically valid permutation test. This result is formulated in the following theorem.

Theorem 3. *The studentized permutation distribution of Q_N^π in (3.1) conditioned on the observed data \mathbf{Y} weakly converges to the central χ_f^2 distribution in probability, where $f = \text{rank}(\mathbf{H})$.*

Remark 3.1. **Theorem 3** states that the permutation distribution asymptotically provides a valid approximation of the null distribution of the test statistic Q_N in (2.5). To be concrete, this means that for any underlying parameters $\boldsymbol{\mu} \in \mathbb{R}^T$ and $\boldsymbol{\mu}_0 \in \mathcal{H}_0(\mathbf{H})$ with $\mathbf{H}\boldsymbol{\mu}_0 = \mathbf{0}$ we have convergence in probability, viz.

$$\sup_{x \in \mathbb{R}} \left| \Pr_{\boldsymbol{\mu}}(Q_N^\pi \leq x | \mathbf{Y}) - \Pr_{\boldsymbol{\mu}_0}(Q_N \leq x) \right| \rightarrow 0. \quad (3.2)$$

Here, $\Pr_{\boldsymbol{\mu}}(Q_N \leq x)$ and $\Pr_{\boldsymbol{\mu}}(Q_N^\pi \leq x | \mathbf{Y})$ denote the unconditional and conditional distribution function of Q_N and Q_N^π , respectively, under the assumption that $\boldsymbol{\mu}$ is the true underlying parameter.

Remark 3.2. A Wald-type permutation test is obtained by comparing the original test statistic Q_N with the $(1 - \alpha)$ -quantile $c_{1-\alpha}^*$ of the conditional distribution of the WTPS Q_N^π given the observed data \mathbf{Y} , i.e., $\varphi_{WTPS} = \mathbf{1}\{Q_N > c_{1-\alpha}^*\}$. More specifically, the numerical algorithm for computation of the p -value is as follows:

1. Given the data \mathbf{Y} , calculate the original Wald-type statistic Q_N for the null hypothesis of interest.
2. Randomly permute the pooled sample \mathbf{Y} (i.e., all univariate observations from each group and each subject) and save them in $\mathbf{Y}^{\pi,1}$.
3. Calculate the studentized Wald-type statistic Q_N^π from Eq. (3.1) with the randomly permuted pooled observations $\mathbf{Y}^{\pi,1}$. Save its value in A_1 .
4. Repeat steps 2 and 3 a large number J (e.g., $J = 1000$) times and obtain values A_1, \dots, A_J .
5. Compute the p -value by the (approximative) conditional permutation distribution (i.e., the empirical distribution of A_1, \dots, A_J) as

$$p\text{-value} = \frac{1}{J} \sum_{j=1}^J \mathbf{1}\{Q_N \geq A_j\}.$$

Theorem 3 implies that this test asymptotically keeps the pre-assigned level α under the null hypothesis and is consistent for any fixed alternative $\mathbf{H}\boldsymbol{\mu} \neq \mathbf{0}$, i.e., it has asymptotically power 1. Moreover, it has the same asymptotic power as the WTS for local alternatives $\mathbf{H}\boldsymbol{\mu} = 1/\sqrt{N} \cdot \mathbf{v}$, i.e., $E_{\mathcal{H}} \mathbf{1}(\varphi_{WTPS}) \rightarrow \Pr(Z > \chi_{f,1-\alpha}^2)$ with $Z \sim \chi_f^2(\delta)$ as in **Theorem 2**.

It follows that the permutation test and the classical Wald-type test are asymptotically equivalent and that both have the same local power under contiguous alternatives. In particular the asymptotic relative efficiency of the WTPS compared to the classical WTS is 1. Moreover, the permutation test based on Q_N^π is finitely exact if the pooled data \mathbf{Y} are exchangeable under the null hypothesis. In comparison, the ATS also leads to a consistent test for fixed alternatives but does not provide an asymptotic level α test since it is only an approximation.

We note that the proof given in the supplement (see **Appendix A**) to this paper indicates that the given permutation technique does not work in the case of the ATS. In particular, a permutation version of the ATS would also possess a weighted χ^2 -limit distribution but with different weights, say $\tilde{\lambda}_{is}$, due to an incorrect covariance structure.

Remark 3.3. Our general framework (2.1) allows for the treatment of different important factorial designs in the context of multivariate repeated measures data analysis. As in [34] the idea is to accordingly split the indices in subindices and to choose an appropriate hypothesis matrix \mathbf{H} . Examples of different cross-classified and hierarchically nested designs are discussed in Section 4 of [28]. For repeated measures, examples are given in Sections 4 and 5 as well as in [6].

4. Simulations

In order to investigate the small sample behavior of the WTPS, we present extensive simulation results for different designs and covariance structures. The procedure is analyzed in different settings with regard to maintaining the pre-assigned type-I error rate ($\alpha = 5\%$). The results for the WTPS are compared to the asymptotic quantiles of the ATS (\mathcal{F} -quantile) and the WTS (χ^2 -quantile).

4.1. Data generation

For our simulation studies, we simulated a split plot design which, in the context of longitudinal data, is a design with a groups, n_i subjects in group i and $t_i = t$ repeated measures Y_{iks} for all $s \in \{1, \dots, t\}$. Let

$$\mathbf{Y}_{ik} = (Y_{ik1}, \dots, Y_{ikt})^\top = \boldsymbol{\mu}_i + B_{ik}\mathbf{1}_t + \mathbf{V}_i^{1/2}\boldsymbol{\epsilon}_{ik},$$

with $\boldsymbol{\mu}_i = \mathbf{E}(\mathbf{Y}_{i1})$ for all $i \in \{1, \dots, a\}$ and let $B_{ik} \sim \mathcal{N}(0, \sigma_i^2)$ denote independent additive subject effects. The i.i.d. random vectors $\boldsymbol{\epsilon}_{ik} = (\epsilon_{ik1}, \dots, \epsilon_{ikt})$ were generated from different standardized distributions by

$$\epsilon_{iks} = \frac{\tilde{\epsilon}_{iks} - \mathbf{E}(\tilde{\epsilon}_{iks})}{\sqrt{\text{var}(\tilde{\epsilon}_{iks})}},$$

where $\tilde{\epsilon}_{iks}$ denote i.i.d. normal, exponential or log-normal random variables.

A simulation setting with $a = 3$ groups and $t = 4, 8$ repeated measures was considered. The null hypotheses investigated are

(1) The hypothesis of *no time effect T*

$$\mathcal{H}_0^\mu(T) : \bar{\mu}_{\cdot 1} = \dots = \bar{\mu}_{\cdot t} \quad \text{or equivalently } \mathbf{H}_T \boldsymbol{\mu} = \mathbf{0}.$$

(2) The hypothesis of *no group \times time interaction effect GT*

$$\mathcal{H}_0^\mu(GT) : \mathbf{H}_{GT} \boldsymbol{\mu} = \begin{pmatrix} \mu_{11} - \bar{\mu}_{\cdot 1} - \bar{\mu}_{\cdot 1} + \bar{\mu}_{\cdot \cdot} \\ \vdots \\ \mu_{at} - \bar{\mu}_{\cdot a} - \bar{\mu}_{\cdot t} + \bar{\mu}_{\cdot \cdot} \end{pmatrix} = \mathbf{0},$$

where $\mathbf{H}_T = 1/a \mathbf{1}_a^\top \otimes \mathbf{P}_t$ and $\mathbf{H}_{GT} = \mathbf{P}_a \otimes \mathbf{P}_t$.

We considered balanced as well as unbalanced designs for the $\mathbf{n} = (n_1, n_2, n_3)$ subjects in group 1–3, respectively. The simulated numbers of subjects were $\mathbf{n}^{(1)} = (30, 20, 10)$, $\mathbf{n}^{(2)} = (10, 20, 30)$ and $\mathbf{n}^{(3)} = (15, 15, 15)$. Furthermore, we simulated three different covariance structures \mathbf{V}_i

Setting 1: $\mathbf{V}_i = \mathbf{I}_t$ for $i \in \{1, 2, 3\}$

Setting 2: $\mathbf{V}_i = \text{diag}(\sigma_1^2, \dots, \sigma_t^2)$ with $\sigma_s^2 = s$ for $t = 4$ and $\sigma_s^2 = \sqrt{s}$ for $t = 8$

Setting 3: $\mathbf{V}_i = \rho_i^{|\ell-j|}$, $(\rho_1, \rho_2, \rho_3) = (0.6, 0.5, 0.4)$ for $i \in \{1, 2, 3\}$.

In Setting 1 and 2 the covariance structures are the same for all groups, whereas in Setting 3 we have an autoregressive covariance structure with different parameters for the different groups. Moreover, we simulated block effects with different variances $\sigma_i^2 \in \{0, 1, 2\}$. However, since the results were almost identical, we here only report the case $\sigma_i^2 = 0$. All simulations were conducted with 10,000 simulation and 1000 permutation runs.

4.2. Type-I error rates

The resulting type-I error rates for the hypotheses of *no time effect T* and *no group \times time interaction GT* are displayed in [Tables 3](#) and [4](#), respectively.

It is obvious that the tests based on the WTS considerably exceed the nominal level for small sample sizes. This behavior becomes worse with an increasing number of repeated measurements and when testing the interaction hypothesis. In some cases, the WTS reaches an empirical type-I error rate of almost 50% when testing the *GT*-interaction. This means that its accuracy is no better than flipping a coin. The ATS, in contrast, keeps the pre-assigned level α pretty well for normally distributed observations, even for small sample sizes. With an increasing number of repeated measurements and/or non-normal data, however, the ATS leads to quite conservative decisions. Furthermore, the ATS leads to slightly conservative decisions when testing the interaction hypothesis, even with normally distributed data. The WTPS is reasonably close to the pre-assigned level α in most situations, even under non-normality and for testing the interaction hypothesis. Despite the dependencies in longitudinal data, the permutation procedure greatly improves the behavior of the WTS in small sample settings. However, when testing the interaction hypothesis for $t = 8$ repeated measurements the WTPS shows a more or less conservative behavior in Setting 3 combined with $\mathbf{n}^{(2)}$, and a slightly liberal behavior for Setting 3 with $\mathbf{n}^{(1)}$.

The simulations show a clear advantage of the permutation procedure as compared to the χ^2 -approximation of the Wald-type statistic. The WTPS controlled the 5% level in most situations, even under non-normality, i.e., in situations where the ATS may lead to quite conservative decisions.

4.3. Additional simulation results

We note that additional simulations for the type-I error can be found in the supplementary material (see [Appendix A](#)) to this paper. There we have compared the above methods with other resampling schemes such as the bootstrap procedures described in [\[28\]](#). Of all procedures analyzed in the simulations, the permutation procedure produced the best results.

Table 3
Results of the simulation studies for the hypothesis of no time effect.

T	Cov. setting	$t = 4$			$t = 8$		
		ATS	WTS	WTPS	ATS	WTS	WTPS
Normal distribution							
1	$\mathbf{n}^{(1)}$	0.046	0.085	0.050	0.040	0.177	0.050
	$\mathbf{n}^{(2)}$	0.046	0.086	0.048	0.040	0.177	0.052
	$\mathbf{n}^{(3)}$	0.050	0.078	0.051	0.043	0.135	0.052
2	$\mathbf{n}^{(1)}$	0.051	0.085	0.050	0.042	0.177	0.051
	$\mathbf{n}^{(2)}$	0.052	0.086	0.051	0.043	0.177	0.052
	$\mathbf{n}^{(3)}$	0.053	0.077	0.051	0.041	0.135	0.052
3	$\mathbf{n}^{(1)}$	0.046	0.092	0.052	0.044	0.198	0.062
	$\mathbf{n}^{(2)}$	0.051	0.080	0.045	0.048	0.155	0.042
	$\mathbf{n}^{(3)}$	0.051	0.078	0.053	0.048	0.136	0.054
Log-normal distribution							
1	$\mathbf{n}^{(1)}$	0.032	0.094	0.051	0.021	0.198	0.047
	$\mathbf{n}^{(2)}$	0.031	0.090	0.052	0.020	0.198	0.046
	$\mathbf{n}^{(3)}$	0.031	0.089	0.051	0.021	0.186	0.048
2	$\mathbf{n}^{(1)}$	0.040	0.110	0.067	0.022	0.207	0.053
	$\mathbf{n}^{(2)}$	0.040	0.107	0.067	0.022	0.203	0.051
	$\mathbf{n}^{(3)}$	0.042	0.107	0.070	0.024	0.197	0.057
3	$\mathbf{n}^{(1)}$	0.033	0.101	0.057	0.024	0.221	0.064
	$\mathbf{n}^{(2)}$	0.037	0.090	0.053	0.033	0.190	0.048
	$\mathbf{n}^{(3)}$	0.036	0.092	0.057	0.031	0.191	0.062
Exponential distribution							
1	$\mathbf{n}^{(1)}$	0.045	0.090	0.048	0.034	0.194	0.051
	$\mathbf{n}^{(2)}$	0.046	0.096	0.053	0.032	0.191	0.048
	$\mathbf{n}^{(3)}$	0.046	0.086	0.054	0.034	0.151	0.050
2	$\mathbf{n}^{(1)}$	0.048	0.093	0.054	0.035	0.194	0.052
	$\mathbf{n}^{(2)}$	0.050	0.101	0.060	0.034	0.193	0.051
	$\mathbf{n}^{(3)}$	0.050	0.088	0.058	0.036	0.154	0.051
3	$\mathbf{n}^{(1)}$	0.049	0.098	0.055	0.042	0.218	0.066
	$\mathbf{n}^{(2)}$	0.050	0.090	0.049	0.046	0.173	0.045
	$\mathbf{n}^{(3)}$	0.050	0.087	0.055	0.042	0.153	0.056

4.3.1. Quality of the approximation

In the following, we denote by F_N the distribution function of Q_N under \mathcal{H}_0 , by F the distribution function of the limiting χ_f^2 -distribution under \mathcal{H}_0 and by F_N^π the distribution function of the WTPS under \mathcal{H}_0 . We can now define

$$KQS = \sup_{0.9 \leq t \leq 0.99} |F_N^{-1}(t) - F^{-1}(t)|$$

as well as

$$KQS^\pi = \sup_{0.9 \leq t \leq 0.99} |F_N^{-1}(t) - (F_N^\pi)^{-1}(t)|$$

in order to compare the distance between the quantile function F_N^{-1} and the limiting quantile function F^{-1} (KQS) with the distance between F_N^{-1} and $(F_N^\pi)^{-1}$, the quantile functions of the test statistic and its permuted version (KQS^π), respectively. We have calculated these distances for all simulation settings described above. Detailed results can be found in Section 10.1 of the supplementary material. It turned out that KQS^π is always smaller than KQS , i.e., the approximation provided by the permutation procedure is considerably better than the asymptotic χ^2 approximation for all simulation settings considered. In our simulations, KQS ranged from 1.991 to 48.11 with a median distance of 9.179, whereas KQS^π ranged from 0.1049 to 7.618 with a median value of 0.8948. Fig. 1 exemplarily shows the plots of the corresponding quantile functions for one of the simulation scenarios.

4.3.2. Large-sample behavior

In this section, we analyze the large sample behavior of the WTS and WTPS. We considered only normally distributed random variables with covariance structure Setting 2 for an unbalanced ($\mathbf{n}^{(1)} = (30, 20, 10)$) as well as a balanced ($\mathbf{n}^{(3)} = (15, 15, 15)$) design with $t = 4, 8$ time points. The sample size was increased by adding $b\mathbf{1}_3$ to the above sample size vectors for $b = \ell 20$ and all $\ell \in \{0, \dots, 10\}$. The results for the type-I error under the null hypothesis of no interaction and covariance setting 2 are presented in Fig. 2. The behavior of the WTS improves with growing sample size but even for 115

Table 4
Results of the simulation studies for the hypothesis of no group \times time interaction.

GT		$t = 4$			$t = 8$		
Cov. setting		ATS	WTS	WTPS	ATS	WTS	WTPS
Normal distribution							
1	$n^{(1)}$	0.049	0.135	0.046	0.033	0.432	0.051
	$n^{(2)}$	0.053	0.142	0.052	0.034	0.433	0.050
	$n^{(3)}$	0.048	0.126	0.049	0.039	0.366	0.051
2	$n^{(1)}$	0.053	0.132	0.050	0.038	0.429	0.052
	$n^{(2)}$	0.053	0.141	0.054	0.038	0.431	0.050
	$n^{(3)}$	0.050	0.122	0.052	0.040	0.366	0.050
3	$n^{(1)}$	0.054	0.141	0.050	0.040	0.465	0.065
	$n^{(2)}$	0.053	0.135	0.045	0.049	0.393	0.037
	$n^{(3)}$	0.051	0.126	0.049	0.045	0.363	0.053
Log-normal distribution							
1	$n^{(1)}$	0.024	0.121	0.047	0.012	0.426	0.053
	$n^{(2)}$	0.022	0.128	0.053	0.013	0.431	0.051
	$n^{(3)}$	0.024	0.118	0.048	0.012	0.406	0.051
2	$n^{(1)}$	0.025	0.129	0.051	0.014	0.427	0.054
	$n^{(2)}$	0.026	0.130	0.054	0.013	0.432	0.052
	$n^{(3)}$	0.023	0.120	0.050	0.013	0.403	0.052
3	$n^{(1)}$	0.029	0.133	0.050	0.020	0.457	0.062
	$n^{(2)}$	0.028	0.121	0.045	0.024	0.399	0.036
	$n^{(3)}$	0.028	0.122	0.049	0.020	0.408	0.053
Exponential distribution							
1	$n^{(1)}$	0.043	0.146	0.054	0.024	0.442	0.054
	$n^{(2)}$	0.041	0.148	0.054	0.024	0.443	0.050
	$n^{(3)}$	0.036	0.122	0.047	0.028	0.397	0.054
2	$n^{(1)}$	0.048	0.151	0.059	0.027	0.444	0.057
	$n^{(2)}$	0.042	0.153	0.059	0.025	0.448	0.052
	$n^{(3)}$	0.034	0.121	0.048	0.029	0.397	0.055
3	$n^{(1)}$	0.047	0.155	0.061	0.032	0.473	0.068
	$n^{(2)}$	0.043	0.140	0.049	0.042	0.406	0.037
	$n^{(3)}$	0.037	0.122	0.047	0.041	0.402	0.058

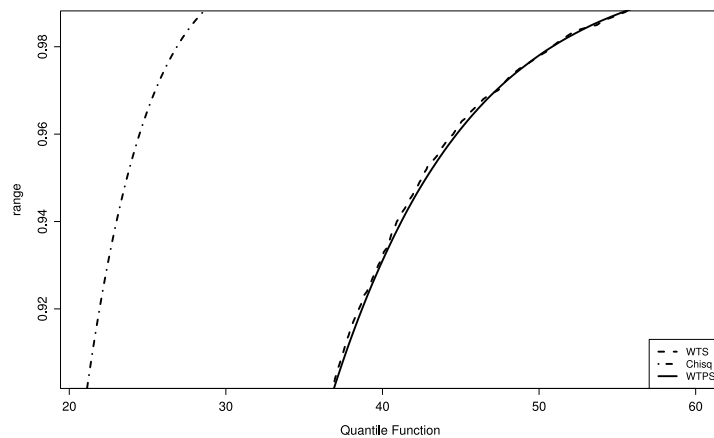


Fig. 1. Quantile functions of the WTS, WTPS and the corresponding χ^2 -distribution Function in the balanced simulation setting with log-normally distributed data, $t = 8$, covariance matrix setting 2 and under the null hypothesis of no interaction.

individuals in all groups, the WTS still exceeds the nominal level. The WTPS, in contrast, is rather close to the pre-assigned level even for small sample sizes.

4.3.3. Power

The power simulations are explained in detail in Section 11 of the supplementary material to this paper. Since the WTS turned out to test on different α -levels (see the simulation results under the null hypothesis), we have excluded it from the

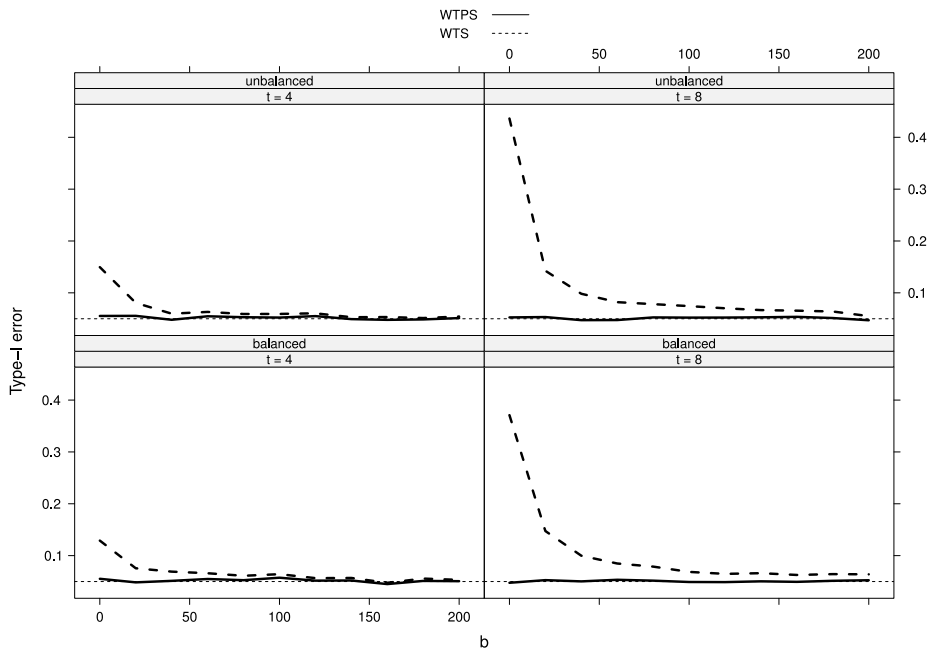


Fig. 2. Type-I error rates under the interaction hypothesis for the WTS and the WTPS, where sample size was increased by adding $b\mathbf{1}_3$, $b = \ell 20$ for all $\ell \in \{0, \dots, 10\}$ to the sample size vectors in a balanced (lower panel) and unbalanced (upper panel) design with $t = 4$ (left panel) and $t = 8$ (right panel) time points under covariance setting 2, i.e., $\mathbf{V}_i = \text{diag}(\sigma_1^2, \dots, \sigma_t^2)$ with $\sigma_s^2 = s$ for $t = 4$ and $\sigma_s^2 = \sqrt{s}$ for $t = 8$.

Table 5
Results of the analysis of the O_2 consumption data.

	ATS	WTS	WTPS
A	0.001	0.001	0.003
B	<0.001	<0.001	<0.001
T	<0.001	<0.001	<0.001
AB	0.110	0.110	0.133
AT	0.009	<0.001	<0.001
BT	0.094	0.115	0.151
ABT	0.117	0.116	0.164

analyses. We additionally considered the approximation described by Lecoutre [30] as well as Hotelling's T^2 [19]. It turns out that the ATS has the highest power for normally distributed data, performing slightly better than the WTPS. For log-normally distributed data, the WTPS has larger power than the other methods and it is the only method controlling the type-I error correctly.

5. Application: analysis of the data example

Finally, we analyze the data example on oxygen consumption of leukocytes in the presence and absence of inactivated staphylococci. In this setting we wish to analyze the effect of the whole-plot factor 'treatment' (factor A, Placebo/Verum, $a = 2$) as well as the sub-plot factors 'staphylococci' (factor B, with/without, $b = 2$) and 'time' (factor T, 6/12/18 min, $t = t_i = 3, i = 1, \dots, ab$). We are also interested in interactions between the different factors. The mean values and empirical standard deviations of the data are given in Table 1 in Section 1.

In the analysis we compared the three tests discussed above: The ATS in (2.4) is compared to the corresponding $\mathcal{F}(\hat{\nu}, \infty)$ -quantile, the WTS in (2.5) to the asymptotic χ_f^2 -quantile as well as the quantile obtained by the permutation procedure (WTPS). The seven different null hypotheses of interest about main and interaction effects can be tested by choosing the related hypotheses matrices. Here, we have chosen $\mathbf{H}_A = \mathbf{P}_a \otimes 1/b \cdot \mathbf{1}_b^T \otimes 1/t \cdot \mathbf{1}_t^T$, $\mathbf{H}_B = 1/a \mathbf{1}_a^T \otimes \mathbf{P}_b \otimes 1/t \mathbf{1}_t^T$ and $\mathbf{H}_T = 1/a \mathbf{1}_a^T \otimes 1/b \mathbf{1}_b^T \otimes \mathbf{P}_t$ for testing the main effect of the three factors A, B, and T. For the interaction terms we used the matrices $\mathbf{H}_{AT} = \mathbf{P}_a \otimes 1/b \mathbf{1}_b^T \otimes \mathbf{P}_t$, $\mathbf{H}_{AB} = \mathbf{P}_a \otimes \mathbf{P}_b \otimes 1/t \mathbf{1}_t^T$ and $\mathbf{H}_{BT} = 1/a \mathbf{1}_a^T \otimes \mathbf{P}_b \otimes \mathbf{P}_t$, and $\mathbf{H}_{ABT} = \mathbf{P}_a \otimes \mathbf{P}_b \otimes \mathbf{P}_t$. The resulting p -values of the analysis are presented in Table 5.

For this example all tests under considerations lead to similar conclusions: Each factor (treatment, staphylococci and time) has a significant influence on the O_2 consumption of the leukocytes. Moreover, there is a significant interaction between treatment and time.

6. Conclusions and discussion

In this paper, we have generalized the permutation idea of Pauly et al. [34] for independent univariate factorial designs to the case of repeated measures allowing for a factorial structure. Here, the suggested permutation test is asymptotically valid and does not require the assumptions of multivariate normality, equal covariance matrices or balanced designs. It is based on the well-known Wald-type statistic (WTS) which possesses the beneficial property of an asymptotic pivot while being applicable for general repeated measures designs. Since it is well known for being very liberal for small and moderate sample sizes, we have considerably improved its small-sample behavior under the null hypothesis by a studentized permutation technique. For univariate and independent observations the idea of this technique dates back to Neuhaus [32] and Janssen [22] and has recently been considered for more complex designs in independent observations by Chung and Romano [11] and Pauly et al. [34]. Extensions of the intriguing methods of Arboretti Giancristofaro et al. [2,3] and Corain et al. [12,13] to our quite general repeated measures design (not requiring any symmetry or homoscedasticity assumptions) would be desirable and will be part of future research.

In addition, we have rigorously proven in [Theorem 3](#) that the permutation distribution of the WTS always approximates the null distribution of the WTS and can thus be applied for calculating data-dependent critical values. In particular, the result implies that the corresponding Wald-type permutation test is asymptotically exact under the null hypothesis and consistent for fixed alternatives while providing the same local power as the WTS under contiguous alternatives.

Moreover, our simulation study indicated that the permutation procedure showed a very accurate performance in all designs under consideration with moderate repeated measures ($t = 4$) and homoscedastic or slightly heteroscedastic covariances. Only in the case of a larger number of repeated measurements ($t = 8$) the WTPS showed a more or less liberal (conservative) behavior when testing the interaction hypothesis in an unbalanced design. However, all other competing procedures considered in the paper and the supplementary material (see [Appendix A](#)) did not perform better in these situations.

Roughly speaking, the good performance of the WTPS for finite samples may be explained by a better approximation of the underlying distribution of the WTS by the permutation distribution as compared to the χ^2 -distribution. This could be seen clearly in the distances between the quantile functions.

Acknowledgments

The authors would like to thank Frank Konietschke for helpful discussions. This work was supported by the Deutsche Forschungsgemeinschaft (grant no. DFG-PA 2409/3-1).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/j.jmva.2016.10.004>.

References

- [1] M.R. Ahmad, C. Werner, E. Brunner, Analysis of high dimensional repeated measures designs: The one sample case, *Comput. Statist. Data Anal.* 53 (2008) 416–427.
- [2] R. Arboretti Giancristofaro, M. Marozzi, L. Salmaso, Repeated measures designs: A permutation approach for testing for active effects, *Far East J. Theor. Stat.* 16 (2005) 303–325.
- [3] D. Basso, M. Chiarandini, L. Salmaso, Synchronized permutation tests in $I \times J$ designs, *J. Statist. Plann. Inference* 137 (2007) 2564–2578.
- [4] G.E.P. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, *Ann. Math. Statist.* 25 (1954) 290–302.
- [5] C. Brombin, E. Midenia, L. Salmaso, Robust non-parametric tests for complex-repeated measures problems in ophthalmology, *Stat. Methods Med. Res.* 22 (2013) 643–660.
- [6] E. Brunner, Asymptotic and approximate analysis of repeated measures designs under heteroscedasticity, in: J. Kunert, G. Trenkler (Eds.), *Mathematical Statistics with Applications to Biometry*, Josef Eul Verlag, Köln, Germany, 2001.
- [7] E. Brunner, Repeated measures under non-sphericity, in: *Proceedings of the 6th St.Petersburg Workshop on Simulation*, 2009, pp. 605–609.
- [8] E. Brunner, A.C. Bathke, M. Placzek, Estimation of Box's ϵ for low- and high-dimensional repeated measures designs with unequal covariance matrices, *Biom. J.* 54 (2012) 301–316.
- [9] E. Brunner, B.M. Becker, C. Werner, Approximate Distributions of Quadratic Forms in High-Dimensional Repeated-Measures Designs. Technical Report, Dept. Medical Statistics, Georg-August-Universität Göttingen, Germany, 2009.
- [10] Y.-Y. Chi, M. Gribbin, Y. Lamers, J.F. Gregory, K.E. Muller, Global hypothesis testing for high-dimensional repeated measures outcomes, *Stat. Med.* 31 (2012) 724–742.
- [11] E. Chung, J.P. Romano, Exact and asymptotically robust permutation tests, *Ann. Statist.* 41 (2013) 484–507.
- [12] L. Corain, S. Ragazzi, L. Salmaso, A permutation approach to split-plot experiments, *Comm. Statist. Simulation Comput.* 42 (2013) 1391–1408.
- [13] L. Corain, L. Salmaso, Improving power of multivariate combination-based permutation tests, *Stat. Comput.* 25 (2015) 203–214.
- [14] C.S. Davis, *Statistical Methods for the Analysis of Repeated Measurements*, Springer, New York, 2002.
- [15] S. Friedrich, E. Brunner, M. Pauly, Supplement to Permuting Longitudinal Data Despite All The Dependencies, 2016.
- [16] S. Friedrich, F. Konietschke, M. Pauly, GFD: Tests for General Factorial Designs. R package version 0.2.2, 2015.
- [17] S. Geisser, S.W. Greenhouse, An extension of Box's result on the use of the F distribution in multivariate analysis, *Ann. Math. Statist.* 29 (1958) 885–891.
- [18] S.W. Greenhouse, S. Geisser, On methods in the analysis of profile data, *Psychometrika* 24 (1959) 95–112.
- [19] H. Hotelling, A generalization of student's ratio, *Ann. Math. Statist.* 2 (1931) 360–378.
- [20] Y. Huang, H. Xu, V. Calian, J.C. Hsu, To permute or not to permute, *Bioinformatics* 22 (2006) 2244–2248.
- [21] H. Huynh, L.S. Feldt, Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs, *J. Educ. Stat.* 1 (1976) 69–82.

- [22] A. Janssen, Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem, *Statist. Probab. Lett.* 36 (1997) 9–21.
- [23] A. Janssen, Resampling Student’s t-type statistics., *Ann. Inst. Statist. Math.* 57 (2005) 507–529.
- [24] A. Janssen, T. Pauls, How do bootstrap and permutation tests work? *Ann. Statist.* 31 (2003) 768–806.
- [25] R. Johnson, D. Wichern, *Applied Multivariate Statistical Analysis*, sixth ed., Prentice Hall, 2007.
- [26] M.G. Kenward, J.H. Roger, An improved approximation to the precision of fixed effects from restricted maximum likelihood, *Comput. Statist. Data Anal.* 53 (2009) 2583–2595.
- [27] H.J. Keselman, et al., Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping, *British J. Math. Statist. Psych.* 53 (2000) 175–191.
- [28] F. Konietzschke, A.C. Bathke, S.W. Harrar, M. Pauly, Parametric and nonparametric bootstrap methods for general MANOVA, *J. Multivariate Anal.* 140 (2015) 291–301.
- [29] F. Konietzschke, M. Pauly, Bootstrapping and permuting paired t-test type statistics, *Stat. Comput.* 24 (2014) 283–296.
- [30] B. Lecoutre, A correction for the \tilde{t} approximative test in repeated measures designs with two or more independent groups, *J. Educ. Stat.* 16 (1991) 371–372.
- [31] P.W. Mielke Jr., K.J. Berry, *Permutation Methods: A Distance Function Approach*, Springer, New York, 2007.
- [32] G. Neuhaus, Conditional rank tests for the two-sample problem under random censorship, *Ann. Statist.* 21 (1993) 1760–1779.
- [33] M. Omelka, M. Pauly, Testing equality of correlation coefficients in an potentially unbalanced two-sample problem via permutation methods, *J. Statist. Plann. Inference* 142 (2012) 1396–1406.
- [34] M. Pauly, E. Brunner, F. Konietzschke, Asymptotic permutation tests in general factorial designs, *J. R. Stat. Soc. - Ser. B* 77 (2015) 461–473.
- [35] M. Pauly, D. Ellenberger, E. Brunner, Analysis of high-dimensional one group repeated measures designs, *Statistics* 49 (2015) 1243–1261.
- [36] F. Pesarin, *Multivariate Permutation Tests. With Applications in Biostatistics*, Wiley, New York, 2001.
- [37] F. Pesarin, L. Salmaso, *Permutation Tests for Complex Data*, Wiley, New York, 2010.
- [38] F. Pesarin, L. Salmaso, Finite-sample consistency of combination-based permutation tests with application to repeated measures designs, *J. Nonparametr. Stat.* 22 (2010) 669–684.
- [39] F. Pesarin, L. Salmaso, A review and some new results on permutation testing for multivariate problems, *Stat. Comput.* 22 (2012) 639–646.
- [40] C. Suo, T. Touloupoulou, E. Bramon, M. Walshe, M. Picchioni, R. Murray, J. Ott, Analysis of multiple phenotypes in genome-wide genetic mapping studies, *BMC Bioinformatics* 14 (2013) 151.
- [41] G. Vallejo, M. Ato, Modified Brown–Forsythe procedure for testing interaction effects in split-plot designs, *Multivariate Behav. Res.* 41 (2006) 549–578.
- [42] G. Verbeke, G. Molenberghs, *Linear Mixed Models for Longitudinal Data*, Springer, New York., 2009.
- [43] G. Verbeke, G. Molenberghs, *Models for Discrete Longitudinal Data*, Springer, New York, 2012.
- [44] C. Werner, Dimensionsstabile Approximation für Verteilungen von quadratischen Formen im Repeated-Measures-Design. Technical Report, Dept. Medical Statistics, Georg-August-Universität Göttingen, Germany, 2004.
- [45] S.S. Wilks, Certain generalizations in the analysis of variance, *Biometrika* 24 (1932) 471–494.
- [46] J. Xu, X. Cui, Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays, *Bioinformatics* 24 (2008) 1056–1062.